Міністерство освіти і науки України

Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»

Математична статистика

Методичні вказівки до лабораторних робіт

Рекомендовано Вченою радою
факультету інформатики та
обчислювальної техніки
НТУУ «КПІ»
Протокол № 3 від 31.10.2016р.

Київ
НТУУ «КПІ»
2016

Математична статистика. Методичні вказівки до лабораторних робіт. /
Уклад.: Т.А.Ліхоузова – К.: НТУУ «КПІ», 2016. – 25 с.

Методичні вказівки призначені для студентів спеціальностей 121 «Інженерія програмного забезпечення» та 151 «Автоматизація та комп'ютерно-інтегровані технології» кафедри технічної кібернетики всіх форм навчання. В посібнику наведена тематика лабораторних занять, дано короткі теоретичні відомості, список корисної літератури, контрольні питання, завдання для робіт, рекомендації щодо виконання робіт та склад звіту.

Укладач Т.А. Ліхоузова, к.т.н., доцент

Відповідальний редактор М.М.Ткач, к.т.н., доцент

Рецензент Ковалюк Т.В., к.т.н., доцент
кафедри АСОІУ ФІОТ

За редакцією укладачів

Зміст

Загальні рекомендації

Роботи можна виконувати в одному з пакетів: R (RStudio), STATISTICA, MathCad, MathLab, Excel. В лабораторії 418а встановлені всі, при бажанні можна користуватись своїм ПК.

Найпоширенішим на даний момент інструментом для статистичних розрахунків є мова програмування R. Щоб зробити роботу з мовою зручнішою, доцільно використовувати RStudio — безкоштовне інтегроване середовище розробки для R.

Скачати мову програмування R:

https://cran.r-project.org

https://m7876.wiki.zoho.com/R-Installation-and-Administration.html

Скачати RStudio (R має бути встановленим раніше):

http://www.rstudio.com/products/rstudio/download/

Рекомендації щодо використання мови:

https://m7876.wiki.zoho.com/Introduction-to-R.html

Блог на русском с массой полезных советов

http://r-analytics.blogspot.ru/

Cookbook for R

http://www.cookbook-r.com/

Курси лекцій по R

Анализ данных в R

https://stepic.org/course/Анализ-данных-в-R-129/syllabus?module=1

Аналіз даних та статистичне виведення на мові R

http://courses.prometheus.org.ua/courses/IRF/Stat101/2016_T3/about

Introduction to R

https://www.datacamp.com

Підручники

Шипунов А.Б., Балдин Е.М. Наглядная статистика. Используем R.

Мастицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R. – Электронная книга.

http://r-analytics.blogspot.com

Ще одним інструментом для статистичних розрахунків на базі мови R є пакет STATISTICA, що має графічний інтерфейс.

Скачати trial-версію STATISTICA (ліцензія на місяць)

http://www.statsoft.ru/products/trial/

Рекомендації щодо використання:

http://hr-portal.ru/statistica/gl1/gl1.php

Підручники

Боровиков В.П. Прогнозирование в системе STATISTICA

http://www.statistica.ru/books/populyarnoe-vvedenie-v-analiz-dannih.php

Электронный учебник по статистике StatSoft

http://www.statsoft.ru/home/textbook/default.htm

ТЕМА 1. Описова статистика

ТЕОРЕТИЧНІ ВІДОМОСТІ

Вся досліджувана сукупність об'єктів (кількістю N) називається *генеральною сукупністю*. Частина об'єктів кількістю n ($n \le N$), випадково відібрана з генеральної сукупності, називається *вибірковою сукупністю* або *вибіркою*.

Нехай над випадковою величиною X проведено n незалежних випробувань, і x_1, x_2, ..., x_n – сукупність її спостережених значень. Цю сукупність називають *статистичним рядом*, а x_i – *варіантами*. Серед варіант можуть бути однакові. Якщо кожній варіанті поставити у відповідність її частоту (кількість повторень) n_i, а також відносну частоту $p_i^* = n_i / n$, а самі варіанти записати у зростаючому (спадаючому) порядку, одержана таким чином таблиця називається *варіаційним рядом*.

Статистичною (*емпіричною*) *функцією розподілу* $F^*(x)$ випадкової величини X називається закон зміни відносної частоти нерівності $X < x$ в даному статистичному матеріалі:

$$F^*(x) = P^*(X < x)$$

Якщо об'єм вибірки n великий, доцільно весь інтервал одержаних значень величини X розбити на часткові (як правило, рівні) інтервали (x_1, x_2), (x_2, x_3), .,(x_k, x_{k+1}). Кількість інтервалів m вибирається в залежності від обсягу вибірки в межах від 7 до 15. Ширину інтервалів можна оцінити за формулою $h = \dfrac{x_{\max} - x_{\min}}{1 + 3.32 \cdot \lg n}$, округливши отримане значення до найближчого зручного.

Позначимо через n_i число значень величини X, що потрапили в інтервал (x_i, x_{i+1}). Для кожного інтервалу визначимо також відносну частоту $p_i^* = n_i / n$. В результаті одержимо таблицю, яку називають *інтервальним варіаційним рядом*:

I	(x_1, x_2)	(x_2, x_3)	…	(x_i, x_{i+1})	…	(x_m, x_{m+1})
N	n_1	n_2	…	n_i	…	n_m
P^*	p_1^*	p_2^*	…	p_i^*	…	p_m

Якщо на кожному з відрізків $[x_i, x_{i+1}]$, як на основі, побудувати прямокутник, висота якого дорівнює p_i^*, одержимо фігуру, що називається *гістограмою*. Вважаємо при цьому, що крок $h = x_{i+1} - x_i = const$. Якщо наближати $h \to 0$, то гістограма буде все більше наближатися до деякої кривої. Ця крива є графіком функції $f^*(x)$ – *щільності розподілу* випадкової величини X.

До числових характеристик (описових статистик) вибірки відносять:

вибіркове середнє

для простого варіаційного ряду (k – кількість різних варіант)

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n};$$

для інтервального варіаційного ряду (m – кількість інтервалів групування, $x_{cp.i}$ – середина i-го інтервалу)

$$\bar{x} = \frac{\sum_{i=1}^{m} x_{cp.i} n_i}{n};$$

дисперсію

$$D = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i}{n} \text{ або } D = \frac{\sum_{i=1}^{m} (x_{cp.i} - \bar{x})^2 n_i}{n};$$

середньоквадратичне відхилення $\sigma = \sqrt{D}$;

моду - значення варіанти з максимальною частотою або для інтервального ряду

$$Mo = x_0 + h \cdot \frac{n_s - n_{s-1}}{(n_s - n_{s-1}) + (n_s - n_{s+1})}$$

де x_0 – початок модального інтервалу, h – крок інтервального ряду, n_s – частота модального інтервалу, n_{s-1} – частота інтервалу, що передує модальному, n_{s+1} – частота наступного за модальним інтервалу;

медіану - значення середнього елемента варіаційного ряду при непарній кількості елементів і середньоарифметичне двох середніх елементів при їхній парній кількості або для інтервального ряду

$$Me = x_0 + h \cdot \frac{n/2 - T_{s-1}}{n_s}$$

де x_0 – початок медіанного інтервалу, тобто інтервалу, в якому утримується середній елемент, h – крок інтервального ряду, n_s – частота медіанного інтервалу, T_{s-1} – сума частот інтервалів, які передують медіанному.

ЛАБОРАТОРНА РОБОТА №1. Описова статистика

Мета роботи: ознайомитись з методикою первинної обробки статистичних даних; проаналізувати вплив способу представлення даних на їх інформативність.

Запитання на допуск до роботи

1. Чим відрізняються генеральна та вибіркова сукупності?
2. Які бувають способи відбору даних?
3. Як побудувати статистичний розподіл вибірки?

4. Що відносять до числових характеристик вибірки?
5. Як визначається мода та медіана?
6. Які бувають способи графічного зображення статистичних розподілів?
7. Що таке емпірична функція розподілу та які її властивості?

ЗАВДАННЯ

Основне завдання

1. Розрахувати:
   1.1. варіаційний ряд для простої вибірки;
   1.2. інтервальний варіаційний ряд для згрупованої вибірки;
   1.3. числові характеристики для простої та згрупованої вибірки.
2. Побудувати для згрупованої вибірки:
   2.1. гістограму частот;
   2.2. емпіричну функцію розподілу.

Додаткове завдання

3. Зробити висновок про інформативність гістограми частот та емпіричної функції розподілу при різній кількості інтервалів групування (дослідити 3-4 варіанти).

Порядок виконання роботи

Створити вибірку відповідно завдання свого варіанту (для створення використовується програма DataGenerator). Завантажити вхідні дані в середовище, що буде використовуватись для їх обробки. Виконати необхідні розрахунки та оформити звіт.

ТЕМА 2. Статистичні оцінки параметрів розподілу

ТЕОРЕТИЧНІ ВІДОМОСТІ

Статистичною оцінкою θ^* невідомого параметра θ випадкової величини X генеральної сукупності (теоретичного розподілу X) називають функцію від випадкових величин (результатів вибірки), що спостерігаються

$$\theta^* = \theta^*(X_1, ..., X_n)$$

Щоб статистичні оцінки давали найкращі наближення параметрів, вони повинні задовольняти певним вимогам.

Статистична оцінка θ^* параметра θ має бути *незсунутою*, тобто $M(\theta^*) = \theta$, бо використання статистичної оцінки, математичне сподівання якої не дорівнює параметру θ, приводить до систематичних (одного знака) похибок.

Статистична оцінка θ^* параметра θ має бути *ефективною*, тобто такою, що при заданому об'ємі вибірки n має найменшу можливу дисперсію.

Статистична оцінка θ^* параметра θ має бути *обґрунтованою*, тобто такою, що при $n \to \infty$ прямує за імовірністю до оцінюваного параметра.

Точковими оцінками параметрів розподілу генеральної сукупності називають такі оцінки, які визначаються одним числом. Основні характеристики вибіркового розподілу ($\bar{x}$, D, σ) є точковими оцінками відповідних параметрів генеральної сукупності: генеральних середнього, дисперсії й середньоквадратичного відхилення. Оскільки вибіркова дисперсія є зміщеною оцінкою генеральної дисперсії, користуються *виправленою дисперсією*

$$s^2 = \frac{n}{n-1} D$$

і відповідним середньоквадратичним відхиленням s.

Визначення точкових оцінок методом моментів

Метод запропоновано К.Пірсоном. Він заснований на тому, що початкові та центральні емпіричні моменти є обґрунтованими оцінками відповідних початкових та центральних теоретичних моментів того ж порядку. Перевага методу – простота.

Нехай випадкова величина задана щільністю розподілу $f(x,\theta_1,\ldots,\theta_r)$, де $\theta_1,\ldots,\theta_r$ - невідомі параметри розподілу. Потрібно знайти точкові оцінки $\theta_1^*,\ldots,\theta_r^*$ цих параметрів, для чого необхідно сформувати r незалежних умов.

Суть методу моментів полягає в тому, що такі умови отримують прирівнюванням довільних r теоретичних моментів (початкових або центральних) відповідним їм емпіричним моментам, які визначаються за вибіркою.

$$\nu_k(\theta_1,\ldots,\theta_r) = \nu_k^*, \; k = \overline{1,m}$$
$$\mu_s(\theta_1,\ldots,\theta_r) = \mu_s^*, \; s = \overline{m+1,r}$$

Вимоги до рівнянь системи:

1. Рівняння повинні бути інформативними (не можна використовувати ν_0, μ_0, μ_1);
2. Рівняння повинні бути незалежними (якщо використано ν_1, ν_2, то не можна використовувати μ_2, бо $\mu_2 = \nu_2 - \nu_1^2$).

Визначення точкових оцінок методом найбільшої подібності

Метод запропоновано Р.Фішером. Нехай X – дискретна випадкова величина, яка в результаті n випробувань набувала значень $x_1,\ldots, x_n$. Припустимо, що вигляд закону розподілу випадкової величини задано, але

невідомо параметр θ, яким визначається цей закон. Потрібно знайти його точкову оцінку.

Позначимо через $p(x_i, \theta)$ імовірність того, що в результаті випробування випадкова величина X прийме значення x_i (i=1, 2,..,n). *Функцією правдоподібності* дискретної випадкової величини X називають функцію аргументу θ, де $x_1, \ldots, x_n$ – фіксовані числа.

$$L(x_1, x_2, \ldots, x_n, \theta) = p(x_1, \theta) \cdot p(x_2, \theta) \cdot \ldots \cdot p(x_n, \theta)$$

Функцією правдоподібності неперервної випадкової величини X називають функцію

$$L(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \ldots \cdot f(x_n, \theta)$$

де $f(x_i, \theta)$ - щільність розподілу в точці x_i.

В якості точкової оцінки параметра θ приймають таке його значення $\theta^* = \theta^*(x_1, \ldots, x_n)$, при якому функція правдоподібності досягає максимуму. Функції L та $\ln L$ досягають максимуму при одному й тому ж значенні θ, тому замість пошуку максимуму функції L шукають (що зручніше) максимум функції $\ln L$.

Визначення інтервальних оцінок

Для вибірок невеликого об’єму точкові оцінки можуть значно відрізнятися від оцінюваного параметра. Тому у випадку малих вибірок застосовують інтервальні оцінки. Інтервальну оцінку даного параметра знаходять, враховуючи задану *надійність* або *довірчу ймовірність*, тобто ймовірність, з якою шуканий інтервал покриває дійсне значення оцінюваного параметра. Одержаний за таких умов інтервал називають *довірчим інтервалом* для оцінюваного параметра.

Нехай задано вибірку об’єму n із деякої генеральної сукупності, основні параметри якої невідомі. Побудову довірчих інтервалів для параметрів починають з визначення їх точкових оцінок: $\bar{x}$, s^2.

Загальна методика визначення інтервальних оцінок параметрів розподілу:

1. Необхідно вибрати підходящу статистику Y для побудови довірчого інтервалу.

   $Y = Y(\theta^*, \theta, \beta)$

   Y – підходяща статистика, якщо:
   - містить параметр θ;
   - містить оцінку параметра θ^*;
   - якщо містить інші параметри β, то вони відомі;
   - має відомий ЗРІ (бажано табличний)

2. За вибраним (заданим) $P_{\text{дов}}$ та законом розподілу статистики Y визначити довірчий інтервал (y_1,y_2), в який потрапить статистика Y з імовірністю $P_{\text{дов}}$ при виконанні експерименту.
3. За виразом $Y=Y(\theta^*, \theta, \beta)$ знайти в загальному вигляді інтервал (θ_1^*, θ_2^*), в який потрапить θ^* при виконанні експерименту.
4. Виконати експеримент, за вибіркою $\{x\}_n$ визначити емпіричне значення $y_{\text{ем}}$ статистики Y та підставити його в п.3.

Довірчий інтервал для математичного сподівання нормального розподілу при відомому σ розраховується за допомогою статистики:

$$Z = Z(\overline{X}, a, \sigma_{\overline{X}}) = \frac{\overline{X} - a}{\sigma_{\overline{X}}} = \frac{\overline{X} - a\sqrt{n}}{\sigma_X}$$

що має нормальний розподіл з параметрами $f(z,0,1)$, тобто це таблична функція Лапласа. За заданим $P_{\text{дов}}$ та законом розподілу статистики Z визначаємо довірчий інтервал (z_1, z_2), в який потрапить статистика Z з імовірністю $P_{\text{дов}}$ при виконанні експерименту. Закон розподілу статистики Z - симетрична функція, тому довірчий інтервал буде найвужчим при $z_1=z_2$.

$$\bar{x} - \frac{z_{Рдов/2}\sigma_X}{\sqrt{n}} < a < \bar{x} + \frac{z_{Рдов/2}\sigma_X}{\sqrt{n}}$$

Довірчий інтервал для математичного сподівання нормального розподілу при невідомому σ розраховується за допомогою статистики T, що має розподіл Стьюдента з $k = n$-1 степенями свободи:

$$T = \frac{\overline{X} - a}{S/\sqrt{n}}$$

Закон розподілу статистики T - симетрична функція, тому довірчий інтервал буде найвужчим при $t_1= t_2$.

$$P\left(\frac{\left|\overline{X} - a\right|\sqrt{n}}{s} < t_\gamma\right) = 2\int_0^{t_\gamma} T(t,n)dt = P_{дов}$$

$$\bar{x}_B - \frac{t_{Рдов}s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_{Рдов}s}{\sqrt{n}}$$

Довірчий інтервал для оцінки середньоквадратичного відхилення σ нормального розподілу розраховується за допомогою статистики χ^2.

$$\chi^2 = \sum_{i=1}^{n} Z_i^2$$

де Z_i (i =1, 2,…, n) – нормальні, нормовані незалежні величини. Ця статистика розподілена по закону χ^2 з $k = n$-1 степенями свободи. За заданим $P_{\text{дов}}$ та законом розподілу статистики χ^2 визначаємо довірчий інтервал, в який потрапить статистика χ^2 з імовірністю $P_{\text{дов}}$ при виконанні експерименту.

$$\sqrt{\frac{s^2(n-1)}{\chi_2^2}} < \sigma < \sqrt{\frac{s^2(n-1)}{\chi_1^2}}$$

Практично для знаходження меж інтервалу користуються таблицею значень q:

$$s(1-q) < \sigma < s(1+q)$$

ЛАБОРАТОРНА РОБОТА №2. Точкові оцінки параметрів розподілу

Мета роботи: ознайомитись з методами визначення точкових оцінок параметрів розподілу; дослідити, що впливає на якість точкових оцінок.

Запитання на допуск до роботи

1. Назвіть основні вимоги до статистичних оцінок.
2. Як визначається проста середньоарифметична вибірки?
3. Як визначається вибіркова середня або зважена середньоарифметична вибірки та які її властивості?
4. Як визначається степенева середня вибірки?
5. Як визначається вибіркова дисперсія та вибіркове середньоквадратичне відхилення? У яких випадках потрібна виправлена вибіркова дисперсія і як її визначити?
6. Для чого потрібні початкові та центральні моменти вибірки та як їх визначити?

ЗАВДАННЯ

Основне завдання

1. Визначити точкові оцінки параметрів розподілу (вважаємо, що випадкова величина розподілена нормально):
   1.1. методом моментів;
   1.2. методом найбільшої подібності.
2. Порівняти отримані оцінки та вказати, яка з них краща та чому.

Додаткове завдання

3. Дослідити залежність оцінок від об’єму вибірки. Зробити висновки про виконання закону великих чисел.

Порядок виконання роботи

Вхідні дані взяти з першої роботи. Розрахунки можна проводити, продовжуючи файл попередньої роботи.

ЛАБОРАТОРНА РОБОТА №3. Інтервальні оцінки параметрів розподілу

Мета роботи: ознайомитись з методикою визначення інтервальних оцінок параметрів розподілу; дослідити, що впливає на якість інтервальних оцінок.

Запитання на допуск до роботи

1. В чому різниця між точковими та інтервальними оцінками параметрів ЗРВ?
2. Що таке надійність (довірча імовірність) оцінки параметра ЗРВ?
3. Чим визначається та для чого використовується довірчий інтервал?
4. Якою повинна бути підходяща статистика для визначення інтервальної оцінки?
5. Порядок визначення інтервальних оцінок математичного сподівання.
6. Порядок визначення інтервальних оцінок дисперсії.

ЗАВДАННЯ

Основне завдання

1. Визначити інтервальні оцінки математичного сподівання та середньоквадратичного відхилення при рівні довіри $P_{\text{дов}} = 0.95$.

Додаткове завдання

2. Дослідити залежність оцінок від рівня $P_{\text{дов}}$.
3. Дослідити залежність оцінок від об’єму вибірки.

Порядок виконання роботи

Вхідні дані взяти з першої роботи. Розрахунки можна проводити, продовжуючи файл попередньої роботи.

ТЕМА 3. Перевірка статистичних гіпотез

ТЕОРЕТИЧНІ ВІДОМОСТІ

Якщо закон розподілу генеральної сукупності невідомий, але є підстави припускати, що він має деякий визначений вид A, то перевіряють нульову гіпотезу: генеральна сукупність розподілена за законом A. Можливі й інші гіпотези: про значення параметру розподілу, про рівність параметрів двох або більше розподілів, про незалежність вибірок тощо. Разом з основною гіпотезою завжди можна розглядати протилежну їй гіпотезу, яку називають альтернативною.

При перевірці статистичної гіпотези за даними випадкової вибірки можна зробити хибний висновок. При цьому можуть бути похибки першого та другого роду. Якщо за висновком буде відкинута правильна гіпотеза, то

кажуть, що це *похибка першого роду*. Якщо за висновком буде прийнята неправильна гіпотеза, то кажуть, що це *похибка другого роду*. Імовірність здійснити похибку першого роду називають *рівнем значущості* α. Найчастіше рівень значущості приймають рівним 0.05 або 0.01.

Перевірка гіпотези здійснюється за допомогою спеціально підібраної величини K – критерію узгодження. Для кожного критерію узгодження є відповідні таблиці. Після обрання певного критерію узгодження, множину усіх його можливих значень поділяють на дві підмножини, що не перетинаються: одна з них містить значення критерію, при яких основна гіпотеза відхиляється, а друга – при яких вона приймається.

Критичною областю називають сукупність значень критерію, при яких основна гіпотеза відхиляється.

Областю прийняття гіпотези (областю допустимих значень) називають множину значень критерію, при яких гіпотезу приймають.

Розрізняють однобічну (правобічну та лівобічну) та двобічну критичні області. Щоб знайти межі критичної області треба розрахувати критичні точки k_{kp}. Для цього задають рівень значущості α, а потім шукають критичні точки з врахуванням вимоги:

$P(K > k_{kp}) = \alpha$ у випадку правобічної критичної області;

$P(K < k_{kp}) = \alpha$ у випадку лівобічної критичної області;

$P(K < k_1) + P(K > k_2) = \alpha$ у випадку двобічної критичної області.

При знаходженні критичної області доцільно враховувати *потужність критерію* – імовірність належності критерію критичній області при умові, що правильна альтернативна гіпотеза.

Коли критична точка вже знайдена, за даними вибірок обчислюють спостережене значення критерію і, якщо виявиться, що $K_{\text{сп}} > k_{kp}$ то нульову гіпотезу відкидають (у випадку правосторонньої критичної області); якщо ж $K_{\text{сп}} < k_{kp}$ то немає підстав, щоб відкинути нульову гіпотезу.

Порядок дій при перевірці статистичних гіпотез

1) сформулювати гіпотезу H_0;
2) обрати статистичну характеристику – критерій узгодження для перевірки гіпотези;
3) визначити або задати допустиму імовірність похибки першого роду – рівень значущості α;
4) визначити гіпотезу H_1, альтернативну до гіпотези H_0;
5) обрати вигляд критичної області з урахуванням наявності альтернативної гіпотези;

6) знайти за відповідною таблицею критичну область (критичну точку) для обраної статистичної характеристики;

7) за вибіркою обчислити емпіричне значення критерію, порівняти його з критичним значенням та зробити висновок про прийняття чи відхилення гіпотези H_0.

Перевірка гіпотез про нормальний розподіл генеральної сукупності. Критерій узгодження Пірсона (χ^2)

Перевірка полягає в порівнянні емпіричних (спостережених) і теоретичних (обчислених за припущенням наявності нормального розподілу) частот. Як критерій перевірки нульової гіпотези (H_0: генеральна сукупність розподілена за нормальним законом) приймається випадкова величина χ^2

$$\chi^2 = \sum_{i=1}^{m} \left(n_i - n_i' \right)^2 / n_i'$$

де n_i – емпіричні частоти; n_i' – теоретичні частоти нормального розподілу; m – кількість інтервалів варіаційного ряду, побудованого за даними вибірки.

Очевидно, що чим менше відрізняються емпіричні та теоретичні частоти, тим менше значення приймає критерій, тому обираємо правосторонню критичну область. Критичне значення $\chi^2_{\text{крит}}(\alpha; k)$ знаходять за таблицею критичних точок розподілу χ^2 з використанням двох параметрів – заданого рівня значущості α і числа степенів свободи $k = m\text{-}1\text{-}r$ (m – кількість інтервалів варіаційного ряду, r – кількість параметрів розподілу, визначених за даними вибірки, для нормального закону розподілу r=2). У випадку, якщо спостережене значення критерію $\chi^2_{\text{спост}}$ виявиться меншим, ніж $\chi^2_{\text{крит}}$, то немає підстав відкидати нульову гіпотезу. Якщо ж $\chi^2_{\text{спост}}$ буде більшим за $\chi^2_{\text{крит}}$, то нульову гіпотезу слід відкинути й прийняти альтернативу H_1: розподіл генеральної сукупності не відповідає нормальному закону.

Критерій узгодження Колмогорова

Перевірка полягає в порівнянні емпіричних і теоретичних значень функції розподілу. Як критерій перевірки нульової гіпотези про вигляд розподілу приймемо статистику $D = \left| F_{ем}(x) - F_{т}(x) \right|$. Критерій Колмогорова має вигляд

$$\lambda(n) = \sqrt{m} \cdot \max(D(m))$$

тобто береться до уваги максимальна різниця між значеннями емпіричної та теоретичної функцій розподілу. Очевидно, що чим менше відрізняються емпіричні та теоретичні значення, тим менше значення приймає критерій, тому обираємо правосторонню критичну область. Критичне значення знаходять за таблицею критичних точок розподілу по заданому рівню

значущості α та кількості інтервалів групування m (або обсягу n для простої вибірки).

m \ α	0.20	0.10	0.05	0.01	m \ α	0.20	0.10	0.05	0.01
1	0.900	0.950	0.975	0.995	25	1.040	1.188	1.320	1.583
2	0.967	1.097	1.191	1.314	30	1,041	1.194	1.325	1.588
3	0.978	1.102	1.226	1.436	40	1.044	1.197	1.328	1.594
4	0.986	1.130	1.248	1.468	50	1.047	1.200	1.330	1.598
5	0.998	1.139	1.260	1.495	60	1.051	1.201	1.332	1.603
6	1.004	1.146	1.271	1.511	70	1.053	1.203	1.335	1.606
8	1.012	1.159	1.284	1.533	80	1.055	1.204	1.338	1.607
10	1.021	1.167	1.293	1.546	90	1.056	1.206	1.340	1.608
15	1.030	1.177	1.309	1.565	100	1.056	1.207	1.340	1.608
20	1.038	1.185	1.315	1.574	∞	1.073	1.224	1.358	1.627

Критерій узгодження Мізеса

Перевірка полягає в порівнянні емпіричних і теоретичних значень функції розподілу. Як критерій перевірки нульової гіпотези про вигляд розподілу приймемо статистику

$$\omega^2 = \int_{-\infty}^{\infty} \left(F_{ем}(x) - F_г(x)\right)^2 f_г(x)dx$$

Критерій Мізеса має вигляд $\mu = n \cdot \omega^2$

тобто беруться до уваги всі різниці між значеннями емпіричної та теоретичної функцій розподілу та додатково зважуються функцією щільності. Очевидно, що чим менше відрізняються емпіричні та теоретичні значення, тим менше значення приймає критерій, тому обираємо правосторонню критичну область. Критичне значення знаходять за таблицею критичних точок розподілу по заданому рівню значущості α та кількості інтервалів групування m (або обсягу n для простої вибірки).

α	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
$\mu_{кр}$	0.118	0.147	0.185	0.241	0.347	0.461	0.620	0.744	1.168

Перевірка гіпотез про параметри розподілу генеральної сукупності.

Перевірка гіпотези про рівність виправленої вибіркової дисперсії гіпотетичній генеральній дисперсії нормальної сукупності

Нехай генеральна сукупність розподілена нормально, при цьому генеральна дисперсія хоча і невідома, але можна припустити, що вона дорівнює певному гіпотетичному значенню σ_0^2. На практиці σ_0^2 встановлюється на підставі попереднього досвіду або теоретично. Із

генеральної сукупності зробили вибірку об'єму n і знайшли виправлену дисперсію S^2. Потрібно по виправленій дисперсії при заданому рівні значущості перевірити гіпотезу H_0: генеральна дисперсія сукупності дорівнює гіпотетичному значенню σ_0^2. Враховуючи, що виправлена дисперсія S^2 є незсунутою оцінкою генеральної дисперсії, нульову гіпотезу можна записати так: $H_0 : M(S^2) = \sigma_0^2$

Як критерій перевірки нульової гіпотези приймемо відношення виправленої дисперсії до гіпотетичного значення генеральної дисперсії, тобто випадкову величину

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Величина має розподіл χ^2 із $k=n$-1степенями свободи. Критична область будується в залежності від вигляду альтернативної гіпотези.

Перевірка гіпотези про рівність вибіркової середньої генеральній середній нормальної сукупності, дисперсія якої відома

Нехай генеральна сукупність розподілена нормально, причому генеральна середня a хоч і невідома, але є підстави вважати, що вона дорівнює гіпотетичному значенню a_0. Дисперсія генеральної сукупності відома, наприклад, з попередніх дослідів або обчислена теоретично. Із сукупності зробили вибірку об'єму n та знайшли вибіркову середню $\bar{x}$. Потрібно по вибірковій середній при заданому рівні значущості перевірити гіпотезу про рівність генеральної середньої a гіпотетичному значенню a_0. Враховуючи, що вибіркова середня є незсунутою оцінкою генеральної середньої, нульову гіпотезу можна записати так: $H_0 : M(\bar{X}) = a_0$

В якості критерію перевірки нульової гіпотези про рівність вибіркових середніх приймемо випадкову величину

$$U = \frac{\bar{X} - a_0}{\sigma(\bar{X})} = \frac{(\bar{X} - a_0)\sqrt{n}}{\sigma}$$

Величина U має нормальний нормований розподіл, $M(U)=0$, $\sigma(U)=1$ при справедливості нульової гіпотези. Критична область будується в залежності від вигляду альтернативної гіпотези.

Перевірка гіпотези про рівність вибіркової середньої генеральній середній нормальної сукупності, дисперсія якої невідома

Нехай генеральна сукупність розподілена нормально, причому генеральна середня a хоч і невідома, але є підстави вважати, що вона дорівнює гіпотетичному значенню a_0. Із сукупності зробили вибірку об'єму n та знайшли вибіркову середню $\bar{x}$ та виправлену вибіркову дисперсію s^2. Потрібно по вибірковій середній при заданому рівні значущості перевірити

гіпотезу про рівність генеральної середньої a гіпотетичному значенню a_0. Враховуючи, що вибіркова середня $\bar{x}$ є незсунутою оцінкою генеральної середньої, нульову гіпотезу можна записати так: $H_0 : M(\bar{x}) = a_0$

В якості критерію перевірки нульової гіпотези про рівність вибіркових середніх приймемо випадкову величину

$$T = (\bar{x} - a_0)\sqrt{n}/S$$

Величина T має розподіл Стьюдента при з $k=n-1$ степенями свободи. Критична область будується в залежності від вигляду альтернативної гіпотези.

ЛАБОРАТОРНА РОБОТА №4. Перевірка статистичних гіпотез про вигляд закону розподілу

Мета роботи: ознайомитись з методами перевірки статистичних гіпотез про вигляд закону розподілу; дослідити, що впливає на ширину критичної області.

Запитання на допуск до роботи

1. Що називають статистичною гіпотезою?
2. Перерахуйте різновиди статистичні гіпотез.
3. Які бувають похибки перевірки гіпотез? Чим вони відрізняються?
4. Що таке критична область і від чого залежить її вигляд?
5. Що впливає на ширину критичної області?
6. Які критерії узгодження можна використовувати для перевірки гіпотез про вигляд розподілу?
7. Перерахуйте обмеження, які накладаються на використання кожного з критеріїв.

ЗАВДАННЯ

Основне завдання

1. Перевірити гіпотези про вигляд закону розподілу при рівні значущості $\alpha=0.05$ за кількома різними критеріями.
2. Якщо за різними критеріями результати перевірки відрізняються, то зробити загальний висновок про прийняття чи спростування гіпотези про вигляд розподілу.
3. Порівняти отримані оцінки та вказати, яка з них краща та чому.

Додаткове завдання

4. Дослідити залежність одного з критеріїв від кількості інтервалів групування, рівня значущості та об’єму вибірки.

Порядок виконання роботи

Вхідні дані взяти з першої роботи. Розрахунки можна проводити, продовжуючи файл попередньої роботи.

ЛАБОРАТОРНА РОБОТА №5. Перевірка статистичних гіпотез про параметри розподілу

Мета роботи: ознайомитись з методами перевірки статистичних гіпотез про параметри закону розподілу; дослідити, що впливає на ширину критичної області.

Запитання на допуск до роботи

1. Які критерії узгодження можна використовувати для перевірки гіпотез про параметри розподілу?
2. Перерахуйте обмеження, які накладаються на використання кожного з критеріїв.

ЗАВДАННЯ

Основне завдання

1. Перевірити гіпотези про параметри розподілу при рівні значущості α=0.05 (альтернативну гіпотезу задати самостійно):
   1.1. гіпотеза про математичне сподівання;
   1.2. гіпотеза про середньоквадратичне відхилення.

Додаткове завдання

2. Дослідити залежність критеріїв від рівня значущості та об’єму вибірки.

Порядок виконання роботи

Вхідні дані взяти з першої роботи. Розрахунки можна проводити, продовжуючи файл попередньої роботи.

ТЕМА 4. Дисперсійний аналіз

ТЕОРЕТИЧНІ ВІДОМОСТІ

Нехай генеральні сукупності X_1, X_2, ..., X_p розподілені нормально і мають однакову, хоч і невідому, дисперсію; математичні сподівання також невідомі, але можуть бути різними. Потрібно при заданому рівні значущості по вибірковим середнім перевірити нульову гіпотезу про рівність всіх математичних сподівань

$$H_0: M(X_1) = M(X_2) = ... = M(X_p)$$

Іншими словами, потрібно встановити, значуще чи незначуще розрізняються вибіркові середні.

На практиці дисперсійний аналіз застосовують, щоб встановити, чи істотно впливає певний якісний фактор F, який має p рівнів F_1, F_2, ..., F_p на величину X, що вивчається.

Основна ідея дисперсійного аналізу полягає в порівнянні «дисперсії фактора», що породжується його дією, і «залишкової дисперсії», обумовленої випадковими причинами. Якщо відмінність між цими дисперсіями значуща, то фактор має істотний вплив на X; в цьому випадку середні спостережуваних значень на кожному рівні (групові середні) відрізняються також значуще. Якщо вже встановлено, що фактор істотно впливає на X, а потрібно з'ясувати, який з рівнів має найбільший вплив, то додатково виконують попарне порівняння середніх.

Нехай на кількісну нормально розподілену ознаку X впливає фактор F, що має p постійних рівнів. Вважаємо, що кількість спостережень (випробувань) на кожному рівні однакове і дорівнює q. Нехай спостережувалось $n=pq$ значень x_{ij} ознаки X, де i – номер випробування (i=1,2,…,q), j – номер рівня фактора (j=1,2,…,p).

Номер випробування	Рівні фактора F_j			
	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
...	…	…	…	
q	x_{q1}	x_{q2}	…	x_{qp}
Групова середня	$\bar{x}_{\text{гр}1}$	$\bar{x}_{\text{гр}2}$	…	$\bar{x}_{\text{гр}p}$

Позначимо загальну суму квадратів відхилень спостережених значень від загальної середньої $\bar{x}$

$$S_{загальне} = \sum_{j=1}^{p}\sum_{i=1}^{q}(x_{ij} - \bar{x})^2$$

факторну суму квадратів відхилень спостережених значень від загальної середньої, яка характеризує розсіювання між групами

$$S_{факт} = q\sum_{j=1}^{p}(x_{ГРj} - \bar{x})^2$$

залишкову суму квадратів відхилень спостережених значень від своєї групової середньої, яка характеризує розсіювання всередині груп

$$S_{зал} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij} - \bar{x}_{ГРj}\right)^2 = S_{загальне} - S_{факт}$$

Поділивши суми квадратів відхилень на відповідне число степенів свободи, отримаємо загальну, факторну та залишкову дисперсії

$$s^2_{загальне} = \frac{S_{загальне}}{pq-1} \qquad s^2_{факт} = \frac{S_{факт}}{p-1} \qquad s^2_{зал} = \frac{S_{зал}}{p(q-1)}$$

де p – число рівнів фактора, q – число спостережень на кожному рівні, (pq-1) – число степенів свободи загальної дисперсії, (p-1) – число степенів свободи факторної дисперсії, p(q-1) – число степенів свободи залишкової дисперсії.

Якщо гіпотеза про рівність середніх справедлива, то всі ці дисперсії є незсунутими оцінками генеральної дисперсії.

Для того, щоб перевірити нульову гіпотезу про рівність групових середніх нормальних сукупностей з однаковими дисперсіями, достатньо перевірити по критерію Фішера гіпотезу про рівність факторної та залишкової дисперсії.

$$F = \frac{s^2_{факт}}{s^2_{зал}}$$

Критичну область у цьому випадку знаходять з урахуванням умови

$$P\left(F > f_\alpha\right) = \alpha$$

де f_a – критичне (і табульоване) значення розподілу Фішера з (p-1) та p(q-1) степенями свободи.

Якщо факторна дисперсія виявиться менше залишкової, то вже звідси слідує справедливість гіпотези про рівність групових середніх і, значить, немає потреби вдаватися до критерію F.

Якщо немає впевненості в справедливості припущення про рівність дисперсій p сукупностей, що розглядаються, то це припущення слід перевірити заздалегідь, наприклад по критерію Кохрена.

ЛАБОРАТОРНА РОБОТА №6. Однофакторний дисперсійний аналіз

Мета роботи: ознайомитись з методикою проведення однофакторного дисперсійного аналізу; навчитись використовувати дисперсійний аналіз для виявлення факторів, що суттєво впливають на тривалість роботи програми.

Запитання на допуск до роботи

1. Для яких задач використовується дисперсійний аналіз?
2. Що розуміють під фактором?
3. Яка різниця між дисперсією фактора, залишковою та загальною дисперсіями?

4. Яка статистика використовується для дисперсійного аналізу?

ЗАВДАННЯ

Основне завдання

1. Створити та завантажити вхідні дані.
2. Виконати однофакторний дисперсійний аналіз.
3. Зробити висновок про суттєвість впливу досліджуваного фактора.

Порядок виконання роботи

Для створення вхідних даних потрібно провести серію експериментів з програмою сортування одномірного масиву одним із методів. В цьому випадку Y – випадкова величина, час роботи програми; X – невипадкова величина, довільний фактор, що може впливати на час роботи програми (обсяг вхідних даних, або їх попередня впорядкованість, або кількість однакових значень тощо). Варіант завдання взяти з таблиці.

На кількісний розподілений за нормальним законом параметр Y впливає фактор X, котрий має m постійних рівнів (задати довільно, від 4 до 6). Кількість спостережень на кожному рівні однакова і дорівнює n (задати довільно, але не менше 10). Спостерігалося $m \cdot n$ значень $y_{i,j}$ параметру Y, де j – номер спостереження (j=1..n), і – номер рівня фактора (i=1..m). Результати спостережень запишіть у файл. Якщо переглядати результати спостережень, то вони мають форму таблиці, у рядках якої наведені спостереження для одного рівня фактору.

Варіанти завдань

№	метод сортування	фактор	№	метод сортування	фактор
1	обміну (бульбашковий)	обсяг даних	16	метод розподіляючого підрахунку	обсяг даних
2	обміну (бульбашковий)	попередня впорядкованість	17	метод розподіляючого підрахунку	попередня впорядкованість
3	обміну (бульбашковий)	кількість однакових значень	18	метод розподіляючого підрахунку	кількість однакових значень
4	вставки	обсяг даних	19	сортування злиттям	обсяг даних
5	вставки	попередня впорядкованість	20	сортування злиттям	попередня впорядкованість
6	вставки	кількість однакових значень	21	сортування злиттям	кількість однакових значень

7	вибору	обсяг даних	22	пірамідальне сортування	обсяг даних
8	вибору	попередня впорядкованість	23	пірамідальне сортування	попередня впорядкованість
9	вибору	кількість однакових значень	24	пірамідальне сортування	кількість однакових значень
10	швидке сортування	обсяг даних	25	порозрядне сортування	обсяг даних
11	швидке сортування	попередня впорядкованість	26	порозрядне сортування	попередня впорядкованість
12	швидке сортування	кількість однакових значень	27	порозрядне сортування	кількість однакових значень
13	метод Шелла	обсяг даних	28	сортування злиттям з відсіканням файлів невеликих розмірів	обсяг даних
14	метод Шелла	попередня впорядкованість	29	сортування злиттям з відсіканням файлів невеликих розмірів	попередня впорядкованість
15	метод Шелла	кількість однакових значень	30	сортування злиттям з відсіканням файлів невеликих розмірів	кількість однакових значень

ТЕМА 5. Задача регресії

ТЕОРЕТИЧНІ ВІДОМОСТІ

Залежність між випадковими величинами X і Y може бути функціональною, статистичною та кореляційною.

Якщо кожному можливому значенню випадкової величини X відповідає одне можливе значення випадкової величини Y, то кажуть, що існує *функціональна* залежність Y від X:

$$Y = Y(X).$$

Статистичною називається залежність, при якій зі зміною однієї з випадкових величин випливає зміна закону розподілу другої випадкової величини.

Умовним середнім $\bar{y}_x$ називається середнє арифметичне значення Y, що відповідає значенню $X = x$. Аналогічно визначається умовне середнє $\bar{x}_y$.

Статистичну залежність називають *кореляційною*, якщо при зміні однієї величини змінюється середнє значення другої величини. Отже, кореляційна

залежність *Y* від *X* - це функціональна залежність умовної середньої $\bar{y}_x$ від *x*, тобто

$$\bar{y}_x = f(x)$$

Це рівняння називається *рівнянням регресії Y* на *X*, а графік функції *f*(*x*) - *лінією регресії*. Аналогічно визначається кореляційна залежність *X* від *Y*.

$$\bar{x}_y = \varphi(y)$$

Основними задачами теорії кореляції є встановлення форми кореляційної залежності, тобто вигляду функції регресії та оцінка тісноти (сили) кореляційного зв'язку.

Якщо обидві функції, *f*(*x*) і $\phi(y)$, - лінійні, то кореляційну залежність (регресію) називають *лінійною*. Метою процедур лінійної регресії є підгонка прямої лінії до деякого набору точок так, щоб мінімізувати квадрати відхилень цієї лінії від спостережуваних точок.

Існують різноманітні види регресійного аналізу - одномірний і багатомірний, лінійний і нелінійний, параметричний і непараметричний.

Для проведення лінійного регресійного аналізу залежна змінна повинна мати інтервальну (або порядкову) шкалу. Бінарна логістична регресія виявляє залежність дихотомічної змінної від якоїсь іншої змінної, що виміряна за будь-якою шкалою. Якщо залежна змінна є категоріальною, але має більше двох категорій, то тут відповідним методом буде поліноміальна логістична регресія. Порядкову регресію можна використовувати, коли залежні змінні виміряні за порядковою шкалою. І, звичайно ж, можна аналізувати і нелінійні зв'язки між змінними, які виміряні за інтервальною шкалою. Для цього призначений метод нелінійної регресії.

Тісноту кореляційного зв'язку між *X* і *Y* можна оцінювати за величиною розсіювання значень *y* навколо умовних середніх $\bar{y}_x$ і значень *x* навколо умовних середніх $\bar{x}_y$.

Рівняння лінійної регресії *Y* на *X* має вигляд

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + ... + b_k \cdot X_k$$

або у випадку однієї незалежної змінної

$$\bar{y}_x - \bar{y} = r_в \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

де $\bar{y}_x$ - умовна середня; $\bar{x}$, $\bar{y}$ - вибіркові середні значення компонент *X* і *Y*; σ_x, σ_y - вибіркові середньоквадратичні відхилення; $r_в$ - вибірковий коефіцієнт кореляції.

Одержане значення вибіркового коефіцієнта кореляції $r_в$ аналізують, використовуючи правило: чим ближче $|r_в|$ до 1, тим тісніший зв'язок між X і Y.

ЛАБОРАТОРНА РОБОТА №7. Оцінка параметрів рівняння лінійної регресії

Мета роботи: ознайомитись з методикою оцінки параметрів рівняння прямої лінії середньоквадратичної регресії; випробувати її для прогнозування.

Запитання на допуск до роботи

1. У чому полягає задача регресії?
2. Які різновиди цієї здачі ви знаєте?
3. Методика оцінки параметрів рівняння лінійної регресії.
4. Якою повинна бути стратегія проведення експериментів (відбору вхідних даних) щоб оцінка параметрів рівняння прямої була якнайкращою?

ЗАВДАННЯ

Основне завдання

1. Визначити точкові оцінки параметрів рівняння лінійної регресії, записати отримане рівняння та побудувати графік.
2. Зробити прогноз для кількох значень за побудованою моделлю та порівняти його з фактичними значеннями.
3. Зробити висновки про якість моделі.

Додаткове завдання

4. Визначити інтервальні оцінки параметрів рівняння лінійної регресії при рівні довіри $P_{\text{дов}} = 0.95$.
5. Дослідити залежність інтервальних оцінок параметрів рівняння регресії від стратегії експерименту (значень незалежної змінної, на базі яких розраховане рівняння регресії).

Порядок виконання роботи

Створити вибірку відповідно завдання свого варіанту. Завантажити вхідні дані. Керуючись здоровим глуздом, вибрати незалежну (фактор) та залежну (відгук) величини. Побудувати діаграму розсіювання з розрахованою лінією регресії та зробити висновок про силу зв'язку між величинами.