Б. С. Незамай, В. В. Бачук

ЕМПІРИЧНІ МЕТОДИ ПРОГРАМНОЇ ІНЖЕНРІЇ

ЛАБОРАТОРНИЙ ПРАКТИКУМ

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Івано-Франківський національний технічний університет нафти і газу

Кафедра математичних методів в інженерії

Б. С. Незамай, В. В. Бачук

ЕМПІРИЧНІ МЕТОДИ ПРОГРАМНОЇ ІНЖЕНРІЇ

ЛАБОРАТОРНИЙ ПРАКТИКУМ

Івано-Франківськ
2016

УДК 004.42: 519.25
ББК 32.973.2–018
Н-44

Рецензент:
Олійник А. П. доктор технічних наук, професор, завідувач кафедри математичних методів у інженерії Івано-Франківського національного технічного університету нафти і газу

Рекомендовано методичною радою університету
(протокол № 13 від 26.02.2016 р.)

Незамай Б.С., Бачук В. В.

Н-44 Емпіричні методи програмної інженерії: лабораторний практикум. – Івано-Франківськ: ІФНТУНГ, 2016. – 44 с.

МВ 02070855-10372-2016

Лабораторний практикум містить методичні вказівки для проведення лабораторних занять з дисципліни «Емпіричні методи програмної інженерії». Розроблений відповідно до робочої програми навчальної дисципліни.

Призначений для підготовки бакалаврів за напрямом 6.050103 – "Програмна інженерія"

УДК 004.42: 519.25
ББК 32.973.2–018

МВ 02070855-10372-2016

ЗМІСТ

ВСТУП

Пропонований практикум містить завдання та короткі теоретичні відомості з дисципліни “Емпіричні методи програмної інженерії”.

Програмна інженерія — це застосування системного, вимірюваного підходу до розроблення, використання та супроводу програмного забезпечення, та дослідження цих підходів, тобто застосування принципів інженерії до програмного забезпечення.

Як і інші інженерні дисципліни, інженерія програмного забезпечення характеризується такими аспектами:

- творчість - інженерія концентрується на проблемах аналізу і проектування;
- інструментальність ключові проблеми в інженерії - це вибір і використання інструментів;
- стандартизація - кращі практичні досягнення інженерії у вигляді інженерних принципів є основою створення стандартів;
- повторне використання - в інженерії повторне використання знань і продуктів фаз життєвого циклу є найважливішим чинником підвищення продуктивності і якості;
- професіоналізм - інженерія програмного забезпечення - це професія.

Емпіричні методи – це методи роботи, методи організації роботи у галузі програмної інженерії. Часто ці методи базуються на методах математичної статистики, або ними обґрунтовується їх ефективність. Крім того, вже класичним є вислів про те, що для того, щоб навчитись програмувати – треба програмувати. Завдання, які лягли в основу даних лабораторних робіт, є завданнями програмної реалізації методів математичної статистики.

Даний лабораторний курс заплановано авторами до подальшого розроблення та вдосконалення.

ЛАБОРАТОРНА РОБОТА № 1

Тема: найпростіша статистична обробка емпіричних даних

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

У лабораторних роботах ми не маємо можливості працювати з якими-небудь дійсними даними, тому необхідною складовою частиною кожної роботи буде генерація вибірки. В основі даної процедури лежить робота з генератором базової випадкової величини.

Базовою випадковою величиною (БВВ) у статистичному моделюванні називають неперервну випадкову величину z, рівномірно розподілену на інтервалі [0, 1].

БВВ моделюється на ЕОМ за допомогою генераторів БВВ. Генератор БВВ – це пристрій або програма, що видає за запитом одне або кілька незалежних значень $z_1,\ldots,z_n$ БВВ.

Генератори БВВ бувають трьох типів: табличні, фізичні та програмні.

Табличний генератор БВВ – це таблиця випадкових чисел. Основний недолік такого генератора полягає в обмеженій кількості випадкових чисел у таблицях.

Фізичний генератор БВВ – це спеціальний радіоелектронний пристрій в ЕОМ, що містить джерело шуму. Шум перетворюється на випадкові числа з рівномірним розподілом. Недоліками є складність тиражування (необхідна додаткова плата) та схемна нестабільність такого генератора.

Програмний генератор БВВ, зазвичай, обчислює значення $z_1,\ldots,z_n$ за рекурентною формулою типу $z_i = f(z_n)$ у разі заданого стандартного значення z_0. Оскільки z_0 повністю визначає всю послідовність, її називають псевдовипадковою. Але її статистичні властивості повністю відповідають властивостям ”справді випадкової”.

Найпростіший метод програмної генерації випадкових чисел – мультиплікативний; у його основі лежить таке рекурентне співвідношення:

$$z_i = (A \cdot z_{i-1} + C) \bmod M,$$

де z_i, z_{i-1} – чергове і попереднє випадкові числа відповідно; А, С – константи; М – велике ціле позитивне число (чим більше М, тим довша неповторювана послідовність).

Дискретна випадкова величина d задається множиною можливих значень $\{d_1, d_2, \dots, d_n\}$ та їх ймовірністю $p_1, \dots p_n$. У випадку, коли результати рівноймовірні ($p_1 = \dots = p_n = 1/n$) алгоритм наступний:

$$d = [nz] + 1, \qquad (1.1)$$

де [ ] – операція обчислення цілої частини числа;

z – базова випадкова величина

Емпіричні дані, що отримані шляхом вимірювання властивостей вибіркових об'єктів – чи отримані методом імітаційного моделювання, як описано вище – повинні пройти первинну обробку. Під первинною обробкою найчастіше розуміють внесення у табличні форми (табуляцію), впорядкування у варіаційні послідовності (або ряди), групування (при побудові інтервального варіаційного ряду), побудова статистичного розподілу, обчислення окремих простих статистичних параметрів (статистик).

Варіаційний ряд – це впорядкована за збільшенням або зменшенням послідовність значень досліджуваної змінної.

Статистичний розподіл – це математична модель у вигляді співвідношення значень змінної, що характеризує властивості вибірки, до частот їх появи. Тобто, стовпець Варіанти та Кількість варіантів утворюють статистичний розподіл – див. таблиця 1. Практично використовують також накопичені, або інтегральні частоти. Крім того, часто використовують не частоту варіантів, а долю в сумі всіх частот, яка рівна відношенню частоти варіанта до загального числа спостережень:

$$\omega_x = \frac{m_x}{n}, \qquad (1.2)$$

де ω_x – частість; у більшості випадків також називають частотою;

m_x – кількість варіанту x;

n – кількість елементів у вибірці.

Використовується також накопичена (інтегральна частість).

Таблиця 1 – Обробка емпіричних даних

Дані		Варіаційний ряд	Варіанти	Частота	Накопичена частота
j	x_j	x_j^*	x_i	m_i	$m_i^{нак}$
1	3	1	1	1	1
2	4	2	2	1	2
3	4	3	3	2	4
4	4	3			
5	3	4	4	5	9
6	2	4			
7	4	4			
8	4	4			
9	5	4			
10	1	5	5	1	10

Хід роботи

Написати програму (заборонено використовувати пакети типу Mathcad, MathLab, Maple, і т. п.), що генерує випадкову послідовність даних з $(N+10)$ елементів, які набувають значень із набору (1, 2, 3, 4, 5). Тут N – номер студента у журналі старости. Одержати та вивести на екран вихідні дані, варіаційний ряд, статистичний розподіл, інтегральну частоту та частість.

Рекомендовано максимально використовувати стандартні бібліотеки та алгоритми.

Звіт повинен містити схему алгоритму сортування (що відповідає використаному алгоритму), код програми з необхідними коментарями, зображення результату роботи програми з екрану, висновки.

Контрольні питання.

1. Що таке статистичний розподіл?
2. Що таке варіаційний ряд?
3. Методи побудови генераторів базової випадкової величини.
4. Що таке генеральна сукупність?
5. Які існують методи вибірок?
6. Що таке базова випадкова величина?
7. Методи одержання базової випадкової величини.

ЛАБОРАТОРНА РОБОТА № 2

Тема: обчислення середніх величин.

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

Статистичні показники, які розкривають властивості вибірки, можна представити такими основними групами:

– емпіричними розподілами, що характеризують структуру досліджуваної області;

– вибірковими показниками (мірами центральної тенденції та мінливості);

– кореляційно-регресійними показниками, які дають можливість встановити приховані взаємозв'язки та закономірності явищ, спрогнозувати розвиток досліджуваних процесів.

Мірами центральної тенденції називають чисельні показники типових властивостей емпіричних даних. Існує невелика кількість таких показників-мір і основними з них є мода, медіана, середнє арифметичне.

Мода Мо – це значення, яке найчастіше трапляється серед емпіричних даних. Так, для ряду значень 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5 мода дорівнює 3.

При визначенні моди дотримуються наступних домовленостей (матеріал, наведений нижче, не є вичерпним, пропонується для виконання роботи обмежитись даними варіантами):

– мода може бути відсутня, наприклад, для даних 2, 2, 3, 3, 4, 4, 5, 5;

– якщо варіанти суміжні і мають однакову частоту, мода визначається як середнє значення сусідніх варіант; наприклад, для ряду 2, 2, 3, 4, 4, 4, 5, 5, 5 мода рівна 4.5;

– якщо варіанти несуміжні, може існувати декілька мод. Так, для даних 2, 2, 3, 3, 3, 4, 5, 5, 5 характерна бімодальність, тобто дві моди Мо1=3, Мо2=5.

Медіана Md – це значення, яке проходить через середину упорядкованої послідовності емпіричних даних. Для непарної кількості даних медіана визначається середнім елементом. Наприклад, для 11-ти значень 4, 4, 4, 5, 5, 5, 5, 5, 6 , 6, 7 медіана Md=5. Якщо кількість значень даних є парною, то медіаною є середнє значення центральних сусідніх елементів $Md = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$. Наприклад, для 12 значень 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 7, 7, медіана Md=(5+6)/2=5.5.

Середнє арифметичне сукупності n значень дорівнює

$$\bar{x} = \frac{x_1 + x_2 + \dots x_n}{n}. \qquad (2.1)$$

Середнім гармонічним чисел називають число, обернене середньому арифметичному їх обернених, тобто

$$A_{-1}(x_1, \dots x_n) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots \frac{1}{x_n}}. \qquad (2.2)$$

Середнім квадратичним двох чисел називають число, рівне квадратному кореню з середнього арифметичного квадратів двох чисел

$$S = \sqrt{\frac{x_1^2 + \dots + x_n^2}{n}}. \qquad (2.3)$$

Середнім геометричним додатних чисел називають таке число, яким можна замінити кожне з цих чисел так, щоб їх добуток не змінився

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}. \qquad (2.4)$$

Середне геометричне двох чисел $\sqrt{a_1 \cdot a_2}$ називають їх середнім пропорційним.

Хід роботи

Написати програму (заборонено використовувати пакети типу Mathcad, MathLab, Maple, і т. п.), що генерує випадкову послідовність даних з $(N+10)$ елементів, які набувають значень із набору (1, 2, 3, …., (N+1)). Тут N – номер студента у журналі старости. Розподіл рівномірний.

Вивести початкову та впорядковану послідовність.

Обчислити моду, медіану, середнє арифметичне сукупності. Обчислення кожної з цих величин має бути реалізоване у вигляді окремого блоку програми (функції/процедури/класу/методу класу залежно від використовуваного середовища для виконання робіт).

Написати програму, яка перевіряє розроблений раніше елемент для обчислення моди. Для цього сформувати таблицю прикладів, які відображають різні варіанти визначення моди та результати при цьому. Здійснити перевірку. Протокол еревірки в зрозумілій формі має відображатись на екрані (і в звіті). До звіту внести як результат першої перевірки (з можливими наявними помилками), так і виправлений варіант.

Зробити висновки.

Контрольні питання

1 Мода.
2 Медіана.
3 Поняття про середнє степеневе.
4 Середнє гармонічне.
5 Середнє арифметичне.
6 Середнє квадратичне.
7 Середнє геометричне.
8 Що таке контроль якості?
9 Визначити поняття емпіричний.
10 Методи моделювання дискретних випадкових величин.
11 Що таке Unit-tests?

ЛАБОРАТОРНА РОБОТА № 3

Тема: основи статистичної обробки неперервних даних.

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

Під час моделювання неперервних випадкових величин із заданим законом розподілу можна використовувати три методи:

– метод нелінійних перетворень;
– метод композицій;
– табличний метод (метод "звернення").

Перші два методи не розглядатимемо, оскільки вони потребують достатньо серйозної математичної підготовки. Третій метод заснований на заміні закону розподілу неперервної випадкової величини спеціальним розрахунковим співвідношенням, яке дає змогу обчислювати значення випадкової величини за значенням випадкового числа, рівномірно розподіленого на інтервалі $(0,1)$. Наведемо співвідношення для показникового та нормального законів розподілу.

Показниковий (експотенційний) розподіл

$$x = -\frac{1}{\lambda} \cdot \ln(z), \tag{3.1}$$

де $\lambda > 0$ – параметр показникового розподілу,
z – рівномірно розподілене випадкове число.

Розрахункове співвідношення для випадкового числа нормального розподілу :

$$x = m + s\left[\sum_{i=1}^{12} z_i - 6\right], \tag{3.2}$$

де m,s – параметри нормального закону розподілу (математичне очікування і середньоквадратичне відхилення); z_i – рівномірно розподілене випадкове число.

Одже, значення, які приймає досліджувана ознака, можуть відрізнятись на як завгодно малу величину. В такому випадку для аналізу значення ознаки групують по інтервалах.

Таблицю, що дозволяє судити про розподіл частот між інтервалами варіювання значень ознаки, називають інтевальним варіаційним рядом. Для побудови інтервального варіаційного ряду необхідно визначити величину інтервалу, встановити шкалу інтервалів, згрупувати результати спостережень.

Для побудови оптимальної величини інтервалу використовують емпіричну формулу Стерджеса:

$$h = \frac{X_{\max} - X_{\min}}{1 + 3.322 \cdot \lg(n)}. \qquad (3.3)$$

Під оптимальною величиною інтервалу розуміють таку величину, при якій побудований варіаційний ряд не буде надто громіздким, але дозволить виявити особливості вибірки. Якщо в результаті обчислень *h* – дробове число, то рекомендується використовувати або близький нескладний дріб, або для величин, що описуються значеннями набагато більшими за 1 близьке ціле число.

Існують загально прийняті способи графічного відображення варіаційних рядів. Використовуються наступні варіанти графічного зображення:

- полігон – як правило використовується лише для зображення дискретного варіаційного ряду. Для його побудови у прямокутній системі координат наносять точки з координатами: (Xi;mX) і з'єднують їх прямими відрізками;

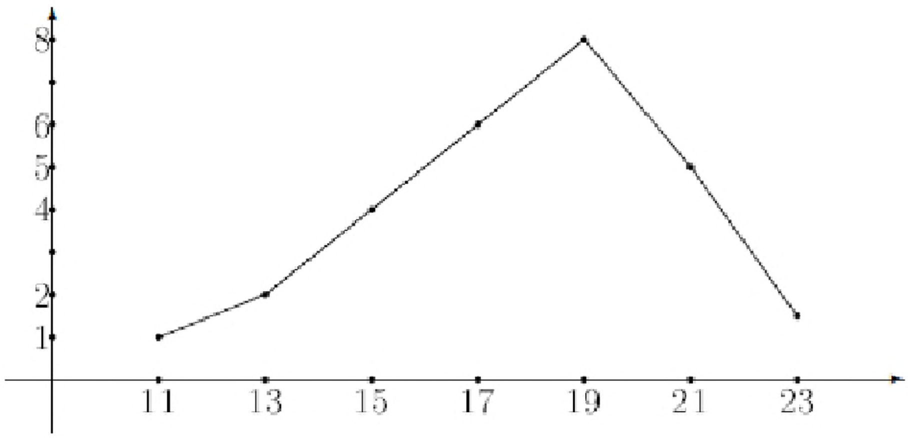

Рисунок 3.1 – Приклад полігона

- гістограма – використовується для зображення інтервального варіаційного ряду. Для її побудови у прямокутній системі координат по осі абсцис (X) відкладають відрізки, що відображають інтервали. На цих відрізках будують прямокутники з висотами рівними частотам або частотностям інтервалів. Іноді інтервальний ряд зображають полігоном. В цьому випадку інтервали замінюють їх серединами.

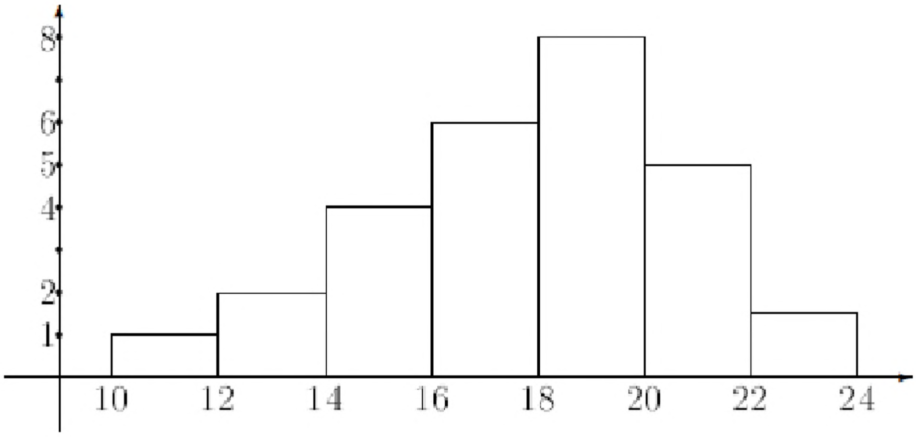

Рисунок 3.2 – Приклад гістограми

- кумулятивна крива – крива накопичення частот або частотностей.
- огіва – якщо в кумулятивної кривої змінити місцями осі координат, то одержана крива називається огівою.

Хід роботи

Задатись двома вибірками із 20N елементів неперервних випадкових величин нормального та показникового розподілу в діапазоні від 0 до N включно. Побудувати гістограми. Для побудови графіки рекомендовано використовувати стандартні бібліотеки.

Контрольні питання

1. Який разподіл називається рівномірним?
2. Який розподіл називається нормальним?
3. Густина розподілу імовірності випалкової величини з нормальним розподілом.
4. Методи графічного зображення варіаційних рядів.
5. Емпірична формула Стерджеса.
6. Які бібліотеки для побудови графіки використовуються з мовою (середовищем) програмування, яке Ви використовуєте?
7. Як розуміти поняття оптимальності для задачі визначення величини інтервалу?
8. Що таке integration tests?
9. Методи моделювання неперервних випадкових величин.

ЛАБОРАТОРНА РОБОТА № 4

Тема: обчислення дисперсії, середнього квадратичного відхилення, коефіцієнтів асиметрії та ексцесу

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

Для оцінювання розміру варіації використовується система абсолютних показників, які розглядаються як абсолютна міра варіації.

Розмах варіації (R) характеризує максимальну амплітуду коливань значень ознаки в сукупності:

$$R = X_{\max} - X_{\min}, \tag{4.1}$$

де: $X_{\max}, X_{\min}$ – відповідно найбільше та найменше значення ознаки.

В інтервальних рядах розподілу розмах варіації визначається як різниця між верхньою межею останнього та нижньою межею першого інтервалу.

Середнє лінійне відхилення ($\bar{l}$), що характеризує середній розмір відхилень значень ознаки від середнього рівня. Для розрахунку за індивідуальними даними використовують середнє лінійне відхилення просте:

$$\bar{l} = \frac{\sum\left|X - \overline{X}\right|}{n}, \tag{4.2}$$

де X – індивідуальні значення ознаки; $\overline{X}$ – середнє значення ознаки; n – кількість одиниць у сукупності.

Дисперсія (σ^2) – це середній квадрат відхилень значень ознаки від її середнього рівня. Для розрахунку за індивідуальними даними використовують дисперсію просту:

$$\sigma^2 = \frac{\sum(X - \overline{X})^2}{n}. \tag{4.3}$$

Середнє квадратичне відхилення (σ) — показує середній розмір відхилень значень ознаки від середнього рівня.

Середнє квадратичне відхилення найчастіше використовують у статистичному аналізі, тому його також називають стандартним відхиленням.

Середнє арифметичне та дисперсія варіаційного ряду є частковими випадками більш загального поняття про моменти варіаційного ряду. Початковим моментом порядку q називають середнє арифметичне q степені варіантів:

$$\tilde{v}_q = \bar{x}^q = \frac{\left(\sum_x x^q \cdot m_x\right)}{\sum_x m_x}, \tag{4.4}$$

де m_x – частота варіанта.

Відповідно до цієї формули початковим моментом нульового порядку буде 1. Початковий момент першого порядку буде рівний середньому арифметичному.

Центральним моментом порядку q називають середнє арифметичне q степенем відхилень варіантів від їх середнього арифметичного:

$$\tilde{m}_q = \frac{\sum_x \left(x - \bar{x}^2\right)^q \cdot m_x}{\sum_x m_x}. \tag{4.5}$$

Центральний момент 1 порядку завжди рівний нулю, центр. момент 2 порядку відповідає дисперсії. Коефіцієнтом асиметрії називають відношення центрального моменту третього порядку до куба середнього квадратичного відхилення.

$$= \frac{\ }{\ } = \frac{\qquad\qquad}{\qquad\qquad}. \tag{4.6}$$

Якщо полігон варіаційного ряду скошений, то такий ряд називають асиметричним. Якщо у варіаційному ряді більше варіантів, менших за середнє арифметичне, то говорять, що є лівостороння асиметрія (відповідно, правостороння асиметрія). Для помірно асиметричних рядів коефіцієнт асиметрії менший одиниці.

Ексцесом (коефіцієнтом крутості) називають зменшене на 3 відношення центрального моменту 4 порядку до 4 степені середнього квадратичного відхилення: $\tilde{E} = \mu^4 / s^4 - 3$. Для кривої, яка представляє собою нормальний розподіл величина ексцесу рівна нулю.

Хід роботи

Задатись вибіркою із 5N елементів (нормального розподілу) в діапазоні від 0 до N включно з довільними параметрами.

Обчислити дисперсію, середньє квадратичне відхилення, коефіцієнти асиметрії та ексцесу на основі змодельованих даних.

Контрольні питання.

1 Розмах варіації

2 Стандартне лінійне відхилення.

3. Середнє квадратичне відхилення. Дисперсія. Зміщені і незміщені оцінки.

4 Поняття про моменти.

5 Асиметрія та ексцес.

6. Поняття життєвого циклу програмного забезпечення.

ЛАБОРАТОРНА РОБОТА № 5

Тема: основи програмування перевірки статистичних гіпотез.

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

Статистичною гіпотезою називається будь-яке припущення щодо властивостей генеральної сукупності на основі оцінок вибірки, припущення щодо виду або параметрів невідомого закону розподілу. Статичну гіпотезу прийнято позначати Н.

Розрізняють прості і складні статистичні гіпотези.

Проста гіпотеза – повністю визначає теоретичну функцію розподілу випадкової величини. Ті що не визначають – називаються складними. Статистичні гіпотези підрозділяються на нульові та альтернативні.

Нульова гіпотеза - позначається H_0 – це гіпотеза про відсутність відмінностей у значеннях ознак.

Альтернативна гіпотеза – позначається H_1 – це гіпотеза є логічним запереченням H_0, тобто – це гіпотеза про існування відмінностей.

Статистичні гіпотези можуть бути також:

- спрямованими – висувають про те, що значення показника в одній сукупності нижче ніж значення показника в іншій. Також називають однобічними.

- неспрямованими – формулюють якщо необхідно довести відмінності форми розподілу або значень показників відхилень. Також називають двобічними.

Перевірка гіпотез здійснюється на основі статистичних критеріїв.

Статистичний критерій – це правило, що забезпечує математично обґрунтоване прийняття істинної і відхилення помилкової гіпотези. Статистичні критерії – практично являють собою метод розрахунку певного числа, яке позначається як емпіричне значення критерію. Дане значення порівнюється з деким критичним значенням для даного критерію.

Співідношення між ними є підставою для підтвердження чи спростування гіпотези.

Критерії поділяються на:

- параметричні – використовуються в завданнях перевірки параметричних гіпотез і включають в свій розрахунок конкретні показники розподілу. Дозволяють безпосередньо оцінити параметри сукупностей чи вибірок. Оцінити середні відмінності в дисперсіях. Такі критерії дають можливість виявити тенденції зміни ознак, оцінити впливи факторів на ознаку.

- непараметричні – оперують частотами, рангами тощо.

Застосування таких критеріїв для прийняття або відхилення статистичних гіпотез завжди здійснюється з певною довірчою ймовірністю, інакше кажучи на певному рівні значущості.

Рівень значущості – це ймовірність того, що ми в результаті застосування критеріїв визнали відмінності істотними, а насправді вони випадкові. Рівень статистичної значущості у більшості випадків прийнятий за 5%. Існує значна кількість різних типів статистичних гіпотез. Ці типи визначаються сукупністю завдань та методів їх розв'язання.

Основні групи статистичних гіпотез за прикладними задачами, яких вони стосуються:

- гіпотези стосовно закону розподілу;

- гіпотези стосовно чисельних показників параметрів розподілів;

- гіпотези стосовно однорідності вибірок;

- гіпотези стосовно рівня ознак досліджуваного явища або процесу.

Незважаючи на різноманітність типів гіпотез і критеріїв загальна схема перевірки статистичних гіпотез така:

1) формулювання нульової та альтернативної гіпотези на основі задачі дослідження;

2) перевірка припущень щодо відповідності розподілам, перевірка параметрів вибірки, та іншої додаткової інформації.

3) прийняття рівня значущості;

4) вибір статистичного критерію;

5) розрахунки емпіричного критерію;

6) визначення області критичних значень критерію;

7) прийняття статистичного рішення;

8) формулювання статистичних висновків;

9) формулювання змістовних висновків.

В статистиці існують 2 підходи стосовно методів перевірки гіпотез. За одним із них обов'язково формулюють і 0, і альтеративну гіпотезу, перевірки яких відбуваються незалежно і повноцінно. При іншому підході формулювання альтернативних гіпотез не відбувається взагалі.

Задачі що вирішуються:

- перевірка гіпотез щодо однорідності вибірок
- перевірка гіпотез про чисельні значення параметрів
- гіпотези про виявлення відмінностей та зсувів в ознаках
- перевірка значущості коефіцієнтів кореляції

Критерій асиметрії та ексцесу застосовують для приблизної перевірки гіпотези про нормальність емпіричного розподілу. Для нормального розподілу коефіцієнти асиметрії та ексцесу рівні 0. Практично щоб одержати оцінку за даним методом обчислюються так звані дисперсії асиметрії та ексцесу:

$$D(A)=\frac{6(n-2)}{(n+1)(n+3)}, \tag{5.1}$$

$$D(E)=\frac{24n(n-2)(n-3)}{(n+1)(n+3)(n+5)}. \tag{5.2}$$

Вважають, що при нормальному розподілі вибіркові показники асиметрії та ексцесу дорівнюватимуть нулю, але реально таке майже не спостерігається. Тому емпіричний розподіл вважають близьким до нормального (приймають нульову гіпотезу), якщо виконуються умови: $|A_x|\le 3\sqrt{D(A)}$ та $|E_x|\le 5\sqrt{D(E)}$. Технологічно при цьому розраховують показники $t_A=\frac{|A_x|}{\sqrt{D(A)}}$ і $t_E=\frac{|E_x|}{\sqrt{D(E)}}$. Про достовірну відмінність емпіричного розподілу від нормального свідчать показники t_A і t_E, якщо приймають значення 3 і більше.

Хід роботи

Задатись вибіркою з (2A+5) елементів із показниковим (експотенційним) розподілом. Задатись вибіркою з 100*(2A+5) елементів із нормальним розподілом (А – номер у журналі старости). Написати програму, яка формує протокол статистичного дослідження для довільних даних на основі критерію асиметрії те ексцесу, результати представити для одержаних вибірок. Для одержання нормального розподілу рекомендується використати спеціальні алгоритми (бібліотеки).

Контрольні питання

1 Що таке статистична гіпотеза?

2 Що таке статистичний критерій?

3 Які є статистичні критерії критерії?

4 Методи генерації вибірок з нормальним розподілом

5 Критерій асиметрії та ексцесу.

6 Які ви знаєте підходи до організації промислового розроблення програмного забезпечення?

7 Назовіть основні положення екстремального програмування.

8 У чому полягає концепція парного програмування?

ЛАБОРАТОРНА РОБОТА № 6

Тема: програмна реалізація критерію узгодженості Пірсона

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості.

Критерій узгодженості Пірсона слугує для перевірки гіпотези про те, що дійсний розподіл ВВ і гіпотетичний розподіл є однаковими.

Нехай у результаті n незалежних спостережень за випадковою змінною y має функцію розподілу $F(x)$, що також відповідає певному гіпотетичному розподілу. На основі цієї вибірки ми хочемо перевірити правдоподібність гіпотези H_o: обидві випадкові величини належать до одного закону розподілу.

Функція розподілу випадкової змінної y повністю описує її, зокрема і простір її можливих значень. Даний простір P розіб'ємо довільно на $r+1$ $(r=1,2,\ldots)$ частину $S_1, S_2, \ldots S_{r+1}$ так, що $S_i \cdot S_j = 0(i \neq j)$.

Нехай у разі такого розбиття в комірку S_i попадає m_i $(i=1,\ldots,r+1)$ елементів вибірки x, де $\sum_{i=1}^{r+1} m_i = n$. Отже, відносна частота трапляння вибіркових значень у комірку, $S_i = \frac{m_i}{n}$. Згідно з гіпотетичною функцією розподілу $F(x)$ імовірність попадання значень випадкової змінної y в цю саму комірку $P(y \in S_i) = p_i$ $(i=1,\ldots,r+1)$, де $\sum_{i=1}^{r+1} p_i = 1$.

Якщо обидві вибірки випадкової величини дійсно належать до одного закону розподілу і керуються функцією розподілу $F(x)$, то за великого n майже напевно (про це існують теореми, доказані Я. Бернуллі та Е. Бореля) $\frac{m_i}{n}$ як завгодно мало відрізняється від p_i. Тому за міру узгодженості висунутої гіпотези природно вибирати величину

$$\chi^2(r,n,F) = \sum_{i=1}^{r+1} C_i \left(\frac{m_i}{n} - p_i \right)^2, \tag{6.1}$$

де c_i – деякі додатні константи.

К. Пірсон показав, що при $c_i = \frac{n}{p_i}$ вибірковий розподіл величини

$$\chi^2(r,n,F) = \sum_{i=1}^{r+1} \frac{(m_i - np_i)^2}{np_i} \qquad (6.2)$$

при $n \to \infty$ прямує до розподілу, який не залежить від виду гіпотетичної функції розподілу $F(x)$ генеральної сукупності.

Отже, алгоритм критерію χ^2 перевірки приналежності вибірки x з гіпотезою H_0 про функцію розподілу $F(x)$ і приналежності двох послідовностей до одного закону розподілу, такий.

Вибираємо рівень значущості α. Фіксуємо розбиття S_i гіпотетичного простору можливих значень змінної y на $r+1$ частину. Визначаємо кількості m_j попадань елементів цієї вибірки в кожну комірку S_i розбиття. Якщо деякі $m_j < 5$, то відповідні комірки об'єднюються з сусідніми так, щоб $m_j > 5$. При цьому змінюється відповідне число ступенів свободи r.

Обчислюємо емпіричне значення $\chi^2_{емп}$ статистики К. Пірсона. За вибраного рівня значущості α та числа ступенів свободи r знаходимо з таблиці "Квантилі статистики " критичне значення $\chi^2_{кр}$ статистики χ^2. При написанні програм використовують апроксимації даної таблиці, наприклад апроксимація Голдштейна, формули якої наведено нижче. Якщо $\chi^2_{емп} > \chi^2_{кр}$, то гіпотезу H_o відкидаємо, а якщо $\chi^2_{емп} < \chi^2_{кр}$, то кажемо, що дані вибірки за рівня значущості α не суперечать висунутій гіпотезі H_o, тобто належать до одного закону розподілу.

Апроксимація Голдштейна полягає в такому алгоритмі:

$$\chi^2_{\alpha,n} = n\left[\sum_{i=0}^{6} n^{-\frac{i}{2}} \cdot d^i \cdot \left(a_i + \frac{b_i}{n} + \frac{c_i}{n^2}\right)\right]^3, \qquad (6.3)$$

де $d = 2.0637 \cdot \left(\ln\frac{1}{1-\alpha} - 0.16\right)^{0.4274} - 1.5774$ при $0.5 \le \alpha \le 0.999$;

$d = -2.0637 \cdot \left(\ln\frac{1}{\alpha} - 0.16\right)^{0.4274} + 1.5774$ при $0.001 \le \alpha \le 0.5$.

Коефіцієнти a, b, c наведені у таблиці:

a	b	c
1.0000886	-0.2237368	-0.01513904
0.4713941	0.02607083	-0.008986007
0.0001348028	0.01128186	0.02277679
-0.008553069	-0.01153761	-0.01323293
0.00312558	0.005169654	-0.006950356
-0.0008426812	0.00253001	0.001060438
0.00009780499	-0.001450117	0.001565326

Хід роботи

Організувати ввід з консолі $(N+20)$ цілих чисел в діапазоні від 0 до 20 включно. Запрограмувати перевірку на допустимі значення. Запрограмувати перевірку апроксимації Голдштейна для різних значень α (контрольні значення – додаток А).

Ввести числа вручну випадковим чином. Організувати перевірку програмним чином на основі критерію χ^2 на відповідність введеної вибірки нормальному та рівномірному розподілу за двома різними рівнями значущості.

Вивести на екран вибірку та результати перевірок.

Контрольні запитання

1 Критерій узгодженості Пірсона

2 Яким чином організована перевірка даних на допустимість.

4 Водоспадна модель ЖЦ

4 Спіральна модель ЖЦ

Лабораторна робота № 7

Тема: програмування перевірки на однорідність.

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

Вибірки називаються однорідними, якщо вони взяті із однієї генеральної сукупності (або так можна вважати).

Для перевірки однорідності незв'язаних вибірок нерідко використовується критерій Стьюдента *t,* статистика якого має вид:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \tag{7.1}$$

де $\overline{X_1}$ і $\overline{X_2}$, s_1^2, s_2^2, n_1 і n_2 – середні, дисперсії та обсяги першої і другої вибірок відповідно.

Критичне значення критерію $t_{кр}$ для заданого рівня значущості *а* й числа ступенів вільності $(n_1 + n_2 - 2)$ можна отримати з таблиць розподілу Стьюдента (додаток Б). Алгоритм одержання (апроксимації) не є тривіальним, при виконанні даної роботи можна скористатись кодом, наведеним у прикладі.

Завдання

1. Задатись нормальним розподілом B_1 з (L+30) елементів з $m = 10$, $\sigma^2 = 10$.
2. Задатись нормальним розподілом B_1 з (L+30) елементів з $m = 15$, $\sigma^2 = 10$.
3. Задатись нормальним розподілом B_1 з (L+30) елементів з $m = 10$, $\sigma^2 = 4$.
4. Перевірити однорідність B_1; B_1 і B_2. Рівень значущості – 0,05.

В коротких теоретичних відомостях звіту представити використаний метод формування нормального розподілу.

Передбачити вивід вибірок та результатів перевірки однорідності відповідних вибірок у текстовий файл.

Приклад виконання

Завдання. Перевірити статистичні гіпотези на рівні значущості 0,05 щодо однорідності двох незалежних вибірок за критерієм Стьюдента.

$$X_1 = \{6\ 7\ 4\ 3\ 4\ 5\ 3\ 6\ 7\ 3\ 7\ 3\ 5\ 4\ 4\ 3\ 5\ 4\}$$

$$X_2 = \{4\ 4\ 5\ 4\ 1\ 5\ 5\ 3\ 3\ 6\ 2\ 3\ 4\ 3\ 7\ 5\ 3\ 2\ 4\ 5\}$$

Послідовність рішення:

- Ситуації відповідає варіант *неспрямованих гіпотез;*
  $H_0: \mu_1 - \mu_2 = 0$ (μ_1 не відрізняється від μ_2);
  $H_1: \mu_1 - \mu_2 \neq 0$ (μ_1 відрізняється від μ_2);
- *Перевірка припущень:* розподіл досліджуваних параметрів, а також дисперсії ***невідомі***; вибірки ***незв'язані;*** обсяги вибірок різні; виміри ***інтервальні***.
- ***Емпіричне значення критерію*** $t_{емп}$ можна оцінити також з елементарних розрахунків:

$$t_{емп} = \frac{4{,}72 - 3{,}90}{\sqrt{\frac{(18-1)*1{,}98 + (20-1)*1.01}{14+20-2}\left(\frac{1}{18}+\frac{1}{20}\right)}} \approx 1{,}77$$

- *Критичне значення* визначається за таблицею $t_{кр} \approx 2{,}03$
- *Прийняття рішення.* Оскільки $t_{емп} < t_{екр}$, тобто 1,77<2,04, нульова гіпотеза H_0 приймається на рівні значущості 0,05;
- ***Формулювання висновків.*** На рівні значущості 0,05 відсутні підстави стверджувати про неоднорідність незалежних вибірок.

```
/*------------------------------------------------------------------------------------------
/* js код повертає t-критерій Стюдента по числу степеней вільності
/* знайдено на просторах інтернету
/* перевірено – працює
/*------------------------------------------------------------------------------------------

function StatCom(q,i,j,b)
{
   var zz = 1, z = zz, k = i;
   while (k <= j){ zz *= (q*k)/(k-b); z += zz; k += 2; }
   return z;
}

function StudentT(t,n)
{
   var w = Math.abs(t)/Math.sqrt(n), th = Math.atan(w);
```

```
  if (n == 1) return ( 1-th/(Math.PI/2) );
  var sth = Math.sin(th), cth = Math.cos(th);
  if ((n%2) == 1) return ( 1-(th+sth*cth*StatCom(cth*cth,2,n-3,-1))/(Math.PI/2) );
  else          return ( 1-sth*StatCom(cth*cth,1,n-3,-1) );
}

function AStudentT(n,alpha)  // повертає t-критерій Стюдента по числу степеней
{                   //вільності та рівні значущості
  var v = 0.5, dv = 0.5, t = 0;
  while (dv > 1e-6)
   { t = 1/v-1; dv /= 2;
    if (StudentT(t,n) > alpha) v -= dv;
    else v += dv;
   }
  return t;
}
```

Контрольні питання

1. Що таке статистична гіпотеза?
2. Які бувають статистичні гіпотези? Алгоритм підтвердження або заперечення статистичної гіпотези.
3. Що таке рівномірний розподіл? Як його описують? Методи одержання рівномірного розподілу (у програмуванні).
4. Що таке нормальний розподіл? Як його описують? Методи одержання нормального розподілу (у програмуванні).

ЛАБОРАТОРНА РОБОТА № 8

Тема: програмна реалізація перевірки на грубі помилки за критерієм Смірнова.

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості.

Будь-які дані можуть містити одне або кілька значень, що помітно відрізняються від решти. Метою аналізу є встановлення того факту, чи є окремі значення вибірки грубими помилками. На цьому базуються методи оцінювання чесності проведення виборів. У окремих видах спорту завідомо відкидаються найкращі та найгірші результати (оцінки).

Для задачі виключення грубих помилок існує ряд критеріїв. Попри використання статистичних критеріїв, будь-яка подібна оцінка залишається доволі суб'єктивною. Практично рекомендується використовувати кілька критеріїв у кожному конкретному випадку.

Для нормально розподіленої випадкової величини часто використовують критерій Н. В. Смірнова.

Якщо генеральна дисперсія достатньо точно відома за поточними вимірюваннями, використовують статистику t_α. Для цього будують варіаційний ряд результатів досліджень, і якщо, одне з крайніх значень сумнівне, обчислюють критерій для сумнівних значень за формулами:

$$t_{розр} = \frac{|x_i - \bar{x}|}{\sigma}.$$

При $t_{розр} > t_\alpha$ результат вважають грубою помилкою і відкидають.

Значення t_α для вибірки з $n > 25$ з достатньою точністю обчислюється за формулами, наведеними в таблиці 8.1

Таблиця 8.1 – Апроксимація критичного значення за критерієм Смірнова

α	t_α при n>25
0,1	0,3053*ln(n)* + 1,6513
0,05	0,2849*ln(n)* + 1,9517
0,01	0,2648*ln(n)* + 2,4839

Хід роботи

Задатись вибіркою з $(N+30)$ елементів з нормальним розподілом та довільними параметрами. Провести дослідження. Замінити у вибірці найменше або найбільше значення на значення, яке :

– випадає з діапазону згенерованої вибірки не більше ніж на 0,3 СКВ;
– випадає з діапазону згенерованої вибірки не більше ніж на 1 СКВ;
– суттєво випадає з діапазону згенрованої вибрки.

Провести дослідження.

Розрахунки провести за допомогою розробленого програмного забезпечення.

Для вирішення поставленого завдання розробити:

– програму, яка генерує текстовий файл з даними відповідно до варіанту;
– програму для аналізу даних. Робота з файлом повинна бути винесена в окремий клас. Передбачити коректну обробку помилок вводу/виводу для роботи з файлом даних.

Продемонстувати вивід програми з явно помилковим водом, вивід програми за відсутності файла даних.

Контрольні питання

1. Критерій Н. В. Смірнова
2. Основні ролі SCRUM
3. Основні артефакти SCRUM.
4. У чому полягає процес SCRUM?
5. Парадигми програмування.
6. Основні поняття ООП.

ЛАБОРАТОРНА РОБОТА № 9

Тема: програмна реалізація перевірки на грубі помилки за критерієм Ірвіна.

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості.

Якщо розподіл досліджуваної вибірки не є нормальний, або його характер є невідомий, для оцінювання викидів можна використовувати критерій Ірвіна.

При цьому будують варіаційний ряд та оцінюють сумнівні значення на одномку чи двох краях ряду. Для цього обчислюють розрахункове значення критерію Ірвіна:

$$\eta_{розр} = \frac{x_k - x_{k\,поп}}{s},$$

де x_k – сумнівне значення;

$x_{k\,поп}$ – попереднє значення в ряду;

s – СКВ.

Одержане розрахункове значення критерію Ірвіна порівнюють з табличним; при автоматизованій обробці зручно використовувати формули, наведені в таблиці 9.1 (для $3 \le n \le 1000$).

Таблиця 9.1 – Апроксимація критичних значень за критерієм Ірвіна

Довірча ймовірність P	$\eta_{табл}$
0,9	$2n^{-0,5}+0,6$
0,95	$2,5n^{-0,5}+0,75$
0,99	$3n^{-0,5}+1,15$

Якщо $\eta_{розр} > \eta_{табл}$, то розглянуте значення відкидається і перевіряється наступне. Перевірку продовжують, поки не одержать $\eta_{розр} > \eta_{табл}$. Якщо крайнє значення ряду за цим критерієм не є грубою помилкою, бажаним є перевірка наступних значень. У випадку, якщо одне з них є грубою помилкою, попередні також вважаються грубою помилкою. Зазвичай перевіряють до 10% варіаційного ряду з кожного боку.

Хід роботи

Задатись вибіркою з $(N+30)$ елементів з нормальним розподілом (близьким до нього). Провести дослідження. Замінити у вибірці найменше або найбільше значення на значення, яке не входить у заданий діапазон. Провести дослідження.

Провести розрахунки з допомогою розробленого програмного забезпечення.

Розробити:

– програму, яка генерує текстовий файл з даними відповідно до варіанту;
– програму для аналізу даних. Робота з файлом повинна бути винесена в окремий клас. Передбачити коректну обробку помилок вводу/виводу для роботи з файлом даних.

Продемонстувати вивід програми з явно помилковим водом, вивід програми за відсутності файла даних.

Контрольні питання

1 Критерій Ірвіна
2 Поняття про RUP
3 Поняття про RAD
4 Поняття про екстремальне програмування
5 Концепція візуального програмування.
6 Концепція парного програмування

ЛАБОРАТОРНА РОБОТА № 10

Тема: проста лінійна регресія
Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості.

Основне завдання регресійного аналізу це встановлення форми і виявлення залежностей і змін.

Одне з головних завдань регресійного аналізу полягає у підборі відповідного виразу Y=f(X), графік якого проходить через емпіричні точки або близько від них. І таким чином зв'язані змінні X та Y. Даний вираз має назву рівняння регресії, функція f(X) називається функцією регресії. Графік даної функції називається лінією даної регресії. Тобто регресійний аналіз виявляє кількісну залежність ознаки фактора від іншої ознаки

Параметри або коефіцієнти емпіричних моделей можуть встановлюватись методом найменших квадратів і методом максимуму правдоподібності.

Типове і основне завдання формулюють таким чином:

Необхідно встановити параметри a_0 і a_1 для залежності

$$y = a_0 + a_1 X$$

на вибірці об`єму n.

У методі найменших квадратів дані параметри визначаються з умови мінімуму наступного критерію

$$R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 \times x_i)^2,$$

на основі чого було одержано наступні формули:

$$a_0 = \frac{\sum_{i=1}^{n} y_i \times \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} x_i \cdot y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)},$$

$$a_1 = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)}.$$

Хід роботи

Задатись довільною лінією у формі $y = kx + b$. Одержати дискретне представлення лінії в деякому невід’ємному діапазоні (N+10) точок, де N – номер у журналі. Додати до цих даних “шум” з нормальним розподілом – розмах варіації (N/5). Знайти рівняння регресії на основі “зашумлених” даних, вивести графіки та коефіцієнти вихідної лінії та регресійної моделі.

Контрольні питання

1. Види програмних вимог.
2. Цілі і задачі регресійного аналізу.
3. Метод найменших квадратів.
4. Що таке інтерполяція.
5. Що таке апроксимація
6. Що таке екстраполяція

ЛАБОРАТОРНА РОБОТА № 11

Тема: елементи кореляційного аналізу та їх програмна реалізація

Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості.

Кореляційні зв'язки можна аналізувати на якісному рівні з діаграм озсіяння емпіричних значень змінних і відповідним чином їх інтерпретувати. Так, наприклад, якщо підвищення рівня однієї змінної супроводжується підвищення рівня іншої, то йдеться про позитивну кореляцію або прямий зв'язок. Якщо ж зростання однієї змінної супроводжується зниженням значень іншої, то маємо справу з негативною кореляцією.

Нульовою називається кореляція за відсутності зв'язку змінних. Проте нульова загальна кореляція може свідчити лише про відсутність лінійної залежності, а не про відсутність залежності між величинами взагалі. Кількісна міра кореляційного зв'язку оцінюється найчастіше за значеннями коефіцієнта кореляції від +1 до -1. Від'ємні значення коефіцієнта кореляції свідчать про зворотний зв'язок, додатні – про прямий.

Нульове значення може свідчити про відсутність зв'язку. Інтенсивність зв'язку (слабкий – помірний – суттєвий – сильний) оцінюється за абсолютним значенням коефіцієнта кореляції.

Методи розрахунку міри кореляційних зв'язків тісно пов'язані із вживаними вимірюваними шкалами.

Лінійний кореляційний зв'язок для емпіричних даних, виміряних за шкалою інтервалів або відношень, оцінюється за допомогою коефіцієнта кореляції Пірсона r_{xy}:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}, \qquad (11.1)$$

де x_i і y_i – значення змінних, $\bar{x}$ та $\bar{y}$ – серединні значення вибірок, n – обсяг вибірок.

Показник кореляції рангів. У тих випадках, коли одиниці досліджуваної часткової сукупності можуть бути у відношенні деякої ознаки розміщені в певному порядку по зростаючим (або спадаючим) номерам, або рангам, в якості статистики зв'язку служить показник кореляції рангів.

Ранг вказує те місце, яке займає дана оодиниця сукупності серед інших одиниць. Якщо б кожна з цих одиниць відрізнялась у відношенні розглядуваної ознаки від всіх інших одиниць сукупності, то ранги являли собою б порядкові номери від 1 до числа n , рівного об'єму сукупності. Якщо ж деякі з одиниць сукупності є однаковими, то ранг всіх цих одиниць приймається середнім з їх відповідніх номерів.

Показник кореляції рангів рівний:

$$\rho = 1 - \frac{6\sum_{h=1}^{n} d_h^2}{n\left(n^2 - 1\right)}, \tag{11.2}$$

де величини d_h являють собою різницю між рангами h_1 та h_2 одиниць, вибраних разом з двох послідовностей:

$$d_h = h_1 - h_2$$

Показник кореляції рангів змінюється від -1 до +1:

$$-1 \le \rho \le +1$$

Чим тісніший буде зв'язок між величинами, тим ближче до одиниці за своєю абсолютною величиною буде показник кореляції рангів; знак вказує, пряма залежність чи зворотна.

Хід роботи

1 Задатись вибіркою А із 100елементів , що відповідає 1-му періоду синусоїди з параметрами Т-секунд, Т – номер студента.

2. Задатись вибіркою Б, що відповідає синусоїді з пункту 1 у протифазі з половинною амплітудою.

3. Задатись вибіркою В, що утворюється з вибірки А шляхом додавання рівномірного шуму амплітудою 25 % від початкової.

4. Здійснити кореляційний аналіз А і Б, А і В.

Контрольні питання

1 Поняття про кореляційний аналіз.
2 Лінійний кореляційний зв язок.
3 Показник кореляції рангів.
4. Множинна кореляція
5 Аналіз вимого до програмного забезпечення
6 Сертифікація вимого до програмного забезпечення

ЛАБОРАТОРНА РОБОТА № 12

Тема: дисперсійний однофакторний аналіз
Технічне забезпечення: ПЕОМ, середовище програмування

Короткі теоретичні відомості

Основною метою дисперсійного аналізу, концепція якого була запропонована Фішером у 1920 р., є дослідження значущості відмінності між середніми декількох груп даних.

Дисперсійний однофакторний аналіз використовується у дослідженнях зміни результативної ознаки під впливом зміни умов або градацій фактора. Суть математичних перетворень дисперсійного методу полягає в тому, щоб зіставити дисперсії за факторами із дисперсією усіх значень отриманих в експерименті. Однофакторний аналіз вимагає не менше трьох градацій фактора і не менше двох випробувань у кожній градації. При проведенні дисперсійного аналізу необхідно перевірити нормальність розподілу досліджуваної випадкової величини і відсутність відмінностей дисперсій сукупностей. Це можна виконати методами перевірки статистичних гіпотез.

Припустимо, що аналізується вплив фактора А на k рівнях А1, А2, …, Аk. Наприклад, в експерименті це можна реалізувати, якщо задіяти k вибірок з різними градаціями умов. На кожному рівні Аі (для кожної вибірки) проведено n спостережень x_{i1}, x_{i2},…, x_{in} (див.таблиця 1).

Розглянемо оцінки різних дисперсій.

Дисперсія s_i^2 для рівня A_i (для певної вибірки) може бути записана як:

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij}-\frac{1}{n}\left(\sum_{j=1}^{n}x_{ij}\right)\right)^2.$$

Дисперсія s_0^2, що характеризує варіативність поза впливом фактора А

$$s_0^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 =$$

$$= \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right].$$

Таблиця №1 – Вихідні дані для дисперсного аналізу

Номери спостережень	Рівні фактора А			
	A_1	A_2		A_k
1	x_{11}	x_{21}	…	x_{k1}
2	x_{12}	x_{22}	…	x_{k2}
…	…	…	…	…
j	x_{1j}	x_{2j}	…	x_{kj}
…	…	…	…	…
n	x_{1n}	x_{2n}	…	x_{kn}
∑	X_1	X_2		X_k

Загальна дисперсія s^2всіх nkспостережень дорівнює $s^2 = \frac{1}{kn-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$, де $\bar{x} = \frac{1}{k}\sum_{i=1}^{k}\bar{x}_i$та $\bar{x}_i = \frac{1}{k}\sum_{j=1}^{n} x_{ij}$. Отже,

$$s^2 = \frac{1}{kn-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{kn}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right).$$

Дисперсія s_A^2, що характеризує зміну середніх $\bar{x}_i$під впливом фактора А:

$$s_A^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2.$$

Перевірка впливу фактора А на зміну середніх може бути зведена до порівняння дисперсій s_A^2 та s_0^2. Вплив фактора А вважатиметься значущим на рівні α, якщо є значущим відношення s_A^2/s_0^2, тобто, якщо:

$$s_A^2/s_0^2 > F[\alpha; k-1; k(n-1)],$$

де k-1; k(n-1) – ступені вільності F-розподілу, F – критерій Фішера.

Завдання

Здійснити дисперсійний аналіз впливу швидкості пред'явлення слів на показники їх відтворення. Вихідні дані (по варіантах) наведено нижче.Для проведення розрахунків можна використовувати програму MicrosoftExel, але стандартний пакет для дисперсійного аналізу використовувати заборонено. Звіт оформити відповідно до прикладу.

1. Низька: 4, 7, 6, 5, 6, 4
   Середня: 5, 6, 5, 4, 5,4
   Висока: 4, 4, 5, 6, 4, 3
2. Низька: 4, 7, 6, 5, 6, 7
   Середня: 5, 6, 5, 5, 5,4
   Висока: 4, 4, 5, 3, 4, 3
3. Низька: 4, 6, 6, 5, 6, 4
   Середня: 5, 6, 5, 4, 5,4
   Висока: 4, 4, 2, 6, 4, 3
4. Низька: 4, 7, 6, 6, 6, 4
   Середня: 5, 6, 5, 5, 5,4
   Висока: 4, 2, 5, 6, 4, 3
5. Низька: 4, 7, 6, 5, 6, 4
   Середня: 5, 6, 2, 4, 5,4
   Висока: 4, 1, 5, 6, 4, 3
6. Низька: 4, 7, 6, 3, 6, 7
   Середня: 5, 6, 5, 4, 5,4
   Висока: 4, 4, 2, 6, 2, 2
7. Низька: 4,8, 6, 5, 6, 5
   Середня: 4, 6, 5, 4, 5,4
   Висока: 3, 3, 5, 6, 4, 3
8. Низька: 4,6, 6, 6, 6, 6
   Середня: 5, 6, 5, 4, 5,4
   Висока: 2, 4, 3, 6, 4, 3
9. Низька: 4,1, 6, 5, 6, 4
   Середня: 5, 6, 2, 4, 2,4
   Висока: 4, 4, 5, 6, 4, 3
10. Низька: 4,2, 6, 5, 6, 4
    Середня: 5, 2, 5, 4, 2,4
    Висока: 4, 1, 5, 5, 4, 3

Приклад виконання роботи

Довести припущення про те, що фактор

Низька швидкість: 6, 7, 6, 5, 6, 4.

Середня швидкість: 5, 6, 5, 4, 4, 5.

Висока швидкість: 4, 4, 4, 3, 5, 3.

Послідовність рішення:

- Формулювання гіпотез.
  Но: фактор швидкості не є більш вираженим, ніж випадковим;
  Н1: фактор швидкості є більш вираженим, ніж випадковим.
- Перевірка припущень.
  Вибірки незв'язані однакових обсягів.
  Виміри за шкалою відношень.
  Вважатиметься, що досліджуваний параметр має нормальний розподіл.
- Введені позначення:
  n=6 – кількість спостережень за фактором,
  k=3 – кількість факторів,
  n*k=6*3=18 – загальна кількість індивідуальних значень,
  j – індекс змінюється від 1 до n
  i – індекс змінюється від 1 до k.
- Математичні розрахунки (приклад обчислень наведено на рисунку 1.)

B18 | f_x =FРАСПОБР(0,01;2;15)

	A	B	C	D	E	F
1		Швидкість пред'явлення				
2		Низька	Середня	Висока		
3		6	5	4		
4		7	6	4		
5		6	5	4		
6		5	4	3		
7		6	4	5		
8		4	5	3		
9	Суми	34	29	23		
10	Середні	5,666667	4,833333	3,833333		
11	n=	6				
12	k=	3				
13	Q1=	432				
14	Q2=	421				
15	Q3=	410,8889				
16	Fемп=	6,893939				
17	F0,05=	3,68232				
18	F0,01=	6,358873				

Рисунок 12.1 – Приклад обчислень для дисперсійного аналізу

– розрахувати значенняза формулами:

$Q_1 = \sum\sum x_{ij}^2, Q_2 = \frac{1}{n}\sum_{i=1}^{k} X_i^2, Q_3 = \frac{1}{kn}\left(\sum_{i=1}^{k} X_i\right)^2$, а саме:

$Q_1 = 6*6 + 7*7 + \cdots = 432$,

$Q_2 = \frac{1}{6}(34*34 + 29*29 + 23*23) = 421$,

$Q_3 = \frac{1}{3*6}(34 + 29 + 23)^2$=410.89;

– розрахувати емпіричний критерій за формулою

$$F_{емп} = \frac{s_A^2}{s_B^2} = \frac{k(n-1)}{k-1}\frac{Q_2 - Q_3}{Q_1 - Q_2}.$$

Після підстановки $F_{емп} \approx 6{,}89$.

- Критичне значення функції F для рівня значущості 0,05 та 0,01 можна одержати використовуючи функцію Microsoft Excel =FРАСПОБР(), або за таблицями відповідного розподілу.
  Кількість ступенів вільності: (k-1)=2 та k(n-1)=15. F(0.05)=3.68, F(0.01)=6.36.
- Прийняття рішення. Нульова гіпотеза відхиляється на рівні значущості 0.01.

- Формулювання висновків. Відмінності в обсязі відтворення слів є більш вираженими, ніж випадковими.

Контрольні питання

1. Однофакторний дисперсійний аналіз. Цілі та задачі, суть методу.
2. Двофакторний дисперсійний аналіз.
3. Специфікація програмних вимог.
4. Затвердження вимог до програм
5. Процес роботи з вимогами

ПЕРЕЛІК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

1. Дональд Кнут Искуство программирования, том 1. Основные алгоримы = The Art of Computer Programming, vol. 1 Fundamental Algorithms. – 3-е изд. – М.: "Вильямс", 2006. – 720с.

2. Дональд Кнут Искуство программирования, том 3. Сортировка и поиск = The Art of Computer Programming, vol. 3 Sorting and Seaching. – 2-е изд. – М.: "Вильямс", 2007. – 824с.

3. Бентли Дж. Жемчужины программирования. 2-е изд. – СПб.: Питер, 2002. – 272с.

4. Жлуктенко В. І. Наконечний С. І. Савіна С. С. Теорія ймовірностей і математична статистика: Навч. метод. посібник: У 2-х ч. – Ч. II. Математична статистика. – К. КНЕУ, 2001. – 336с.

5. Статистика: Підручник / С. С. Герасименко, А. В. Головач, А. М. Єріна та ін.; за наук ред. д-ра екон. наук С. С. Герасименка. – 2-ге вид., перероб. і доп. – К.: КНЕУ, 2000 – 467с.

6. Руденко В. М. Математична статистика. Навчальний посібник / В. М. Руденко – К.: Центр учбової літератури, 2012. – 304с.

7. Гмурман В. Е. Теория вероятностей и математическая статистика : учеб. пособие для бакалавров / В. Е. Гмурман. — 12-е изд. — М.: Издательство Юрайт, 2013. — 479 с. : ил. — Серия : Бакалавр. Базовый курс.

8. Тамре Л. Введение в тестирование программного обеспечения: Пер. с англ. – М.: Изд. дом "Вильямс", 2003 – 368 с.: ил.

9. Черниxівський Є. М. Математичне моделювання телекомунікаційних систем та мереж: навчальний посібник. / Є. М. Черниxівський. – Львів: Видавництво Львівської політехніки, 2011. – 272с.

10. Незамай Б. С. Емпіричні методи програмної інженерії: конспект лекцій / Б. С. Незамай, М. М. Яцишин, Т. В. Дитко – Івано-Франківськ: ІФНТУНГ, 2014. – 74с.

11. Незамай Б. С. Інформаційні системи / Б. С Незамай – Івано-Франківськ: ІФНТУНГ, 2012. – 68с.

Додаток А

Таблиця 1 – Квантилі статистики хіта-квадрат

	0,01	0,025	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,975	0,99
1	0,0002	0,0010	0,0039	0,0158	0,0642	0,1485	0,2750	0,4549	0,7083	1,0742	1,6424	2,7055	3,8415	5,0239	6,6349
2	0,0201	0,0506	0,1026	0,2107	0,4463	0,7133	1,0217	1,3863	1,8326	2,4079	3,2189	4,6052	5,9915	7,3778	9,2103
3	0,1148	0,2158	0,3518	0,5844	1,0052	1,4237	1,8692	2,3660	2,9462	3,6649	4,6416	6,2514	7,8147	9,3484	11,3449
4	0,2971	0,4844	0,7107	1,0636	1,6488	2,1947	2,7528	3,3567	4,0446	4,8784	5,9886	7,7794	9,4877	11,1433	13,2767
5	0,5543	0,8312	1,1455	1,6103	2,3425	2,9999	3,6555	4,3515	5,1319	6,0644	7,2893	9,2364	11,0705	12,8325	15,0863
6	0,8721	1,2373	1,6354	2,2041	3,0701	3,8276	4,5702	5,3481	6,2108	7,2311	8,5581	10,6446	12,5916	14,4494	16,8119
7	1,2390	1,6899	2,1673	2,8331	3,8223	4,6713	5,4932	6,3458	7,2832	8,3834	9,8032	12,0170	14,0671	16,0128	18,4753
8	1,6465	2,1797	2,7326	3,4895	4,5936	5,5274	6,4226	7,3441	8,3505	9,5245	11,0301	13,3616	15,5073	17,5345	20,0902
9	2,0879	2,7004	3,3251	4,1682	5,3801	6,3933	7,3570	8,3428	9,4136	10,6564	12,2421	14,6837	16,9190	19,0228	21,6660
10	2,5582	3,2470	3,9403	4,8652	6,1791	7,2672	8,2955	9,3418	10,4732	11,7807	13,4420	15,9872	18,3070	20,4832	23,2093
11	3,0535	3,8157	4,5748	5,5778	6,9887	8,1479	9,2373	10,3410	11,5298	12,8987	14,6314	17,2750	19,6751	21,9200	24,7250
12	3,5706	4,4038	5,2260	6,3038	7,8073	9,0343	10,1820	11,3403	12,5838	14,0111	15,8120	18,5493	21,0261	23,3367	26,2170
13	4,1069	5,0088	5,8919	7,0415	8,6339	9,9257	11,1291	12,3398	13,6356	15,1187	16,9848	19,8119	22,3620	24,7356	27,6882
14	4,6604	5,6287	6,5706	7,7895	9,4673	10,8215	12,0785	13,3393	14,6853	16,2221	18,1508	21,0641	23,6848	26,1189	29,1412
15	5,2293	6,2621	7,2609	8,5468	10,3070	11,7212	13,0297	14,3389	15,7332	17,3217	19,3107	22,3071	24,9958	27,4884	30,5779
16	5,8122	6,9077	7,9616	9,3122	11,1521	12,6243	13,9827	15,3385	16,7795	18,4179	20,4651	23,5418	26,2962	28,8454	31,9999
17	6,4078	7,5642	8,6718	10,0852	12,0023	13,5307	14,9373	16,3382	17,8244	19,5110	21,6146	24,7690	27,5871	30,1910	33,4087
18	7,0149	8,2307	9,3905	10,8649	12,8570	14,4399	15,8932	17,3379	18,8679	20,6014	22,7595	25,9894	28,8693	31,5264	34,8053
19	7,6327	8,9065	10,1170	11,6509	13,7158	15,3517	16,8504	18,3377	19,9102	21,6891	23,9004	27,2036	30,1435	32,8523	36,1909
20	8,2604	9,5908	10,8508	12,4426	14,5784	16,2659	17,8088	19,3374	20,9514	22,7745	25,0375	28,4120	31,4104	34,1696	37,5662
21	8,8972	10,2829	11,5913	13,2396	15,4446	17,1823	18,7683	20,3372	21,9915	23,8578	26,1711	29,6151	32,6706	35,4789	38,9322
22	9,5425	10,9823	12,3380	14,0415	16,3140	18,1007	19,7288	21,3370	23,0307	24,9390	27,3015	30,8133	33,9244	36,7807	40,2894
23	10,1957	11,6886	13,0905	14,8480	17,1865	19,0211	20,6902	22,3369	24,0689	26,0184	28,4288	32,0069	35,1725	38,0756	41,6384
24	10,8564	12,4012	13,8484	15,6587	18,0618	19,9432	21,6525	23,3367	25,1063	27,0960	29,5533	33,1962	36,4150	39,3641	42,9798
25	11,5240	13,1197	14,6114	16,4734	18,9398	20,8670	22,6156	24,3366	26,1430	28,1719	30,6752	34,3816	37,6525	40,6465	44,3141
26	12,1981	13,8439	15,3792	17,2919	19,8202	21,7924	23,5794	25,3365	27,1789	29,2463	31,7946	35,5632	38,8851	41,9232	45,6417
27	12,8785	14,5734	16,1514	18,1139	20,7030	22,7192	24,5440	26,3363	28,2141	30,3193	32,9117	36,7412	40,1133	43,1945	46,9629
28	13,5647	15,3079	16,9279	18,9392	21,5880	23,6475	25,5093	27,3362	29,2486	31,3909	34,0266	37,9159	41,3371	44,4608	48,2782
29	14,2565	16,0471	17,7084	19,7677	22,4751	24,5770	26,4751	28,3361	30,2825	32,4612	35,1394	39,0875	42,5570	45,7223	49,5879
30	14,9535	16,7908	18,4927	20,5992	23,3641	25,5078	27,4416	29,3360	31,3159	33,5302	36,2502	40,2560	43,7730	46,9792	50,8922

Додаток Б

Значення критерію Стьюдента

f	***p***							
	0.80	**0.90**	**0.95**	**0.98**	**0.99**	**0.995**	**0.998**	**0.999**
1	3.0770	6.3130	12.7060	31.820	63.656	127.656	318.306	636.619
2	1.8850	2.9200	4.3020	6.964	9.924	14.089	22.327	31.599
3	1.6377	2.35340	3.182	4.540	5.840	7.458	10.214	12.924
4	1.5332	2.13180	2.776	3.746	4.604	5.597	7.173	8.610
5	1.4759	2.01500	2.570	3.649	4.0321	4.773	5.893	6.863
6	1.4390	1.943	2.4460	3.1420	3.7070	4.316	5.2070	5.958
7	1.4149	1.8946	2.3646	2.998	3.4995	4.2293	4.785	5.4079
8	1.3968	1.8596	2.3060	2.8965	3.3554	3.832	4.5008	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.780
10	1.3720	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.363	1.795	2.201	2.718	3.105	3.496	4.024	4.437
12	1.3562	1.7823	2.1788	2.6810	3.0845	3.4284	3.929	4.178
13	1.3502	1.7709	2.1604	2.6503	3.1123	3.3725	3.852	4.220
14	1.3450	1.7613	2.1448	2.6245	2.976	3.3257	3.787	4.140
15	1.3406	1.7530	2.1314	2.6025	2.9467	3.2860	3.732	4.072
16	1.3360	1.7450	2.1190	2.5830	2.9200	3.2520	3.6860	4.0150
17	1.3334	1.7396	2.1098	2.5668	2.8982	3.2224	3.6458	3.965
18	1.3304	1.7341	2.1009	2.5514	2.8784	3.1966	3.6105	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.3253	1.7247	2.08600	2.5280	2.8453	3.1534	3.5518	3.8495
21	1.3230	1.7200	2.2.0790	2.5170	2.8310	3.1350	3.5270	3.8190
22	1.3212	1.7117	2.0739	2.5083	2.8188	3.1188	3.5050	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.7251
26	1.315	1.705	2.059	2.478	2.778	3.0660	3.4360	3.7060
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.0360	3.3962	3.8494
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.6460
32	1.3080	1.6930	2.0360	2.4480	2.7380	3.0140	3.3650	3.6210
34	1.3070	1.6909	2.0322	2.4411	2.7284	3.9520	3.3479	3.6007
36	1.3050	1.6883	2.0281	2.4345	2.7195	9.490	3.3326	3.5821
38	1.3042	1.6860	2.0244	2.4286	2.7116	3.9808	3.3190	3.5657
40	1.303	1.6839	2.0211	2.4233	2.7045	3.9712	3.3069	3.5510
42	1.320	1.682	2.018	2.418	2.6980	2.6930	3.2960	3.5370

f	p							
	0.80	**0.90**	**0.95**	**0.98**	**0.99**	**0.995**	**0.998**	**0.999**
44	1.301	1.6802	2.0154	2.4141	2.6923	3.9555	3.2861	3.5258
46	1.300	1.6767	2.0129	2.4102	2.6870	3.9488	3.2771	3.5150
48	1.299	1.6772	2.0106	2.4056	2.6822	3.9426	3.2689	3.5051
50	1.298	1.6759	2.0086	2.4033	2.6778	3.9370	3.2614	3.4060
55	1.2997	1.673	2.0040	2.3960	2.6680	2.9240	3.2560	3.4760
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.9146	3.2317	3.4602
65	1.2947	1.6686	1.997	2.3851	2.6536	3.9060	3.2204	3.4466
70	1.2938	1.6689	1.9944	2.3808	2.6479	3.8987	3.2108	3.4350
80	1.2820	1.6640	1.9900	2.3730	2.6380	2.8870	3.1950	3.4160
90	1.2910	1.6620	1.9867	2.3885	2.6316	2.8779	3.1833	3.4019
100	1.2901	1.6602	1.9840	2.3642	2.6259	2.8707	3.1737	3.3905
120	1.2888	1.6577	1.9719	2.3578	2.6174	2.8598	3.1595	3.3735
150	1.2872	1.6551	1.9759	2.3515	2.6090	2.8482	3.1455	3.3566
200	1.2858	1.6525	1.9719	2.3451	2.6006	2.8385	3.1315	3.3398
250	1.2849	1.6510	1.9695	2.3414	2.5966	2.8222	3.1232	3.3299
300	1.2844	1.6499	1.9679	2.3388	2.5923	2.8279	3.1176	3.3233
400	1.2837	1.6487	1.9659	2.3357	2.5882	2.8227	3.1107	3.3150
500	1.2830	1.6470	1.9640	2.3330	2.7850	2.8190	3.1060	3.3100