УДК 004.89:004.422.6

МЕТОДИ СТАТИСТИЧНОГО АНАЛІЗУ ДАНИХ

Бочарніков І. В.
e-mail: ivan.bocharnikov@nure.ua
Науковий керівник– к.т.н., ст. викладач Путятіна О. Є.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

The thesis discusses the basic concepts and methods of testing statistical hypotheses. The concepts of statistical, null and alternative hypotheses, their content and purpose are described. The classification of statistical significance criteria, including parametric, non-parametric and consistency criteria, is considered. The main distributions used to test hypotheses are presented, including the normal distribution, χ^2, Student's and Fisher's distributions. Particular attention is paid to the practical application of criteria for testing the normal distribution in a sample. The conditions for testing the hypothesis and the methods of analysis used are presented.

Статистична гіпотеза – це будь-яке твердження, що стосується розподілу деякої випадкової величини чи події, яке перевіряється на підставі її вибіркових значень.

Нульова гіпотеза – це початкова гіпотеза, яка передбачає, що між параметрами генеральних сукупностей немає очікуваних розбіжностей, її прийнято позначати H0.

Альтернативна гіпотеза (або конкуруюча) – це гіпотеза, яка передбачає, що варіації генеральних сукупностей статистично неоднакові, її прийнято позначати Ha.

Статистичні критерії значущості бувають трьох типів:

1. Параметричні критерії, які оперують фізичними величинами (м, кг, с).
2. Непараметричні критерії, в яких використовують величини, що не мають розмірностей фізичних величин (місця, ранги).
3. Критерії узгодженості, які використовують для перевірки узгодженості розподілу генеральної сукупності з прийнятою раніше теоретичною моделлю (наприклад, з нормальним розподілом).

Найчастіше на практиці використовують, як критерії, нормальний розподіл, χ^2 -розподіл, розподіли Стьюдента або Фішера. Значенням критерію, що спостерігається, називають його величину, яку розраховують за досліджуваними вибірками.

Для перевірки гіпотези весь вибірковий простір поділяють на дві області, що не перетинаються: область прийняття та критичну. Областю прийняття гіпотези (областю допустимих значень) називають сукупність значень критерію, за яких нульову гіпотезу приймають. Перевірка гіпотези передбачає розрахунок значення критерію і перевірку його потрапляння до області

прийняття гіпотези. Критичною областю називають сукупність значень критерію, за яких нульову гіпотезу слід відхилити.

Перевіримо гіпотезу про нормальний розподіл за вибіркою із n = 55 спостережень, що мають значення, наведені у таблиці 1. Приймемо α = 0,1.

Таблиця 1 – Статистичні дані

18,3	15,4	17,2	19,2	23,3	18,1	21,9
15,3	16,8	13,2	20,4	16,5	19,7	20,5
14,3	20,1	16,8	14,7	20,8	19,5	15,3
19,3	17,8	16,2	15,7	22,8	21,9	12,5
10,1	21,1	18,3	14,7	14,5	18,1	18,4
13,9	19,1	18,5	20,2	23,8	16,7	20,4
19,5	17,2	19,6	17,8	21,3	17,5	19,4
17,8	13,5	17,8	11,8	18,6	19,1	

Для перевірки гіпотези про нормальний розподіл знайдемо оцінки математичного очікування і дисперсії:

$$\widetilde{m} = \tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx 17{,}87$$

$$\widetilde{\sigma} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx 8{,}62$$

Згрупуємо вибірку, розширивши перший та останній інтервали. Результати групування вибірки наведені в другому та третьому стовпчиках таблиці 1.2. У четвертому стовпчику таблиці 1.2 наведені ймовірності pk , що обчислюються за формулою

$$p_k = \mathrm{P}[X \in \Delta_k] = \Phi\left(\frac{b_k - \bar{x}}{s}\right) - \Phi\left(\frac{a_k - \bar{x}}{s}\right), k = 1,2,\dots,7,$$

де a_k і b_k – відповідно нижня та верхня межі інтервалу Δ_k .

Значення функції Лапласа Ф(х) беруться з відповідної таблиці (або з використанням ймовірнісного калькулятора).

Таблиця 2 містить результати групування вибірки для перевірки гіпотези про нормальний розподіл. У ній наведені межі інтервалів, спостережувані частоти, ймовірності потрапляння в інтервал, очікувані частоти та їх скориговані значення після об'єднання інтервалів. Також у таблиці наведені розрахунки для статистики критерію χ^2, яка використовується для оцінки відповідності вибіркових даних нормальному розподілу. Отримане значення χ^2 порівнюється з критичним, що дозволяє зробити висновок про прийняття або відхилення нульової гіпотези.

Таблиця 2 – Результати групування вибірки

№ інтервалу	Межі інтервалу	Спостережувана частота	Імовірність потрапляння в інтервал	Очікувана частота	Обробка отриманих даних		
k	Δ_k	n_k	Δ_k, p_k	np_k	np_k	$n_k - np_k$	$\frac{(n_k - np_k)^2}{np_k}$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	–∞–12	2	0,0228	1,254	5,274	0,725	0,010
2	12-14	4	0,0731	4,020			
3	14-16	8	0,1686	9,273	9,273	-1,273	0,175
4	16-18	12	0,2576	14,168	14,168	-2,168	0,332
5	18-20	16	0,2484	13,662	13,662	-2,338	0,400
6	20-22	10	0,1519	8,354	12,633	0,366	0,011
7	22–+∞	3	0,0778	4,279			
	Сума	55	1,0001	55	55	-	0,928

В п’ятому стовпчику таблиці 2 наведені очікувані частоти np_k, а в шостому – значення np_k після об’єднання перших двох та останніх двох інтервалів. Оскільки після об’єднання залишилось r = 5 інтервалів, а за вибіркою визначені оцінки двох параметрів – математичного очікування і дисперсії, тобто $i = \tilde{2}$, то число вільних ступенів дорівнює $\tilde{5}$ - $\tilde{2}$ - $\tilde{1}$ = $\tilde{2}$. За таблицею квантилів розподілу $\chi 2$ знаходимо $\chi^2_{0,90}$ = 4,61 . Вибіркове значення статистики критерію дорівнює $\chi^2_{0,90}$ = 0,928. Отже, остаточне значення χ^2 порівнюється з критичним значенням з таблиці квантилів. Оскільки отримане значення (0,928) менше критичного (4,61), гіпотеза про нормальний розподіл приймається.

Список використаних джерел:

1. В.О. Гороховатський, І.С. Творошенко (2021) “Методи інтелектуального аналізу та оброблення даних”, стр. 7-11.