УНИВЕРЗИТЕТ У НОВОМ САДУ
ПОЉОПРИВРЕДНИ ФАКУЛТЕТ

СТАТИСТИКА

(за биотехничке смерове)

Др Беба Мутавџић
Мр Емилија Николић Ђорић
Др Драгана Текић
Др Тихомир Новаковић

1

СТАТИСТИКА

ПОЉОПРИВРЕДНИ
ФАКУЛТЕТ
УНИВЕРЗИТЕТ У НОВОМ САДУ

Проф. др Беба Мутавџић
Мр Емилија Николић Ђорић
Др Драгана Текић
Др Тихомир Новаковић

СТАТИСТИКА

(за биотехничке смерове)

Нови Сад, 2023. године

ЕДИЦИЈА: ОСНОВНИ УЏБЕНИК

Оснивач и издавач едиције:

Универзитет у Новом Саду

Пољопривредни факултет Нови Сад

Трг доситеја обрадовића 8

Година оснивања:

1954.

Главни и одговорни уредник едиције:

др Недељко Тица, редовни професор

Декан Пољопривредног факултета

Чланови комије за издавачку делатност:

Др Бранислав Влаховић, редовни професор (председник)

Др Ивана Давидов, ванредни професор

Др Дејан Беуковић, доцент

Др Ксенија Мачкић, доцент

CIP – Каталогизација у публикацији
Библиотека Матице српске, Нови Сад

ISBN 978-86-7520-595-1

COBISS.SR-ID 1

Аутори:

Др Беба Мутавџић, ванреди професор

Мр Емилија Николић Ђорић

Др Драгана Текић, доцент

Др Тихомир Новаковић, доцент

Главни и одговорни уредник:

др Неђељко Тица, редовни професор

Декан Пољопривредног факултета

Уредник:

Др Дејан Јанковић, редовни професор

Директор Департмана за економику пољопривреде и социологију села

Пољопривредни факултет Нови Сад

Рецезенти:

Др Владислав Зекић, редовни професор

Пољопривреди факултет, Универзитет у Новом Саду

Др Отилија Седлак, редовни професор

Економски факултет, Универзитет у Новом Саду

Издавач:

Универзитет у Новом Саду, Пољопривредни факултет Нови Сад

Штампање одобрио:

Комисија за издавачку делатност, Пољопривредни факултет, Нови Сад

Тираж:

20

Одлуком наставно-научног већа Пољопривредног факултет у Новом Саду рукопис је одобрен за издавање као основни уџбеник.

Место и година штампања:

Нови Сад, 2023. године

Предговор

Ова књига је уџбеник за предмет „Статистика”, који се проучава на другој години смера Анимална производња и четвртој години осталих биотехничких смерова Пољопривредног факултета, Универзитета у Новом Саду. Садржај уџбеника у складу је са актуелним акредитованим програмом за наведени предмет и наведене смерове. Књига није намењена само за студенте биотехничких смерова, као основни уџбеник, већ с обзиром на то да је везана за основе статистике, може бити коришћена и од стране студената других смерова Пољопривредног факултета, Универзитета у Новом Саду.

Књига се састоји из шест поглавља, које чине структурну и садржајну целину, која је у складу са акредитованим програмом. Прво поглавље представља ***Увод*** у коме се говори о појму и значају статистике и у оквиру кога су дефинисани основни статистички појмови (статистички скуп, јединице и обележја посматрања, врсте статистичких серија). Друго поглавље је ***Дескриптивна статистика***, у оквиру ког се говори о уређивању и графичком представљању статистичких података, као и о основним статистичким показатељима (показатељи средње вредности, варијабилитета и облика). Треће поглавље књиге односи се на ***Теоријске дистрибуције***, где су најпре дати основни појмови вероватноће, а затим најчешће коришћене прекидне и непрекидне теоријске дистрибуције. Четврто поглавље се односи на ***Инференцијалну статистику,*** где је описан метод узорка у истраживачком раду, као и дистрибуција средина узорака, а наведен је и метод оцена на основу узорака. Пето поглавље обухвата ***Тестирање статистичких хипотеза***, а методолошки су описани и одговарајућим примерима илустровани неки основни тестови аритметичких средина и пропорција. Шесто поглавље је ***Регресиона и корелациона анализа***, у оквиру које је описана проста линеарна регресија и наведене оцене и тестови параметара линеарне регресије.

Аутори се надају да ће ова књига омогућити студентима упознавање са употребом савремених статистичких метода у решавању проблема који су у домену пољопривредних и биолошких наука. Идеја аутора је да се студенти упознају са дескриптивним методама, као и методама анализе резултата огледа. Користећи ову књигу студенти треба да стекну способност за употребу статистичких метода и њихову примену у области свог интересовања. Стечене способности употребе и адекватног коришћења статистике и њених метода омогућиће студентима успешно решавање проблема у даљем раду и стицању образовања.

Захваљујемо се свима који су на директан или индиректан начин помогли израду ове књиге корисним сугестијама, а нарочито рецензентима: проф. др Отилији Седлак и проф. др Владиславу Зекићу.

Нови Сад, АУТОРИ

02.09.2023. године

Садржај

1. УВОД

1.1. Појам и значај статистике

Статистика је саставни део активности научних, образовних, привредних и других институција. Као таква, статистика представља научни метод који се користи за прикупљање, анализу, приказивање и тумачење различитих врста података. Другим речима, статистика представља скуп метода на основу којих је могуће донети веродостојне закључке и одлуке у условима неизвесности.

Порекло речи статистика води од латинске речи *status* – стање, као и *status* – држава. Први пут реч статистика се појављује у првој половини XVIII века у радовима Готфрида Аченвала (*Gottfried Achenwall*), професора Универзитета у Гетингену због чега се сматра оцем статистике. Статистика се односила на скуп нумеричких података о стању посматране појаве.

Једна од основних карактеристика пословног окружења у било којој области су брзе и бројне промене, које прати велика количина података с којима се свакодневно сусрећемо. Познавање извора и квалитета података, њихових карактеристика и правилно тумачење карактеристика, од изузетне су важности у сврху добијања квалитетних информација на основу којих ће се доносити адекватне одлуке. Уколико се до података долази поштујући одређене планске или законски прописане препоруке, прикупљени подаци се сматрају статистичким па је с тога њихово претварање у информације могуће употребом статистичких метода. Сврха примене статистичких метода је доношење закључака о карактеристикама посматраних појава, испитивање различитих претпоставки, процена карактеристичних величина, предвиђање стања и нивоа појава и др.

Примена статистичких метода у облати биолошких истраживања има исту сврху. Анализа резултата генетичких испитивања, извођење и анализа епидемиолошких праћења, дизајн и анализа клиничких истраживања, планирање експеримената у циљу одабира одговарајућег семена, минералног ђубрива, заштитног средства и др., само су неки од случајева који подразумевају примену статистичких метода у биолошким истраживањима.

У складу са претходно наведеним, статистика има два аспекта: *теоријски* и *примењени*. Теоријска или математичка статистика бави се развојем, извођењем и доказивањем теорема, формула, правила и закона, односно усавршавањем нових метода. Теорија вероватноће је фундаментална област на којој је заснована математичка статистика. Примењена статистика подразумева примену нових метода, теорема, формула, правила и закона у решавању реалних проблема.

Статистику делимо на: *дескриптивну* и *инференцијалну* статистику. Дескриптивна статистика обухвата методе прикупљања, сређивања и приказивања података на јасан и разумљив начин, као и израчунавања статистичких параметара. Дескриптивна статистика укључује графичке и нумеричке процедуре за приказивање и анализу података. Инференцијална статистика пружа основу за предвиђање и процену, како би се донели закључци о целокупној популацији на основу података добијених мерења спроведеним на узорку.

1.2. Основни статистички појмови

Предмет истраживања савремене статистике представљају масовне појаве које показују варијабилитет од једног до другог случаја њиховог појављивања. На варијабилитет појаве утиче велики број фактора, при чему сваки од фактора може утицати индивидуално или може имати здружени утицај са другим факторима. У оквиру различитих научних дисциплина варијабилитет посматраних појава анализира се применом адекватне статистичке методологије. Применом одговарајуће методологије стиче се увид у понашање испитиваних појава, уочава њихова повезаност са другим варијабилним појавама, уочавају тенденције у њиховом развоју (кретању) или прогнозирају будуће вредности. Примена статистичке методологије захтева пре свега познавање статистичке терминологије, као и разумевање принципа статистичке анализе.

1.2.1. Статистички скуп

Статистички скуп представља скуп јединица посматрања на основу којих се испитује једно или више променљивих својстава. Посматрана својства другачије се називају: варијабле, обележја, особине или карактеристике. Према обиму, статистички скупови се деле на коначне и бесконачне. Статистички скупови такође могу бити реални и замишљени (хипотетички).

Основни скуп (популација, циљна популација) је скуп података свих јединица (елемената) посматрања чије карактеристике испитујемо. За дефинисање основног скупа (популације) треба да буде позната сврха, односно циљ анализе. Основни скупови се дефинишу појмовно, просторно и временски.

Појмовном дефиницијом скупа утврђује се припадност скупу с обзиром на појам јединице посматрања. Просторном дефиницијом означава се простор којем припадају све јединице основног скупа. Временском дефиницијом одређује се временски интервал или временска тачка за коју су везане све јединице скупа. Број јединица основног скупа назива се величина или обим основног скупа.

У већини случајева, није могуће посматрати или испитати све јединице посматрања, због чега се у анализи користе узорци како би се донели адекватни закључци о основном скупу. *Узорак* представља део основног скупа који је изабран у сврху

извођења статистичке анализе. Неопходно је адекватно одабрати репрезентативан узорак како би се добили прецизни закључци о основном скупу.

1.2.2. Јединице посматрања и њихова обележја

Јединица посматрања основног скупа или узорка, представља одређени субјекат или објекат о којем се прикупљају подаци, односно на којем се одређена појава статистички посматра. Јединице статистичког скупа су појединачни случајеви из којих се статистички скуп састоји. Посматране јединице треба да буду истоврсне али не и истоветне. Циљ посматрања јединица статистичког скупа јесте испитивање диференцираности (различитости) њихових карактеристика (одлика, особина, обележја) и квантитативно изражавање уочених различитости.

Променљива (обележје или варијабла) је особина која се проучава или истражује и која подразумева различите вредности по јединицама посматрања.

Опсервација или податак је вредност променљиве која се односи на једну јединицу посматрања.

Обележја јединица посматрања могу бити: *квалитативна* (атрибутивна, категоријална) и *квантитативна* (нумеричка).

Квалитативна обележја су особине које се уочавају на јединицама посматрања, али се не могу мерити већ се описно исказују. Као такве, не могу узети нумеричке вредности већ се класификују у две или више категорија.

Квантитативна односно нумеричка обележја су особине које се уочавају на јединицама посматрања, могу се мерити и бројчано исказати. Квантитативна (нумеричка) обележја делимо на: *прекидна* (дискретна, дисконтинуирана) и непрекидна (континуирана) обележја.

Прекидна обележја су нумеричка обележја која узимају само одређене вредности са бројне праве. Резултат су пребројавања и узимају целобројне вредности. Непрекидна обележја су нумеричка обележја која могу узети било коју вредност са бројне праве и резултат су мерења. Посматрана обележја могу бити целобројне и/или децималне вредности.

1.2.3. Врсте статистичких серија

Статистички подаци су по правилу многобројни, па није могуће директно доносити закључке о обележјима која се анализирају. Подаци записани редоследом којим се прикупљају пре него што се уреде по величини или групишу, називају се негруписани подаци. Из тог разлога се у првом кораку статистичке анализе приступа уређивању података, а сврха уређивања је да омогући уочавање основних карактеристика анализиране појаве. Уређивањем статистичких података настају статистички низови,

односно статистичке серије. Скупови података се сређују и приказују у виду табела и графикона.

У циљу боље прегледности, поготово ако је број прикупљених података велики, подаци се групишу у класе или групе и одређује се број података у свакој класи односно групи. Приликом груписања података основни проблем је утврђивање критеријума на основу кога ће сви подаци бити сврстани у једнородне групе, које ће бити основа за даљу статистичку анализу. На овај начин добијају се различите врсте статистичких серија.

Груписање података може бити: *географско*, *временско*, *атрибутивно* и *нумеричко*.

Географско груписање може се извести на основу административно-територијалне поделе земље или према неком другом географском критеријуму (на пример, подела на планинске и равничарске крајеве). Овако добијени низови података називају се географске серије података.

Временско груписање података може бити интервално и моментно. Резултат оваквог груписања података представљају временске серије података. Интервалним груписањем, појава се непрекидно прати и региструје. Моментним груписањем добијају се подаци који су резултат посматрања појаве у одређеном моменту времена.

Груписањем сакупљених података по атрибутивним обележјима добијају се атрибутивне (описне) серије података.

Груписањем података по нумеричком обележју настају нумеричке серије података.

Контролна питања:

1. Како се дефинише статистика као научни метод?
2. Како се дели статистика?
3. Шта обухвата дескриптивна статистика?
4. Шта се подразумева под инференцијалном статистиком?
5. Шта је основни скуп или популација?
6. Шта је узорак?
7. Шта је променљива, обележје или варијабла?
8. Како се деле обележја јединица посматрања?
9. Како се деле нумеричка обележја?
10. Навести врсте статистичких серија.

2. ДЕСКРИПТИВНА СТАТИСТИКА

2.1. Формирање дистрибуције фреквенција

Нумеричка статистичка серија представља низ података о обележју које се мери на одређеном броју јединица посматрања и исказује бројчано. Уобичајено је да се бројчане вредности измерене на јединицама посматрања бележе оним редоследом како се до њих долази. На основу таквог низа података тешко је донети било какав закључак о предмету истраживања. Да би се добио прегледнији увид у карактеристике анализиране појаве на основу измерених нумеричких вредности, први задатак је да се утврђене вредности систематизују по неком реду и прикажу у прикладној форми.

Подаци записани редоследом којим се прикупљају пре него што се уреде по величини или групишу, називају се *негруписани* подаци. Ради боље прегледности, поготово ако је њихов број велики, подаци се групишу у класе или групе и одређује се број података у свакој класи односно групи. *Груписани* нумерички подаци називају се дистрибуције фреквенција.

Дистрибуција фреквенција представља табеларно приказивање података, где податке групишемо у две колоне тако да су у првој колони наведене све различите вредности обележја, а у другој колони број јединица наведене вредности обележја. Различите вредности обележја се називају модалитети обележја.

Дистрибуција фреквенција за нумеричке податке садржи два низа података: вредности обележја, приказане појединачним вредностима или групним интервалима и њима одговарајући број јединица посматрања.

У складу са начином на који је исказана вредност обележја разликујемо две врсте дистрибуције фреквенција: *неинтервалне* и *интервалне*. Код неинтервалних вредности обележја наведена је појединачна вредност, док је код интервалне дистрибуције вредност обележја интервал који садржи две или више појединачних вредности.

Број понављања сваке наведене вредности обележја или групе (интервала) обележја назива се *апсолутна фреквенција* (f_i). Апсолутна фреквенција показује колико јединица посматраног скупа има одређени модалитет обележја. На овај начин добија се дистрибуција или расподела фреквенција.

Када обележје има велики број различитих вредности, потребно је груписати их у унапред одређене интервале. Број и величина (ширина) интервала зависе од укупног броја података (N), као и природе самог обележја.

Број групних интервала (k) може се израчунати на основу следећег израза које се још назива и Стургесово правило по америчком статистичару Херберту Стругерсу (*Herbert Sturges*):

$$k = 1 + 3{,}332 \log N$$

где је N укупан број података.

Када је познат потребан број групних интервала, могуће је израчунати и приближно оптималну ширину интервала у ознаци i:

$$i = \frac{X_{max} - X_{min}}{k},$$

где X_{max} и X_{min} представљају највећу и најмању вредност обележја респективно, а k претходно утврђен број групних интервала.

На основу апсолутне фреквенције могу се израчунати *релативне фреквенције* у ознаци p_i и *кумулативне фреквенције* у ознаци F_i.

Релативна фреквенција (структура) добија се као количник апсолутне фреквенције сваке вредности обележја и укупног броја јединица посматрања:

$$p_i = \frac{f_i}{N}\ ,\ \ i = 1,2, \dots , k.$$

Додатно, на основу израчунатих релативних фреквенција може се исказати учешће појединих вредности обележја (s_i) у укупном броју јединица посматрања изражено у процентима:

$$s_i = p_i \times 100\ (\%).$$

За различите потребе анализе, нумеричке серије података је могуће кумулирати тако да се добије нумеричка кумулативна серија, односно *кумулативна фреквенција*. Кумулативна фреквенција одређене вредности обележја добија се сабирањем апсолутних фреквенција свих претходних обележја и апсолутне фреквенције тог обележја:

$$F_i = \sum_{j=1}^{i} f_j\ ,\ \ i = 1,2, \dots , k.$$

У зависности од тога да ли сабирање апсолутних фреквенција почиње од прве или од последње вредности обележја разликују се кумулација испод и кумулација изнад. Самим тим, помоћу кумулативних фреквенција лакше се уочава колики је укупан број јединица посматрања испод или изнад одређене вредности обележја.

На исти начин могуће је изразити и кумулативне вредности релативне фреквенције:

$$F_i^r = \sum_{j=1}^{i} p_j \quad i = 1,2, \dots , k.$$

Пример 1. (прекидно обележје):

Број одвојених парцела земљишта 30 индивидуалних газдинстава дат је у наредној табели:

3	2	4	5	3	1	3	2	6	3	2	3	7	4	3
2	5	1	1	3	4	2	1	3	4	5	3	1	2	6

а) формирати неинтервалну дистрибуцију фреквенција;

б) формирати интервалну дистрибуцију фреквенција ако је $i = 2$;

в) израчунати релативне фреквенције (структуру);

г) формирати кумулативну дистрибуцију фреквенција и кумулацију структуре.

Решење:

а) Први корак приликом формирања неинтервалне дистибуције фреквенција јесте систематизација постојеће серије података, што практично значи поређати вредности од најмање ка највећој. Систематизована серија броја одвојених парцела по газдинству је представљена у наставку:

1	1	1	1	1	2	2	2	2	2	2	3	3	3	3
3	3	3	3	3	4	4	4	4	5	5	5	6	6	7

На основу систематизоване серије података може се уочити да се вредност обележја 1 понавља пет пута, вредност обележја 2 понавља се шест пута, итд. Табеларним приказом с једне стране постојећих вредности обележја (број парцела) и с друге стране фреквенције која представља број понављања појединачних вредности обележја (број газдинстава) добија се неинтервална дистрибуција фреквенција:

Број парцела (X_i)	**Број газдинстава** (f_i)
1	5
2	6
3	9
4	4
5	3
6	2
7	1
Укупно (Σ)	**30**

б) Интервална дистрибуција фреквенција где је ширина интервала $i = 2$, формира ће се тако што ће се у првој колони спојити обележја чије су вредности 1 и 2, 3 и 4, 5 и 6, док ће се обележју 7 прикључити вредност обележја 8 које иако није присутно у серији, обезбеђује прегледност представљене табеле.

Даље ће се у другој колони представити збирно вредности фреквенција за обележја која припадају дефинисаном интервалу. Пример интервалне дистрибуције фреквенција је представљен у наставку:

Број парцела (X_i)	**Број газдинстава** (f_i)
1-2	11
3-4	13
5-6	5
7-8	1
Укупно (Σ)	**30**

в) Релативне фреквенције односно структура серије података, израчунаће се тако што ће се свака фреквенција одговарајућег групног интервала делити са укупним бројем података (p_i). Множењем израчунате структуре са 100, добија се релативна фреквенција изражена у процентима која представља процентуално учешће сваког групног интервала у укупном броју података. У циљу остваривања боље прегледности табеле, колона релативне фреквенције ће бити додата као трећа колона претходно дефинисаној табели интервалне дистрибуције фреквенција, што је представљено у наставку:

Број парцела (X_i)	**Број газдинстава** (f_i)	**Релативна фреквенција (структура)**	
		(p_i)	(s_i)
1-2	11	11/30 = 0,37	0,37×100 = 37%
3-4	13	13/30 = 0,43	0,43×100 = 43%
5-6	5	5/30 = 0,17	0,17×100 = 17%
7-8	1	1/30 = 0,03	0,03×100 = 3%
Укупно (Σ)	**30**	**1**	**100%**

в) На крају, кумулативна дистрибуција фреквенција добија се сабирањем апсолутних фреквенција свих претходних интервала и апсолутне фреквенције посматраног интервала. Као што је већ наведено, разликује се кумулација испод и кумулација изнад. На исти начин се добија и кумулација структуре (испод и изнад), с тим да се овде сабирају релативне фреквенције (у децималном запису или у процентима). Пример формирања кумулативне дистрибуције фреквенција и кумулације структуре представљене су у четвртој и петој колони које су додате последње дефинисаној табели:

Број парцела (X_i)	**Број газдинстава** (f_i)	**Релативна фреквенција (структура)**		**Кумулација**		**Кумулација структуре**	
		(p_i)	(s_i)	**Испод**	**Изнад**	**Испод**	**Изнад**
1-2	11	11/30 = 0,37	0,37×100 = 37%	11	30	37%	100%
3-4	13	13/30 = 0,43	0,43×100 = 43%	11+13=24	30-11=19	80%	63%
5-6	5	5/30 = 0,17	0,17×100 = 17%	24+5=29	19-13=6	97%	20%
7-8	1	1/30 = 0,03	0,03×100 = 3%	29+1=30	6-5=1	100%	3%
Укупно (Σ)	**30**	**1**	**100%**				

Пример 2. (непрекидно обележје):

Бројност ларви по m^2 површине на 20 испитивних парцела представљена је у натавку:

4,6	7,4	0,6	2,8	1,5	3,0	0,5	3,2	3,9	2,5
1,2	1,8	2,3	3,3	4,4	5,5	3,7	6,8	4,3	5,1

Потребно је:

а) Формирати интервалну дистрибуцију ($i = 2$);

б) израчунати релативне фреквенције (структуру);

в) формирати кумулативну дистрибуцију фреквенција и кумулацију структуре.

Решење:

Систематизована серија

0,5	0,6	1,2	1,5	1,8	2,3	2,5	2,8	3,0	3,2
3,3	3,7	3,9	4,4	4,4	4,6	5,1	5,5	6,8	7,4

Интревална дистрибуција, структура и кумулација (структуре):

Број ларви по m^2 (X_i)	**Број парцела** (f_i)	**Релативна фреквенција (структура)**		**Кумулација**		**Кумулација структуре**	
		(p_i)	(s_i)	**Испод**	**Изнад**	**Испод**	**Изнад**
0,01-2,00	5	0,25	25%	5	20	0,25	1,00
2,01-4,00	8	0,40	40%	13	15	0,65	0,75
4,01-6,00	5	0,25	25%	18	7	0,90	0,35
6,01-8,00	2	0,10	10%	20	2	1,00	0,10
Укупно (Σ)	**20**	**1**	**100%**				

2.2. Графичко приказивање статистичких података

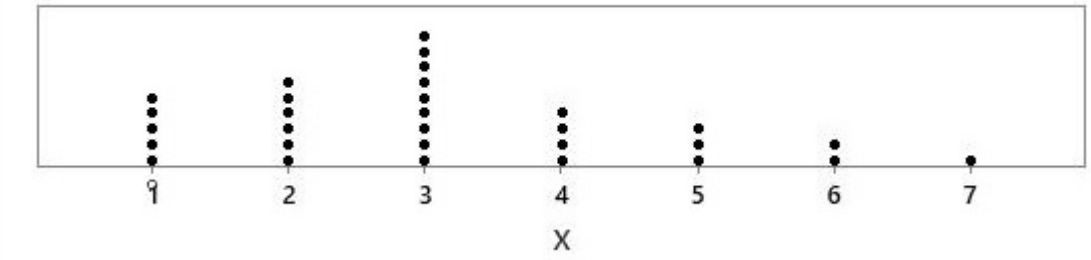
Статистички подаци се приказују помоћу табела и графикона. Графички начин приказивања података омогућава боље уочавање битних карактеристика посматране серије података. Графикони могу бити различитог облика, зависно од природе података односно циља анализе.

Избор одговарајућег графикона зависи од типа расположивих података, начина на који желимо да представимо податке, као и расположивог статистичког софтвера.

У основи разликујемо графиконе који су намењени представљању квантитативних (нумеричких) података, односно квалитативних података (атрибутивне серије). Приликом представљања нумеричких података најчешће се користе *хистограм* и *полигон*. Код атрибутивних серија у употреби су једна или више геометријских површина као што су правоугаоник, квадрат или круг. Најчешће се користе *правоугаоници (стубичасти дијаграм)* и *структурни круг (пита)*. Уколико су предмет анализе временске серије, у употреби је најчешће *линијски дијаграм*. Такође, овде је битно истаћи да постоје и помоћни графикони који могу бити од помоћи приликом систематизације расположивих података, као што су *тачкасти дијаграм* или *дијаграм стабло-лист*. Приликом графичког приказивања података у употреби су и многи други графички прикази, с тим да ћемо се овде ограничити на претходно наведене начине графичког приказивања.

Приликом систематизације расположивих података, где је циљ поређати вредности по реду (од најмање до највеће вредности обележја), узимајући у обзир могућност да се поједине вредности могу понављати, корисни су тачкасти дијаграм и дијаграм стабло-лист. Тачкасти дијаграм се најчешће користи приликом систематизације прекидног обележја. Дијаграм се формира тако што се на *X* осу наносе различите вредности обележја у неопадајућем низу, док се појављивање сваке вредности обележја означава тачком. Тачкасти дијаграм за улазне податке из примера 1. (за прекидно обележје), представљен је у наставку:

Графикон 1. *Тачкасти дијаграм броја парцела по газдинству на основу примера 1*



Слично, приликом систематизације нумеричких података као помоћни дијаграм може послужити дијаграм стабло-лист. Код овог графичког приказа сваки податак делимо на „стабло“ и „лист“. Уколико серију података чине децимални бројеви, стабло обухвата целобројне вредности, а лист вредности децимале. Уколико су вредности серије података двоцифрени бројеви, стабло чине цифре десетица, а листове цифре јединица. Дијаграм стабло-лист може се користити за систематизацију и прекидних и непрекидних нумеричких обележја.

Прва колона дијаграма стабло-лист представља „стабло“. Затим се формираном стаблу придружује колона „лист“. Дијаграм стабло-лист за улазне податке из примера 2. (за непрекидно обележје), представљен је у наставку:

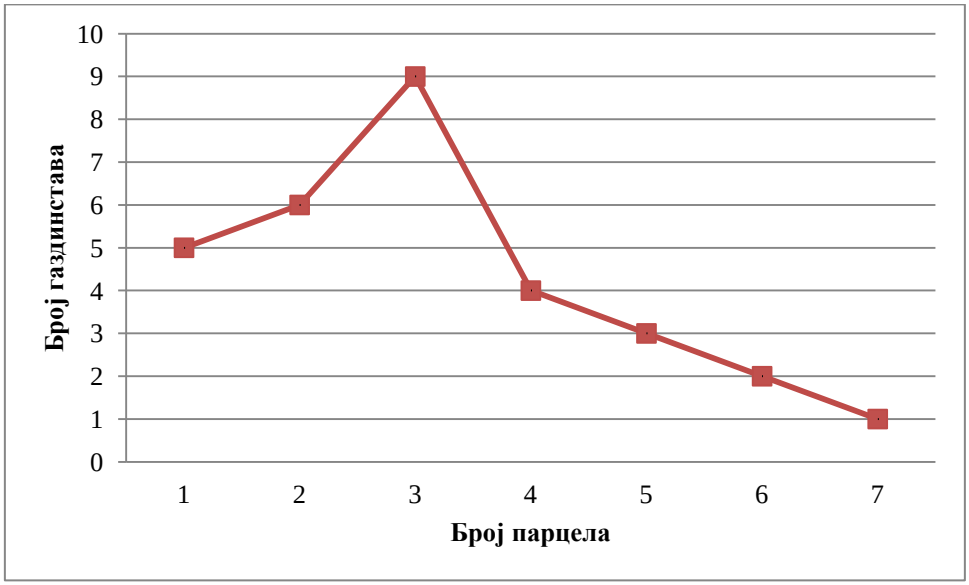
Дијаграм 1. *Дијаграм стабло лист на основуподатака из примера 2*

Стабло	Лист
0	5 6
1	2 5 8
2	3 5 8
3	0 2 3 7 9
4	3 4 6
5	1 5
6	8
7	4

Као што је наведено, приликом графичког приказивања квантитативних (нумеричких) података, у употреби су најчешће хистограм или полигон. Хистограм чине правоугаоници чија је основица једнака величини групног интервала, док висина правоугаоника одговара фреквенцији групног интервала. Полигон је представљен изломљеном линијом која спаја тачке чије су координате вредности обележја или средине групних интервала и одговарајуће фреквенције. Наведени графикони се цртају у правоуглом координатном систему. Код оба начина графичког приказивања нумеричких података, на *X* осу се наносе вредности обележја, док се на *Y* осу наносе вредности фреквенције.

Пример за графички приказ података на основу полигона (пример 1.) представљен је у наставку.

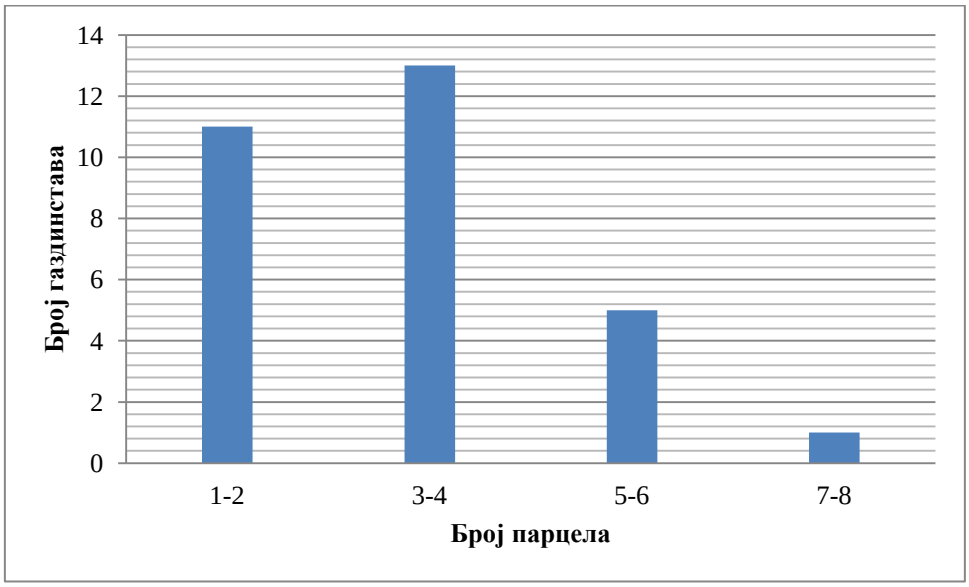
Графикон 2. *Приказ полигона за број парцела по газдинству на основу примера 1*



Извор: *обрада аутора*

Пример за графички приказ података на основу стубичастог (пример 2.) представљен је у наставку:

Графикон 3. *Приказ стубичастог дијаграма за број парцела по газдинству на основу примера 1*



Извор: *обрада аутора*

Атрибутивне серије, односно квалитативни подаци, као што је наведено, најчешће се графички представљају на основу *структурног круга (пита)* или стубичастог дијаграма. Структурни круг представља пун круг који је подељен на више делова у складу са процентуалним учешћем различитих категорија посматраног обележја. Приликом формирања структурног круга потребно је првобитно утврдити процентуално учешће

сваке категорије посматраног обележја. Израчунате проценте је затим потребно помножити са 3,6 (стоти део пуног круга који има 360 степени) како би се добила величина угла, изражена у степенима, кружног исечка стуктурног круга који представља учешће одређене категорије обележја.

С друге стране, стубичасти дијаграм је сличан хистограму за квантитативна обележја. Једина разлика је у томе што се на *X*-оси налазе различите категорије посматраног обележја које је квалитативног типа. Различите категорије посматраног обележја на *X*-оси могу пратити логичан след уколико постоји.

Примери графичког приказивања атрибутивних серија података на основу структурног круга (пите) и стубичастог дијаграма, представљени су у наставку.

Пример 3.

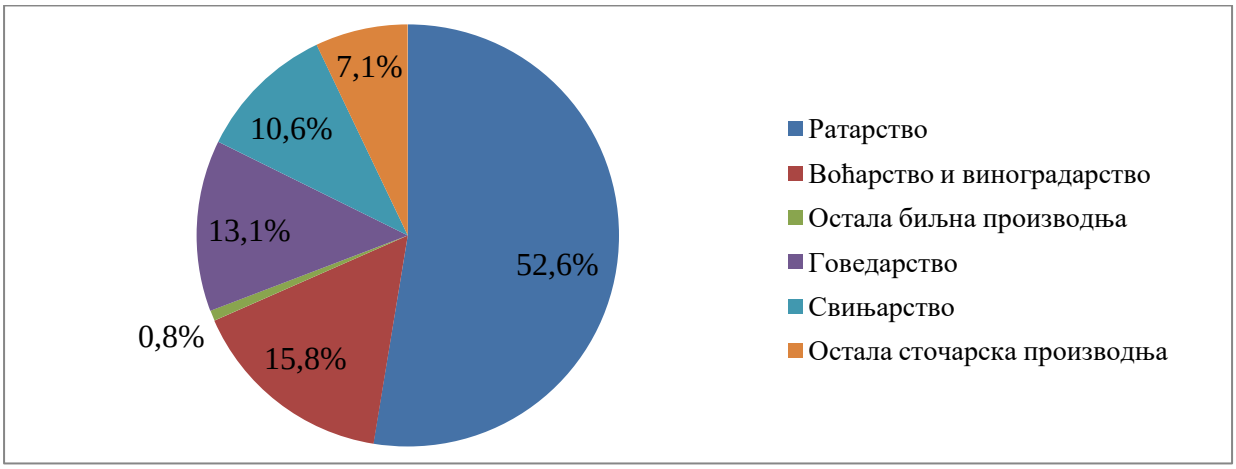
Структура вредности пољопривредне производње у Републици Србији за период 2000-2019. године представљена је у наставку. Потребно је графички представити податке применом структурног круга и хистограм правоугаоника.

Линија производње	**Структура (%)**
Ратарство	52,6
Воћарство и виноградарство	15,8
Остала биљна производња	0,8
Говедарство	13,1
Свињарство	10,6
Остала сточарска произвдња	7,1
Укупно	**100,0**

Извор: обрада аутора на основу података РЗС-а

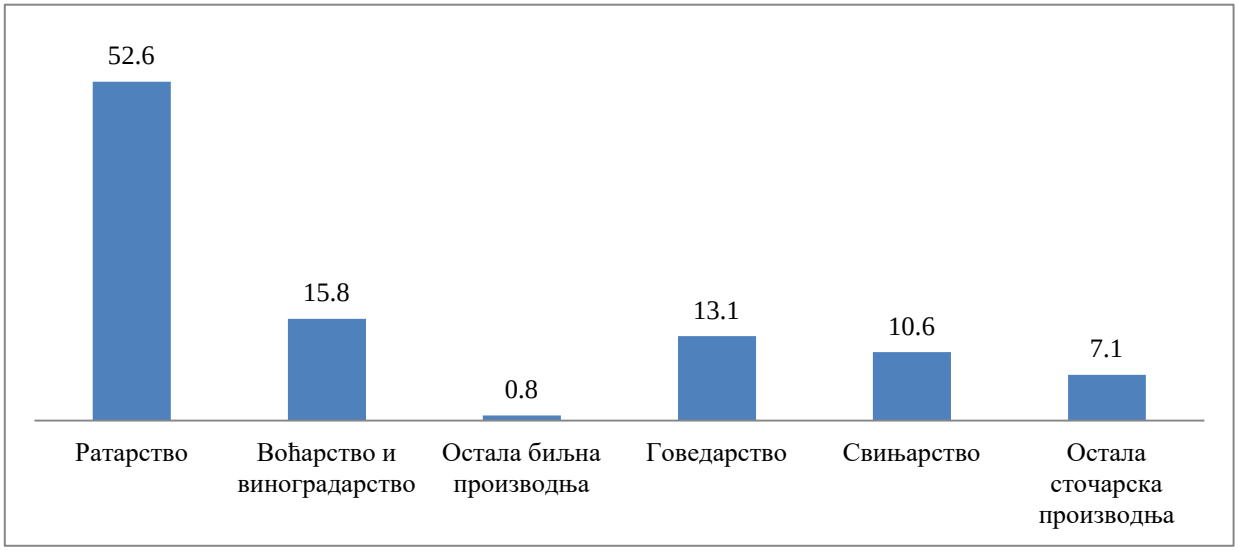
Структура вредности пољопривредне производње у нашој земљи применом структурног круга и стубичастог дијаграма представљена је у наставку.

Графикон 4. Структура вредности пољопривредне производње у Републици Србији за период 2000-2019. године применом структурног круга



Извор: обрада аутора

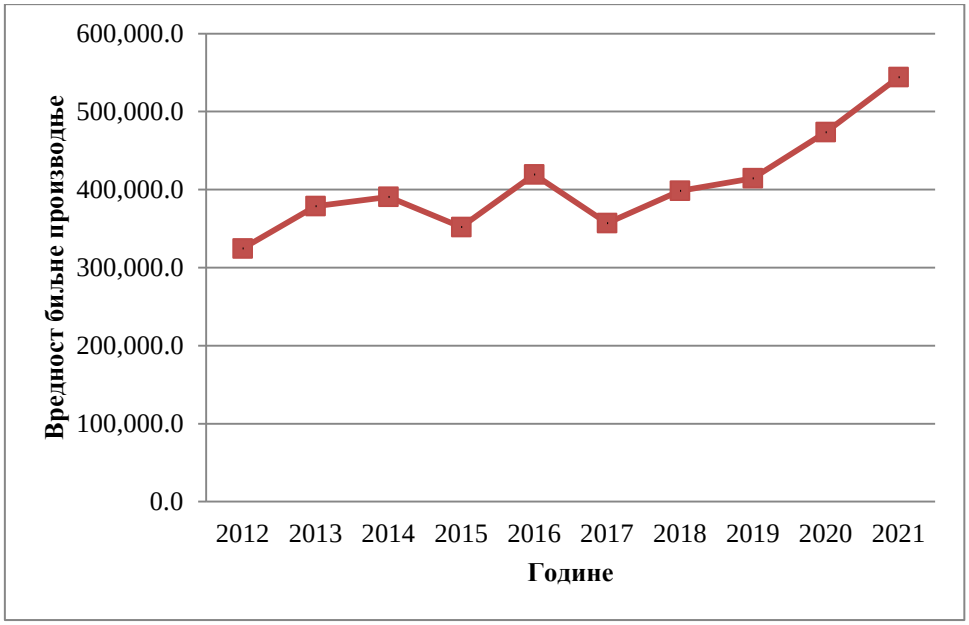
Графикон 5. Структура вредности пољопривредне производње у Републици Србији за период 2000-2019. године применом хистограм правоугоника



Извор: обрада аутора

На крају, графички приказ временских серија, као што је наведено подразумева *линијски дијаграм*. Код линијског дијаграма на *X* осу се наносе посматрани временски периоди (нпр. године), док се на *Y* осу наносе вредности посматране појаве. Пример графичког представљања временске серије података на примеру остварене вредности биљне производње у Републици Србији за период 2012-2021. године представљен је у наставку.

Графикон 6. *Приказ линијског дијаграма на основу вредности биљне производње у Републици Србији за период 2012-2021. године, у милионима РСД*



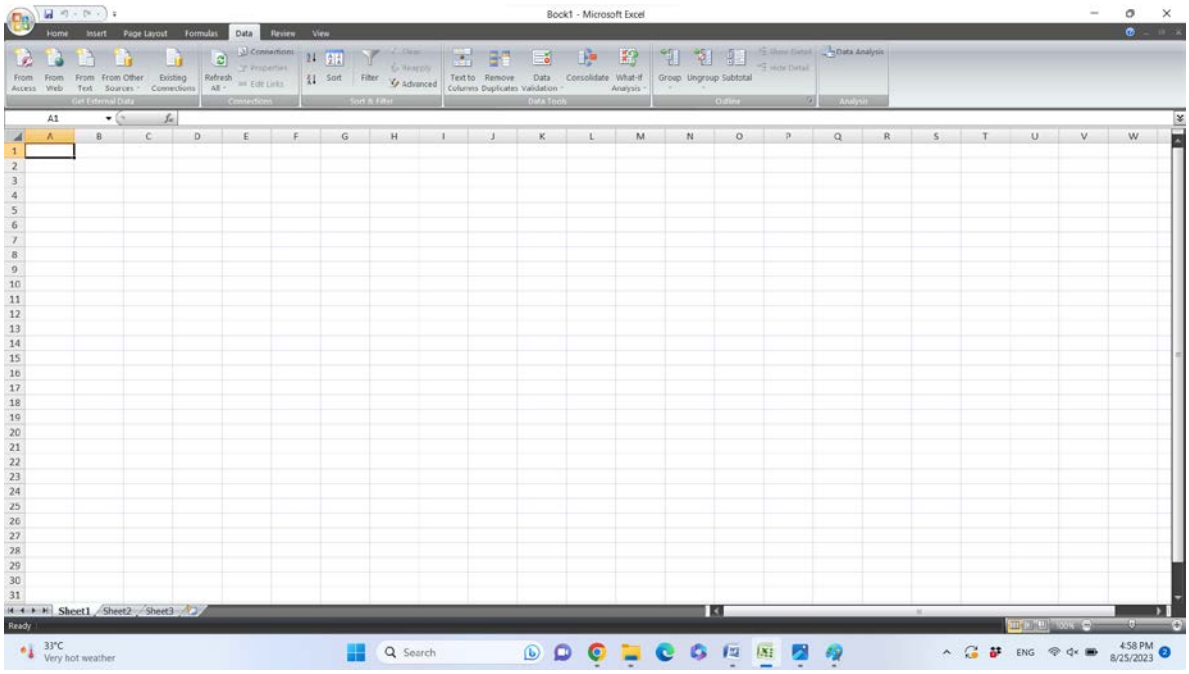
Извор: *Републички завод за статистику Републике Србије*

Вежба 1: Креирање графичких приказа у Microsoft Excel-u

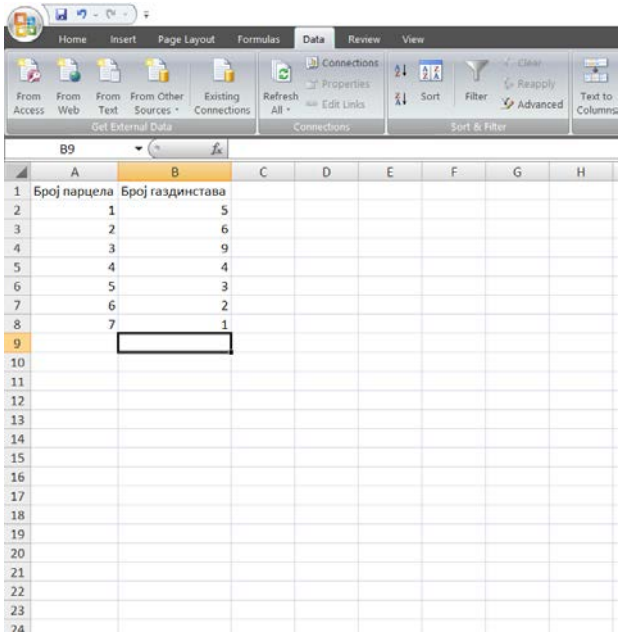
У овој вежби ћемо користити *Microsoft Excel* како би креирали графичке приказе који су објашњени у претходном делу текста.

Извршите следеће кораке:

1. Отворите *Microsoft Excel* празну радну свеску.

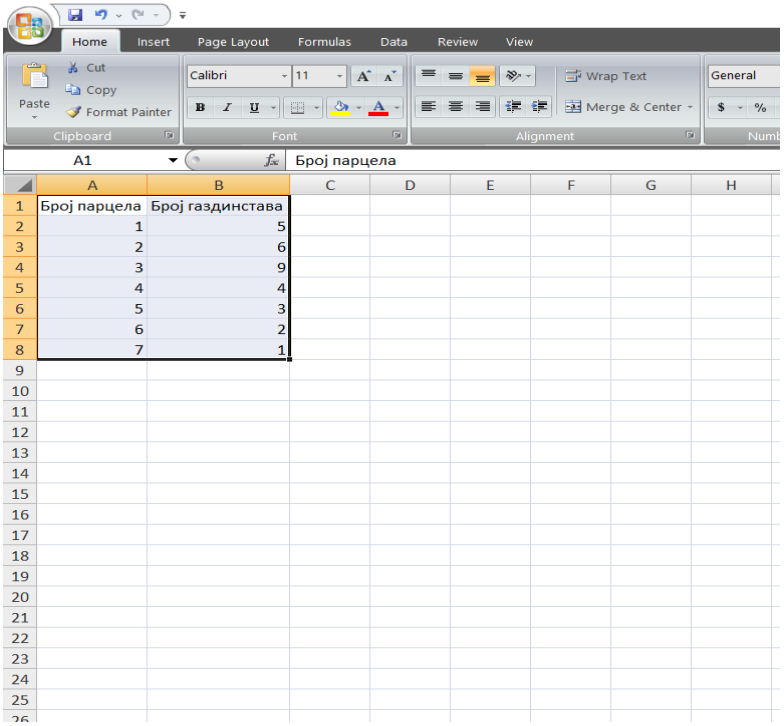


2. Креирајте табелу која ће садржати две колоне. Потребно је да се прва колона односи на вредност обележја, док ће другу колону представљати фреквенција. Искористите податке из примера 1, тако да добијете следећи изглед табеле:



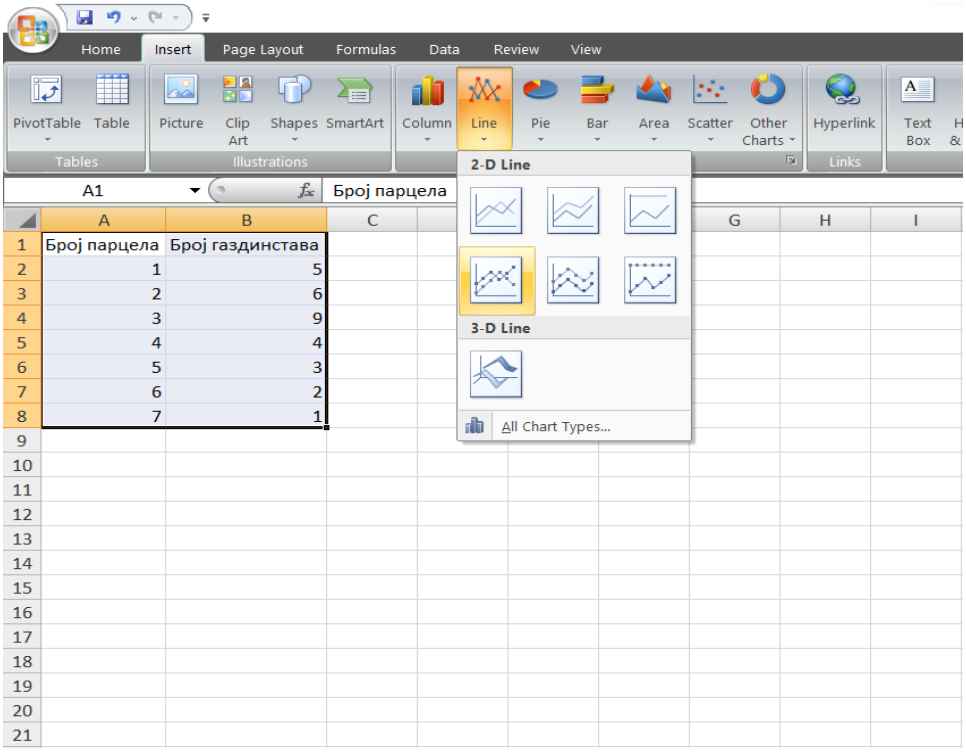
	A	B	C	D	E	F	G	H
1	Број парцела	Број газдинстава						
2	1	5						
3	2	6						
4	3	9						
5	4	4						
6	5	3						
7	6	2						
8	7	1						
9								

3. Како бисте креирали жељени графикон, неопходно је за почетак да обележите све ћелије које обухватају податке на основу којих ће се креирати графички прикази:



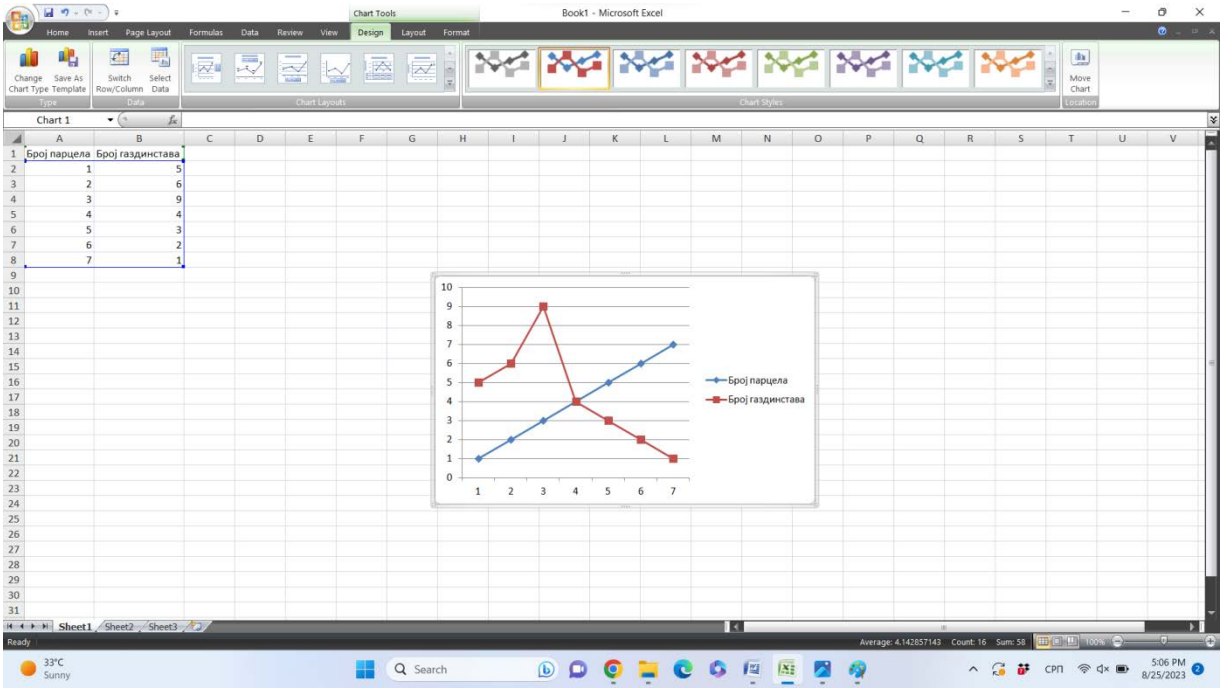
	A	B
1	Број парцела	Број газдинстава
2	1	5
3	2	6
4	3	9
5	4	4
6	5	3
7	6	2
8	7	1

4. Даље је потребно кликнути на картицу *Insert*, а затим у делу који се односи на графичке приказе изабрати жељени графикон. Изабраћемо картицу Line (полигон) са тачкама:

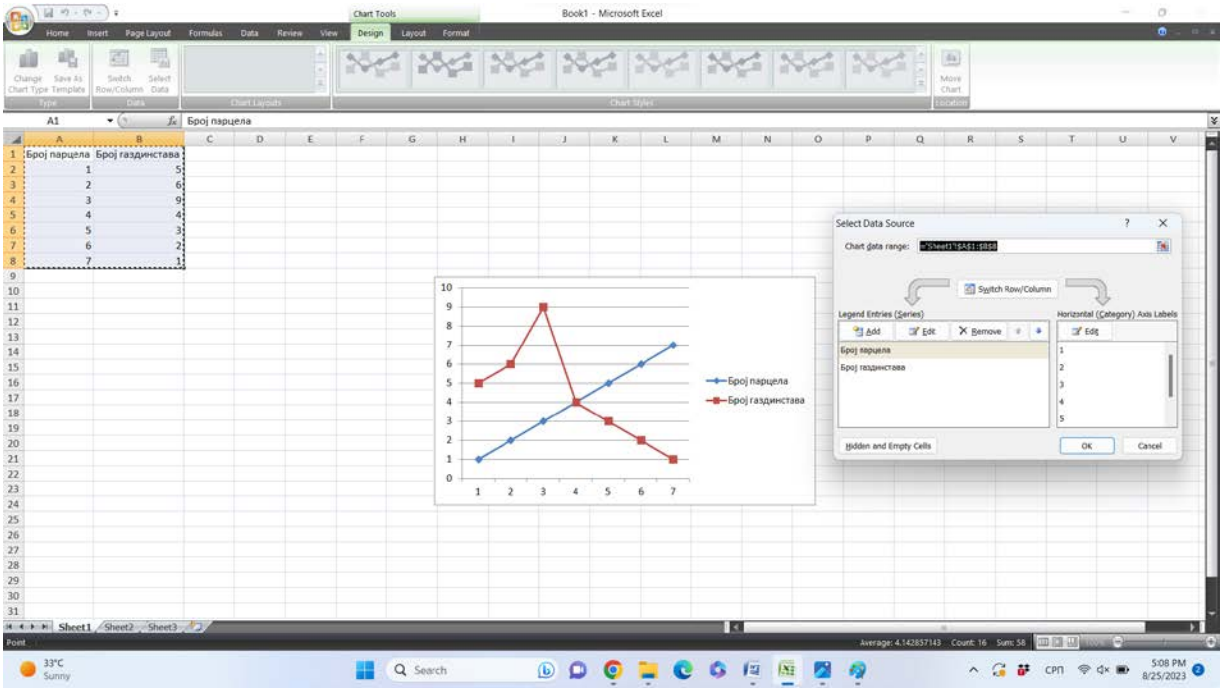


	A	B
1	Број парцела	Број газдинстава
2	1	5
3	2	6
4	3	9
5	4	4
6	5	3
7	6	2
8	7	1

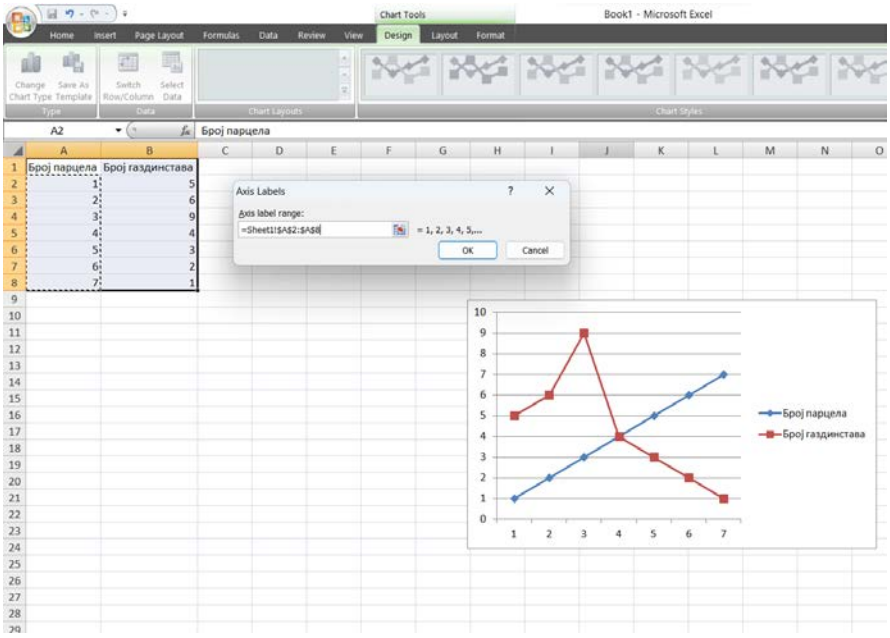
Добије се графички приказ следећег изгледа:



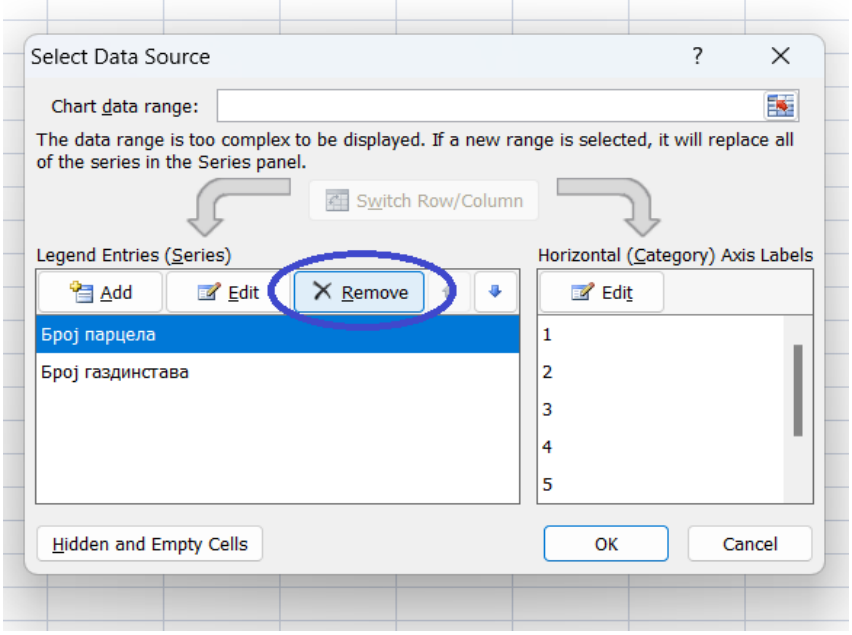
5. Како бисмо уредили посматрани полигон, неопходно је да кликнемо на *Chart Tools*, а затим на *Select Data* и добићемо следећи изглед екрана где ће се отворити нови прозор (*Select Data Source*):



6. У оквиру новоотвореног прозора (*Select Data Source*), неопходно је обратити пажњу на део са заглављем *Horizontal (Category) Axis Labels* који се односи на *X*-осу нашег полигона. Кликом на дугме *Edit* отвара се нови прозор (*Axis Labels*) у оквиру којег је потребно означити нумеричке податке из улазне табле који представљају вредност обележја. У нашем примеру то су подаци од ћелије *A2* до ћелије *A8*. Затим притиснути дугме *OK*:



7. За крај неопходно је у делу са заглављем *Legend Entries (Series)* у оквиру прозора *Select Data Source* кликнути на назив колоне која означава вредност обележја (у нашем случају *Број парцела*) и кликом на дугме *Remove* уклонити:



8. Затим је потребно кликнути на дугме *OK* и добија се полигон следећег изгледа:

Добијени полигон је могуће додатно уредити. Ради прегледности, делове текста у оквиру прозора који приказује полигон могуће је изменити или најједноставније избрисати (означити текст и затим кликнути дугме *Delete*). Такође, кликом на картицу *Design* могуће је додатно променити изглед полигона.

На исти начин могуће је формирати хистограм за квантитативне податке, где ће обележје бити представљено на основу интервалне серије. Када је реч о атрибутивним серијама података (квалитативни подаци), пратећи претходно објашњен поступак, могуће је форморати струтурни круг или хистограм правоугаонике. Код структурног круга, *Microsoft Excel* израчунава проценте учешћа појединих категорија у укупном броју података. Поступак формирања линијског дијаграма за временске серије је исти као поступак формирања полигона, а једина разлика је у томе што ће на *X*-осу бити уписане временске одреднице (нпр. године посматрања).

2.3. Показатељи централне тенденције

Показатељи централне тенденције (средње, просечне вредности) представљају вредности које квантификују тенденцију података у серији према њиховом „центру“, односно средини. Показатељ централне тенденције је репрезентативна вредност која по датим мерилима замењује све вредности обележја у датој серији. Карактерише статистички скуп података и као информација може да замени низ свих вредности серије.

У показатеље централне тенденције убрајају се:

- *Аритметичка средина;*
- *Геометријска средина;*
- *Хармонијска средина;*
- *Медијана;*
- *Квартили;*
- *Модус.*

Према начину утврђивања наведени показатељи централне тенденције деле се у две групе: *израчунате средње вредности* (аритметичка, геометријска и хармонијска средина) и *позиционе средње вредности* (медијана, квартили и модус). Израчунате средње вредности су вредности које се израчунавају на основу свих вредности посматраног обележја, односно свих података у посматраној серији. Позиционе средње вредности су вредности које се утврђују избором конкретне вредности обележја према положају који заузима у посматраној серији података.

Показатељи централне тенденције, односно средње вредности су апсолутни показатељи, њихова вредност се исказује у јединицама мере у којима је исказано и посматрано обележје.

2.3.1. Аритметичка средина

Аритметичка средина је најчешће употребљивани показатељ средње вредности. Разликује се израчунавање просте и пондерисане аритметичке средине. Проста аритметичка средина се утврђује на основу негруписаних нумеричких података, а пондерисана када су подаци груписани у дистрибуцију фреквенција. Аритметичка средина се може израчунавати за податке основног скупа или за податке узорка.

Проста аритметичка средина се израчунава када се све вредности јединица једног посматраног скупа саберу и тај збир подели бројем тих јединица. Аритметичка средина за податке основног скупа означава се са μ и израчунава се на основу следећег израза:

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} \qquad \text{или} \qquad \mu = \frac{\sum_{i=1}^{N} X_i}{N},$$

где је N укупан број података унутар основног скупа, а X_i вредност обележја тако да важи $i = 1,2, \dots, N$.

Артитметичка средина израчуната за податке узорка обележава се са $\bar{X}$, а израчунава се на основу следећег израза:

$$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} \qquad \text{или} \qquad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

где је n ознака за величину посматраног узорка.

Уколико су подаци за анализу груписани, односно ако имамо дистрибуцију фреквенција, тада израчунавамо *пондерисану аритметичку средину*. Пондерисана аритметичка средина добија се на основу збира вредности обележја јединица посматрања које су пондерисане одговарајућим фреквенцијама. Другим речима, неопходно је помножити обележје са припадајућом фреквенцијом, а затим суму свих производа поделити са сумом фреквенција, односно укупним бројем расположивих података.

Пондерисана аритметичка средина основног скупа израчунава се основу израза:

$$\mu = \frac{f_1X_1+f_2X_2+\cdots+f_kX_k}{f_1+f_2+\cdots+f_k} \quad \text{или} \quad \mu = \frac{\sum_{i=1}^{k} f_iX_i}{\sum_{i=1}^{k} f_i},$$

где је f_i фреквенција обележја X_i, k број различитих вредности обележја, а $\sum_{i=1}^{k} f_i = N$.

За податке из узорка, пондерисана аритметичка средина се израчунава на основу сличног израза:

$$\bar{X} = \frac{f_1X_1+f_2X_2+\cdots+f_kX_k}{f_1+f_2+\cdots+f_k} \quad \text{или} \quad \bar{X} = \frac{\sum_{i=1}^{k} f_iX_i}{\sum_{i=1}^{k} f_i},$$

тако да важи: $\sum_{i=1}^{k} f_i = n$.

Особине аритметичке средине су:

1. Приликом израчунавања аритметичке средине учествују све вредности обележја.

2. Аритметичка средина се налази између екстремних вредности обележја, односно већа је од најмање вредности обележја, а мања од највеће вредности обележја у некој серији:

$$X_{min} < \bar{X} < X_{max}$$

3. Уколико су све вредности обележја међусобно једнаке, аритметичка средина је једнака тој вредности.

4. Уколико се свака вредноста обележја увећа или умањи за неку константу (c), аритметичка средина се повећава или смањује за ту константу:

$$X_i^{"} = X_i + c \ (i = 1,2, \dots, n) \rightarrow \bar{X}^{"} = \bar{X} + c;$$

$$X_i^{"} = X_i - c\ (i = 1,2, \dots, n) \rightarrow \bar{X}^{"} = \bar{X} - c.$$

5. Уколико се свака вредност обележја помножи или подели константом (c), аритметичка средина је једнака производу, односно количнику аритметичке средине и те константе:

$$X_i^{"} = X_i \times c\ (i = 1,2, \dots, n) \rightarrow \bar{X}^{"} = \bar{X} \times c;$$

$$X_i^{"} = \frac{X_i}{c}\ (i = 1,2, \dots, n) \rightarrow \bar{X}^{"} = \frac{\bar{X}}{c}.$$

6. Сума одступања свих вредности обележја од њихове аритметичке средине једнака је нули:

$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ (за негруписане податке);

$\sum_{i=1}^{k} f_i\ (X_i - \bar{X}) = 0$ (за груписане податке).

7. Сума квадрата одступања вредности обележја од њихове аритметичке средине је мања од суме квадрата одступања обележја од било које друге вредности c, тако да важи: $c \neq \bar{X}$.

$\sum_{i=1}^{n}(X_i - \bar{X})^2 < \sum_{i=1}^{n}(X_i - c)^2$ (за негруписане податке);

$\sum_{i=1}^{k} f_i(X_i - \bar{X})^2 < \sum_{i=1}^{k} f_i(X_i - c)^2$ (за груписане податке).

2.3.2. Медијана

Медијана представља вредност обележја која сређену серију података дели на два једнака дела. Приликом утврђивања медијане за негруписане податке, претходно је неопходно систематизовати податке, односно рангирати вредности према њиховој величини. Код серије негруписаних података разликује се утврђивање медијане за серије са непарним бројем података (укупан број података није дељив са 2) и за серије са парним бројем података (укупан број података је дељив са 2).

Уколико је број негруписаних података у серији непаран, медијана је једнака средишњој вредности серије и утврђује се на основу следећег израза:

$$Me = X_{\frac{n+1}{2}}.$$

Ако је број негруписаних података у серији паран, медијана је једнака аритметичкој средини два средишња члана:

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}.$$

Код груписаних података (дистрибуција фреквенција), медијана је она вредност обележја која заједно са претходним вредностима садржи бар половину елемената

посматране серије. Утврђивању медијане код дистрибуција фреквенција претходи кумулирање фреквенција.

Уколико су подаци груписани као интервална дистрибуција фреквенција са једнаким групним интервалима, медијану израчунавамо применом кориговане формуле:

$$Me = L + \left(\frac{\frac{n}{2} - F_{med-1}}{f_{med}} \right) \times i \, ,$$

где је:

L - доња граница медијалног интервала;

n - укупан број података;

F_{med-1} - кумулативна вредност интервала који претходи медијалном интервалу;

f_{med} - апсолутна фреквенција медијалног интервала;

i - ширина групног интервала.

2.3.3. Квартили

Квартили су позиционе средње вредности које сређену серију података деле на четири једнака дела (сваки део садржи 25% података серије). Разликујемо три квартила:

Први квартил (Q_1) дели статистичку серију на два дела у размери 1:3. Прецизније, 25% елемената статистичког скупа има вредност мању или једнаку првом квартилу.

Други квартил (Q_2) дели статистичку серију на два једнака дела у размери 1:1. Другим речима, 50% елемената статистичког скупа има вредност мању или једнаку другом квартилу, док 50% елемената статистичког скупа има вредност већу од другог квартила. Други квартил је увек једнак медијани.

Трећи квартил (Q_3) дели статистичку серију на два дела у размери 3:1, односно 75% елемената статистичког скупа има вредност мању или једнаку трећем квартилу.

Код негруписаних података постоје различити начини израчунавања квартила и овде ће бити примењен поступак који се најчешће примењује.

Утврђивању квартила за негруписане податке треба да претходи систематизација, односно рангирање података по њиховој величини. Код серије негруписаних података разликује се утврђивања квартила за серије података када укупан број података није дељив са четири и за серије података када је укупан број података дељив са четири.

У случају да укупан број негруписаних података у серији, није дељив са четири, квартили се утврђују на основу следећих израза:

$$Q_1 = X_{\frac{n}{4}+1} \qquad Q_2 = Me = X_{\frac{n+1}{2}} \qquad Q_3 = X_{\frac{3n}{4}+1} \, .$$

С друге стране, уколико је укупан број негруписаних података у серији дељив са четири, квартити се утврђују на основу следећих израза:

$$Q_1 = \frac{X_{\frac{n}{4}} + X_{\frac{n}{4}+1}}{2} \qquad Q_2 = Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \qquad Q_3 = \frac{X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}}{2} \ .$$

Код груписаних података (дистрибуција фреквенција), први квартил је она вредност обележја која заједно са претходним вредностима садржи 25% елемената посматране серије, други квартил, тј. медијана је она вредност обележја која заједно са претходним вредностима садржи бар половину елемената посматране серије, док је трећи квартил она вредност која заједно са претходним вредностима садржи 75% елемената посматране серије. Утврђивању квартила код дистрибуције фреквенција претходи кумулирање фреквенција.

Ако су подаци груписани као интервална дистрибуција фреквенција са једнаким групним интервалима, квартиле израчунавамо применом коригованих формула:

$$Q_1 = L + \left(\frac{\frac{n}{4} - F_{Q_1-1}}{f_{Q_1}}\right) \times i \qquad \text{и} \qquad Q_3 = L + \left(\frac{\frac{3n}{4} - F_{Q_3-1}}{f_{Q_3}}\right) \times i \ ,$$

где је:

L - доња граница групног интервал који заједно са претходним интервалима садржи четвртину елемената укупног броја података (Q_1) или доња граница групног интервал који са претходним интервалима садржи три четвртине елемената укупног броја података (Q_3).

n - укупан број података;

F_{Q_1-1} - кумулативна вредност интервала који претходи интервалу који садржи први квартил;

F_{Q_3-1} - кумулативна вредност интервала који претходи интервалу који садржи трећи квартил;

f_{Q_1} - апсолутна фреквенција интервала који садржи први квартил;

f_{Q_3} - апсолутна фреквенција интервала који садржи трећи квартил;

i - ширина групног интервала.

2.3.4. Модус

Модус је најучесталија вредност обележја у некој серији података. Модална вредност се може утврдити ако у серији података постоје барем две једнаке вредности обележја. Ако у серији података постоји само једна вредност обележја чија је фреквенција већа од осталих вредности, кажемо да је та серија унимодална. Поред наведеног, серија

података може бити бимодална (два модуса) или може имати више од две модалне вредности.

Код унимодалне интервалне серије дистрибуције фреквенција, приближна вредност модуса може се израчунати на основу следећег израза:

$$Mo = L + \left(\frac{d_1}{d_1 + d_2}\right) \times i\ ,$$

где је:

L - доња граница модалног интервала;

d_1 – разлика између фреквенције модалног интервала и фреквенције интервала који претходи модалном интервалу;

d_2 - разлика између фреквенције модалног интервала и фреквенције интервала који следи након модалног интервала;

i - ширина групног интервала.

Пример 4.

Дати су подаци о приносу пшенице и кукуруза (тоха/хектару) једног пољопривредног предузећа. Принос пшенице је измерен на седам одвојених парцела, док је принос кукуруза измерен на осам одвојених парцела. Израчунати просечан принос (аритметичка средина), медијану, квартиле и модус за пшеницу и кукуруз одвојено.

Принос пшенице (т/ха)	4,0	3,8	4,0	4,2	4,4	4,7	4,6	-
Принос кукуруза (т/ха)	8,0	7,5	7,1	8,1	7,7	7,7	9,0	8,1

Решење:

С обзиром на то да се су посматрани подаци који се односе на принос пшенице и кукуруза у једном пољопривредном предузећу негруписани, први корак јесте систематизација постојећих података. У наставку је представљен обрачун тражених показатеља централне тенденције, најпре за пшеницу, а затим и за кукуруз. Рангиране вредности оствареног приноса пшенице са 7 различитих парцела ($n = 7$) једног пољопривредног предузећа, представљени су у наставку:

Принос пшенице (т/ха)	3,8	4,0	4,0	4,2	4,4	4,6	4,7	-

Проста аритметичка средина (пшеница):

$$\bar{X} = \frac{\sum_{i=1}^{7} X_i}{n} = \frac{3{,}8 + 4{,}0 + 4{,}0 + 4{,}2 + 4{,}4 + 4{,}6 + 4{,}7}{7} = \frac{29{,}7}{7} = 4{,}24 \text{ тона/хектару}$$

Медијана (пшеница):

$$Me = X_{\frac{n+1}{2}} = X_{\frac{7+1}{2}} = X_4 = 4{,}2 \text{ тона/хектару}$$

Квартили (пшеница):

$$Q_1 = \mathrm{X}_{\frac{\mathrm{n}}{4}+1} = \mathrm{X}_{\frac{7}{4}+1} = \mathrm{X}_{2,75} = \mathrm{X}_2 = 4{,}0 \text{ тона/хектару}$$

Приликом утврђивања првог квартила, установљено је да се први квартил налази на 2,75. позицији ($\mathrm{X}_{2,75}$) у уређеној статистичкој серији. Код квартила је правило да се редни број обележја заокружи на постојећи цео број, без обзира на децималну вредност. Самим тим, $\mathrm{X}_{2,75}$ значи да је први квартил X_2.

$$Q_2 = Me = 4{,}2 \text{ тона/хектару}$$

$$Q_3 = \mathrm{X}_{\frac{3\mathrm{n}}{4}+1} = \mathrm{X}_{\frac{3\times 7}{4}+1} = \mathrm{X}_{6,25} = \mathrm{X}_6 = 4{,}6 \text{ тона/хектару}$$

Модус (пшеница):

$$M_O = 4{,}0 \text{ тона/хектару}$$

Рангиране вредности оствареног приноса кукуруза са 8 различитих парцела ($n = 8$) једног пољопривредног предузећа, представљени су у наставку:

Принос кукуруза (т/ха)	7,1	7,5	7,7	7,7	8,0	8,1	8,1	9,0

Проста аритметичка средина (кукуруз):

$$\bar{X} = \frac{\sum_{i=1}^{8} X_i}{n} = \frac{7{,}1 + 7{,}5 + 7{,}7 + 7{,}7 + 8{,}0 + 8{,}1 + 8{,}1 + 9{,}0}{8} = \frac{63{,}2}{8} = 7{,}9 \text{ тона/хектару}$$

Медијана (кукуруз):

$$Me = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{\frac{8}{2}} + X_{\frac{8}{2}+1}}{2} = \frac{X_4 + X_5}{2} = \frac{7{,}7 + 8{,}0}{2} = 7{,}85 \text{ тона/хектару}$$

Квартили (кукуруз):

$$Q_1 = \frac{X_{\frac{n}{4}} + X_{\frac{n}{4}+1}}{2} = \frac{X_{\frac{8}{4}} + X_{\frac{8}{4}+1}}{2} = \frac{X_2 + X_3}{2} = \frac{7{,}5 + 7{,}7}{2} = 7{,}6 \text{ тона/хектару}$$

$$Q_2 = Me = 7{,}85 \text{ тона хектару}$$

$$Q_3 = \frac{X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}}{2} = \frac{X_{\frac{3\times 8}{4}} + X_{\frac{3\times 8}{4}+1}}{2} = \frac{X_6 + X_7}{2} = \frac{8{,}1 + 8{,}1}{2} = 8{,}1 \text{ тона/хектару}$$

Модус (кукуруз):

$$M_o^1 = 7{,}7 \text{ тона/хектару;}$$

$$M_o^2 = 8{,}1 \text{ тона/хектару}$$

С обзиром на то да су у серији података присутне две вредности обележја које се понављају највећи број пута (вредност 7,7 и вредност 8,1 понављају се по два пута),

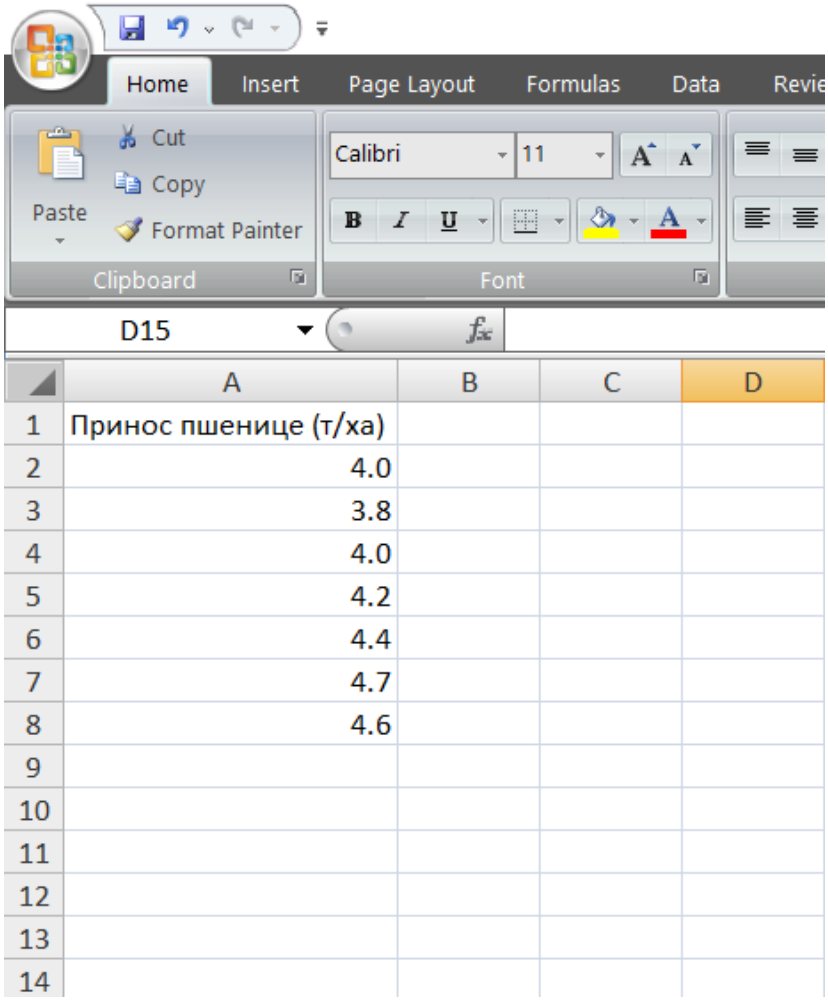
серија података која се односи на принос кукуруза има два модуса (бимодална серија података).

Вежба 2. Дескриптивна статистика за негруписане податке у Microsoft Excel-у

Ради једноставности представљања спроведене анализе у наставку ће бити приказана дескриптивна статистика за податке који се односе на принос пшенице у *Microsoft Excel-у.*

Извршите следеће кораке:

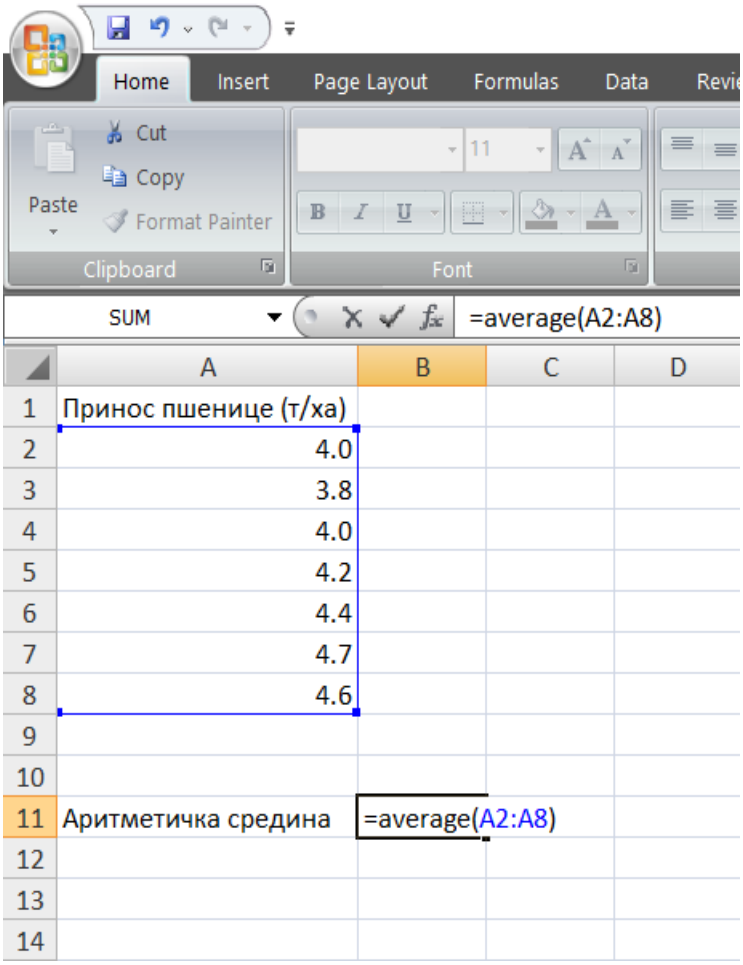
1. Отворите *Microsoft Excel* радну свеску и креирајте табелу која ће садржати вредности оствареног приноса пшенице као што је представљено у наставку:



	A	B	C	D
1	Принос пшенице (т/ха)			
2	4.0			
3	3.8			
4	4.0			
5	4.2			
6	4.4			
7	4.7			
8	4.6			
9				
10				
11				
12				
13				
14				

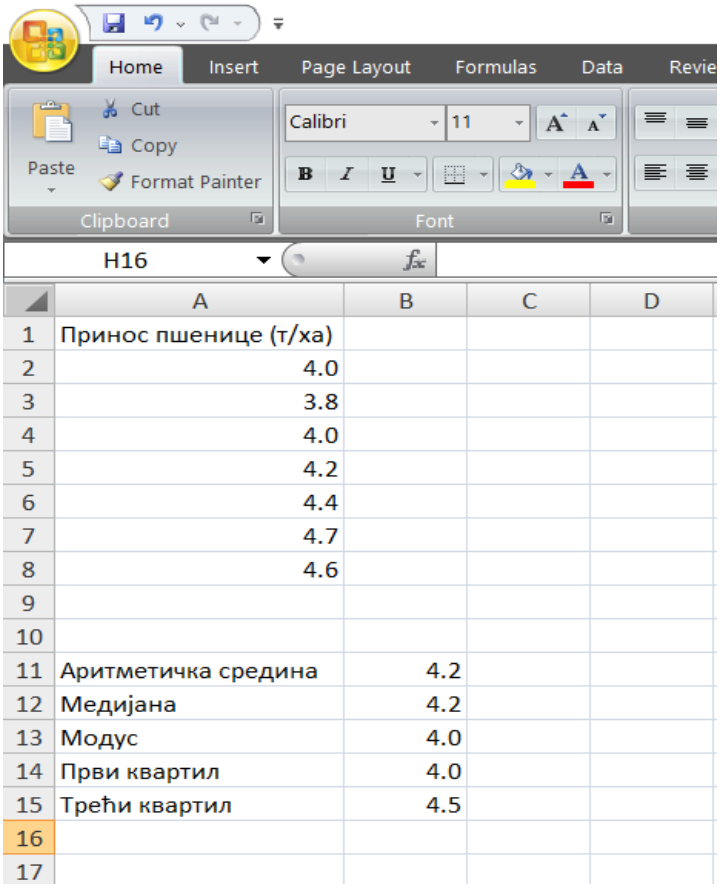
2. У циљу израчунавања основних показатеља дескриптивне статистике, односно аритметичке средине, медијане, квартила и модуса потребно је пратити следеће кораке:

Како би се израчунала **аритметичка средина** потребно је одредити празну ћелију у коју ће у складу са конкретним примером бити уписан код *=average(A2:A8)*, као што је представљено у наставку:



Знак „=“ (или „+“)неопходно је укуцати пре почетка било које рачунске операције у *Microsoft Excel-y*. Ознака „*average*“ је команда за израчунавање аритметичке средине, док се у заграду уписује опсег података који су предмет анализе (у конкретном примеру подаци у колони *A* и то од другог до осмог реда, односно од *A2* до *A8*). Кликом на дугме „*Enter*“ на тастатури, добија се вредност аритметичке средине која износи 4,24 т/ха.

3. На сличан начин могу се израчунати медијана, модус квартили. Приликом израчунавања медијане у конкретном примру, потребно је у празну ћелију уписати код: *=MEDIAN(A2:A8)*. За израчунавање модуса потребно је уписати код: *=MODE(A2:A8)*. Приликом израчунавања квартила, потребно је одвојено израчунати први и трећи квартил, с тим да је други квартил једнак медијани па га није потребно поново рачунати. У циљу израчунавања првог квартила потребно је уписати следећи код: *=QUARTILE(A2:A8,1)*. Једина новина јесте уписивање редног броја квартила унутар заграде у којој је уписан опсег података. Како се ради у првом квартилу, после зареза је уписан број 1. На исти начин могуће је израчунати и трећи квартил, једино че се уместо 1 уписати вредност 3 унутар заграде. На тај начин добијају се вредности као што је представљено у наставку, што уједно представља и крај анализе за конкретан пример:



	A	B	C	D
1	Принос пшенице (т/ха)			
2	4.0			
3	3.8			
4	4.0			
5	4.2			
6	4.4			
7	4.7			
8	4.6			
9				
10				
11	Аритметичка средина	4.2		
12	Медијана	4.2		
13	Модус	4.0		
14	Први квартил	4.0		
15	Трећи квартил	4.5		
16				
17				

У програму *Microsoft Excel* квартили се израчунавају као пондерисана сума $(1-f)\times X_i + f \times X_{i+1}$ где је пондер f одређен разлагањем $(n-1)\times p + 1 = i + f$ при чему је i цео део, f разломљени део броја а p пропорција вредности мањих од квартила.

Тако су у примеру о приносу пшенице вредности квартила:

$(7-1)\times 0{,}25 + 1 = 2{,}5 = 2 + 0{,}5,\ i = 2, f = 0{,}5$, тако да је први квартил:

$$Q_1 = (1-0{,}5)\times X_2 + 0{,}5 \times X_3 = 0{,}5 \times 4 + 0{,}5 \times 4 = 4 \text{ тона/хектару.}$$

$(7-1)\times 0{,}5 + 1 = 4 = 4 + 0,\ i = 4, f = 0$, тако да је други квартил (медијана):

$$Q_2 = 1 \times X_4 + 0 \times X_5 = 4{,}2 \text{ тона/хектару.}$$

$(7-1)\times 0{,}75 + 1 = 5{,}5 = 5 + 0{,}5,\ i = 5, f = 0{,}5$, тако да је трећи квартил:

$$Q_3 = (1-0{,}5)\times X_5 + 0{,}5 \times X_6 = 0{,}5 \times 4{,}4 + 0{,}5 \times 4{,}6 = 4{,}5 \text{ тона/хектару.}$$

У овом примеру су вредности првог и другог квартила исте док се вредност трећег квартила незнатно разликује у поређењу са израчунатом вредношћу у примеру 4 који се односе на принос пшенице. У примеру са кукурузом вредност првог квартила израчуната програмом је $Q_1 = 7{,}65$ тона/хектару и незнатно се разликује од вредности израчунате у примеру 4.

Пример 5.

На основу података о величини поседа 30 пољопривредних газдинстава, израчунати аритметичку средину, медијану, квартиле и модус.

Величина поседа (хектари) (X_i)	**Број газдинстава** (f_i)	**Средина интервала** (X_i')	$f_i \times X_i$	**Кумулатив (испод)**
0,1-2,0	5	1,0	5×1 = 5	5
2,1-4,0	7	3,0	7×3 = 21	12
4,1-6,0	10	5,0	10×5 = 50	22
6,1-8,0	5	7,0	5×7 = 35	27
8,1-10,0	3	9,0	3×9 = 27	30
Σ	**30**		**138**	

Решење:

Пондерисана аритметичка средина:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{138}{30} = 4{,}6 \text{ хектара}$$

Медијана:

$$Me = L + \left(\frac{\frac{n}{2} - F_{med-1}}{f_{med}}\right) \times i = 4{,}0 + \left(\frac{\frac{30}{2} - 12}{10}\right) \times 2 = 4{,}6 \text{ хектара}$$

Квартили:

$$Q_1 = L + \left(\frac{\frac{n}{4} - F_{Q_1-1}}{f_{Q_1}}\right) \times i = 2{,}0 + \left(\frac{\frac{30}{4} - 5}{7}\right) \times 2 = 2{,}71 \text{ хектара}$$

$$Q_2 = Me = 4{,}7 \text{ хектара}$$

$$Q_3 = L + \left(\frac{\frac{3n}{4} - F_{Q_3-1}}{f_{Q_3}}\right) \times i = 6{,}0 + \left(\frac{\frac{3 \times 30}{4} - 22}{5}\right) \times 2 = 6{,}2 \text{ хектара}$$

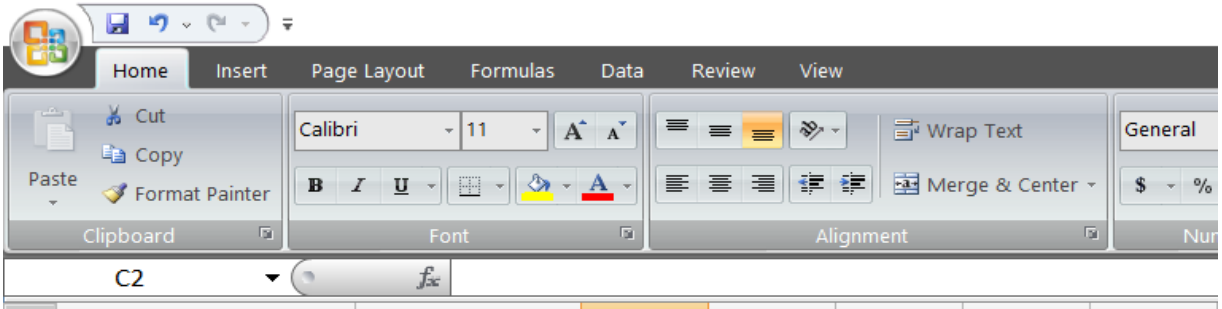
Модус:

$$Mo = L + \left(\frac{d_1}{d_1 + d_2}\right) \times i = 4{,}0 + \left(\frac{3}{3+5}\right) \times 2 = 4{,}75 \text{ хектара}$$

Вежба 3. Дескриптивна статистика за груписане податке у Microsoft Excel-y

Уколико се располаже са груписаним подацима, у *Microsoft Excel-y* није могуће универзалним формулама израчунавати основне показатеље дескриптивне статистике, већ је израчунавање могуће спровести уводећи горе наведене формуле. Цео поступак израчунавања је мало комплекснији и одузима више времена. Ипак, када се располаже са већим бројем подацима, поступак представљен у наставку може бити изузетно користан.

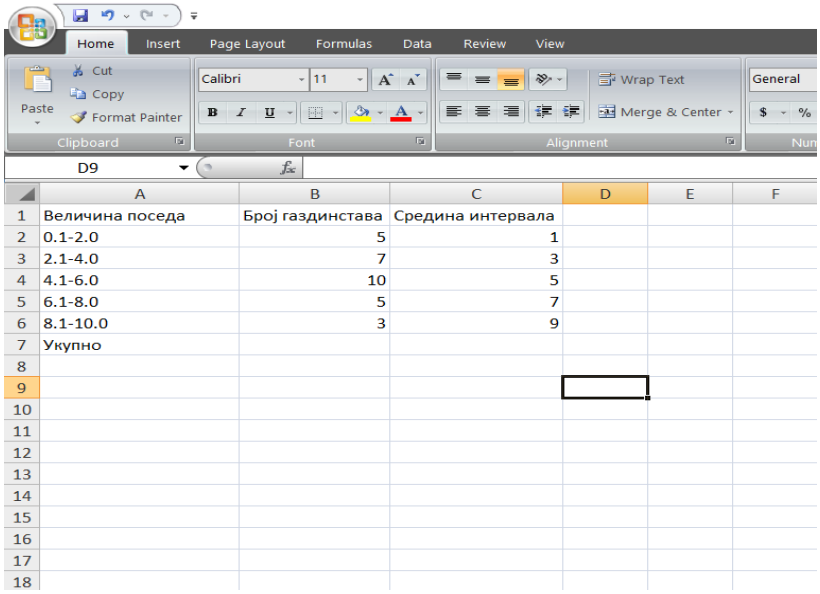
1. У празну *Microsoft Excel* свеску неопходно је ручно унети податке који се односе на вредност обележја и фреквенцију. У нашем примеру, обележје се односи на величину поседа пољопривредних газдинстава, док је фреквенција број газдинстава. Почетни изглед табеле је представљен у наставку:



	A	B	C	D	E	F	G
1	Величина поседа	Број газдинстава					
2	0.1-2.0	5					
3	2.1-4.0	7					
4	4.1-6.0	10					
5	6.1-8.0	5					
6	8.1-10.0	3					
7	Укупно						
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							

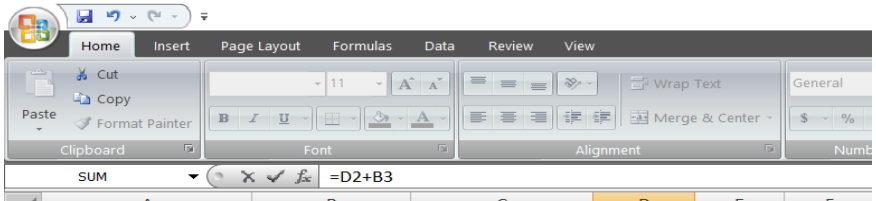
Извођење било каквих рачунских операција са представљеним интервалним вредностима у колони која се односи на обележје није могуће. Разлог је тај што у свакој појединачној ћелији у *Microsoft Excel-y* мора бити уписан искључиво један број. С тим у вези, неопходно је у посебној колони издвојити средине интервала, што ће представљати обележје са којим ће се даље изводити анализа.

2. Средине интервала можемо једноставно израчунати тако што ћемо збир доње и горње границе сваког појединачног интервала поделити са два. Ради једноставности представљене анализе, добијене средине интервала су заокружене на цео број. На тај начин, потребно је добити следећи изглед табеле:



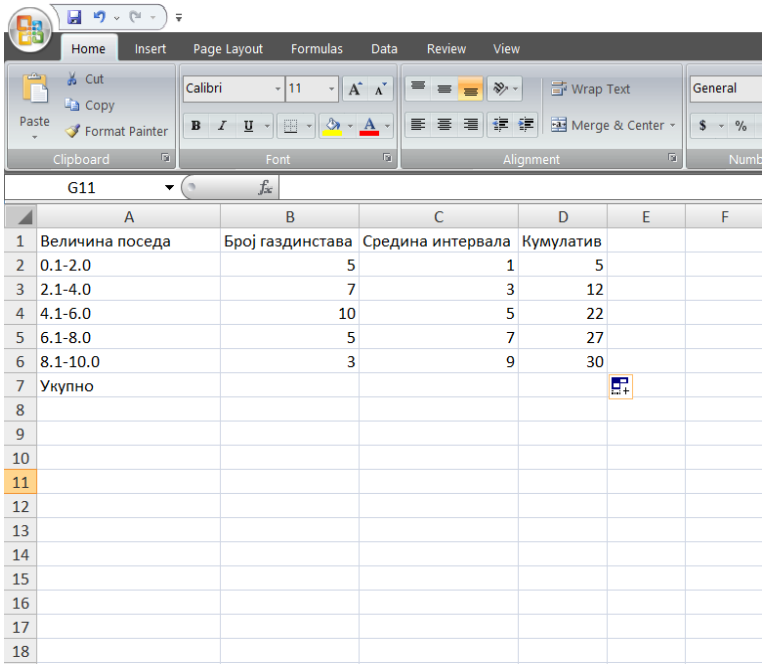
	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала			
2	0.1-2.0	5	1			
3	2.1-4.0	7	3			
4	4.1-6.0	10	5			
5	6.1-8.0	5	7			
6	8.1-10.0	3	9			
7	Укупно					
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

3. Како ће за потребе израчунавања медијане и квартила бити потребна колона кумулатива, неопходно је табелу проширити и за ову колону. Колону кумулатива је могуће израчунати тако што ће се у нову колону (у нашем примеру колону *D*), најпре у поље *D2* преписати вредност фреквенције за први групни интервал, односно уписати команда *=B2*. Затим је потребно у поље *D3* уписати код који ће сабрати претходно поље из колоне *D* (поље *D2*) и фреквенцију другог групног интервала (поље *B3*), као што је представљено у наставку:



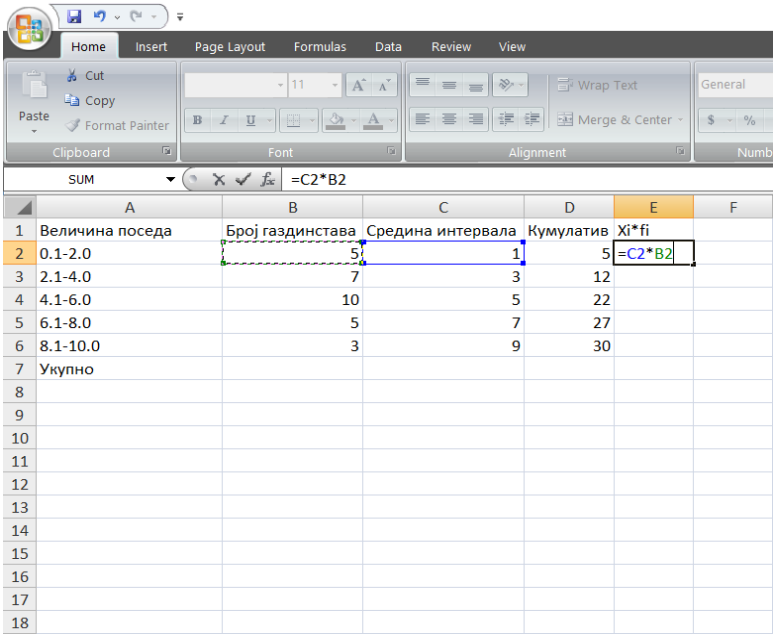
	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив		
2	0.1-2.0	5	1	5		
3	2.1-4.0	7	3	=D2+B3		
4	4.1-6.0	10	5			
5	6.1-8.0	5	7			
6	8.1-10.0	3	9			
7	Укупно					
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

4. Након клика на дугме „*Enter*“, потребно је довести курсор миша у доњи десни угао поља *D3* и превући на доле. На тај начин, *Microsoft Excel* рачуна остале вредности кумулатива у колони *D* на основу формуле уписане у поље *D3*. Добијени изглед табеле је следећи:



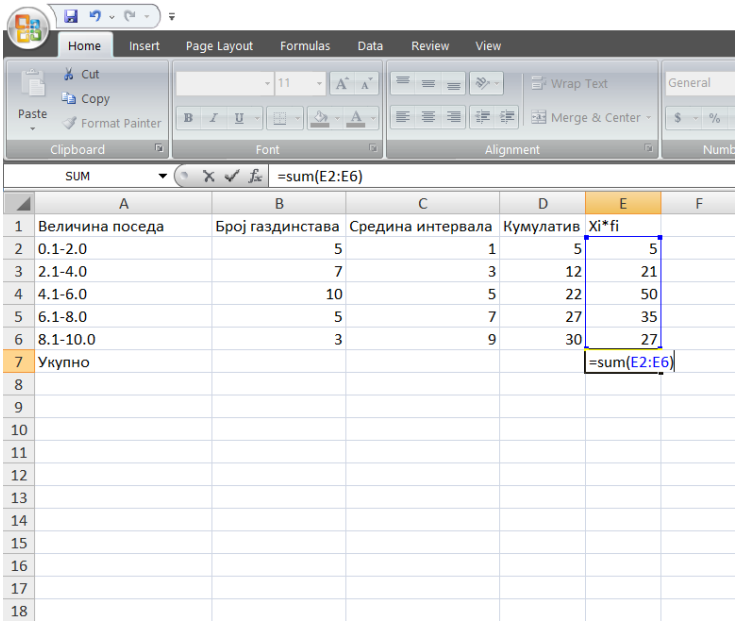
	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив		
2	0.1-2.0	5	1	5		
3	2.1-4.0	7	3	12		
4	4.1-6.0	10	5	22		
5	6.1-8.0	5	7	27		
6	8.1-10.0	3	9	30		
7	Укупно					
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

5. У циљу израчунавања аритметичке средине, с обзиром на то да се ради о груписаним подацима, потребно је искористити следећу формулу: $\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$. Дакле, потребно је формирати колону која представља производ посматраног обележја и припадајуће фреквенције. Слично као и приликом израчунавања кумулатива, потребно је у нову колону (колона *E*), у поље *E2* уписати следећи код: *=C2*B2*. Изглед табеле је следећи:



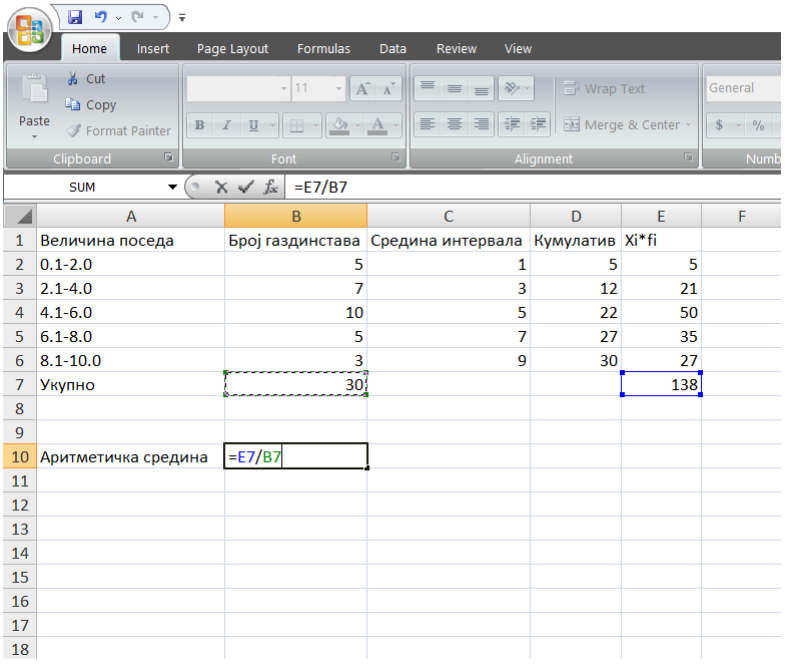
	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив	Xi*fi	
2	0.1-2.0	5	1	5	=C2*B2	
3	2.1-4.0	7	3	12		
4	4.1-6.0	10	5	22		
5	6.1-8.0	5	7	27		
6	8.1-10.0	3	9	30		
7	Укупно					
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

6. Након клика на дугме „*Enter*“, потребно је довести курсор миша у доњи десни угао поља *E2* и превући на доле, као што је учињено у колони која се односи на кумулатив. Затим је потребно израчунати суму колоне $X_i \times f_i$. Потребну суму је могуће добити тако што ће се у поље E7 уписати код: *=sum(E2:E6)*, као што је представљено у наставку. На крају потребно је притиснути дугме „*Enter*“. На исти начин могуће је израчунати и суму колоне фреквенције ($\sum f_i$), која ће у овом примеру бити уписана у поље *B7*.



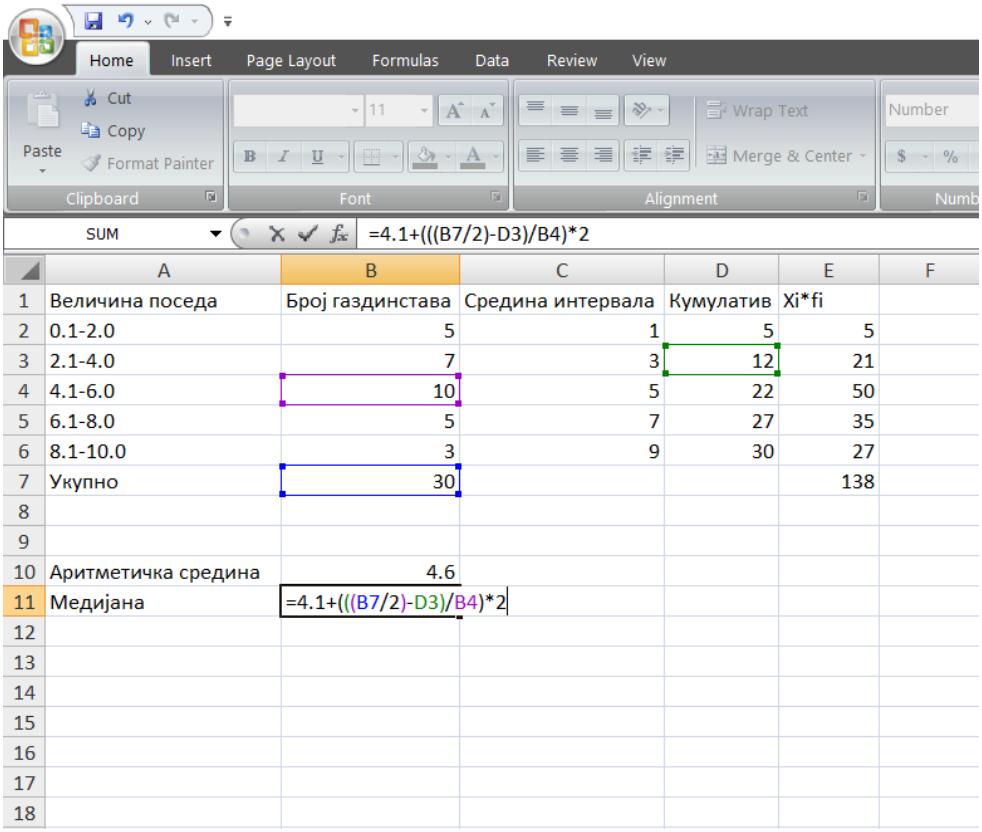
	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив	Xi*fi	
2	0.1-2.0	5	1	5	5	
3	2.1-4.0	7	3	12	21	
4	4.1-6.0	10	5	22	50	
5	6.1-8.0	5	7	27	35	
6	8.1-10.0	3	9	30	27	
7	Укупно				=sum(E2:E6)	
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						

7. Како би се коначно израчунала аритметичка средина, потребно је у неко од слободних поља уписати следећи код: *=E7/B7*. На тај начин, стављене су у однос две претходно израчунате суме и добијена вредност аритметичке средине која износи 4,60 хектара.



	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив	Xi*fi	
2	0.1-2.0	5	1	5	5	
3	2.1-4.0	7	3	12	21	
4	4.1-6.0	10	5	22	50	
5	6.1-8.0	5	7	27	35	
6	8.1-10.0	3	9	30	27	
7	Укупно	30			138	
8						
9						
10	Аритметичка средина	=E7/B7				
11						
12						
13						
14						
15						
16						
17						
18						

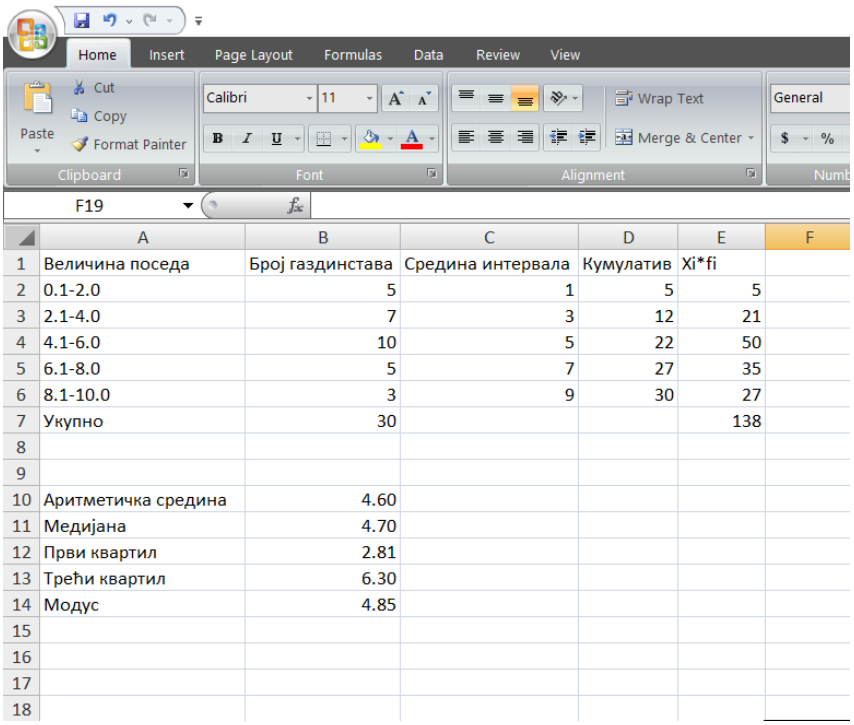
8. Израчунавање медијане, квартила и модуса, неопходно је учинити уписивањем одговарајуће формуле у неко од слободних поља. Примера ради, у оквиру овог конкретног примера за медијану је неопходно уписати следећи код: *=4.0 +(((B7/2)-D3)/B4)*2*. Овде је јасно да је потребно теоријско знање у вези са обрачуном медијане на основу груписаних података, како би се установио медијални интервал, као и доња граница медијалног интервала (вредност 4,0). Вредност 2, на крају кода, представља дужину групног интервала.



	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив	Xi*fi	
2	0.1-2.0	5	1	5	5	
3	2.1-4.0	7	3	12	21	
4	4.1-6.0	10	5	22	50	
5	6.1-8.0	5	7	27	35	
6	8.1-10.0	3	9	30	27	
7	Укупно	30			138	
8						
9						
10	Аритметичка средина	4.6				
11	Медијана	=4.1+(((B7/2)-D3)/B4)*2				
12						
13						
14						
15						
16						
17						
18						

На сличан начин могуће је израчунати квартиле и модус. Код првог и трећег квартила потребно је уписати следеће кодове респективно: *=2.0 +(((B7/4)-D2)/B3)*2* и *=6.0 +(((3*B7/4)-D4)/B5)*2*. На крају у циљу израчунавања модуса, у оквиру представљеног примера потребно је у неко од слободних поља уписати следећи код: *=4.0+((B4-B3)/((B4-B3)+(B4-B5)))*2*.

На крају задатка, *Microsoft Excel* прозор би треблао да изгледа на сличан начин као што је представљено у наставку:



	A	B	C	D	E	F
1	Величина поседа	Број газдинстава	Средина интервала	Кумулатив	Xi*fi	
2	0.1-2.0	5	1	5	5	
3	2.1-4.0	7	3	12	21	
4	4.1-6.0	10	5	22	50	
5	6.1-8.0	5	7	27	35	
6	8.1-10.0	3	9	30	27	
7	Укупно	30			138	
8						
9						
10	Аритметичка средина	4.60				
11	Медијана	4.70				
12	Први квартил	2.81				
13	Трећи квартил	6.30				
14	Модус	4.85				
15						
16						
17						
18						

2.4. Показатељи варијације

У циљу детаљније анализе посматране серије података, поред показатеља централне тенденције, неопходно је утврдити и показатеље варијације (варијабилитета или дисперзије). Две серије података често могу имати исте вредности неког од показатеља централне тенденције, а да истовремено њихове индивидуалне вредности обележја буду у значајној мери различите. Другим речима, варијација између вредности обележја једне серије може бити већа или мања од варијације вредности обележја у другој серији. Самим тим, уколико се не узме у обзир разлика у варијабилитету, може се доћи до погрешног закључка да је посматрана карактеристика у обе серије иста. Због тога је значајно да се утврди и варијабилитет посматране серије.

У показатеље варијације спадају:

- *Интервал (размак) варијације ;*
- *Интерквартилна разлика;*
- *Средње апсолутно одступање;*
- *Стандардна девијација;*
- *Варијанса;*
- *Коефицијент варијације;*
- *Коефицијент интерквартилне варијације;*
- *Стандардизовано (нормализовано) одступање.*

Наведени показатељи варијације могу се поделити на *апсолутне* и *релативне* показатеље. Показатељи варијације чија вредност се исказује у јединицама мере посматраног обележја, називају се апсолутним показатељима. С тим у вези, апсолутни показатељи варијабилитета су: интервал (размак) варијације, интерквартилна разлика, средње апсолутно одступање, стандардна девијација и варијанса. С друге стране, релативни показатељи варијабилитета се не исказују у јединицама мере посматраног обележја. Релативни показатељи варијабилитета су: коефицијент варијације, стандардизовано одступање и коефицијент интерквартилне варијације.

2.4.1. Интервал (размак) варијације

Као најједноставнији показатељ варијације користи се интервал (размак) варијације (*I*). Интервал варијације представља разлику екстремних вредности обележја у некој серији. Код негруписаних података и код неинтервалне серије дистрибуције фреквенција, интервал варијације представља разлику максималне и минималне вредности обележја у серији. Код интервалне дистрибуције фреквенција, интервал варијације представља разлику горње границе последњег и доње границе првог групног интервала. Израчунава се на основу следећег израза:

$$I = X_{max} - X_{min}\,.$$

Недостатак интервала варијације је у томе што искључиво зависи од екстремних вредности у серији и не даје увид у распоред осталих вредности обележја унутар серије.

2.4.2. Интерквартилна разлика

Слично као и интервал варијације, интерквартилна разлика (*IQR*) се дефинише као разлика трећег и првог квартила. Интерквартилна разлика је показатељ апслолутног варијабилитета и израчунава се у сврху елиминисања утицаја екстремних вредности на интервал варјације. Израчунава се наоснову следећег израза:

$$IQR = Q_3 - Q_1.$$

2.4.3. Средње апсолутно одступање

Показатељ варијације који се нешто чешће употребљава од интервала варијације и интерквартилне разлике јесте средње апсолутно одступање (*SO*). Средње апсолутно одступање се утврђује као количник збира апсолутних вредности одступања индивидуалних вредности обележја од њиховог просека и њиховог броја.

Средње апсолутно одступање нумеричког обележја измереног на јединицама основног скупа израчунава се према следећој формули:

$$SO = \frac{\sum_{i=1}^{N}|X_i - \mu|}{N},$$

где је:

X_i - индивидуална вредност обележја;

μ - аритметичка средина основног скупа за посматрано обележје;

N - број јединица посматрања у основном скупу (величина основног скупа).

Уколико су расположиви подаци груписани, средње апсолутно одступање се у случају основног скупа израчунава према формули:

$$SO = \frac{\sum_{i=1}^{k} f_i |X_i - \mu|}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i |X_i - \mu|}{N},$$

где је:

f_i - фреквенција односно број понављања одговарајућег обележја X_i.

За серије негруписаних вредности обележја у случају узорка, средње апсолутно одступање се израчунава на следећи начин:

$$SO = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n},$$

где је:

$\bar{X}$ - аритметичка средина узорка за посматрано обележје;

n - број јединица посматрања у узорку (величина узорка).

Уколико су подаци унутар узорка груписани, у примени је следећа формула:

$$SO = \frac{\sum_{i=1}^{k} f_i |X_i - \bar{X}|}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i |X_i - \bar{X}|}{n}.$$

2.4.4. Стандардна девијација и варијанса

Показатељ варијабилитета који се највише употребљава приликом статистичке анализе података јесте стандардна девијација (σ). Стандардна девијација је квадратни корен из средине квадрата одступања вредности обележја од аритметичке средине. Вредност стандардне девијације показује у којој мери груписане вредности обележја одступају од аритметичке средине.

За негруписане податке основног скупа, стандардна девијација се израчунава на следећи начин:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}.$$

Стандардна девијација се може израчунати и директно из података основног скупа на основу израза:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N}}{N}}.$$

Уколико су подаци основног скупа груписани, сатндардна девијација се израчунава на основу сличних формула где фигурира и фреквенција:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i (X_i - \mu)^2}{N}},$$

односно

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i X_i^2 - \frac{\left(\sum_{i=1}^{k} f_i X_i\right)^2}{\sum_{i=1}^{k} f_i}}{\sum_{i=1}^{k} f_i}} = \sqrt{\frac{\sum_{i=1}^{k} f_i X_i^2 - \frac{\left(\sum_{i=1}^{k} f_i X_i\right)^2}{N}}{N}}.$$

Уколико су предмет анализе подаци из урорка, стандардна девијација у ознаци S за негруписане податке израчунава се на основу следећих израза:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \qquad \text{или} \qquad S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}}.$$

На основу података из узорка који су представљени као дистрибуција фреквенција, стандардна девијација се оцењује на следећи начин:

$$S = \sqrt{\frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{n-1}} \qquad \text{или} \qquad S = \sqrt{\frac{\sum_{i=1}^{n} f_i X_i^2 - \frac{\left(\sum_{i=1}^{n} f_i X_i\right)^2}{n}}{n-1}},$$

где важи да је: $n = \sum_{i=1}^{k} f_i$.

Квадрат стандардне девијације представља варијансу (σ^2). Варијанса такође може да се израчуна за податке основног скупа или да се оцени из података узорка на сличан начин као и стандардна девијација.

За израчунавање варијансе код негруписаних података основног скупа користе се следећи изрази:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \qquad \text{или} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}.$$

Код дистрибуције фреквенција, варијанса се израчунава на основу следећих израза:

$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i(X_i-\mu)^2}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i(X_i-\mu)^2}{N} \quad \text{или} \quad \sigma^2 = \frac{\sum_{i=1}^{k} f_i X_i^2 - \frac{\left(\sum_{i=1}^{k} f_i X_i\right)^2}{\sum_{i=1}^{k} f_i}}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i X_i^2 - \frac{\left(\sum_{i=1}^{k} f_i X_i\right)^2}{N}}{N}.$$

Оцењена варијанса (S^2) на основу негруписаних података из узорка утврђује се на следећи начин:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} \qquad \text{или} \qquad S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}.$$

Уколико се варијанса оцењује на основу груписаних података из узорка, користе се следећи изрази:

$$S^2 = \frac{\sum_{i=1}^{n} f_i(X_i-\bar{X})^2}{n-1} \qquad \text{или} \qquad S^2 = \frac{\sum_{i=1}^{n} f_i X_i^2 - \frac{\left(\sum_{i=1}^{n} f_i X_i\right)^2}{n}}{n-1}.$$

Особине варијансе су:

1. Варијанса је показатељ варијације изражен квадратима јединице мере посматраног обележја. У случају да квадрат јединице нема интерпретацију уз израчунату вредност варијансе, јединица мере у запису се изоставља.

2. Ако су све вредности обележја у некој серији међусобно једнаке, варијанса и стандардна девијација су једнаке нули.

3. Ако свим вредностима обележја у некој серији додамо или одузмемо константу, варијанса нових вредности обележја се не мења:

$$X_i' = X_i \pm C \;\rightarrow\; \sigma_{X_i'}^2 = \sigma_{X_i}^2 \quad (i = 1, 2, \dots, N).$$

4. Ако све вредности обележја у некој серији помножимо константом, варијанса нових вредности обележја биће једнака производу квадрата константе и претходно израчунате варијансе:

$$X_i' = X_i \times C \;\rightarrow\; \sigma_{X_i'}^2 = \sigma_{X_i}^2 \times C^2 \quad (i = 1, 2, \dots, N).$$

Наведене особине варијансе важе и за оцену варијансе S^2.

2.4.5. Коефицијент варијације

Претходно дефинисани апсолутни показатељи варијације зависе од јединица мере у којима су дати посматрани подаци. Приликом упоређивања варијабилитета више серија података изражених у различитим јединицама мере, уколико би се посматрали искључиво апсолутни показатељи варијабилитета, може доћи до погрешног закључка. Како би се избегли погрешни закључки приликом анализе, израчунавају се релативни показатељи варијабилитета.

Најчешће коришћен релативни показатељ варијабилитета, који служи за упоређивање варијабилитета појава које имају различиту јединицу мере јесте коефицијент варијације (V).

Коефицијент варијације у случају основног скупа израчунва се на основу следећег израза:

$$V = \frac{\sigma}{\mu} \times 100(\%).$$

Уколико се подаци односе на узорак, коефицијент варијације се рачуна на следећи начин:

$$V = \frac{S}{\bar{X}} \times 100(\%).$$

2.4.6. Коефицијент интерквартилне варијације

Коефицијент интерквартилне варијације (IQR_{VR}) је релативни показатељ варијабилитета који се израчунава у сврху елиминисања утицаја екстремних вредности. Као и коефицијент варијације, користи се приликом упоређивања варијабилитета појава које имају различиту јединицу мере. Израчунава се на следећи начин:

$$IQR_{VR} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100(\%).$$

2.4.7. Стандардизовано (нормализовано) одступање

Стандардизовано (нормализовано) одступање (Z_i) је мера удаљености појединих вредности обележја од аритметичке средине исказана у односу на стандардну девијацију. Стандардизовано одступање је такође релативни показатељ дисперзије обележја. Његова вредност се у случају основног скупа израчунава на следећи начин:

$$Z_i = \frac{X_i - \mu}{\sigma} \quad (i = 1,2, \dots, N).$$

У случају да су дати подаци узорка, стандардизовано одступање се израчунава на основу следећег израза:

$$Z_i = \frac{X_i - \bar{X}}{S} \quad (i = 1,2, \dots, n).$$

За разлику од осталих показатеља варијације, стандардизовано одступање показује варијабилитет појединачних вредности обележја. Вредност стандардизованог одступања може бити позитивна или негативна, у зависности од тога да ли је вредност обележја већа или мања од аритметичке средине. Удаљеност вредности обележја од аритметичке средине је исказана бројем стандардних девијација обележја.

Аритметичка средина стандардизованог обележја је увек 0, док су варијанса и стандардна девијација увек 1.

Пример 6 (негруписани подаци):

Подаци из узорка који се односе на висину седам биљака пшенице (цм) били су: 79; 81; 82; 83; 92; 95; 97; 98. Израчунати показатеље интервал варијације, интерквартилну разлику, средње апсолутно одступање, стандардну девијацију, варијансу, коефицијент варијације, коефицијент интерквартилне варијације, стандардизовано одступање.

Решење:

Висина (цм) $(\boldsymbol{X_i})$	$X_i - \bar{X}$	$\lvert X_i - \bar{X}\rvert$	$(X_i - \bar{X})^2$	${X_i}^2$
79	79 – 88,4 = -9,4	9,4	88.36	6.241
81	81 – 88,4 = -7,4	7,4	54.76	6.561
82	82 – 88,4 = -6,4	6,4	40.96	6.724
83	83 – 88,4 = -5,4	5,4	29.16	6.889
92	92 – 88,4 = 3,6	3,6	12.96	8.464
95	95 – 88,4 = 6,6	6,6	43.56	9.025
97	97 – 88,4 = 8,6	8,6	73.96	9.409
98	98 – 88,4 = 9,6	9,6	92.16	9.604
Σ707	**0**	**57**	**435,88**	**62.917**

Интервал варијације: $I = X_{max} - X_{min} = 98 - 79 = 19$ цм.

Интерквартилна разлика:

$n = 8$

$$Q_1 = \frac{\mathrm{X}_{\frac{\mathrm{n}}{4}} + \mathrm{X}_{\frac{\mathrm{n}}{4}+1}}{2} = \frac{\mathrm{X}_{\frac{8}{4}} + \mathrm{X}_{\frac{8}{4}+1}}{2} = \frac{\mathrm{X}_2 + \mathrm{X}_3}{2} = \frac{81 + 82}{2} = 81{,}5 \text{ цм};$$

$$Q_1 = \frac{\mathrm{X}_{\frac{3\times\mathrm{n}}{4}} + \mathrm{X}_{\frac{3\times\mathrm{n}}{4}+1}}{2} = \frac{\mathrm{X}_{\frac{3\times8}{4}} + \mathrm{X}_{\frac{3\times8}{4}+1}}{2} = \frac{\mathrm{X}_6 + \mathrm{X}_7}{2} = \frac{95 + 97}{2} = 96{,}0 \text{ цм};$$

$$IQR = Q_3 - Q_1 = 96 - 81{,}5 = 14{,}5 \text{ цм}.$$

Средње апсолутно одступање:

$$\bar{X} = \frac{\sum_{i=1}^{7} X_i}{n} = \frac{707}{8} = 88{,}4 \text{ цм};$$

$$SO = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} = \frac{57}{8} = 7{,}13 \text{ цм}.$$

Стандардна девијација:

I начин:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{435{,}88}{8-1}} = \sqrt{62{,}27} = 7{,}89 \text{ цм};$$

II начин:

$$S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}} = \sqrt{\frac{62.917 - \frac{707^2}{8}}{8-1}} = \sqrt{62{,}27} = 7{,}89 \text{ цм}.$$

Варијанса:

I начин:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{435{,}88}{8-1} = 62{,}27 \text{ цм};$$

II начин:

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1} = \frac{62.917 - \frac{707^2}{8}}{8-1} = 62{,}27 \text{ цм}.$$

Коефицијент варијације:

$$V = \frac{S}{\bar{X}} \times 100 = \frac{7{,}89}{88{,}4} \times 100 = 8{,}93\%.$$

Коефицијент интерквартилне варијације:

$$IQR_{VR} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 = \frac{96 - 81{,}5}{96 + 81{,}5} \times 100 = 8{,}17\%.$$

Стандардизовано одступање:

$$Z_1 = \frac{X_1 - \bar{X}}{S} = \frac{79 - 88{,}4}{7{,}89} = -1{,}19;$$

$$Z_2 = \frac{X_2 - \bar{X}}{S} = \frac{81 - 88{,}4}{7{,}89} = -0{,}93;$$

$$Z_3 = \frac{X_3 - \bar{X}}{S} = \frac{82 - 88{,}4}{7{,}89} = -0{,}81;$$

$$Z_4 = \frac{X_4 - \bar{X}}{S} = \frac{83 - 88{,}4}{7{,}89} = -0{,}68;$$

$$Z_5 = \frac{X_5 - \bar{X}}{S} = \frac{92 - 88{,}4}{7{,}89} = 0{,}46;$$

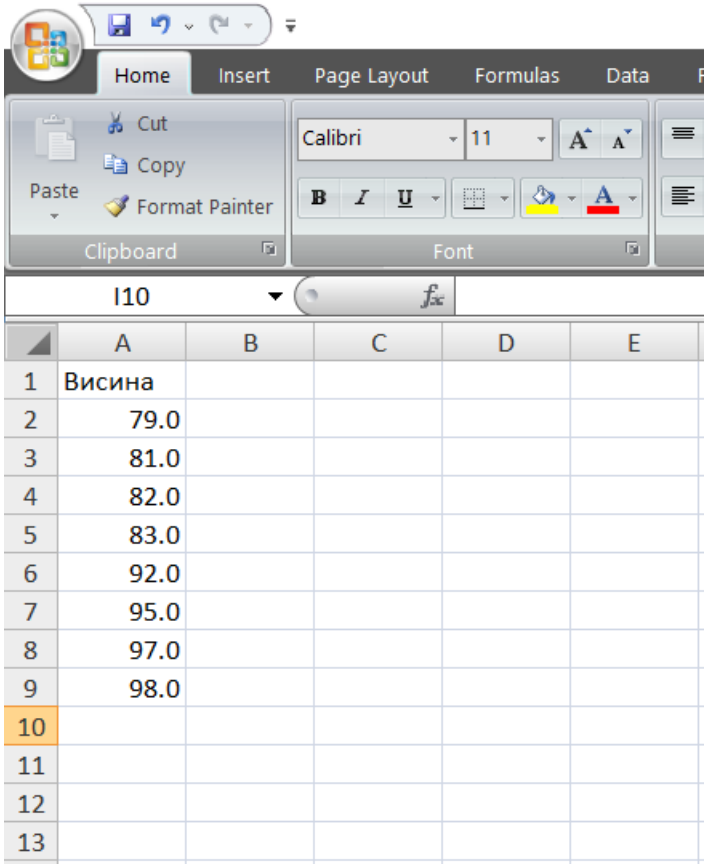
$$Z_6 = \frac{X_6 - \bar{X}}{S} = \frac{95 - 88{,}4}{7{,}89} = 0{,}84;$$

$$Z_7 = \frac{X_7 - \bar{X}}{S} = \frac{97 - 88{,}4}{7{,}89} = 1{,}09;$$

$$Z_8 = \frac{X_8 - \bar{X}}{S} = \frac{98 - 88{,}4}{7{,}89} = 1{,}22.$$

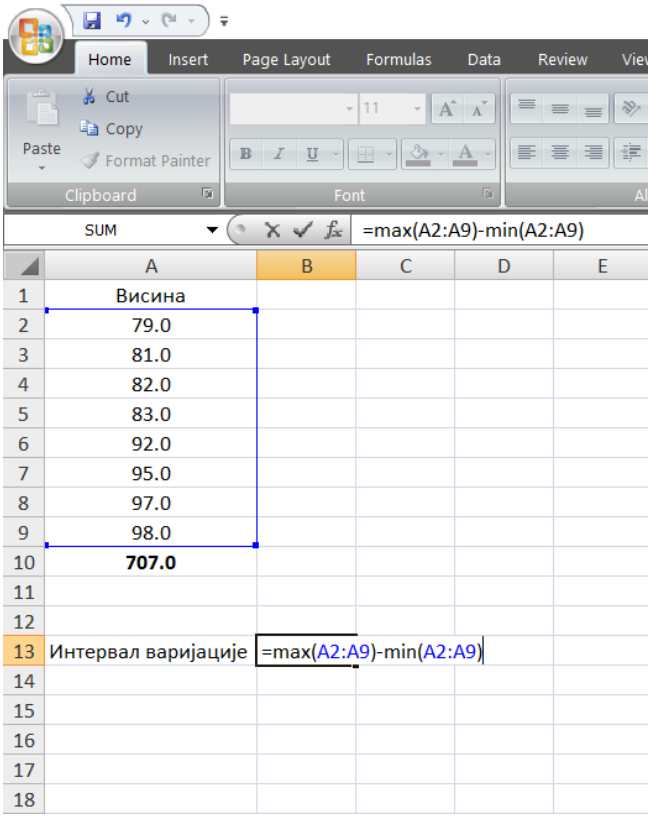
Вежба 4. Показатељи варијабилитета у Microsoft Excel-у

1. Као што и до сада био случај, у овој вежби ће се представити начин на који се могу израчунати основни показатељи варијабилитета на примеру за негруписане податке. Почетни изглед табеле, у складу са примером, у *Microsoft Excel* представљен је у наставку:

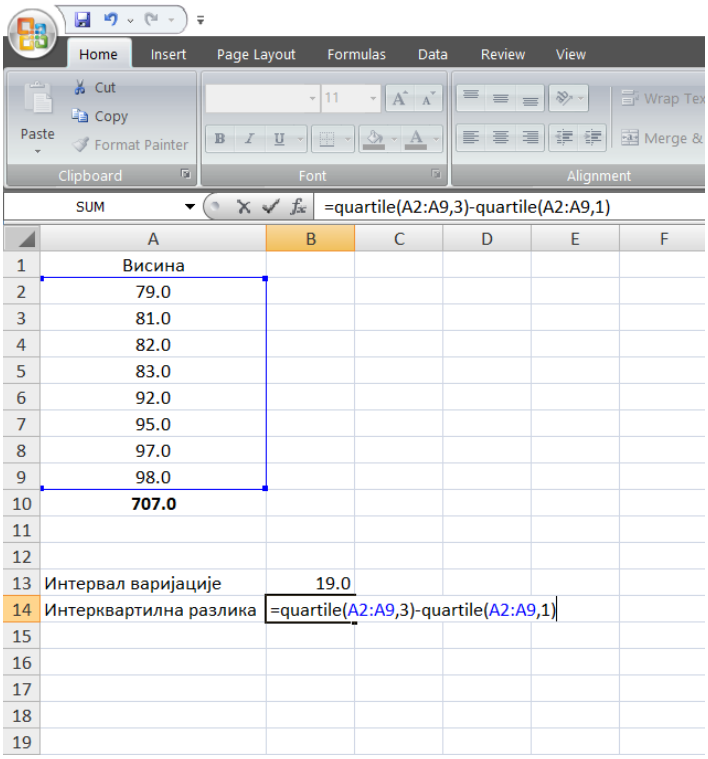


	A	B	C	D	E
1	Висина				
2	79.0				
3	81.0				
4	82.0				
5	83.0				
6	92.0				
7	95.0				
8	97.0				
9	98.0				
10					
11					
12					
13					

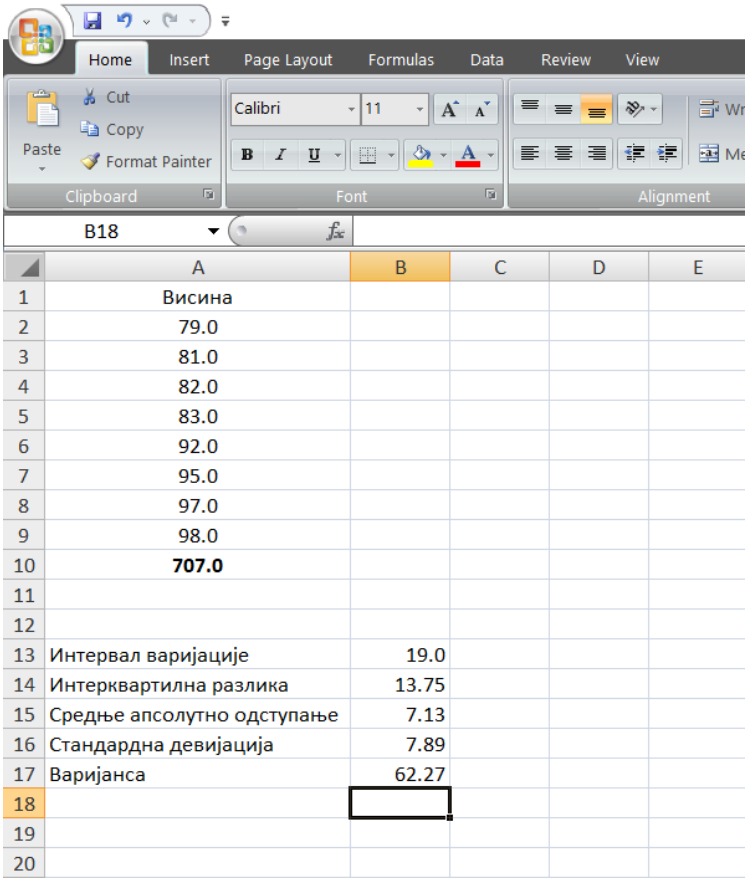
2. Приликом израчунавања интервала варијације, потребно је претходно утврдити максималну и минималну вредност обележја, а затим израчунати њихову разлику. С тим у вези, изглед *Microsoft Excel* прозора приликом израчунавања интервала варијације на примеру који се односи на висину пшенице је следећи:



3. На сличан начин могуће је израчунати и интерквартилну разлику, с тим што ће се сада рачунати разлика између трећег и првог квартила, као што је представљено у наставку:



4. На сличан начин могуће је израчунати и остале показатеље варијабилитета. Конкретно, приликом израчунавања средњег апсолутног одступања, за податке који се односе на конкретан пример, потребно је укуцати следећу команду: *=AVEDEV(A2:A9)*. Приликом израчунавања стандардне девијације и варијансе узорка, потребни су следећи кодови респективно: *=STDEV(A2:A9)* и *=VAR(A2:A9)*. *Microsoft Excel* прозор би сада требао да изгледа на следећи начин:



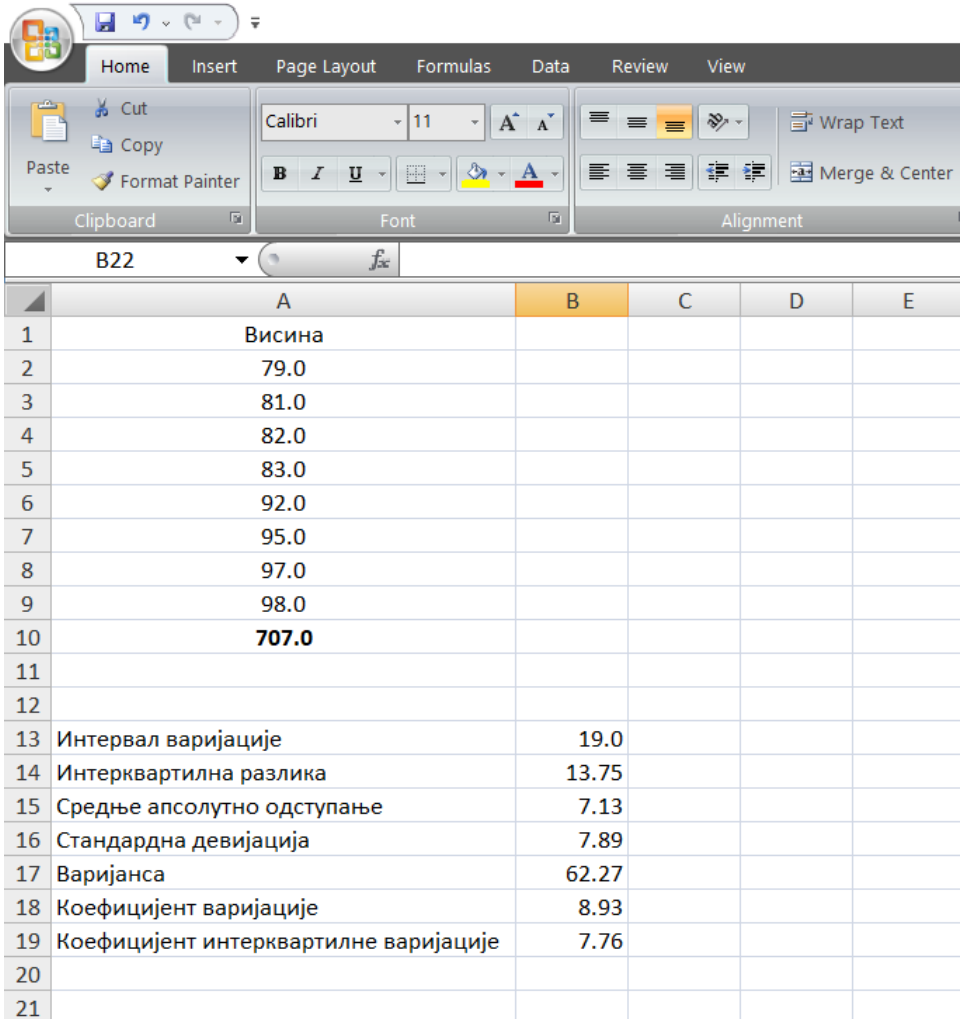
	A	B	C	D	E
1	Висина				
2	79.0				
3	81.0				
4	82.0				
5	83.0				
6	92.0				
7	95.0				
8	97.0				
9	98.0				
10	**707.0**				
11					
12					
13	Интервал варијације	19.0			
14	Интерквартилна разлика	13.75			
15	Средње апсолутно одступање	7.13			
16	Стандардна девијација	7.89			
17	Варијанса	62.27			
18					
19					
20					

Стандардна девијација и варијансе популације израчунавају се применом наредби: *=STDEVP(A2:A9)* и *=VARP(A2:A9)*.

5. Како би се израчунали релативни показатељи варијабилитета, потребно је ставити у однос неке од претходно израчунатих показатеља, с тим да је потребно израчунати и неке нове показатеље као што је аритметичка средина.

Конкретно, приликом обрачуна коефицијента варијације, претходно израчуната стандардна девијација (поље *B16*) дели се са аритметичком средином, а потом се све помножи са 100 како би се добила вредност изражена у процентима. Самим тим, код ће бити следећи: *=(B16/AVERAGE(A2:A9))*100*. Слично, приликом израчунавања коефицијента интерквартилне варијације, пратећи формулу, потребно је у неко од слободних поља уписати следећи код: *=((QUARTILE(A2:A9,3)-QUARTILE(A2:A9,1))/(QUARTILE(A2:A9,3)+QUARTILE(A2:A9,1)))*100*.

На крају, како је потребно израчунати и стандардизована одступања, формираће се нова колона. У поље *B1*, биће уписан следећи код: *=(A2-AVERAGE($A$2:$A$9))/$B$16*. У наставку је потребно наместити курсор на доњи десни угао поља *B1* и превући вредности на доле. Знак *$* је коришћен како би се поља фиксирала. Коначан изглед табеле је следећи:



	A	B	C	D	E
1	Висина				
2	79.0				
3	81.0				
4	82.0				
5	83.0				
6	92.0				
7	95.0				
8	97.0				
9	98.0				
10	**707.0**				
11					
12					
13	Интервал варијације	19.0			
14	Интерквартилна разлика	13.75			
15	Средње апсолутно одступање	7.13			
16	Стандардна девијација	7.89			
17	Варијанса	62.27			
18	Коефицијент варијације	8.93			
19	Коефицијент интерквартилне варијације	7.76			
20					
21					

Пример 7. (груписани подаци):

На основу дистрибуције фреквенција која је формирана на основу броја ларви на 20 одвојених пацела, потребно је израчунати показатеље варијабилитета. Стандардизовано одступање израчунати само за најучесталију вредност обележја (модус).

Број ларви (X_i)	**Број парцела** (f_i)	**Кумулатив**	***Средина интервала*** (X_i')	$f_i \times X_i$	$\lvert X_i - \bar{X}\rvert$	$f_i\lvert X_i - \bar{X}\rvert$	$f_i(X_i - \bar{X})^2$	$f_i \times X_i^2$
0-99	4	4	50	200	125	500	62.500	10.000
100-199	9	13	150	1.350	25	225	5.625	202.500
200-299	5	18	250	1.250	75	375	28.125	312.500
300-399	2	20	350	700	175	350	61.250	245.000
Σ	**20**			**3.500**		**1.450**	**157.500**	**770.000**

Решење:

Интервал варијације: $I = X_{max} - X_{min} = 399 - 0 = 399$ ларви.

Интерквартилна разлика:

$$Q_1 = L + \left(\frac{\frac{n}{4} - F_{Q_1-1}}{f_{Q_1}}\right) \times i = 100 + \left(\frac{\frac{20}{4} - 4}{9}\right) \times 100 = 111{,}11 \text{ ларви}$$

$$Q_3 = L + \left(\frac{\frac{3n}{4} - F_{Q_3-1}}{f_{Q_3}}\right) \times i = 200 + \left(\frac{\frac{3 \times 20}{4} - 13}{5}\right) \times 100 = 240 \text{ ларви}$$

$$IQR = Q_3 - Q_1 = 240 - 111{,}11 = 128{,}89 \text{ ларви}$$

Средње апсолутно одступање:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{3.500}{20} = 175 \,{}^{\text{ларви}}/_{\text{парцели}};$$

$$SO = \frac{\sum_{i=1}^{k} f_i |X_i - \bar{X}|}{n} = \frac{1.450}{20} = 72{,}5 \text{ ларви.}$$

Стандардна девијација:

I начин:

$$S = \sqrt{\frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{157.500}{20-1}} = \sqrt{8.289{,}47} = 91{,}05 \text{ ларви;}$$

II начин:

$$S = \sqrt{\frac{\sum_{i=1}^{n} f_i X_i^2 - \frac{(\sum_{i=1}^{n} f_i X_i)^2}{n}}{n-1}} = \sqrt{\frac{770.000 - \frac{3.500^2}{20}}{20-1}} = \sqrt{8.289{,}47} = 91{,}05 \text{ ларви.}$$

Варијанса:

I начин:

$$S^2 = \frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{n-1} = \frac{157.500}{20-1} = 8.289{,}47 \text{ ларви;}$$

II начин:

$$S^2 = \frac{\sum_{i=1}^{n} f_i X_i^2 - \frac{(\sum_{i=1}^{n} f_i X_i)^2}{n}}{n-1} = \frac{770.000 - \frac{3.500^2}{20}}{20-1} = 8.289{,}47 \text{ ларви.}$$

Коефицијент варијације:

$$V = \frac{S}{\bar{X}} \times 100 = \frac{91{,}05}{175} \times 100 = 52{,}03\%.$$

Коефицијент интерквартилне варијације:

$$IQR_{VR} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 = \frac{240 - 111{,}11}{240 + 111{,}11} \times 100 = 36{,}71\%.$$

Стандардизовано одступање:

Како модус представља најучесталију вредност обележја, потребно је претходно утврдити модалну вредност, а затим израчунати стандардизовану вредност.

Модус:

$$Mo = L + \left(\frac{d_1}{d_1 + d_2}\right) \times i = 100 + \left(\frac{5}{5 + 4}\right) \times 100 = 155{,}56 \text{ ларви};$$

Дакле, потребно је израчунати стандардизовану вредност за обележје 155,56.

$$Z_i = \frac{X_i - \bar{X}}{S} = \frac{155{,}56 - 175}{91{,}05} = -0{,}21.$$

С обзиром на то да су подаци груписани, приликом израчунавања показатеља варијабилитета у *Microsoft Excel-y*, потребно је формирати радну табелу као што је урађено у примеру. Даље је потребно у одговарајућа поља дефинисати потребне формуле, на сличан начин као што је урађено у вежби 2. Разлика ће бити у томе што ће се сада користити формуле за показатеље варијабилитета.

2.5. Показатељи облика дистрибуције

Облик дистрибуције подразумева сагледавање две карактеристике посматраних података: *асиметричност* и *спљоштеност*. Најчешће коришћени показатељи облика дистрибуције су:

- *I Пирсонов коефицијент* (β_1) – показатељ асиметричности;
- *II Пирсонов коефицијент* (β_2) – показатељ спљоштености.

Приликом израчунавања наведених коефицијената, потребно је претходно израчунати *централне моменте*. Под централним моментом *k*-тог реда, подразумева се средина суме одступања вредности обележја од аритметичке средине степенована на *k*-ти степен.

Код негруписаних података, у случају основног скупа, централни моменти се израчунавају на основу следећег израза:

$$\mu_k = \frac{\sum (X_i - \mu)^k}{N}, \quad k = 0{,}1{,}2{,}3, \ldots$$

Уколико су подаци из основног скупа груписани, централни моменти се израчунавају на основу следећег израза:

$$\mu_k = \frac{\sum f_i (X_i - \mu)^k}{N}, \quad k = 0,1,2,3, \ldots$$

С друге стране, уколико се подаци односе на узорак, централне мометне за негруписане односно груписане податке могуће је израчунати на основу следећих израза респективно:

$$\mu_k = \frac{\sum (X_i - \bar{X})^k}{n}, \quad k = 0,1,2,3, \ldots \quad \text{и} \quad \mu_k = \frac{\sum f_i (X_i - \bar{X})^k}{n}, \quad k = 0,1,2,3, \ldots$$

За израчунавање првог и другог Пирсоновог коефицијента потребни су централни моменти другог, трећег и четвртог степена. С тим у вези, I Пирсонов коефицијент (β_1), израчунава се на основу следећег израза:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}.$$

Код симетричних расподела података, важи да је трећи централни моменат једнак 0 ($\mu_3 = 0$), одакле следи да је и $\beta_1 = 0$. Уколико је вредност показатеља β_1 већа од нуле, расподела је асиметрична. Предзнак трећег централног момента у овом случају показује да ли је реч о позитивној или о негативној асиметричности.

Други начин, испитивања асиметричности расподеле података јесте израчунавање коефицијента асиметричности α_3, тако да важи $\alpha_3 = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}$, где је σ вредност стандардне девијације за серију расположивих података. За разлику од I Пирсоновог коефицијента који указује само на присуство асиметричности, на основу вредности коефицијента асиметричности α_3 може се утврдити да ли је асиметрија позитивна или негативна. Поред наведеног, на основу вредности коефицијента асиметричности α_3, може се утврдити и јачина асиметрије и то наледећи начин:

- $|\alpha_3| < 0,1$ – нема асиметрије;
- $0,1 \leq |\alpha_3| < 0,25$ – асиметрија је мала;
- $0,25 \leq |\alpha_3| < 0,5$ – асиметрија је средње величине;
- $0,5 \leq |\alpha_3|$ – асиметрија је јака.

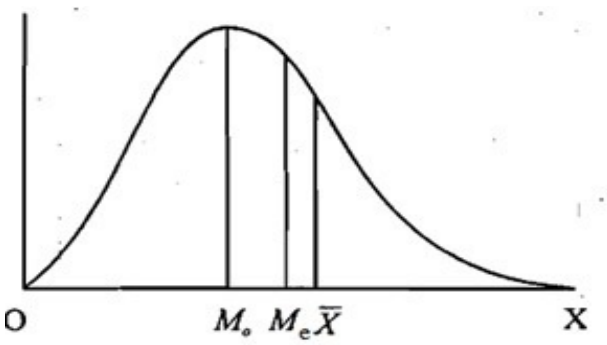
Поред наведеног, симетричност посматране расподеле може се установити и на основу односа између показатеља централне тенденције као што су аритметичка средина, медијана и модус. Када је дистрибуција фреквенција симетрична, аритметичка средина, модус и медијана се поклапају ($\bar{X} = Mo = Me$).

Ипак, из поклапања ова три показатеља не мора да следи симетричност дистрибуције. Потребно је да се на основу графичког приказа дистрибуције или показатеља асиметрије изврши додатно испитивање.

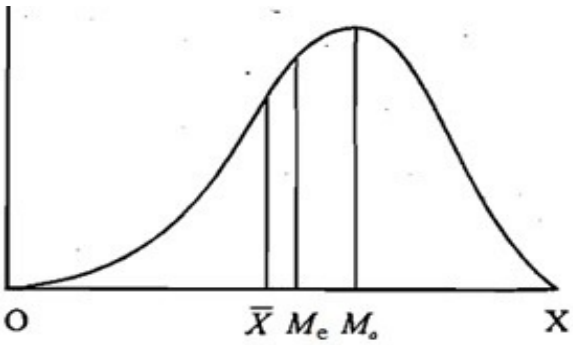
У случају унимодалних дистрибуција тј. дистрибуција које имају један модус, ако је модус по вредности већи од медијане и аритметичке средине, серија је негативно асиметрична или асиметрична улево. Уколико је вредност аритметичке средине већа од вредности медијане и модуса серија је позитивно асиметрична или асиметрична удесно.

У наставку следи илустративни приказ претходно наведеног.

Позитивна асиметрија $Mo < Me < \bar{X}$



Негативна асиметрија $\bar{X} < Me < Mo$



II Пирсонов коефицијент (β_2), који представља показатељ спљоштености може се израчунати на основу следећег израза:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Слично као и код коефицијента асиметричности, могуће је издвојити и коефицијент спљоштености у ознаци α_4, с тим да важи $\alpha_4 = \beta_2$.

Уколико је вредност показатеља $\beta_2 = 3$, дистрибуција података је нормално спљоштена. Другима речима, дистрибуција података има исту спљоштеност као и теоријска нормална расподела. Уколико је $\beta_2 > 3$, каже се да је дистрибуција издуженија у односу на наормалну расподелу, док уколико је $\beta_2 < 3$, дистрибуција је спљоштенија у односу на нормалну расподелу.

У статистичким програмима се као показатељ спљоштености израчунава коефицијент спљоштености који се дефинише као $k = \beta_2 - 3$, тако да ако је $k > 0$ дистрибуција је издуженија у односу на наормалну расподелу, уколико је $k < 0$, дистрибуција је спљоштенија у односу на нормалну расподелу, док ако је $k = 0$ дистрибуција је исте спљоштености као нормална дистрибуција.

Пример 7 (негруписани подаци):

На основу података о приносу кукуруза (тона/хектару) на 5 различитих парцела, израчунати показатеље облика дистрибуције:

Принос (т/ха) $(\boldsymbol{X_i})$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^3$	$(X_i - \bar{X})^4$
4,8	-0,52	0,2704	-0,1406	0,0731
5,3	-0,02	0,0004	0,000008	0,0000002
5,4	0,08	0,0064	0,0005	0,00004
5,4	0,08	0,0064	0,0005	0,00004
5,7	0,38	0,1444	0,0549	0,0209
Σ 26,6	**0**	**0,4280**	**-0,0847**	**0,0941**

Решење:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{26{,}6}{5} = 5{,}32 \text{ т/ха}$$

Централни моменти:

$$\mu_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} = \frac{0{,}4280}{5} = 0{,}0856;$$

$$\mu_3 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n} = \frac{-0{,}0847}{5} = -0{,}0169;$$

$$\mu_4 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n} = \frac{0{,}0941}{5} = 0{,}0188.$$

Пирсонови коефицијенти:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-0{,}0169)^2}{(0{,}0856)^3} = 0{,}4554;$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{0{,}0188}{(0{,}0856)^2} = 2{,}5657.$$

На основу израчунатих вредности Пирсонових коефицијената, може се установити да је дистрибуција негативно асиметрична ($\mu_3 < 0$) и спљоштенија у односу на нормалну расподелу. Вредност коефицијента асиметричности $\alpha_3 = -\sqrt{\beta_1} = -\sqrt{0{,}4554} = -0{,}6748$ указује на то да се асиметрија може оценити као јака.

Такође, како важи да је модална вредност која износи 5,4 т/ха већа од аритметичке средине која износи 5,32 т/ха, још једном се може потврдити закључак да је разматрана серија података која се односи на принос кукуруза са 5 одвојених парцела, негативно асиметрична.

Пример 8 (груписани подаци):

На основу серије података која се односи на број трактора по радним организацијама испитати облик расподеле:

Број тракотра (X_i)	Број радних организација (f_i)	$f_i \times X_i$	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$f_i(X_i - \bar{X})^2$	$f_i(X_i - \bar{X})^3$	$f_i(X_i - \bar{X})^4$
10	3	30	-12,2857	150,9388	452,8163	-5.563,1720	68.347,5419
12	5	60	-10,2857	105,7959	528,9796	-5.440,9329	55.963,8817
20	27	540	-2,2857	5,2245	141,0612	-322,4257	736,9729
24	20	480	1,7143	2,9388	58,7755	100,7580	172,7280
30	15	450	7,7143	59,5102	892,6531	6.886,1808	53.121,9658
Σ	**70**	**1.560**			**2.074,2857**	**-4.339,5918**	**178.343,0904**

Решење:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{1.560}{70} = 22{,}2857 \text{ трактора}/\text{радној организацији}$$

Централни моменти:

$$\mu_2 = \frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^2}{n} = \frac{2.074{,}2857}{70} = 29{,}6327;$$

$$\mu_3 = \frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^3}{n} = \frac{-4.339{,}5918}{70} = -61{,}9942;$$

$$\mu_4 = \frac{\sum_{i=1}^{n} f_i (X_i - \bar{X})^4}{n} = \frac{178.343{,}0904}{70} = 2.547{,}7580.$$

Пирсонови коефицијенти:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-61{,}9942)^2}{(29{,}6327)^3} = 0{,}1477;$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2.547{,}7580}{(29{,}6327)^2} = 2{,}9015.$$

На основу израчунатих вредности Пирсонових коефицијената, може се установити да је дистрибуција негативно асиметрична и спљоштенија у односу на нормалну расподелу.

Вредност коефицијента асиметричности $\alpha_3 = \sqrt{\beta_1} = \sqrt{0{,}1477} = 0{,}3843$ указује на то да је асиметрија средње величине.

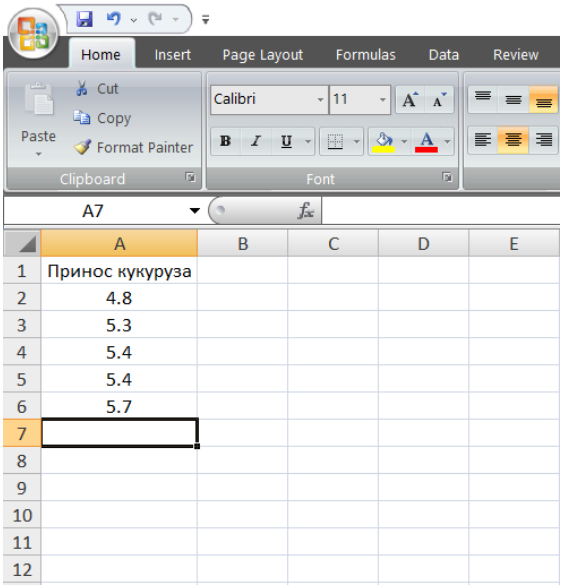
Вежба 5. Показатељи облика расподеле у Microsoft Excel-у

Када је реч о показатељима варијабилитета код негруписаних података, *Microsoft Excel* пружа могућност обрачуна показатеља *Skewness* (показатељ асиметрије) и *Kurtosis* (показатељ спљоштености). Овде је неопходно имати на уму да је интерпретација добијених показатеља мало другачија у односу на Пирсонове коефицијенте.

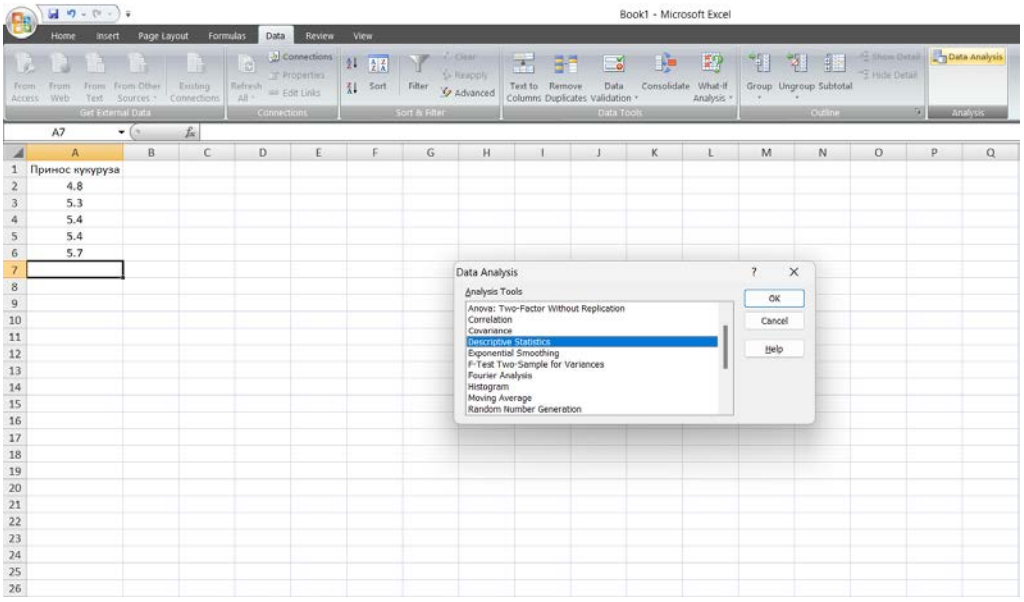
Поред наведеног, могуће је поступно израчунати показатеље β_1 и β_2 (по угледу на решење вежбе 2), формирајући радну табелу како за груписане податке тако и за негруписане податке.

До показатеља *Skewness* и *Kurtosis* у *Microsoft Excel*-у, могуће је доћи на следећи начин:

1. Почетни изглед табеле, у складу са примером, у *Microsoft Excel-y* представљен је у наставку:

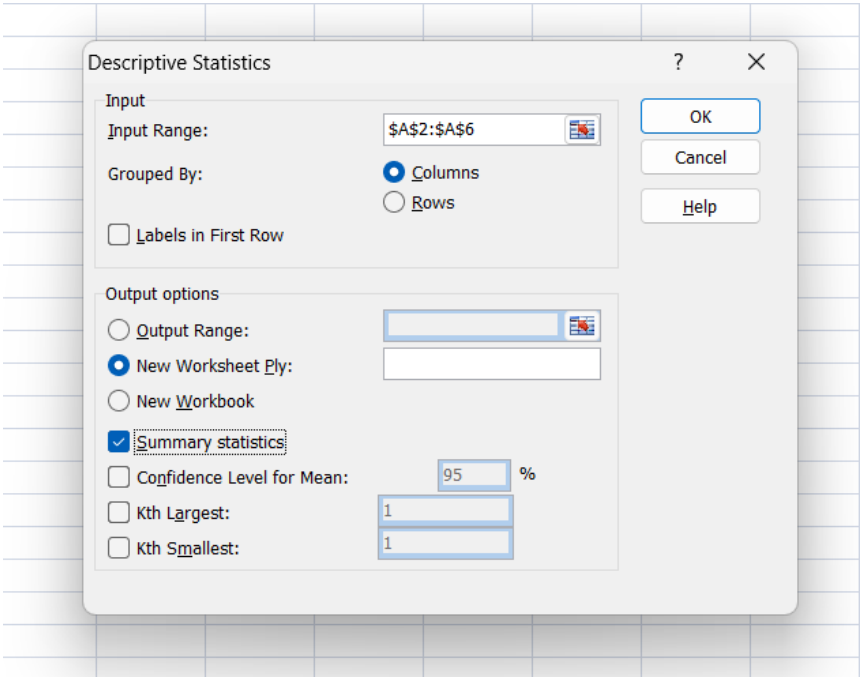


2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys*[1] у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *Descriptive statistics*, као што је представљено у наставку:

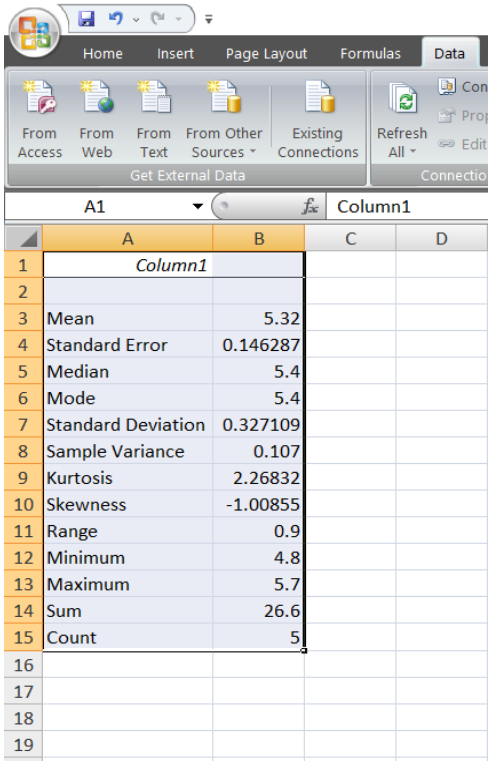


[1] *Уколико немате опцију Data Analsys потребно је да је инсталирате. Поступак инсталације опције Data Analsys приказан је у Прилогу 1.*

3. Кликом на дугме *OK* отвара се нови прозор, где је у делу који се односи на *Input Range* потребно означити податке који су предмет анализе. Такође, потребно је штиклирати *Summary statistics*. Остало може остати као што је одређено као полазно подешавање. Изглед новог прозора је представљен у наставку:



Кликом на дугме *OK* отвара се нови *Sheet*:



	A	B	C	D
1	*Column1*			
2				
3	Mean	5.32		
4	Standard Error	0.146287		
5	Median	5.4		
6	Mode	5.4		
7	Standard Deviation	0.327109		
8	Sample Variance	0.107		
9	Kurtosis	2.26832		
10	Skewness	-1.00855		
11	Range	0.9		
12	Minimum	4.8		
13	Maximum	5.7		
14	Sum	26.6		
15	Count	5		
16				
17				
18				
19				

Овде је потребно обратити посебну пажњу на вредности показатеља *Skewness = -1,0086* и *Kurtosis = 2,2683*. Као што је већ наведено, *Skewness* је показатељ асиметрије. Уколико је вредност показатеља *Skewness* мања од -1 или већа од +1, може се рећи да је присутна изражена негативна односно позитивна асиметрија. Услучају да се вредност показатеља *Skewness* креће у интервалу од -1 до -0,5 или од +0,5 до +1, присутна је умерена негативна односно позитивна асиметрија. На крају, уколико је вредност показатеља *Skewness* између -0,5 и 0, односно 0 и +0,5, може се рећи да је дистрибуција података приближна симетричној расподели. Како је добијена вреднос у посматраном примеру -1,0086 закључак је да је присутна изражена асиметрија, закључак који је сличан изведеном закључку на основу обрачуна α_3 показатеља.

Када је реч о показатељу *Kurtosis* који се односи на спљоштеност вредност која износи 2,2683 указује на то да је дистрибуција издуженија у односу на нормалну расподелу. Наиме, уколико је показатељ *Kurtosis* мањи од 0 дистрибуција је спљоштеннија у односу на нормалну расподелу. Вредност блиска нули указује на то да је дистрибуција нормално спљоштена, док вредност која је већа од 0 указује на закључак који је добијен у посматраном примеру. Формуле за *Skewness* и *Kurtosis* су оцене на основу узорка и разликују се од коефицијента асиметричности (α_3) и спљоштености (k) који су параметри популације.

Поред наведеног, међу бројним показатељима које је *Microsoft Excel* израчунао може се уочити да су представљени и неки од показатеља централне тенденције (аритметичка средина, медијана, модус), односно показатеља варијабилитета (интервал варијације, стандардна девијација и варијанса узорка). Поред наведеног ту су вредности минималног и максималног обележја, укупан број података, сума свих вредности, као и стандардне грешке о којој ће више речи бити касније.

Контролна питања

1. Дефинисати дистрибуцију фреквенција.
2. Дефинисати релативну фреквенцију.
3. Навести шта је кумулативна фреквенција и врсте кумулатива.
4. Објаснити графички приказ стабло-лист.
5. Шта је хистограм и када се користи.
6. Шта је полигон и када се користи.
7. Дефинисати аритметичку средину и навести њене особине.
8. Дефинисати позиционе средње вредности.
9. Навести показатеље варијабилитета и њихову поделу.
10. Навести показатеље облика и тумачење њихових израчунатих вредности.

3. ТЕОРИЈСКЕ ДИСТРИБУЦИЈЕ

3.1. Основни појмови вероватноће

Статистичка теорија је заснована на теорији вероватноће. Теорија вероватноће је грана математике која се бави анализом случајних појава. Резултати посматрања или експеримента називају се *елементарни догађаји*. С друге стране, скуп који садржи све елементарне догађаје назива се *простор елементарних догађаја*.

Случајни догађај је подскуп скупа (простора) елементарних догађаја. Случајни догађаји се обележавају великим словима латинице: A, B, C, D, итд или A_1, A_2, A_3, итд. Случајни догађај A садржи оне елементарне догађаје којима се дефинише догађај A. Сваком догађају A одговара супротан догађај $\bar{A}$ (нон A) који се остварује онда када се не оствари догађај A.

Вероватноћа случајног догађаја је израз могућности јављања тог догађаја. Вероватноћа се исказује бројем који варира од 0 (немогућ догађај) до 1 (сигуран догађај). Утврђивање вероватноће зависи од полазне теорије вероватноће.

Класична дефиниција вероватноће за нпр. догађај A, може се дефинисати као однос броја елементарних догађаја који сачињавају догађај A и броја свих могућих елементарних догађаја простора елементарних догађаја.

$$P(A) = \frac{m(A)}{n}.$$

Класична дефиниција је заснована на претпоставци да су сви елементарни догађаји подједнако могући тј. полази од претпоставке симетричности (хомоген новчић, хомогена коцка и сл.). По овој дефиницији појам вероватноће је апстрактно заснован и не зависи од искуства. Зато се овако уведена вероватноћа назива *вероватноћа априори*.

Пример класичне дефиниције вероватноће

Експеримен се састоји у бацању хомогене коцкице. Одредити простор елементарних догађаја и случајан догађај A (A: добијен је паран број). Израчунати вероватноћу реализације догађаја A.

Решење:

Простор елементарних догађаја је:

S = { ⚀, ⚁, ⚂, ⚃, ⚄, ⚅ }.

Догађај A је:

$A = \{$ ⚀, ⚃, ⚅ $\}$.

$$P(A) = \frac{m(A)}{n} = \frac{3}{6} = 0{,}5.$$

Статистичка дефиниција вероватноће може се дефинисати као гранична вредност фреквенције посматраног догађаја у оквиру n експеримената када n неограничено расте. Статистичка дефиниција вероватноће може се изразити на следећи начин:

$$P(A) = \lim_{n\to\infty} \frac{f}{n}.$$

Како би се одредила вероватноћа догађаја, потребно је понављати експеримент велики број пута под истим условима. Овако дефинисана вероватноћа је заснована на искуству и назива се вероватноћа *апостериори* или *статистичка вероватноћа*.

У случају да није могуће израчунати вероватноћу, она се оцењује на основу релативне фреквенције:

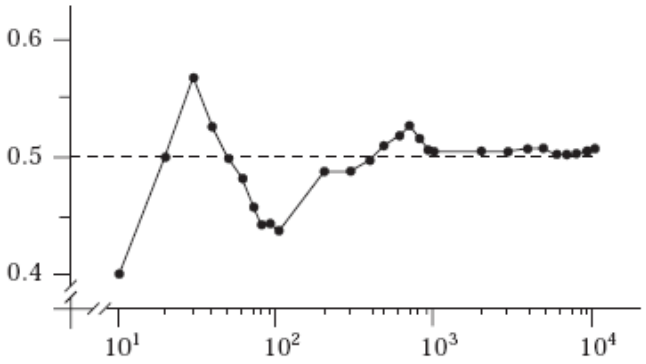
$$P(A) = \frac{f}{n}.$$

Релативне фреквенције нису вероватноће већ су апроксимације вероватноће. Ако се експеримент понавља велики број пута ове апроксимације вероватноће неког исхода теже вероватноћама исхода на основу закона великих бројева.

Пример статистичке дефиниције вероватноће

Уколико се број бацања хомогеног новчића увећава, релативна фреквенција ликова тежи вредности 0,5. Статистичар Карл Пирсон (енгл. Karl Pearson) бацао је новчић 24.000 пута и добио 12.012 ликова, тј. релативну фреквенцију која износи 0,5005.

Релативна фреквенција ликова уколико је број понављања 10.000 може се представити на следећи начин:



Поред класичне и статистичке дефиниције вероватноће неопходно је дефинисати и појам *субјективне вероватноће.* Субјективна вероватноћа је вероватноћа додељена неком догађају на основу субјективне процене, информације, искуства или веровања.

Без обзира која дефиниција вероватноће се примењује, збир вероватноћа свих елементарних догађаја је 1.

Ради лакше анализе у области вероватноће, пожељно је да све елементарне догађаје изражавамо помоћу реалних бројева, који ће самим тим садржати и информацију о вероватноћи појављивања елементарних догађаја које представљају.

Једнодимензионална случајна променљива је функција која сваки елементарни догађај статистичког експеримента пресликава у један реалан број, коме се придружује вероватноћа једнака збиру вероватноћа појављивања свих елементарних догађаја који се у њега сликају. Случајна променљива може бити *прекидна* (*дискретна)* и *непрекидна*.

Прекидна (дискретна) случајна променљива је случајна променљива која узима коначан број вредности или пребројиво бесконачан број вредности.

Непрекидна случајна променљива је случајна променљива која може да узме било коју вредност из једног или више интервала. Непрекидна случајна променљива има непребројиво много вредности.

Квантитативна карактеристика случајног догађаја назива се случајна променљива. Сваки елементарни догађај из простора елементраних догађаја S пресликава се у вредност са бројне праве. Први корак приликом дефинисања случајне променљиве је дефинисање простора елементарних догађаја S, односно дефинисање и исписивање свих могућих елементарних догађаја. За сваку случајно променљиву може се дефинисати закон вероватноће (*закон расподеле*), као и *функција расподеле.*

Дистрибуције које су формиране груписањем опажања или елемената скупа према неком обележју називају се *емпиријске (оригиналне, опажене) дистрибуције (расподеле).*

Насупрот емпиријским дистрибуцијама постоје дистрибуције које се могу очекивати у складу с искуством или на основу неких претпоставки. У питању су *теоријске дистрибуције (расподеле).*

Појму обележја код емпиријских дистрибуција одговара појам случајна променљива код теоријских дистрибуција. Одређивању релативних фреквенција код емпиријских дистрибуција фреквенција претходи пребројавање опсервираних вредности обележја тј. одређивање апсолутних фреквенција. Појму релативна фреквенција код теоријских дистрибуција одговара појам вероватноћа. Вероватноће се израчунавају као одређене функције вредности случајне променљиве.

Свака теоријска дистрибуција има свој закон вероватноће по ком су дистрибуиране вредности случајне варијабле X. Осим функције вероватноће, дистрибуције имају: *функцију расподеле, математичко очекивање, варијансу, коефицијент асиметрије* и *коефицијент спљоштености*.

Функција расподеле се дефинише као кумулативна вероватноћа случајне променљиве $F(x) = P(X \leq x)$ тако да увек важи $0 \leq F(x) \leq 1$. Функција расподеле одговара појму кумулације структуре код емпиријских дистрибуција. Поред наведеног, битно је истаћи да су теоријске дистрибуције основа инференцијалне статистике.

3.2. Прекидне теоријске дистрибуције

3.2.1. Биномна дистрибуција

Биномна дистрибуција је једна од најважнијих прекидних теоријских дистрибуција. Биномна дистрибуција се заснива на сукцесивним догађајима који имају два исхода. Другим речима, биномна дистрибуција може се дефинисати преко *Бернулијевог експеримента*. Бернулијев експеримент је случајни експеримент који има следеће карактеристике:

1. експеримент има два исхода, „успех“ и „неуспех“;

2. приликом сваког понављања експеримента, вероватноћа исхода „успех“ обележава се са p и не мења се од експеримента до експеримента. С друге стране, вероватноћа исхода „неуспех“ једнака је $q=1-p$;

3. експерименти су независни;

4. исход сваког експеримента или процеса је случајан.

Број „успеха“ приликом n понављања Бернуливег експеримента је случајна променљива X која има биномну расподелу. Како је број „успеха“ сваки цео број у интервалу од 0 до n, вредности случајне променљиве која има биномну расподелу X: 0,1,2,3, ... , n. Број модалитета случајне променљиве је $n + 1$.

Вероватноћа $P(X = i)$ за $i = 0,1,2,3, \dots, n$ дата је иразом:

$$p(i) = \binom{n}{i} \times p^i \times q^{n-i} ,$$

где је:

n – број модалитета обележја умањен за 1;

p – вероватноћа „успеха“;

q – вероватноћа „неуспеха“.

Биномна дистрибуција зависи од два параметра n и p и може да се означи са $B(n,p)$.

Вредности основних показатеља су:

- Аритметичка средина: $\bar{X}_{BD} = n \times p$;
- Варијанса: $\sigma_{BD}^2 = n \times p \times q$;
- Стандардна девијација: $\sigma_{BD} = \sqrt{n \times p \times q}$;
- Модус $n \times p - q \leq (Mo = k) \leq n \times p + p$;
- I Пирсонов коефицијент: $\beta_1 = \frac{(q-p)^2}{n \times p \times q}$;
- II Пирсонов коефицијент: $\beta_2 = 3 + \frac{1-6 \times p \times q}{n \times p \times q}$.

Код биномне расподеле, варијанса је увек мања од аритметичке средине. Биномна расподела може имати један или два модуса. Уколико $(n+1)\times p$ није цео број, дистрибуција има један модус. С друге стране, дистрибуција има два модуса уколико је $(n+1)\times p$ цео број, тако да важи: $Mo^1 = n\times p - q = (n+1)\times p - 1$ и $Mo^2 = n\times p + p = (n+1)\times p$.

У случају да је $p = q = 0{,}5$ биномна дистрибуција је симетрична. Уколико важи да је $p<q$, дистрибуција је позитивно асиметрична. Обрнуто, када је $p>q$, дистрибуција је негативно асиметрична. У зависности од вредности параметра p, дистрибуција може бити исте спљоштености, спљоштенија или издуженија у поређењу са нормалном дистрибуцијом. Ако број понављања n неограничено расте, биномна дистрибуција тежи н стандардној нормалној расподели.

Биномна дистрибуција има честу примену у статистици приликом описивања могућег броја случајева појављивања догађаја у низу понављања експеримента. Примењује се и у статистичком закључивању код расподеле пропорције узорка.

Биномна дистрибуција се користи у контроли квалитета робе и контроли производног процеса. Поред наведеног има широку примену у биолошким истраживањима, посебно у генетици.

Пример 8.

Одредити расподелу случајне променљиве *X*, где важи да је променљива *X* број женских телади у три узастопна тељења. Претпоставка је да се при сваком тељењу добија једно теле и да су оба пола подједнако вероватна. Израчунати очекивану вредност, модус, варијансу и показатеље облика расподеле (први и други Пирсонов коефицијент).

Решење:

Случајна променљива *X* има биномну расподелу где је број понављања експеримента $n = 3$ и вероватноћа реализације успеха $p = 0{,}5$, тако да се може записати следеће: *B(3;0,5)*.

Вероватноћа да се приликом три узатопна тељења не добије ниједно женско теле, односно да i узме вредност 0, може се израчунати на следећи начин:

$$p_{(0)} = \binom{3}{0} \times 0{,}5^0 \times 0{,}5^3 = 0{,}1250.$$

На сличан начин могуће је израчунати вероватноћу да се приликом три узастопна тељења отеле једно, два или три женска телета, што значи да ће i узимати вредности 1,2 и 3 респективно:

$$p_{(1)} = \binom{3}{1} \times 0{,}5^1 \times 0{,}5^2 = 0{,}3750;$$

$$p_{(2)} = \binom{3}{2} \times 0{,}5^2 \times 0{,}5^1 = 0{,}3750;$$

$$p_{(3)} = \binom{3}{3} \times 0{,}5^3 \times 0{,}5^0 = 0{,}1250.$$

С обзиром на то да смо у претходном делу представили закон расподеле случајне променљиве X, можемо констатовати да је збир свих могућих вероватноћа неког експеримента једнака 1. У конкретном примеру важи:

$$p_{(0)} + p_{(1)} + p_{(2)} + p_{(3)} = 0{,}1250 + 0{,}3750 + 0{,}3750 + 0{,}1250 = 1.$$

У наставку су представљене вредности аритетичке средине, модуса, варијансе и Пирсонових коефицијената респективно:

- Аритметичка средина: $\bar{X}_{BD} = n \times p = 3 \times 0{,}5 = 1{,}5$;
- Модус:

$$n \times p - q \leq Mo \leq n \times p + p$$
$$3 \times 0{,}5 - 0{,}5 \leq Mo \leq 3 \times 0{,}5 + 0{,}5$$
$$1 \leq Mo \leq 2$$
$$\text{Mo}^1 = 1 \text{ и } \text{Mo}^2 = 2;$$

- Варијанса: $\sigma_{BD}^2 = n \times p \times q = 3 \times 0{,}5 \times 0{,}5 = 0{,}75$;
- I Пирсонов коефицијент: $\beta_1 = \frac{(q-p)^2}{n \times p \times q} = \frac{(0{,}5-0{,}5)^2}{3 \times 0{,}5 \times 0{,}5} = 0$;
- II Пирсонов коефицијент: $\beta_2 = 3 + \frac{1-6 \times p \times q}{n \times p \times q} = 3 + \frac{1-6 \times 0{,}5 \times 0{,}5}{3 \times 0{,}5 \times 0{,}5} = 2{,}33.$

3.2.2. Поасонова дистрибуција

Поасонову дистрибуцију први пут је дефинисао француски математичар Симеон Денис Поасон (Siméon Denis Poisson), 1837. године. Поасонова дистрибуција је у примени од прве половине 19. века и то као веома значајна у неким специфичним истраживањима. Поасонова расподела се често назива закон малих бројева и модел је за расподелу догађаја који се ретко појављују са константном вероватноћом. Поасонова дистрибуција се примењује у контроли квалитета робе или неисправних производа у производним процесима одређене величине, испитивањима саобраћајних удеса, контроли пристизања превозних средстава у станице, итд. У биолошким истраживањима примењује се у моделирању броја мутација гена, броју ретких животиња на одређеној територији, броју микроорганизама на микроскопском пољу, броју ретких обољења. Сва ова испитивања имају заједничку карактеристику да се региструју као прекидне варијабле.

Поасонова дистрибуција је теоријска дистрибуција која се односи на прекидна обележја. Вредности обележја X јесу цели ненегативни бројеви $0,1,2,\dots,n,\dots$. За разлику од биномне дистрибуције број модалитета случајне променљиве је бесконачан. Вероватноће Поасонове дистрибуције зависе од једног параметра и то је параметар m. Параметар m у дистрибуцији представља просечан број наступања неког догађаја у одређеном временском интервалу, јединици површине или запремине.

Вероватноће Поасонове дистрибуције дате су следећим изразом:

$$p_{(i)} = e^{-m} \times \frac{m^i}{i!},$$

где је:

e - Ојлеров број (Напиерова константа) основна природног логаритма ($e \approx 2{,}71828$);

m – позитиван број, параметар Поасонове дитрибуције.

Вредности основних показатеља Поасонове дистрибуције:

- Аритметичка средина: $\bar{X}_{PD} = m$;
- Варијанса: $\sigma_{PD}^2 = m$;
- Стандардна девијација: $\sigma_{PD} = \sqrt{m}$;
- Модус $m - 1 \leq (Mo = k) \leq m$;
- I Пирсонов коефицијент: $\beta_1 = \frac{1}{m}$;
- II Пирсонов коефицијент: $\beta_2 = 3 + \frac{1}{m}$.

Код Поасонове расподеле аритметичка средина и варијанса су једнаке. У случају да параметар m није цео број, Поасонова расподела има један модус, док у случају да је m цео број постоје два модуса ($Mo^1 = m - 1$ и $Mo^2 = m$). Поасонова расподела је позитивно асиметрична и издужена у поређењу са нормалном расподелом.

Поасонова дистрибуција је гранични облик биномне дистрибуције. Када се број експеримената у Бернулијевом процесу повећава, јавља се проблем израчунавања вероватноће да варијабла X узме одређену вредност према формули за биномну дистрибуцију. За биномну дистрибуцију вероватноће се могу апроксимирати Поасоновом формулом ако је вероватноћа наступања неког догађаја p мала, ако је n велико и ако важи: $m = n \times p < 10$.

Пример 9.

Познато је да је 2% мишева оболело од канцера. Израчунати вероватноћу да у узорку од 100 мишева, више од једног миша има канцер.

Решење

Број оболелих мишева има биномну расподелу B(100;0,02). Како је вероватноћа обољења мала ($p = \frac{2}{100} = 0{,}02$) и $m = n \times p = 100 \times 0{,}02 = 2 < 10$, биномна расподела се може апроксимирати Поасоновом расподелом. Вероватноћа да више од једног миша има канцер може се израчунати на следећи начин:

$$P(X > 1) = 1 - p_{(0)} - p_{(1)} = 1 - e^{-2} - 2 \times e^{-2} = 1 - 0{,}1353 - 0{,}2707 = 0{,}5940.$$

3.3. Непрекидне теоријске дистрибуције

3.3.1. Нормална дистрибуција

Најважнији модел теоријске дистрибуције вероватноће је *нормална* или *Гаусова дистрибуција*. Значај наведеног облика дистрибуције у статистичкој теорији и статистичким истраживањима огледа се у томе што се многе емпиријске појаве моделирају нормалном дистрибуцијом. Нормална дистрибуција има значајну примену у статистичкој инференцији. Параметарска статистика је заснована на претпоставци да основни скуп коме припада узорак има нормалну дистрибуцију.

Нормалну расподелу је први дефинисао Абрахам де Моивре (Abraham de Moivre), 1733. године, као гранични облик биномне дистрибуције посматрајући шта се дешава са биномном расподелом када број експеримената бесконачно расте.

У другој половини XVIII века, овакав облик дистрибуције проучавао је и француски математичар Пјер де Лаплас (*Pierre de Laplace*). Почетком XIX века, немачки математичар Карл Фридрих Гаус (*Carl Friedrich Gauss*) је писао о карактеристикама и применама нормалне расподеле у моделирању случајних грешака приликом мерења у астрономији. Управо због овог доприноса, нормална расподела назива се и Гаусова расподела.

Pierre de Laplace
(1749–1827)

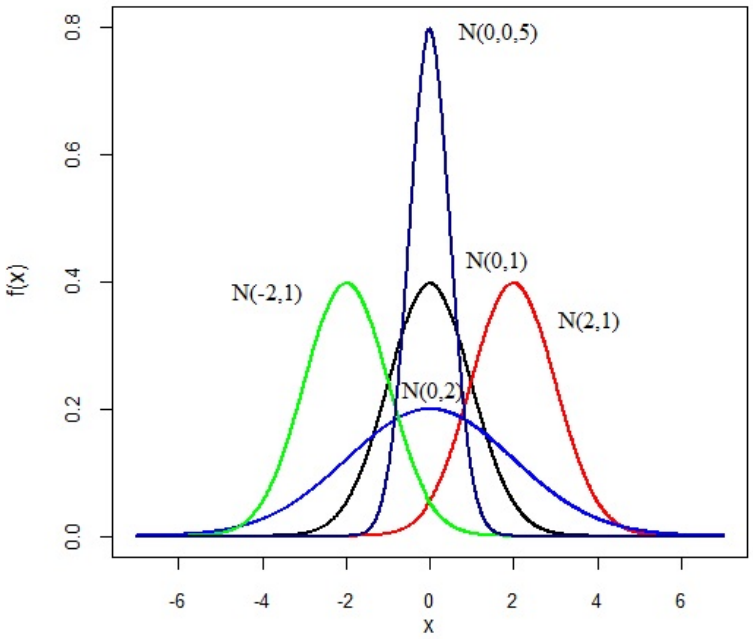
Carl Friedrich Gauss
(1777–1855)

Нормална дистрибуција је непрекидна теоријска дистрибуција. Непрекидна случајна променљива има нормалну расподелу ако је $X \in (-\infty, +\infty)$ и ако је закон вероватноће (функција густине вероватноће):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{1}{2}\times\left(\frac{x-\mu}{\sigma}\right)^2} .$$

Закон вероватноће нормалне дистрибуције зависи од два параметра, аритметичке средине μ и стандардне девијације σ. Стандардна нормална дистрибуција има аритметичку средину 0 и стандардну девијацију 1. Нормална дистрибуција је графички представљена континуираном заобљеном кривом која у односу на X осу има звонасти облик.

***Графикон 7**. Функција густине нормалне расподеле за различите вредности μ и σ*



Особине нормалне расподеле су следеће:

- Површина коју крива заклапа са X-осом представља збир вероватноћа и износи 1.
- Нормална расподела је симетрична у односу на вредност $x = \mu$ тако да је $P(X < \mu) = P(X > \mu) = 0{,}5$.
- Максимум функције густине вероватноће је у тачки μ.
- Аритметичка средина, модус и медијана се поклапају и имају вредност μ.
- Први Писонов коефицијент β_1 је једнак 0, док је други Пирсонов коефицијент $\beta_2 = 3$.
- Уколико $X \to \pm\infty$, функција $f(X) \to 0$.
- Како би се израчунала вероватноћа вероватноћа $P(a < X < b)$ случајне променљиве X која има стандардну нормалну дистрибуцију, користе се таблице нормалне дистрибуције. У таблицама су приказане вероватноће $\Phi(\mathrm{a}) = P(0 < X < \mathrm{a})$.

Уколико случајна променљива X нема стандардну нормалну дистрибуцију, њена очекивана вредност (аритметичка средина) није 0 и стандардна девијација није 1. У том случају, како би се могле користити статистичке таблице, потребно је прво извршити

трансформацију (стандардизација) случајне променљиве X у стандардизовану случајну променљиву Z. Трансформација се изоди на основу следећег израза:

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

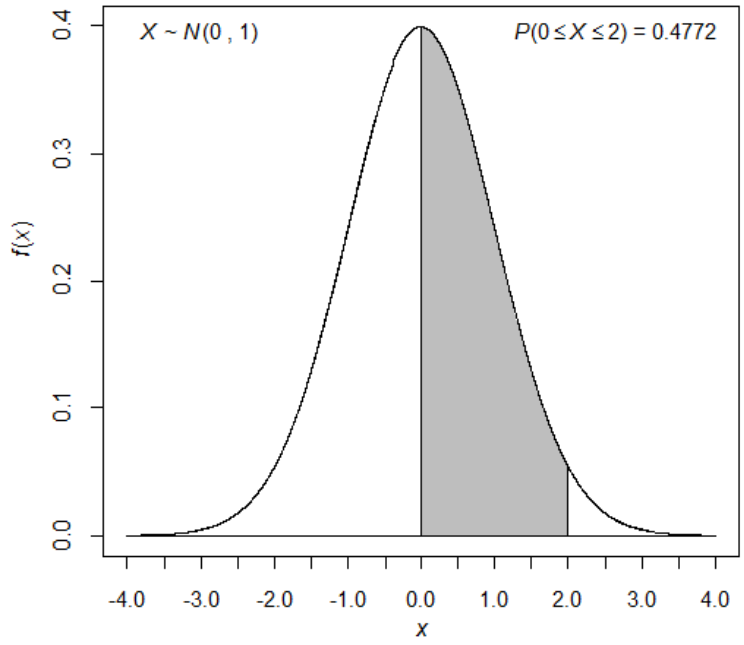
тако да важи:

$$f(Z) = \frac{1}{\sqrt{2\pi}} \times e^{-\frac{1}{2} \times Z^2}.$$

Пример 10:

а) Израчунати вероватноћу да ће случајно променљива X узети вредност између 0 и 2, уколико је познато да случајно променљива X поседује стандардну нормалну расподелу ($\mu = 0, \sigma = 1$).

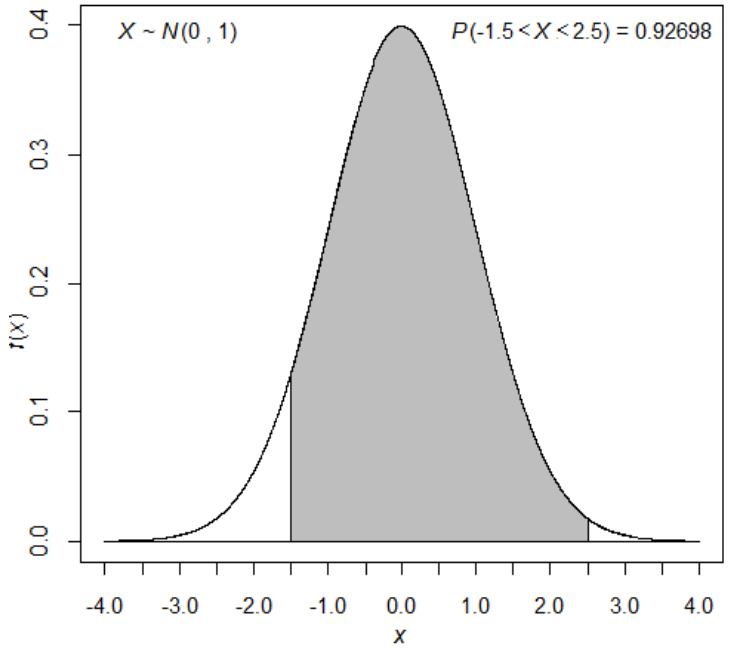
Ако случајна променљива X поседује стандардну нормалну расподелу, вероватноћа да ће X узети неку вредност у интервалу од 0 до 2 ($P(0 \leq X \leq 2)$), може се израчунати на основу таблице нормалне расподеле. Осенчена површина на графикону представљеном у наставку која може да се очита из табица нормалне расподеле репрезентује тражену вероватноћу: $P(0 \leq X \leq 2) = 0{,}4772$.



б) Израчунати вероватноћу да ће случајно променљива X узети вредност између -1,5 и 2,5, уколико је познато да случајно променљива X поседује стандардну нормалну расподелу.

Уколико случајна променљива X има нормалну расподелу, обрачун вероватноће заједно са графичким приказом, да ће случајно променљива X узети неку вредност у интервалу од -1,5 до 2,5 представљен је у наставку:

$$P(-1{,}5 \leq X \leq 2{,}5) = P(-1{,}5 \leq X \leq 0) + P(0 \leq X \leq 2{,}5) = 0{,}4332 + 0{,}4938 = 0{,}9270.$$



в) Израчунати вероватноћу да ће случајно променљива X узети неку вредност у интервалу од 8 до 9 уколико је очекивана вредност једнака 12, а стандардна девијација 2.

Како је нарушен услов да је очекивана вредност 0 и стандардна девијација 1, неопходно је извршити трансформацију вредности обележја, а затим израчунати тражену вероватноћу помоћу таблица нормалне расподеле.

$$P(8 < X < 9) = P\left(\frac{8-12}{2} < X < \frac{9-12}{2}\right) = P(-2 < X < -1{,}5) = 0{,}4772 - 0{,}4332 = 0{,}0440.$$

3.3.2. Студентова *t* – дистрибуција

Студентову или t – дистрибуцију, први је дефинисао енглески хемичар и статистичар Вилијам Госет (William Gosset) 1908. године. Госет је својевремено своје научне радове потписивао псеудонимом „Студент“, тако да се усталио назив Студентова t – дистрибуција.

William Sealey Gosset
(1876 – 1937)

Студентова дистрибуција се односи на случајно променљиву t која представља трансформисано обележје дато изразом:

$$t = \frac{\bar{X} - \mu}{S_{\bar{X}}},$$

где је:

$\bar{X}$ - аритметичка средина узорка;

μ - очекивана вредност (средина основног скупа);

$S_{\bar{X}}$ - оцењена стандардна грешка аритметичке средине.

Оцењена стандардна грешка аритметичке средине добија се на основу оцењене стандардне девијације основног скупа S, применом следећег израза:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}},$$

где је:

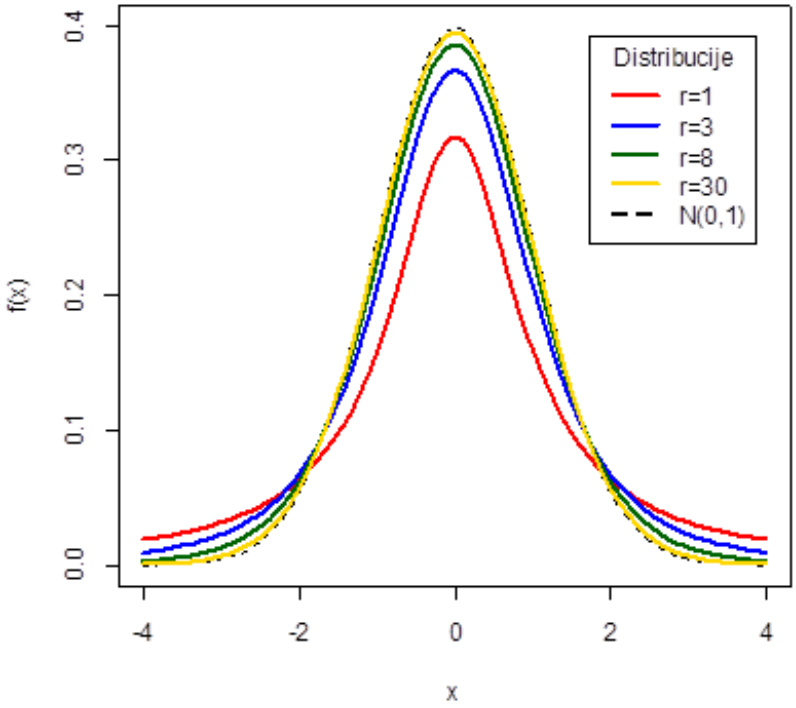
$$S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \ .$$

Количник t има Студентову расподелу уколико се претпостави да обележје X има нормалну расподелу независно од величине узорка, или уколико је величина узорка већа од 30.

Параметар који дефинише Студентову дистрибуцију је степен слободе r ($r = n - 1$). Са порастом степени слободе (r), Студентова дистрибуција се по својим

карактеристикама приближава стандардној нормалној дистрибуцији. Уколико је r = 30, разлика између Студентове и нормалне дистрибуције је незнатна.

Графикон 8. *Студентова расподела за различите степене слободе*



Неке од карактеристика Студентове дистрибуције су следеће:

- Функција густине вероватноће зависи од једног параметра који се назива степен слободе;
- Студентова дистрибуција има сличан облик као стандардна нормална дистрибуција само што је шира и положенија тј. има већу вероватноћу екстремних вредности;
- Како расте број степени слободе, обликом је све сличнија нормалној расподели;
- Примењује се у израчунавању интервала поузданости и тестирању хипотеза о разлици између два узорка уколико обележје има нормалну расподелу и варијансе основних скупова нису познате.

Основни показатељи t – дистрибуције са r степени слободе су:

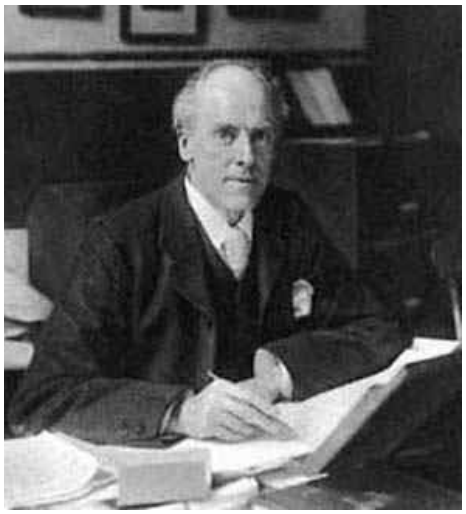
- Аритметичка средина: $\bar{X}_{tD} = 0, r > 1$;
- Варијанса: $\sigma_{tD}^2 = \frac{r}{r-2}, r > 2$;
- Стандардна девијација: $\sigma_{tD} = \sqrt{\frac{r}{r-2}}$;
- Модус $Mo = 0$;
- I Пирсонов коефицијент: $\beta_1 = 0$;

- II Пирсонов коефицијент: $\beta_2 = 3 + \frac{6}{r-4}, r > 4.$

У таблицама Студентове дистрибуције дате су вредности случајне промењиве X за дату вредност степена слободе r и вредност $\alpha = P(|X| > t_{n-1;\alpha})$. Уколико је нпр. $r = 5$, важи да је за $\alpha = 0{,}05$, $t_{5;0,05} = 2{,}571$.

3.3.3. χ^2– дистрибуција

$\boldsymbol{\chi^2}-$ дистрибуција је непрекидна теоријска дистрибуција коју је први пут увео Еби (*Ernst Carl Abbe*) 1863. године. Независно је дефинисао и применио Карл Пирсон (*Karl Pearson*) енглески математичар, биометричар и статистичар, 1900. године.



Karl Pearson
(1857 – 1936)

Закон вероватоће ове дистрибуције зависи од једног параметра и то је број степени слободе $\boldsymbol{r}$ ($\boldsymbol{r} \in \boldsymbol{N}$).

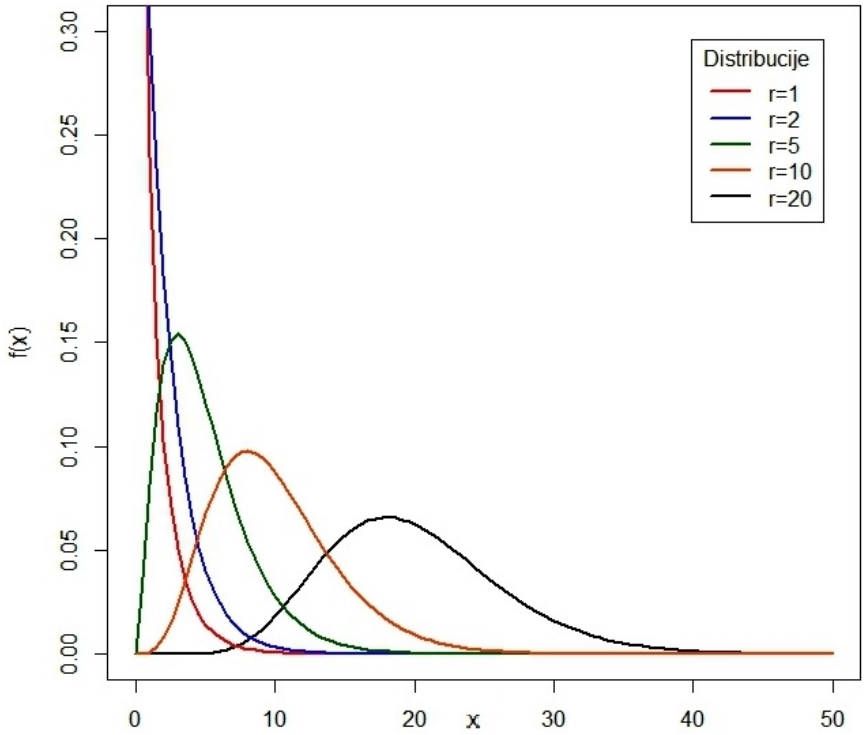
Ако су $\boldsymbol{X_1, X_2, \dots, X_r}$ независне случајне промењиве које имају стандардну нормалну расподелу ($\boldsymbol{\mu = 0}$ и $\boldsymbol{\sigma = 1}$), тада $\boldsymbol{X_1^2 + X_2^2 + \cdots + X_r^2}$ има $\boldsymbol{\chi^2}$ расподелу са $\boldsymbol{r}$ степена слободе.

Основни показатељи $\boldsymbol{\chi^2}-$ дистрибуције са r степени слободе су:

- Аритметичка средина: $\bar{X}_{\chi^2{}_D} = r$;
- Варијанса: $\sigma^2_{\chi^2{}_D} = 2 \times r$;
- Стандардна девијација: $\sigma_{\chi^2{}_D} = \sqrt{2 \times r}$;
- Модус $Mo = 0\ (r \leq 2), Mo = r - 2\ (r > 2)$;
- I Пирсонов коефицијент: $\beta_1 = \frac{8}{r}$;
- II Пирсонов коефицијент: $\beta_2 = 3 + \frac{12}{r}$.

Како је дефинисана као збир квадрата, χ^2- дистрибуција је увек ненегативна. Минимална вредност дистрибуције је нула, док уколико вредност случајне променљиве X тежи бесконачности, χ^2- дистрибуција асимптотски тежи нули. Њена крива је асиметрична у десно, а са порастом r смањују се асиметричност и издуженост **у** поређењу са нормалном дистрибуцијом. Тако, за велике вредности r, χ^2- дистрибуција је приближна нормалној дистрибуцији са параметрима $\mu = r$ и $\sigma = \sqrt{2 \times r}$.

Графикон 9. χ^2*– дистрибуција за различите степене слободе*

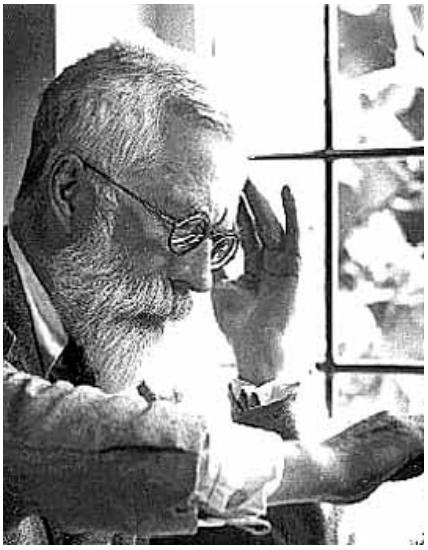


χ^2- дистрибуција се примењује у тестирању статистичких хипотеза. Најчешће се примењују код непараметарских тестова *сагласности емпиријске и теоријске* дистрибуције и *теста независности*.

У таблицама χ^2 дистрибуције дате су вредности случајне променљиве X за дату вредност степена слободе r и вредност $\alpha = P(X > \chi^2_{r;\alpha})$. Уколико је нпр. r =6, важи да је за $\alpha = 0{,}05$, $\chi^2_{6;0,05} = 12{,}59$. То значи да ће случајна променљива која има χ^2 расподелу са 6 степена слободе у 95% случајева добити вредност мању, а у 5% случајева већу или једнаку вредност од 12,59.

3.3.4. Фишерова *F* – дистрибуција

Фишерова (Фишер – Снедекорова) дистрибуција припада групи непрекидних теоријских дистрибуција. Добила је име по познатом енглеском статистичару и генетичару Роналд Фишеру (*Ronald Fisher*) који је по први пут дефинисао 1924. године.



Sir Ronald Aylmer Fisher
(1890 – 1962)

Случајна променљива *F* дефинисана је као количник оцењених варијанси два независна случајна узорка чије су величине n_1 и n_1:

$$F = \frac{S_1^2}{S_2^2},$$

где је:

$$S_1^2 = \frac{\sum(X_{1i} - \bar{X}_1)^2}{n_1 - 1},$$

односно:

$$S_2^2 = \frac{\sum(X_{2i} - \bar{X}_2)^2}{n_2 - 1}.$$

Фишерова дистрибуција зависи од два параметра, односно два степена слободе r_1 и r_2 ($r_1 = n_1 - 1$; $r_2 = n_2 - 1$).

Како је дефинисана као количник две суме квадрата, *F* – дистрибуција је увек ненегативна. Минимална вредност Фишерове дистрибуције је нула. Уколико вредност случајне поменљиве *X* тежи бесконачности, Фишерова дистрибуција асимптотски тежи нули. Фишерова дистрибуција је изразито асиметрична у десно, а са порастом степени слободе, односно величине узорака, тежи ка симетричности.

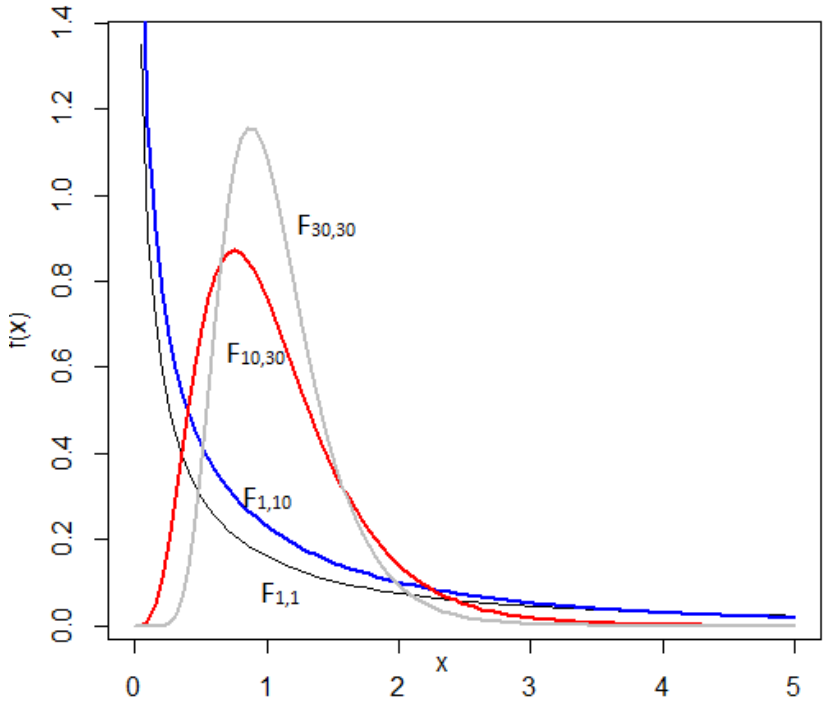
Фишерова дистрибуција може да се дефинише и као количник две независне случајне променљиве које имају $\chi^2 -$ дистрибуцију. Наиме, ако су X_1 и X_2 независне случајне

променљиве које имају χ^2 – дистрибуцију са r_1 и r_2 степена слободе, количник $F = \frac{X_1/r_1}{X_2/r_2}$ има Фишерову дистрибуцију са r_1 и r_2 степена слободе.

Фишерова дистрибуција има широку примену, а најчешће се користи код тестирања једнакости две варијансе и код тестирања разлика три или више аритметичких средина, односно у примени метода анализе варијансе.

Таблице F – дистрибуције су формиране за различите прагове значајности $\alpha = P(X > F_{r_1;r_2;\alpha})$. Најчешће се користе таблице за праг значајности $\alpha = 0{,}05$ и $\alpha = 0{,}01$. У таблицама су бројеви у заглављу вредности првог степена слободе r_1, док су бројеви у предколони вредности другог степена слободе r_2,. Тако нпр. за степене слободе 3 и 16 и $\alpha = 0{,}05$, таблична вредност је 3,24 и означава вредност на X-оси тако да је $P(X > 3{,}24) = 0{,}05$.

Графикон 10. *Фишерова F – дистрибуција за различите степене слободе*



Контролна питања

1. Како дефинишемо емпиријске, а како теоријске дистрибуције?
2. Како се деле теоријске дистрибуције?
3. Навести неке од прекидних теоријских дистрибуција.
4. Навести неке од непрекидних теоријских дистрибуција.
5. Навести карактеристике Биномне дистрибуције.
6. Навести карактеристике Поасонове дистрибуције.
7. Навести карактеристике Нормалне дистрибуције.
8. Навести карактеристике Студентове дистрибуције.
9. Навести карактеристике Фишерове дистрибуције.
10. Када се у статистичкој инференцији користе табличне вредности Нормалне, а када табличне вредности Студентове дистрибуције?

4. ИНФЕРЕНЦИЈАЛНА СТАТИСТИКА

Приликом истраживања, често се дешава да је потребно извести закључке о основном скупу, односно популацији, а да при том немамо на располагању све њихове податке. Део статистике који се бави доношењем закључака о основном скупу на основу дела његових јединица назива се *инференцијална статистика*.

Испитивање дела неког скупа ради оцене карактеристика целокупног скупа изводи се методом узорка. На основу анализе јединица узорка, процењује се вредност параметра основног скупа. Основни скуп често испољава карактеристике које су приближне карактеристикама неке од теоријских дистрибуција. Уколико се зна којој теоријској дистрибуцији се расположиви подаци најбоље прилагођавају, лакше се долази до закључака о самом основном скупу.

4.1. Метод узорка у истраживачком раду

Део основног скупа који је издвојен како би се на њему изводила статистичка анализа, назива се *узорак*. Статистичка теорија узорака дели се на теорију малог и теорију великог узорка, при чему као основа за поделу служи број јединица у узорку. Малим узорком сматра се узорак величине до тридесет јединица, а узорак чија је величина већа од тридесет јединица сматра се великим узорком.

Узорак који у највећој мери одражава карактеристике основног скупа назива се репрезентативни узорак. Репрезентативност узорка постиже се правилним постављањем плана узорка и правилним начином избора јединица у узорак.

Методе за избор јединица узорка можемо поделити на: *методе избора на основу вероватноће* и *методе избора без примене вероватноће*.

Методе избора на основу вероватноће подразумевају да се примени поступак избора који не фаворизује ни једну јединицу посебно, односно да све јединице имају унапред познату вероватноћу да буду изабране у узорак. Применом ових метода добијају се следећи планови узорка:

- прост случајни узорак;
- систематски случајни узорак;
- стратификовани случајни узорак;
- кластер случајни узорак.

Методе избора без примене вероватноће, засноване су на поступцима избора јединица који не зависе од теорије вероватноће. На овај начин добијају се узорци формирани на основу слободне процене истраживача или на основу сврхе истраживања.

Прост случајни узорак је узорак који се добија тако што све јединице основног скупа имају исту вероватноћу да буду изабране у узорак, при чему избор једне не утиче на

избор осталих јединица. Прост случајни узорак може бити изабран са или без понављања (враћања). Узорак са понављањем подразумева да једна јединица основног скупа може да се појави у узорку више пута. Узорак без понављања подразумева да једна јединица основног скупа може да се појави у узорку само једном. Избор јединица из популације у узорак може се извести помоћу таблице случајних бројева, техником лутријског избора или уз помоћ рачунара.

Систематски узорак је узорак код кога се јединице из основног скупа бирају једнаким интервалима времена, простора или поретка (вакциницање деце одређене године старости, награда за сваког стотог купца неког производа, итд.).

Стратификовани и *кластер узорак* су узорци који се добијају када се основни скуп (популација) подели на стратуме или кластере, након чега се случајно бирају јединице из сваког стратума, односно из сваког кластера. Стратификовани узорак се бира у случају када су варијације унутар стратума мале у односу на варијације између стратума, а кластер узорак у супротном случају. Поред наведеног, битно је истаћи да се разликују стратификовани узорак са пропорционалним распоредом и диспропорционални стратификовани узорак.

4.2. Дистрибуција средина узорака

Сваки параметар узорка има своју дистрибуцију. Познавање карактеристика те дистрибуције доприноси бољем разумевању оцена и тестова на основу узорка.

Полазећи од основног скупа који броји N јединица, уколико изаберемо један прост случајан узорак од n јединица, на основу њега се може извести оцена непознатих параметара основног скупа. Уколико се претпостави да се из основног скупа изаберу сви могући узорци чији је број k и израчунају њихове аритметичке средине, на основу издвојених аритметичких средина може се формирати дистрибуција фреквенција аритметичких средина.

Број узорака величине n јединица који може да се добије из једног основног скупа величине N јединица утврђује се на основу следећих израза:

- узорци са понављањем: $k = N^n$;
- узорци без понављања: $k = \binom{N}{n} = \frac{N!}{n!(N-n)!}$.

Дистрибуција аритметичких средина простих случајних узорака величине n има нормалан распоред уколико и основни скуп има нормалан распоред, без обзира на величину узорка.

Када основни скуп има распоред произвољног облика са аритметичком средином μ и варијансом σ^2, распоред аритметичких средина свих простих случајних узорака тежи

нормалном распореду уколико величина узорка $n \to \infty$. Претходно наведено представља једну од најзначајнијих теорема у статистици – ***централна гранична теорема***.

За дистрибуцију средина узорака могу се израчунати и њени показатељи (аритметичка средина и варијанса). Аритметичка средина дистрибуције средина узорака израчунава се на следећи начин:

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{X}_i}{k},$$

где су $\bar{X}_i$ аритметичке средине узорака, док k представља број узорака. На тај начин, аритметичка средина дистрибуције средина узорака једнака је аритметичкој средини основног скупа:

$$\bar{\bar{X}} = \mu.$$

Варијанса дистрибуције средина узорака израчунава се на основу следећег израза:

$$\sigma_{\bar{X}}^2 = \frac{\sum_{i=1}^{k} (\bar{X}_i - \mu)^2}{k}.$$

Уколико је позната вредност варијансе основног скупа, варијанса дистрибуције аритметичких средина узорака, у случају простих случајних узорака без понављања, може се израчунати на следећи начин:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \times \frac{N-n}{N-1},$$

где је:

σ^2 - варијанса основног скупа;

n - величина узорка;

$\frac{N-n}{N-1}$ - корективни фактор.

У случају да су примењени прости случајни узорци са понављањем, варијанса аритметичких средина узорка је:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.$$

Вредност варијансе дистрибуције средина узорака је мања од варијансе основног скупа. Са повећањем величине узорка (n), вредност варијансе дистрибуције средина узорака се смањује и тежи нули. Што је већа величина узорка боље и прецизније се може оценити параметар основног скупа.

Како важи да је $\frac{N-n}{N-1} < 1$ и $n > 1$, варијанса аритметичких средина узорака без понављања је мања од варијансе аритметичких средина узорака са понављањем. У случају да је N велико у поређењу са n, важи: $\frac{N-n}{N-1} \approx 1$.

Стандардна девијација дистрибуције аритметичких средина узорака назива се стандардна грешка аритметичке средине и утврђује се на основу следећег израза:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum_{i=1}^{k}(\bar{X}_i - \mu)^2}{k}}.$$

Уколико је позната варијанса или стандардна девијација основног скупа, стандардна грешка аритметичке средине у случају узорака без понављања може се израчунати на следећи начин:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n} \times \frac{N-n}{N-1}}.$$

У случају узорка са понављањем, грешка аритметичке средине може се израчунати на следећи начин:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}.$$

4.3. Оцене на основу узорка

У практичном раду, у сврху доношења закључака о карактеристикама основног скупа, узима се само један узорак довољне величине, на основу којег се даље оцењују, односно процењују непознати параметри основног скупа. Оцена параметара основног скупа примењује се у следећим случајевима:

- када је непозната величина основног скупа, односно укупан број јединица N;
- када се не могу утврдити све вредности обележја основног скупа;
- када је основни скуп бесконачан.

Вредности израчунате на основу узорка нису тачне, праве вредности, већ су то приближне вредности, односно оцене одговарајућих параметара основног скупа. Вредности неког параметра израчунате на основу узорка су тачкасте оцене параметара основног скупа. Оцена непознатог параметра основног скупа биће тачнија, односно ближа правој вредности, што је узорак већи и што је варијабилитет појаве која се анализира мања. У случају да појава која се анализира не варира, оцењена вредност аритметичке средине из узорка од само једне вредности обележја представљала би тачну вредност средине основног скупа. Како појаве које се у практичном раду

анализирају, показују већи или мањи варијабилитет, тачкасте оцене из узорка нису довољне да би се оценила вредност непознатог параметра основног скупа, већ се у обзир мора узети и стандардна грешка као показатељ варијабилитета. Другим речима, за оцену непознатих параметара основног скупа на основу узорка, користе се интервалне оцене које у обзир узимају и варијабилитет посматране појаве.

У теорији оцењивања се разликују појмови *оценитељ* и *оцена*. Оценитељ је функција узорка (статистика), док је оцена израчуната вредност оценитеља на основу изабраног узорка. Оценитељ је случајна променљива, док је оцена константа. Табела 1 представљена у наставку даје преглед ознака за параметре основног скупа и паралелно ознаке за оцене на основу узорка.

Табела 1. Ознаке за параметре основног скупа и одговарајуће оцене на основу узорка

Параметар	Основни скуп	Оцена на основу узорка
Аритметичка средина	μ	$\bar{X}$
Стандардна девијација	σ	S
Варијанса	σ^2	S^2
Стандардна грешка аритметичке средине	$\sigma_{\bar{X}}$	$S_{\bar{X}}$

Извор: обрада аутора

Оцена параметара основног скупа на основу узорка, заснована је на теорији да је пожељно да оценитељ поседује нека статистичка теоријска својства. Својства која је пожељно да има оценитељ су: *непристрасност, конзистентност, ефикасност* и *егзостивност*. Сви наведени принципи оцене параметара су пожељне али не и неопходне особине оценитеља.

Оценитељ параметара је *непристрасан* када је његова очекивана вредност једнака параметру основног скупа. Аритметичка средина из узорка је непристрасан оценитељ аритметичке средине основног скупа, јер важи: $E(\bar{X}) = \mu$.

С друге стране, оценитељ $S^{2*} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ није непристрасан оценитељ варијансе основног скупа. Варијанса оцењена на основу узорка постаће непристрасан оценитељ варијансе основног скупа ако делилац у следећем изразу буду степени слободе *n – 1*:

$$E\left(\frac{n}{n-1} \times S^{2*}\right) = E(S^2) = \sigma^2.$$

За оценитељ из узорка се каже да ***конзистентно*** оцењује параметар основног скупа, уколико у ситуацији када n тежи бесконачности, оценитељ из узорка тежи вредности параметра основног скупа уз вероватноћу 1. Ако је оценитељ из узорка конзистентан, са повећањем величине узорка његова вредност се приближава вредности параметра основног скупа. Да би оценитељ био конзистентан, није неопходно да буде непристрасан. Тако су S^2 и S^{2*} конзистентни оценитељи варијансе основног скупа σ^2.

Параметар основног скупа може да се оцени на различите начине. Непристрасан оценитељ тог параметра је *ефикаснији* када је његова вредност приближнија правој вредности параметра основног скупа тј. када има мањи варијабилитет. Релативна ефикасност се изражава односом варијанси оценитеља и то односом мање варијансе оценитеља према већој.

Пример: Аритметичка средина и медијана су непристрасне оцене аритметичке средине основног скупа. Уколико се претпостави да је основни скуп нормално распоређен, аритметичка средина је ефикаснији оценитељ јер релативна ефикасност ова два оценитеља произилази из односа:

$$\sigma_{\bar{X}}^2/\sigma_{Me}^2 = 0{,}64 < 1.$$

Оцењени параметар је *егзостиван* ако садржи сва потребна обавештења о параметру основног скупа. Да би један оценитељ из узорка био егзостиван он треба да је функција параметра основног скупа. Егзостивни оценитељи су аритметичка средина и пропорција узорка.

4.3.1. Израчунавање стандардне грешке аритметичке средине

Стандардна грешка аритметичке средине, у случају када је познат варијабилитет основног скупа (познате вредности стандардне девијације или варијансе), може се израчунати на основу следећег израза:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n} \times \frac{N-n}{N-1}},$$

где је $\frac{N-n}{N-1}$ корективни фактор који се користи ако је позната величина основног скупа N и уколико се примењује узорак без понављања (без враћања).

Ако је узорак узет из великог основног скупа или бесконачног основног скупа, стандардна грешка аритметичке средине се своди на следећи израз:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}}\ .$$

С обзиром на то да су стандардна девијација и варијанса основног скупа најчешће непознате, замењују се оценом из узорка, односно *оцењеном стандардном девијацијом или варијансом*. На основу израчунате оцењене стандардне девијације или варијансе израчунава се оцењена стандардна грешка аритметичке средине на основу следећих израза у зависности од тога да ли је примењен прост случајан узорак без или са понављањем респективно:

$$S_{\bar{X}} = \sqrt{\frac{S^2}{n} \times \frac{N-n}{N-1}} \qquad \text{и} \qquad S_{\bar{X}} = \sqrt{\frac{S^2}{n}}\ .$$

Стандардна грешка аритметичке средине може се израчунати и директно из података узорка на основу радних формула. За негруписане податке оцењена стандардна грешка аритметичке средине утврђује се на следећи начин:

$$S_{\bar{X}} = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n\times(n-1)} \times \frac{N-n}{N}} \quad \text{или} \quad S_{\bar{X}} = \sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n\times(n-1)} \times \frac{N-n}{N}}.$$

Код дистрибуције фреквенција, за оцену стандардне грешке аритметичке средине користе се следећи изрази:

$$S_{\bar{X}} = \sqrt{\frac{\sum f_i(X_i-\bar{X})^2}{n\times(n-1)} \times \frac{N-n}{N}} \quad \text{или} \quad S_{\bar{X}} = \sqrt{\frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{n}}{n\times(n-1)} \times \frac{N-n}{N}}.$$

Стандардна грешка аритметичке средине налази примену у израчунавању интервала поверења за непознату средину основног скупа, као и код теста значајности једне средине.

4.3.2. Интервал поверења за оцену непознате средине основног скупа

Интервал поверења (поузданости) непознатог параметра основног скупа је интервал у коме се са одређеном сигурношћу налази параметар основног скупа. У практичном раду интервал поверења се најчешће утврђије на бази 95 % или 99 %, што значи да је могућност погрешке 5 %, односно 1 %. Могућност погрешке датог интервала назива се праг значајности, а обележава се као $\alpha=0,05$ или $\alpha=0,01$. То би практично значило да од 100 интервала поверења које утврдимо на основу 100 независних простих случајних узорака изабраних из основног скупа, њих 95, односно 99 садржи праву вредност параметра основног скупа, док 5, односно 1 неће садржати праву вредност посматраног параметра.

У случају познатих вредности стандардне девијације или варијансе основног скупа, интервал поверења за оцену непознате средине основног скупа има следећи облик:

$$\bar{X} - Z_\alpha \times \sigma_{\bar{X}} < \mu < \bar{X} + Z_\alpha \times \sigma_{\bar{X}},$$

где је Z_α вредност која се одређује из услова да је случајна променљива Z која има стандардизовану нормалну расподелу налази у интервалу $\pm Z_\alpha$ са вероватноћом $1-\alpha$, што се може изразити као:

$$P(-Z_\alpha < Z < Z_\alpha) = 1-\alpha.$$

У случају да варијанса основног скупа није позната, интервал поверења има следећи облик:

$$\bar{X} - t_{n-1;\alpha} \times S_{\bar{X}} < \mu < \bar{X} + t_{n-1;\alpha} \times S_{\bar{X}},$$

где се $t_{n-1;\alpha}$ одређује из услова да се случајно променљива t која има t – дистрибуцију налази у интервалу $\pm t_{n-1;\alpha}$ са вероватноћом $1-\alpha$, што се може записати као:

$$P\left(-t_{n-1;\alpha} < t < t_{n-1;\alpha}\right) = 1-\alpha.$$

У случају великог узорка, t – дистрибуција се може апроксимирати стандардном нормалном дистрбуцијом, тако да је $(1-\alpha)\times 100(\%)$ интервал поверења за непознату аритметичку средину основног скупа (μ) приближно:

$$\bar{X} - Z_\alpha \times S_{\bar{X}} < \mu < \bar{X} + Z_\alpha \times S_{\bar{X}}.$$

У практичном раду, сматра се да је узорак величине $n > 30$ велики узорак, док се узорци мањи од 30 јединица посматрања сматрају малим узорцима.

Сваки интервал поверења има своју доњу (L_1) и своју горњу границу (L_2). На основу оцењеног интервала поверења може се оценити и тотал основног скупа на основу следећег израза:

$$N \times L_1 < N\mu < N \times L_2.$$

Производ табличне вредности и стандардне грешке у изразу за $(1-\alpha)\times 100(\%)$ интервал поверења назива се *маргинална грешка* или *грешка узорка* и представља процену одстојања вредности параметра од његове оцене.

Пример 11:

Дата је потрошња воћа (кг) 11 домаћинастава. Оценити просечну потрошњу воћа у основном скупу на основу узорка за праг значајности $\alpha = 0{,}05$ (95% интервал поверења) ако је: а) $\sigma = 3$; б) непознат варијабилитет основног скупа

Потрошња воћа (кг) X_i	X_i^2
10	100
8	64
3	9
17	289
28	784
6	36
7	49
10	100
18	324
9	81
5	25
Σ121	1.861

Решење:

а) Како је познат варијабилитет основног скупа 95% интервал поверења ће имати следећи облик:

$$\bar{X} - Z_{0,05} \times \sigma_{\bar{X}} < \mu < \bar{X} + Z_{0,05} \times \sigma_{\bar{X}} .$$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{121}{11} = 11; \quad \sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{3^2}{11}} = \sqrt{0{,}82} = 0{,}9045; \quad Z_{0,05} = 1{,}96.$$

$$11 - 1{,}96 \times 0{,}9045 < \mu < 11 + 1{,}96 \times 0{,}9045$$

$$9{,}23 < \mu < 12{,}77.$$

Дакле, аритметичка средина основног скупа (просечна потрошња у основном скупу) креће се у интервалу од 9,23 до 12,77 кг, што је установљено на основу 95% интервала поверења.

б) Како варијабилитет основног скупа није познат, исти је неопходно оценити, што значи да је потребно искористити следећи 95% интервал поверења за оцену непознате аритметичке средине основног скупа на основу малог узорка ($n \leq 30$):

$$\bar{X} - t_{n-1;\alpha} \times S_{\bar{X}} < \mu < \bar{X} + t_{n-1;\alpha} \times S_{\bar{X}}$$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{121}{11} = 11; \quad S_{\bar{X}} = \sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n \times (n-1)}} = \sqrt{\frac{1.861 - \frac{121^2}{11}}{11 \times (11-1)}} = 2{,}1950;$$

$$t_{11-1;0,05} = 2{,}228.$$

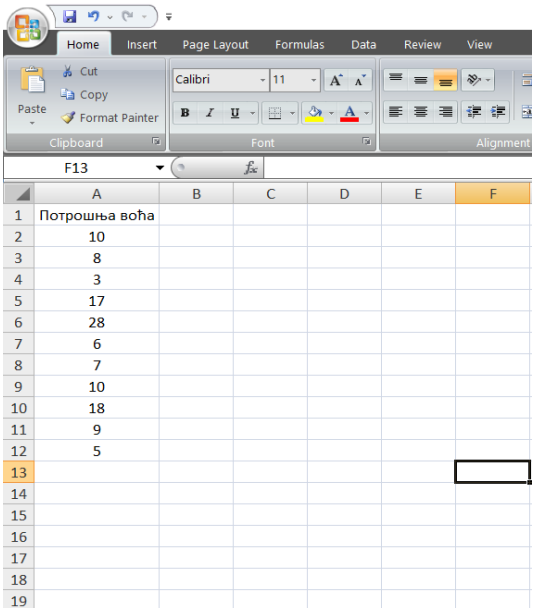
$$11 - 2{,}228 \times 2{,}1950 < \mu < 11 + 2{,}228 \times 2{,}1950;$$

$$6{,}11 < \mu < 15{,}89.$$

Вежба 6. Оцена непознате аритметичке средине основног скупа на основу узорка применом Microsoft Excel-а

Приликом оцене непознате аритметичке средине основног скупа на основу узорка, *Microsoft Excel* пружа могућност оцене у случају када је непознат варијабилитет основног скупа, што је у практичном раду много чешћи случај. Конкретно, *Microsoft Excel* ће приликом дефинисања овог интервала дати једну вредност, коју је потребно одузети и додати од аритметичке средине како би се добиле горња и доња граница интервала. Другим речима, *Microsoft Excel* даје оцену маргиналне грешке која представља производ оцењене стандардне грешке ($S_{\bar{X}}$) и критичне вредности из одговарајуће таблице (у овом примеру $t_{n-1;\alpha}$).

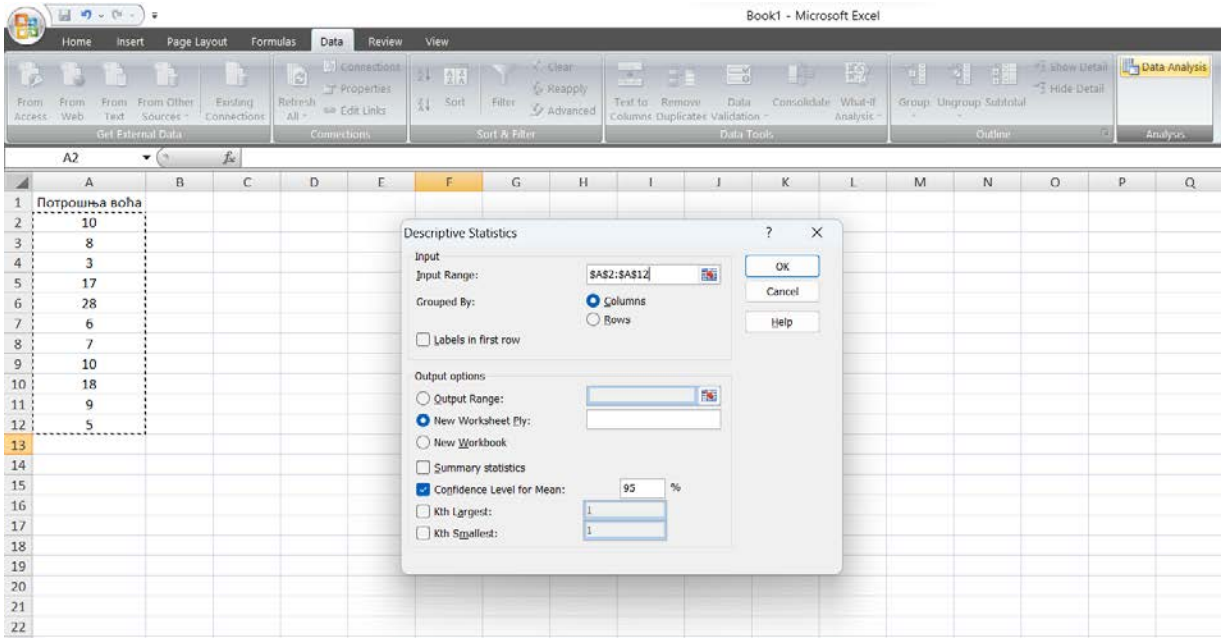
1. Почетни изглед табеле у *Microsoft Excel*-у има следећи изглед:



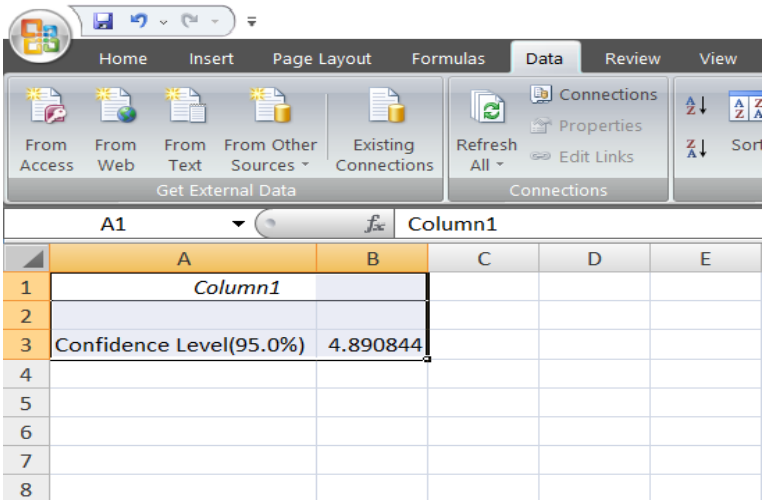
	A	B	C	D	E	F
1	Потрошња воћа					
2	10					
3	8					
4	3					
5	17					
6	28					
7	6					
8	7					
9	10					
10	18					
11	9					
12	5					
13						
14						
15						
16						
17						
18						
19						

2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys* у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *Descriptive statistics*, као што је показано у вежби 3 приликом израчунавања показатеља облика расподеле.

У оквиру прозора *Descriptive statistics* потребно је обележити податке који су предмет анализе (ставка *Input/Input Range*). У делу који се односи на *Output options* потребно је штиклирати *Confidence Level for Mean*, док се са десне стране може одредити жељени интервал. Како је у посматраном примеру циљ оценити 95% интервал поверења, уписана је вредност 95. Претходно наведено је илустровано у наставку:



3. Кликом на дугме *ОК*, отвара се следећи прозор:



Добијена вредност 4,8908 представља маргиналну грешку, односно производ критичне вредности из таблице t-расподеле и оцењене стандардне грешке аритметичке средине: $t_{n-1;\alpha} \times S_{\bar{X}} = 2{,}228 \times 2{,}1950 = 4{,}4809$. Други речима, добијену вредност је потребно одузети од аритметичке средине како би се добила доња граница интервала, односно сабрати са аритметичком средином како би се добила горња граница интервала:

$$11 - 4{,}4809 < \mu < 11 + 4{,}4809;$$

$$6{,}11 < \mu < 15{,}89.$$

Пример 12:

Маса плода једне сорте вишње приказана је у табели. Потребно је оценити 99% интервал поверења за аритметичку средину основног скупа.

Маса (g) X_i	**Број плодова** f_i	$X_i f_i$	$X_i^2 f_i$
4,78	5	23,90	114,24
4,81	12	57,72	277,63
4,82	15	72,30	348,49
4,86	18	87,48	425,15
4,87	7	34,09	166,02
4,90	3	14,70	72,03
Σ	**60**	**290,19**	**1403,56**

У циљу оцене непознате масе вишње у основном скупу на основу великог узорка ($n > 30$), у случају када није познат варијабилитет основног скупа потребно је дефинисати следећи 99% интервал поверења:

$$\bar{X} - Z_{0,01} \times S_{\bar{X}} < \mu < \bar{X} + Z_{0,01} \times S_{\bar{X}}\ .$$

$$\bar{X} = \frac{\sum X_i \times f_i}{n} = \frac{290{,}19}{60} = 4{,}84;$$

$$S_{\bar{X}} = \sqrt{\frac{\sum X_i^2 \, f_i - \frac{(\sum X_i f_i)^2}{n}}{n \times (n-1)}} = \sqrt{\frac{1.403{,}56 - \frac{290{,}19^2}{60}}{60 \times (60-1)}} = 0{,}0041;$$

$$Z_{0{,}01} = 2{,}58.$$

99% интервал поверења има следећи облик:

$$4{,}84 - 2{,}58 \times 0{,}0041 < \mu < 4{,}84 + 2{,}58 \times 0{,}0041;$$

$$4{,}83 < \mu < 4{,}85\,.$$

Уколико је жеља да се оцена интервала изврши у *Microsoft Excel*-у, неопходно је имати на уму да то није могуће учинити применом неких од *Microsoft Excel* опција. Ипак, могуће је формирати радну табелу као што је представљено у претходном примеру и извршити потребне прерачуне.

4.3.3. Интервал поверења за оцену непознате пропорције основног скупа

Пропорција је специфичан начин изражавања неке карактеристике (својства) у основном скупу или узорку и може се дефинисати као удео јединица посматрања са одређеном карактеристиком у укупном броју јединица посматрања (удео неисправних производа у укупном броју производа, број жена у укупном броју становника, удео болесних животиња у укупном броју животиња итд.). Вредност пропорције показује релативни удео посматране карактеристике у основном скупу. Пропорција основног скупа означава се са p и утврђује се као однос броја јединица које поседују жељену карактеристику (особину) и укупног броја јединица у основном скупу:

$$p = \frac{A}{N}\,,$$

где је:

A - број јединица посматрања основног скупа које поседују тражену карактеристику;

N - укупан број јединица посматрања у основном скупу.

Релативни удео јединица које не поседују неку карактеристику у основном скупу обележава се са q, а представља однос броја јединица основног скупа које не поседују тражену карактеристику (B) и укупног броја јединица основног скупа (N):

$$q = \frac{B}{N}\,.$$

У складу са претходно наведеним произилази да је $p + q = 1$, односно $q = 1 - p$.

С обзиром на то да је пропорција основног скупа најчешће непозната, оцењивање се изводи на основу узорка. Пропорција израчуната на основу узорка је оцена пропорције основног скупа. Пропорција оцењена из узорка означава се као $\hat{p}$, а израчунава се на следећи начин:

$$\hat{p} = \frac{a}{n},$$

тако да важи: $\hat{q} = 1 - \hat{p}$.

Оцењена стандардна грешка пропорције у случају простог случајног узорка без понављања једнака израчунава се на основу следећег израза:

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n} \times \frac{N-n}{N}}.$$

Код узорка са понављањем или у случају великог основног скупа где је $\frac{N-n}{N} \approx 1$ (непозната величина основног скупа) важи следеће:

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Ако је узорак довољно велики, може се на основу пропорције из узорка и њене стандардне грешке одредити интервал поверења у коме се очекује да ће се уз одређену вероватноћу налазити непозната пропорција основног скупа.

Уколико се оцена непознате пропорције основног скупа изводи на основу великог узорка (n>30) и ако важе неједнакости $n \times p > 5$ и $n(1-p) > 5$, интервал поверења има следећи облик:

$$\hat{p} - Z_\alpha \times S_{\hat{p}} < p < \hat{p} + Z_\alpha \times S_{\hat{p}}.$$

Облик интервала поверења следи из чињенице да под наведеним условима $\hat{p}$ има приближно нормалну расподелу са параметрима: $\mu_{\hat{p}} = n \times p$ и $\sigma_{\hat{p}}^2 = \frac{p \times (1-p)}{n}$. У случају малог узорка, $\hat{p}$ има биномну расподелу, тако да се наведени интервал поверења не може применити.

На основу утврђеног интервала поверења може се оценити и тотал основног скупа на основу следећег израза:

$$N \times L_1 < Np < N \times L_2.$$

Тотал основног скупа за пропорцију пружа информацију о броју јединица основног скупа које имају посматрану, односно тражену карактеристику.

Пример 13:

На основу података о промеру главе сунцокрета (цм) потребно је одредити границе 95% интервала поверења за прорпрцију глава сунцокрета које имају промер већи од 28 цм.

Промер главе X_i	**Број глава** f_i
18,1-20,0	14
20,1-22,0	24
22,1-24,0	28
24,1-26,0	48
26,1-28,0	35
28,1-30,0	32
30,1-32,0	19
Σ	**200**

Како би се оценила непозната пропорција основног скупа на основу података из узорка, потребно је дефинисати следећи 95% интервал поверења:

$$\hat{p} - Z_{0,05} \times S_{\hat{p}} < p < \hat{p} + Z_{0,05} \times S_{\hat{p}}.$$

Пропорција из узорка $\hat{p}$ може се израчунати стављајући у однос број глава сунцокрета која имају промер већи од 28 цм (*32+19=51*) и укупан број глава сунцокрета ($\sum f_i = 200$). Дакле оцена пропорције из узорка је:

$$\hat{p} = \frac{a}{n} = \frac{51}{200} = 0{,}2550.$$

Критична вредност из таблице нормалне расподеле износи $Z_{0,05} = 1{,}96$, док је оцена стандардне грешке пропорције:

$$S_{\hat{p}} = \sqrt{\frac{\hat{p} \times \hat{q}}{n}} = \sqrt{\frac{0{,}2550 \times (1 - 0{,}2550)}{200}} = 0{,}0308\ .$$

Тражени интервал поверења је следећи:

$$0{,}2550 - 1{,}96 \times 0{,}0308 < p < 0{,}2550 - 1{,}96 \times 0{,}0308;$$

$$0{,}19 < p < 0{,}32.$$

Дакле, оцена је се учешће глава са промером већим од 28 цм креће у интервалу од 19 до 32%,

Као и у досадашњем делу анализе, када је реч о груписаним подацима, потребна израчунавања је могуће спровести и формирањем радне табеле у *Microsoft Excel*-у.

Контролна питања

1. Шта је узорак?
2. Како се постиже репрезентативност узорка?
3. Навести неке планове узорака.
4. Како гласи централна гранична теорема.
5. Која својства је пожељно да поседује оценитељ из узорка?
6. Наведите две врсте статистичког оцењивања и њихове карактеристике.
7. Зашто се интервална оцена користи чешће од тачкасте?
8. Зашто је величина узорка значајна у статистичком оцењивању?
9. На основу којих елемената се оцењује непозната средина основног скупа на основу узорка?
10. На основу којих елемената се оцењује непозната пропорција основног скупа на основу узорка?

5. Тестирање статистичких хипотеза

Под хипотезом се подразумева научна претпоставка заснована на познатим чињеницама ради извођења неког закључка. Поступком тестирања у статистичком закључивању проверава се претпоставка о пробабилистичком моделу који генерише податке. С тим у вези, у параметарској статистици проверавају се претпоставке о вредностима параметара основног скупа. Користећи податке из узорка доноси се закључак да ли се полазна претпоставка прихвата или не.

Поступак или правило којим се доноси одлука о прихватању или неприхватању тврђења о вредностима параметара основног скупа на основу података из случајног узорка назива се *тестирање статистичких хипотеза*. Статистички тестови се деле на *параметарске* и *непараметарске*.

Параметарски тестови полазе од датог облика и карактеристика дистрибуције нумеричког обележја у основном скупу. За примену непараметарских тестова није потребно дати облик дистрибуције нумеричког обележја, а примењиви су и код квалитативних обележја.

Тестирање подразумева поступак провере одређене претпоставке коју зовемо *нулта хипотеза*. Нулта хипотеза је тврђење о неком параметру основног скупа које се сматра истинитим све док се не докаже супротно. *Алтернативна хипотеза* је тврђење о неком параметру основног скупа које ће бити истинито уколико је нулта хипотеза нетачна.

Приликом тестирања хипотеза треба узети у обзир да је поступак тестирања заснован на узорку, где бирање узорка подлеже правилима случајности. Претходно наведено подразумева да се може десити да на основу два изабрана узорка, закључци о истој тврдњи буду супротни. Приликом тестирања нулте хипотезе (H_0), против алтернативне хипотезе (H_1) могу настати две грешке.

Грешка типа I настаје када се одбаци истинита односно тачна нулта хипотеза. Вероватноћа јављања ове грешке представља ниво (праг) значајности и означава се са α. Вероватноћа (1-α) је вероватноћа не одбацивања нулте хипотезе када је она тачна и назива се ниво поверења теста

Грешка типа II се јавља када се неистинита (нетачна) нулта хипотеза прихвати. Вероватноћа јављања грешке типа II означава се са β. Вредност (1-β) назива се јачина (снага, моћ) теста и представља вероватноћу да се не јави грешка типа II. Табела 2 представљена у наставку даје преглед могућих грешака приликом статистичког закључивања.

***Табела 2**. Грешке I и II врсте приликом статистичког закључивања*

Одлука	**Стварно стање**	
	$\boldsymbol{H_0}$ **је тачна**	$\boldsymbol{H_0}$ **је погрешна**
$\boldsymbol{H_0}$ **се одбацује**	ПОГРЕШАН ЗАКЉУЧАК $P(H_1\|H_0) = \alpha$ Грешка I врсте	ТАЧАН ЗАКЉУЧАК $P(H_1\|H_1) = 1 - \beta$ *ЈАЧИНА ТЕСТА*
$\boldsymbol{H_0}$ **се прихвата**	ТАЧАН ЗАКЉУЧАК $P(H_0\|H_0) = 1 - \alpha$ ПОВЕРЕЊЕ	ПОГРЕШАН ЗАКЉУЧАК $P(H_0\|H_1) = \beta$ Грешка II врсте

Извор: Обрада аутора

Поступак статистичког тестирања састоји се из више етапа (фаза). Конкретно, поступак тестирања обухвата три фазе:

1. Формулисање полазне претпоставке – нулте хипотезе;
2. Поступак провере постављене хипотезе;
3. Закључак о постављеној хипотези .

У зависности од начина на који је формулисана алтернативна хипотеза у поступку тестирања могуће је применити три врсте теста: двострани, горњи једнострани и доњи једнострани тест.

Предмет статистичког тестирања могу бити различити параметри, а најчешће су то аритметичка средина и пропорција.

5.1. Тестови аритметичких средина

Приликом тестирања аритметичких средина могуће је издвојити три основна теста:

1. Упоређивање аритметичке средине узорка са аритметичком средином основног скупа или неком хипотетичком вредношћу – тест значајности једне средине;
2. Упоређивање две аритметичке средине из два независна узорка – тест значајности разлике две средине;
3. Упоређивање више од две средине из више од два узорка – метод анализе варијансе.

Сви параметарски тестови аритметичких средина који ће бити разматрани, засновани су на претпоставци да је дистрибуција нумеричког обележја у основном скупу нормална.

5.1.1. Тестирање нулте хипотезе о аритметичкој средини основног скупа

Тест значајности једне средине заснива се на тестирању нулте хипотезе о једнакости аритметичке средине узорка са хипотетичком вредношћу аритметичке средине основног скупа. Овај тест може се извести у случају када је познат варијабилитет основног скупа и у случају када се цео поступак тестирања заснива једино на резултатима узорка, односно када није познат варијабилитет основног скупа.

5.1.1.1. Тестирање нулте хипотезе о аритметичкој средини основног скупа – познат варијабилитет основног скупа

Као што је наведено у претходном делу, у првој фази тестирања неопходно је формулисати две супротстављене хипотезе, нулту и алтернативну. Код теста значајности једне средине полази се од претпоставке (нулта хипотеза) да је непозната аритметичка средина основног скупа μ, једнака претпостављеној (или хипотетичкој вредности), μ_0. Алтернативна хипотеза насупрот нултој, претпоставља да непозната аритметичка средина основног скупа није једнака претпостављеној вредности μ_0. С тим у вези, у случају двостраног теста, нулта и алтернативна хипотеза гласе: $H_o: \mu = \mu_0$ и $H_1: \mu \neq \mu_0$ респективно. Код једностраног теста, нулта хипотеза се може односити на горњу или доњу границу непознате аритметичке средине. Нулте и алтернативне хипотезе које се односе на горњу и доњу границу непознате аритметичке средине респективно су: $H_o: \mu \geq \mu_0$; $H_1: \mu < \mu_0$ и $H_o: \mu \leq \mu_0$; $H_1: \mu > \mu_0$.

Провера постављене нулте хипотезе изводи се израчунавањем одговарајуће тест статистике:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}},$$

где се права стандардна грешка аритметичке средине израчунава на основу следеће формуле:

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n} \times \frac{N-n}{N-1}}\,.$$

Корективни фактор $\frac{N-n}{N-1}$ фигурише у формули за обрачун стандардне грешке аритметичке средине једино уколико је примењен прост случајан узорак без понављања и позната величина основног скупа.

Претпостављајући нулту хипотезу, *Z*-статистика има стандардну нормалну расподелу *N(0,1)*. Приликом доношења закључка о прихватању или одбацивању нулте хипотезе, код теста значајности једне средине када је познат варијабилитет основног скупа, користе се таблице нормалне дистрибуције без обзира на величину узорка на основу ког се тестирање изводи. Закључак се најчешће доноси на прагу значајности α*=0,05* или

α=0,01. Код прага значајности 0,05 критична вредност у таблици нормалне дистрибуције износи 1,96, док за праг значајности 0,01 критична вредност износи 2,58. У складу са претходно наведеним, уколико је апсолутна вредност израчунатог количника *Z*, једнака или већа од наведених критичних вредности има основа за одбацивање нулте хипотезе као нетачне и прихватање алтернативне хипотезе. Насупрот томе, ако је апсолутна вредност израчунатог количника Z мања од критичних вредности, нулта хипотеза се може прихватити.

Овде је битно истаћи, да се прихватањем нулте хипотезе, не може закључити да је нулта хипотеза тачна, већ да постојећи докази против нулте хипотезе нису довољно јаки. Формулација прихвата се нулта хипотеза значи да резултати узорка подржавају нулту хипотезу и да се она не може одбацити.

Нулта хипотеза код теста значајности једне средине у случају када је познат варијабилитет основног скупа, може се проверити и на основу интервала поверења:

$$\bar{X} - Z_\alpha \times \sigma_{\bar{X}} < \mu < \bar{X} + Z_\alpha \times \sigma_{\bar{X}}.$$

На овај начин се утврђује интервал у оквиру којег се налази очекивана вредност аритметичке средине основног скупа. Уколико се хипотетичка вредност параметра налази унутар граница утврђеног интервала поверења, нулта хипотеза се може прихватити. Супротно, уколико је хипотетичка вредност параметра изван граница утврђеног интервала, нулта хипотеза се одбацује.

Пример 14:

На основу података озорка од 100 хектара утврђен је просечан принос грашка од 3,45 тона по хектару. Од раније је позната варијанса основног скупа која износи 1,46. Да ли се може усвојити претпоставка да је просечан принос грашка по хектару 3,69 тона по хектару?

Решење

Имајући у виду да је на располагању један узорак ($n = 100$), где је потребно упоредити аритметичку средину из узорка ($\bar{X} = 3{,}45$) са претпостављеном вредношћу ($\mu_0 = 3{,}69$), јасно је да се ради о тесту значајности једне средине. Како је познат варијабилитет основног скупа ($\sigma^2 = 1{,}46$), неопходно је и тај податак узети у обзир.

1) $H_o: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

Уколико узорак припада основном скупу са претпостављеном средином $\mu_0 = 3{,}69$, аритметичка средина узорка која је пре извлачења узорка случајна променљива, не одступа значајно од претпостављене вредности. То је разлог што се нулта и алтернативна хипотеза могу написати и на следећи начин:

$H_o: \bar{X} = \mu_0$
$H_1: \bar{X} \neq \mu_0$

У првом запису једнакост код нулте хипотезе је идентитет, док је у другом случају означава да су вредности приближно једнаке, њихова разлика је резултат случајне варијације.

2)

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{3{,}45 - 3{,}69}{0{,}1208} = -1{,}99$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{1{,}46}{100}} = 0{,}1208$$

3) Критичне вредности из таблице нормалне расподеле за праг значајности $\alpha = 0{,}05$ и $\alpha = 0{,}01$ су 1,96 и 2,58 респективно. Како је апсолутна вредност Z-количника 1,99, закључак је да се нулта хипотеза одбацује за праг значајности $\alpha = 0{,}05$, док се исти закључак не може извести за праг значајности $\alpha = 0{,}01$.

$$|Z| > 1{,}96 \rightarrow H_1$$

$$|Z| < 2{,}58 \rightarrow H_0$$

Други начин провере полазне претпоставке подразумева претходно дефинисање одговарајућехг интервала поверења. У складу са расположивим подацима, интервал поверења који је неопходно дефинисати је следећи:

$$\bar{X} - Z_\alpha \times \sigma_{\bar{X}} < \mu < \bar{X} + Z_\alpha \times \sigma_{\bar{X}}.$$

95% интервал поверења је следећи:

$$3{,}45 - 1{,}96 \times 0{,}1208 < \mu < 3{,}45 + 1{,}96 \times 0{,}1208$$

$$3{,}21 < \mu < 3{,}69$$

$$\mu_0 \notin (L_1, L_2) \rightarrow H_1;$$

Како вредност за претпостављену аритметичку средину ($\mu_0 = 3{,}69$) не припада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна.

99% интервал поверења је следећи:

$$3{,}45 - 2{,}58 \times 0{,}1208 < \mu < 3{,}45 + 2{,}58 \times 0{,}1208$$

$$3{,}14 < \mu < 3{,}76$$

$$\mu_0 \in (L_1, L_2) \rightarrow H_0;$$

С обзиром на то да претпостављена аритметичка средина ($\mu_0 = 3{,}69$) припада дефинисаном интервалу, нулта хипотеза се прихвата.

5.1.1.2. Тестирање нулте хипотезе о аритметичкој средини основног скупа – непознат варијабилитет основног скупа

Уколико није позната вредност стандардне девијације или варијанса основног скупа, полазна хипотеза се проверава израчунавањем *t*-количника, с тим да формулација полазне претпоставке остаје непромењена ($H_o: \mu = \mu_0$ и $H_1: \mu \neq \mu_0$).

Количник *t* се израчунава на следећи начин:

$$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}},$$

где је $S_{\bar{X}}$ оцењена стандардна грешка аритметичке средине која се може израчунати на више начина. У зависности од тога да ли је изабрани узорак без или са понављањем, оцењена стандарна грешка аритметичке средине се изводи на основу следеће две формуле респективно:

$$S_{\bar{X}} = \sqrt{\frac{S^2}{n} \times \frac{N-n}{N}} \quad \text{и} \quad S_{\bar{X}} = \sqrt{\frac{S^2}{n}},$$

где је S^2 оцењена варијанса добијена на основу узорка.

У складу са претходно наведеним, оцењена стандардна грешка аритметичке средине $S_{\bar{X}}$ може се израчунати на основу следећих формула за негруписане податке:

$$S_{\bar{X}} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n \times (n-1)} \times \frac{N-n}{N}} \quad \text{или} \quad S_{\bar{X}} = \sqrt{\frac{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}{n \times (n-1)} \times \frac{N-n}{N}}.$$

У случају да су подаци из узорка груписани, у употреби су следеће формуле:

$$S_{\bar{X}} = \sqrt{\frac{\sum f_i(X_i - \bar{X})^2}{n \times (n-1)} \times \frac{N-n}{N}} \quad \text{или} \quad S_{\bar{X}} = \sqrt{\frac{\sum f_i {X_i}^2 - \frac{(\sum f_i X_i)^2}{n}}{n \times (n-1)} \times \frac{N-n}{N}}.$$

Без обзира на то да ли се ради о груписаним или негруписаним подацима, корективни фактор $\frac{N-n}{N}$ ће изостати у наведеном формулама уколико није позната величина основног скупа N.

Претпостављајући нулту хипотезу *t*-количник има Студентову *t*-расподелу са $n-1$ бројем степени слободе.

Апсолутна вредност израчунатог количника *t* упоређује се са критичном вредношћу из *t*-дистрибуције. Критична вредност $t_{n-1;\alpha}$ задовољава услов да је $P(|t| > t_{n-1;\alpha}) = \alpha$ и добија се коришћењем таблица *t*-дистрибуције. Уколико се провера хипотезе изводи на основу великог узорка (*n>30*), израчунати количник *t* има приближно стандардну нормалну расподелу, тако да се могу користити и таблице *N(0,1)* дистрибуције. Уколико је апсолутна вредност израчунатог количника *t* једнака или већа од критичне вредности

$t_{n-1;\alpha}$, постоји основа за одбацивање нулте хипотезе и прихватање алтернативне хипотезе. Насупрот томе, ако је апсолутна вредност израчунатог количника t мања од критичне вредности, нулта хипотеза се може прихватити.

Тестирање хипотезе о значајности једне средине у случају непознатих параметара основног скупа, може да извести и израчунавањем одговарајућег интервала поверења:

$$\bar{X} - t_{n-1;\alpha} \times S_{\bar{X}} < \mu < \bar{X} + t_{n-1;\alpha} \times S_{\bar{X}} ,$$

који се у случају великог узорка ($n>30$) може апроксимирати са:

$$\bar{X} - Z_{\alpha} \times S_{\bar{X}} < \mu < \bar{X} + Z_{\alpha} \times S_{\bar{X}}.$$

Уколико хипотетичка вредност μ_0 припада $(1-\alpha)\times 100\%$ интервалу поверења, прихвата се нулта хипотеза на прагу значајности α. У случају да не припада, прихвата се алтернативна хипотеза. Претходно наведено, може се записати на следећи начин:

$$\mu_0 \in (L_1, L_2) \rightarrow H_0;$$

$$\mu_0 \notin (L_1, L_2) \rightarrow H_1.$$

Поред до сада разматраних приступа доношења одлуке на основу поређења тест критеријума са критичном вредношћу и на основу интервала поверења, примењује се и приступ заснован на вредности вероватноће ($p-$вредности) који је заступљен у статистичком софтверу. Предпостављајући да је нулта хипотеза тачна, $p-$вредност се дефинише као вероватноћа да тест критеријум одступа у смеру алтернативне хипотезе бар толико колико реализована вредност статистике узорка. Нпр. Ако је $t_{\text{изр}}$ израчуната вредност тест критеријума, $p = P(|t| > t_{\text{изр}})$ код двостраног теста, односно $p = P(t > t_{\text{изр}})$ или $p = P(t < t_{\text{изр}})$ код једностраног теста. Код овог приступа, нулту хипотезу прихватамо ако је $p \geq \alpha$, одбацујемо уколико је $p < \alpha$, где је α одабрани праг значајности.

Пример 15:

Подаци из узорка од 2.400 регистрованих пољопривредних газдинстава односе се на површину под воћњацима и виноградима у Војводини:

Површина (ха) X_i	**Број газдинстава** f_i	$X_i f_i$	$X_i^2 f_i$
0,6	140	84	50,40
0,8	240	192	153,60
1,0	500	500	500,00
1,2	700	840	.008,00
1,4	160	224	313,60
1,6	455	728	1.164,80
1,8	110	198	356,40
2,0	95	190	380,00
Σ	**2.400**	**2.956**	**3.926,80**

Уколико је познато да је укупан број регистрованих пољопривреднх газдинстава у оквиру којих се налазе воћњаци и виногради 20.000, да ли се може прихватити нулта хипотеза да је просечна површина газдинстава 1,86 ха?

1) У посматраном примеру на располагању је један велики узорак (*n=2.400*) у оквиру осноног скупа чија је величина такође позната (*N=20.000*). Како је непознат варијабилитет основног скупа, поступак провере нулте хипотезе у склопу теста значајности једне средине је следећи:

1) $H_o: \mu = \mu_0$

$H_1: \mu \neq \mu_0$

2)

$$Z = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \frac{1{,}23 - 1{,}86}{0{,}0066} = -95{,}29$$

$$\bar{X} = \frac{\sum X_i f_i}{n} = \frac{2.956}{2400} = 1{,}23\text{ха};$$

$$S_{\bar{X}} = \sqrt{\frac{\sum f_i {X_i}^2 - \frac{\left(\sum f_i X_i\right)^2}{n}}{n \times (n-1)} \times \frac{N-n}{N}} = \sqrt{\frac{3.926{,}8 - \frac{2.956^2}{2.400}}{2.400 \times (2.400-1)} \times \frac{20.000 - 2.400}{20.000}}$$

$$= 0{,}0066\text{ ха}$$

3) Критичне вредности из таблице нормалне расподеле за праг значајности $\alpha = 0{,}05$ и $\alpha = 0{,}01$ су 1,96 и 2,58 респективно. Како је апсолутна вредност Z-количника 95,29, закључак је да се нулта хипотеза одбацује за оба прага значајности ($\alpha = 0{,}01$ и $\alpha = 0{,}05$) и прихвата алтернативна хипотеза.

$$|Z| > 1{,}96 \rightarrow H_1$$

$$|Z| > 2{,}58 \rightarrow H_1$$

Други начин провере полазне претпоставке подразумева претходно дефинисање одговарајућехг интервала поверења. У складу са расположивим подацима, интервал поверења који је неопходно дефинисати је следећи:

$$\bar{X} - Z_\alpha \times S_{\bar{X}} < \mu < \bar{X} + Z_\alpha \times S_{\bar{X}}.$$

95% интервал поверења је следећи:

$$1{,}23 - 1{,}96 \times 0{,}0066 < \mu < 1{,}23 + 1{,}96 \times 0{,}0066$$

$$1{,}22 < \mu < 1{,}24\text{ ха}$$

$$\mu_0 \notin (L_1, L_2) \rightarrow H_1$$

Како вредност за претпостављену аритметичку средину ($\mu_0 = 1{,}86$) не припада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна на прагу значајности 5%.

99% интервал поверења је следећи:

$$1{,}23 - 2{,}58 \times 0{,}0066 < \mu < 1{,}23 + 2{,}58 \times 0{,}0066$$

$$1{,}22 < \mu < 1{,}25 \text{ ха}$$

$$\mu_0 \notin (L_1, L_2) \rightarrow H_1$$

Како вредност за претпостављену аритметичку средину ($\mu_0 = 1{,}86$ ха) не припада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна и на прагу значајности 1%.

Када је реч о *Microsoft Excel-y,* тестирање је могуће спровести на основу интервала поверења који је могуће израчунати као што је урађено у вежби 4 у случају негруписаних података. Уколико су подаци груписани, потребно је формирати радну табелу како би се добиле одговарајуће суме које ће касније бити употребљене приликом израчунавања Z или t количника. Израчунате количнике је затим потребно упоредити са одговарајућим критичним вредностима како би се добио одговарајући закључак.

5.1.2. Тестирање нулте хипотезе о једнакости аритметичких средина два основна скупа

У практичном раду често се изводи експеримент са два третмана. При томе проверава се да ли су просечне вредности основних скупова једнаке, односно да ли је $\mu_1 = \mu_2$, где је μ_1 аритметичка средина једног основног скупа, док је μ_2 аритметичка средина другог основног скупа. Провера једнакости средина основних скупова изводи се на основу два случајна узорка. Овај тест назива се тест значајности разлике две средине, а заснива се на упоређивању две аритметичке средине из два узорка који могу бити независни или зависни.

У наставку ће бити представљени тестови када су посматрани узорци независни, док се тестирање на основу зависних узорака неће детаљније анализирати. Код теста значајности разлике две средине полазна хипотеза гласи да су аритметичке средине основних скупова међусобно једнаке, односно: $H_o: \mu_1 = \mu_2$. Поред наведеног, у оквиру нулте хипотезе може се претпоставити да су аритметичке средине међусобно независних узорака једнаке, односно: $H_o: \bar{X}_1 = \bar{X}_2$. Код овог записа једнакост подразумева да је разлика аритметичких средина резултат случајне варијације настале при избору узорака из истог основног скупа. На тај начин, нултом хипотезом се претпоставља да у утицају два испитивана третмана на експерименталне јединице узорака нема статистички значајне разлике.

Супротна претпоставка, односно алтернативна хипотеза код овог теста гласи $H_1: \mu_1 \neq \mu_2$ или како се још може записати $H_1: \bar{X}_1 \neq \bar{X}_2$. Алтернативна претпоставка се заснива на тврдњи да је утицај испитиваних третмана статистички значајно различит.

Као и код теста једне средине, код теста значајности разлике две средине разликују се случајеви када су познати варијабилитети основних скупова и када нису познати. Разлика је у томе што сада фигурирају два основна скупа и два независна узорка која произилазе из основних скупова. Без обзира на то да ли су познати варијабилитети основних скупова или нису, нулта и алтернативна хипотеза ће гласити исто.

5.1.2.1. Тест значајности разлике две средине – познат варијабилитет основних скупова

Код првог случаја, када су познате варијансе или стандардне девијације основних скупова, нулта хипотеза се проверава израчунавањем количника *Z*. Како би се спровео тест, потребно је најпре израчунати стандардну грешку разлике аритметичких средина на основу следећег израза:

$$\sigma_{(\bar{X}_1-\bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

где су σ_1^2 и σ_2^2 варијансе основних скупова, док су n_1 и n_2 величине независних узорака.

На основу аритметичких средина утврђених на основу узорака и израчунате стандардне грешке, добија се вредност количника *Z*:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{(\bar{X}_1-\bar{X}_2)}}.$$

У циљу доношења закључка у вези са постављеном хипотезом, израчуната вредност количника се упоређује са одговарајућим критичним вредностима из таблице нормалне дистрибуције, најчешће за прагове значајности 5% и 1%. Уколико је апсолутна вредност израчунатог количника Z једнака или већа од критичних вредности, постоји основа за одбацивање нулте хипотезе и прихватање алтернативне хипотезе. У случају да је израчунати Z количник мањи од критичних вредности, нулта хипотеза се одбацује и прихвата се алтернативна.

Провера хипотезе о значајности разлике две средине може се извести и израчунавањем одговарајућег $(1-\alpha) \times 100\%$ интервала поверења, следећег облика:

$$(\bar{X}_1 - \bar{X}_2) - Z_\alpha \times \sigma_{(\bar{X}_1-\bar{X}_2)} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_\alpha \times \sigma_{(\bar{X}_1-\bar{X}_2)}.$$

Нулта хипотеза $\mathrm{H}_0: \mu_1 = \mu_2$, еквиваентна је тврђењу $\mathrm{H}_0: \mu_1 - \mu_2 = 0$. Самим тим, уколико границе израчунатог интервала укључују нулу има основа за прихватање нулте

хипотезе. У супротном, ако нула не припада границама утврђеног интервала, одбацује се нулта и прихвата алтермнативна хипотеза.

Пример 16:

Циљ експеримента је да се утврди да ли постоји статистички значајна разлика у просечном приносу две сорте пшенице. Изабрана су два независна случајна узорка. Приноси две различите сорте пшенице представљени су у табели. Ако су од раније познати варијабилитети посматраних сорти $\sigma_1^2 = 1$ и $\sigma_1^2 = 0{,}75$, тестирати нулту хипотезу да не постоји статистички значајна разлика у просечном приносу посматраних сорти.

Сорта 1	6,3	6,6	7,0	5,9	5,5	5,7	5,8	5,6	6,0	6,1	6,3	6,2	6,2	6,4	6,9
Сорта 2	5,5	5,4	5,1	5,7	5,9	6,2	6,3	5,2	5,7	5,6	-	-	-	-	-

1) С обзиром на то да је потребно тестирати једнакост две средине (два просечна приноса) из два независна узорка који се односе на сорту пшенице 1 ($n_1 = 15$) и сорту пшенице 2 ($n_2 = 10$), нулта хипотеза је следећа:

$H_o: \bar{X}_1 = \bar{X}_2$;

$H_1: \bar{X}_1 \neq \bar{X}_2$.

2) У другом кораку, потребно је формирати Z тест. Пре тога потребно је израчунати аритметичке средине из узорака и стандардну грешку разлике две средине имајући у виду да су познати варијабилитети основних скупова.

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{(\bar{X}_1 - \bar{X}_2)}} = \frac{6{,}17 - 5{,}66}{0{,}3764} = 1{,}3461$$

$$\bar{X}_1 = \frac{\sum X_{1i}}{n_1} = \frac{92{,}5}{15} = 6{,}17 \text{ т/ха};$$

$$\bar{X}_2 = \frac{\sum X_{2i}}{n_2} = \frac{56{,}6}{10} = 5{,}66 \text{ т/ха}.$$

$$\sigma_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1}{15} + \frac{0{,}75}{10}} = 0{,}3764 \text{ т/ха}.$$

3) У трећем кораку се доноси закључак. Апсолутна вредност Z-количника се пореди са критичним вредностима из таблице нормалне расподеле ($Z_{0{,}05} = 1{,}96$ и $Z_{0{,}01} = 2{,}58$).

$$|Z| < 1{,}96 \rightarrow H_0$$

$$|Z| < 2{,}58 \rightarrow H_0$$

Како је $|Z|$ мање од обе критичне вредности на прагу значајности *α=0,05* и *α=0,01*, нулта хипотеза се прихвата. Закључак је да нема статистички значајне разлике у приносима две сорте пшенице.

Полазна претпоставка се може проверити и на основу следећег интервала поверења:

$$(\bar{X}_1 - \bar{X}_2) - Z_\alpha \times \sigma_{(\bar{X}_1 - \bar{X}_2)} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_\alpha \times \sigma_{(\bar{X}_1 - \bar{X}_2)}.$$

95% интервал поверења је следећи:

$$(6{,}17 - 5{,}66) - 1{,}96 \times 0{,}3764 < \mu_1 - \mu_2 < (6{,}17 - 5{,}66) + 1{,}96 \times 0{,}3764.$$

$$-0{,}23 < \mu_1 - \mu_2 < 1{,}24 \text{ т/ха}$$

$$0 \in (L_1, L_2) \rightarrow H_0$$

Како вредност 0 припада дефинисаном интервалу, нулта хипотеза се прихвата.

99% интервал поверења је следећи:

$$(6{,}17 - 5{,}66) - 2{,}58 \times 0{,}3764 < \mu_1 - \mu_2 < (6{,}17 - 5{,}66) + 2{,}58 \times 0{,}3764.$$

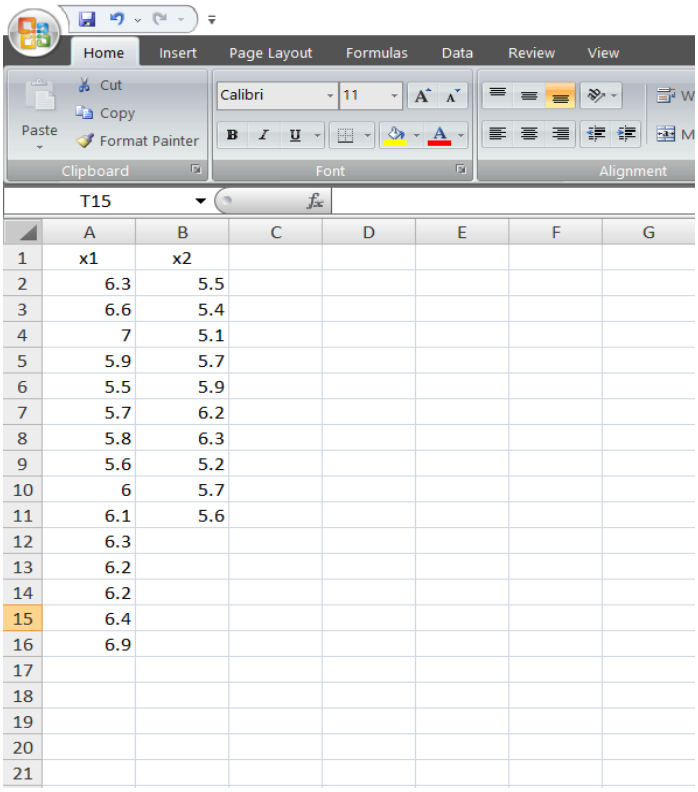
$$-0{,}46 < \mu_1 - \mu_2 < 1{,}48 \text{ т/ха}$$

$$0 \in (L_1, L_2) \rightarrow H_0$$

Како опет вредност 0 припада дефинисаном интервалу, нулта хипотеза се прихвата и за праг значајности α=0,01.

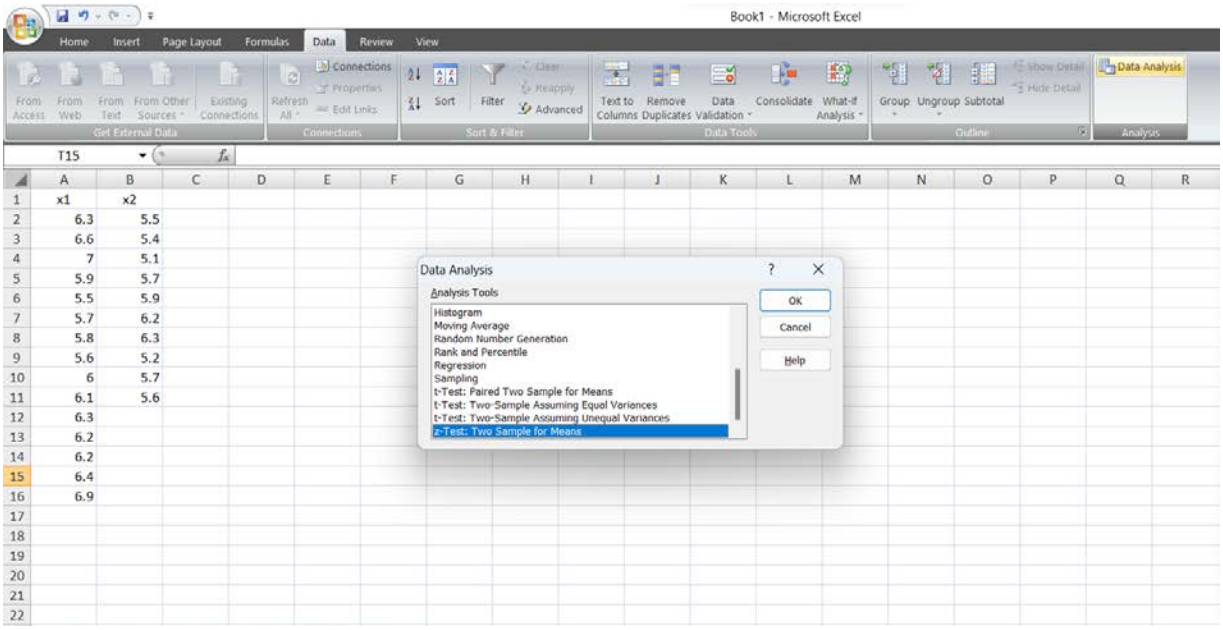
Вежба7. Тест значајности разлике две средине (познати варијабилитети основних скупова) применом Microsoft Excel-a

1. Полазни изглед табеле је следећи:

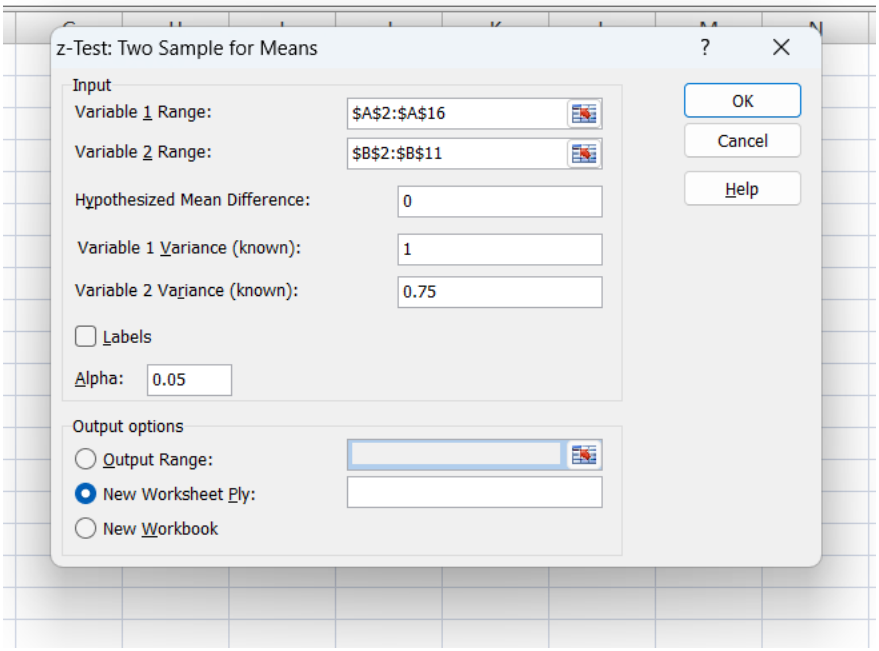


	A	B	C	D	E	F	G
1	x1	x2					
2	6.3	5.5					
3	6.6	5.4					
4	7	5.1					
5	5.9	5.7					
6	5.5	5.9					
7	5.7	6.2					
8	5.8	6.3					
9	5.6	5.2					
10	6	5.7					
11	6.1	5.6					
12	6.3						
13	6.2						
14	6.2						
15	6.4						
16	6.9						
17							
18							
19							
20							
21							

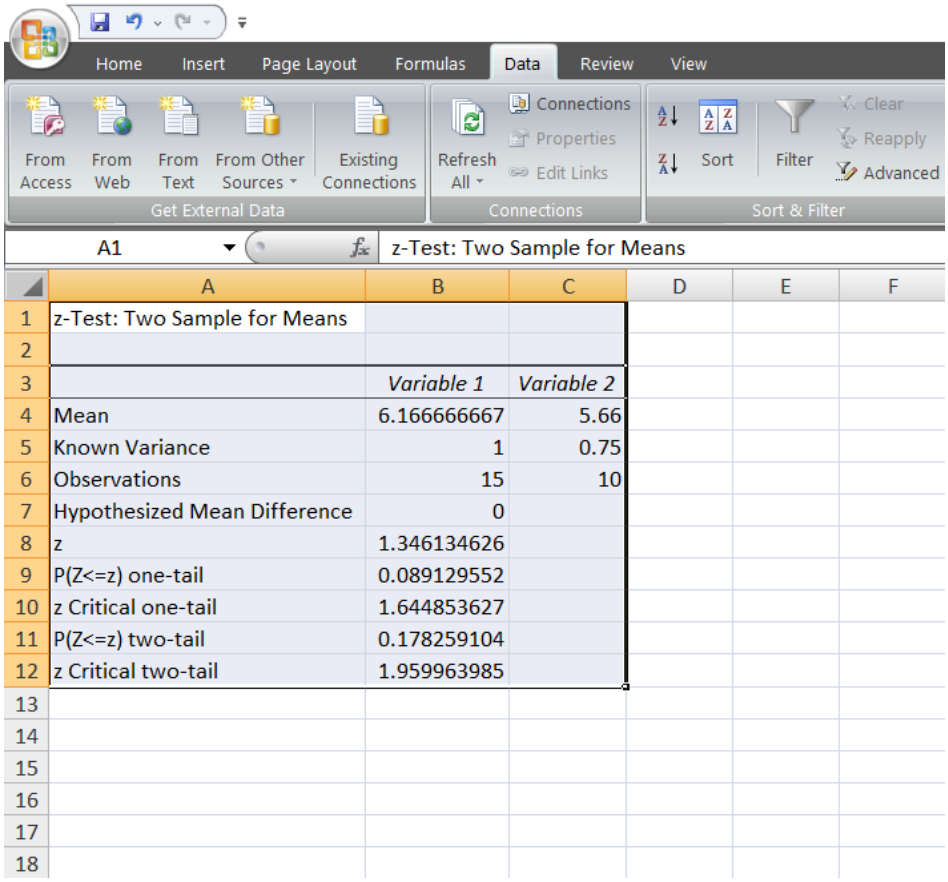
2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys* у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *z-Test:Two Sample for Means*, као што је показано у наставку:



3. Кликом на дугме *OK* отвара се нови прозор. У блоку који се односи на *Input*, у делу који се односи на *Variable 1 Range* и *Variable 2 Range* потребно је обележити податке који припадају првом и другом узорку респективно. У делу у којем пише *Hypothesized Mean Difference* потребно је уписати вредност 0, јер ако је нултом хипотезом претпостављено да нема разлике између посматраних средина, другим речима њихова разлика је једнака 0. С обзиром на то да су познати варијабилитети основних скупова, познате варијабилитете је потребно уписати у делу који се односи на *Variable 1 Variance (known)* и *Variable 2 Variance (known)*. На крају у делу који се односи на *Alpha* неопходно је уписати жељени праг значајности α. У примеру биће изабран праг значајности $\alpha=0,05$. Поглед на *Microsoft Excel* сада изгледа на следећи начин:



4. Кликом на дугме *OK* добијају се резултати теста у новом *Sheet*-у:



	A	B	C	D	E	F
1	z-Test: Two Sample for Means					
2						
3		*Variable 1*	*Variable 2*			
4	Mean	6.166666667	5.66			
5	Known Variance	1	0.75			
6	Observations	15	10			
7	Hypothesized Mean Difference	0				
8	z	1.346134626				
9	P(Z<=z) one-tail	0.089129552				
10	z Critical one-tail	1.644853627				
11	P(Z<=z) two-tail	0.178259104				
12	z Critical two-tail	1.959963985				
13						
14						
15						
16						
17						
18						

У прва четири реда наведени су подаци који се везују за природу теста (тестиране аритметичке средине, познате варијансе, величина узорака, нулта хипотеза). Затим је представљена вредност *Z*-количника који износи 1,3461. У наставку се пружа могућност

поређења добијене вредности количника са критичним вредностима из таблице нормалне расподеле или са одговарајућом *p*-вредношћу. Посебну пажњу треба посветити изразима *P(Z<=z) two-tail* и *z Critical two-tail*. Вредност *P(Z<=z) two-tail* представља добијену *p*-вредност. Као што је било речи, уколико је добијена *p*-вредност већа од 0,05 нулта хипотеза се прихвата. С друге стране, *z Critical two-tail* представља критичну вредност из таблице нормалне расподеле за праг значајности који смо навели дефинишући полазне параметре теста. У овом случају критична вредност је 1,96, јер смо код *Alpha* у претходном кораку уписали 0,05. Како је *Z*-количник мањи од критичне вредности из таблице нормалне расподеле и како је *p*-вредност већа од 0,05, закључак је да се нулта хипотеза прихвата.

5.1.2.2. Тест значајности разлике две средине – непознат варијабилитет основних скупова

Код теста значајности разлике две средине, након уобичајеног првог корака који се односи на дефинисања нулте хипотезе, потребно је у другом кораку израчунати одговарајућу тест статистику. У практичном раду, тест критеријум има следећи облик:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{(\bar{X}_1 - \bar{X}_2)}}.$$

Наведени количник t заснива се на томе да су непознате варијансе основних скупова, замењене њиховим оценама на основу узорака. Претпоставља се да су непознате варијансе основних скупова једнаке (хомогене): $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Утврђена вредност количника t упоређује се са критичним вредностима t-дистрибуције за $[(n_1 - 1) + (n_2 - 1)]$ број степени слободе.

Израчунавање оцењене стандардне грешке разлике две средине $S_{(\bar{X}_1 - \bar{X}_2)}$ условљено је величином узорака на основу којих се оцена изводи. Уколико су узорци неједнаких величина, где важи $n_1 \neq n_2$, израчунавање си изводи на основу следеће формуле:

$$S_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{S_{1+2}^2 \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

С друге стране, уколико су узорци једнаке величине ($n_1 = n_2 = n$), користи се следећи израз:

$$S_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{2 \times S_{1+2}^2}{n}}.$$

У оба случаја да би се добила вредност оцењене стандардне грешке две средине, потребно је претходно израчунати здружену варијансу S_{1+2}^2, која је оцена непознате варијансе σ^2. Здружена варијанса се може израчунати на два начина:

- на бази одступања вредности обележја од просека;

- директно из података.

У складу са претходно наведеним, здружена варијанса, уколико су подаци у узорцима негруписани, може се утврдити на следеће начине:

$$S_{1+2}^2 = \frac{\sum(X_{1i}-\bar{X}_1)^2+\sum(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2} \quad \text{или} \quad S_{1+2}^2 = \frac{\sum X_{1i}^2-\frac{(\sum X_{1i})^2}{n_1}+\sum X_{2i}^2-\frac{(\sum X_{2i})^2}{n_2}}{n_1+n_2-2}.$$

Уколико су подаци у узорцима дати као дистрибуција фреквенција, здружена варијанса се израчунава помоћу следећих израза:

$$S_{1+2}^2 = \frac{\sum f_1(X_{1i}-\bar{X}_1)^2+\sum f_2(X_{2i}-\bar{X}_2)^2}{n_1+n_2-2} \quad \text{или} \quad S_{1+2}^2 = \frac{\sum f_1X_{1i}^2-\frac{(\sum f_1X_{1i})^2}{n_1}+\sum f_2X_{2i}^2-\frac{(\sum f_2X_{2i})^2}{n_2}}{n_1+n_2-2}.$$

Код груписаних података у узорцима потребно је имати у виду да величина узорака представља суму фреквенција, односно: $n_1 = \sum f_1$ и $n_2 = \sum f_2$.

Здружена варијанса такође може да се изрази као пондерисана аритметичка средина оцена варијанси основних скупова на основу узорака:

$$S_{1+2}^2 = \frac{(n_1-1)\times S_1^2 + (n_2-1)\times S_2^2}{n_1+n_2-2}.$$

Апсолутна израчуната вредност количника *t* упоређује се са критичним вредностима из таблица Студентове *t*-дистрибуције за $[(n_1-1) + (n_2-1)]$ број степени слободе и различите прагове значајности, обично 5% и 1%. Уколико је апсолутна вредност израчунатог количника мања од одговарајућих критичних вредности има основа за прихватање нулте хипотезе. У супротном, вредност количника већа или једнака од одговарајућих критичних вредности резултира прихватањем алтернативне хипотезе. У циљу доношења закључка о постављеној хипотези, вредност количника *t* може се упоредити и са критичним вредностима из таблица нормалне дистрибуције ако важи да је $n_1+n_2-2 > 30$.

Тестирање постављене нулте хипотезе и у овом случају могуће је извести израчунавањем одговарајућих интервала поверења. Уколико је n_1+n_2-2 мање од 30, потребно је применити следећи интервал поверења:

$$(\bar{X}_1-\bar{X}_2) - t_{n_1+n_2-2;\alpha}\times S_{(\bar{X}_1-\bar{X}_2)} < \mu_1-\mu_2 < (\bar{X}_1-\bar{X}_2) + t_{n_1+n_2-2;\alpha}\times S_{(\bar{X}_1-\bar{X}_2)}.$$

С друге стране, уколико је степен слободе $t_{n_1+n_2-2;\alpha}$ већи од 30, важи да је $t_{n_1+n_2-2;\alpha} \approx N(0,1)$, тако да је потребно применити следећи интервал:

$$(\bar{X}_1-\bar{X}_2) - Z_\alpha\times S_{(\bar{X}_1-\bar{X}_2)} < \mu_1-\mu_2 < (\bar{X}_1-\bar{X}_2) + Z_\alpha\times S_{(\bar{X}_1-\bar{X}_2)}.$$

Основ за прихватање нулте хипотезе је да нула припада границама израчунатог интервала.

Пример 17:

При испитивању две врсте хране у исхрани јунади постављен је оглед са две групе грла, чији су резултати дати у табели. Утврдити да ли различита исхрана утиче на просечан дневни прираст грла (кг).

Прираст расе А X_1	***Прираст расе Б*** X_2	X_1^2	X_2^2
1,40	1,31	1,9600	1,7161
1,38	1,37	1,9044	1,8769
1,41	1,39	1,9881	1,9321
1,35	1,33	1,8225	1,7689
1,42	1,35	2,0164	1,8225
1,37	1,36	1,8769	1,8496
1,38	1,33	1,9044	1,7689
1,40	1,33	1,9600	1,7689
1,39	1,35	1,9321	1,8225
1,37	1,38	1,8769	1,9044
13,87	**13,50**	**19,2417**	**18,2308**

1)

$$H_o: \bar{X}_1 = \bar{X}_2;$$

$$H_1: \bar{X}_1 \neq \bar{X}_2.$$

2)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{(\bar{X}_1 - \bar{X}_2)}} = \frac{1{,}387 - 1{,}350}{0{,}0005} = 3{,}5439$$

$$\bar{X}_1 = \frac{\sum X_{1i}}{n_1} = \frac{13{,}87}{10} = 1{,}387 \text{ кг};$$

$$\bar{X}_2 = \frac{\sum X_{2i}}{n_2} = \frac{13{,}50}{10} = 1{,}350 \text{ кг}.$$

Собзиром на то да су узорци једнаке величине ($n_1 = n_2 = n$), користи се следећи израз приликом израчунавања оцењене страндардне грешке разлике две средине:

$$S_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{2 \times S_{1+2}^2}{n}} = \sqrt{\frac{2 \times 0{,}0104}{10}} = 0{,}0005 \text{ кг}$$

$$S_{1+2}^2 = \frac{\sum X_{1i}^2 - \frac{(\sum X_{1i})^2}{n_1} + \sum X_{2i}^2 - \frac{(\sum X_{2i})^2}{n_2}}{n_1 + n_2 - 2} = \frac{19{,}2417 - \frac{13{,}87^2}{10} + 18{,}2308 - \frac{13{,}50^2}{10}}{10 + 10 - 2} = 0{,}0104$$

3) Имајући у виду да је $(n_1 + n_2 - 2) < 30$, приликом провере нулте хипотезе, коришћене су таблице t-расподеле. С тим у вези, важи да је $t_{n_1+n_2-2;0,05} = 2{,}101$ и $t_{n_1+n_2-2;0,01} = 2{,}878$. Како је $|t|$ веће од обе критичне вредности, нулта хипотеза се одбацује за оба прага значајности α. Другим речима, закључак је да постоји високо статистични значајна разлика између посматраних аритметичких средина (прираст није исти за различите врсте примењеног хранива).

$$|t| > 2{,}101 \rightarrow H_1$$

$$|t| > 2{,}878 \rightarrow H_1$$

Полазна претпоставка се може проверити и на основу одговарајућег интервала поверења:

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha} \times S_{(\bar{X}_1-\bar{X}_2)} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;\alpha} \times S_{(\bar{X}_1-\bar{X}_2)}$$

95% интервал поверења је следећи:

$$(1{,}387 - 1{,}50) - 2{,}101 \times 0{,}0005 < \mu_1 - \mu_2 < (1{,}387 - 1{,}50) + 2{,}101 \times 0{,}0005 \text{ кг.}$$

$$0{,}0359 < \mu_1 - \mu_2 < 0{,}0381 \text{ кг}$$

$$0 \notin (L_1, L_2) \rightarrow H_1$$

Како вредност 0 не припада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна.

99% интервал поверења је следећи:

$$(1{,}387 - 1{,}50) - 2{,}878 \times 0{,}0005 < \mu_1 - \mu_2 < (1{,}387 - 1{,}50) + 2{,}878 \times 0{,}0005.$$

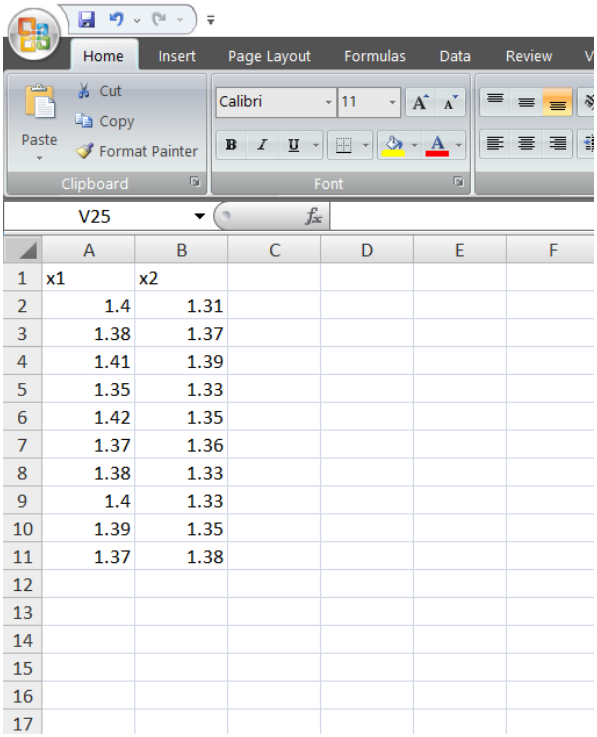
$$0{,}0356 < \mu_1 - \mu_2 < 0{,}0384 \text{ кг}$$

$$0 \notin (L_1, L_2) \rightarrow H_1$$

Како вредност 0 не припада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна.

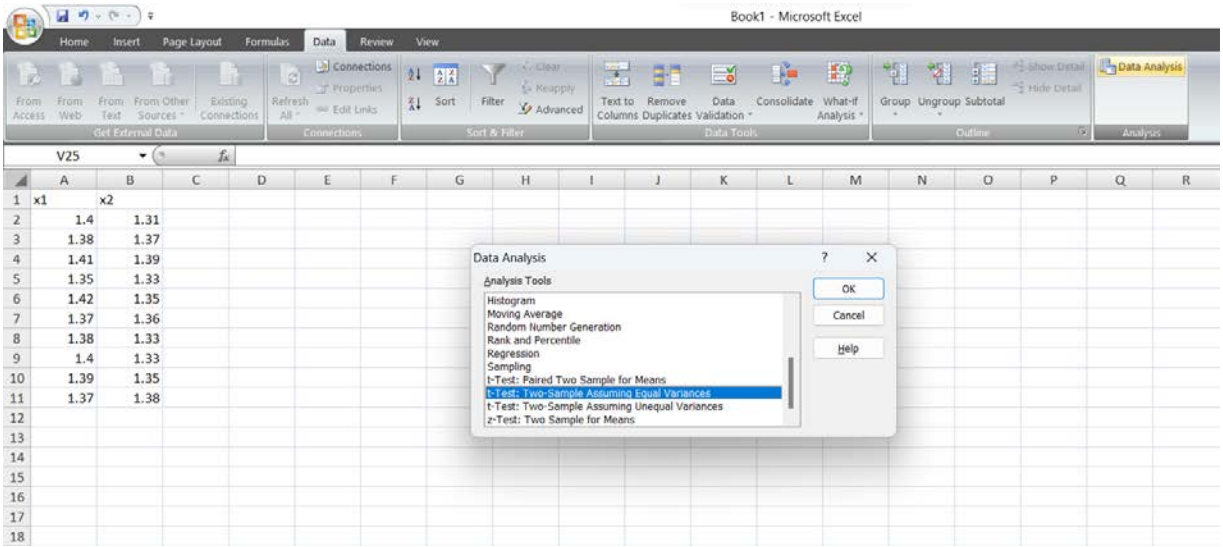
Вежба 8. Тест значајности разлике две средине (непознати варијабилитети основних скупова) применом Microsoft Excel-a

1. Полазни изглед табеле је следећи:

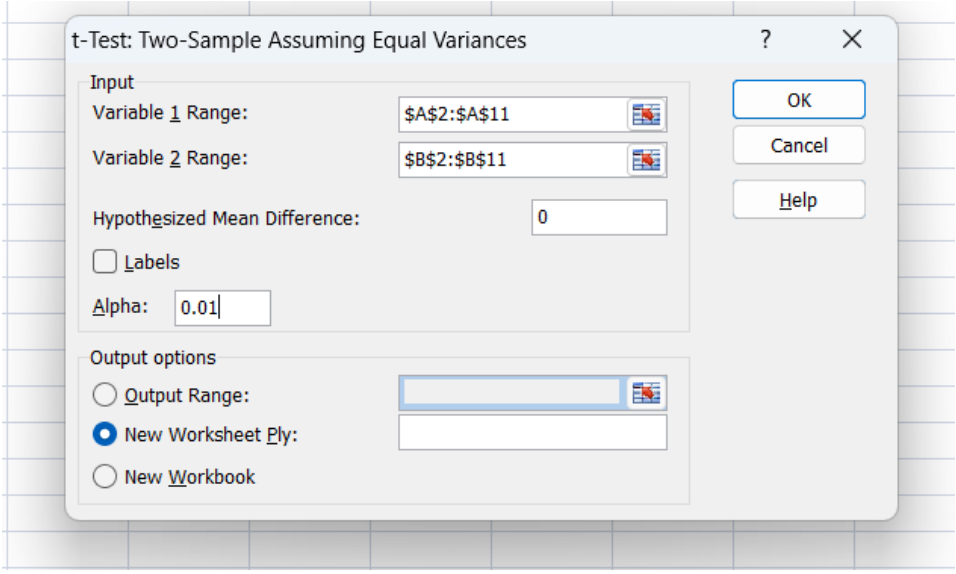


	A	B	C	D	E	F
1	x1	x2				
2	1.4	1.31				
3	1.38	1.37				
4	1.41	1.39				
5	1.35	1.33				
6	1.42	1.35				
7	1.37	1.36				
8	1.38	1.33				
9	1.4	1.33				
10	1.39	1.35				
11	1.37	1.38				
12						
13						
14						
15						
16						
17						

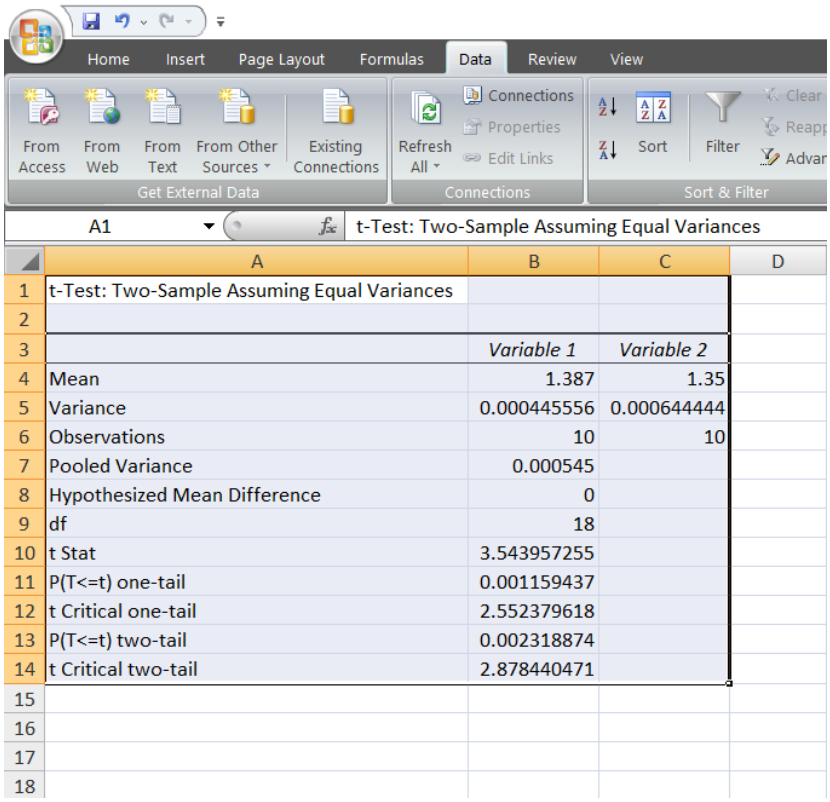
2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys* у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *t-Test:Two Sample Assuming Equal Variances* (изабрана је ставка *t-Test:Two Sample Assuming Equal Variances* јер се претпоставља да су варијансе основних скупова из којих су изабрани прости случајни узорци једнаке), као што је показано у наставку:



3. Кликом на дугме *OK* отвара се нови прозор. У блоку који се односи на *Input*, у делу који се односи на *Variable 1 Range* и *Variable 2 Range* потребно је обележити податке који припадају првом и другом узорку респективно. У делу у којем пише *Hypothesized Mean Difference* потребно је уписати вредност 0. На крају у делу који се односи на *Alpha* неопходно је уписати жељени праг значајности α. У примеру биће изабран праг значајности $\alpha=0,01$. Поглед на *Microsoft Excel* сада изгледа на следећи начин:



4. Кликом на дугме *OK* добијају се резултати теста у новом *Sheet*-у:



	A	B	C	D
1	t-Test: Two-Sample Assuming Equal Variances			
2				
3		*Variable 1*	*Variable 2*	
4	Mean	1.387	1.35	
5	Variance	0.000445556	0.000644444	
6	Observations	10	10	
7	Pooled Variance	0.000545		
8	Hypothesized Mean Difference	0		
9	df	18		
10	t Stat	3.543957255		
11	P(T<=t) one-tail	0.001159437		
12	t Critical one-tail	2.552379618		
13	P(T<=t) two-tail	0.002318874		
14	t Critical two-tail	2.878440471		
15				
16				
17				
18				

У првих шест редова наведени су подаци који се везују за природу теста (тестиране аритметичке средине, оцењене варијансе на основу узорка, величина узорака, здружена варијанса, нулта хипотеза, број степени слободе). Затим је представљена вредност *t*-количника који износи 3,5439. У наставку се пружа могућност поређења добијене вредности количника са критичним вредностима из таблице Студентове *t*-расподеле или са одговарајућом *p*-вредношћу. Посебну пажњу треба посветити изразима *P(T<=t) two-tail* и *t Critical two-tail*. Вредност *P(T<=t) two-tail* представља добијену *p*-вредност. С друге стране, *t Critical two-tail* представља критичну вредност из таблице Студентове *t*-расподеле за праг значајности који смо навели дефинишући полазне параметре теста. У овом случају критична вредност је 2,878, јер смо код *Alpha* у претходном кораку уписали 0,01. Како је *t* -количник већи од критичне вредности из таблице нормалне расподеле и како је *p*-вредност мања од 0,01, закључак је да се нулта хипотеза одбацује и прихвата се алтернативна.

Пример 18:

У огледу са две сорте шећерне репе добијену су следећи приноси по јединици површине (вагона/хектару):

Сорта А		**Сорта Б**		X_1f_1	X_2f_2	$X_1^2f_1$	$X_2^2f_2$
Принос X_1	**Површина f_1**	**Принос X_2**	**Површина f_2**				
4,2	3	4,6	2	12,6	9,2	52,99	42,32
4,6	4	4,9	4	18,4	19,6	84,64	96,04
5,1	6	5,3	5	30,6	26,5	156,06	140,45
5,3	5	5,9	4	26,5	23,6	140,45	139,24
5,8	2	6,2	3	11,6	18,6	67,28	115,32
Σ	**20**	**Σ**	**18**	**99,7**	**97,5**	**501,35**	**533,37**

1)

$$H_o: \bar{X}_1 = \bar{X}_2;$$

$$H_1: \bar{X}_1 \neq \bar{X}_2.$$

2)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{(\bar{X}_1 - \bar{X}_2)}} = \frac{4{,}985 - 5{,}417}{0{,}1677} = -2{,}57$$

$$\bar{X}_1 = \frac{\sum f_1 X_1}{n_1} = \frac{99{,}7}{20} = 4{,}985;$$

$$\bar{X}_2 = \frac{\sum f_2 X_2}{n_2} = \frac{97{,}5}{18} = 5{,}417.$$

Собзиром на то да су узорци неједнаке величине ($n_1 \neq n_2$), користи се следећи израз приликом израчунавања оцењене страндардне грешке разлике две средине:

$$S_{(\bar{X}_1-\bar{X}_2)} = \sqrt{S_{1+2}^2 \times \left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{S_{1+2}^2 \times \left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0{,}2664 \times \left(\frac{1}{20}+\frac{1}{18}\right)} = 0{,}1677$$

$$S_{1+2}^2 = \frac{\sum f_1 X_{1i}^2 - \frac{(\sum f_1 X_{1i})^2}{n_1} + \sum f_2 X_{2i}^2 - \frac{(\sum f_2 X_{2i})^2}{n_2}}{n_1 + n_2 - 2} = \frac{501{,}35 - \frac{99{,}7^2}{20} + 533{,}37 - \frac{97{,}5^2}{18}}{20 + 18 - 2} = 0{,}2664$$

3) Имајући у виду да је $(n_1 + n_2 - 2) > 30$, приликом провере нулте хипотезе, коришћене су таблице нормалне расподеле ($Z_{0,05} = 1{,}96$ и $Z_{0,01} = 2{,}58$). Како је $|t|$ веће од критичне вредности за праг значајности *α=0,05*, нулта хипотеза се одбацује и прихвата се алтернативна хипотеза. С друге стране, како је $|t|$ мање од критичне вредности за праг значајности *α=0,01*, нулта хипотеза се прихвата. Другим речима, може се закључити да постоји статистички значајна разлика између посматраних средина, али само за праг значајности *α=0,05*.

$$|t| > 1{,}96 \rightarrow H_1$$

$$|t| < 2{,}58 \rightarrow H_0$$

Полазна претпоставка се може проверити и на основу одговарајућег интервала поверења:

$$(\bar{X}_1 - \bar{X}_2) - Z_\alpha \times S_{(\bar{X}_1-\bar{X}_2)} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_\alpha \times S_{(\bar{X}_1-\bar{X}_2)}$$

95% интервал поверења је следећи:

$$(4{,}985 - 5{,}417) - 1{,}96 \times 0{,}1677 < \mu_1 - \mu_2 < (4{,}985 - 5{,}417) + 1{,}96 \times 0{,}1677.$$

$$-0{,}7607 < \mu_1 - \mu_2 < -0{,}1033$$

$$0 \notin (L_1, L_2) \rightarrow H_1$$

Како вредност 0 неприпада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна.

99% интервал поверења је следећи:

$$(4{,}985 - 5{,}417) - 2{,}58 \times 0{,}1677 < \mu_1 - \mu_2 < (4{,}985 - 5{,}417) + 2{,}58 \times 0{,}1677.$$

$$-0{,}8647 < \mu_1 - \mu_2 < 0{,}0007$$

$$0 \in (L_1, L_2) \rightarrow H_0$$

Како вредност 0 припада дефинисаном интервалу, нулта хипотеза се прихвата.

Приликом теста значајности разлике две средине (непознати варијабилитети основних скупова и неједнаке варијансе) применом *Microsoft Excel*-а, могуће је пратити поступак који је објашњен у вежби 6. Једина разлика је у томе што ће се у другом кораку уместо

t-Test:Two Sample Assuming Equal Variances изабрати опција *t-Test:Two Sample Assuming Unequal Variances*. Наравно, овакву процедуру је могуће применити уколико су подаци негруписани. За груписане податке потребно је формирати радну табелу и постепено израчунати све потребне параметре, као што би то био случај код решавања претходних задатака. Предност рада у *Microsoft Excel*-у је у томе што је цео поступак брже извести уз мању вероватноћу да се направи грешка у рачуну.

5.2. Тестови пропорција

Приликом тестирања пропорција издвајају се два основна вида тестирања:

- тест значајности једне пропорције;
- тест значајности разлике две пропорције.

5.2.1. Тестирање хипотезе о пропорцији основног скупа

Код тестирања хипотезе о пропорцији основног скупа полази се од следеће нулте и алтернативне хипотезе:

$$H_0: p = p_0;$$

$$H_1: p \neq p_0$$

Тестирање се изводи упоређивањем пропорције из узорка са претпостављеном пропорцијом основног скупа:

$$H_0: \hat{p} = p_0;$$

$$H_1: \hat{p} \neq p_0.$$

У циљу провере нулте хипотезе израчунава се количник Z на следећи начин:

$$Z = \frac{\hat{p} - p_0}{S_{\hat{p}}}.$$

Претпостављајући нулту хипотезу, Z количник има приближно нормалну расподелу уколико се тестирање изводи на основу великог узорка и ако важи да је $n \times p > 5$ и $n \times q > 5$. У том случају израчунати количник се упоређује са вредностима из таблице норалне дистрибуције. Основ за одбацивање нулте хипотезе јесте вредност израчунатог количника који је већа или једнак од критичне вредности из одговарајуће таблице. У случају малог узорка, у тестирању се користи чињеница да пропорција $\hat{p}$ из узорка има биномну расподелу.

Провера хипотезе о значајности једне пропорције може се извести и израчунавањем интервала поверења:

$$\hat{p} - Z_\alpha \times S_{\hat{p}} < p < \hat{p} + Z_\alpha \times S_{\hat{p}}.$$

Уколико претпостављена вредност пропорције основног скупа p_0 припада границама утврђеног интервала, има основа за прихватање нулте хипотезе као тачне. У супротном случају, нулта хипотеза се одбацује и прихвата се алтернативна хипотеза.

Пример 19:

У узорку величине *n=100*, изникло је 80 биљака. Тестирати значајност разлике пропорција изниклих биљака у узорку и претпостављене пропорције у основном скупу p_0*=0,90*.

1)

$H_0: \hat{p} = p_0$;

$H_1: \hat{p} \neq p_0$.

2)

$$\hat{p} = \frac{a}{n} = \frac{80}{100} = 0{,}80;$$

$$Z = \frac{\hat{p} - p_0}{S_{\hat{p}}} = \frac{0{,}8 - 0{,}9}{0{,}04} = 2{,}50$$

$$S_{\hat{p}} = \sqrt{\frac{\hat{p} \times \hat{q}}{n}} = \sqrt{\frac{0{,}80 \times (1 - 0{,}80)}{100}} = 0{,}04.$$

3) Критичне вредности из таблице нормалне расподеле за праг значајности $\alpha = 0{,}05$ и $\alpha = 0{,}01$ су 1,96 и 2,58 респективно. Како је апсолутна вредност Z-количника 2,50, закључак је да се нулта хипотеза прихвата тек за праг значајности $\alpha = 0{,}01$, док се исти закључак не може извести за праг значајности $\alpha = 0{,}05$.

$$|Z| > 1{,}96 \rightarrow H_1$$

$$|Z| < 2{,}58 \rightarrow H_0$$

Други начин провере полазне претпоставке подразумева претходно дефинисање одговарајућег интервала поверења. У складу са расположивим подацима, интервал поверења који је неопходно дефинисати је следећи:

$$\hat{p} - Z_\alpha \times S_{\hat{p}} < p < \hat{p} + Z_\alpha \times S_{\hat{p}}.$$

95% интервал поверења је следећи:

$$0{,}80 - 1{,}96 \times 0{,}0400 < p < 0{,}80 + 1{,}96 \times 0{,}0400$$

$$0{,}72 < p < 0{,}88$$

$$p_0 \notin (L_1, L_2) \rightarrow H_1$$

Како вредност за претпостављену вредност пропорције ($p_0 = 0{,}90$) не припада дефинисаном интервалу, нулта хипотеза се одбацује и прихвата се алтернативна.

99% интервал поверења је следећи:

$$0{,}80 - 2{,}58 \times 0{,}0400 < p < 0{,}80 + 2{,}58 \times 0{,}0400$$

$$0{,}70 < p < 0{,}90$$

$$p_0 \in (L_1, L_2) \rightarrow H_0$$

С обзиром на то да претпостављена пропорција ($p_0 = 0{,}9$) припада дефинисаном интервалу, нулта хипотеза се прихвата.

5.2.2. Тестирање хипотезе о једнакости пропорција два основна скупа

Ако су p_1 и p_2 релативне фреквенције заступљености неког својства у два основна скупа, полази се од следеће нулте и алтернативне хипотезе:

$$H_0: p_1 = p_2;$$

$$H_1: p_1 \neq p_2.$$

На основу два независна случајна узорка из посматраних скупова, израчунавају се оцене пропорција и утврђује да ли се статистички значајно разликују:

$$H_0: \hat{p}_1 = \hat{p}_2;$$

$$H_1: \hat{p}_1 \neq \hat{p}_2.$$

Уколико су узорци велики уз испуњене услове где важи: $n_1 \times p_1 > 5$; $n_2 \times p_2 > 5$; $n_1 \times q_1 > 5$ и $n_2 \times q_2 > 5$, полазна хипотеза се проверава израчунавањем следећег количника:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{S_{(\hat{p}_1 - \hat{p}_2)}},$$

где је $S_{(\hat{p}_1 - \hat{p}_2)}$ оцењена стандардна грешка разлике пропорција.

Дакле, како би се дошло до вредности количника на основу ког ће се проверити полазна претпоставка, потребно је прво утврдити оцењене вредности пропорција узорака ($\hat{p}_1$ и $\hat{p}_2$), као и стандардну грешку разлике две пропорције ($S_{(\hat{p}_1 - \hat{p}_2)}$). Оцењене пропорције из узорака добијају се на основу следећих израза:

$$\hat{p}_1 = \frac{a_1}{n_1} \quad \text{и} \quad \hat{p}_2 = \frac{a_2}{n_2}.$$

Оцењена стандардна грешка разлике две пропорције може се израчунати применом два израза:

$$S_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{2 \times \bar{p} \times \bar{q}}{n_1 + n_2}} \quad \text{или} \quad S_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\bar{p} \times \bar{q} \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

где су $\bar{p}$ и $\bar{q}$ просечне пропорције на основу два узорка:

$$\bar{p} = \frac{\hat{p}_1 \times n_1 + \hat{p}_2 \times n_2}{n_1 + n_2}; \quad \bar{p} = \frac{a_1 + a_2}{n_1 + n_2}; \quad \bar{q} = 1 - \bar{p}.$$

Израчунати количник Z упоређује се са критичним вредностима из таблица нормалне дистрибуције. Ако је израчуната вредност количника мања од одговарајућих критичних вредности из таблице нормалне расподеле, има основа да се прихвати полазна, односно нулта хипотеза и закључи да је посматрана карактеристика подједнако заступљена у основним скуповима из којих су изабрани узорци. Уколико се прихвати алтернативна хипотеза, што значи да је израчунати количник већи од критичних вредности из таблице нормалне расподеле, закључак је да постоји статистички значајна разлика у застпуљености посматране каратеристике у основним скуповима.

Тест значајности разлике две пропорције може се извести и израчунавањем одговарајућег интервала поверења:

$$(\hat{p}_1 - \hat{p}_2) - Z_\alpha \times S_{(\hat{p}_1 - \hat{p}_2)} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_\alpha \times S_{(\hat{p}_1 - \hat{p}_2)}.$$

Нулта хипотеза се може прихватити као тачна уколико границе израчунатог интервала укључују вредност 0. Супротно, ако 0 не припада границама израчунатог интервала, одбацује се нулта и прихвата алтернативна хипотеза.

$$0 \in (L_1, L_2) \rightarrow H_0;$$

$$0 \notin (L_1, L_2) \rightarrow H_1.$$

Пример 20:

У узорцима од по 110 грла две расе говеда, оболела грла учествују са 6% и 13%. Утврдити да ли је отпорност две посматране расе говеда према испитиваној болести иста.

1)

$$H_0: \hat{p}_1 = \hat{p}_2;$$

$$H_1: \hat{p}_1 \neq \hat{p}_2.$$

2.1) Први начин обрачуна оцењене стандардне грешке и Z количника, представљена је у наставку:

$$Z_1 = \frac{\hat{p}_1 - \hat{p}_2}{S_{(\hat{p}_1 - \hat{p}_2)}} = \frac{0{,}06 - 0{,}13}{0{,}028} = -2{,}5$$

$$S_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{2 \times \bar{p} \times \bar{q}}{n_1 + n_2}} = \sqrt{\frac{2 \times 0{,}095 \times 0{,}905}{110 + 110}} = 0{,}028$$

$$\hat{p}_1 = \frac{6}{100} = 0{,}06; \hat{p}_2 = \frac{13}{100} = 0{,}13$$

$$\bar{p} = \frac{\hat{p}_1 \times n_1 + \hat{p}_2 \times n_2}{n_1 + n_2} = \frac{0{,}06 \times 110 + 0{,}13 \times 110}{110 + 110} = 0{,}095$$

$$\bar{q} = 1 - \bar{p} = 1 - 0{,}095 = 0{,}905$$

2.2) Други начин обрачуна оцењене стандардне грешке и Z количника, представљена је у наставку:

$$Z_2 = \frac{\hat{p}_1 - \hat{p}_2}{S_{(\hat{p}_1 - \hat{p}_2)}} = \frac{0{,}06 - 0{,}13}{0{,}0396} = -1{,}768$$

$$S_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{0{,}095 \times 0{,}905 \times \left(\frac{1}{110} + \frac{1}{110}\right)} = 0{,}0396,$$

$$\hat{p}_1 = \frac{6}{100} = 0{,}06;\ \hat{p}_2 = \frac{13}{100} = 0{,}13$$

$$\bar{p} = \frac{\hat{p}_1 \times n_1 + \hat{p}_2 \times n_2}{n_1 + n_2} = \frac{0{,}06 \times 110 + 0{,}13 \times 110}{110 + 110} = 0{,}095$$

$$\bar{q} = 1 - \bar{p} = 1 - 0{,}095 = 0{,}905$$

3.1) На основу обрачунатог количника Z_1 може се закључити да постоји статистички значајна разлика између посматраних пропорција али само за праг значајности *α=0,05*, не и за праг значајности *α=0,01*.

$$|Z_1| > 1{,}96 \rightarrow H_1$$

$$|Z_1| < 2{,}58 \rightarrow H_0$$

3.2) На основу обрачунатог количника Z_2 може се закључити да не постоји статистички значајна разлика између посматраних пропорција. Другим речима, нулта хипотеза се прихвата.

Када је реч о интервалима поверења, могу се издвојити две групе инетрвала у складу са коришћеном формулом за обрачун оценњене стандардне грешке разлике пропорција. Интервал поверења је следећи:

$$(\hat{p}_1 - \hat{p}_2) - Z_\alpha \times S_{(\hat{p}_1 - \hat{p}_2)} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + Z_\alpha \times S_{(\hat{p}_1 - \hat{p}_2)}$$

I начин

95% интервал поверења:

$$(0{,}06 - 0{,}13) - 1{,}96 \times 0{,}028 < p_1 - p_2 < (0{,}06 - 0{,}13) + 1{,}96 \times 0{,}028$$

$$-0{,}1249 < p_1 - p_2 < -0{,}0151$$

$$0 \notin (L_1, L_2) \rightarrow H_1.$$

99% интервал поверења:

$$(0{,}06 - 0{,}13) - 2{,}58 \times 0{,}028 < p_1 - p_2 < (0{,}06 - 0{,}13) + 2{,}58 \times 0{,}028$$

$$-0{,}1422 < p_1 - p_2 < 0{,}0022$$

$$0 \in (L_1, L_2) \rightarrow H_0.$$

II начин

95% интервал поверења:

$$(0{,}06 - 0{,}13) - 1{,}96 \times 0{,}0396 < p_1 - p_2 < (0{,}06 - 0{,}13) + 1{,}96 \times 0{,}0396$$

$$-0{,}1476 < p_1 - p_2 < 0{,}0076$$

$$0 \in (L_1, L_2) \rightarrow H_0.$$

99% интервал поверења:

$$(0{,}06 - 0{,}13) - 2{,}58 \times 0{,}0396 < p_1 - p_2 < (0{,}06 - 0{,}13) + 2{,}58 \times 0{,}0396$$

$$-0{,}1722 < p_1 - p_2 < 0{,}0322$$

$$0 \in (L_1, L_2) \rightarrow H_0.$$

Када је реч о коришћењу *Microsoft Excel-a* у циљу тестирања разлике између две пропорције, као што је било речи до сада, могуће је једино формирати радну табелу и поступно спровести тестирање.

5.3. Анализа варијансе (АНОВА)

У истраживачком раду често се проверава постојање разлика између више од две аритметичких средина истог или различитих основних скупова. Статистички поступак код оваквих истраживања познат је под називом *анализа варијансе (АНОВА)*. Статистичар и генетичар Роналд Фишер је увео термин 1918. и развио метод анализе варијансе 1925. године.

Анализа варијансе се састоји у испитивању варијабилитета аритметичких средина из више случајно одабраних узорака, при чему се укупан варијабилитет (укупна варијанса), раздваја на саставне делове, односно на варијабилитет који настаје услед утицаја примењених третмана и на случајан варијабилитет.

Пре примене анализе варијансе важно је да се размотре одређене претпоставке како би се осигурали тачност и поузданост резултата. Кључне претпоставке за примену анализе варијансе су:

1. Основни скупови из којих се бирају узорци имају приближно нормалну расподелу. Претпоставка се обично проверава помоћу графикона нормалности и статистичких тестова нормалности (нпр. Колмогоров-Смирнов тест, Шапиро Вилков тест);

2. Основни скупови из којих се бирају узорци имају једнаке (хомогене) варијансе. То значи да разлике међу групама нису резултат велике варијабилности унутар исте групе. Да би се проверила ова претпоставка примењује се неки од тестова хомогености варијансе (нпр. Бартлетов тест, Левенов тест);

3. Узорци изабрани из различитих основних скупова су случајни и независни;
4. Пожељно је да су узорци из основних скупова приближно исте величине (балансиран дизајн).

5.3.1. Анализа варијансе потпуно случајног распореда

У анализи варијансе потпуно случајног распореда полази се од k узорака (третмана) израчунавањем њихове аритметичке средине. Аритметичка средина сваког од k узорка дефинисана је на следећи начин:

$$\bar{X}_{i\cdot} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i},$$

где је:

n_i - број јединица у узорку;

X_{ij} - вредност обележја j-те јединице i-тог третмана.

Поред аритметичке средине сваког од k узорака, израчунава се и општа средина свих N јединица из свих узорака:

$$\bar{X}_{i\cdot\cdot} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{\sum_{i=1}^{k} n_i} = \frac{T}{N}, \qquad N = \sum_{i=1}^{k} n_i.$$

Збир вредности обележја свих N јединица назива се *тотал* и означава се са T.

Ако су сви узорци једнаке величине, односно са једнаким бројем понављања (n), укупан број јединица у анализи варијансе (N) може се исказати на следећи начин:

$$n = n_1 = n_2 = \cdots = n_k;$$

$$N = n \times k.$$

Варијабилитет који настаје применом одабраних k третмана на N јединица, у анализи варијансе исказује се на основу одступања сваке индивидуалне вредности обележја од опште средине:

$$\left(X_{ij} - \bar{X}_{\cdot\cdot}\right) = \left(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}\right) + \left(X_{ij} - \bar{X}_{i\cdot}\right).$$

Уколико се наведени израз квадрира, добијају се одговарајуће суме квадрата:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{..}\right)^2=\sum_{i=1}^{k}(\bar{X}_{i.}-\bar{X}_{..})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_1}\left(X_{ij}-\bar{X}_{\boldsymbol{i}.}\right)^2,$$

где је:

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_{..}\right)^2$ – сума квадрата тотала Q;

$\sum_{i=1}^{k}(\bar{X}_{i.}-\bar{X}_{..})^2$ – сума квадрата третмана Q_T;

$\sum_{i=1}^{k}\sum_{j=1}^{n_1}\left(X_{ij}-\bar{X}_{\boldsymbol{i}.}\right)^2$ – сума квадрата погрешке Q_P;

$Q = Q_T + Q_P$.

На основу дефинисаних сума квадрата проверава се полазна хипотеза у примени метода анализе варијансе. Полазна претпоставка код анализи варијансе потпуно случајног распореда, којом се претпоставља да нема разлике у утицају различитих третмана гласи:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k.$$

Алтернативна хипотеза дефиисана је на следећи начин:

$$H_1: \exists(i,j)\ \ i \neq j\ \ \mu_i \neq \mu_j\ \ 1 \leq i \leq k.$$

Дакле, нултом хипотезом се тврди да су аритметичке средине *k*-основних скупова једнаке, док је тврђење алтернативне хипотезе да постоји бар један пар аритметичких средина који се разликује.

Полазна претпоставка проверава се извођењем *F*-теста. За потребе извођење овог теста формира табела анализе варијансе која је представљена у наставку:

Извори Варијације	**Степени слободе**	**Суме квадрата**	**Средина сума квадрата**	***F*-однос**	***F*-таблично $r_1 = k-1$; $r_2 = N-k$**	
					α = 0,05	***α = 0,01***
Третрмани	k-1	Q_T	S_T^2	S_T^2/S_P^2		
Погрешка	N-k	Q_P	S_P^2			
Тотал	N-1	Q				

Суме квадрата се у практичном раду израчунавају применом следећих радних формула:

$$Q=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-C,$$

где је *C* ознака за коректив фактор који се израчунава на следећи начин:

$$C=\frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}\right)^2}{N}.$$

Ако се уведе да је $\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = T$, корективни фактор се може исказати као $C = \frac{T^2}{N}$. У практичном раду, сума квадрата третмана Q_T се издваја као следећа сума квадрата коју је потребно израчунати. У општем случају уколико су третмани примењени на различитом броју јединица, Q_T се израчунава на следећи начин:

$$Q_T = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i} - C.$$

Како важи да је $\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = T$, средина сума квадрата третмана може се израчунати и на следећи начин:

$$Q_T = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C\,.$$

На основу израчунате суме квадрата тотала и суме квадрата третмана долази се до вредности суме квадрта погрешке:

$$Q_P = Q - Q_T\,.$$

Средине сума квадрата, односно варијансе, израчунавају се као количник сума квадрата и одговарајућих степени слободе. Када је реч о варијанси третмана, потребно је ставити у однос суму квадрата третмана и одговарајући број степени слободе третмана:

$$S_T^2 = \frac{Q_T}{k-1}\,.$$

Варијанса погрешке која подразумева количник између суме квадрата погрешке и броја степени слободе погрешке једнака је:

$$S_P^2 = \frac{Q_P}{N-k}\,.$$

За проверу полазне претпоставке израчунава се F-однос који представља количник израчунатих варијанси и увек је вредност већа од нуле:

$$F = \frac{S_T^2}{S_P^2}\,.$$

У циљу доношења закључка о полазној претпоставци, F-количник се упоређује са критичним вредностима из таблица Фишерове дистрибуције, које се очитавају за праг значајности α и степене слободе r_1 и r_2. Уколико је вредност израчунатог F-количника мања од критичних вредности из таблица Фишерове расподеле, нулта хипотеза се може прихватити као тачна.

Прихватање полазне хипотезе указује на то да између примењених третмана не постоје статистички значајне разлике у дејству третмана на експерименталне јединице и тиме се анализа варијансе завршава.

С друге стране, уколико је вредност израчунатог количника F већа од критичних вредности из таблица Фишерове расподеле, одбацује се нулта и прихвата алтернативна хипотеза као тачна. Ако се полазна хипотеза не прихвати, већ се утврди постојање значајних или врло значајних разлика између било која два примењена третмана, анализа варијансе се даље наставља, како би се утврдило између којих третмана постоје статистички значајне разлике. У наставку анализе варијансе примењују се тестови парова третмана, на основу којих се утврђује између којих третмана постоје значајне или врло значајне разлике. Приликом тестирања разлика између средина третмана најчешће се користе следећи тестови:

- тест парова третмана (t – тест);
- тест најмање значајне разлике – НЗР тест;
- вишеструки тест интервала – Данканов тест.

Применом наведених тестова могуће је извести више упоређења између аритметичких средина третмана. Број могућих упоређења условљен је бројем испитиваних третмана, а може се одредити на основу следећег израза:

$$\frac{k \times (k-1)}{2}.$$

5.3.1.1. Тест парова третмана t – тест

Полазна и алтернативна хипотеза код теста парова третмана гласе:

$$H_0: \mu_i = \mu_j;$$

$$H_1: \mu_i \neq \mu_j$$

$$(i < j), \qquad 1 \leq i \leq k, \qquad 1 \leq j \leq k.$$

Нултом хипотезом се претпоставља да нема разлике између посматраних аритметичких средина, док алтернативна хипотеза указује на то да статистички значајна разлика ипак постоји. Полазна претпоставка се проверава израчунавањем t – количника:

$$t = \frac{\bar{X}_i - \bar{X}_j}{S_{(\bar{X}_i - \bar{X}_j)}},$$

где су:

$\bar{X}_i$ и $\bar{X}_j$ - аритметичке средине испитиваних тертмана;

$S_{(\bar{X}_i - \bar{X}_j)}$ - оцена стандардне грешке разлике две аритметичке средине.

Уколико су третмани примењени на једнаком броју јединица посматрања, односно ако је реч о једнаком броју понављања код сваког испитиваног третмана ($n_i = n_j = n$), оцена стандардне грешке разлике две средине израчунава се на основу варијансе погрешке из табеле анализе варијансе применом следећег израза:

$$S_{(\bar{X}_i - \bar{X}_j)} = \sqrt{\frac{2 \times S_P^2}{n}}.$$

Уколико су испитивани третмани примењивани на неједнаком броју понављања што је неопходно узети у обзир приликом израчунавања оцене стандардне грешке разлике две средине, израчунава се следећа формула:

$$S_{(\bar{X}_i - \bar{X}_j)} = \sqrt{S_P^2 \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Израчунати количник t упоређује се са критичним вредностима из таблице Студентове дистрибуције очитаним за праг значајности α и степен слободе погрешке ($N - k$). Апсолутна вредност количника t која је мања од критичних вредности из таблице Студентове расподеле, подразумева прихватање нулте хипотезе о једнаком дејству два испитивана третмана. У супротном, ако је вредност количника већа од критичних вредности прихвата се алтернативна хипотеза и закључује да између два посматрана третмана постоје статистички значајне разлике у дејству на експерименталне јединице.

5.3.1.2. Тест најмање значајне разлике – НЗР тест

Нулта и алтернативна хипотеза код теста најмање значајне разлике (НЗР тест) је иста као и код теста парова третмана:

$$H_0: \mu_i = \mu_j;$$

$$H_1: \mu_i \neq \mu_j$$

$$(i < j), \qquad 1 \leq i \leq k, \qquad 1 \leq j \leq k.$$

У циљу израчунавања НЗР теста потребно је израчунавање најмање значајне разлике која представља производ одговарајуће критичне вредности из таблице Студентове расподеле и стандардне грешке разлике две средине.

$$NZR_\alpha = t_{N-k;\alpha} \times S_{(\bar{X}_i - \bar{X}_j)}.$$

Затим се формира помоћна табела у коју се у прву колони унесу вредности аритметичке средине третмана, уређене према величини у вертикалном низу од максималне до минималне вредности. У следеће колоне се уносе разлике аритметичких средина третмана, које су увек позитивне вредности. Разлике средина третмана упоређују се са израчунатим најмање значајним разликама.

Уколико су добијене разлике мање од NZR_α прихвата се нулта хипотеза. У супротном, уколико су израчунате разлике између аритметичких средина веће од NZR_α, нулта хипотеза се одбацује и прихвата се алтернативна хипотеза.

Примера ради, ако је у анализи варијансе примењено четири третмана А, Б, Ц и Д где највећу просечну вредност има третман А, па третман Б, затим третман Ц и најмању вредност средине третман Д. Табела за НЗР тест има следећи изглед:

Третмани		$\bar{X}_i$	$\bar{X}_i - \bar{X}_Д$	$\bar{X}_i - \bar{X}_Ц$	$\bar{X}_i - \bar{X}_Б$
А	max	$\bar{X}_A$	•	•	•
Б	·	$\bar{X}_Б$	•	•	
Ц	·	$\bar{X}_Ц$	•		
Д	min	$\bar{X}_Д$			

Правило одлучивања је:

$$\bar{X}_i - \bar{X}_j < NZR_\alpha \rightarrow H_0;$$

$$\bar{X}_i - \bar{X}_j \geq NZR_\alpha \rightarrow H_1.$$

5.3.1.3. Вишеструки тест интервала – Данканов тест

Нулта и алтернативна хипотеза су и код вишеструког теста интервала формулисане на исти начин као и код претходна два теста:

$$H_0: \mu_i = \mu_j;$$

$$H_1: \mu_i \neq \mu_j$$

$$(i < j), \qquad 1 \leq i \leq k, \qquad 1 \leq j \leq k.$$

Извођењу Данкановог теста претходи израчунавање оцене стандардне грешке аритметичке средине на основу варијансе погрешке из табеле анализе варијансе као и броја понављања у третманима:

$$S_{\bar{X}} = \sqrt{\frac{S_P^2}{n}}.$$

Претпоставка за примену Данкановог теста јесте једнак број понављања код сваког испитиваног третмана. У наставку је потребно формирати две табеле за два прага значајности (*α=0,05* и *α=0,01*) следећег облика:

Интервал	**2**	**3**	**4**	**5**	**...**	**k**
Критична вредност						
Најмање значајни интервал						

У табели у првом реду (друга колона) уписују се могући интервали на основу броја посматраних третмана. Затим се очитавају критичне вредности из таблица за вишеструки тест интервала за дате прагове значајности α и степене слободе погрешке из табеле анализе варијансе. За сваки интервал идући од 2, 3, 4, ..., k, уписују се критичне вредности у други ред табеле. Очитане и уписане критичне вредности множе се са израчунатом оценом стандардне грешке аритметике средине, а производ представља вредност најмањег значајног интервала и њега уписујемо у трећи ред табеле.

Разлике аритметичких средина третмана упоређују се са најмање значајним интервалима. Аритметичке средине третмана рангирају се у помоћној табели у хоризонталном низу од минималне до максималне вредности.

Примера ради, уколико ако је у анализи варијансе примењено четири третмана А, Б, Ц и Д где највећу просечну вредност има третман А, па третман Б, затим третман Ц и најмању вредност средине третман Д. Табела представљена у наставку илуструје претходно наведено:

Третман	**Д** **min**	•	**Ц**	•	**Б**	•	**А** **max**
$\bar{X}_i$	$\bar{X}_Д$		$\bar{X}_Ц$		$\bar{X}_Б$		$\bar{X}_A$

Највећа критична вредност користи се код поређења аритметичких средина између којих је $k - 1$ интервала (у овом примеру 3 интервала, односно $\bar{X}_A - \bar{X}_Д$). Прва мања критична вредност се користи код поређења средина између којих је $k - 2$ интервала (у овом примеру то су поређења $\bar{X}_A - \bar{X}_Ц$ и $\bar{X}_Б - \bar{X}_Д$). У складу са примером, најмања критична вредност је потребна за поређење средина између којих је 1 интервал. У питању су поређења: $\bar{X}_A - \bar{X}_Б$, $\bar{X}_Б - \bar{X}_Ц$ и $\bar{X}_Ц - \bar{X}_Д$. Уколико су разлике између аритметичких средина мање од одговарајућих критичних вредности, неопходно је прихватити нулту хипотезу. У обрнутом случају прихвата се алтернативна хипотеза.

Битно је још истаћи да су *t*-тест и НЗР тест еквивалентни, што значи да се њиховом применом долази до истог закључка. Уколико број понављања третмана није исти, примењује се *t*-тест. У случају једнаког броја третмана, због прегледнијег приказивања резултата, чешће се примењује НЗР тест. Уколико је велики број поређења, ова два теста нису објективна јер је вероватноћа да се погрешно закључи да је разлика два третмана статистички значајна већа од изабраног прага значајности α. У том случају се препоручује вишеструки интервални тест.

Пример 21:

На релативно хомогеном пољу, подељеном на 20 једнаких парцела, засејане су 4 сорте пшенице, свака на 5 парцела. Размештај сорти по парцелама је потпуно случајан. Резултати експеримента који се односе на принос (00кг/хектару) дати су у следећој табели:

Редни број	**СОРТЕ**			
	А	**Б**	**В**	**Г**
1	32,3	33,3	30,8	29,3
2	34,0	33,0	34,3	26,0
3	34,3	36,3	35,3	29,8
4	35,0	36,8	32,3	28,0
5	36,5	34,5	35,8	28,8
$\boldsymbol{T_i}$	**172,1**	**173,9**	**168,5**	**141,9**
$\boldsymbol{\bar{X}_i}$	**34,42**	**34,78**	**33,70**	**28,38**

а) тестирати нулту хипотезу о једнаком просечном приносу појединих сорти.

б) Тестирати значајност разлике парова третмана применом: *t*-теста, теста најмање значајне разлике (НЗР тест) и вишеструког интервалног (Данкановог) теста.

Решење:

$$H_0: \bar{X}_A = \bar{X}_\text{Б} = \bar{X}_\text{В} = \bar{X}_\Gamma$$

$$H_1: \exists(i,j)\ \ \bar{X}_i \neq \bar{X}_j\ \ (i \neq j, \qquad i,j = \text{А, Б, В, Г})$$

Извори Варијације	**Степени слободе**	**Суме квадрата**	**Средина сума квадрата**	***F*-однос**	***F*-таблично** $r_1 = k - 1$; $r_2 = N - k$	
					α = 0,05	***α = 0,01***
Третрмани	k-1=4-1=3	Q_T=134,45	S_T^2=44,82	S_T^2/S_P^2=15,02	3,24	5,29
Погрешка	N-k=20-4=16	Q_P=47,72	S_P^2=2,98			
Тотал	N-1=20-1=19	Q=182,17				

Обрачун укупне суме квадрата изводи се на следећи начин:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - C = 32{,}3^2 + 34{,}0^2 + 34{,}3^2 + \cdots + 28{,}8^2 - 21.543{,}05 = 182{,}17$$

$$C = \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}\right)^2}{N} = \frac{T^2}{N} = \frac{656{,}4^2}{20} = 21.543{,}05.$$

$$T = \sum_{i=1}^{4} T_i = 172{,}1 + 173{,}9 + 168{,}5 + 141{,}9 = 656{,}4\ (00\text{кг/ха})$$

Сума квадрата третмана Q_T израчунава се на следећи начин:

$$Q_T = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i} - C = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - C = \frac{172{,}1^2 + 173{,}9^2 + 168{,}5^2 + 141{,}9^2}{5} - 21.543{,}05 = 134{,}45.$$

На основу израчунате суме квадрата тотала и суме квадрата третмана долази се до вредности суме квадрта погрешке:

$$Q_P = Q - Q_T = 182{,}17 - 134{,}45 = 47{,}72.$$

Приликом онрачуна варијансе третмана, потребно је ставити у однос суму квадрата третмана и одговарајући број степени слободе третмана:

$$S_T^2 = \frac{Q_T}{k-1} = \frac{134{,}45}{3} = 44{,}82.$$

Варијанса погрешке подразумева количник између суме квадрата погрешке и броја степени слободе погрешке једнака је:

$$S_P^2 = \frac{Q_P}{N-k} = \frac{182{,}17}{16} = 2{,}98.$$

Како би се проверила полазна претпоставка израчунава се *F*-однос који представља количник претходно израчунатих варијанси:

$$F = \frac{S_T^2}{S_P^2} = \frac{44{,}82}{2{,}98} = 15{,}02.$$

Критичне вредности из таблице Фишерове таблице за $r_1 = k - 1 = 4 - 1 = 3$ и $r_2 = N - k = 20 - 4 = 16$ степени слободе. С тим у вези, у конкретном примеру важи:

$$F_{3;16(0{,}05)} = 3{,}24;$$

$$F_{3;16(0{,}01)} = 5{,}29.$$

Како је F-количник већи од обе критичне вредности, може се донети закључак да се нулта хипотеза одбацује и прихвата се алтернативна хипотеза која гласи да постоји статистички значајна разлика у утицају између посматраних третмана.

С обзиром на то, да алтернативна хипотеза указује само на то да постоји статистички значајна разлика између посматраних аритметичких средина, додатним тестирањем је неопходно установити између којих конкретно аритметичких средина постоји статистички значајна разлика.

Тест парова тертмана (t-тест)

$$H_0: \bar{X}_i = \bar{X}_j;$$

$$H_1: \bar{X}_i \neq \bar{X}_j$$

Полазна претпоставка се проверава израчунавањем t – количника:

$$t = \frac{\bar{X}_i - \bar{X}_j}{S_{(\bar{X}_i - \bar{X}_j)}},$$

$$S_{(\bar{X}_i - \bar{X}_j)} = \sqrt{\frac{2 \times S_P^2}{n}} = \sqrt{\frac{2 \times 2{,}98}{5}} = 1{,}0923.$$

$$\bar{X}_A = 34{,}42; \bar{X}_\text{Б} = 34{,}78;\ \bar{X}_\text{В} = 33{,}70;\ \bar{X}_\Gamma = 28{,}3.$$

$$t_1 = \frac{\bar{X}_\text{А} - \bar{X}_\text{Б}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{34{,}42 - 34{,}78}{1{,}0923} = -0{,}3296;$$

$$t_2 = \frac{\bar{X}_\text{А} - \bar{X}_\text{В}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{34{,}42 - 33{,}70}{1{,}0923} = 0{,}6592;$$

$$t_3 = \frac{\bar{X}_\text{А} - \bar{X}_\Gamma}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{34{,}42 - 28{,}38}{1{,}0923} = 5{,}5296;$$

$$t_4 = \frac{\bar{X}_\text{Б} - \bar{X}_\text{В}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{34{,}78 - 33{,}70}{1{,}0923} = 0{,}9887;$$

$$t_5 = \frac{\bar{X}_\text{Б} - \bar{X}_\Gamma}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{34{,}78 - 28{,}38}{1{,}0923} = 5{,}8592;$$

$$t_6 = \frac{\bar{X}_\text{В} - \bar{X}_\Gamma}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{33{,}70 - 28{,}38}{1{,}0923} = 4{,}8705.$$

Критичне вредности из таблице t – расподеле су следеће:

$$t_{0{,}05;16} = 2{,}120 \quad \text{и} \quad t_{0{,}01;16} = 2{,}921.$$

Како су количници t_3, t_5 и t_6 већи од критичних вредности из таблице t – расподеле, закључак је да постоји високо статистички значајна разлика у утицају између третмана А и Г, Б и Г, В и Г.

С друге стране, није установљена статистички значајна разлика између третмана А и Б, А и В, Б и В.

Тест најмање значајне разлике (НЗР тест)

$$H_0: \bar{X}_i = \bar{X}_j;$$

$$H_1: \bar{X}_i \neq \bar{X}_j.$$

Као што је било речи, најмање значајна разлика се израчунава на следећи начин:

$$NZR_{\alpha} = t_{N-k;\alpha} \times S_{(\bar{X}_i - \bar{X}_j)}.$$

Самим тим, у конкретном примеру важи:

$$NZR_{0,05} = 2{,}120 \times 1{,}0923 = 2{,}32;$$

$$NZR_{0,01} = 2{,}921 \times 1{,}0923 = 3{,}19.$$

Третмани	$\bar{X}_i$	$\bar{X}_i - \bar{X}_{\Gamma}$	$\bar{X}_i - \bar{X}_{\text{В}}$	$\bar{X}_i - \bar{X}_{\text{А}}$
Б	$\bar{X}_{\text{Б}} = 34.78$	$6{,}40^{**}$	1,08	0,36
А	$\bar{X}_{\text{А}} = 34{,}42$	$6{,}04^{**}$	0,72	
В	$\bar{X}_{\text{В}} = 33{,}70$	$5{,}32^{**}$		
Г	$\bar{X}_{\Gamma} = 28{,}38$			

Као и код теста парова третмана, може се констатовати да постоји високо статистички значајна разлика у утицају између третмана Г и осталих третмана.

Вишеструки интервални Данканов тест

$$H_0: \bar{X}_i = \bar{X}_j;$$

$$H_1: \bar{X}_i \neq \bar{X}_j$$

Оцењена стандардна грешка коју је потребно помножити са критичним вредностима из таблица вишеструког интервала:

$$S_{\bar{X}} = \sqrt{\frac{S_P^2}{n}} = \sqrt{\frac{2{,}98}{5}} = 0{,}7724.$$

Табличне вредности из таблице за вишеструки интервални тест *α=0,05*:

Интервал	**2**	**3**	**4**
Критична вредност	3,00	3,15	3,23
Најмање значајни интервал	3,00×0,7724=2,32	3,15×0,7724=2,43	3,23×0,7724=2,49

Табличне вредности из таблице за вишеструки интервални тест *α=0,01*:

Интервал	**2**	**3**	**4**
Критична вредност	4,13	4,34	4,45
Најмање значајни интервал	4,13×0,7724=3,19	4,34×0,7724=3,35	4,45×0,7724=3,44

Између аритметичких средина $\bar{X}_{\text{Б}}$ и $\bar{X}_{\Gamma}$ постоје три интервала. Самим тим, разлика између ове две аритметичке средине поредиће се са трећим по реду вредностима најмање значајним интервалима.

$$\bar{X}_{\text{Б}} - \bar{X}_{\Gamma} = 34{,}78 - 28{,}38 = 6{,}40 \ > 2{,}49(3{,}44).$$

Закључак је да постоји високо статистички значајна разлика између средина $\bar{X}_{\text{Б}}$ и $\bar{X}_{\Gamma}$.

Између средина $\bar{X}_{\text{Б}}$ и $\bar{X}_{\text{В}}$, $\bar{X}_{\text{А}}$ и $\bar{X}_{\Gamma}$ постоје два интервала. С тим у вези, разлика између ових аритметичких средина поредиће се са другим по реду вредностима најмање значајним интервалима.

$$\bar{X}_{\text{Б}} - \bar{X}_{\text{В}} = 34{,}78 - 33{,}70 = 1{,}08 \ < 2{,}43(3{,}35);$$

$$\bar{X}_{\text{А}} - \bar{X}_{\Gamma} = 34{,}42 - 28{,}38 = 6{,}04 \ > 2{,}43(3{,}35).$$

Закључак је да постоји високо статистички значајна разлика између средина $\bar{X}_{\text{А}}$ и $\bar{X}_{\Gamma}$, док између средина $\bar{X}_{\text{Б}}$ и $\bar{X}_{\text{В}}$ нема статистички значајне разлике.

Између средина $\bar{X}_{\text{Б}}$ и $\bar{X}_{\text{А}}$, $\bar{X}_{\text{А}}$ и $\bar{X}_{\text{В}}$, $\bar{X}_{\text{В}}$ и $\bar{X}_{\Gamma}$ постоји један интервал. С тим у вези, разлика између ових аритметичких средина поредиће се са првом по реду вредностима најмање значајним интервалима.

$$\bar{X}_{\text{Б}} - \bar{X}_{\text{А}} = 34{,}78 - 34{,}42 = 0{,}36 \ < 2{,}32(3{,}19);$$

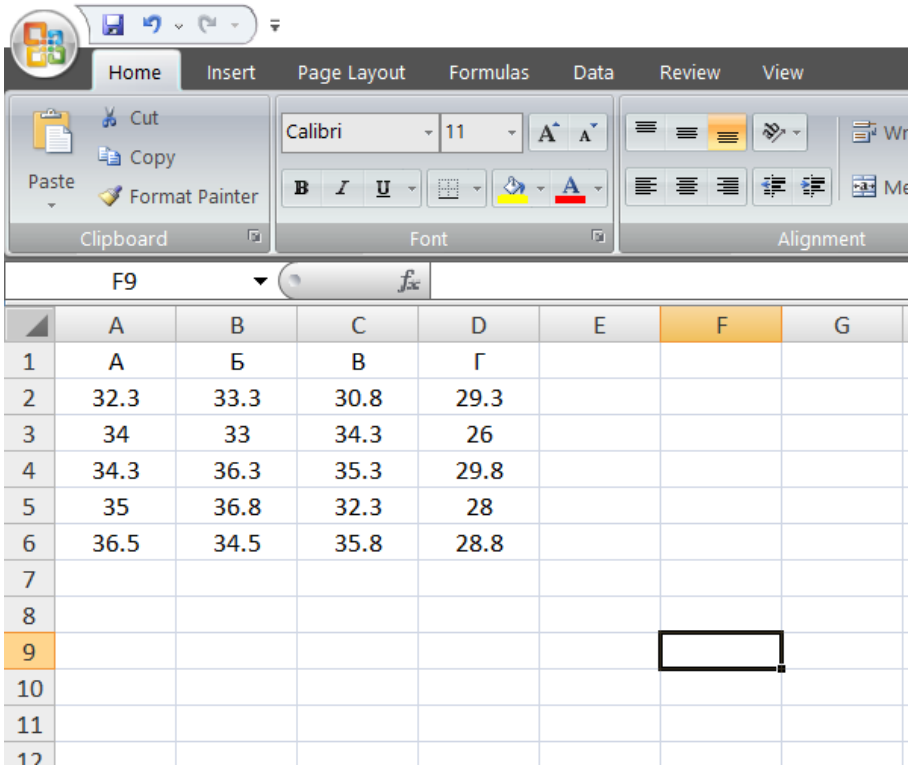
$$\bar{X}_{\text{А}} - \bar{X}_{\text{В}} = 34{,}42 - 33{,}70 = 0{,}72 \ < 2{,}32(3{,}19);$$

$$\bar{X}_{\text{В}} - \bar{X}_{\Gamma} = 33{,}70 - 28{,}38 = 5{,}32 \ > 2{,}32(3{,}19).$$

Закључак је да постоји високо статистички значајна разлика између средина $\bar{X}_{\text{В}}$ и $\bar{X}_{\Gamma}$, док између средина $\bar{X}_{\text{Б}}$ и $\bar{X}_{\text{А}}$, $\bar{X}_{\text{А}}$ и $\bar{X}_{\text{В}}$, нема статистички значајне разлике.

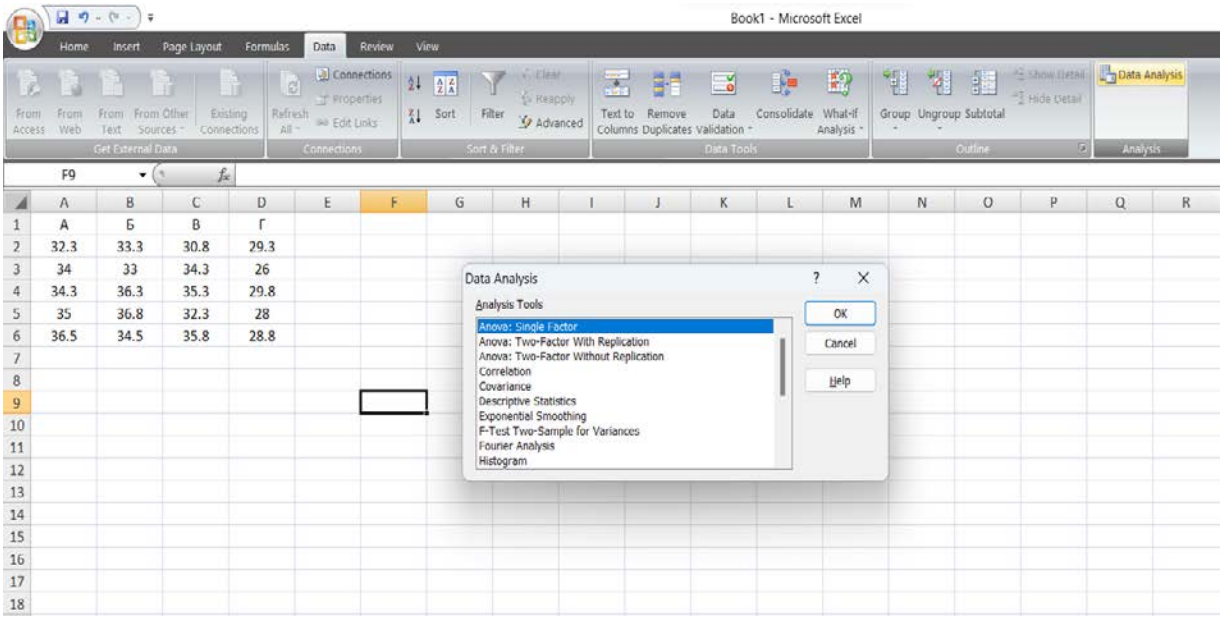
Вежба 9. Анализа варијансе применом Microsoft Excel-a

1. Полазни изглед табеле је следећи:

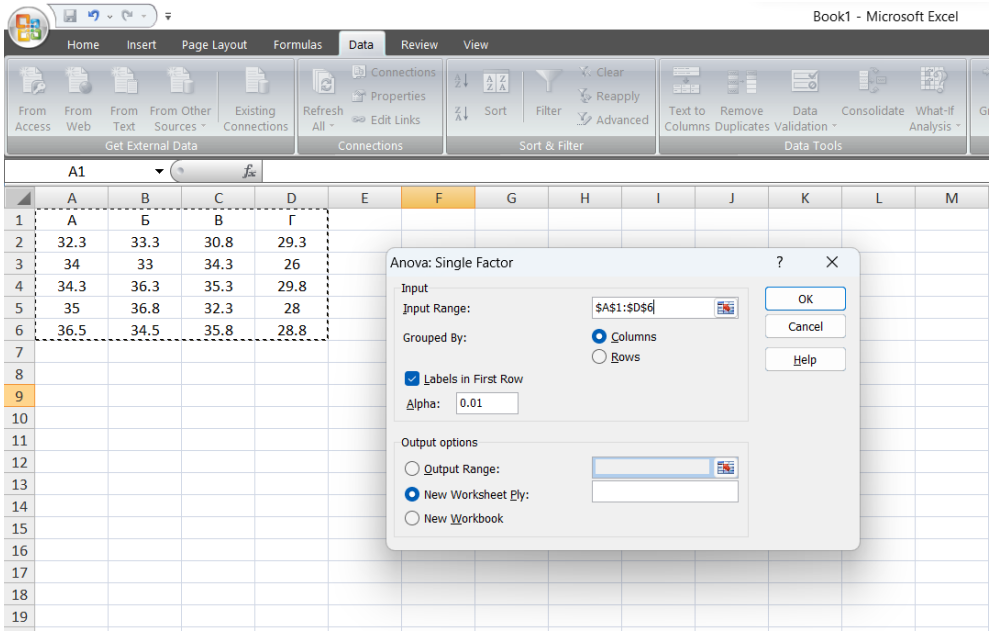


	A	B	C	D	E	F	G
1	А	Б	В	Г			
2	32.3	33.3	30.8	29.3			
3	34	33	34.3	26			
4	34.3	36.3	35.3	29.8			
5	35	36.8	32.3	28			
6	36.5	34.5	35.8	28.8			
7							
8							
9							
10							
11							

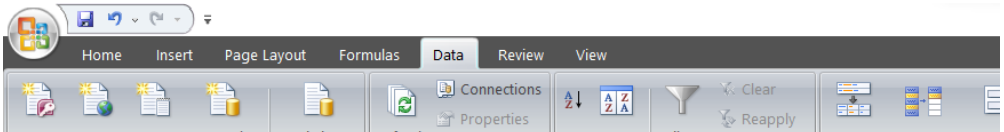
2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys* у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *Anova: Single Factor*, као што је показано у наставку:



3. Кликом на дугме *OK* отвара се нови прозор. У блоку који се односи на *Input*, у делу који се односи на *Input Range* потребно је обележити податке који су предмет анализе. Уколико се обухвати и заглавље, неоподно је штиклирати ставку *Labels in First Row*. У делу у којем пише *Alpha* неопходно је уписати жељени праг значајности *α*. У примеру биће изабран праг значајности *α=0,01*. Поглед на *Microsoft Excel* сада изгледа на следећи начин:



4. Кликом на дугме *ОК* добијају се резултати анализе варијансе у новом *Sheet*-у:



J18

	A	B	C	D	E	F	G	H	I
1	Anova: Single Factor								
2									
3	SUMMARY								
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*				
5	А	5	172.1	34.42	2.337				
6	Б	5	173.9	34.78	2.957				
7	В	5	168.5	33.7	4.425				
8	Г	5	141.9	28.38	2.212				
9									
10									
11	ANOVA								
12	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*		
13	Between Groups	134.448	3	44.816	15.02506077	6.48581E-05	5.292214052		
14	Within Groups	47.724	16	2.98275					
15									
16	Total	182.172	19						
17									
18									

Прва табела приказује основне информације у вези са подацима који су у оквиру различитих третмана предмет анализе. Друга табела представља табелу анализе варијансе, где ознака *SS* представља суму квадрата, *df* број степени слободе, а *MS* средину сума квадрата. Такође, у табели фигурира и *p*-вредност. Једини недостатак је тај што се не може истовремено проверити нулта хипотеза за два или више прагова значајности α.

Када је реч о *post hoc* тестовима, све тестове је могуће спровести пратећи претходно објашњене кораке у решењу разматраног примера. Поред наведеног, могуће је спровести појединачне тестове значајности разлике две средине на начин који је објашњен у вежби 6. Неслагање у односу на решење, постојаће код резултата *t*-количника (јер ће *Microsoft Excel* рачунати сваки пут оцењену стандардну грешку разлике две средине, за разлику од решења где фигурира заједника стандардна грешка разлике две средине за сваки *t*-тест), као и критичних вредности из таблице *t*–расподеле. Ипак, за очекивати је да се добију исти закључци као на основу помињаног теста парова третмана.

5.4. Блок систем

У претходном делу је објашњена анализа варијансе где постоји само један критеријум класификације јединица посматрања. У случају да постоји потреба да се у анализу уврсти дводимензионална класификација, неопходно је увести појам случајни *блок систем*.

Као пример може се узети неки пољски оглед са сортама, где се услед различитог квалитета земљишта појављује већа или мања неуједначеност у приносима између појединих парцела. У том случају, логично је присутне варијације настале услед хетерогености земљишта издвојити као посебну компоненту.

На тај начин, укупан варијабилитет је могуће поделити на варијабилитет који настаје услед дејства третмана, затим варијабилитет између неког познатог фактора што омогућава прецизније спровођење експеримента (варијабилитет између блокова) и случајан варијабилитет, односно варијабилитет погрешке.

Код случајног блок система неопходно је дефинисати две нулте хипотезе. Прва на основу које се претпоставља да нема статистички значајне разлике између аритметичких средина третмана и друга где се претпоставља да нема статистички значајне разлике између аритметичких средина по блоковима. Прихватањем нултих хипотеза, истраживање се завршава. Одбацивањем нулте и прихватањем алтернативне хипотезе за одговарајући праг значајности α, анализа се наставља где је потребно установити између којих тачно третмана односно блокова постоји статистички значајна разлика.

Приликом провере полазних претпоставки као и у претходном делу, неопходно је формирати радну табелу следећег изгледа:

Извори Варијације	**Степени слободе**	**Суме квадрата**	**Средина сума квадрата**	***F*-однос**	***F*-таблично**	
					α = 0,05	***α = 0,01***
Блокови	b-1	Q_b	S_b^2	S_b^2/ S_P^2	$F_{b-1;(b-1)(t-1)}$	$F_{b-1;(b-1)(t-1)}$
Третрмани	t-1	Q_T	S_T^2	S_T^2/S_P^2	$F_{t-1;(b-1)(t-1)}$	$F_{t-1;(b-1)(t-1)}$
Погрешка	(b-1)(t-1)	Q_P	S_P^2			
Тотал	N-1	Q				

Суме квадрата се у практичном раду израчунавају применом следећих радних формула:

$$Q = \sum_{i=1}^{t}\sum_{j=1}^{b} X_{ij}^2 - C,$$

где је C ознака за коректив фактор који се израчунава на следећи начин:

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{b} X_{ij}\right)^2}{N} = \frac{T^2}{N},$$

где је $\sum_{i=1}^{t}\sum_{j=1}^{b} X_{ij} = T$.

У практичном раду, сума квадрата третмана Q_T се издваја као наредна сума квадрата коју је потребно израчунати:

$$Q_T = \sum_{i=1}^{t} \frac{\left(\sum_{j=1}^{b} X_{ij}\right)^2}{b} - C = \sum_{i=1}^{t} \frac{T_j^2}{b} - C,$$

где је T_j сума *j*-тог третмана.

Сума квадрата блокова Q_b израчунава се на следећи начин:

$$Q_b = \sum_{j=1}^{b} \frac{\left(\sum_{j=1}^{t} X_{ij}\right)^2}{t} - C = \sum_{i=1}^{b} \frac{B_i^2}{t} - C.$$

где је B_i сума *i*-тог блока.

На основу израчунатих сума квадрата тотала, суме квадрата третмана и суме квадрата блокова, долази се до вредности суме квадрта погрешке:

$$Q_P = Q - (Q_T + Q_b)\,.$$

Средине сума квадрата, односно варијансе, израчунавају се као количник сума квадрата и одговарајућих степени слободе. Када је реч о варијанси третмана, потребно је ставити у однос суму квадрата третмана и одговарајући број степени слободе третмана:

$$S_T^2 = \frac{Q_T}{t-1}.$$

Средина сума квадрата блокова израчунава се на сличан начин:

$$S_b^2 = \frac{Q_b}{b-1}$$

На крају, варијанса погрешке која подразумева количник између суме квадрата погрешке и броја степени слободе погрешке једнака је:

$$S_P^2 = \frac{Q_P}{(t-1)(b-1)}.$$

За проверу полазне претпоставке израчунава се *F*-однос који представља количник израчунатих варијанси. Провера нулте хипотезе којом се претпоставља једнакост аритметичких средина по третманима, заснива се на F-количнику који се рачуна на следећи начин:

$$F = \frac{S_T^2}{S_P^2}.$$

F-количник на основу којег се проверава нулта хипотеза којом се претпоставља једнакост аритметичких средина по блоковима израчунава се на следећи начин:

$$F = \frac{S_b^2}{S_P^2}.$$

У циљу доношења закључка о полазној претпоставци, F-количник се упоређује са критичним вредностима из таблица Фишерове дистрибуције, које се очитавају за праг значајности α и степене слободе r_1 и r_2. Уколико је вредност израчунатог F-количника мања од критичних вредности из таблица Фишерове расподеле, нулта хипотеза се може прихватити као тачна. Прихватање полазне хипотезе указује на то да између примењених третмана, односно блокова, не постоје статистички значајне разлике у дејству на експерименталне јединице и тиме се анализа варијансе завршава.

С друге стране, уколико је вредност израчунатог количника F већа од критичних вредности из таблица Фишерове расподеле, одбацује се нулта и прихвата алтернативна хипотеза као тачна. Ако се полазна хипотеза не прихвати, већ се утврди постојање значајних или врло значајних разлика између било која два примењена третмана или блока, анализа варијансе се даље наставља, како би се утврдило између којих третмана односно блокова постоје статистички значајне разлике. Приликом тестирања разлика између средина третмана или блокова користе се исти *post hoc* тестови који су представљени у делу 5.3, с тим да је неопходно имати на уму да ће сада број степени слободе погрешке бити $(t-1)(b-1)$ уместо $N-k$.

Пример 22:

У експерименту се испитује ефекат стандардног (T0) и три алтернативна система узгајања (T1, T2, T3) на укупну суву материју кеља. Експеримент је изведен по плану случајног блок система. У сваком од пет блокова сваки је третман примењен само једанпут на случајно одабрану парцелу.

Принос (т/ха) уређен по третманима дат је у табели:

Блокови	**T0**	**T1**	**T2**	**T3**	$\boldsymbol{B_j}$	$\boldsymbol{\bar{X}_j}$
1	7,2	5,4	6,9	5,5	25,0	6,25
2	5,8	4,1	5,3	5,6	20,8	5,20
3	6,9	5,3	6,6	4,5	23,3	5,83
4	7,0	5,0	7,2	6,1	25,3	6,33
5	5,8	4,8	6,2	5,7	22,5	5,63
$\boldsymbol{T_i}$	32,7	24,6	32,2	27,4	116,9	-
$\boldsymbol{\bar{X}_i}$	6,54	4,92	6,44	5,48	-	-

$H_0: \bar{X}_{T0} = \bar{X}_{T1} = \bar{X}_{T2} = \bar{X}_{T3}$;

$H_0: \bar{X}_{B1} = \bar{X}_{B2} = \bar{X}_{B3} = \bar{X}_{B4} = \bar{X}_{B5}$.

У циљу провере полазних претпоставки неопходно је формирати радну табелу:

Извори Варијације	**Степени слободе**	**Суме квадрата**	**Средина сума квадрата**	***F*-однос**	***F*-таблично**	
					α = 0,05	***α = 0,01***
Блокови	b-1 = 5-1=4	Q_b=3,4370	S_b^2=0,8592	S_b^2/ S_P^2 = 0,8592/0,2569= =3,34	3,26	5,41
Третрмани	t-1=4-1=3	Q_T=9,1295	S_T^2=3,0432	S_T^2/S_P^2= 3,0432/0,2569= =11,84	3,49	5,95
Погрешка	(b-1)(t-1)= 4×3=12	Q_P=3,0830	S_P^2=0,2569			
Тотал	N-1=20-1=19	Q=15,6495				

Рачун за укупну суму квадрата представљен је у наставку:

$$Q = \sum_{i=1}^{t}\sum_{j=1}^{b} X_{ij}^2 - C = 7{,}2^2 + 5{,}8^2 + 6{,}9^2 + \cdots + 5{,}7^2 - 683{,}2805 = 15{,}6495$$

$$C = \frac{\left(\sum_{i=1}^{t}\sum_{j=1}^{b} X_{ij}\right)^2}{N} = \frac{T^2}{N} = \frac{(32{,}7 + 24{,}6 + 32{,}2 + 27{,}4)^2}{20} = 683{,}2805$$

Сума квадрата третмана Q_T:

$$Q_T = \sum_{i=1}^{t}\frac{\left(\sum_{j=1}^{b} X_{ij}\right)^2}{b} - C = \sum_{i=1}^{t}\frac{T_j^2}{b} - C = \frac{32{,}7^2 + 24{,}6^2 + 32{,}2^2 + 27{,}4^2}{5} - 683{,}2805 = 9{,}1295$$

Сума квадрата блокова Q_b израчунава се на следећи начин:

$$Q_b = \sum_{j=1}^{b}\frac{\left(\sum_{j=1}^{t} X_{ij}\right)^2}{t} - C = \sum_{i=1}^{b}\frac{B_i^2}{t} - C = \frac{25{,}0^2 + 20{,}8^2 + 23{,}3^2 + 25{,}3^2 + 22{,}5^2}{4} - 683{,}2805 = 3{,}4370$$

На основу израчунатих сума квадрата тотала, суме квадрата третмана и суме квадрата блокова, долази се до вредности суме квадрта погрешке:

$$Q_P = Q - (Q_T + Q_b) = 15{,}6495 - (9{,}1295 + 3{,}4370) = 3{,}0830$$

Средине сума квадрата:

$$S_T^2 = \frac{Q_T}{t-1} = \frac{9{,}1295}{3} = 3{,}0432;$$

$$S_b^2 = \frac{Q_b}{b-1} = \frac{3{,}4370}{4} = 0{,}8592;$$

$$S_P^2 = \frac{Q_P}{(t-1)(b-1)} = \frac{3{,}0830}{12} = 0{,}2569.$$

Како би се проверила прва полазна претпоставка израчунава се *F*-однос који представља количник претходно израчунатих варијанси:

$$F = \frac{S_T^2}{S_P^2} = \frac{3{,}0432}{0{,}2569} = 11{,}84.$$

Критичне вредности из таблице Фишерове таблице за $r_1 = t - 1 = 4 - 1 = 3$ и $r_2 = (t - 1)(b\text{-}1) = 3 \times 4 = 12$ степени слободе. С тим у вези, у конкретном примеру важи:

$$F_{3;12(0{,}05)} = 3{,}49;$$

$$F_{3;12(0{,}01)} = 5{,}95.$$

Како је *F*-количник већи од обе критичне вредности, може се донети закључак да се нулта хипотеза одбацује и прихвата се алтернативна хипотеза која гласи да постоји статистички значајна разлика у утицају између посматраних третмана.

С обзиром на то, да алтернативна хипотеза указује само на то да постоји статистички значајна разлика између посматраних аритметичких средина, додатним тестирањем је неопходно установити између којих конкретно аритметичких средина постоји статистички значајна разлика.

Поред наведеног, како би се проверила друга нулта хипотеза израчунава се *F*-однос који представља количник претходно израчунатих варијанси:

$$F = \frac{S_B^2}{S_P^2} = \frac{0{,}8592}{0{,}2569} = 3{,}34.$$

Критичне вредности из таблице Фишерове таблице за $r_1 = b - 1 = 5 - 1 = 4$ и $r_2 = (t - 1)(b\text{-}1) = 3 \times 4 = 12$ степени слободе. С тим у вези, у конкретном примеру важи:

$$F_{4;12(0{,}05)} = 3{,}26;$$

$$F_{4;12(0{,}01)} = 5{,}41.$$

Како је *F*-количник већи само од критичне вредности за праг значајности α, може се донети закључак да се нулта хипотеза одбацује за наведени прга значајности. С тим у вези, има смисла даље испитати између којих конкретно блокова постоји статистички значајна разлика.

Тест парова тертмана (t-тест)

$$H_0: \bar{X}_i = \bar{X}_j;$$

$$H_1: \bar{X}_i \neq \bar{X}_j$$

Полазна претпоставка се проверава израчунавањем *t* – количника:

$$t = \frac{\bar{X}_i - \bar{X}_j}{S_{(\bar{X}_i - \bar{X}_j)}},$$

$$S_{(\bar{X}_i - \bar{X}_j)} = \sqrt{\frac{2 \times S_P^2}{b}} = \sqrt{\frac{2 \times 0{,}2569}{5}} = 0{,}3206.$$

$$\bar{X}_{TO} = 6{,}54; \bar{X}_{T1} = 4{,}92;\ \bar{X}_{T2} = 6{,}44;\ \bar{X}_{T3} = 5{,}48.$$

$$t_1 = \frac{\bar{X}_{TO} - \bar{X}_{T1}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{6{,}54 - 4{,}92}{0{,}3206} = 5{,}0530;$$

$$t_2 = \frac{\bar{X}_{TO} - \bar{X}_{T2}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{6{,}54 - 6{,}44}{0{,}3206} = 0{,}3119;$$

$$t_3 = \frac{\bar{X}_{T0} - \bar{X}_{T3}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{6{,}54 - 5{,}48}{0{,}3206} = 3{,}3063;$$

$$t_4 = \frac{\bar{X}_{T1} - \bar{X}_{T2}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{4{,}92 - 6{,}44}{0{,}3206} = 2{,}3830;$$

$$t_5 = \frac{\bar{X}_{T1} - \bar{X}_{T3}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{4{,}92 - 5{,}48}{0{,}3206} = 2{,}8004;$$

$$t_6 = \frac{\bar{X}_{T2} - \bar{X}_{T3}}{S_{(\bar{X}_i - \bar{X}_j)}} = \frac{6{,}44 - 5{,}48}{0{,}3206} = 2{,}9944.$$

Критичне вредности из таблице *t* – расподеле су следеће:

$$t_{0{,}05;12} = 2{,}179 \quad \text{и} \quad t_{0{,}01;12} = 3{,}055.$$

Како су количници t_1, t_3 и већи од обе критичне вредности из таблице *t* – расподеле, закључак је да постоји високо статистички значајна разлика у утицају између третмана Т0 и Т1, као и Т0 и Т3. Такође забележена је статистички значајна разлика (за праг значајности *α=0,05*) између третмана Т1 и Т2, Т1 и Т3, као и Т2 и Т3.

Статистички значајна разлика не ппостоји само између третмана Т0 и Т2.

Тест најмање значајне разлике (НЗР тест)

$$H_0: \bar{X}_i = \bar{X}_j;$$

$$H_1: \bar{X}_i \neq \bar{X}_j$$

Као што је било речи, најмање значајна разлика се израчунава на следећи начин:

$$NZR_{\alpha} = t_{(b-1)(t-1);\alpha} \times S_{(\bar{X}_i - \bar{X}_j)}.$$

Самим тим, у конкретном примеру важи:

$$NZR_{0,05} = 2{,}179 \times 0{,}3206 = 0{,}6986;$$

$$NZR_{0,01} = 3{,}055 \times 0{,}3206 = 0{,}9794.$$

Третмани	$\bar{X}_i$	$\bar{X}_i - \bar{X}_{T1}$	$\bar{X}_i - \bar{X}_{T3}$	$\bar{X}_i - \bar{X}_{T2}$
T0	$\bar{X}_{TO} = 6{,}54$	$1{,}62^{**}$	$1{,}06^{**}$	0,10
T2	$\bar{X}_{T2} = 6{,}44$	$1{,}52^{**}$	$0{,}96^{*}$	
T3	$\bar{X}_{T3} = 5{,}48$	0,56		
T1	$\bar{X}_{T1} = 4{,}92$			

На основу НЗР теста може се извести мало другачији закључак у односу на тест парова третмана. Установљена је високо статистички значајна разлика између третмана T0 и T1, T0 и T3, као и T2 и T1. Статистички значајна разлика је установљена између T2 и T3, док између T0 и T2, као и T3 и T1 није забележена разлика.

Вишеструки интервални Данканов тест

$$H_0: \bar{X}_i = \bar{X}_j;$$

$$H_1: \bar{X}_i \neq \bar{X}_j$$

Оцењена стандардна грешка коју је потребно помножити са критичним вредностима из таблица вишеструког интервала:

$$S_{\bar{X}} = \sqrt{\frac{S_P^2}{b}} = \sqrt{\frac{0{,}2569}{5}} = 0{,}2268.$$

Табличне вредности из таблице за вишеструки интервални тест *α=0,05* и $(b-1)(t-1)$ број степени слободе:

Интервал	**2**	**3**	**4**
Критична вредност	3,08	3,23	3,33
Најмање значајни интервал	3,08×0,2268=0,70	3,23×0,2268=0,73	3,33×0,2268=0,75

Табличне вредности из таблице за вишеструки интервални тест *α=0,01* и $(b-1)(t-1)$ број степени слободе:

Интервал	**2**	**3**	**4**
Критична вредност	4,32	4,55	4,68
Најмање значајни интервал	4,32×0,2268=0,98	4,55×0,2268=1,03	4,68×0,2268=1,06

Између аритметичких средина $\bar{X}_{T0}$ и $\bar{X}_{T1}$ постоје три интервала. Самим тим, разлика између ове две аритметичке средине поредиће се са трећим по реду вредностима најмање значајним интервалима.

$$\bar{X}_{T0} - \bar{X}_{T1} = 6{,}54 - 4{,}92 = 1{,}62 > 0{,}75(1{,}06).$$

Закључак је да постоји високо статистички значајна разлика између средина $\bar{X}_{T0}$ и $\bar{X}_{T1}$.

Између средина $\bar{X}_{T0}$ и $\bar{X}_{T3}$, $\bar{X}_{T2}$ и $\bar{X}_{T1}$ постоје два интервала. С тим у вези, разлика између ових аритметичких средина поредиће се са другим по реду вредностима најмање значајним интервалима.

$$\bar{X}_{T0} - \bar{X}_{T3} = 6{,}54 - 5{,}48 = 1{,}06 > 0{,}73(1{,}03);$$

$$\bar{X}_{T2} - \bar{X}_{T1} = 6{,}44 - 4{,}92 = 1{,}52 > 0{,}73(1{,}03).$$

Закључак је да постоји високо статистички значајна разлика између средина $\bar{X}_{T0}$ и $\bar{X}_{T3}$, као и $\bar{X}_{T2}$ и $\bar{X}_{T1}$.

Између средина $\bar{X}_{T0}$ и $\bar{X}_{T2}$, $\bar{X}_{T2}$ и $\bar{X}_{T3}$, $\bar{X}_{T3}$ и $\bar{X}_{T1}$ постоји један интервал. С тим у вези, разлика између ових аритметичких средина поредиће се са првом по реду вредностима најмање значајним интервалима.

$$\bar{X}_{T0} - \bar{X}_{T2} = 6{,}54 - 6{,}44 = 0{,}10 < 0{,}70(0{,}98);$$

$$\bar{X}_{T2} - \bar{X}_{T3} = 6{,}44 - 5{,}48 = 0{,}96 > 0{,}70;$$

$$\bar{X}_{T2} - \bar{X}_{T3} = 6{,}44 - 5{,}48 = 0{,}96 < 0{,}98;$$

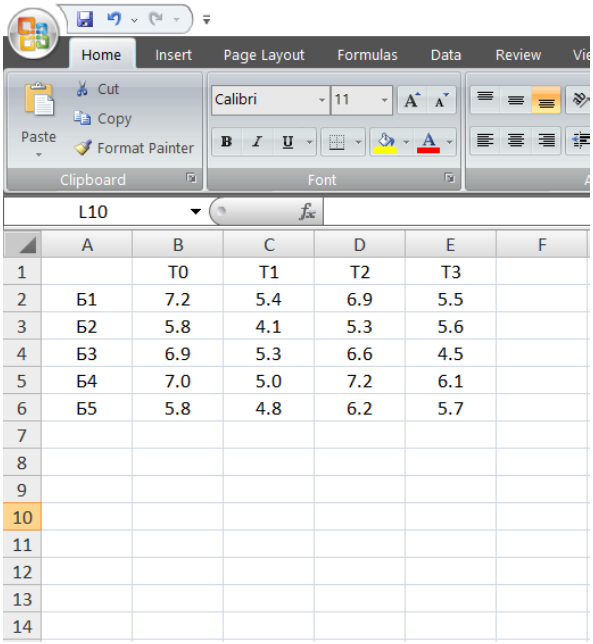
$$\bar{X}_{T3} - \bar{X}_{T1} = 5{,}48 - 4{,}92 = 0{,}56 < 0{,}70(0{,}98).$$

Закључак је да постоји статистички значајна разлика између средина $\bar{X}_{T2}$ и $\bar{X}_{T3}$, док између средина $\bar{X}_{T0}$ и $\bar{X}_{T2}$, $\bar{X}_{T3}$ и $\bar{X}_{T1}$, нема статистички значајне разлике.

На исти начин је могуће тестирати једнакост аритметичких средина блокова ($\bar{X}_{B1} = 6{,}25; \bar{X}_{B2} = 5{,}20;\ \bar{X}_{B3} = 5{,}83;\ \bar{X}_{B4} = 6{,}33;\ \bar{X}_{B5} = 5{,}63$), с тим да је неопходно имати на уму да ће сада стандардна грешка разлике аритметичких средина бити $S_{(\bar{X}_i - \bar{X}_j)} = \sqrt{\frac{2 \times S_P^2}{t}} = \sqrt{\frac{2 \times 0{,}2569}{4}} = 0{,}3584$, док ће стандардна грешка аритметичке средине бити $S_{\bar{X}} = \sqrt{\frac{S_P^2}{t}} = 0{,}2534.$

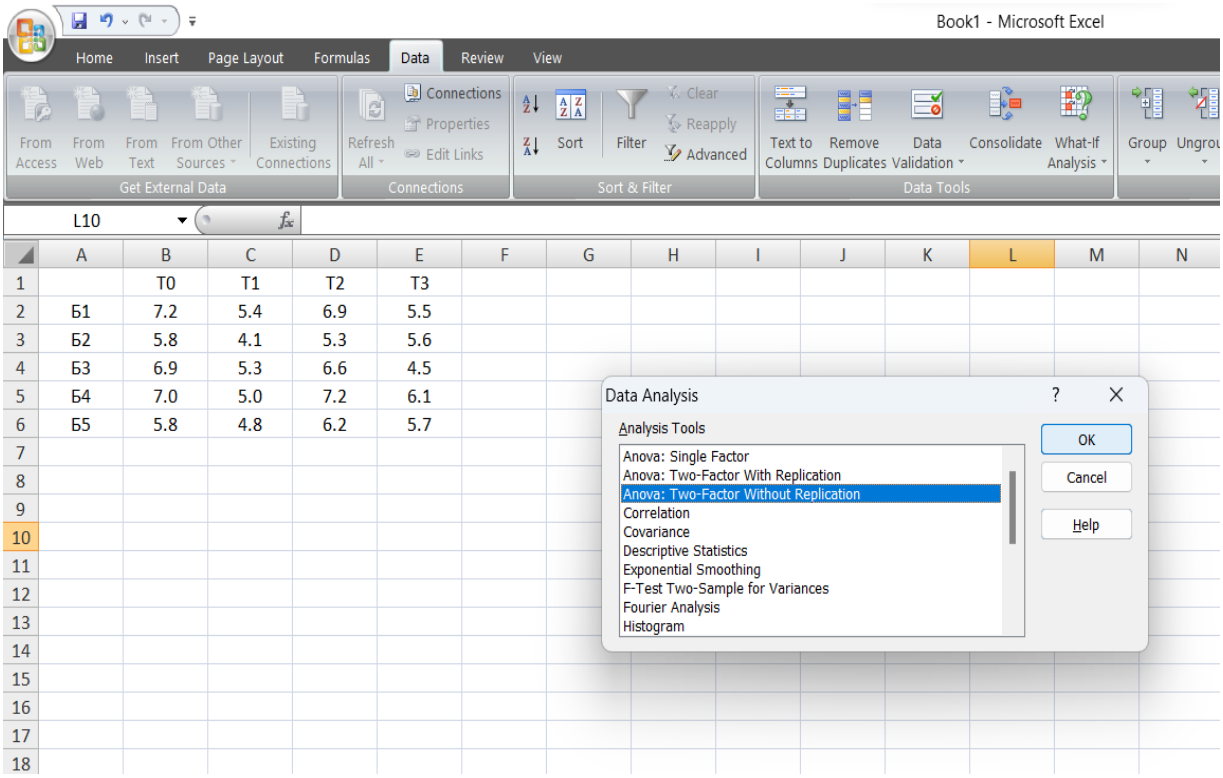
Вежба 10. Анализа варијансе (блок систем) применом Microsoft Excel-а

1. Полазни изглед табеле је следећи:



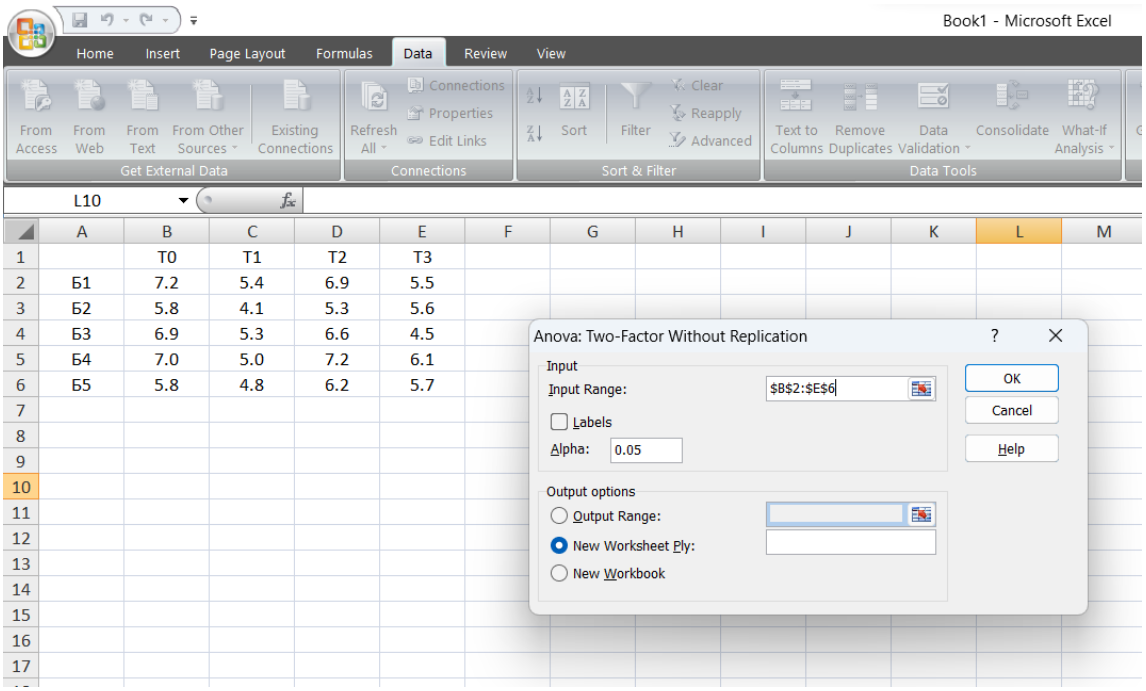
	A	B	C	D	E	F
1		T0	T1	T2	T3	
2	Б1	7.2	5.4	6.9	5.5	
3	Б2	5.8	4.1	5.3	5.6	
4	Б3	6.9	5.3	6.6	4.5	
5	Б4	7.0	5.0	7.2	6.1	
6	Б5	5.8	4.8	6.2	5.7	
7						
8						
9						
10						
11						
12						
13						
14						

2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys* у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *Anova: Two-Factor Without Replication*, као што је показано у наставку:



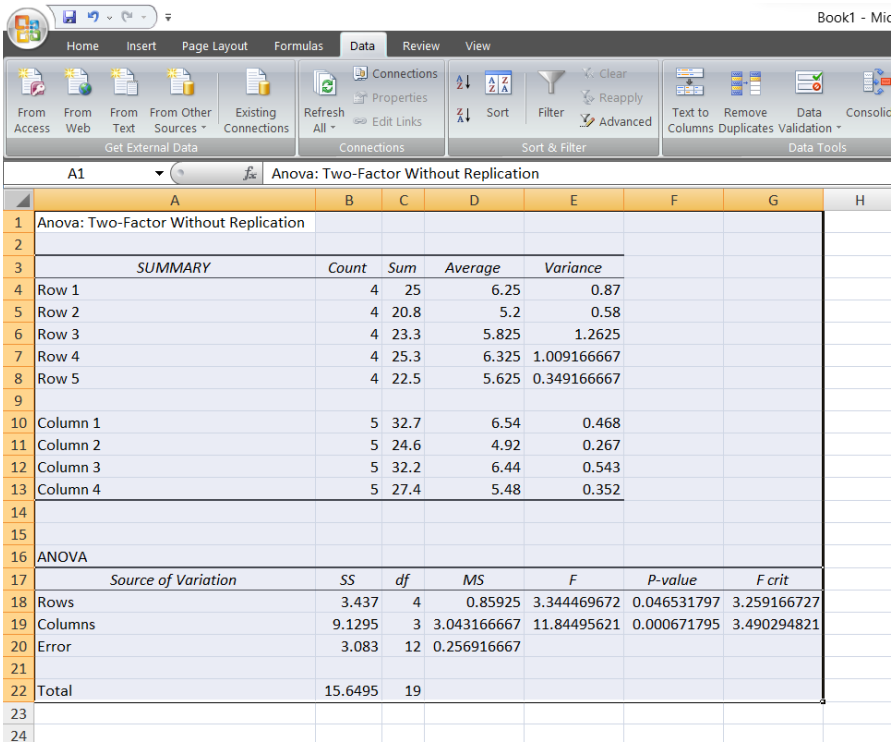
	A	B	C	D	E
1		T0	T1	T2	T3
2	Б1	7.2	5.4	6.9	5.5
3	Б2	5.8	4.1	5.3	5.6
4	Б3	6.9	5.3	6.6	4.5
5	Б4	7.0	5.0	7.2	6.1
6	Б5	5.8	4.8	6.2	5.7

3. Кликом на дугме *OK* отвара се нови прозор. У блоку који се односи на *Input*, у делу који се односи на *Input Range* потребно је обележити податке који су предмет анализе. Уколико се обухвати и заглавље, неопходно је штиклирати ставку *Labels*. У делу у којем пише *Alpha* неопходно је уписати жељени праг значајности α. У примеру биће изабран праг значајности $\alpha=0,05$. Поглед на *Microsoft Excel* сада изгледа на следећи начин:



	A	B	C	D	E
1		T0	T1	T2	T3
2	Б1	7.2	5.4	6.9	5.5
3	Б2	5.8	4.1	5.3	5.6
4	Б3	6.9	5.3	6.6	4.5
5	Б4	7.0	5.0	7.2	6.1
6	Б5	5.8	4.8	6.2	5.7

4. Кликом на дугме *OK* добијају се резултати анализе варијансе у новом *Sheet*-у:



	A	B	C	D	E	F	G	H
1	Anova: Two-Factor Without Replication							
2								
3	*SUMMARY*	*Count*	*Sum*	*Average*	*Variance*			
4	Row 1	4	25	6.25	0.87			
5	Row 2	4	20.8	5.2	0.58			
6	Row 3	4	23.3	5.825	1.2625			
7	Row 4	4	25.3	6.325	1.009166667			
8	Row 5	4	22.5	5.625	0.349166667			
9								
10	Column 1	5	32.7	6.54	0.468			
11	Column 2	5	24.6	4.92	0.267			
12	Column 3	5	32.2	6.44	0.543			
13	Column 4	5	27.4	5.48	0.352			
14								
15								
16	ANOVA							
17	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
18	Rows	3.437	4	0.85925	3.344469672	0.046531797	3.259166727	
19	Columns	9.1295	3	3.043166667	11.84495621	0.000671795	3.490294821	
20	Error	3.083	12	0.256916667				
21								
22	Total	15.6495	19					
23								
24								

Прва табела приказује основне информације у вези са подацима који су предмет анализе. Друга табела представља табелу анализе варијансе, која је готово иста као што је објашњено у претходном делу. Једини додатак јесте присуство *p*-вредности, док је највећи недостатак тај што се не могу истовремено проверити нулте хипотезе за два или више прагова значајности α.

Када је реч о *post hoc* тестовима, све тестове је могуће спровести пратећи претходно објашњене кораке у решењу разматраног примера. Поред наведеног, могуће је спровести појединачне тестове значајности разлике две средине на начин који је објашњен у вежби 6. Неслагање у односу на решење, постојаће код резултата *t*-количника (јер ће *Microsoft Excel* рачунати сваки пут оцењену стандардну грешку разлике две средине, за разлику од решења где фигурира заједника стандардна грешка разлике две средине за сваки *t*-тест), као и критичних вредности из таблице *t*–расподеле. Ипак, за очекивати је да се добију исти закључци као на основу помињаног теста парова третмана.

5.5. χ^2-тест

Непараметарска статистика се односи на статистичке методе које се примењују на податке за које се не претпоставља одређена дистрибуција или параметарски модел. Укључују методе које су мање осетљиве на претпоставке о облику дистрибуције података или карактеристикама популације.

χ^2-тест (хи-квадрат) тест је непараметарски тест који који се примењује када се испитује у којој мери се фреквенције појединих група са квалитативним карактеристикама подударају са теоријским фреквенцијама сваке групе.

5.5.1. Тестирање нулте хипотезе о подударности емпиријских и теоријских фреквенција

Уколико се са f_i означе емпиријске фреквенције појединих модалитета обележја, а са F_i теоријске (очекиване) фреквенције где је $(i = 1, \dots, k)$, нулта и алтернативна хипотеза су:

$$H_0: f_i = F_i;$$
$$H_1: f_i \neq F_i.$$

Тест критеријум који се примењује у тестирању је: $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - F_i)^2}{F_i}$, који претпостављајући нулту хипотезу има χ^2_{k-1} диструбуцију.

Пример 23:

Дати су подаци о броју случајева једне болести на 5 фарми. Тестирати нулту хипотезу о једнакој заступљености обољења на посматраним фармама. Како се претпоставља да је заступљеност обољења једнака на свим фармама, све очекиване фреквенције су једнаке

и износе $F_i = {}^{26}/_{5} = 5{,}2, (i = 1, \dots, 5)$. Израчунавање тест критеријума је приказано у радној табели:

Фарме	$\boldsymbol{f_i}$	$\boldsymbol{F_i}$	$\frac{(\boldsymbol{f_i} - \boldsymbol{F_i})^2}{\boldsymbol{F_i}}$
1	2	26/5	$(2\text{-}5{,}2)^2/5{,}2$
2	10	26/5	$(10\text{-}5{,}2)^2/5{,}2$
3	8	26/5	$(8\text{-}5{,}2)^2/5{,}2$
4	2	26/5	$(2\text{-}5{,}2)^2/5{,}2$
5	4	26/5	$(4\text{-}5{,}2)^2/5{,}2$
Сума	**26**	**26**	**10,16**

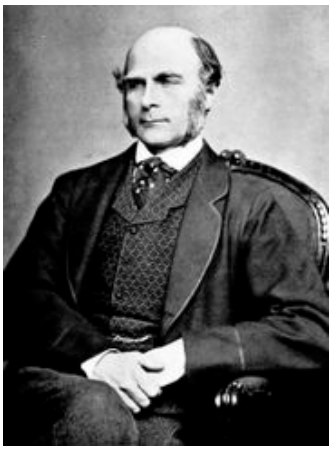
Израчуната вредност тест критеријума је $\chi^2 = 10{,}16$ и већа је од критичне вредности $\chi^2_{4,0,05} = 9{,}49$, тако да се на прагу значајности $\alpha = 0{,}05$ нулта хипотеза одбацује. Како је на прагу значајности $\alpha = 0{,}01$ израчуната вредност мања од критичне ($\chi^2_{4,0,05} = 13{,}28,$) нулта хипотеза се прихвата. Према томе може да се закључи да постоји статистички значајна, али не и високо значајна разлика између емпиријских и очекиваних фреквенција обољења на посматраним фармама.

Контролна питања

1. Дефинисати нулту и алтернативну хипотезу.
2. Дефинисати грешку типа I и грешку типа II.
3. Навести фазе тестирања.
4. Навести основне тестове аритметичких средина.
5. Навести основне тестове пропорција.
6. Када се приликом тестирања изводи Z тест?
7. Када се приликом тестирања изводи t-тест?
8. Када се примењује анализа варијансе?
9. Који тест се изводи у основи анализе варијансе? Која хипотеза се овим тестом проверава?
10. Када се примењује блок систем?

6. Регресиона и корелациона анализа

Концепт корелације и регресије увео је енглески антрополог, географ, генетичар, психометричар и статистичар Френсис Галтон 1888. године.



Sir Francis Galton
(1822–1911)

Значајно место у методама статистичке анализе припада испитивању утицаја и зависности између променљивих. Анализа може да се односи на две или више променљивих за које се зна или претпоставља да су у међусобној вези. На основу емпиријских података могуће је међузависности променљивих исказати математичком функцијом која ће исказати просечну или тзв. очекивану зависност (везу). Ако се ради о две променљиве од којих је једна зависна *(Y)*, а друга независна *(X)*, релација ових променљивих се може исказати функцијом: $Y = f(X)$.

Задатак регресионе анализе је да открије функционални облик (регресиони модел), коме се највише приближава квантитативно слагање варијација посматраних појава. Другим речима, циљ је утврдити како се зависно променљива мења у односу на независне променљиве и на основу степена слагања њихових варијација омогући оцену и предвиђање понашања зависне променљиве. Регресиона анализа може се дефинисати и као оцена вредности зависно променљиве на основу једне или више независних променљивих.

У поступку примене регресионе анализе могу се разликовати три фазе и то: *планирање*, *техника израчунавања параметара* (*развитак модела*) и *провера модела*. Фаза планирања подразумева јасно дефинисање циља истраживања и дефинисање променљивих које треба укључити у модел. Како би се јасно дефинисао циљ истраживања, потребна је анализа претходних истраживања из посматране области, као и дискусије са компетентним лицима која су се бавила истраживањима из посматране области. Друго важно питање у фази планирања јесте питање избора променљивих које треба укључити у анализу. То подразумева спецификацију зависно и независно променљивих као и одређивање њиховог броја.

Након дефинисања зависно променљиве и независно променљивих, приступа се избору модела. Избор модела одређен је пре свега циљем истраживања, али и самим подацима на којима се заснива анализа. Изабрани модел треба да што боље прикаже понашање зависно променљиве појаве у зависности од посматраних чинилаца, односно од одабраних независно променљивих. Такође, потребно је да модел буде основа на којој ће се моћи предвидети промене зависно променљиве. Један јединствен модел не може увек да задовољи све захтеве па се у неком испитивању користи више могућих модела.

Спецификација модела подразумева математичку формулацију утицаја и веза одабраних независно променљивих на зависно променљиву појаву. Теорија области примене и статистичка теорија могу сугерисати одређени облик математичке зависности међу посматраним променљивим. Као критеријуми у избору адекватног модела користе се ранија искуства из анализиране области, резултати оцењеног модела, односно његова прилагођеност подацима, као и тежња да модел буде што једноставнији.

У даљем излагању ће бити разматрана регресија са једном независном променљивом, односно проста линеарна регресија.

У циљу сагледавања међузависности између променљивих, потребно је располагати паровима променљивих измереним на n јединица случајног узорка, као што је представљено у наставку:

$X_i: X_1, X_2, X_3, \dots, X_i, \dots, X_n;$

$Y_i: Y_1, Y_2, Y_3, \dots, Y_i, \dots, Y_n, \qquad i = 1,2,3, \dots, n.$

Уколико независно променљива X условљава величину зависно променљиве Y, тада се ради о *регресији*. Ако се испитује међузависност две променљиве, тада се ради о *корелацији*.

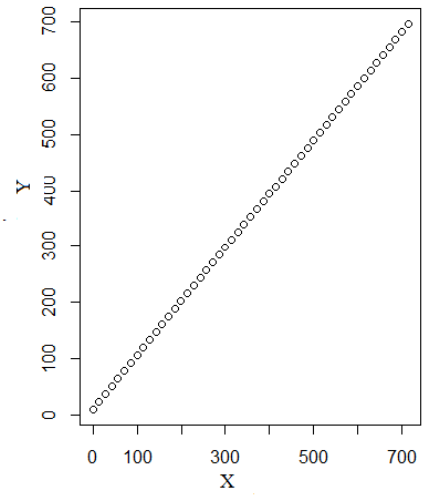
Циљ регресионе анализе је да омогући сагледавање очекиване вредности зависно променљиве на основу дате независно променљиве. Регресија се сагледава на основу *једначине регресије* и *стандардне грешке регресије*.

Циљ корелационе анализе је сагледавање јачине везе између две променљиве. Корелација се сагледава на основу *коефицијента корелације* и *коефицијента детерминације*.

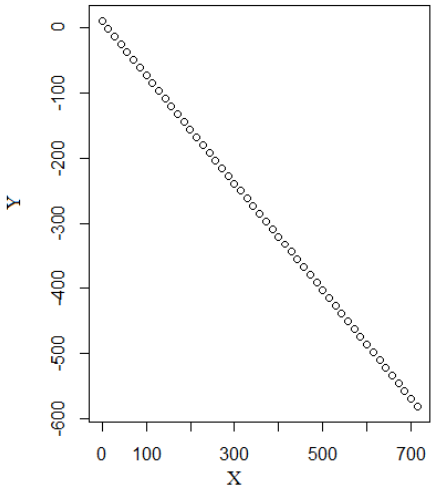
Регресиону и корелациону анализу корисно је започети анализом *дијаграма растурања*. Дијаграм растурања се формира у правоуглом координатном систему, где се на апсцисну осу наносе вредности независно променљиве X, а на ординатну осу вредности зависно променљиве Y. На дијаграм се уносе тачке са координатима (X_i, Y_i), $i = 1,2,3,\dots,n$. Ове тачке могу бити распоређене (расуте) према одређеној законитости. Дијаграм растурања садржи онолико тачака колико је заступљено парова вредности променљивих. Дијаграм растурања омогућава утврђивање зависности или везе између

променљивих, као и сагледавање карактера те везе која може бити линеарна или криволинијска.

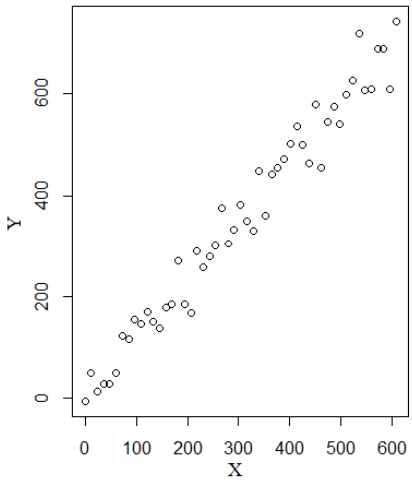
Линеарна веза може бити позитивна или негативна. Код позитивне везе, растом независно променљиве *X*, расте и зависно променива *Y*. С друге стране, код негативне везе, зависно променљива *Y* опада како расте независно променљива *X*. У наставку су представљени различити облици дијаграма растурања.



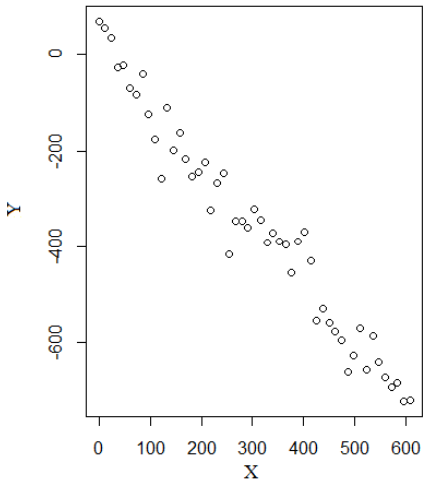
Funkcionalna pozitivna linearna veza



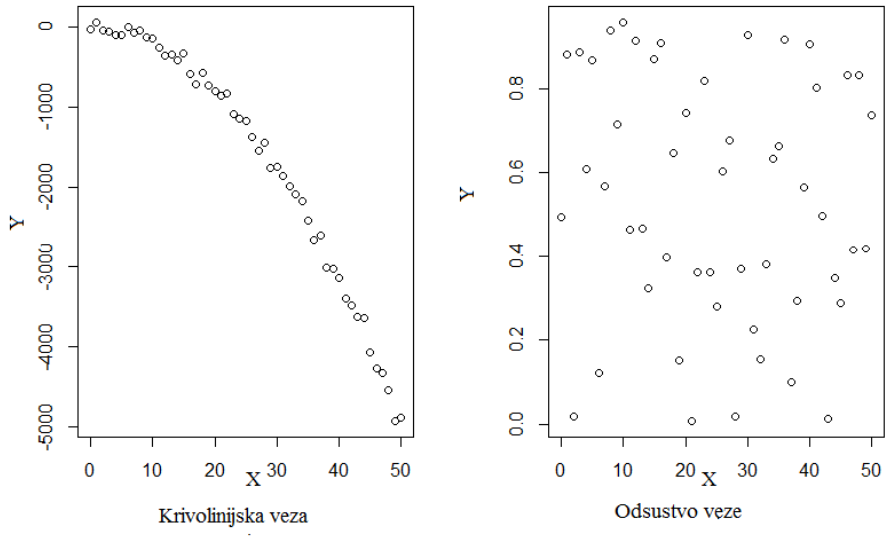
Funkcionalna negativna linearna veza



Stohastička linearna pozitivna veza



Stohastička linearna negativna veza



6.1. Проста линеарна регресија

Најједноставнији облик регресије је проста линеарна регресија помоћу које се сагледава утицај једне независно променљиве X на зависно променљиву Y. Линеарна регресија је исказана функцијом која гласи:

$\hat{Y}_i = a + b \times X_i$,

где је:

$\hat{Y}_i$ – оцењена или очекивана вредност зависно променљиве Y_i;

X_i – независно променљива;

a и b – непознати параметри регресије које је потребно оценити.

Параметар a представља просечни почетни ниво зависно променљиве Y. Другим речима, параметар a показује вредност зависно променљиве у тачки пресека линије регресије и ординатне осе.

Параметар b или коефицијент регресије показује просечну промену зависно променљиве Y за јединицу промене независно променљиве X. Код растуће регресије, параметар b има позитивну вредност ($b > 0$), док код опадајуће регресије има негативну вредност ($b < 0$). Параметри a и b исказују се у јединицама мере зависно променљиве Y.

Рачунски поступак приликом израчунавања параметара регресионог модела, заснован је на методу најмањих квадрата и састоји се у решењу система нормалних једначина. У практичном раду примењују се следећи радни поступци за израчунавање параметара a и b:

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad \text{или} \quad b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}},$$

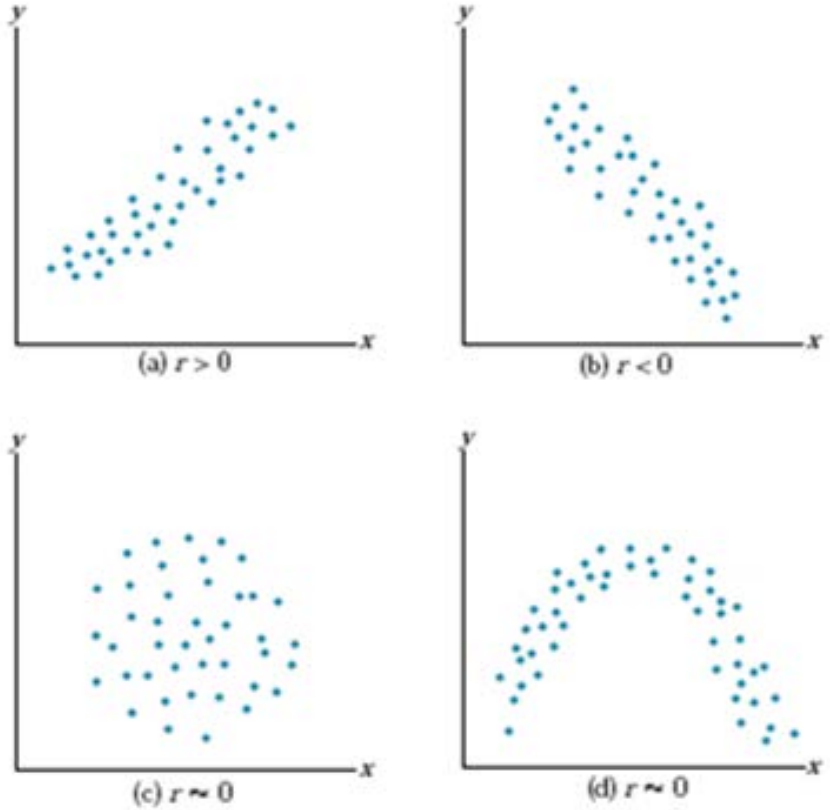
$$a = \bar{Y} - b\bar{X}\,.$$

Стандардна грешка регресије је показатељ дисперзије индивидуалних вредности зависно променљиве Y од линије регресије. Линија регресије представља графички приказ оцењеног регресионог модела и накнадно се уноси на дијаграм растурања. Конкретно, стандардна грешка регресије је показатељ просечног одступања или варијације оригиналних вредности зависно променљиве Y у односу на њихове оцењене вредности (линија регресије). Стандардна грешка регресије исказује се у јединицама мере зависно променљиве Y. У практичном раду утврђује се оцењена стандардна грешка регресије применом следећег обрасца:

$$S_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}}\,.$$

Коефицијент линеарне корелације је показатељ квантитативног слагања две променљиве. Коефицијент линеарне корелације је релативни показатељ корелације, независан од јединица мере променљивих X и Y. Вредност овог коефицијента се креће у интервалу од -1 до 1. Код позитивне корелације, коефицијент корелације се креће у интервалу од 0 до 1, док се код негативне корелације креће у интервалу од 0 до -1. Коефицијент корелације је 1 уколико је веза између X и Y функционална (детерминистичка) и то позитивно линеарна, док је вредност -1 у случају функционалне негативно линеарне везе. Уколико је веза стохастичка, вредности коефицијента корелације блиске 1 указују на позитивну, док вредности блиске -1 указују на негативну линеарну везу. Уколико је вредност коефицијента корелације блиска нули може се само закључити да веза променљивих није линеарна. Дијаграм растурања, између осталог показује да ли постоји нелинеарна веза променљивих или не постоји веза.

Вредност коефицијента корелације није довољна да се закључи да ли је веза променљивих линеарна. Вредност коефицијента корелације може бити блиска ±1 и у случају нелинеарне везе или у случају да један или више парова тачака одступа у односу на остале податке. Дијаграм растурања помаже у правилном тумачењу везе променљивих. У наставку су представљени дијаграми растурања за различите вредности коефицијента корелације.



У практичном раду најчешће се утврђује коефицијент линеарне корелације (r) који се израчунава применом обрасца:

$$r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}} \quad \text{или} \quad r = \frac{\sum X_iY_i-\frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\left[\sum X_i^2-\frac{(\sum X_i)^2}{n}\right]\left[\sum Y_i^2-\frac{(\sum Y_i)^2}{n}\right]}}.$$

Корелациона анализа се допуњује утврђивањем и интерпретацијом *коефицијента детерминације*. Коефицијент детерминације (r^2) представља квадрат коефицијента корелације и најчешће се исказује у процентима. Овај коефицијент се креће у интервалу од 0 до 1 или од 0 до 100%.

Интерпретација овог коефицијента указује на то да је коефицијент детерминације показатељ удела утицаја одабране независно променљиве *X* на варијабилност зависно променљиве *Y*, узимајући у обзир да је укупна варијабилност зависно променљиве *Y* један (100%).

На основу израчунатог коефицијента детерминације може се исказати и коефицијент алијенације, односно коефицијент недетерминације (k^2). Коефицијент недетерминације показује утицај осталих неиспитиваних независно променљивих на варијабилност

зависно променљиве Y, узимајући да је укупна варијабилност зависно променљива Y један (100%). Израчунава се на следећи начин: $k^2 = (1 - r^2) \times 100(\%)$.

6.3. Оцена и тестирање параметара линеарне регресије

Оцена параметара линеарне регресије, подразумева одређивање интервала поверења за коефицијент регресије основног скупа β чија је оцена из узорка параметар b. Интервал поверења има следећи облик:

$$b - t_{n-2;\alpha} \times S_b < \beta < b + t_{n-2;\alpha} \times S_b \ ,$$

где је:

S_b – оцена стандардне грешке коефицијента регресије.

Стандардна грешка коефицијента регресије израчунава се на основу варијансе оцењеног модела на следећи начин:

$$S_b = \sqrt{\frac{S_e^2}{\sum(X_i - \bar{X})^2}},$$

где је:

S_e^2 – квадрирана вредност стандардне грешке модела.

Регресиона анализа се употпуњује извођењем инференције о параметрима регресије. При томе се највећа пажња посвећује тестирању значајности коефицијента регресије b. Нулта и алтернативна хипотеза приликом тестирања значајности коефицијента регресије b гласе:

$$H_0: \beta = 0;$$

$$H_1: \beta \neq 0.$$

Провера полазне хипотезе изводи се помоћу t – теста који се изводи на следећи начин:

$$t = \frac{b - 0}{S_b} = \frac{b}{S_b}.$$

Израчуната вредност t која се упоређује са критичном табличном вредношћу из таблица Студентове дистрибуције ($t_{n-2;\alpha}$), указује на то да ли је t–тест статистички значајан. Уколико је апсолутна вредност израчунатог количника t већа или једнака од критичне вредности из таблице Студентове расподеле, нулта хипотеза се одбацује и закључује се да је вредност коефицијента b статистички значајна на прагу значајности α. Другим речима, може се закључити да постоји статистички значајан утицај независно променљиве X на зависно променљиву Y. Уколико је $|t| < t_{n-2;\alpha}$, нулта хипотеза се прихвата и закључује се да вредност коефицијента b није статистички значајна. На тај

начин, доноси се закључак да не постоји статистички значајан утицај независно променљиве *X* на зависно променљиву *Y*.

Нулта и алтернативна хипотеза приликом тестирања значајности коефицијента линеарне корелације *r* гласе:

$$H_0: r = 0;$$

$$H_1: r \neq 0.$$

Полазна хипотеза за тестирање значајности коефицијента линеарне корелације *r* може се проверити израчунавањем одговарајућег количника на следећи начин:

- у случају великог узорка (*n>30*)

$$Z = \frac{r}{S_r},$$

где је:

r – оцена коефицијента корелације основног скупа на основу узорка;

S_r – оцена стандардне грешке коефицијента корелације на основу узорка.

Стандардна грешка коефицијента корелације на основу великог узорка израчунава се на следећи начин:

$$S_r = \frac{1}{\sqrt{n}}.$$

Апсолутна вредност израчунатог количника Z упоређује се са одговарајућом вредношћу из таблица нормалне дистрибуције (Z_α). Ако је $|Z|$ веће од Z_α, полазна хипотеза се одбацује и прихвата се алтернативна хипотеза. Прихватањем алтернативне хипотезе долази се до закључка да је коефицијент корелације статистички значајан. Супротно, уколико је $|Z|$ мање од Z_α, нулта хипотеза се прихвата на прагу значајности α и закључак је коефицијент корелације није статистички значајан. Другим речима, закључак је да нема статистички значајне корелације између посматраних променљивих.

- у случају малог узорка (*n< 30*) тест критеријум је:

$$t = \frac{r}{S_r}$$

Стандардна грешка коефицијента корелације на основу малог узорка израчунава се на следећи начин:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}.$$

Приликом доношења закључка о полазној хипотези, израчунати количник се упоређује са табличним вредностима Студентове дистрибуције $t_{n-2;\alpha}$.

Уколико је $|t|$ веће од $t_{n-2;\alpha}$ нулта хипотеза се одбацује и прихвата се алтернативна хипотеза на основу које се доноси закучак о статистичкој значајности коефицијента корелације r. У обрнутом случају, уколико је $|t|$ мање од $t_{n-2;\alpha}$, нулта хипотеза се прихвата, што значи да коефицијент корелације није статистички значајан.

Пример 24:

Дати су подаци о утрошку хербицда (кг/ха) и приносу кукуруза (т/ха) на девет случајно одабраних поља:

Утрошак хербицида (кг/ха)	1,22	1,56	1,76	1,84	2,04	1,46	1,91	1,86	2,06
Принос кукуруза (т/ха)	5,15	5,65	5,24	5,74	6,11	5,45	6,24	6,00	6,88

Потребно је:

а) Формирати дијаграм растурања;

б) Оценити линеарни регресиони модел, израчунати оцењене вредности зависно промењиве и уцртати линију регресије на дијаграм растурања;

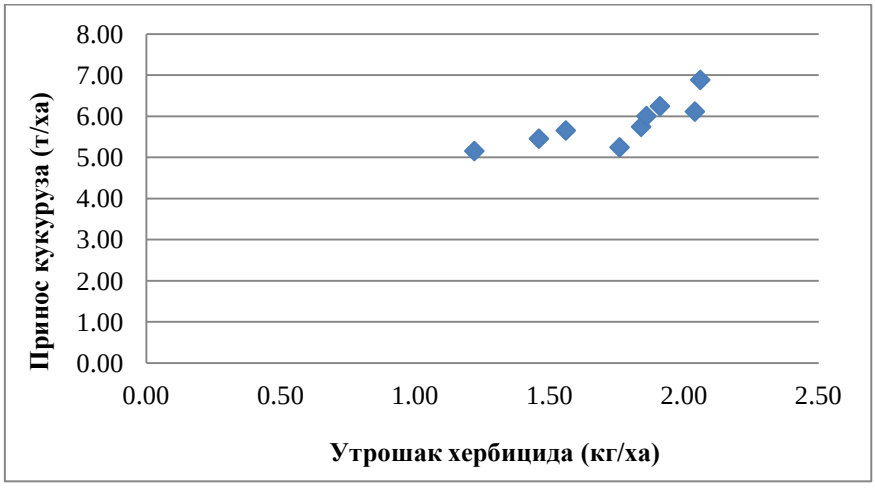
в) Израчунати стандардну грешку регресије;

г) Израчунати коефицијент корелације, детерминације и недетерминације;

д) Колики се принос може очекивати ако је утрошено 2,2 кг/ха хербицида.

а) Дијаграм растурања

Дијаграм растурања се формира тако што се у правоуглом координатном систему унесу сви парови вредности независно и зависно промењиве. На X-оси представљене су вредности независно промењиве (утрошак хербицида), док су на Y-оси представљене вредности зависно промењиве.



б) Оцена регресионог модела $\hat{Y}_i = a \pm b \times X_i$ подразумева претходно израчунавање параметара a и b. Како би се израчунали регресиони параметри неопходно је формирати радну табелу.

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{0{,}9818}{0{,}6270} = 1{,}5657$$

$$a = \bar{Y} - b\bar{X} = 5{,}83 - 1{,}5657 \times 1{,}75 = 3{,}0958;$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{52{,}46}{9} = 5{,}83;$$

$$\bar{X} = \frac{\sum X}{n} = \frac{15{,}71}{9} = 1{,}75.$$

Дакле, оцењени регресиони модел је следећег облика:

$$\hat{Y} = 3{,}0958 + 1{,}5657X_i.$$

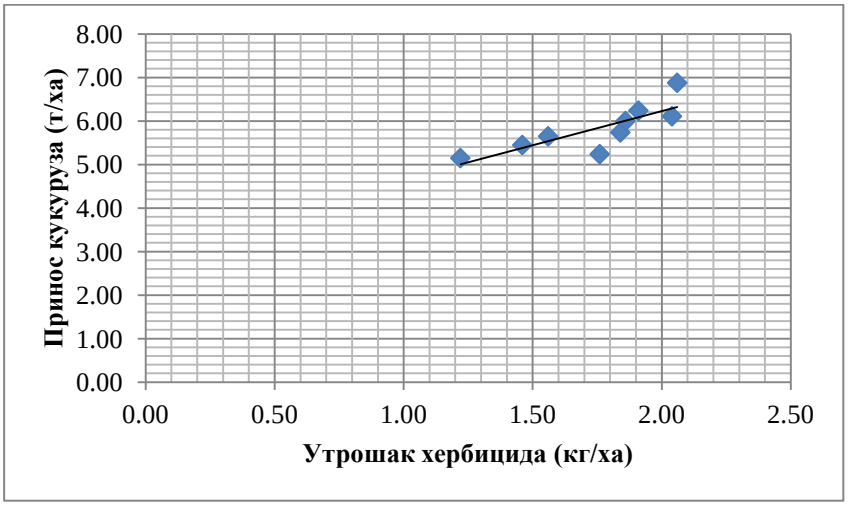
На основу оцењеног регресионог модела, може се констатовати да је за очекивати да се принос кукуруза повећа за 1,5657 т/ха уколико се употреба хербицида повећа за 1 кг/ха.

Оцењене вредности зависно промењљиве ($\hat{Y}$) израчунавају се тако што се у оцењени регресиони модел уврсте вредности независно промењљиве X_i. Нпр. прва оцењена вредност зависно промењљиве може се израчунати на следећи начин:

$$\hat{Y}_1 = 3{,}0958 + 1{,}5657 \times 1{,}22 = 5{,}01.$$

Преостале оцењене вредности зависно промењљиве представљене су у радној табели.

Линија регресије представља графички приказ оцењеног регресионог модела. Како би се уцртала линија регресије неопходно је у дијаграм растурања унети парове вредности независно промениве X и оцењене вредности зависно промењљиве $\hat{Y}$. На основу линије регресије може се визуелно сагледати одступање оригиналних од оцењених вредности зависно промењљиве:



в) Стандардна грешка регресије се израчунава на основу следеће формуле:

$$S_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}} = \sqrt{\frac{0{,}8361}{9-2}} = 0{,}3456 \text{ т/ха.}$$

г) Коефицијент корелације r рачуна се на следећи начин:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} = \frac{0{,}9818}{\sqrt{0{,}6270 \times 2{,}3733}} = 0{,}8048;$$

Коефицијент детерминације r^2 представља квадрат коефицијента корелације:

$$r^2 \times 100(\%) = 0{,}8048^2 \times 100 = 64{,}77\%.$$

С друге стране, коефицијент недетерминације је могуће израчунати на следећи начин:

$$k^2 = (1 - r^2) \times 100(\%) = (1 - 0{,}8048^2) \times 100(\%) = 35{,}23\%.$$

Израчунати коефицијент детерминације указује на то да се 64,77% варијабилитета приноса кукуруза може објаснити употребом хербицида. С друге стране, на основу коефицијента недетерминације може се закључити да преосталих 35,23% представља утицај неких других фактора који нису предмет анализе.

д) Прогнозирање вредности зависно променљиве за дату вредност независо променљиве може утврдити убацивањем одговарајуће вредности независно променљиве у оцењени регресиони модел. Конкретно, прогноза приноса кукуруза услед апликације 2,2 кг/ха хербицида може се утврдити на седећи начин:

$$\hat{Y}_{2,2} = 3{,}0958 + 1{,}5657 \times 2{,}2 = 6{,}54 \text{ т/ха.}$$

Радна табела:

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$	$\hat{Y}$	$Y_i - \hat{Y}$	$(Y_i - \hat{Y})^2$
1,22	5,15	-0,53	-0,68	0,2762	0,4609	0,3568	5,01	0,14	0,0207
1,56	5,65	-0,19	-0,18	0,0344	0,0320	0,0332	5,54	0,11	0,0125
1,76	5,24	0,01	-0,59	0,0002	0,3468	-0,0085	5,85	-0,61	0,3739
1,84	5,74	0,09	-0,09	0,0089	0,0079	-0,0084	5,98	-0,24	0,0561
2,04	6,11	0,29	0,28	0,0867	0,0790	0,0828	6,29	-0,18	0,0324
1,46	5,45	-0,29	-0,38	0,0815	0,1436	0,1082	5,38	0,07	0,0047
1,91	6,24	0,16	0,41	0,0270	0,1690	0,0676	6,09	0,15	0,0236
1,86	6,00	0,11	0,17	0,0131	0,0293	0,0196	6,01	-0,01	0,0001
2,06	6,88	0,31	1,05	0,0989	1,1048	0,3305	6,32	0,56	0,3122
15,71	**52,46**	**0,00**	**0,00**	**0,6270**	**2,3733**	**0,9818**	**52,46**		**0,8361**

Сви ови израчунати показатељи могу се израчунати и применом друге групе формула. Када је реч о оцењени регресионог модела односно израчунавање регресионих параметара b и a.

$$b = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{92{,}55 - \frac{15{,}71 \times 52{,}46}{9}}{28{,}05 - \frac{28{,}05^2}{9}} = 1{,}5657.$$

$$a = \bar{Y} - b\bar{X} = 5{,}83 - 1{,}5657 \times 1{,}75 = 3{,}0958;$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{52{,}46}{9} = 5{,}83;$$

$$\bar{X} = \frac{\sum X}{n} = \frac{15{,}71}{9} = 1{,}75.$$

Дакле, оцењени регресиони модел је следећег облика:

$$\hat{Y} = 3{,}0958 + 1{,}5657 X_i.$$

Оцењене вредности зависно промењљиве и стандардна грешка регресије рачунају се на начин који је објашњен у првом делу решења.

Радна табела:

X	Y	XY	X^2	Y^2
1,22	5,15	6,28	1,49	26,52
1,56	5,65	8,81	2,43	31,92
1,76	5,24	9,22	3,10	27,46
1,84	5,74	10,56	3,39	32,95
2,04	6,11	12,46	4,16	37,33
1,46	5,45	7,96	2,13	29,70
1,91	6,24	11,92	3,65	38,94
1,86	6,00	11,16	3,46	36,00
2,06	6,88	14,17	4,24	47,33
15,71	**52,46**	**92,55**	**28,05**	**308,16**

Коефицијент корелације r рачуна се на следећи начин:

$$r = \frac{\sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}}{\sqrt{\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right]\left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right]}} = \frac{92{,}55 - \frac{15{,}71 \times 52{,}46}{9}}{\sqrt{\left[28{,}05 - \frac{15{,}71^2}{9}\right]\left[308{,}16 - \frac{52{,}46^2}{9}\right]}} = 0{,}8048$$

Коефицијент детерминације r^2 представља квадрат коефицијента корелације:

$$r^2 \times 100(\%) = 0{,}8048^2 \times 100 = 64{,}77\%.$$

С друге стране, коефицијент недетерминације је могуће израчунати на следећи начин:

$$k^2 = (1 - r^2) \times 100(\%) = (1 - 0{,}8048^2) \times 100(\%) = 35{,}23\%.$$

Инференција у оквиру линеарне регресионе анализе подразумева тестирање статистичке значајности регресионог коефицијента b, дефинисање интервала поверења за параметар β и тестирање коефицијента корелације r.

Тестирање статистичке значајности регресионог коефицијента b:

$$H_0: \beta = 0;$$

$$H_1: \beta \neq 0.$$

Провера полазне хипотезе изводи се помоћу *t* – теста који се изводи на следећи начин:

$$t = \frac{b - 0}{S_b} = \frac{b}{S_b} = \frac{1{,}5657}{0{,}4365} = 3{,}59$$

$$S_b = \sqrt{\frac{S_e^2}{\sum(X_i - \bar{X})^2}} = \sqrt{\frac{0{,}3456^2}{0{,}6270}} = 0{,}4365$$

$t_{7;0,05} = 2{,}365$

$t_{7;0,01} = 3{,}499$

$$|t| > 2{,}365 \rightarrow H_1$$

$$|t| > 3{,}499 \rightarrow H_1$$

С обзиром на то да је апсолутна вредност t-колчника већа од обе критичне вредности из t-таблице, нулта хипотеза се одбацује и прихвата се алтернативна. Закључак је да је коефицијент регресије b високо статистички значајан, што значи да има смисла испитивати утицај хербицида на принос кукуруза.

Интервала поверења за параметар β:

Приликом оцене вредности параметра β неопходно је дефинисати следећи интервал поверења:

$$b - t_{n-2;\alpha} \times S_b < \beta < b + t_{n-2;\alpha} \times S_b.$$

95% интервал поверења:

$$1{,}5657 - 2{,}365 \times 0{,}4365 < \beta < 1{,}5657 + 2{,}365 \times 0{,}4365$$

$$0{,}5334 < \beta < 2{,}5980$$

$$0 \notin (L_1, L_2) \rightarrow H_1$$

99% интервал поверења:

$$1{,}5657 - 3{,}499 \times 0{,}4365 < \beta < 1{,}5657 + 3{,}499 \times 0{,}4365$$

$$0{,}0384 < \beta < 3{,}0930$$

$$0 \notin (L_1, L_2) \rightarrow H_1$$

Тестирање коефицијента корелације r:

$$H_0: r = 0;$$

$$H_1: r \neq 0.$$

Провера полазне хипотезе изводи се помоћу *t* – теста који се изводи на следећи начин:

$$t = \frac{r}{S_r} = \frac{0{,}8048}{0{,}2243} = 3{,}5874$$

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0{,}8048^2}{9 - 2}} = 0{,}2243$$

$t_{7;0,05} = 2{,}365$

$t_{7;0,01} = 3{,}499$

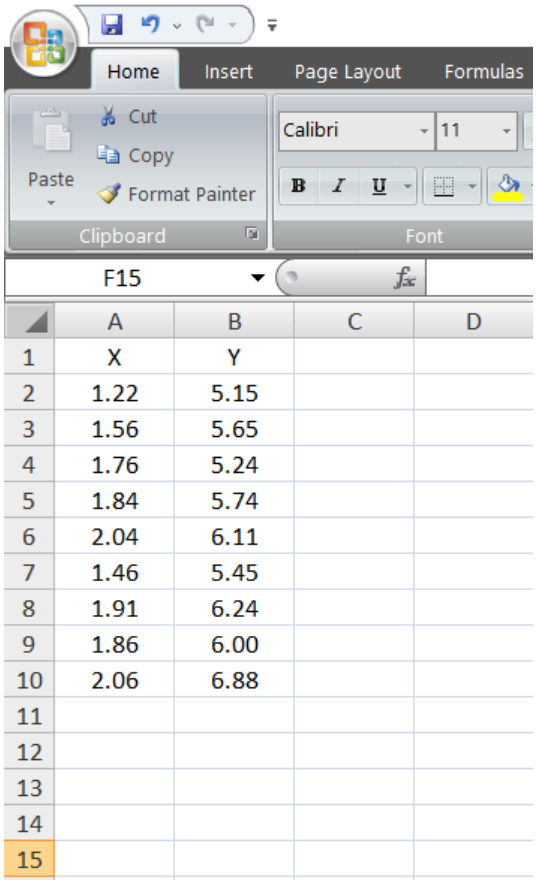
$$|t| > 2{,}365 \rightarrow H_1$$

$$|t| > 3{,}499 \rightarrow H_1.$$

Како нулта хипотеза одбацује и прихвата се алтернативна за оба прага значајности, може се закључити да је линеарна веза између приноса кукуруза и утрошка хербицида високо статистички значајна.

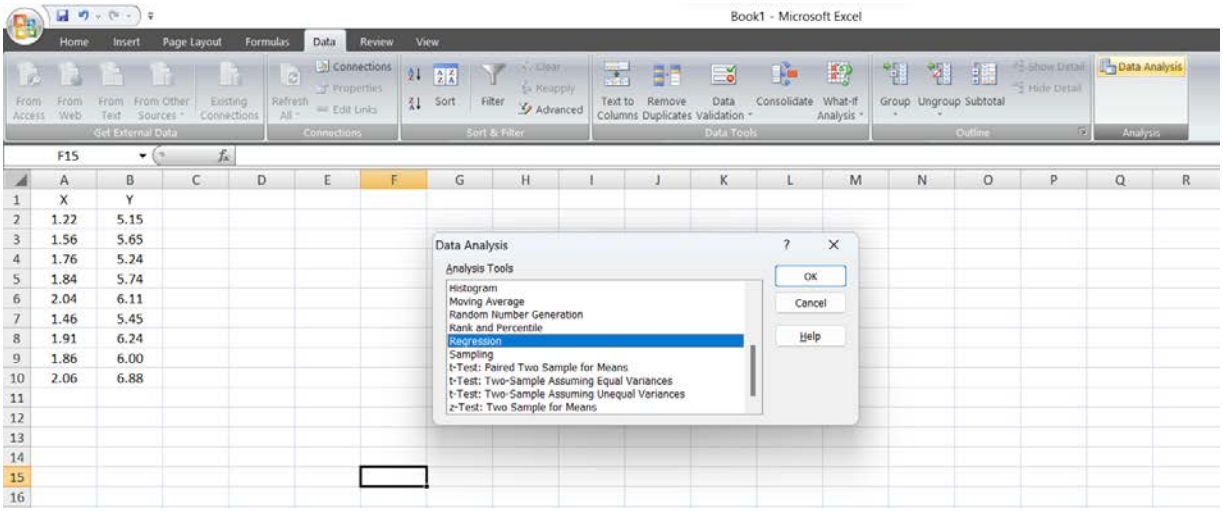
Вежба 11. Регресиона анализа применом Microsoft Excel-a

1. Полазни изглед табеле је следећи:

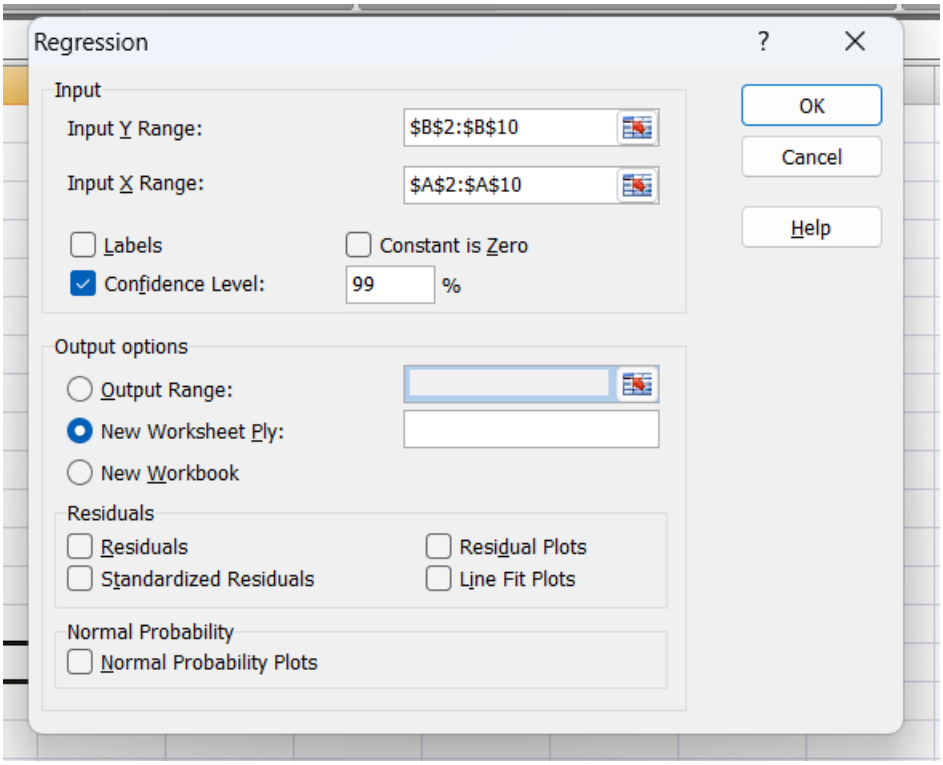


	A	B	C	D
1	X	Y		
2	1.22	5.15		
3	1.56	5.65		
4	1.76	5.24		
5	1.84	5.74		
6	2.04	6.11		
7	1.46	5.45		
8	1.91	6.24		
9	1.86	6.00		
10	2.06	6.88		
11				
12				
13				
14				
15				

2. Следеће што је потребно урадити јесте у картици *Data* клинкути на *Data Analsys* у блоку који се односи на *Analysis*. Отвориће се нови прозор у којем је потребно наћи ставку под називом *Regression*, као што је показано у наставку:



3. Кликом на дугме *OK* отвара се нови прозор. У блоку који се односи на *Input*, у делу који се односи на *Input Y Range* и *Input X Range* потребно је обележити податке који односе на зависно и независно променљиву респективно. У истом блоку штиклираћемо *Confidence Level* и уписати 99 са десне стране. На тај, начин смо задали команду да нам *Microsoft Excel* дефинише доњу и горњу границу 99% интервала поверења за параметар β (95% интервал поверења је већ део коначног извештаја). Поглед на *Microsoft Excel* сада изгледа на следећи начин:



У првој табели са заглављем *Regression Statistics* дате су редом вредности: коефицијента корелације, коефицијента детерминације, кориговани коефицијент детерминације, стандардна грешка регресије и укупан број опсервација.

У другој табели са заглављем *ANOVA* дат је тест регресионог модела у целини, односно анализа варијансе регресије. Код анализе варијансе регресије нулта хипотеза теста је истао као и код теста значајности коефицијента регресије. Поступак анализе варијансе је сличан објашњеном поступку у делу 5.3.. Уколико је вредност у колони *Significance F* мања од 0,01 може се сматрати да је регресиони модел високо статистички значајан.

Прва колона треће табеле у низу односи се на ознаке за параметар *a* (*Intercept*) и коефицијент регресије *b* (*X Variable 1*). У другој колони дате су вредности наведених параметара, док су у трећој колони представљене вредности њихових оцењених стандардних грешака. Четврта колона приказује *t*-количник који се користи приликом тестирања статистичке значајности наведених параметара (*t Stat*), док пета колона показује одговарајуће *p*-вредности (*P-value*), на основу којих се доноси закључак о

статистичкој значајности регресионих параметара. У шестој и седмој колони наведене су доња и горња граница 95% интервала поверења за регерсионе параметре. Осма и девета колона се добију уколико се у другом кораку штиклира *Confidence Level*. Како смо том приликом уписали 99, *Microsoft Excel* нам је оценио 99% интервал поверења за посматране регресионе параметре.

Контролна питања

1, Шта је циљ примене регресионе анализе?

2, Шта је циљ примене корелационе анализе?

3, На основу чега се сагледава регресиона анализа?

4, На основу чега се сагледава корелациона анализа?

5, Како се формира дијаграм растурања и која му је сврха?

6, Шта показује коефицијент правца регресије?

7, Дефинисати коефицијент корелације,

8, Дефинисати коефицијент детерминације,

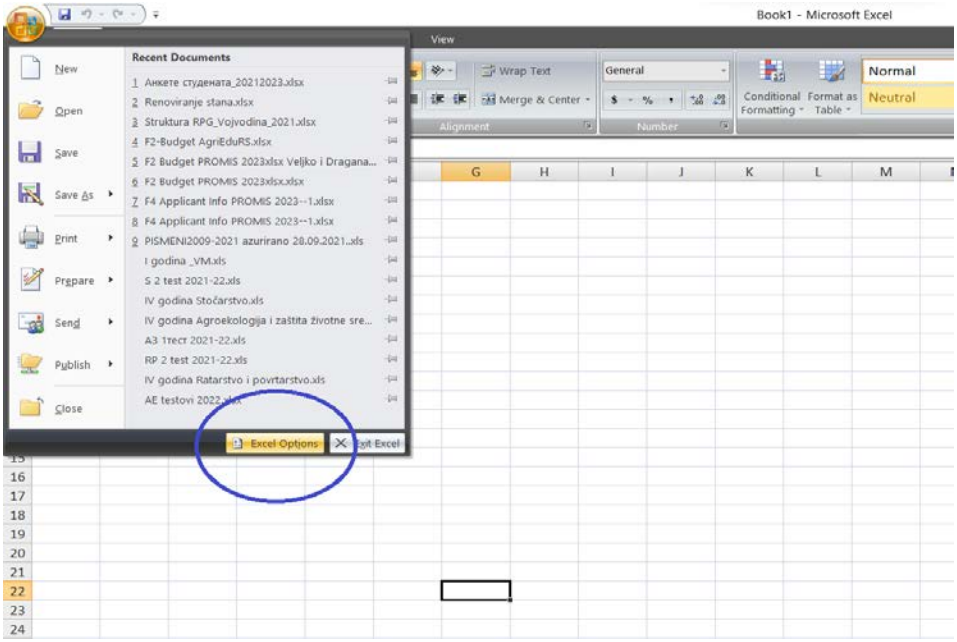
9, На који начин се проверава значајност оцењеног коефицијента регресије?

10, На који начин се проверава значајност оцењеног коефицијента корелације?

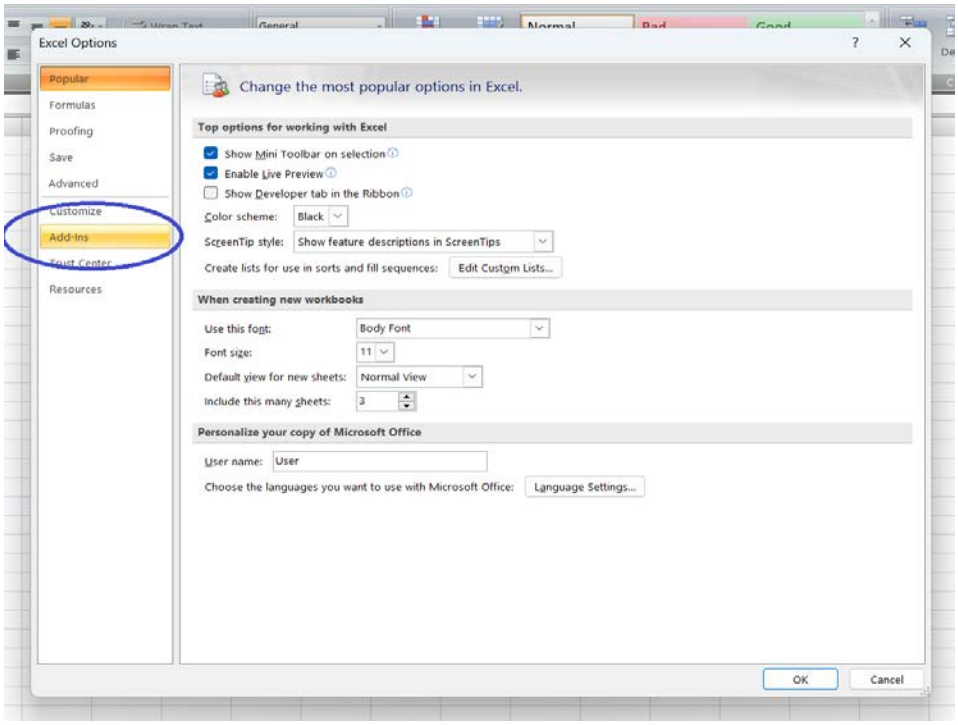
ПРИЛОЗИ

Прилог 1, Инсталација Microsoft Excel пакета Data Analysis

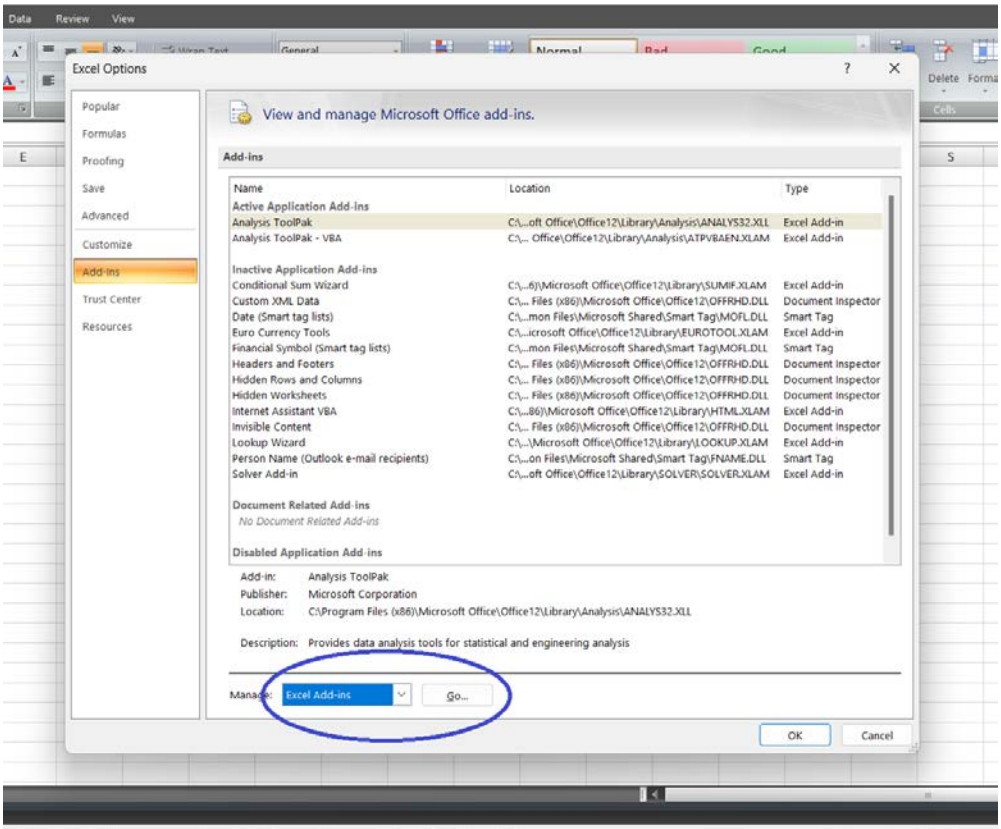
1, Отворите *Microsoft Excel*. У падајућем менију изаберите опцију *File* (новије верзије *Microsoft Excel*-а) или круг са логоом у горњем левом углу. Изаберите опцију *Excel options*.



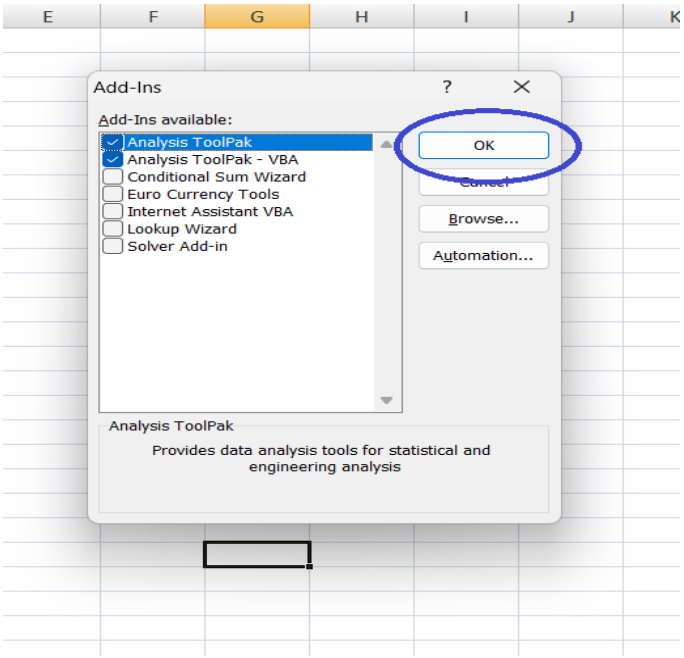
2, Отвриће се нови прозор у којем са леве стране треба да изаберете опцију *Add-Ins*:



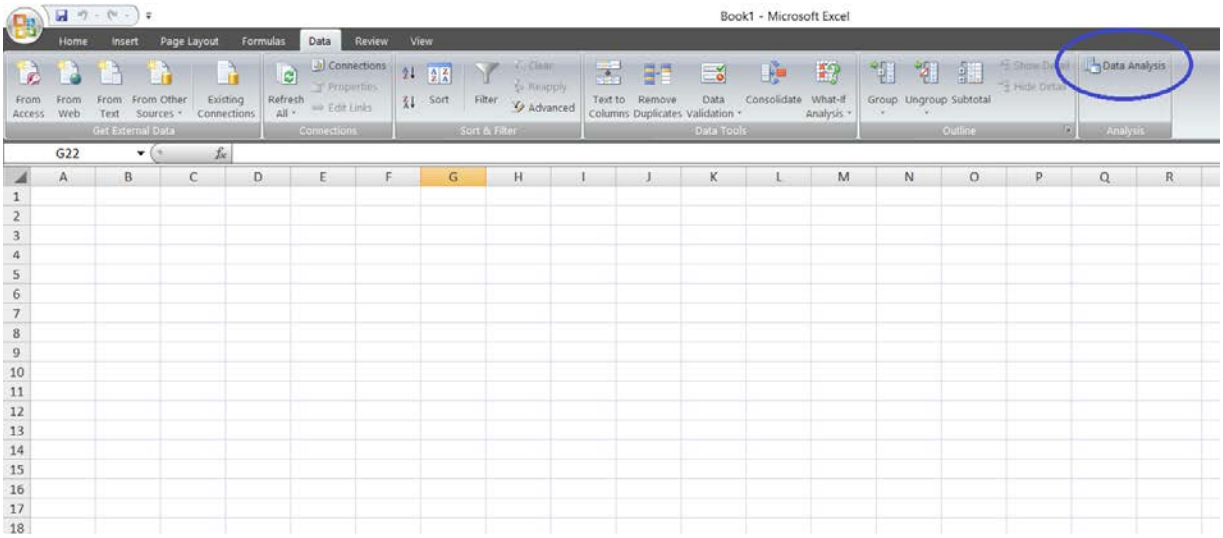
3, Кликом на опцију *Add-Ins* отвара се нови прозор. У доњем делу прозора код ставке *Manage* потребно је да из падајућег менија изаберете опцију *Excel Add-Ins* и затим да кликнете на *Go...* десно од падајућег менија.



4, Отвориће се нови прозорчић у којем је потребно штиклирати *Analysis ToolPak* и *Analysis ToolPak – VBA*. За крај кликните на дугме *ОК*.



У оквиру картице *Data*, на крају са десне стране требало би да се појави инсталирани пакет *Data Analysis*,



Прилог 2. Таблица нормалне расподеле

Z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4094	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Извор: Newbold i sar., 2010

Прилог 3. Таблица Студентове t - расподеле

Степени слободе	**α**				
	0,100	**0,050**	**0,025**	**0,010**	**0,005**
1	3,078	6,314	12,706	31,821	63,567
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	0,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
∞	1,282	1,645	1,960	2,326	2,576

Извор: Newbold i sar., 2010

Прилог 4. Таблица χ^2 расподеле

Степени слободе	**α**									
	0,995	**0,990**	**0,975**	**0,950**	**0,900**	**0,100**	**0,050**	**0,025**	**0,010**	**0,005**
1	$0,0^{4}393$	$0,0^{3}157$	$0,0^{3}982$	$0,0^{2}393$	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	74,40	79,08	93,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,4	104,2
80	51,17	53,54	57,15	60,39	64,28	96,58	101,9	106,6	112,3	116,3
90	59,20	61,75	65,65	69,13	73,29	107,6	113,1	118,1	124,1	128,3
100	67,33	70,06	74,22	77,93	82,36	118,5	124,3	129,6	135,8	140,2

Извор: Newbold i sar., 2010

Прилог 5. Таблица Фишерове F – расподеле (α=0,05)

Степени слободе r_2	**Степени слободе r_1**														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**20**	**30**	**50**	**∞**
1	161	200	216	225	230	234	237	239	241	242	244	248	250	252	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,41	19,44	19,46	19,47	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74	8,66	8,62	8,58	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,80	5,74	5,70	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68	4,56	4,50	4,44	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,87	3,81	3,75	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57	3,44	3,38	3,32	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,28	3,15	3,08	3,03	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07	2,93	2,86	2,80	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91	2,77	2,70	2,64	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,79	2,65	2,57	2,50	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,69	2,54	2,46	2,40	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,60	2,46	2,38	2,32	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,53	2,39	2,31	2,24	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,48	2,33	2,25	2,18	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,28	2,20	2,13	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,38	2,23	2,15	2,08	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,19	2,11	2,04	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,31	2,15	2,07	2,00	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,28	2,12	2,04	1,96	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,09	2,00	1,93	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,23	2,07	1,98	1,91	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,20	2,04	1,96	1,88	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,18	2,02	1,94	1,86	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24	2,16	2,00	1,92	1,84	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,09	1,93	1,84	1,76	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,00	1,84	1,74	1,66	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,95	1,78	1,69	1,60	1,44
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,89	1,72	1,62	1,53	1,35
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,85	1,68	1,57	1,48	1,28
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,75	1,57	1,46	1,35	1,00

Извор: Newbold i sar., 2010

Прилог 6. Таблица Фишерове F – расподеле (α=0,01)

Степени слободе r_2	Степени слободе r_1														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**20**	**30**	**50**	**∞**
1	4,052,2	4,999,5	5,403,4	5,624,6	5,763,7	5,859,0	5,928,4	5,981,1	6,022,5	6,055,9	6,106,3	6,208,7	6,260,7	6,302,0	6,365,9
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,45	99,47	99,48	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,69	26,51	23,35	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,02	13,84	13,69	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,55	9,38	9,24	9,02
6	13,75	10,93	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,40	7,23	7,09	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,16	5,99	5,85	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,36	5,20	5,06	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,81	4,65	4,51	4,31
10	0,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,41	4,25	4,12	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,10	3,94	3,80	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	3,86	3,70	3,56	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,67	3,51	3,37	3,17
14	8,86	6,52	5,56	5,04	4,70	4,46	4,28	4,14	4,03	3,94	3,80	3,51	3,35	3,21	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,90	3,81	3,67	3,37	3,21	3,07	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,26	3,10	2,96	2,75
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,16	3,00	2,86	2,65
18	8,29	6,01	5,09	4,58	4,25	4,02	3,84	3,71	3,60	3,51	3,37	3,08	2,92	2,78	2,57
19	8,19	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,00	2,84	2,70	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	2,94	2,78	2,63	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	2,88	2,72	2,58	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,83	2,67	2,53	2,31
23	7,88	5,66	4,77	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,78	2,62	2,48	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,74	2,58	2,44	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13	2,99	2,70	2,54	2,40	2,17
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,55	2,39	2,24	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,67	2,37	2,20	2,05	1,81
50	7,17	5,06	4,20	3,72	3,41	3,18	3,02	2,88	2,78	2,70	2,56	2,26	2,10	1,94	1,68
70	7,01	4,92	4,08	3,60	3,29	3,07	2,91	2,77	2,67	2,59	2,45	2,15	1,98	1,82	1,53
100	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,36	2,06	1,89	1,73	1,43
∞	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,19	1,88	1,70	1,52	1,00

Извор: Newbold i sar., 2010

Прилог 7. Таблица за нови вишеструки интервални тест (α=0,05)

Степени слободе погрешке	Број средина за интервал који се тестира													
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**14**	**16**	**18**	**20**
1	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0	18,0
2	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09	6,09
3	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50	4,50
4	3,93	4,01	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02	4,02
5	3,64	3,74	3,79	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83	3,83
6	3,46	3,58	3,84	3,68	3,68	3,68	3,68	3,68	3,68	3,68	3,68	3,68	3,68	3,68
7	3,35	3,47	3,54	3,58	3,60	3,61	3,61	3,61	3,61	3,61	3,61	3,61	3,61	3,61
8	3,26	3,39	3,47	3,52	3,55	3,56	3,56	3,56	3,56	3,56	3,56	3,56	3,56	3,56
9	3,20	3,34	3,41	3,47	3,50	3,52	3,52	3,52	3,52	3,52	3,52	3,52	3,52	3,52
10	3,15	3,30	3,37	3,43	3,46	3,47	3,47	3,47	3,47	3,47	3,47	3,47	3,47	3,48
11	3,11	3,27	3,35	3,39	3,43	3,44	3,45	3,46	3,46	3,46	3,46	3,46	3,47	3,48
12	3,08	3,23	3,33	3,36	3,40	3,42	3,44	3,44	3,46	3,46	3,46	3,46	3,47	3,48
13	3,06	3,21	3,30	3,35	3,38	3,41	3,42	3,44	3,45	3,45	3,46	3,46	3,47	3,47
14	3,03	3,18	3,27	3,33	3,37	3,39	3,41	3,42	3,44	3,45	3,46	3,46	3,47	3,47
15	3,01	3,16	3,25	3,31	3,36	3,38	3,40	3,42	3,43	3,44	3,45	3,46	3,47	3,47
16	3,00	3,15	3,23	3,30	3,34	3,37	3,39	3,41	3,43	3,44	3,45	3,46	3,47	3,47
17	2,98	3,13	3,22	3,28	3,33	3,36	3,38	3,40	3,42	3,44	3,45	3,46	3,47	3,47
18	2,97	3,12	3,21	3,27	3,32	3,35	3,37	3,39	3,41	3,43	3,45	3,46	3,47	3,47
19	2,96	3,11	3,19	3,26	3,31	3,35	3,37	3,39	3,41	3,43	3,44	3,46	3,47	3,47
20	2,95	3,10	3,18	3,25	3,30	3,34	3,36	3,38	3,40	3,43	3,44	3,46	3,46	3,47
22	2,93	3,08	3,17	3,24	3,29	3,32	3,35	3,37	3,39	3,42	3,44	3,45	3,46	3,47
24	2,92	3,07	3,15	3,22	3,28	3,31	3,34	3,37	3,38	3,41	3,44	3,45	3,46	3,47
26	2,91	3,06	3,14	3,21	3,27	3,30	3,34	3,36	3,38	3,41	3,43	3,45	3,46	3,47
28	2,90	3,04	3,13	3,20	3,26	3,30	3,33	3,35	3,37	3,40	3,43	3,45	3,46	3,47
30	2,89	3,04	3,12	3,20	3,25	3,29	3,32	3,35	3,37	3,40	3,43	3,44	3,46	3,47
40	2,86	3,01	3,10	3,17	3,22	3,27	3,30	3,33	3,35	3,39	3,42	3,44	3,46	3,47
80	2,83	2,98	3,08	3,14	3,20	3,24	3,28	3,31	3,33	3,37	3,40	3,43	3,45	3,47
100	2,80	2,95	3,05	3,12	3,18	3,22	3,26	3,30	3,32	3,36	3,40	3,42	3,45	3,47
∞	2,77	2,92	3,02	3,09	3,15	3,19	3,23	3,26	3,9	3,34	3,38	3,41	3,44	3,47

Извор: Newbold i sar., 2010

Прилог 8. Таблица за нови вишеструки интервални тест (α=0,01)

Степени слободе погрешке	Број средина за интервал који се тестира													
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**14**	**16**	**18**	**20**
1	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0	90,0
2	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0	14,0
3	8,26	8,50	8,60	8,70	8,80	8,90	8,90	9,00	9,00	9,00	9,00	9,20	9,30	9,30
4	6,51	6,80	6,90	7,00	7,10	7,10	7,20	7,20	7,30	7,30	7,40	7,40	7,50	7,50
5	5,70	5,96	6,11	6,18	6,26	6,33	6,40	6,44	6,50	6,60	6,60	6,70	6,70	6,80
6	5,24	5,51	5,65	5,73	5,81	5,88	5,95	6,00	6,00	6,10	6,20	6,20	6,30	6,30
7	4,95	5,22	5,37	5,45	5,53	5,61	5,69	5,73	5,80	5,80	5,90	5,90	6,00	6,00
8	4,74	5,00	5,14	5,23	5,32	5,40	5,47	5,51	5,50	5,60	5,70	5,70	5,80	5,80
9	4,60	4,86	4,99	5,08	5,17	5,25	5,32	5,36	5,40	5,50	5,50	5,60	5,70	5,70
10	4,48	4,73	4,88	4,96	5,06	5,13	5,20	5,24	5,28	5,36	5,42	5,48	5,54	5,55
11	4,39	4,63	4,77	4,86	4,94	5,01	5,05	5,12	5,15	4,24	5,28	5,34	5,38	5,39
12	4,32	4,55	4,68	4,76	4,84	4,92	4,95	5,02	5,07	5,13	5,17	5,22	5,24	5,26
13	4,26	4,48	4,62	4,69	4,74	4,84	4,88	4,94	4,98	5,04	5,08	5,13	5,14	5,15
14	4,21	4,42	4,55	4,63	4,70	4,78	4,83	4,87	4,91	4,96	5,00	5,04	5,06	5,07
15	4,17	4,37	4,50	4,58	4,64	4,72	4,77	4,81	4,84	4,90	4,94	4,97	4,99	5,00
16	4,13	4,34	4,45	4,54	4,60	4,67	4,72	4,76	4,79	4,84	4,88	4,91	4,93	4,94
17	4,10	4,30	4,41	4,5	4,56	4,63	4,68	4,72	4,75	4,80	4,83	4,86	4,88	4,89
18	4,07	4,27	4,38	4,46	4,53	4,59	4,64	4,68	4,71	4,76	4,79	4,82	4,84	4,85
19	4,05	4,24	4,35	4,43	4,50	4,56	4,61	4,64	4,67	4,72	4,76	4,79	4,81	4,82
20	4,02	4,22	4,33	4,4	4,47	4,53	4,58	4,61	4,65	4,69	4,73	4,76	4,78	4,79
22	3,99	4,17	4,28	4,36	4,42	4,48	4,53	4,57	4,60	4,65	4,68	4,71	4,74	4,75
24	3,95	4,14	4,24	4,33	4,39	4,44	4,49	4,53	4,57	4,62	4,64	4,67	4,70	4,72
26	3,93	4,11	4,21	4,30	4,36	4,41	4,46	4,50	4,53	4,58	4,62	4,65	4,67	4,69
28	3,91	4,08	4,18	4,28	4,34	4,39	4,43	4,47	4,51	4,56	4,60	4,62	4,65	4,67
30	3,89	4,06	4,16	4,22	4,32	4,36	4,41	4,45	4,48	4,54	4,58	4,61	4,63	4,65
40	3,82	3,99	4,10	4,17	4,24	4,30	4,34	4,37	4,41	4,46	4,51	4,54	4,57	4,59
80	3,76	3,92	4,03	4,12	4,17	4,23	4,27	4,31	4,34	4,39	4,44	4,47	4,50	4,53
100	3,71	3,86	3,98	4,06	4,11	4,17	4,21	4,25	4,29	4,35	4,38	4,42	4,45	4,48
∞	3,64	3,80	3,90	3,98	4,04	4,09	4,14	4,17	4,20	4,26	4,31	4,34	4,38	4,41

Извор: Newbold i sar., 2010

Коришћена литература

1. Aho, K. A., *Foundational and Applied Statisstics for Biologist Using R*, CRC Press, Taylor & Francis Group, 2014.
2. Bruce, P.C., Bruce, A. G., *Practical Statistics for Data Scientists*, O'Reilly Media, Inc., USA, 2016.
3. Daniel, W. W., Cross, C. L., *Biostatistics, A Foundation for Analysis in the Health Sciences*, Tenth Edition, Wiley, 2013.
4. Dumičić, K., Bahovec, V., Čižmešija, M., Kurnoga Živadinović, N., Čeh Časni, A., Jakšić, S., Palić, I., Sorić, P., Žmuk, B., *Poslovna statistika*, Element d.o.o., Zagreb, 2011.
5. Darlington, B.R., Hayes,F.A., *Regression Analysis and Linear Models*, Ebook, The Guilford Press, New York, London, 2017.
6. Diggle, P.J., Chetwynd, A. G., *Statistics and Scientific Method, An Introduction for Students and Researchers*, Oxford University Press, Inc., New York, 2011.
7. Hadživuković, S., *Statistika*, Privredni pregled Beograd, 1989.
8. Hadživuković, S., *Statistički metodi*, Drugo prošireno izdanje, Poljoprivredni fakultet, Novi Sad,1991.
9. Jazbec, A., *Osnove statistike*, Šumarski fakultet, Zagreb, 2008.
10. Kaps, M., Lamberson, W.R., *Biostatistics for Animal Science*, Third edition, CABI Publishing, UK, 2017.
11. Le, C. T., Eberly, L. E., *Introductory Biostatistics*, John Wiley & Sons, Inc., New Jersey, 2016.
12. Lozanov-Crvenković, Z., *Statistika*, PMF Novi Sad, 2012.
13. Maletić, R., *Statistika*, Poljoprivredni fakultet, Beograd-Zemun, 2005.
14. Mann, P. S., *Uvod u Statistiku*, Ekonomski fakultet, Beograd, 2009.
15. Mutavdžić Beba, Emilija Nikolić-Đorić, Statistika (za smer Veterinarska medicina), Poljoprivredni fakultet, Univerzitet u Novom Sadu, 2018.
16. Newbold P., Carlson W.L., Thorne Betty, *Statistika za poslovanje i ekonomiju*, MATE d.o.o. Zagreb, 2010.
17. Pagano, R.R., *Understanding Statistics*, Tenth Edition, Wadsworth, Cengage Learning, 2013.
18. Petrie, A., Watson, P., *Statistics for Veterinary and Animal Science*, Third edition, Wiley Blackwell, John Wiley & Sons, Ltd., Publication, 2013.
19. Petz, B., Kolesarić, V., Ivanec, D., *Petzova statistika, Osnovne statističke metode za nematematičare*, Naklada Slap, Jastrebarsko, 2012.

20. Rao, G. N., *Statistics for Agricultural Sciences*, Second Edition, BS Publications, Hyderabad, 2007.

21. Riffenburgh, H.R., *Statistics in Medicine*, Third Edition, Elsevier, 2012.

22. Sokolovska, V., *Deskriptivna Statistika*, Univerzitet u Novom Sadu, Centar za primenjenu statistiku, Novi Sad, 2013.

23. Stanković**,** J., Ralevi**ć,** N., L**j**ubanović-Ralević**,** I., *Statistika sa primenama u poljoprivredi*, Mladost Biro, Beograd, 2002.

24. Vasilj, Đ., *Biometrika i eksperimentiranje u bilinogojsvu*, Hrvatsko agronomsko društvo, Zagreb, 2000.

24. Weiss, A. N., *Introductory Statistics*, 9th Edition, Addison-Wesley, 2012.

25. Wilcox, R. R., *Understanding and Applying Basic Statistical Methods Using R*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2017.

26. Zar, J.H., *Biostatistical Analysis*, 5th Edition, Pearson Education, Ltd. London, 2010.

27. Žižić, M., Lovrić, M., Pavličić, D., *Metodi statističke analize*, Šesnesto izdanje, Centar za izdavačku delatnost Ekonomskog fakulteta, Beograd, 2006.