Практична робота № 4

Тема: Аналіз варіаційних рядів кількісної та якісної мінливості

4.1. Кількісна мінливість

4.2. Обчислення статистичних характеристик малої вибірки.

4.3 Обчислення статистичних характеристик великої вибірки.

4.4 Якісна мінливість

Мета: Навчись обчислювати статистичні характеристики малої і великої вибірки та показники мінливості якісної ознаки

4.1. Кількісна мінливість

Під кількісною мінливістю розуміють таку, в якій відмінність між варіантами виражається кількісними показниками: кількістю зерен, масою, висотою тощо. Розрізняють неперервну та перервну кількісну мінливість. До неперервної мінливості належать ті об'єкти, які виражають і цілими, і дробовими числами - це маса, розмір і об'єм досліджуваних об'єктів (маса врожаю або плоду, довжина і висота рослин, площа листя, об'єм коріння чи бульби). До перервної мінливості належать об'єкти, які обліковують поштучно (кількість колосків, листків, коренеплодів, рослин тощо) і виражаються лише цілими числами.

Конкретний хід аналізу варіаційних рядів кількісної мінливості залежить від об'єму вибірки.

Основними статистичними характеристиками кількісної мінливості для малих і великих вибірок є середня арифметична (x), дисперсія (S^2), стандартне відхилення (S), помилка середнього арифметичного (S_x), коефіцієнт варіації (V) і відносна помилка вибіркової середньої (S_x %).

4.2. Обчислення статистичних характеристик малої вибірки

Для малих вибірок обчислюють такі статистичні характеристики: середні арифметичні, дисперсії, стандартні відхилення, коефіцієнти варіювання, похибки вибіркових середніх, відносні похибки тощо.

Середня арифметична (x). Для обчислення цієї характеристики варіюючі ознаки позначають знаком X, а кількість повторностей – n. Тоді середню арифметичну розраховують за формулою:

$$\bar{x} = \frac{\sum x}{n}$$

Для прикладу скористаємось даними досліду з кукурудзою на полі з рівномірним і нерівномірним внесенням добрив (табл. 1).

Таблиця 1.

Кількість листя на одній рослині залежно від рівномірності внесення добрив

Внесення добрив	Повторність			
	I	II	III	IV
Рівномірне	14	18	13	15
Дуже нерівномірне	16	7	20	17

Для варіанта з рівномірним внесенням добрив середня арифметична

$$\overline{X_1} = \frac{\sum X}{n} = \frac{14 + 18 + 13 + 15}{4}$$

Для варіанта з нерівномірним внесенням добрив середня арифметична:

$$\overline{X_2} = \frac{\sum X}{n} = \frac{16 + 7 + 20 + 17}{4}$$

Отже, середні арифметичні однакові, але розмах варіювання (R) в різних варіантах буде неоднаковим. При рівномірному внесенні добрив $R_2=X_{max}-X_{min}=18-13=5$, а при дуже нерівномірному - $R_2=X_{max}-X_{min}=20-7=13$, тобто кількість листя сильніше варіює при нерівномірному внесенні добрив, коли створюється строкатість родючості грунту.

Середня арифметична є основною статистичною характеристикою кожного варіаційного ряду, а всі інші характеристики лише пояснюють основну. До останніх належить і *дисперсія* (S^2) - середній квадрат відхилень кожного члена варіаційного ряду (X_1, X_2 ... X_n) від середньої арифметичної. Це показник, який повніше за розмах варіації характеризує варіаційні ряди. Дисперсію обчислюють за формулою

$$S^2 = \frac{\Sigma(X - x)^{-2}}{n - 1}$$

Для її розрахунків складається допоміжна таблиця 2, в яку заносяться дані з попередньої таблиці.

Таблиця 2

Обчислення квадратів відхилень від середньої арифметичної

Внесення добрив							
рівномірне				дуже нерівномірне			
Повтор-ність	*X*	*X*-$\bar{x}$	$(X-\bar{x})^2$	Повтор-ність	*X*	*X*-$\bar{x}$	$(X-\bar{x})^2$
1	14	-1	1	1	16	1	1
2	18	3	9	2	7	-8	64
3	13	-2	4	3	20	5	25
4	15	0	0	4	17	2	4
	x=15		$\Sigma(X-\bar{x})^2=14$		x=15		$\Sigma(X-\bar{x})^2=94$

Підставивши одержані дані у вище наведену формулу, матимо такі дисперсії:

$$S_1^2 = \frac{14}{4-1} = 4{,}67$$

$$S_2^2 = \frac{94}{4-1} = 31{,}3$$

Їх порівняння між собою показує, що в другому варіанті, де добрива вносились дуже нерівномірно, дисперсія була майже в 7 разів більшою, ніж при рівномірному внесенні добрив.

Дисперсія використовується не тільки для характеристики варіювання досліджуваних показників, а й для обчислення стандартного відхилення (S) за формулою:

$$S = \sqrt{S^2}$$

Для першого варіанта $S_1 = \sqrt{4{,}67} = 2{,}16$ шт.

для другого - $S_1 = \sqrt{31{,}3} = 5{,}59$ шт.

Стандартне відхилення виражається у таких самих одиницях, як і середня арифметична.

Якщо у дослідженнях порівнюють мінливість ознак, які мають різні одиниці виміру (центнери, штуки, сантиметри, квадратні сантиметри тощо), то

дисперсія і стандартне відхилення для таких порівнянь непридатні. У таких випадках доцільно користуватися коефіцієнтом варіювання.

Коефіцієнт варіювання (V) - це відношення стандартного відхилення до середньої арифметичної, яке виражають у процентах і обчислюють за формулою:

$$V\,\% = \frac{S \cdot 100}{\bar{x}}.$$

Для першого варіанту $V_1 = \frac{2{,}16 \times 100}{15} = 14{,}4\%$

для другого $V_2 = \frac{5{,}59 * 100}{15} = 37{,}3\%$

Отже, строкатість родючості грунту призводить до невирівняності росту рослин, тому що варіювання кількості листя значно більше при дуже нерівномірному внесенні добрив.

Варіювання умовно вважають незначним, якщо коефіцієнт варіювання становить 10 %, середнім – 10-20, значним - понад 20 %.

Варіювання врожаю більшості польових культур становить 8- 12 %. Менше варіювання врожаю у культур звичайного рядкового способу сівби, більше - у просапних.

Середні арифметичні мають свої похибки, які спричинюються внаслідок неповного представництва вибіркової сукупності. Ці похибки властиві лише вибірковому методу досліджень, а їх чисельне значення залежить від ступеня мінливості досліджуваних ознак і обсягу вибірки. *Похибку вибіркової середньої (S_x)* обчислюють за формулою

$$S_{\bar{x}} = \sqrt{\frac{\sum (X - \bar{x})^2}{n(n-1)}}.$$

В наведеному вище прикладі для першого варіанта значення похибки

$$S_{x1} = \sqrt{\frac{14}{4(4-1)}} = 1{,}08$$

а для другого

$$S_{x2} = \sqrt{\frac{94}{4(4-1)}} = 2{,}8$$

Значення похибок використовують для граничної оцінки середніх арифметичних за формулою $x \pm tS_x$. При оцінці на рівні ймовірності $P_{0,95}$ значення t дорівнює 2, а при $P_{0,99}$ - 3. З врахуванням цих показників межове значення середньої арифметичної для першого варіанта становитиме відповідно 15±2•1,08 та 15±3•1,08. Це значить, що на рівні $P_{0,95}$ межові величини середньої арифметичної будуть у межах 12,84÷17,16, а на рівні $P_{0,99}$ -11,76÷18,24.

Відношення похибки вибіркової середньої до її середньої арифметичної, виражене у процентах, називається *відносною похибкою вибіркової середньої* (S_x *%*). Визначають її за формулою

$$S_{\bar{x}}\% = \frac{S_{\bar{x}} \cdot 100}{\bar{x}}.$$

Для варіанта з рівномірним внесенням добрив відносна похибка вибіркової середньої $S_x\% = \frac{1{,}08 \times 100}{15} = 7{,}2\%$, а для варіанта з дуже нерівномірним внесенням добрив $S_x\% = \frac{2{,}8 \times 100}{15} = 18{,}7\%$

Залежно від значення відносної похибки роблять висновки про точність досліду - це різниця між 100 % і відносною похибкою (T = 100% - S_x *%*). Умовно точність вважають високою, якщо значення не перевищує 3%, середньою - становить 3-6 і низькою - перевищує 7 %. Проте у дослідах, проведених на грунтах із значною строкатістю родючості, значення відносної похибки буває і більшим. Для підвищення точності таких досліджень треба дбати про те, щоб досліди проводили на вирівняних за родючістю площах.

4.3 Обчислення статистичних характеристик великої вибірки

Для великих вибірок (n > 30) статистичні характеристики можна обчислювати способом добутків. Слід зазначити, що він досить трудомісткий, особливо для багатозначних чисел та для дуже великих обсягів вибірки. Найбільш зручним є спосіб умовної середньої, тобто від довільного початку (A). Для цього одержані дані розміщують у зростаючому порядку, поділяють їх на групи з певним інтервалом (i) і в їх межах визначають частоти (f), тобто кількість членів у кожній групі варіаційного ряду.

Приклад. Беруть підряд 40 колосів озимої пшениці (n = 40), підраховують в кожному з них кількість зерен (шт.) і дані розміщують в зростаючому порядку: 27, 31, 33, 33, 35, 36, 36, 36, 36, 36, 36, 36, 36, 36, 37, 37, 37, 37, 37, 37, 37, 37, 37, 38, 38, 38, 38, 38, 38, 38, 39, 39, 39, 40, 40, 41, 41, 42, 43, 44.

Перед обробкою варіаційного ряду визначають число груп (*Чг*) за формулою $Чг = \sqrt{n} = \sqrt{40} \approx 6$ груп.

Як правило, коли n знаходиться у межах 30-60, беруть 6-7 груп, 60-100 - 7-8, більше 100 - 8-15 груп. Після цього обчислюють інтервал групи (*i*) за формулою

$$i = \frac{X_{max} - X_{min}}{Чг} = \frac{44 - 27}{6} \approx 3$$

Значення інтервалу групи, якщо він виражений не цілим числом, для зручності розрахунків часто заокруглюють до цілих чисел, а наступні розрахунки ведуть згідно форми таблиці 3.

Таблиця 3

Допоміжна таблиця для обробки варіаційного ряду великої вибірки кількості зерен у колосі

Інтервал групи	Середнє значення групи, *X*	Частота. *f*	Відхилення від довільного початку, *X* - *A*	*f* (*X*-*A*)	(*X*-*A*)	*f* (*X*-*A*)^2
27-29	28	1	-9	-9	81	81
30-32	31	1	-6	-6	36	36
33-35	34	3	-3	-9	9	27
36-38	37(A)	25	0	0	0	0
39-41	40	7	3	21	9	63
42-44	43	3	6	18	36	108
		Σ(*f*)=40		Σ*f*(*X*-*A*) = 15		Σ*f*(*X*-*A*)^2 = 315

Перша група починається найменшим значенням кількості зерен в колосі - 27 шт., до якого додається інтервал групи «*i*» без одиниці (3 - 1 = 2). Обчислюють середнє значення групи. Одне з цих значень вибирають за доцільний початок (*A*), користуючись таким принципом: він повинен займати середнє положення серед всіх середніх значень і належати до групи з найбільшою частотою. Таким значенням є 37.

Сума всіх частот повинна дорівнювати обсягу вибірки *(n)*: $\Sigma f = n = 40$. Віднімаючи значення довільного початку від середніх значень груп матимемо відхилення X - A з відповідним знаком («+» чи «-»). Знак «+» можна опускати.

Статистичні характеристики великих вибірок обчислюють за відповідними формулами в такій послідовності:

1. Довільний момент першого ступеня $в = \frac{\sum f(X-A)}{n} = \frac{15}{40} = 0,37$.

2. Середня арифметична $\bar{x} = A + в = 37 + 0,37 \approx 37,4$ шт.

3. Коректуючий фактор $C = \frac{\left[\sum f(X-A)\right]^2}{n} = \frac{15^2}{40} = \frac{225}{40} = 5,63$.

4. Дисперсія $S^2 = \frac{\sum f(X-A)^2 - C}{n-1} = \frac{315-5,63}{40-1} = \frac{309,37}{39} = 7,93$.

5. Стандартне відхилення $S = \sqrt{S^2} = \sqrt{7,93} = 2,82$ шт.

6. Коефіцієнт варіації $V = \frac{S \cdot 100}{\bar{x}} = \frac{2,82 \cdot 100}{37,4} = 7,5\%$.

7. Похибка вибіркової середньої $S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{2,82}{\sqrt{40}} = \frac{2,82}{6,32} = 0,45$ шт.

8. Відносна похибка середнього арифметичного

$$S_{\bar{x}}\% = \frac{S_{\bar{x}} \cdot 100}{\bar{x}} = \frac{0,45 \cdot 100}{37,4} = 1,2\%.$$

9. Інтервальна оцінка середньої арифметичної для двох рівнів імовірності

9.1.Обчислюють число ступенів вільності

$$v = n - 1 = 40 - 1 = 39.$$

9.2. За таблицею Стьюдента (додаток 2) знаходять критерії $t_{0,95}(2,04)$ та $t_{0,99}(2,75)$, які підставляють у формулу $x \pm Sx \cdot t$

Добуток називають ділянкою індивідуального розсіювання:

$t_{0,95} = 0,45 \cdot 2,04 = 0,92$;

$t_{0,99} = 0,45 \cdot 2,75 = 1,24$.

Межі допустимого коливання кількості зерен в колосі встановлюються відніманням значення від середнього арифметичного і додаванням цього значення до нього. Виходячи з цих розрахунків для рівня $P_{0,95}$ коливання кількості зерен в колосі повинне бути в межах 36,5÷38,3, а для рівня ймовірності $P_{0,99}$ - 36,2÷38,6.

Але в зв'язку з тим, що кількість зерен в колосі являє собою переривчасту (дискретну) кількісну мінливість і виражається цілими числами, то і розраховані межі повинні виражатись такими ж числами. Після заокруглення для рівнів $P_{0,95}$ $P_{0,99}$ ними відповідно будуть 37÷38 і 36÷39. Значення кількості

зерен в колосі, які не вмішаються в ці межі, бракуються і не приймають участі в подальших дослідженнях.

10. Зобразимо варіаційний ряд графічно (рис. 1)

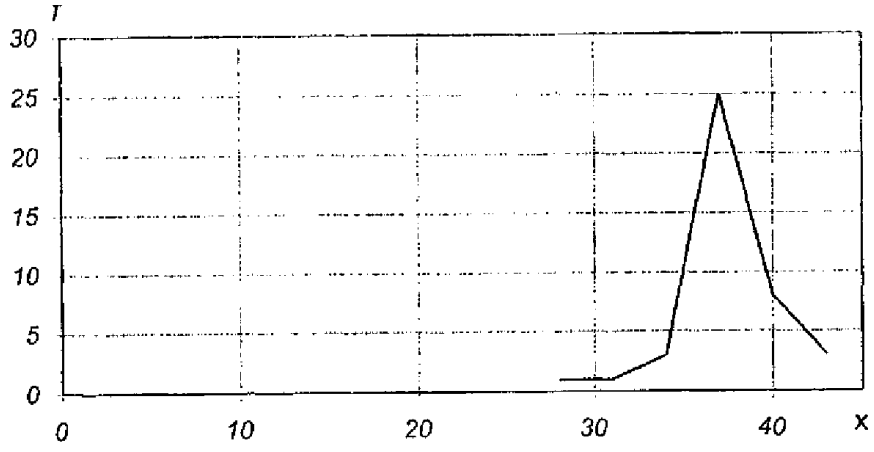

Рис. 1. Графічне зображення варіаційного ряду

Висновки

1. Середня арифметична кількості зерен в колосі дорівнює 37,4 шт.
2. Коефіцієнт варіації 7,5 % свідчить про незначне варіювання кількості зерен в колосі пшениці.
3. Значення відносної похибки 1,2 % свідчить про досить високу точність обчислення середньої арифметичної.
4. До даного варіаційного ряду на рівні $P_{0,95}$ належать колоси з кількістю зерен 37+38 шт., а на рівні $P_{0,99}$– 36÷39 шт. Всі інші дати, які не ввійшли в інтервал оцінки на обох рівнях надійної імовірності, не належать до даного варіаційного ряду і вважаються не характерними для нього.

На графіку крива варіаційного ряду має одну вершину, що свідчить про однорідність вибірки.

4.4 Якісна мінливість

В експериментальній роботі мають справу не тільки з кількісною, але і з якісною мінливістю ознак, яка не має градацій. У кожного окремо взятого члена вибірки така ознака може або бути, або вона відсутня. Наприклад: рослини можуть бути здоровими і хворими; озима пшениця може бути зимостійкою і ні, стійкою до вилягання і нестійкою, зерно її може мати певне забарвлення і не мати.

Групування даних якісної мінливості зводиться до розподілу сукупності об'єктів спостереження на групи з різними якісними ознаками.

Для аналізу варіаційних рядів якісної мінливості обчислюють такі статистичні характеристики: частку наявності ознаки (p); частку відсутності ознаки (∂); показник мінливості якісної ознаки - стандартне відхилення (S); коефіцієнт варіації (V_p) і помилку частки (S_p). Частка наявності ознаки - це відношення кількості об'єктів і даною ознакою (n) до загального обсягу вибірки (N). Обчислюють її за формулою $p = n : N$.

Приклад. Після збирання картоплі виявилось, що у сорту Луговська із 100 бульб (N_1) не уражених фітофторозом було 85 (n), а у сорту Слов'янка із 100 бульб (N_2) не ураженими виявилось 97 (n_2). Визначимо частку наявності ознаки (p_1 і p_2) у кожного з цих сортів:

$p_1 = n_1 : N_1 = 85 : 100 = 0{,}85$ частки або 85 %;

$p_2 = n_2 : N_2 = 97 : 100 = 0{,}97$ частки або 97 %;

Частка відсутності ознаки - це різниця між одиницею і часткою наявності ознаки. Цю частку обчислюють за формулою $q = 1 - p$. Для досліджуваних сортів картоплі Луговська і Слов'янка частка відсутності ознаки (або частка ураження бульб) q_1 і q_2 відповідно становитиме:

$q_1 = 1 — p_1 = 1 - 0{,}85 = 0{,}15$ частки або 15 %;

$q_2 = 1 - p_2 = 1 - 0{,}97 = 0{,}03$ частки або 3 %.

Показники мінливості якісної ознаки. Якщо досліджуваний об'єкт має лише дві градації, як у вищенаведеному прикладі, - кількість здорових (якісних) і хворих бульб картоплі - показник мінливості (S) обчислюють за формулою $S = \sqrt{pq}$. Для сорту Луговська $S_1 = \sqrt{p_1 q_1} = \sqrt{0{,}85 \cdot 0{,}15} = 0{,}35$, а для сорту Слов'янка $S_2 = \sqrt{p_2 q_2} = \sqrt{0{,}97 \cdot 0.03} = 0{,}17$. Максимальне значення (0,5) мінливості будемо мати за умови, якщо $p = q = 0{,}5 (S_{max} = \sqrt{pq} = \sqrt{0{,}5 \cdot 0{,}5} = 0{,}5$.

Якщо об'єкт досліджень має не дві, а більше градацій, то для такої вибірки показник якісної мінливості (S) обчислюють за формулою $S = \sqrt{p_1 \cdot p_2 \cdot p_n}$, де $p_1, p_2, \dots p_n$ - частки ознак із загального обсягу вибірки; n - кількість градацій ознак.

Показник мінливості використовується для визначення *коефіцієнта варіювання* (V_p) як відношення показника мінливості (S) до його максимального значення (S_{max}), виражене в процентах. Обчислюється коефіцієнт варіювання в процентах за формулою.

Для сорту Луговська коефіцієнт варіювання становитиме

$$V_{p1} = \frac{S_1 \times 100}{S_{max}} = \frac{0{,}35 \cdot 100}{0{,}5} = 70\%$$

для сорту Слов´янка

$$V_{p2} = \frac{S_{12} \times 100}{S_{max}} = \frac{0{,}17 \cdot 100}{0{,}5} = 34\%$$

Максимальне значення коефіцієнта варіювання -100 % буває при $S = S_{max} = 0{,}5$.

Для оцінки точності в визначенні вибіркових середніх арифметичних при якісній мінливості вираховують *похибку вибіркової середньої арифметичної* за формулою $S_p = \frac{S}{\sqrt{N}}$. Стосовно до альтернативної мінливості вона матиме такий вигляд $S_p = \frac{S}{\sqrt{N}} = \frac{\sqrt{pq}}{\sqrt{N}} = \sqrt{\frac{pq}{N}}$

для сорту Луговська $S_{p_1} = \sqrt{\frac{p_1 q_1}{N_1}} = \sqrt{\frac{0{,}85 \cdot 0{,}15}{100}} = 0{,}036$,

а для сорту Слов янка-

$$S_{p_{21}} = \sqrt{\frac{p_2 q_2}{N_2}} = \sqrt{\frac{0{,}97 \cdot 0{,}03}{100}} = 0{,}017$$

Інтервальну оцінку помилки вибіркової середньої роблять за формулою $p \pm tS_p$. На рівні імовірності $P_{0,95}$ значення $t = 1{,}98$, а на рівні $P_{0,99}$ $t = 2{,}63$.

Для сорту Луговська ці інтервали становитимуть для рівня $P_{0,95}$ 0,85 - 1,98 • 0,036 = 0,78 і 0,85 + 1,98 • 0,036 = 0,092, тобто 0,78÷0,92.

Для сорту Слов'янка на цьому рівні нижчий інтервал становитиме 0,97 - 1,98 • 0,017 = 0,94, а вищий 0,97 + 1,18 • 0,017 = 1,0, тобто 0,94÷1,00.

Результати спостережень дозволять вважати, що частка здорових бульб картоплі сорту Луговська може становити 77-92 %, а сорту Слов'янка – 94-100 %. Це значить, що сорт Слов'янка стійкіший до хвороби, ніж сорт Луговська.

Для визначення достовірності різниці між частками наявності ознак обчислюють фактичний критерій Стьюдента ($t_ф$):

$$t_ф = \frac{p_2 p_1}{\sqrt{\frac{p_2 q_2}{N_2} + \frac{p_1 q_1}{N_1}}} = \frac{0{,}97 - 0{,}85}{\sqrt{\frac{0{,}97 \cdot 0{,}03}{100} + \frac{0{,}85 \cdot 0{,}15}{100}}} = \frac{0{,}92}{0{,}03957} = 3{,}03$$

Фактичний критерій Стьюдента порівнюють з теоретичним, який беруть із додатку 2 за числом ступенів вільності (v_p)

$v_p = (N_2 - 1) + (N_1 - 1) = (100 - 1) + (100 - 1) = 99 + 99 = 198.$

критеріїв При v_p, яке дорівнює 100 і більше, критерій $t_{0,95} = 1{,}95$, а $t_{0,99} = 2{,}58$.

Якщо критерій фактичний дорівнює теоретичному або більший за нього, то різниця достовірна. Користуючись цим правилом роблять висновок -

оскільки критерій Стьюдента фактичний між варіантами становить 3,03, що значно більше теоретичних на обох рівнях надійної ймовірності, то у картоплі сорту Слов'янка спостерігається достовірне зниження ураження бульб фітофторозом порівняно з сортом Луговська.

Запитання для самоперевірки

1. Що називається варіаційним рядом і які його складові?
2. Які статистичні показники обчислюють при аналізі варіаційних рядів?
3. Які основні статистичні показники якісної мінливості?