Statističke funkcije MS Excela

Topol, Matea

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Preddiplomski sveučilišni studij strojarstva

Završni rad

STATISTIČKE FUNKCIJE MS EXCELA

Rijeka, srpanj 2022.

Matea Topol

0069085195

SVEUČILIŠTE U RIJECI

TEHNIČKI FAKULTET

Preddiplomski sveučilišni studij strojarstva

Završni rad

STATISTIČKE FUNKCIJE MS EXCELA

Mentor: doc. dr. sc. Ivan Dražić

Rijeka, srpanj 2022.

Matea Topol

0069085195

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
POVJERENSTVO ZA ZAVRŠNE ISPITE

Rijeka, 15. ožujka 2022.

Zavod: **Zavod za matematiku, fiziku i strane jezike**
Predmet: **Inženjerska statistika**
Grana: **1.01.08 teorija vjerojatnosti i statistika**

ZADATAK ZA ZAVRŠNI RAD

Pristupnik: **Matea Topol (0069085195)**
Studij: Preddiplomski sveučilišni studij strojarstva

Zadatak: **Statističke funkcije MS Excela // Statistical functions in MS Excel**

Opis zadatka:

U radu je potrebno dati kratki uvod u softverski paket MS Excel te detaljno opisati biblioteku funkcija koje se koristi kod statističke analize. Potrebno je opisati temeljne tehnike deskriptivne statističke analize u vidu formiranja tablice frekvencija, izračuna statističkih pokazatelja te izrade različitih grafičkih prikaza, kao i temeljne tehnike inferencijalne statistike u vidu testiranja statističkih hipoteza te korelacijske i regresijske analize. Svaku od spomenutih tehnika treba potkrijepiti primjerom iz inženjerske struke, s naglaskom na primjenu u strojarstvu. Sve navedene primjere potrebno je riješiti primjenom adekvatnih funkcija MS Excela.

Rad mora biti napisan prema Uputama za pisanje diplomskih / završnih radova koje su objavljene na mrežnim stranicama studija.

Matea Topol

Zadatak uručen pristupniku: 21. ožujka 2022.

Mentor:

Doc. dr. sc. Ivan Dražić

Predsjednik povjerenstva za završni ispit:

Prof. dr. sc. Kristian Lenić

IZJAVA

Ja, Matea Topol, izjavljujem da sam prema članku 8. Pravilnika o završnom radu, završnom ispitu i završetku studija, samostalno izradila završni rad naslova „Statističke funkcije MS Excela“ pod vodstvom doc. dr. sc. Ivana Dražića.

ZAHVALA

Zahvaljujem se mentoru doc. dr. sc. Ivanu Dražiću na uputama, vodstvu i uloženom vremenu tijekom izrade ovog završnog rada. Jednako se zahvaljujem doc. dr. sc. Loredani Simčić uz čiju pomoć sam se odlučila za ovu temu završnog rada.

Posebno se zahvaljujem svojoj obitelji i prijateljima koji su mi bili velika podrška tijekom cijelog studija i bez kojih sve ovo ne bi bilo moguće.

SADRŽAJ

1. UVOD 1
1.1. Što je statistika? 1
1.2. Primjena statistike 1
1.3. Podjela statistike 1
1.4. Statistički softveri 2
2. DESKRIPTIVNA STATISTIKA U MS EXCELU 3
2.1. Određivanje numeričkih karakteristika statističkog skupa 6
2.1.1. Pokazatelji centralne tendencije 6
2.1.1.1. Aritmetička sredina 6
2.1.1.2. Mod 7
2.1.1.3. Medijan 8
2.1.2. Pokazatelji rasapa podataka u statističkom skupu 9
2.1.2.1. Raspon podataka 10
2.1.2.2. Interkvartilni raspon, koeficijent kvartilne devijacije 11
2.1.2.3. Varijanca, standardna devijacija, koeficijent varijacije 12
2.1.3. Pokazatelji oblika 14
2.1.3.1. Koeficijent asimetrije 14
2.1.3.2. Koeficijent spljoštenosti 15
3. SLUČAJNE VARIJABLE 17
3.1. Diskretne slučajne varijable 17
3.1.1. Poissonova razdioba 17
3.1.2. Binomna razdioba 19
3.1.3. Hipergeometrijska razdioba 21
3.2. Kontinuirane slučajne varijable 22
3.2.1. Normalna razdioba 23
3.2.1.1. Interval povjerenja (Normalna razdioba) 26

3.2.1.2. Interval povjerenja (Studentova „t“-razdioba) 28
4. TESTIRANJE STATISTIČKIH HIPOTEZA 30
4.1. T-test 30
4.2. F-test 32
4.3. Hi-kvadrat test 34
5. KORELACIJA I REGRESIJA 37
6. ZAKLJUČAK 43
LITERATURA 44
POPIS STATISTIČKIH FUNKCIJA 45
POPIS SLIKA 46
POPIS TABLICA 48
POPIS PRIMJERA 49
SAŽETAK 50
ABSTRACT 51

1. UVOD

1.1. Što je statistika?

Statistika je znanstvena disciplina i grana matematike koja se bavi prikupljanjem, analiziranjem, opisivanjem i proučavanjem skupa podataka. Analizom prikupljenih statističkih podataka te njihovom usporedbom možemo pratiti kako se određene situacije mijenjaju te tako dobiti rezultat. Rezultat se može očitovati kao numerička vrijednost ili kao pretpostavka o daljnjem razvoju određene situacije. Temeljem rezultata mogu se sigurnije ponuditi rješenja zadanog problema te donositi odluke. Upravo radi toga, statistika se može primijeniti u raznim prirodnim i društvenim znanostima, poput ekonomije, zdravstvenih znanosti, itd.

1.2. Primjena statistike

Značajnost statistike za čovjeka iskazuje se i u tome da se ona primjenjuje već više od 2000 godina. Prvi puta se pojavila na području drevne Mezopotamije gdje su je ljudi upotrebljavali prilikom popisivanja stanovništva, poljoprivrednih prinosa te materijalnog bogatstva. Daljnjim razvojem stanovništva u pogledu obrazovanja, došlo je i do napredovanja u svim područjima pa je tako statistika kojom se mi primjenjujemo danas puno složenija i opširnija i usporedbi sa statistikom koja se koristila u prošlosti. Kao rezultat razvoja statistike, danas imamo mogućnost donošenja puno preciznijih zaključaka, zbog veće dostupnosti podataka te možemo na temelju složenijih izračuna bolje predvidjeti razvoj neke skupine elemenata.

1.3. Podjela statistike

Statistiku možemo podijeliti na deskriptivnu statistiku i inferencijalnu statistiku.

Deskriptivna statistika bavi se organiziranjem prikupljenih podataka, odnosno razvrstavanjem, uređivanjem i određivanjem numeričkih pokazatelja u statističkom skupu.

Inferencijalna statistika bavi se statističkim postupcima koji nam omogućuju testiranje istraživačkih hipoteza, odnosno zaključivanje. Na temelju podataka koje smo dobili na

pojedinačnim slučajevima (uzorku) donosimo zaključke o cijeloj populaciji. Nemoguća je sto postotna sigurnost prilikom statističkog zaključivanja, ali je cilj što više se približiti navedenom postotku.

1.4. Statistički softveri

Danas postoje različiti softveri kojima se statističari služe prilikom vršenja svojih izračuna. Prema namjeni softvere možemo podijeliti na statističke pakete, elektronske tabele i biblioteke koda.

Namjena statističkih paketa jest pružanje pomoći korisniku pri obradi statističkih podataka. Oni izvršavaju sve ili veći dio statističkih funkcija, a metode koje se koriste pri proizvodnji outputa su precizne. Najpoznatiji statistički paketi na tržištu su R, SPSS, Gephi, Epilnfo, itd.

U skupinu elektronskih tabela, čija osnovna namjena nije statistička obrada podataka, spada i MS Excel čije su statističke funkcije glavna tema ovoga rada. Iako statistička obrada nije temelj takvih programa, oni sadrže skup statističkih funkcija koje se korištenjem skript jezika mogu proširiti te približiti funkcionalnostima koje pružaju statistički paketi. Elektronske tabele, uz MS Excel, su još i OpenOffice Calc te LibreOffice Calc.

Biblioteke koda u programskim jezicima C, C++, Fortan su implementacija statističkog algoritma u programskom jeziku. Kako bi prešle u program, na njih se treba nadograditi korisnički interfejs koji će omogućiti korisniku da u interakciji s implementiranim algoritmom kreira program koji će se moći izvršavati u određenom softverskom i hardverskom okruženju.

2. DESKRIPTIVNA STATISTIKA U MS EXCELU

Kao što smo već ranije spomenuli u uvodnom dijelu rada, deskriptivna statistika bavi se uređivanjem i razvrstavanjem podataka te određivanjem numeričkih pokazatelja za dani skup podataka. Dakle, nakon dobivanja empirijskih podataka izvođenjem statističkih pokusa, ti podaci se dalje moraju grupirati, odnosno svrstati u razrede i tabelirati. Statistički podaci mogu poprimiti numeričke vrijednosti pa tada govorimo o kvantitativnim obilježjima, a mogu poprimiti i opisne vrijednosti pa takva obilježja nazivamo atributivna. Kvantitativna (numerički opisana) obilježja nadalje možemo podijeliti na diskretna i kontinuirana obilježja.

Primjer 2.1. Diskretno statističko obilježje

Bacajući igraću kockicu trideset puta za redom, dobiveni su sljedeći podaci:

1 3 5 3 1 4 6 2 4 3 5 1 1 4 6 2 4 2 6 5 3 1 4 2 4 1 5 6 6 2

Primjer 2.2. Kontinuirano statističko obilježje

Prateći temperaturu zraka (u °C) u 12:00 sati tijekom razdoblja od 30 dana, dobiveni su sljedeći podaci:

12.31 12.26 13.17 11.92 11.79 10.92 12.63 13.17 12.42 13.68

11.52 11. 96 10.98 10.79 10.54 9.82 9.18 10.14 11.47 11.83

11.92 12.55 11.77 10.70 10.35 10.12 9.84 9.95 10.32 10.81

Razlika između diskretnih i kontinuiranih statističkih obilježja jest u tome što diskretne varijable mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti dok je skup mogućih vrijednosti kontinuiranih varijabli cijeli skup realnih brojeva ili neki interval u skupu realnih brojeva.

Na temelju navedenih primjera možemo urediti i razvrstati podatke koristeći se programom MS Excel te njegovim funkcijama iz područja statistike. Radi jednostavnosti, funkcije ćemo prikazati koristeći se primjerom 1, tj. diskretnim statističkim obilježjem iako se sve navedene funkcije mogu koristiti i pri razvrstavanju kontinuiranih statističkih obilježja. Prva stvar koju ćemo napraviti prije korištenja bilo koje funkcije jest tabeliranje podataka pri čemu će se u stupcu X_i

nalaziti sve moguće vrijednosti obilježja. Prema navedenom primjeru (2.1.), naša tablica će izgledati ovako:

X_i
1
2
3
4
5
6

Slika 2.1. Tabeliranje podataka iz primjera 2.1.

Frekvencija nekog podatka jest broj pojavljivanja tog podatka u nekom skupu podataka.

Primjer 2.3. Frekvencija

Odrediti frekvenciju brojeva 0, 1, 2, 3, 4 i 7 u skupu podataka: 777332330442312.

Iako bismo u zadanom primjeru sami mogli izbrojati koliko se puta pojavljuju zadani brojevi, u praksi se najčešće susrećemo s puno većim skupovima podataka gdje bi određivanje frekvencije na takav način iziskivalo mnogo vremena. U tu svrhu, može se koristiti funkcija Frequency u MS Excelu.

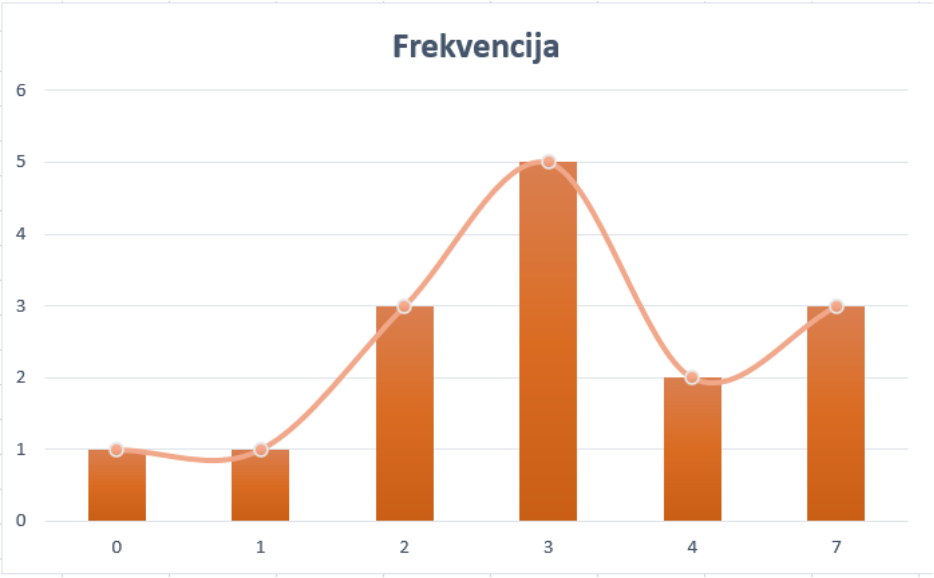
FUNKCIJA FREQUENCY

Da bismo izračunali frekvenciju zadanih brojeva, odabiremo ćelije u rasponu od D25 do D30 (stupac f_i) u kojima bismo htjeli da nam se prikaže podatak o frekvenciji brojeva koji se nalaze u stupcu X_i. Zatim ispisujemo **=FREQUENCY(*x:y;m:n*)**. Raspon *x:y* predstavlja raspon liste i obuhvaća sve zadane podatke dok raspon m:n obuhvaća podatke čija nas frekvencija zanima. U našem slučaju, raspon *x:y* obuhvaća podatke od ćelije A24 do ćelije A38, a raspon *m:n* obuhvaća podatke od ćelije C25 do ćelije C30. Kako bi nam se prikazali traženi rezultati, funkciju pokrećemo pritiskom na CTRL + SHIFT + ENTER pošto je izlazna vrijednost ove funkcije tabelarna.

	A	B	C	D	E	F	G	H
23								
24	7		X_i	f_i				
25	7		0	1				
26	7		1	1				
27	3		2	3				
28	3		3	5				
29	2		4	2				
30	3		7	3				
31	3							
32	0							
33	4		FREQUENCY: =FREQUENCY(A24:A38;C25:C30)					
34	4							
35	2							
36	3							
37	1							
38	2							

Slika 2.2. Funkcija FREQUENCY

Frekvencija ili učestalost pojavljivanja broja 0 jest 1. Frekvencija broja 1 jest također 1. Vrijednost 2 pojavljuje se 3 puta u zadanoj listi podataka, itd. MS Excel nam također omogućava prikazivanje zadanih vrijednosti i njihovih pripadajućih frekvencija pomoću grafa. Slika 2.2. prikazuje graf frekvencija vezan za primjer 2.3. Narančastim stupcima prikazane su zadane vrijednosti dok su narančastom linijom povezane njihove frekvencije. Iz grafa se jasno može iščitati frekvencija svake vrijednosti označena narančastim markerom.



Slika 2.3. Graf funkcije FREQUENCY

Također, moguće je još i izračunati relativnu frekvenciju koja je jednaka omjeru frekvencije i ukupnog broja podataka, kumulativnu frekvenciju pri čemu je prva kumulativna frekvencija jednaka aktualnoj frekvenciji f_i, dok je svaka sljedeća kumulativna frekvencija jednaka zbroju prethodne kumulativne i aktualne frekvencije (f_i) te relativno kumulativnu frekvenciju koju dobijemo kao omjer kumulativne frekvencije i ukupnog broja podataka.

2.1. Određivanje numeričkih karakteristika statističkog skupa

Numeričke karakteristike statističkog skupa koristimo kako bismo na sažeti način prikazali karakteristike promatranog skupa podataka. U skupinu numeričkih karakteristika statističkog skupa spadaju pokazatelji centralne tendencije, pokazatelji rasapa i pokazatelji oblika.

2.1.1. Pokazatelji centralne tendencije

Pokazatelji centralne tendencije koriste se u slučajevima kada je zadan veliki broj podataka koji imaju tendenciju grupiranja oko neke srednje vrijednosti. Srednja vrijednost jest konstantna vrijednost koja služi kao predstavnik niza velikog broja podataka.

2.1.1.1. Aritmetička sredina

Aritmetička sredina predstavlja prosjek ukupnog skupa podataka. Ona je jednaka omjeru zbroja svih podataka i ukupnog broja podataka.

FUNKCIJA AVERAGE

U MS Excelu aritmetičku sredinu računamo pomoću funkcije „Average“. U zasebnu ćeliju upisujemo **=AVERAGE(*x:y*)** pri čemu ***x:y*** predstavlja raspon podataka koji se koristi u proračunu. Klikom na tipku enter, MS Excel u odabranoj ćeliji ispisuje vrijednost koja odgovara korištenoj funkciji.

	A	B	C	D	E
11					
12	1	3	5	3	1
13	4	6	2	4	3
14	5	1	1	4	6
15	2	4	2	6	5
16	3	1	4	2	4
17	1	5	6	6	2
18					
19	AVERAGE: =AVERAGE(A12:E17)				3,4

Slika 2.4. Funkcija AVERAGE

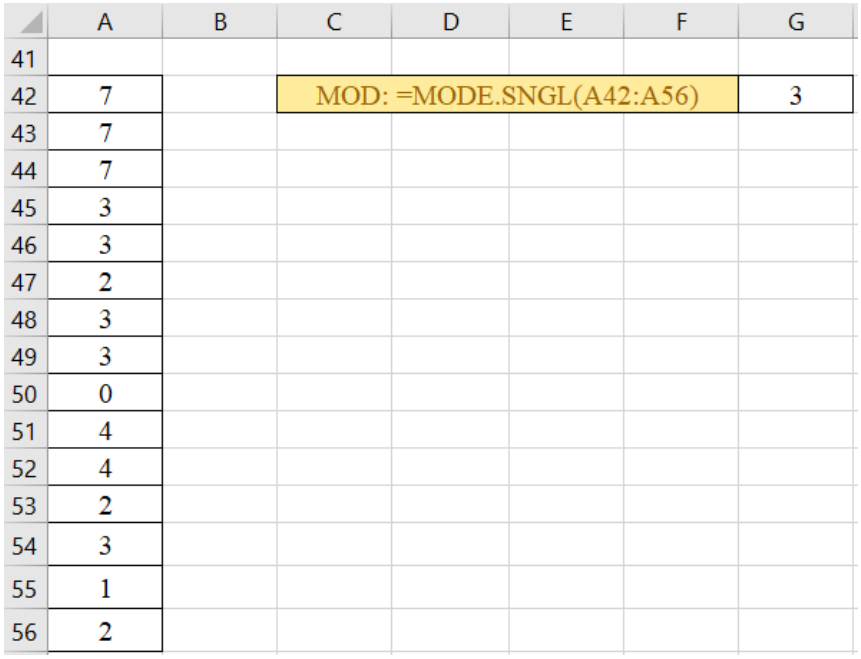
Na slici 2.4. prikazano je korištenje funkcije Average u MS Excelu na već ranije spomenutom primjeru (Primjer 2.1.) iz drugog poglavlja. Pri izračunu, koristio se raspon podataka od ćelije A12 do ćelije E17 te aritmetička sredina iznosi 3,4.

2.1.1.2. Mod

Mod predstavlja dominantnu vrijednost odnosno, podatak koji se najčešće pojavljuje/koji ima najveću frekvenciju.

FUNKCIJA MODE.SNGL

Računanje moda koristeći MS Excel provodi se tako da u odabranu ćeliju upišemo **=MODE.SNGL(*x:y*)** pri čemu *x:y* predstavlja raspon zadanih podataka.



Slika 2.5. Funkcija MODE.SNGL

Prikaz računanja moda na slici 2.5. temelji se na primjeru vezanom za frekvenciju (Primjer 2.3.). Raspon koji smo koristili (*x:y*) jest od ćelije A42 do ćelije A56. Mod, odnosno, vrijednost koja se najčešće pojavljuje u zadanom primjeru jest broj 3. Ako se vratimo na primjer 3. gdje smo računali frekvencije zadanih brojeva, možemo vidjeti da broj 3 doista ima najveću frekvenciju te je mod ispravno određen. Ako funkcija mode kao rezultat da grešku, znači da ne postoji mod danog skupa podataka, odnosno da se svi podaci u skupu javljaju točno jednom. Ako skup podataka ima više modova (skup je polimodalan), funkcija mode.sngl će kao rezultat dati onaj koji se javlja prvi u danom skupu podataka. U Excelu postoji i funkcija mode.mult, koja kao rezultat daje više modova (ako se radi o polimodalnom skupu podataka). Mode.mult je, kao i funkcija frequency, funkcija polja, te je pri njenom korištenju potrebno prvo označiti sve ćelije u koje želimo upisati rezultat, a funkciju potvrditi kombinacijom tipki CTRL + SHIFT + ENTER.

2.1.1.3. Medijan

Medijan predstavlja pozicionu srednju vrijednost. To je veličina od koje je barem pola zadanih podataka u statističkom skupu manjih, tj. barem pola većih od njega.

FUNKCIJA MEDIAN

U MS Excelu, medijan računamo upisivanjem **=MEDIAN(*x:y*)** u željenu ćeliju. Raspon *x:y* predstavlja raspon zadanih podataka.

	A	B	C	D	E	F	G
41							
42	7		MEDIJAN: =MEDIAN(A42:A56)				3
43	7						
44	7		0, 1, 2, 3, 3, 3, **3**, 3, 4, 4, 7, 7, 7				
45	3						
46	3						
47	2						
48	3						
49	3						
50	0						
51	4						
52	4						
53	2						
54	3						
55	1						
56	2						

Slika 2.6. Funkcija MEDIAN

Funkciju Median također smo prikazali koristeći primjer 2.3. kao i funkciju Mod. Raspon podataka na temelju kojih smo odredili medijan jest od ćelije A42 do ćelije A56. Radi lakše provjere, u ćeliji C44 - G44 ispisane su sve zadane vrijednosti te je vidljivo da je od broja 3 koji je označen crvenom bojom točno 6 vrijednosti manjih i 6 vrijednosti većih. Zaključujemo da je medijan zadanih vrijednosti broj 3.

2.1.2. Pokazatelji rasapa podataka u statističkom skupu

Pokazatelji rasapa podataka predočuju raspršenost podataka oko srednje vrijednosti u statističkom skupu te opisuju širenje ili varijabilnost podataka. U pokazatelje rasapa podataka oko aritmetičke sredine ubrajamo varijancu i standardnu devijaciju dok pokazateljima rasapa oko medijana pripadaju raspon podataka i interkvartilni raspon.

2.1.2.1. Raspon podataka

Raspon podataka dobijemo kao razliku najvećeg i najmanjeg podatka. Da bismo odredili raspon podataka, potrebno je najprije odrediti koja je najveća, a koja najmanja vrijednost u zadanom rasponu vrijednosti. Za određivanje minimalne i maksimalne vrijednosti, koristimo se funkcijama MIN i MAX.

FUNKCIJE MIN I MAX

Kako bismo odredili najmanju (minimalnu) i najveću (maksimalnu) vrijednost iz zadanih podataka koje je potrebno prethodno tabelirati, koristimo funkcije MIN i MAX tako da u zasebnoj ćeliji pišemo **=MIN(*x:y*)** za dobivanje minimalne odnosno, **=MAX(*x:y*)** za dobivanje maksimalne vrijednosti pri čemu *x:y* predstavlja raspon podataka na temelju kojih nastojimo odrediti minimalnu/maksimalnu vrijednost.

	A	B	C	D	E	F	G
1							
2		X_i					
3		1		MIN: =MIN(B3:B8)			1
4		2		MAX: =MAX(B3:B8)			6
5		3					
6		4					
7		5					
8		6					

Slika 2.7. Funkcije MIN i MAX

Na slici 2.7. prikazan je postupak kojim smo koristeći Microsoft Excel odredili minimalnu i maksimalnu vrijednost. Za ovaj prikaz koristili smo se tabeliranim podacima iz primjera 2.1. Raspon podataka koji se pri tome koristio bio je od ćelije B3 do ćelije B8, a dobiveni rezultati odgovaraju opisu funkcije: broj 1 doista jest minimalna vrijednost dok je broj 6 maksimalna vrijednost među ponuđenim podacima.

Na temelju toga, moguće je odrediti raspon koji u ovom primjeru prema formuli:

$$R = \max\ (X_i) - \min\ (X_i)$$

iznosi $R = 6 - 1 = 5$.

2.1.2.2. Interkvartilni raspon, koeficijent kvartilne devijacije

Interkvartilni raspon jest pokazatelj rasapa oko medijana te je jednak razlici između 3. kvartila i 1. kvartila. Vrijedi da je 2. kvartil (Q_2) jednak medijanu jer on predstavlja podatak od kojeg je manje 50% podataka. Prema tome, 1. kvartil (Q_1) je podatak od kojeg je manje 25% podataka, a 3. kvartil (Q_3) jest podatak od kojega je manje 75% podataka.

Da bismo odredili interkvartilni raspon, potrebno je odrediti koliko iznosi svaki od kvartila.

FUNKCIJA QUARTILE.INC

Računanje kvartila u MS Excelu izvodi se upisivanjem **=QUARTILE.INC(*x:y;z*)** u željenu ćeliju. Pri tome, *x:y* predstavlja raspon zadanih podataka, a *z* predstavlja broj (0, 1, 2, 3, 4) koji označava koji od kvartila želimo izračunati. Nulti kvartil predstavlja najmanju vrijednost, a četvrti kvartil najveću vrijednost statističkog skupa.

	A	B	C	D	E	F	G	H	I	J	K
41											
42	7		MEDIJAN: =MEDIAN(A42:A56)				3				
43	7										
44	7		0, 1, 2, 3, 3, 3, **3**, 3, 4, 4, 7, 7, 7								
45	3										
46	3		1. Kvartil (Q_1)			1. Kvartil: =QUARTILE.INC(A42:A56;1)					2
47	2		2. Kvartil (Q_2) = Medijan			2. Kvartil: =QUARTILE.INC(A42:A56;2)					3
48	3		3. Kvartil (Q_3)			3. Kvartil: =QUARTILE.INC(A42:A56;3)					4
49	3										
50	0										
51	4										
52	4										
53	2										
54	3										
55	1										
56	2										

Slika 2.8. Funkcija QUARTILE.INC

Kako bismo prikazali način računanja kvartila u MS Excelu, koristimo se primjerom 2.3. na kojem smo već ranije izračunali medijan. U primjeru, raspon podataka koji koristimo za određivanje kvartila obuhvaća sve ćelije između ćelija A42 i A56 dok broj nakon raspona mijenjamo u ovisnosti od toga koji kvartil računamo. Možemo primijetiti da je 2. kvartil doista jednak medijanu kojega smo ranije odredili, 1. kvartil iznosi 2, a 3. kvartil jednak je 4.

Prema tome, interkvartilni raspon iznosi $Q_3 - Q_1 = 4 - 2 = 2$.

Zatim, možemo odrediti koeficijent kvartilne devijacije (V_q) koji je pokazatelj reprezentativnosti medijana kao omjer interkvartilnog raspona i zbroja 1. i 3. kvartila. Ako je koeficijent kvartilne devijacije manji od 0.3, tada medijan smatramo reprezentativnim pokazateljem promatranog skupa podataka.

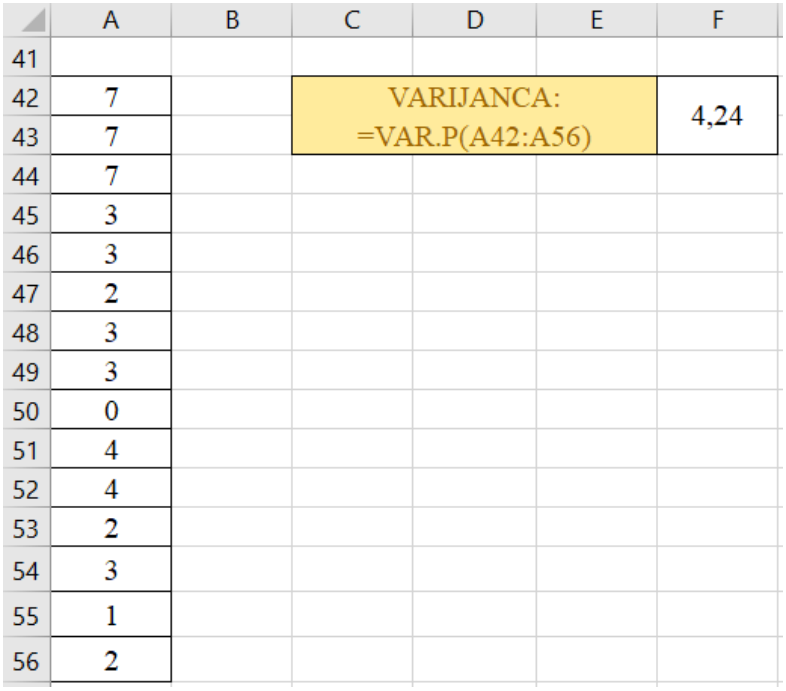
2.1.2.3. Varijanca, standardna devijacija, koeficijent varijacije

Varijanca (σ^2) predstavlja srednje kvadratno odstupanje podataka statističkog skupa od aritmetičke sredine. Što je veće odstupanje od aritmetičke sredine, to je varijanca veća.

FUNKCIJE VAR.S I VAR.P

Varijancu u MS Excelu možemo odrediti pomoću funkcija VAR.S i VAR.P. Funkciju VAR.S koristimo kada nam dani skup podataka predstavlja uzorak pa na temelju danog uzorka procjenjujemo varijancu cijele populacije dok funkcija VAR.P računa varijancu danog skupa podataka, odnosno koristimo ju kada nam dani skup podataka predstavlja populaciju. Obje varijance se određuju na isti način u MS Excelu. Na primjeru 2.3. opisat ćemo određivanje varijance funkcijom VAR.P.

U željenu ćeliju upisujemo **=VAR.P(*x:y*)**. Raspon *x:y* predstavlja raspon zadanih podataka.



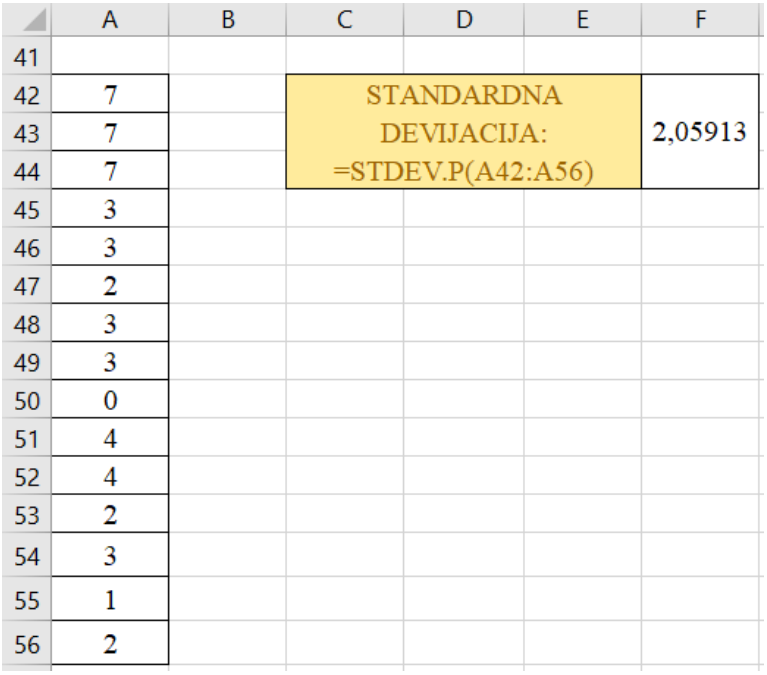
Slika 2.9. Funkcija VAR.P

Srednje kvadratno odstupanje za zadani statistički skup u rasponu od ćelije A42 do ćelije A56 iznosi 4,24.

Standardna devijacija (σ) jest mjera pomoću koje prikazujemo gustoću grupiranosti rezultata nekog mjerenja oko aritmetičke sredine, a jednaka je korijenu iz varijance. Za razliku od varijance, standardna devijacija ima istu mjernu jedinicu kao podaci.

FUNKCIJE STDEV.S I STDEV.P

Standardnu devijaciju možemo odrediti u MS Excelu pomoću funkcija STDEV.S, kada nam dani skup podataka predstavlja uzorak na temelju kojeg procjenjujemo standardnu devijaciju cijele populacije i STDEV.P, kada nam dani skup podataka predstavlja populaciju. Za određeni raspon *x:y* određujemo standardnu devijaciju upisujući **=STDEV.S(*x:y*)** ili **=STDEV.P(*x:y*)** u ćeliju u kojoj želimo prikazati rezultat. Primjer određivanja standardne devijacije koristeći se funkcijom STDEV.P također ćemo prikazati na primjeru 2.3. kojeg smo koristili kod određivanja frekvencije.



Slika 2.10. Funkcija STDEV.P

U navedenom primjeru koristili smo raspon ćelija od A42 do A56 te smo dobili rezultat standardne devijacije koji iznosi 2,05913. Iznos standardne devijacije mogli smo također dobiti korjenovanjem iznosa varijance jer vrijedi da je varijanca jednaka kvadratnoj vrijednosti standardne devijacije.

Koeficijent varijacije dobijemo kao omjer standardne devijacije i aritmetičke sredine. Njega koristimo kako bi ocijenili reprezentativnost aritmetičke sredine. Ako je koeficijent varijacije manji od 0.3, tada aritmetičku sredinu smatramo reprezentativnim pokazateljem promatranog skupa podataka.

U praksi pokazatelj smatramo reprezentativnim ako je relativna mjera rasapa manja od 0.3.

2.1.3. Pokazatelji oblika

Pod pokazatelje oblika svrstavamo koeficijent asimetrije te koeficijent spljoštenosti. Naime, asimetrija se razmatra u odnosu na aritmetičku sredinu dok se spljoštenost razmatra u odnosu na graf funkcije gustoće normalne razdiobe poznat kao Gaussova krivulja.

2.1.3.1. Koeficijent asimetrije

Funkcija SKEW u MS Excelu se koristi kako bismo odredili koeficijent asimetrije. Ako je koeficijent asimetrije jednak 0, tada je distribucija simetrična. Ako je koeficijent asimetrije negativan, tada je distribucija nagnuta u desno, odnosno ako je koeficijent asimetrije pozitivan, distribucija je nagnuta u lijevo.

FUNKCIJA SKEW

Koeficijent asimetrije odredimo pritiskom tipke enter nakon što smo u željenoj ćeliji zapisali **=SKEW(*x:y*)** pri čemu *x:y* predstavlja raspon zadanih podataka na temelju kojih želimo odrediti koeficijent asimetrije.

2.1.3.2. Koeficijent spljoštenosti

Pomoću funkcije KURT određujemo kurtozis odnosno, spljoštenost krivulje koja predstavlja distribuciju danog skupa podataka. Što je krivulja šiljastija, to je njezin koeficijent zaobljenosti veći (pozitivan je). Kurtozis je jednak nuli kada krivulja poprima zvonolik oblik (Gaussova krivulja, Normalna razdioba). Kada je krivulja plosnatija, tada je njezin koeficijent spljoštenosti negativan.

FUNKCIJA KURT

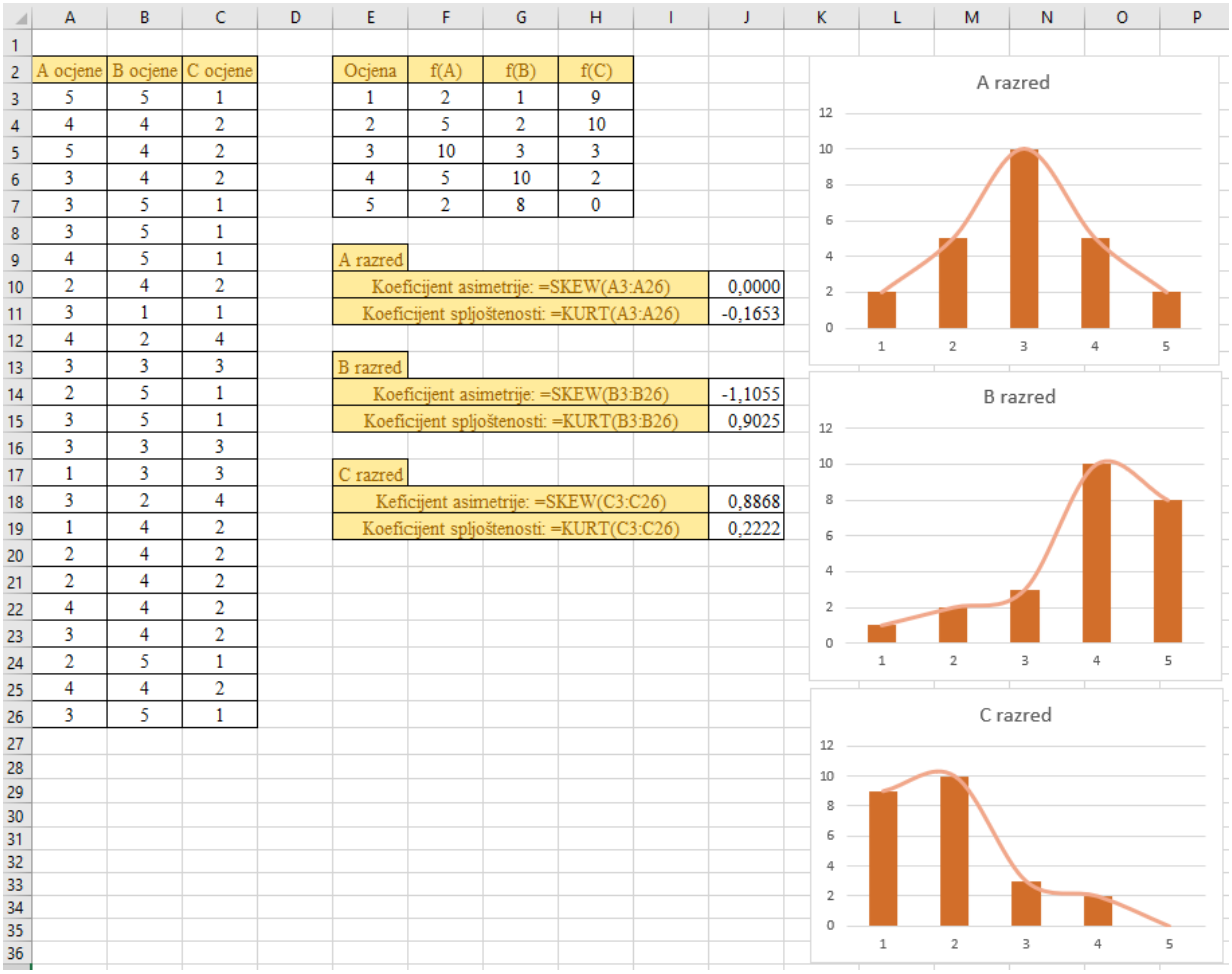
Funkciju KURT određujemo upisivanjem **=KURT(*x:y*)** u ćeliju i pritiskom na tipku enter. Raspon *x:y* predstavlja raspon zadanih podataka.

Primjer 2.4. Koeficijent asimetrije i spljoštenosti

U tablici su prikazani rezultati ispita iz statistike učenika A, B i C razreda. Na temelju zadanih podataka, potrebno je izračunati koeficijente asimetrije i spljoštenosti za svaki pojedini razred te grafički prikazati frekvencije ocjena i objasniti izgled grafova.

Tablica 2.1. Ocjene učenika iz A, B i C razreda (Primjer 2.4.)

A ocjene	5	4	5	3	3	3	4	2	3	4	3	2	3	3	1	3	1	2	2	4	3	2	4	3
B ocjene	5	4	4	4	5	5	5	4	1	2	3	5	5	3	3	2	4	4	4	4	4	5	4	5
C ocjene	1	2	2	2	1	1	1	2	1	4	3	1	1	3	3	4	2	2	2	2	2	1	2	1



Slika 2.11. Koeficijenti asimetrije i spljoštenosti

Pomoću primjera prikazanog na slici 2.11. gdje se nalaze podaci o ocjenama učenika A, B i C razreda nakon ispita iz statistike, opisat ćemo koeficijente asimetrije i spljoštenosti. U A razredu, bio je podjednak broj učenika koji su dobili ocjene 1 i 5 (njih je bilo najmanje), podjednak broj učenika koji su dobili ocjene 2 i 4, te je najviše bilo učenika koji su dobili ocjenu 3. Prema tome, graf poprima zvonolik oblik, tj. distribucija je simetrična pa je koeficijent asimetrije jednak 0. Što se tiče spljoštenosti na primjeru A razreda, ona je približno jednaka Gaussovoj krivulji

(koeficijent spljoštenosti je jako mali), ali pošto je negativan zaključujemo da je krivulja plosnatija od Gaussove krivulje. U B razredu, većina učenika dobila je ocjene 4 i 5 pa vidimo na slici 2.11. da je distribucija B razreda nagnuta u desno, odnosno aritmetička sredina je manja od medijana stoga je koeficijent asimetrije negativan. Koeficijent spljoštenosti je pozitivan što znači da je krivulja šiljastija od Gaussove krivulje. U C razredu, većina učenika je dobila ocjene 1 i 2 stoga je distribucija nagnuta u lijevo, aritmetička sredina je veća od medijana pa je koeficijent asimetrije pozitivan.

3. SLUČAJNE VARIJABLE

Varijabla predstavlja matematičku ili fizikalnu veličinu čija je vrijednost promjenjiva, tj. može poprimiti bilo koju vrijednost iz područja definicije. Varijablu nazivamo slučajnom ako su svi njezini ishodi realni brojevi, ali nije jednoznačno određeno koja će se vrijednost realizirati u statističkom eksperimentu u određenim uvjetima. Razlikujemo diskretne i kontinuirane slučajne varijable. Diskretne slučajne varijable mogu poprimiti vrijednosti unutar diskretnog skupa (konačan skup, skup prirodnih brojeva...), dok kontinuirane slučajne varijable poprimaju vrijednosti unutar neprebrojivog skupa. Kako bismo odredili kolika je vjerojatnost da slučajna varijabla poprimi određenu vrijednost, koristimo se teorijskim razdiobama.

3.1. Diskretne slučajne varijable

Diskretne slučajne varijable susrećemo u eksperimentima u kojima imamo konačan/prebrojiv skup ishoda. Primjeri diskretnih slučajnih varijabli su: broj koji smo dobili bacanjem igraće kockice, broj ispravnih dijelova u skupu od 100 dijelova, slučajno generirana vrijednost u rasponu od 0 do 20, itd. Dakle, to su varijable koje poprimaju jednu vrijednost iz skupa koji se sastoji od određenog broja ishoda. Neke od teorijskih razdiobi diskretnih slučajnih varijabli su: Binomna razdioba, Poissonova razdioba te Hipergeometrijska razdioba.

3.1.1. Poissonova razdioba

FUNKCIJA POISSON.DIST

Kako bismo razumjeli Poissonovu razdiobu, potrebno je najprije definirati parametre Poissonove razdiobe. Grčkim slovom lambda (λ) označavamo intenzitet pojavljivanja nekog događaja u određenom vremenskom intervalu ili na određenoj prostornoj domeni, a X predstavlja broj realizacije tog događaja u promatranom vremenskom intervalu/prostornoj domeni. Pomoću Poissonove razdiobe možemo odrediti vjerojatnost da X poprimi određenu vrijednost ako nam je poznat intenzitet pojavljivanja promatranog događaja (λ). U MS Excelu postoji funkcija Poisson.dist koja tu vjerojatnost računa unosom 3 podatka. Naime, u ćeliju upisujemo **=POISSON.DIST(*X*;*λ*;True/False)**. Argument „True“ se koristi ako nas zanima kolika je vjerojatnost da X poprimi vrijednost koja je manja ili jednaka od zadane vrijednosti, dok argument „False“ koristimo kada računamo vjerojatnost da X poprima točno određenu vrijednost.

Primjer 3.1. Poissonova razdioba

U trgovinu tijekom jednog sata u prosjeku uđe 8 ljudi. Potrebno je odrediti vjerojatnost da će:

a) broj ljudi koji uđe u trgovinu tijekom jednog sata biti 6
b) broj ljudi koji uđe u trgovinu tijekom jednog sata biti manji od 6
c) broj ljudi koji uđe u trgovinu tijekom jednog sata biti između 3 i 7.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
2														
3		λ =	8											
4		P(X=6)		Poissonova razdioba: =POISSON.DIST(6;C2;FALSE)									0,12214	12%
5		P(X<=6)		Poissonova razdioba: =POISSON.DIST(6;C2;TRUE)									0,31337	31%
6		P(3<X<7)		Poissonova razdioba: =POISSON.DIST(6;C2;TRUE)-POISSON.DIST(3;C2;TRUE)									0,27099	27%

Slika 3.1. Funkcija POISSON.DIST

U zadatku nam je zadan intenzitet pojavljivanja događaja, λ=8. U a) dijelu zadatka, kada računamo kolika je vjerojatnost (P) da će u trgovinu tijekom jednog sata ući točno 6 ljudi (X=6) koristimo argument FALSE. U b) dijelu zadatka kada računamo kolika je vjerojatnost (P) da u trgovinu tijekom jednog sata uđe manji ili jednak broj ljudi od 6 (X<=6) koristimo argument TRUE. U c) dijelu zadatka potrebno je od vjerojatnosti (P) da u trgovinu tijekom jednog sata uđe manje od 7 ljudi (X<7) oduzeti vjerojatnost (P) da u trgovinu tijekom jednog sata uđe broj ljudi veći od 3 (3<X) pa stoga taj zadatak rastavljamo na dva dijela. Prvi dio u kojem računamo vjerojatnost da u trgovinu tijekom jednog sata uđe broj ljudi manji ili jednak od 6 (X<=6) što je

zapravo jednako $X<7$ i drugi dio u kojem računamo vjerojatnost da u trgovinu tijekom jednog sata uđe broj ljudi manji ili jednak od 3 ($X<=3$) pa stoga u oba dijela zadatka koristimo argument TRUE.

3.1.2. Binomna razdioba

Pretpostavimo da izvodimo slučajni pokus koji možemo ponoviti n puta, a ishodi svakog pokusa su međusobno nezavisni. Neka slučajna varijabla X predstavlja broj realizacije događaja A u n ponavljanja pokusa. Ako ($X=k$) predstavlja događaj da se promatrani događaj A realizirao k puta u n ponavljanja pokusa, slučajna varijabla X poprima vrijednosti iz skupa $\{0, 1, ..., n\}$ i ravna se po binomnoj razdiobi. Da bismo odredili vjerojatnost događaja ($X=k$), mora biti poznata vjerojatnost uspjeha pri jednom ponavljanju pokusa (p) koja je konstantna.

FUNKCIJA BINOM.DIST

Ukoliko nam je poznat ukupni broj ponavljanja pokusa n i vjerojatnost uspjeha pri jednom ponavljanju pokusa p, vjerojatnost događaja ($X=k$) možemo izračunati koristeću se funkcijom BINOM.DIST u MS Excelu. U odabranu ćeliju upisujemo **=BINOM.DIST(*k*;*n*;*p*;True/False)**. Argument „True“ se koristi ukoliko nas zanima kolika je vjerojatnost da X poprimi vrijednost koja je manja ili jednaka od zadane vrijednosti ($X<=k$), dok argument „False“ koristimo kada računamo vjerojatnost da X poprima točno određenu vrijednost ($X=k$).

Primjer 3.2. Binomna razdioba

Neki stroj proizvodi 10% neispravnih proizvoda. Ukoliko se proizvodi isporučuju u kutijama od 75 dijelova, kolika je vjerojatnost da se:

a) u kutiji pronađe 10 neispravnih proizvoda?
b) u kutiji pronađe broj neispravnih proizvoda manji od 16?

	A	B	C	D	E	F	G	H	I
1									
2		p =	0,1						
3		n =	75						
4		a)	P(X=10)	Binomna razdioba: =BINOM.DIST(10;C3;C2;FALSE)					0,08796
5		b)	P(X<16)	Binomna razdioba: =BINOM.DIST(15;C3;C2;TRUE)					0,99729

Slika 3.2. Funkcija BINOM.DIST

U zadatku nam je zadan ukupni broj proizvoda po kutiji $n=75$ i vjerojatnost da će stoj proizvesti jedan neispravan proizvod $p=0,1$. U a) dijelu zadatka potrebno je izračunati vjerojatnost događaja (P) da se u kutiji pronađe 10 neispravnih dijelova ($X=k=10$). S obzirom da se traži vjerojatnost da broj neispravnih dijelova točno 10, koristimo argument FALSE. U b) dijelu zadatka potrebno je izračunati vjerojatnost (P) da je broj neispravnih dijelova u kutiji manji od 16 ($X=k<16$) što je jednako vjerojatnosti da je broj neispravnih dijelova u kutiji manji ili jednak 15 ($X=k<=15$) pa stoga koristimo argument TRUE.

Ovaj zadatak možemo riješiti i koristeći se Poissonovom razdiobom gdje će intenzitet pojavljivanja nekog događaja (λ) biti jednak umnošku ukupnog broja ponavljanja pokusa (n) i vjerojatnosti uspjeha pri jednom ponavljanju pokusa (p)

	A	B	C	D	E	F	G	H	I	J
1										
2		p =	0,1							
3		n =	75							
4		a)	P(X=10)	Binomna razdioba: =BINOM.DIST(10;C3;C2;FALSE)					0,08796	
5		b)	P(X<16)	Binomna razdioba: =BINOM.DIST(15;C3;C2;TRUE)					0,99729	
6										
7		Poissonova razdioba, općenito: =POISSON.DIST(X;λ;TRUE/FALSE)								
8		a)	P(X=10)	Poissonova razdioba: =POISSON.DIST(10;C3*C2;FALSE)						0,08583
9		b)	P(X<16)	Poissonova razdioba: =POISSON.DIST(15;C3*C2;TRUE)						0,99539

Slika 3.3. Aproksimacija binomne razdiobe poissonovom

Vidimo da su rezultati vjerojatnosti korištenjem binomne i poissonove razdiobe slični, ali nisu isti. Naime, javlja se greška u aproksimaciji koja se povećava s porastom rezultata kvadrata umnoška ukupnog broja ponavljanja pokusa (n) i vjerojatnosti uspjeha pri jednom ponavljanju

pokusa (*p*). Relativnu i apsolutnu grešku aproksimacije binomne razdiobe poissonovom možemo također izračunati pomoću funkcije ABS, slika 3.4.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2		p =	0,1											
3		n =	75											
4		a)	P(X=10)	Binomna razdioba: =BINOM.DIST(10;C3;C2;FALSE)					0,08796					
5		b)	P(X<16)	Binomna razdioba: =BINOM.DIST(15;C3;C2;TRUE)					0,99729					
6														
7		Poissonova razdioba, općenito: =POISSON.DIST(X;λ;TRUE/FALSE)												
8		a)	P(X=10)	Poissonova razdioba: =POISSON.DIST(10;C3*C2;FALSE)						0,08583				
9		b)	P(X<16)	Poissonova razdioba: =POISSON.DIST(15;C3*C2;TRUE)						0,99539				
10														
11														
12		a)	Relativna greška	Relativna: =ABS((vrijednost binomne r. - vrijednost poissonove r.)/vrijednost binomne r.))										0,0242
13			Apsolutna greška	Apsolutna: =ABS(vrijednost binomne r. - vrijednost poissonove r.)										0,00213
14		b)	Relativna greška	Relativna: =ABS((vrijednost binomne r. - vrijednost poissonove r.)/vrijednost binomne r.))										0,0019
15			Apsolutna greška	Apsolutna: =ABS(vrijednost binomne r. - vrijednost poissonove r.)										0,00189

Slika 3.4. Relativna i apsolutna greška aproksimacije binomne razdiobe poissonovom

3.1.3. Hipergeometrijska razdioba

Pretpostavimo da imamo skup od *N* elemenata od kojih *M* elemenata ima neko promatrano obilježje. Iz promatranog skupa biramo uzorak od *n* elemenata. Slučajna varijabla *X* predstavlja broj izabranih elemenata s promatranim obilježjem u uzorku od *n* elemenata. . Vrijednosti koje slučajna varijabla *X* može poprimiti su iz skupa *{0, 1, 2..., min{n,M}}* te se ravna po hipergeometrijskoj razdiobi.

FUNKCIJA HYPGEOM.DIST

Ukoliko nam je poznat ukupan broj elemenata u skupu *N*, broj elemenata s promatranim obilježjem *M* te od koliko se elemenata sastoji uzorak *n*, vjerojatanost da će slučajna varijabla *X* poprimiti vrijednost broja elemenata s promatranim obilježjem u uzorku (*k*) možemo izračunati pomoću funkcije HYPGEOM.DIST. U ćeliju upisujemo **=HYPGEOM.DIST(*k*;*n*;*M*;*N*;True/False)**. Argument „True“ koristi se kod računanja vjerojatnosti da *X* poprimi vrijednost koja je manja ili jednaka od zadane vrijednosti (*X*<=*k*), dok se argument „False“ koristi kod računanja vjerojatnosti da *X* poprima točno određenu vrijednost (*X*=*k*).

Primjer 3.3. Hipergeometrijska razdioba

U skupu proizvoda koji se sastoji od 80 proizvoda, njih 55 je ispravno dok su ostali neispravni. Slučajnim odabirom izabiremo 5 proizvoda. Koja je vjerojatnost da se u uzorku nađe:

a) točno 4 ispravna proizvoda
b) manje od 3 ispravna proizvoda

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		N =	80									
3		M =	55									
4		n =	5									
5		a)	P(X=4)	Hipergeometrijska razdioba: =HYPGEOM.DIST(4;C4;C3;C2;FALSE)								0,354674
6		b)	P(X<3)	Hipergeometrijska razdioba: =HYPGEOM.DIST(2;C4;C3;C2;TRUE)								0,173227

Slika 3.5. Funkcija HYPGEOM.DIST

Kao što je prikazano na slici 3.5., iz zadatka možemo odrediti ukupan broj elemenata u uzorku N=80, broj elemenata s promatranim obilježjem M=55 i broj elemenata u uzorku n=5. U a) dijelu zadatka treba se izračunati vjerojatnost da je u uzorku od 5 elemenata točno 4 elementa ispravno (X=4). Radi toga, u izračunu se koristi argument FALSE. U b) dijelu zadatka potrebno je izračunati vjerojatnost da je broj ispravnih proizvoda u uzorku manji od 3 (X<3), tj. manji ili jednak od 2 (X<=2) pa stoga pri izračunu koristimo argument TRUE.

3.2. Kontinuirane slučajne varijable

Kontinuirane slučajne varijable su varijable koje mogu poprimiti nebrojivo mnogo vrijednosti. Za razliku od diskretnih slučajnih varijabli gdje se svaka vrijednost poprima s nekom pozitivnom konačnom vjerojatnosti, kod kontinuiranih slučajnih varijabli svaka vrijednost će imati infinitezimalnu vjerojatnost (približno jednaku nuli). Radi toga, kontinuiranoj varijabli vjerojatnost možemo pridružiti nekom intervalu. Kod kontinuirane slučajne varijable uvodimo funkciju gustoće vjerojatnosti. To je funkcija kojom možemo opisati relativnu vjerojatnost da kontinuirana slučajna varijabla poprimi određenu vrijednost. Neke od razdioba kontinuiranih slučajnih varijabli: Normalna razdioba, Studentova t-razdioba.

3.2.1. Normalna razdioba

Normalna (Gaussova) slučajna varijabla je kontinuirana slučajna varijabla. Određena je s dva parametra, očekivanjem μ i standardnom devijacijom σ, tj. kvadratom standardne devijacije σ^2 (varijancom).

Funkcija gustoće vjerojatnosti normalne slučajne varijable jednaka je:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{X-\mu}{\sigma})^2} \ , X \in \boldsymbol{R},$$

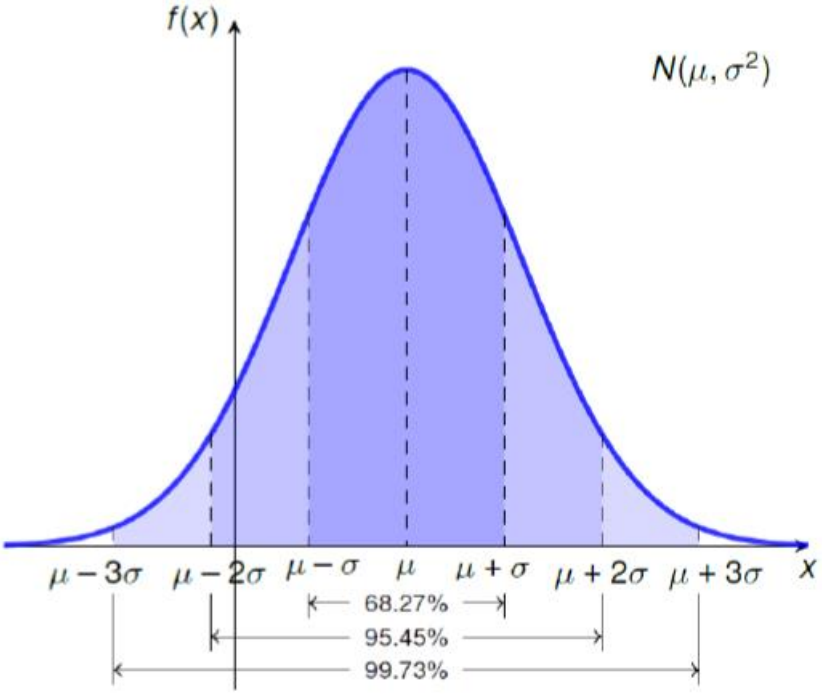
pri čemu su μ i σ realni brojevi, $\sigma > 0$.

U MS Excelu funkcija gustoće vjerojatnosti računa se upisivanjem u željenu ćeliju **=NORM.DIST(*X*;*μ*;*σ*;FALSE)**. Argument FALSE u naredbi NORM.DIST koristimo kako bi dobili vrijednost funkcije gustoće normalne razdiobe u određenoj točki ($X=k$).

Funkcija distribucije predstavlja vjerojatnost da slučajna varijabla X primi vrijednost iz intervala $(-\infty,x] \in R$. Funkcija distribucije jednaka je:

$$F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt \ , X \in \boldsymbol{R}.$$

Funkciju distribucije u MS Excelu dobijemo naredbom **=NORM.DIST(*X*;*μ*;*σ*;TRUE)**. Argument TRUE u naredbi NORM.DIST koristimo kako bi dobili vrijednost funkcije distribucije normalne razdiobe u određenoj točki ($X=k$).



Slika 3.6. Graf normalne razdiobe

Karakteristike grafa normalne (Gaussove) razdiobe: krivulja je simetrična u odnosu na pravac $X=\mu$ te je zvonolikog oblika. U $X=\mu$ dosegnuta je maksimalna vrijednost, točke infleksije se nalaze u $X=\mu\pm\sigma$ te se gotovo sve vrijednosti nalaze unutar 3 standardne devijacije od pravca $X=\mu$ (3σ pravilo).

Za jediničnu normalnu razdiobu vrijedi $\mu=0$ i $\sigma=1$. Ako slučajna varijabla X nije distribuirana po jediničnoj normalnoj razdiobi, tada koristimo varijablu Z koja je jednaka:

$$Z = \frac{X-\mu}{\sigma}$$

te je distribuirana po jediničnoj normalnoj razdiobi N(0,1). Tada vrijedi:

$$P(X<a)=F_X(a)=F_Z(\frac{a-\mu}{\sigma}).$$

FUNKCIJA NORM.DIST

Primjenu funkcije NORM.DIST objasniti ćemo na primjeru:

Primjer 3.4. Normalna razdioba

U tablici se nalaze vrijednosti napona testiranih baterija. Ako je dokazano da su vrijednosti napona u kontingentu normalno distribuirane odredite:

a) Postotak baterija čiji je napon manji od 1.6 V
b) Postotak baterija čiji je napon veći od 1.3 V
c) Postotak baterija čiji je napon između 1.35 V i 1.52 V

Tablica 3.1. Vrijednosti napona testiranih baterija (Primjer 3.4.)

1,1359	2,1781	1,6497	1,3966	1,2197	1,2540	1,3932	1,7136	1,6050	1,3057
1,6681	1,6693	1,2040	1,3754	1,1272	1,7647	1,4675	1,4217	2,1921	1,3761
1,3478	1,6205	1,4696	1,2364	2,0589	0,9972	1,1094	0,7955	1,8041	1,7970
1,2368	1,1676	1,7647	1,2970	0,9691	1,3605	1,1796	1,3537	1,4727	1,7791
1,4361	1,7435	1,2559	1,7915	1,3473	1,4297	1,3916	1,7808	1,6428	1,2240
1,4758	1,7807	1,9140	1,1770	1,5087	1,1966	1,7347	1,2809	1,3903	1,6266
1,8859	1,6003	1,8603	2,0939	1,0815	1,3540	1,4432	1,7713	1,5113	1,1617
1,6277	1,8956	1,2370	1,2104	1,6446	1,4859	1,4685	1,6437	1,3225	2,0548
1,1804	1,4239	1,2474	1,7894	1,1342	1,4393	1,5258	1,1880	1,6548	1,5557
1,7386	1,8973	1,8855	1,4651	1,8917	1,5745	1,7088	1,9429	1,8111	1,1250

Da bismo izračunali navedene postotke, prvo je potrebno odrediti čemu su jednaki parametri normalne razdiobe σ i μ. Parametar μ (očekivanje) najbolje se može procjeniti aritmetičkom sredinom stoga njega određujemo pomoću funkcije AVERAGE. Parametar σ, odnosno standardnu devijaciju računamo koristeći funkciju STDEV.S. Kada smo odredili parametre σ i μ, tražene postotke možemo izračunati na način da u željenu ćeliju upišemo **=NORM.DIST(*k*;*μ*;*σ*;True/False)**. Argument TRUE u ovom zadatku koristiti ćemo kada nas zanima vrijednost funkcije distribucije normalne razdiobe u određenoj točki ($X=k$), dok ćemo argument FALSE koristiti kako bi dobili vrijednost funkcije gustoće normalne razdiobe u određenoj točki ($X=k$).

1,1359	2,1781	1,6497	1,3966	1,2197	1,2540	1,3932	1,7136	1,6050	1,3057	
1,6681	1,6693	1,2040	1,3754	1,1272	1,7647	1,4675	1,4217	2,1921	1,3761	
1,3478	1,6205	1,4696	1,2364	2,0589	0,9972	1,1094	0,7955	1,8041	1,7970	
1,2368	1,1676	1,7647	1,2970	0,9691	1,3605	1,1796	1,3537	1,4727	1,7791	
1,4361	1,7435	1,2559	1,7915	1,3473	1,4297	1,3916	1,7808	1,6428	1,2240	
1,4758	1,7807	1,9140	1,1770	1,5087	1,1966	1,7347	1,2809	1,3903	1,6266	
1,8859	1,6003	1,8603	2,0939	1,0815	1,3540	1,4432	1,7713	1,5113	1,1617	
1,6277	1,8956	1,2370	1,2104	1,6446	1,4859	1,4685	1,6437	1,3225	2,0548	
1,1804	1,4239	1,2474	1,7894	1,1342	1,4393	1,5258	1,1880	1,6548	1,5557	
1,7386	1,8973	1,8855	1,4651	1,8917	1,5745	1,7088	1,9429	1,8111	1,1250	

μ =	Aritmetička sredina: =AVERAGE(B2:K11)	1,5060	
σ =	Standardna devijacija: =STDEV.S(B2:K11)	0,2894	
a)	X<=1,6	Normalna razdioba: =NORM.DIST(1,6;G13;G14;TRUE)	0,6274
b)	X>=1,3	Normalna razdioba: =1-NORM.DIST(1,3;G13;G14;TRUE)	0,7617
c)	1,35<=X<=1,52	Normalna razdioba: =NORM.DIST(1,52;G13;G14;TRUE)-NORM.DIST(1,35;G13;G14;TRUE)	0,2243

Slika 3.7. Funkcija NORM.DIST

U a) i c) dijelu zadatka pisali smo $X<=1.6$ i $1.35<=X<=1.52$ iako se u zadatku traži postotak baterija čiji je napon manji (ne manji ili jednak) zbog toga što kod normalne razdiobe nije bitno da li je granica intervala uključena u interval ili nije.

3.2.1.1. Interval povjerenja (Normalna razdioba)

Interval povjerenja predstavlja raspon mogućih vrijednosti unutar kojega se s izvjesnom vjerojatnosti nalazi statistička mjera populacije koju promatramo.

FUNKCIJA CONFIDENCE.NORM

U primjeru 3.4. također je potrebno odrediti intervalnu procjenu matematičkog očekivanja vrijednosti napona baterije na sljedećim razinama pouzdanosti:

a) 95%
b) 98%
c) 99%

Prije samog rješenja zadatka, objasnimo postupak kojim ćemo se voditi prilikom rješavanja. U zadatku je potrebno odrediti interval povjerenja u granicama <*a*, *b*>. S obzirom na to da je riječ o normalnoj razdiobi, gotovo sve vrijednosti koje slučajna varijabla poprima nalaze se unutar intervala od 3σ lijevo od očekivanja do 3σ desno od očekivanja. Dakle, u sredini imamo pravac $X=\mu$. S obzirom na to da je najbolji procjenitelj očekivanja aritmetička sredina, početak intervala „*a*“ bit će jednak razlici aritmetičke sredine i parametra „*c*“, dok će završetak intervala „*b*“ biti jednak zbroju aritmetičke sredine i parametra „*c*“. Parametar „*c*“ izračunati ćemo u MS Excelu koristeći se naredbom **=CONFIDENCE.NORM(*α*;*σ*;*n*)**. Kako bi odredili parametar *c*, potrebno je poznavati koliko iznosi koeficijent pouzdanosti α, standardna devijacija σ i kolika je veličina uzorka koji promatramo (*n*). Pri određivanju veličine uzorka u MS Excelu koristit ćemo se funkcijom COUNT. Bitno je da je odabrani uzorak randomiziran i stratificiran. Koeficijent pouzdanosti α jednak je:

$$1-(1-\alpha),$$

pri čemu veličina (1-α) predstavlja razinu pouzdanosti koja je u primjeru 3.4. zadana. Tipično se promatra 90%, 95% i 99%-tna pouzdanost, a ako razina pouzdanosti nije zadana, pretpostavlja se pouzdanost od 95%. Možemo zaključiti: za najveću razinu pouzdanosti, interval povjerenja je najširi. Širina intervala (*d*) može se izračunati kao 2**c*.

2	1,1359	2,1781	1,6497	1,3966	1,2197	1,2540	1,3932	1,7136	1,6050	1,3057					
3	1,6681	1,6693	1,2040	1,3754	1,1272	1,7647	1,4675	1,4217	2,1921	1,3761					
4	1,3478	1,6205	1,4696	1,2364	2,0589	0,9972	1,1094	0,7955	1,8041	1,7970					
5	1,2368	1,1676	1,7647	1,2970	0,9691	1,3605	1,1796	1,3537	1,4727	1,7791					
6	1,4361	1,7435	1,2559	1,7915	1,3473	1,4297	1,3916	1,7808	1,6428	1,2240					
7	1,4758	1,7807	1,9140	1,1770	1,5087	1,1966	1,7347	1,2809	1,3903	1,6266					
8	1,8859	1,6003	1,8603	2,0939	1,0815	1,3540	1,4432	1,7713	1,5113	1,1617					
9	1,6277	1,8956	1,2370	1,2104	1,6446	1,4859	1,4685	1,6437	1,3225	2,0548					
10	1,1804	1,4239	1,2474	1,7894	1,1342	1,4393	1,5258	1,1880	1,6548	1,5557					
11	1,7386	1,8973	1,8855	1,4651	1,8917	1,5745	1,7088	1,9429	1,8111	1,1250					
13	μ =	Aritmetička sredina: =AVERAGE(B2:K11)						1,5060							
14	σ =	Standardna devijacija: =STDEV.S(B2:K11)						0,2894							
15	a)	X<=1,6	Normalna razdioba: =NORM.DIST(1,6;G13;G14;TRUE)											0,6274	
16	b)	X>=1,3	Normalna razdioba: =1-NORM.DIST(1,3;G13;G14;TRUE)											0,7617	
17	c)	1,35<=X<=1,52	Normalna razdioba: =NORM.DIST(1,52;G13;G14;TRUE)-NORM.DIST(1,35;G13;G14;TRUE)											0,2243	
19	n =	Veličina uzorka: =COUNT(B2:K11)				100									
21	1 - α	α	c						a			b		d: =2*c	
22	0,95	0,05	c: =CONFIDENCE.NORM(0,05;G14;G19)					0,0567	a: =G13-I22		1,4492	b: =G13+I22	1,5627	0,1135	
23	0,98	0,02	c: =CONFIDENCE.NORM(0,02;G14;G19)					0,0673	a: =G13-I23		1,4386	b: =G13+I23	1,5733	0,1347	
24	0,99	0,01	c: =CONFIDENCE.NORM(0,01;G14;G19)					0,0746	a: =G13-I24		1,4314	b: =G13+I24	1,5805	0,1491	

Slika 3.8. Određivanje intervala povjerenja

a) Interval povjerenja za koeficijent pouzdanosti α=0.05 je (1.4492, 1.5627). Širina intervala povjerenja za razinu pouzdanosti 1-α=0.95 iznosi 0.113455.
b) Interval povjerenja za koeficijent pouzdanosti α=0.02 je (1.4386, 1.5733). Širina intervala povjerenja za razinu pouzdanosti 1-α=0.98 iznosi 0.134663.
c) Interval povjerenja za koeficijent pouzdanosti α=0.01 je (1.4314, 1.5805). Širina intervala povjerenja za razinu pouzdanosti 1-α=0.99 iznosi 0.149105.

Kao što smo već prije naveli pa sada i pokazali zadatkom, što je razina pouzdanosti veća, to je interval povjerenja širi.

3.2.1.2. Interval povjerenja (Studentova „t"-razdioba)

FUNKCIJA CONFIDENCE.T

Studentove se razdiobe razlikuju prema veličini promatranog uzorka n=1, 2, 3, ... Kako se n povećava tako se Studentova t-razdioba približava jediničnoj normalnoj razdiobi i za n=30 praktički joj je jednaka. Normalna razdioba ima dvije nezavisne varijable (očekivanje μ i varijancu σ^2), dok Studentova razdioba ima samo jednu nezavisnu varijablu (t). Također, Studentova razdioba se uglavnom koristi samo kod procjene parametara i kod testiranja hipoteza na osnovi uzorka. U MS Excelu, interval povjerenja Studentove razdiobe računamo na isti način kao i interval povjerenja normalne razdiobe, jedina je razlika u veličini uzorka. Naredba kojom računamo parametar „c" kod t-razdiobe glasi **=CONFIDENCE.T(α;σ;n).**

Primjer 3.5. Studentova „t"-razdioba

Prilikom rješavanja ispita učenici su postigli sljedeće rezultate:

20.6, 21.4, 23.2, 22.2, 21.7, 23.9, 20.2, 24.1, 23.5, 22.9, 24.2, 24.4, 23.4, 22.3, 21.7, 24.4, 22.8, 21.5, 20.9, 23.7

a) Ako je učenik na istom ispitu ostvario rezultat od 20.2 boda, možemo li taj rezultat smatrati prosječnim ili ispodprosječnim?

b) Koliki je minimalni broj bodova za koji možemo reći da opisuje iznadprosječan rezultat?

1												
2	Bodovi:											
3	20,6		n =	Veličina uzorka: =COUNT(B3:B22)				20				
4	21,4											
5	23,2		μ =	Očekivanje: =AVERAGE(B3:B22)					22,7			
6	22,2		σ =	Standardna devijacija: =STDEV.S(B3:B22)					1,32784			
7	21,7		1-α =	95%								
8	23,9		α =	0,05								
9	20,2		c =	c: =CONFIDENCE.T(0,05;I6:I3)					0,62145			
10	24,1		a =	a: =I5-H9	22,0786							
11	23,5		b =	b: =I5+H9	23,3214							
12	22,9											
13	24,2		<a,b> = <22.08, 23.32>		Unutar intervala <a,b> nalaze se prosječna rješenja.							
14	24,4											
15	23,4		ODGOVOR:									
16	23,3		a) 20.2 nalazi se unutar intervala <22.08, 23.32> stoga je to rješenje proječno.									
17	21,7		b) Minimalni broj bodova za koji možemo reći da opisuje iznadprosječan rezultat je 23.4.									
18	24,4											
19	22,8											
20	21,5											
21	20,9											
22	23,7											

Slika 3.9. Studentova „t“-razdioba

Prilikom rješavanja primjera 3.5., razina pouzdanosti nije bila zadana stoga smo pretpostavili pouzdanost od 95%.

4. TESTIRANJE STATISTIČKIH HIPOTEZA

Statistička hipoteza je tvrdnja o populaciji čija se ispravnost može testirati pomoću slučajnog uzorka. Postupak testiranja polazi od postavljanja nulte hipoteze (H_0) i alternativne hipoteze (H_1), pri čemu sadržaj alternativne hipoteze uvijek proturječi sadržaju nulte hipoteze. Provođenjem statističkih testova nastojimo doći do statističkog zaključka na temelju kojeg s određenom vjerojatnošću prihvaćamo ili odbacujemo nultu hipotezu. U postupku odlučivanja da li je neka statistička hipoteza ispravna, mogu se pojaviti sljedeće greške:

a) Odbacivanje istinite hipoteze H_0; Vjerojatnost da dođe do takve pogreške (Pogreška tipa I) pri kojoj kao ispravnu odaberemo alternativnu hipotezu H_1 koja je zapravo neispravna, a odbacimo nultu hipotezu H_0 koja je ispravna, jednaka je „α“.
b) Prihvaćanje neistinite hipoteze H_0; Vjerojatnost da dođe do takve pogreške (Pogreška tipa II) pri kojoj kao ispravnu odaberemo nultu hipotezu H_0 koja je zapravo neispravna, a odbacimo alternativnu hipotezu H_1 koja je ispravna, jednaka je „β“.

Prilikom testiranja statističkih hipoteza, bitan nam je parametar P-vrijednost. Taj parametar predstavlja signifikantnu/statistički značajnu razliku te se može odrediti kod bilo kojeg testiranja. Ako je P-vrijednost manja od α , tada odbacujemo nultu hipotezu H_0 i prihvaćamo alternativnu hipotezu H_1. Također, ako je P-vrijednost veća od α, tada prihvaćamo nultu, a odbacujemo alternativnu hipotezu. Što je P-vrijednost manja, imamo jači dokaz protiv hipoteze H_0.

Najčešće korišteni testovi su Z-test, T-test, F-test, ANOVA test i Hi-kvadrat test. U ovom radu opisat ćemo korištenje T-testa, F-testa i Hi-kvadrat testa u MS Excelu.

4.1.T-test

FUNKCIJA T.TEST

Primjer 4.1. T-test

U tablici su prikazani podaci o težini osmero ljudi prije i nakon odlaska na dijetu koja je trajala četiri mjeseca. Ispitajte je li se težina nakon dijete statistički značajno smanjila.

Tablica 4.1. Težina osmero ljudi prije i nakon odlaska na dijetu (Primjer 4.1.)

Osoba	Težina prije dijete	Težina nakon dijete

A	75	61
B	100	83
C	92	80
D	79	65
E	112	98
F	60	54
G	88	73
H	82	76

Prvo što je potrebno napraviti prilikom rješavanja ovakvog zadataka jest propisati nultu i alternativnu hipotezu. Nulta hipoteza H_0 bit će da je aritmetička sredina težina prije dijete jednaka aritmetičkoj sredini težina nakon dijete ($\mu_1=\mu_2$), pri čemu eksponent 1 označava težinu prije dijete, a eksponent 2 težinu nakon dijete. Alternativna hipoteza bit će da je aritmetička sredina težina prije dijete veća od aritmetičke sredine težina nakon dijete ($\mu_1>\mu_2$) zbog toga što nas zanima da li se težina nakon dijete statistički značajno smanjila. Nakon toga, potrebno je izračunati očekivanja μ_1 i μ_2 pomoću funkcije za aritmetičku sredinu AVERAGE te napisati čemu je jednaka vrijednost α (u ovom slučaju, vrijednost α nije posebno navedena stoga se pretpostavlja da iznosi 0.05). P-vrijednost možemo izračunati korištenjem funkcije T.TEST tako da u željenu ćeliju upisujemo **=T.TEST(*x*:*y*;*m*:*n*;1/2;1/2/3).** Raspon *x:y* sastoji se od podataka o težini prije dijete dok se raspon *m:n* sastoji od podataka o težini nakon dijete. Nakon toga upisuje se broj 1 ako nas u H_1 hipotezi zanima da li je aritmetička sredina težina prije dijete manja ili veća od aritmetičke sredine težina nakon dijete ($\mu_1 < \mu_2$; $\mu_1 > \mu_2$), odnosno upisujemo broj 2 ako nas zanima da li su navedene aritmetičke sredine različite ($\mu_1 \neq \mu_2$). U navedenom primjeru, mi ćemo upisati broj 1 jer nas zanima da li je aritmetička sredina težina prije dijete veća od aritmetičke sredine težina nakon dijete. Nadalje upisujemo broj 1 ako je riječ o uparenim podacima, broj 2 ako je riječ o neuparenim podacima s istom varijancom, tj. broj 3 ako je riječ o neuparenim podacima s različitom varijancom. U primjeru 4.1. imamo uparene podatke pa ćemo kao zadnji argument upisati broj 1.

	A	B	C	D	E	F	G	H	I	J	K
1	Osoba	Težina prije dijete		Težina nakon dijete							
2	A	75		61							
3	B	100		83			H_0 ->	$\mu_1=\mu_2$			
4	C	92		80			H_1 ->	$\mu_1>\mu_2$			
5	D	79		65							
6	E	112		98			μ_1	μ_1: =AVERAGE(B2:B9)			86
7	F	60		54			μ_2	μ_2: =AVERAGE(D2:D9)			73,75
8	G	88		73							
9	H	82		76			α =	0,05			
10											
11											
12				P-vrijednost		P-vrijednost: =T.TEST(B2:B9;D2:D9;1;1)				0,0000319	
13											
14				P-vrijednost < α							
15											
16				P-vrijednost <= α		Odbacujemo H_0, prihvaćamo H_1 ($\mu_1>\mu_2$)					
17				P-vrijednost > α		Odbacujemo H_1, prihvaćamo H_0					
18											
19				Odgovor: Težina nakon dijete se statistički značajno smanjila.							

Slika 4.1. Rješenje primjera 4.1.

4.2. F-test

FUNKCIJA F.TEST

F-test nam služi kako bismo u slučaju neuparenih podataka odredili jesu li varijance statistički značajno različite ili nisu. Nulta i alternativna hipoteza u F-testu uvijek su iste. Nulta hipoteza H_0 uvijek predstavlja tvrdnju da su varijanca 1 i varijanca 2 jednake, dok alternativna hipoteza H_1 uvijek pretpostavlja da se varijanca 1 i varijanca 2 razlikuju. P-vrijednost određujemo koristeći naredbu **=F.TEST(*x*:*y*;*m*:*n*)** pri čemu raspon *x:y* obuhvaća podatke jednog skupa dok raspon *m:n* obuhvaća podatke drugog skupa.

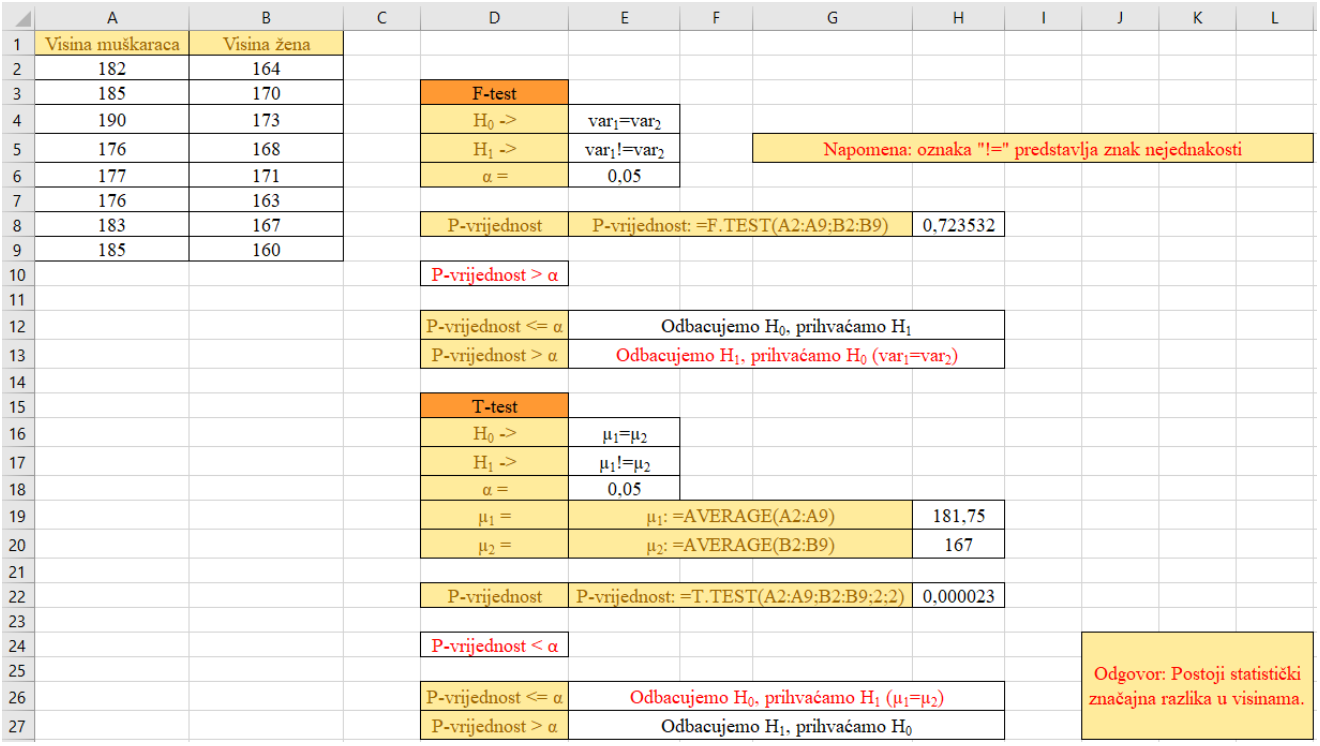
Primjer 4.2. F-test, T-test

U tablici su prikazani podaci o visini muškaraca i žena u centimetrima. Ispitajte postoji li statistički značajna razlika u visinama.

Tablica 4.2. Visina muškaraca i žena (Primjer 4.2.)

Visina muškaraca	Visina žena
182	164
185	170
190	173
176	168
177	171
176	163
183	167
185	160

Da bismo došli do rješenja zadatka i odgovorili na pitanje postoji li statistički značajna razlika u visinama prvo moramo pomoću F-testa odrediti jesu li varijance zadanih skupova podataka jednake te s tom informacijom pomoću T-testa odgovoriti na pitanje.



Slika 4.2. Rješenje primjera 4.2.

Pomoću F-testa došli smo do podatka da su varijanca 1 i varijanca 2 jednake te smo se dalje vodili postupkom koji je opisan u poglavlju 4.1. T-test kako bismo odgovorili na zadano pitanje. S obzirom na to da je P-vrijednost manja od vrijednosti α, prihvaćamo hipotezu H_1 i zaključujemo da postoji statistički značajna razlika u visinama.

4.3. Hi-kvadrat test

Hi-kvadrat test je test kojime testiramo da li se slučajna varijabla ravna prema određenoj razdiobi. Primjenu Hi-kvadrat testa u MS Excelu prikazat ćemo primjerom.

Primjer 4.3. Hi-kvadrat test

Brojevi koje možemo dobili bacanjem igraće kockice su 1, 2, 3, 4, 5 i 6. U tablici se nalaze rezultati koje smo dobili bacanjem igraće kockice 60 puta. Ako pretpostavimo da slučajna varijabla X predstavlja rezultat bacanja kockice, potrebno je testirati hipotezu da se slučajna varijabla X ravna prema uniformnoj razdiobi.

Tablica 4.3. Brojevi dobiveni bacanjem igraće kockice (Primjer 4.3.)

5	5	4	4	4	1
4	6	4	1	6	3
5	3	3	4	4	3
6	5	1	4	2	1
2	5	1	3	3	5
3	1	4	3	4	6
5	3	1	6	3	3
3	3	2	3	5	5
2	4	4	3	3	5
3	2	4	3	4	2

Ono što mi želimo ispitati jest ima li slučajna varijabla X uniformnu razdiobu za koju vrijedi $P(X=1) = P(X=2)= P(X=3) = P(X=4) = P(X=5) = P(X=6) =1/6$, odnosno vjerojatnost da X poprimi bilo koju od mogućih vrijednosti je jednaka 1/6

Prvi korak prilikom testiranja Hi-kvadrat testom jest odrediti nultu i alternativnu hipotezu. U našem slučaju, nulta hipoteza H_0 pretpostavljat će da se slučajna varijabla X ravna prema uniformnoj razdiobi dok će alternativna hipoteza pretpostavljati suprotno, tj. H_1 će pretpostavljati da se slučajna varijabla X ne ravna prema uniformnoj razdiobi. Sljedeći korak je određivanje vrijednosti H. Vrijednost H računa se naredbom $\mathbf{=SUM((f_{ei}-f_{ti})^2/f_{ti})}$. F_{ei} predstavlja eksperimentalne frekvencije, dok su f_{ti} teoretske frekvencije. Eksperimentalne frekvencije

računamo preko naredbe **=FREQUENCY(*x*:*y*;*m*:*n*)** pri čemu raspon *x:y* obuhvaća sve zadane vrijednosti, a raspon *m:n* obuhvaća vrijednosti iz stupca X_i (sve vrijednosti koje može poprimiti slučajna varijabla). Teoretske frekvencije računamo kao umnožak vjerojatnosti da slučajna varijabla poprimi neku vrijednost (p_i) i ukupnog broja zadanih podataka (n). Teoretska frekvencija svakog razreda mora biti veća od 5. Ako nije, spajamo razrede dok se ne postigne vrijednost koja je veća od 5. Nakon što smo izračunali H, potrebno je izračunati vrijednost kritičnog područja (h_{krit}).

FUNKCIJA CHISQ.INV

Vrijednost h_{krit} određujemo naredbom **=CHISQ.INV(1-*α*;*d*$_f$)**. Prvi argument jest razina pouzdanosti, a drugi argument „d_f“ predstavlja stupanj slobode. Stupanj slobode dobije se tako da se od broja razreda (r) nakon eventualnog spajanja oduzme broj procijenjenih parametara (m) i broj 1:

$$d_f = r - m - 1$$

Ako vrijedi da je vrijednost H veća ili jednaka h_{krit}, tada odbacujemo H_0 u korist H_1 (vrijednost statistike H upada u kritično područje, tj. X nije uniformna).

Ako vrijedi da je vrijednost H manja od vrijednosti h_{krit}, tada prihvaćamo H_0 i odbacujemo H_1 (što bi značilo da se slučajna varijabla X ravna po uniformnoj razdiobi).

	A	B	C	D	E	F
1	5	5	4	4	4	1
2	4	6	4	1	6	3
3	5	3	3	4	4	3
4	6	5	1	4	2	1
5	2	5	1	3	3	5
6	3	1	4	3	4	6
7	5	3	1	6	3	3
8	3	3	2	3	5	5
9	2	4	4	3	3	5
10	3	2	4	3	4	2

	H	I	J	K	L
1	Hi-kvadrat test				
2					
3	H_0 ->	Slučajna varijabla X ravna se prema uniformnoj razdiobi			
4	H_1 ->	Slučajna varijabla X ne ravna se prema uniformnoj razdiobi			
5	α =	0,05			
6	n =	n: =COUNT(A1:F10)		60	
7					
8	X_i	p_i	$f_{ti} = p_i*n$	f_{ei}	$(f_{ei} - f_{ti})^2/f_{ti}$
9	1	0,1667	10	7	0,9
10	2	0,1667	10	6	1,6
11	3	0,1667	10	18	6,4
12	4	0,1667	10	14	1,6
13	5	0,1667	10	10	0
14	6	0,1667	10	5	2,5
15					
16	H =	H: =SUM(L10:L15)		13	
17					
18	m =	0			
19	r =	6			
20	d_f = r - m - 1		5		
21					
22	h_{krit} =	h_{krit}: =CHISQ.INV(1-I6;J21)		11,0705	
23					
24	H >= hkrit	Odbacujemo H0, prihvaćamo H1			
25	H < h_{krit}	Ne možemo odbaciti H0 u korist H1			
26					

Slika 4.3. Rješenje primjera 4.3.

S obzirom na to da je vrijednost H veća od vrijednosti h_{krit}, prihvaćamo alternativnu hipotezu H_1 te zaključujemo da se slučajna varijabla X ne ravna prema uniformnoj razdiobi.

5. KORELACIJA I REGRESIJA

Korelacija u statistici predstavlja ovisnost između statističkih varijabla. Kod linearne korelacije, jakost korelacije izražena je Pearsonovim koeficijentom *r*, koji može poprimati vrijednosti između -1 i +1.

Kada govorimo o jačini korelacije, ako se koeficijent korelacije po apsolutnoj vrijednosti nalazi u intervalu od 0 do 0.2, tada koeficijent korelacije smatramo nekoreliranim (ovisnost između varijabli je zanemarivo mala). Ako se koeficijent korelacije nalazi u intervalu od 0.2 do 0.5 tada je ovisnost između varijabli slaba. Za koeficijent korelacije koji poprima vrijednosti od 0.5 do 0.8 kažemo da je jačina korelacije umjerena dok jaku korelaciju imamo ako koeficijent korelacije poprima vrijednosti od 0.8 do 1.

Ako je vrijednost koeficijenta korelacije pozitivna, to bi značilo da porastom nezavisne varijable raste i zavisna varijabla (rastući graf). Pri negativnoj vrijednosti koeficijenta korelacije, s porastom nezavisne varijable, zavisna varijabla pada (padajući graf).

Pomoću regresijskih tehnika kvantitativno možemo izraziti zavisnost varijabla (korelaciju). U najjednostavnijem slučaju imamo linearnu zavisnost jedne varijable (Y) o jednoj nezavisnoj varijabli (X). Ako zavisna varijabla (Y) ovisi o više nezavisnih varijabli, tada govorimo o multilinearnoj regresiji.

Sve vezano za korelaciju i regresiju u MS Excelu objasnit ćemo na primjeru zadatka s multilinearnom regresijom.

Primjer 5.1. Multilinearna regresija

U tablici su dani podaci o vremenima postignutima na polumaratonima (u minutama), te o broju kilometara koje je trkač prešao u mjesecima uoči utrke, o elevaciji (visinskoj razlici) staze utrke, kao i o srednjoj temperaturi za vrijeme utrke.

a) Ispitajte utječu li posljednja tri faktora značajno na rezultat ostvaren na utrci, koristeći multilinearnu regresiju.
b) Navedite regresijsku jednadžbu te prokomentirajte kako koji parametar utječe na rezultat.
c) Koristeći regresijsku jednadžbu odredite vrijeme na utrci ako trkač odradi 1800 km u pripremnom periodu, ako je elevacija na utrci 100 m, te ako je srednja temperatura za vrijeme utrke 12 stupnjeva.

Tablica 5.1. Rezultati polumaratonske trke (Primjer 5.1.)

VRIJEME	KM	ELEVACIJA	TEMPERATURA
79	1420	150	18
86	1680	350	18
109	920	350	27
78	1510	200	7
117	1170	450	19
74	1600	0	15
75	1670	50	17
98	850	200	16
90	860	150	5
122	460	50	16
93	1220	250	15
110	880	450	21
93	1220	350	6
85	1650	450	6
99	1180	350	9
83	1660	350	6
75	1740	50	24
102	800	400	16
114	730	300	13
85	1250	100	6
120	660	350	24
115	910	300	20
79	1140	0	11
86	1040	50	16
112	630	50	23

Da bismo za zadane podatke dobili prikaz svih parametara kao što je prikazano na slici 5.1., potrebno je glavnom izborniku u MS Excelu odabrati „Data“, zatim „Data analysis“ i „Regresion“. U izborniku koji nam otvori pritiskom na „Regresion“ potrebno je odabrati raspon zavisne varijable (od ćelije A1 do ćelije A26) i nezavisnih varijabli (od ćelije B1 do ćelije D26) te odabrati opciju "labels“ kako bi nam se u tablicama prikazala imena varijabla (nije nužno potrebno).

	A	B	C	D	E	F	G	H	I	J	K	L
1	VRIJEME	KM	ELEVACIJA	TEMPERATURA								
2	79	1420	150	18			Jednadžba multilinearne regresije: Y=113.6034-0.0308*X1+0.0401*X2+0.5250*X3					
3	86	1680	350	18								
4	109	920	350	27								
5	78	1510	200	7			SUMMARY OUTPUT			Zavisna varijabla (Y): VRIJEME		
6	117	1170	450	19						Nezavisne varijable:		
7	74	1600	0	15			Regression Statistics			X1:	KM	1800
8	75	1670	50	17			Multiple R	0,9354		X2:	ELEVACIJA	100,0000
9	98	850	200	16			R Square	0,8750		X3:	TEMPERATURA	12,0000
10	90	860	150	5			Adjusted R Square	0,8572				
11	122	460	50	16			Standard Error	5,9568		Y = a0 + a1*X1+a2*X2 + a3*X3		68,5012
12	93	1220	250	15			Observations	25,0000		Y: =G19+G20*K5+G21*K6+G22*K7		
13	110	880	450	21								
14	93	1220	350	6			ANOVA					
15	85	1650	450	6				df	SS	MS	F	Significance F
16	99	1180	350	9			Regression	3,0000	5218,2124	1739,4041	49,0205	0,0000
17	83	1660	350	6			Residual	21,0000	745,1476	35,4832		
18	75	1740	50	24			Total	24,0000	5963,3600			
19	102	800	400	16								
20	114	730	300	13				Coefficients	Standard Error	t Stat	P-value	
21	85	1250	100	6		a0=	Intercept	113,6034	5,9471	19,1023	0,0000	
22	120	660	350	24		a1=	KM	-0,0308	0,0033	-9,2537	0,0000	
23	115	910	300	20		a2=	ELEVACIJA	0,0401	0,0079	5,0766	0,0001	
24	79	1140	0	11		a3=	TEMPERATURA	0,5250	0,1946	2,6973	0,0135	
25	86	1040	50	16								
26	112	630	50	23				Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%	
27							Intercept	101,2357	125,9711	101,2357	125,9711	
28							KM	-0,0377	-0,0239	-0,0377	-0,0239	
29							ELEVACIJA	0,0237	0,0566	0,0237	0,0566	
30							TEMPERATURA	0,1202	0,9298	0,1202	0,9298	

Slika 5.1. Multilinearna regresija

Opći oblik multilinearne regresije glasi:

$$Y=a_0+a_1X_1+a_2X_2+...+a_nX_n$$

pri čemu Y predstavlja zavisnu varijablu, X_1, X_2, ..., X_n su nezavisne varijable, dok su a_0, a_1, ..., a_n koeficijenti. U našem primjeru:

Zavisna varijabla (Y): VRIJEME	
Nezavisne varijable:	
X1:	KM
X2:	ELEVACIJA
X3:	TEMPERATURA

		Coefficients
a0=	Intercept	113,6034
a1=	KM	-0,0308
a2=	ELEVACIJA	0,0401
a3=	TEMPERATURA	0,5250

Slika 5.2. Parametri multilinearne regresije

Pomoću funkcije „Regresion“ automatski su se generirale vrijednosti koeficijenta, stoga je naša regresijska jednadžba jednaka:

Jednadžba multilinearne regresije: Y=113.6034-0.0308*X1+0.0401*X2+0.5250*X3

Slika 5.3. Regresijska jednadžba

Ako nam je poznata regresijska jednadžba, možemo odrediti iznos zavisne varijable „VRIJEME“ za bilo koje vrijednosti nezavisnih varijabli. U c) dijelu zadatka potrebno je odrediti vrijeme na utrci ako trkač odradi 1800 km (X_1) u pripremnom periodu, ako je elevacija na utrci 100 m (X_2), te ako je srednja temperatura za vrijeme utrke 12 stupnjeva (X_3). Ubacimo li te vrijednosti u regresijsku jednadžbu dobiti ćemo rezultat: 68.5012.

Zavisna varijabla (Y): VRIJEME		
Nezavisne varijable:		
X1:	KM	1800
X2:	ELEVACIJA	100,0000
X3:	TEMPERATURA	12,0000
Y = a0 + a1*X1+a2*X2 + a3*X3		68,5012
Y: =G19+G20*K5+G21*K6+G22*K7		

Slika 5.4. Prikaz podataka iz c) dijela zadatka

Ako su koeficijenti ($a_1, a_2, ..., a_n$) negativni, to znači da se njihovim porastom za 1 zavisna varijabla smanjuje za njihovu vrijednost. U našem primjeru, ukoliko bi koeficijent a_1 uz nezavisnu varijablu „KM“ povećali za 1, zavisna varijabla „VRIJEME“ smanjila bi se za 0.0308. Ako su koeficijenti pozitivni, njihovo povećanje prati povećanje zavisne varijable. Ako koeficijente a_2 uz nezavisnu varijablu „ELEVACIJA“ i a_3 uz nezavisnu varijablu „TEMPERATURA“ povećamo za 1, tada se zavisna varijabla „VRIJEME“ poveća za 0.0401 (ELEVACIJA) i 0.5250 (TEMPERATURA).

FUNKCIJA CORREL

Funkciju CORREL koristimo kako bi odredili koeficijent korelacije. U ćeliju upisujemo **=CORREL(*x*:*y*;*m*:*n*)**. Raspon *x:y* odnosi se na raspon vrijednosti zavisne varijable *Y* dok se raspon *m:n* odnosi na raspon vrijednosti nezavisne varijable *X*.

FUNKCIJA INTERCEPT

Intercept predstavlja koeficijent a_0 koji je konstantan. Njega u MS Excelu možemo izračunati koristeći funkciju INTERCEPT tako da u željenu ćeliju upišemo **=INTERCEPT(*x*:*y*;*m*:*n*)** pri čemu raspon *x:y* obuhvaća vrijednosti zavisne varijable *Y*, a raspon *m:n* obuhvaća vrijednosti nezavisne varijable *X*.

FUNKCIJA SLOPE

Ako želimo izračunati vrijednost koeficijenta *a* uz nezavisnu varijablu *X*, koristimo se funkcijom SLOPE. U ćeliju upisujemo **=SLOPE(*x*:*y*;*m*:*n*)**. Raspon *x:y* sastoji se od vrijednosti zavisne varijable *Y*, a raspon *m:n* od vrijednosti nezavisne varijable *X*.

Parametar R Square (R-kvadrat), također poznat kao koeficijent određivanja, koristi se za procjenu prilagodbe modela regresijske jednadžbe. On predstavlja postotak varijance zavisne varijable koji je objašnjen nezavisnom varijablom. Ako je vrijednost R Square-a manja od 0.5 tada nezavisne varijable nisu dobre u predviđanju zavisne varijable. Ako R Square poprima vrijednost veću od 0.5, tada nezavisne varijable dobro predviđaju zavisnu varijablu. U primjeru koji razmatramo, R Square vrijednost jednaka je 0.8750 što znači da zadane nezavisne varijable dobro predviđaju zavisnu varijablu, tj. 87.5% varijabilnosti zavisne varijable objašnjeno je nezavisnim varijablama.

Slika 5.5. R Square

Parametar P-value testira hipotezu da li se koeficijenti razlikuju od nule. Ako je P-value neke nezavisne varijable manji ili jednak od 0.05, tada se koeficijent uz tu nezavisnu varijablu značajno razlikuje od nule što znači da ta nezavisna varijabla ima utjecaja na zavisnu varijablu. Ako je P-value nezavisne varijable veći od 0.05, tada se prihvaća da koeficijent približno jednak nuli pa je i sama nezavisna varijabla koja se množi s navedenim koeficijentom jednaka nuli i

zaključujemo da ta nezavisna varijabla nema utjecaja na zavisnu varijablu. U našem primjeru, P-value svih nezavisnih varijabla manja je od 0.05 što znači da sve nezavisne varijable imaju utjecaja na zavisnu varijablu.

		Coefficients	P-value
a0=	Intercept	113,6034	0,0000
a1=	KM	-0,0308	0,0000
a2=	ELEVACIJA	0,0401	0,0001
a3=	TEMPERATURA	0,5250	0,0135

Slika 5.6. P-value

6. ZAKLJUČAK

U ovome radu navode se samo neke od mnogobrojnih statističkih funkcija koje su dostupne unutar programa Microsoft Excel. Iako statističke funkcije nisu temelj MS Excela, program sadržava osnovni paket funkcija koje su potrebne za statističku analizu. Po potrebi, te se funkcije mogu proširiti i približiti onima koje su sadržane u statističkim paketima. Na temelju objašnjenih statističkih funkcija kroz rad, možemo primijetiti da je program jednostavan za korištenje i pregledan. Svi podaci mogu se razvrstati u tablice i grafikone te se njihovom analizom mogu otkriti trendovi.

S obzirom na moje iskustvo rada u MS Excelu, smatram da je njegova glavna prednost, u usporedbi s drugim programima u kojima sam do sada radila, u tome što je multifunkcionalan. Može se koristiti za upravljanje podataka, analizu podataka, složene matematičke izračune, itd. Kada uz to uzmemo u obzir dostupnost i cijenu, smatram da MS Excel može konkurirati velikom broju programa za statistički analizu.

Nastavno, MS Excel jedan je od osnovnih i najrasprostranjenijih programa u Microsoftovom paketu koji nudi čitav niz različitih usluga te smatram da bi se upravo iz tog razloga trebalo posvetiti više pažnje prema edukaciji budućih generacija u radu u navedenom programu.

LITERATURA

[1] „Vrste statističkog softvera“, s interneta, https://e-statistika.rs/Article/Display/vrste-statistickog-softvera, 24.05.2022.

[2] Begović Kovač E.; Jerković M.: „Statističke i numeričke metode“, s interneta, http://matematika.fkit.hr/novo/statistika_i_vjerojatnost/vjezbe/cjeline/Statistika_skripta.pdf, 24.05.2022.

[3] „Studentova razdioba“, s interneta, http://struna.ihjj.hr/naziv/studentova-razdioba/29915/, 24.05.2022.

[4] Hrvatska enciklopedija, mrežno izdanje: „Regresija“, s interneta, https://www.enciklopedija.hr/Natuknica.aspx?ID=52268, 24.05.2022.

[5] „Statističke funkcije“, s interneta, https://support.microsoft.com/hr-hr/office/statisti%C4%8Dke-funkcije-referenca-624dac86-a375-4435-bc25-76d659719ffd, 24.05.2022.

[6] Benšić M.; Šuvak N.: „Statistika – radni materijali“, s interneta, http://www.mathos.unios.hr/ptfstatistika/Vjezbe/00_statistika_20102011.pdf, 24.05.2022

POPIS STATISTIČKIH FUNKCIJA

Funkcija Frequency 4
Funkcija Average 6
Funkcija Mode.sngl 7
Funkcija Median 9
Funkcije Min i Max 10
Funkcija Quartile.inc 11
Funkcije Var.s i Var.p 12
Funkcije Stdev.s i Stdev.p 13
Funkcija Skew 15
Funkcija Kurt 15
Funkcija Poisson.dist 17
Funkcija Binom.dist 19
Funkcija Hypgeom.dist 21
Funkcija Norm.dist 24
Funkcija Confidence.norm 26
Funkcija Confidence.t 28
Funkcija T.test 30
Funkcija F.test 32
Funkcija Chisq.inv 35
Funkcija Correl 40
Funkcija Intercept 41
Funkcija Slope 41

POPIS SLIKA

Slika 2.1. Tabeliranje podataka iz primjera 2.1. 4
Slika 2.2. Funkcija FREQUENCY 5
Slika 2.3. Graf funkcije FREQUENCY 5
Slika 2.4. Funkcija AVERAGE 7
Slika 2.5. Funkcija MODE.SNGL 8
Slika 2.6. Funkcija MEDIAN 9
Slika 2.7. Funkcije MIN i MAX 10
Slika 2.8. Funkcija QUARTILE.INC 11
Slika 2.9. Funkcija VAR.P 13
Slika 2.10. Funkcija STDEV.P 14
Slika 2.11. Koeficijenti asimetrije i spljoštenosti 16
Slika 3.1. Funkcija POISSON.DIST 18
Slika 3.2. Funkcija BINOM.DIST 20
Slika 3.3. Aproksimacija binomne razdiobe poissonovom 20
Slika 3.4. Relativna i apsolutna greška aproksimacije binomne razdiobe poissonovom 21
Slika 3.5. Funkcija HYPGEOM.DIST 22
Slika 3.6. Graf normalne razdiobe 24
Slika 3.7. Funkcija NORM.DIST 26
Slika 3.8. Određivanje intervala povjerenja 27
Slika 3.9. Studentova „t“-razdioba 29
Slika 4.1. Rješenje primjera 4.1. 32
Slika 4.2. Rješenje primjera 4.2. 33
Slika 4.3. Rješenje primjera 4.3. 36
Slika 5.1. Multilinearna regresija 39
Slika 5.2. Parametri multilinearne regresije 39
Slika 5.3. Regresijska jednadžba 40

Slika 5.4. Prikaz podataka iz c) dijela zadatka 40

Slika 5.5. R Square 41

Slika 5.6. P-value 42

POPIS TABLICA

Tablica 2.1. Ocjene učenika iz A, B i C razreda (Primjer 2.4.) 16

Tablica 3.1. Vrijednosti napona testiranih baterija (Primjer 3.4.) 25

Tablica 4.1. Težina osmero ljudi prije i nakon odlaska na dijetu (Primjer 4.1.) 30

Tablica 4.2. Visina muškaraca i žena (Primjer 4.2.) 33

Tablica 4.3. Brojevi dobiveni bacanjem igraće kockice (Primjer 4.3.) 34

Tablica 5.1. Rezultati polumaratonske trke (Primjer 5.1.) 38

POPIS PRIMJERA

Primjer 2.1. Diskretno statističko obilježje 3
Primjer 2.2. Kontinuirano statističko obilježje 3
Primjer 2.3. Frekvencija 4
Primjer 2.4. Koeficijent asimetrije i spljoštenosti 15
Primjer 3.1. Poissonova razdioba 18
Primjer 3.2. Binomna razdioba 19
Primjer 3.3. Hipergeometrijska razdioba 22
Primjer 3.4. Normalna razdioba 25
Primjer 3.5. Studentova „t“-razdioba 28
Primjer 4.1. T-test 30
Primjer 4.2. F-test, T-test 32
Primjer 4.3. Hi-kvadrat test 34
Primjer 5.1. Multilinearna regresija 37

SAŽETAK

Zadatak završnog rada „Statističke funkcije MS Excela“ jest detaljno opisati biblioteku funkcija MS Excela koje se koriste kod statističke analize. Motivacija za pisanje ovog rada proizlazi iz pitanja može li MS Excel svojim statističkim funkcijama konkurirati drugim programima koji se koriste za provedbu statističke analize.

U radu će biti objašnjene funkcije iz područja deskriptivne statistike u pogledu formiranja tablica frekvencija, izračuna statističkih pokazatelja te izrade grafičkih prikaza, kao i funkcije koje se koriste pri testiranju statističkih hipoteza te korelacijske i regresijske analize.

Svaka funkcija bit će detaljno objašnjena kroz određeni primjer. Svi primjeri bit će riješeni uporabom adekvatne funkcije MS Excela čija će primjena biti detaljno objašnjena riječima i popraćena slikom rješenja u MS Excelu.

Ključne riječi: statističke funkcije, MS Excel, statistička analiza, deskriptivna statistika, statističke hipoteze, korelacijska analiza, regresijska analiza

ABSTRACT

The task of the final paper „Statistical functions of MS Excel“ describes in detail the library of MS Excel functions used in statistical analysis. The motivation for writing this paper starts from the question of whether MS Excel can compete with its statistical functions against other programs used for statistical analysis.

The paper will explain the functions in the field of descriptive statistics in terms of forming a frequency table, calculating statistical indicators and making graphical representations, as well as the functions used in testing statistical hypotheses and correlation and regression analysis.

Each function will be explained in detail through a specific example. All examples will be solved using appropriate functions of MS Excel, the application of which will be explained in detail in words and accompanied by a picture of the solution in ms excel.

Keywords: statistical functions, MS Excel, statistical analysis, descriptive statistics, statistical hypotheses, correlation analysis, regression analysis