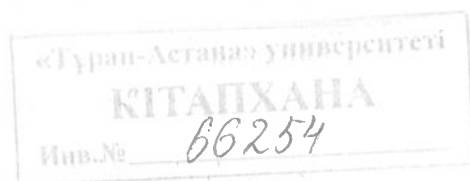


**З.П. Айдынов  
С.А. Джумабаев  
Н.С. Нуркашева**

# **«STATISTICA» ЖҮЙЕСІНДЕ МӘЛІМЕТТЕРДІ ТАЛДАУ**

*Оқу құралы*



**Алматы, 2019**

ӘОЖ 311(075)

КБЖ 60.6я73

А 31

*Нархоз университетінің Ғылыми кеңесі басылымға ұсынды.  
(Хаттама №7, 27 ақпан 2019 жыл)*

**Сын пікір жазғандар:**

**Даулетбақов Б.Д.** – экономика ғылымдарының докторы., профессор;

**Жантаева А.А.** – PhD докторы;

**Алайдарқызы Қ.** – экономика ғылымдарының кандидаты.

**А 31 «STATISTICA» жүйесінде мәліметтерді талдау.** Оқу құралы. **Айдынов.З.П., Джумабаев С.А., Нуркашева Н.С.** Алматы: «Фортуна Полиграф» баспасы ЖШС, 2019- 1076.

ISBN 978-601-329-104-8

Оқу құралы жоғары және арнаулы оқу орындарының «Статистика», «Қаржы», «Ақпараттық жүйелер», «Есеп және аудит», «Экономика», «Менеджмент» және т.б. экономика бағытындағы мамандықтар бойынша білім алушы студенттер мен тыңдаушыларға «STATISTICA» жүйесін игеруге және оларды мәліметтерді талдау мен қорытынды жасауға үйретуде қолдануға ұсынылады.

ӘОЖ 311(075)

КБЖ 60.6я73

ISBN 978-601-329-104-8

© З.П. Айдынов, С.А. Джумабаев

Н.С. Нуркашева, 2019

© «Фортуна Полиграф» баспасы, ЖШС, 2019

## МАЗМҰНЫ

АЛҒЫ СӨЗ.....	4
<b>1 бөлім. STATISTICA бағдарламасын мәліметтерді талдауда қолдану.....</b>	<b>6</b>
1.1 STATISTICA жұмыс терезесі.....	7
1.2 STATISTICA графикалық мүмкіндіктері.....	13
<b>2 бөлім. Корреляциялық талдау.....</b>	<b>21</b>
2.1 Корреляциялық талдау негіздері.....	21
2.2 STATISTICA бағдарламасы арқылы корреляция коэффициентін есептеу.....	23
<b>3 бөлім. Регрессиялық талдау.....</b>	<b>25</b>
3.1 Регрессиялық талдау негіздері.....	25
3.2 Регрессияның параметрлерін анықтау.....	28
3.3 STATISTICA бағдарламасы арқылы регрессиялық талдау.....	29
3.4 Регрессия параметрлерін есептеу.....	32
3.5 Функциялық батырмалар.....	33
3.6 Өзіндік жұмыстар.....	34
3.7 Көптік регрессияны STATISTICA бағдарламасы арқылы талдау.....	36
<b>4 бөлім. Кластерлік талдау.....</b>	<b>44</b>
4.1 Кластерлік талдау негіздері.....	44
4.2 Иерархиялық кластерлік алгоритмдер.....	46
4.3. STATISTICA бағдарламасын кластерлік талдауда пайдалану.....	51
<b>5 бөлім. Уақыт қатарларын талдау.....</b>	<b>54</b>
5.1 Уақыт қатарларын талдау негіздері.....	54
5.2 STATISTICA-да уақыт қатарларын талдау.....	55
5.3 STATISTICA-да уақыт қатарларын болжау.....	58

<b>6 бөлім. Дискриминантты талдау.....</b>	<b>71</b>
6.1 Дискриминантты талдау негіздері .....	71
6.2 <i>STATISTICA</i> -да дискриминантты талдауды жүргізу .....	76
6.3 Нәтижелердің гистограммасын қорыту .....	80
6.4 Дискриминантты функцияны таңдау .....	82
6.5 Дискриминантты талдаудың қадамдары .....	84
<b>7 бөлім. Нейрондық желілер және оларды қолдану .....</b>	<b>90</b>
7.1 Нейрондық желілер туралы түсінік .....	90
7.2 Нейрондық желілердің негіздері .....	91
7.3 Нейрондық желіні оқыту .....	93
7.4 Розенблатт моделі .....	93
7.5 Уақыт қатарларын нейрондық желі арқылы болжау .....	94
<b>Пайдаланған әдебиеттер .....</b>	<b>106</b>

## АЛҒЫСӨЗ

Ғылыми, экономикалық және техникалық үрдістер мен құбылыстарды математикалық модель тұрғысынан сипаттап сол арқылы негізді басқару шешімдерін қабылдай білу қазіргі мамандарға қойылатын талаптардың бірі болып табылады. Ғылыми талдау жұмыстарының басым көпшілігі статистикалық мәліметтер негізінде жүргізіледі. Бүгінде ақпаратты өңдеудегі заманауи технологиялардың мүмкіндіктері статистикалық талдауды анығұрлым жеңіл және көрнекі түрде жүзеге асыруға жағдай жасады.

Ұсынылып отырған оқу құралы «Ақпараттық жүйелер» мамандығы бойынша оқытылатын «Жаңа ақпараттық технология» пәнінің бағдарламасына сай дайындалған. Мұнда мәліметтерді талдау және болжау негіздерін көрнекі түрде жүзеге асыратын STATISTICA жүйесінің жұмыс жасау принциптері баяндалады. Аталмыш пән бойынша мемлекеттік тілдегі әдебиет бүгінде жеткіліксіз. Оқу құралын сонымен бірге экономикалық мамандық бойынша оқитын студенттер мен ізденушілер мәліметтерді талдау және болжау әдістерін игеру үшін пайдалана алады.

Бүгінде әртүрлі бағыттағы үрдістерді басқарудың сапасы барған сайын талданып отырған құбылыстарды сипаттайтын нақты сандық шамаларды анықтаумен байланысты болып отыр. Осы орайда пайдаланылатын арнайы бағдарламаларды терең білу және оны кәсіби түрғыдан қолдану маңызды. Аталмыш мәселе арнайы курстар және сапаны оқу құралдары арқылы шешіледі.

STATISTICA жүйесінің мүмкіндіктері өте зор. Ол арқылы корреляциялық-регрессиялық, кластерлік, дискриминанттық және нейрондық желілерді жүзеге асыруға болады. Талданып отырған деректердің графикалық түрде сипатталуы мүмкіндіктері жөнінен STATISTICA таңдаулы 100 бағдарламаның қатарына енді. Статистикалық тәсілдерді пайдаланған кезде тек деректерді дұрыс жинап арнайы бағдарламалар арқылы нақты бағалау ғана емес, сонымен қатар оның табиғаты мен өңдеу мақсатын түсіну маңызды.

Оқу құралы 7 бөлімнен тұрады.

Бірінші бөлімде мәліметтерді талдайтын бағдарламалардың ерекшеліктері мен пайдалану бағыттары баяндалды. Мұнда STATISTICA жүйесінің ерекшеліктері қарастырылды. Айтылуға тиіс нәрсе, құралда STATISTICA 10 нұсқасы қарастырылды.



## 1-БӨЛІМ

### *STATISTICA* жүйесін мәліметтерді талдауда қолдану

*STATISTICA* бағдарламасы 1991 жылы АҚШ *Statsoft, Inc* фирмасы жасап шығарған өнім. Қазір ол статистикалық мәліметтерді талдауда қолданылатын алдыңғы қатарлы бағдарламалардың бірі болып табылады. Оның басты құндылығы – деректерді талдауда графикалық тәсілдерге басымдылықтың берілуінде. Аталмыш мүмкіндік басқа статистикалық бағдарламаларға қарағанда елеулі артықшылықтар береді.

Бағдарламаның бірнеше нұсқалары атап айтса, *STATISTICA* 5,6,7, 8 және 10 нұсқалары бар. Біз оқу құралында *STATISTICA* 10-нұсқасын пайдаланамыз.

1994 жылы *WINDOWS* операциялық жүйесі үшін *STAT – ISTICA*-ның арнайы нұсқасы шығарылды және ол бірден бағдарламалардың көшбасшына айналды. 1995 жылы *STATISTICA* әлемдегі 100 ең тәуір бағдарламалардың құрамына енді.

Содан бері *STATISTICA* бағдарламасының бірнеше нұсқалары дүниеге келді. Айта кететін жайт, әрбір нұсқа алдыңғы нұсқалардың жетілдірілген түрі болып табылады. Ал принциптік тұрғыдан көп жағдайда өзгеріс болмайды. Сондықтан егер қолданушы бағдарламаның бір нұсқасын игерсе, ол еш қиындықсыз басқа нұсқаларында да жұмыс жасай алады.

*STATISTICA* жүйесі мынадай негізгі бөліктермен сипатталады:

- деректермен жұмыс жасайтын көпфункционалды жүйе;
- статистикалық талдаудың нәтижелерін графикалық түрде бейнелейтін жүйе.

*STATISTICA* жүйесіндегі деректермен жұмыс жасау.

*STATISTICA*-дағы деректер электрондық кесте түрінде орналасқан. Деректер\* *.sta* тектес файлдарда сақталады. Негізінде *STATISTICA* төрт түрлі құжатпен жұмыс жасайды:

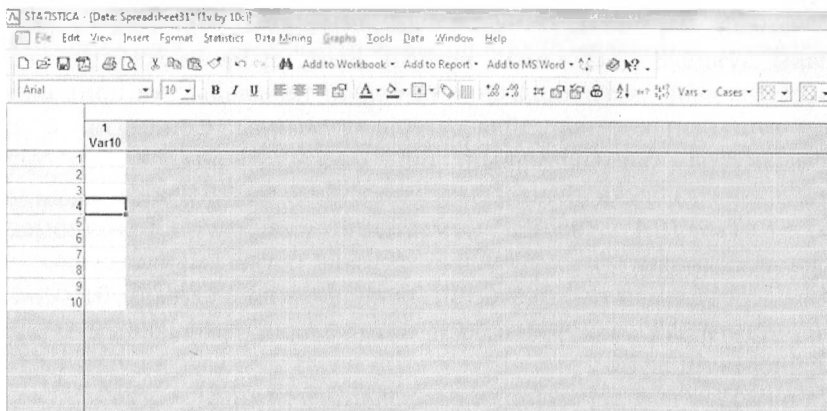
- алғашқы деректерді енгізетін және оны түрлендіретін *SPREADSHEET* электронды кестесі;
- сандық және текстілік нәтижелерді шығаратын *SCROLLSHEET* электронды кестесі;
- сандық деректерді бейнелейтін арнайы форматпен болатын - график;
- мәтіндік және графиктік деректерді сипаттайтын *RTF* форматындағы сесп.

Электрондық кесте бағаналар мен жолдардан тұрады. Мұнда бағаналар айнымалыларды - *VARIABLE*, жолдар айнымалылардың мәндерін- *CASES* сипаттайды. Айнымалылардың мах саны 4092, ал мәндердің саны-20000000 тең.

Аталмыш бағдарламаны қолдану арнайы әдебиетте кеңінен сипатталған[1,2]. Сонымен қатар өзінде де *HELP* жүйесі бар.

### 1.1 *STATISTICA* жұмыс терезесі

*STATISTICA* іске қосылғаннан кейінгі көрінісі 1-суретте берілген:



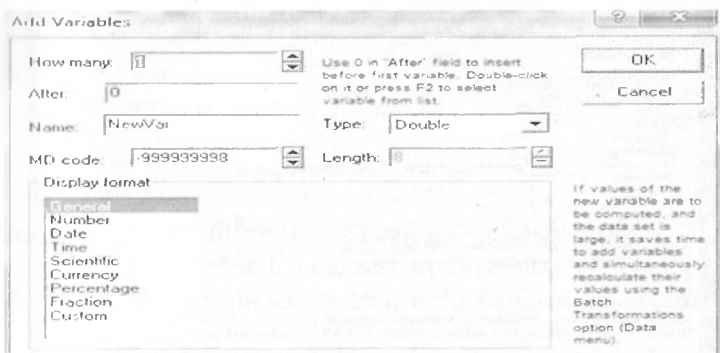
1-сурет. *STATISTICA* көрінісі



Жоғарғы жақта *Windows* бағдарламаларына сай *File, Edit, View, Insert, Format, Statistics, Graphs, Tools, Data, Window, Help* менюі орналасқан.

*STATISTICA* бағдарламасына енгізілетін деректер аты және мәнімен сипатталуға тиісті. Сондықтан мұнда деректердің аты-*Variable*, мәні-*Case* пен белгіленеді.

Айнымалыларды енгізу үшін *Insert => Add variables* командасы шақырылады(2-сурет).



2- сурет. Айнымалыны сипаттау

Мәндерді енгізу үшін *Insert => Add cases* командасы орындалады.

Шыққан панельге қанша айнымалы және қанша мән керек екені енгізіледі, сонымен қатар олардың форматтары нақтыланады. Деректерді енгізу клавиатура және тінтуір арқылы жүргізіледі. Жұмыс EXCEL кестесіндегі жұмыспен өте ұқсас.

*Деректердің файлын ашу.* Ол үшін менюдегі *File => Open Data =>* Файл аты орындалады.

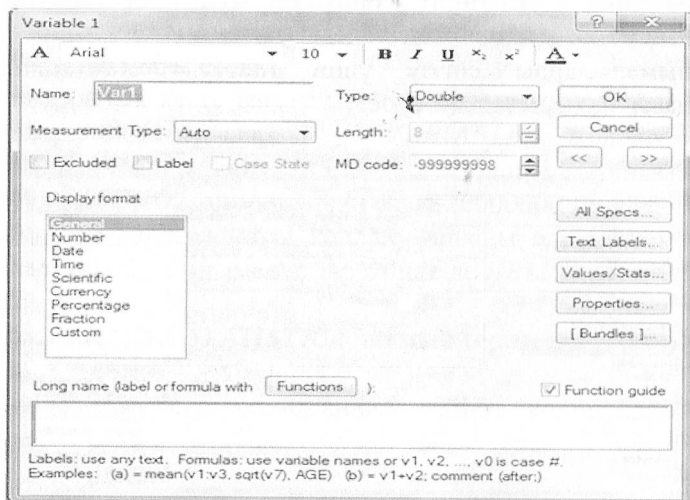
*Файлды сақтау.* *File => Save as. => Сақталатын жер* орындалады.

*Деректерді кестеге енгізу.*

Электрондық кестеге деректерді мынадай жолдармен енгізуге болады:

- клавиатура арқылы;
- енгізілген формула негізінде жаңа деректерді есептеу.

Егер курсорды айнымалы атына алып келіп, екі рет шертсе, онда жаңа терезе ашылып онда басқа айнымалылар арқылы есептейтін формула енгізуге болады (3-сурет):



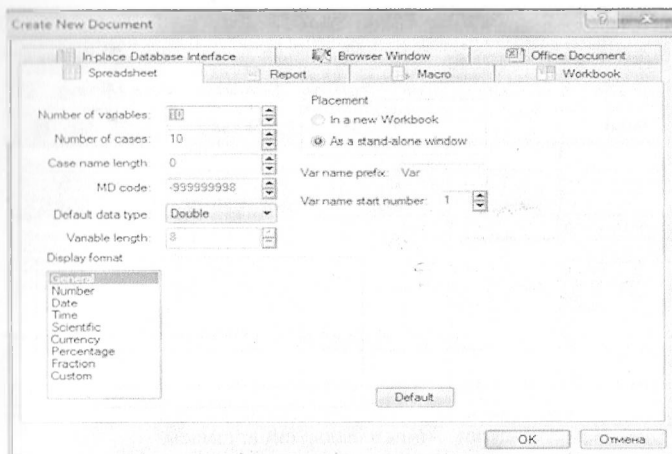
3-сурет. Айнымалылар терезесі

3-суреттегі *Functions* терезесінде басқа айнымалыларды есептеу формуласы орналасады. Сонымен қатар *STATISTICA* бағдарламасында басқа *WINDOWS*-бағдарламаларында өңделген деректерді буфер арқылы көшіруге болады.

*Алғашқы деректермен жұмыс жасау.*

*STATISTICA* жүйесінде жаңа файлды ашу үшін *FILE => NEW* командасын қолдануға немесе *CTRL + N* клавишасын басуға болады.

Аталмыш команда орындалған кезде мынадай көрініс пайда болады (4-сурет):



4-сурет. Жаңа файлды сипаттау

*Number of variables* терезешесіне айнымалылар саны, *Number of cases* терезешесіне мәндердің саны енгізіледі.  
Мысалы: Деректердің мынадай кестесі берілсін (1-кесте).

1 кесте – «*Number of cases*» терезешесіне арналған мәндер

Р/к №	Аты-жөні	Жасы	Мамандық	Мекен-жайы	Еңбекақы мың теңге
1	2	3	4	5	6
1	Ахатов А.Ж.	35	Құрылысшы	Тәуелсіздік, 28	175
2	Исаев М.Қ	49	Инженер	Сейфуллин, 21-42	120
3	Оразова А.Н	28	Дәрігер	Абай,103-47	75
4	Қосанова М.Х.	33	Мұғалім	Бейбітшілік,82-23	80
5	Сейітов Н.Б.	46	Бағдарламашы	Тайманов,21-115	225
6	Жолтаев А.Б.	53	Мұнайшы	Достық,221-57	375
7	Қадыров О.У.	38	Күзетші	Орталық, 48-123	65

1-кестеден көріп отырғандай мұнда 5 айнымалы, 7 мән берілген. Олай болса *Number of variables*-5 мәнін, *Number of cases*-7 мәнін иеленеді. Мәндерді енгізгеннен кейінгі көрініс 5-суретте:

	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5
1					
2					
3					
4					
5					
6					
7					

5-сурет. Айнымалалар сипаттамасы

5-суреттен айнымалылардың саны 5-ке, олардың мәнінің саны 7-ге тең екенін көруге болады. Көп жағдайда олардың құрылымын өзгертуге, яғни ұлғайтуға немесе кемітуге тура келеді.

*Кестенің құрылымын өзгерту.*

Кестенің құрылымын өзгерту әдетте, керек емес айнымалыларды немесе мәндерді жою, немесе қосымша айнымалылар мен мәндерді енгізу арқылы есептеледі.

*Айнымалыларды жою.*

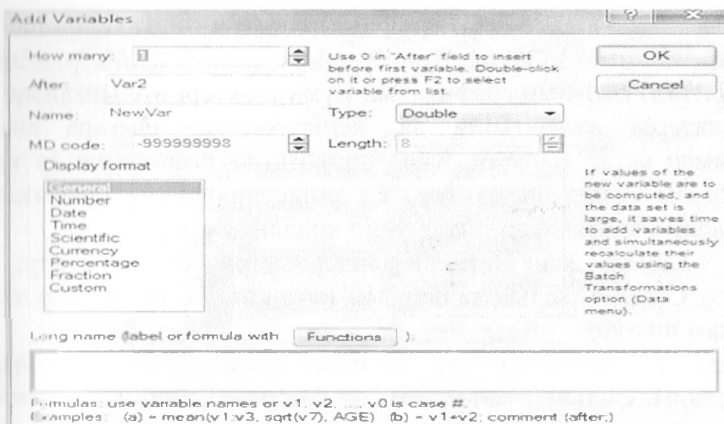
Ол үшін жойылуға тиісті айнымалылар ерекшеленіп алынады да, тіптіреуіннің оң жақ батырмасын басып *Delete – Variables* пунктін таңдайды.

*Мәндерді жою.*

Тіптіреуін арқылы керек емес мәндер ерекшеленіп алынады да, тіптіреуіннің оң жақ батырмасын басып *Delete Cases* пунктін таңдайды.

*Айнымалыларды енгізу.*

Тіптіреуін арқылы кез келген айнымалы ерекшеленеді де, тіптіреуіннің оң жақ батырмасын басып *Add Variables* пунктін таңдайды (6- сурет):

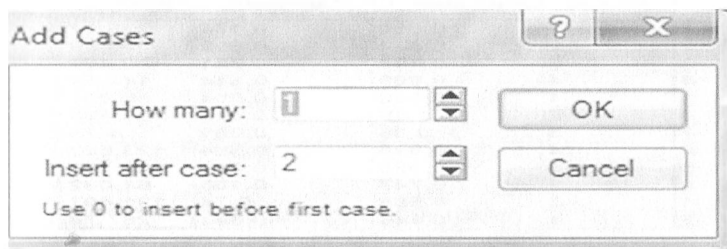


6-сурет. Айнымалыларды енгізу

6-суреттегі *How many* терезесіне қанша айнымалы, ал *After* терезесіне қай айнымалыдан кейін енгізілетіні хабарланады.

*Мәндерді енгізу.*

Тінгіреуін арқылы бірінші бағананың жолы шертіледі де, тінгіреуіннің оң жақ батырмасын басып *Add Cases* пунктін таңдайды (7-сурет):



7-сурет. Мәндерді енгізу

7-суреттегі *How many* терезесіне қанша мән, ал *Insert after case* терезесіне қай мәннен кейін енгізілетіні хабарланады.

## 1.2 STATISTICA графикалық мүмкіндіктері

STATISTICA-ның графикалық мүмкіндіктері өте ыңғайлы. Ол графиктерді жазықтықта да, кеңістікте де шығара алады. Сонымен қатар ғылыми, іскер, танымдық графиктерді әр түрлі гистограммалық, диаграммалық түріндесипаттайды. Графикалық құралдар мынадай мақсаттар үшін қолданылады:

1. Сандық және мәтіндік мәндерді көрнекі етіп шығару;
2. Статистикалық талдаудың нәтижелерін графиктер тізімі арқылы шығару.

Әдетте графиктер\*.stg тегімен сипатталатын файлдарда сақталады. Сонымен қатар егер ол растрлік файл болса \*.bmp, \*.psx файлдарында, метафайл болса \*.wmf файлдарында сақталады.

*Графиктерді тұрғызу.*

*Мысал.* Жарнамаға жұмсалатын шығын мен сатылған тауардың арасындағы байланысты сипаттайтын график тұрғызайық. Мәліметтер 8-суретте бейнеленген. Олар тоқсан бойынша сипатталған. *Жарнама 1-1* типті жарнамаға жұмсалған шығын. *Жарнама 2-2* типті жарнамаға жұмсалған шығын. *Көлем-сәйкес* тоқсандағы сатылған тауардың мөлшері.

	1 Тоқсан	2 Жарнама 1	3 Жарнама 2	4 Көлем
1	1	0.001	0.174	6.97
2	2	0.564	0.859	66.59
3	3	0.193	0.711	20.463
4	4	0.809	0.514	144.64
5	1	0.585	0.304	88.236
6	2	0.48	0.015	18.498
7	3	0.35	0.091	4.75
8	4	0.896	0.364	123.682
9	1	0.823	0.147	150.746
0	2	0.747	0.166	61.082
1	3	0.989	0.78	131.497
2	4	0.446	0.844	267.78

8-сурет. Деректер кестесі

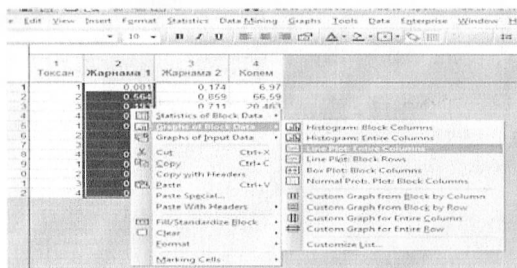
Графикті тұрғызу мынадай кадамдардан тұрады:

*1-қадам.* Қойылған талаптарға сай деректер кестеге енгізіледі.

*2-қадам.* Графигі шығуға тиісті айнымалының аты шертіліп, тіптіреуіннің оң құлағы басылады.

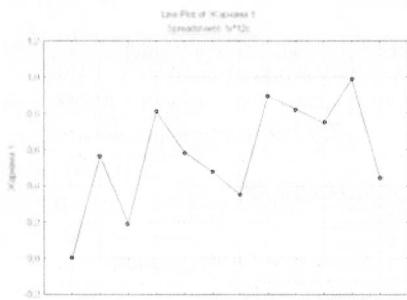
3-қадам.

*Graphs of Block Data* => *Line Plot: Entire Columns* шертіледі (9-сурет):



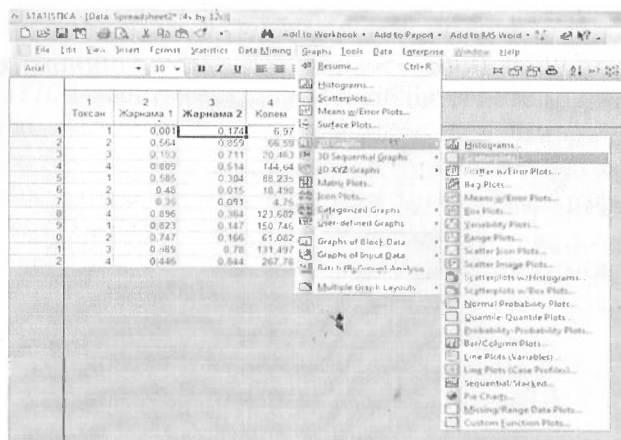
9-сурет. График тұрғызу

Мысалдағы *Жарнама 1* деректерінің шыққан графигі төмендегі 10-суретте бейнеленді:



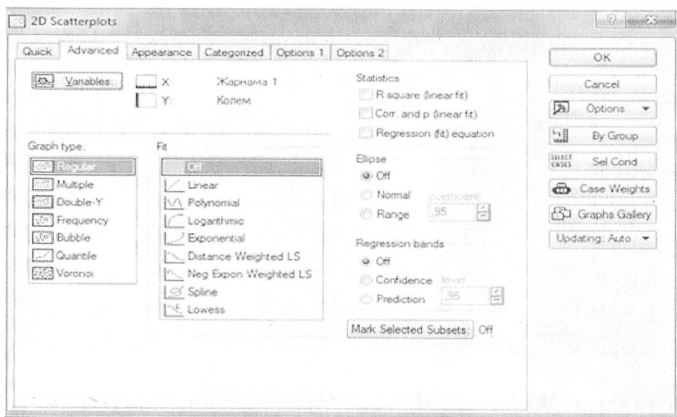
10-сурет. Деректердің графигі

Мысалдағы *Жарнама 1* мен *Көлем* арасындағы байланысты график арқылы сипаттау қажеттігі тусын. Мұнда *Жарнама 1* айнымалысы түсіндіргіш X айнымалысы, ал *Көлем* нәтижелік Y айнымалысы ретінде қарастырылады. Графикті тұрғызу үшін *Graphs* => *2D Graphs* => *Scatterplots* командасы орындалады (11-сурет):



11-сурет. График түргүзү

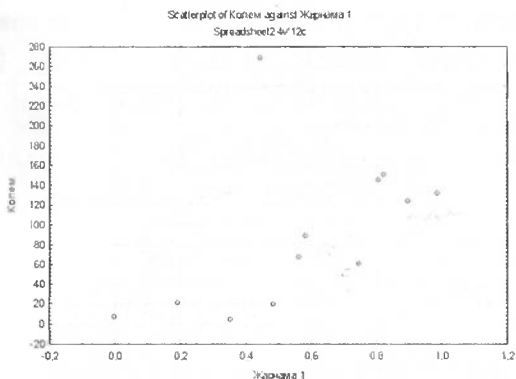
Аталмыш эрекетти орындагандан кийин тәуелсиз  $X$  айнымалысы ретинде *Жарнама 1* айнымалысы, тәуелдi  $Y$  айнымалысы ретинде *Көлем* алынады (12- сурет):



12-сурет. График параметрлерін аныктау

12-суреттегі панельге параметрлер енгізілгеннен кийин «ОК» батырмасын басамыз. Шыккан график 13-суретте бейнеленген





13-сурет. Екі өлшемді график

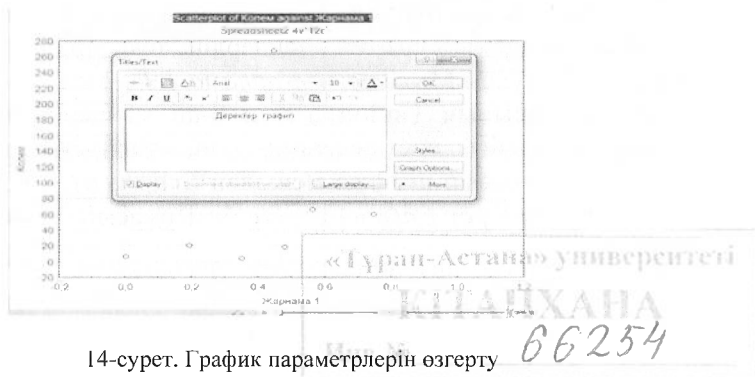
13-суреттен *Жарнама 1* айнымалысының ОХ осі бойынша, *Кoneм* айнымалысының ОУ осі бойынша орналасқанын көруге болады.

*Графиктердің элементтерін түрлендіру.*

*Statistica*-да график элементтерін түрлендіру үшін әртүрлі мүмкіндік бар. Түрлендіру көп жағдайда тінтіреуін арқылы жүзеге асады. Ол үшін тінтіреуін курсорын графиктегі қажет элементтің үстіне алып келіп тез екі рет шертеді немесе ерекшелеп алып оң жақ батырмасын басады.

*Графиктің атын өзгерту.*

Ол үшін графиктің атын екі рет шертеді (14-сурет):

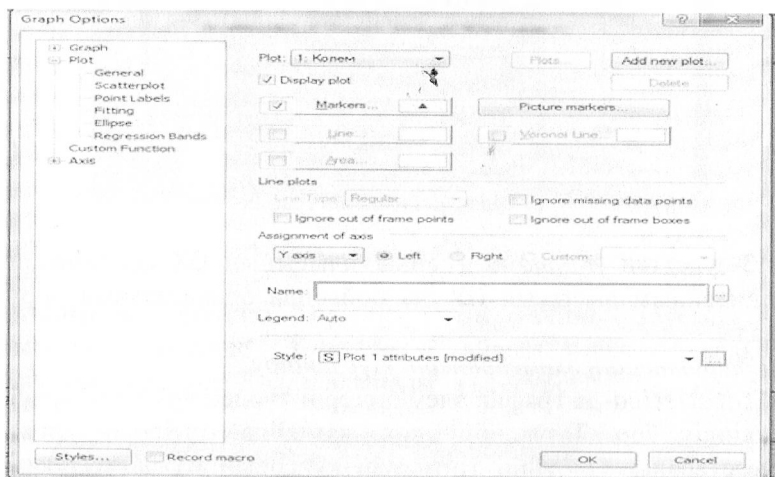


14-сурет. График параметрлерін өзгерту

Жоғарыдағы 14-суреттегі диалогтық терезеге қажетті шрифт немесе өлшемдерді енгізуге болады. Осындай бағытпен X көлденең осі мен Y тік осін өзгертуге болады.

*График маркерін өзгерту.*

Графиктің кез келген нүктесін екі рет шертіп мынадай көрініске қол жеткізуге болады (15-сурет):

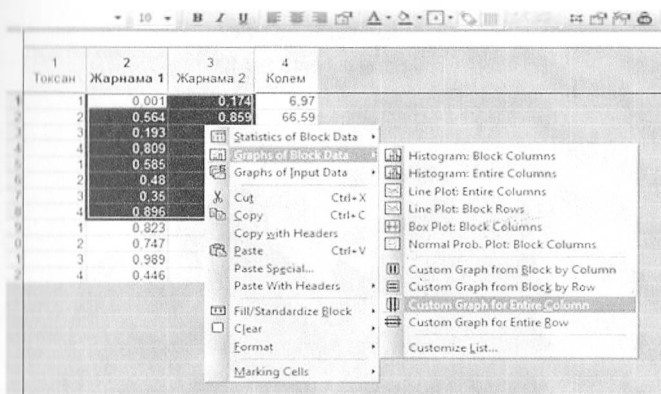


15-сурет. График параметрін өзгерту

15-суреттегі диалогтық терезедегі «*Markers*» батырмасын басып, қажетті түзетулерді енгізуге болады.

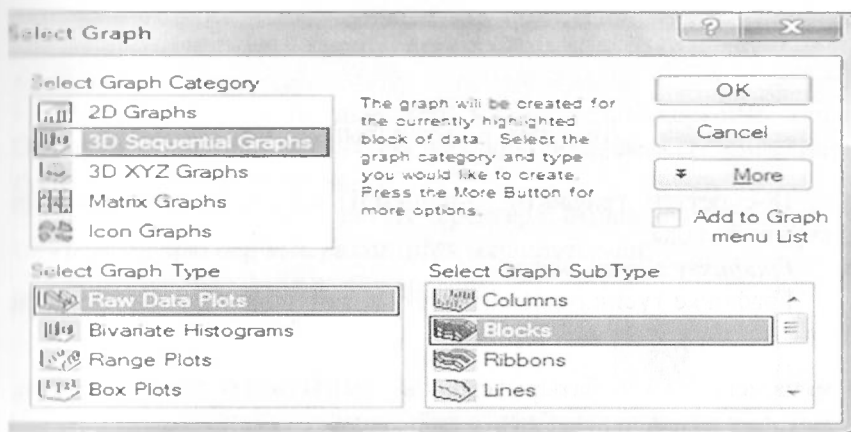
*Қолданушының графигін тұрғызу.*

Көп жағдайда берілген мәндердің тек бір бөлігін талдау қажеттігі туады. Мысалы, *Жарнама 1* мен *Жарнама 2* айнымалыларының 1-ші мән мен 8-ші мәндері ерекшеленсін. Деректер бөлігі ерекшеленген соң, тінтіреуіннің оң жақ батырмасын басып шыққан менюден *Graphs of Block Data => Custom Graph for Entire Column* батырмасын басады (16-сурет).



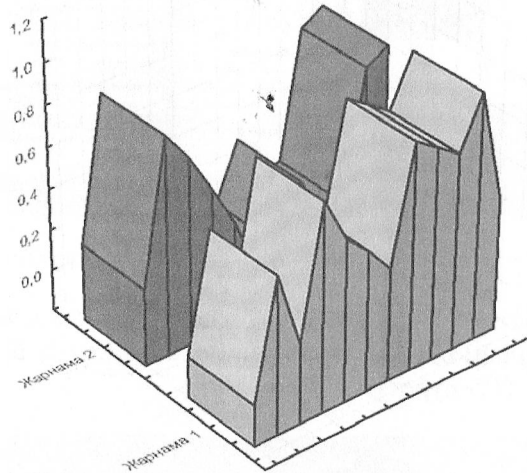
16-сурет. Қолданушының графигін тұрғызу

16-суреттегі командалар тізбегі орындалғаннан кейін *3D Sequential Graphs* => *Raw Data Plots* => *Blocks* әрекеті істелінеді (17-сурет):



17-сурет. Қолданушының графигінің параметрлері

17-суреттегі командалар орындалғаннан кейінгі шыққан график 18-суретте бейнелініп отыр:




18-сурет. Қолданушының графигі

18-суреттен графиктің 3-өлшемді кеңістікте орналасқанын көруге болады.


*Графикке мәтін енгізу.*

Графикке түсініктеме сөз енгізу қажеттігі жиі туады. Ол үшін

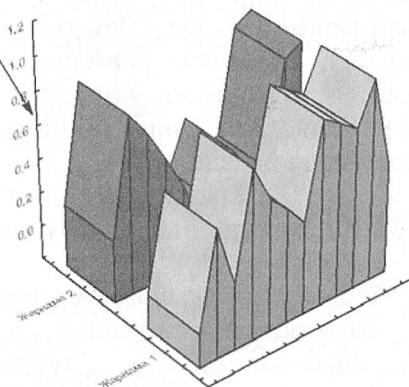
менюдегі  батырмасын басады. Қажетті мәтін енгізіліп, мәтіндік рамка керекті жерге қойылады.

*Ұштықты графикке енгізу.*

Графиктегі элементтерге түсініктеме беру үшін ұштық-бағыт

қолданылады. Оны менюдегі  Arrow –Ұштық батырмасын басу арқылы қолданады (19-сурет):

Деректер көрсеткіші



19-сурет. Ұштықты қолдану

19-суреттегі ұштық деректер мәліметтерін көрсетіп тұр. Осындай бағытпен графиктегі кез келген элементті сипаттап түсініктеме беруге болады.

*Деректердің көздері.* Қажетті деректерді былай алуға болады:

- мысалдары бар *stat/examples* маршрутымен;
- деректері бар файл жасау арқылы;
- импорттау арқылы.

### Сұрақтар мен тапсырмалар:

1. Айнымалылар мен мәндерді қалай енгізеді?
2. Деректерді енгізудің ерекшеліктерін атаңыз.
3. Мысалдары файлының маршрутын көрсетіңіз.

## 2 БӨЛІМ. КОРРЕЛЯЦИЯЛЫҚ ТАЛДАУ

Айнымалылардың арасындағы байланыстың тығыздығын және түрін сипаттау үшін *STATISTICA*-да көптеген мүмкіндіктер пайдаланылады. Мұнда Пирсон корреляция коэффициенті, Спирмен, Кендалл рангілік корреляция коэффициенттері есептеледі. Сонымен қатар мұнда сызықтық және сызықтық емес, жай және көптік регрессия модельдерінің параметрлерін есептеуге болады.

### 2.1 Корреляциялық талдау негіздері

Корреляция айнымалылардың(факторлардың) арасындағы байланыстың тығыздығын сипаттайды. Бірден айту керек корреляция байланыстың себебін түсіндірмейді, ол тек байланыстың қаншалықты мықты немесе әлсіз екенін нақты сан арқылы анықтайды. Әдетте корреляция коэффициенті  $r$  әрпімен белгіленеді және ол -1 мен 1 аралығындағы мәнді қабылдайды, яғни  $r \in [-1;1]$ . Егер мән - 1 немесе 1-ге жақындай түссе онда байланыс күшейеді, ал ол 0-ге жақындаса байланыс әлсіздене түседі. Егер коэффициент теріс таңбаға ие болса, онда айнымалылар арасында кері байланыс, ал оң таңбаға ие болса, онда айнымалылар арасында тура байланыс бар деп есептелінеді.

$X$  және  $Y$  факторларының арасындағы корреляция коэффициентінің мәнін есептеу формуласы қарапайым, ол төмендегідей:

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{S_x S_y} \quad (1)$$

мұндағы:

$n$  – мәлімет саны;

$X_i, Y_i$  –  $x$  және  $y$  айнымалыларының  $i$ -ші мәні;

$S_x, S_y$  –  $x$  және  $y$  айнымалыларының стандарттық ауытқулары;

$\bar{X}, \bar{Y}$  –  $x$  және  $y$  айнымалыларының орташа мәндері.

$r$  – корреляциялық коэффициентінің елеулілігі  $t$  – Студент критеріі арқылы анықталады. Оны анықтаудың жолы бірнеше кезеңнен тұрады:

1-кезең:  $H_0: r$ -корреляция коэффициенті елеусіз деген гипотеза ұсынылады.

2-кезең:  $t_{\text{бақыл}} = \frac{r}{\sqrt{1-r^2}} \sqrt{n-2}$  есептелінеді. Мұндағы  $n$ -

мәлімет саны.

3-кезең: егер  $t_{\text{бақыл}} < t_{\text{криз}}$  болса онда  $H_0$  гипотезасы теріске шығарылмайды, егер  $t_{\text{бақыл}} > t_{\text{криз}}$  болса онда  $H_0$  гипотезасы теріске шығарылады.

Қазіргі қолданылатын статистикалық бағдарламалар арқылы корреляция коэффициенттерінің елеулілігін бірден анықтайды. Одетте, ондай корреляция коэффициенті «қызыл» түске боялып көрсетіледі.

Жеке корреляция коэффициенті басқа факторлардың әсерін тынып тастаған кездегі  $X_i$  факторының  $Y$  факторына әсер ету күшін сипаттайды.

Ол мына формуламен есептелінеді:

$$r_{yX_1 X_2 \dots X_n} = \sqrt{1 - \frac{1 - R^2_{yX_1 \dots X_n}}{1 - R^2_{yX_1 \dots X_{j-1} \dots X_{j+1} \dots X_n}}} \quad (2)$$

мұндағы:

$R^2$  – детерминация коэффициенті.

Жеке корреляция коэффициентінің мәні де – 1 мен 1-дің аралығында жатады.

## 2.2 STATISTICA бағдарламасы арқылы корреляция коэффициентін есептеу

Бой мен салмақтың арасындағы байланысты STATISTICA бағдарламасы арқылы есептеліп көрелік. Ол үшін:

1) Деректер енгізіледі (20 сурет).

	1 Бой	2 Салмақ
1	180	83
2	172	76
3	169	78
4	166	68
5	185	84
6	158	64
7	175	72
8	166	69
9	176	77

20 - сурет. Енгізілген деректер

2) Тінтіреуіннің сол жақ құлағымен «Бой» айнымалысының бірінші мәнін шертіп, жібермеген күйде «Бой» мен «Салмақтың» барлық мәндерін ерекшелеп шығады.

3) *Statistics => Basic Statistics =>*

*Correlation matrices => OK* командалар тізбегі орындалады.

4) *Summary Correlations* – батырмасы басылып нәтижесі шығады (21 сурет).

Variable	Means	Std Dev	Бой	Салмақ
Бой	171.8889	8.176864	1.000000	0.902179
Салмақ	74.5556	6.821127	0.902179	1.000000

21 - сурет. Қорытынды көрсеткіш



21 - суреттен көріп отырғандай «Бой» мен «Салмақтың» арасында корреляция коэффициенті (0,90) орналасқан және босқа түспен боялған. Ол оның елеулі екендігін көрсетеді.

**Сұрақтар мен тапсырмалар:**

1. Корреляциялық талдаудың қажеттілігі қандай?
2. Корреляция коэффициентінің формуласының мәнісі неде?
3. Қандай факторлардың арасынан байланыс іздеуге болады?

### 3 БӨЛІМ. РЕГРЕССИЯЛЫҚ ТАЛДАУ

#### 3.1 Регрессиялық талдау негіздері

Корреляциялық талдау факторлардың арасындағы байланыстың тығыздығын зерттесе, регрессиялық талдау факторлардың арасындағы байланыстың формасын анықтайды. Әдетте алдымен корреляциялық талдау жүргізіледі, себебі факторлардың арасындағы байланыс өте әлсіз немесе жоқ болса, онда оның байланыс формасын да зерттеудің қажеті жоқ.

*Регрессия* деп  $X$  – түсіндіргіш факторы мен  $Y$  - түсіндірілетін фактор арасындағы статистикалық байланысты сипаттайтын функцияны айтады.

Ол былайша белгіленеді:

$$Y = f(x) + e \quad (3)$$

мұндағы:

$e$  - ауытқу.

Түсіндіргіш факторлардың санына орай регрессия *жұптық* және *көптік* болып екіге бөлінеді.

*Жұптық* регрессия деп бір түсіндіргіш  $X$  факторы мен бір түсіндірілетін  $Y$  факторының арасындағы статистикалық байланысты сипаттайтын функцияны айтады.

Мысалы:

$$y = a + bx + e \quad (4)$$

$$y = a + \frac{b}{x} + e \quad (5)$$

$$y = a \cdot x^b + e \quad (6)$$

$$y = a \cdot b^x + e \quad (7)$$

Мұндағы  $a$  және  $b$  регрессия параметрлері болып табылады. Олар түсіндіргіш  $X$  факторының түсіндірілетін  $Y$  факторына қаншалықты әсер беретінін сипаттайды.

*Коптік* регрессия деп бір түсіндірілетін  $Y$  факторы мен  $n$ -түсіндіргіш ( $n = 2, 3 \dots$ )  $X$  - факторларының арасындағы статистикалық функцияны айтады.

Мысалы:

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n + e \quad (8)$$

$$y = a \cdot x_1^{b_1} x_2^{b_2} \dots x_n^{b_n} + e \quad (9)$$

$X$  – түсіндіргіш факторлар мен  $Y$  – түсіндірілетін фактор арасындағы байланыстың формасына орай регрессия *сызықтық* және *сызықтық емес* болып екіге бөлінеді.

*Сызықтық* регрессия деп  $X$  – түсіндіргіш факторының дәрежесі 1-ге тең регрессияны айтады. Оған (4) және (8) өрнектер жатады.

*Сызықтық емес* регрессия деп  $X$  – түсіндіргіш факторының дәрежесі 1-ден өзгеше болатын регрессияны айтады. Оған (5), (6), (7) және (9) өрнектер жатады.

Экономикалық процестер мен құбылыстарды зерттеп талдаған кезде оларға әсер ететін факторлардың қалай және қаншалықты әсер ететінін білу өте маңызды. Осы орайда регрессиялық талдаудың орны бөлек.

*Регрессиялық талдау* – кеңінен тараған статистикалық тәсілдердің біріне жатады. Оның басты мақсаты түсіндірілетін  $Y$  факторы мен түсіндіргіш  $X$  факторы арасындағы байланысты сипаттау болып саналады. Кей жағдайда  $Y$  факторын *тәуелді* айнымалы, ал  $X$  факторын *тәуелсіз* айнымалы немесе *регрессор* деп атайды.

Мысалы, пәтер құнының қандай факторларға тәуелді екен зерттелсін. Олай болса пәтер құны түсіндірілетін  $Y$  факторы ретінде,  $X_1$  – бөлме саны,  $X_2$  - ас-үй ауданы,  $X_3$  - қабат,  $X_4$  – жөнделу т.с.с. алынады.

Аталмыш факторлар негізінде құрылатын регрессия ретінде сипатталады. Ол төменде көрсетілген:

$$Y = f(X_1, X_2, X_3, X_4) + e \quad (10)$$

Мұндағы  $e$  - ауытқу ескерілмеген факторлардың жиынтық күші болып табылады.

Регрессияның сапасын және болжамдық күшін  $R^2$  детерминация коэффициенті анықтайды.

Ол формуламен сипатталады:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y} - \bar{y})^2}{\sum_{i=1}^n (y - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e^2}{\sum_{i=1}^n (y - \bar{y})^2} \quad (11)$$

мұндағы:

$\hat{y}$  - есептеудегі мән;

$\bar{y}$  - орташа мән;

$y$  - фактілік мән;

$e$  - ауытқу.

$R^2$  -  $Y$ -тің  $x$  факторының өзгерісіне қанша пайыз тәуелді екенін білдіреді. Регрессияның сапасы  $R^2$ -1-ге жақындаған сайын артады, 0-ге жақындаған сайын кемиді.

### 3.2 Регрессияның параметрлерін анықтау

Әдетте  $a$  және  $b$  регрессия параметрлерін кіші квадраттар тәсілі арқылы анықтайды. Оны қолданудың принципі мынада:

Алдымен төмендегі анықталады:

$$e = y - \hat{y} \quad (12)$$

Регрессия параметрлерін анықтау үшін  $e$ -ауытқуларының квадраттарының қосындысы минимизациялануы керек.

Яғни,

$$\text{Ал} \quad S = \sum_{i=1}^n (y - \hat{y})^2 \rightarrow \min$$

$$\hat{y} = a + bx \quad \text{— екенін ескерсек,}$$

$$\text{Онда} \quad S = \sum_{i=1}^n (y - a - bx)^2 \rightarrow \min \quad \text{— функцияның}$$

минимумын табу  $a$  және  $b$  параметрлері бойынша туындыны тауып оны 0-ге теңестіруді қажет етеді.

$$\begin{cases} \frac{\partial S}{\partial a} = -2 \sum (y - a - bx) = 0 \\ \frac{\partial S}{\partial b} = -2 \sum (y - a - bx)x = 0 \end{cases} \quad (13)$$

Аталмыш жүйеден және белгілі түрлендіруден кейін төмендегі теңдеулер жүйесін құрамыз:

$$\begin{cases} na + b \sum x = \sum y \\ b \sum x + b \sum x^2 = \sum yx \end{cases} \quad (14)$$

Алынған теңдеудің екі жағын  $n$ -ге бөлетін болсақ онда төмендегі теңдеу қорытылады:

$$\begin{cases} a + b\bar{x} = \bar{y} \\ a\bar{x} + b\bar{x}^2 = \bar{xy} \end{cases} \quad (15)$$

мұндағы:

$a$  - коэффициенті  $X$  факторы 0-ге тең болған жағдайда  $Y$ -тің орташа есеппен қаншаға тең болатынын сипаттайды.

$b$  - коэффициенті  $X$  факторы бір бірлікке ұлғайғанда  $Y$  факторының орташа есеппен қанша бірлікке өзгеретінін білдіреді.

Мысалы:  $Y = 5,2 + 3,7x$   $R^2 = 0,67$  регрессиясы берілсін. Берілген параметрлерді пайымдау керек.

Мұнда:

$a = 5,2$ ;  $b = 3,7$   $a$  - параметрінің пайымдалуы:  $X$  факторы 0-ге тең болған жағдайда (яғни  $X$  факторы болмаса)  $Y$  факторы орташа есеппен 5,2-ге тең болады.

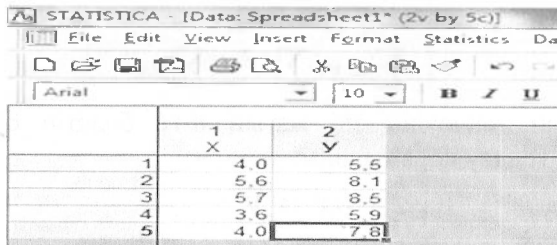
$b$  - параметрінің пайымдалуы:  $X$  факторы бір бірлікке ұлғайса,  $Y$  факторы орташа есеппен 3,7 бірлікке ұлғаяды.

$R^2$  - коэффициенті  $Y$  факторының  $X$  факторының өзгеруіне 67% тәуелді екенін білдіреді.

### 3.3 Statistica бағдарламасы арқылы регрессиялық талдау

Ол үшін мынадай іс-қимылдар жасалады:

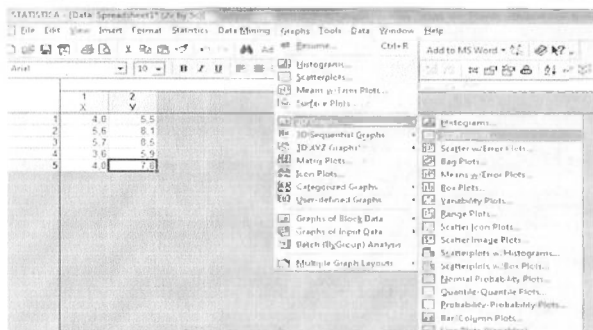
1) Деректер файлы ашылады. Керек емес бағаналар (*Var – Delete*) арқылы, керек емес бақылау жолдары (*Cases – Delete*) арқылы жойылады.  $Y$ -айнымалысы тәуелді (*Dependent*),  $X$ -айнымалысы тәуелсіз (*Independent*) арқылы сипатталады. Төмендегі кестеге қажет деректер енгізіледі (22-сурет).



	1	2
	X	Y
1	4.0	5.5
2	5.6	8.1
3	5.7	8.5
4	3.6	5.9
5	4.0	7.8

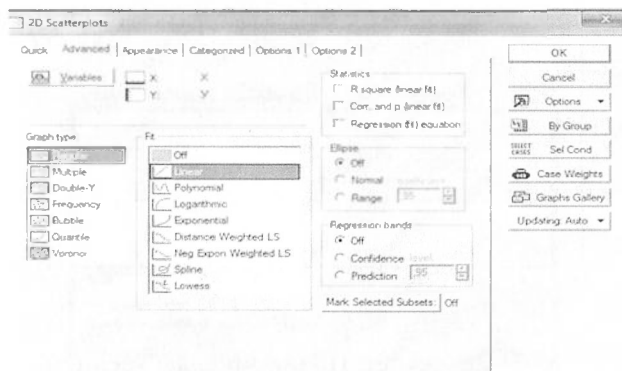
22 - сурет. Деректердің енгізілуі

Енгізілген деректердің графигі тұрғызылады. Ол үшін *Graphs => 2DGraphs => Scatterplots* командасын қолдинамыз. Ол төмендегі 23-суретте бейнеленген.



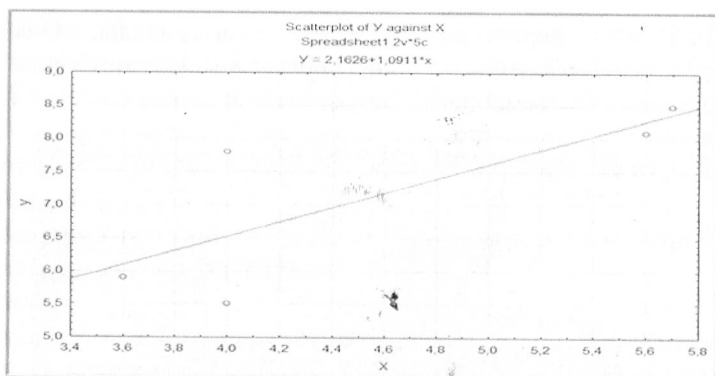
23 - сурет. Statistica-да график тұрғызу

Егер сызықтық тендеудің параметрлері қажет болс онда, *Regular => Linear => OK* командалары орындалады (24-сурет).



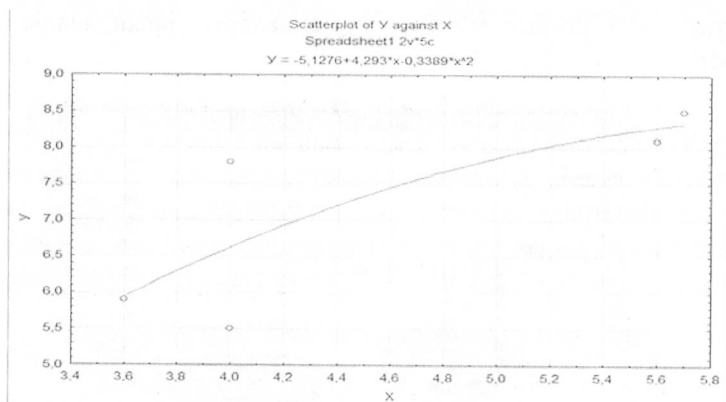
24 - сурет. График тұрғызу

Шыққан кескін 25 - суретте көрсетілді



25 - сурет. Сызықтық тренд түзуі

Егер сызықтық емес тендеудің параметрлері қажет болса онда, *Regular* => *Polynomial* => *OK* командалары орындалады (26-сурет).



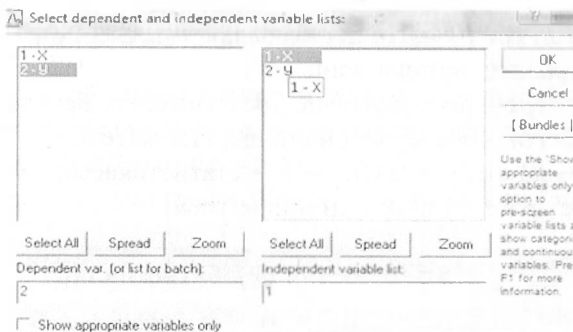
26 – сурет. Полиномиалды тренд

Жоғарыда көрсетілген 25-26 – суреттерден графиктердің барынша нүктелерді қамтитынын көруге болады.



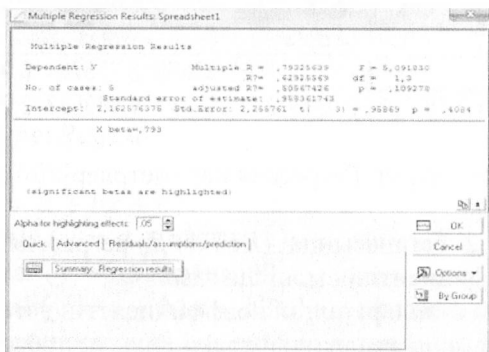
### 3.4 Регрессия параметрлерін есептеу

Регрессия параметрлерін анықтау үшін *Statistics => Multiple Regression* командасы теріледі. Мұнда  $Y$  – *Dependent* (тәуелді)  $X$  – *Independent* (тәуелсіз) айнымалылар ретінде қарастырылады (27-сурет).



27 - сурет. Тәуелді және тәуелсіз айнымалыларды анықтау

Барлығы орындалғаннан кейінгі нәтиже 28-суреттен көрсетіліп тұр.



28 - сурет. Регрессия параметрлерін анықтау

- Суреттегі нәтижелік көрсеткіштерді талдайық:
- $Y$  – тәуелді (*dep var*) айнымалы;
  - *No of cases* – бақыланған мәндердің саны ( $n=5$ );
  - *Multiple R* – көптік корреляция коэффициенті;
  - *R – square R* - детерминация коэффициенті;
  - *Adjusted R – square* - түзетілген детерминация коэффициенті;
  - *Std Error of estimate*-бақылау қателерінің квадраттық ауытқуларының орташа мәні;
  - *Intercept* - регрессияның бос мүшесінің бағасы;
  - *Std Error* - бос мүшенің стандартты қатесі;
  - *t(n – k) and p – value – t* – статистикасының бақыланған мәні және *p* - елеулілік деңгейіндегі мән.

### 3.5 Функциялық батырмалар

*Summary regression* батырмасын басқан кезде мынадай көрініс пайда болады (29-сурет).

Regression Summary for Dependent Variable: Y (Spreadsheet1)						
R= .79325639 R <sup>2</sup> = .62925569 Adjusted R <sup>2</sup> = .50567426						
F(1,3)=5.0918 p< .10928 Std Error of estimate: .95836						
	Beta	Std Err. of Beta	B	Std Err. of B	t(3)	p-level
N=5						
Intercept			2.162576	2.255761	0.958690	0.408439
X	0.793256	0.351542	1.091141	0.483553	2.256508	0.109278

29 - сурет. Регрессия параметрлерінің кестесі

Кестенің 2 бағанасында (*BETA*)  $\beta_0$ -регрессияның стандартталған коэффициентінің мәні шығады.

3-бағанада стандартталған коэффициенттің қатесі орналасқан.

4-бағанада кіші квадраттар тәсілі арқылы табылған регрессия коэффициенттері түзілген.

5-бағанада регрессия коэффициенттерінің стандартты қателері көрсетілген.

6-бағанада  $t$ -статистикасының есептелген деңгейлері берілген.

7-бағанада елеулілік деңгейлері берілген. Мұнда  $p > \alpha$  бұл регрессияның елеусіз екендігін білдіреді. Себебі  $\alpha=0,05$ .

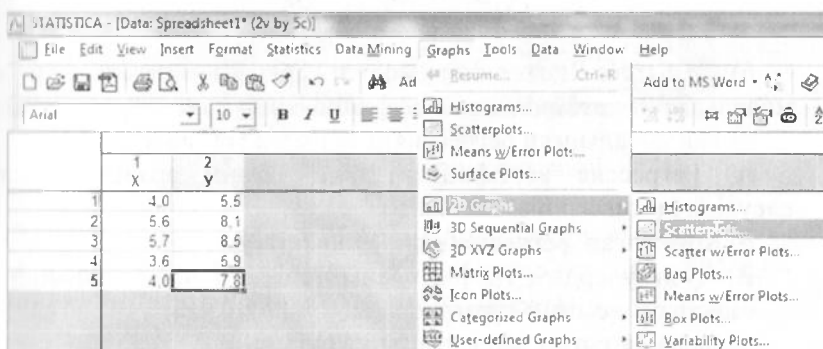
### 3.6 Өзіндік жұмыстар

STATISTICA бағдарламасын қолдана отырып регрессиялық талдау жүргізіңіз.

1)  $X$  және  $Y$  айнымалыларының мәндері енгізіледі. Мұнда  $X$  тәуелсіз  $Y$  тәуелді айнымалы ретінде қарастырылады.

2) *Graphs* => *2Dgraphs* => *Scatterplots*.

3) Айнымалылардың графигінің түрі таңдалынады (30-сурет).



30 - сурет. График тұрғызу

4) *Graph Type: Regular*

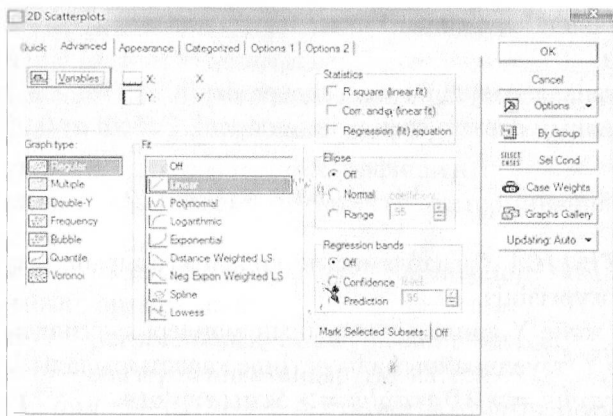
5) *Fit* (модель):

- *Linear* ( $y = a + bx + e$ );

- *Logarithmic* ( $y = a + blgx + e$ );

- *Exponential* ( $y = abexp(bx)\%$ )

- *Polinomial* ( $y = a + b1x + b2x2 + .. + bnxn + e$ ) модельдерін қарап екі барынша ұқсайтын модельді таңдаңыз (31 - сурет).



31 - сурет. Регрессия моделін таңдау

6) *Multiple Regression* модулі арқылы таңдалынған екі модельдің  $R^2$  детерминация коэффициентінің үлкенін анықтаңыз.

7) Таңдап алынған регрессияға талдау жүргізіңіз:

А) Регрессия коэффициенттерінің сенім аралығын және елеулілігін тексеріңіз.

Б) Қортылған регрессияның  $R^2$  анықтаңыз.

В) Ауытқулардың графигін анықтаңыз. Мұнда олардың:

- дисперсиясының тұрақтылығын;
- арасында тығыз байланыстың жоқтығын;
- орташа мәні 0-ге тең екендігін;
- калыпты тарағанын тексеріңіз.

### 3.7 Көптік регрессияны *Statistica* бағдарламасы арқылы талдау

Аталмыш жұмыс бірнеше қадамнан тұрады.

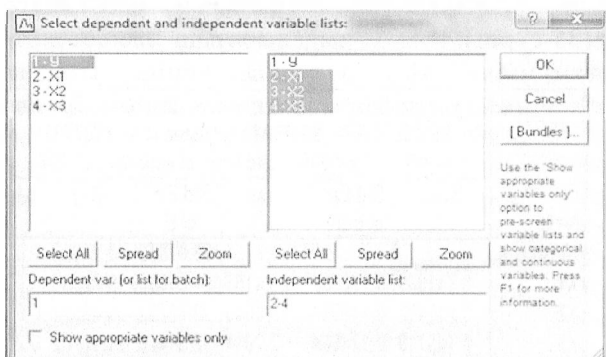
1 қадам. Деректер енгізіледі (32 - сурет):

	1 Y	2 X1	3 X2	4 X3
1	79.3	2.5	10.0	3.0
2	200.1	5.5	8.0	6.0
3	163.2	6.0	12.0	9.0
4	200.1	7.9	7.0	16.0
5	146.1	5.2	8.0	15.0
6	177.7	7.6	12.0	9.0
7	30.9	2.0	12.0	8.0
8	291.9	9.0	5.0	10.0
9	160.0	4.0	8.0	4.0
10	339.4	9.6	5.0	16.0
11	159.6	5.5	11.0	7.0
12	88.3	3.0	12.0	8.0
13	237.5	6.0	6.0	10.0
14	107.2	5.0	10.0	-1.0
15	155.0	3.5	10.0	-1.0

32- сурет. Деректердің енгізілуі

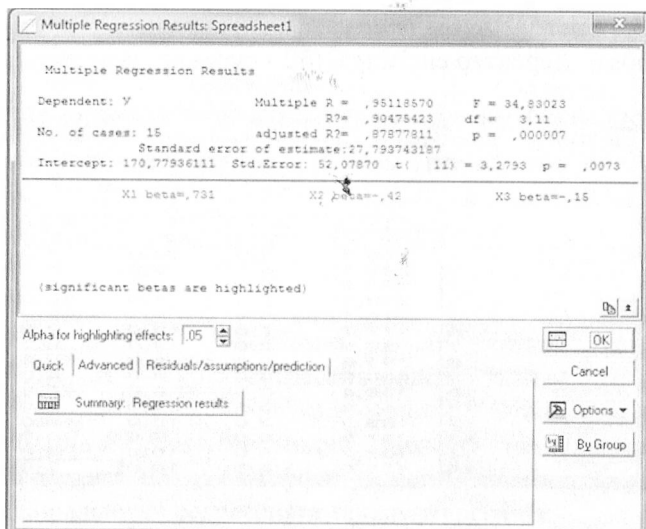
2 қадам. *Statistics* => *Multiple Regression* теріледі.

3 қадам. *Variables* – Y – dependent, X1, X2, X3 – independent => OK таңдалынады (33 - сурет).



33-сурет. Айнымалылардың анықталуы

4 қадам. Регрессия қорытындысы бөлек панельге шығады (34 - сурет).



34 - сурет. Регрессия параметрлерінің анықталуы

5 қадам. Панельдегі «Summary Regression results» батырмасы басылады.

6 қадам. Қорытылған регрессия параметрлері келесі панельге шығады (35-сурет).

Regression Summary for Dependent Variable: Y (Spreadsheet1)						
R= .95118570 R <sup>2</sup> = .90475423 Adjusted R <sup>2</sup> = .87877811						
F(3,11)=34.830 p<.00001 Std Error of estimate: 27.794						
N=15	Beta	Std. Err. of Beta	B	Std. Err. of B	t(11)	p-level
Intercept			170.7794	52.07870	3.27926	0.007343
X1	0.731089	0.139897	25.4163	4.86350	5.22592	0.000283
X2	-0.415192	0.118994	-13.0049	3.72722	-3.48918	0.005066
X3	-0.145719	0.124864	-2.7014	2.31476	-1.16702	0.267878

35 - сурет. Регрессия параметрлерінің кестесі

Көптік регрессия  $Y = 170,77 + 25,41X_1 - 13,0X_2 - 2,7X_3$  ретінде сипатталынады. Кестеден  $b_1$  және  $b_2$  коэффициенттерінің елеулі, ал  $b_3$  коэффициентінің елеусіз екені көрініп тұр. Себебі оған сәйкес  $p=0,26 > \alpha = 0,05$ . Сонымен қатар елеулі параметрлер басқа түске боялып тұрады.

Жалпы регрессияның елеулілігі Фишер критерісі негізінде анықталады.

Мұнда  $F_{\text{бақыл}}$  мәні  $F_{\text{криз}}$  мәнімен салыстырылуы керек. Салыстыру дисперсиялық талдау арқылы тексеріледі.

Ол үшін *Analysis of Variance (ANOVA)* батырмасы басылады (36-сурет).

Analysis of Variance, DV: Y (Spreadsheet1)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	80718.24	3	26906.08	34.83023	0.000007
Residual	8497.41	11	772.49		
Total	89215.66				

36 - сурет. Дисперсиялық талдау кестесі

36 - суреттен көріп отырғандай  $F_{\text{бақыл}}=34,83$  ал  $F_{\text{криз}0,05}=3,59$  демек  $p < \alpha=0,05$  негізінде  $F_{\text{бақыл}} > F_{\text{криз}}$ , яғни регрессия елеусіз деген гипотеза теріске шығарылады.

Регрессиядағы үшінші фактор елеусіз болғандықтан, оны регрессиядан шығарып тастап екі фактормен қайта есептейміз (37 - сурет).

Regression Summary for Dependent Variable: Y (Spreadsheet1)						
R= .94496647 R <sup>2</sup> = .89296162 Adjusted R <sup>2</sup> = .87512189						
F(2, 12)=50.055 p<.00000 Std Error of estimate: 28.210						
N=15	Beta	Std. Err. of Beta	B	Std. Err. of B	t(12)	p-level
Intercept			159.9005	52.00457	3.07474	0.009632
X1	0.643932	0.120065	22.3863	4.17404	5.36321	0.000170
X2	-0.400151	0.120065	-12.5338	3.76075	-3.33279	0.005967

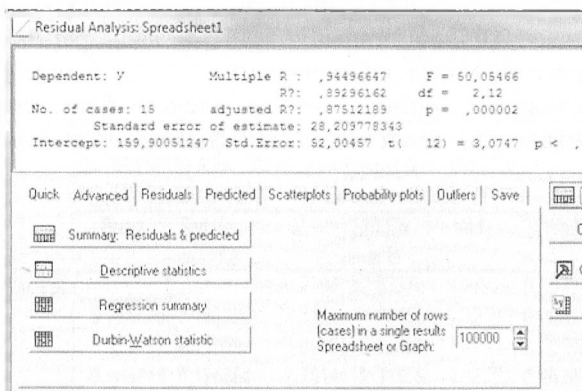
37 - сурет. Регрессия параметрлерінің кестесі

Регрессия теңдеуі  $Y = 159,9 + 22,38x_1 - 12,53x_2$

Екі фактордың да елеулі екені кестеден көрініп тұр. Регрессияның тиянақтылығын ауытқулар арқылы дәлелдеуге болатынын айттық. Ол үшін *Residual Analysis* батырмасын пайдаланамыз.

Алдымен ауытқулардың арасында тығыз байланыс болмау керектігін тексерейік. Ол үшін Дарбин-Уотсон статистикасы қолданады.

Ол *Residual Analysis* => *Durbin - Watcon statistic* командасы арқылы орындалады (38 - сурет).



38 - сурет. Ауытқуларды талдау



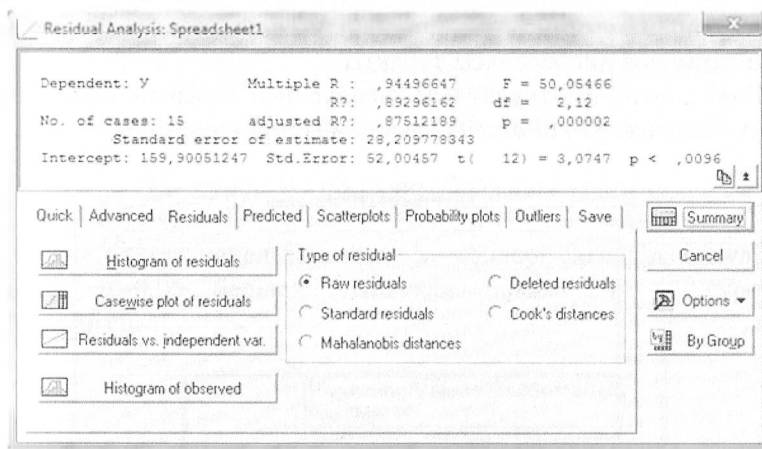
Дарбин-Уотсон статистикасының нәтижесі 39 - суретте берілген:

	Durbin-Watson d (Spre and serial correlation of	
	Durbin- Watson d	Serial Corr.
Estimate	1.898009	-0.059352

39 - сурет. Дарбин-Уотсон статистикасы

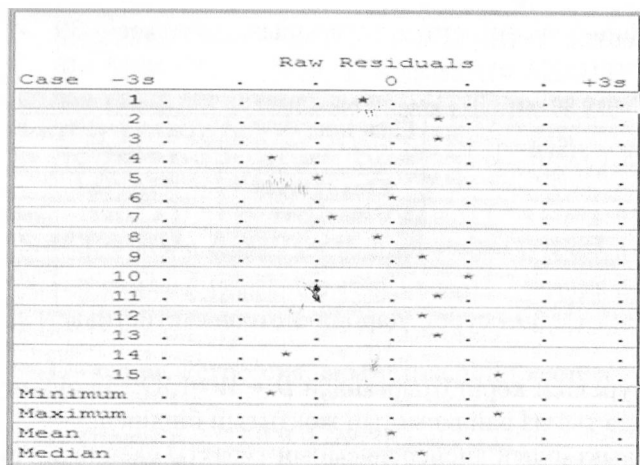
39-суреттен көріп отырғандай  $D - W = 1,89$ . Ол ауытқулардың арасында тығыз байланыстың жоқтығын білдіреді.

Қалдықтардың дисперсиясының тұрақты екендігін тексереміз. Ол үшін қалдықтардың графигі тұрғызылады. Ол *Residuals => Casewise plot* командасы арқылы жүзеге асады (40-сурет).



40 - сурет. Ауытқуларды талдау

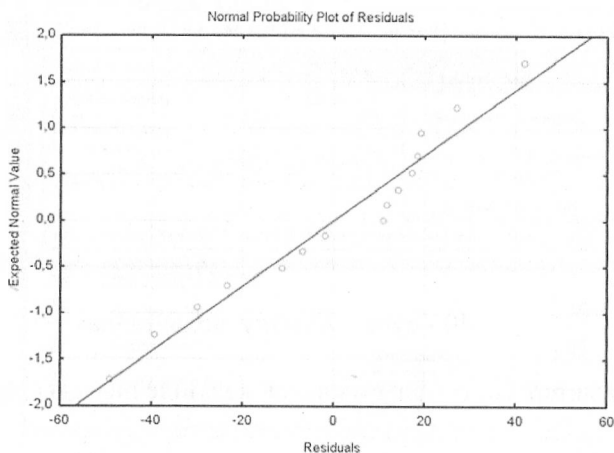
Қалдықтардың графигі мынадай кескінде шығады (41 - сурет).



41 - сурет. Ауытқулардың графигі

41 - суреттен қалдықтардың барлығы нольдік сызықтан  $\pm 2S$  ауытқуының ортасында жатқанын көруге болады, демек қалдықтардың дисперсиясы тұрақты.

Ендігі шарт ауытқулардың таралуының қалыпты болуы. Ол үшін панельден *Probability* кескін 42- суретте көрсетілді:

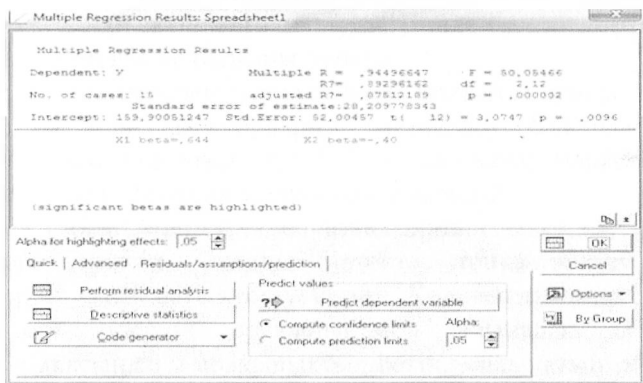


42 - сурет. Ауытқулардың қалыпты болу графигі

42-суреттен нүктелердің нормаль түзуінің бойына жақын орналасқанын көруге болады. Олай болса қалдықтар қалыпты әрімен таралған. Аталмыш жағдайларды айта келіп модельдің тиімді екенін дәлелдедік.

Енді тиісінше модельді болжау жұмысына пайдалануға болады.

Ол үшін *Multiple Regression => Predict Dependent Var* командасы орындалады (43 - сурет).



43 - сурет. Регрессияны болжауға қолдану

Пайда болған панельге  $x_1$  және  $x_2$  мәндері енгізіліп, елеулілік деңгейі  $\alpha=0,05$  беріледі. Шыққан нәтиже 44 - суретте бейнеленген:

Variable	Predicting Values for (Spreadsheet variable) Y		
	B-Weight	Value	B-Weight * Value
X1	22.3863	2.50000	55.966
X2	-12.5338	10.00000	-125.338
Intercept			159.901
Predicted			90.528
-95.0%CL			62.393
+95.0%CL			118.664

44 - сурет. Болжамдық мән кестесі

Мұнда болжанған мән *Predicted* (Y) жолында көрсетілген. Ол 90,528-ге тең. 95 пайыз сеніммен болжанған мән 62,4 пен 118,6 аралығында жатады.

**Сұрақтар мен тапсырмалар:**

1. Регрессияның мәнісі мен түрлерін сипаттаңыз.
2. Регрессия параметрлерін табу жолын көрсетіңіз.
3. Регрессия елеулілігін қалай табады?

## 4-БӨЛІМ. КЛАСТЕРЛІК ТАЛДАУ

### 4.1 Кластерлік талдау негіздері

Объектілерді зерттеп талдаған кезде олардың ерекшеліктеріне қарап белгілі бір топқа немесе жиынға сұрыптаудың қажеттігі туып отырады. Топтарға жіктеудің бірнеше тәсілдері бар. Солардың бірі *кластерлік* тәсіл.

Кластерлік талдаудың міндеті «класс», «таксон» деп аталатын топтарға ұқсас объектілерді жинақтау. Пайда болған топтардың белгілі бір мазмұнға ие болғаны дұрыс.

Кластерлік талдаудың нәтижесін көп жағдайда пайдаланады. Мысалы, медицинада ауруларды кластерлеу оларды емдеу жолдарын анықтауға алып келеді. Сонымен қатар маркетингтік зерттеулерде де кластерлік талдау қолданылады.

Бізге  $p$ -белгілерімен сипатталынатын  $x_1, x_2, \dots, x_n$  объектілері берілсін. Оларды біртектес топқа жіктеу үшін объектілердің ұқсастығын білу қажет. Көп жағдайда ұқсастық ретінде объектілердің арақашықтығы алынады. Ол екі объектінің арасындағы арақашықтық ретінде алынатын  $\rho(x_i, x_j)$  арқылы сипатталынады.

Мұнда мынадай шарттар орындалуға тиісті:

- а)  $\rho(x_i, x_j) = 0$  егер  $x_i = x_j$ ;
- б)  $\rho(x_i, x_j) = \rho(x_j, x_i)$ ;
- в)  $\rho(x_i, x_j) \leq \rho(x_i, x_k) + \rho(x_k, x_j)$ .

Ара қашықтық ұғымы «метрика» сөзімен сипатталады. Объектілердің арақашықтығын анықтау үшін бірнеше тәсіл қолданылады. Тәсілді таңдау деректердің ерекшелігіне, зерттеу мақсатына тәуелді болады. Жиі қолданылатын тәсілдерді қарастырайық.

1. Евклидтік арақашықтық.

Ол мұндағы  $x_{ij}$  –  $i$ -ші объектінің  $j$ -ші белгісі болып саналады:

$$\rho(x_j, x_j) = \sqrt{\sum_{i=1}^k x_{ij} - x_{jk}}^2 \quad (20)$$

Егер объектілердің өлшем бірліктері әртүрлі болса онда евклидтік өлшем қисынсыз болып кетуі мүмкін. Сондықтан оларды «нормалау» (стандарттау) керек.

Ол мына формулалары арқылы орындалады:

$$z_i = \frac{x - \bar{x}}{\sigma}; \quad z_i = \frac{x}{\bar{x}}; \quad z_i = \frac{x - \bar{x}}{x_{\max} - x_{\min}} \quad (21)$$

мұндағы,  $z_i$  – нормаланған мән;

$x$  – берілген мән;

$\bar{x}$  – орташа мән;

$x_{\max}, x_{\min}$  – ең үлкен және ең кіші мән;

$\sigma$  – стандартты ауытқу.

1. «Таразыланған» евклидтік ара қашықтық мына формуласымен есептелінеді:

$$\rho(x_i, x_j) = \sqrt{\sum_{k=1}^n \omega_k (x_{ik} - x_{jk})^2} \quad (22)$$

мұндағы  $\omega_k$  –  $k$  белгісінің салмағы, әдетте  $0 \leq \omega_k \leq 1$ . Мұнда әрбір белгінің маңыздылығына орай «салмақ» беріледі.

1. Махаланобисара қашықтығы:

$$\rho(x_i, x_j) = \sqrt{(X_i - X_j)^T \Delta^{-1} (X_i - X_j)} \quad (23)$$

мұндағы:

$\Delta$  – мәліметтер алынатын басты жиынның ковариациялық матрицасы;

$\Delta$  – диагоналда орналасатын «таразылық» коэффициенттердің симметриялық матрицасы.

City – block арақашықтығы мына формуласымен есептеледі:

$$\rho(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (24)$$

Ара қашықтықтар мен таразы коэффициентін анықтау кластерлік талдаудағы маңызды кезең. Ұқсастықты анықтау үшін сызықты корреляция және рангілік корреляция коэффициентін пайдалануға болады.

## 4.2 Иерархиялық кластерлік алгоритмдер

«Иерархия» сөзі «деңгей» немесе «терек» деген ұғымды білдіреді. Иерархиялық алгоритмнің *агломеративті* және *дивизимді* екі түрі бар. Агломеративті тәсілдің мәнісі бір элементтен тұратын кластерлерді ұқсастығына орай біріктіре келіп, оларды үлкен бір кластерге топтау. Дивизимдік тәсілдің мәнісі керісінше үлкен кластерді белгілеріне орай кіші кластерлерге бөліп, соңында бір элементтен тұратын кластерлерге әкелу.

Агломеративті алгоритмді қарастырайық.

Мұнда алдымен қарастырылатын  $X_i (i=1,2,3...n)$  объектісі жеке кластер ретінде қарастырылады. Кейін алгоритмнің әрбір қадамында екі жақын кластер біріктіріледі және кластерлердің арақашықтығы қайта есептеліп отырады. Аталмыш жұмыс барлық объектілер үлкен бір кластерге біріктірілгенде аяқталады.

Кластерлік талдау объектілердің арасындағы арақашықтықты да пайдаланады. Ол қашықтық бірнеше тәсілмен есептеледі.

Мысалы  $S_i$ -іші класс(топ, кластер) болсын,  $n_i - i$ -ші кластағы элементтер саны,  $\bar{x}(i)$ -і-ші кластың салмағының «орташасы». Ол мына формуласымен есептелінеді:

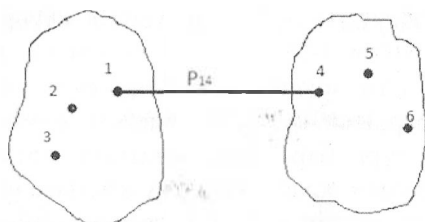
$$\bar{x}_j(i) = \frac{1}{n} \sum_{k=1} x_{jk} \quad (25)$$

Кластерлердің арақашықтығы мынадай тәсілдермен есептеледі:

1. «Жақын көрші»

$$\rho_{\min}(S_i, S_n) = \min \rho(x_i, y_j) \quad (26)$$
$$x \in S_i; y \in S_n$$

Мұнда екі кластердің арақашықтығы екі кластерге тиесілі ең жақын орналасқан элементтердің арақашықтығы арқылы анықталады(45-сурет).



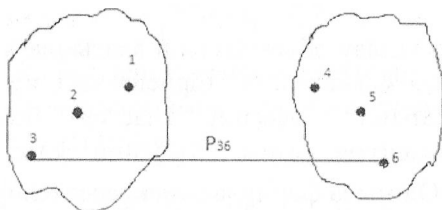
45-сурет. Жакын көрші тәсілі

2. «Алыс көрші» тәсілі

$$\rho \max(S_i, S_n) = \max \rho(X_i, Y_j) \quad (27)$$

$$X_i \in S_i; Y_j \in S_n$$

Мұнда арақашықтық екі кластерге тиесілі ең алыс орналасқан объектілердің арақашықтығы арқылы есептеледі(46-сурет).



46-сурет. Алыс көрші тәсілі

3. «Орташа байланыс» тәсілі.

Мұнда кластерлердің ұқсастығы олардың элементтерінің ұқсастығының орташа мәнімен есептеледі:

$$\rho(S_n, S_m) = \frac{1}{n_i m_m} \sum \sum \rho(X_i Y_i) \quad (28)$$

Кластерлерді біріктіру  $dus$  мәндері арқылы жүргізіледі.

Ол  $dus = \alpha p d p s + \alpha q d q s + \beta d p q + \gamma |d p s - d q$

мұндағы:

$d p s, d q s, d p q$  кластерлердің арасындағы арақашықтық;

$\alpha p, \alpha q, \beta, \gamma$ -кластерлерді біріктіру кезіндегі алгоритмге байланысты параметрлер. Олардың мәндері 2-кестеде берілген.



2 кесте - Кластерлік параметрлер кестесі

Р/к №	Тәсіл	Параметрлер
1	2	3
1	Жақын көрші	$\alpha_p = \frac{1}{2}, \alpha_q = \frac{1}{2}, \beta = 0 \left\{ \begin{array}{l} \gamma = -\frac{1}{2} \\ \gamma = \frac{1}{2} \end{array} \right.$
2	Алыс көрші	$\alpha_p = \alpha_q = \frac{1}{2}, \beta = 0 \left\{ \begin{array}{l} \gamma = -\frac{1}{2} \\ \gamma = -\frac{1}{2} \end{array} \right.$
3	Орташа байланыс	$\alpha_p = \frac{n_p}{n_p + n_q}, \alpha_q = \frac{n_q}{n_p + n_q}, \beta = \gamma = 0$

Мысал. Үш айнымалысы бар  $X_1$ -негізгі қорлардың орташа жылдық құны, млрд теңге,  $X_2$ -өндірілген өнімнің 1 теңгеге кеткен шығыны, тиын,  $X_3$ -өндірілген өнімнің көлемі, млрд теңге бес кәсіпорынды кластерлеу керек.

Мәліметтер 3-кестеде берілген.

3 кесте - Мәліметтер кестесі

№	$X_1$	$X_2$	$X_3$
2			
1	120,0	94,0	164,0
2	85,0	75,2	92,0
3	145,0	81,0	120,0
4	78,0	76,8	86,0
5	70,0	75,9	104,0
Орташа мән	99,6	80,6	113,2
Орташа квадраттық ауытқу	28,4	10,9	27,9

Айнымалылар әртүрлі өлшем бірліктерімен сипатталатын болғандықтан берілген деректерді нормалайды.

Ол мына формуласымен есептеледі:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (29)$$

Нормаланған мәліметтердің матрицасы төмендегідей:

$$Z = \begin{pmatrix} 0,718 & 1,229 & 1,821 \\ -0,514 & -2,238 & -0,760 \\ 1,598 & 0,037 & 0,244 \\ -0,760 & -0,349 & -0,975 \\ -1,042 & -0,431 & -0,330 \end{pmatrix} \quad (30)$$

Объектілердің арақашықтығын сипаттайтын матрица евклидік ара қашықтықпен есептеледі. Мысалы бірінші мен екінші объектінің ара қашықтығы:

$$d_{12} = \left\{ [0,718 - (-0,514)]^2 + [1,229 - (-2,238)]^2 + [1,821 - (-0,760)]^2 \right\}^{1/2} = 4,49 \quad (31)$$

$D_0$ -матрицасы жеке кластер болып табылатын әрбір объектінің арасындағы арақашықтықты сипаттайды:

$$D_0 = \begin{pmatrix} 0 & 4,49 & 2,16 & 3,53 & 3,24 \\ & 0 & 3,26 & 1,92 & 1,93 \\ & & 0 & 2,68 & 2,74 \\ & & & 0 & 0,71 \\ & & & & 0 \end{pmatrix} \quad (32)$$

$D_0$  матрицасын қарасақ ең жақын объект болып  $n_4$  және  $n_5$  объектілері табылады  $d_{4,5}=0,71$ . Оларды біріктіріп кластерге  $S_4$ -нөмірін беруге болады. Енді қалған объектілердің  $S_4$ -кластермен

ари қашықтығын «алыс көрші» тәсілімен есептеуге болады. Мысалы  $n_1$  объектісі мен  $S_4$  кластердің арақашықтығы  $d(S_1, S_4) = \max\{d_{14}, d_{15}\} = \max\{3,53; 3,24\} = 3,53$

Шыққан матрица:

$$D_1 = \begin{pmatrix} 0 & 4,49 & 2,16 & 3,53 \\ & 0 & 3,26 & 1,93 \\ & & 0 & 2,74 \\ & & & 0 \end{pmatrix} \quad (33)$$

$D_1$  матрицасында ең жақын кластерлер болып  $S_2$  және  $S_4$  табылады. Себебі  $d_{24} = 1,93$ . Сондықтан  $S_2$  мен  $S_4$  біріктіріледі демек  $n_2 n_4 n_5$  объектілері болады және  $S_2$  кластері ретінде белгілейді. Нәтижесінде үш кластер  $S_1(1)$ ,  $S_2(2,4,5)$  және  $S_3(3)$  пайда болады.  $d_{12} d_{23}$  ара қашықтығын есептеу арқылы  $D_2$  матрицасын қорытамыз.

$$d_{12} = \max\{d_{1,2}, d_{1,4}\} = \max\{4,49; 3,53\} = 4,49.$$

$$d_{23} = \max\{d_{2,3}, d_{3,4}\} = \max\{3,26; 2,74\} = 3,26.$$

$$D_2 = \begin{pmatrix} 0 & 4,49 & 2,16 \\ & 0 & 3,26 \\ & & 0 \end{pmatrix} \quad (34)$$

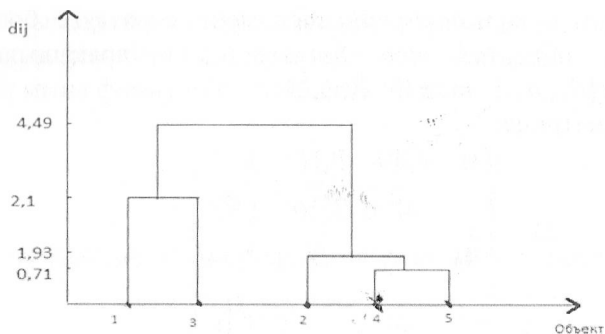
$D_2$  матрицасына қарап  $S_1$  және  $S_3(d_{1,3}=2,16)$  бір кластерге біріктіріліп  $S_1$  нөмірін береміз. Енді екі кластер берілген:

$S_1$ -кластері ( $n_1, n_3$ )  $S_2$ кластері ( $n_2, n_4, n_5$ )

$$d_{12} = \max\{d_{1,2}, d_{3,2}\} = \max\{4,49; 3,26\} = 4,49.$$

$$D_3 = \begin{pmatrix} 0 & 4,49 \\ & 0 \end{pmatrix} \quad (35)$$

Соңғы қадамда  $S_1$  және  $S_2$  кластерлері 4,49 арақашықтығы бойынша бір кластерге бірігеді. Оның дендрограммасы 45-суретте берілген.



45-сурет. Кластердендрограммасы

Сонымен қатар объектілерді кластерге біріктіруде мынадай тәсілдер қолданылады.

1. «Дара байланыс» тәсілі. Мұнда кластерлер «жақын көрші» тәсілімен бірігеді. Топтардың ара қашықтығы жақын объектілердің ара қашықтығымен анықталады.

2. «Толық байланыс» тәсілі. Мұнда топтардың арасындағы қашықтық «алыс көрші» принципімен анықталады. Біріктірілетін кластерлердің өзара қашықтығы екеуін қамтитын сфераның ең кіші диаметріне тең болады.

#### 4.3 STATISTICA бағдарламасын кластерлік талдауда пайдалану

Бағдарламаны пайдаланып иерархиялық *Joining (tree clustering)* операцияларын жүргізу үшін төмендегідей іс-қимылды жүргізеді:

1. Талданатын айнымалыларды таңдау (*Variables*);

2. Енгізілетін деректердің түрін анықтау (*input*) мұнда объектілердің координатасы бар кестені (*raw data*) немесе объектілердің арасындағы арақашықтық матрицасын енгізуге болады (*distance matrix*);

3. Кластерленетін объектілерді анықтау: Олар айнымалылар (бағаналар) (*variables/columns*) немесе бақыланатын мәліметтер (жолдар) *cases (rows)* болуы мүмкін;

4. Кластерлердің арасындағы ара қашықтықты сипаттайтын метриканы анықтау *Amalgamation (linkage) rule*;

5. Объектілердің арасындағы ара қашықтықты сипаттайтын метриканы анықтау *Distance measure*.

Кластерленудің нәтижесі мынадай болады:

1) Көлденең немесе тік дендрограмма тұрғызылады;

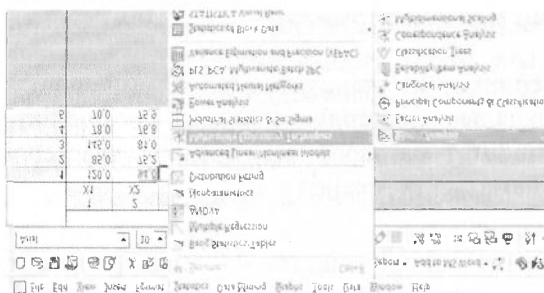
2) Берілген объектілердің ара қашықтығын сипаттайтын матрица тұрғызылады (*Distance measure*);

3) Берілген әрбір объектінің орташа және орташа квадраттар ауытқуы шығарылады (*Distriptive statistics*).

Қарастырылған мысалды *Statistica* бағдарламасы арқылы шығарып көрелік. Ол үшін мынадай іс-қимыл жүргізіледі.

1. Менюден

*Statistics => Multivariate Exploratory Tecniques => Cluster Analysis* командасы орындалады (46-сурет).



46-сурет. Кластерлік талдауды таңдау

2. *Joining(tree clustering) => OK*

3. *Variables => Select All => OK*

4. *Input file => Raw data => OK*

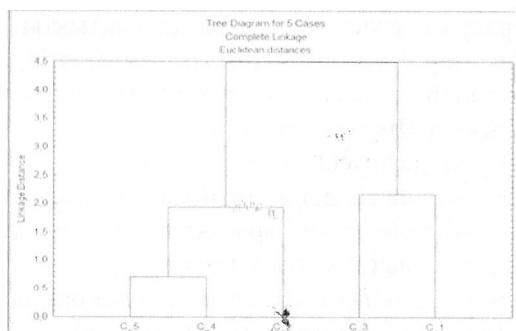
5. *Cluster => Cases(rows)*

6. *Amalgation(linkage) rule => Complete linkage*

7. *Distance measure => Euclidean distances => OK*

8. *Joining Results* терезесі шығады

9. Дендрограмманың түрін таңдағанда *Vertical icicle plot => OK* (47- сурет).



47- сурет. Кластер дендрограммасы

10. Егер дара байланыс түрі қарастырылатын болса *Amalgation (linkage)rule* => *Single linkage* таңдалынады.

### Сұрақтар мен тапсырмалар:

1. Кластердің мәнісі және оның талдануын сипаттаңыз
2. Иерархиялық кластерлік талдаудың ерекшеліктерін атаңыз
3. *STATISTICA* бағдарламасы арқылы кластерлік талдауды жүргізудің бағытын көрсетіңіз

## 5 БӨЛІМ. УАҚЫТ ҚАТАРЛАРЫН ТАЛДАУ

### 5.1 Уақыт қатарларын талдау негіздері

Көптеген экономикалық процестер мен құбылыстар уақыт қатарларымен сипатталады. Уақыт қатарлары деп бірдей уақыт аралығындағы байқалған көрсеткіштердің жиынтығын айтады. Ол  $y_1, y_2, \dots, y_n$  арқылы белгіленеді.

Уақыт қатарлары ретінде:

- а) қараша айындағы дүкендегі түсім ( $y_1, y_2, \dots, y_{30}$ )
- б) облыстағы ай сайынғы салықтың жиналуы ( $y_1, y_2, \dots, y_{12}$ )
- с) кинотеатрға соңғы аптада келген көрерменнің саны ( $y_1, y_2, \dots, y_7$ )

Уақыт қатарын талдаудағы негізгі мақсат мыналар болып саналады:

- 1) Өткен шақты ескеріп болжам жасау.
- 2) Қатардың ерекшеліктерін қысқаша түрде сипаттау.
- 3) Сипатталатын факторлардың сандық мәнін анықтау.

Уақыт қатары әдетте детерминдік (анықталған) факторлар мен кездейсоқ факторлардың әсерінен құралады.

Детерминдік факторларға мыналар жатады:

- а) Тренд- уақыт аралығында болатын ағым. Т-әрпімен белгіленеді.
- б) Маусымдық тербеліс - 1 жылдың ішінде тренд бойында болатын периодтық ауытқу. S - әрпімен белгіленеді.
- с) Циклдық тербеліс - Ұзақ мерзімді (10-15 жыл) болатын ауытқулар. C- әрпімен белгіленеді.

Кездесок фактор ретінде :

- д) Есептелмеген факторлардың жиынтық күшін сипаттайтын ауытқуды қарастырады. E-әрпімен белгіленеді.

Жоғарыдағы тұжырымдарды айта келіп, уақыт қатарларының моделін  $Y = f(T, S, C, E)$  функциясы ретінде қарастырады.

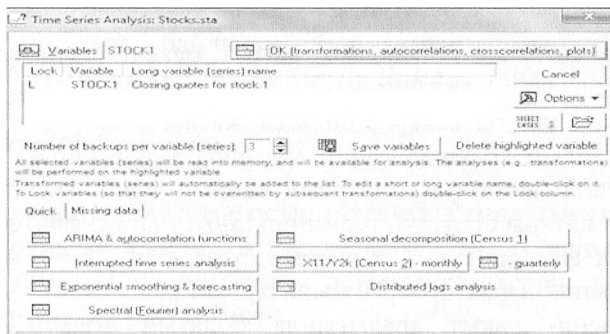
Көп жағдайда C факторының мәні есептеле бермейді, себебі ол ұзақ мерзімді фактор болғандықтан мәліметтерді жинау уақытты талап етеді, екіншіден оны есептеу маусымдық фактордың мәнін есептеумен ұқсас келеді.

## 5.2 STATISTICA - да уақыт қатарларын талдау

STATISTICA-да уақыт қатарларын талдау *Statistics => Advanced Linear/Nonlinear Models => Time Series/Forecasting* командасы арқылы жүзеге асырылады (48 сурет).

Шыққан панель екі: жоғарғы және төменгі бөліктен тұрады. Жоғарғы бөлік талданатын айнымалылардан, ал төменгі бөлік функционалдық батырмалардан тұрады.

*Statistica*-да жұмыс пайдаланатын деректердің файлы ашудан және талдануға тиісті айнымалыларды таңдаудан басталады. Ол үшін мына батырмалар қолданылады.



48– сурет. Уақыт қатарларын талдау

*Open data* - Деректерді ашу;

*Open Examples* - Мысалдар файлы ашу;

*Select cases* - Мәндерді таңдау;

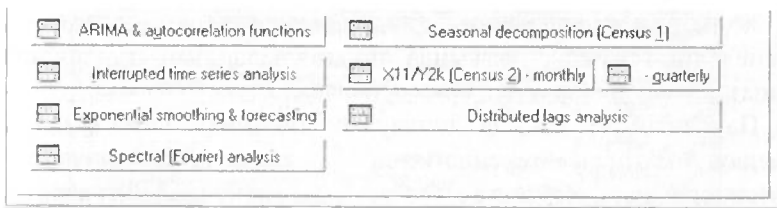
*Save variable*-Айнымалыларды сақтау;

*Delete highlighted variable*-Ерекшеленген айнымалыларды жою;

*Cancel* – Панельді жабу.

Төменгі жақта орналасқан функционалдық батырмалар мынаны сипаттайды (49 сурет).





49 - сурет. Функционалдық батырмалар

Панельдегі батырмалардың сипатталуы:

*ARIMA*-Авторегрессия және енгізілген жылжымалы орташа тәсілі.

*Interrupted time series analyses* - Үзілген уақыт қатарларының талдануы.

*Exponential smoothing & forecasting* - Экспонентті тегістеу және болжау.

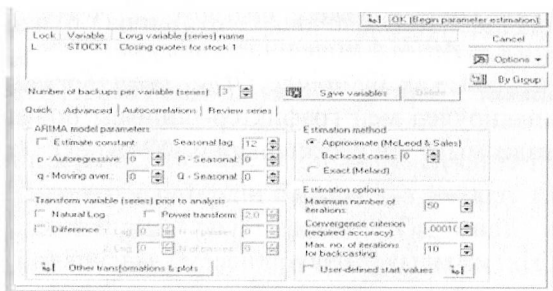
*Spectral (Fourier) analysis* – Спектрлік (Фурье) талдау.

*Seasonal decomposition (Census 1)*-Маусымдық композиция.

*X11(Census 2) – monthly – quarterly – XII* тәсіл – айлық-тоқсандық.

*Distributed lags analysis*-Таралған лагтарды талдау.

*ARIMA* опциясын таңдаймыз(50-сурет). Панельдің бас жағында ОК – параметрлерді бағалауды бастау батырмасы орналасқан. Оны модельді және оның ретін анықтағаннан кейін басу керек.



50 - сурет. *ARIMA* мүмкіндігі

Жүйе тек ерекшеленген айнымалылармен жұмыс жасайды. Панельдің төменгі жағында айнымалылардың түрлендіруін сипаттайтын ақпараттық өріс орналасқан.

Панельдің ортасында талдаудың бөліктері бар: *Quick*-тез талдау, *Advanced* – тереңдетілген талдау, *Autocorrelation*- автокорреляция, *Review series*-қатарды және графикті қарау.

*Advanced* опциясында маңызды батырмалар орналасқан. *ARIM Amodel parametr*s– *ARIMA* моделінің параметрлері, *Transform variable prior to analysis* – айнымалыларды алдын ала түрлендіру (51 - сурет).

ARIMA model parameters:

Estimate constant      Seasonal lag: 12

p - Autoregressive: 0      P - Seasonal: 0

q - Moving aver.: 0      Q - Seasonal: 0

Transform variable (series) prior to analysis:

Natural Log       Power transform: 2.0

Difference 1. Lag: 0      N of passes: 0

   2. Lag: 0      N of passes: 0

Other transformations & plots

51 - сурет. *ARIMA* параметрлерінің формасы

*Estimate constant*-константаны бағалау;

*Seasonal lag* - маусымдық лаг;

*p* – *Autoregressive*- авторегрессия параметрі;

*P* – *Seasonal*-авторегрессияның маусымдық параметрі;

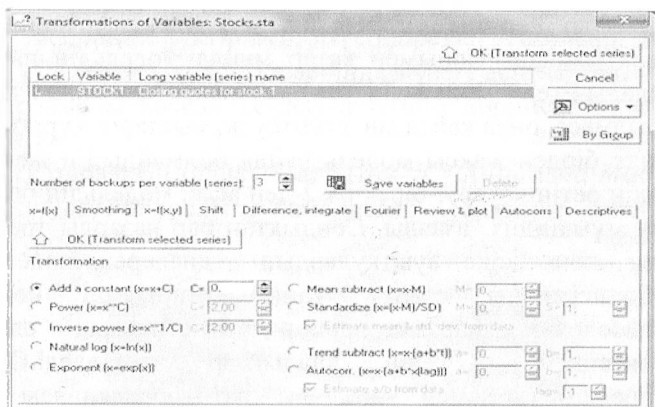
*q* – *Moving average*-жылжымалы орташаның параметрі;

*Natural Lag*- қатарды логарифмдеу;

*Power transform*-дәрежеге шығару;

*Difference* – *ARIMA* моделіндегі айырма.

Панельдің сол жақ бүйірінде *Other transformation&plots*-басқа түрлендірулер мен графиктер опциясы орналасқан. Оны таңдаған кезде мынадай кескін шығады(52-сурет).



## 52 -сурет. Айнымалыларды түрлендіру

Панельдегі  $x = f(x)$  опциясы бір айнымалыны түрлендірудің мүмкіндіктерін сипаттайды. Онда:

- *Add a constant*- константаны қосу;
- *Power*-дәрежеге шығару;
- *Inverse power*-  $x = x^c$  - дәрежесіне шығару;
- *Natural log*- логарифмдеу( $\ln x$ );
- *Exponent*-экспонентті түрлендіру( $e^x$ );
- *Mean subtract*-орташаны жою;
- *Standardize*-стандарттау;
- *Trend subtract* –тренді жою;
- *Autocorr*–автокорреляциялықтүрлендіру;
- *Estimate a/b from data*– тренд параметрлерін бағалау.

### 5.3 STATISTICA-да уақыт қатарын болжау

Деректердің қатарын болжау үшін алдымен онда тренд, цикл және маусымдық тербелістердің барына көз жеткізу қажет. Осы орайда қалдықтардың ролі ерекше. Ол үшін қалдықтар стационар күйге арнайы түрлендірулер арқылы жеткізіледі.

Қалдықтар негізінде модельдің реті идентификацияланады және нақтыланады. Сонымен қатар мұнда модельдің параметрі анықталады. Егер модельдің статистикалық елеулілігі қанағаттандырылмаса, онда қайтадан есептеу жұмыстары жүргізіледі.

Әдетте бірден жақсы модель пайда болмайды. Ең жеңіл жол модельдің ретін көтеру, бірақ ол 2-ден асса, модельдің болжамға қолдану мүмкіндігі азаяды. Сондықтан бар назарды тренді қай тәсілмен жою және ауытқулардың дисперсиясына аудару қажет. Дисперсияны азайту үшін логарифмді қолдануға болады. Модельдің параметрлерінің статистикалық елеулілігі дәлелденгеннен кейін қалдықтар зерттеледі. Олардың автокорреляциясы мен жеке автокорреляциясы сенім аралығында жатуы тиіс және нормальді таралуы керек.

Осыдан кейін болжам жасалады.

Уақыт қатарлары кез келген зерттелетін құбылыс немесе үрдістердің бірдей уақыт аралығындағы тіркелген көрсеткіштерінің жинағы болып табылады. Оның ерекшеліктерін әрі өзгеру ағымын талдау және соның негізінде болжау жұмыстарын жүргізу зерттеушіден тыңғылықты білікті, машықты және шығармашылықты талап етеді. Соңғы кезде уақыт қатарларын модельдеудің тәсілдері мен мүмкіншіліктері ақпараттық технологиялардың ересен дамуына орай барынша толығып көбеюде.

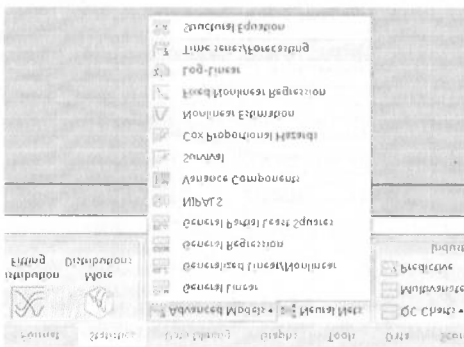
Экономикалық үрдістер мен құбылыстар көпжағдайда уақыт қатарларымен сипатталады, олай болса оған әсер тетін факторлардың жиынтық күшін сипаттау әрі маусымдық факторлардың өлшемін білу арнайы тәсілдер мен амалдарды қажет етеді. Сонымен бірге зерттеліп отырған құбылысты белгілі бір дәлдікпен болжай білу, кейінгі болатын басқару шешімдерінің нақты болуына септігін тигізеді. Кез келген болжамдық көрсеткіш белгілі бір ықтималдықпен берілуі тиіс. Сонымен бірге болжамдық модель де тиянақты болуы керек.

Осы орайда *ARIMA* тәсілінің орны ерекше. «*ARIMA*» сөзі ағылшын тілінен аударғанда «*AR – autoregression*» - авторегрессия - мен «*MA – moving average*» - жылжымалы

оргаша тәсілдің біріктірілуін түсіндіреді. Аталмыш модель *STATISTICA* бағдарламасы кешенінде жүзеге асырылған.

Бағдарламаны қосқан кездегі алғашқы көрініс 53 - суретте бейнеленген.

Уақытқатарларын талдау *Statistics => Advanced Models => Time series/Forecasting* командалары негізінде жүзеге асады(53 - сурет).



53 – сурет. Уақыт қатарларын талдау

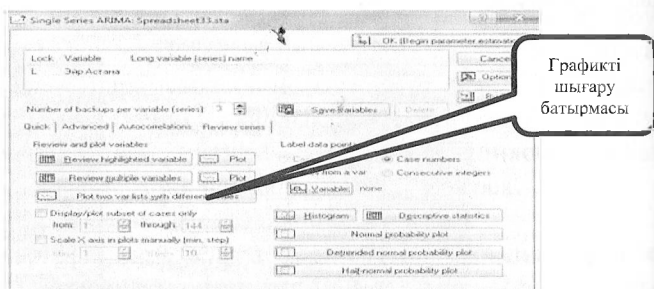
53 суреттегідей іс-қимылды атқарғаннан кейін көрініс 54-суретте бейнеленген.



54 - сурет. Уақыт қатарын талдау панелі

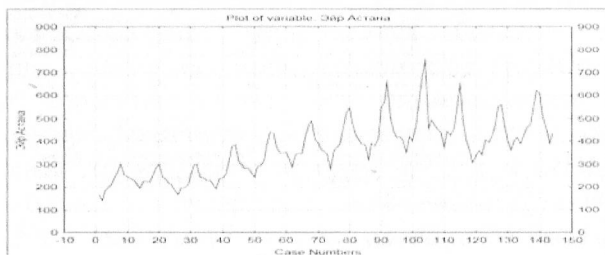
Жоғарыдағы 54-суреттен уақыт қатарын талдаудағы *STATISTICA* бағдарламасының мүмкіндіктерін көруге болады.

Бізге қажетті ең бірінші *ARIMA* & *autocorrelation functions* батырмасы. Оны басу келесі көрініске алып келеді(55 -сурет).



54 - сурет. *ARIMA* модельдеудің көрінісі

Әдетте уақыт қатарларын талдау оның графиктік кескінімен танысудан басталады. График зерттеліп отырған үрдістің немесе құбылыстың ерекшеліктерін атап айтса, сызықтық немесе сызықтық емес екенін сипаттап, онда периодтық тербелістердің бар екендігін көрсете алады. Ол өте маңызды кезең, себебі кейінгі талдау жұмыстары осы графиктен алынған ақпарат бойынша жүргізіледі. Ол үшін 54-суреттегі батырманы басу керек. Сондағы көрініс 55-суретте бейнеленген:



55 сурет - «Эйр Астана» компаниясының көрсеткіштері

55-суреттен компанияның көрсеткіштерінің зерттеукезеңінде компанияның өсу бағытында болғанын және онда белгілі бір периодтық тербелістер бар екендігін көруге болады. Ол периодтық тербеліс маусымдық факторға тәуелді екенін білдіреді.

Ендігі мақсат осы графикті ескере тырып алынған мәліметтерді сипаттайтын *ARIMA* моделінің параметрлерін анықтауға жет. Оны анықтау сайып келгенде модельді толықтай сипаттауды білдіреді.

Әдетте *ARIMA* моделі  $(p, d, q)$  үш параметрімен сипатталады.

Мұндағы:  $p$ -авторегрессиялық модельдің ретін көрсетеді. Ол үш мәнге 0, 1, 2 ие бола алады, 0-ге тең болуы авторегрессиялық модельдің жоқ екендігін, яғни оның параметрінің болмайтындығын білдіреді. Егер 1 немесе 2 мәніне ие болса, онда авторегрессияның соған сәйкес араметрлердің болу керектігін сипаттайды.

$d$ -мәні уақыт қатарларының көршілес мәндерінің айырма ретін білдіреді.

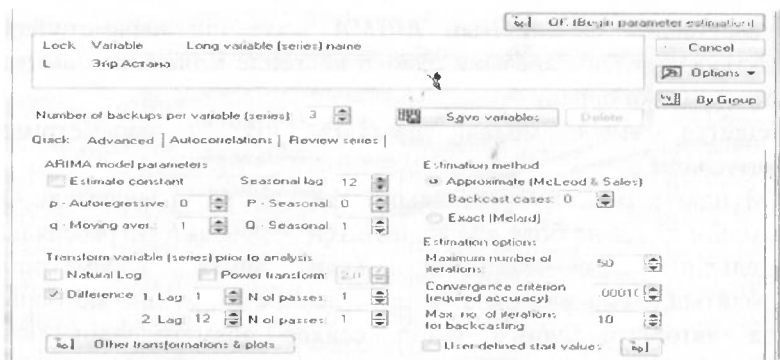
Егер,  $d=1$  болса, онда  $y_i - y_{i-1}$ , егер  $d=2$  болса, онда  $y_i - y_{i-2}$  болатындығын сипаттайдымұндағы:  $y_i$  — уақыт қатарлары болып табылады. Әдетте ол айырма уақыт қатарларын стационар күйге келтіруге қолданылады.

$q$ -мәні жылжымалы орташа моделінің ретін сипаттайды. Ол да 0, 1, 2 мәндеріне ие бола алады. Қалған сипаттама авторегрессиялық модельдің анықтамасына ұқсас.

Стационарлық деректер деп өзінің ерекшеліктерін уақыттың өзгергеніне қарамастан сақтап қала алатын мәліметтерді айтады. Әдетте оның орташа мәні нақты бір санға ие, ауытқуы тұрақты болады [6].

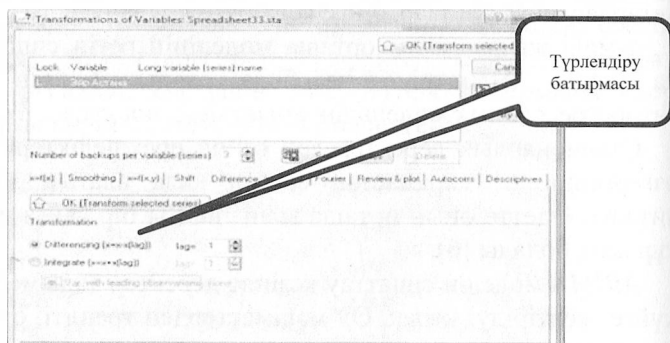
*ARIMA* моделін сипаттау кезінде деректер алдымен стационар күйге келтірілуі керек. Ол мәліметтерден трендті, орташа мәнді алу немесе стандарттық күйге келтіру арқылы жасалады. Көп жағдайда ол көршілес мәндерді бір бірінен алу, яғни  $y_i - y_{i-1}$  арқылы жүзеге асады. Кез келген белгілі бір өсу, не кему ағымы бар уақыт қатарлары стационар деректер бола алмайды.

56-суреттен уақыт қатарларының стационар деректер емес екені анық көрініп тұр, себебі өсу ағымы бар. Демек, оныстационар күйге келтіру қажет. Ол үшін алдымен  $y_i - y_{i-1}$  айырмасын қолданамыз және ол үшін 56- суреттегі *Other transformations&plots* батырмасын басы керек.



56 - сурет. Уақыт қатарларындағы «Advanced» қалтасы

56-суреттегі батырманы басқаннан кейінгі көрініс төмендегі 57-суретте көрсетілген.

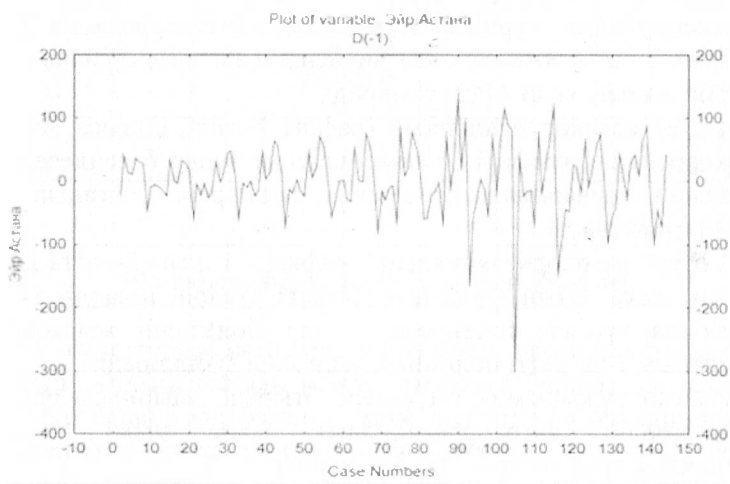


56 - сурет. Айырма функциясын жүзеге асыру



56-суреттегідей күйде берілген деректерден айырма *Difference* функциясы арқылы, яғни  $y_i - y_{i-1}$  формуласы негізінде ОК батырмасын басу арқылы жүзеге асады. Мұндағы басты мақсат берілген мәліметтерді стационар күйге келтіру.

Алынған жаңа деректердің графигі 57-суретте кескінделген.



57 – сурет. Бірінші айырмадан алынған деректердің көрінісі

57-суреттен деректердің бір қалыпқа келгенін байқауға болады. Себебі орташа мәні 0-ге тең және мәліметтердің одан ауытқуы белгілі бір тұрақтылықты сипаттайды. Өсу ағымы жоқ. Дегенменде, деректердің стационар екенін нақты екенін білу үшін арнайы статистикалық тәсілдер қолданылады. Соларға автокорреляция және жеке автокорреляция амалдары жатады.

Автокорреляция - ол уақыт қатарларында болатын өзара байланыстылық. Әдетте ол белгілі бір лагпен сипатталады. Жеке автокорреляция ол уақыт қатарларындағы периодтық тербелістерді сипаттау үшін қолданылады.

Автокорреляция мен жеке автокорреляцияның графиктері зерттеліп отырған уақыт қатарларының қандай модельге жататынын және олардың ретін анықтауға мүмкіндік береді.

Ол үшін мынадай тұжырымдар қолданылады:

1) Егер автокорреляцияның графигі экспоненциалды түрде бәсеңдеп, жеке автокорреляция графигі 1-ші лагта ытқыса, онда модельдің авторегрессияның 1-ші ретті болғанын, яғни  $AR(1)$  сипаттайды.

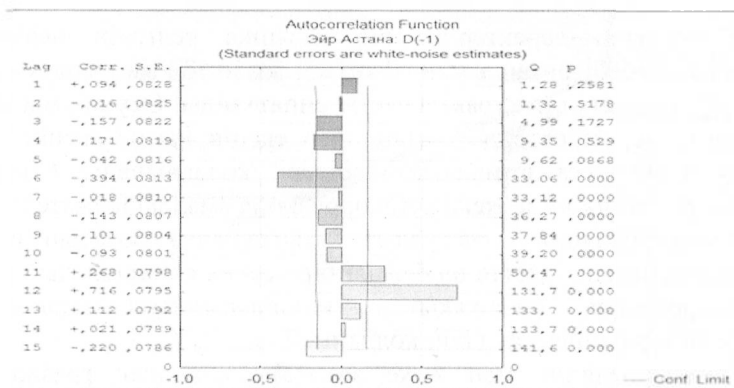
2) Егер автокорреляцияның графигі экспоненциалды немесе синусоида түрінде бәсеңдеп, жеке автокорреляция графигі 1 және 2-лагта ытқыса, онда модельдің авторегрессияның 2-ші ретті болғанын, яғни  $AR(2)$  білдіреді.

3) Егер автокорреляцияның графигі 1-лагта ытқыса, ал жеке автокорреляция графигі экспоненциалды түрде бәсеңдесе, онда модельдің жылжымалы орташаның 1-ші ретті болғанын, яғни  $JO(1)$  сипаттайды.

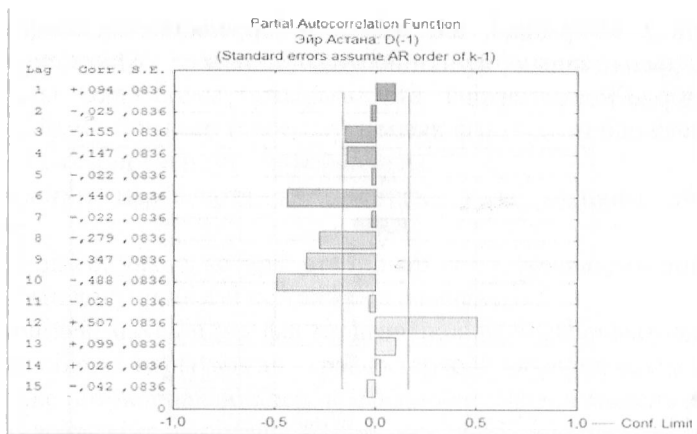
4) Егер автокорреляцияның графигі 1 және 2-лагта ытқып шығып, жеке автокорреляция графигі экспоненциалды немесе синусоида түрінде бәсеңдесе, онда модельдің жылжымалы орташаның 2-ші ретті болғанын, яғни  $JO(2)$  білдіреді.

Аталған тұжырымдарды ескере отырып, айырмасы алынған мәліметтердің автокорреляциялық және жеке автокорреляциялық графиктерін қарастырайық.

Автокорреляциялық және жеке автокорреляциялық функцияның графигі 58-59 суреттерде көрсетілген.

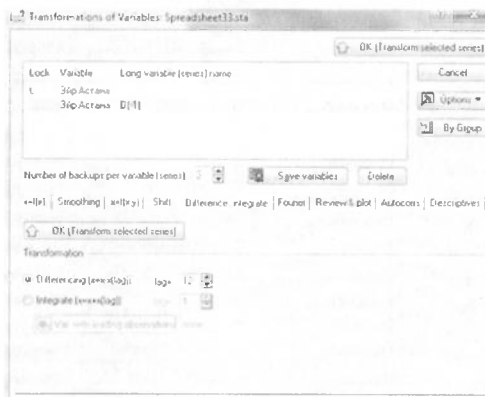


58 – сурет. Автокорреляция функция графигі



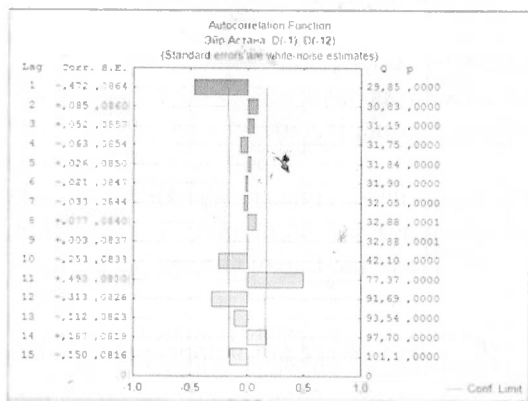
59 - сурет. Жеке автокорреляция графигі

58-59 -суреттерден автокорреляциялық коэффициенттердің 12-ші лагта ытқып шыққан мәнді анықтауға болады. Демек, ол белгілі бір периодтың барын және деректердің әлі стационар емес екендігін білдіреді. Сондықтан деректер стационар болу үшін периодтық тербелістен арылу керек, ол үшін тағы да  $y_i - y_{i-12}$  формула бойынша, яғни лаг=12 негізінде айырма алынады(60-сурет).

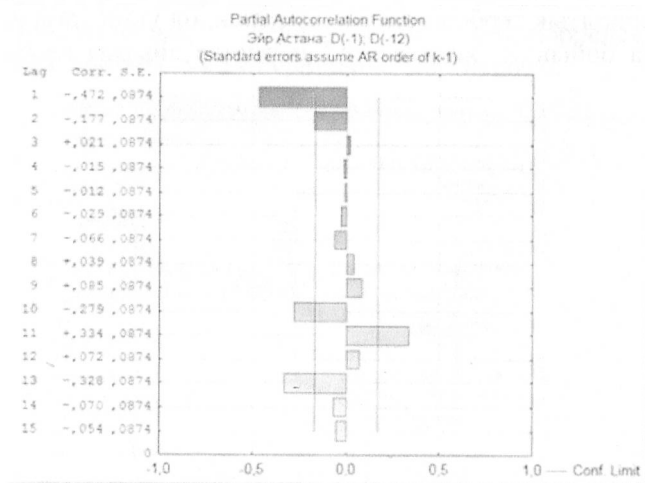


60 – сурет. Екінші айырманы алу

Екінші айырмадан алынған автокорреляциялық және жеке автокорреляциялық функциялардың графигі 61 және 62 - суреттерде көрсетілген.



61 – сурет. Екінші айырманың автокорреляциялық функциясының графигі



62 - сурет. Екінші айырманың жеке автокорреляция графигі

Жоғарыдағы 61 және 62-суреттерден 1-ші лагтың ытқуы автокорреляция коэффициентінің графигінде нақтырақ байқалады. Олай болса тұжырымдарға сай модель ЖО бірінші ретіне сәйкес, сонымен бірге маусымдық фактордың бар екендігі лагтың 12-ге тең екендігінен байқалады.

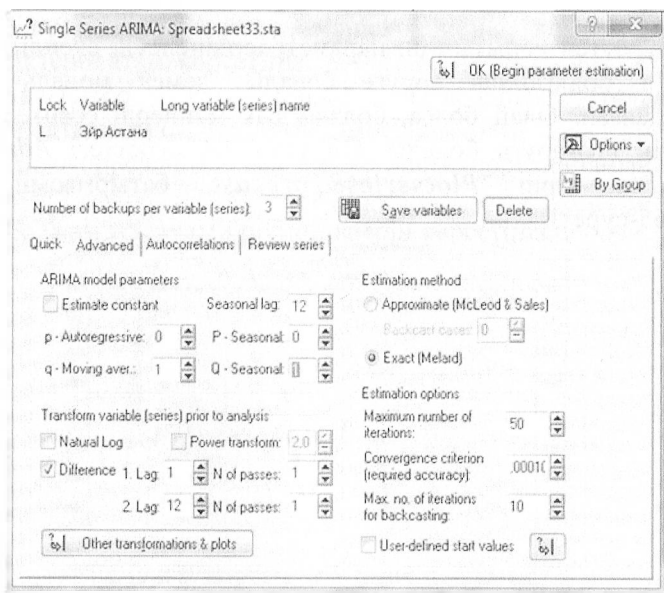
Байқалған нәрселерді тұжырымдай келе мынаны айтуға болады:

Біріншіден, уақыт қатары стационар емес, сондықтан оны ол күйге алып келу үшін екі рет айырма алынады.

Екіншіден, маусымдық фактордың ықпалы бары анықталды.

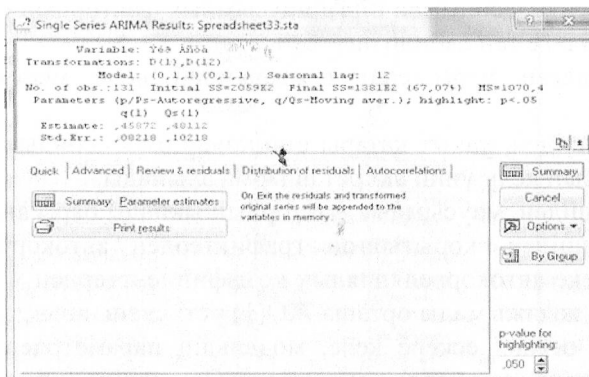
Үшіншіден, қорытылған графиктерден автокорреляциялық және жеке автокорреляциялық коэффициенттерден модельдің 1-ші ретті жылжымалы орташа ЖО(1) реті екені анықталды.

Енді осыны ескере келе, модельдің параметрлерін есептеп, болжамдық мәндерді анықталық. Ол үшін ұстаным 63-ші суреттегідей болуға тиісті.



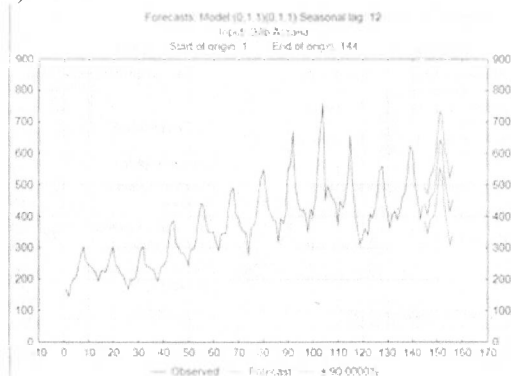
63 – сурет. Модельді анықтау режимі

63 - суреттегідей нақтылап алғаннан кейін ОК батырмасын басады. Алынған нәтиже 64- суретте бейнеленген.



64 - сурет. Қорытылған модельдің параметрлері

64-суреттен қорытылған параметрлердің елеулі екендігі басқа түспен боялғаннан көруге болады, демек модель дұрыс пайымдалған. Олай болса, болжамдық мәндерді есептеп оның графигін шығаруға болады. Ол үшін 64 суреттегі *Advanced* қалтасына кіріп *Plotseries&forecast* батырмасын басы керек(65-сурет).



65 – сурет. Болжамдық көрсеткіштердің графигі

66-суреттен болжамдық көрсеткіштердің белгілі бір дәрежеде шындыққа жанасатынын байқауға болады, мұндасенім аралығы 90% ықтималдығымен берілген. Уақыт қатарларын *ARIMA* арқылы модельдеу біраз машықты талап етеді. Мұнда зерттеуші әртүрлі моделді қарастырып, олардың ерекшеліктеріне және айқындалу параметрлеріне баса назар аударуы қажет.

Қазіргі кезде уақыт қатарларын барынша дәл сипаттау және нақты болжаудың өзектілігі барған сайын артуда. Себебі экономикалық үрдістерді басқаруда қабылданатын шешімдерді тез әрі негізді түрде қабылдау қажеттігі өткірілетусуде. Сондықтан уақыт қатарларын модельдеудің заманауи әдіс амалдарын игеру әрі қолдана білу қазіргі кездестратегиялық және шұғыл шешімдерді қабылдау және белгілі бір тәуекелдерді ескеру үшін маңызды болып отыр.

#### **Сұрақтар мен тапсырмалар:**

1. Уақытқатарларының мәнісі мен компоненттерін ерекшеліктерін атап көрсетіңіз.
2. *Statistica* –да уақыт қатарларын талдаудың бағытыннақтылаңыз.
3. Автокорреляция мен жеке автокорреляцияның ролі қандай?
4. Болжам жасаудың принциптерін көрсетіңіз.

## 6-БӨЛІМ. ДИСКРИМИНАНТТЫ ТАЛДАУ

### 6.1 Дискриминантты талдау негіздері

Дискриминантты талдау бақыланатын объектілерді берілген белгі бойынша сұрыптайтын әдіс-амалдарды сипаттайтын математикалық- статистикалық бөлім. «Дискриминант» сөзі латыннан аударғанда «айыру», «бөду» деген мағынаны білдіреді.

Мысалы, өндірістік көрсеткіштер бойынша кәсіпорындарды топтарға жіктеу үшін дискриминантты талдауды қолданады. Дискриминантты талдау медицинада, психологияда, әлеуметтануда, экономикада кеңінен қолданылады.

Сонымен қатар дискриминантты талдауды жеке тұлғаларды тестілеуде (тексеруде) қолданады. Оларды іріктеу кезінде берілген белгі бойынша «жарамды» және «жарамсыз» деп екі топқа бөлуге болады.

Экономикада дискриминантты талдауды қолданудың мүмкіндігін қарастырайық. Мысалы банк несиені клиенттерге берместен бұрын оларды бірнеше белгі бойынша «сенімді», «сенімсіз» тобына бөледі.

Дискриминантты талдаудың барлық тәсілдерін екі топқа бөлуге болады. Бірінші топқа алдын-ала берілген кластардың(топтардың) айырмашылықтарын анықтайтын тәсілдер жатқызылса, екінші топқа жаңа объектілерді қай класқа жатқызу керек екенін анықтайтын тәсілдер кіреді.

Берілген кластардың бір бірінен айырмашылығын білдіретін белгілерді дискриминантты айнымалылар деп атайды. Аталмыш айнымалылар аралық шкала (өлшем), не қатынастар өлшемімен сипатталады.

Объектілерді классификацияламас бұрын екі берілген топты бөліп тұратын шекараны анықтап алу қажет. Ол  $\bar{f}_1$  және  $\bar{f}_2$  мәндерінен бірдей қашықтықта болатын сандардың жиыны болуға тиісті. Мұндағы  $\bar{f}_1$  -бірінші жиынның орташа мәні,  $\bar{f}_2$  - екінші жиынның орташа мәні.



Яғни,

$$\begin{aligned}\bar{f}_1(x) &= a_1\bar{x}_{11} + a_2\bar{x}_{12} \\ \bar{f}_2(x) &= a_1\bar{x}_{21} + a_2\bar{x}_{22}\end{aligned}\quad (36)$$

$\bar{f}_1$  және  $\bar{f}_2$  -ден бірдей қашықтықта тұратын шаманы

$$C = \frac{1}{2}(\bar{f}_1 + \bar{f}_2) \quad (37)$$

сипаттайды және оны дискриминантты константа деп атайды. Жаңа объектілерді алдын-ала берілген топтардың қайсысына жататынын анықтау үшін мынадай кезеңдер қолданылады:

1) Берілген матрицалардың ковариациялық матрицалары есептелінеді;

2) Ковариациялық матрицалардың қосындысы

$$\hat{S} = \frac{1}{n_1 + n_2 - 2} (n_1 S_x + n_2 S_y) \quad (38)$$

формуласымен есептелінеді, мұндағы  $n_1$  және  $n_2$  бірінші және екінші топтағы элементтердің саны;

3)  $\hat{S}^{-1}$  -кері матрицасы есептелінеді;

4) Дискриминантты функция коэффициенттерімына формула арқылы анықталады:

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}) \quad (39)$$

5) Алдын ала берілген деректердің матрицасы дискриминантты функцияның мәндері мына формула арқылы есептелінеді:

$$\hat{U}_x = Xa \quad \hat{U}_y = Ya \quad (40)$$

6) Дискриминантты функция бағаларының орташа мәндері мына формула арқылы есептелінеді:

$$\bar{\hat{u}}_x = \frac{1}{n_1} \sum \hat{u}_{x_i} \quad \bar{\hat{u}}_y = \frac{1}{n_2} \sum \hat{u}_{y_j} \quad (41)$$

7)  $\hat{C}$  константасы формуласы арқылы есептелінеді:

$$\hat{C} = \frac{1}{2}(\bar{u}_x + \bar{u}_y) \quad (42)$$

*Мысалы:* Кәсіпорындардың қызметі мынадай үш түрлі көрсеткішпен сипатталады:

- негізгі өндірістік қорлардың жылдық орташа құны млн.теңге (НӨК);

- өндірістік-өнеркәсіптік персоналдың тізімінің саны (ӨӨП);

- баланстық пайда млн.теңге.

Зерттеу жұмысы жүргізілген салада озат (X) және қалыс (Y) кәсіпорындардың екі тобы құралған. Озат топта 4-кәсіпорын, қалыс топта 5-кәсіпорын бар. Олардың көрсеткіштері 4-кестеде берілген:

4 кесте - Кәсіпорынның мәліметтері

Рк №	Кәсіпорынның түрлері	НӨК құны, млн. теңге	ӨӨП саны	Баланстық пайдасы млн. теңге
1	2	3	4	5
1	Озат кәсіпорындар	224,228	17,115	22,981
		151,827	14,904	21,481
		147,313	13,627	28,669
		152,253	10,545	10,199
2	Қалыс кәсіпорындар	46,757	4,428	11,124
		29,033	5,51	6,091
		52,134	4,214	11,842
		37,05	5,527	11,873
		63,979	4,211	12,86

Салаға жаңа Z-кәсіпорны берілген. Оның көрсеткіштері мынадай: НӨК- 55, 451; ӨӨП-9, 592; баланс пайдасы-12, 84. Жаңа Z кәсіпорнын қай топқа жатқызу керек екенін анықтау қажет.

Шешімі:

1. Берілген алғашқы деректер X және Y матрицасы ретінде сипатталады:

$$X = \begin{pmatrix} 224,228 & 17,115 & 22,981 \\ 151,827 & 14,904 & 21,481 \\ 147,313 & 13,627 & 28,669 \\ 152,253 & 10,545 & 10,199 \end{pmatrix}$$

$$Y = \begin{pmatrix} 46,751 & 4,428 & 11,124 \\ 29,033 & 5,51 & 6,091 \\ 52,134 & 4,214 & 11,842 \\ 37,05 & 5,527 & 11,873 \\ 63,979 & 4,211 & 12,86 \end{pmatrix} \quad (43)$$

мұндағы:

$$n_1 = n_x = 4; n_2 = n_y = 5;$$

$$Z \text{ матрица жолы: } Z^T = (55,451 \quad 9,592 \quad 12,84)$$

2. Орташа мәндері мынадай

$$\bar{X} = \begin{pmatrix} 168,92 \\ 14,04 \\ 20,83 \end{pmatrix} \quad \bar{Y} = \begin{pmatrix} 45,79 \\ 4,77 \\ 10,76 \end{pmatrix} \quad (44)$$

3. Ковариациялық матрицаның бағалары:

$$S_x = \begin{pmatrix} 1023,94 & 55,62 & 28,91 \\ & 5,647 & 10,27 \\ & & 44,88 \end{pmatrix}$$

$$S_y = \begin{pmatrix} 145,86 & -6,61 & 22,78 \\ & 0,37 & -0,9 \\ & & 5,75 \end{pmatrix} \quad (45)$$

4. Ковариациялық матрицалардың қосындысы есептелінеді:

$$\hat{S} = \frac{1}{4 + 5 - 2} (4S_x + 5S_y);$$

$$\hat{S} = \begin{pmatrix} 689,3 & 27,06 & 32,8 \\ & 3,49 & 5,22 \\ & & 29,75 \end{pmatrix} \quad (46)$$

5.  $\hat{S}$  - матрицасына кері матрица есептелінеді:

$$\hat{S}^{-1} = \begin{pmatrix} 0,001 & -0,01 & 0,0007 \\ & 0,286 & -0,05 \\ & & 0,034 \end{pmatrix} \quad (47)$$

6. Дискриминация коэффициенттерінің мәндері есептелінеді:

$$a = \hat{S}^{-1}(\bar{X} - \bar{Y}) = \hat{S}^{-1} \begin{pmatrix} 123,11 \\ 9,26 \\ 10,07 \end{pmatrix} = \begin{pmatrix} 0,07 \\ 2,14 \\ 0,33 \end{pmatrix} \quad (48)$$

7. Дискриминантты функцияның мәндері есептелінеді:

$$\hat{U}_x = Xa = \begin{pmatrix} 62,07 \\ 51,15 \\ 50,49 \\ 38 \end{pmatrix} \quad \hat{U}_y = Ya = \begin{pmatrix} 16,93 \\ 16,16 \\ 17,13 \\ 18,78 \\ 18,4 \end{pmatrix} \quad (49)$$

8. Дискриминантты функцияның орташа мәндері есептелінеді:

$$\bar{\hat{u}}_x = 50,43 \quad \bar{\hat{u}}_y = 17,48 \quad (50)$$

9. Қортылған константаның мәні:

$$\hat{C} = \frac{1}{2}(50,43 + 17,48) = 33,96 \quad (51)$$

10. Z-кәсіпорнының қай топқа жататыны анықталады. Ол мына формуласы бойынша есептелінеді:

$$\hat{u}_z = a_1 z_1 + a_2 z_2 + a_3 z_3$$

яғни,

$$\hat{u}_z = 0,07 \cdot 55,45 + 2,14 \cdot 9,59 + 0,33 \cdot 12,84 = 29,28 \quad (52)$$

$\hat{u}_z$  - дискриминантты функциясының орташа мәні  $\hat{C}$  - константасынан кіші

$$\begin{aligned} \text{яғни, } \hat{u}_z &< \hat{C} \\ 29,28 &< 33,96(53) \end{aligned}$$

Олай болса Z-кәсіпорны озат топқа жатқызылмайды.

## 6.2 STATISTICA-да дискриминантты талдауды жүргізу

Дискриминантты талдаудың басты мақсаттарының бірі объектілерді немесе құбылыстарды зерттеп талдаған кезде оларды бір бірінен ажырататын басты факторларды (айнымалыларды) табу болып табылады. Ондай мәселе биологияда, элеуметтануда, медицинада т.с.с. көптеп кездеседі.

*Мысал.* Үш түрлі ирстің гүлшелері мен гүлтабандарының ұзындығы мен ені сипатталған 150 гүлдің деректері бар файл берілген. Ирстің бірінші түрі-*Setosa*, екінші түрі-*Versicol*, үшінші түрі-*Virginic* деп аталады.

Талдаудың мақсаты осы гүлдердің берілген ұзындықтары мен ені арқылы оларды ажырататын басты айнымалыларды табу. Деректердің файлы *Irisdat.sta* файлында орналасқан. Оны ашу:

*Open Examples => Datasets => Irisdat.sta* арқылы жүзеге асады (66-сурет).

STATISTICA - [Data: Irisdat.sta (5v by 150c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools

Arial 10

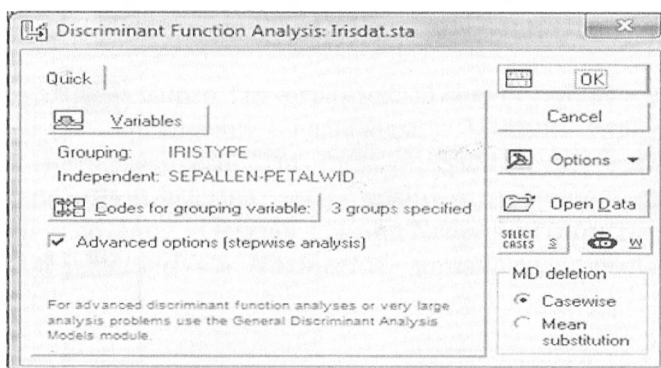
	Fisher (1936) iris data: length & width of sepals and petals, 3					
	1	2	3	4	5	
	SEPALLEN	SEPALWI	PETALLE	PETALWI	IRISTYP	
1	5.0	3.3	1.4	0.2	SETOSA	
2	6.4	2.8	5.6	2.2	VIRGINIC	
3	6.5	2.8	4.6	1.5	VERSICO	
4	6.7	3.1	5.6	2.4	VIRGINIC	
5	6.3	2.8	5.1	1.5	VIRGINIC	
6	4.6	3.4	1.4	0.3	SETOSA	
7	6.9	3.1	5.1	2.3	VIRGINIC	
8	6.2	2.2	4.5	1.5	VERSICO	
9	5.9	3.2	4.8	1.8	VERSICO	
10	4.6	3.6	1.0	0.2	SETOSA	
11	6.1	3.0	4.6	1.4	VERSICO	
12	6.0	2.7	5.1	1.6	VERSICO	
13	6.5	3.0	5.2	2.0	VIRGINIC	
14	5.6	2.5	3.9	1.1	VERSICO	
15	6.5	3.0	5.5	1.8	VIRGINIC	
16	5.8	2.7	5.1	1.9	VIRGINIC	
17	6.8	3.2	5.9	2.3	VIRGINIC	
18	5.1	3.3	1.7	0.5	SETOSA	
19	5.7	2.8	4.5	1.3	VERSICO	
20	6.2	3.4	5.4	2.3	VIRGINIC	
21	7.7	3.8	6.7	2.2	VIRGINIC	
22	6.3	3.3	4.7	1.6	VERSICO	
23	6.7	3.3	5.7	2.5	VIRGINIC	
24	7.6	3.0	6.6	2.1	VIRGINIC	
25	4.9	2.5	4.5	1.7	VIRGINIC	
26	5.5	3.5	1.3	0.2	SETOSA	
27	6.7	3.0	5.2	2.3	VIRGINIC	
28	7.0	3.2	4.7	1.4	VERSICO	
29	6.4	3.2	4.5	1.5	VERSICO	
30	6.1	2.8	4.0	1.3	VERSICO	

66-сурет. *Irisdat* файлының көрінісі

Алдыңғы екі айнымалы *Sepallen*, *Sepawid* гүлтабанның ұзындығымен енін, одан кейінгі екі айнымалы *Petallen*, *Petawid* гүлшелердің ұзындығы мен енін сипаттайды. Соңғы айнымалы иристің түрін білдіреді. Файлда 150 гүлдің түрі бар.

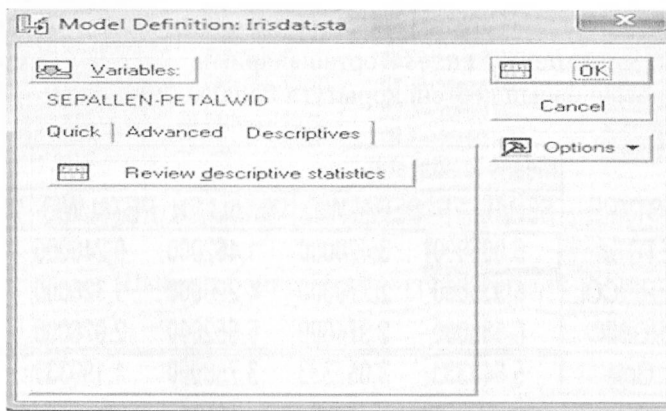
Дискриминантты талдау

*Statistics => Multivariate Exploratory Techniques => Discriminant Analysis* командасы арқылы тәуелсіз айнымалы ретінде *Sepallen – Petawid* айнымалыларын көрсетіп, топтаушы айнымалы ретінде *Iristype* айнымалысы қабылданады. *Advanced options* опциясына жалауша қойылады(67-сурет).



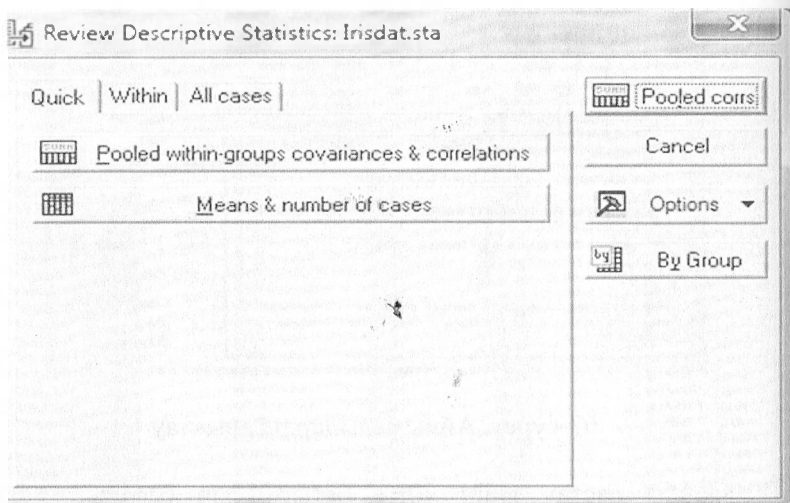
67-сурет. Айнымалыларды анықтау

Талдауды бастау үшін «OK» батырмасын басады. Экран бетіне «*Model Definition*» (модельді анықтау) панелі шығады (68-сурет).



68-сурет. Модельді анықтау

Дискриминантты функцияны таңдамас бұрын айнымалылардың таралу заңдылықтары мен ерекшеліктерін білу үшін *Review descriptive statistics* (статистика сипаттамаларын карау) батырмасын басады (69-сурет).



69-сурет. Статистика сипаттамаларын қарау

Шыққан панельдегі «Means & number of cases» (орташалардың мәндері) батырмасын басып кестені қорытуға болады (70-сурет).

IRISTYPE	Means (Irisdat.sta)				Valid N
	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SETOSA	5.006000	3.428000	1.462000	0.246000	50
VERSICOL	5.936000	2.770000	4.260000	1.326000	50
VIRGINIC	6.588000	2.974000	5.552000	2.026000	50
All Grps	5.843333	3.057333	3.758000	1.199333	150

70-сурет. Орташалардың көрсеткіштері



### 6.3 Нәтижелердің гистограммасын қорыту

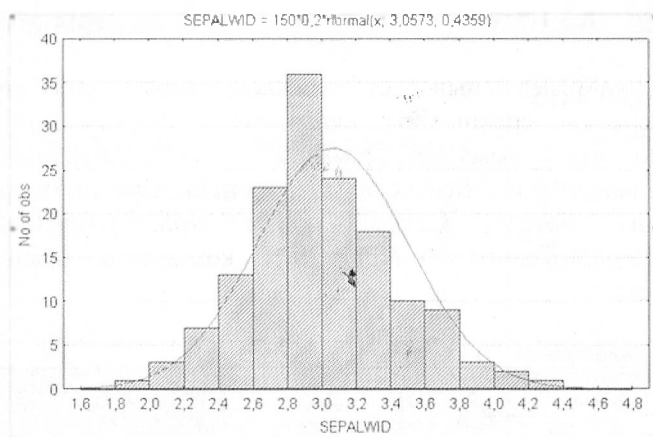
Қажет айнымалының гистограммасын қорыту үшін алдымен айнымалы ерекшеліні палынады. Мысалы, *Sepalwid* – *Wersicol* гистограммасын шығару үшін курсорды екі айнымалының қиылысына алып келіптің тіреуіннің оң жақ құлағынш ертеді. Осыдан кейін *Graphs of inputData => Histogram Sepalwid => NormalFit* командасы орындалады (71-сурет).

IRISTYPE	Means (Insdatt sta)				Valid N
	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SETOSA	5.006000	3.428000	1.462000	0.246000	50
VERSCOL	5.936000	2.770000	1.260000	1.260000	50
VIRGINIC	6.588000	2.974000			50
All Gips	5.843333	3.0573			150

71-сурет. Деректердің гистограммасын шығару

Шыққан бейне 72-суретте көретілді:

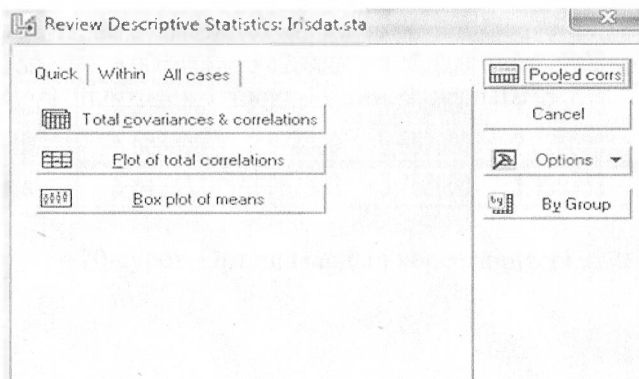


72-сурет. Деректердің гистограммасы

Осындай бағытпен барлық айнымалылардың гистограммаларын қорытуға болады.

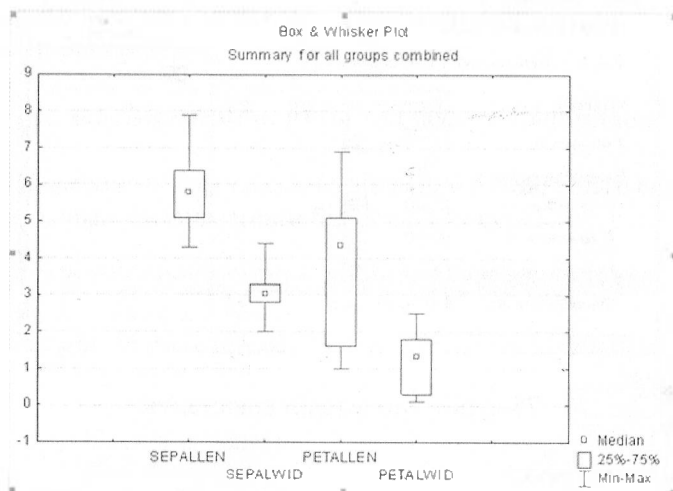
*Аумақтық диаграмма.*

Талданатын тәуелсіз айнымалылардың орташа мәні мен аумағын анықтау үшін графикалық мүмкіндік бар. Ол үшін *All cases* бетіндегі *Boxplot of means* батырмасын басады (73-сурет).



73-сурет. Деректердің аумағын талдау

Мұнда  $Mean/SD/1,96 * SD$  режимі таңдалынады (74-сурет).



74-сурет. Айнымалылардың аумақтырының графиктері

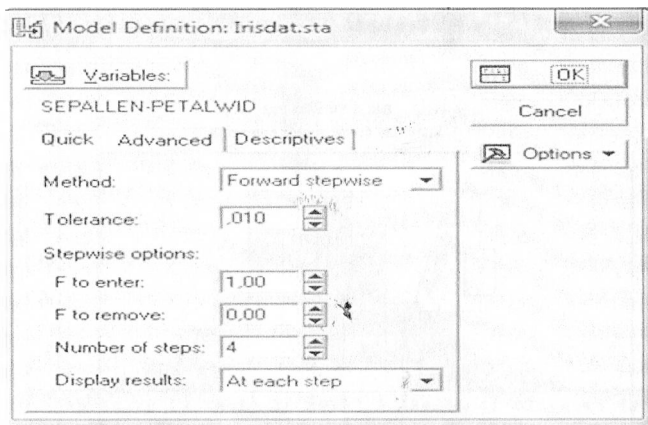
Графиктер үш түрлі компонентпен сипатталады:

1. Орталық нүкте немесе сызық айнымалының орташа мәнін немесе медианасын сипаттайды.
2. Төртбұрыш 25%-75% мәндердің қамтылуын көрсетеді.
3. Сызықшалар мәндердің ең үлкен және ең кіші мәнін анықтайды.

#### 6.4 Дискриминантты функцияны таңдау

Ол үшін *Cancel* батырмасын басып *Model Definition* панеліне келеміз. Дискриминантты талдаудың әрбір қадамында не болып жатқанын білу үшін *Advanced* бетіндегі тәсілдердің түріне *Forward stepwise* (қадам сайын енгізу) таңдалынады (75-сурет).

Аталмыш тәсіл бойынша бағдарлама модельге дискриминацияға үлесі үлкен айнымалыларды біртіндеп енгізеді.



75-сурет. Модельдің анықталуы

#### *Тоқтату ережесі.*

*Statistica* айнымалыларды модельге енгізу немесе модельден шеттетудің қадамдық режимін мынадай жағдайлар болғанда тоқтатады.

1. Барлық айнымалылар модельге енгізілгенде немесе модельден шығарылғанда.

2. Қадамдар саны мах мәніне жаткен кезде.

3. Модельден тыс айнымалылардың мәні *F to enter* (енгізу) артық болмаған жағдайда немесе айнымалылардың мәні *F to remove* (шеттеу) кем болған жағдайда.

4. Қайсы бір айнымалының мәні келесі қадамда панельде көрсетілген толлерантық (төзімділік) мәннен кем болса.

#### *Енгізу/шеттеу.*

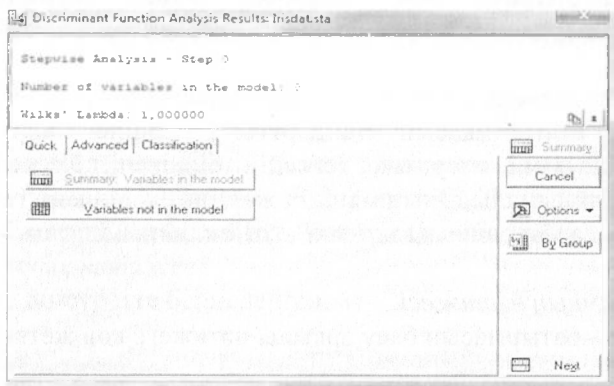
Қадамдық талдау кезінде бағдарлама дискриминацияға үлесі зор айнымалыларды таңдайды. Енгізілуге тиісті айнымалылардың *F* мәні панельде көрсетілген мәннен жоғары болуға тиісті. Ал модель құрамынан шеттету үшін ол айнымалының *F* мәні панельде көрсетілген мәннен төмен болуы керек.

#### *Толлеранттылық (төзімділік).*

Әрбір айнымалы модельге енгізілген сайын оның  $R$ -көптік корреляция коэффициенті есептеліп отырады. Толлеранттық мән  $1-R^2$  арқылы есептеледі. Ол айнымалының модельге сыйысу шамасын білдіреді.

## 6.5 Дискриминантты талдаудың қадамдары

*0-дік қадамның нәтижесі.* 0-дік қадам әлі модельге ешқандай айнымалының енгізілмегенін білдіреді (76-сурет).



76- сурет. 0-ші қадамның нәтижесі

Әлі модельге ешқандай айнымалы енгізілген жоқ, демек *Variable not in the model* (модельден тыс айнымалылар) батырмасын басып, олардың мәндерін білуге болады (77-сурет).

N=150	Variables currently not in the model (Irisdat.sta)					
	Wilks' Lambda	Partial Lambda	F to enter	p-level	Toler.	1-Toler. (R-Sqr.)
SEPALLEN	0.381294	0.381294	119.265	0.000000	1.000000	0.00
SEPALWID	0.599217	0.599217	49.160	0.000000	1.000000	0.00
PETALLEN	0.058628	0.058628	1180.161	0.000000	1.000000	0.00
PETALWID	0.071117	0.071117	960.007	0.000000	1.000000	0.00

77-сурет. Модельден тыс айнымалылардың мәндері

*Уилкс лямбдасы.*

Олқарастырылып отырған модельдің дискриминациялық қуатын білдіретін стандарттық статистикасы. Оның мәні 1.0 ден (ешқандай дискриминация жоқ) 0.0-ге (толық дискриминация бар) дейін өзгеріп отырады. Бірінші бағанадағы көрсеткіштер айнымалы модельге енгізілген бойда модельдің қаншалықты дискриминациялық қуаты артқанын білдіреді.

*Жеке Уилкс лямбдасы.*

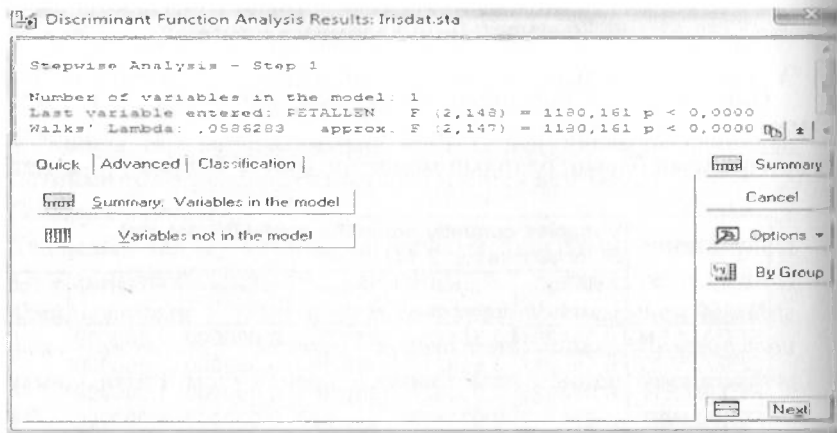
Бұл әрбір айнымалының дискриминацияға қосқан жеке үлесін білдереді. Мұнда мән неғұрлым төмен болған сайын айнымалылардың дискриминацияға қосқан үлесі жоғарылай береді.

*F енгізу және p мәні.*

Уилкс статистикасын стандартты F мәніне айналдыруға болады. P мәні елеулілік деңгейін білдіреді. Кестеге қарап *Petalen* айнымалысының мәнінің жоғары екенін көруге болады. Олай болса бірінші қадамда *Petalen* айнымалысы модельге енеді.

*1-қадамның нәтижесі.*

«Next» батырмасын басу арқылы нәтижеге қол жеткіземіз(78-сурет).



78-сурет. 1-ші қадамның нәтижесі

Суреттен *Petalen* айнымалысының модельге енгенін көруге болады. Енді *Variable not in the model* (модельден тыс айнымалылар) батырмасын бассак мынадай кестені қорытамыз(79-сурет):

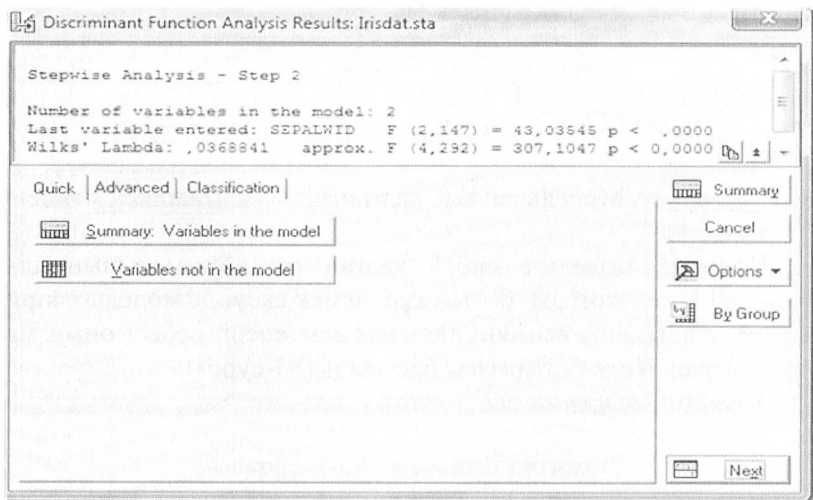
Variables currently not in the model (Irisdat sta)						
Df for all F-tests 2,146						
N=150	Wilks' Lambda	Partial Lambda	F to enter	p-level	Toler.	1-Toler. (R-Sqr)
SEPALLEN	0.039878	0.680189	34.32311	0.000000	0.428216	0.571784
SEPALWID	0.036884	0.629118	43.03545	0.000000	0.857179	0.142821
PETALWID	0.043777	0.746683	24.76568	0.000000	0.765300	0.234700

79-сурет. Модельден тыс айнымалылардың мәндері

Кестеден модельге енетін келесі айнымалы *Sepawid* екенін көруге болады. Себебі  $F \text{ to enter} = 43,03$  қалғандарына карағанда жоғары. Тағы да *Next* батырмасын басамыз.

2-қадамның нәтижесі.

Нәтиже 80-суретте бейнеленген.



80-сурет. 2-ші қадамның нәтижесі

Иристердің айнымалылар арқылы айырмашылығы едәуір. Уилкс лямбдасы 0,037 тең. Енді модельге енген айнымалылардың мәндерін қарау үшін *Summary Variable in the model* батырмасын басады(81-сурет).

Discriminant Function Analysis Summary (Irisdat sta)						
Step 2. N of vars in model: 2 Grouping: IRISTYPE (3 grps)						
Wilks' Lambda: 0.03688 approx. F (4 292)=307.10 p<0.0000						
N=150	Wilks' Lambda	Partial Lambda	F-remove (2,146)	p-level	Toler.	1-Toler. (R-Sqr)
PETALLEN	0.599217	0.061554	1112.954	0.000000	0.857179	0.142821
SEPALWID	0.058628	0.629118	43.035	0.000000	0.857179	0.142821

81-сурет. Модельге енген айнымалылардың мәндері

Енді модельге енбей қалған айнымалыларды қарау үшін *Variable not in the model* батырмасын басамыз (82-сурет).

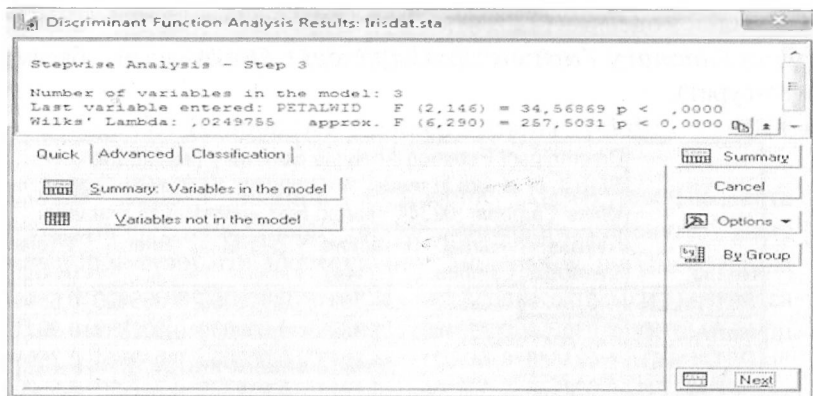
Variables currently not in the model (Irisdat sta)						
Df for all F-tests: 2 145						
N=150	Wilks' Lambda	Partial Lambda	F to enter	p-level	Toler.	1-Toler. (R-Sqr)
SEPALLEN	0.031546	0.855271	12.26848	0.000012	0.358493	0.641507
PETALWID	0.024976	0.677135	34.56869	0.000000	0.668905	0.331095

82-сурет. Модельден тыс қалған айнымалылардың мәндері

Кестеден модельге енбей қалған екі айнымалының да F мәндері 1-ден жоғары болып тұр, демек екеуі де модельге кіруге тиісті. Алдымен *Petalwid* айнымалысы енеді, себебі оның мәні жоғарырақ. *Next* батырмасы басылады(83-сурет).

*3-қадамның нәтижесі.*

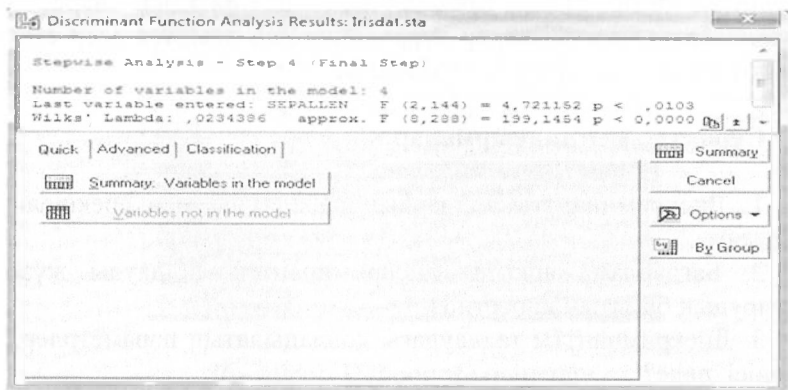




83-сурет. 3-ші кадамның нәтижесі

Уилкс лямбдасы 0,024 ке тең ол 0-ге жақын, демек айнымалының дискриминацияға әсері зор. *Next* батырмасын басамыз(84-сурет).

4-қадамның нәтижесі. Бұл қадам соңғы қадам, барлық төрт айнымалы модельге енді.



84-сурет. 4-ші кадамның нәтижесі

Енді модельге енген айнымалылардың мәнін анықтау үшін *Summary Variables in the model* батырмасын басамыз (85-сурет).

Discriminant Function Analysis Summary (Irisdat sta)						
Step 4. N of vars in model: 4. Grouping: IRISTYPE (3 grps)						
Wilks' Lambda: .02344 approx. F (8 288)=199.15 p<0.0000						
N=150	Wilks' Lambda	Partial Lambda	F-remove (2 144)	p-level	Toler.	1-Toler. (R-Sqr.)
PETALLEN	0.035025	0.669206	35.59018	0.000000	0.365126	0.634874
SEPALWID	0.030580	0.766480	21.93593	0.000000	0.608859	0.391141
PETALWID	0.031546	0.743001	24.90433	0.000000	0.649314	0.350686
SEPALLEN	0.024976	0.938464	4.72115	0.010329	0.347993	0.652007

85-сурет. Модельге енген айнымалылардың мәні

Кестеден *PetalLEN* айнымалысы жалпы дискриминацияға көп үлес қосатынын көруге болады, себебі оның мәні 0,66-ға тең, Алдында жеке Уилкс лямбдасы аз болған сайын көп үлес қосады деп атап кеткенбіз. Екінші орында *Petalwid* оның мәні 0,74. Үшінші орында *Sepalwid*, мәні 0,76. Соңғы орында *Sepallen* жайғасады. Қорыта келгенде айтатын нәрсе ол гүлшелердің өлшемдері әр түрлі иристерді айыруға әсер ететін басты айнымалылар болып табылады.

### Сұрақтар мен тапсырмалар:

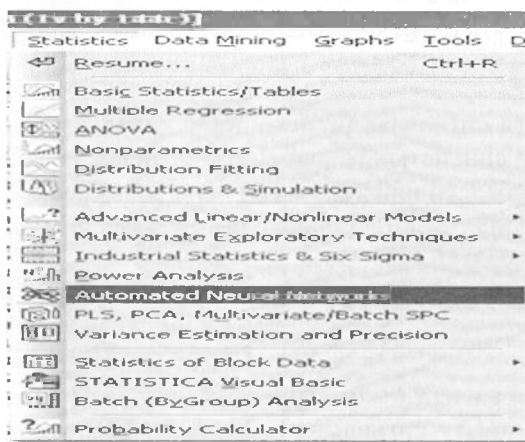
1. Дискриминантты талдаудың қажеттілігі мен ерекшелігін айтыңыз.
2. Бағдарлама арқылы дискриминантты талдауды жүзеге асырудың бағытын көрсетіңіз.
3. Дискриминантты талдаудағы қолданылатын параметрлердің мәнісі неде?

## 7 БӨЛІМ. НЕЙРОНДЫҚ ЖЕЛІЛЕР ЖӘНЕ ОЛАРДЫ ҚОЛДАНУ

### 7.1 Нейрондық желілер туралы түсінік

Соңғы кездері нейрон желілеріне қызығушылық арта түсуде. Нейрон желілері адам миының үйренуге және қателіктерді түзететін икемділігін модельдеу негізінде пайда болды. Егер классикалық талдау тәсілдері жұмыс жасамаса немесе нәтиженің дәлдігі қанағаттандырмаса, онда зерттеуші нейрон тәсілдерін қолдана алады.

Нейрон желілерін әсіресе сұрыптау, болжамдау процестерінде немесе сызықтық емес байланысты сипаттауда табысты қолдана алады. STATISTICA жүйесінде нейрон желілерін құруды қарапайым қимылдармен атқаруға болады. Ол үшін *Automated Neural Networks* модулін ашу керек (86-сурет).



86 - сурет. Нейрон желісін ашу

Нейрондық желілерді пайдалануда белгілі бір шектеулер бар, оны кез келген бағытта қолдана алмайсыз. Дегенмен де, ол жиналған алғашқы деректер құрылымын анықтауда немесе

мәліметтердің арасындағы байланысты нақтылап, негізді жорамал жасауға септігін тигізе алады.

## 7.2 Нейрондық желілердің негіздері

Алдымен нейронның не екенін түсініп алған дұрыс. Адамның миы 10 миллиардқа жуық қарапайым жасуша(клетка) – *нейрондардан* тұрады. Олар бір біріне ұдайы түрде электрохимиялық белгілер береді. Басқаша айтқанда ойлаунемесе сезу дегеніміз аталмыш сигналдардың жұмысы. Әрбір нейрон ақпаратты қабылдайтын канал *дендриттен, ядродан* және ақпаратты шығаратын *аксоннан* тұрады.

Клетканың аксондары басқа клетканың дендриттеріне синапстар арқылы жалғасады. Нейрон қозған кезде аксонға белгі жібереді. Нейрон қай уақытта қозуы мүмкін? Ол нейронға келетін жалпы сигналдардың көлемі белгілі бір белсенділік деңгейінен асқанда қозады. Мұнда сигналдардың белсенділігі синапстардың белсенділігімен байланысты екенін атап өту керек.

Сонымен:

- нейронға сигналдар бірнеше кіріс каналдарынан енеді және әрбір сигнал синапстың белсенділігіне байланысты болады, әрбір синапстың белсенділігі(салмағы)  $w_i$  деп белгіленеді мұнда  $1 \leq i \leq N$  сонда ағымдағы нейронның жағдайы мына формула арқылы есептеледі.

$$S = \sum x_i * w_i \quad (54)$$

- алынған сигнал  $f$  - белсенділік функциясы арқылы түрленіп нейронның шығу сигналына айналады. Шығу сигналын мына формула арқылы белгілейді.

$$Y = f(s) \quad (55)$$

Әдетте, ол сызықтық емес функцияретінде сипатталаынады.

Егер нейронның ағымдағы жағдайы Т-белгілі бір белсенділік деңгейінен кем болса, онда шығу сигналы 0-ге тең, яғни тежелу орын алады.

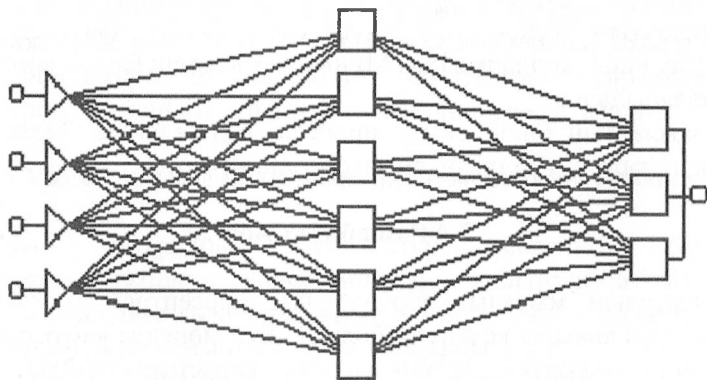
Егер нейронның ағымдағы жағдайы Т-белгілі бір белсенділік деңгейінен артық болса, онда шығу сигналы 1-ге тең яғни козу орын алады.

Нейрон түсінігін анықтап алғаннан кейін нейрондық желілер туралы айтуға болады. Егер нейрондар бір бірімен байланысса, онда олар нейрондық желілерді құрады.

Нейрондық желілер арнайы әдебиеттерде кеңінен сипатталған [10, 11]. Нейрондық желілердің кіру және шығу каналдары болады.

Мысалы, адамды алатын болсақ, оның көзі немесе құлағы ақпараттың кіру органы болса, қолы мен аяғы ақпараттың шығу органы іспетті. Сонымен бірге желіде жасырын нейрондар да болады.

Қарапайым желіде сигналдар кіру каналынан жасырын нейрон арқылы шығу каналдарына өтеді (86-сурет).



86 - сурет. Сигналдың берілу желісі

86-суреттен 4-кіріс каналын 1-шығыс каналын және 2-жасырын нейрон қабатын көруге болады. Сонымен бірге, әрбір

нейронның өзінен бұрынғы және өзінен кейінгі қабаттағы әрбір нейронмен байланыста болатынын көреміз.

### 7.3 Нейрондық желіні оқыту

Адам және миы бар жан-жануар белгілі бір дәрежеде оқып үйрене алады. Мысал ретінде цирктегі үйретілген аңдарды айтсақ жеткілікті. Ол өте маңызды нәрсе.

Оқытудың (үйретудің) мәнісі мынада: синапстар мен Т-белсенділік деңгейінің арасындағы байланыстардың күшін өзгерту арқылы оқыту моделін жасауға болады. Мұны былай түсіндіруге болады:  $w_{ij}$  синапстық белсенділікті мұндағы  $i$ -нейрон нөмірі,  $j$ -желідегі нейрон қабатының нөмірі және Т-белсенділік деңгейін өзгерте отырып, желі жұмысының нәтижесін модельдейді.

Желіні оқытып (үйретіп) болғаннан кейін оны нақты мәліметтерді талдауға қолдануға болады. Талдаудағы ең басты мәселе жиналған деректерді оқытатын (үйрететін), бақылайтын және тексеретін бөліктерге бөлетіндігі.

Оқытатын (үйрететін) бөліктер желінің параметрлерін анықтау үшін қолданылады. Мұнда деректердің басым көпшілігі пайдаланылады.

Бақылайтын бөлік оқыту процесін қадағалайды. Тексеретін бөлік нәтиженің дәлдігін анықтау үшін қажет.

### 7.4 Розенблатт моделі

Оқытудың мысалын бір қабатты перцептрон (*perceptio-*пайымдау) арқылы келтіруге болады. Бұл моделде қабат саны 1-ге тең, сондықтан екінші индекс қарастырылмайды, яғни  $w(i)$ ,  $1 \leq i \leq N$  сипатталады.

Кіріс сигналы А класына немесе Б класына жатады. Желінің кіруіне  $(x_1, x_2, \dots, x_N)$  сигналы беріледі. Түсінікті болуы үшін  $N=2$  алынады.

*Бірқабатты перцептронды оқыту.*

0-қадам. Алдын-ала мәндер беру. Мұнда  $w_1(1), w_2(1)$  және  $T$ -белсенділік мәндері кездейсоқ мәнге ие болады.  $t$ -оқыту қадамы енгізіледі:  $t=0$

1-ші қадам.  $t = t + 1$ . Желіге  $(x_1(t), x_2(t))$  кіру сигналы беріледі. Егер кіру сигналы  $A$  класына тиесілі болса, онда  $d(t) = 1$ , ал кіру сигналы  $B$  класына тиесілі болса, онда  $d(t) = -1$  болсын.

2-ші қадам.  $t$  уақытындағы нейронның  $s(t) = w_1(t)x_1(t) + w_2(t)x_2(t) - T$  жағдайы арқылы есептеледі.

3-ші қадам.  $t$  уақыты мезетіндегі  $y(t)$  нейронның шығу сигналы  $y(t) = \text{sign}(s(t))$ .

4-ші қадам. Жаңа белсенділік салмағы мынадай формуламен есептеледі.

$$w_1(t) = w_1(t-1) + r(y(t) - d(t)), w_2(t) = w_2(t-1) + r(y(t) - d(t)) \quad (56)$$

мұндағы:  $r$  - оқыту қадамы.

5-ші қадам. Егер  $r < L$  болса,  $L$  - оқыту көлемі, онда 1-ші қадамға барады, ол шарт орындалмаса оқытуды аяқтайды.

Нәтижесінде  $w_1(L), w_2(L), T$  салмақтық мәндері бар оқытылған перцептрондықорытып аламыз. Оқытылған бірқабатты перцептрон қарапайым сұрыптау есептерін шығара алады.

## 7.5 Уақыт қатарларын нейрондық желі арқылы болжау

Уақыт қатарлары деп бір белгінің бірдей уақытаралықтарындағы көрсеткіштерінің жиынтығын айтады. Мысалы, күн сайынғы ауа-райының температурасы, ай сайынғы салық түсімі, жыл сайынғы жалпы ішкі өнім т.с.с.

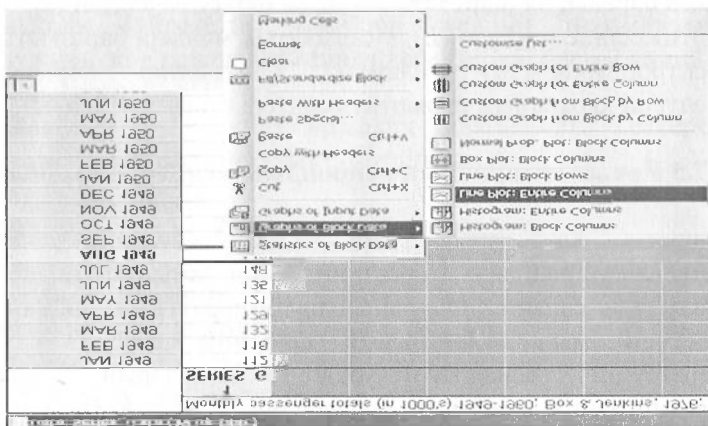
Зерттеудің мақсаты уақыт қатарларын берілген көрсеткіштерінің негізінде болжау болып табылады. Ол үшін бірнеше қадамнан тұратын есептеу жұмысы жүргізіледі.

1-қадам. `File => Openexamples => Datasets => Series_G.sta` командасы арқылы файл ашылады. Файлда АҚШ әуе жолаушыларының ай сайынғы тасымал мәліметтері көрсетілген(87-сурет).

Data: Series_G.sta (1x by 144)	
	Monthly passage
	SERIES G
JAN 1949	112
FEB 1949	118
MAR 1949	132
APR 1949	129
MAY 1949	121
JUN 1949	135
JUL 1949	148
AUG 1949	148
SEP 1949	136
OCT 1949	119
NOV 1949	104
DEC 1949	118
JAN 1950	115
FEB 1950	126
MAR 1950	141
APR 1950	135
MAY 1950	125
JUN 1950	148

87- сурет. Series\_G файлының деректері

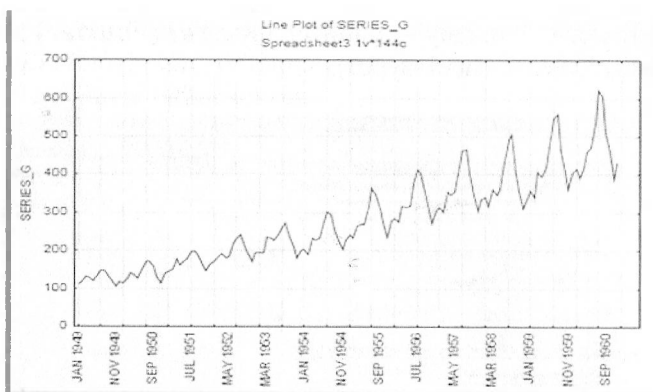
2-қадам. Шыққан мәліметтердің үстіне тінтіреуінді апарып, оң жақ құлағын басып график шығарылады (88-сурет).



88 - сурет. Деректердің графигін шығару

88-суреттегі командаларды орындағаннан кейін шығатын график 89-суретте берілген.



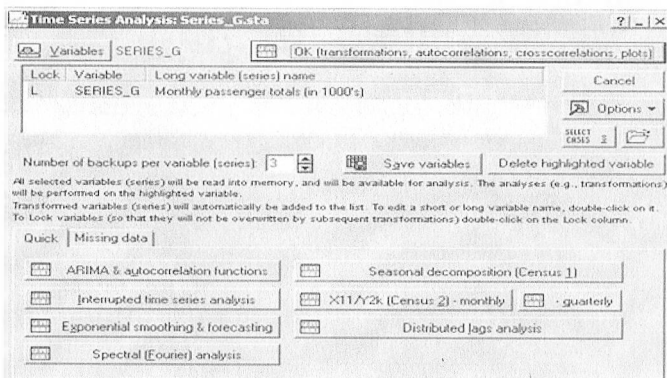


89 - сурет. Деректердің графигі

89 - суреттен уақыт қатарларының өсу ағымында екенін және онда периодтық тербеліс барын көруге болады.

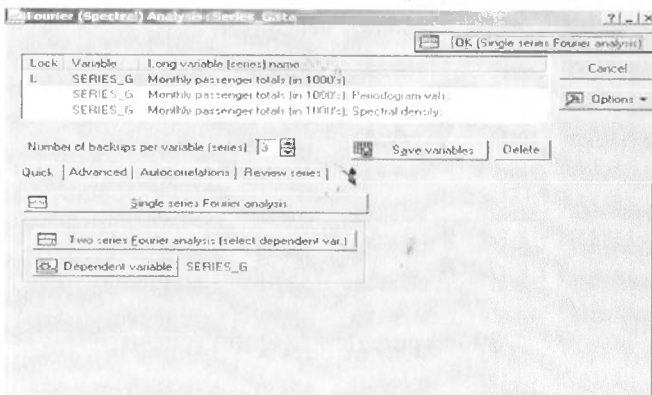
3-қадам. Уақыт қатарларынан жиілікті табу үшін спектралдық талдауға көшу қажет. Ол деректердегі жиіліктің периодын табуға мүмкіндік береді.

*Advanced Linear/ Nonlinear Models* => *Time series/ Forecasting* командасы арқылы Уақыт қатарлары модулін шақыртуға болады (90-сурет):



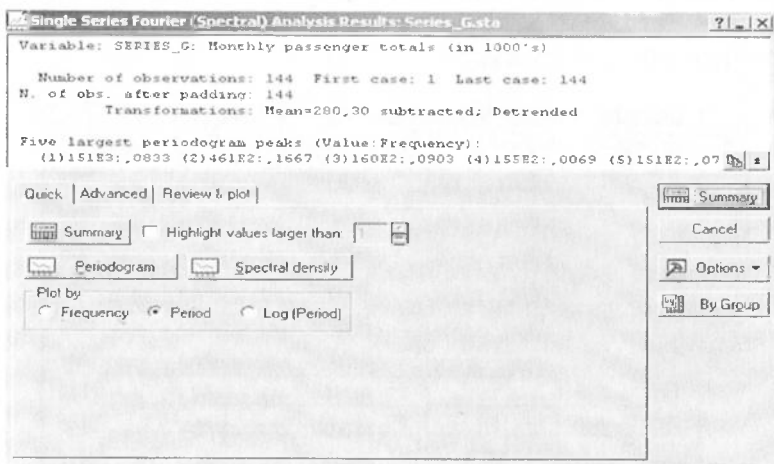
90-сурет. Уақыт қатарларының модулі

Спектралды талдау үшін *Spectral (Fourier) analysis* батырмасы басылады(91-сурет).



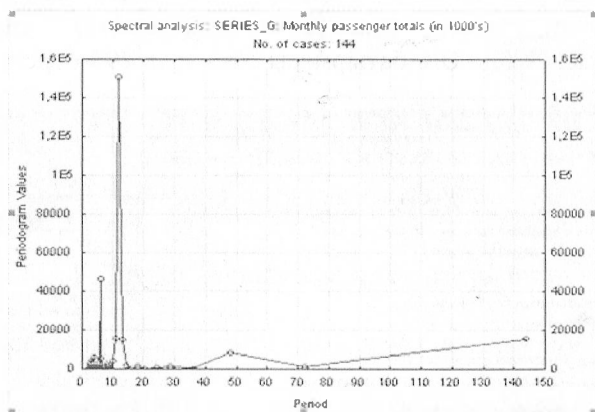
91- сурет. Спектралды талдау терезешесі

91-суреттегі *Single series Fourier analysis* батырмасын басқанда келесі көрініс пайда болады (92-сурет).



92 - сурет. Спектралды талдау мүмкіндігі

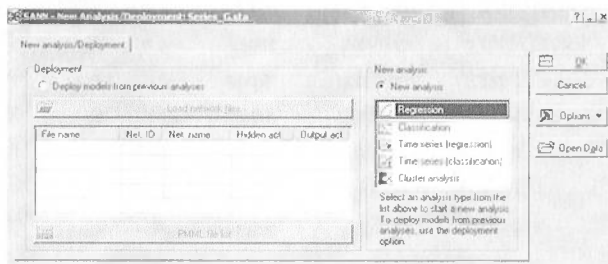
92-суреттегі *Plot by* менюіндегі таңдау нүктесін *Period*-қа қояды. *Periodogram* батырмасын басқаннан кейін шығатын көрініс 93-суретте бейнеленген.



93 - сурет. Деректердің жиілік периодограммасы

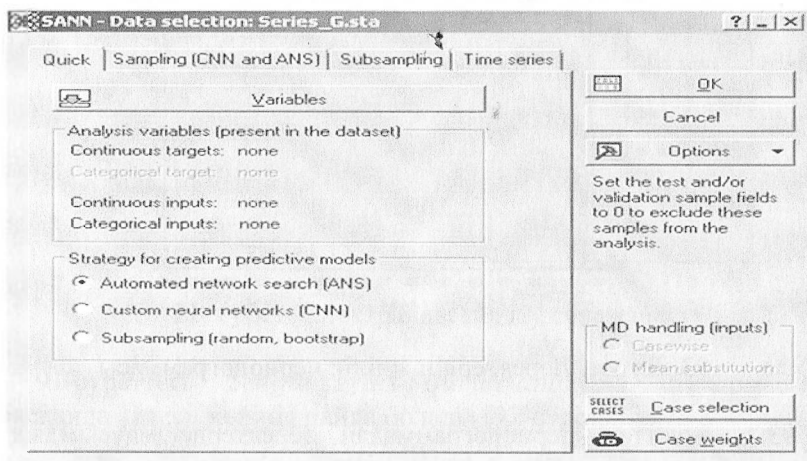
93 - суреттегі периодограммадан деректерде маусымдық период барын және оның 12-ге тең екендігін көруге болады. Демек деректерде жылдық период бар.

4-қадам. Енді «Автоматтық нейрон желілерді» қолдануға болады. Ол үшін *Statistics => Automated Neural Networks* командасы теріледі (94-сурет).



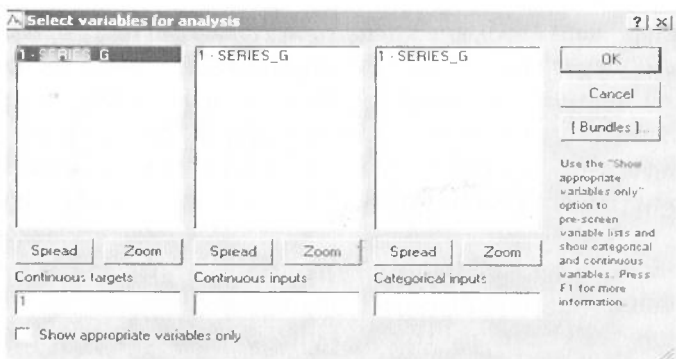
94 - сурет. Нейрон желісінің терезешесі

Нейрондық желі модулінің бастапқы терезесінде талдау түрлері көрсетіледі, яғни мұнда регрессияны, сұрыптауды, кластерлеуді және уақыт қатарларын болжауды жүзеге асыруға болады. Мұнда біз уақыт қатарларының регрессиясын, яғни *Time series (regression)* таңдаймыз. Болжам жасау үшін ОК батырмасы басылады (95-сурет).



95 - сурет. Нейрон желісіндегі талдау

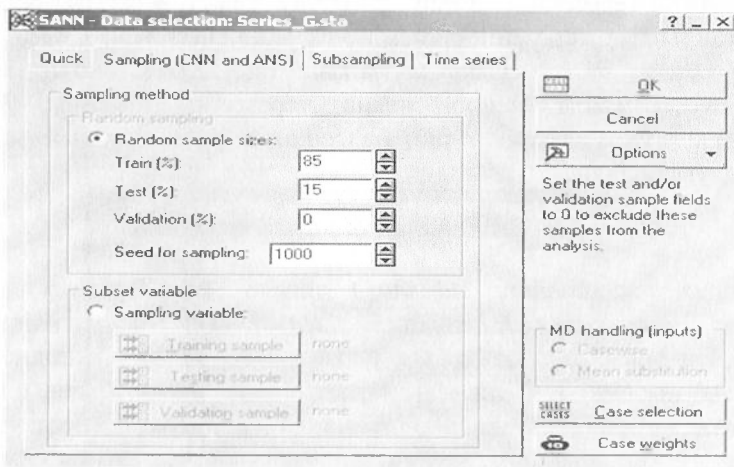
5-қадам. Айнымалыларды таңдау. 95-суреттегі *Variables* батырмасын басып айнымалылардың түрлері нақтыланады. Мұнда үзілмейтін мақсатты (*continuous targets*), үзілмейтін кіріс (*continuous inputs*) және категориялық кіріс (*categorical inputs*) айнымалылары болады. *Series\_G* айнымалысы үзілмейтін мақсатты айнымалысы болып табылады (96 -сурет).



96 -сурет. Айнымалыларды нақтылау

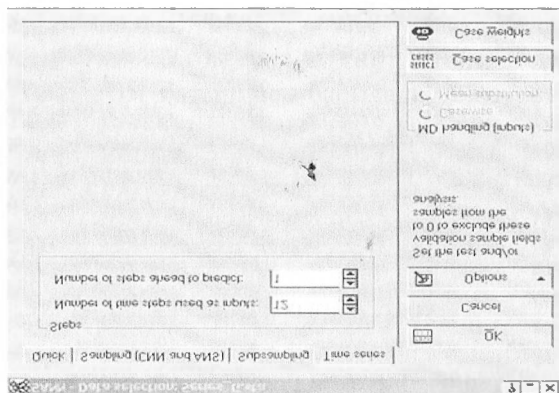
*6-қадам. Оқытатын, бақылайтын және тексеретін жиынды анықтау.*

Уақыт қатарларын талдаған кезде оның әрбір  $t$ -уақыт мезетіндегі мәні  $t - 1, t - 2$  уақыт мезеттеріндегі мәнге тәуелді болады. Мұнда оқытатын мәндерге - 85%, бақылайтын мәндерге-15%, тексеретін мәндерге-0% беріледі (96-сурет)



96 - сурет. Деректердің түрлерін нақтылау

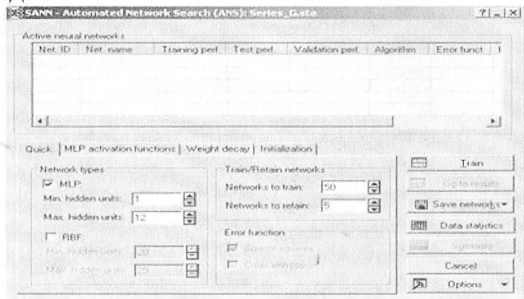
Одан әрі қарай *уақыт қатарлары (time series)* қалтасына көшеміз(97-сурет):



97 - сурет. Деректер ауқымын нақтылау

Болжамдау кезінде нейрондық желіде қанша кіріс каналы болатын терезенің және болжамдық мәндердің көкжиегін көрсететін терезенің өлшемін сипаттау қажет. Жылдық период болғандықтан терезенің өлшеміне 12 мәні енгізіледі.

7-қадам. *Желі архитектурасын таңдау.* ОК батырмасын басып автоматты нейрон желісі терезесіне шығамыз (98 - сурет). Мұнда желі түрін, белсенділік функцияларын т.с.с. таңдауға болады.



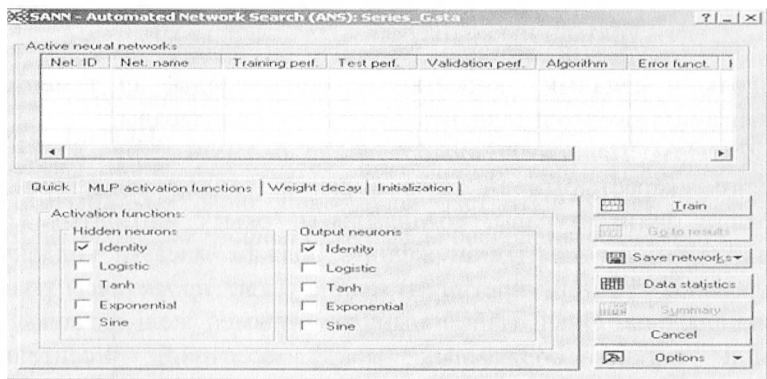
98 - сурет. Нейрон желісінің терезесі

Болжамдық есептерде көпқабатты персептрон пайдаланылады. *Жылдам(quick)* қалтасында желі түрін, жасырын нейрон қабаттарының максимал және минимал мәнін, оқыту және сақтау желілерінің санын көрсетуге болады. Мұнда оқытатын желі саны -50 сақтау желісі -5 болып белгіленеді.

Енді «*Белсенді функциялар*» (*MLP activation functions*) қалтасына көшеміз.

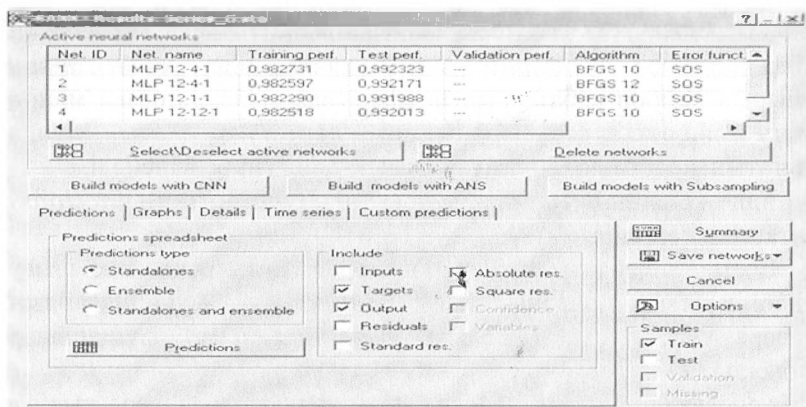
Мұнда көпқабатты персептрондардағы жасырын және шығу нейрондарына қолданылатын белсенділік функциялардың түрлері көрсетіледі. Мысалда *ұқсас (identily)* функциялар таңдалды (99-сурет).

Эксперимент ретінде басқа функцияларды да қарастыруға болады.



99 - сурет. Белсенділік функцияларын қолдану

8 - қадам. *Желіні оқыту*. Сонымен нейрондық желінің құрылымы анықталды. Ол көпқабатты персептрон, онда белсенділіктің *ұқсас* функциялары қолданылады және нейрон саны 1 ден 12 дейін. Енді «*Оқыту*» (*train*) батырмасын басып оқыту процесіне көшеміз. Процестің соңында *Нейрондық желі – Нәтижелер* терезесі шығады (100-сурет).



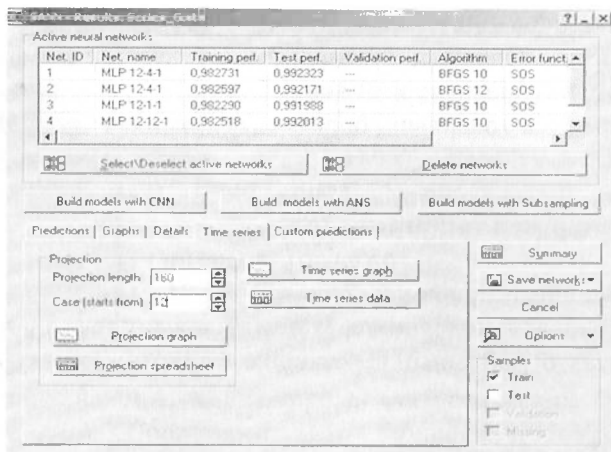
100 - сурет. Нейрондық желі нәтижелер терезесі.

100 - суреттің жоғары жағында ең мықты желілер мен модельдің сапалық көрсеткіштері сипатталады, ал төменінде алынған нәтижелерді талдайтын қалталар орналасады.

*9-қадам. Нәтижені талдау және болжам құру.* Жүйе 50 желіні қарастырып ең тәуір 5 желіні таңдап алды. Терезеде желінің архитектурасы, жылдамдығы, оқыту алгоритмі т.с.с. сипатталады. Оқытылған желіні болжам жасауға қолдануға болады. Оқытылған (үйретілген) желіде уақыт *проекциясы* ұғымы қолданылады. Оның мәнісі мынада: алдымен желі алғашқы 12 мәнді енгізіп болжамдық мәнді есептейді. Есептелген болжамдық мән алдыңғы 11 мәнмен қайта желінің кірісіне беріледі. Осындай алгоритмді қалағанша істеуге болады, бірақ қайталаған сайын болжамның сапасы төмендеуі мүмкін екенін айту қажет.

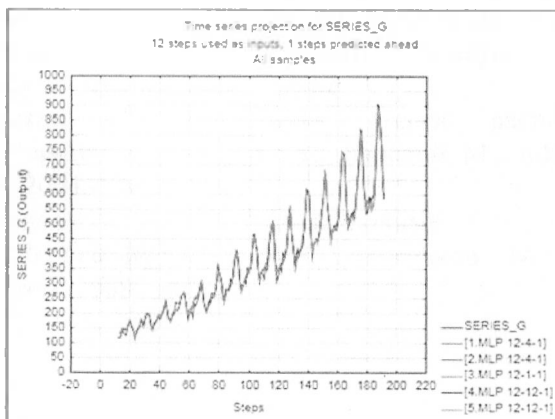
Нейрондық желі – «*Нәтижелер*» терезесінен «Уақыт қатарлары» қалтасын ашамыз (101-сурет).





101 - сурет. Уақыт проекциясын шығару

101 - суреттегі терезешелерді пайдаланып уақыт проекциясын құруға болады. Ол үшін екі параметр қажет *проекцияның ұзындығы* және *бақылауды бастайтын мән*. Проекцияның ұзындығына -180 мәні, бақылауды бастайтын мәнге -12 мәні енгізіліп *проекция графигі* батырмасы басылады (102 - сурет).



102 - сурет. Қорытылған желілердің графигі

102 - суреттен 5-ең тәуір деген желілердің графигі бейнеленген. Олардың барлығының *Series\_G* файлының деректерімен үндесіп тұрғаны көрініп тұр.

Олай болса олардың болжамдық мәндері негізді болып табылады.

### **Сұрақтар мен тапсырмалар**

1. Нейронның мәнісі неде?
2. Розенблатт моделінің ерекшелігін сипаттаңыз.
3. *Series\_G* файлының деректеріне болжам жасаңыз.

## ПАЙДАЛАНЫЛҒАН ӘДБИЕТТЕР

1. Боровиков В.П. Прогнозирование в системе *STATISTICA* в среде *WINDOWS* / В.П.Боровиков, Г.И.Ивченко. –М.: Финансы и статистика, 2000. –383 с.

2. Боровиков В.П. *STATISTICA* Статистический анализ и обработка данных в среде *Windows* /В.П.Боровиков, И.П. Боровиков. –М.: ФИЛИН, 1997.- 603 с.

3. Дубров А.М. Многомерные статистические методы / А.М. Дубров, В.С.Мхитарян, Л.И. Трошин. –М.: Финансы и статистика, 1998. –352 с.

4. Сошникова, Л.А. Многомерный статистический анализ в экономике /Л.А. Сошникова, В.Н. Тамашевич, Г.Уебе –М.: ЮНИТИ, 1999.- 421 с.

5. Айвазян С.А. Прикладная статистика. Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д.Мешалкин.- М.: Финансы и статистика, 1983.-387 с.

6. Льюис, К.Д. Методы прогнозирования экономических показателей /К.Д.Льюис. - М.: Финансы и статистика, 1986.-112 с.

7. Сигел, Эндриу Ф. Практическая бизнес-статистика / Эндриу Ф. Сигел. –М.: Вильямс, 2002.-1055 с.

8. Басовский, Л.Е. Прогнозирование и планирование в условиях рынка /Л.Е. Басовский. –М.:ИНФРА-М, 2006.-178 с.

9. Вуколов, Э.А. Основы статистического анализа / Э.А.Вуколов.-М.: Форум, 2012. - 464 с.

10. Барский А.Б. Нейронные сети: распознавание, управление, принятие решений /А.Б.Барский М.: «Финансы и статистика», 2004.-176 с.

11.Боровиков В.П.Нейронные сети. *STATISTICA Neural Networks*/ В.П.Боровиков М.: Горячая линия –Телеком, 2008.-392 с.