УДК 378.147

С.М. Шевченко, Ю.Д. Жданова,
С.О. Спасітєлєва, О.В. Адамович

СТАТИСТИЧНА ОБРОБКА ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ ЯК ОДНА З ФОРМ НАУКОВО-ДОСЛІДНОЇ РОБОТИ СТУДЕНТІВ СПЕЦІАЛЬНОСТІ «КІБЕРБЕЗПЕКА»

У статті піднімається проблема щодо змісту, форм та організації науково-дослідної роботи студентів у вищих технічних навчальних закладах. Спираючись на дослідження у психолого-педагогічній літературі, доведено, що теоретичне знання, здобуте завдяки власній діяльності і професійно спрямоване, набуває особистісного змісту. Проаналізовані основні підходи до статистичної обробки експериментальних даних, як початкового складника наукового дослідження. Запропоновано практичне застосування даних методів у науково-дослідній роботі студентів спеціальності «Кібербезпека»

Ключові слова: науково-дослідна робота, статистична обробка експериментальних даних, студенти спеціальності «Кібербезпека», проект.

Вступ

Національна система вищої професійної освіти має забезпечити умови для формування та розвитку майбутніх спеціалістів, які зможуть швидко адаптуватися до стрімких змін техніки та технологій у сучасному світі. Теоретична підготовка студентів (навіть найбільш глибока та повна) не зможе дати того, дає людині досвід власної діяльності. Тому особливого значення у технічній вищій школі набувають ті методи та форми навчальної діяльності, які стимулюють самостійну творчість студентів, зокрема їх науково-дослідна робота. Вона розвиває аналітичне мислення, індивідуальні здібності, дослідні навички студентів, формує наукову інтуїцію, творчий підхід до сприйняття знань і практичне застосування їх для вирішення завдань і наукових проблем. Наукове дослідження у галузі «Інформаційної безпеки» (як у будь-якій іншій) починається зі збору та обробки даних, на основі чого створюються інформаційні моделі. Їх аналіз та моніторинг здійснюється за допомогою методів статистики. Тільки на основі аналізу великої кількості мікростанів (подій) та моніторингу результатів відповідні технології дозволять спрогнозувати та оцінити загальну кібербезпеку об’єкту. Оцінка методів захисту від кібератак зосереджена передусім на інформуванні та прогнозуванні. Саме це формує актуальність дослідження даної проблеми у межах навчального процесу студентів спеціальності «Кібербезпека».

Мета статті – охарактеризувати зміст, форми та організацію науково-дослідної роботи студентів, зокрема представити приклад обробки емпіричних даних статистичними методами з використанням інформаційних технологій.

Основна частина

Науково-дослідна робота – найцікавіший вид занять, який у повній мірі дозволяє розкритися творчому потенціалу студентів і на практиці закріпити отримані знання [2]. Педагогічні дослідження відомих науковців таких, як А. Алексюк, Н. Дем’яненко, І. Зязюн, О. Мартиненко, О. Пєхота, С. Гончаренко, В. Шейко, Н. Кушнаренко та інших, присвячені різним аспектам організації та проведення науково-дослідної роботи студентів [3; 4]. Узагальнюючи основні моменти у даних дослідженнях, можна виділити наступні підходи до визначення та суті науково-дослідної роботи:

1) науково-дослідна робота, як засіб розвитку творчої особистості, передбачає стимулювання активності, самостійності та відповідальності, творче застосування знань та умінь;

2) науково-дослідна робота, як самостійна навчальна діяльність, розвиває уміння здійснювати аналіз проблеми, виділяти головне, істотне, висувати гіпотези та будувати моделі їх розв’язання, виводити наслідки;

3) науково-дослідна робота, як здійснення соціального замовлення суспільства, спрямована на підготовку спеціалістів, здібних ефективно діяти, адаптуватися до змін умов праці та технологій протягом всього життя.

Ці точки зору щодо суті науково-дослідної роботи не суперечать одна одній, бо розглядаються в тісному зв'язку і доповнюють суттєві якості цього поняття.

Розвиток сучасного суспільства автоматизував практично всі сфери людської діяльності. У зв'язку з цим за останній час виросла кількість злочинів у комп'ютерній сфері, так званих кіберзлочинів. Ефективне протистояння цим злочинам, забезпечення безпечного життя кожної людини та держави в цілому – основна задача захисту інформації.

Методи та засоби захисту інформації знаходяться у перерізі різних наукових областей. Проте основу тут закладає математика. Математичні методи, які використовуються у інформаційній безпеці, різноманітні і практично виділяються з усіх областей математики (математичний аналіз, лінійна алгебра та аналітична геометрія, дискретний аналіз, теорія ймовірностей, математична статистика, алгебраїчні структури, неархімедовий аналіз та інші).

Таким чином, статистична обробка емпіричних даних є однією з форм науково-дослідної роботи студентів галузі «Інформаційна безпека». Надалі розглянемо організацію такої роботи у процесі навчальної діяльності.

Ми згодні з науковцями, які вважають, що науково-дослідна робота студентів ґрунтується на принципах проектування, де дослідницький проект є інтелектуальною та особистісною цінністю самого студента [1]. Метод проектів – спосіб досягнення дидактичних цілей через детальну розробку проблеми, яка допомагає визначити реальний практичний результат і, головне, є обов'язково професійно спрямованим. Організована таким чином навчальна діяльність мотивує студентів на вивчення математичних дисциплін.

Пропонуємо як зразок виконання такого проекту студентом групи БСД-21 Адамович О.

Методи статистичного опису результатів спостережень

Завдання 1. Представлено вибірку за 100 днів з 1.02.17-10.05.17 кількості веб-загроз, що відбулися на території України протягом дня (дані отриманні з веб-сайту лабораторії Касперського https://securelist.ru/statistics/):

41222	47391	64532	65927	72848	70285	69032	72839	76483	75843
80332	89632	87438	79843	63422	60874	57673	43998	43562	44567
51995	52009	54322	60439	58999	57333	54323	43942	45099	61843
68932	70342	69574	65475	60055	61988	62010	61034	52843	53765
54326	50342	49654	41879	42314	40675	54786	50467	75243	78357
79167	79032	86432	80345	73567	71567	60533	58903	56932	66985
59023	54765	57890	63845	42001	43276	55031	60943	57354	56236
65998	47392	52381	68842	61766	56593	58219	62976	64601	74091
67898	64619	65412	69148	64231	47809	53079	57467	41392	43965
49650	53001	52021	58736	42047	40431	40601	61841	71855	86508

На підставі наведених вибіркових даних:

1) згрупувати дані і побудувати інтервальний статистичний ряд. При цьому область реалізацій розбити на сім однакових інтервалів;

2) побудувати таблицю частот групованої вибірки;

3) побудувати гістограму і полігон відносних частот;

4) побудувати гістограму і полігон накопичених відносних частот;

5) знайти емпіричну функцію розподілу і накреслити її графік;

6) знайти моду $\widetilde{\text{Mo}}$ (аналітично і графічно), медіану $\widetilde{\text{Me}}$ (аналітично і графічно),

7) знайти методом умовних варіант вибіркову середню $\bar{x}$, вибіркову дисперсію $\widetilde{D}$,

8) знайти вибіркове середнє квадратичне відхилення, вибірковий коефіцієнт варіації $\widetilde{\text{V}}$,

9) вибіркові коефіцієнти асиметрії та ексцесу $\widetilde{\text{As}}$ і $\widetilde{\text{Es}}$;

10)побудувати довірчий інтервал для математичного сподівання

Розв язання.

1) **Варіаційним рядом** називається ранжований в порядку зростання (чи спадання) ряд варіант із відповідними їм вагами (частотами та частостями), у нашому випадку варіаційний ряд має вигляд:

40431	40601	40675	41222	41392	41879	42001	42047	42314	43276
43562	43942	43965	43998	44567	45099	47391	47392	47809	49650
49654	50342	50467	51995	52009	52021	52381	52843	53001	53079
53765	54322	54323	54326	54765	54786	55031	56236	56593	56932
57333	57354	57467	57673	57890	58219	58736	58903	58999	59023
60055	60439	60533	60874	60943	61034	61766	61841	61843	61988
62010	62976	63422	63845	64231	64532	64601	64619	65412	65475
65927	65998	66985	67898	68842	68932	69032	69148	69574	70285
70342	71567	71855	72839	72848	73567	74091	75243	75843	76483
78357	79032	79167	79843	80332	80345	86432	86508	87438	89632

Згрупуємо вибіркові дані, для чого розрахуємо довжину інтервалу групування:

$$\text{h} = \frac{\text{x}_{\max} - \text{x}_{\min}}{\text{l}} = \frac{89632 - 40431}{7} \approx 7029$$

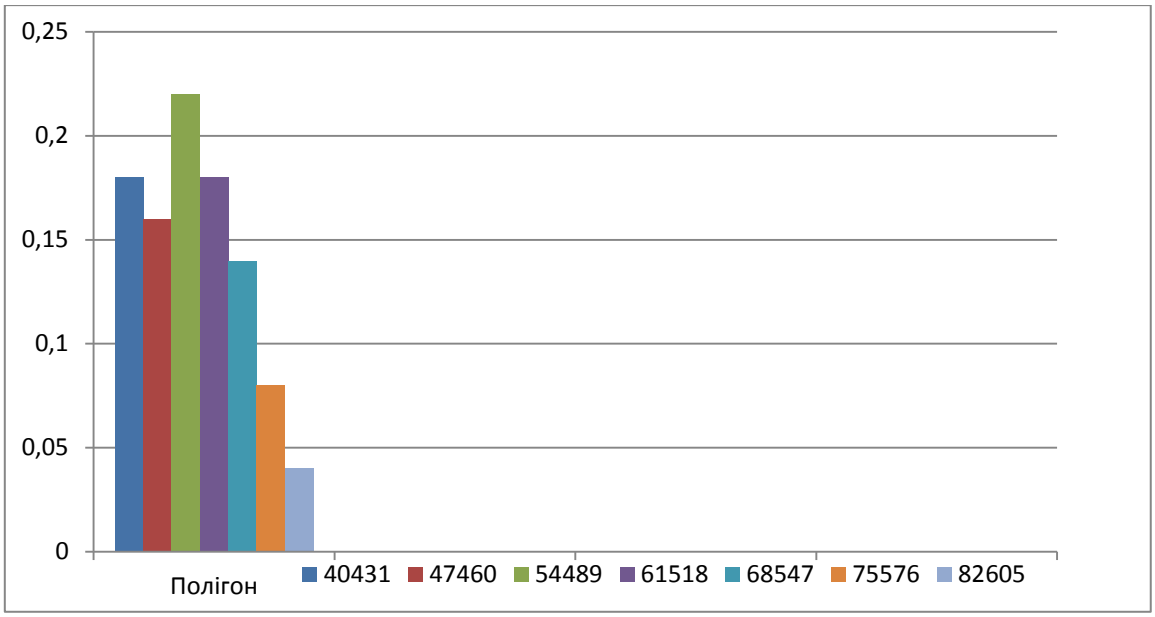
Інтервали	40431-47460	47460-54489	54489-61518	61518-68547	68547-75576	75576-82605	82605-89634
частоти	18	16	22	18	14	8	4

2) Побудуємо таблицю частот групованої вибірки

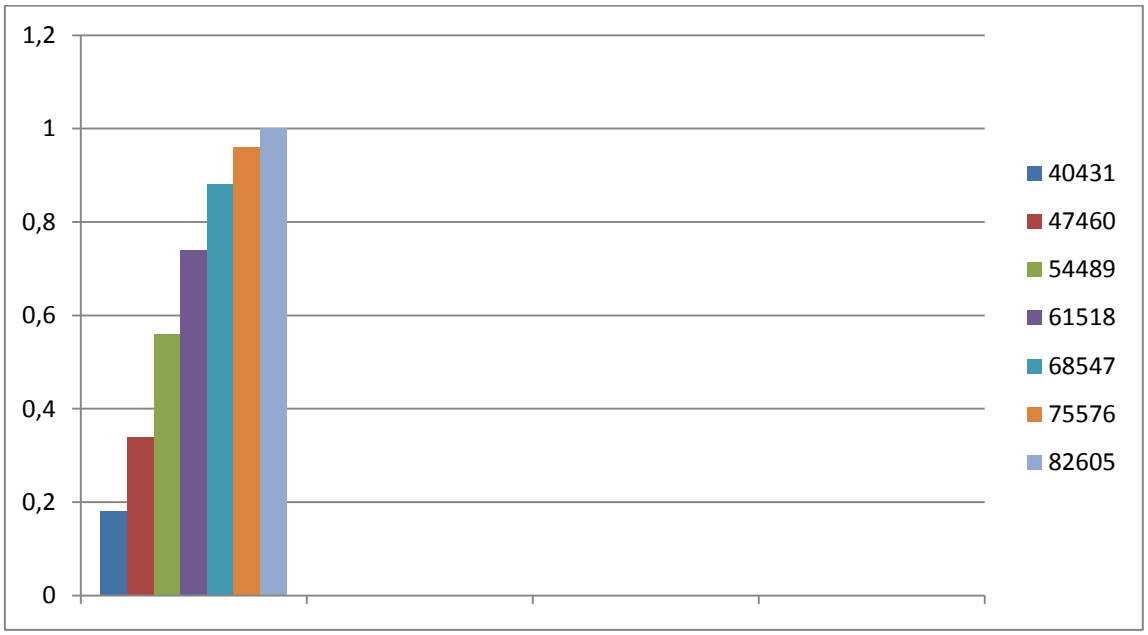
№ інтервалу	1	2	3	4	5	6	7	∑
Межі інтервалів $x_i - x_{i+1}$	40431-47460	47460-54489	54489-61518	61518-68547	68547-75576	75576-82605	82605-89634	
Середина інтервалу x_i^*	43945,5	50974,5	58003,5	65032,5	72061,5	79090,5	86119,5	
Частота інтервалу n_i^*	18	16	22	18	14	8	4	100
Накопичені частоти $\sum_{j=1}^{i} n_j^*$	18	34	56	74	88	96	100	

Відносні частоти $\frac{n_i^*}{n}$	0,18	0,16	0,22	0,18	0,14	0,08	0,04	1,00
Накопичені відносні частоти $\sum_{j=1}^{i} \frac{n_j^*}{n}$	0,18	0,34	0,56	0,74	0,88	0,96	1,00	

3) Побудуємо гістограму і полігон частот, скориставшись 2-м і 6-м рядками таблиці частот:



4) Побудуємо гістограму і полігон накопичених відносних частот, скориставшись 2-м і 7-м рядками таблиці частот:

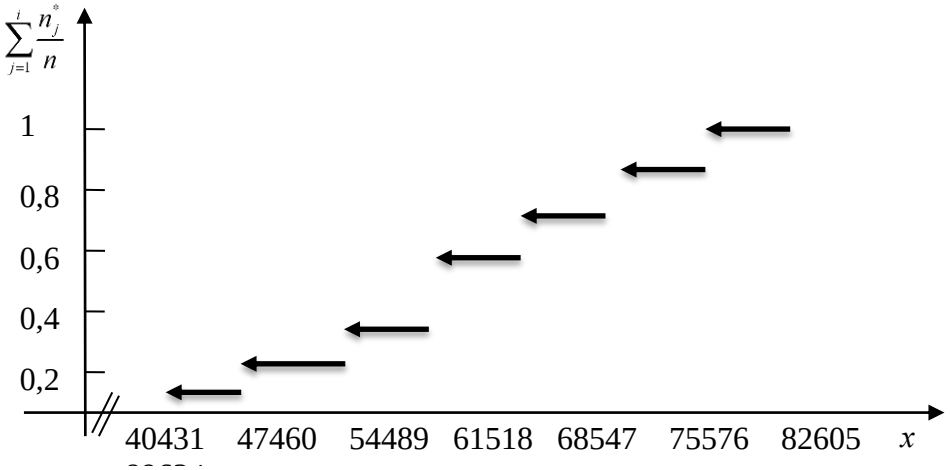


5) Знайдемо емпіричну функцію розподілу:

Емпіричною функцією розподілу *Fn*(*x*) називається відносна частота (частість) того, що ознака (випадкова величина X) прийме значення, менше заданого x, тобто:
$_n(x)=w(X<x)=w_i^{max}$.

$$_n^*(x)=\frac{n_x}{n}=\begin{cases}0, & x\le 40431\\ 0.18, & 40431<x\le 47460\\ 0.34, & 47460<x\le 54489\\ 0.56, & 54489<x\le 61518\\ 0.74, & 61518<x\le 68547\\ 0.88, & 68547<x\le 75576\\ 0.96, & 75576<x\le 82605\\ 1, & 82605<x\le 89634\end{cases}$$

І накреслимо її графік



6) Вибірковою модою називається значення випадкової величини, що зустрічається найчастіше в сукупності спостережень.

Для визначення моди інтервального статистичного розподілу знайдемо модальний інтервал і застосуємо формулу:

$$\widetilde{Mo}=x_{Mo(\min)}+h\left(\frac{n_{Mo}^*-n_{Mo-1}^*}{2n_{Mo}^*-n_{Mo-1}^*-n_{Mo+1}^*}\right)$$

Модальним інтервалом є інтервал № 3: 54489-61518, звідки $x_{Mo(min)}=54489$, h=7029, $n_{Mo}^*=22, n_{Mo-1}^*=16, n_{Mo+1}^*=18$. Отже,

$$Mo^*=54489+7029\left(\frac{22-16}{2\cdot 22-16-18}\right)=58706{,}4$$

Вибірковою медіаною називається: значення варіантів яке поділяє варіаційний ряд на дві рівні частини. Для визначення медіани інтервального статистичного розподілу знайдемо медіанний інтервал і застосуємо формулу:

$$\widetilde{Me}=x_{Me(\min)}+h\left(\frac{\frac{n}{2}-(n_1^*+n_2^*+\cdots+n_{Me-1}^*)}{n_{Me}^*}\right)$$

Медіанним інтервалом також є інтервал № 3: 54489-61518, звідки $x_{Me(min)} = 54489$,h=7029, $\frac{n}{2} = 50$, $n^*_{Me} = 22$, $(n^*_1 + n^*_2) = 34$. Отже,

$$\widetilde{Me} = 54489 + 7029\left(\frac{50-34}{22}\right) = 59601.$$

7) Для обчислення інших числових характеристик вибірки використаємо метод умовних варіант. В таблиці розподілу частот групованої вибірки найбільшу частоту має інтервал № 3. Візьмемо за умовний нуль середину цього інтервалу $x^*_{3515} = 58003{,}5$. Перетворимо дані за формулою

$$u^*_i = \frac{x^*_i - d_x}{h} = \frac{x^*_i - 58003{,}5}{7029}, i = 1{,}2, \dots, 7$$

Будемо мати

x^*_k	43945.5	50974.5	58003.5	65032.5	72061.5	79090.5	86119.5	$\sum$
n^*_k	18	16	22	18	14	8	4	100
u^*_k	-2	-1	0	1	2	3	4	
$u^*_k n^*_k$	-36	-16	0	18	28	24	16	34
u^{*2}_k	4	1	0	1	4	9	16	
$u^{*2}_k n^*_k$	72	16	0	18	56	72	64	298

Обчислимо середню $\bar{u}$ і дисперсію $\widetilde{D_u}$ перетворених даних:

$$\bar{u} = \frac{1}{n}u^*_k n^*_k = \frac{1}{100}(34) = 0.34;$$
$$\widetilde{D_u} = \frac{1}{n}u^{*2}_k n^*_k - \bar{u}^2 = \frac{1}{100}(298) - (0.34)^2 = 2.8644.$$

Знайдемо вибіркову середню і вибіркову дисперсію початкових даних:

$$\bar{x} = h\bar{u} + d_x = (0.34)7029 + 58003{,}5 = 60393{,}36\ см$$
$$\widetilde{D_x} = h^2\widetilde{D_u} = 7029^2 \cdot 2.8644 = 141520955.3604$$

8) Середнє квадратичне відхилення

$$\tilde{\sigma} = \sqrt{\widetilde{D}} = \sqrt{141520955.3604} \approx 11896{,}26;$$

9) Обчислимо вибіркові центральні моменти 3-го і 4-го порядку за формулою:

$$\widetilde{\mu_k} = \frac{1}{n}\sum_{i=1}^{l}(x^*_i - \bar{x})^k n^*_i, \quad k = 3{,}4.$$

$$\widetilde{\mu_3} = \frac{1}{n}\sum_{i=1}^{l}(x^*_i - \bar{x})^3 n^*_i = \frac{1}{100}((43945.5 - 60393{,}36)^3 \cdot 18 + +(50974.5 - 60393{,}36)^3 \cdot 16 + (58003.5 - 60393{,}36)^3 \cdot 22 + (65032.5 - 60393{,}36)^3 \cdot 18 + (72061.5 - 60393{,}36)^3 \cdot 14 + (79090.5 - 60393{,}36)^3 \cdot 8 + (86119.5 - 60393{,}36)^3 \cdot 4) = \frac{1.777420 \cdot 10^{13}}{100} \approx 5.066857 * 10^{11}$$

$$\widetilde{As} = \frac{\widetilde{\mu_3}}{\tilde{\sigma}^3} = \frac{5.066857 * 10^{11}}{11829.83^3} = 0.306057$$

Оскільки $\widetilde{As} > 0$, то варіаційний ряд має лівосторонню асиметрію.

$$\widetilde{\mu_4} = \frac{1}{n}\sum_{i=1}^{l}(x_i^* - \bar{x})^4 n_i^* = \frac{1}{100}((43945.5 - 60393{,}36)^4 \cdot 18 + (50974.5 - 60393{,}36)^4 \cdot 16 + (58003.5 - 60393{,}36)^4 \cdot 22 + (65032.5 - 60393{,}36)^4 \cdot 18 + (72061.5 - 60393{,}36)^4 \cdot 14 + (79090.5 - 60393{,}36)^4 \cdot 8 + (86119.5 - 60393{,}36)^4 \cdot 4) = \frac{4.335676 \cdot 10^{18}}{100} \approx 4.441619 * 10^{16}$$

$$\widetilde{Es} = \frac{\widetilde{\mu_4}}{\tilde{\sigma}^4} - 3 = \frac{4.441619 * 10^{16}}{11829.83^4} - 3 = -0.732082$$

Оскільки $\widetilde{Es} < 0$, то розподіл плосковершинний.

10) Побудуємо довірчий інтервал для математичного сподівання $\bar{X} = m$ з надійністю $\gamma = 0{,}99$, вважаючи генеральну сукупність розподіленою за нормальним законом з невідомою дисперсією. Для побудови довірчого інтервалу при невідомій генеральній дисперсії необхідно знати $\bar{x}$, $\tilde{\tilde{\sigma}}$, n, t.. З умови задачі маємо:

$$\bar{x} = 60393{,}36$$

$$\tilde{\tilde{\sigma}} = \sqrt{\frac{n}{n-1}\widetilde{D}} = \sqrt{\frac{100}{99}141520955.3604} = 11956{,}189190$$

n=100⇒$\sqrt{n} = 10$

Величину t знаходимо з таблиці розподілу Стьюдента (t -розподілу):
$t = t(100; 0{,}99) = 2{,}627$.

Знайдемо числові значення кінців довірчого інтервалу :

$$\bar{x} - t\frac{\tilde{\tilde{\sigma}}}{\sqrt{n}} = 60393{,}36 - 2.627\frac{11956.189190}{10} = 57252{,}469099$$

$$\bar{x} + t\frac{\tilde{\tilde{\sigma}}}{\sqrt{n}} = 60393{,}36 + 2.627\frac{11956.189190}{10} = 63534{,}2509$$

Таким чином, маємо: $57252.469099 < \bar{X} < 63534.2509$

Отже, з надійністю 0,99 інтервал $(57252{,}469099; 63534{,}2509)$покриває оцінюваний параметр $\bar{X}$.

Перевірка статистичних гіпотез

За заданим інтервальним статистичним розподілом випадкової величини X — кількість веб-загроз при рівні значущості $\alpha = 0{,}01$ перевірити правильність гіпотези про нормальний закон розподілу ознаки X .

Розв'язання. При рівні значущості α=0,01 перевіримо гіпотезу про нормальний закон розподілу генеральної сукупності ознаки X — кількості сигналів протягом дня. За формою гістограми відносних частот можемо припустити, що ознака X має нормальний закон розподілу. Отже, висуваємо нульову гіпотезу H_0: ознака X має нормальний закон розподілу, альтернативна гіпотеза: ознака X не має нормальний закон розподілу. Використаємо критерій узгодженості Пірсона.

Обчислимо теоретичні частоти $n_i = np_i$, i=1,…,8 для чого складемо розрахункову таблицю

x_i	x_{i+1}	n_i^*	$z_i = \frac{x_i - \bar{x}}{\tilde{\sigma}}$	$z_{i+1} = \frac{x_{i+1} - \bar{x}}{\tilde{\sigma}}$	$\Phi(z_i)$	$\Phi(z_{i+1})$	$n_i' = np_i = n(\Phi(z_{i+1}) - \Phi(z_i))$
40431	47460	18	–1.678036	–1.087178	–0.4525	–0.3599	10
47460	54489	16	–1.087178	–0.496320	–0.3599	–0.1879	18
54489	61518	22	–0.496320	0.094537	–0.1879	0.0359	23
61518	68547	18	0.094537	0.685395	0.0359	0.2517	22
68547	75576	14	0.685395	1.276253	0.2517	0.3980	16
75576	82605	8	1.276253	1.867111	0.3980	0.4686	8
82605	89634	4	1.867111	2.457969	0.4686	0.4931	3

Обчислимо спостережуване значення критерію

$$X^2_{\text{спост}} = \sum \frac{(n_i - n_i')^2}{n_i'}$$

для чого складемо розрахункову таблицю

n_i^*	$n_i' = np_i$	$n_i^* - n_i'$	$(n_i^* - n_i')^2$	$\frac{(n_i^* - n_i')^2}{np_i}$	n_i^{*2}	$\frac{n_i^{*2}}{n_i'}$
18	10	8	64	6,4	324	32,4
16	18	–2	4	0,22	256	14,22
22	23	–1	1	0,04	484	21.04
18	22	–4	16	0,73	324	14,73
14	16	–2	4	0,25	196	12,25
8	8	0	0	0	64	8
4	3	1	1	0,34	16	5,34
$\sum$ 100	100			7,98		107,98

Отже, маємо

$$X^2_{\text{спост}} = \sum_{i=1}^{7} \frac{(n_i - n_i')^2}{n_i'} = 7{,}98$$

Для контролю правильності обчислень використаємо співвідношення

$$X^2_{\text{спост}} = \sum_{i=1}^{7} \frac{n_i^2}{n_i'} - n \qquad 7{,}98 \approx 107{,}98 - 100$$

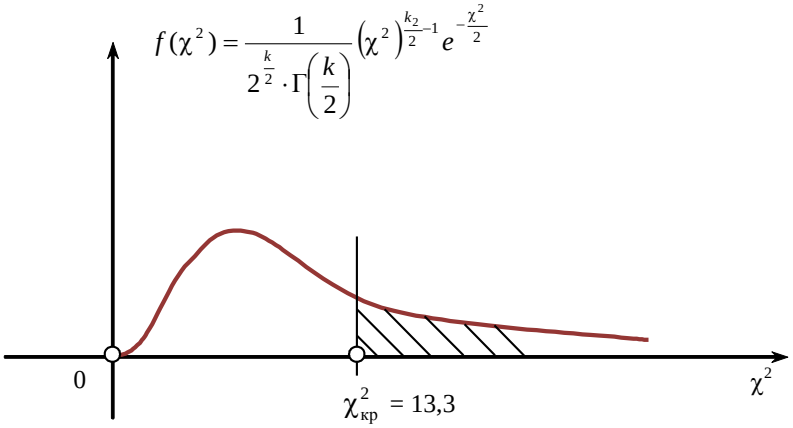
Отже, обчислення виконані правильно.

Знайдемо число степенів свободи, враховуючи, що число груп вибірки: $l = 7$; а число параметрів нормального розподілу $s = 2$: $k = l - s - 1 = 7 - 2 - 1 = 4$.

В таблиці критичних точок розподілу χ^2 за рівнем значущості $\alpha = 0{,}01$ та числом степенів свободи $k = 4$ знаходимо $\chi^2_{кр}(0{,}01;\ 4) = 13{,}3$.

Оскільки $\chi^2_{сп} < \chi^2_{кр}$, то відмінність емпіричних та теоретичних частот незначуща. Отже, дані спостережень узгоджуються з гіпотезою про нормальний розподіл генеральної сукупності.

Правостороння критична область зображена на малюнку:



Статистичне рішення. Таким чином, нульова гіпотеза про нормальний закон розподілу ймовірностей ознаки X приймається. Це означає, що надалі ми з відповідною ймовірністю можемо прогнозувати та аналізувати процес щодо веб-загроз в Україні.

Висновок

Підсумовуючи вище сказане, вважаємо, що науково-дослідна робота у даному випадку виступає як особливе відношення студента до засвоєння соціального досвіду, як особистісна та професійна самореалізація. Слід відмітити, що більшість студентів складають програми різними мовами програмування для розв'язання таких завдань за допомогою ІКТ. Кращі роботи обговорюються на студентських конференціях. Зазначимо, що створення таких проектів (задач прикладного характеру) мотивує наших студентів на вивчення математики, на її значущість, формує інтерес до математичної діяльності.

Список використаної літератури

1. Жданова Ю.Д., Шевченко С.М., Шевченко Г.В. Усвідомлення абстрактності через прикладну спрямованість дискретної математики / Ю.Д. Жданова, С.М. Шевченко, Г.В. Шевченко // Матеріали V Міжнародної науково-практичної конференції «Математика в сучасному технічному університеті», 29-30 грудня 2016 р., Київ: Матеріали конф. – К.: НТУУ «КПІ», 2016. – С. 147 – 149.

2. Крупський Я.В. Тлумачний словник з інформаційно-педагогічних технологій: словник/ Я.В. Крупський, В.М. Михалевич. – Вінниця: ВНТУ, 2010. – 72 с.

3. Шевченко С.М., Жданова Ю.Д. Математичні компетенції майбутніх фахівців інформаційної безпеки / С.М. Шевченко, Ю.Д. Жданова // Сучасний захист інформації. – К.: ДУТ, 2016. – № 4. – С. 90 – 96.

4. Шейко В.М., Кушнаренко Н.М. Організація та методика науково-дослідницької діяльності: Підручник. – 2-ге вид., перероб. і доп. / В.М. Шейко, Н.М. Кушнаренко. – К.: Знання – Прес, 2002. – 295 с.

Надійшла 21.01.2017 р. Рецензент: д.пед.н., доцент Присяжнюк С.І.