O.M. TSARENKO Y.A. ZLOBIN
V.G. SKLYAR S.M. PANCHENKO

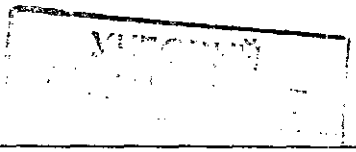
COMPUTER METHODS IN AGRICULTURE AND BIOLOGY

UNIVERSITY BOOK
SUMY
2000

О.М. ЦАРЕНКО Ю.А. ЗЛОБІН
В.Г. СКЛЯР С.М. ПАНЧЕНКО

КОМП'ЮТЕРНІ МЕТОДИ В СІЛЬСЬКОМУ ГОСПОДАРСТВІ ТА БІОЛОГІЇ

Допущено
Міністерством аграрної політики України
як навчальний посібник для студентів агрономічних
спеціальностей вищих аграрних закладів
освіти III-IV рівнів акредитації



СУМИ
2000

УДК 681.3:[631+57](075.8)
ББК 40.7я73
К 63

Рецензенти:
М.І. Стеблянко, кандидат сільськогосподарських наук,
професор, завідувач кафедри ботаніки СДПУ;
К.К. Карпенко, кандидат біологічних наук,
доцент кафедри ботаніки СДПУ

Друкується за рішенням Методичної ради СДАУ

Серія заснована в 2000 році

Комп'ютерні методи в сільському господарстві та біології:
К 63 Навчальний посібник/О.М. Царенко, Ю.А. Злобін, В.Г. Скляр, С.М. Панченко. - Суми: Видавництво «Університетська книга», 2000. - 203 с.

ISBN 966-7550-25-7

Навчальний посібник для студентів сільськогосподарських вузів III та IV рівнів акредитації та спеціалістів сільського господарства. Подаються відомості з математичної статистики, необхідні при комп'ютерному обробленні даних в галузі агрономії, зооінженерії, економіки сільського господарства, переробки продукції сільського господарства та ветеринарної медицини. Пропонується цикл практичних робіт, які дозволяють оволодіти навичками обробки даних польових дослідів і спостережень на основі найбільш поширеного і потужного статистичного пакета Statistica for Windows 5.0. Розглядаються методи програмування динамічних процесів, необхідні для оцінки сільськогосподарських технологій в умовах ринкової економіки.

ББК 40.7я73

ISBN 966-7550-25-7

ВСТУП

Математичне оброблення результатів польових дослідів, обліків і спостережень на базі сучасної комп'ютерної техніки є необхідною складовою будь-якого сільськогосподарського та біологічного дослідження. Вимоги до сучасних дипломних і дисертаційних робіт, досліджень, публікацій в журналах, а тим більше до книг і брошур неодмінно передбачають комп'ютерне опрацювання кількісних показників. На сьогодні таке оброблення ведеться за допомогою комп'ютерів на основі спеціальних пакетів, які звичайно мають досить широкій набір методів математичної статистики.

Набір пакетів математичної статистики на світовому ринку дуже великий і різноманітний. Та деякі з них взагалі не придатні для обробки матеріалів, отриманих під час проведення сільськогосподарських і біологічних досліджень, тому що призначені для інших спеціальностей, не мають усіх необхідних статистичних методів і орієнтовані на обробку даних, які отримуються спеціально, наприклад, результатів соціологічних досліджень. Деякі із таких пакетів взагалі не дозволяють використовувати графіку і кирилицю.

Але є й статистичні пакети, які не мають цих недоліків. У тій чи іншій мірі фахівці сільського господарства та біологи мають змогу використовувати такі пакети.

Statgraphics for DOS. Це один з найбільш розповсюджених у 1980-1990-х роках пакетів. Його переваги – це широкий набір методів математичної статистики, зручний інтерфейс і можливість застосування на простих комп'ютерах з невеликим обсягом оперативної пам'яті та малою ємністю вінчестера. На жаль, працює він повільно, графіка погано піддається масштабуванню, кирилиця не підтримується.

Statgraphics for Windows. Це той же пакет за вмістом математичних методів, але його інтерфейс повністю змінений фірмою SPSS. Працює він тільки в середовищі *Windows*. На відміну від DOSівської версії пакет має доволі потужного "Радника" (Adviser). Система вікон не дуже зручна, під час друку графіки її важко розмістити на папері у тому місці, де потрібно, і навіть виконання підписів у таблицях і графіках кирилицею складає непросту проблему.

SPSS. За набором методів математичної статистики це, мабуть, один з самих потужних пакетів. Робота з ним вимагає високої кваліфікації користувача. Слабким місцем пакета є графіка і

кирилиця. Не дуже зручним є і перегляд результатів на розповсюджених у нас моніторах з діагоналлю 14 і 15 дюймів.

SigmaStat. Пакет порівняно малої потужності. В звичайному користуванні мало зручний. Єдиною його перевагою є повна сумісність з графічними пакетами **SigmaPlot**, які розробленні цією фірмою.

MiniTab. Один з перших пакетів, які працюють в операційній системі **Windows**. Його інтерфейс зручний, але є проблема з графікою і кирилицею. З цього пакета повністю відсутня література, що ускладнює його використання дослідниками-початківцями.

Excel. Користується популярністю, тому що входить до складу широко розповсюдженого MSOffice. Це комерційні та бухгалтерські електронні таблиці, але в нього вмонтований невеликій набір програм для статистичної обробки матеріалу. Набір цих методів обмежений, частина їх не відповідає природі кількісних даних, які отримуються в умовах польових і лабораторних дослідів (що природно для бухгалтерських електронних таблиць). Для використання в біології, агрономії, зооінженерії та механізації сільського господарства його рекомендувати не можна. Але цей пакет може успішно використовуватись економістами, бухгалтерами і спеціалістами в галузі фінансів. Його популярність обумовлена тим, що він русифікований.

Statistica for Windows. Другий за розповсюдженістю після *Statgraphics for DOS пакет*. Сьогодні використовуються його п'ята та шоста версії. При конкурсному оцінюванні, яке проводиться провідними комп'ютерними журналами, він незмінно посідає перше місце. Переваги у пакета значні: в нього дуже зручний інтерфейс, графіки і таблиці легко масштабуються під час друку, всі підписи можна виконувати кирилицею, тобто українською або російською мовами. За набором методів математичної статистики він не має собі суперників. Пакет *Statistica for Windows* відповідає міжнародним стандартам відносно статистичної обробки матеріалу. З цього пакета існує література на російській мові.

З урахуванням переваг і недоліків різних пакетів, які можуть використовуватись для оброблення масових сільськогосподарських і біологічних даних, в цьому навчальному посібнику всі практичні роботи орієнтовані на пакет **Statistica for Windows**. В кінці книги подається переклад основних англійських термінів, які використовуються в пакеті "Статистика", на українську мову. Освоєння цього пакета дозволяє порівняно легко перейти до будь-якого з інших пакетів, тому що принципи їх орієнтації, як і основні методи, або ж подібні, або ж взагалі ідентичні.

Пропонований навчально-методичний посібник вміщує матеріал, який дозволить освоїти основні методи математичної стати-

стики, які застосовуються в сучасній біології та сільському господарстві, і навчитися використовувати їх за допомогою комп'ютерної техніки. Ряд практичних робіт спрямований на оволодіння навичками побудови зведених таблиць, графіків і схем, необхідних як ілюстративний матеріал в курсових, дипломних, магістерських і дисертаційних роботах. Даний навчальний посібник відповідає програмі навчального курсу "Комп'ютерні методи обробки даних", який вводиться на старших курсах усіх спеціальностей сільськогосподарських вузів.

Передбачається, що студент, який розпочинає роботу з курсу "Комп'ютерні методи обробки даних", уже засвоїв загальний курс "Інформатики" й ознайомлений з основними прийомами роботи з комп'ютером і операційними системами: DOS, Norton Commander, Windows 3.1, Windows 95 та Windows 98.

Для поновлення і закріплення навичок в роботі з комп'ютером виконайте практичну роботу № 1.

ОСНОВНІ ОСОБЛИВОСТІ ПАКЕТА STATISTICA

Статистичний пакет Statistica є представником сучасних комп'ютерних програм, побудованих на основі нових технологій обробки даних. Він спрощує і прискорює звичайні рутинні операції та дозволяє користувачеві зосередитись на розумінні характеру даних і поясненні результатів їх статистичної обробки. Пакет не потребує від користувача знання тонкощів математичної статистики та всіх вживаних формул, але, звичайно ж, елементарні базові знання методів математичної статистики та сфери їх застосування необхідні для свідомого користування ним.

Пакет Statistica був створений фірмою StatSoft (США) у 1984 році і неодноразово удосконалювався. На сьогодні вже створена шоста версія пакета, але найбільш розповсюдженою є версія 5.0 (вийшла у 1995 році) і в запропонованому посібнику використовується саме вона. Можна користуватись також версіями 4.5 і 6.0 – між ними та версією 5.0 немає принципових відмінностей.

Пакет Statistica вигідно відрізняється такими перевагами:

1. Має модульну структуру і дозволяє вести обробку матеріалу в окремих модулях, що значно зменшує вимоги до обсягу пам'яті комп'ютера.

2. За рахунок підтримки механізму перенесення даних, а також механізму сполучення та впровадження об'єктів (OLE) пакет дозволяє вести обмін даними з іншими прикладними програмами. Зокрема, результати розрахунків і графіки легко переносяться й вставляються у текст, набраний в найбільш популярному редакторі – Word.

3. Пакет Statistica практично не має обмежень на об'єм числової чи текстової інформації, яка вводиться. Якщо такі масиви дуже великі, то для їх введення і обробки передбачено спеціальний модуль "Менеджер мегафайлів".

4. Підтримується кирилиця, що забезпечує виконання всіх надписів до графіків і діаграм українською чи російською мовами.

5. Має потужну і зручну систему побудови графіків і діаграм, набір видів яких практично не обмежений.

6. Має систему вікон для введення даних і виводу результатів з переходом від вікна до вікна одним натиском мишки.

7. У пакет вмонтована потужна система допомоги користувачу.

Єдиним недоліком пакета Statistica є те, що на сьогодні найбільш розповсюдженим є його не русифікований варіант. Хоча фірма Stat Soft уже почала випускати русифіковану версію "Статистики".

У пакеті Statistica передбачено три основні типи вікон:

Spreadsheet – електронна таблиця для вводу даних.

Scrollsheet – вікно для виводу результатів обробки у вигляді таблиці чи тексту.

Graphic – вікно для виводу графіків.

Зміст кожного із вікон можна зберегти у вигляді самостійного файла (файли з табличними даними отримують розширення *.sta*, файли з результатами статистичної обробки даних, які мають вид таблиць, – розширення *.scr*, а файли з графікою – розширення *.stg*), а також направляти зміст будь-якого активного вікна на принтер для отримання твердої копії.

У склад пакета Statistica входять такі основні модулі, котрі можна викликати нарізно для обробки матеріалу в залежності від того, який метод оброблення вам необхідний:

1. **Basic Statistics/Tables** (Основні статистики/Таблиці). Він вміщує широкий набір методів первинної обробки матеріалу – знаходження точкових і інтервальних оцінок для статистичних рядів, кореляцію, критерій Ст'юдента, однофакторний дисперсний аналіз та ін. Модуль завантажується командою Sta_bas.exe.

2. **Multiple Regression** (Множинна регресія) має широкий набір методів лінійної і нелінійної регресії, включаючи покроковий регресійний аналіз і часткову регресію. Модуль завантажується командою Sta_reg.exe.

3. **ANOVA/MANOVA** (Однофакторний і багатофакторний дисперсійний аналіз). У цьому модулі наведений широкий набір методів дисперсійного аналізу, який дозволяє проводити всебічну оцінку результатів польового досліду, виконаного за самою складною схемою. Модуль завантажується командою Sta_man.exe.

4. **Nonparametrics/Distribution** (Непараметричні методи/Розподіли). Модуль містить в собі набір методів непараметричної статистики, які застосовуються для обробки даних, що не відповідають нормальному статистичному розподілу. Модуль завантажується командою Sta_non.exe.

5. **Factor Analysis** (Факторний аналіз). Служить для виявлення загальних факторів, які впливають на явища і структури, що ви спостерігаєте. Модуль завантажується командою Sta_fac.exe.

6. **Reliability and Item Analysis** (Аналіз надійності) має широкий набір методів, які застосовуються у промисловій статистиці, в тому числі засоби аналізу результатів діагностики окремих видів

механізмів і машин та якость продукції. Завантажується модуль командою Sta_rel.exe.

7. **Cluster Analysis** (Кластерний аналіз) вміщує широкий набір методів для групування об'єктів за рядом кількісних ознак у подібні групи. Модуль завантажується командою Sta_clu.exe.

8. **Time Series/Forecasting** (Ряди динаміки/Прогнозування) призначений для обробки рядів динаміки і прогнозування процесів. Модуль завантажується командою Sta_tim.exe.

9. **Experimental Design** (Планування дослідів і експериментів). Має розгалужений набір прийомів для планування польових дослідів будь-якої складності, включаючи дробові репліки та композиційні плани. Є також спеціальні методи планування економічних спостережень і промислових експериментів. Модуль завантажується командою Sta_exp.exe.

Окрім цього, пакет Statistica вміщує ряд інших методів (їх загальна кількість дорівнює 23) для проведення багатовимірного статистичного аналізу, для контролю за якістю продукції статистичними методами та ін. Користувач може вільно переключатись між модулями із будь-якого із них, а також викликати будь-який модуль на початку роботи.

Система вводу даних у пакет дуже гнучка. Їх можна вводити в електронні таблиці безпосередньо з клавіатури. При необхідності зміни первинних даних їх можна перетворювати за допомогою формул. Можна перенести у пакет дані з другої прикладної програми, для чого існує послідовність команд File-Import Data. Значні можливості маються й для виводу даних практично в будь-якому з форматів. Для цього використовують послідовність команд File-Export Data.

Найчастіше користувачі в комп'ютерних лабораторіях вищих навчальних закладів і науково-дослідних центрів звертаються до пакета Statistica, інсталяція якого й загальна настройка вже проведена програмістом-фахівцем. За наявності деяких навичок таку настройку можна зробити й самому.

Щоб оволодіти методами роботи з пакетом "Статистика", необхідно послідовно виконувати наведені нижче практичні роботи. Крім того, суттєву допомогу може надати знайомство з книгою В.П.Боровикова та І.П.Боровикова " Statistica. Статистический анализ и обработка данных в среде Windows" (М.: Филинъ, 1997), в якій докладно подано опис інсталяції пакета, настройки і способів виконання основних видів статистичного аналізу.

Пакет Statistica має дуже потужний і просто написаний Help (Допомога), але він вимагає знання англійської мови.

Для ознайомлення з інтерфейсом пакета "Статистика" виконайте практичну роботу № 2, а для оволодіння навичками створення комп'ютерної бази даних – практичну роботу № 3.

ГРАФІЧНІ МОЖЛИВОСТІ СИСТЕМИ STATISTICA

Статистичний пакет "Статистика" дає можливість не тільки проводити різнобічні комп'ютерні розрахунки, але й створювати розмаїття графічних матеріалів – графіків і діаграм. У 1995 році за цим показником пакет "Статистика" посів перше місце на відкритому міжнародному конкурсі статистичних і графічних програм. Пропонована пакетом "Статистика" можливість візуалізації даних і результатів розрахунків дуже важлива, тому що дозволяє наочно представити особливості процесів і явищ і дати їм оцінку.

Всі графіки пакет "Статистика" виводить в окремому графічному вікні, яке постійно зв'язане з електронною таблицею, і таким чином усі зміни в таблиці даних автоматично відображаються в графіках. При відкритті вікна графіки панель інструментів змінюється – в ній з'являється велика кількість нових піктограм, які дозволяють виконувати в графіці велику кількість перетворень. Самі по собі графіки не тільки дуже різноманітні, але й піддаються настройкам по всім уявним параметрам: розміру, пропорціям, шрифтам, товщині ліній, конфігурації і багато іншому. Створені графіки можна зберігати окремо від таблиць у файлах з розширенням .STG.

Важливою перевагою пакета "Статистика" є можливість виконувати всі надписи на діаграмах і графіках, використовуючи кирилицю, тобто українську і російську мови. При підготовці графіків дуже зручним є режим збільшення, який включається через піктограму **(збільшувальне скло)** і дозволяє в більшому масштаб: відредагувати найтонші або ж захаращені частини графіка чи діаграми. Після цього за рахунок режиму зменшення **(Graphics Zoom Out)** графік легко повернути до потрібного розміру. Є можливість розміщення заголовків графіка в будь-якому необхідному місці, а також зробити підписи до осей графіка. Користувач також може вставити будь-які підписи прямо в середину графіка чи діаграми. Мається панель інструментів для малювання, яка дозволяє вставляти в графік доволі складні фігури за допомогою миші (звичайно, якщо у вас тверда рука і ви вмієте малювати).

Для друку графіки легко переносяться у текстові процесори (наприклад, у редактор MSWord) за допомогою механізму OLE. На одному аркуші паперу можна розмістити як один графік, так і декілька – вони масштабуються. Мається спеціальна піктограма

Mapping Options (Розміщення), яка дозволяє встановлювати положення графіка на аркуші паперу під час його друку.

В залежності від потреб користувача пакет "Статистика" репрезентує три альтернативні шляхи створення графіків: а) **Quick stats Graph** (швидкі статистичні графіки) – для створення експрес-методом порівняно простих графіків і діаграм; б) **Stats Graph** (статистичні графіки) – для побудови спеціальних графіків, пристосованих до різних типів статистичної обробки даних і в) **Custom Graphs** (графіки користувача) – коли графік створюється не з усього масиву даних, а із попередньо вибраного блоку даних. Вибір типу графіка програмується тим, що на піктограмах наведені схематичні рисунки цих графіків.

Графічна система не тільки дає можливість візуалізувати дані як такі, але й дозволяє проводити безпосередньо в графічному режимі ряд складних обчислень: підгонку кривих, співставлення дослідних даних з кривими різних статистичних розподілів і т.п. За допомогою особливого механізму **Brushing Tool**, його називають "Пензлик", стає можливим безпосередньо на графіку помітити деякі точки, і тоді вони будуть виключені із аналізу.

Для переходу в режим побудови ілюстрацій можливі різні шляхи. Можна увійти в цей режим через опцію **Graphs** у верхній панелі піктограм, а можна використати піктограму **Statistica Graph Gallery** (Галерея графіків). Мається також окрема панель з піктограми графіки, яку користувач може розмістити в будь-якому місці основного робочого вікна (звичайно вона розміщується зліва).

При частому виконанні графічних робіт в межах пакета "Статистика" його можна налаштувати на свої потреби. Для цього при відкритому вікні будь-якого графіка потрібно вибрати з верхнього меню опцію **Option** (Опції) **Global default** (Загальні установки по умовчанню), які будуть зберігатись в майбутньому до їх зміни за будь-яких сеансів роботи.

Для засвоєння навичок візуалізації підсумків статистичного аналізу й створення ілюстрацій різного типу за допомогою пакета "Статистика" виконайте практичну роботу № 13.

ТЕОРЕТИЧНА ЧАСТИНА

РОЛЬ МАТЕМАТИЧНОЇ СТАТИСТИКИ В СІЛЬСЬКОГОСПОДАРСЬКИХ І БІОЛОГІЧНИХ ДОСЛІДЖЕННЯХ

У сучасному сільському господарстві та біології рідко можна зіткнутися з дослідженням, результати якого мали б суто якісний характер. У переважній більшості випадків результатом праці є кількісні показники: величини врожаїв, продуктивність тварин, результати вимірів розмірів рослин, параметри, які характеризують властивості грунту, і т.п. Їх оцінка, згортка, комплектування та інтерпретація проводиться на основі методів математичної статистики.

Математика і математична статистика, зокрема, вже внесли свій суттєвий внесок у такі науки, як фізика, астрономія, хімія та ін. Бурхливий розвиток цих наук і прикладних галузей, які спираються на них (атомна енергетика, виробництво пластика), були наслідком математизації.

Аналіз розвитку практично всіх наук показує, що вони проходять три етапи: *емпіричний* – накопичення дослідних і описових даних; *теоретичний* – це якісний синтез і узагальнення емпіричних даних, і *математико-статистичний*, на рівні якого досліджуються кількісні закономірності й створюються точні математичні моделі структур і процесів, що є об'єктом даної науки. На порозі такої математизації знаходиться й сільське господарство, яке на межі XXI сторіччя все більше застосовує кількісні підходи і комп'ютерні технології. В біологічних дослідженнях масове використання методів математичної статистики відбувається вже півтора десятиріччя.

Стають все більш справедливими слова видатного вченого і художника епохи Відродження Леонардо да Вінчі: "Ніякої достовірності немає в науках там, де не можна застосувати жодної з математичних наук". Справді, застосування математичної статистики вносить у сільське господарство і біологію:

а) точність і однозначність матеріалів і висновків;

б) дає можливість оцінювати ступінь вірогідності й надійності всіх висновків і пропонувань;

в) дозволяє глибше проникати в сутність науково-виробничих завдань, виявляти раніше невідомі закономірності

і ставити нові проблеми, які раніше не стояли перед сільським господарством.

Важливо мати на увазі, що математична статистика як наука виходить із потреб практики. Так, один із класиків математичної статистики, який розробив, зокрема, метод дисперсійного аналізу, - Рональд Фішер виконав усі свої дослідження, коли працював на Ротамстедській сільськогосподарській дослідній станції (Англія) і очолював відділ селекції в Кембріджському університеті. Не випадково, таким чином, більшість методів математичної статистики є ідеально пристосованими для аналізу економічних і агрономічних проблем та проблем тваринництва. В сучасних умовах на шляху математизації й комп'ютеризації все ж стоїть немало перешкод. Головні з них такі:

1. В студентів і практиків сільського господарства часто-густо виявляється слабкою загальна математична підготовка, яка перешкоджає швидкому розумінню сутності методів, що використовуються при комп'ютерній обробці матеріалу. Вона ж заважає й правильно вибрати метод, який би відповідав поставленому завданню. Слабке володіння ідеями та методами математичної статистики в кінцевому підсумку інколи приводить до того, що комп'ютерна обробка матеріалу перетворюється в деякий "косметичний" засіб для надання "пристойного виду" тим чи іншим роботам.

2. Користувачі не знають повного набору сучасних статистичних пакетів і відчувають скруту при виборі методу, який би найбільше відповідав розв'язанню конкретної задачі.

3. Комп'ютерній, як і суто ручній обробці вихідного матеріалу, нерідко заважає низька якість цих початкових даних. Вони не витримують елементарних вимог до того, якими мають бути ці кількісні дані, якщо планується їх обробка методами математичної статистики.

4. Користувачі гадають, що ЕОМ "думає" за них. У такому випадку в комп'ютер уводять безглузді дані і отримують такі ж безглузді відповіді.

5. У зв'язку з тим, що на сучасному ринку поки що майже немає статистичних пакетів, перекладених на російську чи українську мови, користувачі із поганим знанням англійської мови відчувають значні труднощі, хоча набір команд на англійській мові, необхідний для роботи із статистичним пакетом, незначний і ним не складно оволодіти.

6. Виявляється й інерційна орієнтація на спеціалістів сільського господарства старшого покоління, які в свій час не мали доступу до комп'ютерної техніки і вели виробництво в інших умовах,

спираючись на спрощені схеми розрахунків, що робились вручну. Але відчуття, що "так було завжди і так буде в майбутньому", є ілюзією, яка знаходиться в повному протиріччі з міжнародними стандартами ведення сільськогосподарських і біологічних досліджень сучасного типу.

Та все ж, не зважаючи на комплекс цих труднощів, математизація й комп'ютеризація всіх галузей сільського господарства - процес неминучий. Посилання на використання комп'ютерних технологій у сільськогосподарських працях стає все частіщим, хоча вони ще інколи й носять "косметичний" характер. В.П.Леонов і П.В.Іжевський (1998) у зв'язку з цим справедливо відзначили, що метою авторів є спроба за допомогою статистичної термінології та найменувань "комп'ютерів надати роботі більш респектабельний і вагомий вигляд". Зрозуміло, якість досліджень від цього не покращується і прогрес технологій не прискорюється. Математична статистика і комп'ютерні технології у XXI сторіччі є неодмінною умовою прогресу сільського господарства, але тільки за умови їх умілого застосування.

ПОНЯТТЯ ПРО СТАТИСТИЧНІ РЯДИ. КОМП'ЮТЕРНІ БАЗИ ДАНИХ

Необхідність застосування методів математичної статистики в сільському господарстві та біології полягає в самій суті цих професій. Об'єктами фахової діяльності в цих галузях є в першу чергу живі організми – рослини у природних популяціях і посівах, які мають властивість формувати фітомасу і врожай, і тварини – носії функцій продуктивності. А для всього живого властива мінливість, тобто варіювання організмів по формі, величині та багатьом іншим ознакам. При цьому в загальній амплітуді мінливості одні значення ознак спостерігаються частіше, тобто з більшою імовірністю, а другі рідше – з меншою імовірністю. Імовірність появи деяких ознак може бути такою незначною, що вимагає для виявлення цих ознак збору доволі значного матеріалу з обстеженням десятків і сотень рослин, тварин чи господарств. Тільки методи математичної статистики в цій ситуації дають можливість правильно оцінити значення ознак, визначити амплітуду їх варіювання й вичислити можливість їх виявлення за тих чи інших обставин.

В усіх випадках обстеження рослин і посівів агрономом, тварин зооінженером чи ветеринаром фахівець в якісній чи частіше за все в кількісній формі враховує ті чи інші ознаки цих об'єктів. **Ознака – це будь-який кількісний або якісний параметр, будь-яка властивість рослин, тварин, грунтів, полів чи виробничих механізмів**. При одноразовому обчисленні ознаки ми отримуємо деяке її значення – x_i. В силу непостійності будь-яких об'єктів, з якими має справу спеціаліст сільського господарства, однократний облік ознаки не розкриває її типового значення і можливої амплітуди варіювання. Тому необхідний повторний облік цієї ознаки в об'єктах цієї ж категорії. Такі повторні обчислення складають набір даних, які називають **статистичним рядом**. Його членами будуть $x_1, x_2, x_3, ..., x_n$.

Можна собі уявити такий випадок, коли обчисленням охоплюється увесь набір об'єктів даного типу. Тоді обчислені дані складають **генеральну сукупність**. У генеральних сукупностях значення облічувальної ознаки частіше всього розподіляються строго закономірно - деякі із них зустрічаються частіше, інші - рідше. Якщо частоту зустрічей значень виражати як імовірність,

то генеральній сукупності буде відповідати статистичний ряд, властивий **нормальному статистичному розподілу**. Загальний вигляд такого ряду представлений на рис.1. На горизонтальній осі цього графіка показані можливі числові значення, які може приймати змінна в генеральній сукупності, а на вертикальній – імовірність у частках одиниці.

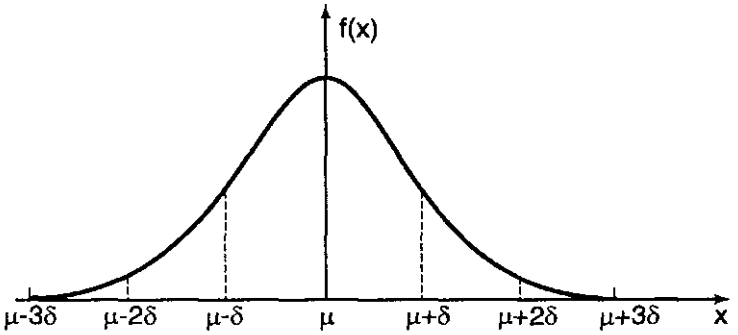


Рис. 1. Крива нормального статистичного розподілу

Але генеральні сукупності частіше всього такі великі, що їх не можна охопити обліком. Кількість рослин даного сорту, наприклад, незчисленно велика. Тому доводиться брати для обліку деяку частину генеральної сукупності – вибірку. **Вибірка – це частина генеральної сукупності, взята для вивчення**. Звичайно фахівець сільського господарства має справу зі статистичними рядами, які характеризують не генеральну сукупність, а вибірку.

Вибірка – це набір значень урахованої ознаки. Ці значення попадають у вибірку у тому ж порядку, в якому фахівець робив вимірювання й обліки і заносив їх результати в журнал обліку чи відомість. При комп'ютерній обробці матеріалу ці значення можуть уводитись в комп'ютер у цьому ж порядку. Але інколи для зручності значення ознаки можуть розміщуватися в новому порядку: від менших значень до великих або ж навпаки. Таку процедуру, коли первинні дані статистичного ряду розміщуються упорядковано, називають **ранжуванням**, або **сортуванням**. У пакеті Statistica є можливість автоматичного проведення сортування в порядку, який встановлює користувач.

Якщо облікові дані мають якісний характер (наприклад, це може бути стать рослин – тичинкові або маточкові форми, колір квіток або плодів і т.д.), то їх можна вводити в комп'ютер у текстовій

формі. Для виконання розрахунків з такими даними нерідко виконується **процедура кодування**, при якій кожна якісна ознака замінюється цифрою. Наприклад, маточкові квітки будуть позначатися цифрою 1, а тичинкові – цифрою 2. Процедура кодування часто полегшує й прискорює впровадження якісних даних, так можна спочатку ввести короткі коди, а потім їх замінити необхідними термінами.

Слід мати на увазі, що фахівця сільського господарства завжди цікавлять властивості генеральної сукупності, а не вибірки. Так, випробувавши на дослідній ділянці дію нового гербіциду на бур'яни, які там ростуть, агроном вважає, що даний гербіцид буде придушувати такі ж бур'яни на будь-яких полях, а не тільки на даній дослідній ділянці. Якщо спеціаліст визначив на прикладі деякої вибірки, що деяка нова порода тварин продуктивніша від старої, він сподівається, що нова порода виявить властивість підвищеної продуктивності в цілому, а не тільки в групі тварин, яку він вивчав. Все це означає, що **вибірка має бути такою, щоб вона цілком відображала властивості генеральної сукупності**. Інакше вся дослідна справа стане безглуздою.

В математичній статистиці і дослідній справі розроблені певні правила, дотримуючись яких одержують показові, надійні вибірки. Тільки до таких вибірок має сенс застосовувати комп'ютерні методи обробки.

Основні правила отримання вибірок у сільському господарстві такі:

1. Рандомізація, тобто відбір зразків для вимірювання проводиться у випадковому порядку. Випадковість означає, що кожний зразок як член генеральної сукупності має імовірність попасти у вибірку і ця імовірність дорівнює частоті зустрічі з ним у генеральній сукупності. Наприклад, у полі є 5000 маків з білими квітками і 5000 – з червоними. Всього 10 тисяч рослин. Для дослідження вирішено взяти вибірку всього з 10 рослин. У цьому випадку імовірність зустрічі рослин з червоними і білими квітками однакова і дорівнює $P=5000/10000=0,5$. Таким чином, при обсязі вибірки у 10 рослин до неї повинні попасти 5 рослин з червоними і 5 рослин з білими квітками. Тільки тоді дана вибірка буде правильно відображати генеральну сукупність.

Але так просто звичайно не буває: дослідник не знає особливостей генеральної сукупності і його завдання якраз полягає у тому, щоб встановити ці особливості за вибіркою. Так, агронома цікавить довгостебельність нового сорту пшениці. Потрібно вибрати для виміру деяку кількість рослин. Але якщо вибирати суб'єктивно типові для даного посіву рослини, то можна непомітно для себе віддати

перевагу або низьким, або, навпаки, високим рослинам. **Вибір типового завжди суб'єктивний і не точний**. Об'єктивність може забезпечити тільки випадковий відбір. Для реалізації рандомізованого – випадкового відбору існують спеціальні таблиці. Сучасні кишенькові калькулятори часто генерують такі числа. Генерує їх і пакет Statistica.

В країнах Західної Європи і США рандомізація є "залізним" правилом. При цьому для її забезпечення реалізують декілька варіантів: а) повний випадковий вибір, б) відбір за системою, тобто, наприклад, кожна десята рослина, в) пошаровий випадковий вибір, коли обстежувану сукупність заздалегідь поділяють на декілька підсукупностей і випадковий вибір роблять у кожній з них; кількість зразків, яка береться з кожної під сукупності, залежить від її відносної величини. Вибір за системою і пошаровий вибір - об'єктивні, але часто не точні. **Об'єктивність у поєднанні з точною відповідністю вибірки генеральній сукупності забезпечує тільки випадковий відбір зразків для аналізу**.

2. Якісна однорідність матеріалу, яка полягає в тому, що всі досліджувані зразки мають однакову категорію якості. Вона визначається фаховими знаннями і сумлінністю спеціаліста. Звичайно, що при вивченні будь-яких особливостей рослин деякого сорту всі вони повинні відноситись до даного сорту. У вибірку не повинні попасти рослини іншої сортової належності. При відборі зразків грунту із орного горизонту для визначення кислотності в число зразків не повинен попасти грунт з-під орного горизонту і т.п.

Такі "чужаки" в вибірках часто мають кількісні показники, які різко виділяються із загальної маси і на графіках "вискакують" за межі кривої нормального статистичного розподілу. Пакет Statistica має спеціальний підмодуль – "нормальний імовірнісний аркуш", який дозволяє виявляти такі вискакуючі значення і при необхідності вилучати їх із вибірки. При реалізації цієї процедури на нормальному імовірнісному аркуші проводиться пряма лінія, а облікові значення кладуться вздовж неї. Якщо вони всі попали на пряму лінію, то вибірка відповідає нормальному статистичному розподілу, якщо ж вони далеко відхиляються від неї, то необхідне вилучення даних, які попали помилково, а то й повторний облік у полі чи на фермі. На рис. 2 наведено приклад перевірки вибірки на нормальному імовірнісному аркуші, видно окремі "вискакуючі" значення. Пакет Statistica має й напівнормальний імовірнісний аркуш. Його використовують тоді, коли необхідно проаналізувати тільки позитивну частину кривої нормального розподілу, тобто тоді, коли дослідника цікавлять тільки самі залишки без урахування їх знаків.

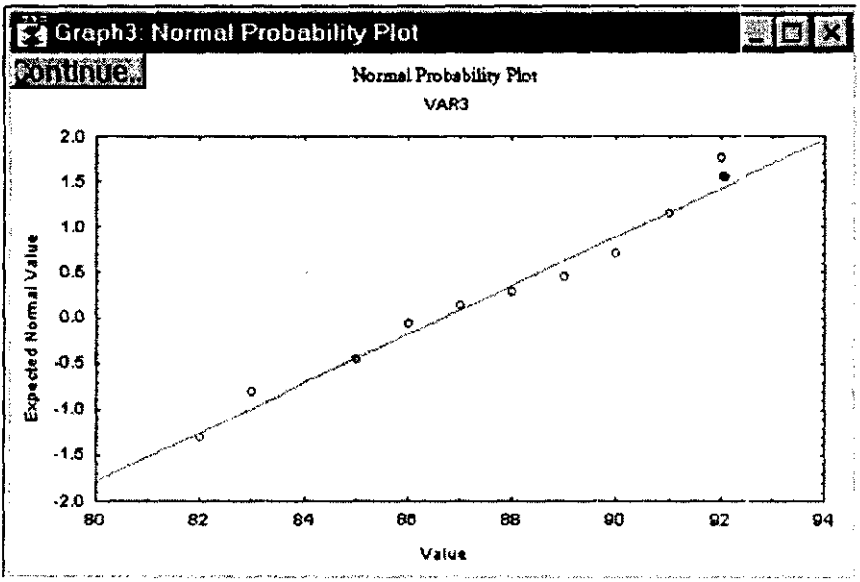


Рис. 2. Приклад аналізу статистичного ряду на відповідність нормальному статистичному розподілу на нормальному імовірнісному аркуші

3. Репрезентативність полягає в тому, що вибірка повинна відображати генеральну сукупність у всіх її важливих особливостях. Певна річ, чим більша величина вибірки, тим шанси на те, що вона відображає властивості генеральної сукупності більше. Але отримання і обробка дуже великих вибірок вимагає значних затрат праці, часу, а часто й коштів. Тому доводиться зменшувати величину вибірок.

Формально вибірки ділять на великі й маленькі. Великі вибірки – це вибірки, які містять більше 20-30 зразків чи позторних вимірювань. Маленькі вибірки – менше 20.

Теоретично відомо, що необхідний обсяг вибірки залежить від ступеня мінливості враховуваної ознаки. Чим більше варіює ознака, тим більше буде необхідно брати зразків для її об'єктивної оцінки. Тож, для того щоб встановити кількість тичинок у квітці жита, необхідно подивитись всього одну-дві рослини. Тичинок завжди три і, крім явних каліцтв, їх кількість не змінюється. Але якщо враховувати кількість квіток у колоску, то доведеться взяти набагато більшу вибірку.

Необхідний обсяг вибірки знаходять за формулою:

$$\hat{n} = \frac{t^2_{0,05} \cdot v^2}{\varepsilon^2},$$

де $t_{0,05}$ *– значення критерію Ст'юдента для рівня значущості 0,05, що дорівнює 1,96;*

v *– коефіцієнт варіації;*

ε *- допустима погрішність у відсотках (у сільському господарстві її звичайно приймають рівною 1,3 або 5%).*

Для використання цієї формули спочатку потрібно встановити мінливість ознаки у вигляді коефіцієнта варіації.

Звичайно розрахунки за цією формулою показують, що вибірки повинні бути великими і містити більше 20 повторень. На жаль, в агрономії і тваринництві досліди нерідко проводять у повторності всього 3-4. Це часто дуже мала повторність для виявлення особливостей генеральної сукупності й отримання статистично достовірних висновків. Тому для опрацювання матеріалів таких дослідів потрібно відноситись дуже пильно.

При комп'ютерній обробці матеріалу перш за все в спеціальні таблиці – їх називають табличний процесор, або електронна таблиця – вводять ті дані, які утворюють статистичні ряди. Статистичний ряд для самостійної ознаки складає **змінну (Variable)**, а члени статистичного ряду утворюють **повторення**, або випадки **(Cases)**. Пакет Statistica має зручні електроні таблиці й широкий сервіс для роботи з ними. Набір даних у вигляді файлів, які містять змінні і повторення, утворює **комп'ютерну базу даних**. Такі бази даних можуть зберігатися на дискетах, їх можна переносити на інші комп'ютери і поповнювати по мірі появи нового матеріалу.

Для того, щоб оволодіти навичками створення електронної бази даних і перевірки статистичних рядів на відповідність нормальному статистичному розподілу, виконайте практичні роботи № 3 і 4.

СТАТИСТИЧНА ХАРАКТЕРИСТИКА ВИБІРКИ

При вивченні тих чи інших процесів чи явищ фахівець сільського господарства може охопити всі об'єкти того чи іншого виду. Частіше всього це буває в економіці, наприклад, коли обстежуються всі фермерські господарства України. Отримані облікові дані у цьому випадку складають, як уже відмічалось, генеральну сукупність. Але звичайно для обліку і вивчення у зв'язку з дуже великим обсягом генеральної сукупності із неї береться тільки деяка частина - вибірка. Вона складає вибіркову сукупність.

У результаті вимірювань і обліків в генеральній сукупності чи у вибірці рослин, тварин, механізмів, грунтів дослідник отримує ряд числових значень, які тією чи іншою мірою відрізняються одне від одного. Ці значення складають статистичний ряд для даної ознаки чи, як часто говорять, даної змінної **(Variable)**, а кожне окреме значення є повторенням, або випадком **(Case)**.

Судження про вибірку по всьому статистичному ряду зробити звичайно важко. Необхідна його «згортка», виділення в ньому суттєвих властивостей і особливостей. Комплекс процедур, розроблених для цього в математичній статистиці, отримав назву описових статистик. Отримувані при цьому оцінки дістали назву **статистичних параметрів**. Описова статистика ділиться на два підрозділи: а) **точкове оцінювання**, при якому статистичний ряд характеризується кількома одиночними узагальнюючими оцінками, і б) **інтервальне оцінювання**, коли статистичний ряд оцінюється деяким інтервалом «від» і «до», в межах якого знаходяться його основні типічні та статистично вірогідні значення.

Точкове оцінювання

Для точкового оцінювання генеральної сукупності використовуються наступні оцінки, які позначають літерами грецького алфавіту:

μ - середнє арифметичне. Воно знаходиться в середин ряду і в ранжованому ряді немовби ділить його навпіл.

σ - стандартне відхилення. Воно дозволяє з більшою чи меншою імовірністю виділити ту зону ранжованого статистичного ряду, в якій знаходиться та чи інша кількість даних. Так, в зоні $\pm 1\sigma$ міститься 68,26%, в зоні $\pm 2\sigma$ - 95,45%, а в зоні $\pm 3\sigma$ - 99,73% усіх даних.

σ^2 - дисперсія. Вона являє собою квадрат стандартного відхилення і зручна тим, що, якщо для характеристики розкиду даних складати стандартні відхилення, завжди виходить 0, яким би не був цей розкид, а дисперсія завжди позитивна (так, наприклад, $2^2=4$ і $(-2)^2=4$) і її абсолютна величина тим більша, чим більше розкид даних у статистичному ряду.

У тому випадку, коли дослідник має справу з вибіркою, то для зручності ці оцінки замінюються латинськими літерами:

$\bar{x}$ - вибіркове середнє арифметичне;

s - вибіркове стандартне відхилення;

s^2 - вибіркова дисперсія;

$s_{\bar{x}}$ - похибка вибіркового середнього арифметичного.

Одна з основних точкових оцінок вибірки - це середнє арифметичне. Воно знаходиться за формулою:

$$\bar{x} = \frac{\sum x_i}{N},$$

де $\bar{x}$ *- середнє арифметичне,*

$\sum x_i$ *- сума всіх членів статистичного ряду;*

N *- число членів у статистичному ряду.*

До середнього арифметичного в математичній статистиці пред'являють багато вимог, і воно вирізняється великою інформативністю. Ці вимоги такі:

1. Переконливість, яка означає що $\bar{x} \approx \mu$, тобто вибіркове середнє приблизно дорівнює середньому арифметичному генеральної сукупності. Завдяки цьому при додаванні нових даних у вибірку її середнє арифметичне майже не змінюється. Напроти, таке додавання веде до все більшого наближення вибіркової середньої до генеральної середньої.

2. Незміщенність, яка означає, що вибіркове середнє лежить на самій вершині кривої нормального розподілу, розділяючи її на дві рівні частини.

3. Ефективність, яка припускає, що для даної вибірки середнє арифметичне займає таке положення, при якому дисперсія мінімальна, тобто розкид точок (значень) навколо середньої є найменшим.

4. Робастність, яка означає, що хоча вибірка у деякій невеликій мірі відхиляється від нормального статистичного розподілу, але всі оцінки для неї можуть здійснюватись на основі цього розподілу і будуть справедливими.

При всіх позитивних якостях вибіркового середнього в нього є один серйозний недолік: середнього арифметичного може просто не існувати у природі. Так, при підрахунку кількості зерен у колоску для вибірки може бути отримане середнє арифметичне, яке дорівнює 13,4, тоді як кількість зерен у колоску завжди виражається цілим числом і не може бути дробовою. Цю особливість середнього арифметичного слід мати на увазі під час інтерпретації даних.

Стандартне відхилення як міра розсіювання вичислається за формулою:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}.$$

Обидва підкореневі вирази у формулі тотожні. Для обчислень за допомогою калькуляторів використовується друга половина формули як більш зручна. Перша (ліва) половина формули наочно показує, що стандартне відхилення обчислюється на основі обліку відхилення кожного спостереження від загальної середньої.

Дисперсія являє собою просто квадрат стандартного відхилення і є підкореневим виразом у наведених формулах. Уже відзначалось, що вона зручна як міра розкиду тому, що гасить знак мінус у тієї половини значень стандартного відхилення, які розміщуються в лівій частині кривої нормального статистичного розподілу.

Стандартна похибка середнього арифметичного знаходиться за формулою:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

і записується у підсумкових таблицях і в тексті у вигляд: $x \pm s_{\bar{x}}$. При цьому у похибці прийнято записувати на один десятичний знак більше, ніж у середньому арифметичному. Наприклад 5,0±0,12.

Коефіцієнт асиметрії оцінює скошеність статистичного ряду і зміщення вершини розподілу вправо чи вліво по відношенню до типової кривої нормального розподілу. Вичисляється за складною формулою:

$$A = \frac{n\sum x_i^3 - 3\sum x_i \sum x_i^2 + 2(\sum x_i)^3 / n}{s^3 (n-1)(n-2)}.$$

Ексцес оцінює провал чи пік у центрі розподілу у порівнянні з положенням найбільш високої частини кривої нормального статистичного розподілу. Для його знаходження використовується формула:

$$E = \frac{((n+1)(n\sum x_i^4) - 4\sum x_i \sum x_i^3 + (6(\sum x_i)^2 \sum x_i^2 / n) - 3(\sum x_i)^4 / n^2)}{(s^4(n-1)(n-2)(n-3)) - 3(n-1)^2 / ((n-2)(n-3)))}.$$

Наведені дві формули досить громіздкі і при ручних обчисленнях на них витрачається багато часу. У пакеті Statistica обчислення цих параметрів займають всього декілька секунд.

Коефіцієнт варіації у вітчизняних агрономічних роботах нерідко використовується як додаткова характеристика розкиду членів статистичного ряду. По суті цей показник є стандартним відхиленням, нормованим по відношенню до середнього арифметичного, і виражений у відсотках:

$$v = \frac{s}{\overline{x}} \cdot 100\%.$$

У сучасних статистичних комп'ютерних пакетах коефіцієнт варіації, як правило, не наводиться в числі обчислюваних статистичних параметрів. Це пов'язано з тим, що в його основі лежить припущення, що дисперсія зростає при збільшенні середнього арифметичного. Теоретично це припущення не доведено. Розглянемо приклад. Припустимо, що для деякої сільськогосподарської культури коефіцієнт варіації складає 30% при середньому врожаї 3 ц/га. Тоді 30=100s/3 і стандартне відхилення s дорівнює 0,9. Якщо ж урожайність культури складає 300 ц/га, то при цьому рівні варіювання стандартне відхилення повинно складати 90. Практично високоврожайні культури ніколи не відрізняються великим розкидом, великим варіюванням урожаїв, ніж низьковрожайні. Теж саме можна сказати й про високопродуктивних тварин. Теоретична основа коефіцієнта варіації очевидно не вірна. Коефіцієнтом варіації для порівняння розкиду вибірок розумно користуватися тільки тоді, коли знаменники цієї формули приблизно однакові, тобто при близьких середніх арифметичних для порівнюваних статистичних рядів. Певним недоліком коефіцієнта варіації є також відсутність у нього верхньої межі: він змінює своє значення від 0 до нескінченності.

Показник точності досліду. Він був запропонований ще у 20-х роках А.А.Сапегіним і знаходиться за формулою:

$$p = \frac{100 s_{\bar{x}}}{\bar{x}} = \frac{100 s}{\bar{x}\sqrt{n}}.$$

Вважаються прийнятними для достовірності висновків по польовому досліду значення точності досліду менше 5%. Як видно з формули, показник точності досліду сконструйований емпірично і подібно до коефіцієнта варіації виходить з передумови, що чим більший урожай, тим будуть більші й стандартне відхилення, й дисперсія.

В подальших дослідженнях це припущення не підтвердилось, тому у більшості статистичних пакетів точність досліду не вираховується. Якщо ж все-таки необхідно з деяких причин вирахувати точність досліду, то вона знаходиться вручну по значенням середнього арифметичного і похибки (чи стандартного відхилення), які видає пакет Statistica.

Інколи для оцінки середини статистичного ряду використовують ще дві другі оцінки:

Медіана (Ме) - ділить ранжований статистичний ряд точно в середині і буває особливо зручною при оцінці рядів якісних даних.

Мода (Мо) - це є випадок, який найбільш часто зустрічається. Даний показник нерідко використовується у селекційних дослідженнях.

Інтервальне оцінювання

Ідея інтервального оцінювання грунтується на вже розглянутій вище обставині, що середнє арифметичне з певною вірогідністю повинно знаходитись в інтервалі, який в узагальненій формі виглядає так:

- в інтервалі $\pm 3 \cdot s_{\bar{x}}$ середнє арифметичне знаходиться з імовірністю 99%;
- в інтервалі $\pm 2 \cdot s_{\bar{x}}$ середнє арифметичне знаходиться з імовірністю 95%;
- в інтервалі $\pm 1 \cdot s_{\bar{x}}$ середнє арифметичне знаходиться з імовірністю 68%.

Фактично середнє знаходиться в інтервалі $\pm t \cdot s_{\bar{x}}$, де t - значення критерію Ст'юдента, яке залежить від обсягу вибірки і табульоване. При великих обсягах вибірки для 95% імовірності воно дорівнює 1,96.

Зона довіри для імовірності 95% таким чином знаходиться шляхом наступних розрахунків. Нехай середнє арифметичне дорівнює 16,0, а помилка середнього - 0,35. Тоді

$$16{,}0\text{-}1{,}96\cdot 0{,}35 < \bar{x} < 16{,}0+1{,}96\cdot 0{,}35$$

або

$$15{,}31 < \bar{x} < 16{,}69,$$

це означає, що середнє арифметичне з імовірністю 95% і знаходиться між 15,31 і 16,69.

Довірчі оцінки часто бувають більш інформативними, ніж точкові. Вони добре зображаються графічно. В пакеті Statistica для цього є спеціальна опція - «ящик з вусами» **(Box and Whiskers)**. У якості прикладу такий графік представлений на рис. 3.

Пакет «Статистика» будує декілька типів «ящиків з вусами». Їх принципова подібність полягає в тому, що увазі користувача пропонується рисунок, основу якого складає двомірна система координат. На осі «*y*» відкладаються абсолютні значення ознаки, яка аналізується, а на осі «*x*» вказуються номери або назви досліджуваних статистичних рядів. Смислове навантаження «ящика», точки в його центрі і прибудованих вусів може бути різним. При інтервальному оцінюванні в якості ілюстрації використовується тип «ящика з вусами» Mean/SE/1.96·SE (як на рис. 3). У такому випадку точка в центрі «ящика» буде відповідати значенню середнього арифметичного, величини $\bar{x}+S_{\bar{x}}$ та $\bar{x}-S_{\bar{x}}$ будуть визначати відповідно

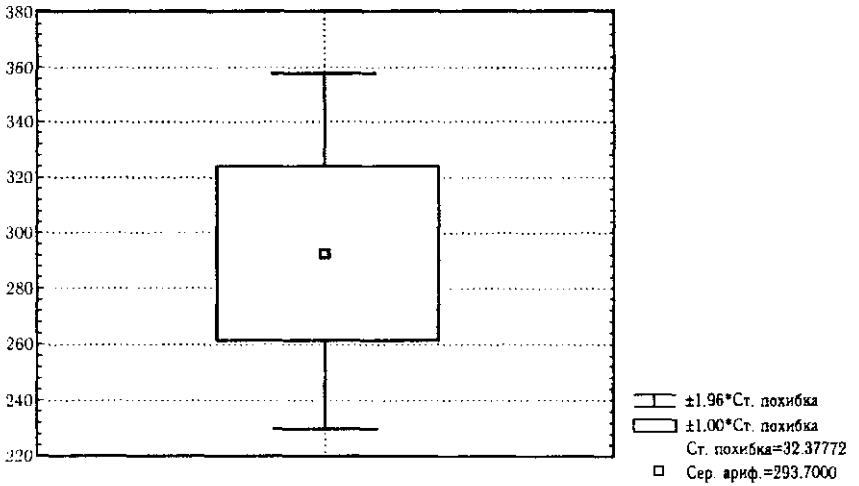


Рис. 3. Подання основних параметрів вибірки у вигляді "ящика з вусами"

верхню і нижню границі «ящика». Значення $+1{,}96S_{\bar{x}}$ та $-1{,}96S_{\bar{x}}$ - відповідно довжину верхнього і нижнього «вуса», тобто межі довірчого інтервалу для 95-ти відсоткової імовірності.

При побудові «ящика з вусами» типу Mean/SE/SD точка в його центрі буде також відповідати величині середнього арифметичного, верхня і нижня границі «ящика» визначаються значеннями $\bar{x} \pm S_{\bar{x}}$, а довжина «вусів» - величиною стандартного відхилення. Пакет «Статистика» будує ще й «ящики з вусами» типу Mean/SD/1,96·SD і Median/Quart/Range. У правому нижньому кутку кожного представленого рисунка завжди дається пояснення, які параметри взяті за основу даного «ящика з вусами».

В цілому користування цією опцією стає особливо інформативним, якщо в один рисунок розміщується декілька «ящиків з вусами», відповідних двом або ж більшій кількості статистичних рядів, які порівнюються.

Для отримання навичок виконання точкового і інтервального оцінювання статистичних рядів послідовно виконайте роботи №5 та № 6.

ПІДГОТОВКА ВИХІДНИХ ДАНИХ ДО СТАТИСТИЧНОЇ ОБРОБКИ

Як правило, кількісні дані піддаються різним видам статистичної обробки безпосередньо у тій формі, в якій вони були отримані дослідником.

Але в деяких випадках вихідні дані перед обробкою піддаються спеціальній обробці - трансформації. Частіше за все це доводиться робити, коли отриманий статистичний ряд суттєво відхиляється від нормального статистичного розподілу, невеликий за обсягом або містить у собі «вискакуючі» значення. Такі вихідні статистичні ряди можуть бути наближені до нормального розподілу штучними прийомами.

Якщо вибірка досить велика, то найбільш доцільно вибраковку «вискакуючих» значень учинити шляхом переводу даних на нормальний імовірнісний аркуш і видаленням «вискакуючих» значень, які знаходяться далеко від прямої лінії, інструментом «Пензлик» (**Brush**). Ця процедура грунтується на критерії τ (тау) і відповідає формулам:

$$\tau_{\max} = \frac{x_i - \overline{x}}{\sigma},$$

$$\tau_{\min} = \frac{\overline{x} - x_i}{\sigma}.$$

Якщо $\tau_{факт.} > \tau_{табл.}$, то цей член ряду викидається. Інколи за пропозицією Діксона з ряду викидаються саме маленьке і саме велике значення.

У тих випадках, коли статистичний ряд не дуже великий і недоцільно повністю губити окремі його значення, використовують перетворення статистичних рядів. Перетворення (трансформації) ведуть до згладжування ряду, зменшують дисперсію без втрат вихідних даних і, не перекручуючи статистичних параметрів вибірки, роблять точкові й інтервальні оцінки більш об'єктивними і надійними.

В математичній статистиці в залежності від характеру вихідних даних розроблені різні методи їх перетворення.

Звичайно, коли дані являють собою результати підрахунків кількості культурних рослин і бур'янів на облікових ділянках, добрі результати дає перетворення добуванням квадратного кореня за формулою:

$$x' = \sqrt{x_i}\ .$$

При обробці даних по величині врожаїв використовується логарифмічне перетворення за однією з таких формул:

$$x' = \log_{10}(x_i + 1),$$
$$x' = \log_{10}(x_i),$$
$$x' = \ln_e(x_i).$$

При обробці даних за схожістю насіння, які виражаються у відсотках, використовують арксинус-перетворення:

$$x' = \arcsin\sqrt{\frac{x_i}{100}}\ ,$$

або

$$x' = \arcsin\sqrt{x_i}\ .$$

У всіх цих випадках трансформації послідовно перетворюється кожне значення та із одержаних похідних складається новий статистичний ряд. Пакет Statistica повністю автоматизує цю процедуру, і вона зводиться лише до вибору відповідного виду перетворення.

Друге незалежне завдання при підготовці статистичних рядів до обробки полягає в одержанні із двох чи декількох статистичних рядів похідного ряду з дорученням цієї роботи комп'ютеру. Так, наприклад, нерідко виникає проблема, подібна наступній: в одному статистичному ряду наведені дані про кількість листків на рослинах, а в другому - середня площа окремого листка на кожній із цих рослин, для аналізу ж необхідно мати площу листків на кожній рослині. Процедура обробки дуже проста: необхідно перемножити пару значень. При ручному обчисленні ця процедура займає багато часу. Комп'ютер виконує такі й більш складні завдання швидко і точно. Якщо перший ряд введений в VAR1, другий - в VAR2, то пакет Statistica дозволяє легко одержати новий стовпчик даних - VAR3, в якому кожне значення дорівнює VAR1 помноженому на VAR2.

Щоб набути навичок виконання трансформації статистичних рядів, виконайте роботу № 7. У роботі № 8 показані способи одержання з двох чи більшого числа вихідних статистичних рядів одного похідного ряду. Робота № 9 присвячена виявленню «вискакуючих» значень і їх видаленню із статистичного ряду.

СТАТИСТИЧНА ДОСТОВІРНІСТЬ КІЛЬКІСНИХ ДАНИХ. ДОВІРЧІ РІВНІ

Однією з важливих переваг математичної статистики є можливість не тільки узагальнювати результати дослідів і спостережень або представляти їх у згорненому вигляді, але також оцінювати їх надійність, ступінь достовірності. Це виключно важливо для прийняття відповідальних виробничо-ділових рішень, пов'язаних з фінансовими затратами та фінансовим ризиком. Оскільки всі біологічні природні явища мають імовірнісний характер, для таких оцінок використовують підходи, розроблені теорією імовірності.

В математичній статистиці імовірність оцінюють у частках одиниці або (рідше) у відсотках. Повна імовірність події складає при цьому 1,0 або 100%, абсолютно неймовірна подія відповідно 0,0 або 0%. Частки легко переводяться у відсотки: 0,1 - 10%; 0,05 - 5%; 0,005 - 0,5% і т.п.

Вибіркові дані за умови правильного збору матеріалу звичайно відповідають нормальному статистичному розподілу, що дозволяє зв'язати їх з кривою цього розподілу, для якої відомі імовірні появи тих чи інших значень. Це дозволяє виносити обгрунтовані судження про розкид даних і оцінити імовірність очікуваних значень статистичних параметрів. Повна частота зустрічей подій визначається площею під кривою (рис. 4 і рис. 5), а імовірність зустрічі конкретних подій пов'язана з величиною середнього арифметичного і його стандартною похибкою.

Так, якщо $\bar{x}$ дорівнює 12,0, а $S_{\bar{x}}$ дорівнює 2, то:

з 99% імовірністю середнє лежить в інтервалі $12 \pm 3 \cdot 2$, тобто 6...18,

з 95% імовірністю середнє знаходиться в інтервалі $12 \pm 2 \cdot 2$, тобто 8 ... 16,

з 68% імовірністю середнє знаходиться в інтервалі $12 \pm 1 \cdot 2$, тобто 10...14.

Це цілком типовий приклад. Його аналіз показує, що ширина зони довіри і імовірність зв'язані між собою так, що при підвищенні імовірності зона довіри, в якій фактично може лежати середнє арифметичне, виявляється більш широкою. У відповідності з цим, коли

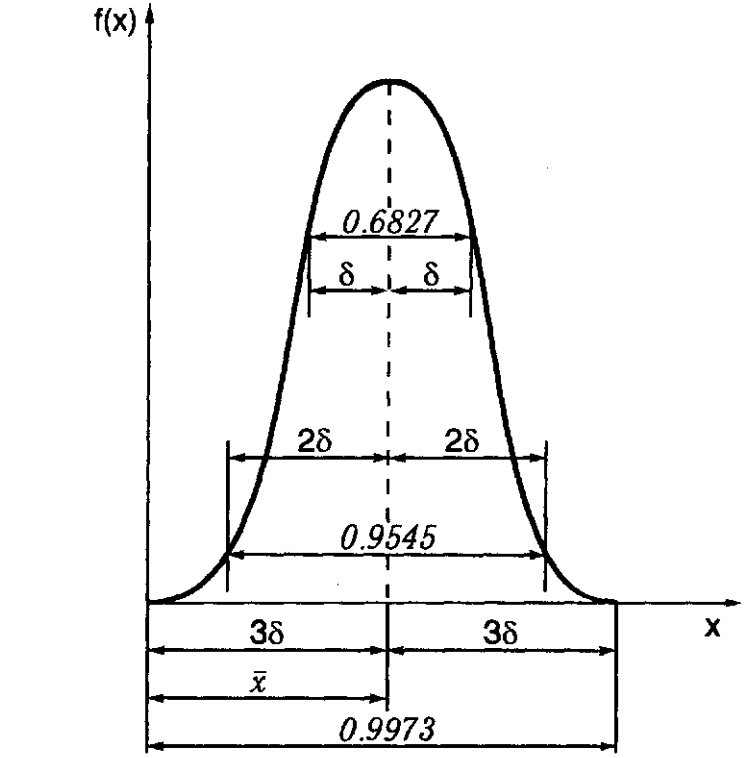


Рис. 4. Щільність нормального статистичного розподілу

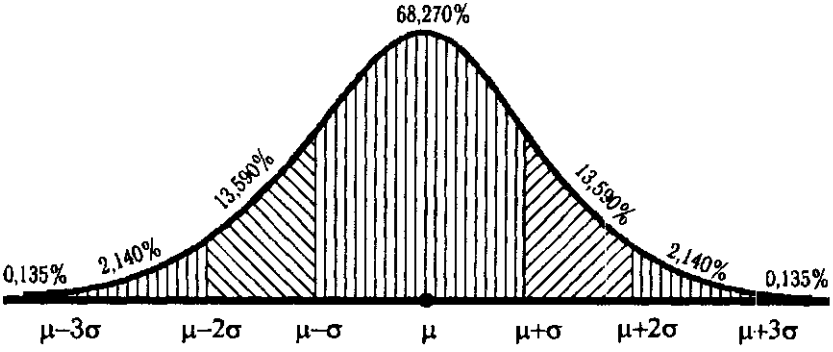


Рис. 5. Щільність нормального статистичного розподілу з середнім арифметичним, стандартним відхиленням і зонами довіри

зона довіри береться вузькою, то імовірність попадання в неї статистичних параметрів, і в першу чергу середньої, робиться низькою. *Ціною збільшення статистичної надійності є втрата точності.*

Для з'ясування цієї важливої залежності розглянемо такий приклад. На іподромі ви укладаєте грошовий заклад на перемогу коней. Біжить всього 10 коней, з яких 5 білих і 5 гнідих, і кожний кінь має свій номер і свою кличку. Умови закладу, які вам можуть запропонувати, можуть бути такими:

1. Б'юсь об заклад, що один з цих коней добіжить до фінішу!

Навряд чи ви погодитесь на цей заклад і будете стверджувати, що жоден з коней не добіжить. Зона довіри в цьому випадку широка - всі коні, тобто охоплює весь простір подій, - і це значить, що той, хто пропонує вам побитись об заклад, може чекати із 100% імовірністю, що його прогноз збудеться. Хоча б одна з цих «кляч» до фінішу добереться.

2. Б'юсь об заклад, що першим прибіжить гнідий кінь.

В цьому випадку (простір подій звузився у два рази!) ви легко підрахуєте, що, оскільки гнідих коней 5, імовірність перемоги одного з них складає 5/10=0,5, або 50%. На цих умовах, коли шанси учасників закладу абсолютно рівні, ви, будучи дуже азартною людиною, можливо й захочете взяти участь закладі. Але ваші шанси й шанси вашого супротивника перемогти у цій суперечці абсолютно однакові. Все вирішить випадковість.

3. Б'юсь об заклад, що першим прибіжить кінь під № 4 по кличці «Жвавий».

На таких умовах суперечки той, хто її пропонував, звузив простір імовірної події до 1/10=0,1, або 10%. Якщо імовірність його прогнозу тільки 10%, то шанси на користь його супротивника будуть складати 90%. Правда, якщо ваш супротивник заздалегідь узнав у конюхів, що кінь № 4 - самий швидкий на всю конюшню, то ви можете й втратити свої гроші. Але це вже не проблема імовірностей, а проблема інформації.

В практичній сільськогосподарській роботі при оцінці даних дослідів і обліків необхідно зупинятись на якомусь визначеному рівні імовірності подій. У звичайних умовах в агрономії й зооінженерії вважається допустимим для прийняття висновку 95% «за» і 5% «проти». Це означає, що технологія чи явище, яке ви вивчаєте й досліджуєте, при перевірці, наприклад, на протязі 100 років у 95 з них підтвердиться і тільки в 5 роках можливо, що вони не дадуть очікуваного результату.

Але при прийнятті того чи іншого довірчого рівня все залежить від характеру дослідів і процесів, що вивчаються. Якщо ви вивчаєте

новий сорт, то можливо й погодитесь з висновком про його переваги й при 80% імовірності, що буде означати: за 10 років вирощування у восьми випадках цей сорт дасть прибавку і тільки у 2 роки із цих десяти, можливо, не виявить своїх переваг. З точки зору економіки й величини прибавки врожаю від нового сорту такі умови можуть бути прийнятними.

Але якщо перевіряється новий пестицид і враховується його безпечність для здоров'я працюючих, то звичайний 95% рівень буде означати, що із 95 працюючих 5 отримають тяжке отруєння. Такий рівень, звичайно ж, буде недопустимим і прийдеться брати 99,9% рівень, а то й більш високий. **Таким чином, припустимий довірчий рівень у кожному випадку визначається сутністю проблеми і вашим рішенням, а не комп'ютером.**

Цей же імовірнісний підхід застосовується у випадку будь-яких порівняльних оцінок. Практично в усіх видах польових досліджень та інших сільськогосподарських дослідженнях фахівець перевіряє деякі нові сорти, технології чи прийоми стосовно до уже існуючих. В математичній статистиці і методиці дослідницької справи в усіх цих випадках прийнято **одне універсальне правило: вихідне припущення (його називають нульовою гіпотезою і позначають H_0) полягає в припущенні про відсутність відмінності між новими сортами або технологіями і вже існуючими.** Це записується так:

H_0:x=y,

де x та y відповідно нова і стара технології.

Така нульова гіпотеза загальноприйнята внаслідок її однозначності. Дійсно в цьому випадку x-y=0. Якщо взяти будь-яку іншу нульову гіпотезу

H_0:x≠y,

то з'являється багато додаткових варіантів x>y, x<y і наскільки менше чи більше. Проблема статистичної перевірки такої нульової гіпотези стає розмитою і вимагала б маси додаткових розрахунків на основі прийняття цілої серії послідовних нових нульових гіпотез.

У математичній статистиці і в сучасних статистичних пакетах при оцінці достовірності матеріалу завжди прийнято вираховувати **шанси на користь нульової гіпотези**. Ці шанси вираховуються в частках одиниці і позначаються через *p*. В цій формі вони і видаються комп'ютером.

Так, наприклад, отримавши довірчий рівень у p=0,05, ми будемо знати, що на користь нульовій гіпотезі про відсутність відмінностей

або переваг нового сорту або нової технології лише 5% шансів, а 95% проти. Таку нульову гіпотезу при загальноприйнятому підході можна відкидати і стверджувати, що з імовірністю 95% новий метод, сорт, технологія суттєво відрізняються від попередніх. Такий висновок слід зробити також і при будь-якому p меншому, ніж 0,05.

У випадках, коли довірчий рівень більше 0,05, нульова гіпотеза не відкидається і, отже, новий сорт чи технологія не мають переваг і відмінностей від старих.

Звичайно для формулювання висновків за такої статистичної обробки матеріалу пишуть:

«з імовірністю ...% можна стверджувати, що середнє арифметичне знаходиться в інтервалі від ... до ...» або

«з імовірністю ...% можна стверджувати, що середнє арифметичне не менше ... і не більше...».

У всіх таких випадках переконливість вашого висновку вища, коли зона довіри вужча. Як її звузити? З наведених вище формул видно, що для цього є тільки один спосіб: зменшити похибку середнього арифметичного $S_{\bar{x}}$. Але сама по собі похибка залежить від двох факторів і діяти треба на один з них або на обидва:

а) від природної мінливості, властивої даному явищу. Якщо це строкатість властивостей грунту на дослідній ділянці, то досліди треба перенести на нове поле, якщо це нестабільність властивостей сорту, то від нього скоріше за все прийдеться відмовитись і т.п.;

б) від випадкових помилок, які виникають під час проведення досліду - нерівномірність посіву, недбалість обліку та інше, то в цьому випадку треба підвищувати акуратність виконання всіх польових робіт і обліків.

Імовірнісний підхід показує не тільки шанси на користь правильного рішення, але й він вказує на можливість зробити помилку при прийнятті рішення. Ці помилки підрозділяються на дві категорії:

а) **Помилка першого роду**. Нульова гіпотеза звичайна H_0:x=y. Ця гіпотеза в даному досліді правильна, але із-за малої величини вибірки або її неякісності гіпотеза відкидається і робиться помилковий висновок x≠y. Цю ситуацію й називають помилкою першого роду. Імовірність зробити помилку першого роду позначають літерою α і називають **рівнем значущості критерію**.

б) **Помилка другого роду**. Нульова гіпотеза звичайна H_0:x=y. Але в цьому випадку нульова гіпотеза неправильна і фактично x≠y. Ви ж її приймаєте і тим самим «губите» реально існуючу відмінність варіантів досліду. Така помилка називається помилкою другого роду. Імовірність помилки другого роду позначається β і величина $1-\beta$ називається **потужністю критерію**.

Ще одна тонкість зв'язана з так званими одно- і двосторонніми критеріями. Коли дослідника цікавить відмінність сама по собі і неважливо більше чи менше значення варіанта у відношенні значення контролю, лиш би вони відрізнялись один від одного, то використовують двосторонній критерій **(Two-tail)** (рис. 6). Але якщо вас цікавлять конкретно ті варіанти, які більші, ніж контроль (чи менші - друга альтернатива), то використовується односторонній критерій **(One-Tail)** (рис. 7), взятий з відповідного боку кривої нормативного статистичного розподілу.

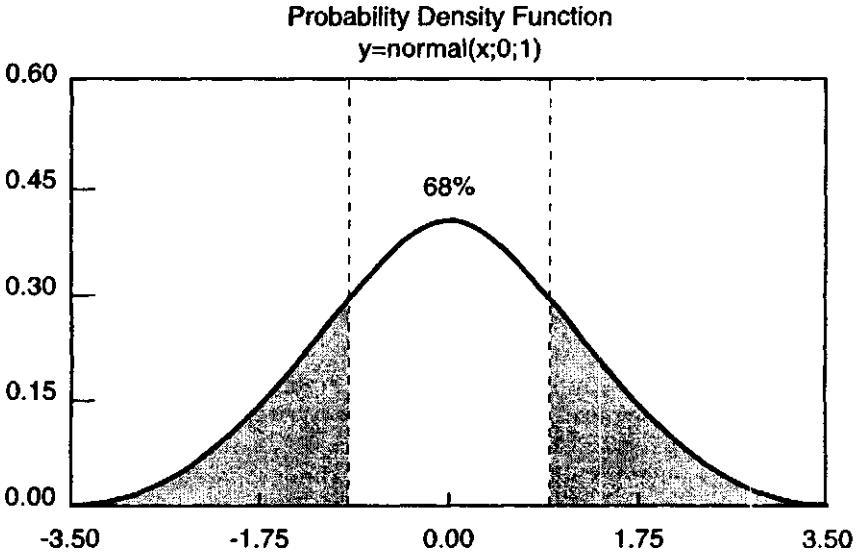


Рис. 6. Двосторонній статистичний критерій. Довірчий інтервал знаходиться між двома заштрихованими зонами

У відповідних випадках статистичний пакет Statistica задасть вам питання: який з цих критеріїв слід застосовувати для обчислення довірчого рівня? Двосторонній критерій більш суворий і в сумнівних випадках краще використати його.

В цілому використання концепції довірчих рівнів в її сучасному вигляді дозволяє відійти від точкових оцінок вибірок, вірогідність яких не відома, до інтервального оцінювання і отриманням імовірнісних оцінок прийнятих рішень. Інтервальне оцінювання - це новий спосіб мислення. Він веде до висновків нового типу, котрі можуть формулюватися таким чином: «новий сорт дає середній

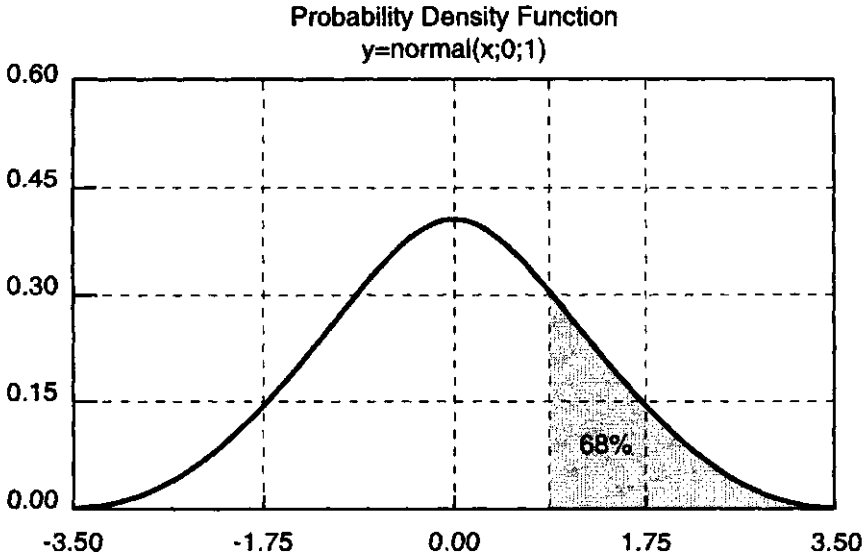


Рис. 7. Односторонній статистичний критерій. Довірчий інтервал знаходиться ліворуч від заштрихованої зони

урожай 40 ц/га, і з імовірністю 95% я стверджую, що за 10 років досліджень його врожайність буде не нижча, наприклад, 35 ц/га, але й не вища 45 ц/га.

Для одержання навичок роботи із знаходженням довірчих рівнів виконайте практичну роботу № 10.

МІНЛИВІСТЬ

Мінливість характерна для всіх біологічних об'єктів, технологічних операцій і економічних процесів. Розрізняють мінливість кількісну і якісну. До кількісної мінливості належать об'єкти, які мають масу, розмір, об'єм або їх можна лічити поштучно. Якісна мінливість - це варіювання кольору пелюсток, наявність у рослин тичинкових або маточкових квіток і т. д.

Крім того, розрізняють неперервну та перервну кількісну мінливість. До неперервної мінливості належать ті об'єкти, які виражають в основному дробовими числами, - це маса, розмір, об'єм досліджуваних об'єктів. До перервної мінливості належать об'єкти, які обліковуються поштучно: кількість листків, суцвіть, рослин, тварин тощо.

Власну амплітуду мінливості мають усі основні ознаки культурних рослин та бур'янів: кількість зерен у колосі, кількість листків на рослині, висота рослин і т.п. Не постійні й основні ознаки домашніх тварин. Не є абсолютно стабільним і режим роботи сільськогосподарських машин і механізмів. У цих випадках значення ознак змінюються від особини до особини, від агрегату до агрегату і створюють мінливість.

Спостерігається й інша категорія мінливості – це нестійкість ознак у часі. Якщо вирощувати один і той же сорт декілька років підряд, то його врожаї будуть різними.

Особливу категорію мінливості показників і параметрів складають помилки виміру, які виникають із-за неточності приладів або помилок операторів.

За будь-якої форми мінливості вона потребує оцінки. Для оцінки мінливості використовують декілька різних показників. Головні із них три:

Розмах – це різниця між самою великою і самою малою величинами в статистичному ряду:

$$R=x_{max}-x_{min}.$$

Дисперсія – точкова параметрична оцінка статистичного ряду, яка є мірою розкиду окремих значень навколо середнього арифметичного. Вона найчастіше використовується для оцінки мінливості показників, зв'язаних з помилками окремих обліків.

Коефіцієнт варіації, який при його очевидних недоліках все-таки нерідко використовують для оцінки мінливості. В агрономії і тваринництві при цьому звичайно приймають таке групування величин коефіцієнта варіації:

від 0 до 10% - незначне варіювання,
від 10 до 20% - невелике варіювання,
від 20 до 40% - середній рівень варіювання,
від 40 до 60% - велике варіювання,
від 60 і більше – дуже велике варіювання.

Коефіцієнт Джині є мірою мінливості і приймає значення від 0 до 1. Він знаходиться на основі так званої кривої Лоренца. В пакеті Statistica він відсутній і може вираховуватися за спеціальною програмою.

Для одержання навичок знаходження мінливості виконайте практичні роботи № 11 і 12.

ПОПАРНЕ ПОРІВНЯННЯ СТАТИСТИЧНИХ РЯДІВ. КРИТЕРІЙ СТ'ЮДЕНТА

В сільськогосподарських дослідженнях часто зустрічається необхідність порівняння двох вибірок для установлення статистичної значимості їх рівності або нерівності. Наприклад, порівнюють два сорти по врожайності для установлення достовірності більш високої врожайності одного із них, порівнюють два варіанти польового досліду між собою або з контролем для установлення достовірності переваг тієї чи іншої технології (добрив, гербіцидів і т.п.), порівнюють стан рослин у два терміни спостережень для установлення достовірності приростів біомаси, порівнюють продуктивність двох порід тварин і т.п.

Для розв'язання всіх таких задач розроблений спеціальний метод, який грунтується на критерії Ст'юдента (Ст'юдент - це псевдонім англійського вченого Госсета). По-іншому цей метод інколи називають t-тест, тому що величину критерію Ст'юдента прийнято позначати літерою t. Критерій Ст'юдента широко використовується для попарного порівняння статистичних рядів і їх середніх арифметичних. Саме для таких цілей і був розроблений цей метод. Але його застосування вимагає обережності.

Це пов'язано з такими обставинами:

а) критерій Ст'юдента чутливий до нормальності статистичних рядів і тому перед його використанням ряди слід перевірити на нормальність, використовуючи нормальний імовірнісний аркуш, а при необхідності зробити їх перетворення;

б) слід враховувати, що критерій Ст'юдента придатний тільки для парних порівнянь, одночасно порівнювати три і більше середніх на основі цього критерію не можна;

в) розрахунок критерію Ст'юдента робиться за різними формулами в залежності від рівності або нерівності дисперсій порівнюваних рядів, що слід ураховувати.

Зате критерій Ст'юдента дозволяє швидко проводити порівняння двох середніх арифметичних, двох коефіцієнтів кореляції чи двох значень, виражених у відсотках.

При вживанні критерію Ст'юдента користузачу необхідно знати середнє арифметичне обох виборок ($\bar{x}_1$ та $\bar{x}_2$), величини дисперсій

для цих вибірок (s_1^2 та s_2^2) і величину кожної із вибірок (n_1 та n_2). Під час комп'ютерних обчислень ці величини знаходяться автоматично на наявній базі даних.

Використання критерію Ст'юдента не є зовсім формальним завданням, тому що для різних ситуацій розрахункові формули неоднакові. Користувачу необхідно прийняти рішення з таких питань:

1. Перш за все необхідно знати залежні *(dependent)* чи незалежні *(independent)* порівнювані вибірки. Це завжди змістовне рішення користувача, бо він зобов'язаний дати правильну відповідь на запитання комп'ютера. Так якщо ви порівнюєте два сорти за врожайністю, то вибірки, очевидно, незалежні. Але якщо ви ураховували висоту рослин одного сорту у два різних терміна, то вибірки залежні, тому що ці рослини з одного поля, їх величина у другий термін залежить від того, якою вона була у перший термін. Для кожної такої ситуації формального рішення немає, воно - результат сільськогосподарського, а не математичного аналізу.

2. Треба оцінити рівні чи не рівні величини порівнюваних вибірок, тобто кількість спостережень у кожному варіанті. Комп'ютер це виконає самостійно.

3. Треба оцінити рівні між собою чи не рівні дисперсії порівнюваних вибірок. Це перевіряється за критерієм Фішера:

$$F= s_1^2 / s_2^2 \text{ при } s_1^2 > s_2^2$$

і ступенях свободи (df)

$$df_1=n_1-1 \text{ і } df_2=n_2-1.$$

По таблиці критерію Фішера визначається статистична достовірність відмінності дисперсії (при цьому більша дисперсія відповідає стовпчику у таблиці, а менша - в рядку). Дисперсії рівні, якщо $F_{\text{фактич.}}$ рівне чи менше, ніж $F_{\text{таблич.}}$ при даних ступенях свободи.

При комп'ютерних обчисленнях цю роботу також повністю виконує програма і користуватися ручними обчисленнями і таблицею немає потреби.

4. Вибрати, виходячи із результатів пунктів 1-3, відповідну формулу для обчислення критерію Ст'юдента. Для вибору формули корисно користуватися схемою-ключем і списком формул. Вони наводяться в будь-яких підручниках з математичної статистики. При комп'ютерних обчисленнях вибір формули - це також формальна процедура, яку виконує сама програма.

При користуванні пакетом «Статистика» для реалізації методу Ст'юдента, таким чином, потрібно:

а) створити базу даних, у типових випадках кожній вибірці відповідає свій стовпчик. Але можна скласти базу даних й інакше: у першому стовпчику код варіанту (слово чи цифра), у другому стовпчику - всі дані, але так, щоб кожне із них стояло проти свого коду;

б) вибрати підменю критерію Ст'юдента (він називається t-test);

в) зробити оцінку залежні чи незалежні вибірки, які ви порівнюєте;

г) перейти до обчислень, які комп'ютер виконає сам.

У результаті роботи програми користувача не буде в загальному випадку цікавити величина t сама по собі (хоча цю величину звичайно ж виписують). Уся змістовна інформація знаходиться в другій величині - це величина p довірчого рівня. Якщо вона більше 0,05, то нульова гіпотеза (а нульова гіпотеза завжди однакова - вибірки зроблені із однієї генеральної сукупності і їх середні арифметичні однакові) не відкидається. А якщо p=0,05 або менше, то нульова гіпотеза відкидається, і це означає, що середні арифметичні двох порівнюваних вибірок не рівні і між ними існує статистично достовірна відмінність.

Для наочності ви можете побачити порівнювані вибірки через графік «ящик з вусами» і роздрукувати його.

Слід підкреслити, що за методом Ст'юдента завжди порівнюються тільки дві вибірки. На практиці частіше доводиться порівнювати декілька вибірок одну з одним. Це завдання розв'язується просто - вибірки порівнюють попарно. В межах пакета «Статистика» зробити це дуже просто.

Найменша істотна різниця

Видно, що користування критерієм Ст'юдента доволі складне при обчисленнях вручну, хоча тривіально просте при комп'ютерних. У радянській літературі в зв'язку з комп'ютерним відставанням було прийнято порівнювати вибірки за так званою найменшою істотною різницею (НІР).

НІР - це деяке критичне число, виражене в абсолютних числах (тобто у тих же одиницях, що й дані в вибірці - центнери для врожаю, см - для висоти рослин, літри для надоїв і т.п.). Якщо абсолютна різниця середніх арифметичних вибірок менше НІР, то вони не відрізняються, а якщо більше, то вони статистично достовірно зідмінні. Треба тільки знайти різницю $\bar{x}_1 - \bar{x}_2$ і знати НІР.

Але виявляється правильно обчислити НІР не так просто. Для дуже великих вибірок воно обчислюється за формулою

$$НІР = t \cdot s_{\bar{x}_1 - \bar{x}_2}, \text{ де } s_{\bar{x}_1 - \bar{x}_2} = (s^2_{\bar{x}_1} + s^2_{\bar{x}_2})^{1/2}.$$

Але з великими вибірками агрономи і зооінженери працюють рідко і тоді потрібно використовувати ті ж формули, що й для обчислення критерію Ст'юдента, перевіряючи рівність дисперсій і рівність вибірок. А оскільки це громіздко, то нерідко обчислення НІР «женуть» за самою простою формулою, до того ж ще й обчислюють НІР не для кожної пари порівнюваних вибірок, а для цілої їх групи. Таке огрублення розрахунків веде до одержання неправильних результатів. Сенсу в них немає.

Тому критерієм НІР в зарубіжних країнах не користуються і в статистичних пакетах він за звичай відсутній.

Швидкий тест Вайєра (Weir, 1960)

При відсутності комп'ютера для швидкого порівняння двох статистичних рядів слід користуватися швидким тестом Вайєра. Цей тест обчислюється за формулою:

$$W=|x_1-x_2|/(\{[(n_1-1)s_1^2+(n_2-1)s_2^2]/(n_1+n_2-4)\}(1/n_1+1/n_2))^{1/2}.$$

Якщо W дорівнює або більше 2.0, то вибірки статистично достовірно відрізняються на рівні 95%.

Для реалізації можливостей критерію Ст'юдента виконайте практичну роботу № 16.

КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Кореляція виявляє зв'язок між ознаками, об'єктами або явищами. Якщо порівнюють дві ознаки, зв'язок (кореляцію) називають **парним**, якщо порівнюється багато ознак – **множинним**. Найбільш розповсюджене використання парної кореляції. Для її вимірювання використовують такі показники:

а) **коефіцієнт парної кореляції Пірсона** – r. Його значення знаходяться в амплітуді від −1 до +1. Значення коефіцієнта кореляції, яке дорівнює 0, вказує на відсутність зв'язку. Використовують коефіцієнт кореляції при прямолінійній залежності ознак одне від одного.

б) **коефіцієнт рангової кореляції Спірмана** – ρ. Має ті ж особливості, але використовується при вивченні зв'язку якісного характеру і у випадках, коли статистичні ряди не відповідають нормальному статистичному розподілу;

в) **кореляційне відношення** - η. Використовується у випадках, коли зв'язок ознак не має прямолінійного характеру. Знаходиться в амплітуді від 0 до +1.

У сільськогосподарських дослідженнях найбільш часто вживають коефіцієнт парної кореляції Пірсона. При його використанні слід мати на увазі, що значення коефіцієнта кореляції установлюють тільки міру зв'язку. Сам же по собі зв'язок може бути наслідком причинної залежності однієї ознаки від другої, а може бути й результатом простого співпадання значень ознак. Причинність або випадковість зв'язку визначає сам дослідник.

Так, під час дослідження площі прапорового листка у злаків і ваги зерен у колосі коефіцієнт кореляції виявить міру зв'язку і буде відображати причинну залежність цих ознак, оскільки відомо, що прапоровий лист у злаків дає до 75% органічних речовин, необхідних для наливання зерна. Розглянемо інший приклад - при обстеженні полів на забур'яненість можна встановити, що має̀ться висока кореляція між великою кількістю на облікових ділянках хвоща Equistum arvense і щавлю Rumex acetosella. Але в цьому випадку велике значення коефіцієнта кореляції не буде відображати причинний зв'язок явищ, а виявиться наслідком того, що обидва види віддають перевагу кислим грунтам і їх сумісний ріст є наслідком співпадання екологічних вимог. Величина коефіцієнта кореляції в цих двох прикладах може бути однаковою, але механізм явища цілком відмінний.

У сільському господарстві проведення кореляційного аналізу дозволяє вирішити багато важливих питань, але під час інтерпретації результатів кореляційного аналізу необхідно уважно розглядати природу явища.

Коефіцієнт парної кореляції Пірсона обчислюють за формулою:

$$r = (n\cdot\Sigma xy-\Sigma x\cdot\Sigma y)/\{[n\Sigma x^2-(\Sigma x)^2]\cdot[n\Sigma y^2-(\Sigma y)^2]\}^{1/2},$$

яка при обчисленні на настільних калькуляторах вимагає великого обсягу обрахувань. При використанні ж пакета «Статистика» розрахунок не займає й декількох секунд.

Статистичну значимість для коефіцієнта кореляції знаходять по спеціальних таблицях при числі ступенів свободи

$$df = n-2.$$

При даному довірчому рівні $r_{фактич.}$ повинно бути більше, ніж $r_{табл.}$

Інколи обчислюють помилку коефіцієнта кореляції, але її наведення у виразі $r\pm s_r$ не несе корисної інформації. Розподіли значень коефіцієнтів кореляції завжди різко асиметричні, і помилка «не вирізає» довірчий інтервал з можливих його значень.

При комп'ютерних обчисленнях пакет Statistica безпосередньо приводить рівень достовірності коефіцієнта кореляції у вигляді значень *p*, якими й слід користуватися (при нульовій гіпотезі –

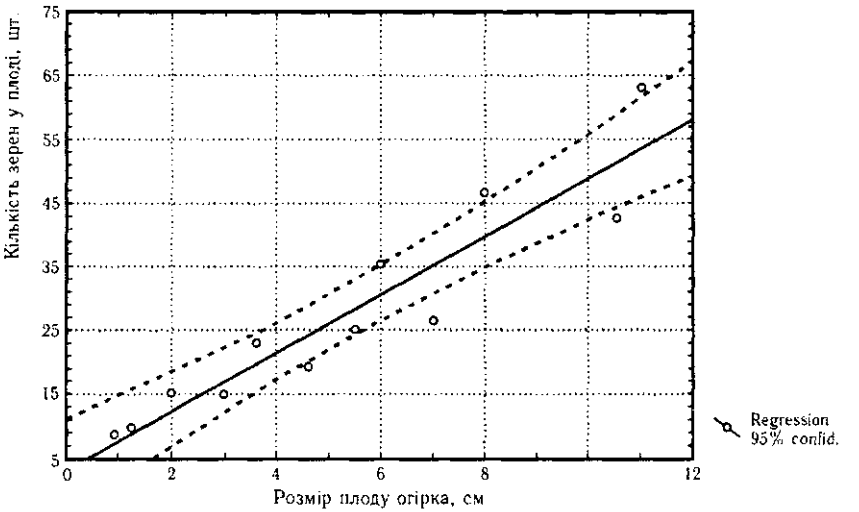


Рис. 8. Еліпс розсіювання точок для випадку прямолінійної залежності

кореляція відсутня). Крім того, для зручності користувача значення коефіцієнтів кореляції статистично значимі на рівні $p<0,05$ виділяються червоним кольором.

Перед обчисленням коефіцієнта кореляції корисно розглянути графік для так званого еліпса розсіювання точок. На рисунку 8 такий графік представлений для позитивного значення коефіцієнта кореляції, коли точки лежать приблизно уздовж прямої лінії. В цьому випадку вирахування коефіцієнта кореляції Пірсона цілком обгрунтоване. Якщо еліпс розсіювання має форму вигнутої ділянки, як на рис. 9, і не співпадає за положенням з прямою лінією, то слід відмовитись від вирахування коефіцієнта парної кореляції Пірсона і використати одну із форм криволінійної залежності.

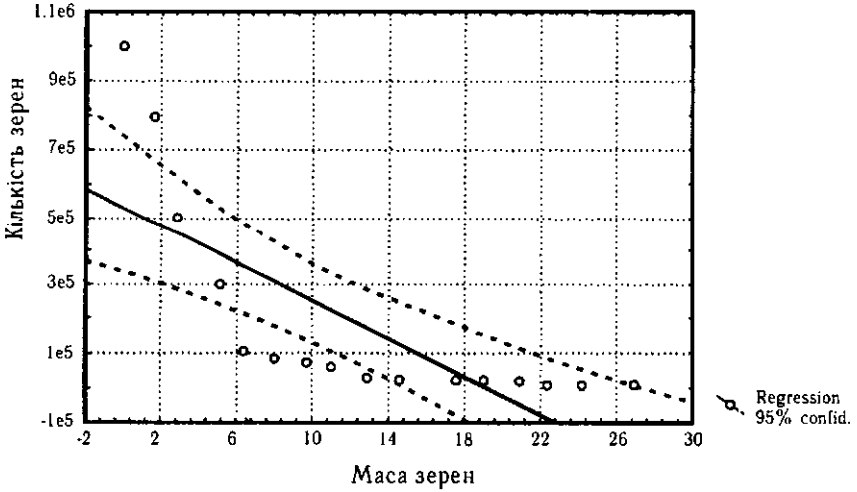


Рис. 9. Еліпс розсіювання точок для випадку криволінійної залежності, коли її апроксимація прямою лінією не є ефективною

При проведенні кореляційного аналізу дослідника звичайно цікавить зв'язок між багатьма ознаками. При проведенні такого множинного кореляційного аналізу результат у пакеті "Статистика" одержують відразу у вигляді кореляційної матриці такого виду:

	Var1	Var2	Var3	Var4
Var1	1.00			
Var2	*	1.00		
Var3	*	*	1.00	
Var4	*	*	*	1.00

У таких таблицях там, де змінна корелює сама з собою, проставлена одиниця, а в останніх пересіченнях, позначених зірочкою, комп'ютер проставляє конкретні значення коефіцієнтів кореляції. Матриця симетрична в відношенні головної діагоналі і тому має сенс заповнити тільки одну її половину.

Для оволодіння навичками комп'ютерного кореляційного аналізу слід виконати роботи № 14 і 15.

Для засвоєння методики використання критерію Ст'юдента для порівняння середніх арифметичних, коефіцієнтів кореляції і даних, виражених у відсотках, виконайте роботу № 20.

ОДНОФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ (ANOVA)

Теоретичні основи методу

Дисперсійний аналіз є одним із найбільш поширених у сільському господарстві методів математичної статистики. Цей метод активно використовується також і в біології. Він дозволяє знаходити відповідь на питання: чи вірогідний вплив того або іншого фактора (добрива, поливу, режиму годування тварин, нової технології і т.п.) на рослини, які вивчаються, та їх урожай, на сільськогосподарських тварин або на результати впровадження тих чи інших технологій. Він також дає можливість порівнювати між собою декілька системно зв'язаних вибірок і визначати, чи маються між ними статистично вірогідні відмінності і яка імовірність цих відмінностей.

Типовими випадками застосування дисперсійного аналізу в сільському господарстві та біології є: а) порівняння декількох сортів однієї культури або декількох порід домашніх тварин за будь-якою із кількісних чи якісних ознак; б) установлення реакції рослин або тварин на той або інший зовнішній вплив (пестициди, добрива, харчі, способи догляду, особливості місцезростання або заселення і т.п.).

У всіх моделях дисперсійного аналізу перевіряється дія деякого загального фактора (в однофакторному дисперсійному аналізі - одного просто фактора, у двохфакторному або трьохфакторному - одночасно двох або трьох факторів) на об'єкт. В якості такого загального фактора можуть бути геном рослин (сорт) або тварин (порода), добрива, спосіб обробітку грунту і т.п.

Фактор у загальному випадку - це та чи інша форма впливу на об'єкт, а також ознака або властивість об'єкта.

Для з'ясування вірогідності дії цього фактора на об'єкт фактор обов'язково повинен бути розбитим на дози або види впливу. У дисперсійному аналізі як методі математичної статистики ці підрозділи факторів називають *рівнями*, або *градаціями*. В сільському господарстві такі розбивки за звичай називають *варіантами* досліду. Кількість рівнів фактора, тобто варіантів досліду, повинно бути не менше двох. Верхню межу кількості варіантів (рівнів) «Статистика» не обмежує.

Так, під час випробування сортів в якості «дози» буде виступати кожний сорт і потрібен набір сортів (не менше двох) для проведення дисперсійного аналізу. При вивченні вірогідності дії азотних

добрив на деяку культурну рослину потрібно це добриво розбити на дози, наприклад, N_{40}, N_{60} і N_{80}. В цьому випадку варіанти складуть ці дози і доз також повинно випробовуватись не менше 2. В якості доз можуть виступати види добрив або пестицидів, а також типи обробки грунту - звичайна оранка, глибока оранка, плантажна оранка. В якості факторів можуть також використовуватись окремі бригади або виробничі підрозділи, поля, календарні дати, способи лікування тварин та ін.

В сільському господарстві, як уже відзначалось, рівні фактора часто називають *варіантами*. Один з варіантів може при цьому розглядатися як *контроль*, або стандарт, але на розрахункову процедуру це не впливає. Виділення контролю робиться тільки для зручності інтерпретації результатів роботи.

В об'єктів, що знаходяться під впливом фактора, який вивчається, або факторів, ураховується деяка важлива для дослідника ознака. Її називають *відгуком*. Відгук у дисперсійному аналізі, як правило, один.

Установивши дози, ми повинні одержати для аналізу вибірку (облік результату по одиничному об'єкту був би частіше всього ненадійним), і для цього кожний варіант реалізується декілька разів. Це називають *повторностями досліду*, або *повторенням*. Але можливе проведення деяких видів дисперсійного аналізу і без повторень.

У сукупності фактори з їх рівнями, ураховувана ознака (відгук) і повторення складають *дисперсійний комплекс*.

Дуже важливо правильно сконструювати дисперсійний комплекс і чітко виділити діючий фактор і його дози. Наприклад, грубою помилкою буде такий дисперсійний комплекс, де в якості доз виступають: 1 - мілка оранка; 2 - звичайна оранка і 3 - добрива. Перші два і третій фактори явно «із різних опер». Комп'ютер, звичайно, буде рахувати, але це гра з цифрами, а не сільськогосподарський аналіз. На етапі складання дисперсійного комплексу помилки зустрічаються найбільш часто. Вибір фактора і його розчленування на рівні (варіанти) залежить від професійної підготовки фахівця і не має типового формального рішення. **Вибір типу градієнта (фактора) і розбивка його на дози (градації), і підбір суттєвого відгуку - це завжди змістовне творче завдання.**

Математична теорія дисперсійного аналізу підрозділяє його на ряд основних моделей або типів. Облікові процедури для них різні, але потужні статистичні пакети можуть автоматично підбирати до даних правильну модель при незначній допомозі дослідника.

Модель I. В цьому випадку рівні факторів установлюються дослідником і є фіксованими. Такій моделі відповідає дисперсійний аналіз, наприклад, з дозами якого-небудь добрива.

Модель II, або модель-компонент дисперсії. В цьому випадку рівні фактора визначаються його природною мінливістю. Вони немовби стихійні. Так, якщо в якості фактора обирається білковість зерен пшениці, то, очевидно, рівень білковості властивий кожному окремому зерну або партії зерна і його доводиться ураховувати, а не установлювати.

Модель III. Використовується тільки в двох- або багатофакторному аналізі. Її ще називають змішаною, тому що в цьому випадку один фактор відповідає моделі I, а другий - моделі II.

Ієрархічна модель, яку також називають *вибірка із вибірок*. У цій моделі, яка використовується в двох- або багатофакторному аналізі, один фактор більш важливий, а другий - другорядний і якби включається в нього. Так, в якості першого фактора можуть виступати заводи, а в якості другого - цехи. В другому випадку першим фактором можуть бути поля, а другим - ділянки на цих полях.

Будь-яка із цих моделей при її реалізації може бути упорядкованою або рендомізованою. *Упорядкована модель* має дисперсійний комплекс з строго визначеним взаємним розміщенням факторів та їх рівнів. *Рандомізована модель* відрізняється їх випадковим розміщенням. Кожний фактор і кожний рівень у такій моделі має рівну імовірність і немає потреби фіксувати їх положення стосовно один до одного.

Засновник методу дисперсійного аналізу Р.Фішер встановив, що метод добре працює, якщо кількість варіантів (доз) дорівнює числу повторень. Будемо позначати варіанти літерами латинського алфавіту, а повторення цифрами. Тоді ми одержимо, записуючи варіанти в рядках, а повторення в стовпчиках, наступну схему дисперсійного комплексу:

a1	a2	a3
b1	b2	b3
c1	c2	c3

Такі дисперсійні комплекси називають *блочними плачами*, а сама схема одержала назву *латинського квадрата*. Вона обов'язкова для планування будь-якого сільськогосподарського досліду чи спостереження, яке проводиться з метою подальшого вживання дисперсійного аналізу. Всередині латинського квадрата вікна можуть розміщуватись упорядковано, як у наведеній схемі, а можуть розміщуватись стосовно один до одного стохастично, тобто

випадково. Рандомізовані блочні плани кращі, ніж систематичні. Кожний блок розбивається на стільки ділянок, скільки рівнів установлено у діючого фактора.

У дисперсійному комплексі кількість повторень по кожному рівню фактора може бути однакова. Такі комплекси називають *рівномірними*. Якщо кількість повторень неоднакова (це часто буває наслідком неможливості одержати дані через випадкову смерть дослідної тварини, знищення врожаю градом або шкідниками тільки на одній або декількох ділянках і т.п.), дисперсійний комплекс є *нерівномірним*. Нерівномірні комплекси можна піддавати дисперсійному аналізу, а можна спочатку відновити в них дані, що випали, за допомогою спеціальних методів.

В латинському квадраті число варіантів дорівнює числу повторень. На щастя, метод володіє деяким запасом тривкості і незначні відхилення від латинського квадрата допустимі. Так, можливо використати 4 варіанти при 3-х повтореннях кожного. Але порівнювати 10 варіантів при 3-х повтореннях кожного вже не припустимо.

Небезпечність даної проблеми полягає в тому, що комп'ютер буде лічити все, що ви вводите, тільки результат буде безглуздим. Справді, деякі статистичні пакети при порушенні принципу латинського квадрата запитують користувача: чи лічити цей безглуздий матеріал?

Навпаки, точність дисперсійного аналізу підвищується, якщо кількість повторень більша кількості варіантів. У цей бік ви можете відхилятись без усяких побоювань.

Так що **правильно складений дисперсійний комплекс - це такий, де кількість варіантів дорівнює кількості повторень або кількість повторень більша, ніж кількість варіаитів.**

Зрозуміло, окремі блоки не обов'язково повинні розміщуватись поряд. Вони можуть бути і на сусідніх полях чи фермах. Таким чином, у таких дисперсійних комплексах з'являється другий фактор - блоки. Такий дисперсійний аналіз за необхідністю стає двохфакторним. Звісно, самі ефекти блоків для дослідника мало цікаві, але дуже важливо, щоб їх вплив був мінімальний.

В цілому **дисперсійний аналіз може бути застосований тільки до матеріалу, який правильно зібраний, і дослід (експеримент) був заздалегідь спланований в розрахунку на обробку методом дисперсійного аналізу.** До будь-якого випадкового набору даних застосовувати цей метод не можна.

В дисперсійному аналізі, як і взагалі в математичній статистиці, вживається нульова гіпотеза H_0:a = b = c = d = ..., тобто між варіантами немає різниці, і фактор, що вивчається, на даний об'єкт не

чинить статистично значущого впливу. Завдання і часто бажання дослідника полягають у тому, щоб відкинути нульову гіпотезу і знайти об'єктивний доказ впливу фактора і відмінності варіантів. Нульова гіпотеза відкидається, і фактор, що вивчається, приймається як впливаючий на враховану ознаку в сільському господарстві звичайно при $p<0,05$.

Математична ідея дисперсійного аналізу по суті дуже проста. Від варіанту до варіанту ми, звісно, маємо деякий розкид даних - дисперсію $s^2_{\text{варіантів}}$. Буде такий розкид - дисперсія, і від повторення до повторення - $s^2_{\text{повторень}}$. Якщо

$$s^2_{\text{варіантів}} > s^2_{\text{повторень}},$$

то фактор, що вивчається, вагомо впливає на об'єкт. У протилежному випадку цього впливу немає, і ми спостерігаємо суто вільне варіювання.

Наведені дві дисперсії розраховують за різними формулами. Виходячи з нульової гіпотези H_0: a = b = c = d = ..., обчислюється дисперсія $s^2_{\text{варіантів}}$ за формулою:

$$s^2_{\text{варіантів}} = \frac{1}{k-1}\sum_{j=1}^{k} n_j(x_j - \bar{x})^2 .$$

Зовсім іншим способом і незалежно від H_0 обчислюється дисперсія повторень:

$$s^2_{\text{повторень}} = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{ij})^2 .$$

Тоді, коли мається статистично значуща відмінність між варіантами, дисперсія варіантів має тенденцію до зростання. У протилежному випадку дисперсія варіантів і дисперсія повторень приблизно дорівнюють одна одній. Виходячи із цієї закономірності, знаходять величину, яка одержала назву критерію Фішера і позначається літерою F:

$$F = s^2_{\text{варіантів}} / s^2_{\text{повторень}}.$$

Якщо $F_{\text{фактичне}} > F_{\text{табличне}}$, то з визначеною вірогідністю (p) фактор, що вивчається, впливає на об'єкт, тобто сорти відрізняються врожайністю, рослини реагують на добрива і т.п. Якщо цього немає, то в даному досліді вплив фактора не доведено і він відсутній.

Використовуючи пакет «Статистика», користувач одержує безпосередньо величину *p*, яка й служить критерієм - прийняти чи відкинути нульову гіпотезу.

Результати і форма їх представлення

Розрахунки дисперсійного аналізу дуже громіздкі і при розрахунках вручну вимагають багато часу. Пакет «Статистика» виконує їх швидко і видає основні результати. Це

df Effect - ступені свободи для діючого фактора;
SSEffect - сума квадратів по ефектах;
MS Effect - середні квадрати для діючого фактора (варіантів);
SSError - сума квадратів для похибки;
df Error - ступені свободи для похибки;
MS Error - середні квадрати для похибки (повторень);
F - величина критерію Фішера;
p-level - рівень вірогідності в частках одиниці на користь нульової гіпотези.

	df Effect	MS Effect	df Error	MS Error	F	*p*-level
Фактор						

Ці результати завжди оформляються у вигляді таблиці (у випадку використання модуля ANOVA/MANOVA).

Найбільш важливий тут один показник - навіть не критерій Фішера, хоча зовнішня мета аналізу якраз є в його знаходженні, - а рівень вірогідності. При $p<0,05$ на користь нульової гіпотези менше 5% шансів, і вона відкидається. Цим доводиться, що фактор, який вивчається, впливає на об'єкт з імовірністю не нижче 95%. Робиться висновок, що цей фактор статистично значуще впливає на об'єкт. Якщо *p* дорівнює або більше 0,05, то нульову гіпотезу відкинути не можна. Роль фактора залишається недоведеною.

Користувача все ж таки цікавить і деяка інша додаткова інформація, і дисперсійний аналіз її представляє. Перш за все тут легко одержати середні арифметичні для кожного варіанту. Далі стає можливим порівняти варіанти попарно і установити, які із них відрізняються один від одного статистично значуще і який рівень цієї вірогідності. Для цього пакет «Статистика» пропонує два методи: LSD-тест і Scheffe-тест. Можна одержати й загальні параметри моделі.

Потрібно відзначити, що в докомп'ютерний період і в зв'язку з бідністю обчислювальної техніки для порівняння варіантів пропонувався один критерій HIP (найменша істотна різниця). Він обчислювався в абсолютних одиницях, і ступінь його довіри не визначався. Крім того, HIP будувався з розрахунку, що всі варіанти рівноправні, чого не буває не практиці. В розвинутих країнах у такій формі HIP ніколи не застосовувався, тому що рівні імовірності «*p*», які видає комп'ютер для кожної пари варіантів, несуть саме ту інформацію, яка потрібна фахівцю. В статистичних пакетах відсутній і показник точності досліду, як не маючий теоретичної основи. Зате для порівняння варіантів в абсолютних одиницях у пакеті «Статистика» мається критерій Дункана.

В даному посібнику орієнтація йде на теоретичну базу математичної статистики і сучасні статистичні пакети, тому такого виду статистики не рекомендуються.

При введенні даних у комп'ютер дисперсний аналіз має свої особливості. Користувач має варіанти і повторення і повинен якимось чином дати знати комп'ютеру, де що знаходиться. Самий простий спосіб (хоча їх існує декілька) - це відвести перший стовпчик (VAR1) для вводу коду варіантів або, простіше кажучи, їх номерів (1, 2, 3, ...). Другий стовпчик (VAR2) відводять для запису самих облікових даних, але так, щоб числа для одного варіанту опинились проти свого коду (порядок при цьому не має ніякого значення!).

Залишки

Будь-який дисперсійний комплекс містить в собі після його аналізу випадковий залишок. Він являє собою варіабельність, яка була не врахована дослідником. Залишки необхідно перевіряти. Вони повинні відповідати низці вимог і в тому числі: а) бути взаємно незалежними; б) мати однакову дисперсію і в) розподілятись у відповідності з нормальним статистичним розподілом.

Якщо залишки великі й нерівномірні або в них прослідковується деяка тенденція зміни - регулярність, то це вказує на те, що модель була підібрана неправильно або дисперсійний комплекс взагалі був складений неправильно. Звичайно, такі ситуації виникають, коли дослідник упускає з виду важливі, які ведуть до результативної ознаки, фактори, а вивчає суто другорядні, або тоді, коли вихідні дані не відповідають нормальному розподілу. У першому випадку дослід треба проводити заново, а в другому - використати перетворення даних.

Перетворення вихідних даних

Вихідні дані повинні відповідати нормальному статистичному розподілу. Дисперсійний аналіз в його стандартних процедурах був розроблений саме для цього випадку.

Якщо дані все ж не відповідають нормальному розподілу, то справу можна поправити, використовуючи їх перетворення. Теорія таких перетворень була розроблена Бартлеттом. Так, у випадку відповідності даних розподілу Пуассона, корисно використати перетворення методом квадратного кореня, тобто з кожного значення вихідної ознаки необхідно здобути квадратний корінь Пуассона. При відповідності даних біноміальному розподілу слід використати арксинус перетворення (див. практичну роботу № 7).

У практичній частині навчального посібника буде розглянуто дві основні моделі однофакторного дисперсійного аналізу. Крім того, ви можете провести дисперсійний аналіз, коли дані деяких повторень відсутні, коли кількість повторень в різних варіантах не однакова і таке ін. При необхідності ці моделі легко освоїти самостійно, використовуючи Help пакета.

Для оволодіння навичками комп'ютерного дисперсійного аналізу необхідно виконати роботи № 17-19.

ДВОФАКТОРНИЙ ДИСПЕРСІЙНИЙ АНАЛІЗ (MANOVA)

Принципові основи двофакторного або взагалі багатофакторного дисперсійного аналізу, який скорочено називають MANOVA, аналогічні однофакторному дисперсійному аналізу. В цьому випадку кожний з факторів, що вивчаються, також підрозділяється на дози, або градації. Кількість таких доз у кожного фактора повинна бути не менше двох. Необхідна також повторність кожного із варіантів досліду, тобто кожної дози. Результативна ознака (відгук) в цьому випадку знову ж таки тільки одна і загальна для всіх факторів, що вивчаються.

Підсумком двофакторного дисперсійного аналізу є оцінка дії не тільки кожного з факторів, що вивчаються, на результативну ознаку, але й оцінка їх взаємної дії на цю ознаку. У випадку дисперсійного аналізу при двох факторах - А і Б – отримується три оцінки: критерій Фішера для фактора А, критерій Фішера для фактора Б і критерій Фішера для взаємодії АхБ. Кожному критерію Фішера відповідає свій рівень вірогідності – p.

При трьохфакторному дисперсійному аналізі, коли фактори можна позначити як А, Б і В, підсумком буде шість рядків відповідей:

А
Б
В
АхБ
АхВ
ВхБ

Видно, що при нарощуванні кількості факторів, що вивчаються, стає важко пояснити сутність їх дії – загальна дисперсія немовби розпиляється між багатьма факторами та їх комбінаціями. Тому в практиці дисперсійного аналізу рідко використовують більше ніж трьохфакторні схеми.

Підсумком двохфакторного дисперсійного аналізу, як і при однофакторному аналізі, є обчислення критерію Фішера й рівня його статистичної значущості для нульової гіпотези про відсутність дії кожного із факторів, що вивчаються. Нульова гіпотеза звичайно відкидається, і фактор вважається статистично вірогідно діючим при $p<0,05$. Додатково вичисляють силу впливу фактора і виявляють

за допомогою критерію Шеффе чи інших (Дункана, LSD) варіанти досліду, які вірогідно відрізняються один від одного.

Двофакторний аналіз і багатофакторний дисперсійний аналіз можна проводити й на основі дисперсійних комплексів, в яких відсутні повторення. Але тоді не можна виявити взаємодію факторів, що вивчаються.

Виконайте практичні роботи № 21 і 22.

РЕГРЕСІЙНИЙ АНАЛІЗ. ЛІНІЙНА РЕГРЕСІЯ

В сільському господарстві та біології часто виникає потреба вивчати дію тих або інших факторів на рослини, тварин або грунти. При цьому фактори за звичай розбивають на дози і саму дію виконують шляхом використання наростаючих або убуваючих доз фактора, що вивчається. Метою таких робіт є знаходження "відгуків" живих організмів на такі дії (звичайно це величина врожаю, стан рослин, продуктивність тварин та ін., які ураховуються за якоюсь практично важливою ознакою). Наприклад, саме так вивчається дія добрив і пестицидів, норм поливної води, лікарських препаратів і т.п. Задачі такого типу розв'язує регресійний аналіз.

При регресійному аналізі установлюють, як змінюється відгук по мірі наростання дози діючого фактора, на скільки значні такі зміни і який ступінь їх статистичної вірогідності. Самі по собі відгуки для наочності і більш правильного аналізу звичайно зображають графічно. Вони можуть мати вигляд прямої лінії або кривої і в тому числі одно- і багато вершинної (рис. 10). Нами буде розглянуто тільки найбільш простий випадок – аналіз відгуків, які можна показати прямою лінією. Цей різновид регресійного аналізу називається **лінійною регресією**.

Проведення прямої лінії через хмаринку точок на графіку має назву "апроксимація прямою лінією". Вона дуже широко використовується в сільському господарстві, тому що з великою статистичною надійністю виявляє загальну тенденцію в дії того чи іншого фактора на об'єкт, що вивчається. Але в випадку чітких криволінійних залежностей користуватись лінійною апроксимацією не слід.

Вихідний матеріал для регресійного аналізу має вид парних даних: x_1-y_1, x_2-y_2, x_3-y_3 і т.п., де x – доза фактора, а y – відгук. Загальну кількість таких пар позначають через n. Вона повинна бути не менше 5-7. Кожному "x" повинен відповідати як мінімум один "y", а краще – декілька "y". При графічному зображенні дози фактора, як правило, відкладають на осі абсцис, а відгук – на осі ординат.

У підсумку регресійного аналізу одержують аналітичний вираз для прямої лінії регресії з визначенням кількісних значень рівняння регресії, а також зображенням їх у вигляді графіка. При знаходженні параметрів регресії широко використовується метод найменших квадратів (LSM). Він грунтується на тому, що лінія регресії

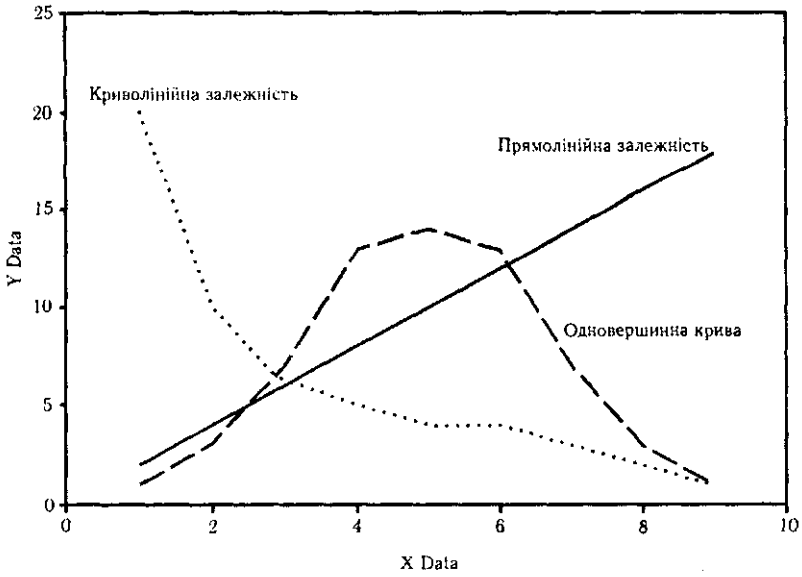


Рис. 10. Основні види регресійних залежностей

через хмаринку точок повинна проходити так, щоб відстані всіх точок від неї були найменшими. Оскільки "+" і "-" відхилення при цьому методі гасять один одного, то для оцінки залишкової дисперсії відхилення підносять у квадрат. Ці дві процедури й дали назву методу – метод найменших квадратів.

Загальна модель регресії має вид:

$$y = f(x) + \varepsilon,$$

де y - відгук, x - діючий фактор, ε *- випадкова помилка.*

При простій лінійній регресії аналітичний вираз для прямої лінії виглядає як

$$y = a + bx + \varepsilon \text{ або } y = b_0 + b_1x + \varepsilon.$$

Частіше використовується перша форма запису. Коефіцієнт *b* являє собою оцінку dy/dx, тобто зміну *y* на кожну одиницю *x*. Це середня реакція відгуку на зміну дози фактора. Англійською мовою цей параметр називають **slope** (нахил). Вільний член рівняння *a* **(intercept)** – це оцінка y при $x = 0$. Це ситуація, коли стан відгуку вільний від дії фактора. Знаки при коефіцієнті рівняння вказують на нахил прямої лінії: її зростання чи падіння при збільшенні доз фактора.

Вільний член і коефіцієнт вираховують за такими рівняннями:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}, \quad a = \frac{\sum y}{n} - \frac{b \sum x}{n}.$$

Регресійний аналіз дозволяє знаходити довірчі рівні лінії регресії. Розрахунок грунтується на параметричних оцінках рівняння регресії. Довірчий інтервал знаходять за наступними рівняннями:

$$s^2_e = \frac{1}{n-2}\{\sum y^2 - \frac{(\sum y)^2}{n} - \frac{b}{n}[n \sum xy - (\sum x)(\sum y)]\},$$

$$\text{var}\, y = s^2_e \cdot \frac{N^2}{n}(1 + \frac{1}{n})(1 - \frac{n}{N}),$$

$$s_{\bar{y}} = \sqrt{\text{var}\, y}$$

при df = n-2 і де n – кількість y, N – кількість x.

Обсяг розрахунків такий великий, що їх комп'ютеризація цілком виправдана.

Довірчі інтервали при графічній побудові показують пунктирними лініями зверху і знизу лінії регресії. Звичайно із-за малої кількості спостережень при невеликих дозах фактора й великих дозах фактора в цих зонах межа довіри розширюється (рис. 11). Довірча зона визначає ту ділянку на графіку, в межах якої з 95%-ю імовірністю находяться значення відгуку.

При практичному виконанні регресійного аналізу необхідно керуватись такими правилами:

1. Дані, тобто *x і y*, повинні бути незалежними один від одного, так, наприклад, кількість листя і вага листя – ознаки незалежності, і стосовно їх зв'язку можна проводить регресійний аналіз. Але якщо дослідник знаходив вагу листків не шляхом їх зважування, а методом перемноження ваги "середнього листка" на їх кількість, то дані будуть залежними і їх регресійний аналіз не має сенсу.

2. Виміри *y* (відгуку) повинні мати однакову точність по всій амплітуді відгуку. Результат буде спотворений, якщо наприклад, при

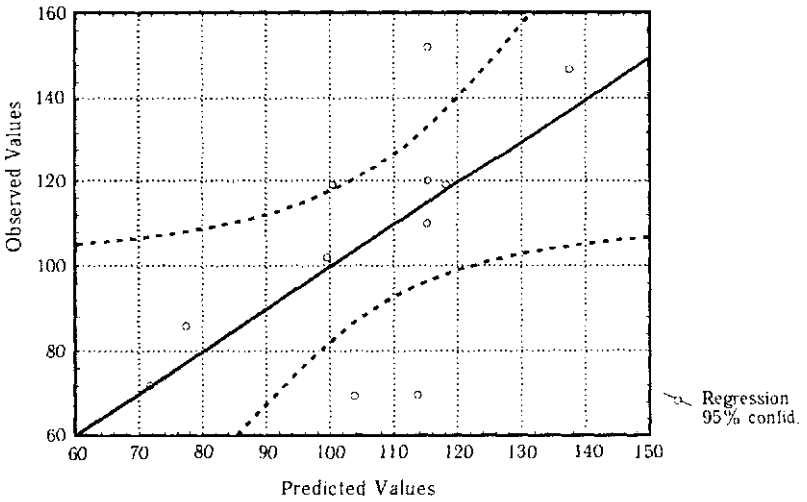


Рис. 11. Лінія регресії та 95%-ві довірчі інтервали (пунктирні лінії)

зважуванні малої ваги користувалися вагами, які не придатні для цього й не розраховані на малі навішення.

3. Після проведення регресійного аналізу дослідник обов'язково повинен переглянути графік залишків. Якщо залишки мають нормальний статистичний розподіл, то регресія увібрала в себе всю основну дію фактора. В протилежному випадку треба припустити, що деякий провідний фактор був не врахований, а регресійному аналізу піддався другорядний фактор.

4. Регресійна модель повинна бути перевірена методом дисперсійного аналізу при нульовій гіпотезі $H_0 : b = 0$, тобто коли дія фактора на реагуючу ознаку відсутня.

5. Регресійна модель має бути оцінена значенням коефіцієнта детермінації R^2. Чим він ближче до 100%, тим модель підібрана більш правильно.

Регресійний аналіз – потужний метод, який дає чіткі й наочні відповіді на поставлені питання, але його використання, як і застосування інших методів математичної статистики, вимагає кваліфікації й уважності.

Виконайте практичні роботи № 23, 24.

МНОЖИННИЙ ПОКРОКОВИЙ РЕГРЕСІЙНИЙ АНАЛІЗ

У сільському господарстві, як уже відзначалось, часто доводиться вивчати дію тих чи інших факторів на рослини, тварин або грунти в наростаючому дозуванні. При цьому в умовах виробничих дослідів далеко не завжди можна виділити який-небудь один фактор як провідний й який визначає кінцевий ефект. Частіше дослідник стикається з ситуацією, коли "під підозрою" знаходиться відразу декілька таких факторів. Серед цього комплексу потрібно виділити найважливіші фактори й бажано упорядкувати їх по убуванню важливості для об'єкта, що вивчається. Задачі такого типу розв'язує метод покрокового множинного регресійного аналізу.

Рівняння множинної регресії має такий вид:

$$y = a+b_1x_1+b_2x_2+b_3x_3+...+b_nx_n+\varepsilon,$$

де y – відгук, x_1-x_n - діючі фактори, розбиті на дози,
ε - випадкова помилка,
*a – вільний член рівняння (**intercept**),*
b – коефіцієнт регресії.

Величина коефіцієнта *b* при цьому чи іншому x є прямо пропорційною внеску цього фактора у відгук і тому дозволяє оцінити вагомість цього ж фактора. Знання всіх коефіцієнтів *b* дає можливість розмістити фактори в порядку їх значущості й "відсортувати" другорядні фактори від найбільш важливих. Такі другорядні фактори надалі можна не включати в аналіз. Відкидання таких другорядних факторів здійснюється по-одному на основі того або іншого обраного критерію, тому метод й одержав назву "покрокового".

Рівняння множинної регресії при ручному обліку вимагають великих затрат часу. За рахунок комп'ютеризації задачі такого роду одержують швидке й ефективне розв'язання.

В цілому за рахунок покрокового множинного регресійного аналізу досягаються такі цілі:

1. Створюється рівняння регресії, за яким можна вираховувати кінцеві ефекти для будь-якого поєднання вивчених факторів.

2. Знайти коефіцієнти регресії та їх статистичну вірогідність і на цій основі упорядкувати фактори за їх важливістю. При цьому в

якості критеріїв оцінки удається використати такий потужний апарат, як дисперсійний аналіз і значення критерію Фішера.

3. В залежності від установленого порога значущості можна виключати із рівняння другорядні фактори, які мало впливають на відгук, а лише створюють зайвий інформаційний шум.

Для оволодіння навичками комп'ютерного покрокового регресійного аналізу слід виконати практичні роботи № 25 і 26.

КЛАСТЕРНИЙ АНАЛІЗ

Для наукових досліджень в галузі сільського господарства та біології типовою є задача групування об'єктів (рослин, тварин або грунтів) на основі їх схожості, а також задача класифікації об'єктів за комплексом ознак. Наприклад, без допомоги комп'ютера можна розмістити сорти якої-небудь сільськогосподарської культури за величиною врожаю в ряд у порядку збільшення врожаю, а потім згрупувати сорти в дві групи - низьковрожайні та високоврожайні. Але тоді, коли таке групування доводиться робити одразу за декількома ознаками (такими як величина врожаю, термін визрівання, кількість білка, кількість вітамінів і т.п.) і ці ознаки не змінюються узгоджено, то така задача для ручного аналізу стає практично нездійсненною.

Такі задачі - групування і класифікація об'єктів одразу за безліччю ознак - розв'язує один з методів багатовимірної статистики - **кластерний аналіз**. Кластерний аналіз одержав свою назву від англійського слова «кластер», що означає гроно, тісна сукупність. Дійсно, при кластерному аналізі схожі об'єкти об'єднуються в групи на основі їх подібності відразу за багатьма ознаками. В підсумку кластерного аналізу видаються відстані між об'єктами в багатовимірному просторі й так звані **дендрограми**, які наочно показують близькість і віддаленість об'єктів один від одного відразу за низкою їх властивостей. Кластерний аналіз має великі перспективи для використання в селекційній роботі й для фахівців із захисту рослин. Він знаходить широке застосування при класифікації грунтів і рослинності.

До останнього часу використання кластерного аналізу в сільському господарстві було незначним, бо він потребує виконання великого обсягу складних розрахунків, які практично не можливі на настільних калькуляторах. З появою сучасних комп'ютерних програм, і зокрема пакета «Статистика», кластерний аналіз став доступним для широкого кола спеціалістів.

Розрахункові процедури кластерного аналізу можуть бути різними. В одних випадках групування об'єктів може виконуватись на основі метрик їх схожості за комплексом ознак **(агломеративний спосіб)**, а в інших - за відмінностями за цими ж ознаками **(розділювальний спосіб)**. Вихідний фактичний матеріал повинен при цьому представляти матрицю даних виду:

$$X = \begin{pmatrix} x_{11} & x_{12} \dots & x_{1p} \\ x_{21} & x_{22} \dots & x_{2p} \\ x_{n1} & x_{n2} \dots & x_{np} \end{pmatrix},$$

де стовпчики відповідають об'єктам, а рядки - ознакам. Шляхом використання метрик подібності або відстаней така матриця перетворюється в матрицю подібності або відмінності виду:

$$D = \begin{pmatrix} d_{11} & d_{12} \dots & d_{1p} \\ d_{21} & d_{22} \dots & d_{2p} \\ d_{n1} & d_{n2} \dots & d_{np} \end{pmatrix}.$$

В якості метрики схожості-відстані (ці два показники легко перетворюються один в одного) найчастіше використовується **евклідова відстань**.

$$d(x_i, x_j) = [\sum_{h=1}^{p} (x_{hi} - x_{hj})^2]^{1/2}.$$

Для того щоб придати більшої ваги об'єктам, які значно віддалені за своїми властивостями від основної маси об'єктів, можна використати квадрат евклідової відстані. Є можливість використати метрику, яка називається «міські квартали» або **манхеттен-метрика (City-block (Manhattan) distance)**. Вона має вигляд

$$d(x_i, x_j) = \sum | x_{hi} - x_{hj} |$$

і дещо змазує великі відстані, бо дані не підносяться до квадрата, як у першій метриці.

Ступенева метрика (Power distance) дозволяє змінювати вагу і розмірність простору ознак:

$$d(x_i, x_j) = (\sum | x_{hi} - x_{hj} |^p)^{1/r},$$

де p контролює індивідуальну розмірність, а r - вага, яка використовується для оцінки відстаней між об'єктами. Ці параметри може

задати користувач. При $p=2$ і $r=2$ формула ступеневої метрики еквівалентна евклідовій відстані.

В деяких випадках як метрику схожості можна використати коефіцієнт кореляції Пірсона r, або його абсолютну величину шляхом відкидання знака, тобто $|r|$. Замінивши r на $1-r$, ми одержуємо міру відмінності між ознаками.

Із багатьох різновидів розрахункових процедур кластерного аналізу користувачеві рекомендується спочатку зупинятись на евклідовій метриці або квадраті евклідової метрики.

Різні технології обліку можуть надавати різним ознакам неоднакову «вагу», що веде до деяких відмінностей в одержаних результатах. Якщо вихідні ознаки вимірюються в різних одиницях, то доцільно до знаходження матриці подібності ці дані стандартизувати. Пакет «Статистика» виконує це за формулою:

Нове значення = [(Старе значення - Середнє по стовпчиках)/ Стандартне відхилення].

Розрахунки в кластерному аналізі мають покроковий характер. Усього реалізується $n-1$ кроків. На першому кроці знаходяться два об'єкти, відстань між якими за сукупністю ознак є мінімальною. Після цього із таблиці відстаней викидається два стовпчика й два рядки і процедура повторюється стосовно першої пари знайдених об'єктів, а потім їх групи, поки всі об'єкти не будуть вичерпані.

В побудові підсумкової дендрограми важливу роль відіграють правила побудови зв'язків між самими кластерами. Можна використовувати простий зв'язок **(Single linkage)**, коли із усіх відстаней береться їх середня, медіанний зв'язок **(Weighted pair-group centroid (median))** та ін. «Простий зв'язок», або інакше метод найближчого сусіда», ефективний , коли кластери самі по собі більш менш чітко віддалені один від одного. Якщо кластери розміщені близько один до одного, то цей зв'язок має тенденцію до їх об'єднання в один кластер. При розв'язанні таких задач краще використовувати інші типи зв'язків.

Деякі максимізуючі зв'язки мають тенденцію робити кластери однакового розміру, навіть якщо це й не відповідає фактичному положенню справи. Алгоритм «найближчого сусіда» напрости, як відзначалось, часто об'єднує в один кластер і несхожі об'єкти. Доцільно до набуття досвіду використовувати простий зв'язок або медіанний, які працюють найбільш об'єктивно.

У пакеті «Статистика» доступні різні методи розрахунків кластерного аналізу з видачею як графіка у вигляді діаграми, так і всіх «відстаней» між об'єктами в багатоознаковому просторі. Можливо

також проводити кластеризацію як об'єктів, так і ознак, уводячи відповідно опцію або **Variables**, або **Cases**. Кількість кластерів може видаватися об'єктивно за встановленим критерієм або задаватися дослідником. Перший варіант кращий. Потужність кластерного аналізу в пакеті «Статистика» унікально велика: можна робити класифікацію до 2100 об'єктів за 600 ознаками.

Пояснення (інтерпретація) підсумків кластерного аналізу вимагає певного досвіду і добрих знань особливостей вихідного фактичного матеріалу. Такий матеріал повинен бути репрезентативним, вибірки достатньо великими, набір ознак для аналізу продуманим. Деякі несуттєві ознаки при їх включенні до аналізу можуть тільки створювати «шум» і утрудняти розуміння результатів.

У якості кінцевих результатів кластерного аналізу розглядаються матриці відстаней та вертикальні дендрограми.

Для оволодіння навичками кластерного аналізу слід виконати практичні роботи № 27 і 28.

КОМП'ЮТЕРНЕ ПРОГНОЗУВАННЯ. МЕТОД ARIMA

В сучасному сільському господарстві нерідко доводиться аналізувати дані, які є результатом повторних спостережень. Це й динаміка росту рослин і накопичення біомаси, послідовність урожаїв за період часу в декілька років або десятків років, динаміка цін на сільськогосподарську продукцію і таке інше. Метою аналізу таких даних є, по-перше, виявлення закономірностей динаміки їх зміни й, по-друге, прогнозування процесу або явища на той чи інший по тривалості період часу. Особливо важливі прогнози, які дають можливість завчасно підготуватися до того чи іншого перебігу подій.

В основі комп'ютерного прогнозування лежать **ряди динаміки**, або, як їх звичайно називають, **time series**. Рядом динаміки називають таку послідовність цифр, в якій по мірі наростання однієї змінної t (звичайно це час, але це й необов'язково) інша змінна x змінюється певним чином.

У будь-якому динамічному ряду розрізняють декілька складових, які впливають на його тип:

а) тенденція, або тренд, що відображає загальну закономірність розвитку явища чи процесу в залежності від його автономних властивостей;

б) циклічні довготривалі коливання;

в) короткочасні (сезонні) коливання;

г) випадкові коливання (білий шум), коли сусідні члени ряду динаміки не корелюють, а їх середні дорівнюють нулю.

Завданням аналізу рядів динаміки й прогнозування є виділення тенденцій при усуненні всіх випадкових впливів на неї.

Такий аналіз не є простою й формальною задачею. Вона не може бути розв'язана за рахунок застосування того чи іншого стандартного алгоритму. Дослідникові перш за все потрібно проаналізувати ряд динаміки по суті, впевнитись у наявності тенденції в явищі або процесі, які вивчаються, підшукати їй реальне пояснення і лише тільки потім звертатися до математичного прогнозування.

Слід мати на увазі, що **прогнозувати можна тільки закономірні процеси**. Випадкові процеси, також як і процеси, які піддаються якісним зламам, не можуть забезпечуватись надійним прогнозуванням.

Технологія прогнозування динамічних рядів звичайно полягає в дослідженні декількох моделей з вибором на основі статистичних критеріїв тієї моделі, яка найбільш адекватно описує динаміку явища й, отже, дає найбільш вірогідний прогноз.

У рядах динаміки з чітко вираженою тенденцією прогнозування можна робити на основі звичайних прямолінійних і криволінійних регресійних моделей. Але частіше в сільському господарстві зустрічаються випадки, коли закономірність процесу виражена не дуже чітко й значно затушована випадковими коливаннями. Саме в цих випадках звертаються до спеціальних методів аналізу рядів динаміки.

Для проведення аналізу й прогнозування такі ряди доводиться підготувати заздалегідь. Із них видаляють випадкові коливання і не стаціонарність.

Є.Є. Слуцький (1960) встановив, що для динамічних процесів усіх типів (соціальних, економічних, метеорологічних, біологічних та ін.) властиві періоди підйому і спаду. Вони одержали назву **«хвиль Слуцького»**. В таких хвилях сусідні дані корелюють один з одним. Такі кореляції одержали назву **автокореляцій**. Досліднику для виявлення загальної тенденції ряду потрібно усунути як хвилі Слуцького, так і автокореляції. Автокореляції усувають підбором параметрів моделі.

Крім того, нестаціонарні ряди динаміки потрібно перетворити в стаціонарні. Стаціонарним рядом називається такий ряд, в якому середні значення $\bar{x}$ та їх дисперсії s^2 не є функцією від t і помітно не змінюються в часі. Приведення ряду до стаціонарності полегшує виділення в ньому загальної тенденції змін.

Для надання ряду потрібних властивостей при необхідності його трансформують, використовуючи звичайні методи. Найбільш корисне перетворення натуральними логарифмами з введенням зсуву **(Lag)** на деяке фіксоване число.

Пакет «Статистика» має широкий набір процедур аналізу й прогнозування динамічних рядів. Найбільш проста і в той же час ефективна із них - це процедура **ARIMA - авторегресія і ковзаючі середні.** Ця процедура включає в себе методи: а) попередньої оцінки рядів динаміки, б) приведення їх до стаціонарності і в) підбір оптимальної моделі.

Перш за все аналізований ряд слід перевірити на нормальному імовірнісному аркуші. Така процедура включена в ARIMA. Якщо аналізований ряд не відповідає нормальному статистичному розподілу, то його слід перетворити таким чином, щоб він до нього максимально наблизився. Звичайно досліджують і перевіряють декілька таких перетворень. Такі перетворення також включені в ARIMA.

В повністю випадковому ряду динаміки коефіцієнти автокореляції при різних зсувах (лагах) стають близькими до нуля і знаходяться в межах 95%-го довірчого інтервалу. Якщо в автокореляціях (а вони представлені в кореляції стовпчиками на діаграмі) є деяка закономірність, то бажано виявити номер лага, з якого вона починається, і ряд перетворити за рахунок введення відповідного зсуву. Наявність автокореляцій часом свідчить про існування в процесі **«сезонності»** (це не обов'язково коливання показника по сезонах вегетаційного періоду, а невеликі коливання взагалі). Щоб подавити сезонність, в модель ARIMA доводиться вставляти відповідний параметр.

Модель ARIMA сама по собі характеризується шістьма головними параметрами:

Estimate constant (Постійний член моделі)
Seasonal lag (Сезонний зсув)
p-autoregressiv. (Авторегресія)
q-moving aver. (Параметр для ковзаючої середньої)
P-Seasonal (Сезонність)
Q-Seasonal (Сезонний параметр для ковзаючої середньої)

Ці параметри позначаються цифрами 0, 1, 2 ... Після назви ARIMA в дужках звичайно вказують три такі цифри у вигляді ARIMA (x,x,x). В першій позиції в дужках стоїть параметр авторегресії. У другій - порядок для ковзаючої середньої, на третій - порядок диференціювання. Якщо виставляється 0, то це означає, що відповідний параметр не включено в модель. Наприклад, ARIMA (1,0,0) означає, що в модель включений тільки параметр авторегресії.

Надання наведеним вище параметрам моделі тих чи інших значень - завдання дослідника (хоча комп'ютер по умовчанню інколи пропонує деякі такі значення, але далеко не завжди вдало). Поки немає значного досвіду, потрібні параметри знаходять просто шляхом їх перебору, хоча це дещо довший шлях. Після гарного засвоєння теорій методу параметри знаходять на основі попереднього дослідження ряду динаміки. Спочатку слід працювати з параметрами нижчих порядків: 0,1,2, рідко 3. Ці порядки найчастіше й ведуть до одержання оптимальних моделей.

Після підбирання моделі ARIMA починається останній етап роботи - оцінка моделі. Справа в тому, що оптимальну модель удається звичайно підібрати не одразу. Досліджується декілька моделей і із них вибирається краща, на її основі й робляться прогнози

Критеріїв перевірки якості моделі багато, але всі вони зводяться до перевірки залишків і статистичної вірогідності параметрів моделі. Залишки повинні бути мінімальні, не закономірні і, як правило,

відповідати випадковому нормальному розподілу. Залишки автокореляції також повинні бути випадковими, не мати чітко видимого тренда і вкладатися в середину довірчої зони. Для перевірки моделі пакет «Статистика» також має відповідні опції.

Після прийняття моделі по таблиці визначаються прогнозні значення членів ряду і їх довірчі рівні. Видається й графік, на якому прорисовується лінія прогнозу і її довірчі межі (звичайно 90%-ві).

Слід мати на увазі, що будь-які методи комп'ютерного прогнозування вимагають від користувача високої фахової й математичної грамотності. Складання гарного, тобто дійсного прогнозу - це й наука, й одночасно мистецтво, що включає елемент інтуїції. Не випадково фахівці з прогнозування у всіх країнах високо ціняться й найбільш оплачуються.

Більш детально прогнозування динамічних рядів описано в книзі В.П. Боровікова та Г.І. Івченко «Прогнозирование в системе STATISTICA в среде WINDOWS» (М.: Изд-во Финансы и статистика, 2000).

Для оволодіння навичками комп'ютерного прогнозування рядів динаміки слід виконати практичні роботи № 29 і 30.

ПРАКТИЧНА ЧАСТИНА

ПРАКТИЧНА РОБОТА № 1

Мета роботи: Набути та відновити навички в роботі з персональним комп'ютером. Виконання основних процедур в середовищі Norton Commander, Windows 3.11, Windows 95 та Windows 98. На комп'ютері, з яким ви працюєте, спочатку має бути інстальований пакет Statistica.

Виконайте такі операції:

1. Увімкніть комп'ютер. В залежності від конфігурації комп'ютера ви можете опинитися в середовищі Norton Commander, Windows 3.11, Windows 95 та Windows 98. Це залежить від настройки конкретного комп'ютера.

2. Якщо першою завантажується оболонка Norton Commander, то вам необхідно загрузити Windows 3.11/95/98 командою Win, яка записується в командний рядок DOS.

3. Розгляньте інтерфейс Windows 3.11/95/98 і ознайомтесь з його основними елементами.

4. У Windows 3.11 знайдіть піктограму, яка звичайно називається Statistica і розкрийте її, клацнувши двічі лівою клавішею миші (ЛКМ). У вікні, яке з'явилось, знайдіть модуль **STA_DAT.EXE**. В залежності від мети роботи для первісного завантаження можна використати модуль **Basic Statistics/Tables.** Також двічі швидко клацніть ЛКМ по піктограмі запускаючого модуля. На її місці повинен з'явитись пісочний годинник, який вказує на початок запуску програми. Чекайте запуску програми.

5. В середовищі Windows 95/98 для запуску пакета "Статистика" достатньо зробити два швидких коротких клацання ЛКМ по піктограмі, яка називається Statistica. На місці піктограми також повинен з'явитись пісочний годинник, який вказує на початок запуску програми. Чекайте запуску програми.

6. В середовищі Windows 95/98 цей запуск здійснюється також і через команди:

Пуск
Виконати
STA_DAT.EXE

При деяких типах настройки конкретного комп'ютера вам необхідно вказувати повний шлях до нього і тоді потрібно знати, саме в якій папці находиться файл STA_DAT.EXE. Для знаходження

шляху запустіть диспетчер файлів і знайдіть, де міститься потрібний вам модуль. Можна скористатися й послугою опції "пошук".

7. Для виходу з пакета у верхньому рядку основного меню, клацнувши один раз ЛКМ, розкрийте підменю **File** і далі ЛКМ задайте команду

Exit

Ця команда в підменю стоїть останньою і інколи із-за довгого списку файлів її не видно. В цьому випадку потрібно провести відповідну настройку комп'ютера: розкрийте ЛКМ в рядку основного меню підменю **Options** і далі задайте команду **General**. У новому вікні в зоні з назвою **Recently used file list** зменшіть число до 5. Потім натисніть клавішу **OK**.

Вихід з пакета Статистика можна здійсніть за допомогою комбінацій клавіш ALT+F4.

8. Навчіться правильно виходити із системи Windows. Це робиться в Windows 3.11. через підменю послідовним виконанням команд:

Файл
Вихід з Windows
OK

В Windows 95/98 через команди

Пуск
Завершення робіт
Виключити комп'ютер

9. Після виходу із Windows ви можете виключити комп'ютер клавішею на передній панелі.

ПРАКТИЧНА РОБОТА № 2

Мета роботи: Знайомство з основними елементами інтерфейсу пакета Statistica й виконання найпростіших операцій за допомогою меню, підменю і піктограм.

Виконайте такі операції:

1. Загрузіть пакет Statistica. Для цього у вікні Windows, клацнувши лівою клавішею миші, активізуйте піктограму STA_DAT.EXE. Ви можете також завантажити цей пакет і іншими способами, які розглянуті в роботі № 1.

2. Відкриється основний модуль пакета **Data Management** (Керування даними) і допоміжний підмодуль. В цьому допоміжному вікні перелічені операції по роботі з електронною таблицею, яка поки що не активна і знаходиться на другому плані. Звичайно пакет настроюють так, що в основне вікно автоматично завантажується саме той файл даних, з яким користувач працював останній раз.

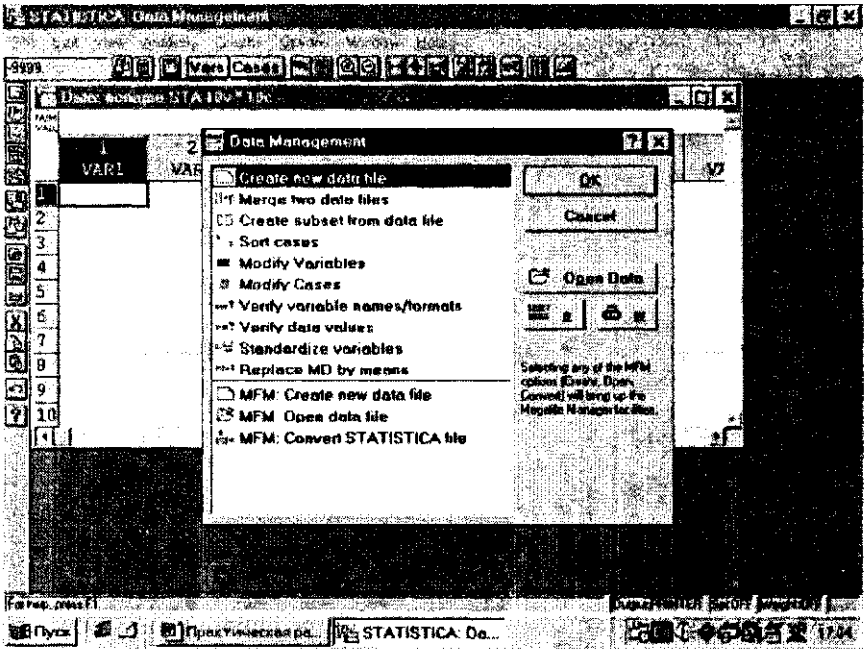


Рис. 12. Головні операції модуля Data Management

Головними операціями активного підвікна є такі (рис. 12):

Create new data file (Створити новий файл)

Merge two data files (З'єднати дані із двох різних файлів в один)

Modify Variables (Модифікувати змінні)

Modify Cases (Модифікувати повторення)

Standardize variables (Стандартизувати дані змінних)

Replace MD by means (Записати в пусті вікна для даної змінної значення середніх арифметичних цього статистичного ряду)

В цьому активному підвікні на початку роботи для вас буде важлива одна опція: **Create new data file** (Створити новий файл) – вона вас веде в підменю модуля створення бази даних і має містити в собі власне ім'я у вигляді імені файла з розширенням .STA, і параметри стовпців і рядків в електронній таблиці. (В залежності від настройки комп'ютера спочатку може появитись не підменю Data Management, а якесь інше.)

3. Прибрати це підменю, яким би воно не було, клацнувши клавішею миші по правій верхній піктограмі із значком хрестика X. Це ж саме ви можете зробити по-іншому: клацніть по лівій піктограмі, поміченій горизонтальною стрілкою. Відкриється підменю, в якому ви виберете опцію "Закрити" й, клацнувши по ній, закриєте це підменю.

4. Перед вами, як правило, відкриється файл, з яким користувач працював останній раз. Зручна послуга, коли з одними й тими ж даними працюєш декілька сеансів підряд. Ця база даних нас цікавити поки що не буде.

5. Зверніть увагу на верхній рядок основного модуля **Data Management** (рис.13). Він вміщує два рядки. У верхньому – даються назви "випадаючих меню" (File, Edit та ін.), а в нижньому – знаходиться серія піктограм, які дозволяють швидко виконувати деякі дії. Розглянемо перш за все можливості випадаючих меню. Для цього, клацнувши мишею, послідовно зліва направо викличте випадаючі меню та вивчіть їх можливості:

File – *в ньому містяться основні прийоми роботи з файлами*:

New Data (Нові дані) – дозволяють створити нову електронну таблицю, дати їй ім'я і потім уводити в неї власний матеріал.

Import Data (Імпорт даних) – за її допомогою можна ввести в пакет дані, які були створені в іншому програмному пакеті.

Save (Зберегти) – записує файл даних на диск для постійного схову, зберігаючи те ім'я, під яким файл уже існує.

Save file as (Зберегти як...) – записує файл на диск з новим ім'ям. В цьому випадку відкривається вікно, куди ви маєте це ім'я

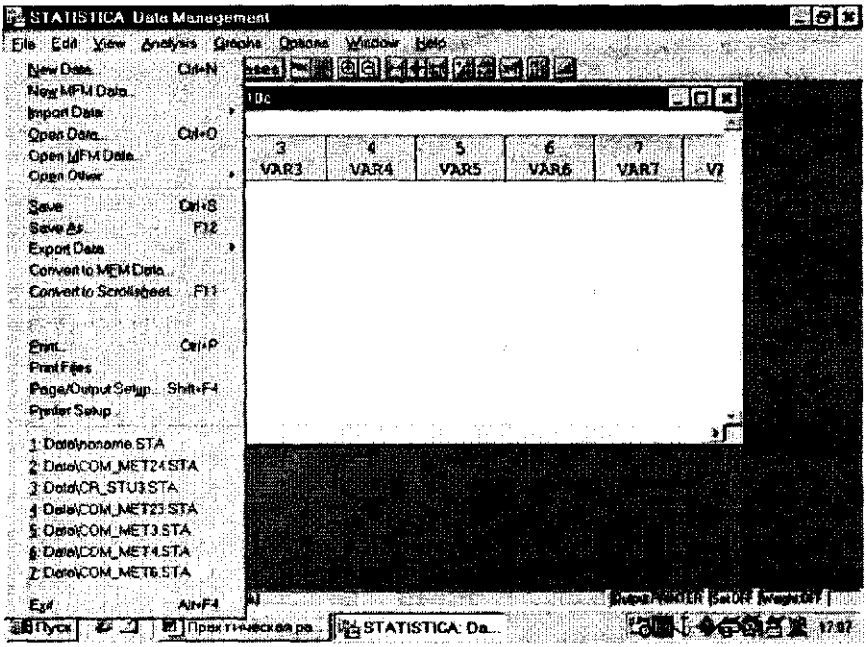


Рис. 13. Можливості випадаючих меню. Показане випадаюче меню підменю File

записати у відповідний рядок. Необхідно до цього імені зберегти розширення .STA.

Print (Друк) – веде у підмодуль друку, де ви можете настроїти принтер і дати команду для друку вашого матеріалу.

Edit *(Редагувати). В цьому випадаючому меню для вас важливі такі можливості:*

Cut (Вирізати) – вирізає вічко, в якому стоїть курсор або позначений блок тексту. Позначають блоки вікон рухом миші при натиснутій її лівій клавіші. Вирізану ділянку потім можна вставити в будь-яке місце електронної таблиці, тому що вона зберігається у буфері пам'яті. Для цього в цьому ж підменю обирається опція **Paste** (Вставити).

Copy (Копіювати) – копіює вічко або блок тексту у буфер пам'яті, не видаляючи його із попереднього місця. Скопійовану ділянку ви можете перенести в будь-яке нове місце

Clear (Очистити) – видаляє вічко або позначений блок.

Paste (Вставити) – вставляє із буфера пам'яті в місцеположення курсора ту частину, яка була вирізана командою Cut або знаходиться в буфері по команді Copy.

View *(Вид) – тут найбільш важливі опції:*

Fonts (Шрифти) – дозволяє вибрати зручний тип і величину шрифту.

Colors (Кольори) – змінює кольорову гаму інтерфейсу, якщо вже існуюча не влаштовує користувача.

Analysis *(Статистичні аналізи) – виводить список найбільш часто використовуваних аналізів.* Тут важливо уміти користуватися опціями:

Startup Panel (Вікно завантаження) – дозволяє відновити допоміжні меню й підменю, якщо ви їх раніше закрили.

Other Statistics (Інші статистичні методи) – веде в список модулів пакета Statistica. Для того щоб в нього попасти, спочатку потрібно відкрити його, а потім у списку вибрати опцію **Customize list** (Список для користувача).

В списку ви побачите всі модулі пакета Statistica і можете завантажити будь-який з них. При цьому база даних залишиться відкритою, і ви можете її обробляти тим модулем, який викликали. Для його виклику потрібно послідовно виконати такі опції:

Позначити модуль, клацнувши по ньому лівою клавішею миші

Replace (Замістити існуючий)

Switch to (Переключитися)

При деяких настройках комп'ютера команда **Switch to** очевидно не спрацьовує. Для її активізації слід виконати операції:

Analysis

Startup Panel

Graphs *(Графіка) – її підкоманди дають змогу будувати різні типи графіків.*

Options *(Настройки) – настройки пакета, які виконує досвідчений користувач або професіональний програміст.*

Windows *(Вікна) – настройка зображення вікон на екрані: поряд, одне поверх другого і т.п.*

Help *(Допомога) – вводить у вікна дуже потужну і добре організовану систему підказок.* Користуйтесь. Тільки вона англійською мовою.

6. Для зручної роботи із системою підменю мається цілий ряд кнопок-піктограм, які одним клацанням миші дозволяють увійти у відповідне підменю. Про сенс цих кнопок можна дізнатися, якщо

поставити на кнопку курсор миші і потримати 2-3 сек. – з'явиться пояснюючий напис. Найбільш важливі із них такі піктограми:

Vars (Змінні – це стовпчики в таблиці). Випадаюче меню вміщує багато корисних опцій для роботи зі стовпчиком таблиці даних.

Cases (Випадки, або Повторності – це рядки таблиці).

Ці дві пікто; ами дозволяють швидко зробити такі види робіт:

Add (Додати) – додати в базу даних нові стовпчики або рядки, якщо їх вам не вистачило при вводі даних.

Move (Перемістити) – пересунути дані або їх частину на нове місце.

Delete (Знищити) – стерти дані.

Current Specs (Поточні специфікації) – дозволяє змінити багато які властивості бази даних і файлу.

В наборі піктограм є ще дві корисні клавиші:

Increase Decimal та **Decrease Decimal**. Вони дозволяють збільшити або зменшити кількість знаків після коми в цифрах, які ви вводите в стовпчик або в уже введеному стовпчику.

На завершення практичної роботи потренуйтесь у користуванні всіма описаними опціями і піктограмами, використовуючи чисту електронну таблицю та будь-яку базу даних, яка є в пакеті.

ПРАКТИЧНА РОБОТА № 3

Мета роботи: Навчитися заповнювати даними електронну таблицю (створювати базу даних) та зберігати ці дані на вінчестері у вигляді файлу з власним оригінальним іменем.

Ці дані такі: два статистичних ряди, в кожному із них по 19 чисел-спостережень:

	VAR1	VAR2
1	21.2	15.6
2	37.7	15.1
3	1.3	8.1
4	8.1	16.5
5	21.5	7.2
6	1.9	11.4
7	13.3	2.6
8	11.2	17.0
9	1.0	4.5
10	43.9	1.8
11	3.5	4.9
12	33.3	7.9
13	22.9	1.2
14	28.6	8.5
15	8.0	4.0
16	13.7	11.9
17	47.0	4.1
18	5.3	19.2
19	3.5	12.5

1. Включити комп'ютер і після завантаження Windows увійти в середовище пакета Statistica за допомогою команди STA_DAT.EXE, тобто виконати операції 1-3 з роботи № 2.

2. Клацнувши ЛКМ, розкрийте випадаюче меню **Analysis** і далі виконайте такі команди:

Other statistics
Customize list
Data Management
Replace
Switch to

Можливо, що для активізації модуля вам необхідно буде виконати додатково команди:

Analysis
Startup Panel

Якщо ваш комп'ютер має такий тип настройки, то цю особливість потрібно враховувати при виконанні всіх подальших робіт.

3. Відкриється допоміжний модуль Data Management. У підменю цього ж модуля знайдіть рядок **Create new data file** (Створити файл даних). Активізуйте цей рядок ЛКМ і натисніть **OK**. У вікні (рис. 14), що з'явилось, вам потрібно перш за все дати ім'я створюваному новому файлу даних. Ім'я має записуватись латинськими літерами і бути не більше 8 літер. Наприклад, ви можете дати ім'я: LIDA. Для цього клацніть ЛКМ по кнопці **New file name** (Нове ім'я файлу), і в вікні **Specify file name** (Оригінальне ім'я файлу), що відкрилось, в рядку ІМ'Я ФАЙЛУ поставте курсор строго перед крапкою в імені файлу, і клавішею Backspace забийте слово **noname** (немає імені). Потім уведіть ім'я файлу. У вас утвориться lida.sta. Зверніть увагу на обов'язковість збереження розширення .STA. Клацніть ЛКМ по клавіші OK.

4. У вікні **Create new data file** з'явилось ім'я. Вам тепер необхідно визначити кількість змінних (стовпчиків) у вашому файлі даних, поставивши відповідну цифру (в нашому прикладі це цифра

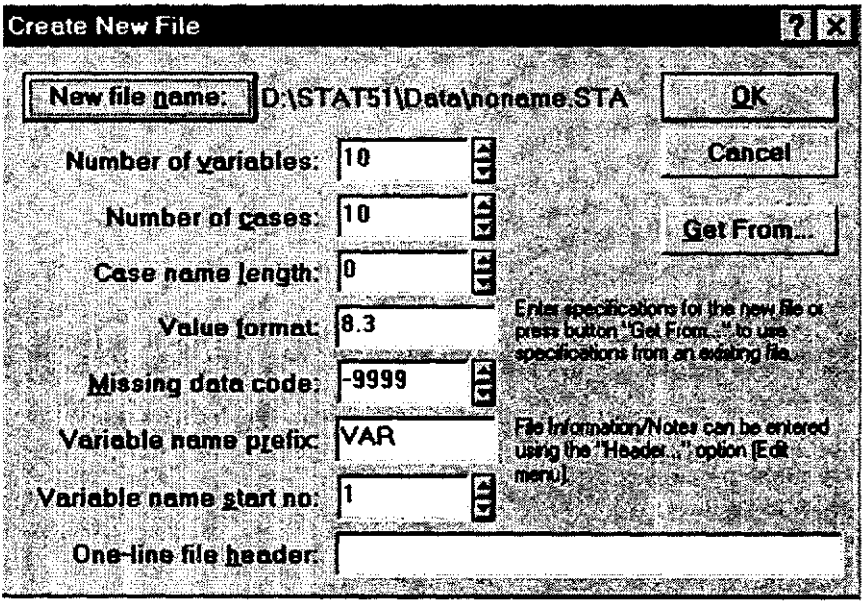


Рис. 14. Підмодуль, що дозволяє описувати властивості створюваного файлу з даними

2!) в рядку **Number of variables** стрілками або за допомогою ЛКМ. Потім визначте, якої довжини будуть ваші стовпчики, тобто кількість змінних у кожному статистичному ряду. Для цього установіть таким же способом це число (19!) у полі **Number of cases**. Останні поля не змінюйте і натисніть клавішу **OK**.

5. Ви одержали бланк електронної таблиці під оригінальним іменем і встановленого вами формату. Клавішею-піктограмою у формі квадратика у верхньому його лівому кутку його можна розгортати на весь екран або згортати – як вам буде зручно. Зверніть увагу, що змінні мають ім'я VAR і номер при ньому. Часто так можна і працювати. Але ви маєте можливість дати їм оригінальні імена. Для цього послідовно активізуйте піктограми для одного із стовпчиків даних:

VARS (змінні)

Current Specs... (Поточні специфікації)

В підменю в рядку **Name** наберіть ім'я змінної. Ви можете скористатись латинськими літерами, а можете дати ім'я кирилицею – але не більше 8 літер.

6. Тепер уведіть свої дані в електронну таблицю. Робиться це по вертикальним стовпчикам. Використовуйте для зручності в роботі також і праву цифрову клавіатуру, яку попередньо включіть клавішею **Num Lock**. Для розділення цілої і дробової частини чисел використайте крапку, а не кому. Проте це залежить від настройки системи Windows, з якою потрібно познайомитись.

7. Збережіть створений вами файл даних на вінчестері командами:

File

Save

з рядка основного меню Statistica.

8. Для перевірки своєї роботи вийдіть з Statistica, а потім знову увійдіть у неї. Оскільки ваш файл був створений останнім, він повинен відкритися автоматично. Але ви можете його викликати (якщо він не буде останнім) командами

File

Open data

У вікні Open data знайдіть свій файл, позначте його ЛКМ і натисніть клавішу **OK** (або відкрити).

9. Тепер навчіться роздруковувати на папері свій файл. Для цього вікно із даними повинно бути активним (його заголовок виділяється синім кольором). Послідовно активізуйте у випадаючому меню:

File

Print...

По настройці пакета вивід на принтер може бути закритим. І тоді комп'ютер запитає вас, чи відкрити такий тип виводу. Відповідайте – "так".

У полі модуля **Page/Output/Setup** (Настройки виводу сторінок) необхідно мати крапку в полі проти рядка **Text/...**, у полі **Output** позначте ЛКМ слово **Printer**. У полі ж **Output Header** (Вивід заголовка) ви можете будь-яким шрифтом набрати заголовок таблиці, яку ви хочете роздрукувати. Можна підібрати тип і розмір шрифту у підменю **Fonts** цього ж вікна. В підменю **Fonts** у полі "Набір символів" повинно стояти слово "*Кирилиця*", *якщо ви бажаєте дати заголовок таблиці українською чи російською мовами.*

Натисніть ОК.

У вікні **Print data**, що з'явилось, можна перевірити настройки виводу, змінювати поки що нічого не потрібно – просто натисніть **ОК**. Зрозуміло, принтер необхідно спочатку включити й заправити папером.

10. По закінченні роботи вийдіть з пакета Statistica і правильно виключіть комп'ютер.

ПРАКТИЧНА РОБОТА № 4

Мета роботи: Перевірка статистичних рядів на відповідність нормальному статистичному розподілу, яка дозволяє проводити їх подальшу обробку методами математичної статистики.

1. Включіть комп'ютер і після завантаження Windows увійдіть в середовище пакета Statistica за допомогою команди STA_DAT.EXE

2. Завантажте файл, створений вами на попередньому занятті, командами

File

Open data

Якщо файлу з готовими даними немає, створіть його, використовуючи матеріал, який був одержаний при обліці відсотка гемоглобіну у курей породи Бетнамки. Вони такі:

87-91-85-82-92-90-82-82-86-88-85-85-91-89-86-83-90

3. В основному меню виберіть випадаюче меню **Analysis**, а потім послідовно виконайте команди:

Other statistics (Інші статистики)

Customize list (Список користувача – це перелік модулів пакета)

Basic statistics – позначте її ЛКМ

Replace

Switch to...

У меню підмодуля Basic statistics виберіть:

Descriptive Statistics (активізуйте цей рядок ЛКМ)

OK

4. У великому вікні модуля натисніть клавішу **Variables** (Змінні) і в списку змінних (у вас їх дві), що з'явився, виберіть VAR1 (або те ім'я, яке ви самі дали цьому стовпчику даних) і позначте її ЛКМ.

OK

5. Знайдіть у вікні модуля **Descriptive Statistics** клавішу **Half-normal probability plots** (Напівнормальний імовірнісний графік) і активізуйте її ЛКМ. Ви одержите графік, на якому у випадку відповідності вашого статистичного ряду нормальному розподілу всі точки (це окремі x_i) *лежать на діагональній прямій лінії або поблизу неї.*

Розгляньте ваш графік і зробіть висновок про можливості обробки цих даних методами математичної статистики.

6. Натисніть на графіку клавішу **Continue** (Продовження роботи). Це поверне вас в основне вікно модуля **Descriptive Statistics**. Повторіть всю цю роботу стосовно другої змінної вашого масиву даних (якщо вона є). Зробіть висновок і про цю змінну.

7. Ви можете провести й більш жорстку перевірку ваших даних на нормальність. Для цього необхідно використати не клавішу **Half-normal probability plots**, а клавішу **Normal probability plots** (Нормальний імовірнісний графік). Виконайте цю роботу.

8. Зробіть підсумкові висновки про придатність ваших даних для статистичної обробки (рис. 15).

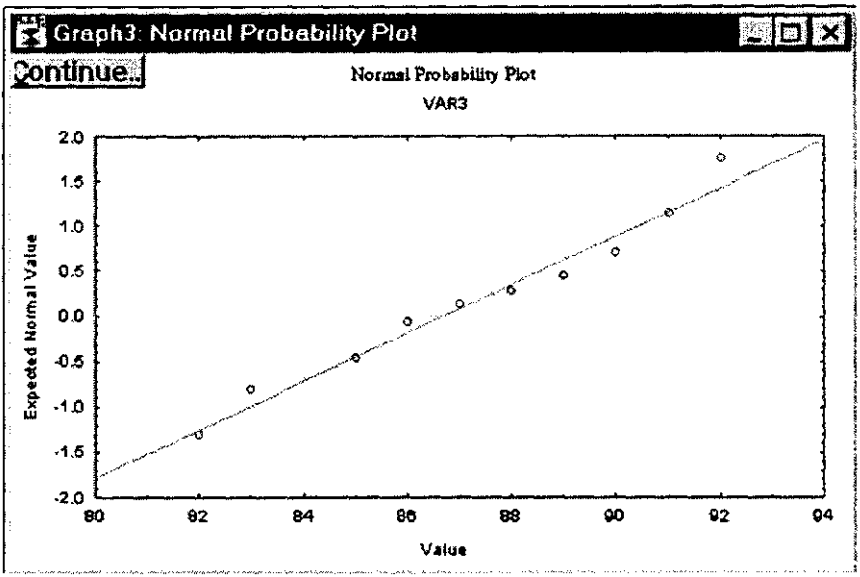


Рис. 15. Результати перевірки на відповідність нормальному статистичному розподілу даних практичної роботи № 4

9. Закрийте всі робочі вікна пакета Statistica знайомою вам клавішею-піктограмою "хрестик" або знаком "-".

10. По закінченні роботи вийдіть з пакета Statistica і правильно виключіть комп'ютер.

ПРАКТИЧНА РОБОТА № 5

Мета роботи: Навчитись одержувати основні статистичні параметри для вибіркових рядів (точкове оцінювання). Такими параметрами є: кількість членів в ряду, середнє арифметичне, стандартне відхилення, дисперсія, стандартна похибка середнього арифметичного, коефіцієнт асиметрії (скошеність ряду) й ексцес.

Для виконання роботи виконайте такі процедури:

1. Включити комп'ютер і після завантаження Windows увійти в середовище пакета Statistica за допомогою команди STA_DAT.EXE

2. Завантажити файл, створений вами при виконанні практичної роботи № 3, командами

File

Open data

Завважте, що файл даних можна відкрити й іншим способом: у випадаючому меню File в нижній його частині міється список з 3-5 (це залежить від настройки пакета) файлів, з якими ви працювали на останніх сеансах. Тому звертатись до команди **Open data** не обов'язково – можна просто активізувати потрібний вам файл в меню File і загрузити його ЛКМ. Крім того, останній файл даних, з яким працював користувач, при завантаженні Statistica відкривається автоматично.

3. В основному меню виберіть випадаюче меню **Analysis**, а потім послідовно виконайте команди:

Other statistics (Інші статистики)

Customize list (Список користувача – це перелік модулів пакета)

Basic statistics – позначте її ЛКМ

Replace

Switch to

У меню підмодуля Basic statistics виберіть

Descriptive Statistics (активізуйте цей рядок ЛКМ)

OK

4. У відкрившомуся вікні (рис. 16) підменю перш за все введіть ім'я статистичного ряду, для якого ви маєте обчислити точкові статистичні оцінки. У вас таких рядів два. Звичайно вони мають імена VAR1 і VAR2 (якщо ви не дали їм інших імен).

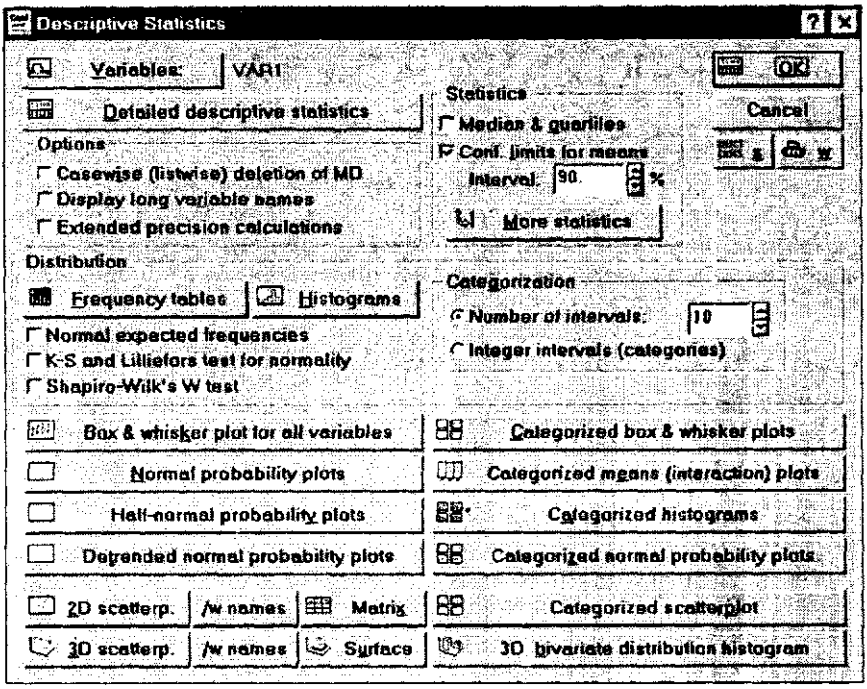


Рис. 16. Вікно підмодуля, що дозволяє вибрати варіанти для аналізу і визначити перелік виведених статистичних параметрів

Для вводу імені натисніть ЛКМ по клавіші Variables, у списку позначте ЛКМ VAR1 і натисніть **OK**.

Знайдіть клавішу **More Statistics** (Більше статистичних параметрів) і натисніть її ЛКМ. У списку статистичних параметрів, що відкрився, позначте галочкою, тобто клацнувши ЛКМ, ті параметри, які ви бажаєте обчислити для вашого статистичного ряду. Приберіть галочки проти тих параметрів, які вам не потрібні.

Вам необхідно позначити:

Valid N (кількість придатних даних)

Mean (середнє арифметичне)

Standard Deviation (стандартне відхилення)

Variance (дисперсія)

Standard Error of mean (стандартна похибка середнього арифметичного)

Skewness (асиметрія ряду)

Kurtosis (ексцес)
OK

Підменю закриється. Натисніть **OK** у вікні **Descriptive Statistics**. У вікні виводу результатів, що відкрилося, прочитайте і випишіть ці результати. Якщо їх всі не видно, то ви можете відкрити вікно на весь екран піктограмою "квадратик" (або знаком трикутника) чи просуватись по вікну за допомогою "повзунків" у правій і нижній частині вікна.

5. Натисніть клавішу **Continue** і проробіть всю цю роботу повторно для другого статистичного ряду VAR2.

Випишіть точкові оцінки для цього ряду і співставте їх з оцінками для першого ряду.

Майте на увазі, що ви можете одержати відразу статистичні параметри обох цих рядів, якщо у вікні **Descriptive Statistics** активізувати в списку одразу обидва ряди: VAR1 і VAR2. Це можна зробити, якщо держати натиснутою клавішу CTRL при позначці ЛКМ ім'я активної змінної.

6. Пакет Statistica не передбачає знаходження для статистичних рядів коефіцієнта варіації. Ви можете легко знайти цей коефіцієнт, знаючи стандартне відхилення і середнє арифметичне, за допомогою калькулятора, який входить до складу стандартних програм Windows.

7. Закрийте всі робочі вікна пакета Statistica знайомою вам клавішею-піктограмою "хрестик".

8. По закінченні роботи вийдіть з пакета Statistica і правильно виключіть комп'ютер.

ПРАКТИЧНА РОБОТА № 6

Мета роботи: Одержання інтервальних статистичних оцінок для статистичних рядів.

Для виконання роботи виконайте такі процедури:

1. Включіть комп'ютер і після завантаження Windows увійдіть в середовище пакета Statistica за допомогою команди STA_DAT.EXE

2. Загрузіть файл, створений вами при виконанні практичної роботи № 3, командами

File

Open data

Якщо у вас немає такого файлу, то створіть його заново із будь-яких даних, які маються у вашому розпорядженні.

3. В основному меню виберіть випадаюче меню **Analysis**, а потім послідовно виконайте команди:

Other statistics (Інші статистики)

Customize list (Список користувача – це перелік модулів пакета)

Basic statistics – позначте її ЛКМ

Replace

Switch to

У меню підмодуля Basic statistics виберіть

Descriptive Statistics (активізуйте цей рядок ЛКМ)

OK

4. У відкрившомуся вікні підменю **Descriptive Statistics** перш за все введіть ім'я статистичного ряду, для якого ви бажаєте обчислити точкові й інтервальні статистичні оцінки. У вас таких рядів два. Звичайно вони мають імена VAR1 і VAR2 (якщо ви не дали їм інших імен).

Для вводу імені натисніть ЛКМ по клавіші **Variables**, у списку позначте ЛКМ VAR1 і натисніть **OK**.

Знайдіть клавішу **More statistics** (Більше статистичних параметрів) і натисніть її ЛКМ. У списку статистичних параметрів, що відкрився, позначте галочкою, тобто клацнувши ЛКМ, ті параметри, які ви бажаєте мати для вашого статистичного ряду (рис. 17). Приберіть галочки проти тих параметрів, які вам не потрібні:

Вам необхідно позначити:

Valid N (кількість придатних даних)

Mean (середнє арифметичне)

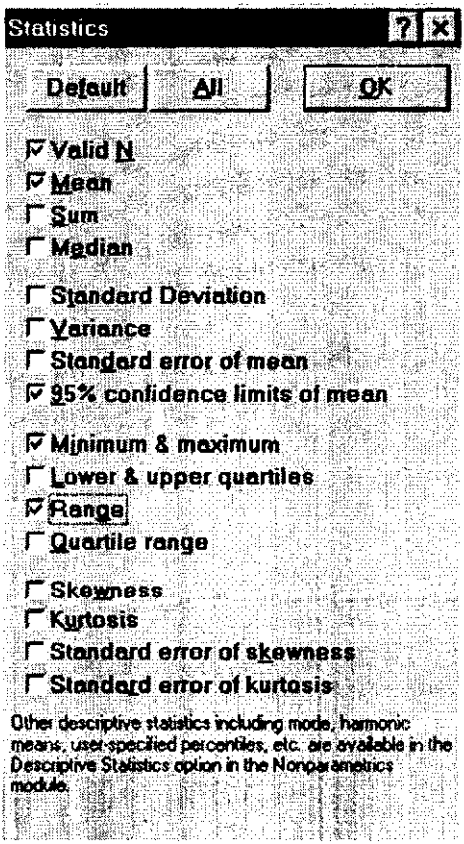


Рис. 17. Список статистичних параметрів, які можна обрахувати для даних, що опрацьовуються, в підмодулі More Statistics

95% confidence limits (95% довірчі інтервали)
Minimum @ maximum (найбільше і найменше значення)
Range (розмах)
Натисніть клавішу **OK**

5. Ви повернулись у підменю **Descriptive Statistics**. Натисніть у ньому **OK**.

6. В новому вікні ви одержали потрібні вам результати для ряду VAR1. Запишіть їх.

7. Повторіть роботу для VAR2, використовуючи клавішу **Continue**. Запишіть результати.

8. Проаналізуйте результати. Впевніться, що з надійністю 95% середнє арифметичне для VAR1 (воно дорівнює 17,2) знаходиться в інтервалі від 10,0 до 24,4, а середнє арифметичне для VAR2 (воно дорівнює 9,2) – в інтервалі від 6,4 до 11,9.

9. Інтервальні оцінки мають ту перевагу, що вони є наочними. Ви можете за допомогою опції **Box and Whisker plot** ("ящик з вусами") зобразити положення середніх, ширину розмаху за похибкою середнього арифметичного і довірчим інтервалом. Особливо це інформативно при порівнянні двох або декількох статистичних рядів.

10. Для одержання графіка "ящик з вусами" вам потрібно повернутися у вікно **Descriptive Statistics** і через клавішу **Variables** позначити як активні обидва ряди: VAR1 і VAR2.

Потім натисніть клавішу **Box and Whisker plot**. Відкриється нове підменю, в якому потрібно ЛКМ позначити опцію

Mean/SE/1,96*SE (Середнє, Стандартна похибка, Довірчий інтервал)

OK

Відкриється вікно з графіком (рис. 18), в якому наочно видимі взаємні положення аналізованих статистичних параметрів. Розгляньте

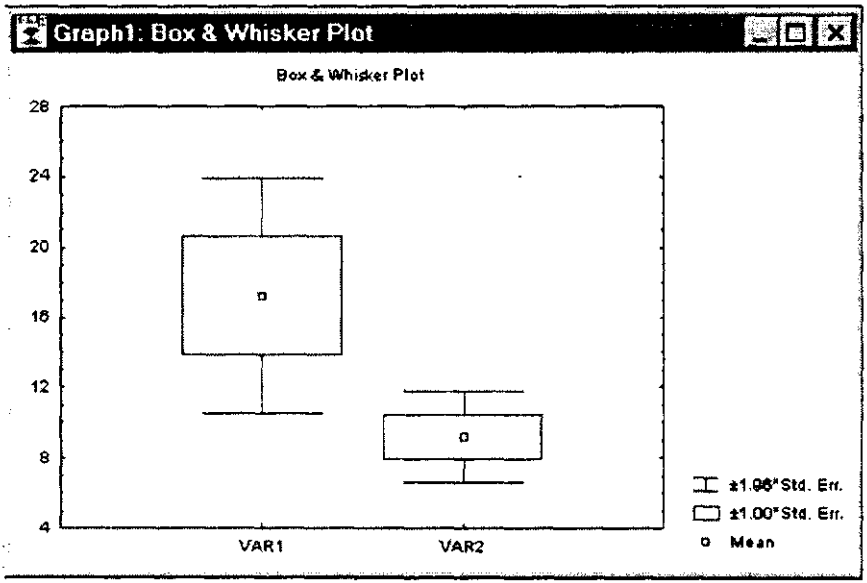


Рис. 18. "Ящик з вусами" для даних з практичної роботи № 6

і проаналізуйте їх. Зверніть увагу, що VAR1 явно менш надійний, ніж VAR2. Розкид у першому випадку ширший. Видно, що ряди самостійні, але все ж таки найменші значення одного ряду накладються на найбільші другого ряду, тобто їх довірчі інтервали перекриваються.

Натиснувши клавішу **Continue**,повернiться у вікно Descriptive Statistics. Знову активізуйте клавішу **Box and Whisker plot** і ознайомтесь з іншими типами "ящиків з вусами".

Такі графіки зручно зображати як ілюстрацію ваших даних при виконанні курсових і дипломних робіт, а також журнальних і книжкових публікацій.

11. Закрийте всі робочі вікна пакета Statistica знайомою вам клавішею-піктограмою "хрестик".

12. По закінченні роботи вийдіть з пакета Statistica і правильно виключіть комп'ютер.

ПРАКТИЧНА РОБОТА № 7

Мета роботи: Оволодіти навичками перетворення статистичних рядів для наближення їх властивостей до нормального статистичного розподілу.

1. Увімкнути комп'ютер, загрузити систему Windows і пакет Statistica.

2. За допомогою відповідних процедур завантажити будь-який файл даних, створений вами на попередніх заняттях. Можна просто увести будь-який статистичний ряд. Наприклад:

1-2-3-4-5-6-7-8-9-10-11

3. За допомогою піктограми **Vars** відкрийте випадаюче меню і знайдіть у ньому рядок

Current Spec... (Поточні специфікації)

Відзначте його ЛКМ й активізуйте.

4. В підмодулі, що відкрився, у нижньому вікні з ім'ям **Long Name** (Довгі імена) можна записувати формули, за допомогою яких виконуються перетворення будь-якого із статистичних рядів (стовпчиків) даних. Такі формули завжди починаються зі знаку рівняння і можуть містити математичні функції, перелік яких можна переглянути, активізувавши клавішу **Functions**.

5. Виконайте перетворення квадратним коренем. Для цього, використавши курсор миші, впишіть в це вікно таку команду (без пропусків!):

= sqrt (var1)

Вона означає, що для першого стовпчика з ім'ям **var1** ви даєте команду перетворення існуючих значень методом добутку квадратного кореня з кожного рядка цього стовпчика.

Натисніть **OK**. Відкриється вікно, в якому повинно бути записане

Expression OK. Recalculate the variable now? (Команда вами записана правильно. Тепер виконати перетворення?).

Натисніть клавішу **"Так"**. В стовпчику негайно з'являться нові значення. Знаючи алгебру, ви зможете легко повернутись до попередніх значень цього стовпчика, якщо для **var1** дасте нову команду:

= (var1) * (var1)

Спробуйте зробити це.

6. Проведіть перетворення статистичного ряду **var2** шляхом здобуття десяткового логарифма із кожного значення цього ряду.

Для цього виконайте необхідні підготовчі процедури, які були раніше описані, і в командному вікні **Long Name** запишіть команду:

= log10 (var2)

Виконайте подальші процедури і переконайтесь, що ряд дійсно перетворений.

Після цього поверніть статистичний ряд (коли це необхідно) до вихідного виду. Для цього слід записати команду:

= exp ((var2)*log(10))

7. Подібно до цього здійснюється перетворення методом добутку натуральних логарифмів. Для цього в командному вікні **Long Name** уводиться команда:

= log (var1)

Зворотне перетворення в цьому випадку виконується шляхом запису команди:

= exp ((var1)*log(2.71828))

В цьому виразі фігурує величина 2.71828, оскільки основа натуральних логарифмів е дорівнює 2.71828.

8. В стовпчиках з даними після перетворення ви, можливо, захочете побачити більше чи менше значущих цифр після коми. Така зміна здійснюється за допомогою двох піктограм з основного їх ряду: **increase decimal** або **decrease decimal**. Знайдіть їх.

Після цього зробіть активним потрібний вам стовпець даних, клацнувши по його заголовку. Далі ви одним клацанням по відповідній піктограмі збільшуєте чи зменшуєте кількість видимих знаків після коми.

Майте на увазі, що таке подання чисел має сенс тільки для їх зорового огляду або для того, щоб правильно роздрукувати таблицю з даними. Насправді ЕОМ пам'ятає до 24 знаків після коми проводить обчислення з усіма.

ПРАКТИЧНА РОБОТА № 8

Мета роботи: Навчитись утворювати з двох (або декількох) статистичних рядів один похідний. Зокрема, така задача зустрічається при вивченні темпів росту рослин чи тварин та в багатьох інших випадках.

1. Створіть базу даних і збережіть її у вигляді файлу. Пропонуємо такий приклад: враховувалась висота рослин соняшника в два строки з інтервалом у 10 днів. Були одержані такі дані:

Перший строк: 7.5-6.8-4.5-5.9-7.1-7.3-6.5-6.4-5.7-8.0

Другий строк: 11.5-9.8-6.6-7.3-11.5-11.0-10.9-8.6-8.9-12.6

2. У вас перший строк матиме ім'я VAR1, а другий - VAR2. Тепер вам необхідно створити нову колонку, в якій повинні розміститись дані про добові прирости кожної з 10 рослин. Виконайте такі процедури за допомогою ЛКМ:

Активізуйте клавішу-піктограму **VARS**

Активізуйте опцію **Add** (Додати)

3. У вікні, що відкрилось (рис.19), заповніть:

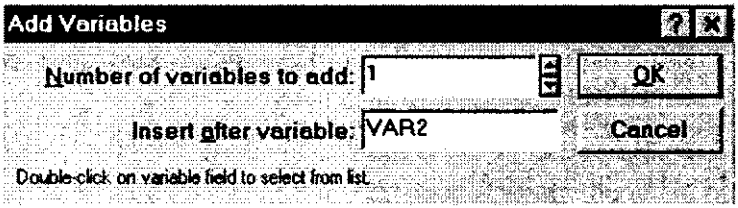


Рис. 19. Вікно, що дозволяє збільшити кількість стовпців в електронній таблиці

Number of variables to add (Кількість доданих змінних)

Insert after variable (Вставити після змінної)

В першому вікні ви поставите цифру 1, а в другому - VAR2. Натисніть **OK**. Новий стовпець в таблиці з'явився. Він повинен іти слідом за стовпчиком VAR2. Якщо його не видно, то за допомогою ЛКМ відсуньте праву межу вікна праворуч.

4. Нова колонка одержує ім'я автоматично і до нього завжди додається префікс **New**. Таке ім'я не завжди зручне, але ви можете будь-якому стовпцю вашої таблиці дати будь-яке ім'я будь-яким шрифтом. Зробіть це, виконавши такі процедури:

Клацніть ЛКМ по заголовку нового стовпця. Це його активізує. Виконайте команди:

VARS

Current Specs...

У вікні **Name** забийте курсором ЛКМ старе ім'я та наберіть кирилицею нове ім'я. Наприклад, "ПРИРІСТ" (рис.20).

OK

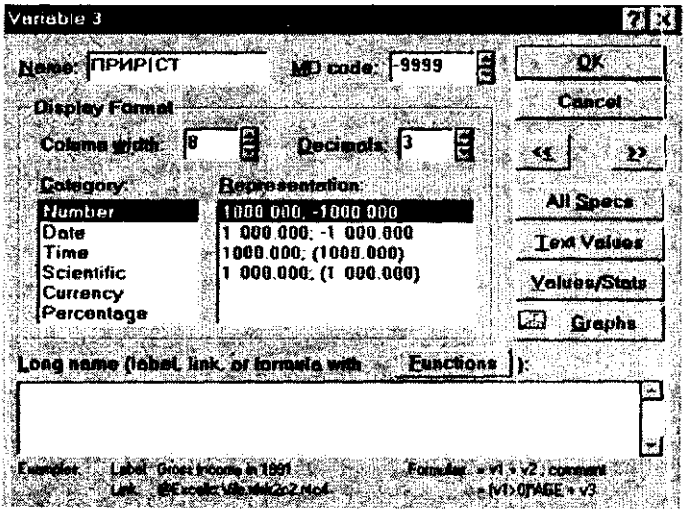


Рис. 20. Вікно поточних специфікацій, що дозволяє змінювати назви стовпців в електронній таблиці

Видно, що колонка придбала нове зручне змістовне ім'я. Так можна перейменувати по черзі всі стовпці. Для підготовки таблиць до роздрукування та при збереженні файлів це дуже зручно.

5. Тепер необхідно заповнити колонку ПРИРІСТ обчисленими значеннями. Зрозуміло, що вам потрібно відняти від висоти кожної рослини другого строку обліку висоту в перший строк і розділити цю різницю на 10, оскільки між двома обліками проминуло 10 днів. Виконайте такі процедури:

Клацніть ЛКМ по заголовку стовпця ПРИРІСТ для його активізації і далі виконайте команди:

VARS

Current Spec...

До вікна *Long name* уведіть команду (рис. 21):

=((var2)–(var1))/10 (без пробілів!)

OK

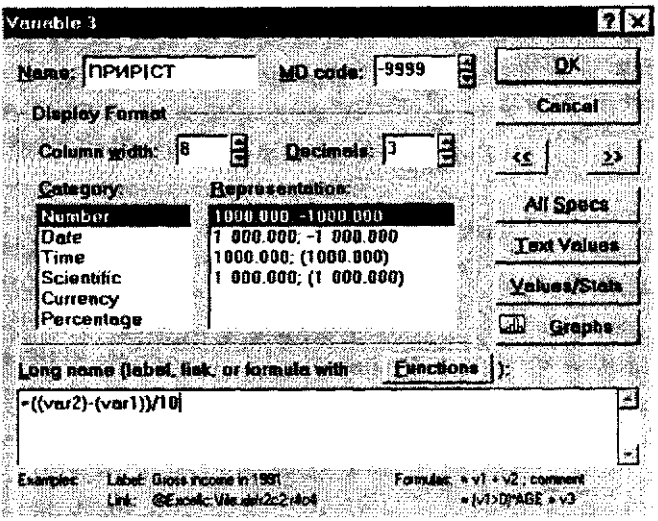


Рис. 21. Вікно поточних специфікацій, що дозволяє виконати різні типи розрахунків з вихідними даними

Коли синтаксис команди вірний, то комп'ютер відзначить цей факт і запитає: чи виконувати розрахунок? Після підтвердження правильності запису натисканням на клавішу **OK**, ви побачите, що комп'ютер записав до колонки ПРИРІСТ потрібні вам значення. Задачу розв'язано.

Statistica має широкий набір команд для виконання різних розрахунків з даними. У вікні **Current Spec... - Long name** ви знайдете клавішу **Functions**. Після її активізації відкривається вікно категорій команд і в правому вікні після кожної категорії – десятки можливих команд. Їх можна не переписуючи вставляти в рядок **Long name** клавішею **Insert**. Впевніться у цій можливості.

Якщо ви сумніваєтесь у правильності запису, чи комп'ютер не приймає вашої команди (звичайно в цьому випадку не відбувається ніяких дій), натисніть клавішу **Syntax**. Відкриється вікно допомоги, в якому можна бачити, як правильно записувати основні команди. Впевніться і в такій можливості.

При виконанні цієї роботи вас може зацікавити середній добовий приріст соняшнику. Ви вже вмієте це робити. Для колонки ПРИРІСТ можна отримати всі необхідні точкові та інтервальні оцінки. Для закріплення практичних навичок зробіть ці обчислення самостійно.

ПРАКТИЧНА РОБОТА № 9

Мета роботи: Виявлення "вискакуючих" значень, що відхиляють вибірку від нормального статистичного розподілу, та їх видалення.

Бажано, аби опрацьовувані статистичні ряди відповідали нормальному статистичному розподілу. В деяких випадках тільки окремі 1-3 облікові дані "вискакують" із загального ряду, погіршуючи рівень його відповідності нормальному статистичному розподілу. Такі дані найчастіше доцільно видаляти, бо вони можуть бути пов'язані з помилками обліку.

Існують алгебраїчні способи вибракування вискакуючих даних, але вони громіздкі і потребують багато часу. Statistica надає можливість робити це швидко та ефективно за допомогою інструменту **ПЕНЗЛИК (Brushing Tool)**. Слід брати до уваги, що цей інструмент допомагає вам визначити і видалити не найменші чи найбільші значення вибірки, а ті, що якнайбільше віддаляють її від нормального статистичного розподілу.

1. Активізуйте один із файлів з вашими даними за допомогою випадаючого меню

File

Open data

Якщо готового файлу даних немає, то створіть його, використовуючи матеріали, одержані при обліку надоїв корів (кг/добу):

13-21-14-22-17-19-20-17

2. Для оволодіння інструментом **Brushing Tool** у вашому файлі з даними активізуйте ЛКМ той стовпець, з яким ви хочете попрацювати.

Тепер відкрийте модуль **Basic Statistics**, а в ньому – підмодуль **Descriptive Statistics**.

3. В меню **Descriptive Statistics** у відповідному вікні вкажіть ім'я обраної для аналізу змінної (стовпця) і натисніть клавішу **Normal probability plots**. Ви побачите, як ваш ряд виглядає на нормальному імовірнісному аркуші. Ті точки (кружальця), які лежать найдалі від прямої лінії, підлягають відбракуванню і видаленню в першу чергу. Знайдіть одну таку точку.

Після цього можна буде працювати з інструментом **Brushing Tool**. Для цього виконайте такі операції:

1. Знайдіть у верхньому меню піктограму **Brushing Tool** – вона має вигляд лупи з прицільною рамкою (не плутайте її з іншою піктограмою – лупа з хрестом всередині, яка слугує для збільшення розміру графіка). Клацніть ЛКМ по цій піктограмі. Ви побачите, що курсор змінив форму, а праворуч з'явилось додаткове вікно **Brushing** (рис. 22).

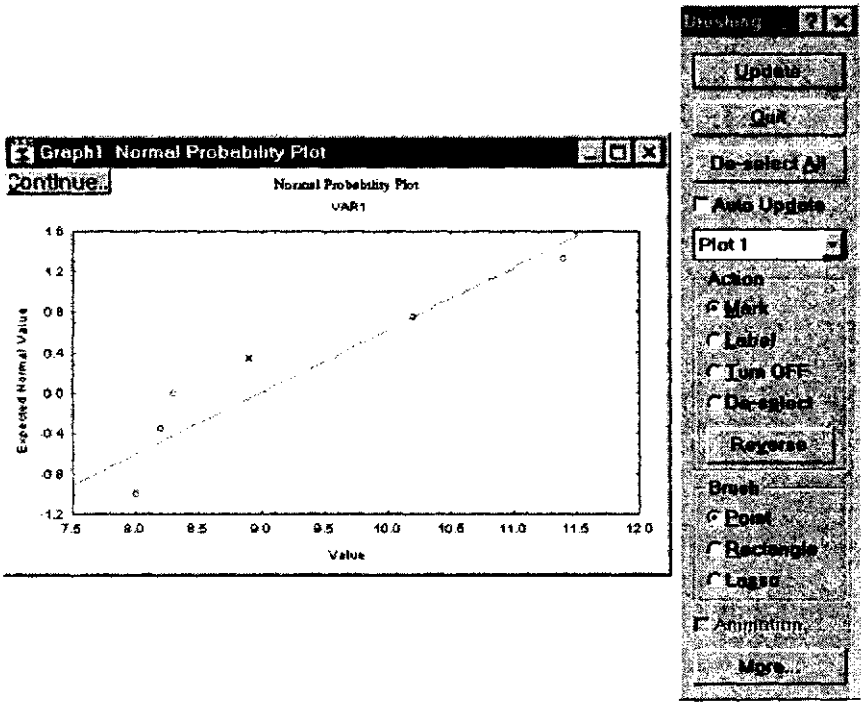


Рис. 22. Інструмент "Пензлик" у вікні Brushing Tool

2. В додатковому вікні **Brushing** поставте крапку проти рядка **Label**.

3. Тепер курсором у вигляді прицільної лупи позначте ту точку, яка є найбільш віддаленою від прямої лінії на графіку. Клацніть ЛКМ. Точка змінить форму: звичайно замість кружальця з'явиться хрестик.

4. Поверніться у вікно **Brushing** і курсором натисніть клавішу **Update** (поновити). На графіку поруч з точкою з'явиться її ім'я і номер у стовпці ваших даних **(Case №...)**. Запишіть цей номер.

5. Подивіться координати цієї точки на графіку. Для цього в рядку основного меню знайдіть **Layouts**. Клацніть по ньому ЛКМ. У випадаючому меню знайдіть рядок

Edit data

Клацніть по ньому ЛКМ. З'явиться ще одне вікно (ви помітили, що Statistica – це багатовіконний пакет, вікна можна відкривати, закривати, активізувати. Це дуже зручно). У новому вікні координати вискакуючої точки позначені червоним. Ви знайшли вашу підозрілу точку і знаєте її номер і координати.

6. Тепер поверніться до вікна з вашими даними. Для цього можна закрити всі вже не потрібні вам вікна, а можна просто активізувати вікно з базою даних звичайним клацанням ЛКМ по ньому (навіть якщо на екрані виглядає бодай найменший клаптик цього вікна!).

Окресливши позначену точку курсором, ви можете її видалити засобами редагування:

Edit

Cut

Але перш ніж видаляти з бази даних будь-який рядок, добре поміркуйте. Це ваше рішення, а не комп'ютера. Необмірковане видалення всіх вискакуючи точок наближає ряд до нормальності, а вам по суті необхідно інше – щоб його оцінки (точкові та інтервальні) максимально відповідали генеральній сукупності. Це різні поняття. Звичайно, 1-2 чи 3 точки можна видалити і подивитись, що вийде.

В багатьох випадках більш правильно для знаходження точного рішення піти в поле чи на ферму і зібрати більше даних, збільшивши обсяг вибірки, або взагалі зробити нову вибірку, дотримуючись усіх правил її здобуття.

Роботу закінчено. Закрийте всі вікна і правильно вийдіть із системи.

ПРАКТИЧНА РОБОТА № 10

Мета роботи: Навчитись встановлювати різні довірчі рівні для точкових та інтервальних оцінок.

1. Увійдіть в систему Windows і відкрийте пакет Statistica.

2. Активізуйте один із файлів з вашими даними за допомогою випадаючого меню:

File

Open data

Якщо готового файлу даних немає, то створіть його, використовуючи матеріали, одержані при обліку надоїв корів (кг/добу):

13-21-14-22-17-19-20-17

3. Завантажте модуль **Basic Statistics**, а ньому підмодуль **Descriptive Statistics**.

4. Клавішею **Variables** уведіть для опрацювання одну із змінних (стовпців) з вашого файла даних. Проти клавіші повинен з'явитись відповідний запис її імені.

5. Відкрийте вікно **More statistics...** і в ньому позначте галочкою **Mean** (середнє арифметичне) та **95% confidence limits for mean** (95% довірчий рівень). Всі інші віконця повинні бути пусті – зайва інформація не потрібна. **OK**

6. Натисніть послідовно ще раз **OK**. Відкрилось вікно з результатами, випишіть з нього значення середнього і той інтервал, в якому воно міститься з імовірністю (надійністю!) 95%. Випишіть ці дані.

7. Повторіть всю цю роботу (пункт 5). Але тепер установіть 90%-ий довірчий рівень у вікні **Descriptive Statistics** у полі **Statistics** в рядку *Interval* для того же стовпця з даними (рис. 23). **OK**. Випишіть одержаний результат.

8. Ще раз повторіть всю цю роботу (пункт 5), але вже для 99%-ого довірчого рівня. Занотуйте результат.

9. Дайте змістовну оцінку вашим результатам, порівнюючи зони довіри для трьох оцінок. Зазначте, що чим більше значення довірчого інтервалу, тим ширша зона, в якій лежить середнє арифметичне. Це об'єктивне правило математичної статистики.

Зважте, що зона довіри повинна бути основою для прийняття будь-яких рішень в галузі сільського господарства. На матеріалі своїх досліджень в кожному випадку вам необхідно визначити, якими зонами довіри ви будете користуватись у тих чи інших випадках.

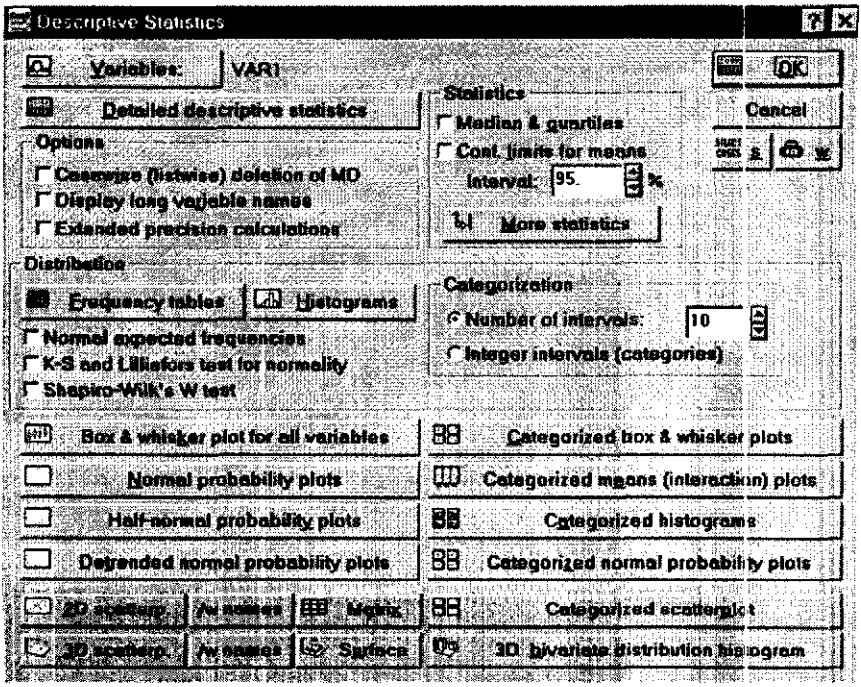


Рис. 23. Використання вікна Descriptive Statistics для встановлення довірчого інтервалу

Широкі зони довіри, звичайно, знецінюють ваші технологічні рішення. Майте на увазі, що звузити зони довіри та, відповідно, одержати більш надійні результати, можна лише одним способом: **більш пильно і грамотно проводити всі польові досліди і спостереження, аби дисперсія та похибка середнього були якнайменші.** Тоді і зони довіри повужчають!

ПРАКТИЧНА РОБОТА № 11

Мета роботи: Навчитись знаходити показники мінливості для різних ознак сільськогосподарських об'єктів та робити висновки про усталеність чи нестабільність цих ознак.

Для виконання роботи використайте дані обліку врожайності двох культур: картоплі та цукрового буряку в Сумській області за 11-річний період (з 1978 по 1988 р.) і зробіть висновок про ступінь адаптованості районованих сортів цих культур до погодно-кліматичних умов Сумської області.

Виконайте такі операції:

1. Увійдіть в систему Windows та відкрийте пакет Statistica.

2. Створіть базу даних так, щоб врожайність картоплі була занесена в один стовпець (VAR1), а цукрового буряку в інший (VAR2).

3. Ця урожайність по роках становила (в ц/га):

картопля: 147-152-70-86-72-110-120-102-119-70-119;

цукровий буряк: 278-247-231-194-186-247-247-225-226-251-245.

Збережіть цей файл під оригінальним ім'ям.

4. Послідовно використовуючи модуль **Basic Statistics** та **Descriptive Statistics**, знайдіть для цих даних основні статистичні оцінки. При цьому у вікні Statistics натисніть клавішу **More Statistics** і позначте "галочкою" відповідні опції (рис. 24):

Mean (Середнє арифметичне);

Minimum and maximum (Мінімум і максимум);

Range (Розмах);

Variance (Дисперсія);

Standard deviation (Стандартне відхилення).

5. Зручніше і швидше одержати ці оцінки для обох рядів одразу, зробивши відповідну позначку в списку змінних. Для цього потрібно при позначенні активної змінної тримати натиснутою клавішу **Ctrl** або клацнути ЛКМ по клавіші **Select All**.

Природно, що середній врожай за 11 років у цих культур був різним і становив:

картопля – середнє арифметичне 106,1, розмах 82, при min=70, max=152

цукровий буряк – середнє арифметичне 234,3, розмах 92, при min=186, max=278.

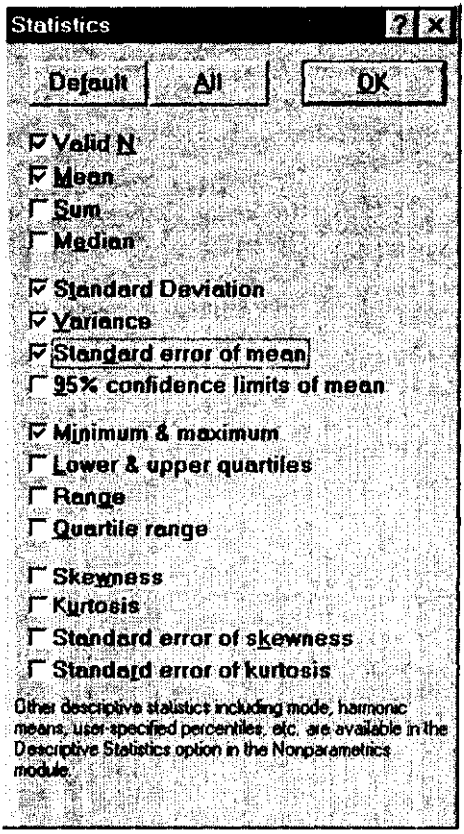


Рис. 24. Використання вікна підмодуля More Statistics для знаходження показників мінливості статистичних рядів

Перевірте свою роботу. Співпали ваші результати з цими цифрами?

6. За першим враженням розмах, а отже і нестійкість врожаїв, вищий у цукрового буряку. Але ж і абсолютні значення врожаю у цієї культури більші, а отже, така думка помилкова. Слід порівнювати інші показники мінливості – дисперсію та коефіцієнт варіації.

7. Дисперсії дорівнюють:

картопля – 851,1

цукровий буряк – 691,0

Доводиться робити інший, і вірний, висновок: розкид врожаїв по роках був вищий у картоплі.

8. Перевіримо цей висновок обчисленням коефіцієнтів варіації. Для цього використайте введений в систему калькулятор. Викликається він через Windows.

9. Порівняйте свої обрахунки з вірними:

Коефіцієнт варіації $_{картопля} = s / \bar{x} \cdot 100 = 29,17 / 106,4 \cdot 100 = 27,5\%$

Коеф. варіації $_{цукр.\ буряк} = s / \bar{x} \cdot 100 = 26,28 / 234,3 \cdot 100 = 11,2\%$

Таким чином, коефіцієнт варіації для картоплі більш як у 2 р. більший. А це значить, що районовані сорти картоплі менш стійкі стосовно коливань погодних умов по роках. Саме у цьому напрямку селекціонерам необхідно покращувати сорти картоплі.

ПРАКТИЧНА РОБОТА № 12

Мета роботи: Закріплення навичок обчислення мінливості ознак сільськогосподарських об'єктів та вміння правильно оцінювати фактичний матеріал за цими стратегічними параметрами.

У посіві ярої пшениці й озимої пшениці був зроблений облік кількості насіння у звичайного бур'яну зернових культур гірчака шорсткого (Polygonum Scabrum) з метою встановлення, в якому із посівів насіннєва продуктивність цього бур'яну була найбільш стійкою, а в якому – нестабільно-мінливою.

Були одержані такі результати:

Вихід насіння гірчака у посіві ярої пшениці (шт./особину) становив:

13.5-13.7-18.9-11.9-10.0-5.4-23.5-25.8-9.3-18.6-11.8-29.5-33.2-4.7-8.1-8.7-22.7-38.3-20.6-29.5-31.3-18.5-16.6-17.3-3.2.

Вихід насіння гірчака в посіві озимої пшениці (шт./особину) становив:

2.8-2.4-1.6-0.8-0-1.9-0-2.2-2.5-3.2-1.9-2.1-1.1-2.7-2.4-0.4-0.5-1.1-2.4-1.7-2.2-2.2-0.5-6.2-4.2.

1. Використовуючи статистичний пакет Statistica, створіть із цих матеріалів базу даних.

2. Знайдіть потрібні статистичні параметри і оцініть, в якому посіві мінливість насіннєвої продуктивності гірчака була вищою, та спробуйте дати пояснення причинам цього явища.

3. Для перевірки Вашої роботи звірте результати з наведеними нижче.

	Для посіву ярої пшениці:	Для посіву озимої пшениці:
Середнє	17,78	1,96
Розмах	35,1	6,2
Дисперсія	91,7	1,84
Станд. відхилення	9,58	1,36
Коеф. варіації	53,8%	68,87%

Причина більш високої мінливості насіннєвої продуктивності гірчака в посівах озимої пшениці досить очевидна. Структура пологу культурних рослин у посівах озимих, як правило, нерівномірна: через загибель частини рослин пшениці в період зимівлі в посіві з'являються прогалини, де гірчак розростається без перешкод і дає високий вихід насіння.

Слід звернути увагу і на інше. Середній вихід насіння гірчака в посівах озимої пшениці істотно нижчий, ніж у посівах ярої. Вочевидь, полог озимої пшениці більш ефективно пригнічує цей бур'ян.

Задачі для самостійного розв'язання

Задача № 12.1. В польовому досліді враховували довжину колосу (см) у двох сортів озимої пшениці: Охтирчанка і Миронівська 808. Були одержані такі результати:

	Охтирчанка	Миронівська 808
1	6.0	6.0
2	6.0	6.2
3	7.0	6.5
4	5.5	6.5
5	6.0	5.5
6	6.0	6.0
7	7.0	5.5
8	5.5	6.5
9	6.0	5.5
10	6.0	6.0
11	7.0	6.0
12	5.0	5.5
13	6.0	4.2
14	4.0	4.0
15	6.0	6.5
16	4.0	6.1
17	5.0	5.5
18	6.0	4.5
19	6.0	6.0
20	4.0	6.0
21	7.0	5.7

Визначте, у якого з цих сортів варіювання довжини колосу більше.

Задача № 12.2. Були виміряні (довжина тіла в см) по 7 особин звичайного та каліфорнійського дощового черв'яка, які використовуються в господарстві для компостування. Визначте, у якої форми черв'яка довжина тіла варіює більше. Вихідні дані:

	Звичайний черв'як	Каліфорнійський черв'як
1	10.2	12.2
2	8.2	10.6
3	8.9	9.9
4	8.0	13.0
5	8.3	8.1
6	8.0	10.8
7	11.4	11.5

ПРАКТИЧНА РОБОТА № 13

Мета роботи: Оволодіти навичками створення кругових діаграм, стовпчастих діаграм і графіків для ілюстрації журнальних і книжкових публікацій, курсових та дипломних робіт, використовуючи пакет "Статистика".

Кругові діаграми

Для створення кругової діаграми перш за все необхідно створити невелику базу даних з того матеріалу, який ви хочете ілюструвати. Так, в господарстві структура посівів становила:

зернові – 816 га, 68% посівних площ,
цукровий буряк – 288 га, 24%,
картопля – 60 га, 5%,
багаторічні трави – 36 га, 3%.

1. Увійдіть в середовище пакета "Статистика" і створіть файл з даними. Всі дані в гектарах слід ввести в одну перемінну (VAR2), відсотки вводити не треба, за вашим бажанням комп'ютер вирахує їх сам. Кожному рядку дайте відповідну назву. Це робиться через клавішу-піктограму

Cases (випадки)

верхнього меню, де в свою чергу необхідно активізувати опцію **Names** (імена).

2. У вікні, що відкрилось, проти відповідних номерів уведіть назви культур. База даних може виглядати приблизно так:

NUMERIC VALUES	1 VAR1	2 VAR2
Зернові		816.000
Цукр. буряк		288.000
Картопля		60.000
Багат. трави		36.000

3. В основному меню виберіть послідовно

Graph

Stats. 2D Graph,

а потім в меню, що відкрилось, активізуйте опцію

Pie Charts (кругові діаграми).

4. У з'явившомуся підмодулі перш за все введіть ім'я змінної **(Variables)**, до якої у вас є кількісні дані.

У вікні **Graph Type** необхідно активізувати опцію **Pie Chart-Values** (кругові діаграми для кількісних даних).

У вікні **Pie legend** позначте опцію **Values**. В цьому випадку біля відповідних секторів стоятимуть вихідні дані (гектари). Якщо ви виберете опцію **Percent** (проценти), то комп'ютер їх вирахує і розмістить проти секторів. Якщо позначити опцію **Off**, то ніяких написів біля секторів не буде. Поекспериментуйте з цими опціями.

5. У вікні **Cases** необхідно виставити номери рядків, для яких будується діаграма. В нашому випадку це рядки 2-5. Впишіть ці номери у відповідні позиції.

OK

Ви одержали кругову діаграму (рис. 25).

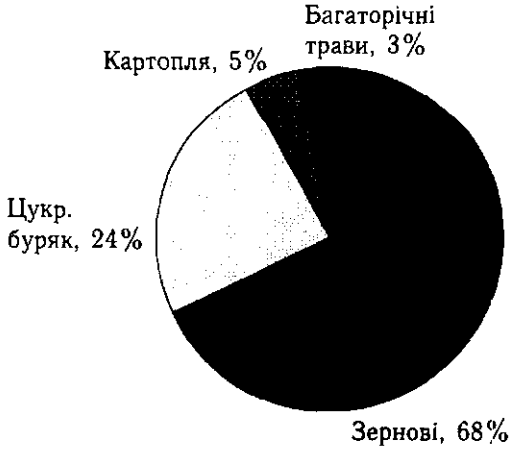


Рис. 25. Приклад кругової діаграми. Показана структура посівних площ

6. Кругову діаграму необхідно забезпечити потрібними підписами. Для цього перш за все видаліть непотрібні вам верхній та нижній підписи. Клацніть ПКМ (правою клавішею миші) по верхньому підпису. У меню, що з'явилось, оберіть **Delete Text** і клацніть по цих словах ЛКМ. Підпис щез.

Тепер клацніть ПКМ по нижньому підпису. З'явиться меню, в якому ви клацнете ЛКМ по **Delete Text** або **Icon Legend Off**. І цей підпис буде стертий.

7. "Статистика" зазвичай розміщує кругову діаграму біля лівого краю аркуша паперу (хоча в цілому це залежить від вихідної настройки пакета і може бути й по-іншому). Якщо ви хочете розмістити її приблизно по центру, то зробіть таке: у верхньому меню натисніть

View (Вид)

Graph Mapping Option... (Вибір положення графіків)

Потім у вікні **Graph Area** (площа графіка) проти поля **Right** поставте замість числа 1000 – число 550.

OK

Розміщення графіка на сторінці змінилось.

8. Непогано створити вільні поля навколо діаграми (якщо вони вузькі). Для цього застосовується піктограма з верхнього ряду **Adjust Margins, Plot Area** (точні поля, площа графіка) - на ній зображена хвиляста лінія.

Активізуйте її. За допомогою повзунків і стрілок ЛКМ навчіться змінювати поля і розміщення діаграми на аркуші паперу. Зробіть достатньо широке поле згори і збоків.

9. Тепер можна зробити підписи. Для цього служить піктограма, на якій над літерами **"AB"** намальований олівець. Вона називається **Graphic text Editor**. Клацніть по ній ЛКМ. З'явилось вікно.

10. В його основному полі ви можете набрати будь-яким шрифтом який завгодно текст. Шрифт вибирається після натискання клавіші **Font** (шрифт). Слід мати на увазі, що не всі шрифти з їх списку підтримують кирилицю. Необхідно підібрати такий, що має цю властивість і задовольняє вас своїм написанням.

Спочатку наберіть слова "Рис. 1. Структура посівних площ"

OK

11. Ви повернулись у поле графіка. Тепер знайдіть місце, куди ви хочете помістити ці слова, і на його початку поставте курсор і клацніть ЛКМ. Ваш текст зайняв необхідну позицію. До речі, тепер його можна пересувати, якщо це місце вам не сподобалось.

Для цього клацніть по ньому ЛКМ. Текст обведе рамка і чорні квадратики. Притиснутою лівою клавішою миші ви можете його вільно переміщати. Цим же методом при необхідності можна на графіку поруч з секторами надписати назви культур.

12. Подібним чином зробіть бокові підписи. Для того, щоб вони займали вертикальну позицію у вікні **Orientation**, позначте опцію **Vertical Left** (вертикально ліворуч од графика) **Vertical right** (вертикально праворуч од графіка)

OK

і повторіть процедуру з курсором.

13. Рисунок готовий. Він може виглядати так, як на рис. 25. Ви можете його надрукувати.

Стовпчасті діаграми

Для створення стовпчастої діаграми можна використати той самий файл даних.

1. Після активізації вікна з цими даними виконайте команди:

Graph

Stats. 2D Graph,

а потім оберіть опцію **Bar/Column Plots** (стовпчасті діаграми).

2. Спочатку введіть назву змінної, в якій знаходяться вихідні дані. Потім заповніть поля підмодулю так, щоб у полі **Graph Type** стояло **Regular**, в полі **Control limits** стояло **Off**, в полі **Cases** проставте номери випадків, для яких будується діаграма (в цьому прикладі це 2-5), в полі **Orientation** поставте **Vertical**.

3. Активізуйте клавішу **Option**. В полі **Displey** (показати) поставте точку проти опції **Cases names** (імена випадків).

OK

OK

4. Ви одержали стовпчасту діаграму. Всі підписи до неї можна прибрати, виправити чи додати тим же чином, як і при побудові кругових діаграм. Ваша діаграма може виглядати приблизно так, як це зображено на рис. 26.

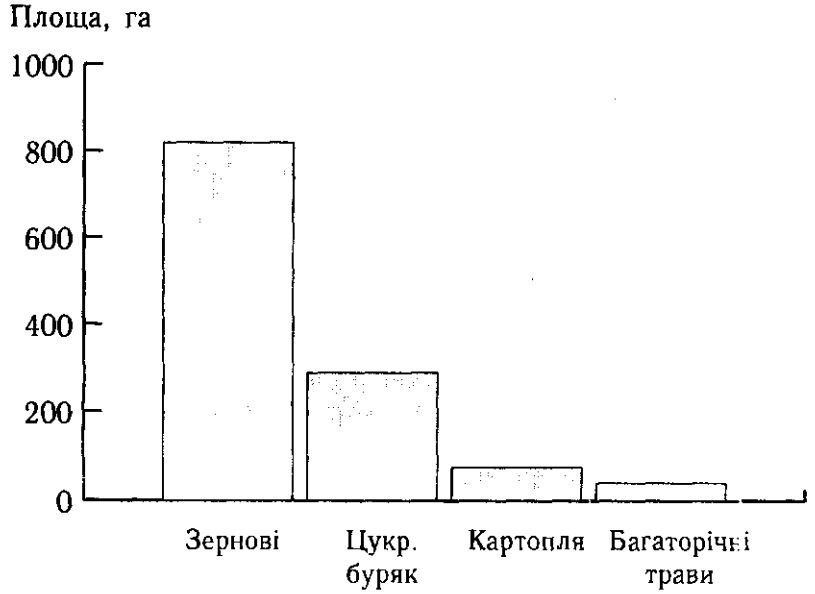


Рис. 26. Приклад стовпчикової діаграми. Показана структура посівних площ

Двовимірні графіки

Пакет "Статистика" дає можливість будувати найрізноманітніші типи графіків за зразками, що є в наявності, а також створювати графіки користувача. Розглянемо побудову найпростішого лінійного графіка.

1. Для цього створіть базу даних із таких даних про динаміку врожайності озимої пшениці в одному з господарств за 1990-1995 роки:

1990 – 42 ц/га
1991 – 36 ц/га
1992 – 28 ц/га
1993 – 21 ц/га
1994 – 29 ц/га
1995 – 26 ц/га

Роки занесіть в один стовпець, врожаї – в інший.

2. Виконайте команди:

Graph
Stats. 2D Graph,

а потім у меню, що відкрилось, активізуйте опцію *Lines plot (Variables)* (лінійний графік).

3. У вікні **Variables** слід позначити обидві змінні. В полі **Graph Type** активізуйте опцію **XY Trace** (XY лінія). В полі **Fit** активізуйте опцію **Off**. В полі **Cases** перевірте правильність розставлення номерів випадків.

OK

4. Ви одержали графік динаміки урожайності по роках.

5. Не виключно, що комп'ютер самовільно зробить роки дробовими числами. Це легко виправити. Клацніть ЛК миші по рядку років на графіку і оберіть опцію **Edit scale Values** (редагувати масштаб) і в полі **Scaling** виставьте у відповідних вікнах **Mode-Manual** (спосіб - ручний) та **Min, Max** відповідно 1990 і 1995. Якщо виставити числа 1989 і 1996, графік виглядатиме краще.

6. Якщо клацнути ПК миші по центру графіка і обрати опцію **Changes Plot Layout** (змінити настройку графіка), то за допомогою відповідних опцій можна змінювати товщину ліній, форму та розмір кружечків на графіку, підбирати кольори та робити багато інших удосконалень у графіку.

7. Всі інші написи видаляються, додаються чи виправляються так само, як і на кругових діаграмах. Графік може виглядати приблизно так, як на рис. 27.

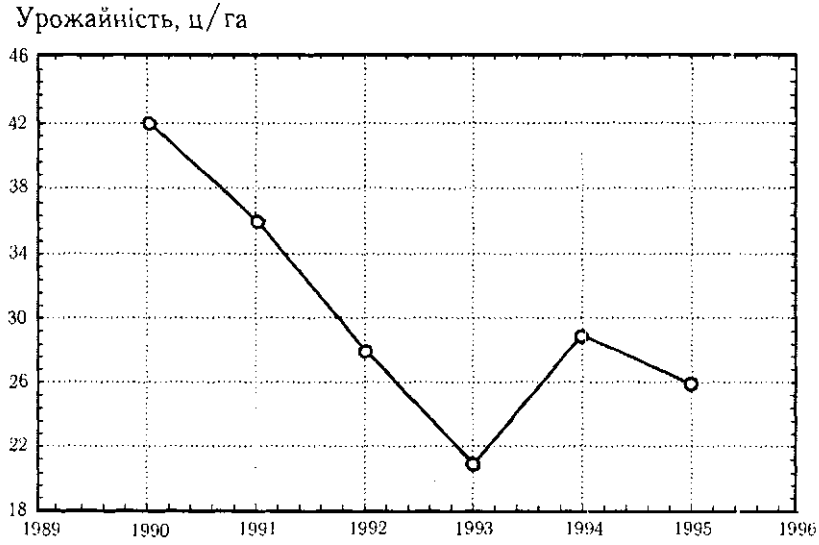


Рис. 27. Приклад двовимірного графіка (2D). Показана динаміка врожайності озимої пшениці по роках

Тривимірні графіки

Цінною властивістю пакета є можливість побудови тривимірних графіків з поверхнею відгуку, що дозволяє наочно визначити, при якому сполученні двох провідних агрономічних факторів буде оптимальним відгук.

Для цього скористаємось даними про вихід насіння люцерни (VAR5) в залежності від ваги особин (VAR1), висоти (VAR2).

	VAR1	VAR2	VAR3	VAR4	VAR5
1	15.62	97	198	16	.960
2	15.20	119	115	14	1.490
3	16.53	117	97	9	2.100
4	7.19	85	74	9	.098
5	11.42	83	135	12	.066
6	2.56	72	23	1	.040
7	17.02	102	12	16	.590
8	4.88	108	42	12	.104
9	8.47	91	107	19	.028
10	3.99	79	54	10	.076

1. Після створення бази даних виконайте команди:

Graph

Stats 3D XYZ Graphs
Surface Plot

2. У вікні підмодуля, що відкрилось, перш за все заповніть вікно **Variables**, визначивши як X – VAR1 (вага рослин), Y – VAR2 (висота рослин), Z – VAR5 (вихід насіння). У вікні **Fit** (підгонка) – **Least Squares** (метод найменших квадратів).

OK

Одержали тривимірний графік. Користуючись вже описаними прийомами, можна забезпечити графік підписами та змінити його параметри. Графік може мати вигляд, як на рис. 28. Видно, що найбільший вихід насіння дають крупні рослини, хоча є вторинні піки, що становлять предмет окремого агрономічного аналізу.

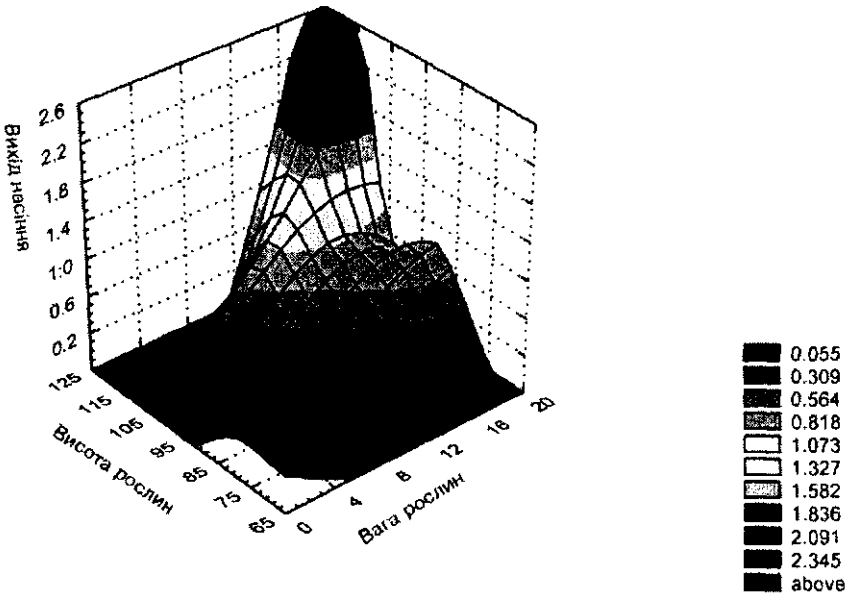


Рис. 28. Приклад тривимірного графіка (3D). Показаний розмір виходу насіння люцерни в залежності від висоти рослин і ваги особин

ПРАКТИЧНА РОБОТА № 14

Мета роботи: Вивчення методів проведення повного кореляційного аналізу.

Для проведення кореляційного аналізу можна використати дані про урожайність картоплі та цукрового буряку за 11-річний період (в ц/га):

	VAR1	VAR2
1	147	278
2	152	247
3	70	231
4	86	194
5	72	186
6	110	247
7	120	247
8	102	225
9	119	226
10	70	251
11	119	245

Необхідно визначити, чи є кореляція в величині врожаю цих двох культур.

1. Увійдіть в середовище пакета "Статистика" і створіть файл з даними про врожайність картоплі і цукрового буряку. Збережіть його під оригінальним іменем.

2. Виконайте такі процедури:

Analysis
Other statistics
Customize list
Basic statistics
Replace
Switch to

Відкрився підмодуль "Основних статистик" (рис. 29). Оберіть **Correlation matrices** (Кореляційні матриці)
OK

3. За допомогою клавіші **One variable list** позначте (рис. 30) для аналізу обидві змінні VAR1 і VAR2.

4. Натисніть клавішу **Correlations**. Відкриється вікно з матрицею коефіцієнтів кореляції (рис.31). В матриці, звичайно, є не цікаві кореляції кожної змінної самої з собою. Такі кореляції завжди

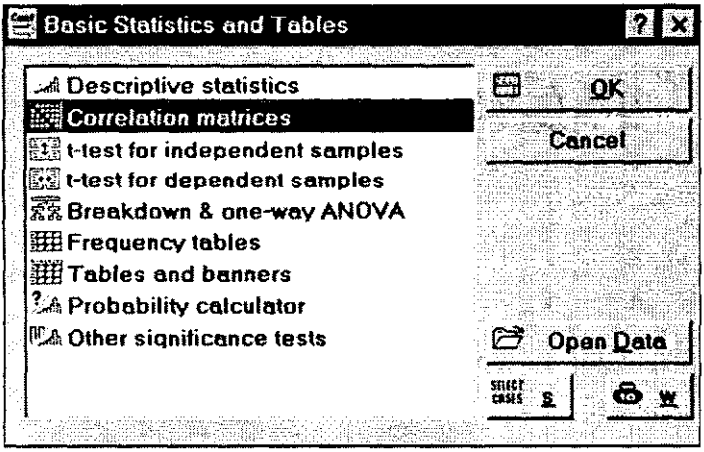


Рис. 29. Вікно підмодуля Basis Statistics and Tables, що використовується при кореляційному аналізі

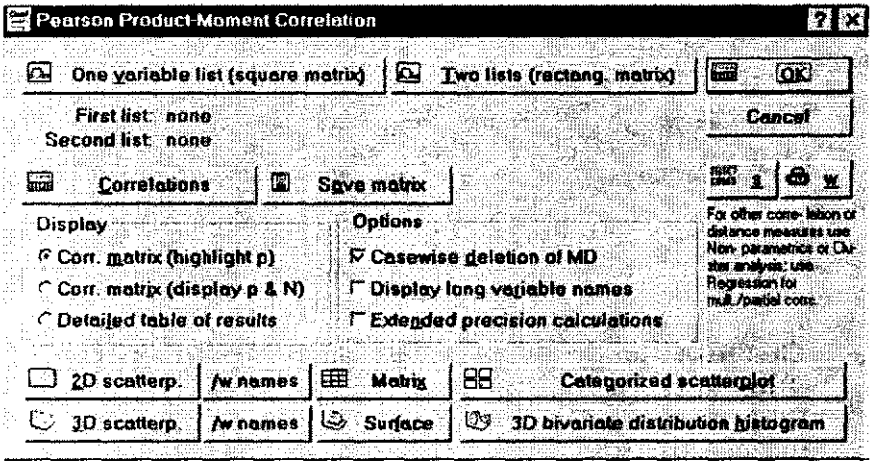


Рис. 30. Вікно для визначення параметрів кореляційного аналізу

дорівнюють 1.0. Потрібну кореляцію VAR1 з VAR2 можна знайти на перетині відповідних стовпців і рядків. В даному випадку коефіцієнт кореляції дорівнює +0,60. Це число виділене червоним – значить при довірчому рівні <0,05 (див. верхній рядок в заголовку вікна видачі результату) цей коефіцієнт статистично достовірний. Такий рівень відповідає достовірності в 95%.

Correlations (corn_met23.sta)

BASIC STATS | Marked correlations are significant at p < .05000
N=11 (Casewise deletion of missing data)

Variable	VAR1	VAR2
VAR1	1.00	.60
VAR2	.60	1.00

Рис. 31. Вид матриці коефіцієнтів кореляції

5. Ви можете поглибити аналіз результату шляхом його графічного подання.

Для цього натисніть клавішу **Continue** і у вікні **Pearson Product-Moment Correlation** натисніть клавішу **2D Scatterplot**. У вікні, що відкрилось, позначте справа одну змінну (VAR1), а зліва – іншу (VAR2). Натисніть **OK**.

6. Ви отримали графік (рис. 32) для залежності між врожаями двох культур із зоною довіри (вона виділена червоним пунктиром). Видно, що точки досить сильно розкидані і частина їх знаходиться взагалі поза зоною довіри.

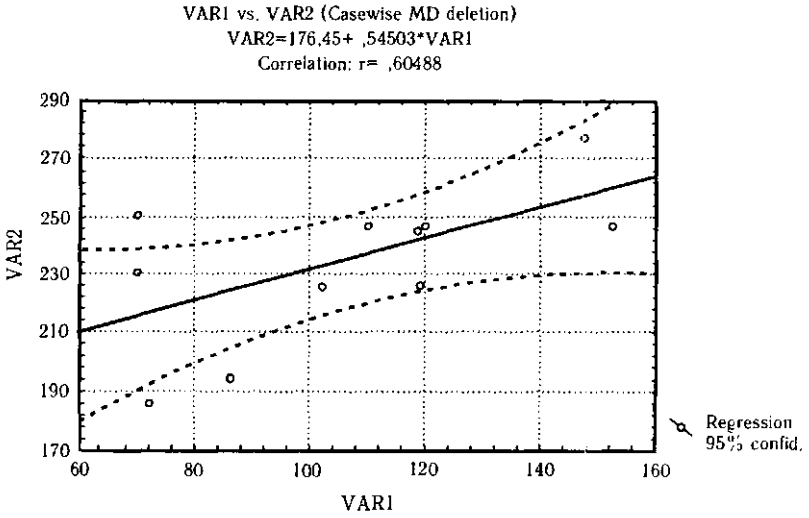


Рис. 32. Приклад графічного подання результатів кореляційного аналізу

7. Можна відзначити точки поза зоною довіри за допомогою інструмента "Пензлик" і провести змістовний аналіз, визначивши, до яких конкретно років вони належать. Зазначте, що в даному випадку видаляти такі точки з бази даних не можна – це конкретна урожайність культур по Сумській області за той чи інший рік. Але визначивши приналежність точки до конкретного року, можна спробувати визначити причини такого надзвичайно низького чи високого врожаю.

Так, для точки № 10, яка відповідає 1987 року (70 на 251), видно, що врожай картоплі був різко знижений, а цукрового буряку – досить високим. Проаналізувавши метеорологічні дані за 1987 рік, можна помітити, що за погодними умовами він був несприятливий для картоплі.

ПРАКТИЧНА РОБОТА № 15

Мета роботи: Закріплення навичок проведення повного кореляційного аналізу.

У картоплі сорту Молодіжний в 1997 році були враховані три селекційно важливі ознаки і одержані такі дані:

Листова поверхня, $см^2$ (VAR1)	Кількість пагонів у кущі, шт. (VAR2)	Вага бульби, г (VAR3)
199,5	1	209,7
1987,7	6	184,4
3971,9	8	409,8
4175,0	3	385,2
6216,2	4	463,8
3141,9	4	199,3
4784,2	6	311,5
4752,1	6	341,2
3597,9	4	238,6
1168,4	4	193,5

1. Створіть базу комп'ютерних даних із цього матеріалу.

2. Проведіть повний кореляційний аналіз цих даних і зробіть висновок про статистично значущі і незначущі кореляції. Продивіться і проаналізуйте графіки для цих кореляцій.

3. Звірте одержані результати з правильними відповідями. Коефіцієнти кореляції становили:

VAR1 з VAR2 = +0,43 Статистично недостовірно

VAR1 з VAR3 = +0,81 Статистично достовірно при p=0,05

VAR2 з VAR3 = +0,30 Статистично недостовірно.

Правомірно зробити висновок, що вага бульби картоплі контролюється потужністю розвитку листової поверхні куща, але не залежить від кількості пагонів у кущі. Очевидно, ця закономірність повинна враховуватись при селекційній роботі з картоплею, а при вирощуванні цієї культури слід забезпечувати якнайкращий розвиток листової поверхні картоплі відповідними технологічними прийомами.

Задачі для самостійного розв'язання

Задача № 15.1. В польовому досліді ураховували кількість зернин в колоссі ячменю (шт.) сорту Роланд і сорту Пироговський. Були одержані такі дані:

	VAR1	VAR2
1	34.0	19.0
2	44.2	44.0
3	33.0	31.0
4	47.0	45.0
5	55.0	51.0
6	41.0	34.0
7	48.0	30.0
8	36.0	31.0
9	48.0	15.0
10	34.0	37.0
11	47.0	34.0
12	35.0	30.0
13	29.0	45.0
14	43.0	42.0
15	5.0	30.0
16	37.0	33.0
17	39.0	35.0
18	55.0	33.0
19	42.0	25.0
20	52.0	33.0
21	40.0	34.0

Визначте, чи є кореляція в озерненості колосся у цих двох видів ячменю.

Задача № 15.2. В польовому досліді на протязі 10 років вивчали врожайність двох сортів ярої пшениці Дніпрянка і Ленінградка. Визначте, чи є кореляція між врожаями цих сортів за цей період.

Вихідні дані:

	VAR1	VAR2
1	28.0	25.7
2	21.3	12.2
3	30.0	15.0
4	15.0	18.0
5	22.0	16.7
6	18.1	17.1
7	16.6	17.0
8	18.0	16.0
9	22.1	19.2
10	22.0	20.0

ПРАКТИЧНА РОБОТА № 16

Мета роботи: Навчитись порівнювати між собою дві вибірки (метод парних порівнянь) за критерієм Ст'юдента (t-тест).

В посіві ураховували суху вагу особин люцерни (W) на двох полях: а) з нормою висіву 6 кг/га та б) на полі з нормою висіву 14 кг/га. Були одержані такі дані:

Норма висіву 6 кг/га:

21.2-37.7-1.3-8.1-21.5-1.9-13.3-11.2-1.0-43.9-3.5-33.3-22.9-28.6-8.0-13.7-47.0-5.3-3.5.

Норма висіву 14 кг/га:

15.6-15.1-8.1-16.5-7.2-11.4-2.6-17.0-4.5-1.8-4.9-7.9-1.2-8.5-4.0-11.9-14.1-19.2-12.5.

Визначте, чи є достовірна статистична різниця між вагою рослин у цих двох варіантах досліду. Нульова гіпотеза H_0 полягає в тому, що такої різниці немає, тобто $H_0: x_1 = x_2$.

1. Для проведення аналізу необхідно виконати такі операції:

Analysis
Other statistics
Customize list

Вибрати зі списку:

Basic statistics
Replace
Switch to...

2. У підменю, що з'явилось, вибрати тип аналізу: **t-test for independent samples** (Тест для незалежних вибірок), бо наші два посіви явно один від одного не залежать.

3. Послідовно заповніть у новому вікні (рис. 33) такі поля:

Input file (вхідний файл), для чого стрілочкою у правій частині рядка розкрийте список і виберіть у ньому опцію:

Each variable contains the data for one group (кожна змінна містить дані для однієї з порівнюваних груп).

Перевірте, чи там правильно внесено: у VAR1 записані дані для одного варіанту досліду, а у VAR2 – для іншого!

4. Заповніть тепер поле **Variables**, клацнувши по ньому ЛКМ. У відкрившомуся вікні лівої частини позначте ЛКМ VAR1, у правій – VAR2. Правильність позначки можна перевірити в нижній частині вікна – які номери стоять проти **First** (перша змінна) та **Second** (друга змінна).

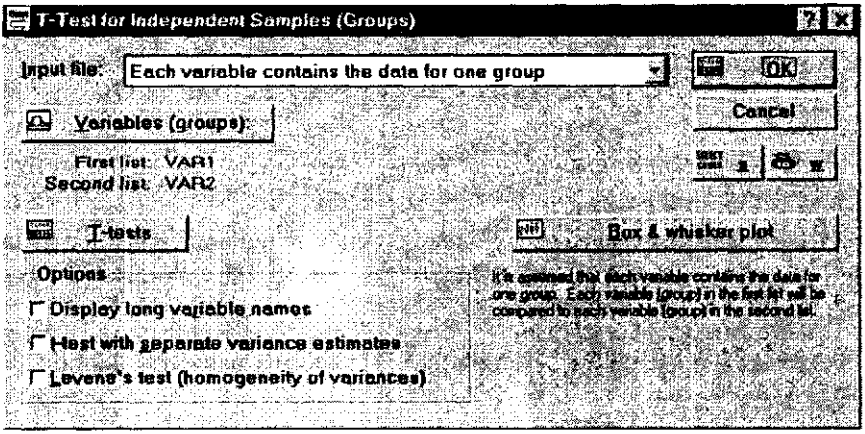


Рис. 33. Вікно, що використовується при оцінці вибірок за допомогою критерію Ст'юдента (t-test)

OK

5. Виберіть **t-test** та натисніть на його клавішу. У вікні з'являться результати розрахунків. Випишіть їх:

Середня арифметична для першої вибірки – 17,20

Середня арифметична для другої вибірки – 9,68

Величина критерію Ст'юдента t = 2,06

Ступінь свободи – 36

Довірчий рівень p = 0,046

Таким чином, на користь того, що знайдені середні не відрізняються одна від одної (наша нульова гіпотеза) всього 0,046 шансу з 1,0. Це менше, ніж 0,05, отже вибірки видрізняються одна від одної статистично достовірно.

6. Розглянемо також співвідношення дисперсії для двох вибірок F. Воно дорівнює в нашому випадку 6,97 при p = 0,00014. Таким чином, на користь того, що дисперсії дорівнюють всього 0,00014 з 1,0. Отже дисперсії розрізняються як і середні. Цю обставину комп'ютер враховував, коли обирав формули для обчислення t.

7. Для наочності порівняння подивіться "ящик з вусами". Для цього

Continue

Box a.whisker plot

8. У вікні вибору позначте

Mean/SE/1.96 SE,

тобто точка на графіку – це середнє арифметичне, "ящик" – межі стандартних похибок, а "вуса" – межі довірчих рівнів при імовірності 95%.

9. Зробіть змістовний підсумок про переваги тієї чи іншої норми висіву насіння люцерни для одержання крупніших рослин.

* * *

Закріпити навички використання методу парних порівнянь ви можете на ще одному прикладі.

Враховувалась висота рослин люцерни (норма висіву 6 кг/га) у два різних строки: 13 червня і 14 липня. Визначте, чи є статистично достовірне збільшення у висоті рослин люцерни за цей термін. Нульова гіпотеза $H_0: x_1 = x_2$.

1. Створіть базу даних, використовуючи наведені нижче дані.

Висота рослин люцерни у перший строк (см):

21.0-14.0-17.5-19.5-12.8-15.0-19.8-5.5-11.5-14.5-9.0-15.1-11.5-23.0--14.5-18.5-10.5-10.0-10.5-8.0

Висота рослин у другий строк (см):

68.0-66.0-60.1-59.1-71.5-52.6-45.9-36.0-71.0-49.5-54.6-52.0-46.4--62.0-37.2-36.0-35.4-57.0-56.0-43.4

2. Для виконання роботи всі процедури аналізу аналогічні раніш описаним. Але слід враховувати, що вибірки – це той самий посів, і висота рослин у другий строк біологічно обумовлена ступенем розвитку рослин в перший строк спостереження.

Тому в меню у відповідному місці слід вибрати опцію:

t-test for dependent samples

3. Випишіть результати. Вони дорівнюють: середні арифметичні відповідно 14,08 і 52,98, критерій Ст'юдента – 15,4, а p = 0,0000. Останнє значить, що на користь нульової гіпотези про відсутність різниці у висоті рослин не має шансів зовсім. Отже, у місячний строк відбулося статистично достовірне (достовірність 100%) збільшення висоти рослин в посіві.

4. Подивіться "ящик з вусами" і роздрукуйте його.

* * *

Розв'яжіть самостійно таку задачу. В ТСХА було проведено польовий дослід із посівом соняшника восени і навесні. Через 15 днів після сівби на обох варіантах провели облік висоти рослин і отримали такі дані, вимірявши по п'ять рослин у кожному варіанті:

Осіння сівба:

14.5-16.0-15.0-14.0-15.5.

Весняна сівба:

21.0-14.0-16.5-19.5-19.0.

Порівняйте свої результати з правильними: середні арифметичні дорівнюють 15 і 18. Критерій Ст'юдента – 2,33, p = 0,0477. Вибірки достовірно відмінні, бо значення довірчого рівня менше, ніж 0,05. F = 12,2, p = 0,032 – достовірно відмінні й дисперсії.

ПРАКТИЧНА РОБОТА № 17

Мета роботи: Вивчення техніки виконання однофакторного дисперсійного аналізу і одержання навичок по його реалізації.

В польовому досліді вивчався вплив ширини міжрядь на кількість листків, що формується на рослинах люцерни. Дослід був закладений у трьох варіантах:

1 варіант – ширина міжрядь 15 см. При обліку (вибірка 10 рослин, тобто ніби 10 повторностей) на рослинах в цьому варіанті кількість листків становить:

24-31-22-37-45-9-116-23-130-23;

2 варіант – ширина міжрядь 45 см. При обліку (вибірка 10 рослин, тобто ніби 10 повторностей) на рослинах в цьому варіанті кількість листя становила:

125-395-109-73-248-51-62-64-83-79;

3 варіант – ширина міжрядь 60 см. При обліку (вибірка 10 рослин, тобто ніби 10 повторностей) на рослинах в цьому варіанту кількість листя становила:

23-30-28-39-40-19-106-20-132-25.

1. Використовуючи ці дані, проведіть однофакторний дисперсійний аналіз з метою визначення наявності статистично достовірного впливу ширини міжрядь на кількість листя люцерни.

2. Відкрийте пакет "Статистика" і створіть файл даних для проведення дисперсійного аналізу. Слід мати на увазі, що для дисперсійного аналізу доводиться вводити дані особливим чином. Першу змінну VAR1 призначають для запису кодів (номерів варіантів) з присвоєнням варіанту 1 коду 1, варіанту 2 – коду 2, варіанту 3 – коду – 3. В VAR2 ми будемо вводити дані обліку кількості листя (повторення) точно по варіантах. Тому база даних виглядатиме приблизно так:

1	24
1	31
1	22
...	...
2	125
2	395
2	109
...	...
3	23
3	30
3	28
...	...

Таким чином, вам знадобиться таблиця, що має два стовпці і 30 рядків (таблиця 3x10). Створіть файл і збережіть його під оригінальним ім'ям.

3. Починайте процедуру дисперсійного аналізу. Для цього послідовно виконайте операції:

Analysis
Other Statistics...
Customize list

Виберіть зі списку **ANOVA/MANOVA** (Однофакторний/багатофакторний дисперсійний аналіз).

Replace
Switch to...

4. У вікні, що відкрилося, модуля "Дисперсійний аналіз" (рис. 34) перш за все слід активізувати клавішу **Variables** і ввести змінні. В ліве поле для **Independent variables** (незалежні змінні) необхідно ввести VAR1, бо вона в нас містить коди варіантів, а це і є незалежний параметр. В праве вікно **Dependent variables** введіть VAR2 – це залежна змінна, бо вона містить дані обліків, які залежать від характеру варіанту, тобто ширини міжряддя.

Перевірте правильність обраних змінних для аналізу по нижніх полях. Якщо все вірно,

OK

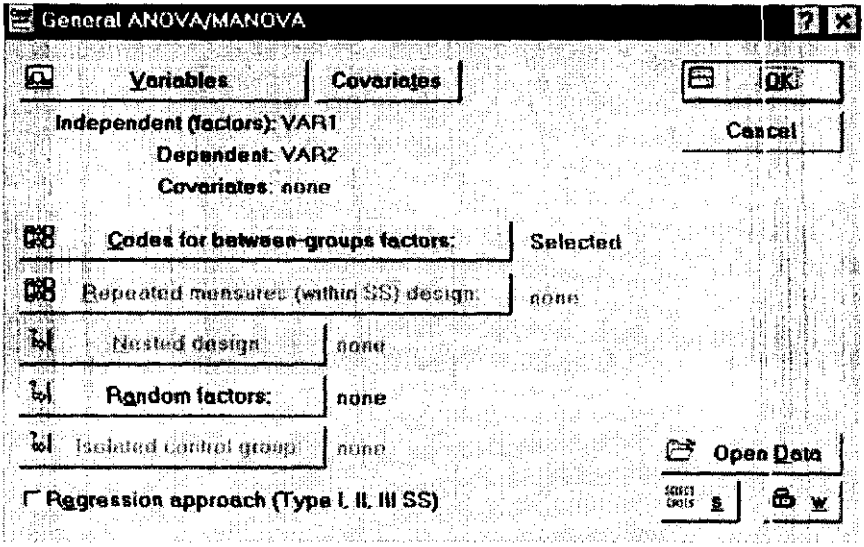


Рис. 34. Вікно, що використовується при проведенні дисперсійного аналізу

5. Тепер активізуйте клавішу **Codes for...** (коди для...). Ми маємо намір провести повний дисперсійний аналіз для всіх трьох варіантів, тому натисніть клавішу **All** (все). З'явиться запис 1-3, тобто комп'ютер буде проводити аналіз 1,2-го й 3-го варіантів.

Інші поля заповнювати не потрібно.

OK

OK

6. З'явилось вікно із записом результатів дисперсійного аналізу. У верхній його частині записано, що проведений 1-факторний дисперсійний аналіз **(1-way ANOVA)**. Наведені також номери кодів, варіантів, включених до аналізу.

7. Для виведення результатів треба натиснути клавішу

All effects (всі ефекти)

Відкрилося вікно з результатами аналізу (рис. 35). В ньому прийняті такі позначення:

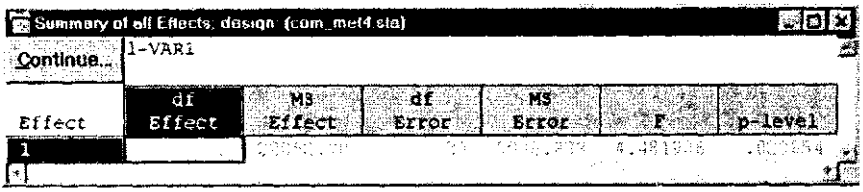


Рис. 35. Вікно з результатами однофакторного дисперсійного аналізу

Df Effect (Ступінь свободи для діючого фактора)
MS Effect (Середні квадрати для діючого фактора)
Df Error (Ступені свободи для похибки)
MS Error (Середні квадрати для похибки)
F (Критерій Фішера)
p-level (Довірчий рівень, тобто шанси на користь нульової гіпотези про те, що різниці між варіантами немає і, отже, фактор, що вивчається, не впливає на стан досліджуваного об'єкта (параметра)).

Запишіть результати дисперсійного аналізу в стандартній формі:

Джерело варіювання	Ступінь свободи для діючого фактора	Середні квадрати для діючого фактора	Ступінь свободи для похибки	Середні квадрати для похибки	Критерій Фішера	Довірчий рівень
Ширина міжрядя	2	22852,9	27	5098,833	4,48	0,02 (2%)

В такій формі результат часто пропонується в дипломних роботах і наукових публікаціях.

8. У випадку, що аналізується, шанси на користь нульової гіпотези, що кількість листків на рослинах люцерни за будь-якої ширини міжрядь однакова, всього 0,02 з 1,0, тобто 2%. Таким чином, 98% проти цієї гіпотези, тобто за те, що різниця в кількості листя є, і фактор ширини міжряддя статистично значущо вплинув на кількість листків на рослинах люцерни.

9. Тепер слід зайнятись порівнянням варіантів як таких. Це поглиблення дисперсійного аналізу. Активізуйте клавішу

Continue

у вікні, що відкрилось, натисніть клавішу

Post hoc comparisons (Пост-хок порівняння)

У вікні, що відкрилось, слід позначити курсором ЛКМ VAR1, бо в ній містяться коди наших варіантів.

OK

З'явиться вікно (рис. 36), в якому представлений список критеріїв, які можна використати для парного порізняння варіантів.

10. Тепер натисніть клавішу

Means (Середні арифметичні)

Вони такі:

1 варіант (15 см) – 46,0 листків на рослині;
2 варіант (45 см) – 128,9;
3 варіант (60 см) – 46,2.

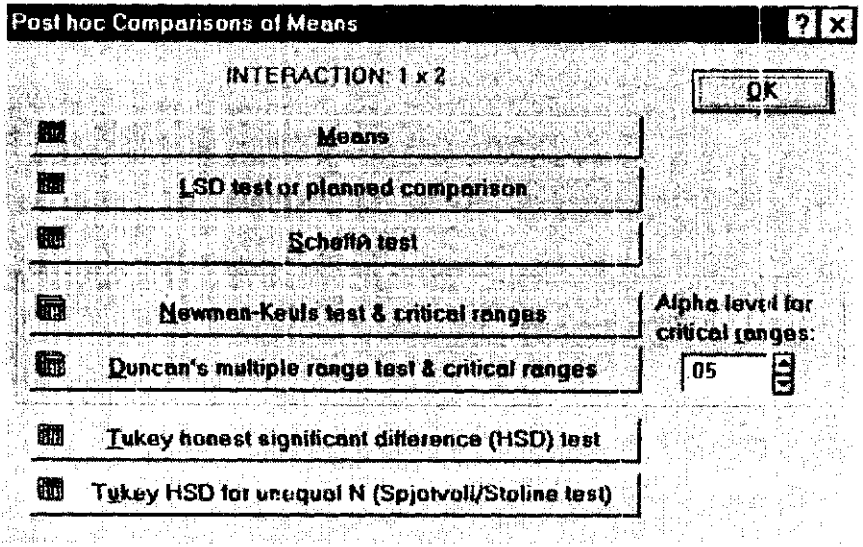


Рис. 36. Вікно пост-хок порівнянь (Post-hoc Comparision)

Тут слід поміркувати над агрономічною суттю результатів. Але спочатку потрібно визначити статистичну достовірність різниці між цими середніми. Пакет "Статистика" пропонує для цього широкий асортимент методів.

11. В типових випадках це робиться так:

Continue

Натисніть клавішу **LSD**. Ви отримаєте інформацію у вигляді матриці імовірностей достовірності різниці між варіантами.

Варіант 1 з варіантом 2 - p = 0,0150, нульову гіпотезу про те, що ці варіанти однакові за результатом, доводиться відкинути. На її користь дуже мало шансів.

Варіант 1 з варіантом 3 - p = 0,995. На користь нульової гіпотези 99,5%. Вона вочевидь вірна, і за кількістю листків на рослинах ці два варіанти не відрізняються.

Варіант 2 з варіантом 3 - p = 0,0152. Тут нульова гіпотеза відкидається – кількість листків на люцерні в цих двох варіантах статистично достовірно відрізняється.

12. Для порівняння варіантів є достатньо широкий набір критеріїв.

Continue

Серед них найбільш жорсткий критерій – критерій Шеффе, який слід використовувати в найбільш відповідальних випадках. Натисніть клавішу **Scheffe-test**. В матриці результатів цього разу інші цифри:

Варіант 1 з варіантом 2 – p = 0,049

Варіант 1 з варіантом 3 – p = 0,999

Варіант 2 з варіантом 3 - p = 0,05003

В цьому випадку для варіанту 1 з 2 різниця залишається достовірною (навіть цифри червоні), а от для варіантів 2 з 3 різниця в кількості листків на рослинах на порозі 95%-достовірністю і трохи нижче цього порогу. Цифри не виділені червоним. Прийняття цієї різниці у відповідальних дослідах дуже сумнівне.

13. Пакет "Статистка" дозволяє оцінити різницю варіантів в абсолютних значеннях. Для цього служить тест Дункана **(Duncan's multiple range...)**. Перед його використанням дорахуєм вручну різниці середніх між варіантами:

- першого з другим – 128,9-46,0 = 82,9;
- другого з третім – 128,9-46,2 = 82,7.

Тепер застосуємо критерій Дункана:

Continue

Duncan's multiple range...

Видно, що на рівні 95% різниця варіантів буде суттєвою, якщо вона більша 65,39 (1 step – м'який критерій на першому кроці). Наші різниці в обох випадках більші, отже статистично достовірні.

14. Випишіть результати всіх пост-хок порівнянь і зробіть висновки. Вони можуть бути в цьому випадку такі:

а) ширина міжрядь в посівах люцерни статистично суттєво впливає на кількість листків на рослинах;

б) максимальну кількість листків люцерна мала при ширині міжрядь 45 см, і вона достовірно відрізнялась від варіантів з шириною міжрядь 15 і 60 см;

в) варіанти з міжряддями 15 і 60 см статистично достовірно між собою не відрізнялись і мали схожу й низьку кількість листків на рослинах. Той, хто ставив дослід і проводив додаткові спостереження, міг відзначити, що у варіанті 15 см рослини люцерни глушили і затіняли одна одну. У варіанті з міжряддям 60 см через рідке розташування рослин буйно розрослися бур'яни, які не вдалось пригнітити культивуванням. Тому тут зменшення кількості листків на рослинах було викликане забур'яненістю посіву.

В цілому висновки дисперсійного аналізу змістовні й однозначні. Їх можна використати для прийняття відповідних технологічних рішень.

ПРАКТИЧНА РОБОТА № 18

Мета роботи: Закріплення навичок проведення однофакторного дисперсійного аналізу. Поглиблення знань про інтерпретацію результатів і оформлення підсумків аналізу.

Було проведено польовий дослід з вирощуванням 7 сортів томата, в 4-х повторностях з метою визначення, чи залежить вміст вітаміну С від сорту томата, а якщо залежить, то у яких сортів він статистично достовірно вищий. Дані обліку кількості вітаміну С (в мг %) наведені нижче в таблиці:

Сорти	Повторності			
	1	2	3	4
1	1,41	0,95	1,0	0,93
2	1,17	1,10	0,84	1,01
3	1,38	1,38	0,91	1,36
4	0,62	0,48	0,43	0,62
5	0,74	0,41	0,41	0,43
6	0,76	0,59	0,74	0,46
7	0,64	1,02	1,04	0,98

1. Складіть базу даних, проведіть однофакторний дисперсійний аналіз і визначте залежність вмісту вітаміну С в томатах від сорту.

2. При складенні бази даних завжди слід продумати, як організувати кодування варіантів. В цьому прикладі все зроблено просто. Варіант у нас відповідає номеру сорту, і у кожного сорту аналіз проводився 4 рази. Тому база даних потребує тільки 2-х стовпців (перший для кодів, другий – для власне даних). Кількість рядків завжди дорівнює добутку кількості варіантів на кількість повторень, тобто 7x4 = 28. Виглядатиме комп'ютерна база даних так:

VAR1 (коди)	VAR2 (дані)
1	1,41
1	0,95
1	1,00
1	0,93
2	1,17
...	...

3. Проведіть дисперсійний аналіз і порівняйте варіанти. Визначте номери сортів, які достовірно відрізняються підвищеним

вмістом вітаміну С, використовуючи алгоритм, який було описано в роботі № 17.

4. Звірте свої результати з правильними, що наведені нижче:

Df Effect (Ступені свободи для діючого фактора) = 6;

MS Effect (Середні квадрати для діючого фактора) = 0,3466;

Df Error (Ступені свободи для похибки) = 21;

MS Error (Середні квадрати для похибки) = 0,03;

F (Критерій Фішера) = 11,205;

p-level (Довірчий рівень, тобто шанси на користь гіпотези про те, що різниці між варіантами немає) = 0,000013.

5. На користь нульової гіпотези про те, що між сортами немає різниці, шансів 0,000013 (з 1,0), і вона, звісно, відхиляється. Ваш висновок очевидний – сорти достовірно відрізняються за вмістом вітаміну С.

6. Тепер розглянемо середні арифметичні **(Means)**. Вони дорівнюють:

1 сорт = 1,07
2 сорт = 1,03
3 сорт = 1,26
4 сорт = 0,54
5 сорт = 0,50
6 сорт = 0,64
7 сорт = 0,92

Дані явно не однозначні. Зрозуміло, що сорт № 5 скоріш за все відрізняється від сорту № 3, але відмінності сортів 1 і 2 чи 4 і 5 не з'ясовані. Розв'язати цю проблему вам допоможуть критерії HIP, Шеффе і Дункана.

7. Почнемо з критерію HIP (LSD). Його результат такий:

LSD test; variable VAR2

Probabilities for Post Hoc Tests

MAIN EFFECT: VAR1

		{1}	{2}	{3}	{4}	{5}	{6}	{7}
		1.072500	1.030000	1.257500	.5375000	.4975000	.6375000	.9200000
1	{1}		.735941	.151723	.000316	.000146	.002143	.233677
2	{2}	.735941		.081587	.000715	.000331	.004764	.386436
3	{3}	.151723	.081587		.000010	.000005	.000062	.013004
4	{4}	.000316	.000715	.000010		.750905	.430348	.005735
5	{5}	.000146	.000331	.000005	.750905		.272977	.002715
6	{6}	.002143	.004764	.000062	.430348	.272977		.033748
7	{7}	.233677	.386436	.013004	.005735	.002715	.033748	

У верхньому рядку наведені середні арифметичні по сорту, а на перехрещенні сортів подані довірчі рівні в формі шансів на користь відсутності відмінності. Червоні цифри (вони відповідають статистично достовірним відмінностям на рівні 95%) в таблиці позначені підкресленням. Слід уважно проаналізувати ці дані. Так, наприклад, 1-й сорт не має за вмістом вітаміну С в плодах достовірних відмінностей від сортів № 2, 7 і 3, але від інших відрізняється достовірно.

8. Тепер звернемось до критерію Шеффе. Видно, що тут достовірність відмінностей менша, бо критерій більш суворий. Він допоможе вам виділити сорти, що відрізняються один від одного, з найбільшою надійністю:

Scheffe test; variable VAR2
Probabilities for Post Hoc Tests
MAIN EFFECT: VAR1

		{1}	{2}	{3}	{4}	{5}	{6}	{7}
		1.072500	1.030000	1.257500	.5375000	.4975000	.6375000	.9200000
1	{1}		.999959	.890481	.025167	.013616	.105220	.953593
2	{2}	.999959		.758725	.047265	.026136	.180378	.990989
3	{3}	.890481	.758725		.001351	.000710	.006710	.331898
4	{4}	.025167	.047265	.001351		.999971	.994586	.203146
5	{5}	.013616	.026136	.000710	.999971		.969191	.123948
6	{6}	.105220	.180378	.006710	.994586	.969191		.539641
7	{7}	.953593	.990989	.331898	.203146	.123948	.539641	

Так, наприклад, відмінності сорту № 7 від інших сортів виявились сумнівними, хоча за критерієм LSD вони були.

9. Критерій Дункана свідчить, що достовірні лише ті розбіжності будь-якої пари сортів, коли різниця в кількості вітаміну С перевищить 0,25-0,29 мг%. Підрахуйте ці різниці самостійно і зробіть необхідні співставлення.

* * *

За результатами проведеного дисперсійного аналізу необхідно скласти підсумкову таблицю. В роботі № 17 був наведений приклад складання підсумкової таблиці так, як це роблять в західноєвропейських країнах та США. У вітчизняній літературі такі таблиці складають більш повними. В цьому випадку робота виконується за два етапи.

Перший етап – заповнення підсумкової таблиці даними, які видала "Статистика". В дужках наведені назви граф в пакеті "Статистика", з яких вам слід брати дані для заповнення відповідної клітинки. Цю інформацію необхідно обов'язково мати в довідковому зошиті:

Джерело змін	Суми квадратів	Ступені свободи	Середні квадрати	Критерій Фішера	Довірчий рівень
Факторіальне (сорт томата)		6 (df Effect)	0,3466 (MS Effect)	11,21	0,00001
Випадкове		21 (df Error)	0,0309 (MS Error)		
Загальне		27 (це просто сума 6+21)			

Для заповнення рядків, що залишились пустими в стовпчику "Суми квадратів", необхідний додатковий розрахунок на звичайному калькуляторі. Він простий.

Сума квадратів факторіальна дорівнює добутку середнього квадрата на кількість ступенів свободи, тобто 6х0,3466 = 2,079.

Сума квадратів випадкова – відповідно добутку середнього квадрата і ступенів свободи в цьому рядку, тобто 21х0,0309 = =0,6489.

Сума квадратів загальна обчислюється як сума квадратів факторіальних і випадкових: 2,079 + 0,6489 = 2,7279.

Тепер таблиця набуває такого вигляду:

Джерело змін	Суми квадратів	Ступені свободи	Середні квадрати	Критерій Фішера	Довірчий рівень
Факторіальне (сорт томата)	2,079	6 (df Effect)	0,3466 (MS Effect)	11,21	0,00001
Випадкове	0,6489	21 (df Error)	0,0309 (MS Error)		
Загальне	2,7279	27 (це просто сума 6+21)			

В такій формі її звичайно наводять в курсових і дипломних роботах, дисертаціях і наукових публікаціях в Україні та Росії.

Під таблицею можна записати НІР. Її беруть з критерію Дункана. При м'якій оцінці НІР = 0,25 мг%, та якщо матеріал цього потребує, ви можете скористатись більш суворим критерієм (скажімо тоді, коли ви оцінюєте кількість вільної синильної кислоти – це сильна отрута – в плодах мигдалю), то НІР – 0,29 мг%.

Під таблицею у нас прийнято записувати і так звану силу впливу фактора. Її позначають грецькою буквою η. Обраховується вона за формулою:

η = [(Сума квад. Фактор.)/(Сума квадр. Загальна)] х 100%.

В нашому прикладі це буде:

η = [2,079/2,7279] х 100% = 76,2%.

Це значить, що на частку сорту як чинника, який визначає вміст вітаміну С в плодах томата, припадає 76,2% (з 100%). А на всі інші впливи (наприклад, добрива, опади і т.ін.) припадає тільки приблизно 24%, що залишились. Сорт, як свідчить цей дослід, дійсно стає найбільш важливим чинником для виходу вітаміну С, і якщо ви хочете мати найбільш вітамінізовані томати, то слід покладати надії не на добрива чи полив, а на правильний підбір сорту.

В підсумку повна таблиця виглядатиме так:

Дисперсійний аналіз вмісту вітаміну С в мг%
у семи сортів томата.

Джерело змін	Суми квадратів	Ступені свободи	Середні квадрати	Критерій Фішера	Довірчий рівень
Факторіальне (сорт томата)	2,079	6	0,3466	11,21	0,00001
Випадкове	0,6489	21	0,0309		
Загальне	2,7279	27			

НІР = 0,25 мг%, сила впливу фактора η = 76,2%.

* * *

Тепер, якщо ви виконали роботи №17 і 18, ви знаєте всі основні тонкощі однофакторного дисперсійного аналізу. Ви вмієте на підставі польових чи лабораторних даних скласти вихідну таблицю з даними, з'ясувати, чи впливає фактор, що вивчається, на цей об'єкт, і визначити, які з об'єктів статистично достовірно відрізняються один від одного, а які ні. Але одержану вами навичку слід закріпити. Для цього проведіть дисперсійний аналіз власних польових матеріалів.

ПРАКТИЧНА РОБОТА № 19

Мета роботи: Закріплення навичок проведення однофакторного дисперсійного аналізу.

В польовому досліді висівали кукурудзу в 5 різних термінів з інтервалом 10 днів з 3 травня по 11 червня та враховували ураженість рослин кукурудзяним метеликом (кількість гусені на 25 стебел). Визначіть на підставі облікових даних, чи залежало ураження рослин кукурудзяним метеликом від терміну посіву і який термін посіву ви вважаєте найбільш сприятливим, а який – найменш сприятливим?

Облікові дані такі:

1 термін: 3-4-5-6-6

2 термін: 30-8-16-11-25

3 термін: 11-14-4-2-3

4 термін: 2-13-4-4-2

5 термін: 0-2-9-6-0

1. Створіть комп'ютерну базу даних і збережіть її під оригінальною назвою.

2. Проведіть однофакторний дисперсійний аналіз у відповідності з алгоритмом робіт № 17 і 18.

3. Порівняйте отримані дані з правильними результатами.

Результати: критерій Фішера дорівнює 5,73 при p = 0.03.

Достовірно несприятливим є другий строк (середня ураженість 18,0), інші терміни посіву статистично достовірно між собою не різнились. Рослини як першого (дуже раннього) терміну, так і рослини 3-5 термінів посіву мали статистично достовірно однаковий рівень ураження.

Задачі для самостійного розв'язання

Задача № 19.1. В польовому досліді з цукровим буряком порівнювали ефективність мінеральних добрив та сидератів. У контроль добрива не вносили і сидерати не заорювали. Дослід проведено в 4-кратній повторності, але контрольних ділянок було 10. Вихідні дані (врожай коренеплодів в ц/га):

- контроль: 301, 320, 319, 306, 244, 333, 327, 325, 308, 292;
- мінеральні добрива $N_{40}P_{40}K_{40}$: *332, 346, 378, 354*;
- сидерити: 380, 369, 406, 382.

Задача 19.2. Проведений польовий дослід по передпосівній обробці насіння гороху стимулятором росту – гетероауксином в різних

дозах. Враховувалась польова схожість гороху у відсотках. Дослід був проведений в 5-кратній повторності. Методом дисперсійного аналізу визначте, чи корисна така обробка.

Вихідні дані:

Варіанти	Повторності				
	1	2	3	4	5
Контроль	92	90	88	87	89
Доза 1	98	94	93	89	95
Доза 2	96	90	91	92	90
Доза 3	97	95	91	90	94

Задача № 19.3. Враховували кількість бур'яну щириці загнутої (шт./кв.м) в посівах вівса в залежності від трьох видів весняного обробітку грунту. Проведіть повний дисперсійний аналіз і визначте, який обробіток доцільний для боротьби з цим бур'яном. Вихідні дані:

Варіанти	Повторності			
	1	2	3	4
Передпосівне культивування	438	442	319	380
Передпосівне культивування + післяпосівне культивування	538	422	377	315
Боронування по сходах	77	61	157	52

ПРАКТИЧНА РОБОТА № 20

Мета роботи: Оволодіння технікою парних порівнянь середніх арифметичних, коефіцієнтів кореляції та даних, зображених у відсотках, на основі критерію Ст'юдента.

Для реалізації парних порівнянь необхідно знати:

а) для середніх їх стандартні відхилення та обсяги вибірок;

б) для коефіцієнтів кореляції - обсяги вибірок;

в) для відсотків необхідно знати обсяги вибірок, а відсотки заздалегідь перевести у частки одиниці.

1. Відкрийте пакет Statistica, а в ньому **Basic Statistics @ Tables**.

2. З підменю виберіть команду (рис. 37) **Other significance tests** (Інші тести істотної різниці).

3. Послідовно виконайте три завдання.

Перше завдання

У вікні, що відкрилося (рис. 38), оберіть опцію "Порівняння двох середніх **(Difference between two means)**". Проведіть слівстав-

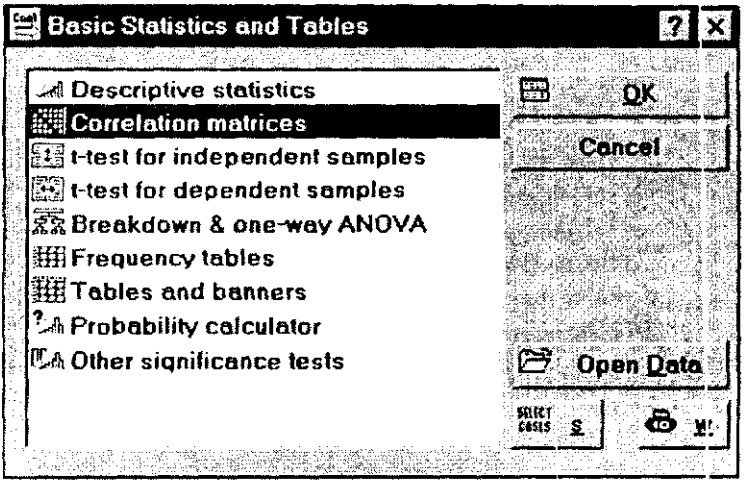


Рис. 37. Підмодуль, що дозволяє проводити оцінку достовірності відмінностей середніх арифметичних, коефіцієнтів кореляції і відсотків на основі критерію Ст'юдента

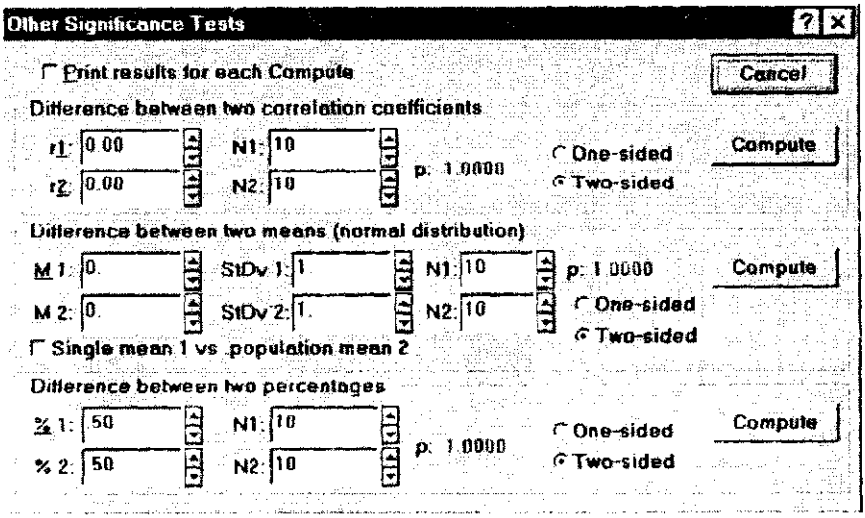


Рис. 38. Підмодуль для визначення статистично достовірної різниці вибірок за допомогою різних тестів

лення двох середніх висоти рослин озимої пшениці у фазі колосіння за такими даними:

1) середнє – 45,1, стандартне відхилення – 4,43, обсяг вибірки – 50;

2) середнє – 49,9, стандартне відхилення – 0,90, обсяг вибірки – 40.

Уведіть ці дані до відповідних вікон: середні – до вікон **M1** і **M2**, стандартні відхилення – до вікон **StDv1** та **StDv2**, обсяги вибірок – до вікон – **N1** та **N2**.

Встановіть для гарантії двосторонній критерій **(Two-sided)** і натисніть клавішу **Compute** (обчислювати).

Ви одержали p = 0,000, тобто на користь нульової гіпотези про те, що висота рослин у цих посівах не відрізняється, є нуль шансів. Гіпотезу слід відкинути і зробити висновок, що середня висота рослин статистично достовірно відмінна.

Друге завдання

Кореляція кількості зерен в колосі з розміром прапорового листа в одного сорту пшениці дорівнює +0,73 при обсязі вибірки 100, а в іншого сорту +0,68 при обсязі вибірки 95. Визначте, однакові чи різні ці кореляції.

Для цього заповніть вікно для коефіцієнтів кореляції **(Difference between two correlation coefficients)**. Коефіцієнти кореляції

внесіть до вікна **r1** і **r2**, а обсяги вибірок – до вікон **N1** і **N2**. Встановте двосторонній критерій **(Two-sided)** і натисніть клавішу **Compute** (обчислювати).

Ви одержите p = 0,4945, тобто на користь нульової гіпотези про відсутність відмінностей в кореляції 50% шансів. Відкидати її не можна. Слід зробити висновок, що скорельованість цих ознак в обох випадках однакова.

Третє завдання

Схожість насіння цукрового буряку на одному полі становила 86% (обсяг вибірки 200), а на іншому – 98% (обсяг вибірки 200). Чи відрізняється схожість цукрового буряку на цих полях?

Для відповіді на це запитання викличте підмодуль, як ви робили раніш, і заповніть поля у вікні "відсотків" **(Defference between two percentages)**, записуючи 86% як 0,86, 98% як 0,98. Значення процентів вводяться в вікна **%1** як **%2**, обсяги вибірок – у вікна **N1** і **N2**.

При двосторонньому критерії p = 0,0000, тобто на користь нульової гіпотези про відсутність відмінностей шансів немає зовсім. Нульову гіпотезу слід відкинути і визнати, що схожість насіння на двох полях статистично достовірно відрізнялась.

ПРАКТИЧНА РОБОТА № 21

Мета роботи: Навчитись використовувати двофакторний дисперсійний аналіз в обробці результатів польового досліду.

Для оволодіння навичками двофакторного дисперсійного аналізу проведіть аналіз результатів польового досліду, в якому вивчалась дія вапнування та фосфорних добрив на врожайність озимої пшениці. Кожний з двох чинників, що вивчався, - вапнування і фосфорні добрива, поділений на дві градації (дози): вапнування – без вапнування, фосфорні добрива – без фосфорних добрив. Повторність досліду була трикратною. Результати занесені до таблиці за загальноприйнятою формою:

Варіанти		Повторення		
		1	2	3
Без вапнування	Без фосфору	58	84	39
	Фосфорні добрива	72	72	64
Вапнування	Без фосфору	49	55	48
	Фосфорні добрива	74	74	85

1. В таких складних випадках для зручності та задля попередження помилок кодування варіантів при заповненні електронної таблиці краще доручати комп'ютеру, а назви варіантів записувати такими, як вони є: Lime (вапнування), Nolime (без вапнування), Phos (фосфорні добрива), Nophos (без фосфорних добрив). Слід лише потурбуватись, щоб для пакета Statistica було зрозуміло, які дані про врожаї відповідатимуть яким співвідношенням чинників.

2. Електронна таблиця повина мати такий вигляд:

	VAR1	VAR2	VAR3
1	Nolime	Nophos	58
2	Nolime	Nophos	84
3	Nolime	Nophos	39
4	Nolime	Phos	72
5	Nolime	Phos	72
6	Nolime	Phos	64
7	Lime	Nophos	49
8	Lime	Nophos	55
9	Lime	Nophos	48
10	Lime	Phos	74

11	Lime	Phos	74
12	Lime	Phos	85

Уважно розберіться в її структурі: перші дві змінні призначені для запису варіантів, в третій - записані облікові дані врожаїв у строгій відповідності до змісту варіантів. Після заповнення таблиці збережіть її у вигляді файлу з оригінальною назвою.

3. Для виклику модуля дисперсійного аналізу виконайте такі процедури:

Analysis
Other statistics
ANOVA/MANOVA
Replace
Switch to

4. У вікно, що відкрилось (рис. 39), перш за все введіть назви варіантів через клавішу **Variables**. Як незалежні **(Independent)**, слід відзначити одразу дві змінні VAR1 і VAR2. Це робиться при натиснутій клавіші **CTRL**. Як залежну **(Dependent)**, позначте VAR3, бо саме в ній містяться дані по врожаях. Натисніть клавішу **OK**.

5. Під клавішею **Codes for** (Коди для ...) слід позначити **All** (Всі) як для першої, так і для другої змінної. У вікнах з'являться відповідні записи варіантів досліду. Інші поля залиште без змін. Після цього натисніть послідовно

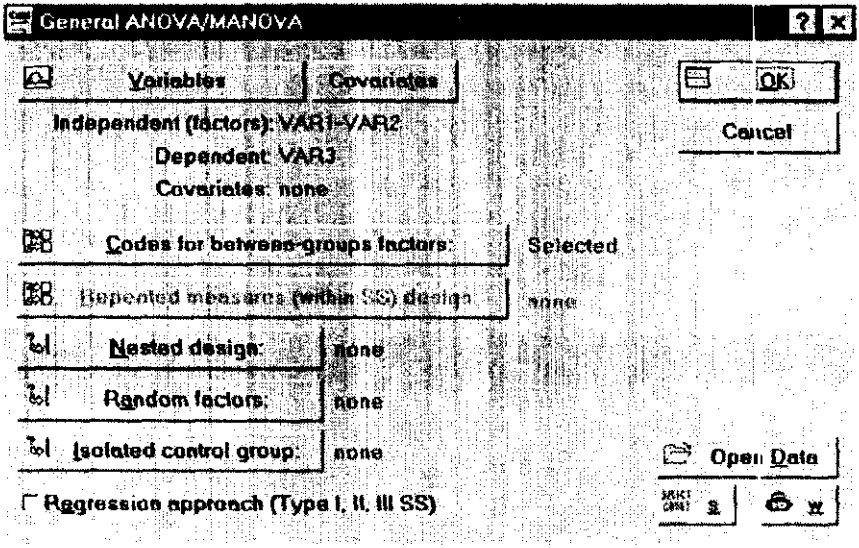


Рис. 39. Підмодуль для проведення двофакторного дисперсійного аналізу

OK
OK

6. Відкриється вікно **ANOVA Results** (Результати дисперсійного аналізу) (рис. 40), у верхній частині якого записані особливості опрацьованого матеріалу. Звірте їх правильність. Після цього можна викликати результат дисперсійного аналізу, активізувавши клавішу **All effects** (Всі ефекти).

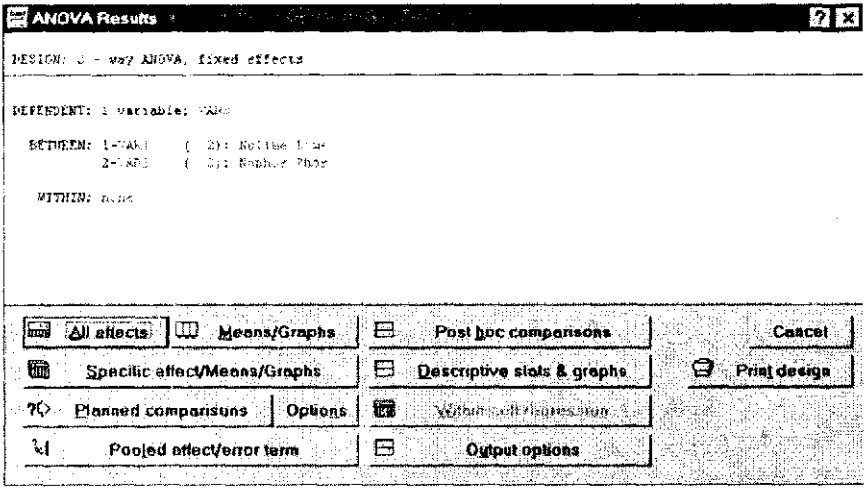


Рис. 40. Вікно з попередніми результатами двофакторного дисперсійного аналізу

7. Розгляньте одержані результати і проаналізуйте їх. Випишіть в робочий зошит або роздрукуйте. Видно, що достовірний вплив на врожай озимої пшениці мали лише фосфорні добрива, вплив на врожай вапнування та вапнування в поєднанні з фосфорними добривами був статистично не достовірним.

8. Для більш поглибленого аналізу результатів активізуйте, послідовно натискаючи, клавіші **Continue** і **Post hoc...** Майте на увазі, що Post hoc порівняння робляться самостійно для кожної із змінних і їх поєднання. Тому необхідно повертатись до вікна **Specify Effect for Post Hoc...** і послідовно проаналізувати VAR1 і VAR2, а потім, позначивши обидві змінні, і їх сполучення (рис. 41).

Позначивши в цьому вікні **VAR1**, а потім, натиснувши клавіші **OK** та **Means**, одержимо значення середніх арифметичних:

	VAR3
Nolime ...	64,83
Lime ...	64,17

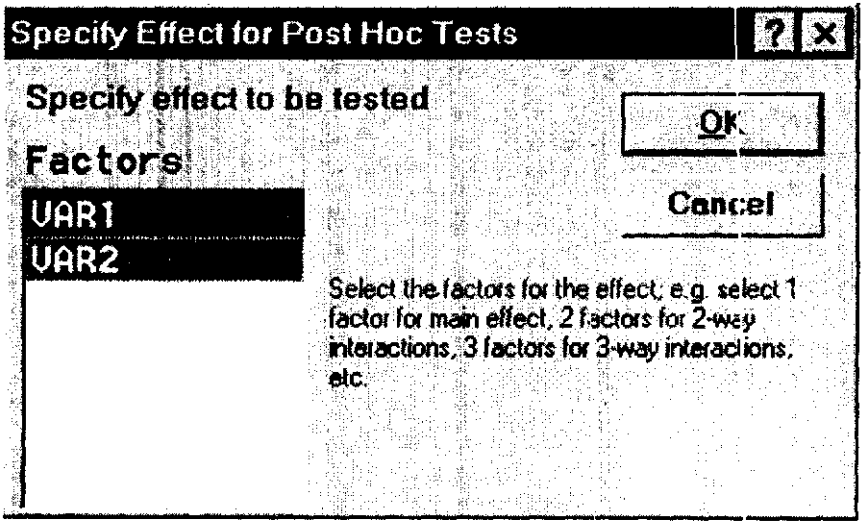


Рис. 41. Вікно, яке дозволяє проводити порівняння варіантів при двофакторному дисперсійному аналізі.

Бачимо, що сама по собі різниця між середнім врожаєм на ділянках з вапнуванням і без вапнування дуже невелика.

Позначивши у вікні Specify Effect for Post Hoc... VAR2, одержимо середні арифметичні:

	VAR3
Nophos	55,50
Phos	73,50

В цьому випадку врожайність по фону фосфорних добрив виявляється набагато вищою, ніж без добрив.

Якщо відзначити одразу обидві змінні, то буде помітний такий порівняльний результат:

	VAR3
Nolime Nophos	60,33333
Nolime Phos	69,33334
Lime Nophos	50,66667
Lime Phos	77,66666

9. Статистичну достовірність відмінностей між варіантами досліду визначте за критерієм Шеффе, активізувавши клавішу **Scheffe test** у вікні **Pos hoc Comparisions of Means** (рис.42). Знов послідовно у вікні **Specify Effect for Post Hoc...** позначайте VAR1,

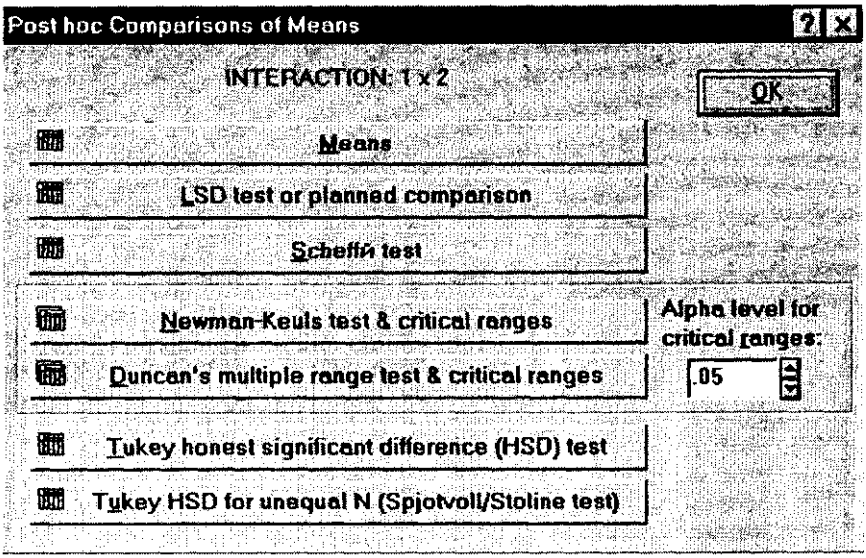


Рис. 42. Підмодуль критеріїв, що використовуються при порівнянні варіантів в двофакторному дисперсійному аналізі

VAR2, а потім їх поєднання. Тест Шеффе покаже, що статистично достовірна лише різниця варіантів з фосфорними добривами та без них, вапнування та поєднання вапнування з фосфорними добривами у всіх випадках не дали статистично достовірного ефекту.

10. Менш суворий тест НІР (LSD) з'ясує, що статистично достовірна відмінність варіантів фосфорні добрива – без фосфорних добрив (p = 0,0328). Під час аналізу VAR1, VAR2 разом з'ясується статистично достовірна відмінність 3-го та 4-го варіантів, тобто фосфор давав ефект лише по фону вапняку.

11. За допомогою критерію Дункана аналогічним чином одержують НІР, виражене в одиницях врожаю. Порогова різниця становить:

- НІР для вапнування – 16,1 ц/га;
- НІР для фосфорних добрив – 16,1 ц/га;
- НІР для поєднання – 22,8 – 24,3 ц/га.

Силу впливу факторів доводиться дораховувати на калькуляторі, використовуючи дані дисперсійного аналізу. Вона дорівнює:

$\eta_{вапн.} = (1,33/2388,3) \times 100 = 0,06\%$;

$\eta_{фосф.}$ = (972,0/2388,3) x 100 = 40,7%;

$\eta_{вапн.+фосф.}$ = (243,0/2388,3) x 100 = 10,1%.

Загальні результати двофакторного дисперсійного аналізу оформіть у вигляді таблиці 1.

Таблиця 1

Джерело змін	Сума квадратів	Ступені свободи	Середні квадрати	Критерій Фішера	Довірчий рівень
Вапнування	1,33	1	1,33	0,009	0,926
Фосфорні добрива	972,0	1	972,0	6,63	0,033
Вапнування+ фосф. добрива	243,0	1	243,0	1,65	0,233
Випадкове	1172,0	8	146,5	-	-
Загальне	2388,3	11	-	-	-

НІР для вапнування – 16,1 ц/га;

НІР для фосфорних добрив – 16,1 ц/га;

НІР для поєднання вапнування з фосфорними добривами – 24,3 ц/га;

сила впливу фактора для вапнування – 0,06%;

сила впливу фактора для фосфорних добрив – 40,7%;

сила впливу для поєднання вапнування з фосфорними добривами – 10,1%.

Зробіть агрономічний висновок за результатами досліду: на врожайність озимої пшениці достовірно вплинули лише фосфорні добрива. На грунтах, характерних для цієї дослідної ділянки, вапнування було неефективним.

ПРАКТИЧНА РОБОТА № 22

Мета роботи: Закріплення навичок проведення двофакторного дисперсійного аналізу.

В польовому досліді з п'ятьма сортами картоплі (їх позначили літерами A, B, C і т.д.) в дні з різною температурою повітря визначали кількість відкладення яєць колорадського жука в шт./м^2 листової поверхні. Повторність досліду 2-кратна. Результати досліду оформлені у вигляді таблиці:

Сорт картоплі	Температура, °С	Повторення	
		1	2
A	20	1	1
	25	0	4
	30	0	1
B	20	11	9
	25	5	5
	30	10	14
C	20	4	3
	25	3	2
	30	1	1
Д	20	10	7
	25	8	6
	30	5	7
E	20	2	0
	25	2	0
	30	2	4

1. Проведіть повний дисперсійний аналіз цього досліду і зробіть висновки.

2. Порівняйте свою роботу з наведеним нижче результатом.

Дисперсійний аналіз досліду:

Джерело змін	Сума квадратів	Ступені свободи	Середні квадрати	Критерій Фішера	Довірчий рівень
Сорт	305,52	4	76,38	33,69	0,0000
Температура	9,26	2	4,63	2,04	0,1640
Взаємодія	63,04	8	7,88	3,47	0,0180
Випадкове	33,9	15	2,26	-	-
Загальне	411,72	29	-	-	-

$НІР_{сорт}$ = 1,85-2,04
$НІР_{темп.}$ = 1,4-1,5
$НІР_{сорт+темп.}$ = 3,2-3,7
$\eta_{сорт}$ = 74,2%
$\eta_{темп.}$ = 2,2%
$\eta_{сорт+темп.}$ = 15,3%

Результати **Post hoc** співставлень

Середні арифметичні по сортах: A – 1,16; B – 9,0; C – 2,3; D – 7,2 та E-1,66.

За LSD тестом достовірно відрізнялись сорти:

A-B, A-D, B-C, B-E, C-D, D-E.

Тест Дункана для сортів дорівнює 1,85-2,04.

Середнє арифметичне по температурі: 20°C – 4,8; 25°C – 3,5; 30°C – 4,5.

За LSD-тестом достовірної різниці в кількості яйцекладки при різних температурах немає. Тест Дункана для температури дорівнює 1,4-1,5.

Можна зробити висновок, що відкладання яєць достовірно залежало від сорту і співвідношення сорту і температури. Але вплив сорту був більш вираженим і важливішим. Особливо складні і цікаві взаємодії сорт – температура повітря. Для їх повної інтерпретації слід врахувати біологію сортів. Бачимо, що найбільш вразливим є сорт B: за будь-якої температури відкладання яєць колорадського жука на ньому були найбільшими.

Задачі для самостійного розв'язання

Задача № 22.1.

В польовому досліді вивчали відкладання яєць непарним шовкопрядом на три різні породи дерев при реєстрації рівня освітленості в місці відкладання яєць. Одержані такі результати:

Варіанти		Повторення		
		1	2	3
Дуб	Освітлене	155	128	194
	Затінене	137	134	73
Клен	Освітлене	194	194	154
	Затінене	25	58	63
Яблуня	Освітлене	117	140	121
	Затінене	128	184	122

Проведіть повний дисперсійний аналіз і визначте роль деревної породи та освітленості для вибору місця відкладання яєць шовкопрядом.

Задача № 22.2.

В польовому досліді з трикратною повторністю вивчали два сорти томатів у два роки з різними метеорологічними умовами. Одержані такі результати врожаю томатів в кг плодів/кущ:

Варіанти		Повторення		
		1	2	3
Сорт А	1997 р.	2	3	2
	1998 р.	2	2	3
Сорт Б	1997 р.	3	3	2
	1998 р.	3	4	2

Проведіть повний дисперсійний аналіз і визначте роль погодних умов для врожаю томатів.

ПРАКТИЧНА РОБОТА № 23

Мета роботи: Оволодіння навичками проведення лінійного регресійного аналізу.

Для проведення регресійного аналізу розгляньте таку задачу. У семи сортів ячменю з різною тривалістю періоду вегетації від скоростиглих до пізньостиглих – визначали середню вагу 1000 зернівок. Були одержані такі дані:

	VAR1	VAR2
1	90	47.50
2	85	46.75
3	80	45.75
4	75	42.85
5	70	44.76
6	65	41.44
7	60	37.00,

де VAR1 – довжина періоду вегетації ячменю в днях, VAR2 – середня вага 1000 зерен в грамах.

1. Почніть проведення регресійного аналізу, як завжди, із створення бази даних і збереження її у вигляді файлу з оригінальною назвою.

2. Відкрийте модуль регресійного аналізу, виконавши такі команди:

Analysis
Other statistics
Customize list
Multiple regression
Replace
Switch to...

3. У вікні регресійного аналізу (рис. 43) знайдіть клазішу

Variables

Активізуйте її.

4. Відкрилось вікно для вибору змінних. Як незалежну змінну (**independent**) оберіть тривалість періоду вегетації ячменю (VAR1), а як залежну (**dependent**) – вагу 1000 зерен (VAR2).

OK

5. Перевірте настройку опцій для проведення регресійного аналізу. В рядку **Input file** повинно бути записане **Raw data** (вихідні

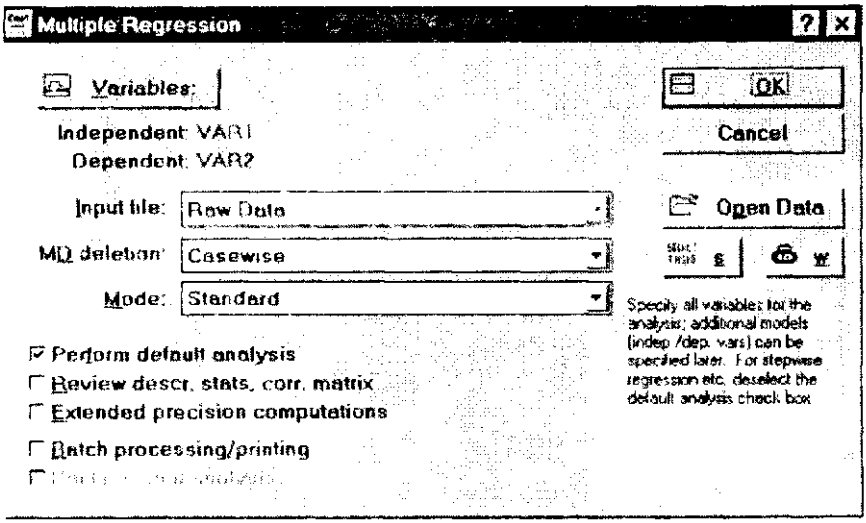


Рис. 43. Підмодуль для проведення регресійного аналізу

дані). В рядку **MD deletion** повинно залишитись **Casewise**. В рядку **Mode** повинно залишитись **Standard**.

OK

6. Відкриється вікно з результатами регресійного аналізу (рис. 44).

Перш за все випишіть і оцініть підсумки дисперсійного аналізу лінії регресії (клавіша **Analysis of variance**) (рис. 45). Вони такі:

F = 24,7

df = 1 і 5

p = 0,004

На користь нульової гіпотези $b = 0$ зовсім мало шансів, отже, прямолінійна регресія статистично достовірна на рівні 95%.

Натисніть клавішу **Continue** і з верхнього вікна загальних результатів випишіть

RI (R^2) = 0,83

Це коефіцієнт детермінації. Загальний розкид тут приймається за 1,0 і *RI* показує частину цього розкиду, враховану регресією. Вона становить 83%, що означає: на частку випадкових факторів, не врахованих регресією, залишилось лише 17%. Це добрий результат.

7. Тепер натисніть клавішу

Regression summary

У стовпці "B" (рис. 46) наведене значення вільного члена рівняння регресії (**intercept**).

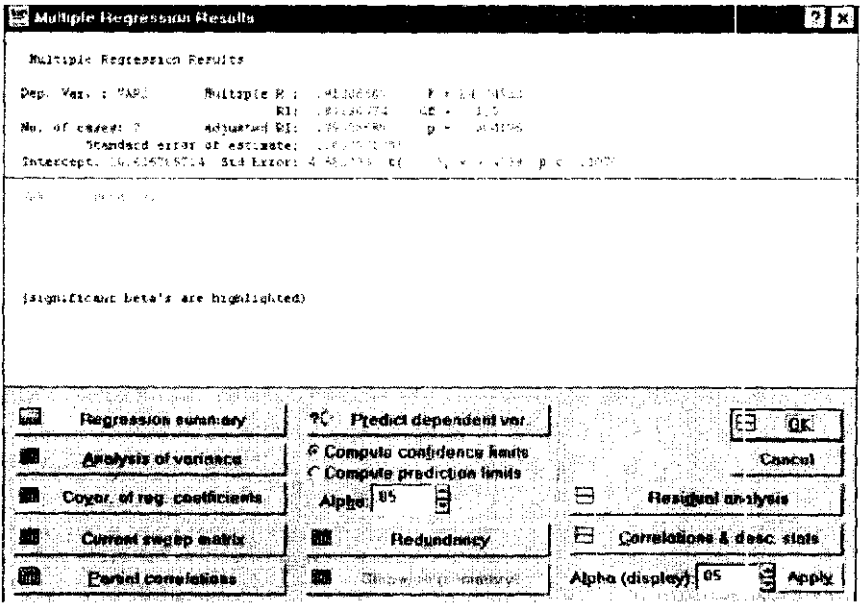


Рис. 44. Вікно попередніх результатів дисперсійного аналізу лінії регресії

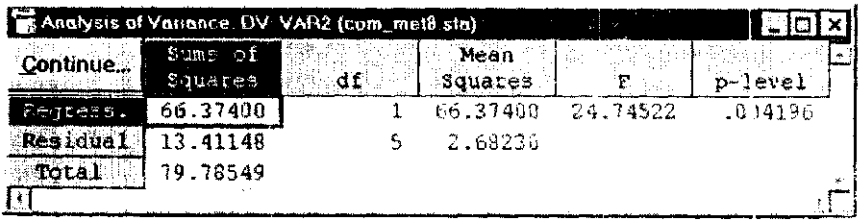


Рис. 45. Вікно остаточних результатів дисперсійного аналізу лінії регресії

Він дорівнює

a = 20,6,

а із стовпця p видно, що шансів на користь нульової гіпотези про те, що a дорівнює нулю, всього 0,006. Отже, рівняння регресії статистично достовірне.

Коефіцієнт b записаний на перетині стовпця "B" і VAR1.

b = 0,308 (при p = 0,004).

Статистично достовірний і цей коефіцієнт. Це дозволяє записати рівняння регресії в такому вигляді:

[Вага 1000 зернин] = 20,6+0,3·[Тривалість вегетаційного періоду].

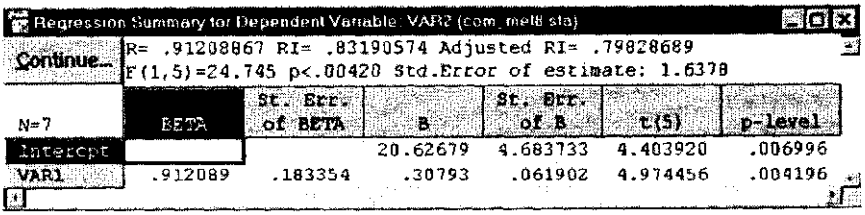


Рис. 46. Параметри лінії регресії

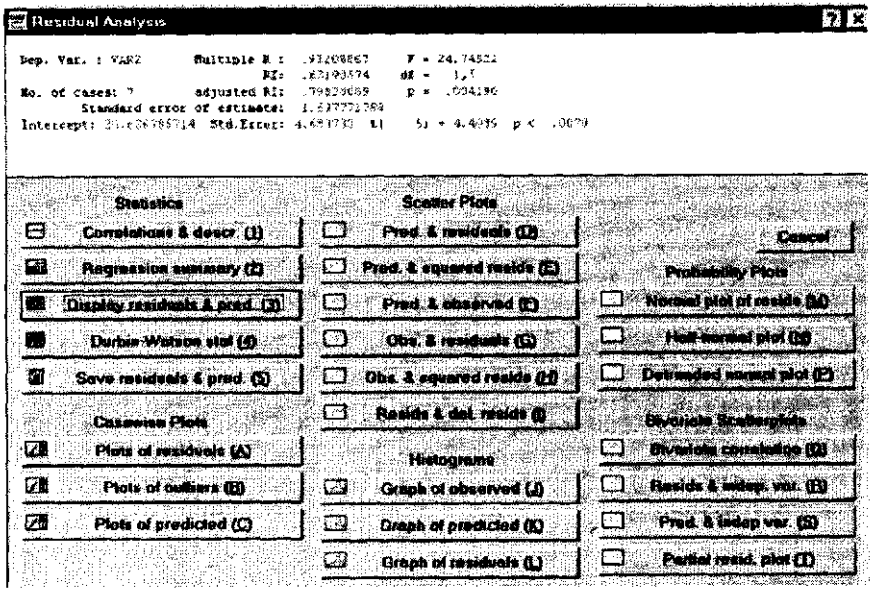


Рис. 47. Підмодуль для аналізу залишків при регресійному аналізі

8. Основне завдання регресійного аналізу розв'язане. За цим рівнянням можна обрахувати вагу 1000 зерен ячменю для сорту з будь-якою тривалістю вегетаційного періоду.

9. Модуль регресійного аналізу надає можливості одержати і деяку іншу корисну інформацію. Для цього активізуйте

Continue

10. У вікні, що з'явилось, активізуйте клавішу **Residual analysis** – аналіз залишків, а потім

Displays residuals a.predict (показати залишки і передбачені значення) (рис. 47).

В таблиці, що відкрилась, в графі **Pred. Value** (прогнозні значення) наведені значення ваги 1000 зерен, що точно відповідають тій чи іншій тривалості вегетаційного періоду. Це ті точки, що лежать точно на прямій лінії регресії.

11. Для перевірки залишків поверніться в попереднє вікно за допомогою клавіші **Continue** і активізуйте клавішу **Normal Plot of resids** (Залишки на нормальному імовірнісному аркуші). Видно, що точки гуртуються навколо прямої лінії. Таким чином, одержана модель цілком припустима, бо залишки випадкові і не не містять прихованих закономірностей.

12. Звичайно результати регресійного аналізу подаються у вигляді графіка. Його можна подивитись, активізувавши у вікні **Residual Analysis** клавішу

Pred.@observed (передбачені значення і ті, що спостерігаються).

На цьому графіку показана лінія регресії та її 95%-ві довірчі інтервали.

13. Можна побудувати цей графік і по-іншому (рис. 48). Для такої альтернативної побудови поверніться до вікна з файлом даних (всі інші вікна слід закрити), активізуйте його (про це свідчитиме яскрава синя смуга у верхній частині вікна). Далі із загального меню виконайте команди

Graphs (графік)
Starts 2D Graph (статистичний двомірний графік)
Scatterplot (точки на графіку)

14. Відкрилось підменю. Під клавішею **Variable** введіть як вісь x – VAR1, а як вісь y – VAR2. Натисніть клавішу **OK**. В зоні **Graph Type** (типи графіків) виділіть **Regular** (регулярний), а в зоні **FIT** (підгонка) – **Linear** (лінійний). В зоні **Coefficient** (в даному випадку це довірчий рівень) поставте 0,95. В зоні **Confidence Bands** (довірча зона) також повинно стояти 0,95.

OK

15. Одержане графічне подання лінії регресії. Видно, що в заголовку графіка записане рівняння регресії, яке раніш було одержане більш складним, але статистично надійним шляхом. Довірчі рівні не показані.

16. На прикладі цього графіка можна оволодіти технологією заміни написів англійською мовою, написами російською чи українською мовами. З цією метою розгорніть графік на весь екран. Потім клацніть правою клавішею миші (ПКМ) по будь-якому рядку тексту. У підменю, що відкрилось, оберіть ЛКМ опцію **Edit text** (редагувати текст). Відкриється вікно з усіма записами, які є на

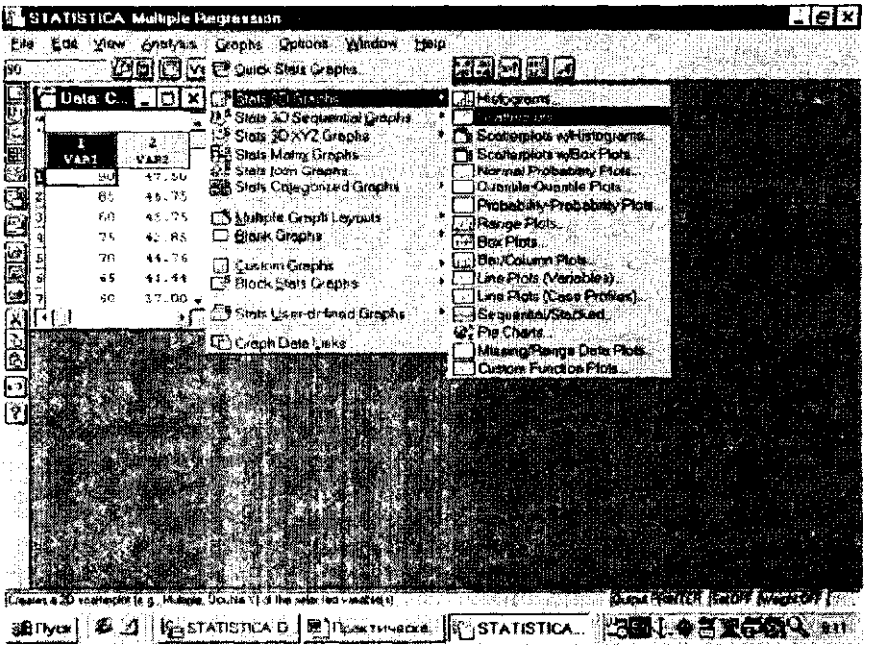


Рис. 48. Приклад використання графічних можливостей пакета "Статистика" для побудови лінії регресії

графіку і які можуть бути до нього додані. Забийте клавішею **Backspace** ці написи і запишіть замість них свої.

Замість напису в першому рядку запишіть "Залежність ваги 1000 зерен ячменю від тривалості вегетаційного періоду". Щоб напис був зроблений кирилицею і мав підходящий розмір, натисніть клавішу **"AB"** і оберіть шрифт Time Roman Cyr-8 (чи будь-який інший, що підтримує кирилицю, звичайно це можуть не всі шрифти). Аналогічним чином можна замінити чи додати всі інші написи.

17. Крім того, можна підписати рисунок знизу. Для цього поверніть графік до віконного розміру і активізуйте клавішу у верхньому меню:

Set to fixed Graph Proportion (Настройка графіка)

За допомогою ЛКМ звільніть під графіком місце для підпису. Потім натисніть клавішу у верхньому меню.

Graphic text editor (Редактор написів графіка).

У полі, що відкрилось, напишіть, наприклад, "Рис. 1. Залежність ваги зерен у семи сортів ячменю від тривалості періоду вегетації". Підберіть потрібний шрифт.

OK

Курсор змінив свою форму і замість стрілки набув форму хрестика. Обережно підведіть його до того місця під графіком, де, на вашу думку, повинен починатись підпис, і клацніть ЛКМ. Підпис став на своє місце. Якщо він вас не влаштовує, його можна виправити: поставте курсор точно на підпис і клацніть ЛКМ. Оберіть опцію **Edit text** і внесіть виправлення.

18. Роздрукуйте одержаний вами графік. Попередньо через опцію **Print preview** перевірте його розміщення на аркуші паперу. Розміщення графіка на аркуші паперу можна настроювати клавішею *Margins (Краї)*.

ПРАКТИЧНА РОБОТА № 24

Мета роботи: Закріплення навичок проведення повного лінійного регресійного аналізу.

В польовому досліді вивчалась дія зростаючих доз азотних добрив (кг д.р./га) на врожай озимої пшениці. Були одержані такі дані:

Дози добрив: 0-10-20-30-40-50;

Урожайність: 26.0-34.0-36.3-37.8-38.6-38.9.

Визначте залежність урожайності від доз азотного добрива.

1. Проведіть повний лінійний регресійний аналіз відповідно з вказівками, наведеними у практичній роботі № 23.

2. Після проведення лінійного регресійного аналізу можна зробити такі висновки (порівняйте свої результати з ними):

1. Експериментальні дані не дуже добре лягають на пряму лінію, виявляється закон затухаючої дії факторів харчування рослин. Але криволінійність виражена слабо, і можливе застосування лінійної апроксимації.

2. Дисперсійний аналіз ліній регресії дав F = 12,86 при p = 0,023, оскільки це менше 0,05, то нульова гіпотеза відкидається і лінія регресії статистично достовірна на 95% рівні.

3. RI(R^2) = 0,76 = 76%. При цьому 24% залишається на інші фактори – це досить багато.

4. Параметри рівняння регресії дорівнюють: b = 0,23 (p = 0,023), a = 29,57 (p = 0,0001). Обидва параметри статистично достовірно відрізняються від нуля. Отже, рівняння регресії виду

[Урожайність] = 29,57 + 0,23·[Доза добрива].

Це рівняння статистично достовірне на рівні 95%.

5. Перевірка залишків на нормальному імовірнісному аркуші показала, що є помітні відхилення від прямої лінії. Це свідчить про приблизний характер виконаної лінійної апроксимації.

6. Графік передбачуваних значень урожайності має вигляд, відповідний рис. 49. Видно, що довірча зона дуже широка. Відносно малих і великих доз добрив прогнози врожаю явно матимуть низьку точність.

Задачі для самостійного розв'язання

Задача № 24.1. В 10 зразках грунту визначили pH сольової витяжки (VAR1) і вміст P_2O_5 мг на 100 г грунту (VAR2). Визначте за допомогою регресійного аналізу залежність вмісту фосфору від pH грунту. Вихідні дані такі:

ПРАКТИЧНА РОБОТА № 25

Мета роботи: Оволодіння навичками проведення покрокового множинного регресійного аналізу.

У 10 рослин люцерни були враховані п'ять ознак: 1 – вага надземної фітомаси, г; 2 – висота рослин, см; 3 – кількість листків, шт.; 4 – кількість бокових гілок, шт.; 5 – вихід насіння, г. В інтересах селекційної роботи визначте, від якої ознаки чи ознак найбільше залежить вихід насіння люцерни.

Були одержані такі дані:

	VAR1	VAR2	VAR3	VAR4	VAR5
1	15.620	96.8	198.0	16.0	0.960
2	15.200	119.0	115.0	14.0	1.490
3	16.530	116.8	97.0	9.0	2.100
4	7.190	85.0	74.0	9.0	0.098
5	11.420	83.0	135.0	12.0	0.066
6	2.560	71.6	23.0	1.0	0.040
7	17.020	102.0	12.0	16.0	0.590
8	4.880	108.3	42.0	12.0	0.104
9	8.470	91.0	107.0	19.0	0.028
10	3.990	78.5	54.0	10.0	0.076

1. Створіть із цих даних базу даних і збережіть її під оригінальним ім'ям.

2. Відкрийте модуль множинного регресійного аналізу, виконавши для цього такі команди:

Analysis
Other statistics
Customize list
Multiple regression
Replace
Switch to...

3. У вікні множинного регресійного аналізу (рис. 50) після активізації клавіші **Variables** як незалежні **(independent)** змінні позначте VAR1- VAR4, а як залежну **(dependent)** – VAR5.

OK
OK

4. Перш за все розгляньте результати дисперсійного аналізу рівняння множинної регресії, наведені у верхній частині вікна (рис. 51). Вони такі:

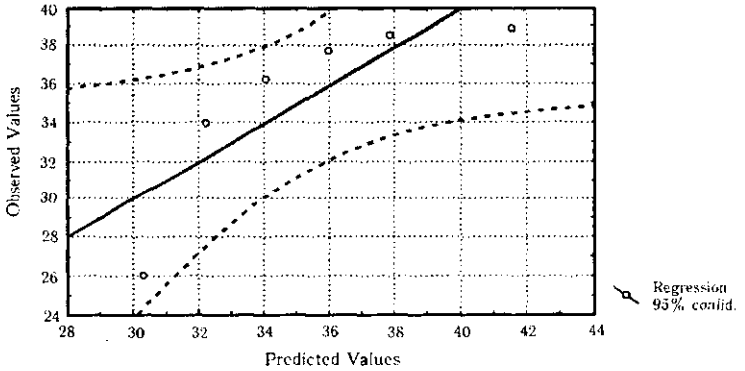


Рис. 49. Приклад прогнозування величини врожайності до практичної роботи № 24

	VAR1	VAR2
1	4.5	10.0
2	4.6	10.1
3	4.7	10.2
4	4.7	10.2
5	4.7	10.4
6	4.8	10.5
7	4.8	10.5
8	4.9	10.7
9	4.9	10.7
10	4.9	10.8

Задача № 24.2. Перед випіканням хліба муку прогрівали від чверті години до 8 годин при температурі 45°С (VAR1). Якість тіста оцінювали по умовній шкалі Q (VAR2). Були отримані такі результати:

	VAR1	VAR2
1	.3	93
2	.5	71
3	.8	63
4	1.0	54
5	1.5	43
6	2.0	38
7	3.0	29
8	4.0	26
9	6.0	22
10	8.0	22

Визначте характер залежності якості тіста від прогрівання муки.

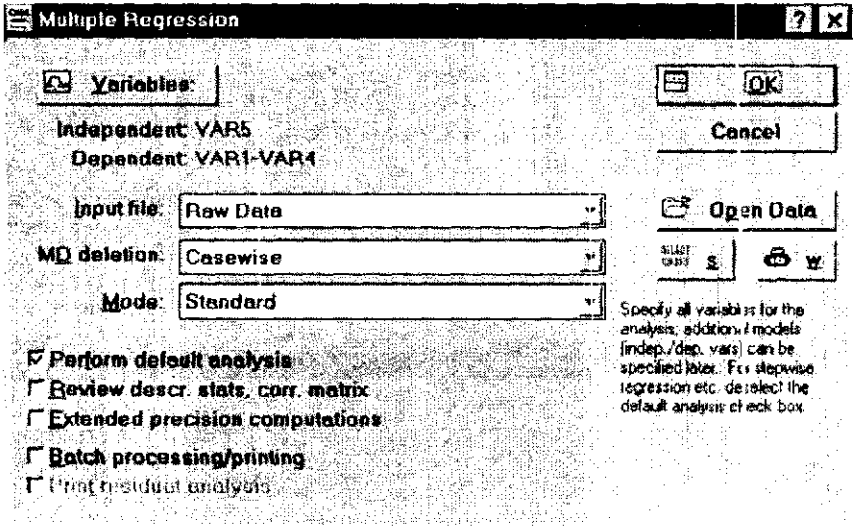


Рис. 50. Підмодуль для проведення покрокового множинного регресійного аналізу

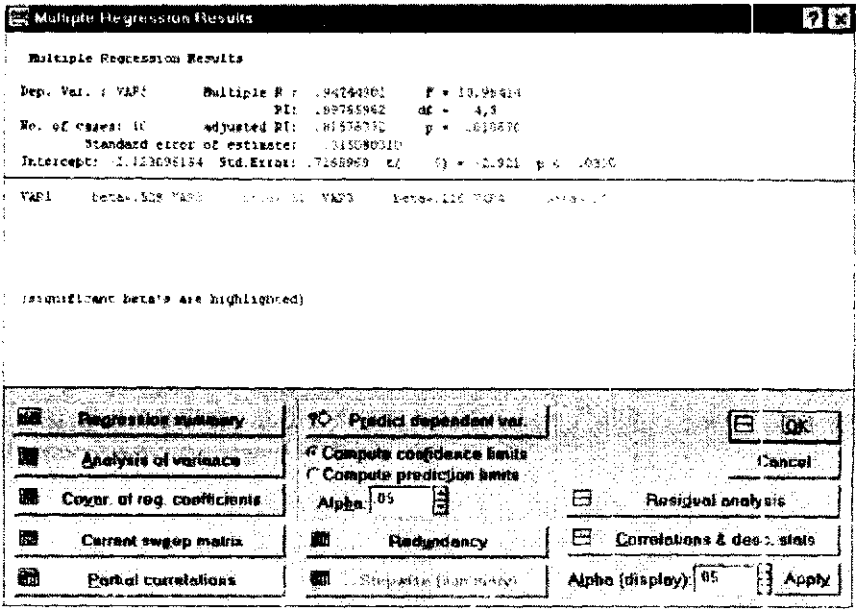


Рис. 51. Вікно попередніх результатів покрокового множинного регресійного аналізу

F = 10,96 і p = 0,0108,

RI = 0,897, або 89,7%.

Таким чином, модель статистично достовірна на рівні 95% і охоплює 89,7% загальної мінливості виходу насіння.

5. Для подальшого аналізу активізуйте клавішу

Regression summary

6. З'явиться нове вікно (рис. 52). Випишіть вільний член і регресійні коефіцієнти. Вільний член і регресійні коефіцієнти беруться з колонки **B**. Загальний вид рівняння виявиться таким:

VAR5 = -2,12 + 0,070VAR1 + 0,028VAR2 + 0,003VAR3 − −0,078VAR4.

Regression Summary for Dependent Variable: VAR5

Continue...

R= .94744901 RI= .89765962 Adjusted RI= .81578732
F(4,5)=10.964 p<.01087 Std.Error of estimate: .31508

N=10	BETA	St. Err. of BETA	B	St. Err. of B	t(5)	p-level
Intercpt			-2.12310	.726897	-2.92077	.032984
VAR1	.528619	.223615	.06962	.029449	2.36398	.064432
VAR2	.612117	.200289	.02775	.009080	3.05617	.028225
VAR3	.216063	.170742	.00280	.002212	1.26543	.261479
VAR4	-.533925	.176941	-.07794	.025830	-3.01754	.029499

Рис. 52. Вікно з оцінками лінії регресії, одержаної після проведення множинного регресійного аналізу

Видно, що деякі коефіцієнти регресії настільки малі, що є зрозумілим, наскільки малий їх внесок у вихід насіння.

До того ж довірчі рівні вільного члена і коефіцієнтів регресії (колонка **p-level**) істотно різняться:

вільний член p = 0,03

VAR1 p=0,06

VAR2 p=0,028

VAR3 p=0,261

VAR4 p=0,029

Коефіцієнти при VAR3 і VAR1 статистично не достовірні, але в інших значеннях p менше 0,05, і коефіцієнти статистично достовірні.

Модель явно потребує уточнень.

7. Для цього виконаємо такі процедури:

Continue

OK

Cancel

8. Відкриється вікно **Model Definition** (рис. 53). В ньому необхідно в вікнах поставити такі опції:

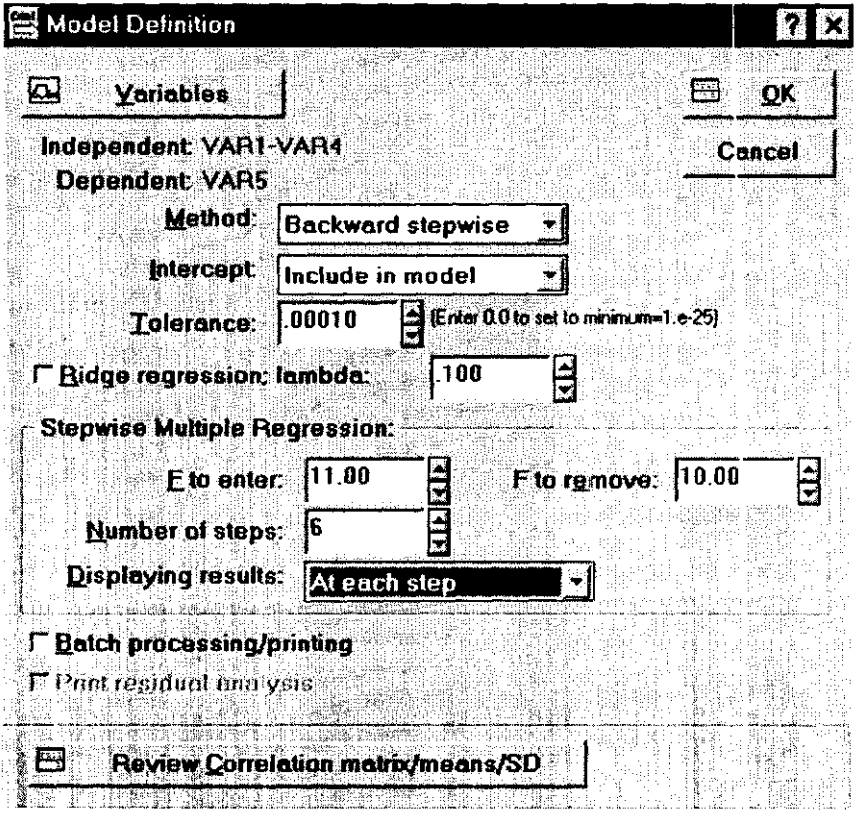


Рис. 53. Підмодуль визначення параметрів і уточнення моделі при виконанні покрокового множинного регресійного аналізу

Method-Backward stepwise
Intercept-Include in model
F to remove – 10,0
Displaying results – Тут пропонуються дві альтернативи:
Summary only
або
At each step

В першому випадку комп'ютер одразу видасть кінцеву модель після виключення тих факторів, для яких F менше 10, а в другому - фактори будуть виключатись послідовно по одному і кожен раз можна буде подивитись проміжні результати. В цьому випадку перехід до слідуючого кроку роблять, натискаючи клавішу **Next**. Оберіть опцію

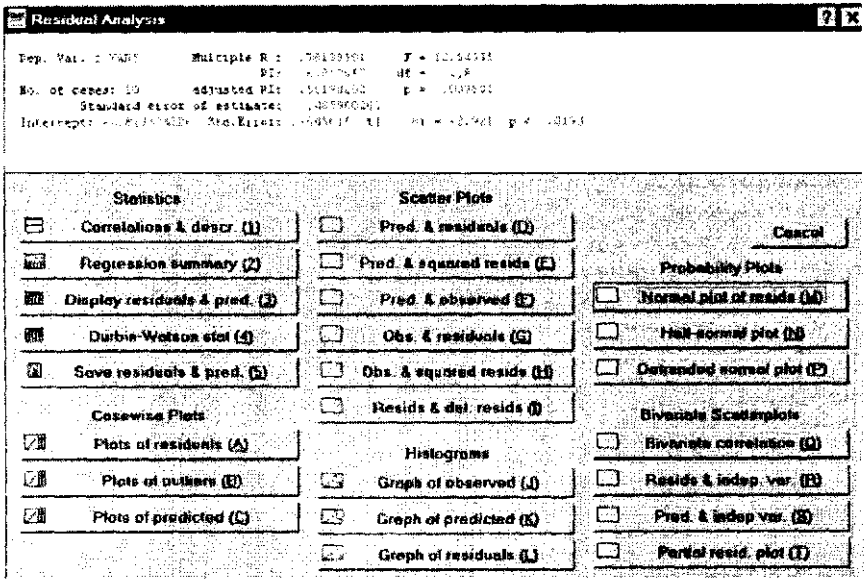


Рис. 54. Вікно підсумкових результатів покрокового множинного регресійного аналізу

At each step
OK
Next

9. Оцінуйте нові результати (рис. 54). F = 12,54, p = 0,007. Якість підгонки явно покращилась. Правда, RI = 61,05% і модель стала охоплювати менше дисперсії, що природно – адже в ній залишився лише один фактор!

1. Для запису рівняння регресії активізуйте клавішу

Regression summary

Використовуючи дані таблиці, можна записати нове рівняння регресії:

VAR5 = -2,81 + 0,035VAR2.

Довірчі рівні для параметрів моделі відповідно становлять $p = 0{,}019$ і $p = 0{,}007$, тобто вільний член і коефіцієнт статистично достовірні. Вихід насіння люцерни в підсумку на 61% визначається висотою рослин. Інші фактори менш суттєві.

Можно включити в рівняння ще й VAR1 (надземна фітомаса рослин) – ця змінна при покроковому перегляді відкидалась останньою.

11. При скороченні кількості членів регресійного рівняння суттєва інформація не втрачається. При розгляді залишків на

нормальному імовірнісному папері (**OK-OK-Normal plot of resids**), вони непогано лягають на пряму лінію.

В цілому задача розв'язана, і селекціонер одержує певні орієнтири для виділення найбільш продуктивних рослин вже на ранніх етапах онтогенезу. Це будуть найбільш високі і міцні рослини. Для них можна передбачити вихід насіння, підставляючи в рівняння регресії замість VAR2 можливі значення висоти рослин.

ПРАКТИЧНА РОБОТА № 26

Мета роботи: Закріплення навичок проведення покрокового множинного регресійного аналізу.

В польовому досліді враховували вагу коренеплодів цукрового буряку в грамах, а також кількість бур'яну (шт./м^2) на цих же пробних ділянках. Були отримані такі результати:

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8
1	94.000	9.000	8.000	7.000	1.000	4.000	0.000	0.000
2	81.000	0.000	1.000	4.000	0.000	9.000	0.000	0.000
3	146.000	0.000	0.000	12.000	1.000	17.000	0.000	0.000
4	94.000	0.000	0.000	7.000	3.000	7.000	0.000	0.000
5	103.000	0.000	0.000	10.000	3.000	10.000	11.000	0.000
6	91.000	2.000	22.000	10.000	2.000	1.000	0.000	1.000
7	166.000	7.000	6.000	1.000	2.000	0.000	0.000	0.000
8	141.000	1.000	12.000	15.000	1.000	6.000	10.000	0.000
9	96.000	0.000	1.000	8.000	5.000	3.000	0.000	0.000
10	80.000	2.000	6.000	5.000	16.000	1.000	41.000	0.000
11	47.000	0.000	0.000	5.000	50.000	0.000	10.000	2.000
12	108.000	9.000	38.000	0.000	2.000	10.000	0.000	2.000

В цій таблиці VAR1 – вага коренеплодів буряку, VAR2 – осот польовий, VAR3 – жабрій, VAR4 – гірчак березковидний, VAR5 – редька дика, VAR6 – лобода біла, VAR7 – мишій, VAR8 – гірчак шорсткий.

1. Знайдіть залежність ваги коренеплодів від кількості бур'янів і визначте, які з них найбільш шкідливі.

2. Послідовно виконайте всі етапи покрокового множинного регресійного аналізу у відповідності з роботою № 25.

1. Порівняйте свої результати з наведеними нижче.

3. За результатами покрокового аналізу одержана модель:

VAR1 = 75,5 + 0,9VAR2 + 5,39VAR3 - 0,77VAR4 + 4,19VAR5+ + 3,3VAR6 - 2,15VAR7 -106,9VAR8.

Але значення F для цієї моделі дорівнює 0,64 при p = 0,71, тобто вона статистично недостовірна.

4. Для її уточнення був проведений покроковий аналіз при пороговому значенні F = 2,0. Ця величина мала бути вписаною у вікно **Model Definition** в поле **F to remove**.

В цьому випадку модель буде такою:

VAR1 = 104,69 + 1,48VAR3 – 29,7VAR8.

Вільний член цієї моделі і коефіцієнт при VAR8 в цьому випадку статистично достовірні.

5. Таким чином, можна припустити, що на вагу коренеплодів цукрового буряку найбільш шкідливо вплизає гірчак шорсткий. Вплив інших бур'янів поки залишається нез'ясований. Для його з'ясування в цьому досліді потрібно збільшити обсяг вибірки.

Задачі для самостійного розв'язання

Задача № 26.1. У 26-ти рослин звіробою були враховані 15 морфогенетичних параметрів: 1-14 – параметри вегетативної і генеративної сфери рослин, 15 – репродуктивне зусилля в частках одиниці. Визначте, від яких параметрів і як саме залежить величина репродуктивного зусилля рослин. Вихідні дані:

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7
1	1.830	3.220	1.220	11.850	2.800	.230	17
2	1.180	6.310	4.920	3.650	3.200	4.000	26
3	1.210	6.400	2.850	4.800	2.450	2.230	17
4	1.090	6.660	.940	1.020	1.080	2.060	19
5	.300	3.250	2.060	2.310	1.360	2.420	14
6	1.220	3.140	1.660	2.960	1.170	.670	20
7	1.530	1.580	.920	1.280	3.750	.160	16
8	1.180	2.250	.800	4.220	3.590	.750	19
9	.420	.930	1.020	1.190	5.000	2.310	9
10	3.000	.270	.660	3.170	3.960	2.900	22
11	1.180	1.260	.900	2.090	2.650	1.920	19
12	.410	1.730	.600	1.080	1.070	.660	11
13	1.890	1.580	.720	1.560	.630	.630	28
14	1.200	1.770	.550	.370	.160	4.050	18
15	1.810	1.100	.790	1.040	.440	3.870	20
16	1.120	.690	.850	.430	.680	2.490	17
17	.510	.370	.320	2.480	4.170	1.280	13
18	1.900	.660	1.100	.420	4.470	.380	16
19	1.180	.310	.410	.430	2.690	1.340	25
20	.420	.530	.480	.940	3.720	.950	13
21	2.670	.610	.320	2.230	.340	.910	21
22	1.110	.200	.240	1.690	.840	3.510	16
23	.340	.190	.360	1.460	3.090	1.280	15
24	1.180	.150	.250	.940	2.180	3.400	22
25	.390	.230	.290	.140	2.590	.040	10
26	.310	.190	1.200	.360	3.310	1.170	15

	VAR8	VAR9	VAR10	VAR11	VAR12	VAR13	VAR14	VAR15
1	31	30	60	48	35	0	.53	.19
2	39	48	38	53	53	11	.14	.04
3	36	38	50	44	46	11	.29	.12
4	35	30	26	38	49	14	.11	.10
5	31	31	47	39	44	8	.30	.22
6	26	32	39	40	33	4	.31	.26
7	33	28	36	50	27	2	.42	.11
8	25	25	40	43	40	5	.65	.18
9	22	35	41	43	47	8	.73	.15
10	21	28	47	36	41	11	.58	.15
11	26	32	30	45	36	5	.22	.08
12	24	30	28	32	33	5	.15	.14
13	20	30	41	26	35	4	.22	.35
14	18	28	30	26	41	12	.05	.31
15	37	30	37	33	43	7	.16	.36
16	20	28	28	25	45	8	.29	.43
17	13	21	49	43	45	9	.76	.18
18	22	30	38	48	37	2	1.12	.25
19	15	21	32	38	35	5	.50	.19
20	19	23	35	48	37	8	1.50	.40
21	18	23	45	35	39	4	.09	.26
22	15	18	46	39	48	10	.04	.05
23	17	22	28	45	33	5	.20	.06
24	10	23	34	43	54	10	.08	.04
25	13	12	22	47	17	1	.17	.07
26	10	32	28	30	51	4	.09	.03

Задача № 26.2. Були розглянуті санітарно-гігієнічні умови і захворюваність дизентерією в 14 населених пунктах з урахуванням таких параметрів: VAR1 – чисельність населення, VAR2 – середня кількість осіб в одному жилому приміщенні, VAR3 – частка осіб, що додержуються правил особистої гігієни, VAR4 – середня площа жилого приміщення на одну людину, VAR5 – кількість змін харчування в їдальнях, VAR6 – наявність каналізації (ні – 1, так – 2), VAR7 - колі-індекс питної води, VAR8 – термін ізоляції хворого на протязі захворювання на дизентерію, VAR9 – кількість лікарів на 1000 чол. населення, VAR10 – середньорічний рівень захворюваності на дизентерію у відсотках. Одержані такі дані:

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8	VAR9	VAR10
1	140	35	.6	1.1	1	2	324	1.0	.714	.444
2	42	14	.7	2.0	1	2	231	2.0	2.380	.444
3	133	33	.7	1.5	1	2	129	6.0	.750	.222
4	160	80	.4	4.0	1	2	333	4.0	0.000	.667
5	480	120	.2	2.2	2	2	315	2.1	.208	1.333
6	45	23	.5	1.5	1	1	315	3.0	0.000	.889
7	188	94	.5	1.2	1	2	167	2.0	0.000	1.111
8	318	53	.7	4.5	1	2	333	2.5	.629	.444
9	130	26	.4	1.9	1	2	129	3.0	0.000	1.333
10	60	60	.6	4.5	1	2	12	3.0	0.000	.444
11	164	41	.5	2.5	2	2	167	3.6	.610	.889
12	160	80	.2	4.0	1	2	333	2.8	0.000	1.556
13	177	22	.7	1.8	1	2	333	1.0	.565	.222
14	240	34	.6	1.5	1	1	273	3.5	.417	.222

Визначте, який з врахованих параметрів найбільшою мірою впливає на захворюваність дизентерією.

ПРАКТИЧНА РОБОТА № 27

Мета роботи: Одержання навичок проведення кластерного аналізу.

В сільськогосподарських дослідженнях нерідко постає завдання порівняти сільськогосподарські культури, сорти рослин чи породи тварин за комплексом ознак. Таке групування об'єктів дозволяє здійснити кластерний аналіз.

1. В таблиці, що далі наводиться, вказані характеристики шести основних зернових культур за шістьма біохімічними ознаками: 1 – гігроскопічна волога, %; 2 – попіл, %; 3 – протеїн, %; 4 – клітковина, %; 5 – жир, %; 6 – безазотисті екстрактивні речовини, %.

Видно, що овес вирізняється високим вмістом попелу, яра пшениця – протеїну, яра та озима пшениці схожі за низьким вмістом клітковини тощо, але підсумкову подібність культур одразу за всіма шістьма групами ознак "на око" оцінити неможливо.

Зробіть це методом кластерного аналізу, створивши попередньо базу даних із наведеної таблиці:

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6
1. Пшениця яра	13.4	1.9	13.6	1.8	2.0	67.3
2. Пшениця озима	13.4	1.8	11.4	1.8	1.9	69.7
3. Жито	15.1	1.7	11.5	2.1	1.8	67.8
4. Ячмінь	12.4	2.6	12.3	4.5	2.5	65.8
5. Овес	12.8	3.0	10.2	10.0	5.3	59.7
6. Кукурудза	13.3	1.5	9.6	2.6	5.1	67.9

В цій таблиці: VAR1 – гігроскопічна волога, %; VAR2 – попіл, %; VAR3 – протеїн, %; VAR4 – клітковина, %; VAR5 – жир, %; VAR6 – безазотисті екстрактивні речовини, %.

2. Збережіть базу даних під оригінальною назвою.

3. Для проведення кластерного аналізу послідовно виконайте команди:

Analysis
Other statistics
Customize list
Cluster analysis
Replace
Switch to

4. У вікні, що відкрилось, оберіть опцію **Joining (tree clustering)** (процедура об'єднання – побудова дерева) (рис. 55). Натисніть **OK**.

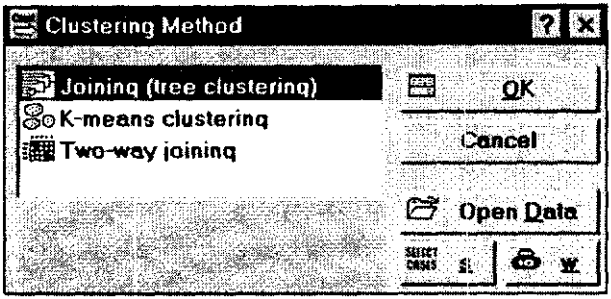


Рис. 55. Підмодуль для проведення кластерного аналізу

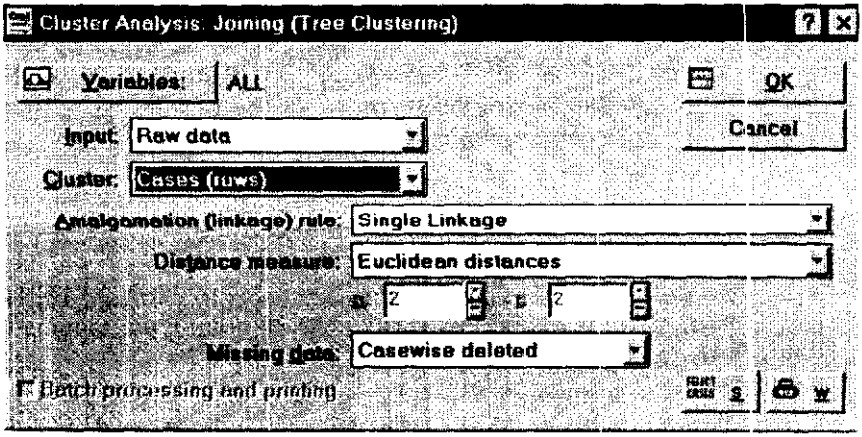


Рис. 56. Підмодуль для вибору методів розрахунків при проведенні кластерного аналізу

5. У підменю, що відкрилось (рис. 56), у вікні **Variables** відкрийте список змінних і оберіть опцію **Select All** (позначити все) для того, щоб аналіз було проведено за всіма шістьма змінними. Натисніть **OK**.

6. У вікні **Input** залишите **Raw data** (вихідні дані), оскільки ви піддаєте кластеризації файл з реальними даними без будь-якого перетворення. У вікні **Cluster** слід зробити дуже важливий вибір – у нашому випадку це **Cases**, оскільки у вихідній таблиці сільськогосподарські культури розташовані саме по строках. Якщо тут відзначити опцію **Variables**, то кластеризації були б піддані біохімічні властивості об'єктів, але не самі об'єкти, що в завдання цього аналізу не входить.

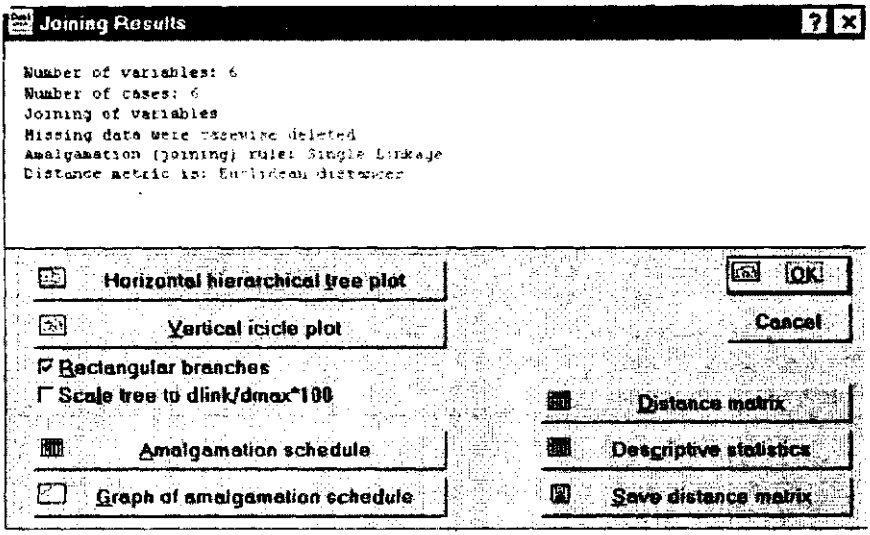


Рис. 57. Підмодуль для вибору форми подання результатів кластерного аналізу

У двох нижче розташованих вікнах слід обрати опції **Single Linkage** і **Euclidean distances** (евклідова відстань), як найбільш прості метрики.

OK

7. В новому вікні (рис. 57) слід позначити опцію **Vertical icicle plot** (вертикальне дерево кластеризації). Таке подання результатів звичайно є більш наочним.

8. Одержаний графік (рис. 58), на якому культури позначені символами від C_1 до C_6. Чим нижче пролягають лінії, що пов'язують окремі культури, тим вони більш схожі за комплексом ознак.

Графік показує, що за комплексом властивостей найбільш ізольованим є зерно вівса (C_5). А найбільш подібними за комплексом ознак є жито (C_3) і озима пшениця (C_2) (обидві озимі культури!). За своїми властивостями до них найближчим є зерно ярої пшениці (C_1). Ізольовані проміжні позиції займають ячмінь і кукурудза.

9. Якщо активізувати опції

Continue,

а потім

Distance matrix,

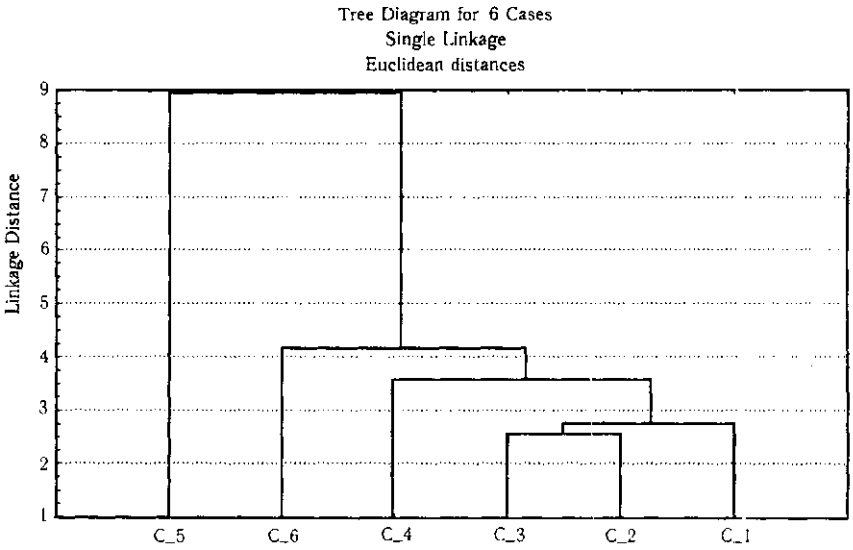


Рис. 58. Приклад вертикальної дендрограми з результатами кластерного аналізу

то можна розглянути і вивчити реальні відстані між всіма сільськогосподарськими культурами в одиницях евклідової відстані. Вони становлять:

Case	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0.0	3.3	2.8	3.60	12.2	5.2
C_2	3.3	0.0	2.6	5.03	13.5	4.2
C_3	2.8	2.6	0.0	4.36	12.2	4.2
C_4	3.6	5.0	4.4	0.00	8.9	4.9
C_5	12.2	13.5	12.2	8.95	0.0	11.2
C_6	5.2	4.2	4.2	4.91	11.2	0.0

Видно, що змінні C_2 і C_3 дійсно знаходяться одна від одної на найменшій з усіх відстані.

ПРАКТИЧНА РОБОТА № 28

Мета роботи: Закріплення навичок проведення кластерного аналізу.

З метою закріплення навичок проведення кластерного аналізу проведіть групування 21 сорту ярої пшениці за п'ятьма ознаками. Ці сорти різного географічного походження вирощувались на Кинельській дослідній сільськогосподарській станції, з кожного географічного регіону були взяті три сорти. Вони позначені номерами. Для кожного сорту наведено: тривалість вегетаційного періоду в днях (VAR1), висота рослин в см (VAR2), вилягання в балах (VAR3), озерненість колосу в шт. (VAR4), маса 1000 зерен в г (VAR5).

	VAR1	VAR2	VAR3	VAR4	VAR5
Індія					
1	68	100	4	33	35
2	70	95	5	30	27
3	70	95	5	35	32
Пакистан					
4	75	95	3	27	35
5	72	105	3	30	30
6	70	100	3	30	30
Австралія					
7	72	110	4	30	30
8	75	115	4	30	34
9	78	115	5	32	29
Кенія					
10	70	110	5	27	35
11	78	100	5	35	24
12	70	100	5	30	32
Канада					
13	81	85	4	35	36
14	80	125	4	30	34
15	81	80	5	30	35
США					
16	82	120	3	32	37
17	80	100	5	25	35
18	75	115	3	30	37

Мексика					
19	75	85	5	25	40
20	78	95	5	25	39
21	70	75	5	25	37

1. Для проведення кластерного аналізу створіть із цих даних базу даних і збережіть її під оригінальною назвою.

2. Викличте модуль кластерного аналізу, послідовно виконавши команди, аналогічні до тих, що були в попередній роботі:

Analysis
Other statistics
Customize list
Cluster analysis
Replace
Switch to

3. У вікні, що відкрилось, оберіть опцію **Joining (tree clustering)** (процедура об'єднання – побудова дерева). Натисніть **OK**.

4. В підменю, що відкрилось, у вікні **Variables** оберіть опцію **Select All** (позначити все) для того, аби включити до аналізу всі змінні. Натисніть **OK**. У вікні **Input** залиште **Raw data** (вихідні дані), оскільки ви піддаєте кластеризації файл з реальними даними без будь-яких їх перетворень. У вікні **Cluster** слід зробити вибір – **Cases**, оскільки у вихідній таблиці сорти розташовані саме по строках. Якби тут була позначена опція **Variables**, то кластеризації були б піддані властивості сортів.

5. У двох нижче розташованих вікнах слід обрати опції **Single Linkage** і ще нижче – **Euclidean distances** (евклідова відстань) - як найбільш прості метрики.

OK

6. В новому вікні слід відзначити опцію: **Vertical icicle plot** (вертикальне дерево кластеризації).

OK

7. Одержали дерево кластеризації (рис. 59). Видно, що при різному географічному походженні чіткого групування в межах розглянутих сортів пшениці немає – всі вони за ознаками, що вивчаються, виявились досить схожі.

Найбільшу подібність мають сорти № 6 (Пакистан) і № 12 (Кенія), а також сорти № 8 (Австралія) і № 18 (США). Найбільш ізольований за комплексом властивостей від усієї групи сорт № 21 (Мексика).

8. Якщо активізувати опції

Continue,

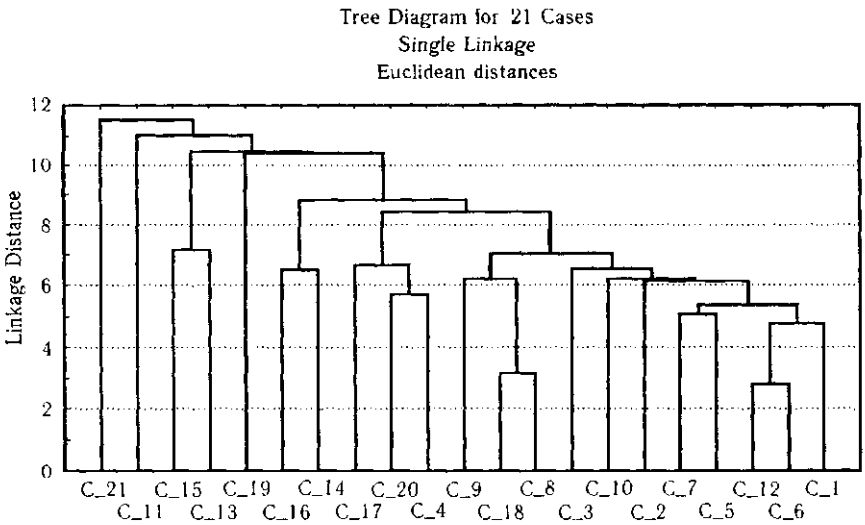


Рис. 59. Вертикальна дендрограма результату кластерного аналізу для даних практичної роботи № 28

а потім

Distance matrix,

то можна одержати і вивчити реальні відстані між кожним із сортів в одиницях евклідової відстані. Виконайте цю роботу і впевніться в подібності основних евклідових відстаней.

Задачі для самостійного розв'язання

Задача № 28.1. В посіві були одібрані 10 рослин і реєструвались їх відмінності за 5 ознаками. З'ясуйте, які рослини найбільш подібні між собою за цими ознаками? Вихідні дані:

	VAR1	VAR2	VAR3	VAR4	VAR5
1	15.62	96.8	198	16	.960
2	15.20	119.0	115	14	1.490
3	16.53	116.8	97	9	2.100
4	7.19	85.0	74	9	.098
5	11.42	83.0	135	12	.066
6	2.56	71.6	23	1	.040
7	17.02	102.0	12	16	.590
8	4.88	108.3	42	12	.104
9	8.47	91.0	107	19	.028
10	3.99	78.5	54	10	.076

Задача № 28.2. На 12 полях була врахована кількість 8 видів бур'янів (в шт./кв.м). Визначте, які поля були найбільш подібні по забур'яненості, настільки, що на них можна було б регулювати кількість бур'янів однаковими заходами. Вихідні дані:

	VAR1	VAR2	VAR3	VAR4	VAR5	VAR6	VAR7	VAR8
1	9.0	8.0	7.0	1.0	4.0	0.0	0.0	29.0
2	0.0	1.0	4.0	0.0	9.0	0.0	0.0	14.0
3	0.0	0.0	12.0	1.0	17.0	0.0	0.0	30.0
4	0.0	0.0	7.0	3.0	7.0	0.0	0.0	17.0
5	0.0	0.0	10.0	3.0	10.0	11.0	0.0	34.0
6	2.0	22.0	10.0	2.0	1.0	0.0	1.0	38.0
7	7.0	6.0	1.0	2.0	0.0	0.0	0.0	16.0
8	1.0	12.0	15.0	1.0	6.0	10.0	0.0	45.0
9	0.0	1.0	8.0	5.0	3.0	0.0	0.0	17.0
10	2.0	6.0	5.0	16.0	1.0	41.0	0.0	71.0
11	0.0	0.0	5.0	50.0	0.0	10.0	2.0	67.0
12	9.0	38.0	0.0	2.0	10.0	0.0	2.0	61.0

ПРАКТИЧНА РОБОТА № 29

Мета роботи: Оволодіння навичками аналізу рядів динаміки методом ARIMA.

Спочатку розглянемо простий ряд динаміки, в якому числа 1-21 можна розглянути як послідовні роки чи місяці, а відповідні їм значення в змінній VAR1 – як реальні значення якого-небудь явища. Завдання полягає в тому, щоб визначити, якими можуть бути значення VAR1 на 22, 23 і наступних етапах цього процесу, тобто скласти його прогноз.

1. Введіть в базу даних наведений нижче ряд і збережіть його під оригінальним ім'ям.

	VAR1
1	5.0
2	7.0
3	8.0
4	8.0
5	9.0
6	11.0
7	12.0
8	14.0
9	16.0
10	16.0
11	18.0
12	20.0
13	25.0
14	20.0
15	25.0
16	26.0
17	28.0
18	30.0
19	31.0
20	33.0
21	35.0

2. Уважно проаналізуйте цифри, що відображають динаміку ряду. На перший погляд здається, що це простий зростаючий ряд, який можна апроксимувати прямою лінією. Але більш ретельний аналіз з'ясує, що статистичний ряд має два невеликі плато: 8,8 і 16,16, а також плато з проваллям: 20, 25, 20, 25. Тому передбачити

примітивне зростання x_i-их на 22-му, 23-му і наступних ступенях ряду динаміки небезпечно. Потрібен більш адекватний метод.

3. Для реалізації комп'ютерного прогнозування методом ARIMA послідовно виконайте такі види робіт. Увійдіть в модуль "Аналіз рядів динаміки і прогнозування", виконавши такі процедури:

Analysis
Other statistics
Customize list
Time Series/Forecasting (Ряди динаміки/Прогнозування)
Replace
Switch to

4. У вікні, що відкрилось (рис. 60), проти клавіші **Variables** вже стоїть VAR1, бо у вашій базі лише один ряд. Якщо ви аналізуєте ряд динаміки із бази даних, де кілька змінних, то потрібну вам зміну ви повинні внести в список для аналізу звичайним способом.

5. Зверніть увагу, що праворуч від VAR1 в полі для змінних з'явилась літера **L (Lock – замок)**, що вказує на закритість змінної для довільних змін. Ви можете трансформувати ряд як завгодно, але вихідні дані будуть захищені від випадкових змін і до них завжди можна повернутись.

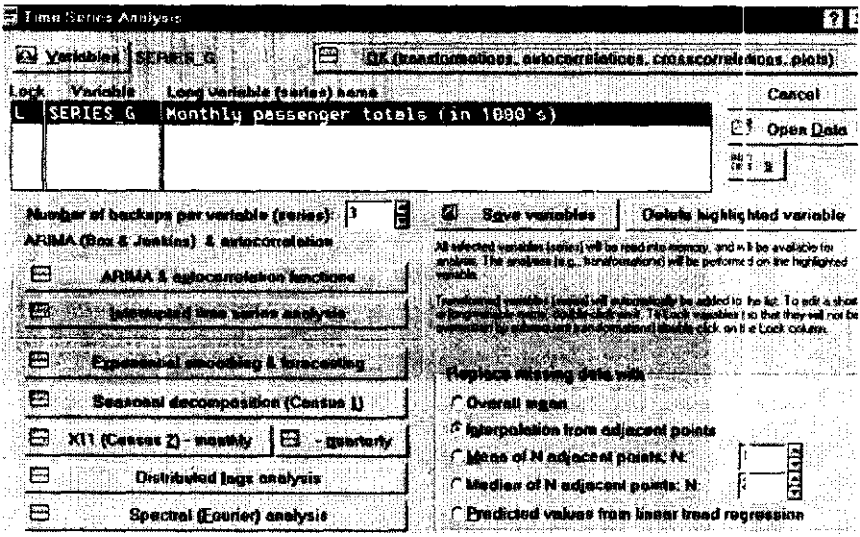


Рис. 60. Підмодуль для встановлення попередніх параметрів при прогнозуванні динаміки процесу методом ARIMA

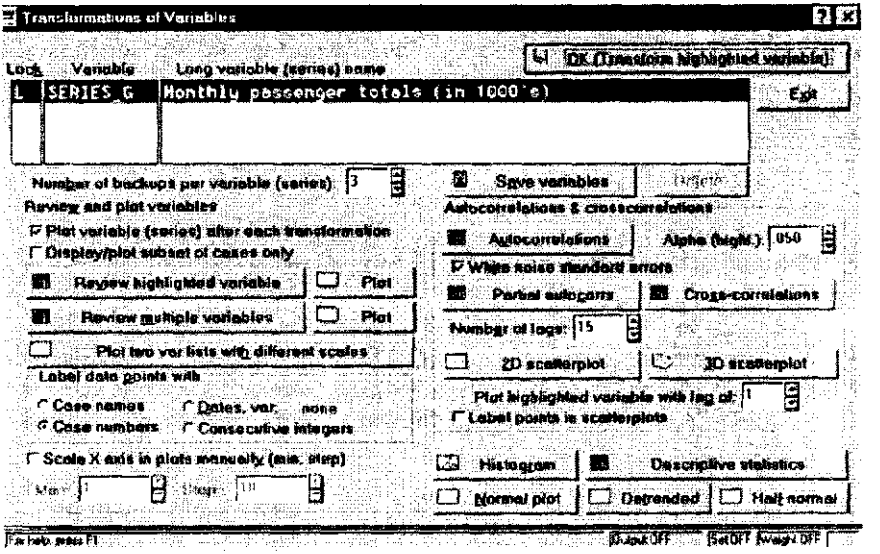


Рис. 61. Підмодуль для трансформації статистичного ряду при оцінці динаміки процесу

6. В правій нижній частині вікна є поле **Replace missing data with** (Заміщення даних, що випали). Річ у тім, що за певний період часу облікові дані можуть бути відсутні через різні причини (облік не був проведений чи втрачений, наприклад, для рядка 12 або іншого). Тоді тут можна обрати один з п'яти способів відновлення таких втрачених даних. Поки не накопичені теоретичні знання про такі способи, краще вибрати найпростіші і найнадійніші з них:

Interpolation from adjacent points (Інтерполяція за найближчими точками)

або

Mean of N adjacent points, N (Середнє з **N** найближчих точок)

і тоді як N ввести число 3 або 5.

В нашому ряду даних, що випали, немає і тому ніяких змін в полі **Replace missing data with** робити не слід.

7. Тепер уважно дослідимо ряд динаміки. Для цього активізуйте клавішу вгорі справа **OK (transformations...)**. Відкриється нове вікно (рис. 61). У верхньому його полі показана змінна, під її ім'ям знаходиться ряд динаміки. Як і раніш, це VAR1.

Перш за все необхідно подивитись графік вихідного ряду динаміки. В полі **Review and plot variables** (Огляд і графік для змінної)

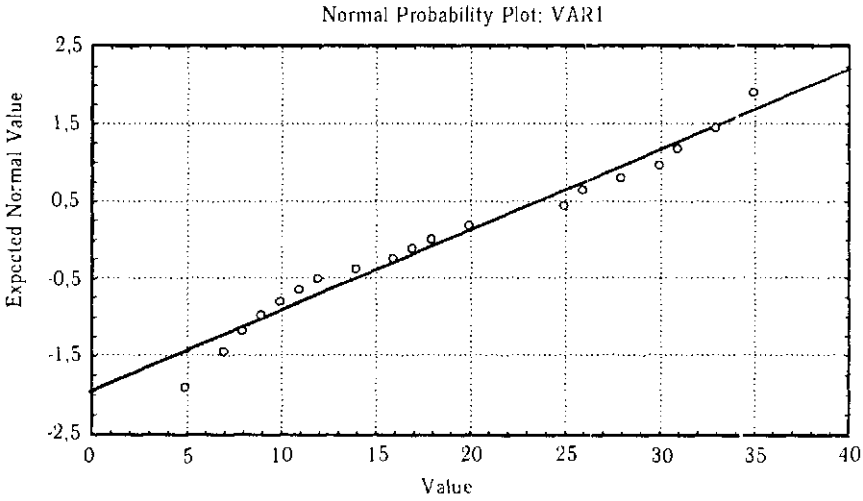


Рис. 62. Перевірка ряду динаміки на відповідність нормальному статистичному розподілу

знайдіть клавішу **Review highlighted variable** (Огляд для змінної, внесеної в заголовок). Натиснувши її, ви побачите числові значення ряду динаміки, а настинувши розташовану поруч клавішу **Plot** (Графік) – його графік. Видно, що графік має плоскі ділянки і провал.

Тепер перевірте, чи відповідає аналізований ряд динаміки нормальному статистичному розподілу. Для цього знайдіть і натисніть в самому низу вікна і в його середній частині клавішу **Normal Plot** (Нормальний імовірнісний аркуш). Ви побачите (рис. 62), що точки досить добре лежать на прямій лінії. Це означає, що перетворювати цей ряд динаміки не слід.

Далі необхідна перевірка автокореляцій досліджуваного ряду. Для виявлення автокореляцій послідовно активізуйте клавіші **Autocorrelations** (Автокореляції) та **Partial autocorrs** (Часткові автокореляції). Видно, що автокореляції присутні (рис. 63 та рис. 64). Як перші, так і другі із них виходять за межі 95%-го довірчого інтервалу (червоні штрихові лінії) і мають певну періодичність. Отже, для цього часового ряду при прогнозуванні необхідно врахувати автокореляційний фактор.

8. Дослідження ряду динаміки закінчене, а оскільки воно не потребує трансформації, то слід просто вийти з цього вікна, натиснувши клавішу **Exit** (верхня частина вікна, справа).

9. Ми повернулися до вікна **Time Series Analysis** (Аналіз ряду динаміки). В правій частині вікна є сім клавіш, кожна з яких

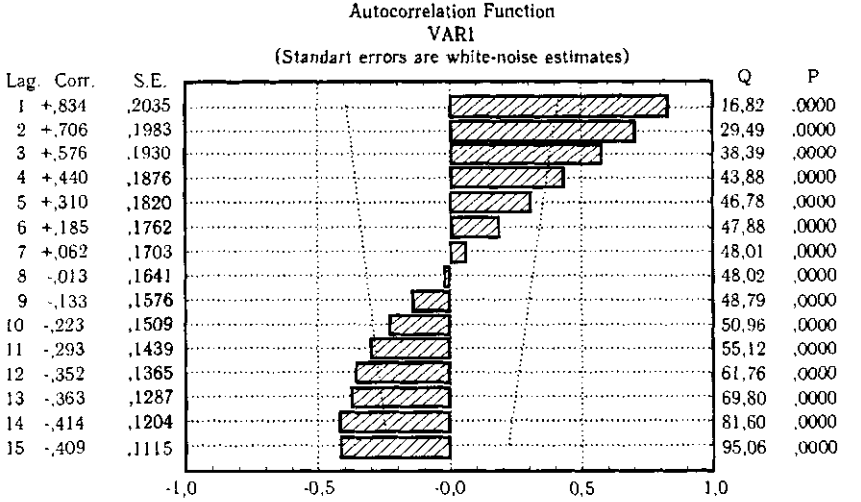


Рис. 63. Результати знаходження автокореляцій

відповідає одному з методів аналізу ряду і складення прогнозу. Скористаємося найбільш простим і надійним методом **ARIMA and autocorrelation functions**. Натисніть цю клавішу. Відкриється вікно **Single Series ARIMA** (Метод **ARIMA** для простих рядів динаміки) (рис. 65).

10. Це вікно надає широкі можливості для підбору адекватної моделі, за якою можна прогнозувати динаміку досліджуваного динамічного процесу.

Робота ведеться в основному з такими вікнами:

Estimate constant (Постійний член моделі)

Seasonal lag (Сезонний зсув)

p-Autoregressive (Авторегресія)

q-Moving aver. (Параметр для ковзаючої середньої)

P-Seasonal (Сезонність)

Q-Seasonal (Сезонний параметр для ковзаючої середньої).

Будемо "експериментувати" з цими вікнами, але це не експеримент наосліп, а ретельний підбір параметрів моделі на підставі особливостей ряду динаміки.

З урахуванням того, що часткова авторегресія (рис. 64) має найбільший член з лагом 1, виставимо у вікні **Seasonal lag** найменший можливий зсув – 2. Включимо до моделі постійний член, поставивши в його вікні хрестик клацанням ЛКМ. Проти чотирьох

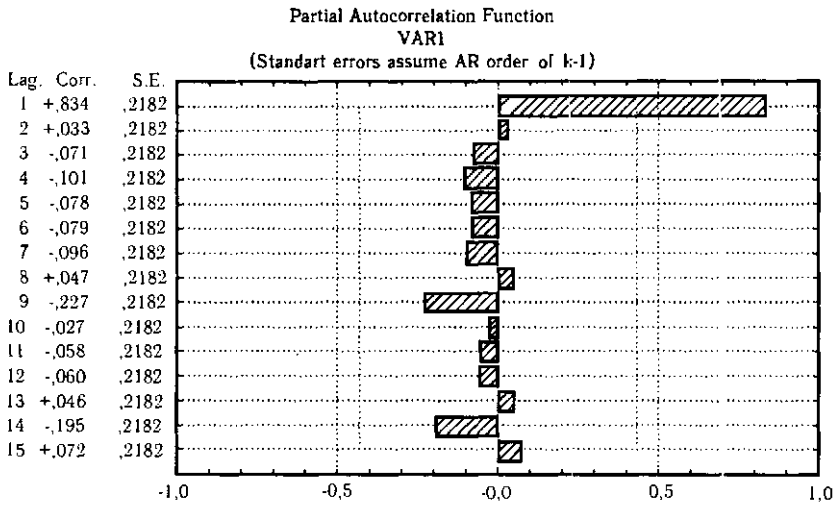


Рис. 64. Результати знаходження часткових автокореляцій

інших параметрів виставимо 1. Натисніть клавішу **OK (Begin parameter estimation)** (Почати визначення параметрів моделі).

11. На етапі розрахунків комп'ютер може декілька разів видавати попередження про те, що в стандартній процедурі обчислень є труднощі, і він шукає обхідний шлях для їх розв'язання. Натискайте в цих випадках клавішу **OK**. Іноді на проміжному етапі розрахунків комп'ютер повідомляє, що з цими параметрами закінчити обчислення неможливо. Тоді доведеться повернутись до початку, краще дослідити динамічний ряд і підібрати інші параметри моделі.

Коли розрахунок буде виконаний, з'явиться поле з клавішею **OK**.

12. Натисніть клавішу **OK**. Розгляньте її результати в верхній частині вікна. Тут найбільш важливі два рядки: рядок **Const., p(1), q(1), P_s, Q_s** і наступний рядок **Estimate** (Оцінки). Всі цифри подані синім кольором. Це означає, що статистично вони не достовірні.

13. Натисніть розташовану нижче клавішу **Parameter estimate**. Видно, що для знайдених параметрів в цій моделі навіть неможливо обрахувати довірчі рівні і статистичну надійність. Всюди стоять риски. Модель не адекватна процесу. Необхідно підбирати нову модель.

14. Натисніть клавішу **Exit** і поверніться до вікна **Single Series ARIMA**. Випробуйте інші числа в настроюванні моделі.

За умови терплячого і ретельного проведення аналізу, ви одержите таку комбінацію:

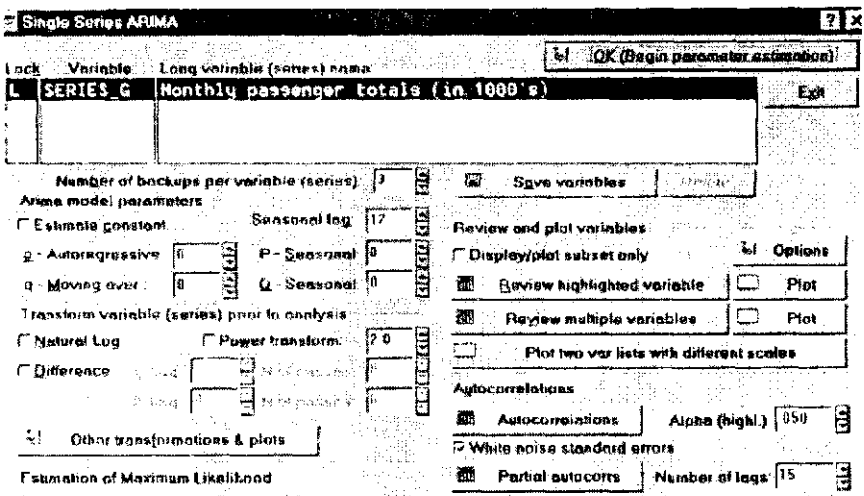


Рис. 65. Вікно визначення базових параметрів для ряду динаміки

Estimate constant X (хрестик)
Seasonal lag – 2
p – Autoregressive – 2
q – Moving aver. – 0
P – Seasonal – 0
Q – Seasonal – 0

Набравши цю комбінацію параметрів, проведіть розрахунок вже описаним способом. Завважте, що в цьому випадку комп'ютер не зіткнувся з труднощами в обчисленні і одразу знайшов параметри моделі.

15. У вікні результатів два параметри виділені червоним кольором – це означає, що вони статистично достовірні. Для більш ретельного вивчення моделі активізуйте клавішу **Parameter estimate**. Видно, що довірчий рівень для p(1) і P(2) дорівнює 0,0 – це 100%-ва надійність. Правда, вільний член моделі статистично недостовірний, але модель в цілому достовірна.

16. Перевіримо достовірність моделі, активізувавши клавіші **Autocorrelations** (Автокореляції) і **Partial autocorrs** (Часткові автокореляції). Дійсно, лише один лаг виходить за межі довірчого рівня, а часткові автокореляції й зовсім вкладаються в довірчий рівень повністю. Модель підібрано.

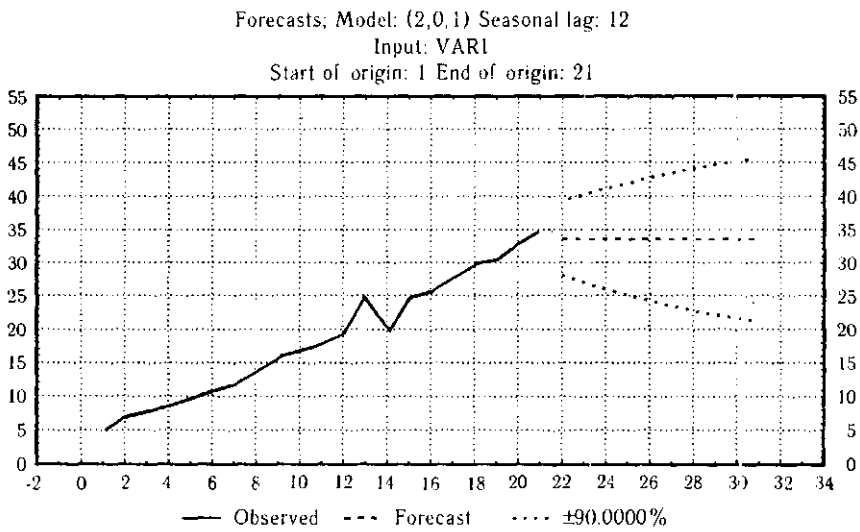


Рис. 66. Графічне подання динаміки статистичного ряду з прогнозом на п'ять ступенів з довірчими рівнями (штрихові лінії)

17. Закрийте допоміжні вікна і у вікні **Single Series ARIMA results** знайдіть клавіші: **Plot series and foresats** і **Forecast cases**. Перша з них покаже графік процесу і лінію прогнозу (червоного кольору) з 90%-ми довірчими інтервалами (вони позначені зеленими пунктирними лініями) (рис. 66).

Видно, що на наступних етапах процесу, для яких складався прогноз, можна очікувати стабілізацію процесу з 90%-вою надійністю, але залишається 10%-ва вірогідність того, що прогнозовані цифри будуть зростати чи зменшуватись. Скоріш за все, ви покладатиметеся на основний 90% прогноз.

Конкретні значення прогнозованих величин ви зможете прочитати, активізувавши клавішу **Forecast cases**. Тут для кожного кроку процесу аж до 31-го наведені можливі значення x_i та його межі.

В цілому процес, який за поверховою оцінкою може здатись зростаючим, має явну тенденцію стати стаціонарним, і його значення зафіксуються на 34-35-му етапах. Якби процес, що аналізується, являв собою, наприклад, вартість певних акцій на біржі, то найрозумніше було б вважати, що подальшого зростання ціни на ці акції не буде.

ПРАКТИЧНА РОБОТА № 30

Мета роботи: Закріплення навичок аналізу рядів динаміки методом ARIMA.

Для закріплення навичок проведення аналізу ряду динаміки і складання прогнозу проведіть аналіз урожайності озимої пшениці в одному з господарств за період в 21 рік. Складіть прогноз урожайності на найближчі 2-3 роки.

1. Введіть дані про величину врожайності і збережіть файл з оригінальним ім'ям. Ці дані такі:

	VAR1
1	35
2	27
3	32
4	35
5	30
6	30
7	30
8	34
9	29
10	35
11	24
12	32
13	36
14	34
15	35
16	37
17	35
18	37
19	40
20	39
21	37

2. Виконайте роботу повністю самостійно і звірте свої результати з наведеними нижче даними.

Перегляд вихідного ряду показує, що чіткої тенденції динамічний ряд не має, тому не доцільно проводити для нього регресійний аналіз з апроксимацією ряду просто прямою лінією.

Дані відповідають нормальному статистичному розподілу. Автокореляції виражені слабо, знаходяться в межах довірчого рівня.

Але в розподілі лагів є очевидна періодичність. Для зняття цієї періодичності пакет "Статистика" підказує трансформацію типу x-24. Вона виявляється ефективною.

В результаті перевірки ряду моделей з різними параметрами найбільш адекватною виявилась модель виду ARIMA (0, C, 1) з такими опціями:

Estimate constant (Постійний член моделі) – X (хрестик)
Seasonal lag (Сезонний зсув) – 2
p – autoregressive (Авторегресія) - 0
q – moving aver. (Параметр для ковзаючої середньої) - 1
P – Seasonal (Сезонність) - 0
Q – Seasonal (Сезонний параметр для ковзаючої середньої) – 0

Ця модель обчислюється без труднощів. Прогнозна урожайність знаходиться на рівні 9-10, але не можна забувати, що ряд був трансформований. Зворотня трансформація виконується за формулою x+24 і легко обчислюється. Після зворотнього перетворення виду 9+24 та 10+24 одержуєм 33 і 34. Це відповідає прогнозній урожайності в 33-34 ц/га зерна.

Цей прогноз може бути статистично надійним при збереженні діючих технологій і умов вирощування озимої пшениці.

Задачі для самостійного розв'язання

Задача 30.1. Наведені дані за 25 років про імміграцію в США. Складіть прогноз про величину імміграції в цю країну на наступні 5 років.

	VAR2
1	108721
2	147292
3	170570
4	188317
5	249187
6	205717
7	265520
8	170434
9	208177
10	237790
11	321625
12	326867
13	253265
14	260686
15	265398
16	271344

17	283763
18	260000
19	300000
20	290212
21	282565
22	199369
23	245899
24	199990
25	295858

Задача 30.2. Наведені дані про врожайність ячменю в Англії. Зробіть прогноз на найближчі 3 роки про величину врожаїв ячменю.

	VAR1	VAR2
1	1946	15.200
2	1947	16.900
3	1948	15.300
4	1949	14.900
5	1950	15.700
6	1951	15.100
7	1952	16.700
8	1953	16.300
9	1954	16.500
10	1955	13.300
11	1956	16.500
12	1957	15.000
13	1958	15.900
14	1959	15.500
15	1960	16.900
16	1961	16.400
17	1962	14.900
18	1963	14.500
19	1964	16.600
20	1965	15.100
21	1966	14.600
22	1967	16.000
23	1968	16.800
24	1969	16.800
25	1970	15.500

ВІДПОВІДІ НА ЗАДАЧІ ДЛЯ САМОСТІЙНОГО РОЗВ'ЯЗАННЯ

12.1. Середнє арифметичне для сорту Охтирчанка дор внює 5,76, дисперсія – 0,865, коефіцієнт варіації – 16,1%. Для сорту Миронівська 808 середнє арифметичне дорівнює 5,72, дисперсія – 0,513, коефіцієнт варіації – 12,5%. Таким чином, при практичнс однаковій довжині колосу він варіював більше у сорту Охтирчанка.

12.2. Для звичайного черв'яка середнє арифметичне дорівнює 9,0, дисперсія – 1,72, коефіцієнт варіації – 14,6%. Для каліфорнійського черв'яка середнє арифметичне дорівнює 10,9, дисперсія – 2,57, коефіцієнт варіації – 14,7%. Таким чином, при більшій середній довжині каліфорнійського черв'яка варіювання довжини їх тіл було майже однакове.

15.1. Коефіцієнт кореляції дорівнює +0,17. Він статистично не достовірний, бо p=0,459 і нульову гіпотезу про відсутність кореляції відкинути на можна.

15.2. Коефіцієнт кореляції дорівнює +0,269. Він статистично не достовірний, бо p=0,452 і нульову гіпотезу про відсутність кореляції відкинути не можна.

19.1. F=17,8 при p=0,0001.

Вплив мінеральних добрив та сидератів на цукровий буряк статистично достовірний на рівні 99,9%. Критерій Шеффе для порівняння варіанта 1 з варіантом 2 дорівнює 0,015, для 1 з 3 – 0,000, а це свідчить, що мінеральні добрива та сидерати достовірно збільшують урожай.

Середні по варіантах:

1 – 307,5

2 – 352,5

3 – 384,3

Критерій Дункана (HIP)=32,2, і за цим критерієм всі варіанти достовірно відрізняються від контролю. При вирощуванн цукрового буряку доцільно використовувати мінеральні добрива та сидерати, останні забезпечували навіть більше зростання врожаю.

19.2. F=3,01 при p=0,060. Критерій Шеффе для всіх варіантів обробки насіння нижчий 95% рівня достовірності.

Середні за дослідженням:

контроль – 89,2

варіант 1 – 93,8
варіант 2 – 91,8
варіант 3 – 93,4

Критерій Дункана (НІР)=3,88, за цим критерієм варіанти 1 і 3 достовірно відрізняються від контролю. Для них р складає відповідно 0,02 і 0,03. Це свідчить: стимуляція насіння є перспективною і необхідні подальші досліди по її вивченню.

19.3. F=27,77 при p=0,0001. Оброблення грунтів статистично достовірно на рівні 99,9% впливає на кількість бур'яну в посіві.

Критерій Шеффе для варіанта 1 з 3 дорівнює 0,0005, 2 з 3 – 0,0003.

Середні по варіантах:
1 – 394,7
2 – 413,0
3 – 86,7

Критерій Дункана (НІР)=116,2, за цим критерієм явно виділяється меншою засміченістю третій варіант – з боронуванням по сходах.

В цілому передпосівні та післяпосівні культивації виявились ефективними лише при поєднанні з післясходовим боронуванням. Слід вивчити можливості пригнічення щириці одним боронуванням по сходах.

21.1. Вид породи дерева: F=1,17, p=0,341
Ступінь освітленості: F=15,91, p=0,0017
Взаємодія порода-освітленість: F=10,99, p=0,0019

Вид породи дерева на відкладання яєць не впливає, має значення лише освітленість та освітленість + вид породи дерева.

За критерієм Шеффе відрізняються достовірно варіанти:
B2 з A1 – 0,013
B2 з B1 – 0,003
B2 з C2 – 0,033
Критерій Дункана (НІР) = 55,0

Таким чином, мінімальна кількість відкладених яєць на клені в затінку. Інші варіанти суттєвої різниці не мають. Очевидно, відкладання яєць відбувається переважно на освітлених місцях.

22.2. Сорт F=1,50, p=0,255
Умови року: F=0,166, p=0,693
Взаємодія: F=0,166, p=0,693

І сорт томата, і умови року достовірного впливу на урожайність не мали.

За критерієм Шеффе достовірно відмінні варіанти не виокремлені.

Критерій Дункана (НІР)=1,42.

Групові середні:
Сорт А, 1997 р. – 2,3
Сорт А, 1998 р. – 2,3
Сорт Б, 1997 р. – 2,6
Сорт Б, 1998 р. – 3,0

Таким чином, польовий дослід не з'ясував достовірного впливу на врожай сорту, умов року та їхньої взаємодії. При розгляді середніх можна вважати, що сорт Б дещо більш врожайніший, але для доказу цієї гіпотези потрібні додаткові досліди.

24.1. Рівняння регресії має вигляд VAR1=-0,160+0,47VAR2. RI=0,92, таким чином, рівняння охоплює 92% загальної дисперсії. F=104,8 при p= 0,000007. Це результат високої статистичної достовірності. Залишки не мають регулярного характеру. Висновок: вміст фосфору в грунті залежить від pH з достовірністю більше 99% і може прогнозуватись і обчислюватись за знайденним рівнянням регресії.

24.2. Рівняння регресії має вигляд VAR2=66,05-7,39VAR1. RI=0,645. F=14,5 при p<0,0051. Якість тіста достовірно змінюється (покращується) при збільшенні строку прогрівання муки.

26.1. Підсумкова модель має вигляд:

Var 15=0,25-0,085 Var5 – 0,35 Var14,

де Var5 – суха вага листя, Var14 – суха вага квітів. Модель охоплює 74,4% загальної дисперсії. Критерій Фішера для неї дорівнює 34,0 при p=0,000, тобто вона має 100%-ву достовірність. Таким чином, репродуктивне зусилля рослин звіробою визначають маса листя і маса квітів на рослині. Інші параметри є другорядними.

26.2. Підсумкове рівняння регресії має вигляд:

VAR10=1,964-2,36VAR3

Для цього рівняння RI=0,79, F=46,11 при p<0,0002. Це означає, що захворювання дизентерією залежить перш за все (на 79%) від дотримання правил особистої гігієни. Надійність цього висновку більша за 99%.

28.1. Найбільш подібними виявились рослини номер C_3 і C_2 (евклідова відстань дорівнює 19,0) і рослини C_10 та C_4 (евклідова відстань дорівнює 21,0).

28.2. Найбільш подібними є поля C_9 та C_4 (евклідова відстань дорівнює 4,7). До них наближаються поля C_2, C_7 і C_1 (евклідова відстань не перевищує 18,6). На цих полях можливе використання однакової технології боротьби з бур'янами.

30.1. Модель

Estimate constant (Постійний член моделі) – X (хрестик)

Seasonal lag (Сезонний зсув) – 2

p – autoregressive (Авторегресія) - 1
q – moving aver. (Параметр для ковзаючої середньої) - 0
P – Seasonal (Сезонність) - 0
Q – Seasonal (Сезонний параметр для ковзаючої середньої) – 0
Прогноз – зниження імміграції в найближчі роки з 290000 до 250000.

30.2. Модель
Estimate constant (Постійний член моделі) – не заповнювати
Seasonal lag (Сезонний зсув) – 2
p – autoregressive (Авторегресія) - 1
q – moving aver. (Параметр для ковзаючої середньої) - 0
P – Seasonal (Сезонність) - 0
Q – Seasonal (Сезонний параметр для ковзаючої середньої) – 1
Прогноз - урожайність 15,4-15,8 ц/га.

АНГЛО-УКРАЇНСЬКИЙ СЛОВНИК ОСНОВНИХ ТЕРМІНІВ, ЯКІ ВИКОРИСТОВУЮТЬСЯ В ПАКЕТІ «СТАТИСТИКА»

A

add - додати
additive - адитивный (який одержується шляхом додавання)
adjust margins - точні краї (поля аркуша паперу)
adviser - порадник (довідка по виконуваних процедурах)
all - усе, все
all effects - усі эфекти
all Specs - усі специфікації (настройки)
analysis - аналіз
analysis of variance - дисперсійний аналіз
analysis of component - компонентний аналіз
ANOVA - однофакторний дисперсійний аналіз
approximate - наближений
area - площа
ARIMA - модель проінтегрованої ковзаючої середньої
ARIMA Results - результати ARIMA
autocorrelation - автокореляція
autocorrelations of residuals - автокореляція залишків
autoregressive model - модель авторегресії
autoregressive moving average model - модель авторегресії ковзаючої середньої
autoregressive process - процесс авторегресії

B

bar/column plots – стовпчикові діаграми
basic statistics – основні статистичні методи
begin parameter estimation – розпочати обчислення параметрів
between – між, поміж
bias - зміщення
boundary – межа, границя
box & whisker plot – графік у вигляді «ящика з вусами»
brushing tool – інструмент «Пензлик»

C

cases - випадок (повторення)
chi-square - хі-квадрат
clear - очистити
cluster analysis - кластерний аналіз
coding - кодування
coefficient - коефіцієнт
coefficient of variation - коефіцієнт варіації
comparison - порівняння
compute - обчислювати
confidence band - довірча зона
confidence level - довірчий рівень
confidence limits for means - довірчі рівні для середніх
confidence interval - довірчий інтервал
contingency - сполученість
continue - продовжити
copy - копіювати, копія
correlation coefficient - коефіцієнт кореляції
correlation matrix - матриця коефициєнтів кореляції
create new file - створити новий файл
current Specs - поточні специфікації (настройки)
curve fitting - підбір кривої
customize list - список користувача
cut - вирізати

D

data management - управління даними
data value - числовое (набране цифрами) значення
decrease column width - зменшити ширину колонки
decrease decimal - зменшити кількість знаків після коми
degrees of freedom (d.f.) - ступені свободи
delete - знищити, стерти
dependent - залежний
descriptive statistics - описові статистичні методи
design - схема, план
detailed - детальний
detrend - видалення тренда
difference - відмінність
differencing - різниця
digits - цифра
display - дисплей, показати

distance - відстань
distance matrix - матриця відстаней
distributed lags analysis - аналіз розподілених лагів
distribution - розподіл
double - подвійний
down - униз

E

edit - редактор, редагування, редагувати
edit scale values - редагувати масштаб
effect - ефект
efficiency - ефективність
error - помилка
error of mean square (EMS) - середньоквадратична похибка
error of prediction - похибка прогнозу
estimate - оцінювати, оцінка
estimation - оцінювання
estimation of maximum likelihood - оцінювання методом максимальної правдоподібності
exit - вихід
exponential smoothing - експоненціальне згладжування
exponential trend - експоненціальний тренд
export data - експорт даних
expression - рівняння, запис

F

F-test - F-тест (критерій Фішера)
file - файл
filtering - фільтрація
find - знайти
find what - знайти що
first - перший
fitting - підбір, підгін
forecasting - прогнозування
forecasting error – похибка прогнозування
forecasting by exponential smoothing - прогнозування методом експоненціального згладжування
frequency tables - таблиці частот
function - функція

G

graph - графік
graph mapping option - настройка положення графіка (у вікні, на папері)
graph type - тип графіка
graphic text editor - редактор написів на графіку

H

half-normal probability plot - графік на напівнормальному аркуші
help - допомога
hypothesis - гіпотеза

I

import data - імпорт даних
include - включати
increase column width - збільшити ширину колонки
increase decimal - збільшити кількість знаків після коми
independent - незалежний
independent variable - незалежна змінна
input file - вхідний файл
interaction - взаємодія
insert after - вставити після
intercept - вільний член (рівняння)
interpolation - інтерполяція
interpolation from adjacent points - інтерполяція по найближчим точкам
interrupted ARIMA - перервана ARIMA (ARIMA з інтервенцією)
interval estimate - інтервальна оцінка
inverse - зворотний
irregular - нерегулярний

J

joining - об'єднання

K

kurtosis - ексцес

L

label - мітка
lag - лаг, зсув, зміщення, затримка