

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ**

Кафедра математичних методів захисту інформації

ЗВІТ

З ВИРОБНИЧОЇ ПРАКТИКИ

Напрямок підготовки: 6.040301 «Прикладна математика»

Тема: «Розробка автоматизованого тестуючого комплексу,
що враховує психологічні особливості студентів»

Виконав студент 4 курсу

групи ФІ-13

Кригін Валерій Михайлович

Науковий керівник:

Доктор фізико-математичних наук, професор

Дороговцев Андрій Анатолійович

(підпис)

Київ 2015

ЗМІСТ

1 Вступ	3
1.1 Обґрунтування та актуальність роботи	3
1.2 Мета та завдання	3
2 Основна частина	4
2.1 Теоретичні відомості	4
2.1.1 Метод головних компонент	4
2.1.2 Гістограма	7
2.1.3 Поліноміальний розподіл.	8
2.1.4 Критерій узгодженості Пірсона χ^2	10
Перелік посилань	13

1 ВСТУП

1.1 Обґрунтування та актуальність роботи

Існуючі на даний момент системи тестування недостатньо гнучкі: вони аналізують лише відповіді на запитання, відносячи їх до вірних або невірних, а на цій базі роблять кінцевий висновок щодо знань студента. Стрімкий розвиток комп'ютерної техніки й інформаційних технологій надає можливість визначати ритм складання тесту, а також індивідуальні особливості людини. Дані психологічних досліджень допоможуть правильно трактувати отримані значення, а добре вивчені та перевірені часом математичні методи надають великі можливості для систематизації та обробки результатів вимірювання.

1.2 Мета та завдання

Завдання наступні:

- 1) Вивчити математичні методи та розділи психології, що дозволять розв'язати поставлену задачу, пояснити та обґрунтувати отримані результати
- 2) Ознайомитися з правилами побудови тестових завдань для найбільш ефективної та об'єктивної процедури оцінки знань студентів
- 3) Розробити програмний комплекс тестування й обробки результатів
- 4) Моделювання

За мету поставлено збільшення об'єктивності тестування, а також покращення якості навчання за допомогою порад студентам і викладачам практичних занять.

2 ОСНОВНА ЧАСТИНА

2.1 Теоретичні відомості

2.1.1 Метод головних компонент

Метод головних компонент (Principal component analysis) — метод, що дозволяє зменшити розмірність досліджуваної вибірки з мінімальними втратами інформації.

Маємо m об'єктів, з яких треба зняти по n певних властивостей. На вході в нас є виборки \vec{X}_k , кожна з яких відповідає сукупності властивостей k -го об'єкту

$$\vec{X}_k = \begin{bmatrix} x_k^1 \\ x_k^2 \\ \vdots \\ x_k^n \end{bmatrix}, \quad k = \overline{1, m}$$

Згрупуємо всі вимірювання в одну матрицю X

$$X = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_m^1 \\ x_1^2 & x_2^2 & \dots & x_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & \dots & x_m^n \end{bmatrix}$$

Спочатку нам знадобиться знайти вибіркові середні значення для кожної властивості

$$a_i = \frac{1}{m} \cdot \sum_{k=1}^m x_k^i, \quad i = \overline{1, n}$$

Маємо вектор вибірових середніх значень

$$\vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Центруємо отримані дані, що містяться в матриці X , віднявши від кожного стовбця вектор вибірових середніх \vec{a}

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1^1 & \tilde{x}_2^1 & \dots & \tilde{x}_m^1 \\ \tilde{x}_1^2 & \tilde{x}_2^2 & \dots & \tilde{x}_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{x}_1^n & \tilde{x}_2^n & \dots & \tilde{x}_m^n \end{bmatrix} = \begin{bmatrix} x_1^1 - a_1 & x_2^1 - a_1 & \dots & x_m^1 - a_1 \\ x_1^2 - a_2 & x_2^2 - a_2 & \dots & x_m^2 - a_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n - a_n & x_2^n - a_n & \dots & x_m^n - a_n \end{bmatrix}$$

Обчислюємо вибірову коваріаційну матрицю властивостей. Вибіркову коваріацію i та j властивості рахуємо за формулою

$$\sigma_i^j = \frac{1}{m} \cdot \sum_{k=1}^m \tilde{x}_k^i \cdot \tilde{x}_k^j = \frac{1}{m} \cdot \sum_{k=1}^m \left[(x_k^i - a_i) \cdot (x_k^j - a_j) \right], \quad i, j = \overline{1, n}$$

Маємо вибірову коваріаційну матрицю

$$K = \begin{bmatrix} \sigma_1^1 & \sigma_1^2 & \dots & \sigma_1^n \\ \sigma_2^1 & \sigma_2^2 & \dots & \sigma_2^n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n^1 & \sigma_n^2 & \dots & \sigma_n^n \end{bmatrix}$$

Щоб отримувати лише потрібну інформацію, ми хочемо знайти таке ортогональне лінійне перетворення L вхідної матриці \tilde{X} , щоб отримати матрицю

$Y = L \cdot \tilde{X}$, яка має діагональну вибірку коваріаційну матрицю K' з незростаючими зверху вниз значеннями. Діагональна вибірка коваріаційна матриця гарантує той факт, що отримані значення Y будуть некорельованими. Рангування значень діагональних елементів матриці K' за величиною дасть більш наглядне представлення про будову досліджуваних об'єктів, адже діагональні елементи — вибірккові дисперсії; а чим більше дисперсія, тим більше відповідна властивість змінюється від об'єкту до об'єкту і тим більше корисної інформації вона нам надає.

Вибіркова коваріаційна матриця K' для $Y = L \cdot \tilde{X}$ має вигляд

$$K' = L \cdot K \cdot L^* = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

З лінійної алгебри відомо, що матриця L складається з координат власних векторів матриці K , а елементи λ_k — її власні числа, які існують і є невід'ємними через невід'ємну означеність матриці K . Вважаємо що числа $\lambda_1, \dots, \lambda_n$ впорядковані від більшого до меншого для зручності подальших дій. Позначимо власний вектор матриці K , що відповідає власному числу λ_k , як \vec{l}_k . Тоді

$$\vec{l}_k = [l_k^1, l_k^2, \dots, l_k^n], \quad k = \overline{1, n}$$

Матриця L має вигляд

$$L = \begin{bmatrix} l_1^1 & l_1^2 & \dots & l_1^n \\ l_2^1 & l_2^2 & \dots & l_2^n \\ \vdots & \vdots & \ddots & \vdots \\ l_n^1 & l_n^2 & \dots & l_n^n \end{bmatrix}$$

Треба зменшити розмірність простору досліджуваних параметрів системи з n до $p < n$, але при цьому втратити якомога менше відомостей про досліджувані об'єкти. Введемо міру інформації, що залишається при зменшенні кількості компонент, що розглядаються

$$I = \frac{\lambda_1 + \dots + \lambda_p}{\lambda_1 + \dots + \lambda_n}$$

Будемо вважати, що діємо продуктивно, тому починаємо обирати з перших компонент, адже саме вони є найбільш інформативними. Також бачимо, що інформативність змінюється в межах від 0 (нічого не дізнаємось) до 1 (зберегли усю інформацію).

Надалі буде розглядатися матриця головних компонент Y

$$Y = \begin{bmatrix} y_1^1 & y_2^1 & \dots & y_m^1 \\ y_1^2 & y_2^2 & \dots & y_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ y_1^p & y_2^p & \dots & y_m^p \end{bmatrix}$$

2.1.2 Гістограма

Для подальшого аналізу потрібно здобути щільність розподілу головних компонент. Оскільки маємо справу з вибіркою і вибірковими характеристиками, потрібно побудувати гістограму, адже це і є вибіркова характеристика, що відповідає щільності.

Побудуємо j -й стовбець гістограми для виборки з k -ї строки матриці Y

$$h_j^k = \frac{1}{m} \cdot \sum_{i=1}^m \mathbb{1}(y_i^k \in I_j^k), \quad j = \overline{1, N}, \quad k = \overline{1, p}$$

де I^k — набір напівінтервалів, що розбиває відрізок $\left[\min_{i=\overline{1, m}} y_i^k; \max_{i=\overline{1, m}} y_i^k \right]$ на N

рівних частин. Для вибору N можна скористатися досить відомою формулою Стюрджеса (Sturges' formula) [1]

$$N = \lfloor \log_2 m \rfloor + 1$$

Маємо матрицю гістограм

$$H = \begin{bmatrix} h_1^1 & h_2^1 & \dots & h_N^1 \\ h_1^2 & h_2^2 & \dots & h_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ h_1^p & h_2^p & \dots & h_N^p \end{bmatrix}$$

і напівінтервалів, що відповідають кожному стовбчику кожної гістограми

$$I = \begin{bmatrix} I_1^1 & I_2^1 & \dots & I_N^1 \\ I_1^2 & I_2^2 & \dots & I_N^2 \\ \vdots & \vdots & \ddots & \vdots \\ I_1^p & I_2^p & \dots & I_N^p \end{bmatrix}$$

2.1.3 Поліноміальний розподіл

Введемо матрицю частот ν

$$\nu = m \cdot H$$

Кожна компонента — кількість елементів вибірки, що потрапили у відповідний напівінтервал

$$\nu_j^k = \sum_{i=1}^m \mathbb{1}(y_i^k \in I_j^k), \quad j = \overline{1, N}, \quad k = \overline{1, p}$$

Розглянемо вектор

$$\nu^k = [\nu_1^k, \dots, \nu_N^k], \quad k = \overline{1, p}$$

Маємо серію з m незалежних експериментів, кожен з яких може закінчитися одним з N результатів E_1^k, \dots, E_N^k , що взаємно виключаються

$$\mathbb{P}(E_i^k \cap E_j^k) = 0, \quad i \neq j, \quad k = \overline{1, p}$$

Якщо випадкові величини y_i^k мають заздалегіть відомий розподіл, який однако-
вий в межах однієї строки Y^k , маємо ймовірності кожного результату експери-
менту

$$\rho_j^k = \mathbb{P}(E_j^k) = \mathbb{P}(y_1^k \in I_j^k) = \dots = \mathbb{P}(y_m^k \in I_j^k), \quad j = \overline{1, N}, \quad k = \overline{1, p}$$

Математичне сподівання і дисперсія кожного елемента співпадає з матема-
тичним сподіванням і дисперсією біноміального розподілу з відповідними хара-
ктеристиками, адже випадкові величини ν_j^k не залежать одна від одної

$$M \nu_j^k = N \cdot \rho_j^k, \quad D \nu_j^k = N \cdot \rho_j^k \cdot (1 - \rho_j^k), \quad j = \overline{1, N}, \quad k = \overline{1, p}$$

Коваріація двох різних елементів вектора ν^k рахується за формулою [2]

$$\text{cov}(\nu_i^k, \nu_j^k) = -N \cdot \rho_i^k \cdot \rho_j^k, \quad i \neq j$$

Введемо дельта-функцію

$$\delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Отримаємо загальну формулу для коваріації

$$\text{cov}(\nu_i^k, \nu_j^k) = N \cdot \rho_i^k \cdot (\delta_{i,j} - \rho_j^k)$$

Отже, коваріаційна матриця A^k вектора ν^k виглядає наступним чином

$$A^k = |\text{cov}(\nu_i^k, \nu_j^k)|_{i,j=1}^N = N \cdot \begin{bmatrix} \rho_1^k \cdot (1 - \rho_1^k) & -\rho_1^k \cdot \rho_2^k & \cdots & -\rho_1^k \cdot \rho_N^k \\ -\rho_2^k \cdot \rho_1^k & \rho_2^k \cdot (1 - \rho_2^k) & \cdots & -\rho_2^k \cdot \rho_N^k \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_N^k \cdot \rho_1^k & -\rho_N^k \cdot \rho_2^k & \cdots & \rho_N^k \cdot (1 - \rho_N^k) \end{bmatrix}$$

2.1.4 Критерій узгодженості Пірсона χ^2

Гістограма може використовуватися не тільки для графічної інтерпретації отриманих даних, але й для віднесення вибірки до якогось відомого розподілу. Відповідь на питання “Чи дійсно вибірка y_1^k, \dots, y_p^k має розподіл F^k ?” може надати критерій узгодженості Пірсона.

Розглянемо вектор

$$S^k = [p \cdot h_1^k, \dots, p \cdot h_N^k] = \left[\sum_{i=1}^p \mathbb{1}(y_i^k \in I_1^k), \dots, \sum_{i=1}^p \mathbb{1}(y_i^k \in I_N^k) \right]$$

Кожна компонента є сумою p результатів бернулівських експериментів, ймовірність успіху якого заздалегіть невідома — саме ця характеристика й визначається припущенням щодо розподілу. Отже, треба визначити ймовірність ρ_i^k того, що випадкова величина ξ^k з функцією розподілу F^k потрапить у напівінтервал I_i^k

$$\rho_i^k = \mathbb{P}(\xi^k \in I_i^k), \quad \mathbb{P}(\xi^k \leq x) = F^k(x), \quad k = \overline{1, p}$$

Тобто вектор S має поліноміальний розподіл — розподіл експерименту, що складається з p випробувань, кожне з яких може мати лише один з N результатів E_1, \dots, E_N [3]. Кількість випробувань з результатом E_i знаходиться в i -ій компоненті вектора, а сума всіх компонент дорівнює p .

Згідно з багатовимірною центральною граничною теоремою маємо

$$S^k \sim N([p \cdot \rho_1^k, \dots, p \cdot \rho_N^k], A), \quad p \rightarrow \infty$$

Коваріаційна матриця A визначається наступним чином

$$A = p \cdot \begin{bmatrix} \rho_1^k \cdot (1 - \rho_1^k) & -\rho_1^k \cdot \rho_2^k & \cdots & -\rho_1^k \cdot \rho_N^k \\ -\rho_2^k \cdot \rho_1^k & \rho_2^k \cdot (1 - \rho_2^k) & \cdots & -\rho_2^k \cdot \rho_N^k \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_N^k \cdot \rho_1^k & -\rho_N^k \cdot \rho_2^k & \cdots & \rho_N^k \cdot (1 - \rho_N^k) \end{bmatrix}$$

Відомо, що

$$\sum_{i=1}^N h_i^k = 1$$

Це означає, що матриця A є виродженою. Тобто, ми не можемо знайти зворотню до неї, але якщо розглянути її мінор (наприклад, A_{NN}), то отримаємо гарну матрицю, яка має визначник і зворотню матрицю. Ми просто не розглядаємо гістограму h_N^k . Тоді обернена до мінора матриця буде виглядати наступним чином

$$A_{NN}^{-1} = \frac{1}{p} \cdot \begin{bmatrix} \frac{1}{\rho_1^k} + \frac{1}{p_N^k} & \frac{1}{\rho_N^k} & \cdots & \frac{1}{\rho_N^k} \\ \frac{1}{\rho_N^k} & \frac{1}{\rho_2^k} + \frac{1}{p_N^k} & \cdots & \frac{1}{\rho_N^k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\rho_N^k} & \frac{1}{\rho_N^k} & \cdots & \frac{1}{\rho_{N-1}^k} + \frac{1}{p_N^k} \end{bmatrix}$$

Якщо розглянути щільність отриманого гаусівського розподілу, то показник експоненти буде

$$R^k = (S^k - M S^k) \cdot A_{NN}^{-1} \cdot (S^k - M S^k)^T = p \cdot \sum_{i=1}^N \frac{(h_i^k - \rho_i^k)^2}{\rho_i^k}$$

Аналіз характеристичної функції величини R^k показує, що її розподіл тим більше прямує до розподілу χ^2 з $N - 1$ ступенями вільності, чим більше розмір

аналізованої вибірки p

$$R^k \xrightarrow{p \rightarrow \infty} \chi_{N-1}^2$$

З таблиці для функції розподілу χ_{N-1}^2 обираємо рівень значущості α і шукаємо відповідне до кількості ступенів вільності r_α . Рівень значущості — ймовірність помилки першого роду, тобто ймовірність того, що буде відкинута вірна гіпотеза

$$\mathbb{P}(\chi_{N-1}^2 \geq r_\alpha) = \alpha$$

Якщо $R^k \leq r_\alpha$, то гіпотеза про те, що вибірка Y^k дійсно має розподіл F^k , не відхиляється. Інакше R^k буде поводитись як \sqrt{p} і достатньо швидко зростати при великих p , а гіпотезу буде відхилено.

Чим більше рівень значущості, тим менше значення r_α , а отже і проміжок, в який дозволяється потрапити значенню R^k . Тобто, більша ймовірність відхилити вірну гіпотезу щодо розподілу, але при цьому є більше впевненості в правильності результату. Зазвичай α обирають рівним 0.1, 0.05, 0.01.

ПЕРЕЛІК ПОСИЛАНЬ

1. *Sturges, Herbert A.* The Choice of a Class Interval / Herbert A. Sturges // *J-AM-STAT-ASSOC.* — 1926. — March. — Vol. 21, no. 153. — Pp. 65–66.
2. *Mukhopadhyay, N.* Probability and Statistical Inference / N. Mukhopadhyay. Statistics: A Series of Textbooks and Monographs. — Taylor & Francis, 2000.
3. *Cramér, H.* Mathematical Methods of Statistics / H. Cramér. Princeton Mathematical Series. — Princeton University Press, 1999.