

**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	11
Перелік посилань	14

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 — вибіркві 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

$$\nu^k = [\nu_1^k, \dots, \nu_N^k], \quad k = \overline{1, p} \quad (2.1)$$

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

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

Ймовірність того, що вектор ν^k буде дорівнювати вектору $X = [x_1, \dots, x_N]$,
рахується за формулою

$$f_{multi}^k(X) = \mathbb{P}(\nu^k = X) = m! \cdot \prod_{j=1}^N \frac{(\rho_j^k)^{x_j}}{x_j!}, \quad \sum_{j=1}^N x_j = m$$

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

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

Додамо всі її строки до першої

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

Пам'ятаємо те, що сума ймовірностей дорівнює одиниці

$$\sum_{j=1}^N \rho_j^k = 1$$

Отже

$$A^k = N \cdot \begin{bmatrix} 0 & 0 & \cdots & 0 \\ -\rho_2^k \cdot \rho_1^k & \rho_2^k \cdot (1 - \rho_2^k) & \cdots & -\rho_2^k \cdot \rho_N^k \\ \vdots & \vdots & \ddots & \vdots \\ -\rho_N^k \cdot \rho_1^k & -\rho_N^k \cdot \rho_2^k & \cdots & \rho_N^k \cdot (1 - \rho_N^k) \end{bmatrix}$$

Отже, матриця A є виродженою і має ранг $N - 1$ [3].

Лагранжем було показано, що функція ймовірності поліноміального розподілу, помножена на $\sqrt{m^{N-1}}$, прямує до щільності ймовірності нормального вектора, що має $N - 1$ компонент, де константа $\sqrt{2 \cdot \pi}$ з'являється завдяки застосуванню формули Стірлінга [4]

$$m^{\frac{N-1}{2}} \cdot f_{multi}^k(X) \xrightarrow{m \rightarrow \infty} f_{norm}^k(H^k) = \frac{\exp \left\{ -\frac{1}{2} \cdot \sum_{j=1}^N \frac{m \cdot (\rho_j^k - h_j^k)^2}{h_j^k} \right\}}{\sqrt{(2 \cdot \pi)^{N-1} \cdot h_1^k \cdots h_N^k}} \quad (2.2)$$

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.

Знову розглянемо вектор (2.1)

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

Пам'ятаємо, що функція ймовірності цього вектора прямує до функції ймовірності нормального $(N - 1)$ -вимірному вектора. Розглянемо квадратичну форму, що знаходиться в експоненті (2.2)

$$R^k = \sum_{j=1}^N \frac{(m \cdot \rho_j^k - \nu_j^k)^2}{\nu_j^k} = m \cdot \sum_{j=1}^N \frac{(\rho_j^k - h_j^k)^2}{h_j^k} \quad (2.3)$$

Спираючись на результати, отримані Лагранжем, Пірсон показав, що квадратична форма має розподіл χ^2 з $N - 1$ ступенями вільності [5] [6].

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

гіпотезу

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

Якщо $R^k \leq r_\alpha$, то гіпотеза про те, що вибірка Y^k дійсно має розподіл F^k , не відхиляється.

Без втрати загальності розглянемо той випадок, коли ймовірність ρ_i^k відгадана невірно. Повернемося до формули (2.3)

$$R^k = \sum_{j=1}^N \frac{(m \cdot \rho_j^k - \nu_j^k)^2}{\nu_j^k}$$

Всі члени суми є невід'ємними. Якщо хоча б один елемент буде завеликим, то великою буде вся сума. Маємо випадкову величину η

$$\eta = \nu_i^k - m \cdot \rho_i^k = \sum_{j=1}^m (\xi_j - \rho_i^k), \quad \mathbb{1}(y_j^k \in I_i^k) = \xi_j$$

Якщо ρ_i^k вгадано невірно, то воно не дорівнює математичному очікуванню індикатора. Додамо та віднімемо справжнє математичне очікування

$$\eta = \sum_{j=1}^m (\xi_j - \mathbb{M} \xi_1 + \mathbb{M} \xi_1 - \rho_i^k) = \sum_{j=1}^m (\xi_j - \mathbb{M} \xi_1) + \sum_{j=1}^m (\mathbb{M} \xi_1 - \rho_i^k)$$

Останній доданок є просто різницею, помноженою на m

$$\eta = \sum_{j=1}^m (\xi_j - \mathbb{M} \xi_1) + m \cdot (\mathbb{M} \xi_1 - \rho_i^k)$$

Поділимо на \sqrt{m} , щоб скористатися центральною граничною теоремою

$$\frac{\eta}{\sqrt{m}} = \frac{1}{\sqrt{m}} \cdot \sum_{j=1}^m (\xi_j - \mathbb{M} \xi_1) + \frac{1}{\sqrt{m}} \cdot m \cdot (\mathbb{M} \xi_1 - \rho_i^k)$$

Перший доданок має розподіл $N(0, \sigma^2)$, де σ^2 — дисперсія випадкової величини ξ_1 для достатньо великих m . Отже, вся сума зростає пропорційно \sqrt{m}

$$\frac{\eta}{\sqrt{m}} = \frac{1}{\sqrt{m}} \cdot \sum_{j=1}^m (\xi_j - \mathbb{M} \xi_1) + \sqrt{m} \cdot (\mathbb{M} \xi_1 - \rho_i^k) \sim \sqrt{m} \cdot (\mathbb{M} \xi_1 - \rho_i^k)$$

Тобто, якщо гіпотеза невірна, то R^k буде зростати пропорційно до величини \sqrt{m} , що є достатньо великою швидкістю.

Чим більше рівень значущості, тим менше значення r_α , а отже і проміжок, в який дозволяється потрапити значенню R^k . В цьому випадку зростає 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-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. *Kendall, M.G.* Теория распределений / M.G. Kendall, A.F. Stuart. — Наука, 1966.
4. Œuvres de Lagrange: Mémoires extraits des recueils de l'Académie de Turin. (Suite.) Mémoires extraits des recueils de l'Académie royale des sciences et belles-lettres de Berlin / J.L. Lagrange, J.A. Serret, G. Darboux et al. Œuvres de Lagrange. — Gauthier-Villars, 1868.
5. *Pearson, Karl.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling / Karl Pearson // *Philosophical Magazine Series 5.* — 1900. — Vol. 50, no. 302. — Pp. 157–175.
6. *Hald, A.* A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935 / A. Hald. *Sources and Studies in the History of Mathematics and Physical Sciences.* — Springer New York, 2010.