МІНІСТЕРСТВО ОСВІТИ І НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ
Запорізький національний технічний університет

Томашевський О.В., Рисіков В.П.

НАВЧАЛЬНИЙ ПОСІБНИК

КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ СТАТИСТИЧНОЇ ОБРОБКИ ДАНИХ

Рекомендовано
Міністерством освіти і науки України
для студентів вищих навчальних закладів

2-е видання, доповнене

2015

ББК 32.973.233
Т56
УДК 681.3.01

Рецензенти
О.М.Горбань., д.ф.-м.н., проф.,. перший проректор Класичного приватного університета;
В.В.Погосов, д.ф.-м.н., проф., Запорізький національний технічний університет.

(Гриф МОНу № 1/11-5523 від 15.04.2014)

Томашевський О.В., Рисіков В.П.
Т56 Комп'ютерні технології статистичної обробки даних / Навчальний посібник. – Запоріжжя: Запорізький національний технічний університет, 2015. – 175 с.

Викладені теоретичні відомості та розглянуті комп'ютерні технології для основних методів статистичної обробки даних. Зроблен аналіз існуючих програм і пакетів статистичної обробки даних. На прикладах вирішення конкретних задач з різних галузей науки і техніки показано використання комп'ютерних технологій.

Посібник призначений для студентів, що навчаються по спеціальності "Якість, стандартизація та сертифікація", та для інших спеціальностей, де викладаються дисципліни по статистичній обробки даних. Також, посібник буде корисним для фахівців, яким у практичній роботі необхідно використання методів статистичної обробки даних.

ISBN 966-7809-55-2

ББК 32.973.233
УДК 681.3.01

ЗМІСТ

ЗМІСТ 6
ВСТУП 4
1 КОРОТКА ХАРАКТЕРИСТИКА ОСНОВНИХ ПАКЕТІВ СТАТИСТИЧНОЇ ОБРОБКИ 6
1.1 Класифікація пакетів статистичної обробки 6
1.2 Універсальні пакети статистичної обробки 6
1.3 Інструментальні програмні засоби 16
1.4 Статистичні експертні системи 17
1.3 Контрольні запитання і завдання 19
2 ПОЧАТКОВА СТАТИСТИЧНА ОБРОБКА ДАНИХ 20
2.1 Поняття про генеральну сукупність і вибірку 20
2.2 Випадкові величини і їхні характеристики 26
2.3 Розподіл випадкових величин 34
2.4 Підготовка даних у пакеті STATISTICA 37
2.5 Комп'ютерні технології статистичної обробки 44
2.6 Контрольні питання і завдання 51
3 ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ 54
3.1 Основні положення 54
3.2 Перевірка гіпотези про закон розподілу 57
3.3 Перевірка гіпотез про рівність дисперсій і середніх 60
3.4 Комп'ютерні технології перевірки статистичних гіпотез 64
Двовибірковий F-тест для дисперсії 73
Двовибірковий t-тест з однаковими дисперсіями 73
3.5 Контрольні питання і завдання 74
4 КОРЕЛЯЦІЙНИЙ АНАЛІЗ 80
4.1 Основні положення 80
4.2 Кореляційне поле 81
4.3 Вибірковий коефіцієнт кореляції 82
4.4 Кореляційне відношення 85
4.5 Власні коефіцієнти кореляції 87
4.6 Рангова кореляція 88
4.7 Комп'ютерні технології кореляційного аналізу 91
4.8 Контрольні запитання і завдання 97
РЕГРЕСІЙНИЙ АНАЛІЗ 98
5.1 Основні положення 98

5.2 Комп'ютерні технології регресійного аналізу 101
5.3 Контрольні запитання і завдання 109
6 ДИСПЕРСІЙНИЙ АНАЛІЗ 111
6.1 Основні положення.....111
6.2 Однофакторний дисперсійний аналіз 111
6.3 Двофакторний дисперсійний аналіз.....113
6.4 Комп'ютерні технології дисперсійного аналізу.....115
6.5 Контрольні питання і завдання 124
7 ПЛАНУВАННЯ ЕКСПЕРИМЕНТУ 128
7.1 Основні положення.....128
7.2 Повний факторний експеримент.....129
7.3 Центральне композиційне планування 143
7.4 Комп'ютерні технології планування експерименту.....155
7.5. Контрольні питання і завдання 162
ДОДАТОК А СТАТИСТИЧНІ ТАБЛИЦІ.....165
Функція стандартного нормального розподілу 165
Критичні точки розподілу Стьюдента 166
Критичні точки розподілу χ^2.....167
Таблиця значень критерія Колмогорова.....168
Відсоткові точки F-розподілу Фішера, якщо α=1%.....169
Відсоткові точки F-розподілу Фішера, якщо α= 5%.....170
ДОДАТОК В - ТЕСТОВІ ЗАВДАННЯ 171
РЕКОМЕНДОВАНА ЛІТЕРАТУРА 174

ВСТУП

Необхідність статистичної обробки отриманих даних виникає в науці, техніці, економіці, промисловості, бізнесі та багатьох інших областях діяльності людини. Більшість реальних явищ, об'єктів, процесів у навколишній дійсності піддаються впливу різних випадкових факторів, під дією яких змінюють свої властивості і характеристики, тому параметри їх стану – це випадкові величини.

Наприклад, експериментальні результати, одержувані при дослідженні фізичних явищ, як правило, є випадковими величинами. До випадкових величин відносяться показники якості та ціни на продукцію, котирування акцій, параметри стану технічних систем. Технологічний процес як об'єкт статистичного дослідження відноситься до складних стохастичних систем. Стохастичні властивості технологічних процесів досить різноманітні, і з них особливо важливе значення мають ті властивості, що визначають якість продукції, що випускається. Перемінні, що характеризують продукцію, а отже, і технологічний процес залежать від розкиду параметрів вихідних матеріалів, впливу на стан технологічного процесу різних контрольованих і неконтрольованих випадкових факторів і також є випадковими величинами.

До основних етапів статистичної обробки даних можна віднести наступні.

1 етап. Попередній аналіз досліджуваної реальної системи (об'єкта). На цьому етапі визначаються:

- основні цілі дослідження на змістовному (неформалізованому) рівні;
- перелік параметрів (показників), що характеризують стан (поводження) досліджуваної системи (об'єкта);
- формалізована постановка задачі, по можливості що включає стохастичну модель досліджуваного явища.

2 етап складається з:

- планування передбачуваного статистичного аналізу, що включає вибір методів;
- розробка форм надання вихідних даних для статистичного аналізу.

3 етап. Збір статистичних даних і їхня підготовка для обробки на комп'ютері.

4 етап. Первинна статистична обробка. На даному етапі виконується:

- аналіз спостережень, що різко виділяються, і їхнє виключення, відновлення пропущених даних, перевірка статистичної незалежності спостережень вибірки;

- дослідження закону розподілу випадкової величини по вибірці з генеральної сукупності.

5 етап. Вибір комп'ютерних технологій для обробки вихідних даних, що включає:

- вибір і обґрунтування програмних продуктів статистичної обробки;

- формулювання вимог до комп'ютерної системи (розрядність, об’єм пам'яті, швидкодія й ін.).

6 етап. Комп'ютерна обробка статистичних даних.

7 етап. Аналіз результатів статистичної обробки, що складається з:

- інтерпретації результатів;

- аналізу досягнення цілей; підготовки звіту.

У даній роботі розглянуті питання, що відносяться до 5 і 6 етапів, тобто статистична обробка та аналіз даних с допомогою сучасних комп’ютерних технологій..

1 КОРОТКА ХАРАКТЕРИСТИКА ОСНОВНИХ ПАКЕТІВ СТАТИСТИЧНОЇ ОБРОБКИ

1.1 Класифікація пакетів статистичної обробки

Комп'ютерні технології для статистичної обробки даних - пакети статистичної обробки (ПСО) знайшли широке застосування в практичній і науково-дослідній роботі в найрізноманітніших областях. Сучасний світовий і вітчизняний ринок програмних продуктів характеризується великою різноманітністю ПСО [2,6,8,17,18]. За даними Міжнародного статистичного інституту, число програмних продуктів для статистичної обробки наближається до тисячі. Тому для користувача важливо правильно орієнтуватися в цьому різноманітті.

Можна розділити існуючі ПСО на такі групи:

- універсальні, з орієнтацією на певні методи статистичної обробки (методо-орієнтовані);
- інструментальні програмні засоби, що містять статистичний компонент;
- статистичні експертні системи.

Розглянемо загальну характеристику, переваги і недоліки найбільш розповсюджених пакетів.

1.2 Універсальні пакети статистичної обробки

. Весь математико-статистичний інструментарій універсальних пакетів статистичної обробки, як правило, організований у вигляді окремих бібліотек модулів. Кожна з бібліотек містить обробні модулі або їхні групи, що реалізують певні функції ПСО.

Бібліотека 1: допоміжні програми. Вона складається з трьох розділів:

- Методи матричної алгебри. Містить у собі модулі, що реалізують методи рішення систем лінійних рівнянь, обчислення власних чисел і власних векторів в узагальненій постановці задачі;
- Оптимізаційні алгоритми. Забезпечує статистичні модулі необхідними методами й алгоритмами пошуку екстремумів різних функціоналів, що визначають критерії якості статистичного методу (наприклад, метод найменших квадратів і т.п.);

- Статистичне моделювання на ПК. Містить у собі модулі, що реалізують процес машинного генерування послідовностей одномірних і багатомірних спостережень, що добуваються з генеральних сукупностей відповідного типу.

Бібліотека 2: описова статистика і розвідницький аналіз вихідних даних. Зміст бібліотеки визначається основними задачами первинної статистичної обробки даних.

Бібліотека 3: статистичне дослідження залежностей. Це сама об'ємна частина пакета, що тематично розпадається на шість розділів: кореляційно-регресійний аналіз; дисперсійний і коваріаційний аналіз; системи одночасних структурних економетричних рівнянь; планування регресійних експериментів і вибіркових обстежень; аналіз тимчасових рядів; аналіз залежностей марковського типу.

Бібліотека 4: класифікація і зниження розмірності. Тематично розділяється на 5 розділів: дискримінантний аналіз; статистичний аналіз сумішей розподілів; кластер-аналіз (таксономія); зниження розмірності відповідно до критерію автоінформативності (без навчання); зниження розмірності відповідно до критерію зовнішньої інформативності (при наявності навчання).

Бібліотека 5: деякі спеціальні методи статистичного аналізу нечислової інформації й експертних оцінок. Доцільність подібної бібліотеки пояснюється специфікою і дуже інтенсивним розвитком математичних моделей експертного оцінювання, що часом апелюють до вихідних даних нечислової природи, а також до методів і понять, що не укладаються в рамки традиційних схем (наприклад, що оперують з так званими нечіткими множинами). У складі використовуваного в ній математико-статистичного інструментарію: аналіз таблиць спряженості, логлінійні моделі, суб'єктивні імовірності, логіт- і пробіт-аналіз, рангові методи і т.п.

Бібліотека 6: планування експерименту і вибіркових обстежень. Містить модулі для планування експериментів і обробки результатів вибіркових обстежень.

Крім перерахованих шести бібліотек, що поєднують так звані обробні модулі, у пакет входить ряд керуючих модулів і програм: організуюча програма (програма-адміністратор), сервісна програма, бібліотека паспортів модулів та ін.

Для рішення науково-технічних задач та обробки економічної інформації найбільш рейтинговими є універсальні продукти: система SAS, пакет SPSS, SYSTAT. Розглянемо особливості цих продуктів.

Система SAS існує і розвивається з 1976 р. і працює на різноманітних платформах під керуванням однієї з 12-ти операційних систем (ОС). Фірма-розроблювач SAS займає одне з провідних місць серед розроблювачів програмних продуктів, має 3200 співробітників, що підтримують більш 3 мільйонів користувачів у 120 країнах. По суті, SAS сьогодні є могутнім комплексом з понад 20-ти різних програмних продуктів, об'єднаних один з одним «засобами доставки інформації» (Information Delivery System чи IDS, так що весь пакет іноді позначається як SAS/IDS). На ринку статистичного програмного забезпечення основним «козирем» SAS є його неперевершена потужність по набору статистичних алгоритмів. Традиційно склалося, що в СРСР, а потім і в СНД основними користувачами системи є підприємства ВПК, великі бізнесмени (деякі великі банки, включаючи Центробанк, біржі, торгові фірми), деякі атомні станції, найбільші медичні і геофізичні центри, великі державні структури. Під поняттям IDS розроблювач SAS розуміє, що її користувачу досить поставити на свій комп'ютер крім ОС систему SAS і цим обмежитися для 100%-ї інформатизації діяльності будь-якої фірми (всі інші функції типу задач, розв'язуваних на основі Excel, Word, кожної із СУБД та ін. цілком візьме на себе SAS/IDS).

SAS/IDS — це інтеграція різноманітних можливостей доступу до даних і керування ними, засобів аналізу даних, способів представлення інформації і генерації звітів. Система має модульну структуру і легко може бути сконфігурована під специфічні особливості її користувача. У систему входять наступні модулі.

Модуль BASE SAS — ядро системи з вбудованою мовою програмування 4GL і мовою роботи з базами даних SQL, засобу керування даними, підтримки індексів для баз даних, можливостями доступу до широкого набору форматів даних, процедури описової статистики і генерації звітів.

Модуль FSP забезпечує повноекранний доступ до даних, уведення, редагування, перетворення даних, генерацію звітів і ділове переписування.

Модуль GRAPH містить ділову, наукову, рекламну графіку, різні шрифти і карти. Дружні до користувача засоби малювання і

редагування підтримують створення складних графічних елементів, таких як складні графіки, тривимірні поверхні, різноманітні стовпчикові чи кругові діаграми з будь-яким ступенем параметризації.

Модуль STAT містить у собі багатофункціональний набір статистичних процедур аналізу даних.

Модуль IML являє собою інтерактивну *матричну* мову програмування для виконання поглиблених математичних, інженерних і статистичних розрахунків. Ця мова дає можливість математику легко програмувати свої власні процедури, використовуючи мову, близьку до мови лінійної алгебри.

Модуль LAB забезпечує користувачу експертну підтримку. Зокрема, тут система підказує користувачу, виконуються чи ні припущення, що лежать в основі того чи іншого методу аналізу даних.

Модуль EIS є меню-керованим інструментом розробки і підтримки могутніх інтерактивних інформаційних систем, що виконуються, *методом об'єктно-орієнтованої* технології. За допомогою цього модуля легко настроїти систему на свої дані і форми представлення результатів.

Модуль ACCESS дає можливість будувати окремі інтерфейси для зв'язку SAS/IDS з найрізноманітнішими СУБД (ADABAS, DB2, ORACLE, SQL/DS і ін.).

Модуль INSIGHT являє собою у високому ступені інтерактивний інструмент для графічного аналізу даних.

Ясно, що з вищеописаних модулів - «цеглин» можна будувати будь-які, «як завгодно високі будинки», однак процес освоєння технології будівництва, самого будівництва, а також одержання ліцензії на «право забудови» зажадає чималих інтелектуальних і матеріальних витрат.

Переваги і недоліки пакета. Основними перевагами SAS є могутнє інтелектуальне ядро, підтримка всіх п'яти архітектур клієнт-сервер, можливість доступу й інтеграції даних з будь-яких джерел і наявність об'єктно-орієнтованої технології швидкої розробки додатків. При цьому, завдяки винятковій гнучкості системи, додаток, створений в одній з ОС, може бути перенесений на кожну з платформ, підтримуваних SAS/IDS, починаючи від суперЕОМ типу CRAY до Mainframe чи робочої станції.

Головні недоліки системи - громіздкість, великі труднощі в освоєнні, високі вимоги до статистичної кваліфікації користувача,

тверді вимоги до апаратної частини ЕОМ, великий її розмір на диску. Особливістю пакета є і його дорожнеча (за кожен модуль близько $300).

Пакет SPSS став відомий у науковому і діловому світі, будучи реалізований на великих машинах. Основними користувачами його «пакетного варіанта» традиційно були вчені, що працюють в академічних інститутах і університетах, а також у різноманітних додатках математичної статистики, наприклад, в області контролю якості.

Як і SAS, пакет призначений у першу чергу для статистиків-професіоналів, тому що має досить могутній апарат статистичного аналізу, цілком порівняний по потужності з SAS.

Завдяки переорієнтації розроблювачів в останні роки на платформу Windows, програма SPSS стала в даний час одним з лідерів серед універсальних статистичних пакетів. SPSS для Windows відрізняється різноманітними можливостями по керуванню даними і маніпулюванню отриманими результатами, по роботі з електронними таблицями, надає досить зручну графіку (більш 50 типів діаграм), а також розвиті засоби підготовки звітів. Аналітичні параметри відображаються на екрані у вигляді простих і ясних меню і діалогових вікон. На основі DDE і OLE технологій фірми Microsoft а також стандарту ODBC-2.0 у SPSS також вирішені питання обміну з іншими Windows-додатками і виконується зв'язок з більшістю форматів баз даних. Так, можна дуже легко переносити отримані тестові чи графічні результати з SPSS у документи Word.

*Переваги і недоліки***.** Після SAS у своїй повній конфігурації SPSS для Windows має дуже повний набір статистичних (усього їх більше 60-ти) і графічних процедур, а також процедур створення звітів. Також творці пакета пишаються інтерфейсом SPSS з користувачем, вважаючи його дуже простим і зручним. Крім того, традиційно пакет відрізняється високою точністю обчислень.

Однак, ліцензійні диски з повним комплектом SPSS коштують чималі суми. Так, модулі «Заглиблена статистика» і CHAID коштують, відповідно, близько $500 і $700 при покупці в дистриб'ютора в РФ; нейромодуль же коштує майже тисячу USD.

Система SYSTAT. Ця система універсального характеру розроблена однойменною фірмою, що з вересня 1994 р. «поглинена» корпорацією SPSS. Пакет SYSTAT відрізняється від ряду інших

універсальних систем типу SAS, SPSS, тим, що він на початковому етапі був спроектований під платформу IBM PC. Головна перевага пакета, як втім і пакетів SAS і SPSS, - широкий діапазон і глибина проробки функціонального наповнення. Тут є широкі можливості для не досить підготовленого в статистиці користувача і для досить досвідченого статистика.

У 90-х роках пакет був одним із кращих серед універсальних пакетів поглибленого статистичного аналізу. Журнал Software Digest (Rating Report), що видається лабораторією NSTL, назвав SYSTAT у травні 1991 р. найкращим статистичним пакетом універсального характеру (general-purpose). Відзначалася висока якість статистичних алгоритмів пакета, його явне домінування в області планування експериментів. Пакет має чудову графіку, по якій ще недавно пакет був одним із кращих у своєму класі. Однак, із сучасних позицій проглядається уже певне відставання в графіку в режимі «високого розрізнення». Документація пакета містить у собі чотири томи. Це ясно і добре написана інструкція до виконання «Як почати роботу», а також посібник з розділів «Графіка», «Статистика». Є і невеликий «Посібник з даних».

Пакет керується через легко використовуване меню. Уведення даних і їхнє редагування здійснюється через затабульоване вікно. Для зручності виклику найбільш частих статистичних процедур у SYSTAT уведені клавіші QuickStat. Робоча таблиця з даними легко активізується і ясна на інтуїтивному рівні. Пакет має прекрасні можливості відображення на екрані вихідних даних і отриманих результатів розвідницького аналізу, маючи у своєму розпорядженні біля 30 різних способів графічного відображення: гістограми, шухляди з «вусами», стебла з листям, іконки, 2-D і 3-D діаграми розсіювання і т.д. Крім того, є матриці діаграм розсіювання, графіки функцій і географічних карт. Є можливість легко комбінувати координатні осі, поверхні, контури, діаграми розсіювання. Крім того, пакет дозволяє породжувати і зображувати складні поверхні, що корисно для візуалізації складних функцій. Натисканням кнопки миші легко обертати навіть складні 3-D графіки з координатними осями по відношенню до площини екрану. Для багатьох графіків є спеціальні засоби типу стрілки, для того щоб досліджувати крапки-викиди, ключі з діапазонами для режиму «лупа» чи «ласо», для дослідження виділеного фрагмента даних. Графіка пакета гнучка, легко керована й

об’єктно орієнтована. Є можливості інтерактивних графічних перетворень даних, що дуже зручно при розвідницькому аналізі. Також є засоби розробки презентацій.

*Переваги та недоліки***.** SYSTAT має добру та заслужену репутацію через його точність, використовуючи багато чудових алгоритмів. Він має досить велике меню з функціональними алгоритмами, включаючи описову та непараметричну статистику, кореляцію, кластерний аналіз, перевірку багатомірних гіпотез для загальної лінійної моделі й таблиці спряженості. Пакет дає можливість чудової роботи у всіх областях статистики, але особливо він сильний в області дисперсійного аналізу й планування експериментів. Має багато додаткових процедур для дискримінантного аналізу, матричної алгебри, логлінійних моделей, планування експериментів, структурного аналізу і карт контролю якості. Також були додані робастні (стійкі) алгоритми, що дають точні й коректні результати при майже вироджених даних. Крім того, ця версія надає користувачу найбільш широкі можливості аналізу загальної лінійної статистичної моделі.

Деяка незручність роботи з пакетом зв'язана з тією обставиною, що частина операцій доступна лише з командного рядка. Іноді пояснення в інструкції для користувача дається для спрощеного варіанта меню, а деталі використовуваного статистичного методу даються тільки як інструкції до командного рядка. Деякі розділи меню містять у собі менше, ніж це було б потрібно для оптимального дружнього інтерфейсу з користувачем. Також, кілька важливих статистичних методів рішуче не є дружніми до користувача (наприклад, по непарному t-критерії й простому однофакторному дисперсійному аналізу).

Пакет STATGRAPHICS призначений в основному для тих користувачів, що вже мають певний досвід у статистиці. Особливо це стосується модуля з багатомірними методами. Для коректного їхнього використання користувач повинен мати базові знання по статистиці й знати допуски й обмеження тих чи інших статистичних критеріїв і багатомірних методів. Пакет на початковому етапі був розроблений для платформи IBM PC і націлений, у першу чергу, на *графічні* можливості комп'ютерної статистики. Однак, постійне його удосконалювання в плані функціональних алгоритмів і способів керування даними істотно підсилило даний пакет і підвищило його

конкурентноздатність. У 1995 р. пакет STSC був визнаний однієї з найбільш ефективних інтегрованих систем статистичного аналізу даних на ПЕОМ. Його найважливішим «плюсом» вважається вдале з'єднання математичного апарата обробки даних із сучасною інтерактивною графікою. Також до його переваг відносяться широкі можливості взаємодії з електронними таблицями і СУБД. Обмін з таблицями в Windows-версії виконується через стандартний буфер обміну (Windows clipboard).

Пакет Statgraphics має 8 великих розділів про методи математичної статистики. Крім того, у плані суміжних задач, пакет містить такі розділи, як Дескриптивна статистика, Розвідницький аналіз, Багатомірний аналіз. Пакет Statgraphics побудований за модульним принципом.

Базовий модуль містить ряд загальних процедур, а також процедури лінійної регресії. У розділі *описової статистики* можна проаналізувати одну чи декілька змінних, підігнати ті чи інші теоретичні функції розподілу, одержати необхідні статистики. Є можливість розрахувати коефіцієнти кореляції Пірсона й рангової кореляції Спірмена, але інших засобів ранжирування змінних немає.

Традиційно пакет має великі і дуже гнучкі графічні можливості: у наявності не тільки 2-D — кольорові, але й 3-D — графіки, діаграми прямокутників з контактами, графіки на сітці, частотні гістограми, діаграми розсіювання, стовпчасті й кругові діаграми.

Багатомірний аналіз винесені, як і ряд інших розділів статистики, у додаткові модулі. Усього їх чотири: контролі якості,

Цікавою новинкою є введення *сеансу роботи StatFolio*, що планування експериментів, тимчасові ряди й багатомірні методи. зовні нагадує папку. У цій «папці» поєднуються вихідні дані, а також застосовані до них аналітичні процедури, отримані вихідні графіки (у виді піктограм). Сеанси можна комбінувати по два або в більшій кількості. Так, наприклад, якщо Ви бажаєте виконати аналіз нових даних, тоді просто додайте їх в обраний сеанс.

*Переваги та недоліки***.** Пакет має широкий діапазон графічних засобів, у пакеті легко настроїти ті чи інші параметри, потрібні для видачі графіка на екран чи на будь-який зовнішній пристрій, включаючи кольоровий принтер. Роботу з графіками легко освоїти, тому що багато чого зрозуміло навіть на інтуїтивному рівні. Таким

чином, саме блискуча графіка є «нішею пакета» на ринку універсальних пакетів.

До недоліків можна віднести те, що в базовому модулі засоби створення звітів слабкі, зокрема, не можна в одному звіті об'єднати текст і графіку. Додаткові модулі підсилюють засоби створення звітів. Убудована система довідки містить простий для вживання словник, що не є цілком гіпертекстовим. Трохи незручна довідкова система. Так, деякі особливості пакета, що роблять його легким у використанні для підготовленого користувача, можуть заводити «не в той степ», недосвідченого користувача. Пакет некоректно працює при проведенні парних порівнянь на основі t-критерію. У плані коректності обчислень, очевидно, пакет трохи уступає своїм найближчим конкурентам SYSTAT і SPSS.

У документації на додаткові модулі багато тестів добре пояснюються, однак ясність і глибина викладу матеріалу явно недостатні. Документація на модуль багатомірного аналізу відразу ж активно використовує ті чи інші методології кластерного аналізу, які зовсім не пояснюються на багатьох її сторінках. Це змушує непідготовленого користувача навчатися основам статистики самостійно, щоб коректно використовувати блок багатомірних методів.

Пакет STATISTICA - це могутній пакет статистичного аналізу, у якому реалізовані всі новітні комп'ютерні і математичні методи аналізу даних. Це універсальна інтегрована система, призначена для статистичного аналізу і візуалізації даних, керування базами даних і розробки користувальницьких додатків. Крім загальних статистичних і графічних засобів, у системі є спеціалізовані модулі, наприклад, для проведення соціологічних чи біомедичних досліджень, вирішування технічних і промислових задач. Пакет орієнтований на користувача, що добре володіє методами статистичних досліджень, але його базові модулі можуть бути використані і новачками в статистиці.

Пакет Statistica за своєю структурою складається зі зв'язаних між собою модулів, що взаємодіють один з одним, маючи однаковий формат системних файлів. Так, якщо потрібен розділ лінійної регресії, тоді необхідно залишити оточення головного модуля і вийти в оточення модуля лінійної регресії. У плані функціонального наповнення пакет відрізняється великою розмаїтістю, включаючи в

себе і ті розділи аналізу, що в інших пакетах (наприклад, Statgraphics) містяться лише в додаткових модулях (що поставляються за додаткову ціну). Зокрема, він містить у собі ряд непараметричних методів аналізу, методи багатомірного аналізу: дискримінантного, факторного кластерного логлінійного й ін.

*Переваги та недоліки***.** Пакет зручний для керування, дані легко ввести в середовище пакета, відносно легко відредагувати, створити нові змінні («ознаки»), вибрати окремі спостереження чи «вирізати» підмножини даних по рядках і/чи по стовпцях таблиці «об'єкт-ознака». Завдяки великій панелі інструментів, для виконання більшості задач досить кілька клацань мишки, тому що майже для усіх функцій пакета тут є піктограми. Крім того, клацанням правої кнопки мишки можна викликати додаткові підменю, що істотно прискорюють роботу з пакетом.

Цікавою особливістю пакета є настроювання функцій під екран, відкритий у даний момент часу. Так, при завантаженні програми в пам'ять машини в активному вікні виникає список модулів, доступних користувачу в даний момент часу. Звідси користувач може самостійно вирішити, якого сорту аналіз йому необхідно зараз виконати. Список модулів і порядок їхнього проходження у вікні можуть бути визначені користувачем, що дає йому додаткові зручності в гнучкості настроювання.

Statistica має можливість роботи в пакетному режимі, використовуючи свою командну мову SCL. Можна використовувати і набори команд, які поєднуються в послідовності чи макроси. За допомогою макрокоманд зручно готувати презентації. Крім того, пакет надає засоби складання звітів.

Пакет можна зв'язувати різними Windows-додатками. Завдяки підтримці DDE, неважко виконати ті чи інші командні сценарії зсередини інших додатків. Наприклад, можна в Excel написати макрос, що запускає пакет Statistica. Після додавання в макрос спеціальних SQL-команд можна імпортувати в пакет дані. Використання OLE технології обміну між Windows-додатками дозволяє легко інтегрувати результати, наприклад, Word і Statistica.

Найбільш сильною стороною пакета є графіка і засоби редагування графічних матеріалів. Представлено сотні типів графіків: типу 2-D чи 3-D (є навіть графіки типу 4-D), матриці і піктограми. Є можливість розробити свій дизайн графіка і додати його в меню.

Засоби керування графіками містять у собі роботу одночасно з декількома графіками, зміну розмірів складних об'єктів, розширені можливості малювання з додаванням художньої перспективи і рядом спеціальних ефектів, розбивку сторінок і швидке перемальовування. Наприклад, 3-D графіки можна обертати, накладати один на одного, стискати чи збільшувати. Передова анімаційна техніка, яка застосовується у версії 5.0 і відноситься скоріше до області мистецтва, дозволяє побачити на графіках, які крапки там змінилися під впливом змін в одній зі змінних.

Крім того, пакет має докладну документацію і коротку інструкцію до виконання, добре і детально описані у монографії [2]. В екранний довідник входить майже вся інформація друкованої документації. Рекомендації, які містяться в документації й екранному довіднику, корисні, але часом недостатньо повні.

До недоліків можна віднести відсутність методів планування експериментів, контролю якості. У цілому пакет Statistica по широті охоплення статистичних методів уступає і SAS, і SPSS.

1.3 Інструментальні програмні засоби

До таких засобів варто віднести, насамперед додаток Excel, пакети Mathcad, Mathematica, що містять різноманітні математичні й статистичні засоби і мають модульну структуру (аналіз часових рядів, модулі фінансового аналізу, розвідницького аналізу, вирішування диференціальних рівнянь, цифрової обробки сигналів та ін.) й інтерактивне середовище S-plus.

Додаток **Excel** дає можливість не дуже підготовленому користувачу виконати початкову статистичну обробку даних, перевірити статистичні гіпотези, здійснити регресійний, дисперсійний і інший види аналізу даних. До складу Microsoft Excel входить набір статистичних функцій і засоби аналізу даних (так званий пакет аналізу), призначений для вирішування статистичних задач. Проведення статистичного аналізу здійснюється досить просто - варто указати вхідні дані і вибрати за допомогою меню потрібну статистичну функцію чи в пакеті аналізу - макрофункцію (інструмент аналізу), результати відобразяться на листі вікна додатка. Докладніше про можливості додатка Excel для проведення статистичного аналізу в п.1.4.

У пакеті **Mathcad** реалізовані основні функції для аналізу експериментальних даних (середнє, медіана, дисперсія, кореляція й ін.), розподілів і проведення регресійного аналізу. MathCAD має зручний користувальницький інтерфейс. У цієї системи є й ефективні засоби типової наукової графіки, вони прості в застосуванні й інтуїтивно зрозумілі, розрахунки в системі MathCAD орієнтовані на масового користувача. Відмінна риса - реалізується зручне і наочне об'єктно-орієнтоване програмування найскладніших задач, при якому програма складається автоматично за завданням користувача, а саме завдання формулюється на природній математичній мові спілкування із системою.

Вбудовані в пакет **Mathematica** «електронні ноутбуки» дозволяють легко організувати текст, результати обчислень і графіку у виразні технічні звіти і презентаційні матеріали. Так само легко одержати 2-D і 3-D графіки і виконати інші способи візуалізації даних.

S-plus являє собою інтерактивне середовище, що містить у собі повноцінний графічний аналіз даних і мову програмування, що є розширюваним і зручним для використання. Середовище S-plus засноване на S мові програмування, що розроблені в AT&T Bell Labs і є об'єктно-орієнтованою мовою, спеціально призначеною для аналізу даних. Саме тому одержувані результати мають необмежену свободу при проведенні досліджень, аналізу і моделювання даних у науці та техніці.

S-plus може бути дуже корисною для статистика-аналітика, що вміє складати свої програми на основі об'єктно-орієнтованої технології. S-plus має у своєму складі більш 1650 функцій, включаючи регресію і дисперсійний аналіз, багатомірні методи, тимчасові ряди, аналіз сигналів та ін. Є і сучасні робастні (стійкі) методи. В плані класифікації є сучасні непараметричні методи: деревоподібні моделі класифікації, моделі цілеспрямованого проектування даних на площину, узагальнені адитивні моделі. Середовище S-plus має могутні засоби візуалізації а також додаткові модулі, орієнтовані під аналіз сигналів чи часових рядів, планування експерименту, аналіз просторової статистики.

1.4 Статистичні експертні системи

Відрізняються наявністю бази знань і механізмом логічного виводу з наявних знань нових знань. Прикладами таких програм по

темі нашої статті є казахстанський пакет СТАТЕКС, американський Statistical Navigator Pro і англійський STAREX.

Дві головні відмінності пакетів, що містять ознаки *експертної системи*:

- вона орієнтована не на методи, а на мету аналізу даних (останні пропонуються системою в процесі роботи з нею);
- користувач може зовсім не розбиратися в механізмі обробки даних, але повинен чітко розуміти зміст його даних і загальну мету аналізу.

Інші специфічні риси:

- діалог орієнтований на користувача-новачка в статистиці;
- професійний статистик може в спеціальному режимі роботи пакета безпосередньо звернутися до методів аналізу, використовуючи назви останніх;
- результати аналізу видаються у вигляді контекстно-орієнтованих екранів, що містять коментарі, що дозволяє розглядати їх як готові рішення;
- *база знань* являє собою набір правил, зв'язаних із властивостями й особливостями застосування статистичних методів; *база даних* дозволяє зберігати їх у вигляді «куба»: таблиця «об'єкт-ознака» і «час».

У системі **СТАТЕКС** у функціональному плані реалізовані наступні групи методів:

- розрахунок стандартних статистичних 1–D-характеристик;
- класифікація об'єктів (комбінаційне групування, кластерний аналіз) а також багатомірне шкалування і візуалізація;
- виявлення й аналіз статистичних залежностей ознак (кореляція, групування ознак, головні компоненти і візуалізація);
- встановлення залежностей між цільовим показником і факторами, що впливають на нього (регресійний аналіз, індексний аналіз і розпізнавання образів);
- прогнозування (економетричні моделі).

Більшість методів СТАТЕКС поряд із класичними містять оригінальні результати її авторів.

Система **Statistical NavigatorPro** допомагає неспеціалісту в математичній статистиці провести кваліфікований аналіз наявних у нього даних. Він консультує користувача по використанню більш, ніж

200 різних методів, включаючи багатомірний аналіз і класифікацію. Statistical Navigator задає ряд питань щодо мети дослідження даних і їхнього характеру, супроводжуючи кожне з питань серією підказок. Пакет фокусує увагу користувача на ключових моментах, підкреслюючи, що необхідно прийняти до уваги для ухвалення вірного рішення.

З усього різноманіття ПСО виділимо і розглянемо детально пакет статистичної обробки Statistica і програму Excel. Такий вибір обумовлений тим, що пакет Statistica має надзвичайно широкі можливості щодо застосування різних методів статистичного дослідження, а розповсюджена програма Excel дозволяє кваліфіковано виконати статистичну обробку даних користовачам з невеликим досвідом роботи у цьому напрямку.

1.3 Контрольні запитання і завдання

1.3.1.Запитання

1. Загальна характеристика основних універсальних пакетів статистичної обробки даних.
2. Інструментальні програмні засоби, що містять статистичний компонент.
3. Статистичні експертні системи.

1.3.2. Завдання

1. Вивчити і дати загальну характеристику статистичного пакета STATGRAPHICS.
2. Описати введення і редагування даних у пакеті STATGRAPHICS.
3. Вивчити і дати загальну характеристику статистичної системи аналізу NCSS.

2 ПОЧАТКОВА СТАТИСТИЧНА ОБРОБКА ДАНИХ

2.1 Поняття про генеральну сукупність і вибірку

Зміна стану різних об'єктів навколишнього середовища, технічних, виробничих і економічних систем відбувається в результаті впливу різних контрольованих і неконтрольованих факторів. Під поняттям дані будемо розуміти результати спостережень за параметрами стану об'єкта (системи) чи результати експериментів над ними.

Дослідження різних об'єктів чи систем здійснюється за допомогою експериментів чи спостережень. Розглянемо типові приклади проведення експериментів для дослідження властивостей об'єктів (систем) і спостережень за параметрами, що характеризують їх стан:

1) кидаються монети, у результаті випадає “орел” чи “решка”;
2) кидається гральна кістка, кожне кидання дає в результаті одне з чисел 1,2,3, ... 6;
3) виміряється ріст і вага представника деякої групи тварин;
4) береться вибірка з партії виготовлених транзисторів, і для кожного виміряються коефіцієнт підсилення і зворотні струми;
5) спостерігається ціна якого-небудь товару через певний проміжок часу, наприклад через тиждень;
6) здійснюється експеримент для установлення впливу певного фактора чи сукупності факторів на параметри досліджуваного об'єкта (системи);
7) беруться вибірки зі сховища даних інформаційної системи, що відбивають продажі деякого товару за розмірностями час і регіон.

Експерименти по киданню монети чи гральної кістки можна виконувати багато раз (абстрактно, звичайно). Результати виміру росту і ваги представника деякої групи тварин можуть бути отримані по тим, що живуть, і що жили раніше індивідам, і кількість індивідів досить велика і визначити її точне значення нереально. Для 4-го прикладу обсяг вибірки і партії являють собою цілком визначені значення. Те ж можна сказати щодо коливання ціни. У 6-му прикладі, навіть якщо прийняти всі міри, що б контролювати умови, що

впливають на експеримент, результати експерименту можуть мінятися при його повторенні. В останньому прикладі повна сукупність даних, що відбивають продажі, постійно змінюється і з часом, і по регіонах.

Дані, одержувані в наведених вище прикладах, а число таких прикладів можна розширити, можуть складати кінцеву чи нескінченну, постійну чи змінну сукупність. При дослідженні властивостей різних об'єктів, процесів, явищ, визначаються поняття генеральної сукупності і вибірки з генеральної сукупності.

Генеральна сукупність – усі мислимі значення (виміри, спостереження), що описують поводження досліджуваного об'єкта, процесу чи явища.

Вибірка з генеральної сукупності – обмежений набір вибіркових з генеральної сукупності значень, що реально спостерігаються і описують досліджуваний об'єкт, процес чи явище. Кількість цих значень називається **обсягом вибірки**.

Висновки про поводження об'єкта, процесу чи явища робляться по вибірці обмеженого обсягу і поширюються на всю генеральну сукупність.

Позначимо кількість об'єктів, що підлягають обстеженню, N. Припустимо, що кожному *i* об'єкту відповідає значення x_i. Згідно даного раніше визначення сукупність усіх можливих значень (теоретично домислюємих), що характеризують *N* об'єктів, називається ***генеральною сукупністю***, а *N* – ***обсягом генеральної сукупності***. Генеральна сукупність може бути кінцевою або нескінченною, тобто N може бути кінцевою або $N \to \infty$.

Нехай кількість реально спостерігаємих об'єктів із *N* дорівнює *n*. Тоді $x_i = \overline{1,n}$ – ***вибірка з генеральної сукупності***, *n* – ***обсяг вибірки***.

Вибірка з генеральної сукупності повинна мати наступні властивості:

- кожен елемент x_i обраний випадково;
- усі x_i мають однакову імовірність потрапити у вибірку;
- *n* повинне бути настільки великим, наскільки це дозволяє вирішувати задачу з необхідною якістю (вибірка повинна бути репрезентативною, представницькою).

Надалі будемо мати справу з вибіркою, що має перераховані властивості. Прийнято вважати, що при $n \geq 60$ вибірка **велика**, чи репрезентативна, а при $n < 60$ – **мала**. Такий розподіл вибірки на велику і малу умовний.

Поняття **репрезентативна вибірка** не завжди можна зв'язати з її обсягом *n*. Частіше це залежить від реально досліджуваного об'єкта чи явища, обсягу генеральної сукупності, трудомісткості і вартості одержання спостережень чи вимірів для формування вибірки. Можливі ситуації, коли генеральна сукупність мала. Наприклад, досліджується експериментальна партія інтегральних схем малого обсягу (генеральна сукупність визначається обсягом N виготовленої партії) або час наробітку до відмовлення унікального устаткування, коли в експлуатації знаходиться свідомо мала кількість його екземплярів (генеральна сукупність визначається кількістю екземплярів N). Доступного для дослідження устаткування (вибірка обсягом *n*) може бути ще менше. Тому вибірка обсягом *n*, близьким до обсягу генеральної сукупності N, може вважатися репрезентативною й одночасно малою.

Приклад. Здійснюється запуск 10 дослідних партій інтегральних схем (ІС) обсягом по 30 пластин, на кожній з який формується по 1000 кристалів. Вихідні пластини були узяті випадковим чином, що забезпечує випадковий розкид значень кристалографічних і електрофізичних параметрів по пластинах. Партії при обробці розподілялися також випадковим чином по робітниках і одиницям однорідного устаткування. Вибірка *n*=100 була утворена кристалами, узятими по10 з кожної партії. Вибірка репрезентативна.

Форми представлення вибірки з генеральної сукупності можуть бути наступними.

1. Представлення вибірки з генеральної сукупності у **негрупованому вигляді** x_i, i=1,n.

Така форма зв'язана з наявністю зведень про кожен елемент вибірки. Наприклад, у таблиці 2.1.1 наведені результати вимірів статичного коефіцієнта передачі струму емітера транзисторів середньої потужності (обсяг вибірки *n*=10).

Табл.2.1.1.

n	1	2	3	4	5	6	7	8	9	10
x_i	0,975	0,967	0,967	0,969	0,972	0,965	0,972	0,985	0,963	0,968

2. Представлення вибірки у вигляді **варіаційного ряду** (в упорядкованому вигляді):

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(i)} \le \ldots \le x_{(n)}.$$

У цьому випадку $x_{(i)}$ – член варіаційного ряду або **варіанта** Індекс (i) указує на порядковий номер елемента у варіаційному ряді. Форма представлення вибірки з генеральної сукупності у вигляді варіаційного ряду не приводить до втрати інформації про кожен елемент вибірки, але спотворює інформацію, установлюючи залежність між сусідніми елементами вибірки. Представлення вибірки попереднього приклада (табл.2.1.1) у вигляді варіаційного ряду показано в табл.2.1.2.

Табл.2.1.2.

i	1	2	3	4	5	6	7	8	9	10
x_i	0,960	0,965	0,967	0,968	0,970	0,975	0,975	0,976	0,977	0,981

3. Представлення вибірки у **групованому вигляді**. Така форма представлення вибірки з генеральної сукупності зв'язана з розбивкою області завдання випадкової величини X на L інтервалів групування. При цьому відома тільки кількість елементів вибірки n_j,, що потрапили в j інтервал, і послідовність границь інтервалів розбивки. Для визначення числа L інтервалів штучного групування користуються формулою Старджеса.

$$L = 1 + 3.322 \lg n . \qquad (2.1.1).$$

Іноді L може бути задано природою досліджуваного явища чи умовами проведення експерименту. У деяких випадках ширина кожного інтервалу може бути відмінною від інших (нерівноточне групування).

Розглянемо послідовність процедури представлення вибірки в груповану вигляді.

1. Формування варіаційного ряду.
2. Виділення мінімального і максимального елементів вибірки

$$x_{min} = x_{(1)},$$
$$x_{max} = x_{(n)}.$$

3. Визначення числа інтервалів групування здійснюється або з розуміння точності і встановлюється емпіричним шляхом у

залежності від обсягу вибірки, або за формулою Старджеса (1.1), або визначається особливостями проведення експерименту. Округлення при визначенні L здійснюється до найближчого цілого числа.

4. Визначення ширини інтервалів гістограми (при рівноточному групуванні)

$$h = \frac{x_{(n)} - x_{(1)}}{L}.$$

Якщо при обчисленні h необхідно округлити результат, варто пам'ятати, що останній інтервал групування буде менше ширини h при округленні у велику сторону і більше h - при округленні в меншу сторону.

5. Формування послідовності границь інтервалів розбивки. Утворений варіаційний ряд границь інтервалів групування буде виглядати як

$$x_{(1)},\ x_{(1)} + h,\ x_{(1)} + 2h,\ \ldots,\ x_{(1)} + (L\text{-}1)\,h,\ x_{(n)}.$$

Іноді, для того щоб $x_{(1)}$ і $x_{(n)}$ потрапили усередину відповідно 1-го і L-го інтервалів групування, границі $x_{(1)}$ і $x_{(n)}$ коректують у такий спосіб:

$$x'_{(1)} = x_{(1)} - h/2,$$
$$x'_{(n)} = x_{(n)} + h/2.$$

Отже, число інтервалів розбивки збільшується на 1:

$$L' = L + 1.$$

При цьому послідовність границь інтервалів розбивки буде представлена у вигляді

$x'_{(1)},\ x'_{(1)} + h,$	$x'_{(1)} + h,\ x'_{(1)} + 2h,$	$,\ldots,$	$x'_{(1)} + L'h,\ x'_{(n)}.$

6. Визначення кількості елементів вибірки n_j, що потрапили в кожен j інтервал.

Групована форма представлення випадкової величини не містить інформації про кожен елемент вибірки. При цьому часто як значення випадкової величини на інтервалі приймається його середина.

Приклад. Результати вимірів зворотних струмів (у наноамперах) на тестових діодних структурах у вибірці обсягом n=50 приведені в табл.2.1.3.

Результати вимірів зворотних струмів діодних структур

Табл.2.1.3

15	19	6	18	21	16	20	17	15	10
16	20	7	19	22	17	21	19	16	11
19	10	8	18	20	8	18	16	20	12
16	21	21	9	19	19	14	18	19	19
12	20	20	8	13	10	18	17	22	18

Представимо вибірку в групованому вигляді.

1. Формуємо варіаційний ряд з даних табл.2.1.3.

Табл.2.1.4.

6	7	8	8	8	9	10	10	10	11
12	12	13	14	15	15	16	16	16	16
16	17	17	17	18	18	18	18	18	18
19	19	19	19	19	19	19	19	20	20
20	20	20	20	21	21	21	21	22	22

2. Знаходимо $x_{(1)} = 6$, $x_{(n)} = 22$.
3. Визначаємо число інтервалів розбивки за формулою Старджеса (2.1.1)

$$L = 1 + 3{,}322\lg 50 = 6.6\,, \quad \text{тобто} \quad L=7.$$

4. Знаходимо ширину інтервалу розбивки h

$$h = (22 - 6) / 7 = 2.2857.$$

Обмежимося двома знаками після коми й одержимо h=2.28. Оскільки h округлене убік зменшення, те останній інтервал буде ширше попередніх.

5. Будуємо варіаційний ряд границь інтервалів групування (без коректування границь першого й останнього інтервалів): [6 ; 8.28], [8.28 ; 10.56], [10.56; 12.84], [12.84; 15.12], [15.12; 17.4], [17.4 ; 19.68], [19.68 ; 22]. Та ж процедура, але з коректуванням границь першого й останнього інтервалів дасть наступні результати:

$$L'=L+1=7+1=8,$$
$$x'_{(1)} = x_{(1)} - h/2 = 6 - 1.14 = 4.86,$$
$$x'_{(n)} = x_{(n)} + h/2 = 22 + 1.14 = 23.14.$$

Одержуємо послідовність границь інтервалів розбивки для L=8: [4.86; 7.14], [7.14; 9.42], [9.42; 11.7], [11.7; 13.98], [13.98; 16.26], [16.26; 18.54], [18.54; 20.82], [20.82; 23.14]. Ширина останнього інтервалу в тому і іншому випадку (L=7 і L□=8) дорівнює h=2.32.

6. Знаходимо кількість елементів вибірки n_j, що потрапили в j інтервал (випадок без коректування границь інтервалів).

j-ий інтервал	1	2	3	4	5	6	7
n_j	4	5	3	4	8	14	12

Знаходимо n_j, (випадок коректування границь інтервалів розбивки).

j-ий інтервал	1	2	3	4	5	6	7	8
n_j	2	4	4	3	8	9	14	6

Будь-які характеристики випадкової величини, отримані по вибірці з генеральної сукупності, називаються ***вибірковими***, або ***емпіричними, характеристиками***, а характеристики, отримані по генеральній сукупності, – ***теоретичними,*** або ***генеральними, характеристиками.***

2.2 Випадкові величини і їхні характеристики

Загальним для наведених у п.2.1. прикладів результатів спостережень є те, що одержання певного чисельного значення росту, коефіцієнт підсилення, ціни, чи продажів, будь-яка подія (випадання "орла", якого-небудь числа грального кістки) характеризується деякою імовірністю.

При визначенні імовірності одержання певного чисельного значення деякої величини чи події використовується поняття частоти

появи даного значення чи події – як відношення числа випадків його появи до загального числа спостережень чи вимірів.

Імовірність визначається як межа, до якої прямує частота появи певного чисельного значення чи частота здійснення даної події при безупинному збільшенні числа спостережень. Теорія імовірностей доводить існування такої межі і збіжність частоти до імовірності при наближенні числа спостережень до нескінченності.

Величина, для якої одержання її конкретного значення зв'язується з імовірністю, називається **випадковою величиною**. Випадкова величина може приймає ті чи інші числові значення в залежності від різних випадкових обставин. Одержувані при спостереженнях чи вимірах значення випадкової величини непередбачені, можна тільки задатися імовірністю одержання того чи іншого значення. Параметри стану, що характеризують досліджувані об'єкти (системи), можуть підпадати під вплив неконтрольованих факторів, зовнішнього середовища, тобто є випадковими величинами. Випадкові величини підрозділяються на дискретні і неперервні:

- дискретні випадкові величини можуть приймати тільки конкретні, заздалегідь обговорені значення (наприклад, значення чисел на верхній грані кинутої гральної кістки чи кількість відмов у партії виробів при випробуваннях);
- неперервні випадкові величини можуть приймати будь-які значення, найчастіше в деякому визначеному інтервалі (наприклад, значення параметрів інтегральної схеми, час безвідмовної роботи транзистора, об’єм продаж у ціновому виразі).

Результати спостережень чи вимірів утворюють вибірку. Будь-які характеристики випадкової величини, отримані по вибірці з генеральної сукупності, називаються ***вибірковими*** або ***емпіричними характеристиками***, а характеристики, отримані по генеральній сукупності, – ***теоретичними*** або ***генеральними характеристиками.***

Вибіркові, (*емпіричні) характеристики.* За результатами вибірки можна оцінити центр групування, розсіювання і розподіл імовірностей випадкової величини.

Для оцінки центра групування вибірки випадкових величин $x_1,x_2,,\ ..\ ,..x_n$ використовуються:

- середнє арифметичне $\overline{x}$;

- медіана;
- мода.

Медіана дорівнює середньому члену вибірки випадкових чисел $x_1,x_2,,\ ..\ ,..x_n$, розташованих у зростаючому порядку, при непарному n. Якщо n - парне, медіана дорівнює напівсумі двох середніх членів. Модою називається таке значення випадкової величини, імовірність одержання якого найбільша.

Розсіювання випадкових величин усередині вибірки $x_1,x_2,,\ ..\ ,..x_n$ характеризується:

- емпіричною дисперсією s^2;
- розмахом.

Розмах визначається як різниця максимального і мінімального членів вибірки.

Статистичним розподілом випадкової величини називається розташована в порядку зростання сукупність значень випадкових величин (варіаційний ряд) із вказівкою імовірності їхнього виникнення.

Статистичний розподіл може бути представлений у вигляді **гістограми або полігона**. Гістограма - це ступінчастий графік, ординати якого відповідають частоті чи частості попадання даного випадкового у визначений інтервал значення, відкладених по осі абсцис. Відмінність полігону від гістограми в тім, що роль ступінчастого графіка грає багатокутник, кути якого відповідають частоті або частості. При побудові гістограми або полігона використовується представлення вибірки в групованому вигляді. Частість f_i (чи відносна частота) визначаються за формулою:

$$f_i = \frac{m_i}{n};$$

де: m_i - кількість попадань значень випадкової величини в i-ий інтервал (абсолютна частота);

n – об'єм вибірки;

Розглянемо приклад побудови гістограми на основі вибірки випадкових величин $x_1,x_2,,\ ..\ ,..x_n$,.де n=100, наведеної в табл.2.2.1.

Результати вимірювань коефіцієнта підсилення транзисторів

Таблиця 2.2.1

	1	2	3	4	5	6	7	8	9	10
1	944,8	993,4	920,8	899,4	1042,0	840,6	958,5	1043,0	1018,4	1034,4
2	867,4	666,5	937,4	763,2	701,3	889,0	795,2	919,6	975,9	883,6
3	800,8	1006,0	998,5	842,4	883,5	1019,6	1002,1	1114,7	937,4	1091,8
4	943,5	917,2	808,4	842,0	981,1	989,3	972,1	754,3	897,2	1052,3
5	745,2	869,5	915,9	991,1	1090,9	1016,4	938,9	942,5	851,4	925,8
6	687,6	891,5	818,5	1028,3	965,0	975,6	770,9	690,5	934,9	888,8
7	950,0	870,7	901,8	1004,3	892,8	946,9	882,7	810,8	827,4	1042,5
8	889,6	954,7	961,4	966,5	1083,6	920,6	785,3	831,3	938,6	1103,1
9	1049,7	976,9	857,6	968,1	828,0	981,4	930,6	835,7	882,9	991,1
10	1046,9	766,0	946,1	876,7	903,2	860,1	930,0	869,5	820,9	874,8

В табл.2.2.2 показано число елементів $m_i^{'}$, *що* потрапили в кожен *і-й* інтервал і результати підрахунку відносних частот f_i.

Таблиця 2.2.2

i	$(X_i \div X_{i+1}]$	$m_i^{'}$	f_i
1	(650 ÷ 700]	3	0,03
2	(700 ÷750]	2	0,02
3	(750 ÷ 800]	6	0,06
4	(800 ÷ 850]	12	0,12
5	(850 ÷ 900]	20	0,20
6	(900 ÷ 950]	21	0,21
7	(950 ÷ 1000]	17	0,17
8	(1000 ÷ 1050]	13	0,13
9	(1050 ÷ 1100]	4	0,04
10	(1100 ÷ 1150]	2	0,02
11	(1150 ÷ 1200]	0	0
	$\sum$	100	1

Відзначимо, що

$$\sum_{i=1}^{L} m_i^{'} = n,$$

де *L*- число всіх інтервалів, наведених у табл. ; *і - номер* інтервалів.

За даними табл.2.2.2 будуємо емпіричну (експериментальну) функцію розподілу, що може бути показана у вигляді гістограми (рис.2.2.1).

При невеликому числі N випробувань іноді багатокутник розподілу виявляється сильно перекрученим. У таких випадках рекомендується збільшити ширину інтервалів.

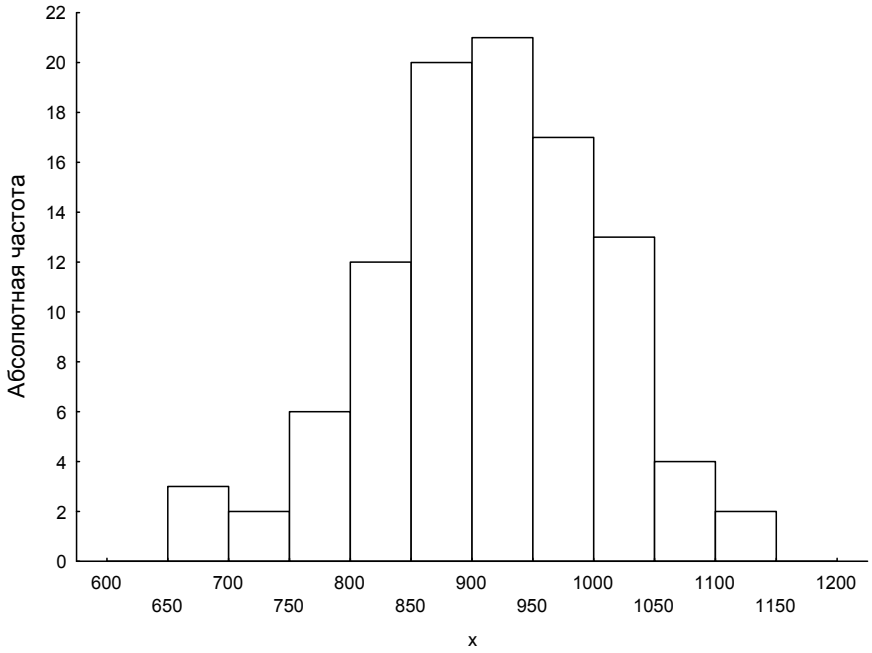


Рисунок 2.2.1 - Емпірична функція розподілу у виді гістограми

Теоретичні (генеральні) характеристики. Функція розподілу випадкової величини, отримана по генеральній сукупності, називається ***теоретичною функцією розподілу*** і позначається *F(x)*.

Розподіл випадкової величини *x* може бути описаний аналітично за допомогою теоретичного розподілу з функцією густини імовірності *f (x)*.

$$f(x) = \frac{dF(x)}{dx}$$

Основними *теоретичними* характеристиками є:

- математичне очікування;
- дисперсія;
- коефіцієнт асиметрії;
- коефіцієнт ексцесу.

Математичне очікування характеризує центр розподілу випадкової величини *x* і визначається за формулою:

$$E(x) = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

Оцінюється *E(x)* середнім арифметичним для значень, отриманих при спостереженнях (вимірах) випадкової величини *x*.

Дисперсія характеризує розсіювання випадкової величини *x* і визначається за формулою:

$$D(x) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$$

Для характеристики форми розподілу використовуються **коефіцієнти асиметрії й ексцесу.** Коефіцієнт асиметрії визначає асиметрію розподілу і задається як

$$в_1^2 = \frac{м_3^2}{м_2^3},$$

де $\mu_{оо}$, $\nu_л$ - центральні і початкові теоретичні моменти порядку *k,* що визначаються виразами:

$$м_k = \int_{-\infty}^{\infty} (x - н_1)^k f(x)dx,$$

$$н_k = \int_{-\infty}^{\infty} x^k f(x)dx.$$

Для симетричного розподілу β_1 (і μ_3) дорівнює 0. Наприклад, для моделі нормального розподілу β_1=0. При $\beta_1<0$ ($\mu_3<0$) розподіл має лівосторонню асиметрію, при $\beta_1>0$ ($\mu_3>0$) - правосторонню.

Коефіцієнт ексцесу β_2 також є характеристикою форми розподілу, а саме його гостровершинності, і визначається виразом:

$$в_1^2 = \frac{м_3^2}{м_2^3},$$

Величина $\gamma=\beta_2$-3 називається ***приведеним коефіцієнтом ексцесу***. Для нормальної моделі випадкової величини $\gamma=0$, *а* $\beta_2=3$.

Теоретичні характеристики оцінюються відповідними їм вибірковими (емпіричними) характеристиками (табл.2.2.2).

Табл.2.2.2

Теоретичні характеристики	Вибіркові характеристики
Математичне очікування *E(x)*	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Дисперсія *D(x)*	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Початковий момент *k-го* порядку $\nu_л$	$н_k^* = \frac{1}{n}\sum_{i=1}^{n} x_i^k$
Центральний момент *k-го* порядку $\mu_{до}$	$м_k^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - н_1^*)^k$
Коефіцієнт асиметрії $(\beta_1)^2$	$(в_1^*)^2 = \frac{(м_3^*)^2}{(м_2^*)^3}$,
Коефіцієнт ексцесу β_2	$в_2^* = \frac{(м_4^*)}{(м_2^*)^2}$,

Для характеристики випадкових величин можуть використовуватися - **квантиль порядку *p*** чи **p-квантиль x_p.** Фізичний зміст квантилі полягає в наступному: x_p – це таке значення випадкової величини, нижче якої лежить *p-а* частина розподілу.

Квантиль рівня $p=0.5$, $x_{0,5}$ називається медіаною. Іноді замість квантилі користуються процентною точкою або процентиллю.

Процентиль $x_{q\%}$ – це таке значення випадкової величини, вище якого лежить q(%) розподілу. Інакше процентною точкою рівня q(%) $0<q<100$ називається таке x_q, при якому виконується умова

$P\{x \geq x_q\} = q/100$.

Теоретична квантиль x_p є функцією, інверсною функції розподілу випадкової величини, і визначається з виразу $F(x_p)=p$.

Якщо p=0.25, 0.5, 0.75, то x_p відповідно **1-а, 2-а і 3-я квартилі**.

2.3 Розподіл випадкових величин

Функція розподілу випадкової величини задає імовірність того, що випадкова величина прийме значення менше, ніж x, тобто імовірність перебування випадкової величини в інтервалі $[-\infty, x]$. У загальному вигляді функція розподілу визначається як $F(x,\theta)$, де θ - вектор параметрів. Для найбільш розповсюджених законів розподілу неперервних випадкових величин $\theta=(a,b,c)$, де a– параметр положення (зрушення), b – параметр масштабу (розсіювання), c – параметр форми. Якщо параметри a і b відомі, то можна використовувати нормовану форму представлення випадкової величини $X_N = (x-a)/b$.

Функція розподілу $F(x,\theta)$ має наступні властивості:

1) $0 \le F(x,\theta) \le 1$;
2) $F(x_2,\theta) \ge F(x_1,\theta)$, якщо $x_2 \ge x_1$;
3) $F(x,\theta) = 0$, при $x \to -\infty$, $F(x,\theta) = 1$, при $x \to \infty$

При відомій функції густини імовірності $f(x,\theta)$:

$$F(x,\theta) = \int_{-\infty}^{x} f(x,\theta)dx.$$

Для неперервної випадкової величини функція густини імовірності $f(x,\theta)$ показує імовірність одержання величини x в інтервалі $[x, x+dx]$, для дискретної випадкової величини функція густини імовірності показує імовірність одержання величини x, тобто дискретному набору значень $x_1, x_2, \ldots$ зіставляються імовірності $f(x_1,\theta)$, $f(x_2,\theta)$, … ...

Найбільш розповсюджені закони розподілу неперервних і дискретних випадкових величин наведені в таблиці 2.3.1.

Основні закони розподілу неперервних і дискретних випадкових величин

Табл.2.3.1.

Вид розподілу	Функція густини імовірності $f(x,\theta)$:
Неперервні розподіли	
Нормальне	$f(x,\theta)=\frac{1}{b\sqrt{2p}}\exp(\frac{(x-a)^2}{2b^2})$
Вейбулла	$f(x,\theta)=\frac{c}{b}\left(\frac{x-a}{b}\right)^{c-1}\exp(-\frac{(x-a)^c}{b})$
Експонентне	$f(x,\theta)=\frac{1}{b}\exp(\frac{(x-a)}{b})$
Рівномірне	$f(x,\theta)=\frac{1}{b-a}\quad,\quad x\in[a,b]$
Дискретні розподіли	
Біноміальне	$f(x,\theta)=C_n^x p^x q^{n-x}$
Пуассона	$f(x,\theta)=\frac{л^x}{x!}\exp(-л)$

Нормальний розподіл розглянутий вперше А.Муавром у 1733р., а в 1809р. знову відкрито незалежно від А.Муавра К.Гауссом. Нормальний розподіл (Муавра-Гаусса) займає провідне місце в теорії і практиці, зокрема, у техніці, економіці, соціології, медицині, біології й ін. При нормуванні випадкових величин функція нормального розподілу представляється функцією Лапласа

$$\text{Ц}(X_N)=\frac{1}{\sqrt{2\text{p}}}\int_0^{X_N}\exp((-z^2)/2)\text{dx}\,.$$

Нормальний розподіл симетричний відносно μ і має наступні характеристики: математичне очікування $M(x)=\mu=a$, дисперсію $D(x)=\sigma^2=b^2$, коефіцієнт асиметрії $\beta_1=0$, коефіцієнт ексцесу $\beta_2=3$, приведений коефіцієнт ексцесу $\gamma=0$.

Розподіл Вейбулла досить часто використовується в теорії надійності для опису імовірності безвідмовної роботи, густина імовірності відмовлень, часу наробітку до відмовлення й ін. для нормованих випадкових величин (a=0, b=1) має математичне очікування $M(x)=\Gamma(c+1)$, дисперсію $D(x)=\Gamma((c+2)/c)-\Gamma^2((c+1)/c)$, де $\Gamma(\cdot)$ – гамма-функція.

Експонентний розподіл добре описує випадкові величини, що характеризують тривалість життя (демографію), інтенсивність відмов (теорія надійності) і ін. і для нормованих випадкових величин має математичне очікування $M(x)=b$, дисперсію $D(x)=b^2$.

Рівномірний (прямокутний) розподіл знаходить застосування при аналізі часу очікування "обслуговування", при точно періодичному, через кожний Т проміжок часу, прибутті (включенні) обслуговуючого пристрою, і при випадковому надходженні заявки на обслуговування в цьому інтервалі часу. Функція рівномірного розподілу визначається як

$$F(x,\theta)=\begin{cases}0, & при \quad x<a,\\ \dfrac{x-a}{b-a}, & при \quad a\le x<b\\ 1 & при \quad x\ge b\end{cases}.$$

Математичне очікування $M(x)=(a+b)/2$, дисперсія $D(x)=(b-a)^{2/12}$.

Біноміальний закон показує імовірність появи деякої події x раз (тобто дискретної випадкової величини, що приймає значення 1,2, …) при n незалежних випробуваннях, за умови, що p – імовірність появи цієї події і $q=1-p$ - імовірність його непояви. Математичне очікування $M(x)=np$, дисперсія D(x)=npq. Біноміальний закон можна використовувати для опису числа дефектних виробів у партії, для оцінки числа покупців у магазині, що зробили покупку й ін.

Закон Пуассона є граничним випадком біноміального при зростанні числа випробувань $n\to\infty$, коли $q\to 0$. Використовується, якщо необхідно описати появу будь-яких подій (наприклад, виникнення відмов) на заданому інтервалі часу. Інтервал часу зв'язується з параметром розподілу λ. Математичне очікування $M(x)=\lambda$, дисперсія $D(x)=\lambda$..

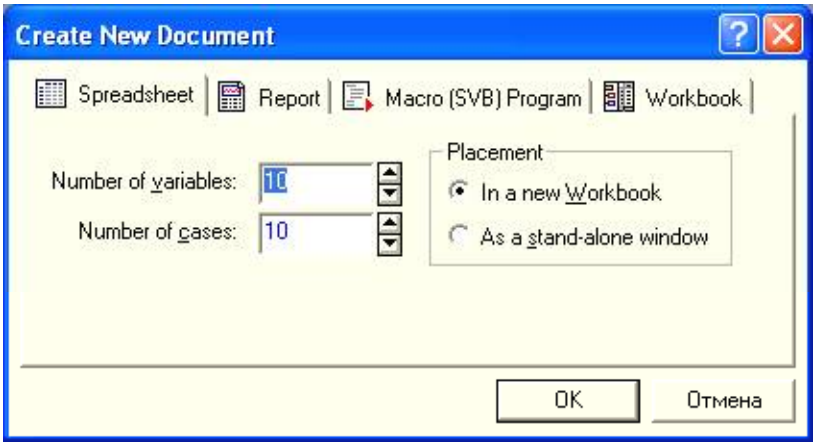
Вигляд функцій густини імовірності для розподілів, наведених у табл.2.3.1, наочно проілюстрований у модулі Probability calculator пакета STATISTICA

2.4 Підготовка даних у пакеті STATISTICA

Вихідні дані в пакеті STATISTICA організовані у вигляді електронних таблиць (Spreadsheets) і складаються з рядків і стовпців. Стовпці таблиці називаються Variables - Перемінні, а рядки Cases - Спостереження.

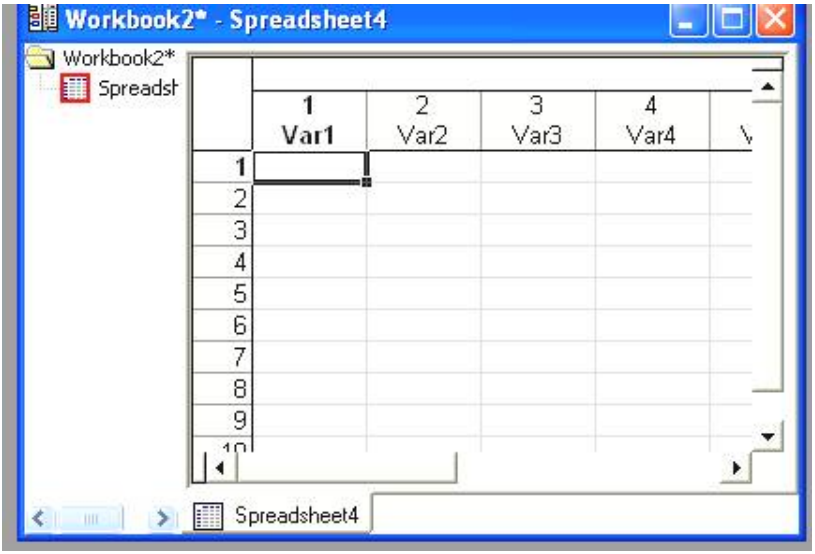
Розглянемо уведення вихідних даних на прикладі версії StatSoft Statistica v.6.0. Вихідні дані можна вводити безпосередньо в пакеті зі створенням файлів вихідних даних у форматі *.stw чи імпортувати дані з інших систем, наприклад, з додатка MS Excel.

Для безпосереднього уведення вихідних даних можна використовувати команду File → New, після чого відкривається вікно Create New Document, де на вкладці Spreadsheet задається число рядків і стовпців.



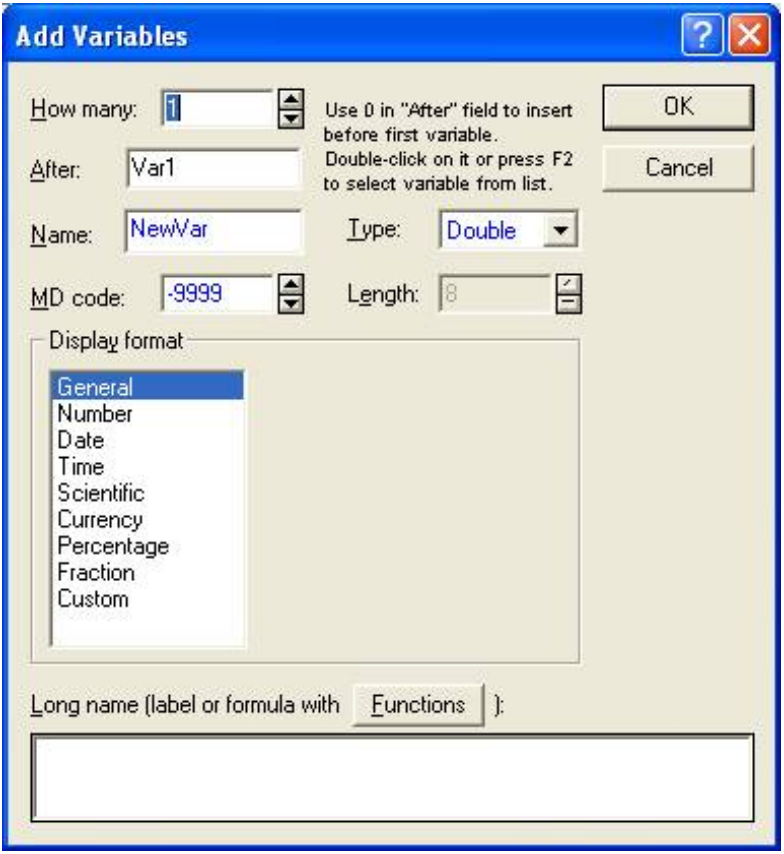
Після клацання по кнопці OK відкривається вікно Workbooks - Spreadsheet з електронною таблицею, уведення даних у яку здійснюється так само, як і в додатку Excel.

Можна збільшити кількість, перемістити, копіювати, видаляти як рядки, так і стовпці цієї таблиці, для чого використовуються команди DATA → cases чи DATA → vars. Після вибору команди DATA → cases → Add на екрані виникне меню, що пропонує наступний вибір: Add…(додати), move…(перемістити), copy…(копіювати), Delete…(видалити).

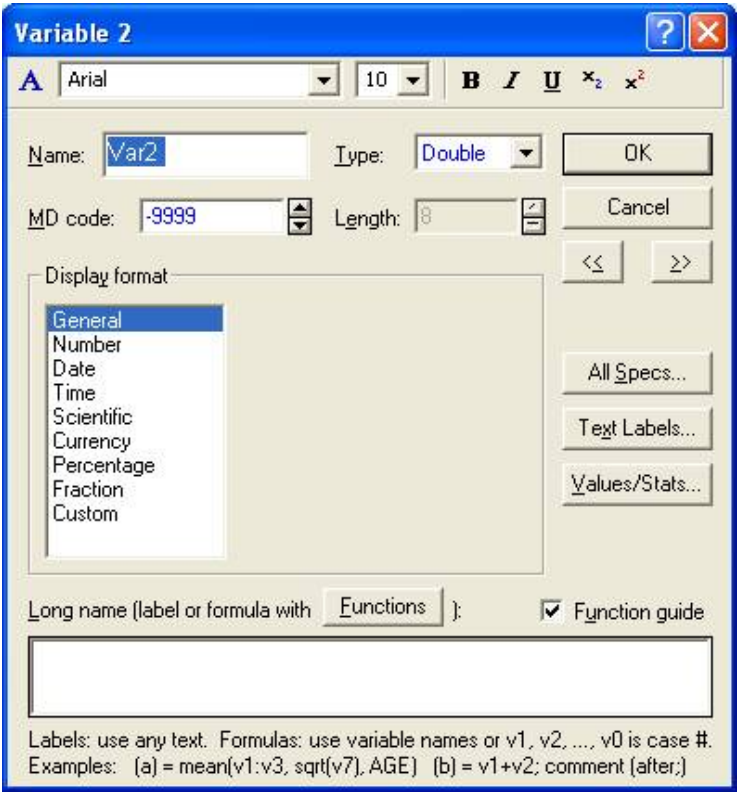


Після вибору, наприклад, Add, з'являється вікно Add Cases, що дозволяє вказати після якого рядка скільки додається рядків.

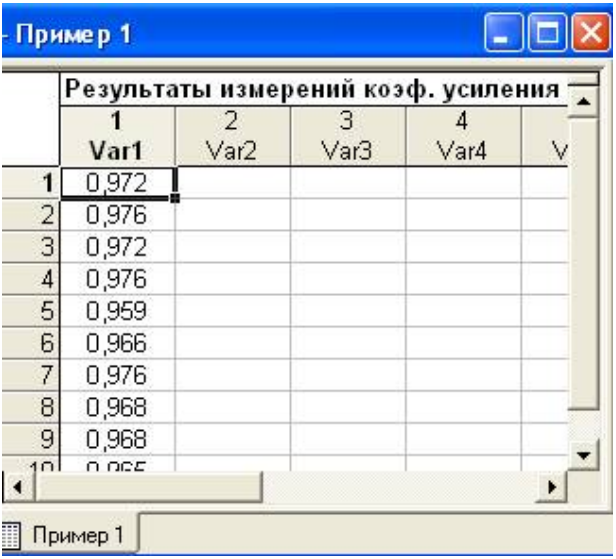
Вибір DATA → vars → Add приводить до появи вікна Add Variables, що дозволяє вказати після якого стовпця скільки додається стовпців, ім'я, тип і формат перемінних у стовпці.



Для того щоб описати перемінну, необхідно двічі клацнути мишею по її імені - наприклад, після клацання по заголовку перемінної 2 (VAR2*)* відкриється вікно Variable2, у якому можна задати її ім'я (чи перейменувати), задати вид і розмір шрифту для цього імені, тип і формат перемінної і т.п.



Для введення заголовка таблиці необхідно двічі клацнути лівою кнопкою миші по порожньому полю вище назв стовпців, і після появи курсору ввести потрібний заголовок. Наприклад, **Результати вимірів коеф. підсилення.**

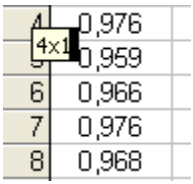


Аналогічно, після подвійного клацання лівою кнопкою миші по номеру рядка вводиться і потрібна для нього назва. Наприклад, для 2-го рядка введемо **2-вимір**

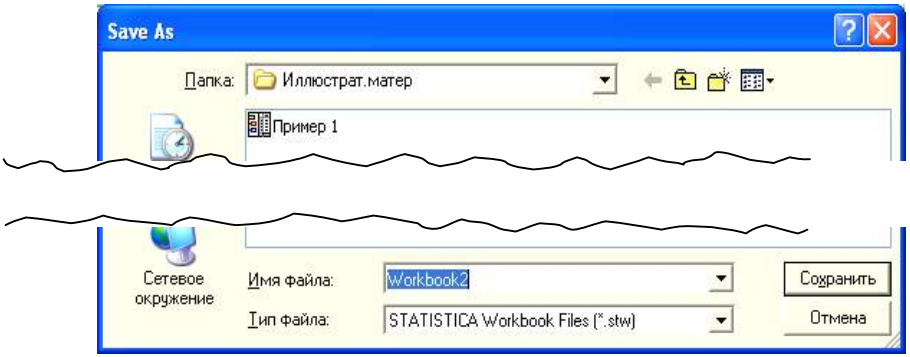
Оскільки пакет STATISTICA є звичайним Windows-додатком, можна легко і швидко імпортувати дані, отримані в системі STATISTICA, в інший Windows-додаток, наприклад у MS Word, Excel.

Найкраще зробити це в такий спосіб: натисніть одночасно кнопки ALT і F3. На екрані замість курсору миші з'явиться значок «приціл». Використовуючи мишу, помістіть приціл у верхній лівий кут таблиці. Потім натисніть ліву кнопку миші, зафіксуйте приціл і, утримуючи кнопку миші, перемістіть приціл у потрібне для копіювання місце таблиці. Виділена

частина таблиці буде відзначена прямокутною заштрихованою рамкою. Після того, як ви відпустите кнопку миші, відзначена частина таблиці буде поміщена в буфер обміну. Якщо тепер відкрити потрібний документ якого-небудь Office-додатка, то обраний сегмент таблиці буде знаходитися в буфері обміну і може бути скопійований. Результат копіювання частини таблиці в MS Word показаний нижче.



Після закінчення введення даних і закриття таблиці при відповіді «так», на запит про необхідність збереження файлу з даними з'являється вікно Save As , за допомогою якого задається ім'я і шлях до файлу. Файл зберігається у форматі *.stw.



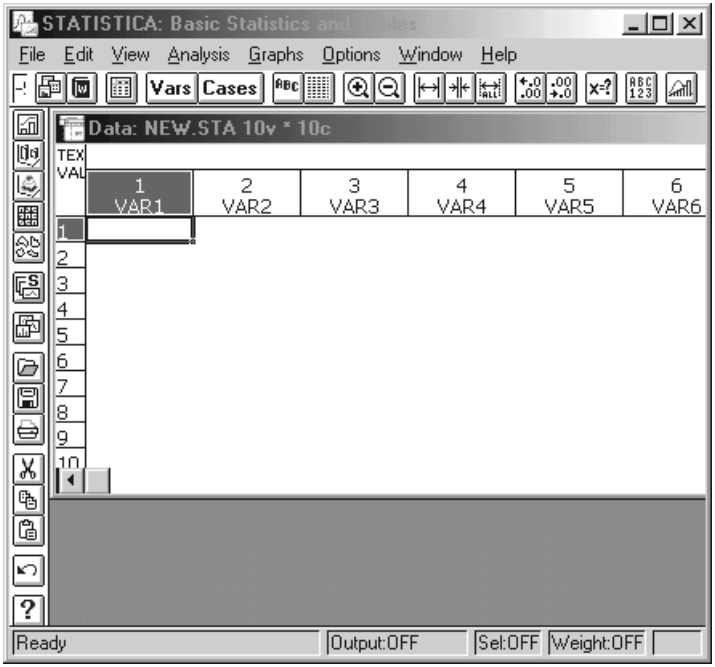
У більш ранніх версіях пакета STATISTICA (5.0 і нижче) уведення вихідних даних можна здійснювати, використовуючи модуль STATISTICA:Basic Statistics and Tables. Для того щоб створити таблицю з даними, необхідно:

- Запустити програму.
- Відкриється меню *Статистичних модулів* (STATISTICA Module Switcher).
- Виберіть з меню модуль *Основні статистики і таблиці* (Basic Statistics and Tables) і клацніть по ньому мишею.

- Тепер ви знаходитесь в модулі *Основні статистики і таблиці,* у якому можете вибрати будь-яку статистичну процедуру, що входить у цей модуль. Але оскільки у вас інша мета, просто клацніть мишею по кнопці Вихід (Cancel).

В основному робочому вікні системи підведіть курсор миші до рядка меню Файл і клацніть лівою кнопкою. В меню виберіть команду *Створити дані.* На екрані комп'ютера відразу ж з'являється вікно *Створення даних.* У цьому вікні можна ввести ім'я файлу, для чого помістіть курсор миші в поле *Filename - Ім'я файлу* і наберіть із клавіатури потрібне ім'я.

Після натискання клавіші *Enter* чи на клавіатурі кнопки *Save* програма створить порожню таблицю, що містить 10 рядків і 10 стовпців (таку ж, як і у вікні Workbooks – Spreadsheet).

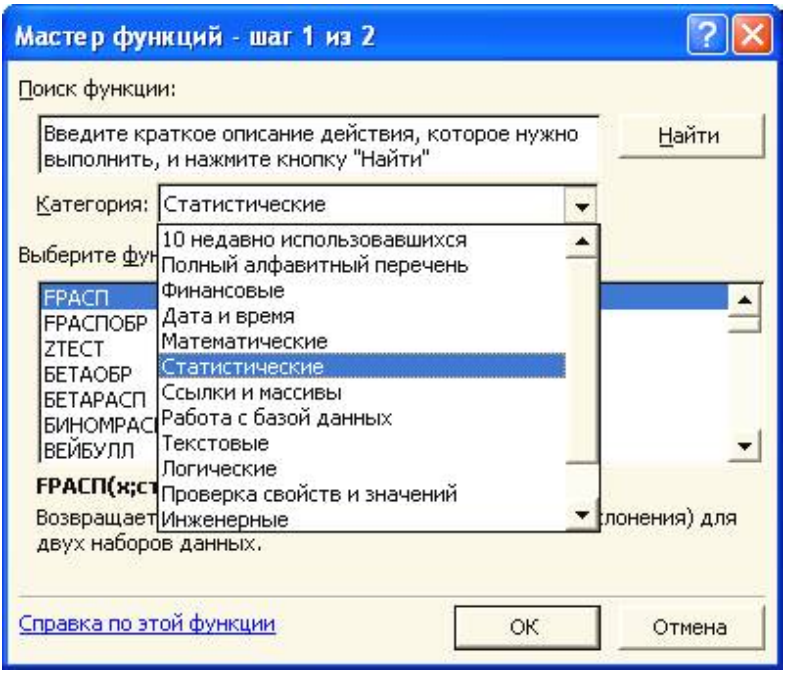


Вікно модуля STATISTICA: Basic Statistics and Tables.

Для введення заголовка таблиці, зміни кількості рядків/стовпців, їхнього переміщення, копіювання, видалення і перейменування використовуються аналогічні прийоми, що описані вище.

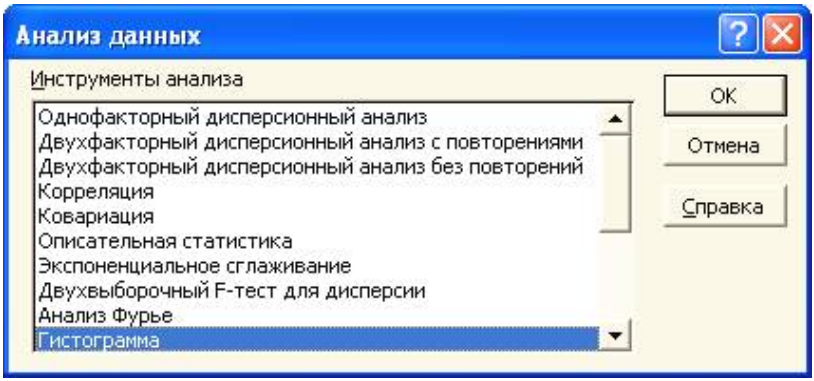
2.5 Комп'ютерні технології статистичної обробки

ВИКОРИСТАННЯ ВМОНТОВАНИХ СТАТИСТИЧНИХ ФУНКЦІЙ І ПАКЕТА АНАЛІЗУ **Microsoft Excel.**. В програмі Microsoft Excel представлене велике число вмонтованих статистичних функцій. З їхньою допомогою можна розрахувати всі статистичні характеристики, що звичайно використовуються на практиці . Для виклику потрібної *вмонтованої статистичної функції* використовується команда Вставка → Функція … чи кнопка Вставка функції, і у вікні, що з'явилося, Майстер функцій вибрати категорію «статистичні» і вказати потрібну

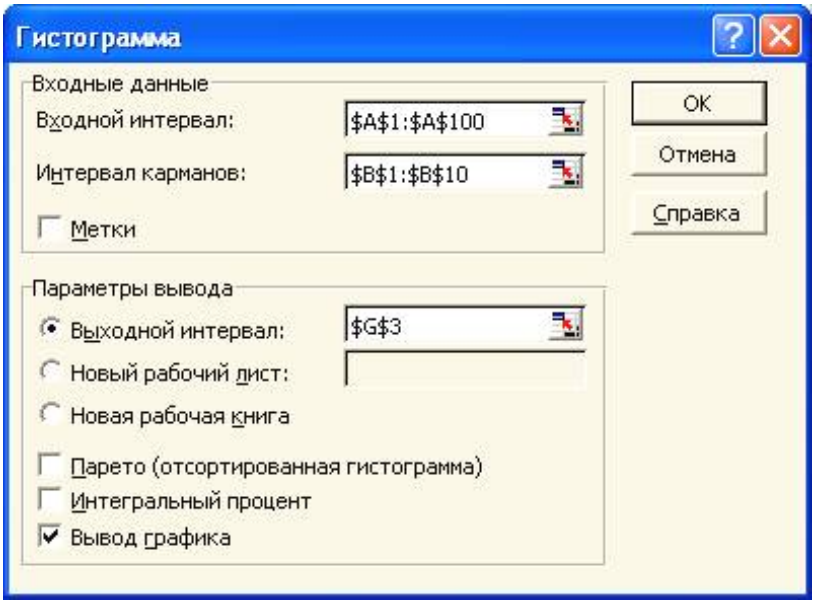


функцію, в підвікні виберіть функцію. Після вибору і клацання по кнопці ОК з'являється вікно Аргументи функції, за допомогою якого здійснюється введення адрес комірок таблиці

зі значеннями оброблюваного масиву і, при необхідності, інших необхідних величин (наприклад, ступенів вільності).



Для побудови *гістограми* використовується команда Сервіс → Аналіз даних, після чого у вікні, що з'явилося, Аналіз даних - вибирається інструмент аналізу – Гістограма.



У вікні Гістограма задається вхідний інтервал, де вказуються адреси комірок з вихідними даними, і інтервал

кишень, де вказуються адреси комірок зі значеннями границь інтервалів. Потім задаються бажані параметри виводу, клацнути кнопку ОК.

ПРИКЛАД 2.5.1.. Розрахувати статистичні характеристики і побудувати гістограму по вихідним даним, наведеним у таблиці 2.2.1.

Рішення за допомогою **Microsoft Excel.**. Для розрахунку основних статистичних характеристик: середнього арифметичного $\overline{x}$, медіани, моди, емпіричної дисперсії s^2 і розмаху, введемо в комірки таблиці вихідні дані і використаємо вмонтовані статистичні функції. Результати розрахунку зведені в таблицю 2.5.1.

Таблиця 2.5.1

срзнач	медіана	мода	дисп	мін	макс
916,4847	923,34109	944,78	9261,16	666,518	1114,7

Для побудови гістограми використаємо команду Сервіс → Аналіз → Гістограма. У вікні Гістограма задаємо вхідний інтервал, де вказуємо адреси комірок з вихідними даними, і інтервал кишень, де вказуємо адреси комірок зі значеннями границь інтервалів. Виводяться таблиця частот і Гістограма (рис. 2.5.1).

Кишеня	*Частота*
650	0
700	3
750	2
800	6
850	12
900	20
950	21
1000	17
1050	13
1100	4
1150	2
Ще	0

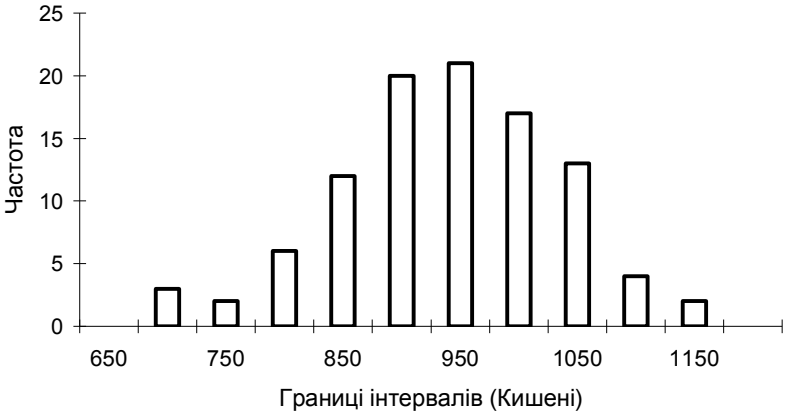


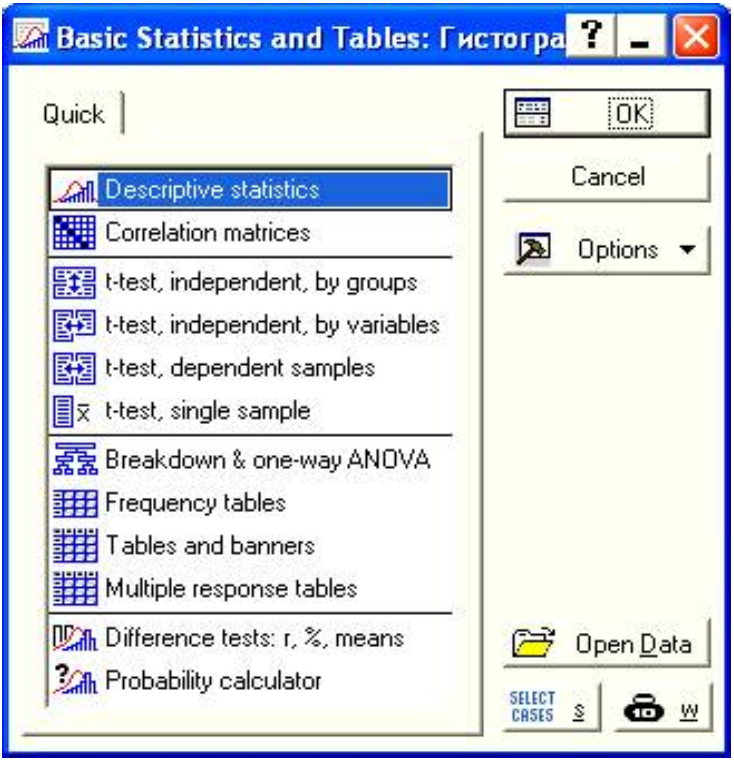
Рисунок 2.5.1. Гістограма, побудована за даними табл. 2.2.1.

***РОЗРАХУНОК СТАТИСТИЧНИХ ХАРАКТЕРИСТИК І ПОБУДОВА ГІСТОГРАМИ В ПАКЕТІ* STATISTICA.** У версії пакета 6.0 розрахунок статистичних характеристик і побудова гістограми визначаються процедурою, що складається з наступних кроків.

Крок 1. Ввести чи імпортувати (наприклад, з Excel) вихідні в робочу книгу (Workbook) системи STATISTICA.

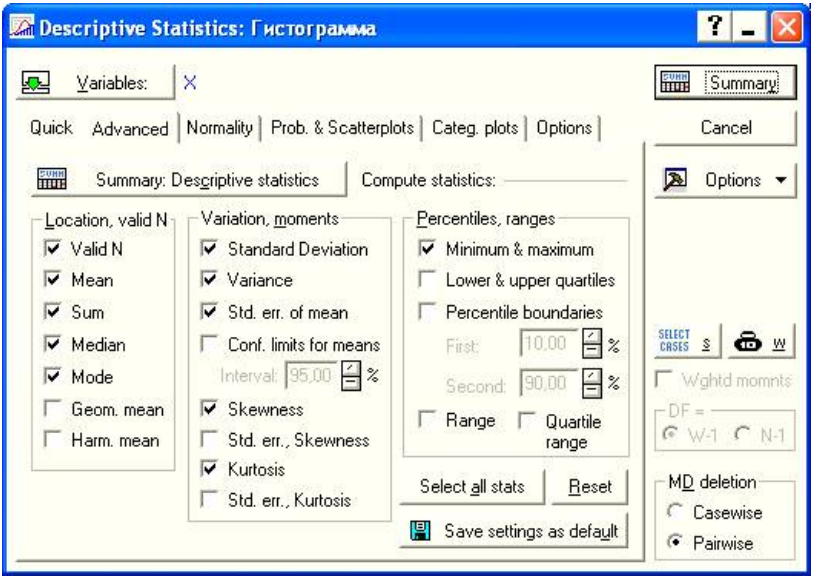
Крок 2. Виділити стовпець із уведеними чи імпортованими даними, для якого необхідно розрахувати статистичні характеристики і побудувати гістограму. При необхідності можна змінити назву стовпця.

Крок 3. Клацнути по кнопці Start menu ..., розташованої в лівому нижньому куті вікна додатка й у меню, що з'явилося, вибрати Statistics →Basic Statistics and Tables → Descriptive Statistics

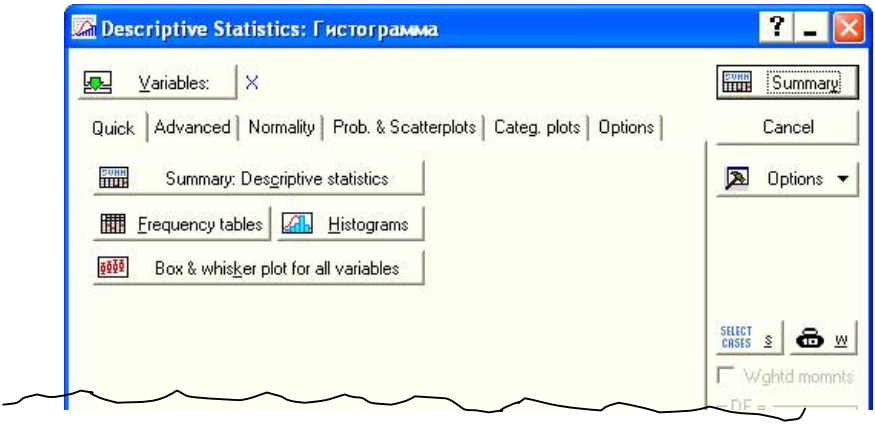


Крок 4. Необхідні статистичні характеристики (описові статистики) задаються у вікні Descriptive Statistics на вкладці Advanced.

Крок 5. Для побудови таблиці з абсолютними і відносними частотами попадання даних в автоматично обраний інтервал і гістограми використовуються кнопки Frequency table, Histograms на вкладку Quick вікна Descriptive Statistics.



Вкладка Advanced вікна Descriptive Statistics.



Вкладка Quick вікна Descriptive Statistics

ПРИКЛАД 2.5.2. Для прикладу виконаємо початкову статистичну обробку для вихідних даних з таблиці 2.2.1. Вихідні дані імпортовані з таблиці Excel попереднього прикладу. Статистичні характеристики, таблиця частот і Гістограма, отримані за технологією початкової статистичної обробки

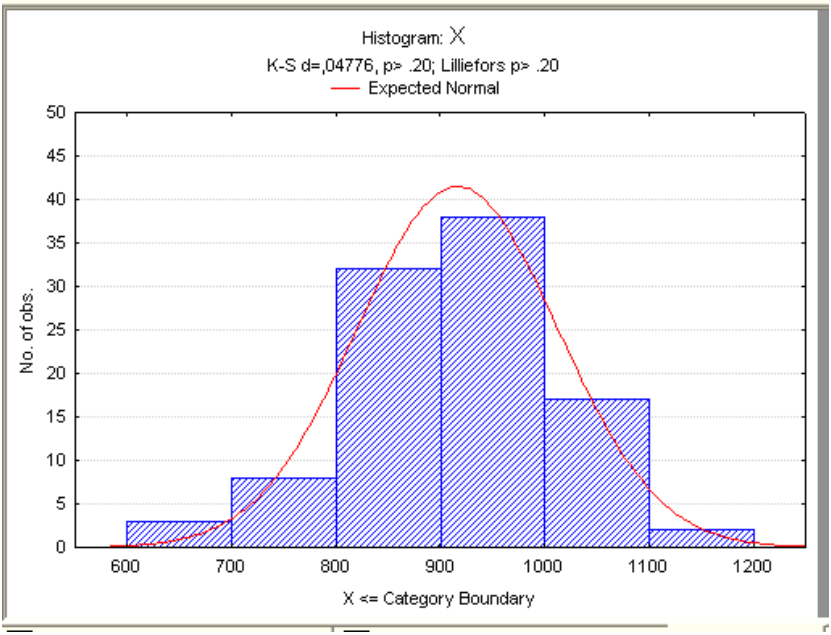
пакета STATISTICA, приведено нижче.

Descriptive Statistics (Гистограмма)

Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Sum	Minimum	Maximum	Variance	Std.Dev.
X	100	916,4847	923,3411	Multiple		91648,47	666,5183	1114,730	9261,165	96,23495

Frequency table: X (Гистограмма)
K-S d=,04776, p> .20; Lilliefors p> .20

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
600,0000<x<=700,0000	3	3	3,00000	3,0000	3,00000	3,0000
700,0000<x<=800,0000	8	11	8,00000	11,0000	8,00000	11,0000
800,0000<x<=900,0000	32	43	32,00000	43,0000	32,00000	43,0000
900,0000<x<=1000,000	38	81	38,00000	81,0000	38,00000	81,0000
1000,000<x<=1100,000	17	98	17,00000	98,0000	17,00000	98,0000
1100,000<x<=1200,000	2	100	2,00000	100,0000	2,00000	100,0000
Missing	0	100	0,00000		0,00000	100,0000



2.6 Контрольні питання і завдання

2.6.1.Питання

1. Поняття про генеральну сукупність і вибірку, репрезентативна вибірка.
2. Форми представлення вибірки, формула Старджеса.
3. Випадкова величина, дискретні і неперервні випадкові величини.
4. Вибіркові характеристики випадкової величини.
5. Теоретичні характеристики випадкової величини (математичне чекання; дисперсія; коефіцієнт асиметрії; коефіцієнт ексцесу).
6. Побудова гістограми і полігона частот.
7. Квантиль, процентиль, квартилі
8. Функція розподілу випадкової величини і її властивості.
9. Нормальний розподіл і його характеристики.
10. Розподіл Вейбулла, його математичне чекання і дисперсія.
11. Біноміальний закон і його характеристики.
12. Закон Пуассона і його характеристики.
13. Уведення даних у пакеті STATISTICA.
14. Додавання, переміщення, копіювання, видалення рядків у пакеті STATISTICA.
15. Додавання, переміщення, копіювання, видалення і перейменування стовпців у пакеті STATISTICA.
16. Копіювання в буфер пам'яті сегмента таблиці у пакеті STATISTICA..

2.6.2. Завдання

При вимірюванні на 100 тестових кристалах ємності МОН-структури (структури метал-окисел-напівпровідник) у режимі збагачення отримані наступні значення (у пФ) наведені у . таблиці 2.6.1

Таблиця 2.6.1

	1	2	3	4	5	6	7	8	9	10
1	704	663	675	672	623	648	645	686	679	684
2	679	672	658	671	694	709	665	693	670	700
3	655	679	710	702	669	662	676	643	684	705
4	700	651	713	682	692	696	663	678	707	652
5	683	672	644	692	648	673	657	643	674	676
6	711	701	662	643	691	691	629	667	693	694
7	692	661	664	657	657	654	663	669	664	647
8	659	684	636	618	634	678	661	620	680	701
9	692	661	640	641	647	701	637	690	674	724
10	677	678	708	687	661	659	708	689	670	704

На підставі моделі, що відбиває вплив на ємність МОН–структури товщини плівки окисла отримано наступні значення математичного очікування **μ** і середньоквадратичного відхилення **s** для генеральних сукупностей, що характеризуються розподілами ємності МОН–структури по тестових кристалах

Таблиця 2.5.2

№	1	2	3	4	5	6	7	8	9	10
μ	200	250	300	350	400	450	500	550	600	650
s	21	22	23	24	25	26	27	28	29	30
№	11	12	13	14	15	16	17	18	19	20
μ	700	720	740	760	780	785	790	795	800	805
s	30	32	33	34	35	36	37	38	39	40

Необхідно:

- розрахувати середнє арифметичне і середньоквадратичне відхилення для вибірки таблиці 2.5.1;
- утворити масив з 100 нормованих випадкових величин, отриманих на основі вибірки таблиці 2.5.1, прийнявши параметр зрушення **a** рівним середньо арифметичному вибірки і параметр масштабу **b** рівним середньоквадратичному відхиленню;

- провести початкову статистичну обробку (розрахувати статистичні характеристики і побудувати Гістограму) для вибірки n=100, отриманої перетворенням нормованих значень до реальних, відповідній генеральній сукупності з визначеним варіантом набору значень **μ** і **s** (варіант відповідає номеру №).

При проведенні початкової статистичної обробки використовувати:

- вмонтовані статистичні функції і Пакет аналізу MICROSOFT EXCEL.
- Пакет аналізу **STATISTICA.**

3 ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ

3.1 Основні положення

При проведенні статистичного дослідження виникає необхідність у формулюванні й експериментальній перевірці деяких *можливих тверджень (гіпотез)*, наприклад:

- чи можна вважати, що конструктивна доробка чи технологічне удосконалення дійсно поліпшило якість виробів;
- чи підкоряється отримана вибірка нормальному чи будь-якому іншому закону розподілу;
- чи можна вважати дві чи більш узятих вибірок приналежних одній генеральній сукупності;
- чи є технологічний процес налагодженим чи розладженим;
- яка величина невідомих параметрів, що визначають стан стохастичної системи і т.п.

Для перевірки зроблених припущень висуваються статистичні гіпотези. Відмінність статистичної гіпотези від просто гіпотези в тім, що *статистична гіпотеза* висувається щодо виду чи розподілу параметрів відомого розподілу випадкової величини.

Процедура обґрунтованого зіставлення висловленої гіпотези з наявною вибіркою $x_1, x_2, ..., x_n$ здійснюється за допомогою того чи іншого статистичного критерію і називається *статистичною перевіркою гіпотез*.

Результат такого зіставлення може бути як негативним (дані спостереження суперечать висунутій гіпотезі, отже, від неї треба відмовитися), так і позитивним (спостереження не суперечать висловленій гіпотезі, і тому її можна прийняти в якості одного з рішень). Прийнята гіпотеза буде розглядатися як досить правдоподібне, не суперечне досвіду, твердження.

Статистичні критерії перевірки гіпотез різноманітні, але є єдина послідовність дій побудови критерію, що укладається в 5 кроків [1].

1-й крок. Висувається *основна* (або що перевіряється) *гіпотеза* H_0. Гіпотеза H_1, що суперечить основній H_0, називається *альтернативною*, або конкуруючою.

2-й крок. Задається *рівень значимості критерію* α. Будь-яке статистичне рішення, прийняте на основі обмеженого ряду спостережень, супроводжується, хоч і малою, імовірністю помилкового висновку. Саме в долі випадків α гіпотеза H_0 може бути відкинута за умови, що вона вірна, чи, навпаки, у долі випадків β ми можемо прийняти гіпотезу H_0, у той час як вона помилкова. При фіксованому обсязі вибірки n величину імовірності α чи β ми можемо вибирати самостійно. Якщо є можливість як завгодно збільшувати n, то теоретично можна домогтися яких завгодно малих помилок α і β при будь-якій фіксованій конкуруючій гіпотезі H_1.

Величину α називають рівнем значимості, величиною критерію або помилкою першого роду. Це імовірність відкинути основну гіпотезу H_0 за умови, що вона вірна.

Чим вагоміші для дослідника втрати від помилкового відкидання гіпотези H_0, тим менше α необхідно вибирати. Звичайно користуються стандартними значеннями α (0.1; 0.05; 0.025; 0.01; 0.005; 0.001).

3-й крок. Задається деяка функція результатів спостереження – *критична статистика*

$$\Psi_{кр.} = \Psi(x_1, x_2, ..., x_n).$$

Як функція результатів спостережень ця критична статистика також є випадковою величиною й у припущенні справедливості H_0 підпорядкована деякому добре вивченому закону розподілу з відомою функцією густини імовірності $W(\Psi_{кр})$.

4-й крок. Зі статистичних таблиць розподілу $W(\Psi_{кр})$ знаходяться квантилі рівня $\alpha/2$ і $1-\alpha/2$ або процентні точки $\Psi_{(1-\alpha/2)\cdot 100\%}$ і $\Psi_{(\alpha/2)\cdot 100\%}$, що є відповідно нижньою $\Psi_{кр.н}$ і верхньою $\Psi_{кр.в}$ *критичними точками (границями)*. Вони поділяють всю область припустимих значень $\Psi_{кр}$ на області:

- неправдоподібно малих (I);
- правдоподібних (II);
- неправдоподібно великих (III).

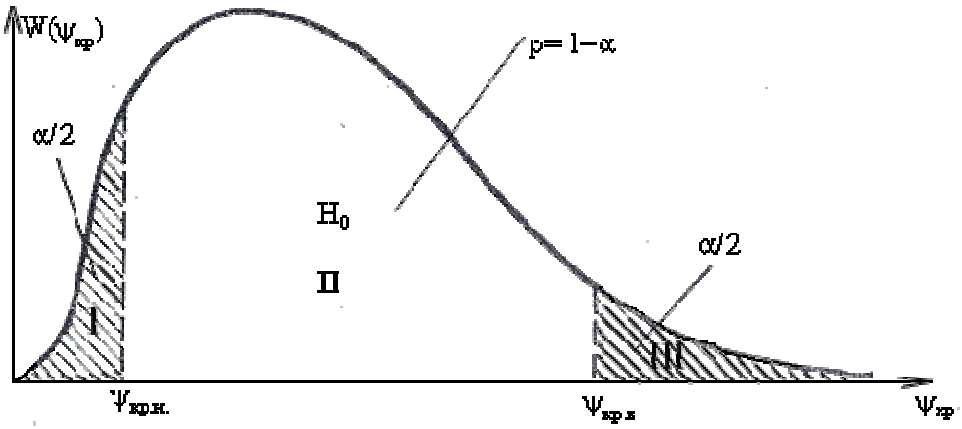


Рис.3.1.1 Області прийняття і відкидання гіпотези H_0.

Область прийняття гіпотези H_0 визначається як довірчий інтервал для $\Psi_{кр}$, що формується на основі розподілу статистики $W(\Psi_{кр})$ при рівні довірчої імовірності $p=1-\alpha$.

Розрізняють *односторонні* і *двосторонні критерії*. Для одностороннього критерію область прийняття основної гіпотези може мати обмеження тільки з однієї сторони (зверху або знизу). При цьому область значень статистики $\Psi_{кр}$ розбивається на дві: область правдоподібних і область неправдоподібно великих чи неправдоподібно малих значень.

Для двостороннього критерію область прийняття гіпотези H_0 має два обмеження – зверху і знизу.

5-й крок. Визначається спостережене (розрахункове) значення критичної статистики $\Psi_{розр.}$ підстановкою в $\Psi_{кр}$ конкретних вибіркових значень $x_1, x_2, ..., x_n$ чи деяких функцій від них. Якщо виявиться, що $\Psi_{розр}$ належить області правдоподібних значень, то гіпотеза H_0 вірна, тобто не суперечить вибірковим даним. У противному випадку H_0 відкидається з *помилкою першого роду* α. Відкидання H_0 означає, що $\Psi_{розр.}$ не підкоряється закону розподілу $W(\Psi_{кр})$. Помилка β може виникнути тоді, коли приймається H_0, у той час коли вона невірна. β називається *помилкою другого роду*, а **(1-β)** *– потужністю критерію*.

Можна сказати, що статистичним критерієм називають випадкову величину з визначеною функцією густини імовірності, що служить для перевірки гіпотез. Якщо гіпотеза, що перевіряється, H_0 зводиться до перевірки точної рівності, то гіпотеза називається простою, в інших випадках ми маємо справу зі складною або вкладеною гіпотезою.

Найбільше часто при статистичному аналізі перевіряються гіпотези з використанням наступних критеріїв:

- χ^2*-критерій* Пірсона, λ-критерій Колмогорова і $n\omega^2$-критерій Мізеса-Смірнова для перевірки гіпотези про закон виробничого розподілу;
- t-критерій Стьюдента для перевірки гіпотез про однорідність середніх двох вибірок;
- F-критерій Фішера для перевірки гіпотез про однорідність дисперсій двох вибірок;
- G-критерій Кохрана і критерій Барлетта для перевірки гіпотез про однорідність декількох дисперсій;
- непараметричний критерій Манна-Уітни для перевірки гіпотези про те, що вибірки узяті з однієї сукупності з визначеним довільним законом розподілу.

Розглянемо перераховані статистичні критерії.

3.2 Перевірка гіпотези про закон розподілу

Нехай дана вибірка незалежних випадкових чисел $x_1, x_2, \ldots, x_n$, що мають емпіричний (експериментальний, виробничий) розподіл $F_n(x)$. Для перевірки узгодження розподілу $F_n(x)$ з теоретичним розподілом $F(x)$ використовуються: критерій Пірсона, λ-критерій Колмогорова і $n\omega^2$-критерій Мізеса-Смірнова. Попередня оцінка вигляду розподілу може бути зроблена порівнянням емпіричного розподілу з теоретичними розподілами.

Розглянемо використання критерію Пірсона чи як його інакше називають, критерію χ^2 (хі-квадрат) для перевірки гіпотези про відповідність експериментального розподілу випадкової величини нормальному закону.

Використання критерію Пірсона для оцінки погодженості досліджуваного розподілу з нормальним

засновано на порівнянні емпіричних $m_i^{'}$ і теоретичних (обчислених у припущенні нормального розподілу) m_i частот.

Спочатку знаходимо середини часткових інтервалів

$$X_i^* = \frac{X_i + X_{i+1}}{2} .$$

Варіанту X_i^* відповідають частоти $m_i^{'}$, тобто число спостережень, що потрапили в *i-й* інтервал. У підсумку одержуємо послідовність рівновіддалених варіант і відповідних їм частот у вигляді табл. 3.2.1.

Таблиця 3.2.1

X_1^*	X_2^*	…	X_k^*
$m_1^{'}$	$m_2^{'}$	…	$m_k^{'}$

Далі обчислюємо середнє значення $\bar{x}$ й оцінку середньоквадратичного відхилення s і нормуємо границі інтервалів за формулами

$$Z_i = \frac{X_i - \bar{x}}{s} , \qquad Z_{i+1} = \frac{X_{i-1} - \bar{x}}{s} .$$

Теоретичні імовірності P_i попадання x в інтервал (X_i, X_{i+1}) обчислюємо за допомогою функції Лапласа $\Phi(Z)$ по рівності

$$P_i = \Phi(Z_{i+1}) - \Phi(Z_i) .$$

Шукані теоретичні частоти нормального розподілу

$$m_i = nP_i .$$

Використання критерію Пірсона засновано на порівнянні емпіричних (що, спостерігаються) $m_i^{'}$- і теоретичних (обчислених у припущенні нормального розподілу) m_i частот. Звичайно $m_i^{'}$ і m_i різні.

Для перевірки гіпотези про відповідність експериментального закону розподілу випадкової величини нормальному при рівні значимості q потрібно перевірити нульову гіпотезу: генеральна сукупність, з якого узята вибірка $x_1, x_2, \ldots, x_n$ розподілена нормально.

Як критерій проверни нульової гіпотези приймається випадкова величина

$$\chi^2 = \frac{(m_1' - m_1)^2}{m_1} + \frac{(m_2' - m_2)^2}{m_2} + \ldots + \frac{(m_k' - m_k)^2}{m_k},$$

чи

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i' - m_i)^2}{m_i},$$

де K - число інтервалів.

Ця величина випадкова, тому що в різних дослідах вона приймає різні, заздалегідь невідомі, значення. Чим менше розрізняються емпіричні і теоретичні частоти, тим менше значення критерію, і, отже, він характеризує близькість емпіричного і теоретичного розподілів. Зведенням у квадрат різниці частот усувається можливість взаємного погашення позитивної і негативної різниці.

Якщо розрахункове значення критерію виявилося менше критичного $\chi^2_{кр}$, котре знаходять по таблиці, наведеній у додатку , для відповідного рівня значимості α і числа ступенів вільності κ

, тобто якщо

$$\chi^2 < \chi^2_{кр} = \chi^2_{\alpha,\kappa},$$

то нема підстави відкинути нульову гіпотезу про нормальність розподілу. У противному випадку (при $\chi^2 > \chi^2_{кр}$) нульова гіпотеза відкидається.

Перевірка за допомогою критерію Колмогорова здійснюється в такий спосіб. Вибіркова послідовність $\{x_i\}$ упорядковується в послідовність $\{x_k\}$, що не убуває. Обчислюється абсолютне значення максимальної розбіжності D_n між виробничими і теоретичними розподілами:

$$D_n = \max_{1 \le k \le n} \left| F_n(x_k) - F(x_k) \right|.$$

За отриманим значенням D_n розраховується величина $\lambda=n^{1/2}\ D_n$. Критичні значення для λ приведені в таблиці , додатка .

Для перевірки узгодження за допомогою критерію Мізеса-Смірнова ($n\omega^2$-критерій), розраховується величина:

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{k=1}^{n}\left[F(x_k) - \frac{2k-1}{2n}\right]^2 .$$

Висновок про істотну розбіжність між виробничими і теоретичними розподілами робиться з довірчою імовірністю 95%, якщо $n\omega^2 \geq 0{,}4614$. Критичні значення для $n\omega^2$-критерію при інших значеннях довірчої імовірності $1-\alpha$ наведені в табл. 3.2.1.

Таблиця 3.2.1 -$n\omega^2$-критерій Мізеса-Смірнова

$1-\alpha$	0.5	0.4	0.3	0.2	0.1
$n\omega^2_{кр}$	0.1184	0.1467	0.1843	0.2412	0.3473
$1-\alpha$	0.05	0.03	0.02	0.01	0.001
$n\omega^2_{кр}$	0.461	0.5489	0.6198	0.7435	1.1679

Обсяг вибірки для перевірки гіпотез про закон розподілу за допомогою χ^2-*критерію* Пірсона і λ-критерію Колмогорова повинний бути не менше 100. При вибірках менших обсягів більш чуттєвий $n\omega^2$-критерій. Для більш певного висновку про закон розподілу досліджуваної величини доцільно скористатися всіма критеріями.

3.3 Перевірка гіпотез про рівність дисперсій і середніх

Перевірка гіпотез про рівність дисперсій при відомих математичних очікуваннях. Нехай є дві вибірки випадкових величин ***A*** і ***B*** обсягом n_A і n_B з нормально-розподілених

генеральних сукупностей з функціями розподілу $N(x, \mu_A, \sigma_A)$, $N(x, \mu_B, \sigma_B)$. Необхідно перевірити гіпотезу про рівність дисперсій випадкових величин за умови, що математичні очікування μ_A і μ_B відомі. Процедура перевірки складається з наступних кроків.

1-й крок. Формування основної H_0 і альтернативної H_1 гіпотез

$$H_0: S^2_A = S^2_B,$$
$$H_1: S^2_A \neq S^2_B.$$

2-й крок. Задання рівня значимості α.

3-й крок. Формування критичної статистики. Як міру розрізнення S_A і S_B обрана величина

$$\Psi_{кр} = S^2_A / S^2_B.$$

Граничний розподіл статистики $\Psi_{кр}$ як випадкової величини наближається до розподілу Фішера $F(\Psi_{кр}; k_1, k_2)$, де k_1 і k_2 –числа ступенів вільності, рівні $k_1 = n_a - 1$ $k2 = n_b - 1$.

4-й крок. Визначення критичних границь. Критична границя відповідна ступеням волі k_1 і k_2 визначається як процентна крапка розподілу Фішера рівня α і знаходиться з таблиці розподілу Фішера (додаток). Позначимо цю критичну границю $F_{кр}(\alpha, k_1, k_2)$

5-й крок. Розрахунок розрахункового значення критичної статистики (критерій Фішера) з виразу

$$F_{розр.} = S^2_A / S^2_B,$$

де S_A^2, S_B^2 - вибіркові дисперсії, розраховані для вибірок **А** і **В.** Причому, у чисельник поміщаємо велику дисперсію, що дає можливість розглядати як критичну границю тільки верхню границю.

6-й крок. Формулювання висновку.

Якщо $F_{розр} \leq F_{кр}(\alpha, k_1, k_2)$, то робиться висновок, що дисперсії однорідні, тобто приймається гіпотеза H_0 про рівність дисперсій.

Якщо $F_{розр} > F_{кр}(\alpha, k_1, k_2)$, то робиться висновок про наявність істотного розходження між дисперсіями S^2_A і S^2_B, тобто приймається гіпотеза H_1.

Перевірка гіпотез про рівність середніх випадкових величин. Нехай дані дві вибірки з нормальних генеральних сукупностей випадкових величин. Розглянемо два випадки:

а) дисперсії S_A^2 і S_B^2 - однорідні,

б) дисперсії S_A^2 і S_B^2 - неоднорідні.

Необхідно перевірити гіпотезу про рівність математичних очікувань.

1-й крок. Формування основної та альтернативної гіпотез (для обох випадків гіпотези однакові)

$$H_0: X_A = X_B,$$
$$H_1: X_A \neq X_B.$$

Така задача ставиться звичайно тоді, коли вибіркові середні виявляються різними. Гіпотеза H_0 затверджує, що це розходження несуттєве, гіпотеза H_1, - розходження істотне, значимо.

2-й крок. Завдання рівня значимості α.

3-й крок. Перевірка гіпотези про рівність середніх. Однорідності проводиться за критерієм Стьюдента (*t*-критерій). Якщо дисперсії однорідні (випадок а), то обчислюються статистика *t* і число ступенів вільності κ за формулами:

$$t = \frac{\left|\overline{x_A} - \overline{x_B}\right|}{S} \cdot \frac{1}{\sqrt{1/n_A + 1/n_B}}$$

$$S = \sqrt{\frac{(n_Б - 1)s_Б^2 + (n_B - 1)s_B^2}{\kappa}} \qquad \kappa = n_A + n_B - 2.$$

Якщо дисперсії істотно розрізняються (випадок б), то для обчислення статистики *t* і числа ступенів вільності κ використовуються формули:

$$t = \frac{|x_A - x_B|}{S}$$

$$S = \sqrt{S_A^2 / n_A + S_B^2 / n_B}, \qquad \kappa = \frac{S_A^2 / n_A + S_B^2 / n_B}{(S_A^2 / n_A)^2 / (n_A + 1) + (S_B^2 / n_B)^2 / (n_B + 1)}$$

4-й крок. Формулювання висновку. Для того, щоб зробити висновок про рівність (однорідність) математичних очікувань, необхідно статистику t порівняти з критичною точкою розподілу Стьюдента (додаток , таблиця). Якщо $t \le t_к$, то робиться висновок про однорідність математичних очікувань, приймається гіпотеза H_0. Якщо $t > t_к$, вважається, що математичні очікування істотно розрізняються H_1.

Порівняння вибірок непараметричними методами. Перевірка приналежності двох вибірок А і В одній сукупності проводиться за критерієм Манна-Уітни (U-критерію) [2].

U-критерій є непараметричним критерієм і може використовуватися для довільних функцій розподілу без пропозицій про їхній вигляд.

Для обчислення статистики U виконуються наступні кроки.

1-й крок. Вибірки А і В упорядковують у порядку зростання. Отриманий номер по порядку для вибіркових значень в об'єднаній вибірці називають ранговим числом.

2-й крок. Кожному рангу приписують до якої вибірки (А чи В) він відноситься.

3-й крок. Обчислюється:

$$U_A = n_A n_B + \frac{n_B(n_B + 1)}{2} - R_A$$

$$U_B = n_A n_B + \frac{n_A(n_A + 1)}{2} - R_B$$

де: n_A, n_B- обсяг вибірок А і В;

R_A,R_B - суми рангів вибірок А і В, відповідно.

4-й крок. Статистика U визначається як найменше зі значень U_A і U_B. Мірою значимості статистики U у вибірках

обсягом більше 8 може служити величина Z, що розраховується за формулою:

$$Z = \frac{U - n_A n_B / 2}{\sqrt{n_A n_B (n_B + n_A + 1)/12}}.$$

5-й крок. Отримане значення Z порівнюється з P%-квантилем нормального розподілу. Якщо $Z \le Z_p$, то з довірчою імовірністю P робиться висновок про однорідність вибірок.

Для визначення Z_p можна використовувати табл..А1 додаеку А

3.4 Комп'ютерні технології перевірки статистичних гіпотез

***ВИКОРИСТАННЯ ВМОНТОВАНИХ СТАТИСТИЧ-НИХ ФУНКЦІЙ І ПАКЕТА АНАЛІЗУ* Microsoft Excel.** Для перевірки статистичних гіпотез можуть бути використані наступні вмонтовані функції:

- ТТЕСТ – для визначення імовірності того, що дві вибірки узяті з генеральних сукупностей з однаковим математичним очікуванням;
- ZТЕСТ – для визначення імовірності того, що вибірка узята з визначеної нормально-розподіленої генеральної сукупності. Можна використовувати цю функцію щоб оцінити імовірність того, що конкретне спостереження узяте з конкретної генеральної сукупності;
- СТЬЮДРАСП – для розрахунку критичних значень розподілу Стьюдента (*t*-розподілу);
- НОРМРАСП - повертає нормальну функцію розподілу для зазначеного середнього і стандартного відхилення.
- ДОВІРИТЬ - повертає довірчий інтервал для середньо визначеної нормально розподіленої генеральної сукупності.
- ФТЕСТ - повертає однобічну імовірність того, що дисперсії вибірок розрізняються.
- ХИ2ТЕСТ - повертає значення для розподілу хі-квадрат (використовується як критичне значення χ^2-*критерій* Пірсона).

У засобах статистичного аналізу, що викликаються командою *Аналіз даних* меню *Сервіс,* для перевірки статистичних гіпотез пропонуються наступні інструменти.

Двовибірковий F-тест для дисперсії. Дозволяє перевірити гіпотезу про рівність дисперсій для двох вибірок.

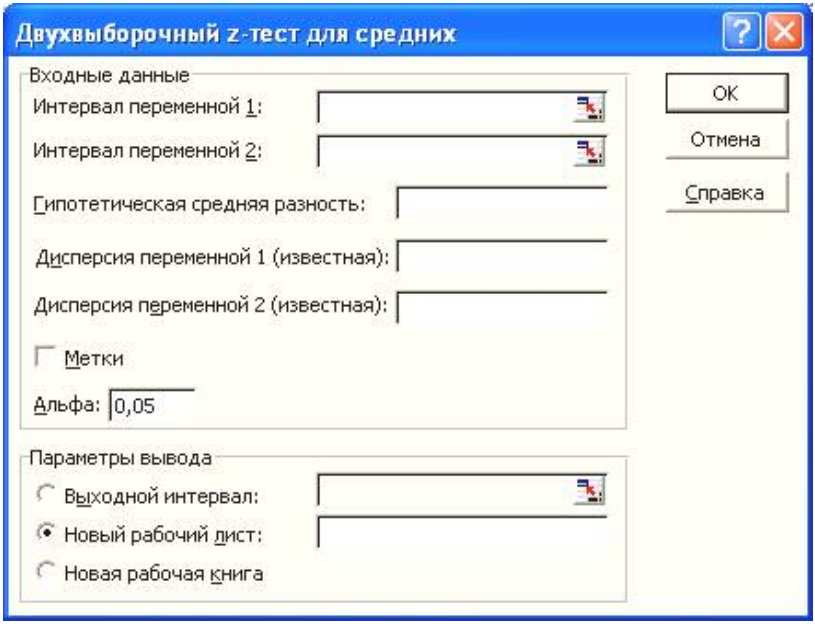
Парний двовибірковий t-тест для середніх. Для перевірки гіпотези про рівність середніх для двох вибірок за допомогою *t*-критерію за умови рівності обсягу вибірок.

Двовибірковий t-тест з однаковими дисперсіями. Двовибірковий *t*-тест Стьюдента служить для перевірки гіпотези про рівність середніх для двох вибірок за умови рівності дисперсій.

Двовибірковий t-тест із різними дисперсіями. Двовибірковий *t*-тест Стьюдента використовується для перевірки гіпотези про рівність середніх для двох вибірок за умови неоднорідності дисперсій.

Вигляд вікон для *t-тестів* аналогічний як і для *F-тест*.

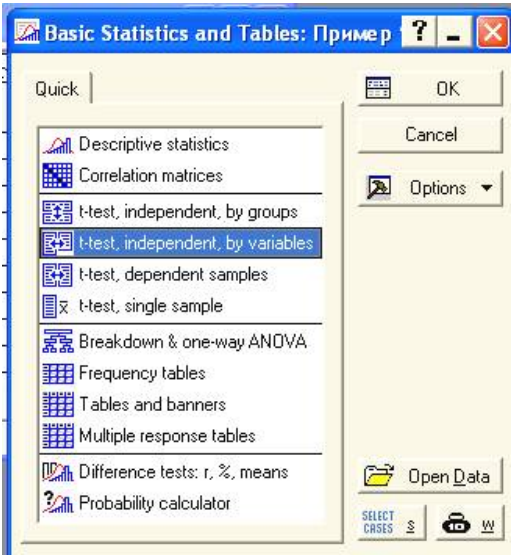
*Z-тест***.** Двовибірковий *z*-тест для середніх з відомими дисперсіями. Використовується для перевірки гіпотези про розходження між середніми двох генеральних сукупностей за умови, що дисперсії відомі, але гіпотези про їхню однорідність не перевірялися.



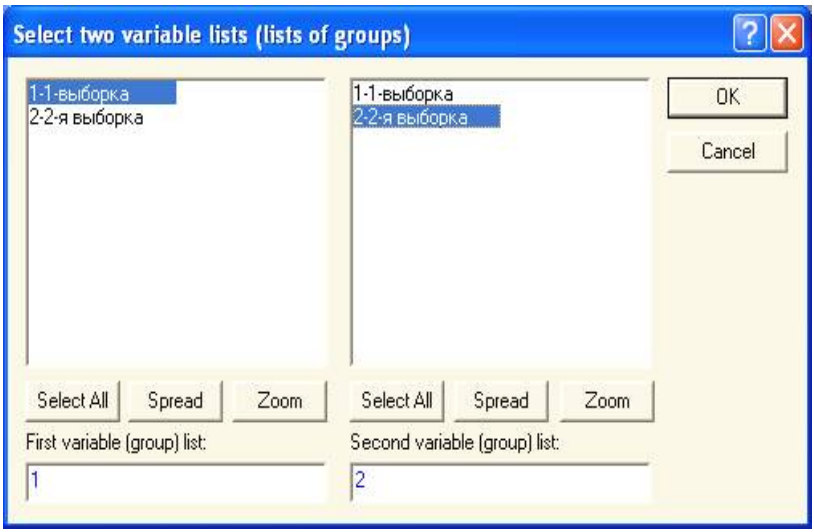
***ПЕРЕВІРКА СТАТИСТИЧНИХ ГІПОТЕЗ У ПАКЕТІ* STATISTICA.** Розглянемо на прикладі перевірки гіпотез про однорідність середніх і дисперсій двох вибірок.

Крок 1. Ввести або імпортувати (наприклад, з Excel) вихідні дані в робочу книгу (Workbook) системи STATISTICA. При перевірці гіпотез у відношенні двох вибірок вводяться два масиви (групи) даних. При необхідності можна ввести назву для таблиці і змінити назву стовпця.

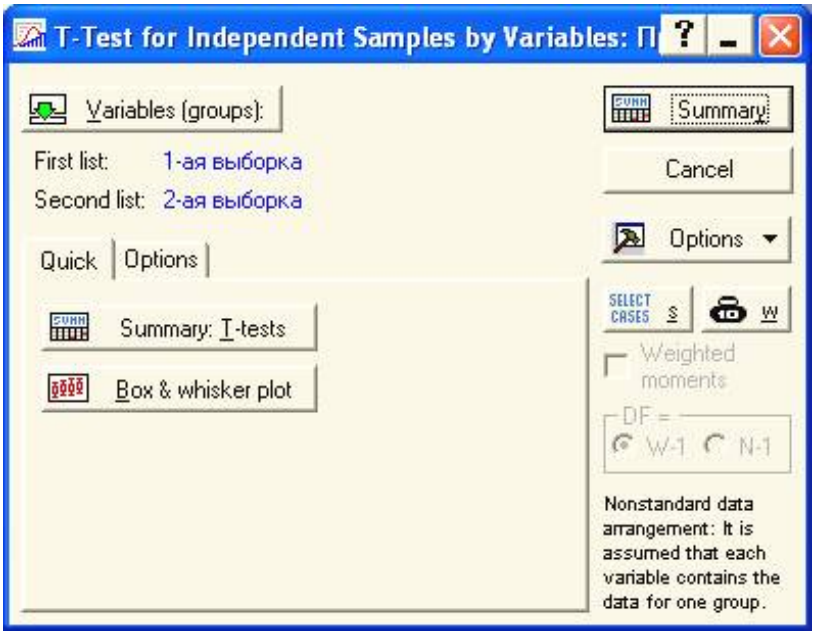
Крок 2. Клацнути по кнопці Start menu ..., розташованої в лівому нижньому куті вікна додатка, й у меню, що з'явилось, вибрати Statistics → Basic Statistics and Tables → t-test, independent, by variables.



Крок 3. У вікні, що з'явилося, T-Test for Independent Samples by Variables після клацання по кнопці Variables (Groups) - указати на стовпці з першою і другою вибіркою.



Крок 4. Для перевірки гіпотез про однорідність дисперсій і середніх вибирається кнопка Summary:T-tests



Крок 5. Після клацання по кнопці Summary:T-tests з'являється вікно зі значеннями середніх, середньоквадратичних, обсягами вибірок і результатами перевірки гіпотез - розрахунковими значеннями *t*-критерію і *F* -критерію.

T-test for Independent Samples (Пример 1 in Пример 1)
Note: Variables were treated as independent samples

Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances

ПРИКЛАД 3.4.1. У таблицях 3.4.1, 3.4.2 наведені дві вибірки з результатами вимірів на тестових кристалах ємності МОН-структур у режимі збагачення (у пФ). Вибірки узяті з двох дослідних партій, виготовлених в різних технологічних режимах.

Таблиця 3.4.1

	1	2	3	4	5	6	7	8	9	10
1	451	460	417	456	460	412	447	464	423	435
2	447	466	460	503	458	443	457	474	476	471
3	445	470	476	448	430	488	518	435	447	490
4	466	454	453	466	444	447	460	461	429	448
5	431	468	417	426	470	484	485	500	430	449
6	420	468	487	435	464	456	448	508	458	422
7	473	430	468	472	440	474	442	491	466	461
8	458	437	440	464	434	442	464	411	459	447
9	449	449	457	438	437	447	443	434	473	486
10	396	468	487	412	476	435	435	461	444	452

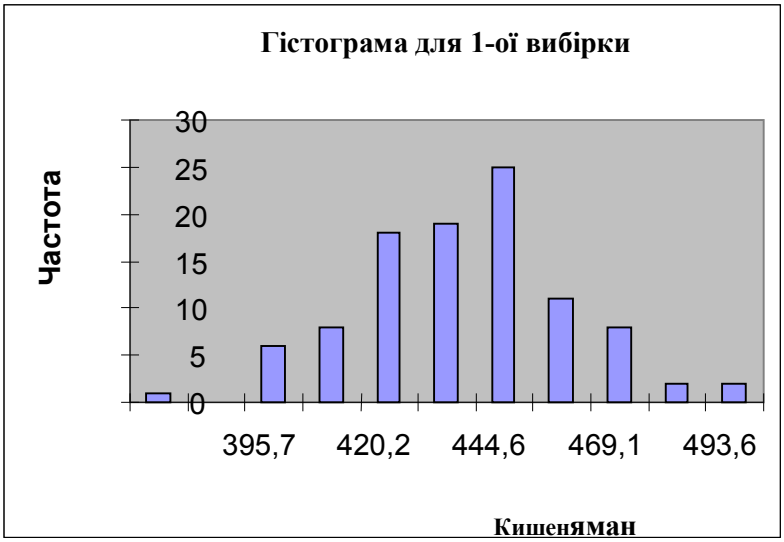
Таблиця 3.4.2

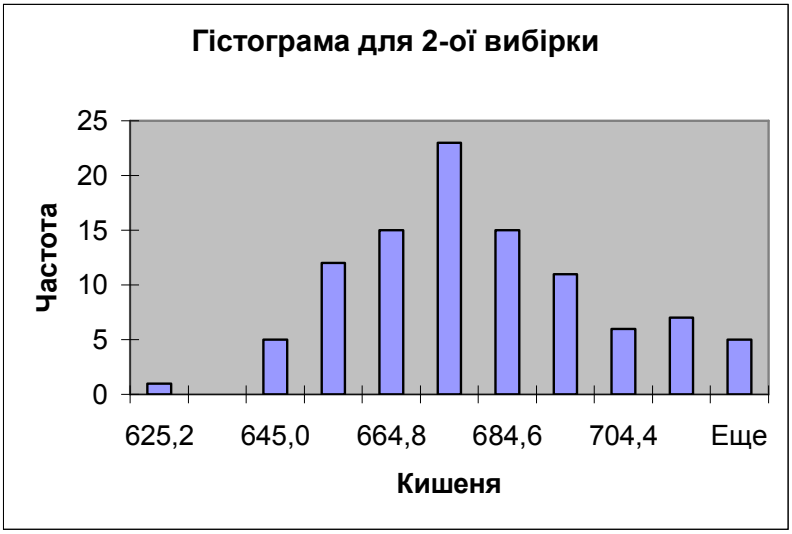
	1	2	3	4	5	6	7	8	9	10
1	670	663	638	640	662	652	657	692	698	677
2	710	699	717	665	653	652	687	677	666	706
3	709	680	694	697	670	674	668	716	646	685
4	666	662	681	674	666	662	667	678	675	644
5	653	660	677	666	716	672	702	625	724	654
6	658	664	668	688	672	665	695	644	672	682
7	672	687	708	654	639	680	648	685	691	661
8	646	658	657	682	701	659	655	669	705	680
9	668	711	647	709	655	688	673	646	720	677
10	687	667	673	662	687	684	678	660	670	683

Ставляться задачі.

- Визначити закон розподілу для випадкової величини - ємність МОН-структури в режимі збагачення. Чи змінюється цей закон при різних технологічних режимах.
- Чи істотне розходження між дисперсіями вибірок.
- Чи істотне розходження між середніми вибірками.

Рішення за допомогою **Microsoft Excel**. Для визначення закону розподілу побудуємо гістограми для 1-ої і 2-ої вибірок за допомогою технології, описаної в 2.4.

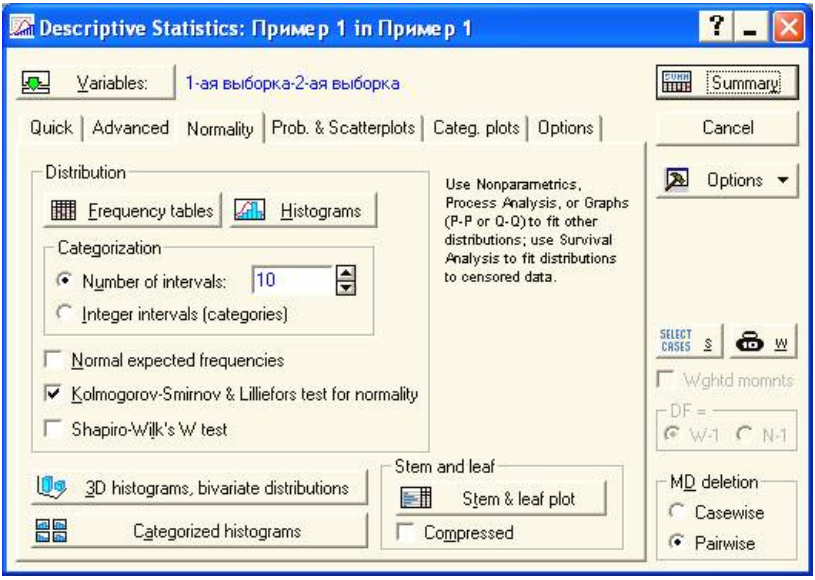




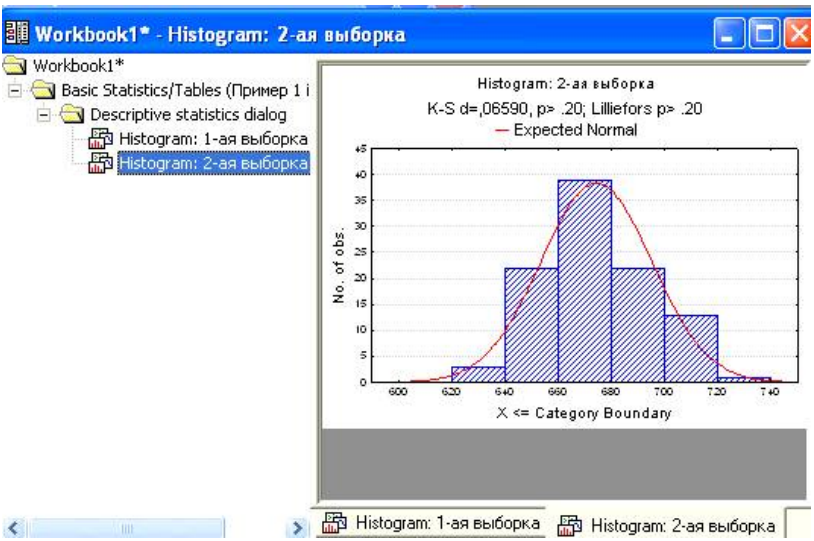
По вигляду гістограм можна припустити нормальний закон розподілу, але додаток Excel не надає засобів для перевірки статистичної гіпотези про нормальний закон.

Здійснити перевірку гіпотези про вигляд закону розподілу можна за допомогою модуля Descriptive Statistics пакета STATISTICA, для чого імпортуємо вихідні дані з таблиць 3.4.1, 3.4.2 у систему STATISTICA.

і викличемо модуль Descriptive Statistics (за технологією, описаною в 2.4), у якому виберемо вкладку Normality.



На вкладці Normality активізуємо прапорець Kolmogorov-Smirnov & Lilliefors test for normality і клацаємо по кнопці Histograms. Виводяться вікна, у яких приведені значення *K-S* d (для 1-ої вибірки *d*=0,04616, для 2-ий вибірки d=0,06590) і за значенням d розраховуємо значення критерію Колмогорова-Смірнова $\lambda=n^{1/2}$ d.



Одержимо для 1-ої вибірки λ=0,4616, що дозволяє прийняти гіпотезу про нормальний закон розподілу з довірчою імовірністю рівної 0,9840, для 2-ий вибірки λ==0,6590 і гіпотеза приймається з довірчою імовірністю 0,7764. (Для оцінки довірчої імовірності використовувалася таблиця , додатки).

Таким чином, випадкова величина - ємність МОН–структури в режимі збагачення підкоряється нормальному закону розподілу з довірчою імовірністю не менш 0,7764 і цей закон не змінюється при зміні технологічного режиму.

Для відповіді на питання чи істотно розходження між дисперсіями вибірок використовуємо *двовибірковий F-тест для дисперсії* додатка Excel.

Двовибірковий F-тест для дисперсії		
	Перемінна 1	*Перемінна 2*
Середнє	454,3039	674,3605
Дисперсія	504,9464	434,4266
Спостереження	100,0000	100,0000
df	99,0000	99,0000
F	1,1623	
P(F<=f) одностор.	0,2278	
F критичне	1,3941	

З приведених вище результатів *двовибіркового F-тест* розраховане значення **F-критерію** 1,1623 менше **F** критичного, рівного 1,3941**,** і гіпотеза про незначуще розходження (однорідності) дисперсій приймається.

Для відповіді на питання чи істотно розходження між середніми вибірок використовуємо *двовибірковий t-тест з однаковими дисперсіями.* додатка Excel.

Двовибірковий t-тест з однаковими дисперсіями		
	Перемінна 1	*Перемінна 2*
Середнє	454,3039	674,3605
Дисперсія	504,9464	434,4266
Спостереження	100	100
Об’єд. дисперсія	469,6865	

Гіпотет.різн.середовищ.	0,0000	
df	198	
t-статистика	-71,7985	
P(T<=t) односто р.	0,0000	
t критич. односто р.	1,6526	
P(T<=t) двостор.	0,0000	
t критич. двостор.	1,9720	

З наведених вище результатів *двовибірковий t-тест з однаковими дисперсіями* розраховане значення t-критерію -71,7985 набагато більше t-критичного, рівного1,6526 для однобічної критичної області і рівного 1,6526 для двосторонньої критичної області**,** і приймається гіпотеза про істотне розходження між середніми з довірчою імовірністю, близькою до 1.

Також питання про наявність істотного розходження між дисперсіями вибірок можна вирішити за допомогою пакета STATISTICA. Використання технології перевірки статистичних гіпотез у пакеті STATISTICA (кроки 1 – 5) привело до наступного результату.

T-test for Independent Samples **Пример 3.4.1.**

Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
454,304	674,360	-71,798	198	0,000	100	100	22,471	20,843	1,1623	0,4556

Значення *t*- і *F*-критеріїв цілком збігаються з отриманими за допомогою Microsoft Excel.

3.5 Контрольні питання і завдання

3.5.1 Питання

1. Поняття про статистичну гіпотезу і процедуру її перевірки.
2. Статистичний критерій, послідовність дій побудови критерію.
3. Помилки першого і другого роду при прийнятті статистичної гіпотези.

4. Області прийняття і відкидання нульової гіпотези H_0.
5. Однобічні і двосторонні критерії прийняття статистичної гіпотези.
6. Перевірка гіпотези про закон розподілу за допомогою
   - 6.1. χ^2 - критерію Пірсона,
   - 6.2. λ-критерію Колмогорова,
   - 6.3. $n\omega^2$-критерію Мізеса-Смірнова.
7. Перевірка гіпотез про рівність дисперсій при відомих математичних очікуваннях.
8. Перевірка гіпотез про рівність математичних очікувань випадкових величин.
9. Порівняння вибірок непараметричними методами (за критерієм Манна-Уітни).

3.5.2. Завдання

Завдання 1**.** При вимірі коефіцієнта підсилення тестових інтегральних транзисторів у вибірці обсягом n=100, узятої із серійної партії, отримані наступні значення (таблиця 3.5.1.).

Необхідно:

1. визначити закон розподілу для вибірок, узятих із серійної і дослідної партій;
2. перевірити гіпотезу про рівність дисперсій для вибірок, узятих із серійної і дослідної партій;
3. перевірити гіпотезу про рівність середніх для вибірок, узятих із серійної і дослідної партій;
4. порівняти вибірки, узяті із серійної і дослідної партій, непараметричними методами;

При виконанні завдання використовувати пакет аналізу MICROSOFT EXCEL.

Таблиця 3.5.1.

	1	2	3	4	5	6	7	8	9	10
1	0,946	0,947	0,956	0,953	0,950	0,942	0,954	0,946	0,948	0,942
2	0,952	0,946	0,943	0,949	0,947	0,946	0,947	0,943	0,944	0,951
3	0,957	0,951	0,947	0,959	0,945	0,944	0,948	0,953	0,950	0,948
4	0,946	0,948	0,944	0,950	0,944	0,952	0,956	0,949	0,949	0,953
5	0,950	0,942	0,953	0,950	0,949	0,956	0,947	0,955	0,951	0,951
6	0,948	0,951	0,958	0,941	0,953	0,949	0,957	0,961	0,943	0,953
7	0,955	0,950	0,952	0,949	0,956	0,952	0,948	0,951	0,950	0,950
8	0,943	0,946	0,951	0,946	0,949	0,953	0,945	0,955	0,946	0,955
9	0,952	0,944	0,954	0,959	0,954	0,953	0,951	0,957	0,943	0,952
10	0,949	0,956	0,947	0,947	0,952	0,954	0,954	0,936	0,958	0,950

Після внесення змін у технологічний процес була проведена дослідна партія, результати вибірки з неї наведені в таблиці 3.5.2.

Таблиця 3.5.2.

	1	2	3	4	5	6	7	8	9	10
1	0,979	0,979	0,979	0,983	0,979	0,984	0,984	0,983	0,982	0,982
2	0,978	0,978	0,972	0,972	0,981	0,974	0,968	0,979	0,978	0,975
3	0,981	0,980	0,976	0,975	0,980	0,978	0,981	0,983	0,981	0,988
4	0,984	0,982	0,979	0,978	0,978	0,982	0,980	0,976	0,975	0,986
5	0,980	0,980	0,977	0,981	0,987	0,983	0,983	0,981	0,983	0,981
6	0,984	0,978	0,977	0,977	0,979	0,976	0,990	0,976	0,986	0,983
7	0,977	0,977	0,981	0,983	0,979	0,975	0,988	0,979	0,977	0,977
8	0,980	0,977	0,980	0,982	0,982	0,984	0,979	0,981	0,985	0,986
9	0,980	0,984	0,981	0,976	0,978	0,983	0,981	0,982	0,977	0,978
10	0,978	0,979	0,975	0,983	0,978	0,980	0,981	0,981	0,982	0,981

Завдання 2**.** Виконати завдання 1 за допомогою пакета STATISTICA**.**

Завдання 3. Побудувати графіки функцій густини імовірності і розподіли імовірності випадкової величини при заданому (у таблиці 3.5.3.) законі розподілу. Значення кожного з параметрів узяти не менш трьох і підібрати їхнє сполучення для відображення відмінних рис досліджуваного розподілу.

Таблиця 3.5.3

Бета	F-розподіл.	Парето	χ^2-розподіл	Коші
Гама	Релея	Експоненц.	Лог-нормальн.	Лапласа
t-розподіл	Екстрем. велич.	Логіст.	Нормальн.	Вейбулла

Рекомендується використовувати пакет STATISTICA**.**

Завдання 4**.** Описати технологію перевірки гіпотези про закон розподілу за допомогою пакета SPSS. Як приклад виконати п.1 (щодо закону розподілу для вибірок) Завдання 1.

Завдання 5**.** Описати технологію перевірки гіпотези про рівність середніх для двох вибірок за допомогою пакета SPSS. Як приклад виконати п.2 (щодо гіпотез про рівність дисперсій) Завдання 1.

Завдання 6**.** Описати технологію перевірки гіпотези про рівність дисперсій двох вибірок за допомогою пакета SPSS. Як приклад виконати п.3 (щодо гіпотез про рівність середніх) Завдання 1.

Завдання 7**.** Описати технологію використання непараметричних критеріїв у пакеті SPSS. Як приклад виконати п.4 (порівняти непараметричними методами вибірки, узятими із серійної і дослідної партій) Завдання 1.

Завдання 8. Описати можливості модуля Descriptive statistics, викликаного зі стартової панелі Basic Statistics and Tables пакета STATISTICA, при вивченні статистичних розподілів**.**

Завдання 9. Описати можливості модулів

a) T-Test for Independent Samples by Groups - Quick Tab,
b) T-Test for Dependent Samples - Quick Tab,

викликаних зі стартової панелі Basic Statistics and Tables пакета STATISTICA при перевірці статистичних гіпотез.

4 КОРЕЛЯЦІЙНИЙ АНАЛІЗ

4.1 Основні положення

Закономірності поводження технічної, виробничої або економічної системи можна представити сукупністю залежностей. Якщо ці залежності стохастичні, то їхній аналіз здійснюється по вибірці з генеральної сукупності, і дана область досліджень відноситься до задач статистичного дослідження залежностей [1,4,5,7,8], які містять у собі кореляційний, регресійний, дисперсійний та коваріаційний аналізи.

У даному розділі розглянуті основні поняття кореляційного аналізу.

Кореляційний аналіз – сукупність методів дослідження випадкових величин, що дозволяє по вибірках з генеральних сукупностей зробити висновки про статистичні зв'язки між цими величинами.

Кореляційний зв'язок - це окремий випадок стохастичного зв'язку. При дослідженні кореляційного зв'язку зважуються три основні задачі, а саме, визначаються:

- наявність зв'язку;
- форма зв'язку;
- сила зв'язку.

Розглянемо загальну схему взаємозв'язків вхідних, вихідних перемінних (параметрів) і випадкових факторів у стохастичній системі, розглянутій як «чорна шухляда» (рис.4.1).

У «чорній шухляді» реалізується механізм перетворення вхідних перемінних у вихідні (відгук).

x_j – *вхідні перемінні*, що описують умови функціонування реальної системи (деякі з них можуть бути піддані регулюванню), ці перемінні часто називають *незалежними*;

$e^{(i)}, i = \overline{1,n}$ – *випадкові (неконтрольовані) фактори*, вплив яких на $y^{(i)}$ важко врахувати (виміряти);

$y^{(i)}, i = \overline{1,n}$ – *вихідні перемінні (відгуки), що* характеризують результат функціонування системи (об'єкта).

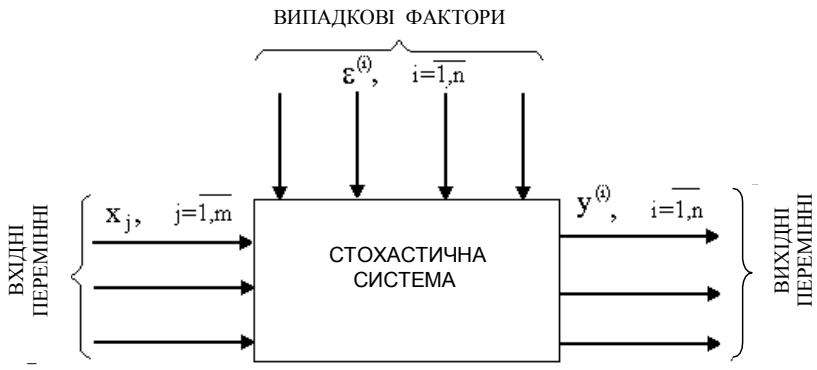


Рисунок 4.1 - Загальна схема взаємозв'язку параметрів при статистичному дослідженні залежностей

Усе це задачі кореляційного аналізу. Для визначення сили (ступеня тісноти) парних зв'язків між кількісними перемінними використовують *коефіцієнт кореляції* (іноді використовують термін **"**коефіцієнт кореляції Пірсона**"),** *кореляційне відношення***.** Іноді використовують *частковий коефіцієнт кореляції***,** застосовуваний для дослідження часток, "очищених" зв'язків, звільнених від опосередкованого одночасного впливу на досліджуваний парний зв'язок інших перемінних.

Якщо статистична інформація про багатомірну ознаку представлена не в кількісній, а в порядковій шкалі, то вимір парних зв'язків здійснюється за допомогою *рангових вибіркових показників ступеня тісноти зв'язку – коефіцієнтів кореляції Кендалла і Спірмена.*

Визначення ступеня тісноти множинного зв'язку між кількісними перемінними можливе за допомогою *множинного коефіцієнта кореляції* (чи *коефіцієнта детермінації*), а між порядковими перемінними – за допомогою *коефіцієнта конкордації***.**

4.2 Кореляційне поле

Попередній аналіз структури зв'язку між перемінними досліджуваної багатомірної ознаки, представлений вибіркою з генеральної сукупності, здійснюють за допомогою *кореляційних полів***.**

Під *кореляційним полем (діаграмою розсіювання)* перемінних (*y,x*) розуміється графічне представлення результатів вимірів $\{(y_1, x_1), \ldots, (y_i, x_i), \ldots, (y_n, x_n)\}$, цих перемінних у площині (*Y, X*). На підставі аналізу кореляційного поля легко вирішити питання про наявність чи відсутність зв'язку, простежити характер зв'язку (лінійна, нелінійна, функціональна чи стохастична) і її тенденцію (позитивна, негативна). На рис.4.2,а показане кореляційне поле, що відповідає відсутності зв'язку, на рис.4.2,б - кореляційне поле, що відповідає наявності зв'язку.

4.3 Вибірковий коефіцієнт кореляції

Вхідні і вихідні перемінні можна розглядати як компоненти двомірної випадкової величини. Нехай досліджується парна кореляційна залежність між випадковими компонентами *X* і *Y* двовимірної двомірної випадкової величини і припустимо, що в результаті експерименту отримана вибірка з двовимірної нормальної генеральної сукупності. Ступінь тісноти статистичного зв'язку між двома досліджуваними компонентами може бути обмірювана за допомогою вибіркового *коефіцієнта кореляції* [1]:

$$r_{xy} = \frac{cov(X,Y)}{s_x s_y} \tag{4.1}$$

де $\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ – оцінка другого змішаного центрального моменту випадкової величини (*X,Y*), що називається коваріацією величин *X,Y*.

Необхідно відзначити, що коефіцієнт кореляції має чіткий сенс як характеристика ступеня тісноти зв'язку тільки у випадку спільного нормального розподілу досліджуваних випадкових величин *X* і *Y*.

Властивості коефіцієнта кореляції. У загальному випадку коефіцієнт кореляції може приймати значення $|r| \le 1$. Зокрема, якщо $|r|=1$ між досліджуваними ознаками існує функціональна лінійна залежність. При $r=-1$ має місце негативна лінійна залежність, при $r=1$ – позитивна. Якщо $r=0$, то параметри X і *Y* некорельовані.

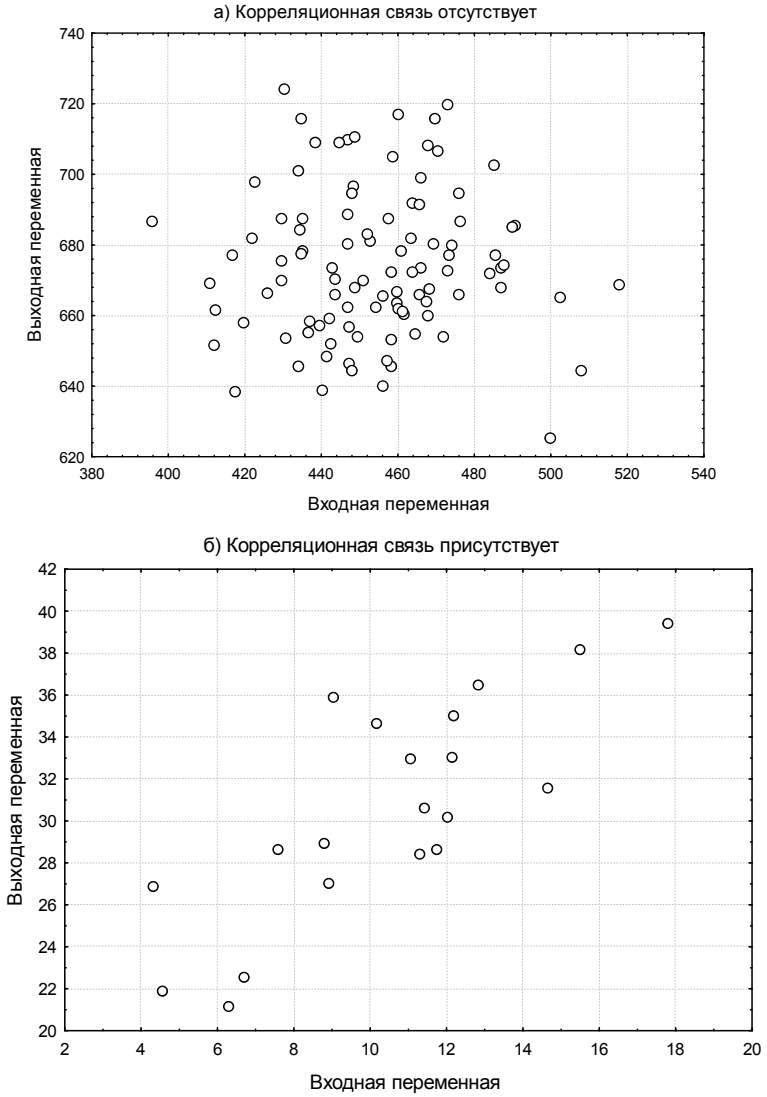


Рисунок 4.2 - Кореляційні поля, що відповідають відсутності зв'язку (а) і наявності зв'язку (б)

Однак це зовсім не означає, що X і Y незалежно, якщо апріорі допускається відхилення цієї залежності від лінійної. Отже,

некорельованість не означає незалежності досліджуваної пари ознак. У той же час незалежність завжди означає і некорельованість X і Y. При r=0 необхідно додаткове статистичне дослідження ступеня відхилення розподілу розглянутих величин від нормального. Коефіцієнт кореляції має властивість симетрії, тобто $r_{xy}=r_{yx}$.

Для випадку багатомірної випадкової величини

$$\{x_i^{(j)}\}, \quad i=\overline{1,n};\ j=\overline{1,p},$$

де p – розмірність. Статистичний аналіз усіх парних зв'язків може бути представлений кореляційною матрицею.

	$x^{(1)}$	$x^{(2)}$	. . .	$x^{(p)}$
$x^{(1)}$	1	$\hat{r}_{12}$	. . .	$\hat{r}_{1p}$
$x^{(2)}$	$\hat{r}_{21}$	1	. . .	$\hat{r}_{2p}$
. . .	. . .	. . .	1	. . .
$x^{(p)}$	$\hat{r}_{p1}$	$\hat{r}_{p2}$	. . .	1

Коефіцієнт кореляції як показник ступеня тісноти парного статистичного зв'язку має чіткий сенс при лінійному зв'язку і спільному нормальному розподілу досліджуваних пар параметрів багатомірного параметра. Парний коефіцієнт кореляції не враховує опосередкованого чи спільного впливу інших факторів.

Надійність оцінки ступеня тісноти кореляційного зв'язку слабшає зі зменшенням обсягу вибірки n, тому важливо уміти визначати мінімальне значення r_{xy}, відхилення якого від нуля можна вважати значимим. Це задача перевірки статистичної гіпотези про значимість лінійного зв'язку [1,7]. Розглянемо процедуру формування і перевірки такої гіпотези.

1-й крок. Формування гіпотези про відсутність статистичного зв'язку

H_0: $r_{xy} = 0$,

H_1: $r_{xy} \neq 0$.

2-й крок. Завдання рівня значимості α.

3-й крок. Як критичну статистику для перевірки нульової гіпотези виберемо величину

$$\psi_{кр} = \frac{r_{xy}\sqrt{n-1}}{\sqrt{1-r_{xy}}}.$$

Розподіл статистики $Y_{кр}$ має *t*-розподіл Стьюдента з (n-2) числом ступенів волі.

4-й крок. Визначення розрахункового значення критерію $T_{розр}=Y_{кр}$ і критичної точки $t_{кр}(\alpha,n-2)$ по таблиці розподілу Стьюдента (додаток таблиця).

5-й крок. Якщо виконується умова $|T_{розр}|<t_{кр,}$ то гіпотеза H_0 вірна з помилкою першого роду α, у противному випадку, H_0 відкидається.

4.4 Кореляційне відношення

Як відзначалося, при відхиленні парної статистичної залежності від лінійної коефіцієнт кореляції втрачає свій зміст як характеристика ступеня тісноти зв'язку. У цьому випадку варто скористатися таким показником (вимірником) зв'язку як *кореляційне відношення*. Кореляційне відношення застосовне в тих випадках, коли:

- між парою досліджуваних перемінних (параметрів) відзначається нелінійна залежність;
- характер вибіркових даних (кількість, щільність розташування на діаграмі розсіювання) допускає, по-перше, їхнє групування по осі вхідної перемінної, по-друге, можливість підрахунку "часток" математичних очікувань усередині кожного інтервалу групування.

Приведемо послідовність методики обчислення кореляційного відношення. Нехай має місце вибірка з двовимірної генеральної сукупності і між X і Y існує нелінійна залежність (див. рис.4.3) і компоненти X і Y мають спільно нормальний розподіл.

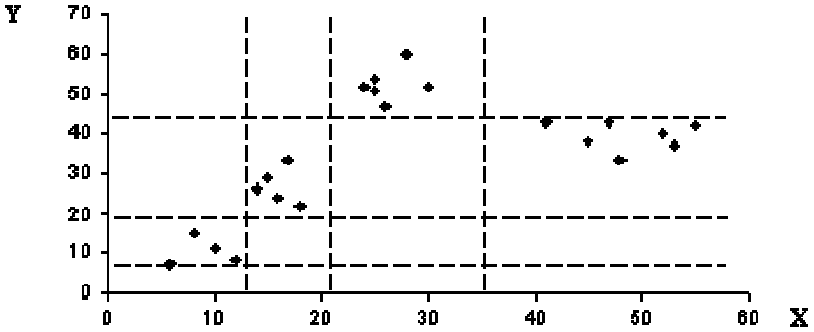


Рисунок 4.3 - Нелінійна залежність між компонентами X і Y двовимірного параметра.

1 Розіб'ємо діаграму розсіювання по перемінній X на L непересічних інтервалів групування, що можуть мати різну довжину.

2 Знайдемо "частки" математичного очікування відгуку Y у кожній з L виділених груп

$$\hat{m}_{yj} = \frac{1}{n_j} \sum_{k=1}^{n_j} y_{ij}, \tag{4.2}$$

де $j = \overline{1,L};\ k = 1,n_j$, і n_j – кількість елементів вибірки в j-ому інтервалі групування.

3 Знайдемо математичне очікування по групованому відгуку, використовуючи "частки" $\hat{m}_{y_i}$

$$\hat{m}_y = \frac{1}{n} \sum_{j=1}^{L} n_j \hat{m}_{yj}, \tag{4.3}$$

4 Одержимо групову дисперсію вихідної перемінної Y

$$\hat{\sigma}^2(\hat{m}_y) = \frac{1}{n} \sum_{j=1}^{L} n_y (\hat{m}_{yj} - \hat{m}_y)^2, \tag{4.4}$$

і дисперсію, отриману по негрупованому відгуку

$$\hat{\sigma}^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_j - \hat{m}_y)^2 . \qquad (4.5)$$

5 Кореляційне відношення залежної перемінної Y по незалежній перемінній X може бути отримане з відношення

$$\hat{\rho}_{yx} = \frac{\hat{\sigma}(\hat{m}_y)}{\hat{\sigma}(y)} , \qquad (4.6)$$

Властивості кореляційного відношення. Кореляційне відношення не має властивості симетрії, тобто $\rho_{xy} \neq \rho_{yx}$. Крім того ρ_{xy} не негативне, оскільки передбачається, що воно є результатом витягу кореня квадратного з $(\rho_{xy})^2$. Кореляційне відношення $0 \leq \rho_{xy} \leq 1$. З $\rho_{xy}=1$ випливає, що між Y і X існує однозначна функціональна залежність. Зворотне твердження в загальному випадку не вірне. Відсутність кореляційного зв'язку між Y і X означає, що умовні середні $\hat{m}_{yj}$ зберігають від групи до групи постійне значення, рівне загальному середньому $\hat{m}_y$, тому $\rho_{xy}=0$. Необхідно також відзначити, що між ρ_{xy} і ρ_{yx} немає якої-небудь визначеної залежності. Некорельованість Y від X не означає некорельованості X від Y.

4.5 Власні коефіцієнти кореляції

Іноді в практичних ситуаціях не вдається інтерпретувати на змістовному рівні виявлений парний зв'язок між досліджуваними компонентами ознаки. Причину цього часто варто шукати в опосередкованому впливі на досліджувані показники деякого третього фактора [1]. Роль опосередкованих факторів, що впливають, можуть грати безліч неврахованих показників. Отже, необхідне введення показників статистичного зв'язку, що були б очищені від такого впливу. Як показник ступеня тісноти зв'язку між перемінними X і Y при фіксованих значеннях інших перемінних використовуються власні (“очищені”) коефіцієнти кореляції.

Нехай є багатомірний нормальний вектор *X={x(1), x(2), ..., x(p)},*

де *x(i)* – компоненти вектора, *p* – його розмірність. Необхідно визначити власний коефіцієнт кореляції r_{ij} між *x(i)* і *x(j)* компонентами вектора при фіксованій множині перемінних *x(i,j),* що доповнюють пари *x(i)* і *x(j)*. За даних умов

$$r_{ij \cdot x(i)} = \frac{-R_{ij}}{R_{ii} \cdot R_{jj}}, \tag{4.7}$$

де *Rij* – алгебраїчне доповнення для елемента r_{ij} у детермінанті кореляційної матриці *R* аналізованих ознак *x(i),* тобто у детермінанті

$$\det R = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & 1 & r_{23} & & r_{2p} \\ \dots & \dots & \dots & & \dots \\ r_{p1} & r_{p2} & r_{p3} & & 1 \end{vmatrix} \tag{4.8}$$

Вираз (4.1.6) за умови *p*=3 буде мати вигляд

$$r_{12 \cdot x(3)} = r_{12 \cdot (3)} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}, \tag{4.9}$$

4.6 Рангова кореляція

Іноді при дослідженні залежностей має місце ситуація, коли шкала кількісного виміру ступеня прояву деякої властивості (ознаки) відсутня (невідома) чи її просто не може бути. Крім того, можлива ситуація, коли інформація має умовний характер і може бути використана тільки для ранжирування об'єктів. Прикладами таких процесів можуть служити показники ефективності функціонування різних соціально-економічних систем, структура споживчого бюджету родини, ступінь прогресивності пропонованого на конкурс проекту. У подібних ситуаціях замість конкретних значень досліджуваної ознаки використовуються її ранги.

Рангова кореляція відбиває статистичний зв'язок між порядковими перемінними. Вихідний статистичний матеріал представлений упорядкуваннями (ранжуваннями) *n* об'єктів по деяких властивостях. Методи рангової кореляції засновані на використанні умовної числової мітки, що позначає місце об'єкта в ряді всіх аналізованих об'єктів, що розташовуються в порядку убування досліджуваної властивості. При цьому під умовною числовою міткою розуміється *ранг об'єкта* по досліджуваній ознаці.

Послідовність рангів елементів варіаційного ряду, що вказують на місце об'єкта в ряді, називається *ранжуванням*.

Під *ранговою кореляцією* розуміється статистичний зв'язок між порядковими перемінними.

Існують методи і показники, що дозволяють оцінити і проаналізувати статистичний парний і множинний зв'язок між декількома параметрами досліджуваного багатомірного об'єкта, якщо вони представлені ранжуваннями.

Для виміру ступеня тісноти парного статистичного зв'язку між ранжуваннями К.Спірмен у 1904 р. запропонував показник, що згодом одержав назву рангового коефіцієнта кореляції Спірмена [1]:

$$\overset{\wedge}{\tau}{}_{kj}^{(s)} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (R_i^{(k)} - R_i^{(j)})^2, \qquad (4.10)$$

де $R_i^{(k)}$ і $R_i^{(j)}$ – *i*-*i* ранги відповідно параметрів *k* і *j*.

Вираз (4.1.9) справедливий при відсутності в ранжуваннях груп об'єднаних рангів. Якщо такі групи є, то $\overset{\wedge}{\tau}{}_{kj}^{(s)}$ визначається з виразу:

$$\overset{\wedge}{\tau}{}_{kj}^{(s)} = \frac{(1/6)(n^3 - n) - \sum_{i=1}^{n} (R_i^{(k)} - R_i^{(j)})^2 - T^{(k)} - T^{(j)}}{\sqrt{[(1/6)(n^3 - n) - 2T^{(k)})][(1/6)(n^3 - n) - 2T^{(j)}]}}, \qquad (4.11)$$

де $T^{(k)}$ і $T^{(j)}$ можуть бути знайдені з

$$T^{(k)} = (1/12)\sum_{i=1}^{m(k)}[(n_i^{(k)})^3 - n_i^{(k)}], \qquad (4.12)$$

де $n_i^{(k)}$ – кількість елементів у групі нерозрізнених рангів, а $m(k)$ – число груп нерозрізнених рангів.

Неважко переконатися, що при співпадаючих ранжуваннях $R_i^{(k)} = R_i^{(j)}$, а $\hat{\tau}_{kj}^{(s)} = 1$, при протилежних $\hat{\tau}_{kj}^{(s)} = -1$. У всіх інших випадках $\hat{\tau}_{kj}^{(s)} \prec 1$. Якщо $\hat{\tau}_{kj}^{(s)} = 0$, то зв'язок між компонентами відсутній. Крім того, очевидно, що ранговий коефіцієнт кореляції має властивість симетрії, тобто $\hat{\tau}_{kj}^{(s)} = \hat{\tau}_{jk}^{(s)}$.

Коефіцієнт конкордації.. Властивості розглянутих вище показників парних зв'язків свідчать про те, що чим тісніше зв'язок, тим більше інформації містить одна перемінна щодо іншої. На практиці буває важливо пояснити поводження однієї перемінної (відгуку) поводженням сукупності інших. Для рішення таких задач використовуються вимірники ступеня тісноти множинного зв'язку [1]

Кендаллом був запропонований показник $\hat{W}(m)$, названий коефіцієнтом конкордації (погодженості), що обчислюється з виразу:

$$\hat{W}(m) = \frac{12}{m^2(n^3 - n)}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} R_i^{(k)} - \frac{m(n+1)}{2}\right)^2, \qquad (4.13)$$

де m – число одночасно аналізованих порядкових перемінних,

$R_i^{(k)}$ – і-ий ранг відібраної для дослідження порядкової перемінної k_j; $j = \overline{1, m}$ – номер цієї перемінної в досліджуваній багатомірній ознаці.

Коефіцієнт конкордації має наступні властивості:

- $0 < \hat{W}(m) < 1$;
- $\hat{W}(m) = 1$ за умови, коли всі m аналізованих упорядкувань збігаються;
- коефіцієнт конкордації, обчислений для двох перемінних, пропорційний парному ранговому коефіцієнту кореляції Спірмена $\hat{\tau}_{kj}^{(s)}$.

4.7 Комп'ютерні технології кореляційного аналізу

***ВИКОРИСТАННЯ ВМОНТОВАНИХ СТАТИСТИЧНИХ ФУНКЦІЙ І ПАКЕТА АНАЛІЗУ* MICROSOFT EXCEL**

Для проведення кореляційного аналізу можуть бути використані ФУНКЦІЇ:

КОВАР - повертає значення коваріації, тобто середнє добутків відхилень для кожної пари точок даних. Коваріація використовується для визначення зв'язку між двома множинами даних.

КОРРЕЛ - повертає коефіцієнт кореляції. Коефіцієнт кореляції використовується для визначення наявності взаємозв'язку між двома вибірками.

ЗАСОБИ СТАТИСТИЧНОГО АНАЛІЗУ доступні через команду *Аналіз даних* меню *Сервіс*. У вікні Аналіз даних для кореляційного аналізу використовуються наступні інструменти.

КОРЕЛЯЦІЯ для кількісної оцінки взаємозв'язку двох наборів даних. Коефіцієнт кореляції вибірки представляє відношення коваріації двох наборів.

КОВАРІАЦІЯ дає можливість установити, чи асоційовані набори даних по величині, тобто великі значення з одного набору даних зв'язані з великими значеннями іншого набору (позитивна коваріація), або, навпаки, малі значення одного набору зв'язані з великими значеннями іншого (негативна коваріація), чи дані двох діапазонів ніяк не зв'язані (коваріація близька до нуля).

ПРИКЛАД 4.1. У таблиці 4.1 дані два набори даних

Таблиця 4.1

1	2,0	2,1
2	3,7	9,4
3	1,4	1,9
4	0,8	-1,6
5	0,8	2,7
6	1,7	2,8
7	3,7	5,8
8	2,0	3,2
9	1,2	4,2
10	1,6	4,7
11	-0,9	5,7
12	-1,1	-3,1
13	3,8	7,0
14	1,1	4,3
15	2,4	5,1
16	-0,3	-1,8
17	-0,3	6,1
18	1,1	3,4
19	4,2	4,1
20	-0,9	-1,9

В результаті обробки даних таблиці 4.1 за допомогою інструмента Кореляція одержимо.

	Рядок 1	*Рядок 2*
Рядок 1	1	
Рядок 2	0,750144	1

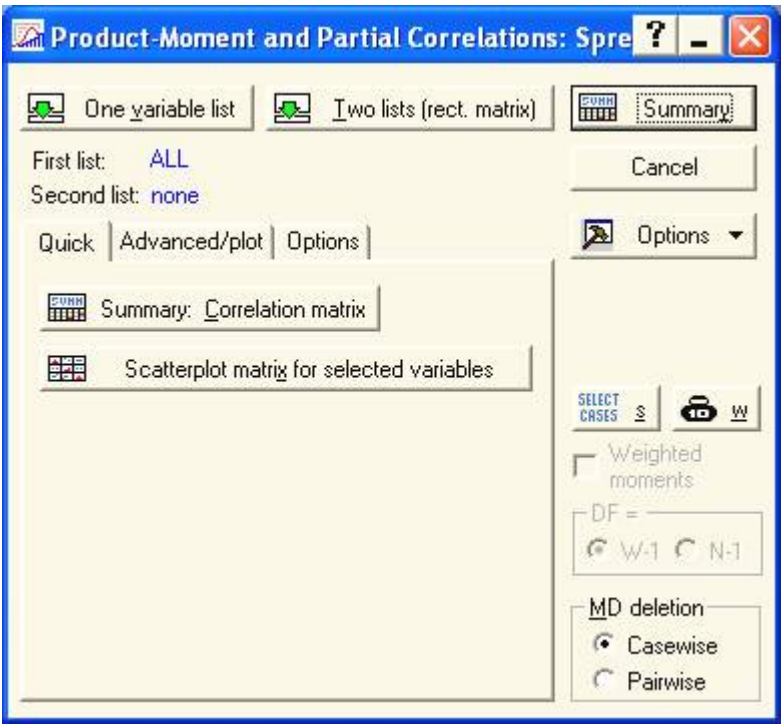
Отже, коефіцієнт кореляції дорівнює 0,75.

КОРЕЛЯЦІЙНИЙ АНАЛІЗ У ПАКЕТІ STATISTICA

Крок 1. Увести або імпортувати (наприклад, з Excel) вихідні дані в робочу книгу (Workbook) системи STATISTICA, виділити їх, увести назву таблиці вихідних даних і назви перемінних.

Крок 2. Клацнути по кнопці Start menu …, розташованій в лівому нижньому куті вікна додатка, й у меню, що з'явилося, вибрати Statistics → Basic Statistics and Tables → Correlation matrices.

Крок 3. У вікні, що з'явилося, Product-Moment and Partial Correlations - Quick Tab



вибрати кнопку Summary: Correlation matrix, після чого з'являється вікно з кореляційною матрицею по обраним перемінним.

ПРИКЛАД 4.7.2. При виготовленні біполярних інтегральних мікросхем однією з найважливіших технологічних операцій є «дифузія фосфору». Дана операція характеризується наступними перемінними.

Вихідна перемінна:

y - глибина дифузії фосфору, мкм.

Вхідні перемінні:

x_1 – час дифузії 1-ої стадії, хв.;

x_2 – час дифузії 2-ий стадії, хв.;

x_3 – поверхневий опір (ом/$\square$) після 1-ої стадії дифузії;

x_4 – поверхневий опір (ом/$\square$) після 2-ої стадії дифузії;

x_5 – температура дифузії;

Результати вимірів за даними 30 партій приведені в таблиці 4.2.

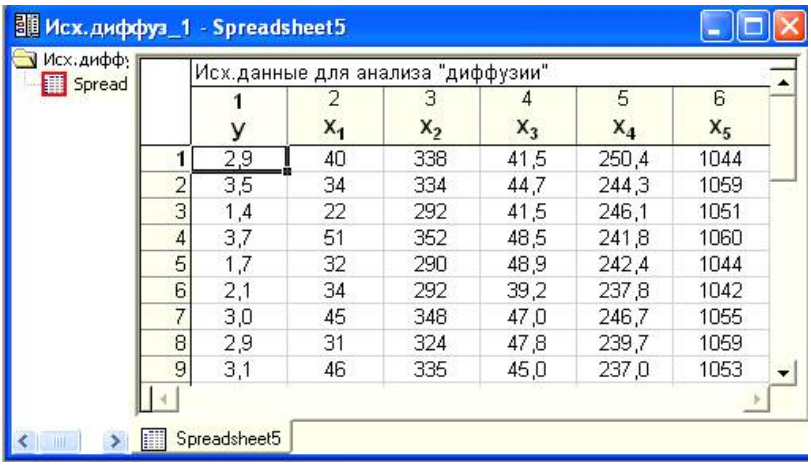
Таблиця 4.2

	y	x_1	x_2	x_3	x_4	x_5
1	2,9	40	338	41,5	250,4	1044
2	3,5	34	334	44,7	244,3	1059
3	1,4	22	292	41,5	246,1	1051
4	3,7	51	352	48,5	241,8	1060
5	1,7	32	290	48,9	242,4	1044
6	2,1	34	292	39,2	237,8	1042
7	3,0	45	348	47,0	246,7	1055
8	2,9	31	324	47,8	239,7	1059
9	3,1	46	335	45,0	237,0	1053
10	2,6	44	326	45,1	247,2	1054
11	2,3	45	336	33,7	255,6	1047
12	2,7	43	304	45,8	248,4	1054
13	3,0	46	344	46,0	244,1	1064
14	1,6	20	308	40,4	231,8	1044
15	2,3	49	288	41,7	246,1	1045
16	1,4	24	315	41,8	244,0	1057
17	2,9	45	330	43,0	250,2	1046
18	2,6	42	316	46,9	250,5	1049
19	1,9	37	325	42,5	242,7	1051
20	1,9	36	319	43,0	238,6	1053

21	1,6	33	284	35,2	235,1	1037
22	1,8	33	275	36,3	248,0	1048
23	4,4	50	351	42,0	224,7	1043
24	1,6	28	289	44,9	229,4	1034
25	3,8	58	353	39,9	231,2	1044
26	2,1	37	318	33,2	229,5	1055
27	2,5	36	323	46,3	232,4	1044
28	0,4	33	260	41,1	240,0	1056
29	3,4	55	349	50,4	234,5	1059
30	3,6	43	352	42,2	229,3	1053

Необхідно: оцінити наявність і силу (тісноту) кореляційних зв'язків між вихідною і вхідними перемінними.

Рішення. Для рішення поставленої задачі може бути використана технологія кореляційного аналізу статистичний пакет STATISTICA. Підготуємо вихідні дані приведені в таблиці 4.2 за технологією, викладеної в п.1.2.



	1 y	2 x_1	3 x_2	4 x_3	5 x_4	6 x_5
1	2,9	40	338	41,5	250,4	1044
2	3,5	34	334	44,7	244,3	1059
3	1,4	22	292	41,5	246,1	1051
4	3,7	51	352	48,5	241,8	1060
5	1,7	32	290	48,9	242,4	1044
6	2,1	34	292	39,2	237,8	1042
7	3,0	45	348	47,0	246,7	1055
8	2,9	31	324	47,8	239,7	1059
9	3,1	46	335	45,0	237,0	1053

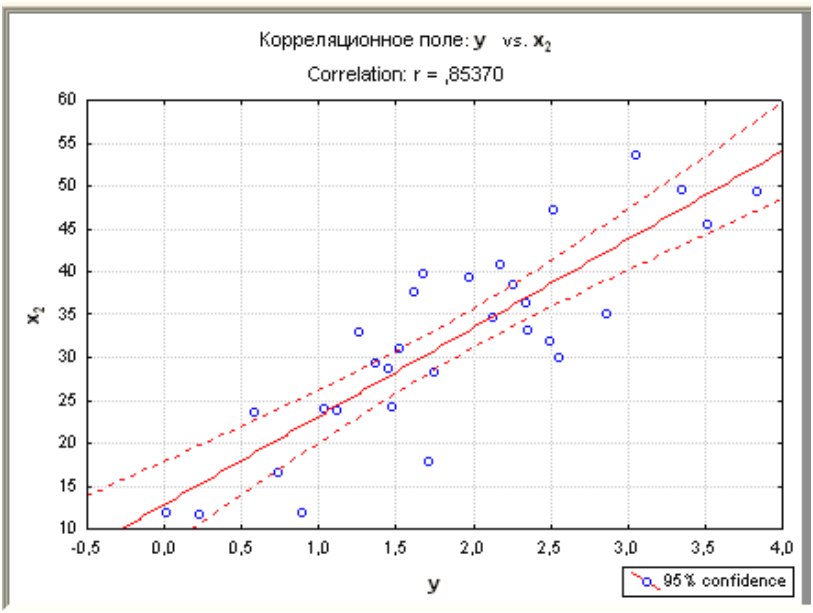
Виконавши команду Statistics →Basic Statistics and Tables →Correlation matrices, кроки 1-4, одержимо кореляційну матрицю

* - Correlations (Spreadsheet5 in Исх.диффуз_1)

Корреляционная матрица
для анализа технологической операции "диффузия"

Перемен.	y	x_1	x_2	x_3	x_4	x_5
y	1,00	0,74	0,85	0,33	-0,16	0,22
x_1	0,74	1,00	0,63	0,20	0,02	0,16
x_2	0,85	0,63	1,00	0,30	-0,11	0,36
x_3	0,33	0,20	0,30	1,00	0,05	0,39
x_4	-0,16	0,02	-0,11	0,05	1,00	0,23
x_5	0,22	0,16	0,36	0,39	0,23	1,00

з якої видно, що між y - глибиною дифузії фосфору і перемінними x_1 – дифузії 1-ої стадії; x_2 – часом дифузії 2-ої стадії існує досить сильний кореляційний зв'язок. Також спостерігаються незначні кореляції між вхідними перемінними. Кореляційне поле y - x_2 указує на лінійний характер зв'язку між цими перемінними.



4.8 Контрольні запитання і завдання

4.8.1 Питання

1. Кореляційний аналіз і його задачі.
2. Загальна схема взаємозв'язків параметрів і факторів у стохастичній системі.
3. Кореляційні поля, їхнє призначення.
4. Вибірковий коефіцієнт кореляції і його обчислення.
5. Властивості коефіцієнта кореляції.
6. Кореляційна матриця.
7. Перевірки статистичної гіпотези про значимість коефіцієнта кореляції.
8. Кореляційне відношення і його властивості.
9. Обчислення кореляційного відношення.
10. Власні коефіцієнти кореляції.
11. Рангова кореляція, ранговий коефіцієнт кореляції Спірмена.
12. Коефіцієнт конкордації Кендалла.

4.8.2 Завдання

Завдання 1**.** За допомогою технології кореляційного аналізу статистичного пакета STATISTICA побудувати кореляційні поля для перемінних з прикладу 4.2:

a) $y \leftrightarrow x_1$;
b) $x_2 \leftrightarrow x_5$;
c) $x_3 \leftrightarrow x_5$.

Завдання 2**.** Описати технологію кореляційного аналізу за допомогою пакета SPSS. Як приклад досліджувати наявність і силу кореляційного зв'язку між вибірками, приведеними в таблиці 3.1 і 3.2 завдання 1 (п. 3.2).

Завдання 3. Описати можливості модуля Product-Moment and Partial Correlations - Quick, викликаного зі стартової панелі командою Statistics → Basic Statistics and Tables → Correlation matrices.

Завдання 4. Описати можливості вкладки Prob. & Scatterplots Tab модуля Descriptive Statistics.

РЕГРЕСІЙНИЙ АНАЛІЗ

5.1 Основні положення

Регресійний аналіз дає більш повну інформацію про зв'язок результуючої випадкової величини (вихідна перемінна) із вхідними перемінними, котрі звичайно називають у регресійному аналізі факторами [1,4,5,7]. У кореляційному аналізі зв'язок між випадковими величинами описується числами (коефіцієнтом кореляції, кореляційним відношенням), а в регресійному аналізі описується за допомогою функцій - рівнянь регресії. Рівняння регресії може бути записане у вигляді:

$$\hat{M}(y / x_1, x_2, ..., x_k) = f(x_1, x_2, ..., x_k), \qquad (5.1)$$

де $\hat{M}(y / x_1, x_2, ..., x_k)$ - оцінка умовного математичного очікування випадкової величини y за умови, що багатомірна випадкова величина, що визначає сукупність вхідних перемінних, приймає значення $x_1, x_2, ..., x_k$.

Нехай в результаті спостережень вихідна перемінна приймає в кожному *і-ому* спостереженні певні значення y_i при відповідному наборі значень вхідних перемінних $x_1,x_2...x_k$. Якщо провести серію з n спостережень, то результати спостережень можна представити в наступному вигляді:

№ спостереження	Вихідна перемінна	Вхідні перемінні
1-е спостереження	y_1	$x_{11}, x_{12},...,x_{1k}$
2-е спостереження	y_2	$x_{21}, x_{22},...,x_{2k}$
3-є спостереження	y_3	$x_{31}, x_{32},...,x_{3k}$
.....	.....	.....
n-е спостереження	y_n	$x_{n1}, x_{n2},...,x_{nk}$

Залежність випадкової величини y від незалежних перемінних $x_1,x_2,...,x_k$, якщо враховувати тільки лінійні ефекти, тобто

знехтувати спільним впливом факторів x_{ij}, x_{ii}, відбивається рівнянням множинної регресії:

$$y=b_0+b_1x_1+b_2x_2+...+b_kx_k; \qquad (5.2)$$

Для одержання оцінок коефіцієнтів $b_0,b_1,b_2,...,b_k$ рівняння регресії (5.2) використовується метод найменших квадратів [1,4]. Оцінки b, $s=0,k$ виходять з розв'язку систем рівнянь, утворених прирівнюванням нулю часткових похідних по b_S форми

$$Q = \sum_{i=1}^{n}(y_i - b_0 - \sum_{s=1}^{k} b_s x_{is}); \qquad (5.3)$$

В результаті розв'язання такої системи рівнянь для оцінки параметра b_0 отримано:

$$b_0 = (1/n)\sum_{i=1}^{n} y_i ; \qquad (5.4)$$

Параметр b_0 називається вільним членом і він визначає перетинання ординати при рівних нулю незалежних перемінних. Вираз для оцінок b_s, $s=\overline{1,k}$ визначається в такий спосіб.

Уведемо позначення:

$$l_{rs} = (1/n)\sum_{i=1}^{n}(x_{ri} - \overline{x_r})(x_{si} - \overline{x_s}); \qquad 1 \le r \le s \le k; \qquad (5.5\ а)$$

$$l_{0s} = (1/n)\sum_{i=1}^{n}(y_i - \overline{y})(x_{si} - \overline{x_s}); \qquad (5.5\ б)$$

Елементи l_{rs} утворюють матрицю, визначник якої позначимо Z і, якщо позначити Z_s визначник матриці, що виходить заміною S стовпця на l_{0s}, $s=\overline{1,k}$, тобто:

$$b_S = \frac{Z_s}{Z}, \qquad (5.6)$$

На основі елементів матриці розраховуються парні коефіцієнти кореляції r_{is} між незалежними перемінними x_s, $i,s=1,k$ і коефіцієнти

кореляції r_{i0} між залежної перемінний y і незалежними перемінними. Елементи r_{is} складають кореляційну матрицю. Виходячи з парних коефіцієнтів кореляції, можна розрахувати коефіцієнт множинної кореляції

$$R = \left[\sum_{i=1}^{k} \gamma_i r_{i0} \right]^{1/2}; \qquad (5.7)$$

де $\gamma_i = \sum_{i=1}^{k} r_{i0} r_{ij}^{-1};$

rij-1-елементи матриці, зворотної до кореляційної матриці.

Оцінка значимості коефіцієнтів регресії здійснюється за допомогою *t*-критерію. Для одержання обчислених значень *t*-критерію t_j, $j = \overline{1,s}$ використовується формула:

$$t_j = b_j / s(b_j); \qquad (5.8)$$

Середньоквадратичне відхилення коефіцієнтів регресії дорівнює:

$$s(b_j) = \sqrt{D_{jj}^{-1} r_{jj}^{-1} s^2}; \qquad (5.9)$$

де $D_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2;$ $\qquad s^2 = \frac{1}{n-m-1} \sum_{i=1}^{n} \varepsilon_i;$

ε_i*-відхилення* фактичних значень y від розрахованих по рівнянню регресії.

Передумови. Передбачається, що "залежна" випадкова величина y підпорядкована нормальному закону розподілу з постійною дисперсією.

Рекомендується перед проведенням регресійного аналізу:

- оцінити вплив неконтрольованих і контрольованих невраховуваних перемінних на розподіл "залежної" перемінної *в*;
- виключити з аналізу лінійно залежні фактори;

- оцінити тимчасову стабільність розподілу *в*, для чого перевірити гіпотези про однорідність середньоарифметичних і дисперсії *в*, розрахованих за результатами вибірок, узятих у різний час.

Також аналіз отриманого рівняння регресії утруднює різнобій в одиницях виміру різних факторів (градуси, тонни, кілометри, ангстреми і т.п.). Для того, щоб по виду рівняння можна було скласти уявлення про ступінь впливу кожного фактора на вихідну перемінну, необхідно перейти до безрозмірних величин. Замість випадкових величин $x_1, x_2, ..., x_k$ беруть центровані, нормовані випадкові величини

$$\overset{\circ}{x}_i = \frac{x_i - \mu}{\sigma}; \tag{5.10}$$

Для нормованих величин рівняння регресії (5.2) прийме вид

$$y = \beta_0 + \beta_1 \overset{\circ}{x}_1 + \beta_2 \overset{\circ}{x}_2 + ... + \beta_k \overset{\circ}{x}_k; \tag{5.11}$$

Приведені (стандартизовані, нормовані) коефіцієнти β_i характеризують реальний внесок кожного з факторів у варіацію вихідної перемінної *y*. Чим більше абсолютна величина β_i, тим цей внесок більше.

5.2 Комп'ютерні технології регресійного аналізу

ВИКОРИСТАННЯ ВМОНТОВАНИХ СТАТИСТИЧНИХ ФУНКЦІЙ І ПАКЕТА АНАЛІЗУ MICROSOFT EXCEL.

ЛИНЕЙН (відомі_значення_y; відомі_значення_x; конст; статистика) - розраховує за допомогою методу найменших квадратів параметри для лінійного рівняння регресії. Синтаксис цієї функції наступний:

відомі значення y і x - це значення вихідної (залежної) і вхідної (незалежної) перемінних;

“конст” - це логічне значення, що вказує, чи потрібно, щоб вільний член рівняння регресії дорівнював 0, якщо аргумент конст має значення НЕПРАВДА, чи обчислюється звичайно, якщо аргумент “конст”— має значення ІСТИНА; “статистика”— це логічне значення,

що вказує, чи потрібно повернути додаткову статистику по регресії, якщо аргумент "статистика" має значення НЕПРАВДА чи опущена, то функція ЛИНЕЙН повертає тільки коефіцієнти рівняння регресії; якщо аргумент "статистика" має значення ІСТИНА, то функція ЛИНЕЙН повертає додаткову регресійну статистику, а саме стандартні значення помилок для коефіцієнтів, стандартну помилку для оцінки *y*, *F*-статистику (*F*-статистика використовується для визначення того, чи є взаємозв'язок між залежною і незалежною перемінними випадковою чи ні), ступеня вільності та ін.

ПРЕДСКАЗ (х;відомі_значення_у;відомі_значення_х) - передбачає значення на основі рівняння лінійної регресії. Значення, що передбачається - це *y*-значення, що відповідає заданому *x*-значенню. Відомі значення - це *x*- і *y*-значення, а нове значення передвіщається з використанням лінійної регресії. Цю функцію можна використовувати для передбачення майбутніх продажів, потреб в устаткуванні чи тенденцій споживання.

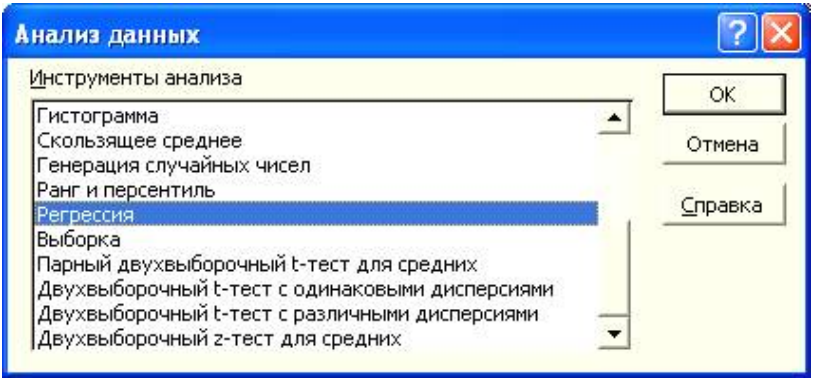
x - це точка даних, для якої передбачається значення.

ТЕНДЕНЦІЯ(відомі_значення_у;відомі_значення_х;нові_значення_х;конст). Повертає значення відповідно до лінійного тренда. Апроксимує прямою лінією (по методу найменших квадратів) _значення_*y* по відомим_значеннях_*x*. Повертає значення *y* відповідно до цієї прямої для заданого масиву нових_значень_*x*.

Відомі_значення_у і х, "конст" визначаються аналогічно як і у функції ЛИНЕЙН.

Нові_значення_х - це нові значення х, для яких ТЕНДЕНЦІЯ повертає відповідні значення у.

ЗАСОБИ СТАТИСТИЧНОГО АНАЛІЗУ доступні через команду *Аналіз даних* меню *Сервіс*. У вікні Аналіз даних для регресійного аналізу використовується інструмент - РЕГРЕСІЯ.



Після вибору інструмент – регресія і клацання по кнопці ОК з'являється вікно Регресія (аналогічне вікнам Двовибірковий F-тест для дисперсії (п.3.4) і Кореляція) , у якому запитуються вихідні дані. Після введення вихідних даних і клацання по кнопці ОК виводяться таблиці з регресійною статистикою, коефіцієнтами регресії, стандартними помилками, *t*-статистикою й ін. При необхідності можна вивести залишки між передбаченими по рівнянню і значеннями, що спостерігаються, залежної перемінної *y*.

ПРИКЛАД 5.1. Використовуємо вихідні дані ПРИКЛАДУ 4.2**.**

Необхідно: Одержати рівняння лінійної регресії, що відбиває вплив на глибину дифузії *(y)* – час дифузії 1-ої стадії ($x_{1)}$.

Рішення. Використовуємо технологію проведення регресійного аналізу за допомогою інструмента - регресія з вікна Аналіз даних MICROSOFT EXCEL. В результаті одержимо

	Коефі-цієнти	*Стандартна помилка*	*t-статистика*
Y-перетинання	-0,2442	0,4796	-0,5091
Перемінна X 1	0,0700	0,0119	5,8570

Регресійна статистика	
Множинний R	0,742
R-квадрат	0,551
Нормований R-квадрат	0,535

Стандартна помилка	0,603
Спостереження	30

Якщо установити відповідний прапорець у вікні Регресія, то виводяться значення *y*, що *передбачаються,* і залишки між передвіщеними і значеннями *y*, що *спостерігаються*.

ВИВЕДЕННЯ ЗАЛИШКУ		
Спостереження	*Передвіщене Y*	*Залишки*
1	2,555	0,345
2	2,135	1,365
3	…..	…….

Рівняння регресії можна відобразити в графічному вигляді за допомогою Майстра діаграм додатка MICROSOFT EXCEL. Графік для розглянутого приклада показаний на рис.5.1.

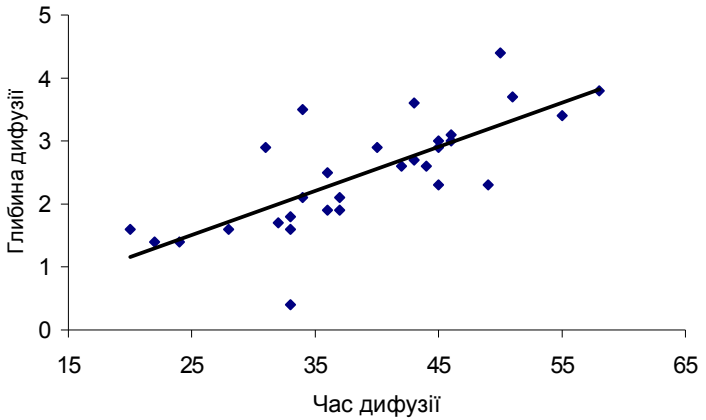
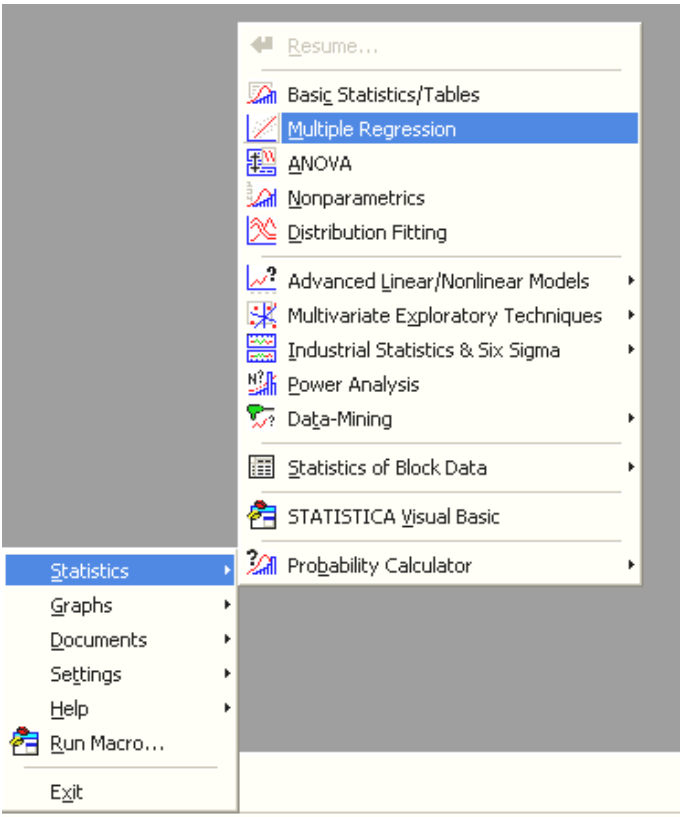


Рисунок 5.1 - Графік рівняння лінійної регресії, що відбиває вплив на глибину дифузії часу дифузії 1-ої стадії

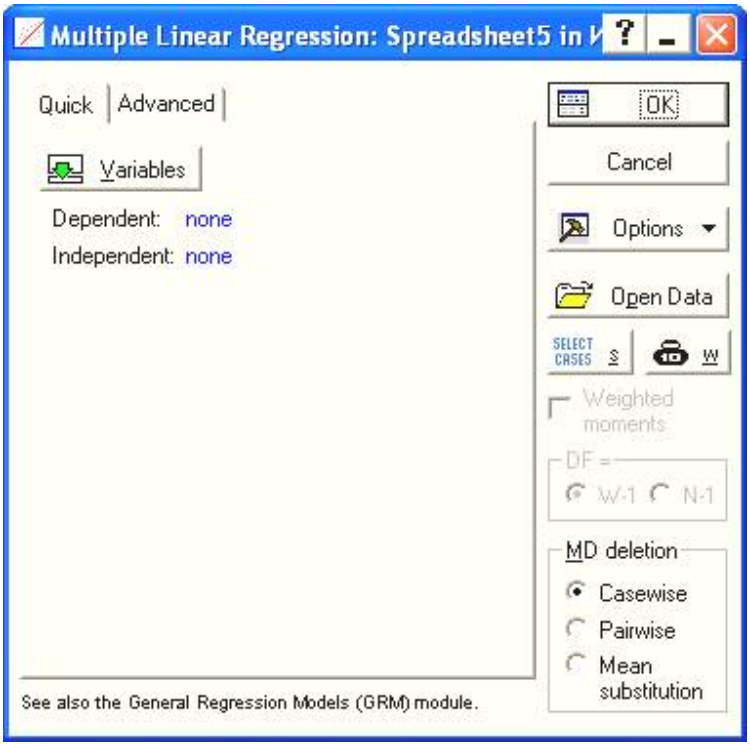
РЕГРЕСІЙНИЙ АНАЛІЗ У ПАКЕТІ STATISTICA

Крок 1. Ввести чи імпортувати (наприклад, з Excel) вихідні дані в робочу книгу (Workbook) системи STATISTICA, виділити їх, увести назву таблиці вихідних даних і назви перемінних.

Крок 2. Клацнути по кнопці Start menu ..., розташованій в лівому нижньому куті вікна додатка, й у меню, що з'явилось, вибрати Statistics → Multiple Regression



Крок 3. У вікні, що з'явилося, Multiple Linear Regression: ...



клацнути по кнопці Variables, після чого з'являється вікно Select dependent and independent variable list, у якому необхідно задати вихідну (залежну) перемінну і вхідні (незалежні) перемінні.

Крок 4. Клацнути по кнопці OK і з'являється вікно Multiple Regression Results, у якому відбиті результати аналізу і містяться вкладки, що дозволяють одержати більш детальну інформацію. За допомогою кнопки Summary: Regression results на вкладці Quick викликається вікно Regression Summary for Dependent Variable, у якому виведені коефіцієнти рівняння регресії і їхніх оцінок.

ПРИКЛАД 5.1 Використовуємо вихідні дані ПРИКЛАДА 4.2**.**

Необхідно: Одержати рівняння множинної регресії, що відбиває вплив на глибину дифузії фосфору технологічних факторів – часів проведення дифузії, поверхневих опорів після кожної стадії дифузії і температури дифузії.

Рішення. Використовуємо технологію проведення регресійного аналізу в пакеті STATISTICA (Кроки 1 – 4). У вікні Select dependent and independent variable list як залежну перемінну задамо глибину дифузії фосфору *y,* а як незалежні

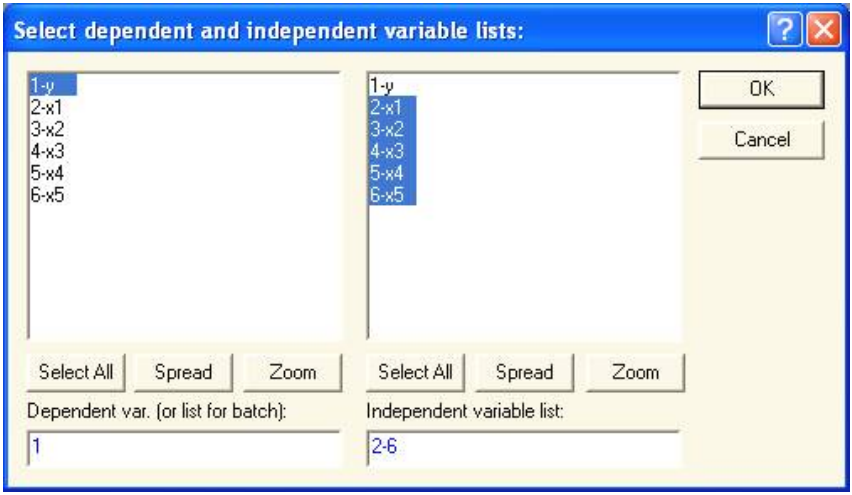
x_1 – час дифузії 1-ої стадії, хв.;

x_2 – час дифузії 2-ої стадії, хв.;

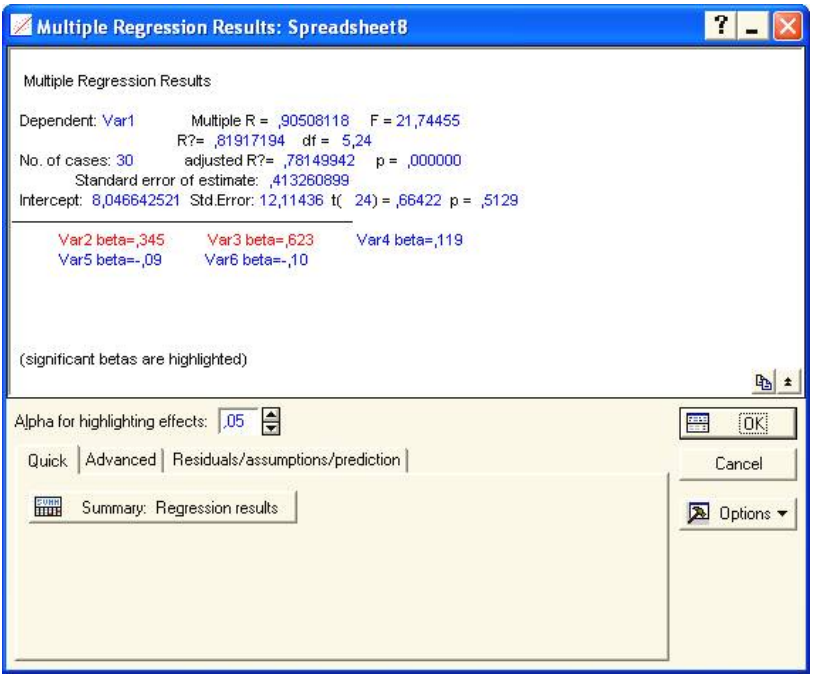
x_3 – поверхневий опір (ом/$\square$) після 1-ої стадії дифузії;

x_4 – поверхневий опір (ом/$\square$) після 2-ої стадії дифузії;

x_5 – температура дифузії;



Далі у вікні Multiple Linear Regression: … клацнути по кнопці OK і з'являється вікно з результатами регресійного аналізу



Вибравши кнопку Summary: Regression results, побачимо вікно Regression Summary for Dependent Variable

з якого видно, що значимими будуть тільки коефіцієнти рівняння при перемінних x_1, x_2. У стовпці Beta показані нормовані (стандартизовані) коефіцієнти, у стовпці B реальні (нестандартизовані) коефіцієнти.

5.3 Контрольні запитання і завдання

5.3.1 Запитання

1 Призначення регресійного аналізу
2 Загальний вид рівняння регресії
3 Одержання оцінок коефіцієнтів рівняння регресії
4 Передумови застосування регресійного аналізу
5 Вид рівняння регресії для нормованих величин
6 Використання убудованих статистичних функцій додатка MICROSOFT EXCEL для проведення регресійного аналізу
7 Використання Пакета аналізу додатка MICROSOFT EXCEL для регресійного аналізу
8 Регресійний аналіз у пакеті STATISTICA

5.3.2 Завдання

Завдання 5.1**.** Для дослідження впливу на товщину термічно вирощених плівок діоксиду кремнію (SiO_2) часу і температури окислювання отримані наступні результати (таблиця 5.1).

Необхідно. Побудувати рівняння регресії, що відбивають вплив на товщину плівки SiO_2 досліджуваних технологічних факторів, проаналізувати отримане рівняння.

Завдання 5.2. За допомогою статистичного пакета STATISTICA, використавши технологію проведення регресійного аналізу, побудувати кореляційні поля для перемінних.

Завдання 5.3. Описати технологію кореляційного аналізу за допомогою пакета SPSS. Як приклад досліджувати наявність і силу

кореляційного зв'язку між вибірками, наведеними в таблиці 3.1 і 3.2 завдання 1 (п. 3.5.2).

Таблиця 5.1

№ виміру	Товщина плівки SiO_2, (ангстрем)	Час окислювання, (хв.)	Температура окислювання, (°С)
1	145	42	695
2	138	45	705
3	157	71	727
4	152	57	701
5	180	86	713
6	135	59	694
7	146	57	691
8	152	55	672
9	149	67	717
10	170	71	719
11	161	71	707
12	156	47	706
13	161	71	701
14	156	78	676
15	158	81	711
16	117	41	673
17	151	56	682
18	113	52	688
19	166	67	724
20	138	37	685
21	149	47	680
22	151	66	712
23	174	97	724
24	138	51	684
25	157	67	701

Завдання 5.4. Описати можливості модуля Product-Moment and Partial Correlations - Quick, викликаного зі стартової панелі командою Statistics → Basic Statistics and Tables → Correlation matrices.

Завдання 5.5. Описати можливості вкладки Prob. & Scatterplots Tab модуля Descriptive Statistics.

6 ДИСПЕРСІЙНИЙ АНАЛІЗ

6.1 Основні положення

Дисперсійний аналіз використовується для перевірки гіпотези рівності середніх декількох вибірок, причому кожній з вибірок відповідає визначений комплекс зовнішніх умов. План дисперсійного аналізу припускає проведення експерименту, що дозволяє оцінити внесок окремих факторів у загальну дисперсію результуючої (вихідної) випадкової величини в тих випадках, коли побудувати регресійну модель неможливо. Наприклад, фактори (вхідні перемінні) є дискретними чи якісними. Дисперсійний аналіз складається в розкладанні загальної дисперсії результуючої випадкової величини на складові, котрі визначаються дією окремих факторів. Чим більше складова, тим більше вплив фактора, що обумовлює цю складову.

Основні передумови застосування дисперсійного аналізу:

- вибірки беруться з генеральних сукупностей з нормальним законом розподілу;

-дисперсія, обумовлена неврахованими неконтрольованими причинами, однорідна.

Дослідження впливу одного чи декількох факторів проводиться однофакторним чи багатофакторним дисперсійним аналізом [1, 9].

6.2 Однофакторний дисперсійний аналіз

Однофакторний дисперсійний аналіз застосовують у випадку, коли припускають, що головною причиною мінливості досліджуваної випадкової величини є один фактор. Відповідно до цього фактора загальна вибірка розбивається на кілька вибірок і оцінюється однорідність середніх отриманих вибірок.

При проведенні однофакторного дисперсійного аналізу сумарний розкид значень досліджуваної випадкової величини розкладається на розкид усередині вибірок і розкид між середніми вибірок.

Для характеристики сумарного розкиду використовується сума

квадратів Q відхилень значень, що *спостерігаються, хij* навколо загального $\bar{x}$ середніх :

$$Q = \sum_i \sum_j (x_{ij} - \bar{x})^2 \qquad (6.1)$$

де: $i=1,2, \ldots m,$ m - число вибірок;

$j=1,2, \ldots n_i,$ n_i - обсяг i-ої вибірки.

Розкид усередині вибірок і між середніми вибірок характеризується, відповідно, сумами квадратів Q_A і Q_W, обумовленим формулам:

$$Q_A = \sum_i n_i (x_i - \bar{x})^2; \qquad (6.2)$$

$$Q_W = \sum_i \sum_j (x_{ij} - x_i)^2; \qquad (6.3)$$

Для перевірки гіпотези про рівність середніх x_i використовується критерій Фішера

$$F = \frac{S_A^2}{S_B^2} \; ; \qquad (6.4)$$

$$S_A^2 = Q_A /(m-1); \qquad S_B^2 = Q_B /(n-1). \qquad (6.5)$$

Формули, використовувані при однофакторному аналізі, зручно звести в таблицю дисперсійного аналізу. При рівних обсягу вибірок $n_1=n_2=\ldots=n$ ці формули приведені в таблиці 6.1.

Прикладом використання однофакторного дисперсійного аналізу є оцінка стабільності технологічного процесу. Нехай стан технологічного процесу оцінюється нормально - розподіленою випадковою величиною x і спостереження за ходом процесу за допомогою узятих вибірок виробляються через визначені проміжки часу. Якщо для процесу характерна сталість вибіркових дисперсій, то стабільність техпроцесу може оцінюватися величиною внеску в загальний розкид значень x розкиду, обумовленого неоднорідністю вибіркових середніх.

Для проведення однорідності середніх використовується метод

однофакторного дисперсійного аналізу.

Таблиця дисперсійного аналізу при однофакторній класифікації

Таблиця 6.1

Джерело дисперсії	Сума квадратів	Число ст. вільності	Середній квадрат	F-критерій
Між групами	$Q_A=\sum_i n_i(x_i-\bar{x})^2;$	*m-1*	$S_A^2=Q_A/(m-1)$	$F=\frac{S_A^2}{S_B^2}$
Усередині груп	$Q_W=\sum_i\sum_j (x_{ij}-x_i)^2;$	*n-m*	$S_B^2=Q_B/(n-1)$	
Повна сума квадратів	$Q=\sum_i\sum_j (x_{ij}-\bar{x})^2;$	*n-1*	$S^2=Q/(n-1);$	

Позначимо: n - обсяг вибірки, m - кількість вибірок, x_i, S^2_i- вибіркової середні і дисперсії.

Якщо обчислене значення F- критерію чи менше дорівнює табличному значенню, то з прийнятою довірчою імовірністю P робиться висновок про несуттєвий вплив на загальну дисперсію розкиду вибіркових середніх.

Якщо обчислене значення F- критерію більше табличного, то вибіркові середні істотно розрізняються. Тоді прийнявши повну суму квадратів відхилень x_{ij} від загального середніх за 100%, можна оцінити внесок розкиду середніх у загальний розкид, виходячи з того, що $Q=Q_A+Q_W$.

6.3 Двофакторний дисперсійний аналіз

Двофакторний дисперсійний аналіз застосовується у випадку, коли передбачається, що мінливість значень досліджуваної випадкової, що спостерігаються, *величини* x обумовлена двома факторами. Умовно позначимо ці фактори A і B. Спостереження x_{ijk} зв'язуються зі значеннями факторів A_i і B_j, індекс *к* - означає рівнобіжні спостереження. Розрізняються два різновиди

двофакторного дисперсійного аналізу. Це дисперсійний аналіз з пересічними факторами і дисперсійний аналіз з угрупованням.

Якщо класифікація спостережень по фактору В має той самий зміст для всіх значень фактора А, то це класифікація для двофакторного аналізу з пересічними факторами. Приклад двофакторної класифікації з пересічними факторами - нехай вироби в процесі виготовлення проходять через установки типу А (фактор А) і потім через установки типу В (фактор В), і вироби, що пройшли через установку типу А попадають на всі наявні установки типу В.

Якщо спостереження представлені таким чином, що значення, що відповідають класифікації В згруповані до визначеного значення класифікації А, те це класифікація двофакторного аналізу з угрупованням. Приклад двофакторної класифікації з угрупованням - уся продукція розбивається на партії (класифікація А), кожна партія складається з пластин (класифікація В) на кожній пластині виробляється ряд рівнобіжних вимірів.

Сумарний розкид Q значень досліджуваної випадкової величини обумовлений наступними причинами:

- фактором А;
- фактором У;
- взаємодією АВ;
- неврахованими факторами.

Якщо для оцінки сумарного розкиду використовується сума квадратів відхилень, то

$$Q=Q_A+Q_B+Q_{AB}+Q_W; \qquad (6.6)$$

де: $Q_A+Q_B+Q_{AB}+Q_W$ - суми квадратів відхилень, обумовлених дією факторів $A,У$, взаємодією AB і неврахованими факторами W.

Оцінка впливу на загальний розкид факторів, яким відповідає класифікація А,В,В(А),(А усередині В), взаємодія АВ відбувається за допомогою критерію Фішера. Формули для розрахунку F- критерію приведені в таблицях 6.2, 6.3. Обчислення середніх здійснюється за наступними формулами:

$$
\begin{aligned}
x &= \frac{1}{n\,n_A n_B}\sum_i\sum_j\sum_k x_{ijk};\\
x_{ij.} &= \frac{1}{n}\sum_k x_{ijk};\\
x_{i..} &= \frac{1}{n\,n_B}\sum_j\sum_k x_{ijk};\\
x_{.j.} &= \frac{1}{n\,n_A}\sum_i\sum_k x_{ijk};
\end{aligned}
\qquad (6.7)
$$

де: x_{ijk} - *к*-е рівнобіжне спостереження, що належить *i*-му рівню класу А і *j*-му рівню класу В, при $k=1,n;\ i=1,n;\ j=1,n$.

Отримані обчислені значення F- критерію порівнюються з критичним значенням F-розподілу, узятих для числа ступенів вільності, що відповідає середньому квадрату. Наприклад, $F^{(A)}=S^2_A/S^2_W$ відповідають числа ступенів вільності $k_1=n_A-1,\ k_2=n_An_B(n-1)$.

Висновок про істотний вплив на загальну дисперсію досліджуваного виду класифікації робиться для рівня значимості α, якщо $F>F_{v1,v2}$ де: $F_{v1,v2}$ – процентна точка (критичне значення) розподілу Фішера.

6.4 Комп'ютерні технології дисперсійного аналізу

ВИКОРИСТАННЯ ПАКЕТА АНАЛІЗУ **MICROSOFT EXCEL.** Інструменти дисперсійного аналізу доступні через команду *Аналіз даних* меню *Сервіс*. Існує кілька видів дисперсійного аналізу. Необхідний варіант вибирається з урахуванням числа факторів і наявних вибірок з генеральної сукупності.

Однофакторний дисперсійний аналіз. Однофакторний дисперсійний аналіз використовується для перевірки гіпотези про подібність середніх значень двох чи більш вибірок, що належать однієї і тієї ж генеральної сукупності.

Таблиця двофакторного дисперсійного аналізу
при класифікація з пересічними факторами

Таблиця 6.2

Джерело дисперсії	Сума квадратів	Число ст. вільності	Середній квадрат	F-критерій
Класифікація А	$Q_A = n_B n \sum_{i=1}^{n_A} (x_{i..} - \bar{x})^2$	n_A-1	$S_A^2 = Q_A/(n_A - 1);$	$F^A = \frac{S_A^2}{S_W^2}$
Класифікація В	$Q_B = n_A n \sum_{j=1}^{n_B} (x_{.j.} - \bar{x})^2$	n_B-1		$F^B = \frac{S_B^2}{S_W^2}$
Взаємодія	$Q_{AB} = n \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (x_{ij.} + \bar{x} - x_{i..} - x_{.j.})^2$	(n_A-1)(n_B-1)	$S_B^2 = Q_B/(n_B - 1);$	$F^{AB} = \frac{S_{AB}^2}{S_W^2}$
Усередині груп	$Q_W = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sum_{k=1}^{n} (x_{ijk} - x_{ij.})^2$	n_A n_B (n-1)	$S_W^2 = \frac{Q_W}{n_A n_B (n-1)};$	
Повна сума квадратів	$Q = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sum_{k=1}^{n} (x_{ijk} - \bar{x})^2$	n_A n_B n-1	$S^2 = \frac{Q}{n_A n_B n - 1};$	

Таблиця двофакторного дисперсійного аналізу при класифікації з групуванням

Таблиця 6.3

Джерело дисперсії	Сума квадратів	Число ст. вільності	Середній квадрат	F-критерій
Класифікація А	$Q_A = n_B n \sum_{i=1}^{n_A} (x_{i..} - \bar{x})^2$	n_A-1	$S_A^2 = Q_A/(n_A - 1);$	$F^A = \frac{S_A^2}{S_W^2}$
Класифікація В усередині А	$Q_{B(A)} = n \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} (x_{ij.} - x_{i..})^2$	$n_A(n_B$-1)	$S_{B(A)}^2 = \frac{Q_{B(A)}}{n_A(n_B - 1)};$	$F^{B(A)} = \frac{S_{B(A)}^2}{S_W^2}$
Усередині груп	$Q_W = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sum_{k=1}^{n} (x_{ijk} - x_{ij.})^2$	n_A n_B (n-1)	$S_W^2 = \frac{Q_W}{n_A n_B (n-1)};$	
Повна сума квадратів	$Q = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sum_{k=1}^{n} (x_{ijk} - \bar{x})^2$	n_A n_B n-1	$S^2 = \frac{Q}{n_A n_B n - 1};$	

Двофакторний дисперсійний аналіз з повтореннями. Являє собою більш складний варіант дисперсійного аналізу з декількома вибірками для кожної групи даних, називаний також дисперсійним аналізом при класифікації з групуванням.

Двофакторний дисперсійний аналіз без повторення. Являє собою двофакторний аналіз дисперсії, що не включає більш однієї вибірки на групу, називаний також дисперсійним аналізом при класифікація з пересічними факторами.

ПРИКЛАД 6.1. Поставлена задача - оцінити стабільність операції "вплавлення" технологічного процесу виготовлення транзисторів середньої потужності, контрольованої по параметру α_f - коефіцієнту підсилення по струму на частоті f. Для дослідження після "вплавлення" узято 10 вибірок по 30 блоків із транзисторною структурою. Результати вимірів параметра α_f на блоках, приведені в таблиці 6.4.

Рішення. Пропонується використати метод однофакторного дисперсійного аналізу.

Позначимо x параметр α_f, яким оцінюють стабільність операції "вплавлення".

В результаті попереднього статистичного аналізу встановлено, що x – це нормально розподілена випадкова величина. Спостереження за ходом технологічного процесу здійснюється за допомогою вибірок, що беруться через визначені постійні проміжки часу. Стабільність техпроцесу може оцінюватися величиною внеску в загальний розкид значень x розкиду, обумовленого неоднорідністю вибіркових середніх. Гіпотеза про однорідність середніх і перевіряється методом однофакторного дисперсійного аналізу.

Дані вибіркових спостережень з таблиці 6.4. оброблені за допомогою інструмента «однофакторний дисперсійний аналіз» з пакета аналізу MICROSOFT EXCEL. Результати обробки приведені в таблицях 6.5, 6.6.

Таблиця 6.4

№	Номер вибірки									
	1	2	3	4	5	6	7	8	9	10
1	0,82	0,79	0,80	0,80	0,79	0,83	0,84	0,81	0,84	0,79
2	0,82	0,79	0,78	0,79	0,77	0,82	0,82	0,76	0,83	0,82
3	0,81	0,81	0,80	0,78	0,81	0,85	0,86	0,81	0,84	0,80
4	0,83	0,84	0,79	0,75	0,75	0,85	0,86	0,79	0,83	0,76
5	0,76	0,82	0,76	0,82	0,77	0,82	0,83	0,83	0,80	0,80
6	0,76	0,78	0,83	0,78	0,76	0,84	0,86	0,84	0,84	0,80
7	0,82	0,84	0,80	0,83	0,73	0,76	0,86	0,77	0,84	0,81
8	0,78	0,80	0,81	0,79	0,72	0,83	0,86	0,81	0,83	0,80
9	0,82	0,83	0,81	0,76	0,71	0,82	0,86	0,80	0,83	0,76
10	0,83	0,79	0,80	0,80	0,79	0,80	0,86	0,80	0,83	0,82
11	0,80	0,81	0,82	0,73	0,80	0,79	0,82	0,79	0,82	0,80
12	0,79	0,85	0,82	0,79	0,75	0,81	0,85	0,81	0,84	0,80
13	0,79	0,77	0,79	0,81	0,76	0,86	0,83	0,78	0,83	0,79
14	0,80	0,83	0,83	0,80	0,77	0,85	0,85	0,82	0,82	0,82
15	0,80	0,83	0,81	0,77	0,76	0,83	0,87	0,83	0,82	0,79
16	0,79	0,82	0,79	0,79	0,79	0,81	0,85	0,80	0,86	0,79
17	0,79	0,80	0,79	0,79	0,72	0,81	0,88	0,79	0,85	0,78
18	0,81	0,83	0,79	0,79	0,80	0,84	0,80	0,82	0,84	0,82
19	0,79	0,78	0,79	0,77	0,75	0,77	0,83	0,80	0,81	0,78
20	0,82	0,79	0,81	0,81	0,78	0,82	0,83	0,77	0,86	0,79
21	0,82	0,85	0,82	0,78	0,82	0,84	0,83	0,79	0,86	0,82
22	0,78	0,80	0,82	0,78	0,78	0,82	0,81	0,83	0,85	0,81
23	0,80	0,82	0,79	0,79	0,76	0,82	0,87	0,81	0,83	0,79
24	0,78	0,80	0,78	0,81	0,74	0,81	0,87	0,78	0,86	0,82
25	0,80	0,80	0,80	0,78	0,77	0,79	0,82	0,78	0,82	0,80
26	0,79	0,81	0,78	0,77	0,80	0,80	0,83	0,83	0,85	0,80
27	0,81	0,83	0,82	0,75	0,75	0,83	0,81	0,81	0,83	0,83
28	0,79	0,81	0,77	0,80	0,78	0,81	0,84	0,77	0,83	0,79
29	0,78	0,80	0,83	0,75	0,79	0,83	0,80	0,81	0,85	0,82
30	0,77	0,84	0,80	0,78	0,79	0,78	0,86	0,81	0,80	0,76

Таблиця 6.5

Результати				
Групи	*Рахунок*	*Сума*	*Середнє*	*Дисперсія*
Стовпчик 1	30	23,90076	0,79669	0,00047
Стовпчик 2	30	24,25154	0,80838	0,00043
Стовпчик 3	30	23,83703	0,79457	0,00053
Стовпчик 4	30	23,35077	0,77836	0,00048
Стовпчик 5	30	22,92961	0,76432	0,00071
Стовпчик 6	30	24,69176	0,82306	0,00060
Стовпчик 7	30	25,31971	0,84399	0,00046
Стовпчик 8	30	24,14639	0,80488	0,00047
Стовпчик 9	30	24,69011	0,82300	0,00027
Стовпчик 10	30	24,03924	0,80131	0,00038

Таблиця 6.6

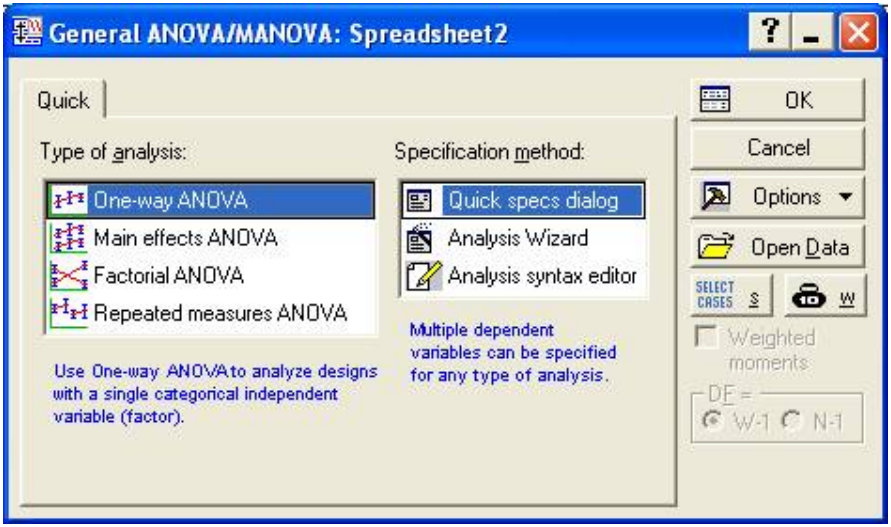
Дисперсійний аналіз						
Джерело варіації	*SS*	*df*	*MS*	*F*	*P-Значення*	*F критичне*
Між групами	0,141749	9	0,0157498	32,79	1,9413E-39	1,912
Усередині груп	0,139283	290	0,0004803			
Усього	0,281031	299				

Обчислене значення F- критерію дорівнює 32,79 і воно набагато більше критичного, рівного 1,912, отже те з прийняті рівнем значимості робиться висновок про істотний вплив на загальну дисперсію розкиду вибіркових середніх. Процес нестабільний. Можна оцінити внесок розкиду середніх у загальний розкид, виходячи з того, що $Q=Q_A+Q_W$. Приймемо повну суму квадратів відхилень x_{ij} від загального середніх за 100%, внесок Q_A і Q_W складають приблизно по 50%.

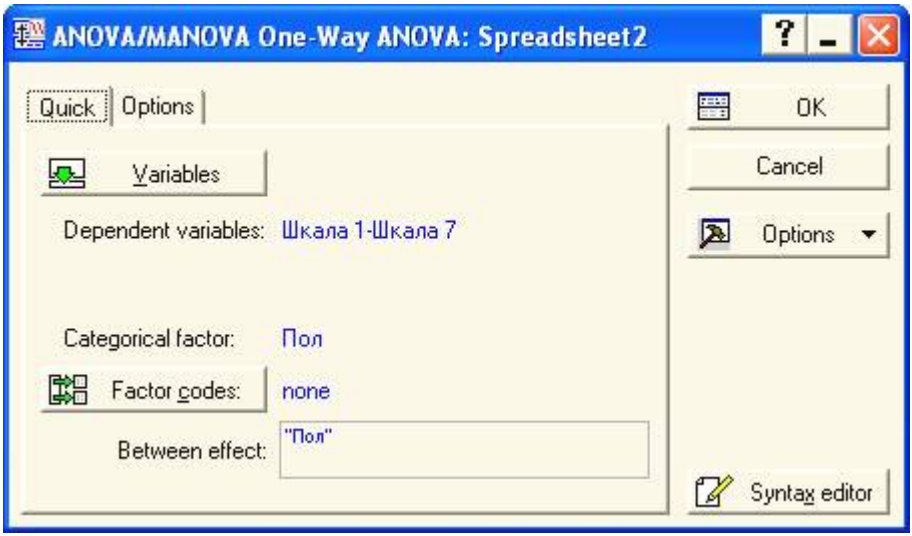
ДИСПЕРСІЙНИЙ АНАЛІЗ У ПАКЕТІ STATISTICA

Крок 1. Ввести або імпортувати (наприклад, з Excel) вихідні дані в робочу книгу (Workbook) системи STATISTICA, виділити їх, увести назву таблиці вихідних даних і назви перемінних.

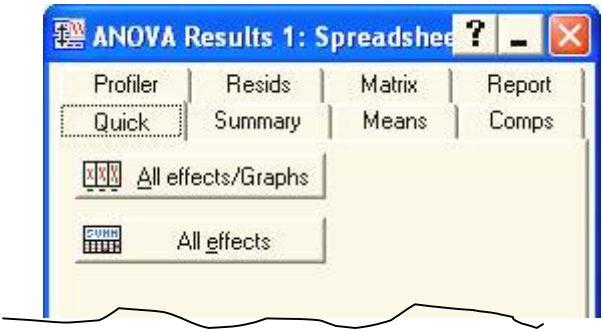
Крок 2. Клацнути по кнопці Start menu ..., розташованій в лівому нижньому куті вікна додатка й у меню, що з'явиться, вибрати Statistics → ANOVA. Після чого з'являється вікно General ANOVA/MANOVA - Quick Tab



Крок 3. Після вибору рядка One-way ANOVA з'являється відповідне вікно, за допомогою якого вибираються залежні (dependent variable(s) і перемінні , що групують (categorical predictor variables).



Крок 4. Потім клацнути по кнопці OK і з'являється вікно ANOVA Results - Quick,



де вибрати кнопку All effects, після чого і з'являється вікно Multivariate Tests of Significance з результатами аналізу.

Приклад 6.2. Як приклад використовуємо вихідні дані соціологічного опитування рекламного дослідження з файлу Adstudy.sta, що знаходиться в папці Examples і поставляється разом із системою STATISTICA. Дані являють собою оцінки респондентів (чоловіків і жінок) якості двох рекламних роликів – ADVERT:1=Coce,

ADVERT:2=Pepsi. Привабливість роликів оцінювалася за різними шкалами, обмежимося 9-ма шкалами (з Measur 1 по Measur 9). У кожній зі шкал респонденти виставляли бали від 0 до 7.

Data: Spreadsheet2* (9v by 50c)

	Результаты социологического исследования качества рекламных роликов								
	1	2	3	4	5	6	7	8	9
	Пол	Ролик	Шкала 1	Шкала 2	Шкала 3	Шкала 4	Шкала 5	Шкала 6	Шкала 7
1	Мужчина	PEPSI	9	1	6	8	1	2	1
2	Мужчина	COKE	6	7	1	8	0	0	6
3	Женщина	COKE	9	8	2	9	8	8	0
4	Мужчина	PEPSI	7	9	0	5	9	9	6
5	Мужчина	PEPSI	7	1	6	2	8	9	6
6	Женщина	COKE	6	0	0	8	3	1	0
7	Женщина	COKE	7	4	3	2	5	7	1
8	Мужчина	PEPSI	9	9	2	6	6	8	7
9	Женщина	PEPSI	7	8	2	3	6	9	1
10	Мужчина	PEPSI	6	6	2	8	3	6	4
11	Женщина	PEPSI	4	6	6	5	6	8	7
12	Мужчина	COKE	7	3	3	7	0	6	5
13	Мужчина	PEPSI	6	2	3	1	8	1	4
14	Мужчина	COKE	7	2	4	8	1	2	6
15	Женщина	PEPSI	6	2	7	5	7	2	5
16	Женщина	PEPSI	3	2	5	4	4	4	3
17	Мужчина	COKE	2	9	9	3	1	4	5
18	Женщина	PEPSI	1	0	7	5	2	4	2

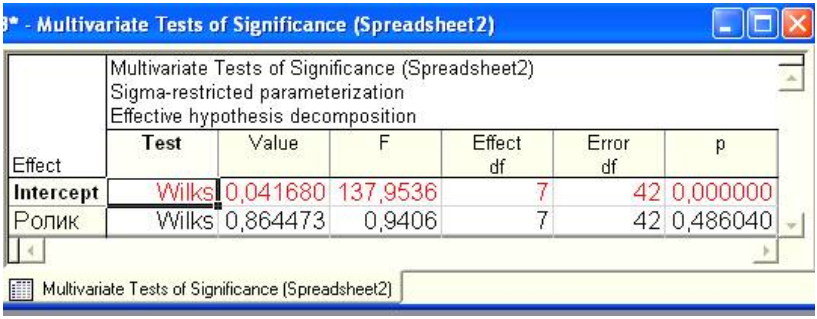
Провівши дисперсійний аналіз відповідно до кроків 1 – 4, одержимо вікно Multivariate Tests of Significance з результатами, що відбивають вплив факторів «стать» і «ролик».

1* - Multivariate Tests of Significance (Spreadsheet2)

Multivariate Tests of Significance (Spreadsheet2)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0,040497	142,1591	7	42	0,000000
Пол	Wilks	0,741519	2,0915	7	42	0,065647

Multivariate Tests of Significance (Spreadsheet2)



6.5 Контрольні питання і завдання

6.5.1 Питання

1 Призначення дисперсійного аналізу.
2 Основні передумови застосування дисперсійного аналізу.
3 Однофакторний дисперсійний аналіз.
4 Використання однофакторного дисперсійного аналізу (приклад).
5 Двофакторний дисперсійний аналіз, його різновиду.
6 Оцінка сумарного розкиду досліджуваного параметра при двофакторній класифікації.
7 Двофакторний дисперсійний аналіз при класифікації з пересічними факторами.
8 Двофакторний дисперсійний аналіз при класифікації з угрупованням.

6.5.2 Завдання

Завдання 1. ПРИ виготовленні ІМС можуть виникати сховані технологічні дефекти. Для діагностики потенційно ненадійних виробів використовуються діагностичні методи контролю. Робиться припущення, що забруднення кристала можна діагностувати за допомогою перевірки струмів витікання $I_{вит}$. В дослідних партіях вимірювались $I_{вит}$ на кожній мікросхемі. Потім зроблені додаткові лабораторні дослідження, що дозволили розбракувати партію на дві групи - з «чистими» і - з «забрудненими» кристалами. Результати по 5

партіям приведені в таблиці 6.7, «чисті» кристали закодовані - «год», «забруднені» - «з».

Необхідно. Перевірити припущення про можливість діагностування забруднення кристала за допомогою перевірки струмів витікання $I_{вит}$ за допомогою однофакторного дисперсійного аналізу.

Результати перевірки струмів витоку $I_{вит}$

Таблиця 6.7

№ п/п	1-а партія		2-а партія		3-ая партія		4-а партія		5-а партія	
	«год»	«з»	«год»	«з»	«год»	«з»	«год»	«з»	«год»	«з»
1	1,45	1,83	1,88	2,07	1,83	2,03	1,51	2,22	1,90	2,15
2	1,63	1,92	1,39	1,85	1,60	2,14	1,56	1,86	1,41	1,84
3	1,71	1,98	1,43	1,92	1,92	2,08	1,92	1,95	1,59	2,06
4	1,54	2,03	1,45	1,93	1,66	2,22	1,64	1,95	1,72	1,92
5	1,84	2,10	1,38	1,92	1,85	1,87	1,60	1,96	1,81	1,90
6	1,75	1,91	1,87	1,62	1,48	2,22	1,56	2,08	1,80	2,02
7	1,80	1,92	1,78	2,09	1,78	2,14	1,59	1,96	1,52	1,93
8	1,78	2,07	1,85	1,86	1,35	2,12	1,75	1,95	1,92	2,13
9	1,55	1,78	1,84	1,81	1,69	1,90	1,50	2,08	1,82	1,96
10	1,44	1,97	1,50	2,04	1,78	1,84	1,75	1,99	1,70	2,03
11	1,81	2,05	1,71	2,00	1,77	1,91	1,74	2,06	1,53	1,54
12	1,78	1,99	1,72	2,22	1,46	1,88	1,56	2,15	1,94	1,85
13	1,76	2,07	1,74	1,87	1,48	2,06	1,81	2,09	1,93	1,91
14	1,73	1,94	1,60	2,20	1,78	1,96	1,52	2,11	1,88	2,03
15	1,80	1,96	1,56	2,14	1,67	2,07	1,53	2,19	1,60	1,96
16	1,56	1,88	1,63	1,98	2,12	2,01	1,63	2,11	1,69	2,03
17	1,85	1,92	1,36	2,08	1,81	1,88	1,65	2,05	1,65	1,78
18	1,78	1,95	1,95	1,84	1,62	1,78	1,77	2,17	1,82	1,99
19	1,69	2,09	1,87	1,99	1,64	2,30	1,56	2,06	1,92	1,83
20	1,84	2,14	1,71	1,93	1,71	1,89	2,10	2,01	1,63	2,12

Завдання 2. Контроль дифузійних операцій при виготовленні ІМС здійснюється за допомогою вимірювання коефіцієнта підсилення транзисторної структури β_N, вимірюваного на тестових комірках. В табл.6.8 наведені результати вимірів у вибірках, взятих на протязі зміни.

Необхідно. Оцінити стабільність дифузійних операцій за зміну.

Результати вимірів β_N на тестових комірках

Таблиця 6.8

№ зм.	Номер вибірки							
	1	2	3	4	5	6	7	8
1	76,0	73,0	56,2	56,6	81,8	82,6	58,9	82,8
2	27,4	50,4	40,6	73,4	73,9	79,7	102,3	72,5
3	80,9	80,8	104,4	77,6	84,5	29,7	46,8	71,9
4	41,9	79,1	75,6	73,8	90,3	32,5	43,1	102,5
5	47,9	48,7	138,9	104,1	95,2	74,6	77,7	106,7
6	58,2	45,0	50,3	34,3	80,1	59,7	58,7	81,3
7	80,2	94,4	44,1	83,9	67,3	24,8	75,2	88,0
8	88,0	73,8	60,8	62,1	59,9	57,2	85,9	66,9
9	82,6	70,7	91,8	80,8	99,7	61,4	59,1	65,1
10	54,4	75,1	118,8	40,1	79,2	78,2	61,9	51,1

Завдання 3**.** Сильно впливає на відсоток виходу годних при виробництві сплавних потужних транзисторів типу П210 операція “травлення”.

Основними виробничо-технологічними факторами цієї операції є робітники-травильники й установки травлення (режими травлення в даному технологічному процесі зафіксовані). Умовно назвемо ці фактори “виконавець” і “установка”. Задача полягає в оцінці впливу якості праці робітників-травильників і однорідності основного технологічного устаткування, установок травлення, на відсоток виходу годних після операції “травлення”. На основі інформації, супровідних листів партій, результати спостережень були класифіковані відповідно до вимог проведення двофакторного дисперсійного аналізу з пересічними факторами. Для дослідження узято 2 “установки”, 10 “виконавців” і проведено по 7 рівнобіжних спостережень. Результати спостережень зведені в таблиці 6.9.

Результати спостережень за виходом годних

після операції "Травлення"

Таблиця 6.9

Уст.	Травильник	Паралельні спостереження						
		1	2	3	4	5	6	7
1	1	0.968	0.989	0.989	0.989	0.984	0.984	0.995
	2	0.989	0.995	0.995	0.995	0.989	0.995	0.989
	3	0.979	0.989	0.984	0.99	0.985	0.995	0.945
	4	0.974	0.989	0.956	0.937	0.974	0.974	0.989
	5	0.974	0.948	0.959	0.933	0.99	0.947	0.964
	6	0.939	0.947	0.914	0.974	0.979	0.984	0.984
	7	0.974	0.938	0.979	0.984	0.989	0.969	0.979
	8	0.995	0.995	0.969	0.989	0.995	0.995	0.995
	9	0.995	0.984	0.984	0.979	0.969	0.984	0.979
	10	0.984	0.984	0.995	0.995	0.995	0.995	0.995
2	1	0.989	0.995	0.984	0.99	0.984	0.984	0.995
	2	0.995	0.984	0.99	0.995	0.99	0.989	0.989
	3	0.985	0.984	0.995	0.995	0.984	0.984	0.995
	4	0.995	0.989	0.989	0.989	0.995	0.99	0.989
	5	0.969	0.954	0.963	0.963	0.918	0.967	0.942
	6	0.964	0.969	0.968	0.968	0.989	0.995	0.989
	7	0.973	0.974	0.99	0.99	0.908	0.979	0.995
	8	0.984	0.995	0.973	0.973	0.984	0.968	0.989
	9	0.984	0.995	0.989	0.989	0.942	0.974	0.989
	10	0.979	0.979	0.979	0.979	0.995	0.984	0.984

Необхідно: виділити з двох факторів - “виконавець” і “установка” той, що найбільш істотно впливає на відсоток виходу годних.

7 ПЛАНУВАННЯ ЕКСПЕРИМЕНТУ

7.1 Основні положення

Під математичною теорією планування експерименту розуміється наука про способи складання ощадливих експериментальних планів, що одночасно дозволяють витягати найбільшу кількість інформації про об'єкт, про способи проведення експерименту, обробки експериментального даних, використання отриманих результатів для оптимізації виробничих процесів. Математичний апарат теорії експерименту побудований на сполученні методів математичної статистики і рішення екстремальних задач [5].

У практиці інженерних досліджень способи накопичення вихідного статистичного матеріалу умовно розділяються на два основних: пасивні експерименти й активний експеримент.

Пасивний експеримент полягає в спостереженні і реєстрації вхідних і вихідних перемінних процесу в режимі звичайної роботи досліджуваного об'єкта, у спостереженні за природними довільними змінами всіх технологічних перемінних без активного втручання дослідника в хід технологічного процесу, без внесення в нього навмисних втручань.

Активний експеримент реалізується за допомогою планування експерименту і полягає в реєстрації перемінних процесу після внесення в досліджуваний процес навмисних втручань технологічних факторів. Він ведеться за спеціальною програмою, називаною матрицею планування, де передбачаються потрібні досліднику діапазони варіювання керованих технологічних факторів. В результаті обробки вихідних даних при активному експерименті повинна бути отримана математична модель процесу

$$y = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + \sum a_{ii} x_{ii}^2 + \varepsilon \qquad (7.1)$$

де y - значення вихідного параметра;

a_i , a_{ij} , a_{ii} ,- невідомі оцінки коефіцієнтів, що потрібно знайти; x_i - перемінні на вході об'єкта; ε - помилка.

Основними напрямками при плануванні експериментів є повний факторний експеримент (ПФЕ) і центральне композиційне планування (ЦКП).

7.2 Повний факторний експеримент

При повному факторному експерименті безпосередньому проведенню експерименту передує:

- вибір факторів і параметра оптимізації;
- вибір координат базової точки і рівнів варіювання;
- складання матриці планування.

Вибір факторів, що підлягають вивченню і включенню в рівняння, повинен задовольняти наступним основним вимогам - фактори повинні бути керовані, функціонально незалежні, сумісні й у розгляд повинні бути включені усі відомі фактори.

Параметр оптимізації, тобто вихідну перемінну, котру часто називають цільовою функцією, що підлягає оптимізації, вибирають на підставі наступних умов.

Цільова функція повинна носити узагальнений характер, тобто відбивати властивості і якості процесу в цілому, а не тільки однієї його сторони, повинна нести в собі істотну інформацію про якість процесу, повинна бути можливість досить точно виміряти цільову функцію.

Координата базової точки ($X_{1,0}; X_{2,0};\ldots;X_{n,0}$), яку називають також центром експерименту, вибирають з наступних міркувань.

Центр експерименту часто вибирають у точці звичайного, номінального ведення процесу.

Базова точка може бути обрана в центрі області обмежень, якщо такі є і якщо немає іншої інформації. Якщо є якісь припущення про місце екстремума або в процесі попередніх досліджень отримана деяка точка з кращим значенням цільової функції, є сенс вибрати центр експерименту поблизу передбачуваного оптимуму - очевидно, це прискорить процес оптимізації.

Вибір рівнів (ступенів) варіювання λ повинен відповідати наступним умовам.

Якщо на вхідні і вихідні перемінні накладаються обмеження, то при варіюванні вхідних перемінних не слід виходити з установлених меж

$$\left.\begin{aligned} X_{i,0} + \lambda_i &\le X_{i\max} \\ X_{i,0} - \lambda_i &\ge X_{i\min} \end{aligned}\right\}. \tag{7.2}$$

Ступінь варіювання λ_i. повинен істотно перервати погрішність вимірювання по X_i. Ступені варіювання λ_i не повинні бути занадто великими, інакше математичний опис поверхні відгуку виявиться занадто грубим, ряд важливих його деталей залишиться непоміченим.

Складання матриці планування. Для одержання ортогональної матриці планування (МП) усі фактори повинні бути пронормовані за формулою

$$Z_i = \|X_i\| = \frac{X_i - X_{i,0}}{л_i}, \tag{7.3}$$

де Z_i.- нормоване значення фактора; $X_{i,0}$ - базове початкове значення фактора; λ_i- ступінь варіювання.

Нормовані значення факторів для базового $Z_{i,0}$, нижнього $Z_{i,H}$ і верхнього $Z_{i,Y}$ рівнів варіювання будуть рівні

$$Z_{i,0} = \frac{X_{i,0} - X_{i,0}}{л_i} = 0; \tag{7.4}$$

$$Z_{i,н} = \frac{X_{i,н} - X_{i,0}}{л_i} = \frac{X_{i,0} - л_i - X_{i,0}}{л_i} = -1; \tag{7.5}$$

$$Z_{i,в} = \frac{X_{i,в} - X_{i,0}}{л_i} = \frac{X_{i,0} - л_i - X_{i,0}}{л_i} = +1; \tag{7.6}$$

де $\left.\begin{aligned} X_{i,н} &= X_{i,0} - л_i \\ X_{i,в} &= X_{i,0} + л_i \end{aligned}\right\}$.

Складемо план експерименту у вигляді МП для випадку, коли здійснюються спробні досліди при активному експерименті з числом факторів n=2 і варіювання здійснюється біля базової точки M_o на двох рівнях для нормованих значень факторів Z_1, Z_2 (табл.7.1). Для того щоб легше перебрати всі можливі комбінації рівнів, застосовують наступне простої правило: у стовпці Z_0 усі рівні тільки +1; у другому стовпці для Z_1 знаки рівнів чергуються в кожнім рядку вектор - стовпці; у третьому векторі-стовпці Z_2 частота зміни знаків рівнів зменшується в 2 рази; частота зміни знаків для Z_3 зменшується в 4 рази і т.д. При складанні МП у матриці часто замість +1 чи -1 ставлять просто знак "+" чи "-"

Таблиця 7.1

g	МП			Взаємодія і члени вищих порядків							Паралельн. зміна вих. перемінних		
	Z_0	Z_1	Z_2	Z_1Z_2	Z_1^2	Z_2^2	Z_1^3	Z_2^3	$Z_1Z_2^2$	$Z_2Z_1^2$	Y_1	Y_2	Y_3
1	+	-	-	+	+	+	-	-	-	-	43	35	39
2	+	+	-	-	+	+	+	-	+	-	90	86	88
3	+	-	+	-	+	+	-	+	-	+	10	16	13
4	+	+	+	+	+	+	+	+	+	+	56	50	53

Власне МП містить у собі тільки вектори-стовпці для незалежних (лінійних) нормованих факторів $Z_i\,(i=\overline{1,n})$. Інші вектори-стовпці потрібні тільки для розрахунку коефіцієнтів при відповідних парних взаємодіях, членах вищих порядків (якщо їх можна незалежно обчислити) і вільного члена a_0, представником якого в МП є фіктивна перемінна Z_0.

У МП для n=2 (див.табл. 7.1) наведені всі можливі сполучення рівнів, їх виявилося чотири. У загальному випадку, коли кількість факторів дорівнює n, а рівнів варіювання k, число всіх можливих сполучень визначається формулою

$$N = k^n$$

При дворівневому плануванні (k=2), мабуть,

$$N = 2^n$$

Якщо реалізуються всі 2^n сполучень рівнів варіювання факторів, то такий експеримент називають повним факторним експериментом (ПФЕ), а МП, у якій передбачені всі *2* сполучень рівнів, - повною матрицею планування (ПМП). Однак на початкових етапах дослідження прагнуть обмежитися тільки постановкою частини ПФЕ - так називаною "дробовою реплікою", тоді експеримент називають дробовим факторним експериментом (ДФЕ), а МП - дробовою матрицею планування (ДМП).

Після заміни X_i на нормовані Z_i за формулою (7.2) рівняння регресії (7.1) приймає вигляд:

$$\vartheta = a_o^* + \sum_{i=1}^{n} a_i^* Z_i + \sum_{\substack{i,l=1 \\ i \neq 1}}^{n} a_{i,l}^* Z_i Z_l + \varepsilon, \qquad (7.7)$$

де $a_0^*, a_i^*, a_{i,l}^*$ - оцінки коефіцієнтів нормованого рівняння регресії.

У рівнянні (7.7) на відміну від рівняння (7.1) відсутні квадратичні члени. Це пояснюється тим, що квадратичні коефіцієнти при плануванні на двох рівнях оцінити роздільно неможливо, тому що вони змішані з вільним членом.

Оскільки a_0^* - величина постійна і не залежна від інших факторів, уводять фіктивну перемінну Z_0, якій у всіх рядках вектор-стовпця Z_0, приписують рівні Z_0=+1. Це дозволяє обчислювати вільний член a_0^* також як і інші коефіцієнти рівняння (7.7).

МП (див.табл.7.1) для факторів Z_0, Z_1, Z_2 ортогональна, ознакою чого є рівність нулю скалярних добутків будь-яких двох вектор-стовпців МП;

$$\sum_{g=1}^{N} Z_{i,g} Z_{l,g} = 0, \qquad (7.8)$$

Крім умови ортогональності (7.8) МП для ПФЕ задовольняє ще двом важливим умовам.

$$\sum_{g=1}^{N} Z_{i,g}^2 = N; \qquad (i = \overline{0, N-1}); \qquad (7.9)$$

$$\sum_{g=1}^{N} Z_{i,g}^2 = 0; \qquad (i = \overline{1, N-1}); (i \neq 0); \qquad (7.10)$$

Рівність (7.9) називають умовою нормування: у МП передбачені тільки або нижні рівні Z=-1, або верхні Z=+1 і ніякі інші. Умова (7.10) виражає властивість симетричності МП: усі спробні експериментальні точки при ПФЕ розташовані симетрично щодо базової точки.

Необхідно забезпечити випадковість помилок, що накладаються на вхідні і вихідні перемінні. Однак, якщо проводити досвіди в тім порядку, у якому випливають рядка МП, то неминуче виникнуть систематичні погрішності, особливо для тих факторів X_i, що представлений останніми вектор - стовпцями МП. Справа в тім, що установка заданих МП нижніх $X_{i,н}$ чи верхніх $X_{i,y}$ рівнів варіювання не може вироблятися абсолютно точно, без помилок, і якщо та сама установка, з тією же помилкою у визначенні рівня не змінюється підряд у декількох досвідах (тобто в декількох рядках МП підряд), те виникає систематична помилка. Щоб залишити тільки випадкову помилку в установці рівнів факторів з нульовим математичним чеканням, усі досвіди виконують у випадковому порядку.

Рандомізацію дослідів, тобто визначення випадкового порядку проведення досвідів, виконують за допомогою таблиць випадкових чисел. У табл.7.2 приведений приклад вибору випадкових чисел m_1 і m_2 для проведення двох серій рівнобіжних експериментів. Для цього використана таблиця рівномірно розподілених випадкових чисел.

Таблиця 7.2

g	Випадкові числа		$S_g^2\{\vartheta\}$	Перевірка адекватності		
	m_2	m_2		$\vartheta_{p,g}$	$(\overline{\vartheta}_{н,g} - \overline{\vartheta}_{в,g})$	$(\overline{\vartheta}_{н,g} - \vartheta_{p,g})^2$
1	4	1	32	41,3	39-41,3	5,29
2	1	2	8	85,9	88-85,9	4,41
3	2	4	18	10,7	13-10,7	5,29
4	3	3	18	55,3	53-55,3	5,29

Перевірка відтворюваності дослідів полягає в оцінці однорідності дисперсій виходу, тобто дисперсій у крапках факторного

простору, обумовлених кожним рядком МП. Порядкові дисперсії обчислюють за формулою

$$S_g^2\{\vartheta\} = \frac{1}{\gamma - 1}\sum_{k=1}^{\gamma}(\vartheta_{g,k} - \overline{\vartheta}_g)^2; (g = \overline{1, N}), \tag{7.12}$$

де γ - число рівнобіжних досвідів; k- номер рівнобіжного досвіду в кожнім рядку МП; $\overline{\vartheta}_g$ - порядкове середнє значення виходу ϑ,

У табл.7.2 приведені результати обчислень незміщених оцінок порядкових дисперсій $S_g^2\{\vartheta\}$ за даними табл.2.1. Однорідність порядкових дисперсій перевіряють за G-критерієм Кохрана, для чого обчислюють дисперсійне відношення

$$G_{\max} = \frac{S_g^2\{\vartheta\}_{\max}}{\sum_{g=1}^{N} S_g^2\{\vartheta\}}, \tag{7.13}$$

де $S_g^2\{\vartheta\}_{\max}$ максимальна з усіх N порядкових дисперсій; $\sum_{g=1}^{N} S_g^2\{\vartheta\}$- сума порядкових дисперсій.

Критерій Кохрана, можна застосовувати для з'ясування однорідності любо групи дисперсій, однак обов'язкова умова для його застосування - однакові числа досвідів γ по який визначені порівнювані дисперсії. У противному випадку (при різних значеннях γ) варто застосовувати F-критерій Фішера, для чого варто вибрати для порівняння найбільшу $S_g^2\{\vartheta\}_{\max}$, і найменшу $S_g^2\{\vartheta\}_{\min}$ з N порівнюваних дисперсій.

Обчислене значення G_{max} (7.I3) порівнюють з табличним (критичним) значенням *G*-статистики, що вибирають з ститстичних таблиць [3] для прийнятого рівня значимості q і для чисел ступенів волі чисельника $f_{числ}$ і знаменника $f_{знам}$

$$\left.\begin{aligned} f_{числ} &= f_1 = \gamma - 1 \\ f_{знам} &= f_2 = N \end{aligned}\right\}, \qquad (7.14)$$

де f_1 - відповідає числу ступенів волі, по якому оцінюють максимальну з порядкових дисперсій $S_g^2\{\vartheta\}_{\max}$; f_2 - відповідає числу N порівнюваних дисперсій, причому f_1 знаходиться в горизонтальному заголовку, угорі таблиці, а f_2 вибирають ліворуч, у вертикальному заголовку таблиці.

За даними табл.7.2 одержимо

$$G_{\max} = \frac{32}{32+8+18+18} = 0{,}421.$$

З таблиці, приведеної в [3], для q=0,05 f_1=1 і f_2=4 знайдемо G=0,9065. Через те, що виконується умова

$$G_{\max} < G_{кр} = G_{табл}, \qquad (7.15)$$

всі порядкові дисперсії визнаються однорідними. У противному випадку рекомендується збільшити число рівнобіжних експериментів додаванням до наявним даної ще 1-2 серій рівнобіжних чи досвідів провести експеримент заново, звернувши особливу увагу на правильність і точність установки рівнів вхідних факторів X_i, а також застосувавши більш точні чи прилади методи виміру.

Обчислення коефіцієнтів рівняння регресії. Завдяки властивостям ортогональності (7.1.8), нормування (7.9) і симетричності (7.10) МП формули для розрахунку коефіцієнтів рівняння регресії (7.7) виявляються дуже простими

$$a_i^* = \frac{1}{N}\sum_{g=1}^{N} Z_{i,g}\vartheta_g, \qquad (7.16)$$

$$(i = \overline{0, N-1}),$$

де $Z_{i,g}$ – дорівнює +1 чи –1 у залежності від номера g рядка МП;

ϑ_g - значення цільової функції, отримане при значеннях факторів Z_i містяться в g-й рядку МП.

Якщо проводиться γ рівнобіжних досвідів, то (7.1.16) видозмінюється:

$$a_i^* = \frac{1}{N}\sum_{g=1}^{N} Z_{i,g}\,\overline{\vartheta}_g. \qquad (7.17)$$

Порядкове середнє значення цільової функції в *g*-му рядку МП $\overline{\vartheta}_g$ визначається за (7.12).

Коефіцієнти рівняння регресії (7.7) можна обчислити, якщо порядкові дисперсії однорідні, інакше не можна гарантувати статистичну значимість оцінок коефіцієнтів $a_i^*, a_{i,l}^*$. Коефіцієнти обчислюють за формулою (7.17), для приклада розрахуємо $a_0^*, a_1^*, a_2^*, a_{1,2}^*$ за даними табл. 7.1

$$\begin{aligned} a_o^* &= \frac{1}{4}(+39+88+13+53) = +48{,}3; \\ a_1^* &= \frac{1}{4}(-39+88-13+53) = +22{,}3; \\ a_2^* &= \frac{1}{4}(-39-88+13+53) = -15{,}3; \\ a_{1,2}^* &= \frac{1}{4}(+39-88-13+53) = -2{,}3. \end{aligned} \qquad (7.18)$$

Число коефіцієнтів a_i^*, оцінених незалежно один від одного, не повинно перевищувати число *N* рядків МП. Відзначимо, що можуть бути оцінені лише коефіцієнти тих факторів Z_i чи їхніх взаємодій, для яких комбінації знаків у вектор-стовпцях виявляються неповторюваними.

Покажемо це на прикладі табл.7.1, у якій незалежними є тільки перші чотири вектор-стовпці: Z_0, Z_1, Z_2, $Z_{1,2}$, і тому тільки коефіцієнти $a_0^*, a_1^*, a_2^*, a_{1,2}^*$ можна оцінити незалежно один від одного. Якби ми спробували одержати, наприклад, коефіцієнти при квадратичних членах Z_1^2, Z_2^2, то знайшли б, що знаки у вектор-стовпцях для Z_1^2, Z_2^2 точно такі ж, як і у вектор-стовпці Z_0, тобто при розрахунках $a_o^*; a_{1,1}^*; a_{2,2}^*$ за формулою (7.16) чи (7.17) ми одержали б однаковий результат. Це означає, що при плануванні типу 2^n оцінки коефіцієнтів

$a_0^*, a_{i,i}^*$ виявляються змішаними, і, щоб одержати ці оцінки роздільно, варто застосувати планування другого порядку.

Таким чином, при плануванні типу $N=2^n$ одержати незалежну, незмішану оцінку можна тільки для a_0^* і то лише у випадку, коли ділянка поверхні відгуку в районі базової точки M_0 відрізняється невеликою кривизною - у цьому випадку квадратичні члени і члени більш високих порядків виявляються незначними. Як можна бачити, при організації планування $N=2^n$ виходять з гіпотези про невелику кривизну поверхні відгуку в районі базової точки M_0, а в справедливості цієї гіпотези переконуються тільки в результаті перевірки адекватності отриманого рівняння регресії (7.7).

Статистична оцінка значимості коефіцієнтів рівняння регресії. Статистичну оцінку значимості коефіцієнтів $a_0^*, a_{i,l}^*$ роблять за *t*-критерієм Стьюдента. Для кожного коефіцієнта обчислюють *t*-відношення

$$t_i = \frac{\left|a_i^* - a_i\right|}{S\left\{a_i^*\right\}} = \frac{\left|a_i^*\right|}{S\left\{a_i^*\right\}}, \tag{7.19}$$

$$\text{де} \quad S\left\{a_i^*\right\} = \sqrt{\frac{S_0^2\left\{\vartheta\right\}}{N\gamma}}, \tag{7.20}$$

де a_i - теоретичний генеральний коефіцієнт, що думається рівним нулю (основна гіпотеза);

$S\left\{a_i^*\right\}$ - оцінка дисперсії коефіцієнтів a_i^*;

$S^2\left\{a_i^*\right\}$ - оцінка середньоквадратичної похибки у визначенні коефіцієнтів a_i^*, що знаходять з оцінки дисперсії коефіцієнтів $S^2\left\{a_i^*\right\}$;

$S_в^2\left\{\vartheta\right\}$ - оцінка дисперсії відтворюваності (помилки експерименту).

Оцінкою генеральної дисперсії відтворюваності $S_в^2\left\{\vartheta\right\}$ може служити кожна з порядкових оцінок дисперсій $S_g^2\left\{\vartheta\right\}$, $(g = \overline{1, N})$,

однак оптимальною є оцінка, обчислена як середня з усіх порядкових дисперсій:

$$S_в^2\{\vartheta\} = \frac{1}{N}\sum_{g=1}^{N} S_g^2\{\vartheta\} \ , \qquad (7.21)$$

Розраховане значення t_i (7.19) порівнюють із критичним (табличним) $t_{кр}=t_{табл}$, що вибирають для прийнятого рівня значимості q і числа ступенів вільності

$$f_в = N(\gamma - 1) \ . \qquad (7.22)$$

Якщо виконується умова

$$t_i > t_{кр} = t_{табл} \ , \qquad (7.23)$$

основна гіпотеза (a_i=0) відкидається, і коефіцієнт a_i^* визнається статистично значимим. У противному випадку основна гіпотеза приймається, a_i^* визнається статистично незначущим, і з рівняння регресії виключається.. З (7.19) випливає, що $a_{кр}^*$ можна розрахувати за формулою

$$a_{кр}^* = t_{кр} S\{a_i^*\} \ , \qquad (7.24)$$

Проведемо оцінку статистичної значимості коефіцієнтів, отриманих за даними табл.7.1. Підсумки зведені в табл.7.3.

Таблиця 7.3

		Фактори Z_i	Z_0	Z_1	Z_2	Z_1Z_2
q	5%	a_i^*	+48,3	+22,3	-15,3	-2,3
f	4	$S^2\{a_i^*\}$	2,38	2,38	2,38	2,38
$t_{кр}$	2,776	$S\{a_i^*\}$	1,54	1,54	1,54	1,54
$\sum S_{кр}^2\{\vartheta\}$	76	t_i	31,4	14,5	9,94	1,49
$\sum S_в^2\{\vartheta\}$	19	$Sign(t_i - t_{кр})$	+1	+1	+1	-1
		Вивід для a_i^*	+48,3	+22,3	-15,3	0

Значення $t_{кр}$ беруть з таблиць, приведених у [3]. З табл.7.3 видно, що при ортогональному факторному плануванні на двох рівнях дисперсії всі коефіцієнти a_i^* $(i = \overline{0, N-1})$ виявляються однаковими - це одне з переваг такого планування. $S^2\{a_i^*\}$ розраховують у такий спосіб:

$$S^2\{a_i^*\} = \frac{S_в^2\{\vartheta\}}{N\gamma} = \frac{19}{4 \cdot 2} = 2{,}38 \ .$$

$S\{a_i^*\}$ обчислюють як квадратний корінь з $S^2\{a_i^*\}$.

За результатами розрахунків видно, що три коефіцієнти виявилися значимими, а один коефіцієнт a_i^* при парній взаємодії - незначущим, і в рівнянні регресій (7.7) він повинен бути замінений нулем.

Можливі причини статистичної незначимості коефіцієнтів a_i^* можуть полягати в наступному.

Ступінь варіювання λ_i обраний для X_i занадто малим. У цьому випадку рекомендується збільшити λ_i або застосувати більш точний прилад чи метод виміру, а також збільшити число рівнозначних дослідів.

Некеровані і неконтрольовані перемінні створюють занадто високий рівень шуму, що не дозволяє виявити вплив X_i при оптимальному ступеню варіювання - тут варто застосувати спеціальні методи дослідження, наприклад еволюційне планування, чи доповнити активний експеримент пасивним.

Базова точка випадково виявилася поблизу власного екстремуму по X_i, Звідси випливає висновок, що незначимість a_i^* у даному ПФЕ не завжди означає неістотність X_i. в усьому факторному просторі, і в цьому випадку необхідно спробувати досліджувати вплив X_i в інших крапках факторного простору або перейти до квадратичної моделі.

Причина незначимості коефіцієнта a_i^* може полягати в тому, що фактор X_i дійсно є несуттєвим, і між X_i і виходом ϑ немає помітного кореляційного зв'язку. У цьому випадку ніякі міри не допоможуть зробити коефіцієнт a_i^* значимим.

Після перевірки значимості коефіцієнтів для кожного значимого a_i^* знаходять довірчий інтервал, у який повинний потрапити істинний генеральний коефіцієнт a_i із прийнятим рівнем значимості *q,* для чого застосовують формулу

$$[a_i^* - t_{кр} S\{a_i^*\} \le a_i \le a_i + t_{кр} S\{a_i^*\}] \ . \qquad (7.25)$$

За результатами розрахунків a_i^* (див.табл.2.3) видно, що отримано наступне рівняння регресії в нормалізованому масштабі:

$$\vartheta = 48{,}3 + 22{,}3Z_1 - 15{,}3Z_2 + \varepsilon \ . \qquad (7.26)$$

Статистична перевірка адекватності рівняння регресії. Статистична перевірка адекватності рівняння (7.26) полягає в перевірці того, наскільки добре апроксимує отримані експериментальні точки $\overline{\vartheta}_g \ (g = \overline{1, N})$. Для цього обчислюють дисперсію неадекватності $S_{ag}^2\{\vartheta\}$ (часто її називають дисперсією адекватності):

$$S_{ag}^2\{\vartheta\} = \frac{\gamma}{N-d} \sum_{g=1}^{N} (\overline{\vartheta}_{н,g} - \vartheta_{p,g})^2 \ , \qquad (7.27)$$

де γ' - число паралельних дослідів, по яким обчислюють $\overline{\vartheta}_{н,g}$;

N - число рядків МП, *d*- число членів рівняння регресії, у рівнянні (7.26) *d*=3; $\overline{\vartheta}_{н,g}$ - середнє порядкове зі значень виходу ϑ, що спостерігаються у g-му рядку табл. 7.1; $\vartheta_{p,g}$ - розрахункове (передбачене) значення виходу ϑ, обчислене за отриманим рівнянням регресії (7.26) для g-го рядка табл. 7.1.

$\vartheta_{p,g}$ розраховують у такий спосіб:

1-я рядок (g=1): $\vartheta_{p,1} = 48{,}3 + 22{,}3(-1) - 15{,}3(-1) = 41{,}3$;

2-я рядок (g=2): $\vartheta_{p,2} = 48{,}3 + 22{,}3(+1) - 15{,}3(-1) = 85{,}9$;

3-я рядок (g=3): $\vartheta_{p,3} = 48{,}3 + 22{,}3(-1) - 15{,}3(+1) = 10{,}7$;

4-я рядок (g=4): $\vartheta_{p,4} = 48{,}3 + 22{,}3(+1) - 15{,}3(+1) = 55{,}3$.

Таким чином, щоб одержати розрахункове (передбачене) значення $\vartheta_{p,g}$ в g-му рядку МП, слід в рівнянні (7.26) замість Z_1 і Z_2 підставити ті нормовані значення факторів, тобто ті знаки, що Z_1 і Z_2 мають в g-му рядку МП. Результати розрахунків зводимо в табл.7.2. Розраховуємо дисперсію неадекватності по (7.27):

$$S^2_{ag}\{\vartheta\} = \frac{2}{4-3}(5{,}29 + 4{,}41 + 5{,}29 + 5{,}29) = 40{,}56 \ .$$

Дисперсію $S^2_{ag}\{\vartheta\}$ порівнюємо з дисперсією відтворюваності $S^2_{в}\{\vartheta\}$ за F-критерієм Фішера:

$$F = \frac{S^2_{ag}\{\vartheta\}}{S^2_{в}\{\vartheta\}} \ . \qquad (7.28)$$

Отримане дисперсійне *F*-відношення порівнюємо з критичним (табличним) значенням *F*-статистики, що вибираємо для прийнятого рівня значимості q і чисел ступенів вільності відповідно чисельника і знаменника:

$$\left.\begin{aligned} f_{числ} &= f_{ag} = f_1 = N - d \\ f_{знам} &= f_{в} = f_2 = N(\gamma - 1) \end{aligned}\right\} , \qquad (7.29)$$

причому f_1 вибираємо в горизонтальному заголовку вгорі таблиці, а f_2 - у вертикальному заголовку ліворуч. Якщо розрахункове *F*-відношення виявилося менше критичного $F_{кр}$, тобто якщо

$$F < F_{кр} = F_{табл} \ , \qquad (7.30)$$

те отримане рівняння регресії з прийнятим рівнем значимості q визнається адекватним експериментальним даним, у противному випадку гіпотеза про адекватність відкидається.

Перевірка адекватності можлива при $f_{ag} = N - d > 0$. Якщо число N варіантів варіювання плану ПФЕ дорівнює числу всіх значимих оцінок коефіцієнтів рівняння регресії (N-d), то для перевірки гіпотези про адекватність математичного опису ступенів вільності не

залишається (f_{ag}=0). Якщо деякі оцінки коефіцієнтів регресії виявилися незначущими, то число *d*-членів рівняння, що перевіряється, у цьому випадку менше числа *N* варіантів варіювання (*N*>*d*), і для перевірки гіпотези про адекватність залишається одна чи кілька ступенів вільності *(fag*>0).

Дисперсія неадекватності $S^2_{ag}\{\vartheta\}$ не повинна значно перевищувати дисперсію відтворюваності $S^2_{в}\{\vartheta\}$, що характеризує помилку експерименту.

Проведемо перевірку адекватності рівняння (7.26). Для цього розрахуємо *F*-відношення за формулою (7.28):

$$F = \frac{40{,}56}{19} = 2{,}13 \ .$$

Для *q*=0,05 і чисел ступенів вільності $f_1 = N - d = 4 - 3 = 1$ і $f_2 = N(\gamma - 1) = 4(2-1) = 4$ знаходимо $F_{кр}$=$F_{табл}$=7,71 [3], тобто рівняння (7.26) можна визнати адекватним.

Для підвищення надійності перевірки адекватності часто ставлять додатково γ паралельних експериментів у базовій $Z_i = 0\,(i = \overline{1,n})$ точці чи в реальному масштабі при X_i=$X_{i,0}$, тоді число точок, по яких оцінюється адекватність рівняння регресії, збільшується на одну і стає *N*+1, тобто збільшується на 1 і число ступенів вільності f_{ag}. Однак слід зазначити, що базова точка при ПФЕ не бере участь у розрахунку коефіцієнтів рівняння.

Можливі причини неадекватності отриманого лінійного рівняння регресії чи рівняння, що включає в себе деякі з взаємодій факторів, полягають у наступному:

- у рівняння не включені деякі істотні перемінні, необхідно повторити експеримент із розглядом додаткових факторів;
- ступіні варіювання $\lambda_i^{"}$ узяті занадто великими, тому не схоплені характерні риси поверхні відгуку реального об'єкта - у цьому випадку варто провести експеримент зі зменшеними ступінями варіювання;
- кривизна поверхні відгуку реального об'єкта в районі базової точки настільки істотна, що навіть при нормально обраних ступінях варіювання лінійне чи неповне квадратичне рівняння

погано апроксимує цю поверхню - тут для одержання адекватного математичного опису варто перейти до планування другого порядку з використанням додаткових рівнів варіювання;

- рівень шумів надзвичайно великий, причому перешкоди несприятливо наклалися на виміри виходу, у цих умовах варто спробувати додати кілька серій паралельних експериментів чи перейти до еволюційного планування.

7.3 Центральне композиційне планування

Повний факторний експеримент дозволяє одержувати незалежні оцінки для всіх лінійних коефіцієнтів рівняння регресії і деяких взаємодій. Квадратичні коефіцієнти при плануванні на двох рівнях оцінити роздільно неможливо, тому що вони змішані з вільним членом

$$a_0^* = a_0 + \sum_{i=1}^{N} a_{i,i}. \qquad (7.31)$$

Дійсно, якщо при ПФЕ спробувати одержати вектор-стовпець для Z_i^2, то у всіх рядках такого вектор-стовпця при будь-якому i будуть лише +1, як і в стовпці для Z_0.

Таким чином, якщо ПФЕ проводиться в тій області поверхні відгуку, де кривизна істотна (що звичайно буває поблизу екстремуму), тобто там, де квадратичні члени відіграють істотну роль, то ПФЕ не дасть можливості одержати незалежну оцінку навіть для вільного члена a_0^*.

Зробити висновок про те, чи є квадратичні коефіцієнти істотними, дозволяє статистична перевірка адекватності отриманої при ПФЕ математичної моделі. Якщо модель виявилася неадекватною, це означає, що всі інші члени - квадратичні і більш високих порядків - відіграють істотну роль. У цьому випадку варто переходити до більш складного планування.

Таким чином, приступаючи до ПФЕ, дослідник заздалегідь висуває гіпотезу про неістотність квадратичних членів, а статистична перевірка адекватності отриманої при цьому моделі дозволяє йому

вирішити, суперечить висунута гіпотеза експериментальним даним чи ні.

Припустимо, було виявлено, що базова точка знаходиться поблизу головного екстремуму. У багатьох практичних випадках обмежуються таким перебуванням координат точки поблизу екстремуму. Однак у ряді випадків є необхідність більш точного дослідження екстремальної області й одержання адекватної математичної моделі цієї області, для чого необхідно в неї включити квадратичні Z_i^2, а іноді і кубічні Z_i^3 члени. Але з теорії апроксимації відомо: щоб одержати лінійну модель з одним вхідним фактором, досить двох експериментальних точок, тобто досить проваріювати цей фактор на двох рівнях (k=2). Очевидно, що для одержання n-факторної лінійної моделі кожний з n факторів потрібно проваріювати на двох рівнях. Для того щоб визначити положення кривої на площині, необхідні, у найпростішому випадку, мінімум три точки. Наприклад відомий спосіб проведення кола на площині, де задані три точки.

Таким чином, для одержання квадратичної моделі потрібний план активного експерименту, що передбачає мінімум три рівні варіювання для кожного фактора. Можна було б просто додати для усіх факторів третій рівень у центрі експерименту, що у нормованому вигляді виражається Z_i=0. Однак ПФЕ з числом рівнів варіювання k=3 мав би занадто велику кількість рядків МП, що виражається формулою

$$N = k^n = 3^n, \qquad (7.32)$$

наприклад, при n=6 N=729, у той час як при дворівневому плануванні число рядків матриці ПФЕ складає всього 2^6=64. Крім того, планування ПФЕ на не є ортогональним.

При трьох рівнях доцільно використовувати центральне композиційне планування (ЦКП). Центральним воно називається тому, що передбачає постановку деякого числа N_0 експериментів у центрі планування, тобто в базовій точці при нормованих рівнях Z_i=0. Композиційним (чи послідовним) таке планування називається тому, що воно складається з декількох частин, що здійснюються послідовно. Кожній частині відповідає група експериментальних точок.

Перша група - точки ПФ (чи ДФЕ), у яких уже були поставлені досліди, але результати яких дали неадекватну математичну модель.

Експеримент у цих точках не потрібно ставити заново, використовуються тільки його результати.

Друга група - зоряні точки, що лежать на всіх n осях нормованого факторного простору по обидві сторони від центра плану, тобто від базової точки B_0 на відстані $\pm\alpha$. Зоряними ці точки називаються тому, що на кресленнях їх звичайно позначають косими хрестиками - зірочками.

Третя група - центральні точки з нормованим рівнем $Z_i=0$.

Загальне число точок ЦКП у факторному просторі

$$N = N_{\Phi} + N_{\alpha} + N_0, \tag{7.33}$$

де N_{ϕ} - число точок факторного планування на двох рівнях (ПФЕ чи ДФЕ);

N_{α} - число зоряних точок; N_0 - число точок у центрі плану.

Очевидно,

$$\left.\begin{aligned} N_{\phi} &= 2^n \\ N_{\alpha} &= 2n \end{aligned}\right\}. \tag{7.34}$$

Число центральних точок N_0 вибирають у залежності від виду ЦКП із метою забезпечення виконання прийнятого критерію оптимальності. Величину зоряного плеча α також вибирають з умови виконання критерію оптимальності.

Таким чином, ЦКП - це планування на п'ятьох рівнях, що у нормованому вигляді можна представити так: 1) $-\alpha$; 2) -1; 3) 0; 4) +1; 5) $+\alpha$.

ЦКП буває двох видів: ортогональне (ОЦКП) другого чи третього порядку і ротатабельне (РЦКП) другого чи третього порядку. Види ЦКП розрізняються критеріями оптимальності.

Для ОЦКП другого порядку критерієм оптимальності є ортогональність усіх вектор-стовпців МП, включаючи і вектор-стовпці для квадратичних членів $Z_i^2\ (i=\overline{1,n})$. Для ортогонального ЦКП третього порядку МП складають так, щоб ортогональними з всіма іншими й один з одним були також і вектор-стовпці кубічних членів Z_i^3. Останнє ОЦКП дуже складно і рідко застосовується.

У більшості практичних випадків повного квадратичного

полінома виявляється досить, щоб адекватно описати поверхню відгуку в екстремальній області.

Для ротатабельного (rotatable - здатний до обертання координат) ЦКП критерієм оптимальності служить однакова точність прогнозування виходу за допомогою рівняння регресії по будь-якій координаті і взагалі по будь-якому напрямку у вивченій області факторного простору на рівній відстані від центра експерименту.

ПФЕ на двох рівнях є й ортогональним і ротатабельним плануванням, але дозволяє одержувати тільки лінійні і неповні квадратичні моделі.

Одержання МП при реалізації ОЦКП другого порядку. В ОЦКП другого порядку повинні бути ортогональними один з одним усі вектор-стовпці, включаючи і вектор-стовпці квадратичних членів Z_i^2 $(i=\overline{1,n})$ і нульового члена Z_0. Ортогональність забезпечується наступними заходами.

По-перше, вибором величини зоряного плеча α і кількості N_0 точок у центрі ОЦКП. Кількість точок факторного простору і розміри зоряних пліч при ОЦКП для різної кількості N факторів приведені в табл.7.4. Навіть мінімальне число центральних точок N_0=1 забезпечує ортогональність усіх вектор-стовпців, включаючи квадратичні.

Таблиця 7.4

Число факторів	2	3	4	5	6
Точок ПФЕ (ДФЕ) N_ϕ	2^2	2^3	2^4	2^5-1[x]	2^6-1[xx]
Зоряних точок N_α	4	6	8	10	12
Центральних точок N_0	1	1	1	1	1
Загальна кількість точок N	9	15	25	27	45
Розмір плеча α	1,00	1,215	1,414	1,547	1,724

[x]При визначальному контрасті 1=Z_1,Z_2, Z_3, Z_4,Z_5.

[xx]При ОК 1=Z_1, Z_2, Z_3, Z_4, Z_5, Z_6.

По-друге, ортогональність забезпечується перетворенням квадратичних членів Z_i^2 у нові штучні члени $\widehat{Z}_i^2$ за формулою

$$\hat{Z}^2_{i,g} = Z^2_{i,g} - \frac{1}{N}\sum_{g=1}^{N} Z^2_{i,g} = Z^2_{i,g} - (\overline{Z}^2), \qquad (7.35)$$

де i - номер фактора $(i = \overline{1,n})$; g - номер рядка МП ОЦКП другого порядку (табл.7.5); N - загальне число рядків МП ОЦКП, визначене за формулою (7.33); $(\overline{Z}^2)$ - усереднене значення рядків для квадратичних стовпців $\overline{Z}^2$ $(i = \overline{1,n})$, ця величина однакова для всіх квадратичних стовпців і залежить тільки від кількості факторів n.

$$(\overline{Z}^2) = \frac{2^n + 2\alpha^2}{N} = \frac{2^n + 2\alpha^2}{2^n + 2^{n+1}}. \qquad (7.36)$$

Значення $(\overline{Z}^2)$ а також значення інших констант ОЦКП приведені в табл.7.6.

Таблиця 7.6

Константи ОЦКП	Число факторів		
	2	3	4
$(\overline{Z}^2)$	0,6667	0,7302	0,8000
ОВ	0,1111	0,0667	0,0400
ІВ	0,1667	0,913	0,0500
ILB	0,2500	0,1250	0,0625
IIB	0,5000	0,2293	0,1250

Таблиця 7.5

№ стовп-ця		1	2	3	4	5	6	7	8	9а	10а	11а	9	10	11
Групи точок	g	Z_0	Z_1	Z_2	Z_3	$Z_1 Z_2$	$Z_1 Z_3$	$Z_2 Z_3$	$Z_1 Z_2 Z_3$	Z_1^2	Z_2^2	Z_3^2	$\hat{Z}_1^2$	$\hat{Z}_2^2$	$\hat{Z}_3^2$
N_ϕ	1	+1	-1	-1	-1	+1	+1	+1	-1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	2	+1	+1	-1	-1	-1	-1	+1	+1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	3	+1	-1	+1	-1	-1	+1	-1	+1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	4	+1	+1	+1	-1	+1	-1	-1	-1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	5	+1	-1	-1	+1	+1	-1	-1	+1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	6	+1	+1	-1	+1	-1	+1	-1	-1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	7	+1	-1	+1	+1	-1	-1	+1	-1	+1	+1	+1	+0,2699	+0,2699	+0,2699
	8	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+0,2699	+0,2699	+0,2699
N_α	9	+1	-1,215	0	0	0	0	0	0	+1,476	0	0	+0,7459	-0,7301	-0,7301
	10	+1	-1,215	0	0	0	0	0	0	+1,476	0	0	+0,7459	-0,7301	-0,7301
	11	+1	0	-1,215	0	0	0	0	0	0	+1,476	0	0,7301	+0,7459	-0,7301
	12	+1	0	+1,215	0	0	0	0	0	0	+1,476	0	0,7301	+0,7459	-0,7301
	13	+1	0	0	-1,215	0	0	0	0	0	0	+1,476	0,7301	-0,7301	+0,7459
	14	+1	0	0	+1,215	0	0	0	0	0	0	+1,476	0,7301	-0,7301	+0,7459
N_0	15	+1	0	0	0	0	0	0	0	0	0	0	0,7301	-0,7301	-0,7301

В табл. 7.5 приведена МП ОЦКП для трьох факторів, вектор-стовпці для лінійних факторів Z_i, для парних і потрійних взаємодій, для неперетворених членів Z_i^2 і для перетворених квадратичних членів $\widehat{Z}_i^2$. Якщо при складанні МП для ПФЕ чи ДФЕ, де застосовують тільки два рівні варіювання, у матриці часто замість +1 чи -1 ставлять просто знак "+" чи "-", то в матриці ЦКП, щоб уникнути непорозумінь, необхідно проставляти і знак, і нормовану величину п'яти рівнів варіювання: 1) -(; 2) -1; 3) 0; 4) +1; 5) +α.

Квадратичні члени в табл.7.5 відповідно до виразу (7.35) і табл.7.6 перетворені в такий спосіб:

$$\widehat{Z}_{i,g}^2 = Z_{i,g}^2 - 0{,}73\,. \qquad (7.37)$$

Варто пам'ятати, що зоряне плече α у табл. 7.4 приведене в нормованому масштабі, а при переході до звичайного масштабу необхідно врахувати величину ступіней варіювання λ_i по кожному факторі Z_i. Нормування здійснюють за формулою

$$Z_i = \|X_i\| = \frac{X_i - X_{i,0}}{\lambda_i}. \qquad (7.38)$$

Величину $X_i(\lambda)$ у звичайному масштабі, необхідну при проведенні реальних експериментів у зоряних точках, варто обчислювати за формулою, що випливає з (3.8):

$$X_i(\pm\alpha) = X_{i,0} \pm \alpha\lambda_i\ , \qquad (7.39)$$

де $X_{i,0}$ - базове значення фактора X_i.

Приклад розрахунку координат зоряних точок у реальному масштабі. Припустимо, центр ПФЕ, у результаті проведення якого була отримана неадекватна математична модель, був обраний у точці трифакторного простору з координатами

$X_{1,0}$=40; $X_{2,0}$=50; $X_{3,0}$=60,

а ступіні варіювання обрані в ПФЕ в такий спосіб:

λ_1=15; λ_2=20; λ_3=10.

Тоді координати зоряних точок будуть такими, як це показано в табл.7.7.

Розрахунок виконують за формулою (7.39):

$$X_1(-\alpha) = 40 - 1{,}215 \cdot 15 = 21{,}8 \approx 22;$$
$$X_1(+\alpha) = 40 + 1{,}215 \cdot 15 = 58{,}2 \approx 58;$$
$$X_2(-\alpha) = 50 - 1{,}215 \cdot 20 = 25{,}7 \approx 26;$$

(7.40)

$$X_2(+\alpha) = 50 + 1{,}215 \cdot 20 = 74{,}3 \approx 74;$$
$$X_3(-\alpha) = 60 - 1{,}215 \cdot 10 = 47{,}9 \approx 48;$$
$$X_3(+\alpha) = 60 + 1{,}215 \cdot 10 = 72{,}1 \approx 72.$$

Таблиця 7.7

Рівень	Нормована величина	Реальний масштаб		
		X_1	X_2	X_3
Базовий рівень $X_{i,0}$	0	40	50	60
Ступінь варіювання λ_i	+1	15	20	10
Нижній рівень ПФЕ $X_{i,н}$	-1	25	30	50
Верхній рівень ПФЕ $X_{i,y}$	+1	55	70	70
Нижня зоряна точка ОЦКП $-\alpha$	-1,215	22	26	48
Верхня зоряна точка ОЦКП $+\alpha$	+1,215	58	74	72

Обчислення коефіцієнтів квадратичного рівняння регресії. Розглянемо дані табл.7.7. Можна переконатися, що її вектор - стовпці задовольняють двом із трьох умов (7.2.6) - (7.2.10):

$$\sum_{g=1}^{N} Z_{i,g} Z_{l,g} = 0; \qquad (7.41)$$

$$\sum_{g=1}^{N} Z_{i,g} = 0, \qquad (7.42)$$

де Z_i, Z_l- будь-які вектор-стовпці табл.7.5, крім Z_1^2, Z_2^2, Z_3^2, причому умова (7.41) - умова ортогональності вектор-стовпців, а (7.42) - умова симетричності МП, що означає,

що рівні, однакові по абсолютній величині, зустрічаються в кожнім вектор–стовпці рівне число раз з різними знаками. Для вектор-стовпців з перетвореними квадратичними членами $\widehat{Z}_i^2$ умова симетричності (7.42) означає, що сума значень рівнів, що мають позитивний знак, дорівнює по абсолютній величині сумі рівнів з негативним знаком.

В МП табл.7.5 не виконується умова нормування, тобто для неї

$$\sum_{g=1}^{N} Z_{i,g}^2 \neq N. \qquad (7.43)$$

В (7.43) під Z_i розуміється будь-який вектор-стовпець, крім нульового, і це приводить до того, що розрахункові формули для a_i^* при ОЦКП трохи відрізняються від формул ПФЕ:

$$a_i^* = \frac{1}{\sum_{g=1}^{N} Z_{i,g}^2} \sum_{g=1}^{N} Z_{i,g}^2 \vartheta_g; \qquad (i = \overline{0, k-1}), \qquad (7.44)$$

де *k*- номер останнього стовпця в МП ОЦКП.

Формула (7.44) дозволяє одержати математичну модель для нормованих факторів, що включає перетворені квадратичні члени $\widehat{Z}_i^2$:

$$\vartheta_p = a_0^* + \sum_{i=1}^{n} a_i^* Z_i + \sum_{\substack{i,l=1 \\ i \neq l}} a_{i,l}^* Z_i Z_l + \sum_{i=1}^{n} a_{ii}^* \widehat{Z}_i^2 + \varepsilon, \qquad (7.45)$$

де ε - помилка одиничного передбачення.

На підставі формули (7.35) одержуємо рівняння регресії для неперетворених квадратичних членів:

$$\vartheta_p = \widehat{a}_0^* + \sum_{i=1}^{n} a_i^* Z_i + \sum_{\substack{i,l=1 \\ i \neq l}} a_{i,l}^* Z_i Z_l + \sum_{i=1}^{n} a_{ii}^* \widehat{Z}_i^2 + \varepsilon, \qquad (7.46)$$

де

$$\widehat{a}_0^* = a_0^* - (\overline{Z}^2)\sum_{i=1}^{n} a_{ii}^*. \qquad (7.47)$$

Розглядаючи формулу (7.44) і табл.7.5, бачимо, що співмножник $1/\sum_{g=1}^{N} Z_{i,g}^2$ у (7.44) неоднаковий для різних груп однорідних коефіцієнтів, але усередині кожної групи має рівне значення для всіх коефіцієнтів групи. Пояснимо сказане на прикладі табл. 7.5. Приведемо розрахунок цього співмножника для чотирьох груп однорідних коефіцієнтів:

стовпець 1 (постійний коефіцієнт a_0^*):

$$1/\sum_{g=1}^{N} Z_{0,g}^2 = 1/N = 1/(2^n + 2n + 1) = OB; \qquad (7.48)$$

стовпці 2-4 (лінійні коефіцієнти a_i^*):

$$1/\sum_{g=1}^{N} Z_{i,g}^2 = 1/(2^n + 2\alpha^2) = IB; \qquad (7.49)$$

стовпці 5-8 (коефіцієнти $a_{i,l}^*$, при парних і потрійних взаємодіях):

$$1/\sum_{g=1}^{N} Z_{i,l,g}^2 = 1/(2^n) = ILB; \qquad (7.50)$$

стовпці 9-11 (коефіцієнти a_{ii}^* при перетворених квадратичних членах):

$$1/\sum_{g=1}^{N} (\hat{Z}_{i,g}^2)^2 = 1/\{2^n[1-(\overline{Z}^2)]^2 + 2[\alpha^2 - (\overline{Z}^2)]^2 + \qquad (7.51)$$
$$+ (2n-1)[0-(\overline{Z}^2)]^2\ \} = IIB$$

Аналізуючи вираз (7.48)-(7.51), можна помітити, що співмножники ОВ, ІВ, ІLВ і ІІВ є константами, що залежать тільки від кількості факторів n. Значення цих констант ОЦКП приведені в табл. 7.6.

При використанні констант ОЦКП вихідна формула (7.46) для розрахунку оцінок коефіцієнтів рівняння регресії має наступний вигляд:

$$\left.\begin{aligned}
a_i^* &= (IB)\sum_{g=1}^{N} Z_{i,g}\overline{\vartheta}_g;\\
a_{i,l}^* &= (ILB)\sum_{g=1}^{N} Z_{i,l,g}\overline{\vartheta}_g;\\
a_{ii}^* &= (IIB)\sum_{g=1}^{N} \bar{Z}_{i,g}\overline{\vartheta}_g = (IIB)\sum_{g=1}^{N}[Z_{i,g}^2 - (\overline{Z}^2)]\overline{\vartheta}_g;\\
a_0^* &= (OB)\sum_{g=1}^{N} \overline{\vartheta}_g\\
\bar{a}_0^* &= a_0^* - (\overline{Z}^2)\sum_{i=1}^{n} a_{ii}^*.
\end{aligned}\right\} \quad (7.52)$$

Дисперсії коефіцієнтів також виявляються однаковими тільки усередині чотирьох груп однорідних коефіцієнтів, перерахованих раніше. Наприклад дисперсії $S^2\{a_i^*\}$ лінійних коефіцієнтів однакові при $(i=\overline{1,n})$ але відрізняються від дисперсій коефіцієнтів $a_0^*, a_{i,l}^*, a_{ii}^*$. Оцінки дисперсій коефіцієнтів обчислюють за наступними формулами:

$$S^2\{a_i^*\} = \frac{(IB)}{\gamma} S_в^2\{\vartheta\}; \quad (7.53)$$

$$S^2\{a_{i,l}^*\} = \frac{(ILB)}{\gamma} S_в^2\{\vartheta\}; \quad (7.54)$$

$$S^2\{a_{ii}^*\} = \frac{(IIB)}{\gamma} S_в^2\{\vartheta\}; \quad (7.55)$$

$$S^2\{a_0^*\} = \frac{(OB)}{\gamma} S_в^2\{\vartheta\}; \quad (7.56)$$

$$S^2\{\hat{a}_0^*\} = S\{a_0^*\} + (\bar{Z}^2)^2 \cdot \sum_{i=1}^{n} S^2\{a_{ii}^*\} =$$

$$= S^2\{a_0^*\} + n(\bar{Z}^2)^2 S^2\{a_{ii}^*\} + n(\bar{Z}^2)^2 \frac{(IIB)}{\gamma} S_в^2\{\vartheta\}, \quad (7.57)$$

де γ - число паралельних дослідів; $S_в^2\{\vartheta\}$- дисперсія відтворюваності дослідів для цільової функції ϑ, що обчислюється за формулою (7.2.21) з урахуванням того, що N визначається за формулами (7.6) і (7.7); i, l - номера лінійних факторів; g - номер рядка МП ОЦКП; N- загальне число рядків МП ОЦКП. обумовлене по формулах (7.3.3) і (7.3.4); n – число лінійних факторів.

Вибір числа γ паралельних дослідів, рандомізація порядку їхнього проведення, статистична оцінка відтворюваності дослідів здійснюється так само, як і при ПФЕ. Однак варто пам'ятати, що якщо вже був виконаний ПФЕ, то рандомізувати і проводити досліди слід тільки в зоряних і центральних точках: Наприклад, в табл.7.5 слід рандомізувати в γ серіях паралельних експериментів тільки рядки МП із 9 по 15, тобто з таблиці випадкових чисел треба вибирати числа з 9 по 15 включно. Але статистичну оцінку відтворюваності (чи статистичну оцінку однорідності порядкових дисперсій) варто робити для всіх N порядкових дисперсій $S_g^2\{\vartheta\}$.

Статистична оцінка значимості отриманих коефіцієнтів рівняння регресії робиться так само, як і при ПФЕ з тією лише різницею, що для чотирьох різних груп коефіцієнтів їхньої дисперсії будуть різними, як це випливає з формул (7.53)-(7.57).

Статистичну перевірку адекватності квадратичного рівняння регресії, отриманого при ортогональному ЦКП, проводять так само, як і при ПФЕ, однак з урахуванням зміненого числа ступенів вільності: у формулах (7.2.29) N буде визначатися зі співвідношеннями (7.33), (7.34).

Якщо квадратичне рівняння (7.46) виявилося адекватним із прийнятим рівнем значимості, то можна досліджувати його відомими аналітичними методами.

Математичний опис області екстремуму можна використовувати для прогнозування цільової функції при заданих значеннях факторів, а також для оптимального керування технологічними процесами.

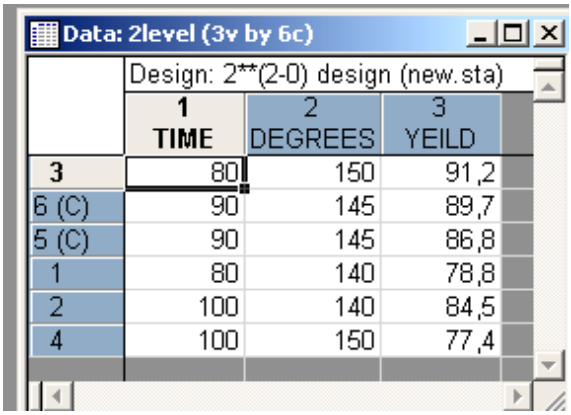
7.4 Комп'ютерні технології планування експерименту

***ПЛАНУВАННЯ ЕКСПЕРИМЕНТУ В ПАКЕТІ* STATISTICA**

В пакеті STATISTICA є могутній модуль планування експериментів, що дозволяє здійснити планування повних і дробових факторних експериментів з варіюванням факторів на 2-х і/чи 3-х рівнях, побудувати центральні композиційні плани, латинські і греко-латинські квадрати, робастні плани і ряд інших планів.

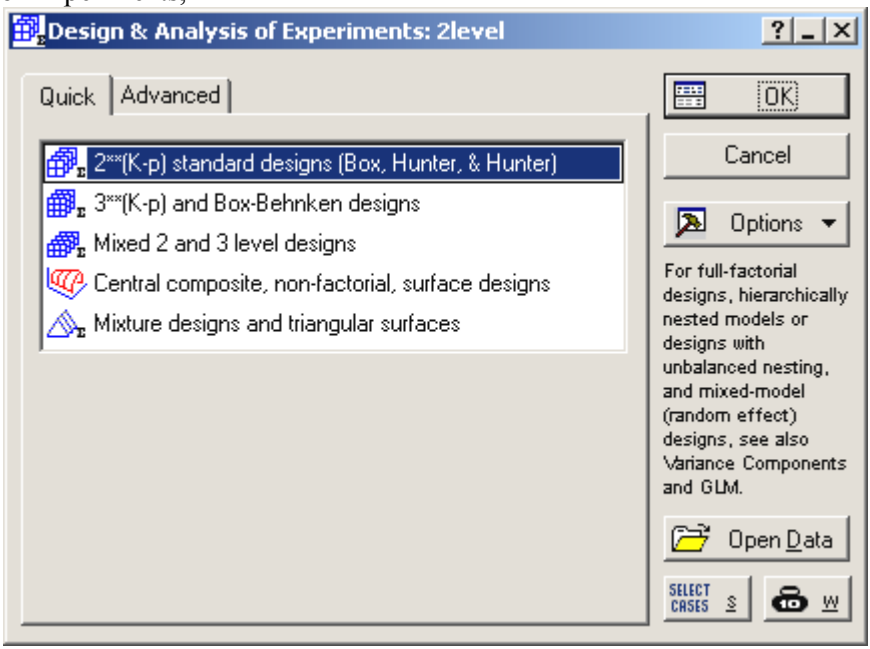
Обмежимося розглядом тільки технології побудови планів ПФЕ та їхнім аналізом.

Крок 1. Ввести або імпортувати (наприклад, з Excel) вихідні дані в робочу книгу (Workbook) системи STATISTICA, виділити їх, увести назву таблиці вихідних даних і назви перемінних. Або можна вибрати файл із вже існуючих, для приклада візьмемо файл 2level.sta, що знаходиться в папці Examples і поставляється разом із системою STATISTICA.



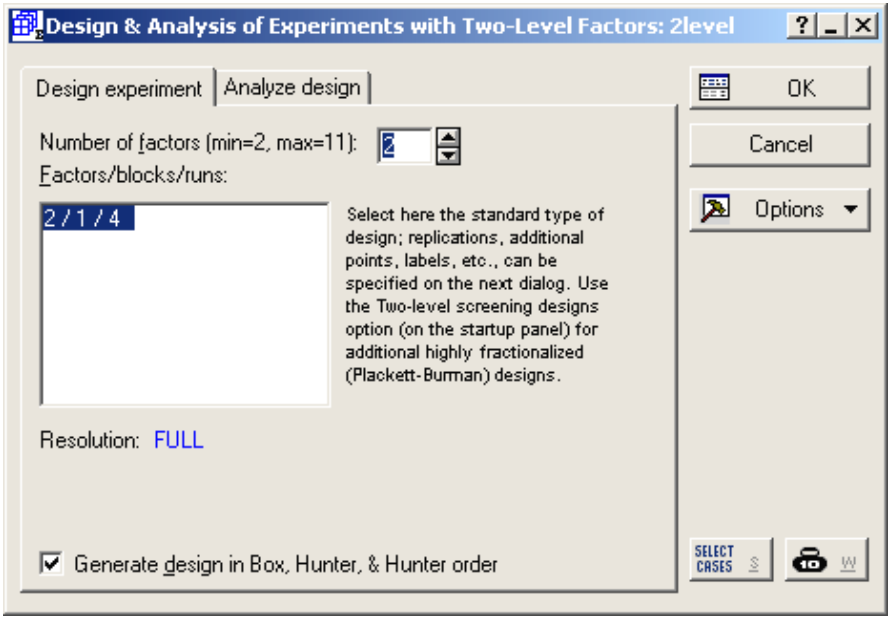
	1 TIME	2 DEGREES	3 YEILD
3	80	150	91,2
6 (C)	90	145	89,7
5 (C)	90	145	86,8
1	80	140	78,8
2	100	140	84,5
4	100	150	77,4

Крок 2. Клацнути по кнопці Start menu ..., розташованої в лівому нижньому куті вікна додатка й у меню, що з'явилось, вибрати Statistics (Industrial Statistics & Six Sigma (Experimental Design (DOE). Після чого з'являється вікно Design and Analysis of Experiments,



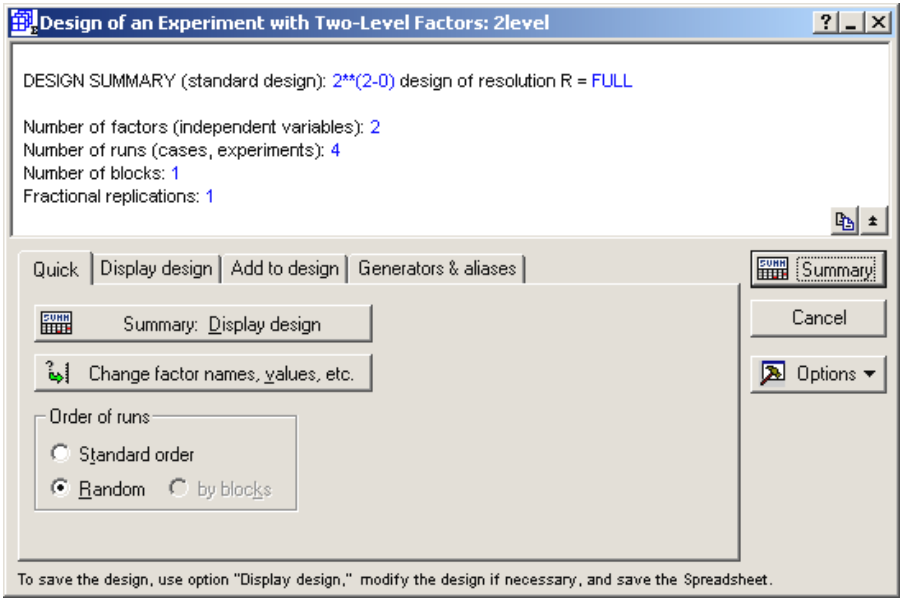
у який, після вибору потрібного плану (у використовуваному прикладі вибирається план **2****(K-p) standard designs (Box, Hunter & Hunter), клацнути по кнопці OK.

Крок 3. У вікні, що з'явилося, Design & Analysis of Experiments with Two-Level Factors - Design Experiment задати на лічильнику Number of factors (min=2, max=11) (у використовуваному прикладі число факторів – 2)

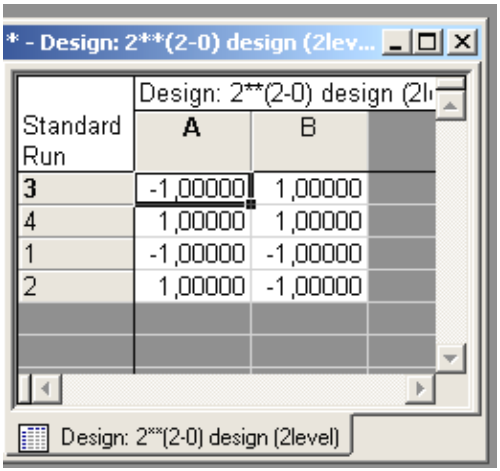


і клацнути по кнопці OK.

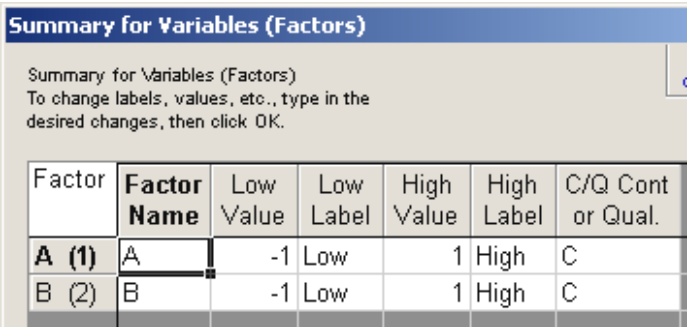
Крок 4. У вікні, що з'явилося, Design of an Experiment with Two-Level Factors - Quick Tab,



щиглик по кнопці Summary: Display design, дозволяє вивести вікно з матрицею планування ПФЭ (у використовуваному прикладі це матриця виду2^2)

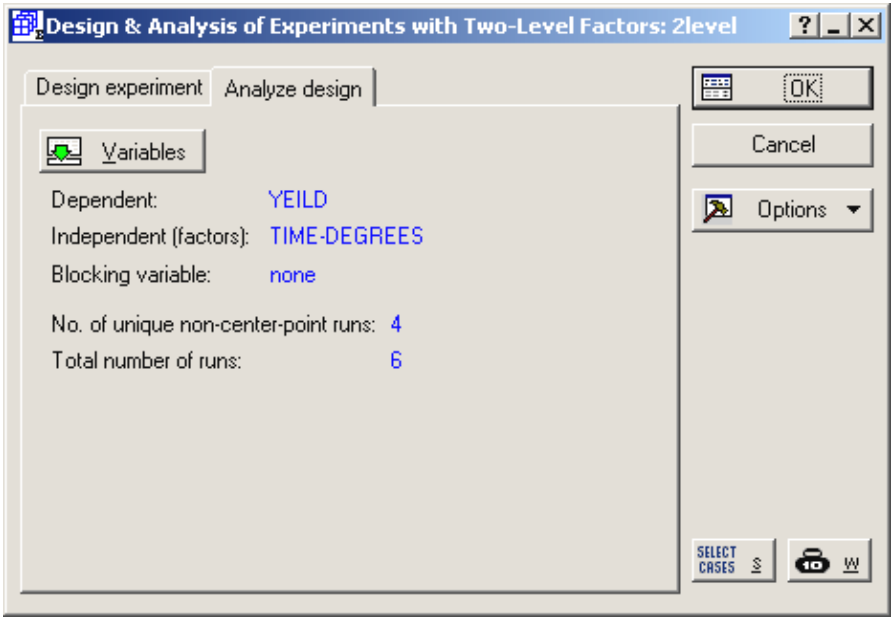


щиглик по кнопці Change factor names, values, etc показує кодування факторів у матриці планування



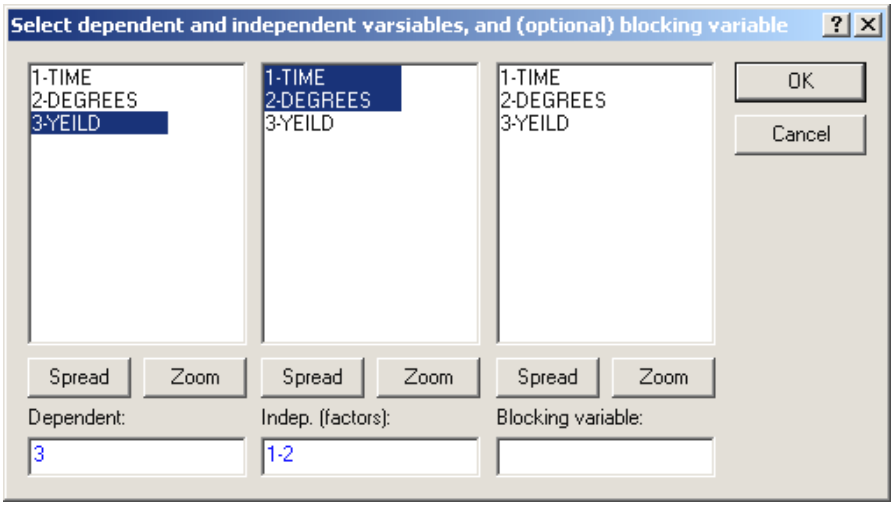
Factor	Factor Name	Low Value	Low Label	High Value	High Label	C/Q Cont or Qual.
A (1)	A	-1	Low	1	High	C
B (2)	B	-1	Low	1	High	C

Крок 5. Повернуться до вікна Design & Analysis of Experiments with Two-Level Factors , у которм выбрыть вкладку - Analyze Design Tab і клацнути по кнопці OK. У вікні, що з'явилося, Design & Analysis of Experiments with Two-Level Factors - Analyze Design Tab



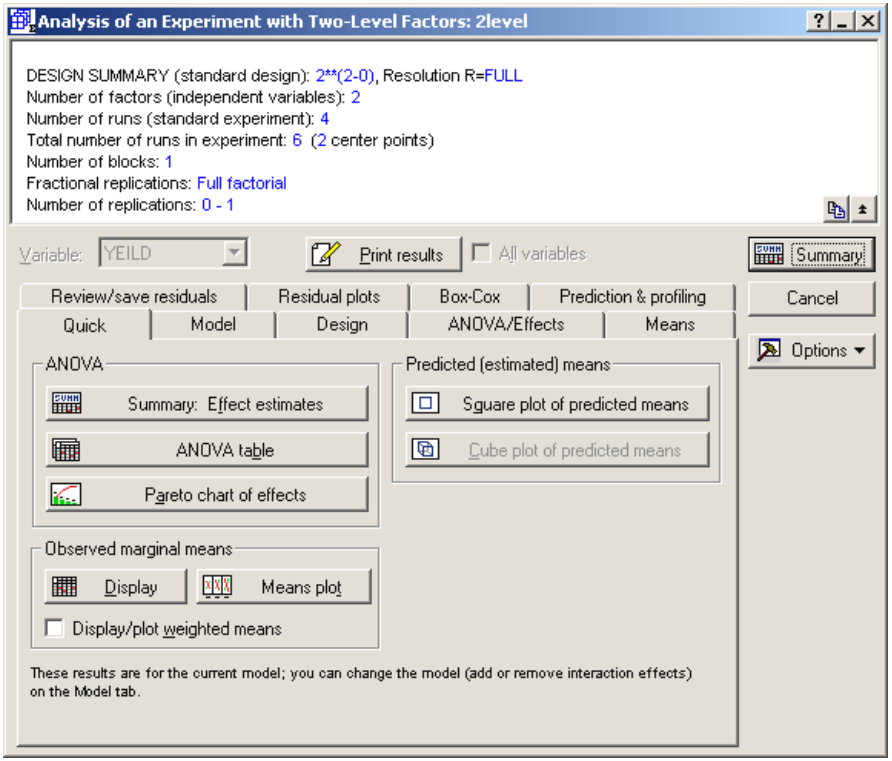
і клацнути по кнопці Variables і визначити залежну перемінну (dependent), відгук і незалежні перемінні (indep.

(factors)), фактори. У використовуваному прикладі визначення перемінних показане в наступному вікні

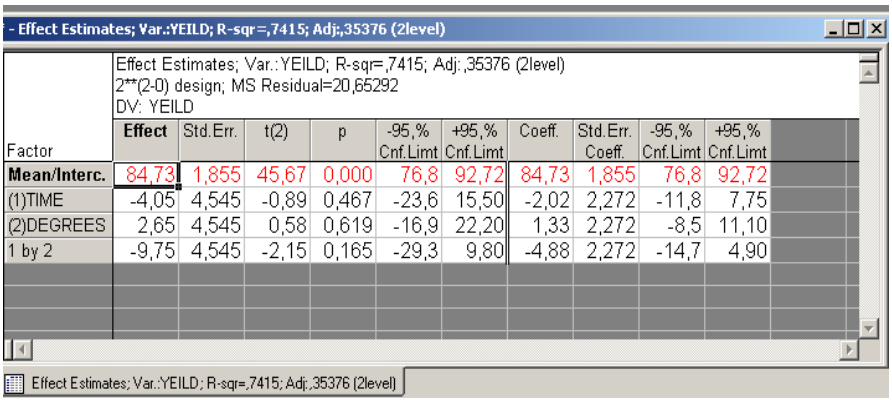


Клацнути по кнопці OK.

Крок 6. За допомогою вікна, що з'явилося, можна зробити всебічний аналіз результатів експерименту за досліджуваним планом. Для використовуваного приклада дане вікно буде мати наступний вид



При щиглику по кнопці Summary: Effect estimates виводиться вікно з ефективними коефіцієнтами моделі.



Factor	Effect	Std.Err.	t(2)	p	-95,% Cnf.Limt	+95,% Cnf.Limt	Coeff.	Std.Err. Coeff.	-95,% Cnf.Limt	+95,% Cnf.Limt
Mean/Interc.	84,73	1,855	45,67	0,000	76,8	92,72	84,73	1,855	76,8	92,72
(1)TIME	-4,05	4,545	-0,89	0,467	-23,6	15,50	-2,02	2,272	-11,8	7,75
(2)DEGREES	2,65	4,545	0,58	0,619	-16,9	22,20	1,33	2,272	-8,5	11,10
1 by 2	-9,75	4,545	-2,15	0,165	-29,3	9,80	-4,88	2,272	-14,7	4,90

7.5. Контрольні питання і завдання

7.5.1.Питання

1 Поняття математичної теорії планування експерименту.
2 Пасивний і активний експеримент.
3 Повний факторний експеримент. Вибір факторів і параметра оптимізації.
4 Вибір координат базової крапки і рівнів варіювання при повному факторному експерименті.
5 Складання матриці планування повного факторного експерименту.
6 Рандомизація досвідів при плануванні експериментів.
7 Перевірка відтворюваності досвідів при плануванні експериментів.
8 Обчислення коефіцієнтів рівняння регресії при факторному експерименті.
9 Статистична оцінка значимості коефіцієнтів рівняння регресії**.**
10 Можливі причини статистичної незначимості коефіцієнтів рівняння регресії.
11 Статистична перевірка адекватності рівняння регресії**.**
12 Можливі причини неадекватності отриманого лінійного рівняння регресії
13 Центральне композиційне планування (ЦКП). Коли доцільно його використовувати.
14 Загальне число крапок ЦКП (зоряні і центральні крапки).
15 Ортогональне і ротатабельне ЦКП.
16 Одержання матриці планування при реалізації ортогонального ЦКП другого порядку.
17 Обчислення коефіцієнтів квадратичного рівняння регресії при ЦКП.
18 Статистична оцінка значимості отриманих коефіцієнтів рівняння регресії при ЦКП.
19 Статистичну перевірку адекватності квадратичного рівняння регресії при ЦКП.

7.5.2. Завдання

Завдання 1. Для побудови математичної моделі технологічної операції «дифузія бора» передбачається використовувати метод планування експерименту. Дана операція контролюється по величині поверхневого опору R_s, що заміриться на тестових осередках контрольних пластин. Факторами, що використовуються для керування операцією є:

X_1 – максимальна температура дифузії, вимірювана в °С;

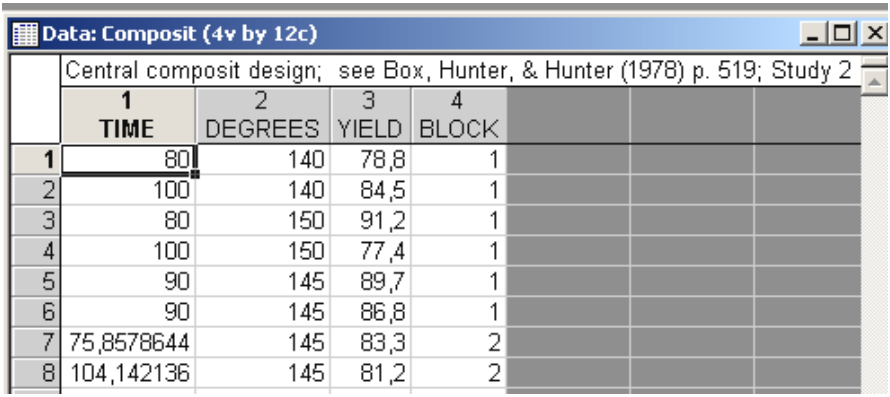
X_2 – час дифузії, вимірюваний у хв.

Необхідно. На підставі результатів експерименту (таблиця 7.5.1.) одержати лінійне рівняння регресії і виконати його аналіз.

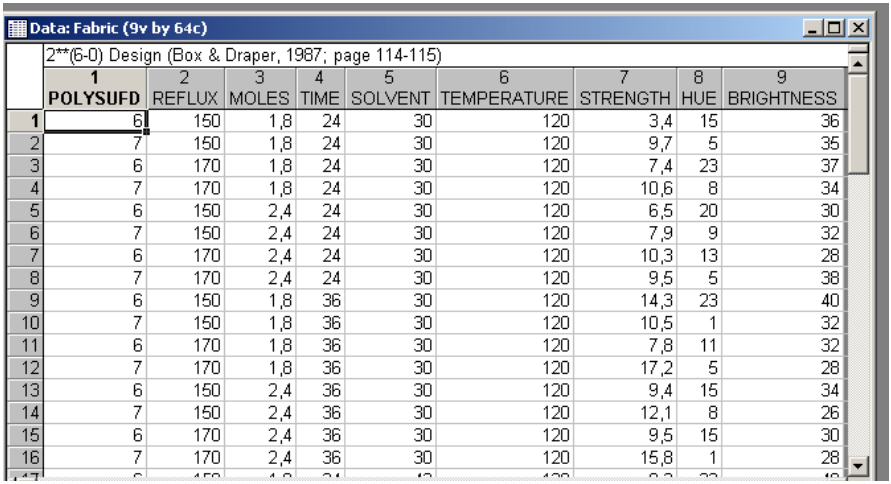
Таблиця 7.5.1

N п/п	X_1	X_2	Значения R_s									
			1	2	3	4	5	6	7	8	9	10
1	1100	15	124	121	118	119	121	123	118	126	119	120
2	900	15	93	95	90	94	93	95	91	96	87	95
3	1100	5	110	112	111	110	117	118	108	101	107	108
4	900	5	67	70	71	68	71	71	72	73	74	70

Завдання 2. Побудувати матрицю планування, одержати рівняння регресії і виконати його аналіз, використовувавши в якості вихідних даних файл composit.sta, що знаходиться в папці Examples і поставляється разом із системою STATISTICA.



Завдання 3. Побудувати матрицю планування, одержати рівняння регресії і виконати його аналіз, використовувавши в якості вихідних даних файл fabric.sta (9v by 64c), що знаходиться в папці Examples і поставляється разом із системою STATISTICA.



Data: Fabric (9v by 64c)

2**(6-0) Design (Box & Draper, 1987; page 114-115)

	1 POLYSUFD	2 REFLUX	3 MOLES	4 TIME	5 SOLVENT	6 TEMPERATURE	7 STRENGTH	8 HUE	9 BRIGHTNESS
1	6	150	1,8	24	30	120	3,4	15	36
2	7	150	1,8	24	30	120	9,7	5	35
3	6	170	1,8	24	30	120	7,4	23	37
4	7	170	1,8	24	30	120	10,6	8	34
5	6	150	2,4	24	30	120	6,5	20	30
6	7	150	2,4	24	30	120	7,9	9	32
7	6	170	2,4	24	30	120	10,3	13	28
8	7	170	2,4	24	30	120	9,5	5	38
9	6	150	1,8	36	30	120	14,3	23	40
10	7	150	1,8	36	30	120	10,5	1	32
11	6	170	1,8	36	30	120	7,8	11	32
12	7	170	1,8	36	30	120	17,2	5	28
13	6	150	2,4	36	30	120	9,4	15	34
14	7	150	2,4	36	30	120	12,1	8	26
15	6	170	2,4	36	30	120	9,5	15	30
16	7	170	2,4	36	30	120	15,8	1	28

ДОДАТОК А
СТАТИСТИЧНІ ТАБЛИЦІ

Функція стандартного нормального розподілу

$$\Phi(x) = \frac{1}{\sqrt{2x}} \int_0^x e^{\frac{-t^2}{2}} dt$$

x	0,00	0,02	0,04	0,05	0,06	0,07	0,08	0,09
0	0	0,008	0,016	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0478	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0871	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1255	0,1331	0,1368	0,1406	0,1443	0,148	0,1517
0,4	0,1554	0,1628	0,17	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1985	0,2054	0,2088	0,2123	0,2157	0,219	0,2224
0,6	0,2257	0,2324	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,258	0,2642	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2888	0,2939	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3212	0,3264	0,3289	0,3315	0,334	0,3365	0,3389
1	0,3413	0,3461	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3686	0,3729	0,3749	0,377	0,379	0,381	0,383
1,2	0,3849	0,3888	0,3925	0,3944	0,3962	0,398	0,3997	0,4015
1,3	0,4032	0,4066	0,4099	0,4115	0,4134	0,4147	0,4162	0,4177
1,4	0,4192	0,4222	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4357	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4474	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4573	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4656	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4726	0,4738	0,4744	0,475	0,4756	0,4761	0,4767
2	0,4772	0,4783	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,483	0,4838	0,4842	0,4846	0,485	0,4854	0,4857
2,2	0,4861	0,4868	0,4875	0,4878	0,4881	0,4884	0,4887	0,489
2,3	0,4893	0,4898	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4922	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4941	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4956	0,4959	0,496	0,4961	0,4962	0,4963	0,4964
2,8	0,4974	0,4976	0,4977	0,4978	0,4979	0,49794	0,498	0,4981
3,0	0,49865	0,49874	0,49881	0,49886	0,49889	0,49893	0,49897	0,499

Критичні точки розподілу Стьюдента

k	α							
	10%	5%	2,50%	1%	0,50%	0,25%	0,10%	0,05%
1	3,078	3,314	12,706	31,821	63,657	127,32	318,30	636,61
2	1,886	2,920	4,303	6,965	9,925	14,089	22,327	31,599
3	1,638	2,353	3,182	4,541	5,841	7,453	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	5,598	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	4,773	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	4,317	5,208	5,959
7	1,415	1,895	2,365	2,998	3,500	4,029	4,785	5,408
8	1,397	1,860	2,306	2,897	3,355	3,833	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	3,690	4,297	4,781
10	1,372	1,813	2,228	2,764	3,169	3,581	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	3,497	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,428	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,373	3,852	4,221
14	1,345	1,761	2,145	2,625	2,977	3,326	3,787	4,141
15	1,341	1,753	2,131	2,603	2,947	3,286	3,733	4,073
20	1,325	1,725	2,086	2,528	2,845	3,153	3,552	3,850
25	1,316	1,708	2,060	2,485	2,787	3,078	3,450	3,725
30	1,310	1,697	2,042	2,457	2,750	3,030	3,385	3,646
40	1,303	1,684	2,021	2,423	2,705	2,971	3,307	3,551
50	1,299	1,676	2,009	2,403	2,678	2,937	3,261	3,496
60	1,296	1,671	2,000	2,390	2,660	2,915	3,232	3,460
70	1,294	1,667	1,994	2,381	2,648	2,899	3,211	3,435
80	1,292	1,664	1,990	2,374	2,639	2,887	3,195	3,416
90	1,291	1,662	1,987	2,369	2,632	2,878	3,183	3,402
100	1,290	1,660	1,984	2,364	2,626	2,871	3,174	3,391
200	1,286	1,653	1,972	2,345	2,601	2,839	3,132	3,340
300	1,284	1,650	1,968	2,339	2,592	2,828	3,118	3,323
400	1,284	1,649	1,966	2,336	2,588	2,823	3,111	3,315
500	1,283	1,648	1,965	2,334	2,586	2,820	3,107	3,310

Критичні точки розподілу χ^2

k	α												
	99%	97,50%	95%	90%	80%	60%	40%	20%	10%	5%	1%	0,50%	0,10%
1	0,03157	0,03982	0,02393	0,0158	0,0642	0,275	0,708	1,642	2,706	3,841	6,635	7,879	10,828
2	0,0201	0,0506	0,103	0,211	0,446	1,022	1,833	3,219	4,605	5,991	9,21	10,597	13,816
4	0,297	0,484	0,711	1,064	1,649	2,753	4,045	5,989	7,779	9,488	13,277	14,86	18,467
6	0,872	1,237	1,635	2,204	3,07	4,57	6,211	8,558	10,645	12,592	16,812	18,548	22,458
8	1,646	2,18	2,733	3,49	4,594	6,423	8,351	11,03	13,362	15,507	20,09	21,955	26,125
10	2,558	3,247	3,94	4,865	6,179	8,295	10,473	13,442	15,987	18,307	23,209	25,188	29,588
15	5,229	6,262	7,261	8,547	10,307	13,03	15,733	19,311	22,307	24,996	30,578	32,801	37,697
20	8,26	9,591	10,851	12,443	14,578	17,809	20,951	25,038	28,412	31,41	37,566	39,997	45,315
30	14,953	16,791	18,493	20,599	23,364	27,442	31,316	36,25	40,256	43,773	50,892	53,672	59,703
40	22,164	24,433	26,509	29,051	32,345	37,134	41,622	47,269	51,805	55,758	63,691	66,766	73,402
50	29,707	32,357	34,764	37,689	41,449	46,864	51,892	58,164	63,167	67,505	76,154	79,49	86,661
60	37,485	40,482	43,188	46,459	50,641	56,62	62,135	68,972	74,397	79,082	88,379	91,952	99,607
70	45,442	48,758	51,739	55,329	59,898	66,396	72,358	79,715	85,527	90,531	100,425	104,215	112,317
80	53,54	57,153	60,391	64,278	69,207	76,188	82,566	90,405	96,578	101,879	112,329	116,321	124,839
90	61,754	65,647	69,126	73,291	78,558	85,993	92,761	101,054	107,565	113,145	124,116	128,299	137,208
100	70,065	74,222	77,929	82,358	87,945	95,808	102,946	111,667	118,498	124,342	135,807	140,169	149,449

Таблиця значень критерія Колмогорова

$$P\left\{\sqrt{n}\cdot D_n \geq \lambda\right\} = 1 - K(\lambda) = 1 - \sum_{k=-\infty}^{+\infty} (-1)^k e^{-k^2\lambda^2}$$

λ	α	λ	α	λ	α	λ	α
λ≤0,29	1,0000	0,47	0,9800	0,65	0,7920	0,83	0,4962
0,30	1,0000	0,48	0,9753	0,66	0,7764	0,84	0,4806
0,31	1,0000	0,49	0,9700	0,67	0,7604	0,85	0,4653
0,32	1,0000	0,50	0,9639	0,68	0,7442	0,86	0,4503
0,33	0,9999	0,51	0,9572	0,69	0,7278	0,87	0,4355
0,34	0,9999	0,52	0,9497	0,70	0,7112	0,88	0,4209
0,35	0,9997	0,53	0,9415	0,71	0,6945	0,89	0,4067
0,36	0,9995	0,54	0,9325	0,72	0,6777	0,90	0,3927
0,37	0,9992	0,55	0,9228	0,73	0,6609	0,91	0,3791
0,38	0,9987	0,56	0,9124	0,74	0,6440	0,92	0,3657
0,39	0,9981	0,57	0,9013	0,75	0,6272	0,93	0,3527
0,40	0,9972	0,58	0,8896	0,76	0,6104	0,94	0,3399
0,41	0,9960	0,59	0,8772	0,77	0,5936	0,95	0,3275
0,42	0,9945	0,60	0,8643	0,78	0,5770	0,96	0,3154
0,43	0,9926	0,61	0,8508	0,79	0,5605	0,97	0,3036
0,44	0,9903	0,62	0,8368	0,80	0,5441	0,98	0,2921
0,45	0,9874	0,63	0,8222	0,81	0,5280	0,99	0,2809
0,46	0,9840	0,64	0,8073	0,82	0,5120	1,00	0,2700

Відсоткові точки F-розподілу Фішера, якщо α=1%

K_2	K_1															
	2	3	4	5	6	7	8	9	10	15	20	24	30	40	60	120
2	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,43	99,45	99,46	99,47	99,47	99,48	99,49
3	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	26,87	26,69	26,60	26,51	26,41	26,32	26,22
4	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,20	14,02	13,93	13,84	13,75	13,65	13,56
5	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,72	9,55	9,47	9,38	9,29	9,20	9,11
6	10,93	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,56	7,40	7,31	7,23	7,14	7,06	6,97
7	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,31	6,16	6,07	5,99	5,91	5,82	5,74
8	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,52	5,36	5,28	5,20	5,12	5,03	4,95
9	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,96	4,81	4,73	4,65	4,57	4,48	4,40
10	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,56	4,41	4,33	4,25	4,17	4,08	4,00
15	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,52	3,37	3,29	3,21	3,13	3,05	2,96
20	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,09	2,94	2,86	2,78	2,69	2,61	2,52
30	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,70	2,55	2,47	2,39	2,30	2,21	2,11
40	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,52	2,37	2,29	2,20	2,11	2,02	1,92
60	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,35	2,20	2,12	2,03	1,94	1,84	1,73
120	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,19	2,03	1,95	1,86	1,76	1,66	1,53

Відсоткові точки F-розподілу Фішера, якщо $\alpha = 5\%$

K_1	K_2													
	2	3	4	5	6	7	8	9	10	20	30	40	60	120
2	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40	19,45	19,46	19,47	19,48	19,49
3	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66	8,62	8,59	8,57	8,55
4	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80	5,75	5,72	5,69	5,66
5	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56	4,50	4,46	4,43	4,40
6	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87	3,81	3,77	3,74	3,70
7	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44	3,38	3,34	3,30	3,27
8	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15	3,08	3,04	3,01	2,97
9	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94	2,86	2,83	2,79	2,75
10	4,10	3,71	3,47	3,33	3,22	3,14	3,07	3,02	2,98	2,77	2,70	2,66	2,62	2,58
15	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,33	2,25	2,20	2,16	2,11
20	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12	2,04	1,99	1,95	1,90
25	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,01	1,92	1,87	1,82	1,77
30	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	1,93	1,84	1,79	1,74	1,68
40	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,84	1,74	1,69	1,64	1,58
60	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,75	1,65	1,59	1,53	1,47
120	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,66	1,55	1,50	1,43	1,35

ДОДАТОК В - ТЕСТОВІ ЗАВДАННЯ

№	ПИТАННЯ Варіанти відповідей (A, B, C, D)			
1	Генеральна сукупність визначається як			
	A) сукупність, кожен елемент якої обраний випадково	**B)** набір значень, що реально спостерігаються і описують досліджувані об'єкти	**C)** усі мислимі значення, що описують досліджуємий об'єкт	**D)** нескінчену сукупність досліджуваних об'єктів, що досліджуються
2	Найбільш повно випадкова величина характеризується			
	A) середнім арифметичним	**B)** дисперсію і середнім арифметичним	**C)** функцію розподілу	**D)** коефіцієнтами асиметрії й ексцесу
3	Квантиль порядку p визначається як			
	A) значення функції розподілу, що відповідає величині p	**B)** значення x_p, що визначається з виразу $F(x_p)=p$	**C)** значення функції розподілу, нижче якого лежить p (%) розподілу	**D)** значення випадкової величини, вище якого лежить p (%) розподілу
4	Функція нормального розподілу визначається формулою			
	A) $\left(\frac{x-a}{b}\right)\exp(-\frac{(x-a)}{b^2}$	**B)** $\frac{1}{b}\exp(\frac{(x-a)^2}{b^2})$	**C)** $\frac{1}{b\sqrt{2p}}\exp(\frac{(x-a)^2}{2b^2})$	**A)** $\frac{1}{b}\exp(\frac{(x-a)}{b})$
5	Гістограма визначається як			
	A) функція розподілу випадкової величини	**B)** ступінчастий графік, де визначенні частоти попадання значень випадкової величини у певний інтервал	**C)** варіаційний ряд із вказівкою імовірності їхнього виникнення	**D)** статистичним розподілом випадкової величини

№	ПИТАННЯ Варіанти відповідей (A, B, C, D)			
6	Вибірковим аналогом математичного очікування є			
	A) абсолютна частота події	**B)** відносна частота події	**C)** вибіркове середнє значення випадкової величини	**D)** імовірність достовірної події
7	Які параметри визначають біноміальний закон, при n незалежних випробуваннях ?			
	A) математичне очікування	**B)** імовірність появи очікуваної події	**C)** інтенсивність потоку подій	**D)** частота появи очікуваної події
8	Які параметри має густина показового закону?			
	A) математичне очікування	**B)** дисперсія	**C)** границі множини значень	**D)** інтенсивність потоку подій
9	Статистична перевірка гіпотез дозволяє			
	A) Прийняти висунуту гіпотезу	**B)** Визначити правильність основної гіпотези з певною достовірністю	**C)** Визначити правильність основної гіпотези з певним рівнем значимості	**D)** Визначити яка гіпотеза вірна з певною імовірностю помилок 1-го і 2-го роду
10	З допомогою яких критеріїв можна найбільш повно перевірили гіпотезу щодо закону розподілу випадкової величини ?			
	A) G-критерій Кохрана	**B)** χ^2-критерій Пірсона, λ-критерій Колмогорова $n\omega^2$-критерій Мізеса-Смірнова	**C)** $n\omega^2$-критерій Мізеса-Смірнова t-критерій Стьюдента	**D)** t-критерій Стьюдента F-критерій Фішера

№	ПИТАННЯ Варіанти відповідей (A, B, C, D)			
11	Яка з наведенних формул відповідає потрібному критерію для перевірки гіпотези про рівність середніх арифметичних двох виборок при умові відомої генеральної диспесії σ^2			
	A) $\frac{\lvert\overline{x_A}-\overline{x_B}\rvert}{\sigma^2}\cdot\frac{1}{\sqrt{1/n}}$	**B)** $\frac{\lvert\overline{x_A}-\overline{x_B}\rvert}{\sigma}\cdot\sqrt{n}$	**C)** $(\sigma_A)^2/(\sigma_B)^2$	**D)** $\frac{\lvert\overline{x_A}-\overline{x_B}\rvert}{\sqrt{n}}\cdot\sigma$
12	З допомогою якого критерія можна перевірили гіпотезу щодо закону розподілу випадкової величини?			
	A) *t*-критерій Стьюдента	**B)** *F*-критерій Фішера	**C)** χ^2-критерій Пірсона	**D)** *U*-критерій Манна-Уітни
13	Для двох випадкових величин сила статистичного зв'язку визначаються			
	A) математичне очікування	**B)** кореляційний момент (коваріацією)	**C)** дисперсія	**D)** коефіцієнтм кореляції
14	Регресією *Y* на *X* називається			
	A) залежність *Y* від *X*	**B)** тіснота зв'язку *Y* і *X*	**C)** математичне очікування *M(X/Y)* при зміні *Y*	**D)** пряма, на якій лежать точки, що відповідають значенням *(X,Y)*
15	З того, що ковариация дорівнює нулю випливає, що			
	A) немає регресії	**B)** немає функціональної залежності	**C)** величини незалежні	**D)** немає лінійної кореляції

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

1. Айвазян С.А. Прикладная статистика. Основы эконометрики. В 2-х т. [Текст]:/С.А. Айвазян – М.:ЮНИТИ-ДАНА, 2001. – Т.1.- .-656с., Т.2 – 432с.
2. Боровиков В. STATISTICA. Искусство анализа данных на компьютере (+CD). Для профессионалов [Текст]:/ В. Боровиков. – 2-е изд., СПб: Питер, 2004. – 688 с.
3. Большев Л.Н. Таблицы математической статистики. [Текст]:/ Л.Н.Большев, Н.В.Смирнов. – М.: Наука, Главная редакция физ.-мат. литературы, 1983. – 416 с.
4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ [Текст]:/ Н. Дрейпер Н. Г. Смит Г. – М.: ФиС, 1986-1987. Кн.1. – 351 с., Кн.2. – 366 с.
5. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке: Методы планирования эксперимента [Текст]:/ Н.Джонсон, Ф.Лион –М.: Мир, 1981. –520 с.
6. Наследов А.Д. IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ даннях [Текст]:/ А.Д. Наследов. – Издательство Питер, 2013 – 416 с.
7. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере [Текст]:/ Ю.Н.Тюрин, А.А.Макаров. – М.: Инфра-М, 2003.- 544 с.
8. Халафян А.А. STATISTICA 6. Статистический анализ даннях [Текст]:/ А.А.Халафян. – М.:Бином-Пресс, 2007. – 512 с.
9. Шеффе Д. Дисперсионный анализ [Текст]:/ Д. Шеффе. – М.: Физматгиз, 1963. – 628 с.
10. Бендерский А.М. Оценка эффективности последовательных процедур обнаружения разладки методом имитационного моделирования / А.М.Бендерский, А.В.Томашевский // Изв.АН СССР. Техн. Кибернетика. – 1988. – №1. – С.193-194.
11. Томашевський О.В.Використання статистичних методів при сертифікаційних випробуваннях інтегрованих мікросхем / О.В.Томашевський, В.В.Погосов, Г.В.Сніжной // Радіоелектроніка, інформатика, управління. – 2009. – № 1. – С. 38-41.
12. Томашевский А.В. Компьютерные технологии анализа медицинской информации. Подготовка данных. Начальная

статистическая обработка. / А.В.Томашевский, В.А.Каширин // Ринологія. – 2010. – №1. – С.30-37.
13. Томашевский А.В. Статистический анализ медицинской информации. Определение нормальности распределения. Проверка статистической гипотезы. / Томашевский А.В., Каширин В.А., Воронцова Л. Л. // Сучасні медичні технології. – 2010. – № 1(5). – С.78-82.
14. Томашевский А.В. Анализ статистических методов обеспечения надежности сложных технических систем / Томашевский А.В., Тевс А.А. // Авиационно-космическая техника и технология. – 2011. – № 9(86). – С.116-119.
15. Каширин В.А., Томашевский А.В. Применение и описание статистических методов в клинических исследованиях / В.А.Каширин, А.В.Томашевский // Актуальні питання медичної науки та практики. Збірник наукових праць. – Запоріжжя:Дике поле, 2012. – Випуск 79, Т.1, кн.2. – С.22-29
16. Томашевский А.В.Статистический анализ качества изготовления турбинных лопаток / А.В.Томашевский, В.И.Фомичева // Авиационно-космическая техника и технология. – 2012. – № 7(94). – С.7-10
17. Программное обеспечение по статистическому анализу данных: методология сравнительного анализа и выборочный обзор рынка. / Айвазян С.А., Степанов В.С. [Електронний ресурс] – Режим доступу: http://ts1.cemi.rssi.ru/rus/publicat/e-pubs/ep97001t.htm
18. Электронный учебник STATSOFT [Електронний ресурс] – Режим доступу: http://StatSoft.ru