PRIMJENA EXCELA U KOLEGIJU VJEROJATNOST I STATISTIKA

Rendulić, Nikolina

VELEUČILIŠTE U KARLOVCU

STROJARSKI ODJEL

SPECIJALISTIČKI DIPLOMSKI STRUČNI STUDIJ STROJARSTVA

Nikolina Rendulić

PRIMJENA EXCELA U KOLEGIJU VJEROJATNOST I STATISTIKA

ZAVRŠNI RAD

KARLOVAC, 2022.

KARLOVAC UNIVERSITY OF APPLIED SCIENCES

MECHANICAL ENGINEERING DEPARTMENT

PROFFESIONAL GRADUATE STUDY OF MECHANICAL ENGINEERING

Nikolina Rendulić

APPLICATION OF EXCEL IN THE COURSE PROBABILITY AND STATISTICS

FINAL PAPER

KARLOVAC, 2022.

VELEUČILIŠTE U KARLOVCU

STROJARSKI ODJEL

SPECIJALISTIČKI DIPLOMSKI STRUČNI STUDIJ STROJARSTVA

Nikolina Rendulić

PRIMJENA EXCELA U KOLEGIJU VJEROJATNOST I STATISTIKA

ZAVRŠNI RAD

Mentor:

Marin Maras, dipl.ing.math

KARLOVAC, 2022.

VELEUČILIŠTE U KARLOVCU Karlovac University of Applied Sciences	Klasa: 602-11/__-01/____
ZADATAK ZAVRŠNOG / DIPLOMSKOG RADA	Datum:

Ime i prezime	Nikolina Rendulić		
OIB / JMBG			
Adresa			
Tel. / Mob./e-mail			
Matični broj studenta	0123420020		
JMBAG			
Studij (staviti znak **X** ispred odgovarajućeg studija)	preddiplomski	**X** specijalistički diplomski	
Naziv studija	Specijalistički diplomski stručni studij strojarstva		
Godina upisa	2020.		
Datum podnošenja molbe			
Vlastoručni potpis studenta/studentice			

Naslov teme na hrvatskom: Primjena Excela u kolegiju Vjerojatnost i statistika	
Naslov teme na engleskom: Application of Excel in the course Probability and Statistics	
Opis zadatka: U radu se koristeći program Microsoft Excel analizira korištenje statističkih funkcija za lakšu analizu podataka iz dijela deskriptivne i inferencijalne statistike.	
Mentor/komentor: Marin Maras, viši predavač	Predsjednik Ispitnog povjerenstva: dr.sc. Tihana Kostadin, profesor visoke škole

IZJAVA

Izjavljujem da sam diplomski rad napisala samostalno, koristeći se znanjem stečenim tijekom obrazovanja na Veleučilištu u Karlovcu, određenom literaturom te uz stručnu pomoć mentora Marina Marasa, dipl.ing.math, kojemu iskreno zahvaljujem na pruženoj pomoći.

Karlovac, 2022. Nikolina Rendulić

SAŽETAK

Diplomski rad prikazuje način rješavanja zadataka u svrhu lakše analize podataka pomoću programa MS Excel koristeći statističke funkcije ili alat “Analiza podataka“. U prvom poglavlju upoznaju se opći statistički pojmovi te faze statističkog istraživanja potkrijepljeno riješenim primjerima u MS Excelu. Od deskriptivne statistike objašnjene su mjere centralne tendencije kao i mjere disperzije, a u poglavlju inferencijalne statistike težište je na linearnoj regresiji i testiranju hipoteza.

KLJUČNE RIJEČI

Statistika, deskriptivna statistika, inferencijalna statistika, MS Excel

SUMMARY

APPLICATION OF EXCEL IN THE COURSE PROBABILITY AND STATISTICS

This final paper presents a way to solve tasks using statistical functions or the tool “Data Analysis“ in MS Excel for easier data analysis. In the first chapter, general data on statistical concepts and phases of statistical research are given, supported by solved examples in MS Excel. From descriptive statistics, there are explained measures of central tendency and measures of dispersion, and in the chapter on inferential statistics, the focus is on linear regression and hypothesis testing.

KEY WORDS

Statistics, descriptive statistics, inferential statistics, MS Excel

SADRŽAJ

ZADATAK ZAVRŠNOG RADA I
IZJAVA II
SAŽETAK III
KLJUČNE RIJEČI III
SUMMARY IV
KEY WORDS IV
SADRŽAJ V
POPIS SLIKA VII
POPIS TABLICA IX
POPIS OZNAKA X
1. UVOD 1
2. OSNOVNI POJMOVI STATISTIKE 2
2.1. Podjela statistike 2
2.2. Osnovni statistički pojmovi 3
2.3. Statističko istraživanje 5
2.3.1. Grupiranje statističkih podataka 6
2.3.2. Tablični i grafički prikaz statističkih podataka 16
2.3.3. Linijski grafikoni 22
3. DESKRIPTIVNA STATISTIKA 24
3.1. Mjere centralne tendencije 24
3.1.1. Aritmetička sredina 24
3.1.2. Geometrijska sredina 28
3.1.3. Harmonijska sredina 28
3.1.4. Mod 29
3.1.5. Medijan 29
3.2. Mjere disperzije 31
3.2.1. Raspon varijacije 31
3.2.2. Interkvartil 32
3.2.3. Varijanca i standardna varijacija 32
3.2.4. Koeficijent varijacije 34
3.2.5. Koeficijent kvartilne devijacije 34
3.3. Procedura *Descriptive Statistics* 35

4. INFERENCIJALNA STATISTIKA 40
4.1. Testiranje hipoteza 40
4.1.1. Z-test 41
4.1.2. T-test 45
4.1.3. $\chi 2$- test 50
4.2. Korelacijska i regresijska analiza 52
4.2.1. Korelacijska analiza 53
4.2.2. Regresijska analiza 55
5. ZAKLJUČAK 66
LITERATURA 67
PRILOZI 68

POPIS SLIKA

Slika 1. Vrste obilježja 4
Slika 2. Unos statističkih podataka na radni list 8
Slika 3. Odabir funkcije Analiza podataka 9
Slika 4. Data Analysis 9
Slika 5. Dijaloški okvir Histogram 10
Slika 6. Ispunjavanje dijaloškog okvira Histogram 11
Slika 7. Dobivena tablica grupiranih podataka 11
Slika 8. Računanje kumulativnog niza "manje od" 12
Slika 9. Računanje kumulativnog niza "više od" 12
Slika 10. Dobiveni podaci kumulativnih nizova 13
Slika 11. Suma apsolutnih frekvencija 13
Slika 12. Računanje relativnih frekvencija 14
Slika 13. Dobivene relativne frekvencije 14
Slika 14. Konačni rezultati 14
Slika 15. Dijagram stablo-list 16
Slika 16. Umetanje grafikona 18
Slika 17. Odabir grafičkog prikaza 18
Slika 18. Strukturni krug 19
Slika 19. Umetanje stupčastog ili trakastog grafikona 19
Slika 20. Jednostavni stupci 20
Slika 21. Položeni stupci 20
Slika 22. Chart Output 21
Slika 23. Dobiveni podaci u grafičkom prikazu histogram 21
Slika 24. Grafički prikaz histogram 21
Slika 25. Podaci za kreiranje linijskog grafikona 22
Slika 26. Linijski grafikon 23
Slika 27. Average 25
Slika 28. Dobivena prosječna plaća 25
Slika 29. Ponderirana aritmetička sredina 27
Slika 30. Izračun medijana 30
Slika 31. Dobiveni rezultati medijana 31
Slika 32. Prikaz upisanih podataka 36

Slika 33. Odabir Descriptive Statistics 36
Slika 34. Dijaloški okvir Descriptive Statistics 37
Slika 35. Ispunjeni podaci 37
Slika 36. Dobiveni podaci 38
Slika 37. Podaci za z-test 42
Slika 39. Dobivena p-vrijednost 43
Slika 40. Područje prihvaćanja i odbacivanja nulte hipoteze za t-test 44
Slika 41. P-vrijednost za dvosmjerni z-test 45
Slika 42. Vrijednost F-testa 46
Slika 43. T-test 47
Slika 44. Izračunati podaci 47
Slika 45. Područje prihvaćanja i odbacivanja nulte hipoteze za t-test 48
Slika 46. Data Analysis - t-test 48
Slika 47. Unos podataka za t-test 49
Slika 48. Dobivene vrijednosti 49
Slika 49. Opažene i očekivane frekvencije 51
Slika 50. Hi kvadrat test 52
Slika 51. a) pozitivna funkcionalna veza, b) pozitivna statistička veza 54
Slika 52. a) negativna funkcionalna veza, b) negativna statistička veza 54
Slika 53. Raspršeni dijagram 58
Slika 54. Dijagram s jednadžbom linearnog regresijskog modela 59
Slika 55. Parametri jednadžbe regresijskog modela 60
Slika 56. Data Analysis – Regression 61
Slika 57. Dobivena rješenja regresijskog modela 61
Slika 58. Izračun =TDIST 64

POPIS TABLICA

Tablica 1. Razlika deskriptivne i inferencijalne statistike.....2
Tablica 2. Varijabilitet elemenata ovisno o postotku koeficijenta varijacije [1] 34
Tablica 3. Varijabilitet središnjih 50% elemenata ovisno o iznosu Vq [1].....35
Tablica 4. Pogreške statističkih testova [4].....40
Tablica 5. Razlika između jednosmjernog i dvosmjernog testa 42
Tablica 6. Rezultati ispitivanja 51
Tablica 7. Jakost veze između varijabli.....55
Tablica 8. Ovisnost trošenja mekog čelika o viskoznosti ulja.....58

POPIS OZNAKA

$\bar{X}$	aritmetička sredina
f	frekvencija
x_i	modaliteti
G	geometrijska sredina
H	harmonijska sredina
Mo	mod
Me	medijan
R	raspon varijacije
I_q	interkvartil
σ^2	varijanca
σ	standardna devijacija
V	koeficijent varijacije
V_q	koeficijent kvartilne devijacije
H_0	nulta hipoteza
H_1	alternativna hipoteza
α	razina značajnosti
β	vjerojatnost prihvaćanja lažne hipoteze
k	stupnjevi slobode
e	rezidualno odstupanje
r	koeficijent korelacije
$\bar{R}^2$	korigirani koeficijent determinacije

1. UVOD

Za statistiku se, jednostavno rečeno, može reći da se odnosi na učenje iz podataka. Njome se podaci prikupljaju, prikazuju i analiziraju da bi se na kraju donijela neka odluka, riješio problem i slično. Poznavanje statistike važno je u svim znanstvenim područjima koji uključuju rad s podacima, a ono što je također bitno je poznavati i program pomoću kojeg se lakše dolazi do analize podataka. Jedan od njih je i Microsoft Excel koji je vodeći softverski program za proračunske tablice na tržištu te napredan alat za vizualizaciju i analizu podataka. U nastavku rada će biti objašnjene njegove statističke funkcije kao i alat ''Analiza podataka".

2. OSNOVNI POJMOVI STATISTIKE

Statistika je znanstvena disciplina koja planira, prikuplja, selektira, grupira, prezentira i analizira informacije ili podatke te interpretira rezultate provedene analize, a u svrhu realizacije postavljenih istraživačkih ciljeva. [1]

Da bi ostvarila sve prethodno navedene zadatke i ciljeve, statistika ima razvijene vlastite metode i tehnike koje su danas primjenjive u raznim područjima poput ekonomije, zdravstva, kulture, politike, sporta, obrazovanja, znanosti itd.

2.1. Podjela statistike

Dva posebna dijela statistike su deskriptivna (opisna) i inferencijalna statistika.

Opisivanjem konkretnih rezultata dobivenih prilikom nekog istraživanja ili mjerenja pri čemu je obuhvaćen statistički skup u potpunosti, bavi se deskriptivna statistika. Njezina zadaća je da opiše dobivene rezultate, tj. sredi ih i sažme da budu što pregledniji i razumljiviji za daljnju analizu i primjenu. U svrhu opisivanja svojstava, deskriptivna statistika koristi mjere centralne tendencije u koje pripadaju aritmetička sredina, medijan, i mod te mjere varijabilnosti od kojih su najpoznatije raspon, standardna devijacija, varijanca, koeficijent varijacije.

Inferencijalna statistika se temelji na nepotpunom obuhvatu statističkog skupa ili populacije. [1] Rezultati se dobivaju ispitivanjem uzorka promatranih elemenata. Dio inferencijalne statistike su statistički modeli: analiza varijance, t-test, hi-kvadrat test, regresijska analiza, testiranje hipoteza.

U Tablica 1. navedene su neke osnovne razlike između deskriptivne i inferencijalne statistike.

Tablica 1. Razlika deskriptivne i inferencijalne statistike

DESKRIPTIVNA STATISTIKA	INFERENCIJALNA STATISTIKA
bavi se opisom populacije koja se proučava	usredotočena na donošenje zaključaka o populaciji na temelju analize uzoraka i promatranja
prikuplja, organizira, analizira i prezentira podatke na smislen način	uspoređuje podatke, hipotezu testiranja i predviđa buduće ishode
konačan rezultat prikazuje pomoću dijagrama, grafikona ili tablice	konačan rezultat prikazuje u obliku vjerojatnosti
opisuje situaciju	objašnjava vjerojatnost pojave događaja

2.2. Osnovni statistički pojmovi

Temelj statističke analize bilo kojeg procesa su statističke informacije, metode i tehnike. Pritom se dobiva uvid u strukturu pojava te njihove međusobne veze i odnose.

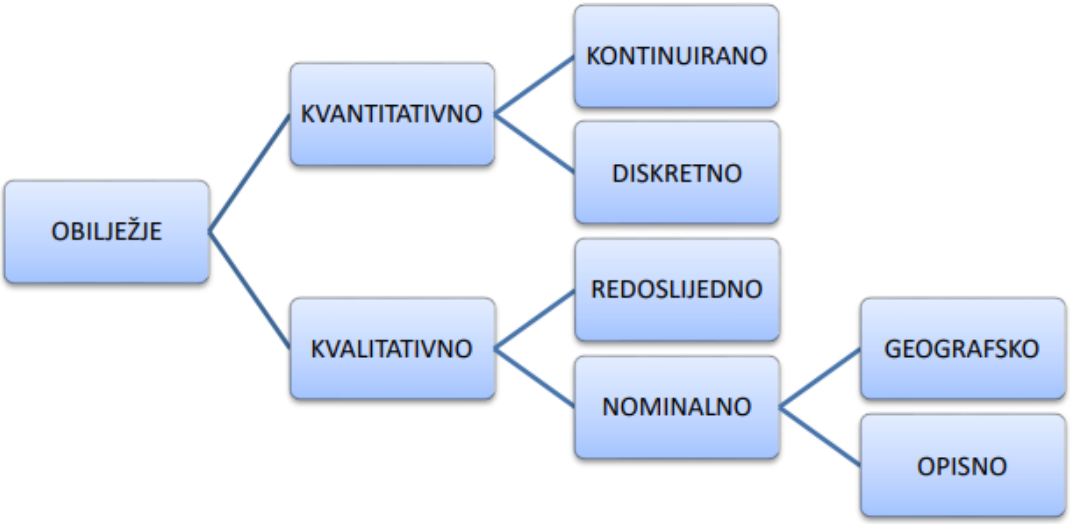
Osnovni pojam u statistici je statistički skup. On se sastoji od statističkih jedinica ili elemenata (osobe, mjesta, stvari, države, regije i sl.) koji imaju barem jedno zajedničko svojstvo tj. obilježje. Ukupan broj takvih elemenata čine opseg statističkog skupa prema kojemu se skupovi dijele na konačne i beskonačne. Konačni su oni koji se sastoje od konačnog broja jedinica, dok su u protivnom beskonačni.

Za svaki element statističkog skupa prikupljaju se podaci, a skup tih prikupljenih podataka naziva se osnovni skup ili populacija.

Po pojmom "uzorak" smatra se dio elemenata statističkog skupa zajedno s odgovarajućim podacima o tim elementima.

Prije samog prikupljanja potrebnih podataka, statistički se skup mora točno definirati i to pojmovno, prostorno i vremenski. [3] Na temelju takve definicije određuje se pripada li neka jedinica tom skupu ili ne. Na primjer, ako se promatra statistički skup kojeg tvore svi studenti 2. godine Specijalističkog studija strojarstva na Veleučilištu u Karlovcu na dan 22.02.2022., pojmovnu definiciju ovog skupa predstavljaju svi studenti 2. godine Specijalističkog studija strojarstva, Veleučilište u Karlovcu predstavlja prostorni dio definicije, dok je naznačeni datum vremenski dio definicije navedenog statističkog skupa.

Kao što je prethodno u poglavlju navedeno, odgovarajuća svojstva po kojima su jedinice statističkog skupa međusobno slične ili različite je njihovo obilježje. Svako se obilježje može pojaviti u više oblika, odnosno modaliteta, uz koje se vežu apsolutne ili relativne frekvencije. Postoji više različitih vrsta obilježja, a na Slika 1. prikazana je njihova podjela.



Slika 1. Vrste obilježja

Općenito se statistička obilježja mogu podijeliti na kvalitativna i kvantitativna. Ukratko, kvalitativna su opisana riječima, a kvantitativna brojkama. Kvantitativna statistička obilježja mogu biti diskretna i kontinuirana. Diskretna su ona koja imaju točno određene, cjelobrojne vrijednosti, npr. broj ljudi (ne može biti 5.8 ljudi, već cijeli broj poput 6,7,8…). Za razliku od njih, kontinuirana imaju beskonačno mnogo vrijednosti te su ona uglavnom prikazana u decimalnom obliku. Primjeri takvog obilježja su visina, težina, duljina i sl. Kvalitativna statistička obilježja dijele se na redoslijedna i nominalna. Redoslijedna se mogu mijenjati prema intenzitetu ili rangu, npr. ocjena na ispitu. U tom slučaju modaliteti su: nedovoljan, dovoljan, dobar, vrlo dobar, izvrstan. Primjer redoslijednog obilježja je i ocjena kvalitete proizvoda (npr. ispravan, neispravan), stručna sprema radnika (nekvalificiran, srednja stručna sprema, visoka stručna sprema). Nominalna mogu biti opisna, odnosno alternativna i atributivna. Opisna obilježja su ona koja opisuju, odnosno iskazuju svojstvo elementa statističkog skupa, npr. boja očiju, boja kose, tip automobila, horoskopski znak itd. Alternativna obilježja su ona koja mogu poprimiti samo dva modaliteta, npr. obilježje spol može imati oblike muški i ženski. Još jedna grupa nominalnih statističkih obilježja su geografska statistička obilježja koja označavaju prostor s kojim su elementi statističkog skupa u vezi. To je npr. mjesto rođenja, mjesto prebivališta.

Svako statističko obilježje vezano je uz određenu mjernu skalu. Mjerna skala daje pravila prema kojima se provode mjerenja svojstava jedinica statističkog skupa. Postoji nominalna, ordinalna, intervalna i omjerna skala. [3]

Za obilježja izražena riječima, odnosno kvalitativna statistička obilježja, koriste se nominalna i redoslijedna skala. Nominalna skala sadrži listu svojstava po kojima se jedinice statističkog skupa razlikuju, a ordinalnu skalu čine redoslijedna obilježja. Poredak modaliteta takvog redoslijednog obilježja je po intenzitetu od najjačeg prema najslabijem mjerenom svojstvu ili obrnuto. Modaliteti redoslijedne skale mogu se samo uspoređivati, točnije s njima je nemoguće provoditi aritmetičke operacije (npr. dvije ocjene dovoljan (2) nisu jednake kao jedna ocjena vrlo dobar (4)).

Kvantitativna statistička obilježja izražena brojčanim vrijednostima, mjere se na intervalnoj i omjernoj skali. Intervalnu skalu čine brojevi kojima se mjeri neko svojstvo na taj način da jednake razlike brojeva na toj skali predstavljaju jednake razlike mjerenog svojstva. [3] Za ovu skalu je karakteristično da je položaj nule unaprijed dogovoren, ali nula u tom slučaju ne znači da promatrana pojava ne postoji. Npr. za intervalnu skalu karakteristična je temperaturna skala te u tom slučaju nula (0°C) ne znači da temperature nema, već upućuje na to da je hladno. Vrijednosti intervalne skale mogu se zbrajati i oduzimati, ali ne i dijeliti. Tako npr. ako je u Karlovcu izmjerena temperatura od 5°C, a u Zadru 15°C, znači da je u Zadru temperatura zraka za 10°C viša nego u Karlovcu, no ne može se tvrditi da je u Zadru tri puta toplije nego u Karlovcu. Omjerna skala se također sastoji od brojeva čije jednake razlike predstavljaju jednake razlike mjerenog svojstva, ali ono što je različito od intervalne skale je to što nula nije utvrđena dogovorom i predstavlja nepostojanje pojave. Npr. broj djece u nekoj obitelji može biti nula što će značiti da djece nema. Varijabla koja se mjeri na omjernoj skali naziva se numeričkom varijablom i s takvim vrijednostima je moguće provoditi aritmetičke operacije. Npr. ako je numeričko obilježje "stanje na računu", za nekoga tko na računu ima 30 000 kn može se reći da ima tri puta veću uštedevinu od nekoga s 10 000 kn na računu ili da ima 20 000 kn više.

2.3. Statističko istraživanje

Svako statističko istraživanje obuhvaća sljedeće faze: [1]

1. Određivanje cilja i razradu plana istraživanja
2. Organizirano prikupljanje statističkih podataka
3. Sređivanje, odnosno grupiranje statističkih podataka
4. Tablično i grafičko prikazivanje statističkih podataka

5. Statističku analizu i interpretaciju rezultata provedene analize

Prva navedena faza koja se odnosi na određivanje cilja i razradu plana istraživanja je ujedno i najvažnija za cjelokupno statističko istraživanje. Prije nego se krene s prikupljanjem podataka, vrlo je važno precizno definirati što se istraživanjem želi postići jer u suprotnom može doći do uzaludnog prikupljanja podataka koje nije moguće podvrgnuti statističkoj analizi.

Što se tiče prikupljanja statističkih podataka, ono općenito uključuje:

1) brojanje
2) mjerenje
3) ocjenjivanje
4) evidentiranje
5) anketiranje ili intervjuiranje

Najpoznatija metoda je anketiranje. Anketa predstavlja skup različitih postupaka kojima se prikupljaju i analiziraju izjave anonimnih ljudi kako bi se dobili podaci o njihovom ponašanju, stavu prema određenom problemu, mišljenju, interesima itd. [2] Uvjet anonimnosti ispitanika je bitan zbog mogućnosti utjecaja na iskrenost prilikom odgovaranja na pitanja koja su dana anketnim upitnikom. Kao rezultat organiziranog prikupljanja podataka dobiju se negrupirani podaci iz kojih je vidljivo da je svakom pojedinom elementu statističkog skupa pridružen element osnovnog skupa.

2.3.1. Grupiranje statističkih podataka

Nakon prikupljanja i dobivanja negrupiranih podataka, iste je potrebno "preurediti" kako bi bili što pregledniji i lakši za daljnje statističko istraživanje. Za taj postupak koristi se grupiranje statističkih podataka čime se za svaki pojedini modalitet utvrđuje kolikom je broju elemenata statističkog skupa pridružen. Prilikom te podjele statističkog skupa na podskupove, potrebno je poštivati dva osnovna načela: isključivost (jedan te isti element statističkog skupa ne može istovremeno biti u dva različita podskupa) i iscrpnost (svaki element statističkog skupa mora biti obuhvaćen pri grupiranju podataka). Grupiranjem kvalitativnih podataka dobiva se učestalost ili frekvencija svakog modaliteta unutar statističkog skupa. Postoji apsolutna frekvencija modaliteta koja je jednaka ukupnom broju pojavljivanja tog modaliteta u osnovnom skupu te relativna frekvencija koja predstavlja omjer odgovarajuće apsolutne frekvencije toga modaliteta i opsega statističkog skupa. Relativna frekvencija najčešće je prikazana u postotcima, proporcijama ili promilima.

Grupiranjem kvantitativnih podataka dobije se razdioba, odnosno distribucija frekvencija koja se definira kao skup uređenih parova, pri čemu prvi element uređenog para predstavlja vrijednost određenog modaliteta, a drugi element frekvenciju ili učestalost te vrijednosti unutar promatranog statističkog skupa. [1] Razlika u grupiranju kvantitativnih podataka u odnosu na kvalitativne je u tome što se apsolutne i relativne frekvencije mogu postupno zbrajati. U tom slučaju dobivaju se kumulativne frekvencije koje se primjenjuju na dva načina. Postoji kumulativna apsolutna frekvencija ''manje od'' koja se definira kao ukupan broj elemenata osnovnog skupa koji su ili strogo manji od dotičnog modaliteta ili jednaki tom modalitetu, te kumulativna apsolutna frekvencija ''veće od'' koja označava ukupan broj elemenata osnovnog skupa koji su ili strogo veći od dotičnog modaliteta ili su mu jednaki.

Funkcija koja se u MS Excelu koristi za grupiranje podataka naziva se *Histogram*. U nastavku je dan primjer grupiranja kvantitativnih podataka u programu MS Excel.

Primjer 2.1. Dani su podaci o broju odsutnih učenika s nastave u 30 razreda srednje škole u jednom danu. Potrebno je grupirati navedene podatke: 2, 3, 3, 7, 6, 3, 2, 4, 5, 1, 5, 2, 3, 1, 1, 3, 1, 2, 4, 3, 4, 7, 2, 1, 3, 4, 2, 2, 7, 2 te izračunati kumulativne i relativne frekvencije.

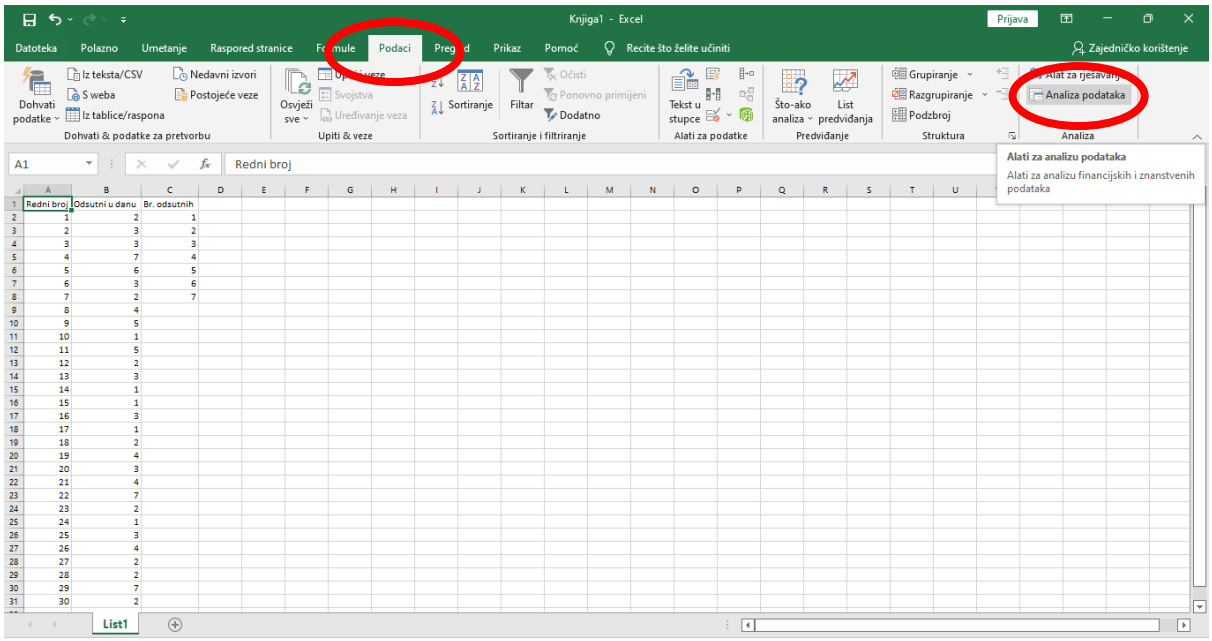
Prvo je potrebno upisati podatke u prazan radni list programa MS Excel kao što je prikazano na Slika 2.

	A	B	C
1	Redni broj	Odsutni u danu	Br. odsutnih
2	1	2	1
3	2	3	2
4	3	3	3
5	4	7	4
6	5	6	5
7	6	3	6
8	7	2	7
9	8	4	
10	9	5	
11	10	1	
12	11	5	
13	12	2	
14	13	3	
15	14	1	
16	15	1	
17	16	3	
18	17	1	
19	18	2	
20	19	4	
21	20	3	
22	21	4	
23	22	7	
24	23	2	
25	24	1	
26	25	3	
27	26	4	
28	27	2	
29	28	2	
30	29	7	
31	30	2	

Slika 2. Unos statističkih podataka na radni list

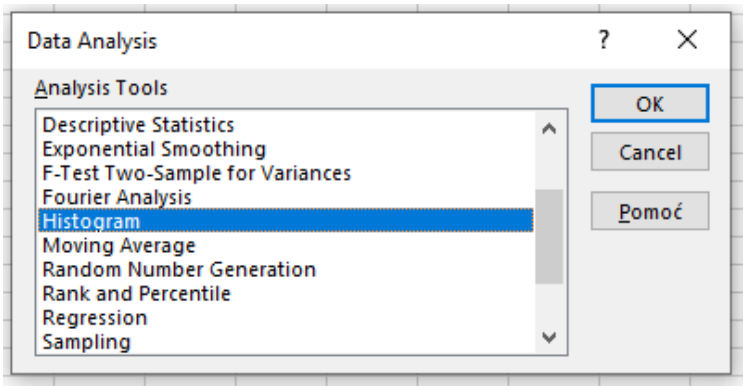
Pritom je u stupcu B upisan broj odsutnih učenika u danu po pojedinom razredu. U Stupcu C navedeni su modaliteti na temelju kojih će se i izvršiti postupak grupiranja, a to su u ovom slučaju brojevi odsutnih učenika koji se pojavljuju u primjeru, točnije brojevi od 1 do 7 (u nekom razredu odsutan je 1 učenik, u nekom 2 itd.).

Nakon unosa podataka, može se krenuti na postupak grupiranja, a kao što je već prethodno navedeno, za to se koristi funkcija *Histogram*. Na padajućem izborniku *Podaci* odaberemo opciju *Analiza podataka* (Slika 3.).

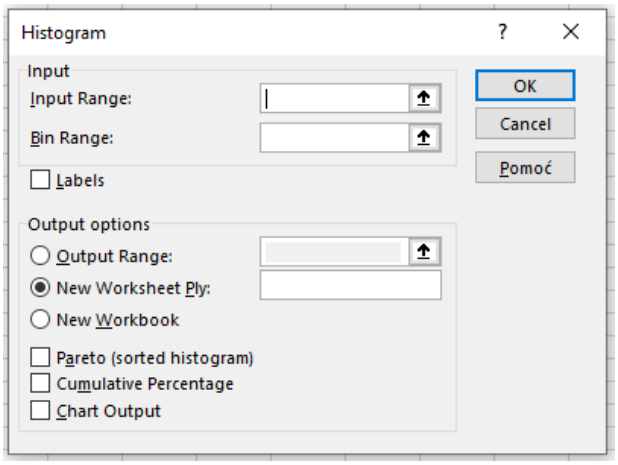


Slika 3. Odabir funkcije Analiza podataka

Dobije se okvir *Data Analysis* u kojem se odabere opcija *Histogram* (Slika 4.) te se klikom na OK na radnom listu otvori dijaloški okvir *Histogram* (Slika 5.).

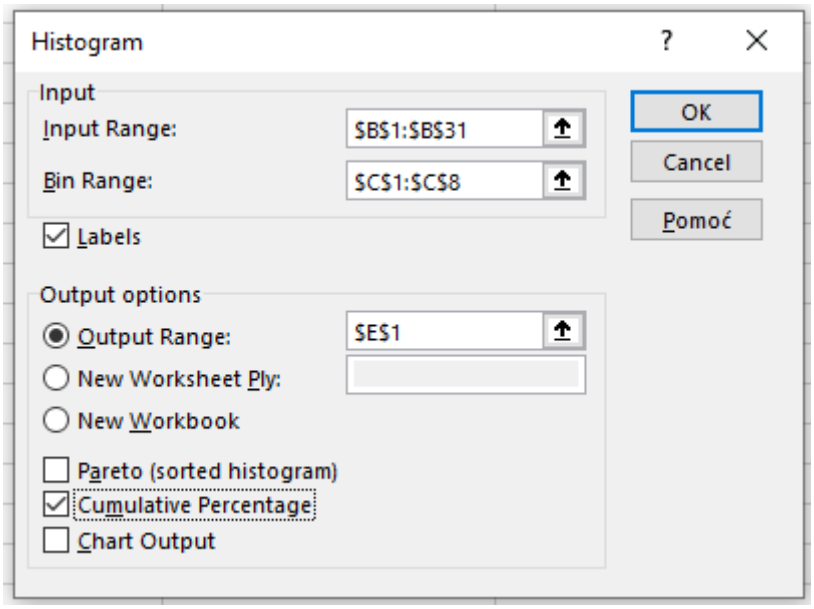


Slika 4. Data Analysis



Slika 5. Dijaloški okvir Histogram

Zatim slijedi popunjavanje dijaloškog okvira *Histogram* koji se sastoji od dva dijela. U prvi dio (*Input*) unose se adrese podataka koje se žele grupirati, a u drugi (*Output options*) odabire se mjesto i vrsta dobivenih rezultata. Za navedeni primjer, unutar okvira *Input Range* unose se adrese ćelija u kojima su negrupirani podaci, a to su u ovom slučaju ćelije od B1 do B31 u kojima su podaci o odsutnim učenicima u danu za pojedine razrede. U okvir *Bin Range* unose se adrese ćelija u kojima se nalaze modaliteti (C1:C8). Pošto su odabrane i adrese ćelija u kojima se nalaze naslovi (B1 i C1), kvačicom je označena i opcija *Labels*. U drugom dijelu dijaloškog okvira nalaze se tri opcije za odabir mjesta na kojem će se pojaviti tablica s grupiranim podacima: *Output Range* ako se želi da tablica bude na istom radnom listu, ali pritom u okviru treba upisati adresu ćelije u kojoj će biti njezin lijevi gornji kut, *New Worksheet Ply* za postavljanje tablice na novi radni list u odabranu adresu ćelije ili *New Workbook* ako je potrebno da tablica bude u novoj radnoj knjizi. Za konkretan primjer odabrana je opcija *Output Range* i adresa ćelije E1. Na kraju dijaloškog okvira postoje tri opcije – *Pareto (sorted histogram)*, *Cumulative Percentage* i *Chart Output*. Pošto se u ovom primjeru radi o kvantitativnim podacima, moguće je izračunati kumulativni niz te je stoga odabrana opcija *Cumulative Percentage*.



Slika 6. Ispunjavanje dijaloškog okvira Histogram

Klikom na opciju OK, na radnom listu se pojavi tablica prikazana na Slika 7.

E	F	G
Br. odsutnih	*Frequency*	*Cumulative %*
1	5	16,67%
2	8	43,33%
3	7	66,67%
4	4	80,00%
5	2	86,67%
6	1	90,00%
7	3	100,00%
More	0	100,00%

Slika 7. Dobivena tablica grupiranih podataka

U stupcu E nalaze se modaliteti koji su i odabrani u dijaloškom okviru *Histogram*, dok se stupac F sastoji od frekvencija njihovog pojavljivanja u skupu podataka. Podaci se interpretiraju na način da npr. u 5 razreda je odsutan samo 1 učenik, u 8 razreda odsutna su 2 učenika, u 7 razreda odsutna su 3 učenika itd.

Kumulativni niz frekvencija formira se postupnim zbrajanjem frekvencija odozgo prema dolje (kumulativni niz "manje od") ili odozdo prema gore (kumulativni niz "više od"). Radi preglednosti, dodan je novi stupac pomakom kumulativnih frekvencija u desno. Za kumulativni niz "manje od" u prvu ćeliju stupca (G2) prepisuje se prva apsolutna

frekvencija (5), a sljedeća vrijednost kumulativnog niza se dobiva zbrajanjem prve vrijednosti sa sljedećom apsolutnom frekvencijom (Slika 8.).

E	F	G	H
Br. odsutnih	*Frequency*	*Kumulativni niz "manje od"*	*Cumulative %*
1	5	5	16,67%
2	8	=G2+F3	43,33%
3	7		66,67%
4	4		80,00%
5	2		86,67%
6	1		90,00%
7	3		100,00%
More	0		100,00%

Slika 8. Računanje kumulativnog niza "manje od"

Za kumulativni niz "više od", u zadnju ćeliju stupca (H9) upiše se vrijednost 0, a ostale vrijednosti dobivaju se zbrajanjem prethodne sa sljedećom apsolutnom frekvencijom čitajući odozdo prema gore (Slika 9.). Također je radi preglednosti umetnut novi stupac za formiranje kumulativnog niza.

E	F	G	H	I
Br. odsutnih	*Frequency*	*Kumulativni niz "manje od"*	*Kumulativni niz "više od"*	*Cumulative %*
1	5	5		16,67%
2	8	13		43,33%
3	7	20		66,67%
4	4	24		80,00%
5	2	26		86,67%
6	1	27		90,00%
7	3	30	=H9+F8	100,00%
More	0	30	0	100,00%

Slika 9. Računanje kumulativnog niza "više od"

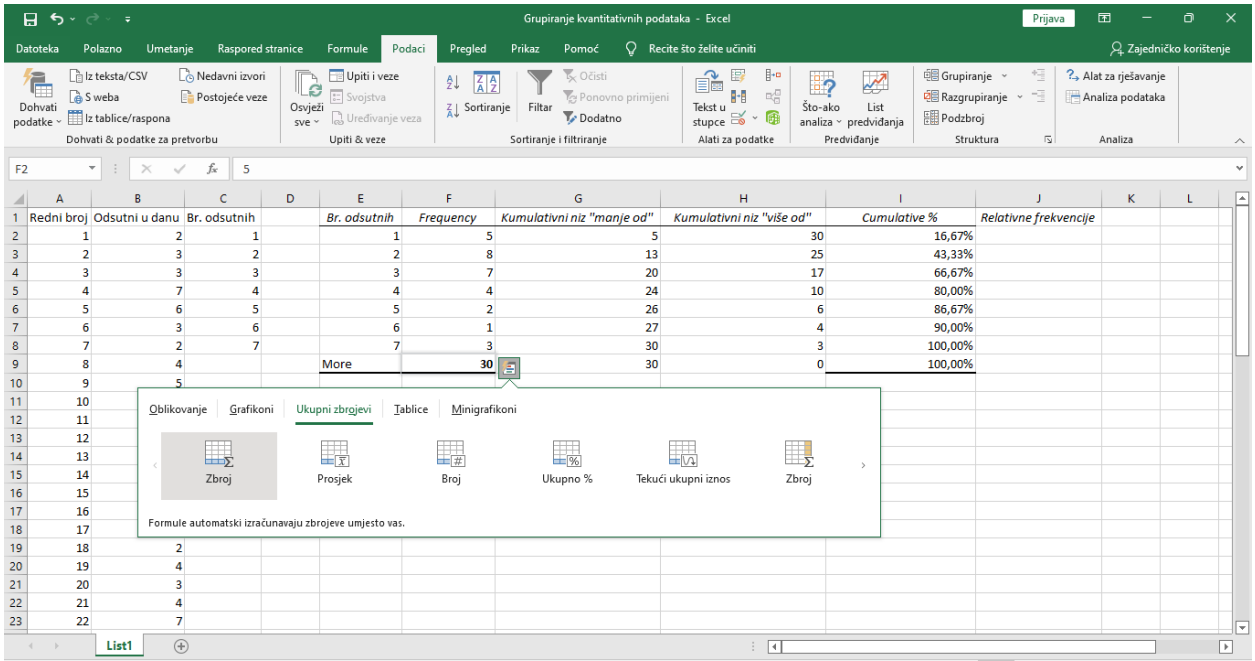
Na Slika 10. prikazani su dobiveni podaci. Zadnja kumulativna frekvencija uvijek mora biti jednaka zbroju svih frekvencija.

E	F	G	H	I
Br. odsutnih	*Frequency*	*Kumulativni niz "manje od"*	*Kumulativni niz "više od"*	*Cumulative %*
1	5	5	30	16,67%
2	8	13	25	43,33%
3	7	20	17	66,67%
4	4	24	10	80,00%
5	2	26	6	86,67%
6	1	27	4	90,00%
7	3	30	3	100,00%
More	0	30	0	100,00%

Slika 10. Dobiveni podaci kumulativnih nizova

Dobivene kumulativne frekvencije interpretiraju se na sljedeći način: u 13 razreda odsutna su 2 ili manje od 2 učenika, u 10 razreda odsutna su 4 ili više od 4 učenika itd.

U nastavku je prikazan izračun relativnih frekvencija. Prije njihovog izračuna, potrebno je izračunati sumu apsolutnih frekvencija koje se nalaze u stupcu F. Označe se ćelije F2:F8, odabere se ikona koja se pojavi u donjem desnom kutu te u okviru koji se pojavi, klikne se na opciju *Zbroj* (Slika 11.). Zatim su izračunate relativne frekvencije prema Slika 12. na način da se svaka apsolutna frekvencija podijelila s njihovim ukupnim zbrojem, a zatim su se dobivene vrijednosti pomoću dijaloškog okvira *Oblikovanje ćelija* prebacile u oblik postotka s dvije decimale (Slika 13.).



Slika 11. Suma apsolutnih frekvencija

E	F	G	H	I	J
Br. odsutnih	*Frequency*	*Kumulativni niz "manje od"*	*Kumulativni niz "više od"*	*Cumulative %*	*Relativne frekvencije*
1	5	5	60	16,67%	=F2/F9
2	8	13	55	43,33%	
3	7	20	47	66,67%	
4	4	24	40	80,00%	
5	2	26	36	86,67%	
6	1	27	34	90,00%	
7	3	30	33	100,00%	
More	30	60	30	100,00%	

Slika 12. Računanje relativnih frekvencija

J
Relativne frekvencije
16,67%
26,67%
23,33%
13,33%
6,67%
3,33%
10,00%
100,00%

Slika 13. Dobivene relativne frekvencije

Malim sređivanjem tablice, na Slika 14. prikazani su konačni rezultati zadanog primjera. Gledajući relativne frekvencije, u 16,67% razreda odsutan je jedan učenik, u 6,67% razreda odsutno je pet učenika itd. Ako se promatraju kumulativne frekvencije, može se reći da 80% razreda ima odsutna četiri ili manje od četiri učenika, a 66,67% tri ili manje od tri odsutna učenika.

E	F	G	H	I	J
Br. odsutnih	**Odsutni u danu**	**Kumulativni niz "manje od"**	**Kumulativni niz "više od"**	**Kumulativne frekvencije**	**Relativne frekvencije**
1	5	5	30	16,67%	16,67%
2	8	13	25	43,33%	26,67%
3	7	20	17	66,67%	23,33%
4	4	24	10	80,00%	13,33%
5	2	26	6	86,67%	6,67%
6	1	27	4	90,00%	3,33%
7	3	30	3	100,00%	10,00%
Ukupno	**30**				100,00%

Slika 14. Konačni rezultati

Ako postoji velik broj modaliteta prilikom grupiranja kvantitativnih podataka, radi preglednosti ih je bolje grupirati u razrede. Svaki razred ima svoju donju i gornju granicu, a one mogu biti neprave (nominalne) ili prave (precizne). [1] Kod nepravih granica postoji određeni razmak između gornje granice jednog i donje granice idućeg razreda te se one najčešće primjenjuju za analizu modaliteta kvantitativnog diskretnog obilježja, dok se prave granice najčešće koriste u slučaju kontinuiranih numeričkih obilježja te je kod njih gornja granica svakog razreda jednaka donjoj granici sljedećeg.

Za lakši, sažeti prikaz kvantitativnih statističkih podataka često se koristi dijagram stablo-list (*stem-leaf*). Njegova osnovna ideja je zapisati svaki numerički podatak u obliku "x.y", gdje je y bilo koja znamenka dekadskog brojevnog sustava, a x ostatak numeričkog podatka. Pritom x predstavlja stablo, a y list. Često se u takvom dijagramu daje odgovarajuća "uputa za uporabu", tj. objašnjenje kako pravilno očitati polaznu numeričku vrijednost. Npr. decimalni broj 40.07 ima dvije znamenke iza decimalne točke te je u tom slučaju list znamenka 7, a stablo 400. Međutim, iz ovakvog zapisa slijedi da je originalan podatak 400.7, što nije točno pa se u ovom slučaju navodi "uputa za uporabu" koja glasi ''podijeliti zapis x.y s 10''. Prije unošenja podataka u tablicu, iste je potrebno poredati po veličini od najmanje do najveće vrijednosti. U nastavku je dan primjer konstruiranje S-L dijagrama.

Primjer 2.2.: Vrijeme od dana primitka narudžbe koju je kupac ispostavio do dana isporuke specijalnog uređaja za zavarivanje bilo je (u danima) kako slijedi: 145, 117, 185, 140, 182, 175, 132, 131, 161, 156, 140, 141, 164, 136, 123, 157, 192, 200, 177, 125, 154, 132. Konstruirati S-L dijagram. [4]

Najprije je potrebno vrijednosti varijable poredati po veličini od najmanje do najveće: 117, 123, 125, 131, 132, 132, 136, 140, 140, 141, 145, 154, 156, 157, 161, 164, 175, 177, 182, 185, 192, 200. Kako se radi od troznamenkastom broju, za stablo će se uzeti prve dvije znamenke, a za list znamenka na mjestu jedinica te se tako i formira tablica. (Slika 15.)

stablo	list
11	7
12	3,5
13	1,2,2,6
14	0,0,1,5
15	4,6,7
16	1,4
17	5,7
18	2,5
19	2
20	0
uputa za uporabu:	*pomnožiti stablo.list s 10*

Slika 15. Dijagram stablo-list

Iz ovako navedene tablice pomoću "upute za uporabu" koja govori da se originalni podaci dobivaju množenjem s 10, lako se očitaju polazni podaci. Npr. za prvi podatak stablo je 11, a list 7 što daje broj 11.7, no množenjem s 10 dobiva se polazna vrijednost 117.

2.3.2. Tablični i grafički prikaz statističkih podataka

Tabličnim prikazom statističkih podataka cilj je jasno i na pregledan način prikazati rezultate prikupljanja i grupiranja podataka. Svaka statistička tablica mora sadržavati naslov, tekstualni dio, numerički dio i izvor podataka. Ako postoji jedan statistički niz koji je nastao sređivanjem podataka prema modalitetima jednog obilježja, on se prikazuje jednostavnom tablicom, dok se više takvih statističkih nizova prikazuju skupnom tablicom. Takve tablice vrlo je jednostavno izraditi u MS Excelu pomoću naredbe *Pivot-table and PivotChart Report* te se kao takve nazivaju pivot-tablice.

Grafički prikaz statističkih podataka najvažniji je dio prezentiranja rezultata statističkog istraživanja, pošto su oni ti koji na svim vrstama izlaganja najprije privuku pozornost slušatelja. Grafički se prikazi mogu podijeliti u četiri osnovne skupine:

1) Površinski grafikoni
2) Linijski grafikoni
3) Točkasti grafikoni
4) Kartogrami

Uz svaki grafički prikaz, kako bi bio potpuno čitljiv, potrebno je navesti i odgovarajuće oznake poput naslova grafikona, nazive kategorija na koordinatnim osima te legendu, tj. dodatno objašnjenje osobitosti grafikona. MS Excel je zasigurno jedan od najboljih programa za kreiranje grafičkih prikaza.

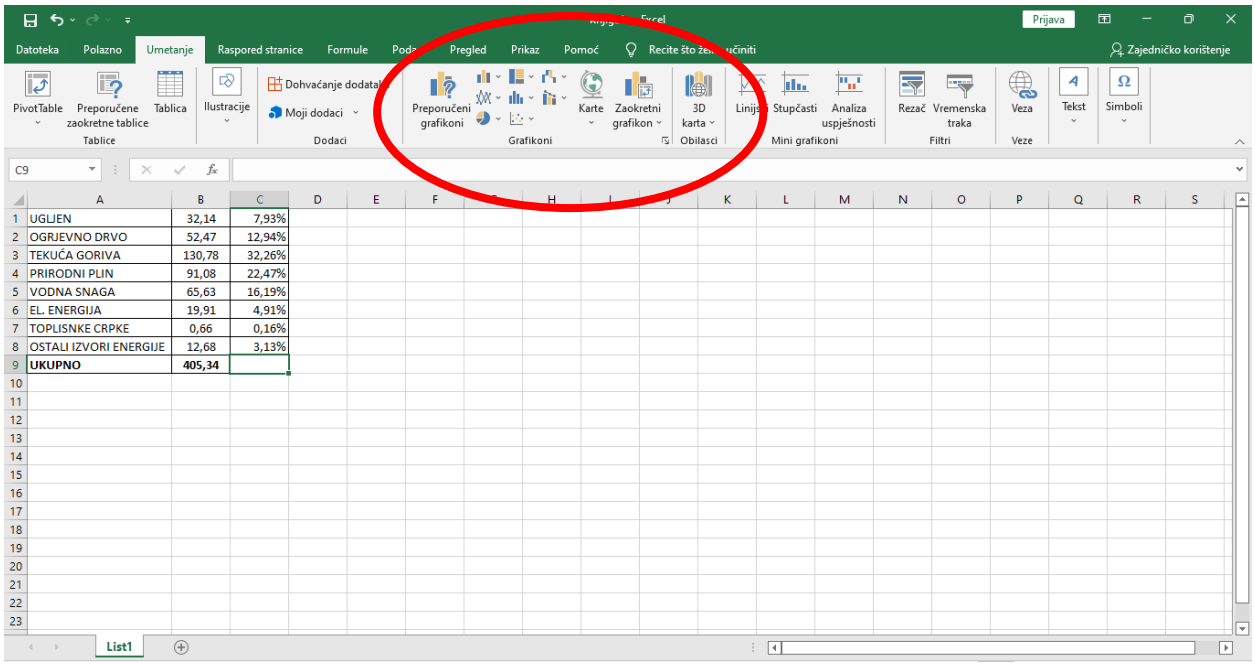
2.3.2.1. Površinski grafikoni

Vrste površinskih grafikona koji se najčešće koriste su različite vrste stupaca (jednostavni, položeni, dvostruki, višestruki, strukturni), strukturni krugovi ili polukrugovi te histogram. U praksi se najčešće kao grafički prikaz koriste strukturni krugovi i jednostavni stupci upravo zbog njihove jednostavnosti i preglednosti. U slučaju kad postoji veliki broj kategorija, grafički prikaz je pregledniji ako se koriste položeni stupci.

Primjer 2.3.. Dani su podaci o ukupnoj potrošnji energije za 2016.g. [5]

	Potrošnja [PJ]
Ugljen	32,14
Ogrjevno drvo	52,47
Tekuća goriva	130,78
Prirodni plin	91,08
Vodna snaga	65,63
El. energija	19,91
Toplinske crpke	0,66
Ostali izvori energije	12,68
Ukupno	405,34

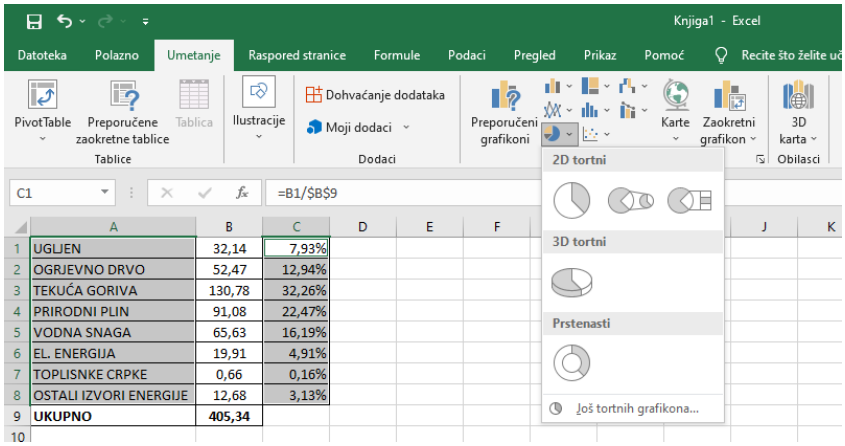
Na navedenom primjeru bit će objašnjeno grafičko prikazivanje podataka u MS Excelu. Prvo su podaci upisani u prazan radni list, a pošto su vrijednosti u petadžulima (PJ) izračunati su i postotci. Na padajućem izborniku *Umetanje* postoji dio *Grafikoni* gdje se odabire željeni grafički prikaz (Slika 16).



	A	B	C
1	UGLJEN	32,14	7,93%
2	OGRJEVNO DRVO	52,47	12,94%
3	TEKUĆA GORIVA	130,78	32,26%
4	PRIRODNI PLIN	91,08	22,47%
5	VODNA SNAGA	65,63	16,19%
6	EL. ENERGIJA	19,91	4,91%
7	TOPLISNKE CRPKE	0,66	0,16%
8	OSTALI IZVORI ENERGIJE	12,68	3,13%
9	**UKUPNO**	**405,34**	

Slika 16. Umetanje grafikona

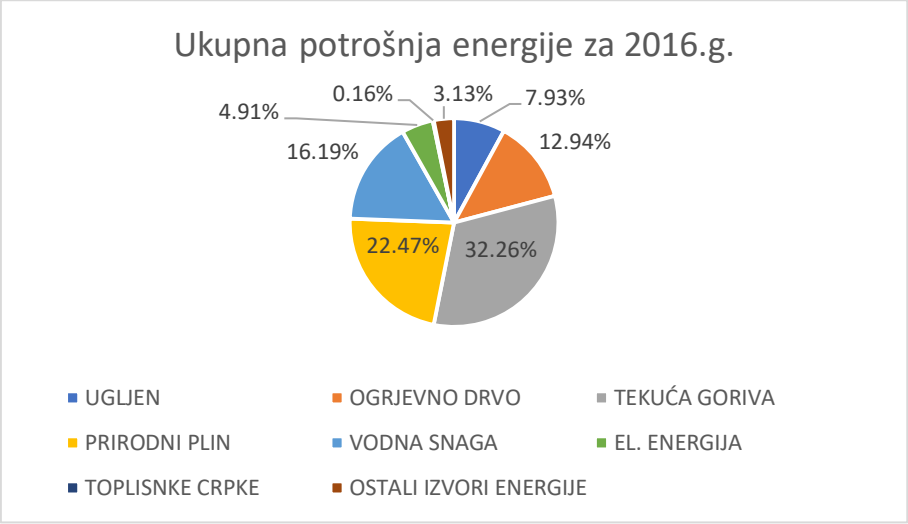
Označe se ćelije s podacima koji trebaju biti prikazani te se u izborniku klikne na željeni grafički prikaz, npr. za strukturni krug odabere se 2D tortni prikaz (Slika 17.)



	A	B	C
1	UGLJEN	32,14	7,93%
2	OGRJEVNO DRVO	52,47	12,94%
3	TEKUĆA GORIVA	130,78	32,26%
4	PRIRODNI PLIN	91,08	22,47%
5	VODNA SNAGA	65,63	16,19%
6	EL. ENERGIJA	19,91	4,91%
7	TOPLISNKE CRPKE	0,66	0,16%
8	OSTALI IZVORI ENERGIJE	12,68	3,13%
9	**UKUPNO**	**405,34**	

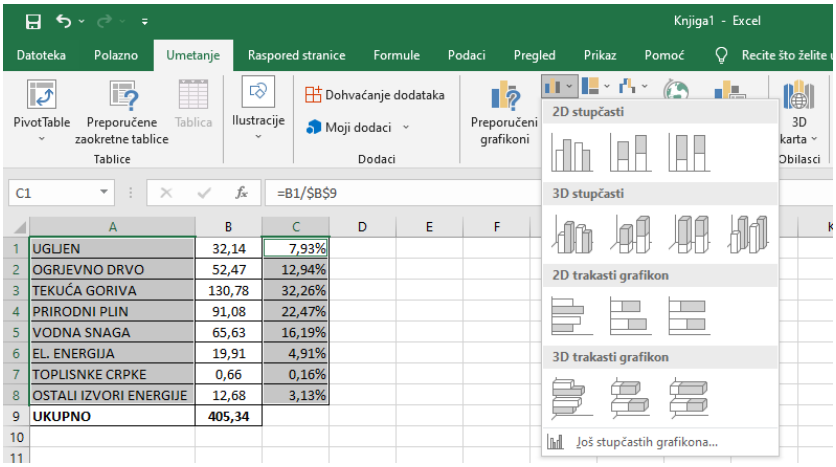
Slika 17. Odabir grafičkog prikaza

Dobije se prikaz kao na Slika 18. na kojem je bilo potrebno upisati i naziv grafikona. Isti se može i uređivati po želji (promjena boje, dizajna) u padajućem izborniku *Dizajn grafikona*.



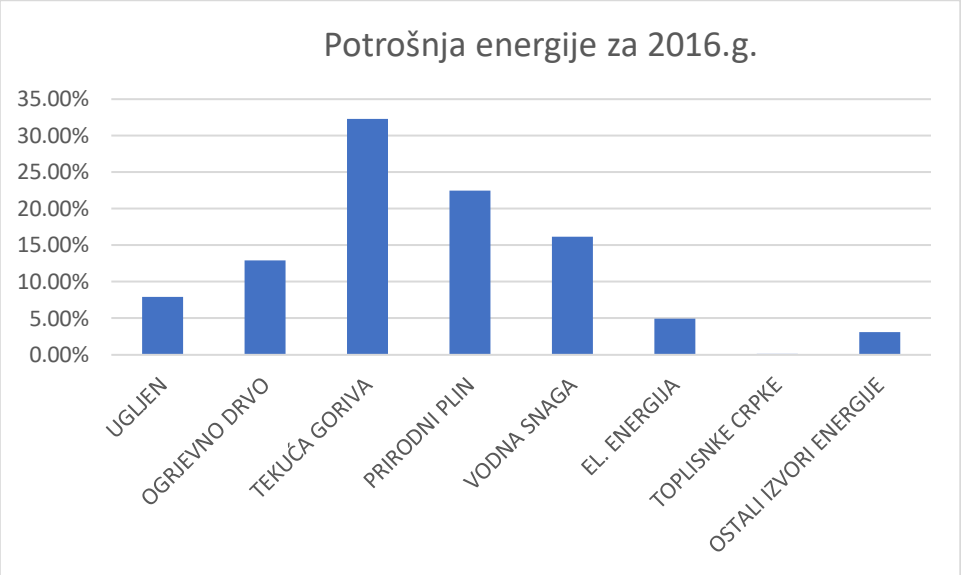
Slika 18. Strukturni krug

Iste podatke moguće je prikazati i pomoću jednostavnih i položenih stupaca koji su za ovaj primjer i prikladniji zbog većeg broja podataka. U padajućem izborniku *Umetanje* odabere se ikona za umetanje stupčastog ili trakastog grafikona (Slika 19.).

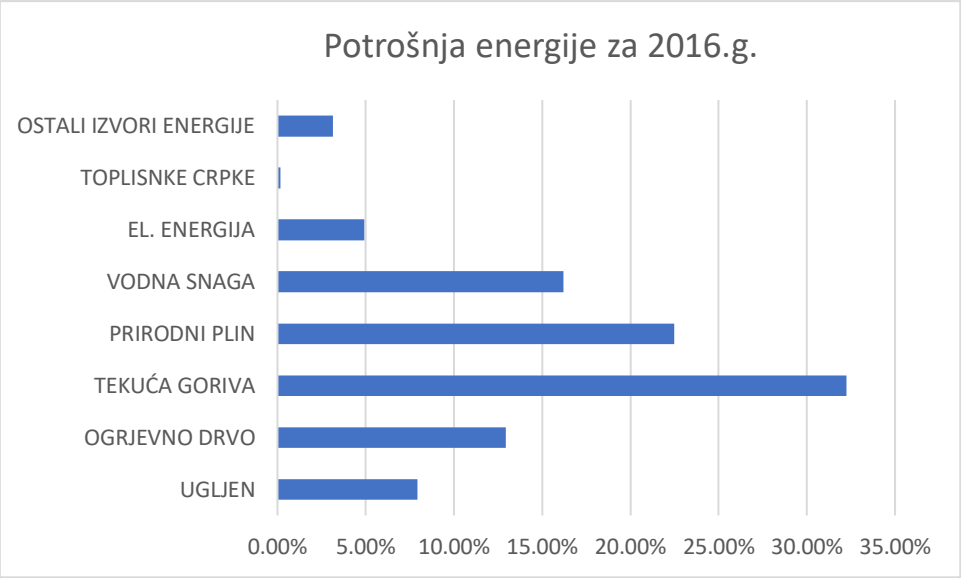


Slika 19. Umetanje stupčastog ili trakastog grafikona

Slika 20. i Slika 21. prikazuju dobivene grafikone koji se također mogu uređivati. Moguće je dodavati, ukloniti ili promijeniti elemente grafikona kao što su naslov, legenda, nazivi osi. Također je moguće promijeniti stil i boje grafikona.

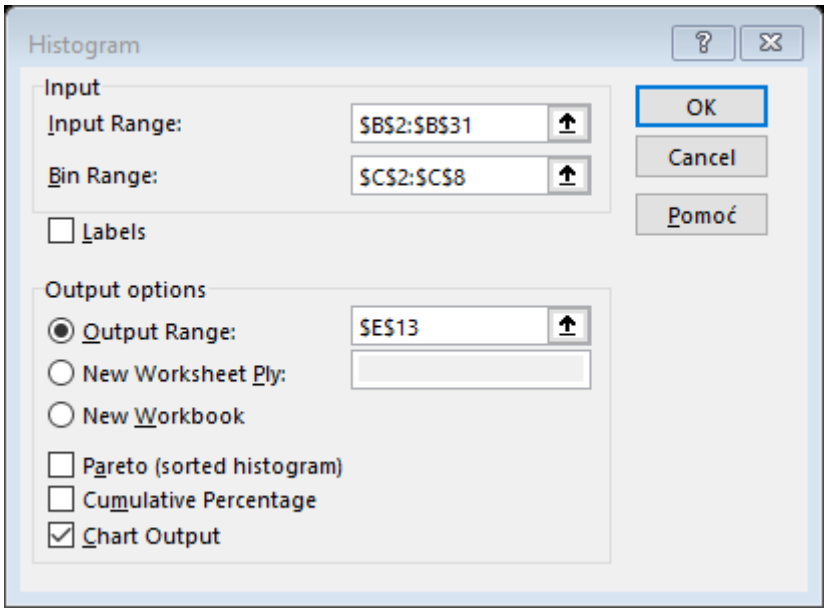


Slika 20. Jednostavni stupci



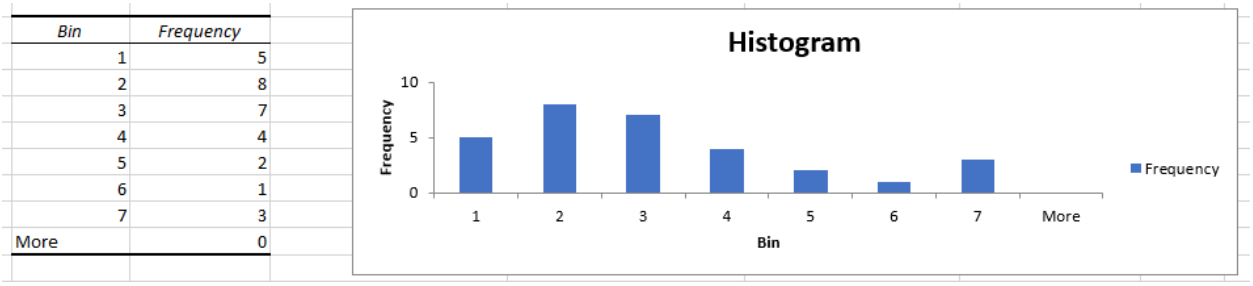
Slika 21. Položeni stupci

Posebna vrsta površinskog stupčastog grafikona je histogram koji je već spomenut prilikom grupiranja podataka. Ta procedura omogućava i direktno dobivanje grafičkog prikaza. Ako se uzmu podaci iz Primjera 2.1., označe se ćelije s vrijednostima koje treba grupirati kao i modaliteti te se u zadnjem dijelu dijaloškog okvira odabere opcija *Chart Output* za grafički prikaz histograma (Slika 22.)



Slika 22. Chart Output

Dobiju se podaci i izgled histograma kao na Slika 23. kojeg je potrebno urediti da bi bio pregledniji (Slika 24.). Na x osi nalaze se modaliteti, odnosno u ovom slučaju broj odsutnih učenika, dok su na y osi frekvencije pojavljivanja modaliteta u skupu podataka.



Bin	Frequency
1	5
2	8
3	7
4	4
5	2
6	1
7	3
More	0

Slika 23. Dobiveni podaci u grafičkom prikazu histogram

Slika 24. Grafički prikaz histogram

Nakon posljednjeg zadanog modaliteta, Excel automatski doda opciju "*More*" ("više") koja može prikazati neki modalitet iza zadnjeg navedenog. U ovom primjeru frekvencija pod opcijom "More" je nula, što znači da nema više razreda u kojima je neki učenik odsutan. Ova opcija može predstavljati i kontrolu za slučaj da dođe do pogrešnog grupiranja podataka.

Razlika između histograma i jednostavnih stupaca je u tome što se histogram koristi za prikaz modaliteta kvantitativnih obilježja koji su grupirani u razrede, dok se jednostavni stupci koriste za prikaz kvalitativnih podataka. Također, jednostavnim stupcima se varijable uspoređuju, a histogramom se prikazuju distribucije varijabli.

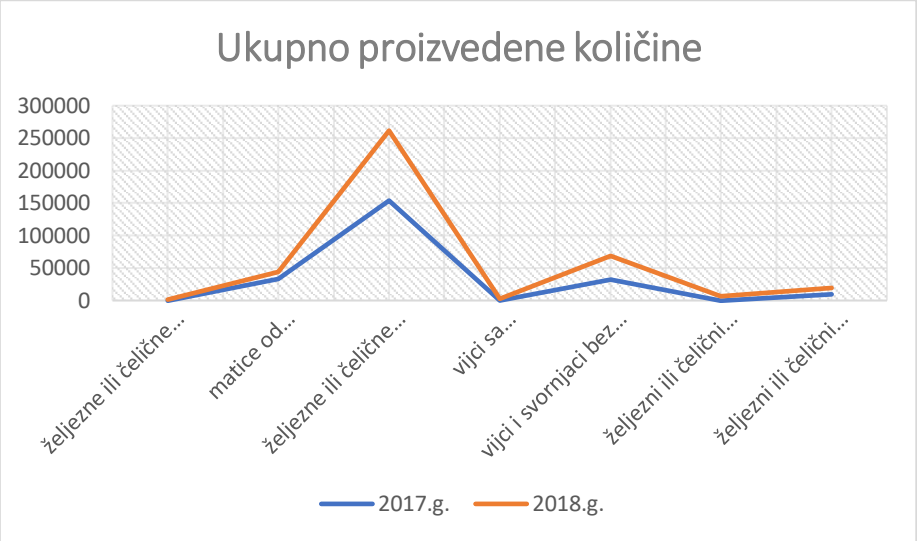
2.3.3. Linijski grafikoni

Ova vrsta grafikona prikladna je za grafički prikaz poligona frekvencija – apsolutnih, relativnih i kumulativnih, za prikaz vremenskih nizova kao i za usporedbu kvantitativnih podataka dvaju nizova.

Primjer 2.4.: Dani su podaci o ukupno proizvedenim količinama pojedinih proizvoda prema statističkom izvješću u 2017. i 2018.g. [6] koji će biti prikazani pomoću linijskog grafikona.

	A	B	C
1		**2017.g.**	**2018.g.**
2	željezne ili čelične plošne spiralne opruge	127	1288
3	matice od nehrđajućeg čelika	33680	10236
4	željezne ili čelične matice	153794	107715
5	vijci sa šesterokutnom glavom od nehrđajućeg čelika	4	2299
6	vijci i svornjaci bez glave od čelika	32243	36043
7	željezni ili čelični svornjaci s glavom	50	6991
8	željezni ili čelični vijci za drvo	9495	9780

Slika 25. Podaci za kreiranje linijskog grafikona



Slika 26. Linijski grafikon

3. DESKRIPTIVNA STATISTIKA

Kao što je već prethodno u radu navedeno, deskriptivna statistika u svrhu opisivanja svojstava koristi mjere centralne tendencije i mjere varijabilnosti. U nastavku rada bit će objašnjena njihova podjela.

3.1. Mjere centralne tendencije

Pošto se pokusi često bave velikim brojem podataka, u praksi se javlja potreba da se ti podaci sažmu i to na način da se niz prikupljenih podataka zamijeni jednom, srednjom vrijednosti. Srednja vrijednost je konstanta kojom se predočuje niz varijabilnih podataka, a predstavlja i vrijednost oko koje se gomila većina podataka numeričkog niza pa se zato naziva mjerom centralne tendencije. Osnovna podjela srednjih vrijednosti je na potpune i položajne. U potpune srednje vrijednosti ubrajaju se aritmetička, geometrijska i harmonijska sredina te se prilikom izračuna ovih vrijednosti koriste svi elementi numeričkog niza. U skupinu položajnih srednjih vrijednosti ubrajaju se medijan i mod. Kako i sam naziv govori, njihova je vrijednost određena položajem unutar numeričkog niza pa u njihovom izračunu ne sudjeluju svi elementi.

3.1.1. Aritmetička sredina

Aritmetička sredina je najraširenija i najčešće korištena srednja vrijednost. Računa se na način da se zbroje vrijednosti numeričke varijable i podijele s njihovim ukupnim brojem. Brojnik aritmetičke sredine tj. zbroj svih elemenata konačnog numeričkog niza, definira se kao total.

Najpoznatija vrsta aritmetičkih sredina je jednostavna aritmetička sredina za čiji se izračun koristi konačan niz negrupiranih kvantitativnih podataka $x_1, x_2, \dots, x_n$, a formula glasi:

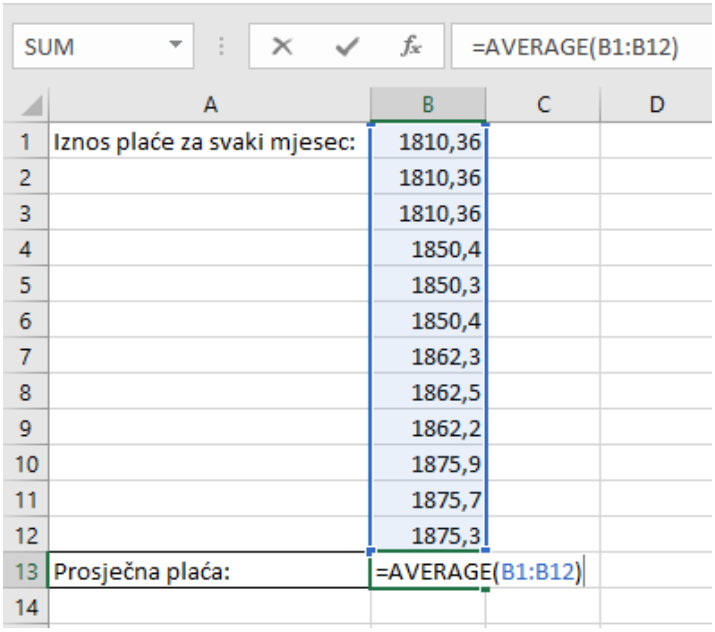
$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

(3.1)

U programu MS Excel aritmetička sredina iz negrupiranih podataka računa se pomoću funkcije =AVERAGE (raspon podataka).

Primjer 3.1. Dane su mjesečne neto-plaće (izražene u kunama) jednog radnika u prethodnoj godini: 1810,36; 1810,36; 1810,36; 1850,4; 1850,3; 1850,4; 1862,3; 1862,5; 1862,2; 1875,9; 1875,7; 1875,3. Potrebno je izračunati njegovu prosječnu mjesečnu plaću protekle godine. [2]

Podaci se prvo unose u prazan radni list programa MS Excel, a pošto se radi o negrupiranim podacima, za izračun aritmetičke sredine koristit će se funkcija =AVERAGE. U praznu ćeliju (B13) upiše se "=AVERAGE" te se označe sve ćelije s vrijednostima koje trebaju biti u izračunu, u ovom primjeru su to ćelije od B1 do B12 (Slika 27.) Pritiskom tipke Enter dobije se konačna vrijednost. (Slika 28.)



Slika 27. Average

	A	B
1	Iznos plaće za svaki mjesec:	1810,36
2		1810,36
3		1810,36
4		1850,4
5		1850,3
6		1850,4
7		1862,3
8		1862,5
9		1862,2
10		1875,9
11		1875,7
12		1875,3
13	Prosječna plaća:	1849,673

Slika 28. Dobivena prosječna plaća

Dakle, prosječna mjesečna neto-plaća u protekloj godini za ovog radnika iznosila je približno 1849,67 kn.

Ako je zadan niz grupiranih podataka, u tom se slučaju računa vagana ili ponderirana aritmetička sredina, a njezina formula glasi:

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

(3.2)

Pritom oznake $x_1, x_2, \ldots, x_k$ označavaju međusobno različite modalitete kvantitativnog obilježja, a $f_1, f_2, \ldots, f_k$ su njima pripadne apsolutne frekvencije za koje se često kaže da imaju ulogu pondera ili težine pojedinog modaliteta pa otuda i sam naziv navedene aritmetičke sredine. U programu MS Excel ne postoji funkcija za direktno računanje ponderirane aritmetičke sredine, već se računa postupno prema formuli.

Primjer 3.2. Ispitani su zastoji na proizvodnoj liniji u jednom poduzeću. Broj zastoja promatrao se po radnim smjenama. Analizom 400 radnih smjena dobivena je sljedeća distribucija: [4]

Broj zastoja	0	1	2	3	4	5	6
Broj smjena	45	95	112	48	15	25	60

Iz formule (3.2) se vidi da su za izračun aritmetičke sredine potrebne apsolutne frekvencije (f_i) i modaliteti (x_i). Pošto su u ovom primjeru podaci već grupirani, broj zastoja predstavlja modalitete, a broj smjena njima pripadne apsolutne frekvencije. Svi podaci upisani su u radni list, izračunata je suma apsolutnih frekvencija pomoću funkcije SUM. Također je izračunat umnožak modaliteta i frekvencija ($f_i * x_i$) te njihov zbroj. Tražena aritmetička sredina dobivena je dijeljenjem zbroja umnožaka modaliteta i frekvencija sa sumom apsolutnih frekvencija te rezultat prikazuje da je prosječan broj zastoja po radnoj smjeni jednak 2,52. (Slika 29.)

D11 =D9/C9

	A	B	C	D
1		Broj zastoja (xi)	Broj smjena (fi)	fi*xi
2		0	45	0
3		1	95	95
4		2	112	224
5		3	48	144
6		4	15	60
7		5	25	125
8		6	60	360
9	**Σ**		**400**	**1008**
10				
11			Arit. sredina:	2,52

Slika 29. Ponderirana aritmetička sredina

U nastavku su navedena osnovna svojstva aritmetičke sredine koja služe kao procjena da li izračunata aritmetička sredina dobro reprezentira numerički niz [1]:

1. Svaki statistički skup u kojem su vrijednosti obilježja pridružene jedinicama skupa na temelju intervalne ili omjerne skale, ima aritmetičku sredinu
2. Sve vrijednosti statističkog skupa uključene su u izračunavanje aritmetičke sredine (potpunost)
3. Numerički niz ima samo jednu aritmetičku sredinu
4. Aritmetička sredina može poslužiti za usporedbu dvaju ili više numeričkih nizova koji su nastali grupiranjem prema istom obilježju
5. Aritmetička sredina nalazi se uvijek između najmanje (X_{min}) i najveće (X_{max}) vrijednosti numeričkog obilježja u distribuciji
6. Aritmetička sredina jedina je mjera centralne tendencije gdje je zbroj odstupanja svih vrijednosti od sredine uvijek jednak nuli:

$$\sum_{i=1}^{N}(x_i - \bar{X}) = 0 \qquad \sum_{i=1}^{N} f_i \cdot (x_i - \bar{X}) = 0$$

(3.3)

7. Zbroj kvadrata odstupanja pojedinih vrijednosti x_i od aritmetičke sredine je minimalan, tj. ako umjesto $\bar{X}$ uzmemo bilo koji broj $a \neq \bar{X}$, zbroj kvadrata odstupanja pojedine vrijednosti x_i od a bit će veći od prethodno spomenutog zbroja
8. Aritmetička sredina neće biti zadovoljavajuće reprezentativna kada u numeričkom nizu postoje ekstremno male ili velike vrijednosti promatranog obilježja. To je slučaj kada npr. u nekom poduzeću radnik ima plaću 2000 kn, a netko na puno višoj poziciji 20000 kn, pritom aritmetička sredina ne opisuje dobro plaće u tom poduzeću.

9. Aritmetička sredina izračunata na temelju distribucije frekvencija u kojoj su modaliteti obilježja predstavljeni razredima najčešće sadrži pogrešku zato što izračunata razredna sredina, koja se koristi za određivanje vagane sredine, predstavlja samo aproksimativnu zamjenu stvarne sredine odgovarajućeg razreda
10. Problem reprezentativnosti aritmetičke sredine dodatno je izražen u slučaju kada postoje otvoreni razredi u distribuciji frekvencija, a osobito kada nije moguće objektivno procijeniti nepoznate granice otvorenih razreda

3.1.2. Geometrijska sredina

Ako $x_1, x_2, \dots, x_n$ označava konačan numerički niz takav da za svaki $i = 1,2, \dots, n$ vrijedi $x_i > 0$, onda se geometrijska sredina (G) definira kao srednja vrijednost koja se dobije kao n-ti korijen iz umnoška svih članova niza. U slučaju kada je barem jedan član numeričkog niza jednak nuli, geometrijska sredina se ne definira.

Kao i aritmetička sredina, geometrijska se također može izračunati iz negrupiranih i grupiranih podataka. Formula za izračunavanje iz negrupiranih podataka je sljedeća:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

(3.4)

U programu MS Excel geometrijska sredina iz negrupiranih podataka računa se pomoću funkcije GEOMEAN.

U slučaju kada su podaci grupirani i postoji ukupno *n* različitih modaliteta $x_1, x_2, \dots, x_n$ te je za svaki $i = 1,2, \dots, n$ s f_i označena apsolutna frekvencija modaliteta x_i, tada je vagana (ponderirana) geometrijska sredina izračunata iz grupiranih podataka izrazom:

$$G = \sqrt[\sum_{i=1}^{n} f_i]{x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_n^{f_n}}$$

(3.5)

3.1.3. Harmonijska sredina

Ako se s $x_1, x_2, \dots, x_n$ označi konačan numerički niz podataka takav da za svaki $i = 1,2, \dots, n$ vrijedi da je $x_i > 0$, harmonijska sredina takvog niza definira se kao recipročna vrijednost aritmetičke sredine recipročnih vrijednosti svih elemenata niza. Ako u numeričkom nizu postoji barem jedan član čija je vrijednost nula, harmonijska sredina se u tom slučaju ne definira.

Formula za izračun harmonijske sredine negrupiranih numeričkih podataka je sljedeća:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

(3.6)

U programu MS Excel harmonijska sredina iz negrupiranih podataka računa se pomoću funkcije HARMEAN.

U slučaju kada su podaci grupirani, harmonijska sredina računa se prema:

$$H = \frac{f_1 + f_2 + \cdots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_n}{x_n}} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

(3.7)

3.1.4. Mod

Često se prilikom analize statističkih nizova želi odrediti modalitet koji se najčešće pojavljuje, tj. modalitet s najvećom apsolutnom ili relativnom frekvencijom. Takav modalitet naziva se mod (Mo), a postoji i naziv dominantna vrijednost (D). Mod je jedina srednja vrijednost koja se može odrediti i za kvalitativna i za kvantitativna obilježja. Kao i sve ostale srednje vrijednosti i mod se može računati iz negrupiranih i grupiranih podataka, ali i iz grafičkih prikaza razdioba. U programu MS Excel, mod iz negrupiranih podataka računa se pomoću funkcije MODE.

Jedno od svojstava moda je da ne mora biti jedinstven, točnije može se dogoditi da barem dva modaliteta imaju jednake apsolutne frekvencije, a da su apsolutne frekvencije svih ostalih modaliteta strogo manje od njih. Iz toga se razlikuju unimodalne razdiobe (razdiobe koje imaju samo jedan mod), bimodalne razdiobe (razdiobe koje imaju točno dva moda) i multimodalne razdiobe (razdiobe koje imaju barem tri moda). Npr. niz 1, 2, 2, 3, 3, 4, 5 ima točno dva moda (2 i 3) te je stoga ovaj niz bimodalan.

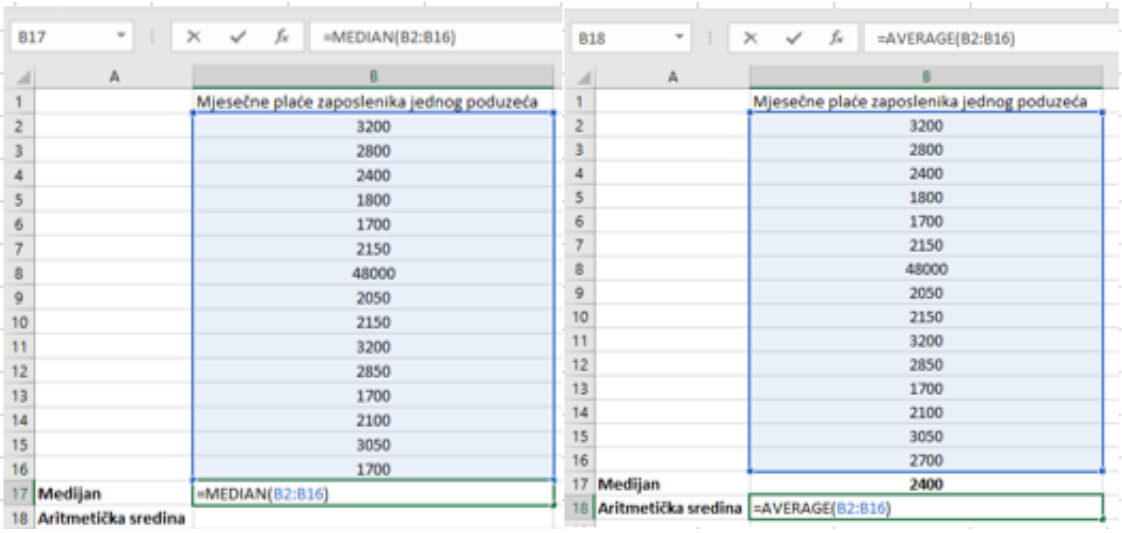
3.1.5. Medijan

Medijan (Me) se definira kao srednja položajna vrijednost numeričkog obilježja koja uređeni statistički niz dijeli na dva jednaka dijela na način da 50% jedinica u nizu ima vrijednost obilježja jednaku ili manju od vrijednosti medijana, a preostalih 50% vrijednost obilježja veću ili jednaku od vrijednosti medijana. U literaturi se ponekad umjesto naziva medijan koristi naziv centralna vrijednost (C). U programu MS Excel, funkcija pomoću koje se računa medijan iz negrupiranih podataka je MEDIAN.

Ako je broj podataka neparan, medijan predstavlja vrijednost varijable središnjeg člana niza uređenog prema veličini, a ako niz ima paran broj podataka, medijan je jednak aritmetičkoj sredini vrijednosti varijable središnjih dvaju članova niza:

- neparan broj podataka: $n = 2k + 1 \rightarrow Me = x_{k+1}$
- paran broj podataka: $n = 2k \rightarrow Me = \frac{x_k + x_{k+1}}{2}$

U situacijama kad u nizu postoje ekstremno niski i visoki rezultati, medijan predstavlja pogodniju srednju vrijednost od aritmetičke sredine. Npr. ako su mjesečne plaće 15 zaposlenika u nekom poduzeću 3200, 2800, 2400, 1800, 1700, 2150, 48000, 2050, 2150, 3200, 2850, 1700, 2100, 3050 i 2700 kn, njihov medijan ispada 2400 što znači da je 50% zaposlenih u tom poduzeću imalo plaću manju od 2400 kn ili 2400 kn, a 50% zaposlenih veću od 2400 kn ili 2400 kn. S druge strane njihova aritmetička sredina je 5456,67 kn što govori da prosječna mjesečna plaća 15 zaposlenih u toj tvrtki iznos 5456,67 kn. Jasno je da je medijan reprezentativnija mjera od aritmetičke sredine jer u konkretnom slučaju plaća zaposlenika koja iznosi 48000 kn znatno utječe na visoku vrijednost aritmetičke sredine te time daje lažnu sliku o visini plaća. Izračun podataka dobiven je u MS Excelu koristeći funkcije MEDIAN i AVERAGE kako je prikazano na Slika 30., dok Slika 31. prikazuje dobivene rezultate.



Slika 30. Izračun medijana

	A	B
1		Mjesečne plaće zaposlenika jednog poduzeća
2		3200
3		2800
4		2400
5		1800
6		1700
7		2150
8		48000
9		2050
10		2150
11		3200
12		2850
13		1700
14		2100
15		3050
16		2700
17	**Medijan**	**2400**
18	**Aritmetička sredina**	**5456,666667**

Slika 31. Dobiveni rezultati medijana

3.2. Mjere disperzije

Same srednje vrijednosti često ne daju dobru sliku o skupu podataka pa je njihovo računanje samo jedan korak u statističkoj analizi. Sljedeće što je potrebno je odrediti koliko dobivena srednja vrijednost dobro opisuje elemente osnovnog skupa. Pokazatelji koji se pritom definiraju nazivaju se mjere disperzije (raspršenja, stupnja varijabilnosti) koje mogu biti apsolutne i relativne.

Apsolutnim mjerama disperzije pripadaju raspon varijacija, interkvartil, varijanca te standardna devijacija. Njihovo zajedničko obilježje je da se iskazuju u jedinicama mjere varijable koja se analizira. U skupinu relativnih mjera disperzije pripadaju koeficijent kvartilne devijacije i koeficijent varijacije, koje se obično izražavaju u postotcima.

Manja mjera disperzije pokazuje veću reprezentativnost srednje vrijednosti i obrnuto. Pomoću mjera disperzije uspoređuju se razlike varijabiliteta dviju ili više distribucija pri čemu vrijedi da ako se radi o istim obilježjima, za usporedbu se koriste apsolutne mjere disperzije, a ako je riječ o distribucijama različitih obilježja, onda se obavezno koriste relativne mjere disperzije.

3.2.1. Raspon varijacije

Raspon varijacije (R) je najjednostavnija mjera disperzije koja se definira kao razlika između najveće (X_{max}) i najmanje (X_{min}) vrijednosti kvantitativnog obilježja:

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

(3.8)

Najmanja vrijednost koja je moguća za raspon varijacije je 0 i to u slučaju kada svi podaci imaju istu vrijednost, a najveća vrijednost nije općenito određena. Ako se radi o grupiranim podacima, za najmanju vrijednost se uzima donja granica prvog razreda, a za najveću gornja granica posljednjeg razreda.

U programu MS Excel ne postoji direktna funkcija za izračun raspona varijacije, ali se na jednostavan način može izračunati pomoću relacije:

=MAX(raspon podataka) – MIN(raspon podataka)

Kao prednost raspona varijacije navodi se jednostavnost njegovog izračuna, a kao nedostatak korištenje samo dva elementa osnovnog skupa.

3.2.2. Interkvartil

Interkvartil (I_q) je apsolutna mjera disperzije koja se definira kao raspon varijacije središnjih 50% elemenata uređenog niza statističkih podataka, a to je zapravo razlika između trećeg (gornjeg) i prvog (donjeg) kvartila:

$$I_q = Q_3 - Q_1$$

(3.9)

Prednost interkvartila u odnosu na raspon varijacije je eliminiranje ekstremnih vrijednosti čime se dobiva točniji opis većine elemenata niza, a nedostatak je također što se u izračunu koriste samo dvije vrijednosti pa se zbog toga smatra nepotpunom mjerom.

Program MS Excel nema funkciju za direktno računanje interkvartila, ali može se izračunati iz sljedeće relacije pomoću funkcija za gornji i donji kvartil:

=QUARTILE(raspon podataka; 3) – QUARTILE(raspon podataka; 1)

3.2.3. Varijanca i standardna varijacija

Varijanca i iz nje izvedena standardna devijacija predstavljaju najvažniji pokazatelj varijabiliteta odnosno raspršenosti modaliteta kvantitativnog obilježja, a razlog tome je što se u izračunu koriste svi elementi statističkog niza pa se pritom može smatrati potpunom mjerom raspršenja. Također, mogu se računati i iz negrupiranih i iz grupiranih podataka.

Varijanca (σ^2) se definira kao aritmetička sredina kvadrata odstupanja vrijednosti kvantitativnog obilježja od aritmetičke sredine svih vrijednosti. [2] Računa se iz formule:

- za negrupirane podatke:

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

(3.10)

- za grupirane podatke:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i \cdot (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i}$$

(3.11)

Varijanca je jednaka nuli samo ako je pripadni statistički niz konstantan.

U programu MS Excel varijancu iz negrupiranih podataka može se izračunati pomoću funkcije: VARP(raspon podataka).

Varijancu je relativno teško interpretirati pošto je ona "kvadratna" mjera disperzije, ali pomoću njezinog drugog korijena dobiva se praktično najprimjenjenija mjera – standardna devijacija (σ). Ona se definira kao drugi korijen iz varijance, no takva definicija se obično zamjenjuje matematički netočnom, ali intuitivno prihvatljivom definicijom koja govori da je standardna devijacija prosječno odstupanje od aritmetičke sredine. [2] Analogno formulama za varijancu, formule za standardnu devijaciju su sljedeće:

- za negrupirane podatke:

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

(3.12)

- za grupirane podatke:

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

(3.13)

Za izračun standardne devijacije u programu MS Excel koristi se funkcija: STDEVP(raspon podataka).

3.2.4. Koeficijent varijacije

Koeficijent varijacije (*V*) također predstavlja jedan od najvažnijih pokazatelja disperzije kvantitativnih podataka uz varijancu i standardnu devijaciju. Pripada relativnim mjerama disperzije, a predstavlja omjer standardne devijacije i aritmetičke sredine pomnožen sa sto, dakle prikazuje se u postotcima:

$$V = \frac{\sigma}{\bar{x}} \cdot 100$$

(3.14)

Najmanja vrijednost koeficijenta varijacije je 0% i to u slučaju ako svi podaci imaju istu vrijednost, a najveća vrijednost nije općenito određena. U Tablica 2. je prikazan kriterij procjene varijabiliteta prema vrijednostima koeficijenta varijacije. Glavni nedostatak mu je loša reprezentativnost u slučaju ekstremnih vrijednosti.

Tablica 2. Varijabilitet elemenata ovisno o postotku koeficijenta varijacije [1]

V(%)	**VARIJABILITET**
0-10	vrlo slab
10-30	relativno slab
30-50	umjeren
50-70	relativno jak
>70	vrlo jak

Program MS Excela nema direktnu funkciju za izračun koeficijenta varijacije, ali se može izračunati prema formuli (3.14) uz korištenje funkcija:

=STDEVP(raspon podataka)/AVERAGE (raspon podataka)*100

3.2.5. Koeficijent kvartilne devijacije

Interkvartilu pripadna relativna mjera disperzije je upravo koeficijent kvartilne devijacije (V_q) koji se definira kao omjer interkvartila i zbroja prvog i trećeg kvartila:

$$V_q = \frac{I_q}{Q_3 + Q_1} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

(3.15)

Može se interpretirati i kao intenzitet varijabiliteta središnjih 50% elemenata uređenog statističkog niza, a može poprimiti bilo koju vrijednost od 0 do 1. Što je varijabilitet središnje polovice niza manji, vrijednost V_q je bliža nuli i obratno, što je varijabilitet središnje polovice veći, to je vrijednost V_q bliža 1. U Tablica 3. je prikazana najčešća interpretacija intenziteta varijabiliteta središnjih 50% elemenata uređenog statističkog skupa ovisno o iznosu koeficijenta kvartilne devijacije.

Tablica 3. Varijabilitet središnjih 50% elemenata ovisno o iznosu Vq [1]

V_q	VARIJABILITET
0-0,1	vrlo slab
0,1-0,2	relativno slab
0,2-0,3	umjeren
0,3-0,5	relativno jak
0,5-1	vrlo jak

U MS Excelu se koeficijent kvartilne devijacije može izračunati prema formuli.. koristeći funkcije:

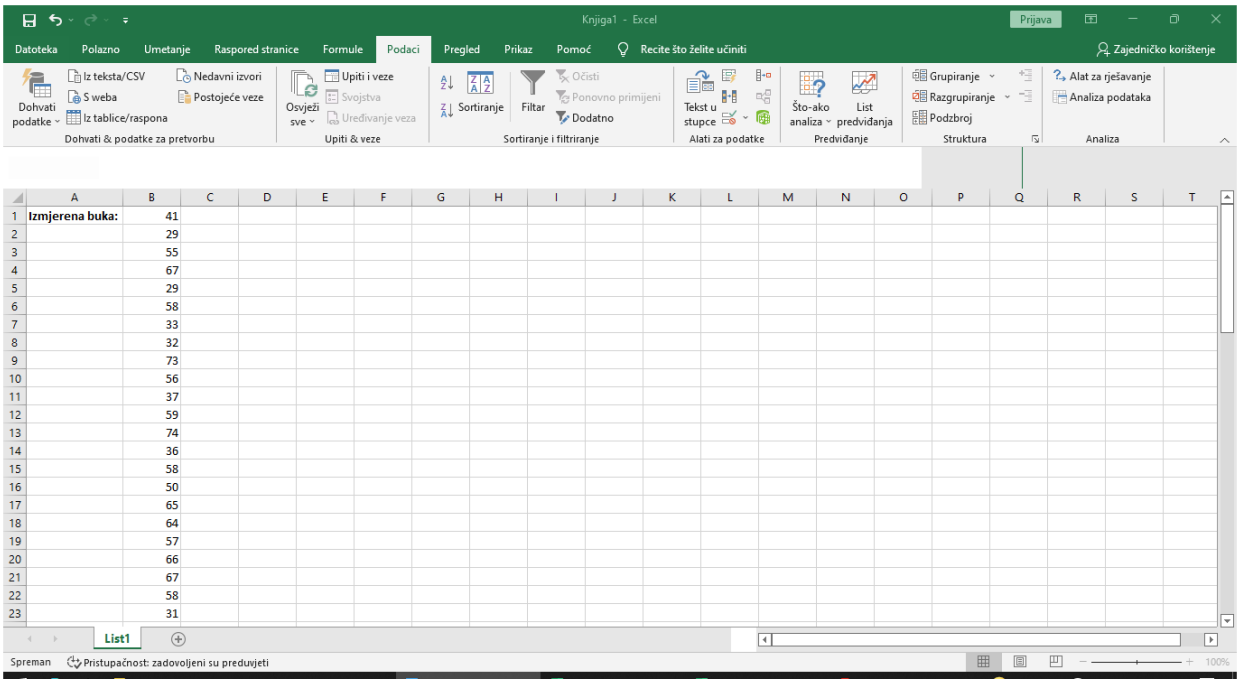
=(QUARTILE(…;3)-QUARTILE(…;1))/(QUARTILE(…;3)+QUARTILE(…;1))

3.3. Procedura *Descriptive Statistics*

Za jednostavnije rješavanje zadataka, u programu MS Excel postoji funkcija *Descriptive Statistics* koja se nalazi unutar alata *Analiza podataka (Dana Analysis)* pomoću koje se u jednom koraku izračunaju najvažniji elementi deskriptivne statistike poput aritmetičke sredine, medijana, moda, varijance, standardne devijacije i ostalih koji će biti prikazani u nastavku rada na primjeru.

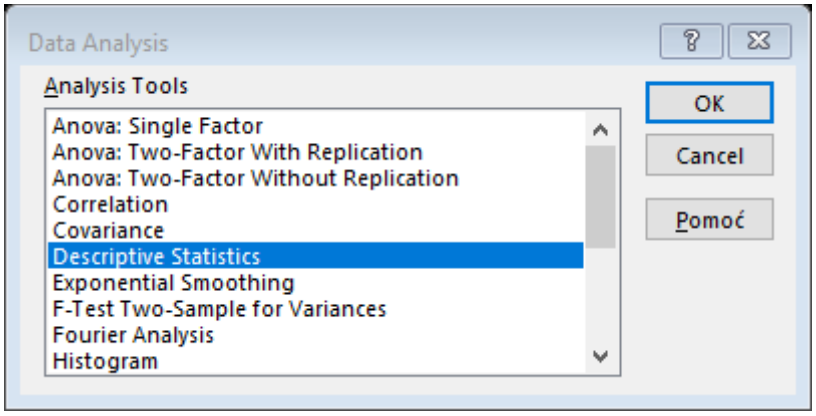
Primjer 3.3.: U poduzeću Metal ispitivani su uvjeti rada u proizvodnom pogonu kako bi se poduzele mjere radi povećanja produktivnosti. Među ostalim, mjerena je i buka na 73 radna mjesta, a dobiveni su sljedeći rezultati (u decibelima): 41, 29, 55, 67, 29, 58, 83, 32, 73, 56, 37, 59, 74, 36, 58, 50, 65, 64, 57, 66, 67, 58, 31, 69, 61, 41, 46, 65, 47, 30, 55, 26, 37, 50, 64, 40, 55, 74, 73, 49, 67, 29, 25, 23, 40, 69, 46, 39, 46, 57, 67, 42, 26, 39, 37, 38, 58, 46, 69, 20, 27, 37, 74, 24, 58, 29, 24, 39, 58, 27, 46, 62, 43. [4]

Za navedeni primjer, najvažniji podaci deskriptivne statistike bit će izračunati pomoću procedure *Descriptive Statistics*. Podaci su najprije upisani na prazan radni list MS Excela kao što je prikazano na Slika 32.



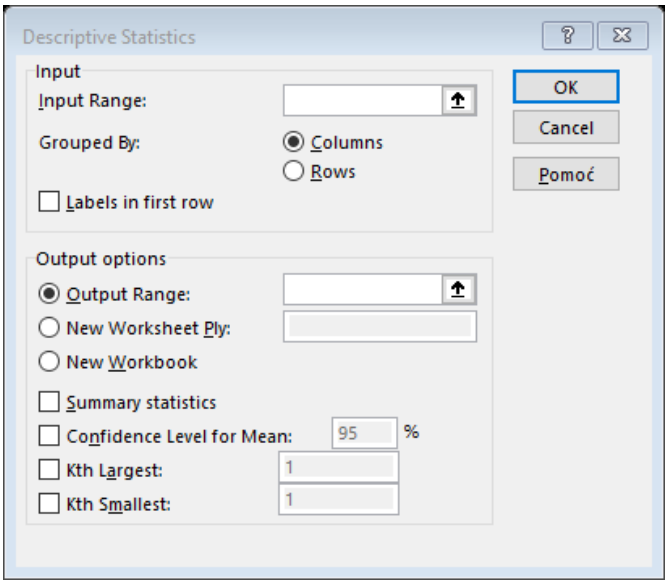
Slika 32. Prikaz upisanih podataka

Nakon toga, u alatnoj traci *Podaci* unutar alata *Analiza podataka*, pokreće se dijaloški okvir *Data Analysis* unutar kojeg se odabire procedura *Descriptive Statistics* (Slika 33).

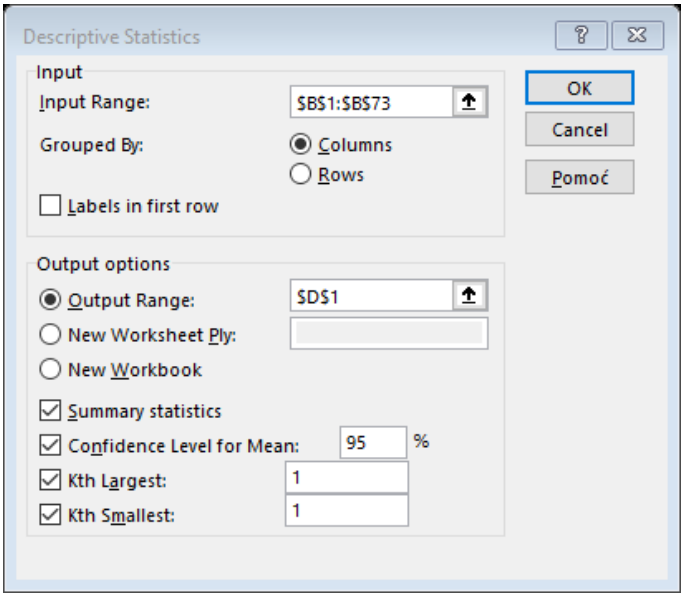


Slika 33. Odabir Descriptive Statistics

Otvara se njezin dijaloški okvir (Slika 34.) unutar kojeg je potrebno unijeti podatke (Slika 35.)



Slika 34. Dijaloški okvir Descriptive Statistics



Slika 35. Ispunjeni podaci

U polje *Input Range* unosi se raspon ćelija unutar kojih su upisani podaci koji sudjeluju u izračunu. Ispod toga moguće odabrati način grupiranja – po stupcima (*Columns*) ili retcima (*Rows*). U dijelu *Output options* odabire se mjesto gdje će biti prikazani podaci. U ovom slučaju odabrana je ćelija D1 na istom radnom listu. Kvačicom je označeno, tj odabrano *Summary statistics* jer je upravo to bit ovog izračuna. *Confidence Level of Mean* (razina povjerenja za aritmetičku sredinu) najčešće se odabire da je 95%.

Kth Largest omogućava prikaz k-te po redu najveće vrijednosti unutar promatranog statističkog skupa. Npr. ako se u polje za *Kth Largest* upiše broj 1, program će izbaciti vrijednost maksimuma. Ako se upiše broj 2, program će izbaciti drugu najveću vrijednost itd. Slično vrijedi i za *Kth Smallest*, ali se u tom slučaju radi o najmanjim vrijednostima pa će za vrijednost 1 u polju program izbaciti minimum. Klikom na *OK* dobivaju se rezultati (Slika 36.).

D	E
Column1	
Mean	48,05479452
Standard Error	1,834255726
Median	46
Mode	58
Standard Deviation	15,67188779
Sample Variance	245,608067
Kurtosis	-1,23181037
Skewness	0,00087208
Range	54
Minimum	20
Maximum	74
Sum	3508
Count	73
Largest(1)	74
Smallest(1)	20
Confidence Level(95,0%)	3,656521962

Slika 36. Dobiveni podaci

Interpretacija dobivenih podataka je sljedeća:

- *Mean* (aritmetička sredina) prikazuje da je prosječna buka 48,05 decibela
- *Standard Error* se odnosi na standardnu pogrešku aritmetičke sredine koja u ovom primjeru iznosi 1,83
- *Median* (medijan) govori da je na 50% radnih mjesta izmjerena buka 46 ili manje od 46 decibela, a na ostalih 50% 46 ili više od 46 decibela
- *Mode* (mod) pokazuje da je najčešći rezultat 58, odnosno da je na najviše radnih mjesta izmjerena buka od 58 decibela
- *Standard Deviation* (standardna devijacija) prikazuje da je prosječno odstupanje od prosječnog rezultata buke jednako 15,67
- *Sample Variance* (varijanca) prikazuje da prosječno kvadratno odstupanje od aritmetičke sredine iznosi 245,61

- *Kurtosis* (mjera zaobljenosti) prikazuje da koeficijent zaobljenosti iznosi -1,23
- *Skewness* (mjera asimetrije) prikazuje da koeficijent asimetrije iznosi 0,00087
- *Range* (raspon varijacije) govori da je raspon rezultata izmjerene buke 54 decibela
- *Minimum* (minimum) govori da je najmanja izmjerena buka 20 decibela
- *Maximum* (maksimum) govori da je najviša izmjerena buka 74 decibela
- *Sum* (zbroj) prikazuje da zbroj svih izmjerenih decibela iznosi 3508
- *Count* (veličina uzorka ili opseg promatranog statističkog skupa) govori da je ukupno 73 radna mjesta na kojima je mjerena buka
- *Largest (1)* zapravo prikazuje maksimalnu vrijednost koja iznosi 74 decibela
- *Smallest (1)* prikazuje minimalnu vrijednost koja iznosi 20 decibela
- *Confidence Level (95,0%)* (procjena intervala aritmetičke sredine osnovnog skupa uz razinu pouzdanosti od 95%) govori da se uz 95% pouzdanosti može tvrditi da se prosječni rezultat izmjerene buke u kompletnom osnovnom skupu (73 radna mjesta) nalazi unutar intervala 48,05±3,66. Ova pouzdanost se računa iz razloga što se većina istraživanja vrši na određenom uzorku, a ne na cjelokupnoj populaciji te samim time dobivena aritmetička sredina nije precizna. Zbog toga se računa interval unutar kojeg se nalazi "prava" aritmetička sredina cijelog skupa uz pomoć razine pouzdanosti

4. INFERENCIJALNA STATISTIKA

Prilikom donošenja zaključaka o populaciji na temelju izrabranog uzorka, inferencijalna statistika koristi procjene, testiranje hipoteza, određivanje veza između varijabli te predviđanja o populaciji.

Da bi uzorak mogao odgovoriti na zadaću da se pomoću njega dobiveni zaključci mogu protegnuti na cijeli osnovni skup, on mora biti reprezentativan, točnije po svojim karakteristikama mora biti kao i sam osnovni skup.

4.1. Testiranje hipoteza

Statistička hipoteza je tvrdnja o veličini parametra ili o obliku distribucije osnovnog skupa čija se vjerodostojnost ispituje pomoću slučajnog uzorka. [4] Rješavanje problema zahtjeva donošenje odluke o prihvaćanju ili ne prihvaćanju hipoteze o nekom parametru, a taj se postupak naziva testiranje hipoteza. Pritom svaki taj postupak testiranja polazi od postavljanja nulte (H_0) i alternativne (H_1) hipoteze da bi se ispitala istinitost postavljene pretpostavke. Nulta hipoteza je ona koja se testira, a alternativna joj se suprotstavlja. Postupci testiranja hipoteza oslanjaju se na korištenje informacija iz slučajnog uzorka pa ako su te informacije u skladu s hipotezom, zaključuje se da je ona istinita, a ako nisu, dolazi se do zaključka da hipoteza nije istinita. No, pošto se ne ispituje cijela populacija već samo dio, istinitost ili neistinitost pojedinih hipoteza se ne može sa sigurnošću znati, već uvijek treba imati na umu vjerojatnost donošenja pogrešnog zaključka. Pritom se može pojaviti pogreška tipa I kada se odbaci istinita nulta hipoteza te pogreška tipa II ako se prihvati lažna nulta hipoteza.

Tablica 4. Pogreške statističkih testova [4]

Odluka	**H_0 je istinita**	**H_0 je lažna**
Prihvatiti nultu hipotezu	Odluka ispravna	Pogreška tipa II
Odbaciti nultu hipotezu	Pogreška tipa I	Odluka ispravna

Pogreška tipa I prikazuje se pomoću razine značajnosti (razine signifikantnosti) α koja se odnosi na vjerojatnost odbacivanja istinite hipoteze, dok je β vjerojatnost da se prihvati lažna nulta hipoteza odnosno da dođe do pogreške tipa II.

Svaki postupak testiranja provodi se u nekoliko koraka: [4]

1. Određivanje sadržaja nulte i alternativne hipoteze
2. Identificiranje izraza za testnu veličinu i izračunavanje njezine vrijednosti

3. Odabir razine značajnosti i određivanje kritičkih granica koje dijele područje prihvaćanja nulte hipoteze od područja njezina odbacivanja
4. Donošenje zaključka o ishodu testa

P-vrijednost je vjerojatnost opažanja podataka kada je nulta hipoteza istinita. Najčešća razina značajnosti je 0,05 pa ako je $P\ vrijednost < 0{,}05$ nulta hipoteza se odbacuje, a u suprotnom prihvaća.

U nastavku rada bit će objašnjeno na koji način se pomoću programa MS Excel može odrediti koja je hipoteza istinita.

4.1.1. Z-test

Za veliki uzorak ($n < 30$) distribucija aritmetičke sredine uzorka kao procjenitelja za očekivanje ne mora biti normalna. Oznaka μ predstavlja nepoznatu aritmetičku sredinu populacije, a μ_0 njezinu pretpostavljenu veličinu.

Primjer 4.1.: Članak u časopisu Materials Engineering (1989, Vol. II, No. 4, str.275-281) opisuje rezultate ispitivanja vlačne adhezije na 22 uzorka legure U-700. Opterećenje pri lomu uzorka je kako slijedi (u megapikselima): 19.8, 10.1, 14.9, 7.5, 15.4, 15.4, 15.4, 18.5, 7.9, 12.7, 11.9, 11.4, 11.4, 14.1, 17.6, 16.7, 15.8, 19.5, 8.8, 13.6, 11.9, 11.4. [7]

Na temelju izmjerenih opterećenja, može li se zaključiti da će prosječno opterećenje biti veće od 12 megapiksela? Potrebno je pronaći 95% interval pouzdanosti za srednji promjer šipke.

Najprije je potrebno postaviti nultu i alternativnu hipotezu:

$$H_0; \mu = 12$$
$$H_1: \mu > 12$$

Zatim se zadani podaci unose u prazan radni list kao što je prikazano na Slika 37.

	A	B	C
1	Redni br.	Opterećenje	μ_0
2	1	19,8	12
3	2	10,1	
4	3	14,9	
5	4	7,5	
6	5	15,4	
7	6	15,4	
8	7	15,4	
9	8	18,5	
10	9	7,9	
11	10	12,7	
12	11	11,9	
13	12	11,4	
14	13	11,4	
15	14	14,1	
16	15	17,6	
17	16	16,7	
18	17	15,8	
19	18	19,5	
20	19	8,8	
21	20	13,6	
22	21	11,9	
23	22	11,4	

Slika 37. Podaci za z-test

Pošto u ovom slučaju alternativna hipoteza govori da je vrijednost veća od pretpostavljene, radi se o jednosmjernom z-testu. U Tablica 5. prikazane su razlike koje općenito vrijede za jednosmjerni i dvosmjerni test.

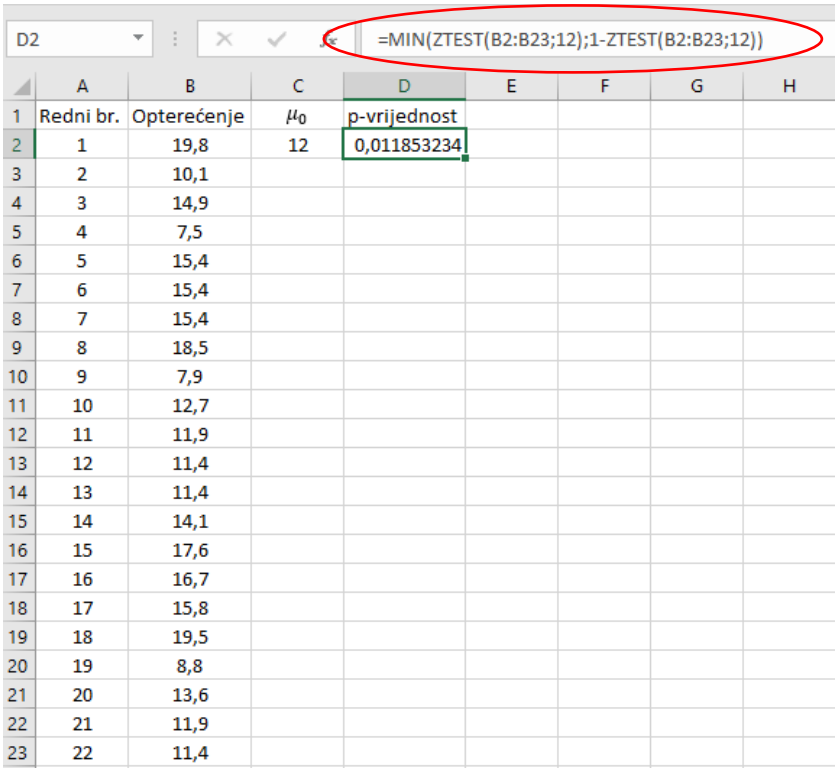
Tablica 5. Razlika između jednosmjernog i dvosmjernog testa

JEDNOSMJERNI TEST	DVOSMJERNI TEST
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	

Funkcija koja se u MS Excelu koristi za jednosmjerni z-test je:

MIN[ZTEST(raspon podataka;μ_0;σ);1-ZTEST(raspon podataka;μ_0;σ)]

Korištenjem te funkcije za navedeni primjer, dobiva se p-vrijednost 0,0119. (Slika 38.) Pošto je zadan interval pouzdanosti od 95%, to znači da p-vrijednost mora biti manja od 0,05 da bi se odbacila nulta hipoteza i prihvatila alternativna. Dobivena p-vrijednost za ovaj primjer je 0,012<0,05 pa se konačni rezultat može interpretirati na sljedeći način: uz 95% pouzdanosti zaključuje se da je razlika među promatranim vrijednostima statistički značajna, odnosno prosječno opterećenje bit će veće od 12 megapiksela.



D2 | =MIN(ZTEST(B2:B23;12);1-ZTEST(B2:B23;12))

	A	B	C	D	E	F	G	H
1	Redni br.	Opterećenje	μ_0	p-vrijednost				
2	1	19,8	12	0,011853234				
3	2	10,1						
4	3	14,9						
5	4	7,5						
6	5	15,4						
7	6	15,4						
8	7	15,4						
9	8	18,5						
10	9	7,9						
11	10	12,7						
12	11	11,9						
13	12	11,4						
14	13	11,4						
15	14	14,1						
16	15	17,6						
17	16	16,7						
18	17	15,8						
19	18	19,5						
20	19	8,8						
21	20	13,6						
22	21	11,9						
23	22	11,4						

Slika 38. Dobivena p-vrijednost

Za ručni izračun, potrebno je izračunati $\bar{x}$ i σ. Aritmetička sredina izračunata je preko Excel funkcije AVERAGE te iznosi 13.71, a standardna devijacija računa se prema formuli:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{4402{,}59 - 22 \cdot 13{,}71^2}{22-1}} = 3{,}568$$

(4.1)

Vrijednost z računa se:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{13{,}71 - 12}{\frac{3{,}568}{\sqrt{22}}} = 2{,}248$$

(4.2)

Zatim je potrebno pronaći kritičnu vrijednost koja se nalazi u statističkoj tablici za normalnu raspodjelu. Za vjerojatnost od 95%, z_α odnosno $z_{0,95}$ iznosi 1,645. Budući da je $z > z_\alpha$ (2,248>1,645) odbacuje se nulta hipoteza. Slika 39. prikazuje područje prihvaćanja i odbacivanja nulte hipoteze za z-test.

Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hip.	Područje odbacivanja H_0
$H_0 ... \theta = \theta_0$	$H_1 ... \theta \neq \theta_0$	$-z_{\alpha/2} < z < z_{\alpha/2}$	$z \leq -z_{\alpha/2}$ ili $z \geqslant z_{\alpha/2}$
$H_0 ... \theta \leq \theta_0$	$H_1 ... \theta > \theta_0$	$z < z_\alpha$	$z \geqslant z_\alpha$
$H_0 ... \theta \geqslant \theta_0$	$H_1 ... \theta < \theta_0$	$z > -z_\alpha$	$z \leq z_\alpha$

Slika 39. Područje prihvaćanja i odbacivanja nulte hipoteze za t-test

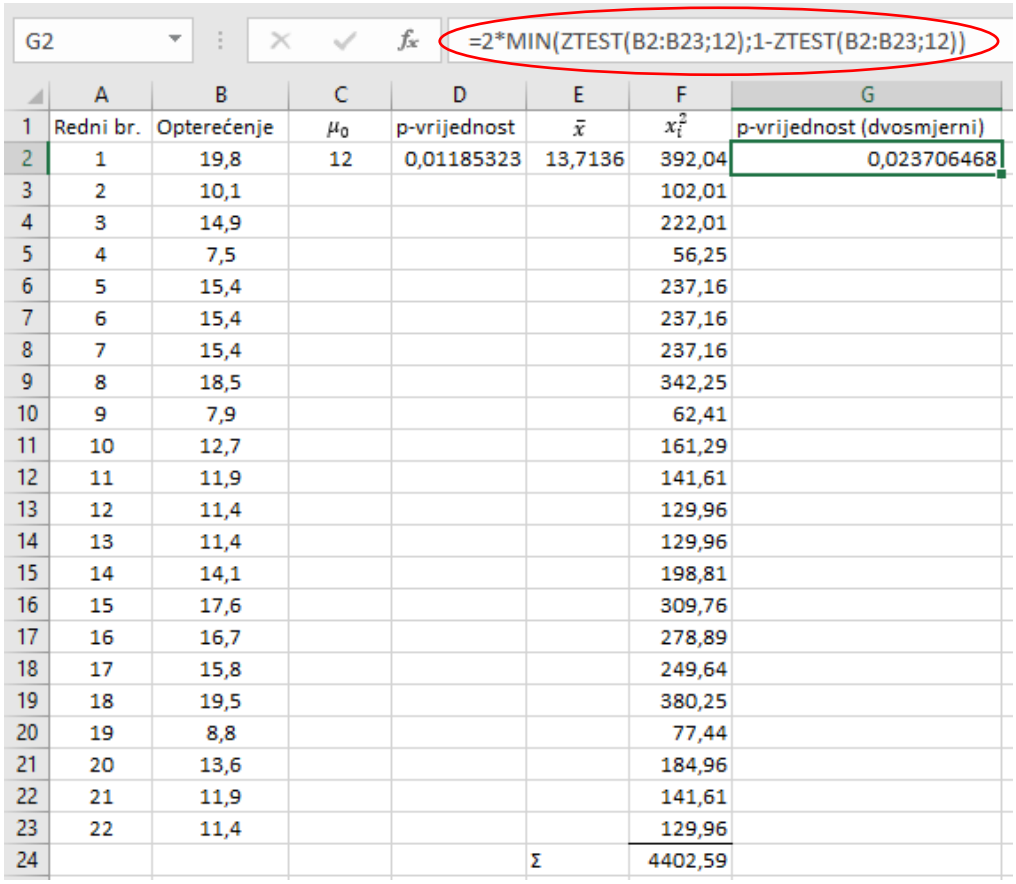
Ako se za navedeni primjer želi provjeriti odstupa li prosječno opterećenje značajno od 12, koristi se dvosmjerni z-test. Funkcija u Excelu za njega je:

2*MIN[ZTEST(raspon podataka;μ_0;σ);1-ZTEST(raspon podataka;μ_0;σ)]

Pritom su hipoteze sljedeće:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

a dobivena p-vrijednost iznosi 0,023 što je manje od 0,05 te se zaključuje da prosječno opterećenje odstupa značajno od 12, odnosno odbacuje se nulta hipoteza. (**Error! Reference source not found.**.)



G2 | =2*MIN(ZTEST(B2:B23;12);1-ZTEST(B2:B23;12))

	A	B	C	D	E	F	G
1	Redni br.	Opterećenje	μ_0	p-vrijednost	$\bar{x}$	x_i^2	p-vrijednost (dvosmjerni)
2	1	19,8	12	0,01185323	13,7136	392,04	0,023706468
3	2	10,1				102,01	
4	3	14,9				222,01	
5	4	7,5				56,25	
6	5	15,4				237,16	
7	6	15,4				237,16	
8	7	15,4				237,16	
9	8	18,5				342,25	
10	9	7,9				62,41	
11	10	12,7				161,29	
12	11	11,9				141,61	
13	12	11,4				129,96	
14	13	11,4				129,96	
15	14	14,1				198,81	
16	15	17,6				309,76	
17	16	16,7				278,89	
18	17	15,8				249,64	
19	18	19,5				380,25	
20	19	8,8				77,44	
21	20	13,6				184,96	
22	21	11,9				141,61	
23	22	11,4				129,96	
24					Σ	4402,59	

Slika 40. P-vrijednost za dvosmjerni z-test

4.1.2. T-test

T-test je statistički postupak kojim se želi odrediti statistička značajnost razlike između dva uzorka, odnosno između dvije aritmetičke sredine. Ovisno o problemu koji se rješava postoji jednosmjerni i dvosmjerni t-test, a u MS Excelu postoje njegova tri tipa. Tip 1 odabire se kad su uzorci zavisni, Tip 2 kada postoje dva uzorka s približno jednakim varijancama, a Tip 3 ako su uzorci s različitim varijancama. Jesu li varijance uzoraka jednake ili različite, može se utvrditi preko F-testa čija funkcija u MS Excelu ima oblik:

FTEST(raspon 1. uzorka; raspon 2.uzroka)

Funkcija za t-test općenito u MS Excelu je oblika:

TTEST(raspon podataka 1.uzorka; raspon podataka 2.uzorka;smjer;tip)

Ako je riječ o jednosmjernom testu, pod "smjer" se upisuje broj 1, a ako se radi o dvosmjernom testu upiše se broj 2. Pod "tip" se upisuju brojevi 1, 2 ili 3 ovisno o tipu t-testa koji su prethodno objašnjeni.

Primjer 4.2.: Želi se provjeriti pretpostavka da beton s čeličnim vlaknima ima veću tlačnu čvrstoću od betona bez čeličnih vlakana. Dobivene su sljedeće vrijednosti tlačne čvrstoće (u N/mm^2) za beton s čeličnim vlaknima: 67, 69, 71, 63, 64, 71, 65, a za beton bez čeličnih vlakana: 62, 60, 58, 63, 64, 62, 66. [8]

Pokazuju li navedeni podaci da beton s čeličnim vlaknima ima veću vlačnu čvrstoću od betona bez čeličnih vlakana? Razina značajnosti testa iznosi 5%.

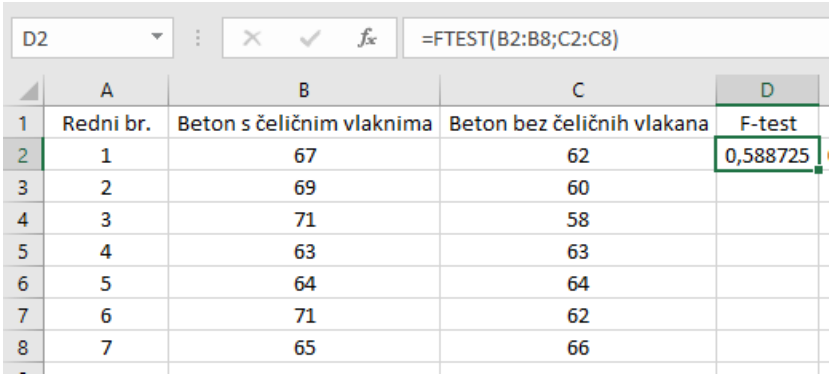
Prvi korak je postavljanje nulte i alternativne hipoteze:

$$H_0: \mu = \mu_0$$
$$H_1: \mu > \mu_0$$

gdje je μ prosječna tlačna čvrstoća betona s čeličnim vlaknima, a μ_0 prosječna tlačna čvrstoća betona bez čeličnih vlakana.

Zatim je potrebno napraviti F-test kako bi se mogao odabrati tip t-testa u daljnjem izračunu. Funkcija u Excelu za F-test također izbacuje p-vrijednost te vrijedi da ako je $p < 0{,}05$), (ili $p < 0{,}01$),) radi se o uzorku s različitim varijancama i odabire se Tip 3. Ako je $p \geq 0{,}05$ (ili $p \geq 0{,}01$), uzorci imaju približno jednaku varijancu te se prilikom izračuna t-testa odabire Tip 2.

U ovom primjeru vrijednost F-testa je 0,5887 što je veće od 0,05 pa će se provoditi Tip 2 t-testa. (Slika 41.)



	A	B	C	D
1	Redni br.	Beton s čeličnim vlaknima	Beton bez čeličnih vlakana	F-test
2	1	67	62	0,588725
3	2	69	60	
4	3	71	58	
5	4	63	63	
6	5	64	64	
7	6	71	62	
8	7	65	66	

Slika 41. Vrijednost F-testa

Sada se može provoditi t-test korištenjem Excelove funkcije. Rezultat i formula prikazani su na Slika 42.

E2 =TTEST(B2:B8;C2:C8;1;2)

	A	B	C	D	E
1	Redni br.	Beton s čeličnim vlaknima	Beton bez čeličnih vlakana	F-test	t-test
2	1	67	62	0,588725	0,004175
3	2	69	60		
4	3	71	58		
5	4	63	63		
6	5	64	64		
7	6	71	62		
8	7	65	66		

Slika 42. T-test

Rezultat p-vrijednosti je 0,0042 što je manje od 0,05 pa se odbacuje nulta hipoteza da su očekivane vrijednosti jednake s mogućnošću pogreške od 5%, odnosno zaključuje se da postoji statistički značajna razlika u tlačnoj čvrstoći betona s čeličnim vlaknima i betona bez takvih vlakana.

Za ručni izračun potrebno je odrediti aritmetičke sredine i standardne devijacije za svaki uzorak. Podaci su prikazani na Slika 43. (vlačne čvrstoće betona s čeličnim vlaknima je uzorak broj 1, a vlačne čvrstoće betona bez čeličnih vlakana je uzorak broj 2).

	A	B	C	D	E	F	G	H	I
1	Redni br.	Beton s čeličnim vlaknima	Beton bez čeličnih vlakana	F-test	t-test	$\bar{x}_1$	$\bar{x}_2$	$x_i^2(1)$	$x_i^2(2)$
2	1	67	62	0,588725	0,004175	67,14286	62,1428571	4489	3844
3	2	69	60					4761	3600
4	3	71	58					5041	3364
5	4	63	63					3969	3969
6	5	64	64					4096	4096
7	6	71	62					5041	3844
8	7	65	66					4225	4356
9							Σ	31622	27073

Slika 43. Izračunati podaci

Vrijednosti standardnih devijacija računaju se prema formuli:

$$s_1 = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{31622 - 7 \cdot 67{,}14286^2}{7-1}} = 3{,}2877$$

(4.3)

$$s_2 = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{27073 - 7 \cdot 62{,}14286^2}{7-1}} = 2{,}6094$$

(4.4)

A vrijednost statistike t:

$$t = \sqrt{n} \cdot \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2} \cdot \sqrt{\frac{s_1^2 + s_2^2}{2}}} = \sqrt{7} \cdot \frac{67{,}143 - 62{,}143}{\sqrt{2} \cdot \sqrt{\frac{3{,}2877^2 + 2{,}6094^2}{2}}} = 3{,}152$$

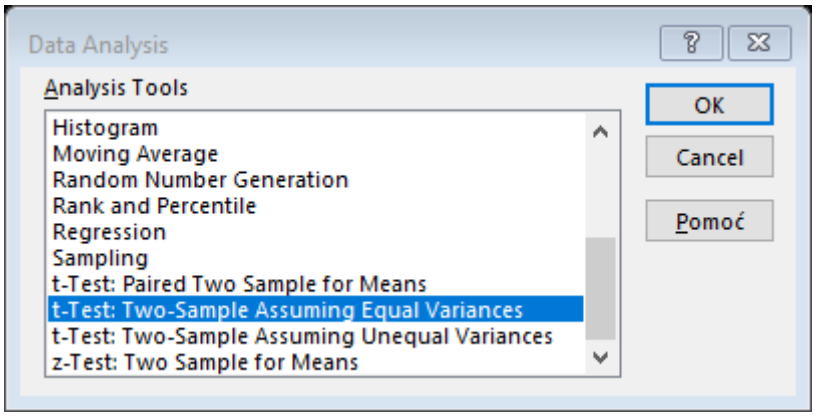
(4.5)

U statističkoj tablici t-raspodjele za $\alpha = 0{,}05$ i $df = 12$ (stupnjevi slobode) očita se $t_\alpha = 0{,}995$. Pošto je $t_\alpha < t$, odbacuje se nulta hipoteza. Na Slika 44. prikazana su područja prihvaćanja i odbacivanja nulte hipoteze za t-test.

Nulta hipoteza	Alternativna hipoteza	Područje prihvaćanja nulte hip.	Područje odbacivanja H_0
$H_0 ... \theta = \theta_0$	$H_1 ... \theta \neq \theta_0$	$-t_{\alpha/2} < t < t_{\alpha/2}$	$t \leq -t_{\alpha/2}$ ili $t \geqslant t_{\alpha/2}$
$H_0 ... \theta \leq \theta_0$	$H_1 ... \theta > \theta_0$	$t < t_\alpha$	$t \geqslant t_\alpha$
$H_0 ... \theta \geqslant \theta_0$	$H_1 ... \theta < \theta_0$	$t > -t_\alpha$	$t \leq t_\alpha$

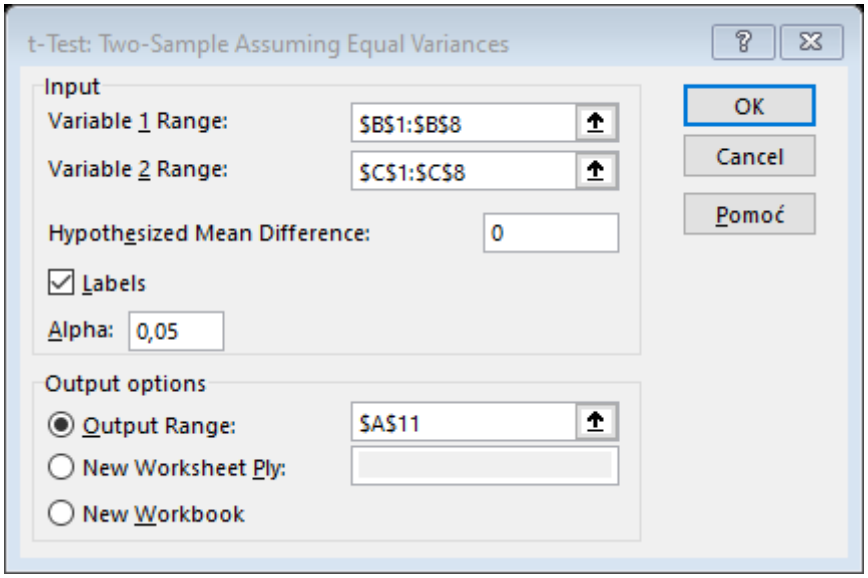
Slika 44. Područje prihvaćanja i odbacivanja nulte hipoteze za t-test

Postoji još jedan način izračuna t-testa u programu MS Excel, a to je putem alata *Analiza podataka*. U prozoru *Data Analysis* odabire se vrsta t-testa koji se želi provesti. Za ovaj primjer odabran je *t-test: Two-Sample Assuming Equal Variances* pošto se radi o uzorcima s približno jednakim varijancama. (Slika 45.)



Slika 45. Data Analysis - t-test

U novom prozoru upisuju se podaci koji su potrebni za izračun. (Slika 46.) Pod *Variable 1 Range* upisuje raspon ćelija u kojima su podaci za prvi uzorak, odnosno vrijednosti tlačnih čvrstoća betona s čeličnim vlaknima, dok se za *Variable 2 Range* unose ćelije s podacima drugog uzroka. Za *Hypothesized Mean Difference* obično se upisuje 0, a *Alpha* je vrijednost razine značajnosti testa koja je u ovom slučaju 0,05. Pritiskom na "OK" dobivaju se podaci prikazani na Slika 47.



Slika 46. Unos podataka za t-test

11	t-Test: Two-Sample Assuming Equal Variances		
12			
13		*Beton s čeličnim vlaknima*	*Beton bez čeličnih vlakana*
14	Mean	66,85714286	62,14285714
15	Variance	8,142857143	5,80952381
16	Observations	7	7
17	Pooled Variance	6,976190476	
18	Hypothesized Mean Difference	0	
19	df	12	
20	t Stat	3,339187124	
21	P(T<=t) one-tail	0,002948482	
22	t Critical one-tail	1,782287556	
23	P(T<=t) two-tail	0,005896965	
24	t Critical two-tail	2,17881283	

Slika 47. Dobivene vrijednosti

Vidljivo je da su dobiveni podaci isti kao ručno izračunati. Pri dnu tablice vidljive su vrijednosti t-testa i p-vrijednosti za jednostrani i dvostrani test. U oba slučaja dolazi se do zaključka da se odbacuje nulta hipoteza.

4.1.3. χ^2- test

Hi kvadrat test je najprimjenjivanija neparametarska metoda metoda testiranja hipoteza. Koristi se u situacijama kada se želi utvrditi da li neke dobivene frekvencije odstupaju od teoretskih, tj. frekvencija koje su očekivane. Očekivane frekvencije su definirane u hipotezi H_0, a ako postoji jedan uzorak, ona se testira preko formule:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - f_{ti})^2}{f_{ti}} \tag{4.6}$$

gdje su: f_i – promatrani broj frekvencija

f_{ti} – očekivani broj frekvencija

U hi-kvadrat testu potrebno je odrediti i stupnjeve slobode *k*, koji se za jedan uzorak računaju kao $k = n - 1$.

X^2 test za dva nezavisna uzorka koristi se kada se želi testirati postoji li značajna razlika između frekvencija ili su one slučajne. Veličina χ^2 se u tom slučaju računa prema formuli:

$$\chi^2 = \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(f_{ij} - ft_{ij})^2}{ft_{ij}} \tag{4.7}$$

Dobivene se frekvencije prvo upišu u tablicu koja ima *r* redaka i *s* stupaca pa su stupnjevi slobode $k = (r - 1) \cdot (s - 1)$.

U programu MS Excel funkcija za izračun hi kvadrat testa je:

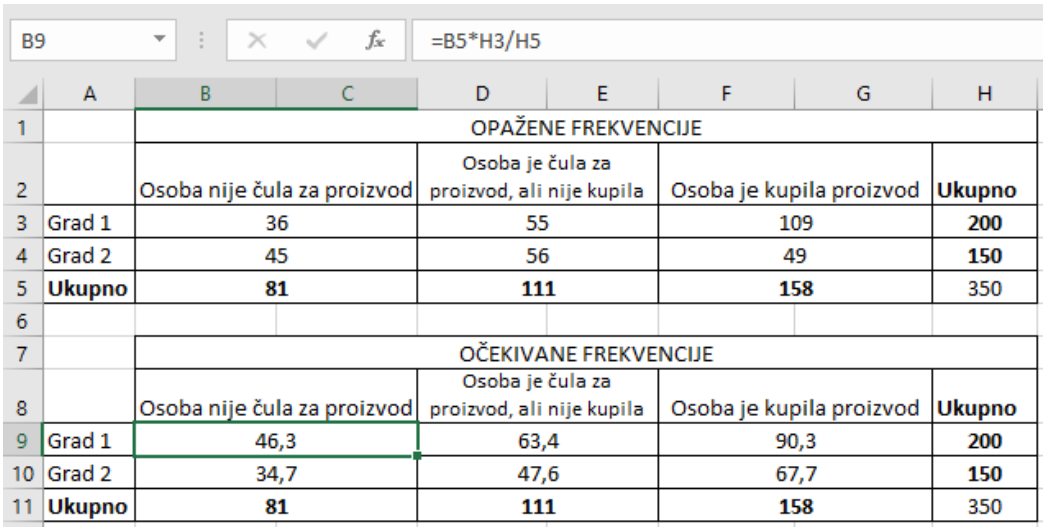
CHITEST(raspon opaženih frekvencija; raspon teoretskih frekvencija)

Primjer 4.3.: Potrebno je odrediti kakva je uspješnost na tržištu novih proizvoda u neka dva izabrana grada. Ispitano je 200 kupaca iz jednog grada i 150 iz drugog grada te su dobiveni podaci kao u Tablica 6. [9]

Tablica 6. Rezultati ispitivanja

	Osoba nije čula za proizvod	Osoba je čula za proizvod, ali nije ga kupila	Osoba je kupila proizvod	Ukupno
Grad 1	36	55	109	200
Grad 2	45	56	49	150
Ukupno	81	111	158	350

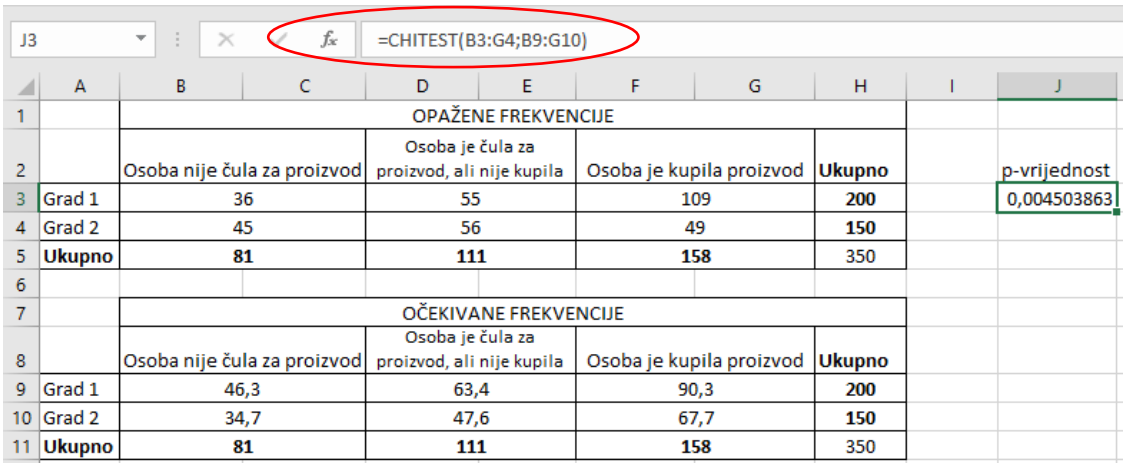
Podaci se unesu u radni list programa MS Excel kako je prikazano na Slika 48. Pritom su napravljene dvije tablice koje se općenito nazivaju tablicama kontigencije. Jedna se odnosi na opažene frekvencije koje su i zadane u samom zadatku, a druga na očekivane frekvencije koje se računaju na način da se suma *s*-tog stupca pomnoži se sa sumom *r*-tog retka i na kraju podijeli s ukupnim brojem frekvencija kao što je i prikazano na Slika 48. za ćeliju B9.



	A	B	C	D	E	F	G	H
1		OPAŽENE FREKVENCIJE						
2		Osoba nije čula za proizvod		Osoba je čula za proizvod, ali nije kupila		Osoba je kupila proizvod		**Ukupno**
3	Grad 1	36		55		109		**200**
4	Grad 2	45		56		49		**150**
5	**Ukupno**	**81**		**111**		**158**		350
6								
7		OČEKIVANE FREKVENCIJE						
8		Osoba nije čula za proizvod		Osoba je čula za proizvod, ali nije kupila		Osoba je kupila proizvod		**Ukupno**
9	Grad 1	46,3		63,4		90,3		**200**
10	Grad 2	34,7		47,6		67,7		**150**
11	**Ukupno**	**81**		**111**		**158**		350

Slika 48. Opažene i očekivane frekvencije

Korištenjem funkcije za χ^2-test, kao rezultat se dobiva vjerojatnost, odnosno p-vrijednost da su eventualne razlike između opaženih i teoretskih frekvencija slučajne. Za konkretan primjer p-vrijednost iznosi 0,0045. (Slika 49.)



	A	B	C	D	E	F	G	H	I	J
1		OPAŽENE FREKVENCIJE								
2		Osoba nije čula za proizvod		Osoba je čula za proizvod, ali nije kupila		Osoba je kupila proizvod		**Ukupno**		p-vrijednost
3	Grad 1	36		55		109		**200**		0,004503863
4	Grad 2	45		56		49		**150**		
5	**Ukupno**	**81**		**111**		**158**		350		
6										
7		OČEKIVANE FREKVENCIJE								
8		Osoba nije čula za proizvod		Osoba je čula za proizvod, ali nije kupila		Osoba je kupila proizvod		**Ukupno**		
9	Grad 1	46,3		63,4		90,3		**200**		
10	Grad 2	34,7		47,6		67,7		**150**		
11	**Ukupno**	**81**		**111**		**158**		350		

Slika 49. Hi kvadrat test

Zaključak se provodi na isti način kao i kod prethodnih testova. U ovom slučaju dobivena p-vrijednost je manja od 0,05 (0,0045<0,05) pa se zaključuje da opažene frekvencije statistički značajno odstupaju od očekivanih frekvencija, odnosno da je uspješnost na tržištu između navedenih gradova različita.

Prema ručnom izračunu, χ^2 iznosi:

$$\chi^2 = \frac{(36-46{,}3)^2}{46{,}3} + \frac{(45-34{,}7)^2}{34{,}7} + \frac{(55-63{,}4)^2}{63{,}4} + \frac{(56-47{,}6)^2}{47{,}6} + \frac{(109-90{,}3)^2}{90{,}3} + \frac{(49-67{,}7)^2}{67{,}7} = 16{,}98$$

a broj stupnjeva slobode:

$$k = (2-1)\cdot(3-1) = 2$$

U statističkoj tablici za hi kvadrat distribuciju, vrijednost χ^2 za ovaj primjer za 2 stupnja slobode uz razinu značajnosti od 5% iznosi 5,99. Pošto je ona manja od izračunate vrijednosti (5,99<16,98), odbacuje se nulta hipoteza.

4.2. Korelacijska i regresijska analiza

Do sad su se u radu razmatrale analize u kojima je predmet bila jedna statistička varijabla, no često je potrebno provoditi istovremenu analizu dviju ili više statističkih varijabli. Te varijable, odnosno pojave, su međusobno povezane, a spoznati njihovu povezanost je svrha upravo korelacijske i regresijske analize.

Veza između dvije varijable može biti ili funkcionalna ili statistička. [1] Kod funkcionalne povezanosti, svakoj vrijednosti jedne varijable odgovara točno određena vrijednost druge varijable, dok kod statističke povezanosti za određenu vrijednost jedne varijable odgovara

više vrijednosti druge varijable. Funkcionalna veza između dvije varijable je oblika $Y = f(x)$, dok se statistička prikazuje u obliku $Y = f(x) + e$ koji predstavlja model jednostavne regresije. *Y* predstavlja zavisnu, a *x* nezavisnu varijablu, dok je s e označena slučajna pogreška.

Korelacijska analiza je prvi korak kojim se želi ispitati stupanj povezanosti između dviju ili više varijabli te ako se isti utvrdi, slijedi regresijska analiza kojom se najbolje opisuje odnos između promatranih varijabli.

4.2.1. Korelacijska analiza

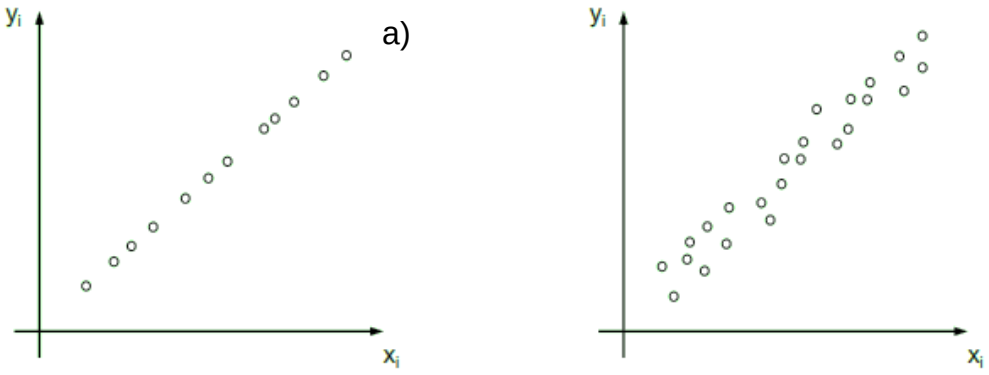
Provođenje korelacijske analize povezanosti dviju varijabli odvija se u tri koraka:

1. Konstrukcija dijagrama rasipanja kojim se grafički prikazuje odnos između varijabli.
2. Računanje koeficijenta korelacije, tj. brojčanog pokazatelja oblika, jakosti i smjera veze među varijablama.
3. Računanje brojčanog pokazatelja statističke značajnosti koeficijenta korelacije, tzv. p-vrijednosti.

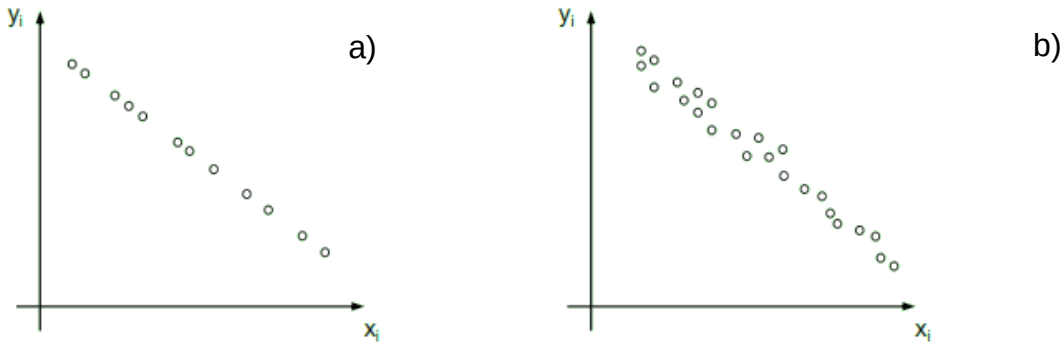
Prilikom spominjanja veze između dvije varijable, razlikuju se nezavisna (X) i zavisna (Y) varijabla koje su prethodno u poglavlju već i spomenute. Kako im i sami nazivi govore, nezavisna varijabla utječe na drugu varijablu, dok je zavisna ona na koju se utječe, tj. njezine vrijednosti ovise o vrijednostima nezavisne varijable. U korelacijskoj analizi nije baš svaki put točno određeno koja je varijabla zavisna, a koja nezavisna.

Dakle, prvi korak u provođenju korelacijske analize je konstrukcija dijagrama rasipanja. U koordinatni sustav se unose vrijednosti varijabli X i Y te se dobivaju točke iz čijeg se rasporeda može približno odrediti postoji li uopće veza između tih varijabli.

Ako se nakon konstrukcije dijagrama utvrdi da postoji veza između promatranih varijabli, može se krenuti na izračun koeficijenta korelacije. Određivanjem smjera veze utvrđuje se da li povećanje vrijednosti jedne varijable uzrokuje povećanje (pozitivna veza) ili smanjenje vrijednosti druge varijable (negativna veza). Slika 50. i Slika 51. prikazuju izgled pozitivne i negativne funkcionalne i statističke veze između dvije varijable. Valja napomenuti da se dijagram rasipanja može crtati isključivo kod povezanosti dviju varijabli, dok se kod povezanosti tri ili više varijabli ovaj korak izostavlja zbog teškog crtanja.



Slika 50. a) pozitivna funkcionalna veza, b) pozitivna statistička veza



Slika 51. a) negativna funkcionalna veza, b) negativna statistička veza

Drugi korak u provođenju korelacijske analize je određivanje koeficijenta korelacije. Označava se s malim slovom *r*, a može poprimiti vrijednosti između -1 i 1. Ako mu je vrijednost pozitivna, to ukazuje na to da rast jedne varijable uzrokuje rast druge i obrnuto. Negativan *r* govori da rast jedne varijable uzrokuje pad druge i obrnuto. Korelacijska analiza zasebno interpretira apsolutnu vrijednosti koeficijenta korelacije (oznaka: $|r|$) koja govori o jakosti veze između varijabli. Što je $|r|$ bliže nuli veza je slabija, a što je bliže jedinici, veza je jača. Mogući kriteriji jakosti veze navedeni su u Tablica 7.

Tablica 7. Jakost veze između varijabli

Vrijednost koeficijenta korelacije	***Jakost veze između varijabli***
1	potpuna
$0.8 \leq \lvert r \rvert < 1$	jaka
$0.5 \leq \lvert r \rvert < 0.8$	srednje jaka
$0.2 \leq \lvert r \rvert < 0.5$	slaba
$0 < \lvert r \rvert < 0.2$	neznatna
0	potpuna odsutnost

U statističkoj analizi najčešće se koristi *Pearsonov koeficijent linearne korelacije* koji se primjenjuje za varijable čije jedinice odgovaraju intervalnoj ili omjernoj skali, a njegova vrijednost se računa prema formuli:

$$r = \frac{\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i\right)^2}}$$

(4.8)

U slučaju kada postoje dvije varijable od kojih je barem jedna redosljedna, za određivanje stupnja linearne veze koristi se *Spearmanov koeficijent korelacije ranga* (oznaka r_s).

Posljednji korak korelacijske analize je procjenjivanje koliko je izračunati koeficijent korelacije statistički značajan. Drugim riječima, tom procjenom želi se utvrditi je li do promjene vrijednosti zavisne varijable zbilja došlo zbog promjene vrijednosti nezavisne varijable ili je do toga došlo slučajno. U tom slučaju, računa se *p-vrijednost* gdje je *p* vjerojatnost da je do promjene zavisne varijable došlo slučajno. Pritom postoje i dva kriterija: prvi ako je $p \geq 0{,}05$, tada se smatra da je procjena izvršena uz razinu pouzdanosti od 95%, a drugi je ako je $p \geq 0{,}01$ kada se smatra da je procjena izvršena uz razinu pouzdanosti od 99%.

P-vrijednost koja odgovara Pearsonovom koeficijentu linearne korelacije računa se pomoću Studentove t-razdiobe.

4.2.2. Regresijska analiza

Nakon korelacijske analize, ono što je potrebno za istraživanje veze između dviju varijabli je regresijska analiza. Njezin temeljni cilj je vezu između promatranih varijabli opisati

pomoću regresijskog modela. Takvim izrazom moguće je objasniti kakva je ovisnost promatranih pojava te procijeniti vrijednosti zavisne varijable Y za određene vrijednosti barem jedne nezavisne varijable X. Ako se u modelu pojavljuje jedna zavisna i jedna nezavisna varijabla, regresija je jednostavna, a ako su u modelu jedna zavisna i barem dvije nezavisne varijable, regresija je višestruka.

Opći oblik jednostavnog regresijskog modela je:

$$Y = f(X) + e$$

gdje je *e* varijabla čiji utjecaji na zavisnu varijablu nisu poznati. Drugim riječima, to je slučajna pogreška kojom su obuhvaćene i varijable koje nisu uključene u postavljeni model, a utječu na zavisnu varijablu.

Postoje dva osnovna cilja koja se žele postići stvaranjem regresijskog modela:

1. Odrediti tip realne funkcije (linearna, kvadratna, eksponencijalna, logaritamska) koja najbolje opisuje vezu između promatranih varijabli.
2. Parametre te funkcije odrediti tako da *e* bude što manji.

Do danas su razvijene razne tehnike regresijske analize poput jednostavne, višestruke, linearne i nelinearne. Najpoznatija je linearna regresija koja za procijenu parametara koristi metodu najmanjih kvadrata.

4.2.2.1. Model jednostavne linearne regresije

Ako se iz dijagrama rasipanja uočava da ravnomjerno povećanje vrijednosti nezavisne varijable X uzrokuje približno ravnomjerno povećanje (ili smanjenje) zavisne varijable Y, regresijski model koji najbolje opisuje navedenu vezu je model jednostavne linearne regresije čiji je opći oblik:

$$Y = \alpha + \beta X + e \qquad (4.9)$$

gdje su $\alpha, \beta \in R$ nepoznati parametri koje je i cilj odrediti tako da *e* bude minimalan. Zapravo, traži se jednadžba pravca koji će najbolje aproksimirati skup točaka dobiven konstrukcijom dijagrama rasipanja. Promatranjem takvog grafičkog prikaza, slučajnu pogrešku *e* predstavlja udaljenost originalnog para točaka varijable X i Y od pravca regresije. Osnovna jednadžba pravca regresije je:

$$y = \alpha + \beta x$$

(4.10)

a njegova procjena je:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

(4.11)

Pritom se *e* računa kao $e_i = y_i - \hat{y}_i$, a naziva se i rezidualno odstupanje.

Metoda kojom se procjenjuju regresijski parametri α i β je metoda najmanjih kvadrata. Njezina osnovna ideja je postići da zbroj kvadrata odstupanja empirijskih vrijednosti zavisne varijable y_i od očekivanih vrijednosti ($\hat{Y}_i$) te varijable bude minimalan. [2]

$$S = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

(4.12)

odnosno:

$$S = \sum_{i=1}^{N} (y_i - (\alpha + \beta x_i))^2$$

(4.13)

Rješavanjem parcijalnih derivacija navedene funkcije i sređivanjem izraza, dobiju se formule za izračunavanje parametara modela jednostavne linearne regresije:

$$\alpha = \bar{y} - \beta \bar{x}$$

(4.14)

$$\beta = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}$$

(4.15)

gdje su $\bar{x}$ i $\bar{y}$ aritmetičke sredine empirijskih vrijednosti varijable x, odnosno varijable y, α predstavlja očekivana vrijednost zavisne varijable kada je nezavisna varijabla jednaka nuli, a β je regresijski koeficijent koji pokazuje prosječnu promjenu zavisne varijable kada se nezavisna varijabla poveća za jedan.

Na taj način, dobiven je i model jednostavne linearne regresije:

$$\hat{y} = \alpha + \beta x$$

(4.16)

Oznaka $\hat{y}$ govori o tome da je riječ o vrijednosti koja je izračunata na temelju regresijskog modela, a ne o vrijednosti dobivenoj statističkim istraživanjem. [2]

Na sljedećem primjeru bit će prikazan i detaljnije objašnjen izračun linearne regresije u programu MS Excel.

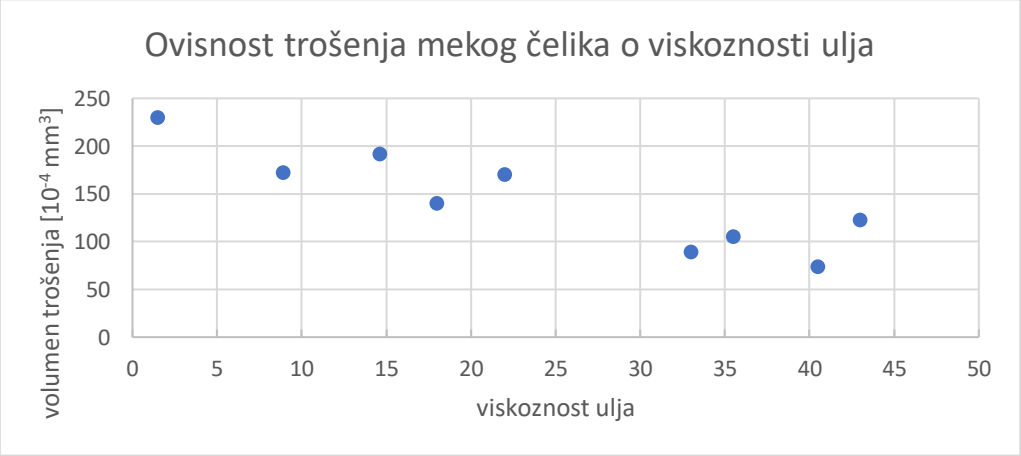
Primjer 4.1..: Prikazani su podaci za viskoznost ulja i trošenje mekog čelika, pri čemu varijabla *x* predstavlja viskoznost ulja, a varijabla *y* volumen trošenja (10^{-4} mm^3). [7]

Tablica 8. Ovisnost trošenja mekog čelika o viskoznosti ulja

y	230	172	192	140	170	105	123	74	89
x	1.5	8.9	14.6	18	22	35.5	43	40.5	33

Nakon što se podaci unesu u radni list programa MS Excel i to redom prema zavisnoj i nezavisnoj varijabli, potrebno ih je prikazati grafički, odnosno nacrtati dijagram rasipanja. To je prvi korak regresijske analize jer se na temelju dijagrama može otprilike procijeniti kolika je povezanost među varijablama.

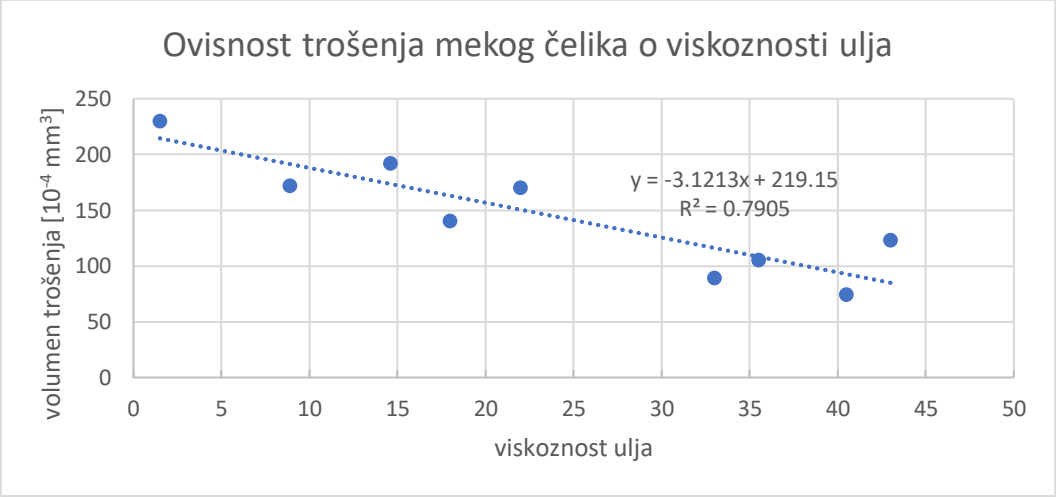
Označe se ćelije sa svim podacima te se u padajućem izborniku *Grafikoni* odabere raspršeni (x,y) dijagram. Nakon sređivanja, dobije se dijagram kao na Slika 52. s devet točaka jer je zadano devet uređenih parova (x_i,y_i).



Slika 52. Raspršeni dijagram

Iz dobivenog dijagrama rasipanja vidljivo je da zamišljeni pravac koji prolazi kroz točke pada, odnosno da je veza između varijabli x i y negativna te će time i regresijski koeficijent β biti negativan. Također, može se koristiti linearni model regresije pošto je vidljivo da smanjenje varijable x prati smanjenje varijable y.

Kako bi se dobila jednadžba linearnog regresijskog modela, sljedeće što je potrebno je izračunati regresijske koeficijente α i β. Za to u MS Excelu postoji više načina, a najjednostavniji je da se u već dobivenom dijagramu doda zamišljeni pravac odabirom opcije *Crta trenda (Add Trendline)* te pod više mogućnosti označi *Prikaži jednadžbu na grafikonu (Display equation on chart)* i *Prikaži R-kvadratnu vrijednost na grafikonu (Display R-squared value on chart)*. Nakon toga, dobije se dijagram kao na Slika 53. na kojemu piše da je jednadžba regresijskog modela $y = -3{,}1213x + 219{,}15$. R^2 je reprezentativnost grafa za koji inače vrijedi da što mu je vrijednost veća, to je graf reprezentativniji (maksimalna vrijednost R^2 je 1).



Slika 53. Dijagram s jednadžbom linearnog regresijskog modela

Drugi način izračuna regresijskih parametara je pomoću funkcija INTERCEPT i SLOPE. Za izračun koeficijenta regresije β koristi se funkcija =SLOPE(raspon varijable Y; raspon varijable X), a za izračun konstantnog člana α =INTERCEPT(raspon varijable Y; raspon varijable X).

U konkretnom primjeru, unošenjem ćelija u kojima se nalaze rasponi varijabli, dobivaju se podaci kao na Slika 54. te je vidljivo da su dobivene vrijednosti za α i β jednake kao i vrijednosti dobivene na dijagramu sa Slika 53. pa se može zapisati da je procijenjena regresijska jednadžba $\hat{y} = -3{,}121x + 219{,}15$.

	A	B
1	x= viskoznost ulja	y=volumen trošenja
2	1,5	230
3	8,9	172
4	14,6	192
5	18	140
6	22	170
7	35,5	105
8	43	123
9	40,5	74
10	33	89
11		
12	α	219,1477387
13	β	-3,121334783

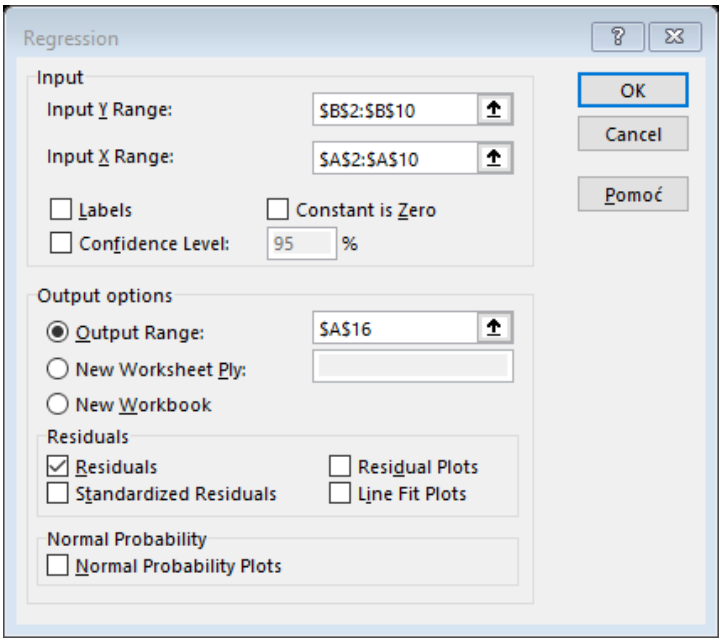
Slika 54. Parametri jednadžbe regresijskog modela

Konkretno značenje dobivenih parametara je sljedeće: konstantan član α=219,15 govori o tome da kada bi viskoznost ulja bila 0, u tom slučaju bi očekivani volumen trošenja mekog čelika bio 219,15 mm^3. Regresijski koeficijent β= -3,121 pokazuje da se za svako smanjenje volumena trošenja mekog čelika za 10^{-4} mm^3 , viskoznost ulja smanji za 3,121.

Uvrštavanjem stvarnih vrijednosti nezavisne varijable *x* u procijenjenu regresijsku jednadžbu, dobiju se regresijske vrijednosti zavisne varijable *y*. Npr. za x=1,5 regresijska vrijednost je $\hat{y} = 219{,}15 - 3{,}121 \cdot 1{,}5 = 214{,}4685$, a interpretira se na način da je za viskoznost ulja od 1,5 očekivan volumen trošenja mekog čelika 214,4685 mm^3. Na isti način se računaju i interpretiraju ostale regresijske vrijednosti.

Što se tiče rezidualnih odstupanja, već je prethodno navedeno da ona predstavljaju razliku stvarnih vrijednosti zavisne varijable *y* od procijenjene $\hat{y}$. U navedenom primjeru za vrijednost y=230, njezina regresijska vrijednost je $\hat{y} = 214{,}47$ te je pritom rezidualno odstupanje $e = 230 - 214{,}47 = 15{,}53$. Analogno se dobivaju i ostala rezidualna odstupanja.

Sva ta i ostala rješenja regresijskog modela, u programu MS Excel mogu se dobiti u "jednom koraku" pomoću funkcije *Analiza podataka (Data Analysis)*. U navedenom okviru odabere se opcija *Regression* te se u dobivenom prozoru unesu potrebni podaci (Slika 55.)



Slika 55. Data Analysis – Regression

Na radnom listu dobiju se podaci prikazani na Slika 56.

16	SUMMARY OUTPUT								
17									
18	*Regression Statistics*								
19	Multiple R	0,93776177							
20	R Square	0,879397137							
21	Adjusted R Square	0,862168156							
22	Standard Error	19,95696437							
23	Observations	9							
24									
25	ANOVA								
26		*df*	*SS*	*MS*	*F*	*Significance F*			
27	Regression	1	20328,9259	20328,9259	51,04173975	0,000186278			
28	Residual	7	2787,962989	398,280427					
29	Total	8	23116,88889						
30									
31		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
32	Intercept	234,0707398	13,74839579	17,02531287	5,91371E-07	201,5609497	266,5805299	201,5609497	266,5805299
33	X Variable 1	-3,508556273	0,491095226	-7,1443502	0,000186278	-4,669811953	-2,34730059	-4,66981195	-2,34730059
34									
35									
36									
37	RESIDUAL OUTPUT								
38									
39	*Observation*	*Predicted Y*	*Residuals*						
40	1	228,4570498	11,54295024						
41	2	201,0903108	-20,09031083						
42	3	179,6881176	13,31188243						
43	4	163,8996143	-8,899614339						
44	5	156,8825018	15,11749821						
45	6	109,5169921	0,48300789						
46	7	83,20282006	29,79717994						
47	8	91,97421075	-16,97421075						
48	9	118,2883828	-24,28838279						

Slika 56. Dobivena rješenja regresijskog modela

U dijelu *Regression Statistics (Regresijska statistika), Multiple R* je koeficijent korelacije kojim se mjeri snaga linearnog odnosa između dvije varijable. To je, u ovom slučaju, pozitivna vrijednost koeficijenta korelacije *r* dok se njegov predznak određuje na temelju predznaka regresijskog koeficijenta uz varijablu X ($r = -0{,}9378$). *R square* je vrijednost R^2 koji je već prije u poglavlju spomenut, a govori o reprezentativnosti grafa. Naziva se još i koeficijentom determinacije te pripada relativnim mjerama reprezentativnosti modela jednostavne linearne regresije. Općenito, za provjeru valjanosti modela koriste se mjere reprezentativnosti koje se dijele na apsolutne i relativne. U apsolutne pripadaju varijanca i standardna devijacija regresijskog modela, a u relativne koeficijent varijacije i koeficijent determinacije regresijskog modela. R^2 u ovom primjeru iznosi 0,8794 odnosno 87,94% što govori o dobroj reprezentativnosti jer je njegova maksimalna vrijednost 1 te vrijedi da što je vrijednost koeficijenta bliže 1, to je reprezentativnost bolja. Može se nalaziti u intervalu $0 \leq R^2 \leq 1$. U MS Excelu se može izračunati i korištenjem funkcije =RSQ(raspon varijable Y; raspon varijable X) ili uvrštavanjem u formulu:

$$R^2 = \frac{SP}{ST} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(4.17)

gdje je SP protumačeni dio sume kvadrata odstupanja, odnosno ukupan zbroj kvadrata razlike procijenjene vrijednosti zavisne varijable Y od aritmetičke sredine te vrijednosti, a ST je ukupan zbroj kvadrata razlike empirijske vrijednosti zavisne varijable Y od aritmetičke sredine te vrijednosti.

Adjusted R square je korigirani koeficijent determinacije koji se primjenjuje u slučaju kad je mali broj uzoraka ($n \leq 30$). Vrijednost mu može biti manja ili jednaka od koeficijenta determinacije, ali i manja od nule što nije poželjno. Računa se prema formuli:

$$\bar{R}^2 = 1 - \frac{n-1}{n-2} \cdot (1 - R^2)$$

(4.18)

Standard Error (Standardna pogreška) je zapravo standardna devijacija regresijskog modela koja je apsolutni pokazatelj reprezentativnosti regresijskog modela. Može se izračunati pomoću funkcije =STEYX(raspon varijable Y; raspon varijable X) ili uvrštavanjem u formulu:

$$\widehat{\sigma_y} = \sqrt{\frac{SR}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}$$

(4.19)

gdje je SR zbroj kvadrata rezidualnih odstupanja, odnosno zbroj kvadrata procijenjenih od odgovarajućih empirijskih vrijednosti, a može se i reći da ovaj član pokazuje ukupnu pogrešku svih *n* procjenjivanja. U ovom primjeru standardna pogreška iznosi 19,96, a izražena je u jedinicama mjere zavisne varijable Y.

Obervations prikazuje ukupan broj promatranja kojih je u ovom primjeru 9.

Unutar tablice *ANOVA* nalazi se analiza varijance, tj. dan je prikaz o razini varijabilnosti regresijskog modela. U prvom stupcu pod *df* prikazani su stupnjevi slobode. *Df/Regression* govori o broju nezavisnih varijabli u regresijskom modelu kojih je u ovom primjeru jedna. *Df/Residual* je razlika ukupnog broja promatranja (9) i broja procjenjivanih varijabli (2), dakle u ovom slučaju 7. Podatak *SS/Regression* je zapravo protumačeni dio ukupne sume kvadrata odstupanja SP, *SS/Residual* je rezidualni (neprotumačeni) dio ukupne sume kvadrata odstupanja SR, a *SS/Total* je ukupna suma kvadrata odstupanja ST. Podatak *MS/Regression* je isti kao *SS/Regression*, dok se *MS/Residual* računa kao omjer vrijednosti SR i preostalih stupnjeva slobode (*df/Residual*) te predstavlja srednju kvadratnu pogrešku. Iz stupca *F* može se očitati vrijednosti empirijskog F-omjera (51,04) te njegova *p*-vrijednost (*Significance F*) koja iznosi 0,00018 i predstavlja statističku značajnost povezanosti između dviju promatranih varijabli. U slučaju jednostavne linearne regresije, ta vrijednost jednaka je p-vrijednosti nezavisne varijable. F-omjer se može izračunati i preko formule:

$$F = \frac{SP}{\frac{SR}{n-2}}$$

(4.20)

Tablica s koeficijentima je zapravo najvažniji dio za izradu jednadžbe regresijskog modela. U prvom stupcu pod *Coefficients* se nalaze vrijednosti parametara α (*Intercept*) i β (*X Variable*) koje su iste kao vrijednosti dobivene računanjem preko funkcija INTERCEPT i SLOPE i preko formula. Stupci *t Stat* i *P-value* odnose se na procjenu kojom se želi utvrditi je li do promjene zavisne varijable došlo slučajno stjecanjem drugih okolnosti. Vidljivo je da je p-vrijednost uz nezavisnu varijablu jednaka p-vrijednosti od F

(*Significance F*) te za nju općenito vrijedi da ako je manja od 0,05, znači da je statistički značajna varijabla u modelu.

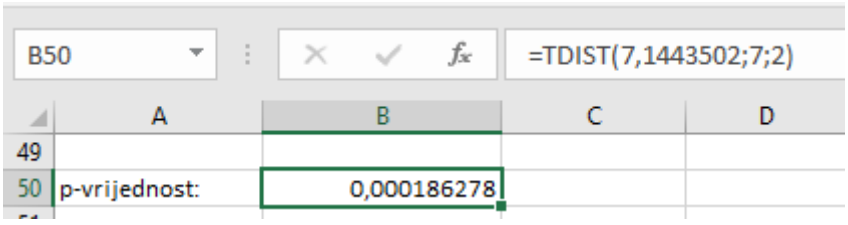
Pošto se radi o malom uzorku od n=9 promatranja, *p*-vrijednost koja odgovara Pearsonovom koeficijentu jednostavne linearne korelacije računa se koristeći Studentovu t-razdiobu. Prvi korak je izračun Pearsonovog koeficijenta jednostavne linearne korelacije *r* prema formuli (4.8) ili preko relacije $r^2 = R^2$ ili pak korištenjem Excelove funkcije =CORREL. Zatim slijedi izračun broja $d = n - 2$, gdje je *n* broj promatranja, a *d* je zapravo broj stupnjeva slobode. Na kraju se izračuna *t*-vrijednost prema formuli:

$$t = r \cdot \sqrt{\frac{d}{1 - r^2}} \tag{4.21}$$

U MS Excelu funkcija pomoću koje se može izračunati *p*-vrijednost je TDIST. Za njezin izračun potrebno je znati t-vrijednost, stupnjeve slobode i krakove, odnosno radi li se o jednokrakoj ili dvokrakoj distribuciji. T-vrijednost izračunata je u tablici s koeficijentima (*t Stat*) i za nezavisnu varijablu iznosi -7,1443502, ali se za TDIST unosi njezina apsolutna vrijednost. Također, može se izračunati i ručno preko formule:

$$t = 0{,}93776 \cdot \sqrt{\frac{7}{1 - 0{,}879397}} = 7{,}14435$$

Broj stupnjeva slobode u ovom slučaju je 7, izračunato preko formule $d = n - 2$, a pošto se radi o dvokrakoj distribuciji, na kraju je upisan broj 2. Izračun u Excelu prikazan je na Slika 57.



Slika 57. Izračun =TDIST

Pošto je u ovom primjeru p=0,00018, može se reći da postoji statistički značajna linearna veza između viskoznosti ulja i volumena trošenja mekog čelika.

Lower 95% i *Upper 95%* prikazuju donju, odnosno gornju granicu za interval povjerenja.

U dijelu *Residual output* nalaze se regresijske vrijednosti zavisne varijable $\hat{y}$ za svako promatranje (*Predicted Y*) kao i rezidualna odstupanja za svakog od njih (*Residuals*) čiji je izračun bio objašnjen prethodno u radu.

5. ZAKLJUČAK

Na temelju napisanog diplomskog rada, zaključuje se da se korištenjem programa MS Excel može brže i jednostavnije odraditi statistička analiza podataka. To je program koji ima širok skup statističkih funkcija, ali i alat za obradu podataka kojim se dobiva niz osnovnih statističkih pokazatelja u samo jednom kliku.

U području deskriptivne statistike brže je i jednostavnije koristiti statističke funkcije koje daje MS Excel prilikom izračuna mjera centralne tendencije kao i mjera disperzije, nego uvrštavati podatke u formulu, pogotovo ako se radi o velikom broju podataka.

U četvrtom poglavlju koje se odnosi na inferencijalnu statistiku objašnjeno je kako pomoću statističkih testova donijeti zaključak što je također jednostavnije prikazati u Excelu.

Najjednostavnije je koristiti alat "Analiza podataka" pomoću kojega se, nakon unošenja potrebnih informacija, u jednom kliku izračunaju najvažniji statistički elementi.

Prikazanim primjerima, vidi se da poznavanje alata i funkcija koje pruža MS Excel za metode statistike, može uvelike pridonijeti kvaliteti i produktivnosti u mnogim područjima pa tako i u strojarstvu.

LITERATURA

[1] Papić, M.: *Primjenjena statistika u Excelu*, Zoro d.o.o., Zagreb-Sarajevo, 2008.

[2] Kovačić, B.: *Poslovna statistika*, interna skripta

[3] Gogala, Z.: *Osnove statistike*, Sinergija, Zagreb, 2001.

[4] Šošić, I.: *Primijenjena statistika,* Školska knjiga, Zagreb, 2004.

[5] *Statistički ljetopis Republike Hrvatske*, Zagreb, 2018.

[6] Statistička izvješća: *Industrijska proizvodnja u 2018.,* Zagreb, 2020.

[7] Montgomery D.C., Runger G.C.: Applied Statistics and Probability for Engineers, John Wiley & Sons, 2003.

[8] https://www.grad.unizg.hr/_download/repository/Testiranje_hipoteza.pdf - pristupljeno 25.5.2022.

[9] Dujman, J: *Uporaba neparametarske statistike u inženjerskim problemima*, završni rad, Zagreb, 2017.

PRILOZI

Statistička tablica standardne normalne raspodjele:

	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Statistička tablica t-raspodjele:

		Significance level					
Degrees of freedom	*Two-tailed test:* / *One-tailed test:*	10% / 5%	5% / 2.5%	2% / 1%	1% / 0.5%	0.2% / 0.1%	0.1% / 0.05%
1		6.314	12.706	31.821	63.657	318.309	636.619
2		2.920	4.303	6.965	9.925	22.327	31.599
3		2.353	3.182	4.541	5.841	10.215	12.924
4		2.132	2.776	3.747	4.604	7.173	8.610
5		2.015	2.571	3.365	4.032	5.893	6.869
6		1.943	2.447	3.143	3.707	5.208	5.959
7		1.894	2.365	2.998	3.499	4.785	5.408
8		1.860	2.306	2.896	3.355	4.501	5.041
9		1.833	2.262	2.821	3.250	4.297	4.781
10		1.812	2.228	2.764	3.169	4.144	4.587
11		1.796	2.201	2.718	3.106	4.025	4.437
12		1.782	2.179	2.681	3.055	3.930	4.318
13		1.771	2.160	2.650	3.012	3.852	4.221
14		1.761	2.145	2.624	2.977	3.787	4.140
15		1.753	2.131	2.602	2.947	3.733	4.073
16		1.746	2.120	2.583	2.921	3.686	4.015
17		1.740	2.110	2.567	2.898	3.646	3.965
18		1.734	2.101	2.552	2.878	3.610	3.922
19		1.729	2.093	2.539	2.861	3.579	3.883
20		1.725	2.086	2.528	2.845	3.552	3.850
21		1.721	2.080	2.518	2.831	3.527	3.819
22		1.717	2.074	2.508	2.819	3.505	3.792
23		1.714	2.069	2.500	2.807	3.485	3.768
24		1.711	2.064	2.492	2.797	3.467	3.745
25		1.708	2.060	2.485	2.787	3.450	3.725
26		1.706	2.056	2.479	2.779	3.435	3.707
27		1.703	2.052	2.473	2.771	3.421	3.690
28		1.701	2.048	2.467	2.763	3.408	3.674
29		1.699	2.045	2.462	2.756	3.396	3.659
30		1.697	2.042	2.457	2.750	3.385	3.646
32		1.694	2.037	2.449	2.738	3.365	3.622
34		1.691	2.032	2.441	2.728	3.348	3.601
36		1.688	2.028	2.434	2.719	3.333	3.582
38		1.686	2.024	2.429	2.712	3.319	3.566
40		1.684	2.021	2.423	2.704	3.307	3.551
42		1.682	2.018	2.418	2.698	3.296	3.538
44		1.680	2.015	2.414	2.692	3.286	3.526
46		1.679	2.013	2.410	2.687	3.277	3.515
48		1.677	2.011	2.407	2.682	3.269	3.505
50		1.676	2.009	2.403	2.678	3.261	3.496
60		1.671	2.000	2.390	2.660	3.232	3.460
70		1.667	1.994	2.381	2.648	3.211	3.435
80		1.664	1.990	2.374	2.639	3.195	3.416
90		1.662	1.987	2.368	2.632	3.183	3.402
100		1.660	1.984	2.364	2.626	3.174	3.390
120		1.658	1.980	2.358	2.617	3.160	3.373
150		1.655	1.976	2.351	2.609	3.145	3.357
200		1.653	1.972	2.345	2.601	3.131	3.340
300		1.650	1.968	2.339	2.592	3.118	3.323
400		1.649	1.966	2.336	2.588	3.111	3.315
500		1.648	1.965	2.334	2.586	3.107	3.310
600		1.647	1.964	2.333	2.584	3.104	3.307
∞		1.645	1.960	2.326	2.576	3.090	3.291

Statistička tablica χ^2 distribucije:

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38