Змістовий модуль 2. Поняття параметричної та непараметричної статистики

Лекція 2.1. Параметрична статистика.

План:

1. Визначення статистичних критерій які забезпечують математичне обґрунтування прийняття істинної і відхилення помилкової гіпотези.
2. Розрахунок показники розподілу: середнього, дисперсії
3. Критерій Стьюдента.
4. Критерій Фішера.

Емпіричні розподіли і числові характеристики спостережень

Угруповання експериментальних даних

Отримані дані в ході експериментальної роботи представлені у вигляді невпорядкованого набору чисел. Для того щоб по ним можна було робити якісь висновки, необхідна первинна їх обробка − угруповання. Розглянемо угруповання на конкретному прикладі.

Приклад 1: У експерименті отримані дані результатів стрибка вгору з місця спортсменів баскетболістів (65 осіб): 59, 48, 53, 47, 57, 64, 62, 62, 65, 57, 57, 81, 83, 48, 65, 76 , 53, 61, 60, 37, 51,51, 63, 81, 60, 77, 71, 57, 82, 66, 54, 47, 61, 76, 50, 57, 58, 52, 57, 40, 53 , 66, 71, 61, 61, 55, 73, 50, 70, 59, 50, 59, 83, 69, 67, 66, 47, 56, 60, 43, 54, 47, 81, 76, 69 см.

В даному прикладі число спостережень склало 65 виміряних значень ознаки (результатів стрибка вгору з місця), n = 65.

Для угруповання наявних даних необхідно весь проміжок (діапазон варіювання ознаки) між найбільшими і найменшими значеннями розбити на ряд інтервалів, або, як їх зазвичай називають, розрядів.

Весь діапазон варіювання варіантів спостережень розташований в проміжку 37-83 см. Далі необхідно визначити число розрядів (R); ширину розряду (h); межі розрядів (нижня $x_{нi}$, верхня $x_{вi}$). Число розрядів можна вибрати, керуючись таблицею 1.

Таблиця 1 - Вибір числа розрядів угруповання

Число спостережень (n)	Число розрядів (R)
30-60	5-8
60-100	7-10
100-200	9-12
200-500	11-16

У нашому прикладі число спостережень, n = 65, приймаємо R = 10, вибравши число розрядів, визначаємо ширину розряду (h) за такою формулою:

$$h = \frac{x_{max-}x_{min}}{R-1}$$

де h − ширина розряду,
x_{max}, x_{min} − найбільше і найменше значення ознаки;
R − число розрядів.

В даному прикладі:

$$h = \frac{83-37}{10-1} = 5{,}1\ см.$$

Оскільки вихідні дані визначені з точністю до сантиметра, то округляємо знайдене значення ширини розряду до необхідної точності (цілого числа). З урахуванням цього приймаємо h = 5 см. Тепер знаходимо межі розрядів угрупування. Рекомендується вибрати межі розрядів таким чином, щоб найменше спостереження виявилося приблизно в середині першого, а найбільше − в середині останнього розряду. Звідси, нижню межу першого розряду (x_{H1}) можна визначити за формулою:

$$x_{H1} = x_{min} - \frac{h}{2}$$

Для розглянутого прикладу:

$$x_{H1} = 37 - \frac{5}{2} = 34{,}5\ см$$

додавши до цієї величини ширину розряду, знайдемо нижню межу другого розряду:

$$x_{H2} = 34{,}5 + 5 = 39{,}5\ см$$

Це буде одночасно і верхня межа попереднього першого розряду ($x_{В1}$). Аналогічно знаходимо x_{H3} = $x_{В2}$ = 39,5 + 5 = 44,5 см і т.д. для всіх десяти розрядів. Іноді знайдені межі розрядів точно збігаються з числовим значенням варіанту спостережень. Виникає питання: в який розряд віднести таке число? В цьому випадку рекомендується зменшити верхню межу всіх розрядів на величину, рівну точності вимірювання ознаки

Далі заповнюємо таблицю даних, що пройшли початкову статистичну обробку (табл. 2).

Таблиця 2 − Табличне представлення даних результатів стрибка вгору з місця спортсменів - баскетболістів

Номер розряду (i)	Межі розрядів (x_{Hi}-x_{Bi})	Серединні значення (x_i)	Розподіл даних	Частоти (n_i)	Частості (w_i)
1	34,5-39,5	37	0	1	0,015
2	39,5-44,5	42	00	2	0,031
3	44,5-49,5	47	000000	6	0,092
4	49,5-54,5	52	00000000000	11	0,169
5	54,5-59,5	57	000000000000	12	0,185
6	59,5-64,5	62	00000000000	11	0,169
7	64,5-69,5	67	00000000	8	0,123
8	69,5-74,5	72	0000	4	0,062
9	74,5-79,5	77	0000	4	0,062
10	79,5-84,5	82	000000	6	0,092
	Сума		**65**	**65**	**1,000**

У першому стовпці міститься номер розряду угруповання (i), у другому – межі розрядів ($x_{нi} - x_{вi}$), в третьому – серединні значення розрядів (x_i), четвертий стовпець візуально показує, скільки містить кожен розряд варіантів спостережень. Маючи перед собою цифровий масив, умовними значками, наприклад гуртками, відзначаємо повторюваність варіантів в кожному розряді, тобто по порядку для кожного з чисел, ставимо умовний значок в рядку табл. 2, яка відповідає розряду угруповання, в якій це число потрапляє. Після того, як вихідні дані будуть вичерпані, підраховуємо число умовних значків в кожному рядку і заносимо їх кількість в п'ятий стовпець. Числа, що показують скільки разів варіанти, які стосуються кожного розряду, зустрічаються в спостереженнях, називаються частотами (n_i). Сума частот завжди дорівнює числу спостережень (n), що можна використовувати для перевірки правильності заповнення таблиці.

У шостий стовпець заноситься величина, що показує частку спостережень, які потрапили в даний розряд, так звана частість (w_i), вона визначається формулою:

$$w_i = \frac{n_i}{n}$$

Сума всіх частотних завжди дорівнює 1. Таким чином, експериментальні дані, представлені в такій формі (у вигляді таблиці) дають первинні статистичні уявлення про результати досліджень.

Графічне представлення експериментальних даних

Емпіричні розподіли експериментальних даних найнаочніше виглядають у вигляді графічних зображень. Найчастіше використовують дві основні форми графічного представлення даних: гістограма (рис. 1) і полігон частот (рис. 2).

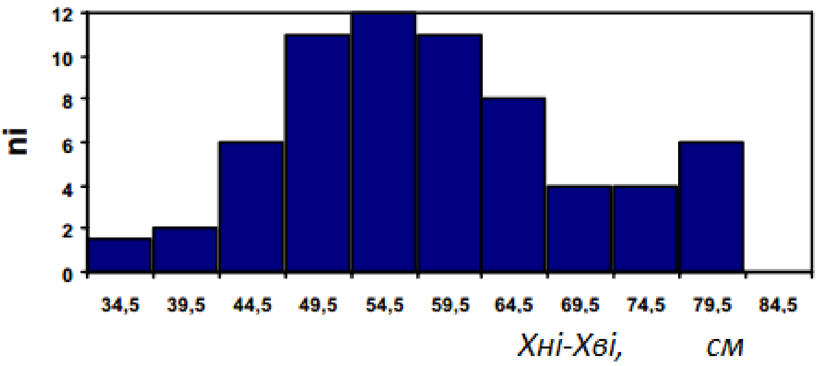

Рис.1 – Гістограма

Гістограма складається прямокутників, що примикають один до одного, основи яких відкладаються по осі абсцис (крайні точки основ прямокутників - межі розрядів), а по осі ординат – висоти прямокутників, що відображають відносну щільність розподілу експериментальних даних і пропорційних відношенню

$$\frac{n_i}{h_i}$$

де n_i – частота і-го розряду, h_i - ширина і-го розряду угруповання.

Полігон частот утворюється ламаною лінією, що з'єднує точки, відповідні серединним значенням розрядів угруповання (x_i) і частотам цих розрядів (n_i). Серединні значення відкладаються по осі абсцис, а частоти − по осі ординат.

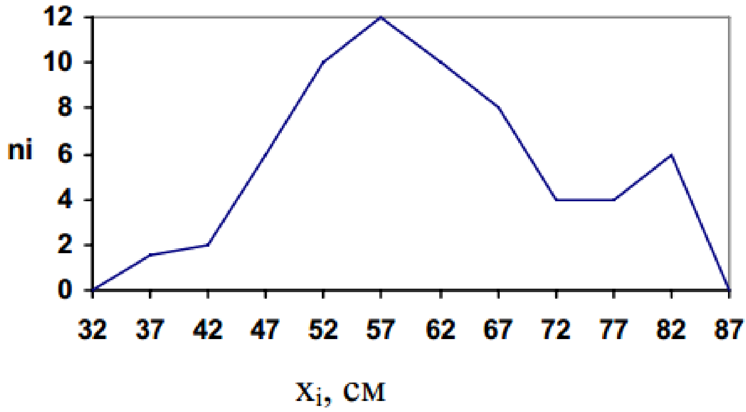

Рис.2 – Полігон частот

У розглянутому вище прикладі гістограма і полігон частот наочно показують, що використовуваний тест (стрибок вгору з місця), як інструмент вимірювання, який досліджує швидкісне силові якості баскетболістів, успішно розрізняє дані, що потрапляють в діапазон 37-72 см. Для вивчення об'єктів спостережень, що стрибають вище 72 см слід або вдосконалити інструмент вимірювання, або посилити коректність проведення експерименту.

Числові характеристики спостережень

Первинна обробка експериментальних даних (угрупування) і графічне їх уявлення наочно показують, як варіює ознака в вибірковій сукупності, але вони недостатні для повної характеристики всього обсягу спостережень. Необхідні узагальнюючі числові характеристики, які показують положення центра емпіричних розподілів (середнє арифметичне ($\bar{x}$); медіана (M_e); мода (M_o)), показники їх розсіювання (дисперсія (s^2); стандартне відхилення (s); коефіцієнт варіації (V) і асиметрії (коефіцієнт асиметрії (As), коефіцієнт ексцесу (Ex)).

Середнє арифметичне

Середнє арифметичне або просто середнє прийнято позначати тією ж літерою, що і варіанти спостережень, але над цією літерою ставиться символ усереднення − риска. Наприклад, якщо позначити досліджувану ознаку через (x), то середнє арифметичне буде позначатися − ($\bar{x}$). Середнє арифметичне може обчислюватися як по необробленим первинним даним, так і по згрупованим показникам. Точність обчислення по необроблених даних завжди вище, але процес обчислення виявляється трудомістким при великому обсязі спостережень. Обчислення середнього арифметичного для негрупованих даних здійснюється за формулою:

$$\overline{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum x_i}{n} \quad (1)$$

де n – обсяг спостережень;

x_i – варіанти спостережень.

Якщо дані згруповані, то застосовується формула:

$$\overline{x} = \frac{\sum n_i x_i}{n} \quad (2)$$

де n_i – частоти розрядів;

x_i – серединні значення розрядів.

Результати розрахунків середнього арифметичного за формулами (1) і (2) не завжди збігаються. Це пов'язано з тим, що в першому випадку беруться вихідні дані, а в другому – підсумовуються добутки частот розрядів і їх серединних значень. Для практичного розрахунку середнього арифметичного $\overline{x}$ скористаємося даними, наведеними в прикладі 1:

а) для незгрупованих даних

$$\overline{x} = \frac{59+48+53+\cdots+69}{65} = \frac{3952}{65} = 60{,}8 \text{ см}$$

б) для згрупованих даних – для наочності проміжні результати розрахунків для знаходження середнього арифметичного за формулою (2) наведені в табл. 3.

Таблиця 3 – Розрахунок середнього арифметичного результату стрибка вгору з місця спортсменів-баскетболістів

Номер розряду (i)	Середині значення (x_i)	Частоти (n_i)	$n_i x_i$
1	37	1	37
2	42	2	84
3	47	6	282
4	52	11	572
5	57	12	684
6	62	11	682
7	67	8	536
8	72	4	288
9	77	4	308
10	82	6	492
	Сума		3965

$$\bar{x} = \frac{37+84+282+\dots+492}{65} = \frac{3952}{65} = 61 \text{ см}$$

Крім середнього арифметичного існують інші характеристики, що визначають положення центра емпіричного розподілу. До них відносяться: медіана Ме − число розділяє упорядкований (по зростанню або зменшенням) ряд експериментальних даних на дві рівні частини; мода Мо − значення ознаки, що зустрічається в спостереженні найбільш часто. Медіана і мода є допоміжними характеристиками спостережень і використовуються рідка.

Характеристики розсіювання

Середнє арифметичне, медіана і мода є одними з найбільш інформативних характеристик розподілу, але вони не дають повної картини про ознаку, що варіює. Для того, щоб побачити в якому діапазоні розсіяні знайдені значення ознаки, обчислюють характеристики розсіювання: дисперсія s^2; середнє квадратичне відхилення або стандартне відхилення s; коефіцієнт варіації V.

Дисперсія для незгрупованих даних обчислюється за формулою:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{3}$$

де $\sum(x_i - \bar{x})^2$ − сума квадратів відхилень значень ознаки x від середнього арифметичного $\bar{x}$;

n-1 − число ступенів свободи, яка дорівнює кількості спостережень без одного.

Ппредставлену формулу важко застосувати на практиці (особливо при ручних методах обчислення), так як при збільшенні обсягу числа спостережень збільшується помилка, що виникає при підсумовуванні округленої середньої арифметичної і, крім того, збільшується небезпека зробити помилку при підсумовуванні багато розрядних значень x_i. Тому пропонується застосовувати цю формулу в перетвореному вигляді:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}, \tag{4}$$

для згрупованих даних:

$$s^2 = \frac{\sum n_i(x_i - \bar{x})^2}{n-1} = \frac{\sum n_i x_i^2 - \frac{(\sum n_i x_i)^2}{n}}{n-1}, \tag{5}$$

де n_i − частоти;

x_i серединні значення розрядів.

Середнє квадратичне відхилення або стандартне відхилення s розраховується за формулою:

$$s = \sqrt{S^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}, \qquad (6)$$

де s^2 – дисперсія.

Розмірність середньоквадратичного або стандартного відхилення на відміну від розмірності дисперсії збігається з одиницями вимірювання експериментальних даних, тому на практиці зазвичай використовують s, а не s^2.

Таблиця 4 - Розрахунок дисперсії результатів стрибка вгору з місця спортсменів-баскетболістів

№ $^n/_n$	X_i, *см*	X_i^2	№ $^n/_n$	X_i, *см*	X_i^2	№ $^n/_n$	X_i, *см*	X_i^2
1	2	3	1	2	3	1	2	3
1	59	3481	23	63	3969	45	61	3721
2	48	2304	24	81	6561	46	55	3025
3	53	2809	25	60	3600	47	73	5329
4	47	2209	26	77	5929	48	50	2500
5	57	3249	27	71	5041	49	70	4900
6	64	4096	28	57	3249	50	59	3481
7	62	3844	29	82	6724	51	50	2500
8	62	3844	30	66	4356	52	59	3481
9	65	4225	31	54	2916	53	83	6889
10	57	3249	32	47	2209	54	69	4761
11	57	3249	33	61	3721	55	67	4489
12	81	6561	34	76	5776	56	66	4356
13	83	6889	35	50	2500	57	47	2209
14	48	2304	36	57	3249	58	56	3136
15	65	4225	37	58	3364	59	60	3600
16	76	5776	38	52	2704	60	43	1849
17	53	2809	39	57	3249	61	54	2916
1	2	3	1	2	3	1	2	3
18	61	3721	40	40	1600	62	47	2209
19	60	3600	41	53	2809	63	81	6561
20	37	1369	42	66	4356	64	76	5776
21	51	2601	43	71	5041	65	69	4761
22	51	2601	44	61	3721	**сума**	**3952**	**248108**

Необхідно також додати, що в практичній статистиці часто потрібно визначити рівень однорідності вибіркових спостережень. Для цього використовується безрозмірний показник - коефіцієнт варіації V:

$$V = \frac{s}{\bar{x}} * 100\% \qquad (7)$$

Вважається, що якщо коефіцієнт варіації не перевищує 10%, то спостереження можна вважати однорідними.

Крім того, коефіцієнт варіації часто використовується при зіставленні (порівнянні) ступеня варіювання різних ознак, виражених в різних одиницях виміру.

Розглянемо розрахунок дисперсії стандартного відхилення і коефіцієнта варіації, використовуючи дані прикладу 1.

На початку обчислимо характеристики розсіювання по незгрупованим даним (проміжні розрахунки наведені в табл. 4).

За формулою (4) знаходимо:

$$s^2 = \frac{248108 - \frac{3952^2}{65}}{65-1} = 122.3 \text{ см}^2.$$

Стандартне відхилення складе: s=$\sqrt{s^2} = \sqrt{122{,}3}$=11 см, звідси коефіцієнт варіації:

$$V = \frac{11}{60{,}8} * 100 = 18{,}1\%$$

Для згрупованих даних (проміжні розрахунки наведені в табл. 5):

Таблиця 5.

x_i, *см*	n_i	$n_i x_i$	$n_i x_i^2$
37	1	37	1369
42	2	84	3528
47	6	282	13254
52	11	572	29744
57	12	684	38988
62	11	682	42284
67	8	536	35912
72	4	288	20736
77	4	308	23716
82	6	492	40344
сума		3965	249875

По формулі (5) знаходимо:

$$s^2 = \frac{249875 - \frac{3965^2}{65}}{65-1} = 125{,}2 \text{ см}^2$$

стандартне відхилення складе: s = $\sqrt{125{,}2}$ = 11,2 см, звідси, коефіцієнт варіації

Коефіцієнт Стьюдента розраховується за формулою

$$t = \frac{x_{ср} - M}{D/\sqrt{n}}$$

Де D=s, а M- математичне сподівання

t-критерій Стьюдента — загальна назва для класу методів статистичної перевірки гіпотез (статистичних критеріїв), заснованих на порівнянні з розподілом Стьюдента. Найчастіші випадки застосування t-критерію пов'язані з перевіркою рівності середніх значень у двох вибірках

Коефіцієнт Фішера

F-тестом або критерієм Фішера — називають будь-який статистичний критерій, тестова статистика якого при виконанні нульової гіпотези має розподіл Фішера (F-розподіл).

Статистика тесту так чи інакше зводиться до відношення вибіркових дисперсії (сум квадратів, ділених на «ступеня свободи»). Щоб статистика мала розподіл Фішера, необхідно, щоб чисельник і знаменник були незалежними випадковими величинами і відповідні суми квадратів мали розподіл Хі квадрат. Для цього потрібно, щоб дані мали ***нормальний розподіл***. Крім того, передбачається, що дисперсія випадкових величин, квадрати яких підсумовуються, однакова.

Тест проводиться шляхом порівняння значення статистики з критичним значенням відповідного розподілу Фішера при заданому рівні значимості. Відомо, що якщо

$$F \sim F(m,n), \text{ то } 1/F \sim F(n,m).$$

Крім того, квантилі розподілу Фішера мають властивість

$$F_{1-\alpha} = 1/F_{\alpha}.$$

Тому зазвичай на практиці в чисельнику бере участь потенційно велика величина, в знаменнику — менша і порівняння здійснюється з «правою»

Квантиль розподілу. Проте тест може бути і двостороннім і одностороннім. У першому випадку при рівні значущості α використовують квантиль

$F_{\alpha/2}$,

А при односторонньому тести F_α

Більш зручний спосіб перевірки гіпотез — за допомогою p-значення p(F) — імовірністю того, що випадкова величина з даними розподілом Фішера перевищить дане значення статистики. Якщо p(F) (для двостороннього тесту — 2p(F) менше рівня значущості α , то нульова гіпотеза відкидається, в іншому випадку приймається.

Формула Фішера для розрахунку

$$F = \frac{S_1^2}{S_2^2}$$