НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

Кафедра комп'ютерних систем і мереж

Методичні вказівки

до виконання лабораторних робіт з дисципліни

Математичні моделі та планування експерименту

для аспірантів денної і заочної форм навчання

КИЇВ 2016

УДК 004.4

Наведено матеріал щодо виконання лабораторних робіт з дисципліни "Математичні моделі та планування експерименту" з використанням програм пакету Microsoft Office табличного процесора MS Office Excel.

Рекомендовано Вченою радою факультету інформаційних технологій Національного університету біоресурсів і природокористування України протокол № 4 від 16.11.2016 року.

Укладачі: доц. Осипова Т.Ю., ас. Савицька Я.А.

Рецензенти:
д.фіз.-мат.н., провідний науковий співробітник Інституту фізики НАН України **І.І.Ясковець**
к.фіз.-мат.н., доц. кафедри економічної кібернетики НУБіП України **Т.В. Коваль**

ЗМІСТ

Лабораторна робота № 1
Вибірковий метод. Точкові оцінки вибіркових статистичних показників

Мета роботи: отримання практичних навичок з визначення точкових оцінок статистичних показників вибіркової сукупності.

Теоретичні відомості

Статичні об'єкти, системи, процеси тощо, як правило, відзначаються складністю, залежністю від часу і великої кількості різноманітних факторів, вплив яких наперед неможливо врахувати, або передбачити. Результати вимірювань ознак таких об'єктів називають *випадковими* і до їх аналізу застосовують відповідний математичний апарат. *Випадковою* називають величину, яка в результаті вимірювань може прийняти одне можливе наперед невідоме значення, що залежить від випадкових чинників, дія яких наперед не може бути врахованою. Розрізняють випадкові величини які можуть приймати лише окремі, ізольовані значення, і випадкові величини, можливі значення яких заповнюють деякий проміжок.

Дискретною (перервною) називають випадкову величину, яка приймає окремі ізольовані значення з визначеними ймовірностями.

Неперервною називають випадкову величину, яка може приймати всі значення із деякого кінцевого або нескінченного проміжку. Результати багаторазових вимірювань певних ознак об'єктів відносяться до *випадкових дискретних величин.*

Генеральна і вибіркова сукупності

В експериментальних дослідженнях використовують поняття *генеральної* і *вибіркової* сукупностей. При проведенні суцільних досліджень, тобто досліджень кожного із об'єктів сукупності відносно ознаки, яка цікавить дослідника, говорять про належність цих об'єктів до генеральної сукупності. *Генеральною* називають сукупність об'єктів кількість яких n прямує до нескінченності, тобто $n \to \infty$. Як правило, проводити суцільне дослідження неможливо і/або недоцільно. Зазвичай від усієї кількості об'єктів певним чином відбирають частину об'єктів і проводять дослідження відносно певної ознаки відібраних об'єктів. Сукупність випадково відібраних із генеральної сукупності об'єктів називають *вибірковою сукупністю,* або *вибіркою. Обсягом* сукупності називають кількість n об'єктів цієї сукупності. Вибірка повинна правильно відтворювати властивості генеральної сукупності, від якої вона відібрана. Тобто вона повинна бути *репрезентативною.*

Визначення основних статистичних показників вибіркової сукупності випадкових величин

Мета математичної обробки результатів багаторазових вимірювань полягає в обчисленні найвірогіднішого значення величини, що визначається, та оцінці його точності і надійності. Така обробка ґрунтується на методах теорії імовірності та математичної статистики, які застосовуються для аналізу випадкових дискретних величин.

Законом розподілу дискретної випадкової величини називають відповідність між можливими її значеннями і ймовірностями їх появи. Закон розподілу може бути заданий *таблично, аналітично* і *графічно.*

За достатньо великої кількості вимірювань випадкових величин їх поява підпорядковується нормальному закону розподілу (закону Гауса), формула якого має вигляд

$$f(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

f(x,μ,σ) - щільність ймовірності; *μ* - математичне сподівання випадкової величини (центр групування її значень**);** σ^2 - дисперсія випадкової величини (міра розсіювання значень випадкової величини відносно центру групування); *σ* середнє квадратичне відхилення випадкової величини (характеристика розсіювання значень випадкової величини відносно центру групування, яка дорівнює кореню квадратному з дисперсії); *e* - основа натурального логарифму. Таким чином, нормальний закон розподілу характеризується лише *двома параметрами, μ* - математичним сподіванням і *σ* - середнім квадратичним відхиленням.

Нормальний закон розподілу може точно описувати лише нескінченно велику кількість випадкових величин (генеральну сукупність).

Однак його застосовують і для опису репрезентативної вибіркової сукупності. У вибірках зі скінченим числом вимірювань п, точне обчислення μ та σ неможливе. Замість них розраховують середнє вибіркове значення $\overline{x}_{в}$, вибіркове середнє квадратичне відхилення σв вибіркову дисперсію $D_в$, та статистичні оцінки відповідних показників генеральної сукупності.

Таким чином, припускаючи, що експериментальні дані підпорядковуються нормальному закону розподілу, обчислюють параметри, що його характеризують: вибіркові середнє значення, дисперсію і середнє квадратичне відхилення.

Середнє вибіркове значення $\overline{x}_{в}$ (середнє арифметичне значення ознаки вибіркової сукупності, що досліджується) визначається за формулою

$$\bar{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2)$$

x_{i} – значення i-ого вимірювання, n – кількість вимірювань.

Вибіркова дисперсія $D_в$ (середнє арифметичне квадратів відхилень значень ознаки x_{i} що досліджується, від середнього вибіркового значення $\bar{x}_{\text{в}}$) визначається за формулою

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^2 \qquad (3)$$

Вибіркове середнє квадратичне відхилення $\sigma_в$ визначається за формулою

$$\sigma_в = \sqrt{D_в} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2} \qquad (4)$$

Для оцінки варіації даних використовують вибірковий коефіцієнт варіації $C_в$, який обчислюється за формулою

$$C_в = \frac{\sigma_в}{x_в}\cdot 100\% \qquad (5)$$

Коефіцієнт варіації $C_в$ застосовують для порівняння варіації рядів спостережень, що відрізняються середніми значеннями і дисперсією. Окрім цього, $C_в$ - величина безрозмірна і може використовуватися для порівняння варіації рядів спостережень, що мають різні одиниці вимірювання.

Оцінка відхилення експериментального розподілу від нормального

Обчислення статистичних показників правомірне за умови підпорядкування експериментальних даних нормальному закону розподілу. При вивченні невідомих (експериментальних) розподілів, або розподілів, що відрізняються від нормального, виникає потреба кількісно оцінити цю відмінність. З цією метою застосовують спеціальні характеристики, зокрема асиметрію (коефіцієнт асиметрії) і ексцес (коефіцієнт ексцесу). Асиметрія показує, наскільки розподіл даних несиметричний відносно нормального розподілу. Якщо асиметрія є величиною додатною, то більша частина даних має значення, що перевищує середнє вибіркове $\bar{x}_{\text{в}}$. Якщо асиметрія менше нуля, то більша частина даних має значення менше за $\bar{x}_{\text{в}}$. Ексцес оцінює крутість, тобто величину більшого, або меншого підйому вершини графіка розподілу експериментальних даних порівняно з вершиною графіка

нормального розподілу. Якщо ексцес є величиною додатною, то вершина графіка експериментального розподілу вище нормального, якщо ексцес менше нуля, то - нижче нормального. Для нормального розподілу ці показники дорівнюють нулю. Якщо для розподілу, що вивчається, асиметрія і ексцес знаходяться в межах ±0,5, то можна припустити, що експериментальний закон розподілу близький до нормального. При цьому припускають, що емпіричний і теоретичний нормальні розподіли мають однакові математичне сподівання (середнє) і дисперсію. За умови малої кількості спостережень, перед використанням асиметрії та ексцесу для оцінки близькості експериментального розподілу до нормального треба оцінити точність визначення вказаних характеристик.

Асиметрія А визначається за формулою

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^2\right)^{\frac{3}{2}}} \qquad (6)$$

Ексцес Е визначається за формулою

$$E = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_{\text{в}}\right)^2\right)^2} - 3 \qquad (7)$$

Застосування вбудованих функцій MS Excel для визначення статистичних показників вибіркової сукупності

У MS Excel основні статистичні характеристики вибіркової сукупності визначаються за допомогою вбудованих функцій категорії **Статистические**. Для їх використання виконують такі дії:

- **Вставка=>Функция**;
- **або** користуються піктограмою *fx* (**Вставка функции**), яка розташована на панелі інструментів;
- у діалоговому вікні **Мастер функций**, що відкриється, у полі **Категория** вибирають **Статистические**;
- у полі **Выберите функцию** вибирають потрібну функцію;
- у діалоговому вікні **Аргументы функции**, що відкриється після вибору функції, вводять діапазон комірок з вхідними даними, або перевіряють, чи правильно він заданий, якщо діапазон комірок був попередньо вибраний;

- натискують кнопку **ОК**.

Перелік і призначення статистичних функцій наведено в табл. 1.

Вид функції	**Призначення**
СЧЕТ	Обсяг вибірки
СРЗНАЧ	Вибіркове середнє значення
ДИСПР	Вибіркова дисперсія
СТАНДОТКЛОНП	Вибіркове середнє квадратичне відхилення
МАКС	Максимальне число вибіркової сукупності
МИН	Мінімальне число вибіркової сукупності
СКОС	Показник асиметрії A
ЭКСЦЕСС	Показник ексцесу E

Визначення вибіркових статистичних показників множини ознак, що досліджуються

Якщо об'єкт, процес, або система, що досліджуються, характеризуються не однією ознакою, а множиною ознак, які представлені своїми рядами спостережень, то для визначення множини статистичних показників виконують дії:

- визначають статистичний показник однієї ознаки за допомогою відповідної вбудованої функції категорії **Статистические**;
- розповсюджують, або копіюють вбудовану функцію на діапазон комірок, що потребують аналогічного виду обчислень.

Виявлення і вилучення анормальних даних

Однією із причин, коли за показниками асиметрії і ексцесу експериментальний закон розподілу не можна вважати близьким до нормального, є наявність у вибірковій сукупності анормальних даних, тобто таких даних, які за своїми значеннями різко відрізняються від решти. Такі дані не можна віднести до нормально розподілених. У разі підозри на наявність анормальних даних перевірці підлягають максимальне і мінімальне вибіркові значення.

Показники анормальності $V_{\max}$ і $V_{\min}$ визначають за формулами

$$V_{\max} = \frac{\left(x_{\max} - \bar{x}_в\right)}{\sigma_в}, \qquad V_{\min} = \frac{\left(\bar{x}_в - x_{\min}\right)}{\sigma_в}, \tag{8}$$

де x_{max}, x_{min} – відповідно максимальне і мінімальне значення експериментальних даних; $\overline{X}_{в}$ - вибіркове середнє значення; $\sigma_в$ – вибіркове середнє квадратичне відхилення.

Для перевірки анормальності даних застосовують правило трьох сигм: якщо випадкова величина розподілена нормально, то абсолютна величина її відхилення від математичного сподівання не перевищує потроєного середнього квадратичного відхилення. Тобто, якщо показники анормальності $V_{max,min} \leq 3$, то відповідне максимальне і/або мінімальне значення не є анормальним і з подальшого аналізу не вилучається. Якщо $V_{max,min} > 3$, то відповідне максимальне і/або мінімальне значення є анормальним і його з подальших досліджень виключають.

Після вилучення анормальних даних числові статистичні показники вибірки обчислюють повторно і перевіряють наступні максимальне і мінімальне значення на анормальність. Якщо при розрахунках за формулами використовувати посилання на адреси комірок з потрібними значеннями, то в Excel перерахунок показників виконається автоматично.

<u>Програма виконання роботи</u>

1. Завантажити табличний процесор MS Excel.
2. Вибрати із табл. 1 результати вимірювань трьох довільних показників якості ґрунту і оформити їх у вигляді таблиці на аркуші Excel.
3. Під останнім значенням першого ряду спостережень за допомогою вбудованих функцій категорії **Статистические**: **СЧЕТ**, **СРЗНАЧ**, **ДИСПР**, **СТАНДОТКЛОНП**, **МАКС**, **МИН**, **СКОС**, **ЭКСЦЕСС** визначити відповідні статистичні показники, розташовуючи їх один під одним.
4. Визначити вибірковий коефіцієнт варіації за формулою (5).
5. Розповсюдити формули на дві комірки праворуч, щоб визначити відповідні показники для інших вибраних показників ґрунту.
6. Оцінити відповідність експериментальних розподілів нормальному за допомогою значень асиметрії та ексцесу (чи знаходяться відповідні значення в межах ± 0,5).
7. Скопіювати одержані результати на новий аркуш.
8. На новому аркуші визначити показники анормальності для мінімального і максимального значень вибірки за формулами (8).
9. Вилучити анормальні значення з вибірки, якщо такі виявлено.
10. Повторити визначення показників анормальності для наступних мінімального і/або максимального значень вибірки.

11. Вилучити анормальні значення з вибірки, якщо такі виявлені повторно.
12. Повторювати перевірку даних на анормальність і їх вилучення до виконання умови $V_{max,min} \leq 3$.
13. Після вилучення всіх виявлених анормальних експериментальних значень звернути увагу на те, як зміняться статистичні числові показники вибірки. Повторно оцінити відповідність експериментального розподілу нормальному розподілу за допомогою значень асиметрії та ексцесу.
14. Порівняти основні статистичні числові характеристики отримані до і після вилучення анормальних даних.
15. Відмітити, які показники при цьому змінилися.
16. Дати пояснення отриманим результатам.
17. Зберегти документ в папці під своїм прізвищем, яку помістити в папку "**Мои документы**".

Запитання для самоперевірки

1. Які величини називають випадковими?
2. Яка різниця між генеральною і вибірковою сукупностями випадкових величин?
3. У чому полягають мета і сутність статистичного аналізу дослідних даних?
4. Які основні статистичні показники застосовуються для характеристики вибіркових даних?
5. Що оцінюють за допомогою вибіркової дисперсії?
6. На припущені відповідності якому закону розподілу випадкових величин ґрунтується статистична обробка даних?
7. Які вбудовані функції табличного процесора MS Excel використовуються для визначення вибіркових статистичних показників? Як вони викликаються?
8. В чому полягають мета і сутність виявлення та вилучення анормальних значень із отриманої в процесі експерименту вибіркової сукупності даних?
9. Як оцінити близькість експериментального розподілу випадкових величин до нормального?
10. Про що свідчать негативні значення показників асиметрії і ексцесу?
11. Як провести розповсюдження статистичних функцій на діапазони комірок?
12. Які у Excel правила розрахунку за формулою?
13. За яким показником оцінюється варіація даних?

14. Який показник використовується для порівняння варіації рядів спостережень, що мають різні одиниці вимірювання?

Таблиця 1 -Показники якості чорнозему типового

№ з/п	Щільність грунту, г/см³	Запаси продуктивної вологи, 0-100 см, мм	Гідролітична кислотність, кг-екв/100 г	Сольова обмінна, pH	Кислотність актуальна, pH, водне	Вміст ввібраних основ, мг-екв/100 г	Вміст гумусу, %	Вміст легкогідролізованого азоту, мг/кг	Вміст рухомого фосфору, мг/кг	Вміст обмінного калію, мг/кг	Вміст бору, мг/кг	Вміст марганцю, мг/кг	Вміст кобальту, мг/кг	Вміст міді, мг/кг
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1,13	117,36	2,61	6,79	5,89	22,31	2,54	123,37	27,85	58,08	0,41	15,96	0,04	2,69
2	1,12	99,73	3,53	6,75	5,92	25,43	2,15	96,05	57,51	48,08	1,02	38,64	1,49	3,53
3	1,13	95,44	1,85	6,50	6,03	4,25	2,54	87,14	310,33	39,18	0,01	13,74	1,39	10,96
4	1,10	168,44	2,69	6,33	0,97	27,24	1,93	11,71	106,99	64,74	0,66	18,48	1,54	2,47
5	1,06	103,41	3,38	6,78	6,33	25,27	2,47	70,84	74,57	42,14	0,61	16,32	1,42	3,85
6	1,16	84,04	0,03	6,69	6,01	26,60	2,32	127,76	71,56	76,24	0,71	14,49	1,76	2,20
7	1,15	123,81	3,22	6,04	6,06	31,53	2,51	92,67	98,31	24,57	0,44	15,51	1,62	3,59
8	1,19	101,34	1,57	7,08	6,85	31,08	8,75	89,62	11,94	72,91	0,77	1,18	1,63	2,98
9	1,29	85,53	2,62	7,03	6,30	27,43	2,45	128,88	100,73	185,87	0,67	15,21	1,56	0,66
10	1,21	18,78	1,56	1,23	7,24	22,75	2,17	103,96	64,86	46,48	0,55	13,91	1,99	1,88
11	1,22	108,62	1,43	7,09	5,42	30,66	3,23	98,42	38,45	39,59	0,60	18,01	1,85	3,52
12	1,21	109,62	2,60	6,97	6,24	19,84	3,47	81,87	116,92	58,76	0,51	14,08	1,56	3,16
13	1,22	114,17	0,96	6,45	6,41	27,72	1,47	109,04	37,48	54,35	0,60	18,12	1,93	2,90
14	1,20	118,59	2,21	6,55	7,01	24,14	3,40	78,71	42,54	50,20	5,84	15,89	1,82	0,50
15	1,10	113,65	2,59	7,01	6,51	25,95	1,68	70,34	164,04	7,35	0,65	14,08	1,91	3,68
16	1,16	95,65	3,33	7,08	5,34	26,56	0,52	118,15	99,05	57,13	0,80	14,87	4,82	2,68
17	1,12	77,51	1,44	6,82	5,51	27,72	2,14	52,47	132,62	47,16	0,80	14,98	1,72	2,36
18	1,20	136,52	2,03	6,47	7,40	29,55	2,38	84,82	102,32	61,09	0,60	14,86	1,48	3,91
19	1,15	127,95	2,63	6,51	6,57	32,13	2,95	81,18	120,79	20,38	0,48	16,19	1,83	4,56
20	1,12	86,46	2,43	6,68	6,20	26,55	2,36	89,51	74,93	38,81	0,52	15,28	1,57	2,74
21	1,12	111,80	2,44	7,12	5,78	26,54	1,40	115,49	142,30	48,36	0,73	14,07	1,71	4,23
22	3,41	142,10	2,66	7,78	5,69	25,01	2,48	116,17	1,58	45,45	0,52	15,01	1,54	3,49

23	1,19	102,76	5,88	6,64	5,39	26,82	2,76	74,80	101,95	66,62	0,63	15,20	1,69	2,44
24	1,15	69,39	2,95	6,94	5,67	27,52	3,05	63,57	82,20	30,18	0,55	16,94	1,11	2,80
25	1,17	81,48	2,93	6,93	6,76	28,03	3,63	280,12	112,42	50,82	0,55	14,99	1,60	4,23
26	1,24	283,16	2,48	7,74	6,67	25,38	2,07	66,93	128,23	24,47	0,68	15,08	1,37	4,57
27	1,12	98,92	1,82	7,49	6,58	27,59	2,60	117,22	85,28	64,58	0,71	16,79	1,46	4,19
28	1,16	110,21	2,63	15,22	5,67	47,36	1,67	99,69	129,84	59,05	0,60	14,88	1,55	3,25
29	1,23	97,38	2,19	7,00	15,99	24,84	0,66	111,95	33,34	72,81	0,85	18,29	1,40	3,05
3O	1,21	102,55	1,70	6,34	7,00	28,20	3,56	127,42	165,28	67,70	0,56	15,58	1,71	3,04
31	0,99	119,51	3,78	6,32	6,21	27,65	2,18	96,44	142,45	62,21	0,85	14,89	1,42	3,07
32	1,03	113,26	1,44	6,65	6,03	23,59	1,24	113,01	95,81	39,52	0,61	15,66	1,31	2,70
33	1,17	116,90	1,11	6,18	5,97	29,37	1,35	140,16	96,87	52,58	0,55	13,92	1,64	5,24
34	1,12	86,46	2,57	6,81	6,40	29,83	2,66	51,36	43,84	58,98	0,61	14,35	1,38	2,97
35	1,09	87,45	1,30	6,08	6,45	20,19	2,86	84,15	53,44	51,92	0,75	13,92	1,62	2,92
36	1,17	87,48	2,01	7,74	6,16	24,67	1,86	103,63	89,04	67,82	0,67	17,64	1,59	1,87
37	1,21	101,06	2,91	4,96	6,24	25,61	3,65	90,72	81,54	66,94	0,72	17,53	1,83	3,65
38	1,18	83,39	1,90	7,12	6,41	31,13	2,92	103,26	127,76	35,69	0,77	15,55	1,88	2,27
39	1,16	53,61	1,66	7,91	5,81	31,90	3,41	116,38	31,75	49,99	0,74	15,55	1,32	3,62
40	1,11	133,89	1,69	7,25	6,13	25,94	2,84	103,97	115,31	58,63	0,73	14,32	1,43	5,35
41	1,14	124,43	3,07	7,31	5,47	27,53	1,61	85,41	84,90	41,67	0,82	15,80	1,58	3,95
42	1,12	111,79	1,58	6,02	6,30	28,09	3,20	86,03	93,12	43,76	0,56	15,67	1,58	3,65
43	1,15	89,38	1,42	5,65	7,04	29,07	1,68	135,60	112,91	36,62	0,67	16,23	1,72	4,47
44	1,22	90,92	2,74	6,88	5,75	28,41	2,48	69,94	53,53	23,21	0,57	18,09	2,21	5,68

Лабораторна робота №2
Оцінка статистичних показників генеральної сукупності, визначення довірчих похибок та інтервалів

Мета роботи: отримання практичних навичок з визначення оцінок статистичних показників генеральної сукупності, довірчих похибок та довірчих інтервалів

Теоретичні відомості

Оцінка статистичних показників генеральної сукупності

Точне визначення статистичних показників генеральної сукупності неможливе, так як кількість вимірювань при цьому $n \to \infty$. Тому використовують *оцінки* відповідних показників.

В якості *оцінки генерального середнього* приймається середнє вибіркове значення.

В якості оцінки дисперсії генеральної сукупності використовують *виправлену вибіркову дисперсію* S^2 *(або* $D_г$*),* яка

визначається за формулою:

$$S^2 = \frac{n}{n-1} D_в = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}_в\right)^2 \qquad (9)$$

де дріб $\frac{n}{n-1}$ називають поправкою Бесселя. За малих значень n поправка суттєво відрізняється від одиниці, при збільшені n вона прямує до одиниці. При $n > 50$ практично немає різниці між S^2 і $D_в$,

Для оцінки середнього квадратичного відхилення генеральної сукупності випадкових величин використовують *виправлене вибіркове середнє квадратичне відхилення S (або* $\sigma_г$*),* яке визначається за формулою:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_в)^2}{n-1}} \qquad (10)$$

Застосування вбудованих функцій MS Excel для визначення оцінок статистичних показників генеральної сукупності

Для визначення оцінок статистичних показників генеральної сукупності використовують вбудовані функції категорії **Статистические:**

ДИСП - для оцінки генеральної дисперсії;

СТАНДОТКЛОН – для оцінки генерального середнього квадратичного відхилення.

Аргументами зазначених функцій є діапазон комірок з вхідними даними.

Визначення довірчих похибок і довірчих інтервалів для статистичних характеристик генеральної сукупності випадкових величин

Визначення довірчих похибок і інтервалів ґрунтується на деяких інтервальних характеристиках випадкових величин.

Довірчим інтервалом називають інтервал, у який потрапляє істинне значення величини, що вимірюється, з заданою ймовірністю γ.

Надійністю результатів вимірювань називають імовірність γ того, що істинне значення величини, що вимірюється, потрапляє в даний довірчий інтервал. Надійність виражається в частках одиниці, або у відсотках.

Рівнем значущості називається величина $\alpha=1-\gamma$, яка характеризує ймовірність помилки, тобто частку ризику в оцінці істинного значення величини, що вимірюється. Рівень значущості – це ймовірність, якою вирішено знехтувати в заданій області досліджень.

Визначення довірчого інтервалу для математичного сподівання Правила побудови довірчого інтервалу для математичного сподівання залежать від того, відома чи не відома дисперсія генеральної сукупності $D_г$. Розглянемо випадок, коли дисперсія відома, генеральна сукупність підпорядковується нормальному закону розподілу. У цьому випадку:

■ Визначається *абсолютна довірча похибка* $\varepsilon_{\bar{x}_в}$ вибіркового середнього за формулою:

$$\varepsilon_{\bar{x}_в} = \sigma_в \frac{t_\gamma}{\sqrt{n}} \qquad (11)$$

де n – кількість спостережень;

$\sigma_в$ – вибіркове середнє квадратичне відхилення;

t_γ – коефіцієнт довіри, що береться із таблиці значень функції Лапласа Ф(t) при Ф(t) $=\gamma/2$, де γ– задана ймовірність.

Для прикладних досліджень приймається γ=0,95 (γ=95%), що відповідає рівню значущості α = 1–γ= 1–0,95 = 0,05. При цьому t_γ =1,96. t_γ ще називають *нормованим* значенням нормально розподіленої випадкової величини.

Абсолютну довірчу похибку ще називають *точністю оцінки.*

■ Визначається *відносна довірча похибка* вибіркового середнього:

$$\delta_{\bar{x}_в} = \frac{t_\gamma \sigma_в}{\bar{x}_в \sqrt{n}} = \frac{\varepsilon_{\bar{x}_в}}{\bar{x}_в} \qquad (12)$$

За потреби забезпечення результатів вимірювань заданою точністю, тобто необхідною відносною довірчою похибкою, із формули (12) визначають потрібне число спостережень *n:*

$$n = \left(\frac{t_\gamma \cdot \sigma_в}{\delta_{\bar{x}_в} \cdot \bar{x}_в} \right)^2 \qquad (13)$$

Наприклад, для визначення кількості вимірювань, яка б забезпечувала відносну довірчу похибку на рівні 5% формула (13) матиме вигляд:

$$n = \left(\frac{t_\gamma \cdot \sigma_в}{0{,}05 \cdot \bar{x}_в} \right)^2 \qquad (14)$$

Межі довірчого інтервалу математичного (генерального середнього) визначаються за формулами:

нижня межа: $\mu_{\min} = \bar{X}_в - \varepsilon_{\bar{x}_в}$ (15)

верхня межа: $\mu_{\min} = \bar{X}_в + \varepsilon_{\bar{x}_в}$ (16)

Ширина довірчого інтервалу *h* генерального середнього значення визначається за формулою:

$$h = \mu_{\min} - \mu_{\max} \qquad (17)$$

$$\bar{X}_в - \varepsilon_{\bar{x}_в} \le \mu \le \bar{X}_в + \varepsilon_{\bar{x}_в} \qquad (18)$$

При цьому, $P(\bar{X}_в - \varepsilon_{\bar{x}_в} \le \mu \le \bar{X}_в + \varepsilon_{\bar{x}_в}) = \gamma$ (19)

Формули (12) і (13) у прикладних дослідженнях займають особливе місце. За ними можна, наприклад, обчислити обсяг вибірки *n* необхідний для оцінки середнього значення нормально розподіленої вибіркової сукупності з заданою надійністю γ і точністю $\varepsilon_{\bar{x}_в}$, а також для заданої точності і відомого обсягу вибірки можна визначити надійність (імовірність).

Для визначення абсолютної довірчої похибки в Excel існує функція **ДОВЕРИТ.** Зазначена функція має три аргумента: **Альфа –** рівень значущості (для прикладних досліджень α=0,05); **Станд_откл –** оцінка середнього квадратичного відхилення (визначається за допомогою вбудованої функції **СТАНДОТКЛОН); Размер –** кількість спостережень *n* (визначається за допомогою вбудованої функції **СЧЕТ).**

Для оцінки точності вимірювань використовують також *стандартну похибку середнього* Δ, яка визначається за формулою:

$$\Delta = \frac{S}{\sqrt{n}} \qquad (20)$$

де S - оцінка генерального середнього квадратичного відхилення, n кількість вимірювань.

Визначення довірчого інтервалу для генеральної дисперсії Зазначений параметр визначається за формулою:

$$\frac{nD_в}{\chi_2^2} \le D_г \le \frac{nD_в}{\chi_1^2} \qquad (21)$$

де $D_в$ - вибіркова дисперсія, n - кількість вимірювань, χ_1^2, χ_2^2 критерії Пірсона.

Критерії Пірсона визначаються із таких положень:

$$P\left(\chi^2 > \chi_1^2\right) = 1 - \frac{\alpha}{2} = 1 - \frac{0{,}05}{2} = 0{,}975 \qquad (22)$$

$$P\left(\chi^2 > \chi_2^2\right) = \frac{\alpha}{2} = \frac{0{,}05}{2} = 0{,}025 \qquad (23)$$

За таблицею χ^2-розподілу, або іншими засобами для числа степенів свободи $f=n-1$ та одержаних імовірностей 0,975 і 0,025 знаходять значення χ_1^2, χ_2^2.

Для обчислення значення критеріюχ^2 в MS Excel є вбудована функція **ХИ20БР** аргументами якої є задана ймовірність і число степенів свободи $f=n-1$.

Визначення довірчого інтервалу для генерального середнього квадратичного відхилення

Зазначений параметр при $n \le 30$ визначається за формулою:

$$\frac{\sqrt{n} \cdot \sigma_в}{\chi_2} \le \sigma_г \le \frac{\sqrt{n} \cdot \sigma_в}{\chi_1} \qquad (24)$$

За великих обсягів вибірки $n>30$ довірчий інтервал для генерального середнього квадратичного відхилення визначається за формулою:

$$\frac{\sqrt{2n} \cdot \sigma_в}{\sqrt{2n-3} + t_\gamma} \le \sigma_г \le \frac{\sqrt{2n} \cdot \sigma_в}{\sqrt{2n-3} - t_\gamma} \qquad (25)$$

де t_γ - нормоване значення нормально розподіленої випадкової величини, яке відповідає заданій надійності γ і визначається за таблицею функції Лапласа Ф(t).

Визначення статистичних показників за допомогою засобу "Описательная статистика"

До методів описової статистики відносять методи опису вибірки за допомогою різних числових показників.

В MS Excel існує засіб **Описательная статистика,** який дозволяє одночасно визначити ряд статистичних показників. Щоб скористатися засобом **Описательная статистика,** виконують дії:

- вибирають меню **Сервис=>Анализ данных=>Описательная статистика;**
- у відповідному діалоговому вікні в поле **Входные данные** вводять діапазон комірок з експериментальними даними;
- у полі **Метки в первой строке** ставлять галочку, якщо перший рядок вхідного діапазону містить заголовок стовпчика, або нічого не ставлять, якщо заголовка немає;
- у групі показників **Параметры вывода** відмічають галочкою пункт **Итоговая статистика**;
- рівень надійності за замовчуванням дорівнює 0,05 – за необхідності зміни рівня надійності, активізують перемикач **Уровень надежности** і вводять потрібне значення;
- натискують кнопку **ОК**;
- в результаті виконаних дій з'явиться таблиця з такими статистичними показниками експериментальних даних:

Назва показника	Зміст показника
Среднее	середнє значення
Стандартная ошибка	стандартна похибка середнього , яка визначається за формулою (20)
Медиана	значення, яке ділить вибірку на дві, рівні за числом вимірювань, частини
Мода	значення, яке має найбільшу частоту появи
Стандартное отклонение	оцінка генерального середнього квадратичного відхилення
Дисперсия выборки	оцінка генеральної дисперсії
Эксцесс	ексцес
Асимметричность	асиметрія
Интервал	інтервал варіювання (розмах вибірки – різниця між максимальним і мінімальним
Минимум	мінімальне значення вибірки
Максимум	максимальне значення вибірки
Сумма	сума усіх значень вибірки
Счет	кількість вимірювань

Програма виконання роботи

1. Скопіювати в новий документ експериментальні дані і результати обчислень із Л_р_№1 – один із показників (після вилучення анормальних значень).

2. За допомогою вбудованих функцій визначити оцінки генеральної

дисперсії і генерального середнього квадратичного відхилення.

3. За формулою (11) визначити абсолютну довірчу похибку середнього.

4. За формулою (12) визначити відносну довірчу похибку середнього.

5. За формулою (20) визначити стандартну похибку середнього.

6. За формулою (13) обчислити необхідну кількість вимірювань для забезпечення відносної похибки середнього значення на рівні 7%.

7. За формулами (15) і (16) обчислити верхню і нижню межі довірчого інтервалу для математичного сподівання.

8. Перевірити правильність визначення абсолютної похибки середнього за допомогою вбудованої функції **ДОВЕРИТ.**

9. За формулою (21) обчислити довірчий інтервал для генеральної дисперсії.

10. За формулами (24) або (25) обчислити довірчий інтервал для генерального середнього квадратичного відхилення.

11. Використовуючи власні отримані дані та вбудовані функції категорії **Текстовые СЦЕПИТЬ** (або оператор &) та **ФИКСИРОВАННЫЙ**, записати межі довірчих інтервалів для математичного сподівання μ, генеральної дисперсії $D_г$ і середнього квадратичного відхилення $\sigma_г$ у такому вигляді:

Межі довірчого інтервалу для $D_г$	$4{,}8 \le D_г \le 7{,}2$

12. Провести перевірку визначення деяких статистичних показників за допомогою засобу **Описательная статистика.**

13. Порівняти результати, одержані шляхом розрахунків, з результатами, одержаними шляхом використання засобу **Описательная статистика.**

Запитання для самоперевірки

1. Що називається поправкою Бесселя?
2. Як співвідносяться вибіркова дисперсія і оцінка дисперсії генеральної сукупності.
3. Від чого залежить величина абсолютної похибки середнього? Що можна зробити, щоб похибка була меншою?
4. Яка вбудована функція застосовується для визначення абсолютної похибки середнього значення?
5. Що називають довірчим інтервалом?
6. Що називають надійністю результатів вимірювань?
7. Що називають рівнем значущості?
8. Які показники обчислюються за допомогою засобу **Описательная статистика?**

9. У яких випадках використовуються вбудовані статистичні функції, а у яких - **Описательная статистика?**

Лабораторна робота № 3
Дослідження експериментальних розподілів

Мета роботи: одержання практичних навичок дослідження експериментальних розподілів шляхом групування та графічного аналізу експериментальних даних, побудови розподілів частот і частостей, накопичених частот і частостей, полігонів, гістограм

Теоретичні відомості

Емпіричним (експериментальним) називають розподіл відносних частот. Для його дослідження використовують апарат математичної статистики.

Теоретичним називають розподіл імовірностей. Для його вивчення застосовують теорію ймовірностей.

Для дослідження експериментального розподілу результати експерименту представляють у вигляді послідовності чисел $x_1,x_2,...x_k$.

Якщо експериментальне значення x_1 спостерігалось n_1 раз, значення x_2 спостерігалось n_2 раз і т. д., то значення x_i називаються *варіантами,* а числа їх спостережень n_i – *частотами.* Процедура підрахунку частот називається *групуванням даних.*

Обсяг вибірки n дорівнює сумі всіх частот n_i.

$$n = \sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k \qquad (26)$$

Відносною частотою (частістю) значення x_i називається відношення частоти спостережень цього значення n_i до загального обсягу вибірки n.

$$w_i(n) = \frac{n_i}{n} \qquad (27)$$

Статистичним розподілом частот (або просто розподілом частот) називається перелік варіант і відповідних їм частот, записаний у вигляді таблиці

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	$n_к$

Розподілом відносних частот називається перелік варіант і відповідних їм відносних частот.

Полігоном частот називають ламану, відрізки якої сполучають точки $(x_1;n_1),(x_2;n_2),...(x_k;n_k)$. Для побудови полігону на осі абсцис відкладають варіанти x_i, а на осі ординат відповідні їм частоти n_i.

Точки ($x_i;n_i$) сполучають відрізками прямих і отримують полігон частот.

Полігоном відносних частот (частостей) називають ламану, відрізки якої сполучають точки $(x_1;w_1),(x_2;w_2)\ldots (x_к;\ w_k)$. Для побудови полігону відносних частот на осі абсцис відкладають варіанти x_i, а на осі ординат – відповідні їм відносні частоти w_i.. Точки (x_i; w_i) сполучають відрізками прямих і отримують полігон відносних частот.

Статистичний розподіл вибірки за частотами може бути графічно зображеним за допомогою гістограми. Для її побудови всі експериментальні значення розбиваються на декілька інтервалів $[x_i,\ x_{i+1})$, які називаються *класовими або частковими інтервалами,* або *кишенями.* Довжина λ *i*-го класового інтервалу дорівнює: $\lambda = x_{i+1}-x_i$. Якщо обсяг вибірки великий, то можна вибрати *к* класових інтервалів однакової довжини $\lambda = (x_{max}-x_{min})/k$, де $x_{\max}$ і $x_{\min}$- найбільша і найменша варіанти відповідно. Кількість класів *(k)* визначають за формулою

$$k=l+3{,}32\cdot\log(n), \qquad (28)$$

де *n* - кількість спостережень (визначається за допомогою вбудованої функції **СЧЕТ);** *log(n)* – визначається за допомогою відповідної вбудованої функції Excel категорії **Математические.**

Гістограмою частот називають фігуру, яка складається з прямокутників, основою яких є класові інтервали довжиною λ, а висоти дорівнюють n_i. Для побудови гістограми частот на осі абсцис відкладають класові інтервали, а над ними проводять відрізки, паралельні осі абсцис на відстані n_i.

Гістограмою відносних частот називають фігуру, яка складається із прямокутників, основами яких служать класові інтервали довжиною λ, а висоти рівні w_i. Для побудови гістограми відносних частот на осі абсцис відкладають класові інтервали, а над ними проводять відрізки, паралельні осі абсцис, на відстані w_t.

Побудова розподілу частот

Проводять сортування вхідних даних за збільшенням. Для цього виокремлюють діапазон комірок з вхідними даними, виконують дії:

- **Данные=> Сортировка;**
- у діалоговому вікні **Сортировка данных** вибирають пункт **По возрастанию;**
- натискують кнопку ОК.

Отриманий ряд дозволяє оцінити максимальне та мінімальне значення варіант і різницю між ними. Ця інформація використовується для підрахунку потрібного числа класів при побудові згрупованого розподілу частот.

У вільній комірці визначають кількість класів *к,* на які треба розподілити дослідні дані за формулою (28). Одержане значення *к,* округлюють до цілого. Визначають величину класового інтервалу за формулою

$$\lambda = \frac{x_{\max} - x_{\min}}{k} \qquad (29)$$

За необхідності округлюють одержане значення до цілого. *Для одержання масиву класових інтервалів* можна задавати їх вручну, або користуватись засобами автозаповнення. В останньому випадку:

■ вводять у вільну комірку мінімальне число вибірки, натискують клавішу **Enter**;

■ активізують комірку з мінімальним значенням вибірки, виконують команду **Правка=>Заполнить=>Прогрессия;**

■ у діалоговому вікні **Прогрессия** у полі **Шаг** установлюють довжину класового інтервалу, в полі **Тип** вказують **Арифметическая,** в полі **Предельное зачение** - максимальне значення вибірки;

■ у полі **Расположение** відмічають **По столбцам;**

■ натискують кнопку ОК.

У результаті виконаних дій буде виведено масив класових інтервалів (кишень). Якщо кінцеве значення масиву класових інтервалів виявилося меншим за максимальне значення вибірки, тоді нижче добавляють ще один класовий інтервал вручну.

Для побудови масиву частот виконують дії:

■ праворуч від діапазону комірок з масивом класових інтервалів виокремлюють комірки стовпчика для виведення частот, причому на одну більше, ніж займає масив класових інтервалів;

■ виконують дії **Вставка=> Функция=> Статистические => Частота;**

■ у діалоговому вікні функції **Частота** у відповідні поля вводять адреси комірок, що містять масив вхідних (експериментальних) даних і масив класових інтервалів;

■ натискують клавішу F2, а потім одночасно клавіші **<Ctrl>+<Shift>+<Enter>;**

■ в результаті у вибраний діапазон комірок буде виведено масив частот;

■ якщо замість масиву частот виведеться одне число, треба впевнитися, що діапазон комірок для виведення частот виокремлено вірно, а потім ще раз натиснути клавішу F2 і одночасно клавіші **<Ctrl>+<Shift>+<Enter>;**

■ правильність підрахунків перевіряють шляхом обчислення суми отриманих частот, яка має дорівнювати кількості експериментальних даних.

Вбудована функція MS Excel **Частота** дозволяє одержати розподіл частот по класовим інтервалам, причому, якщо розподіл класових інтервалів розпочинається зі значення x_i, то до нього входять частоти появи значень, які $\leq x_i$. Якщо наступне значення класового інтервалу x_{i+1}, то до нього входять частоти появи значень, які $x_i <$ і $< x_{i+1}$.

Побудова розподілу відносних частот

Як відомо, відносні частоти (частості) – це частоти, поділені на загальне число спостережень (число експериментальних даних). Стовпчик з відносними частотами будують поруч зі стовпчиком з частотами. Для побудови розподілу відносних частот виконують такі дії:

■ у вільній комірці, якщо це не зроблено раніше, підраховують кількість експериментальних даних (суму частот) за допомогою вбудованої функції **СЧЕТ;**

■ у першу комірку стовпчика для побудови розподілу відносних частот вводять формулу ділення значення першої комірки з частотою на кількість спостережень, використовуючи відповідні посилання на адреси комірок;

■ на адресу комірки з першим значенням частоти робиться відносне посилання, на адресу комірки з сумою частот – абсолютне;

■ натискують клавішу **Enter;**

■ активізують комірку з першим отриманим значенням відносної частоти й розповсюджують формулу на діапазон комірок, призначений для побудови розподілу відносних частот.

■ для перевірки правильності розрахунків визначають суму відносних частот, яка має дорівнювати 1.

Побудова розподілів накопичених частот і частостей

Для заповнення діапазону комірок накопиченими частотами виконують дії:

■ у вибраній комірці після знаку "=" (дорівнює) записують відносне посилання на адресу першої комірки зі значенням частоти - це буде перше значення розподілу накопичених частот;

■ у наступну комірку після знаку "=" вводять адресу другої комірки зі значенням частоти, знак "+" і адресу попередньої комірки зі значенням першої накопиченої частоти;

■ натискують клавішу **Enter;**

■ активізують комірку, в якій з'явилося друге значення накопиченої частоти;

■ методом автозаповнення заповнюють діапазон стовпчика значеннями накопиченої частоти (протягують маркер автозаповнення);

■ останнє значення діапазону має відповідати сумі частот.

Для розрахунку накопичених частостей проводять описані вище дії, використовуючи діапазон комірок не з частотами, а з відносними частотами (частостями). Останнє число діапазону відносних накопичених частот має дорівнювати одиниці.

Побудова полігону, гістограми і кумуляти

Побудова полігонів:

■ виокремлюють стовпчики, які містять класові інтервали (кишені) і частоти;

■ виконують дії: **Вставка=>Диаграмма;**

■ у діалоговому вікні **Мастера диаграмм (шаг 1 из 4): тип диаграммы** на закладці **Стандартные** в групі **Тип** вибирають **Точечная,** в групі **Вид** вибирають **Точечная диаграмма, на которой значення соединены отрезками;**

■ для того, щоб переглянути, як буде виглядати полігон на даному етапі, натискують кнопку **Просмотр результата** в цьому ж діалоговому вікні - в результаті на місці групи **Вид** з'явиться поле **Образец,** в якому буде показано полігон;

■ натискують кнопку **Далее;**

■ у наступному вікні **Мастера диаграмм (шаг 2 из 4): источник данных** нічого не змінюють, треба лише впевнитися, що відмічено **Ряды** в **столбиках** і натиснути кнопку **Далее;**

■ у наступному вікні **Мастера диаграмм (шаг 3 из 4): параметры диаграммы** на закладці **Заголовки** вводять у полі **Название диаграммы** заголовок "Полігон частот"; у полі **Ось X (категорий)** - назву осі X: "Класові інтервали"; у полі **Ось Y (значенний)** -назву осі Y: "Частоти";

■ на закладці **Линии сетки** знімають галочку з перемикача **Ось Y (значений): основные линии;**

■ на закладке **Легенды** знімають галочку з перемикача **Добавить легенду** і натискують кнопку **Готово;**

■ одержану діаграму редагують далі: за допомогою миші рисунку надають квадратної форми;

■ для того, щоб прибрати сірий фон діаграми, натискують двічі мишею в сірій області - в результаті з'явиться вікно **Формат области построения,** в якому в групі **Заливка** відмічають **Прозрачная** і натискують кнопку ОК;

■ у діалоговому вікні **Формат оси,** яке викликається натискуванням

правою кнопкою миші на осі, на закладці **Шкала** за потреби змінюють початкове значення вісі і ціну основних поділок.

Аналогічним чином будується полігон частостей. Необхідно впевнитися, що правильно задані інтервали комірок з класовими інтервалами і частостями. Після дій з форматування діаграми необхідно звернути увагу на те, що числа по вісі *Y* можуть мати різну кількість знаків після коми. Щоб кількість знаків після коми була однаковою виконують дії:

- двічі натискують мишею на даній осі;
- в діалоговому вікні **Формат оси** вибирають закладку **Число;**
- в групі **Числовые форматы** вибирають **Числовой** і встановлюють **Число десятичных знаков: 2** (це число задане за замовчуванням);
- натискують кнопку **ОК.**

Побудова гістограми

В **Пакете анализа** меню **Сервис** є інструмент для швидкої побудови гістограми, який так і називається **Гистограмма.** Для побудови гістограми:

- викликають діалогове вікно **Гистограмма;**
- задають діапазони комірок з вхідними даними і класовими інтервалами (кишенями);
- відмічають галочкою перемикач **Вывод графика;**
- натискують кнопку ОК.

Інструмент **Гистограмма** виводить два стовпчики: **Карманы і Частота.** У стовпчику **Карманы** дублюються задані раніше класові інтервали. У стовпчику **Частота** повторно виводяться обчислені для кожного класового інтервалу частоти. Сама гістограма виводиться справа від стовпчика частот. Форматування гістограми здійснюється звичними для форматування графіків способами. Для заміщення зазорів між прямокутниками на гістограмі викликають меню **Формат рядов данных.** Для цього:

- встановлюють курсор на поле одного із прямокутників і натискують праву клавішу миші;
- у контекстному меню, що відкриється, вибирають **Формат рядов данных=>Параметры;**
- у вікні **Ширина зазора** встановлюють 0;
- натискують кнопку ОК.

Гістограму можна побудувати також за допомогою **Мастера диаграмм,** вибравши відповідний вид графіка.

Побудова кумулятивної кривої

Кумулятивна крива (крива накопичених частот, або крива накопичених частостей) будується таким чином: по осі абсцис відкладають класові

інтервали, а по осі ординат – накопичені частоти, або накопичені частості. Другим варіантом побудови кумулятивної кривої є використання інструменту **Сервис => Пакет анализа => Гистограмма:**

- у діалоговому вікні **Гистограмма** вказують **Входной интервал** - діапазон комірок з вхідними даними, **Интервал карманов** -діапазон комірок з класовими інтервалами (кишенями);
- відмічають галочкою **Интегральный % і Вывод графика;**
- натискують кнопку ОК - в результаті з'явиться таблиця з даними і об'єкт з двома графіками;
- в області графіка натискують праву клавішу миші;
- у контекстному меню, що відкриється, вибирають пункт **Входные данные;**
- у діалоговому вікні, що відкриється, переходять на закладку **Ряд;**
- в однойменному полі **Ряд** вилучають ряд, що має назву **Частота;**
- натискують кнопку ОК;
- натискують праву клавішу миші в області графіка;
- вибирають у контекстному меню пункт **Параметры диаграммы;**
- вводять необхідні виправлення в назви заголовків;
- натискують праву клавішу миші на осі *У*;
- у контекстному меню вибирають **Формат оси;**
- переходять на закладку **Число,** відмічають формат **Числовой;**
- натискують кнопку ОК.

Визначення частоти попадання випадкової величини в заданий класовий інтервал

Розподіл частот використовують для визначення частоти попадання випадкової величини в заданий класовий інтервал. Наприклад, результати дослідження вмісту сухої речовини у цибулі наведені у табл.3

Таблиця 3 - Частотний розподіл вмісту сухої речовини в цибулі, %

Класові інтервали	Частоти	Частості	Накопичені частоти	Накопичені частості
15	1	0,0345	1	0,0345
16	4	0,1379	5	0,1724
17	8	0,2759	13	0,4483
18	10	0,3448	23	0,7931
19	4	0,1379	27	0,9310
20	2	0,0690	29	1,0000

Згідно з представленими даними, для дослідження вмісту сухої речовини в цибулі було проведено 29 вимірювань. Кількість випадків, які

потрапляють в той чи інший класовий інтервал наведено у другій графі (розподіл частот). Наприклад, вимагається визначити, скільки вимірювань вмісту сухої речовини прийме значення >16 і ≤18 %. Підраховуючи частоти, встановлюють, що із 29 випадків вимогам задовольняють 18 значень. Наприклад, вимагається визначити, який відсоток вимірювань буде мати значення вмісту сухої речовини ≤17%. Використовуючи розподіл відносних накопичених частот (накопичених частостей), встановлюють, що вимогам відповідає 0,4483 (44,83%) вимірювань.

Розподіл відносних частот використовують також для порівняння двох рядів спостережень, що мають різну кількість вимірювань.

Програма виконання роботи

1. Відкрити документ MS Excel з даними Л_р_ №4.
2. Скопіювати на новий аркуш дані одного із показників Л_р_№ 4, розподіл якого після вилучення анормальних даних найбільш близький до нормального.
3. Перейменувати аркуш з даними, присвоївши йому ім'я Л_р_6.
4. Провести сортування вхідних даних за збільшенням.
6. Визначити кількість класів *k,* на які треба розподілити дослідні дані, *k* округлити до цілого.
7. Визначити величину класового інтервалу λ. За необхідності округлити отримане значення до цілого.
8. Побудувати розподіл частот .
9. Побудувати розподіл відносних частот (частостей).
10. Побудувати розподіли накопичених частот і частостей .
10. Побудувати полігони частот і відносних частот, гістограму і кумулятивну криву (див. зразки рис.).
11. Визначити, який відсоток вимірювань ≤ середнього значення класових інтервалів ?

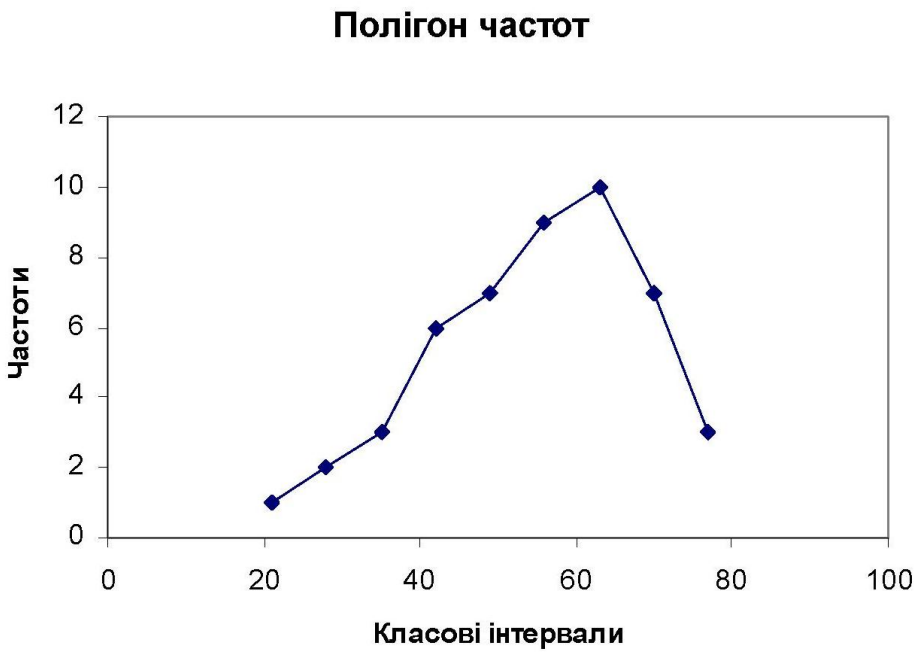
Запитання для самоперевірки

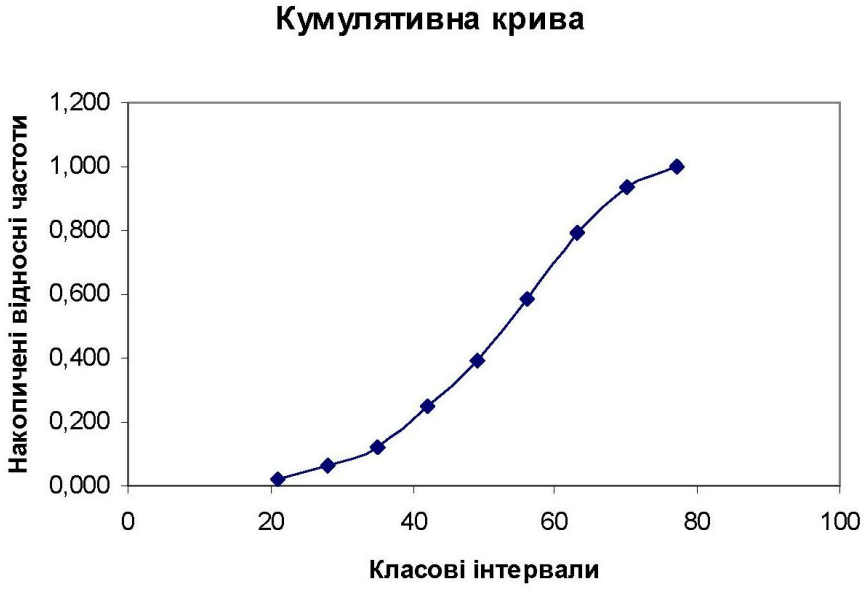
1. Для чого застосовуються розподіли частот, відносних частот (частостей), накопичених частот і накопичених частостей?
2. Як побудувати за допомогою засобів MS Excel розподіл частот?
3. Які особливості побудови розподілів накопичених частот?
4. Які особливості побудови полігонів частот і частостей?
5. За допомогою яких засобів MS Excel будують гістограми?
6. Як визначити частоту попадання випадкової величини в заданий класовий інтервал?
7. Як визначити кількість спостережень, яка не перевищує задану

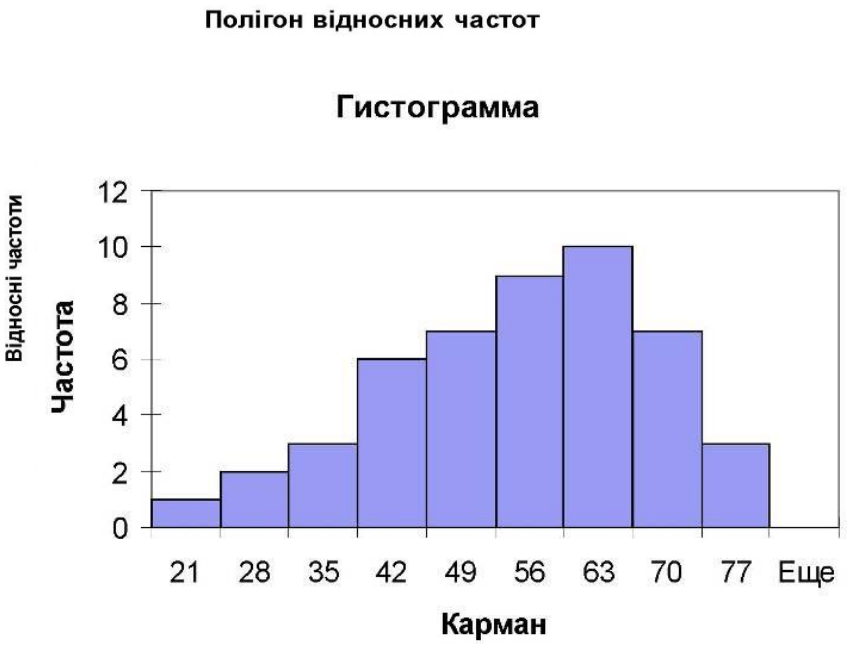
межу?

8. Як порівняти два ряди експериментальних даних, що мають різну кількість вимірювань?

Зразки рисунків побудованих графіків







Лабораторна робота № 4
Графічне порівняння експериментального розподілу з теоретичним

Мета роботи: одержання практичних навичок графічного порівняння експериментального розподілу з теоретичним

Теоретичні відомості

Диференціальна і інтегральна функції нормального розподілу

Теоретичним розподілом називають розподіл ймовірностей появи того чи іншого значення.

Як уже відзначалося, *нормальним законом розподілу імовірностей* (або просто нормальним розподілом) називається закон розподілу неперервної випадкової величини, заданий щільністю розподілу у вигляді формули (1). Наведена функція називається також *диференціальною функцією нормального розподілу.*

Інтегральна функція нормального розподілу визначається за формулою:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}} dx \qquad (1)$$

Засоби MS Excel для визначення диференціальної і інтегральної функцій нормального розподілу

В Excel серед статистичних функцій є функція **НОРМРАСП**, яка дозволяє обчислити як диференціальну функцію нормального розподілу *f(x)*,

так і інтегральну функцію розподілу *F(x).* Ця функція має наступні параметри: **НОРМРАСП**(*x*;*α*;*σ*;***тип***), де *x* - значення змінної, для якої необхідно обчислити функцію; *α*- середнє значення нормального розподілу; *σ* - оцінка стандартного відхилення цього розподілу; ***тип*** - це логічне значення, що визначає тип функції розподілу: ***тип*** - приймає значення ***ИСТИНА*** або ***ЛОЖЬ***. Якщо вказати в якості цього параметру ***ИСТИНА,*** то буде обчислена інтегральна функція *F(x)* нормального розподілу, а якщо - ***ЛОЖЬ,*** то буде обчислена диференціальна функція розподілу *f(x).*

Експериментальний розподіл випадкових величин за достатньо великої кількості спостережень наближається до нормального закону (закону Гауса). Перевірка емпіричного (експериментального) закону розподілу на "нормальність" необхідна для підтвердження коректності виконання статистичного аналізу експериментальних даних, оцінки його достовірності й надійності, вибору статистичних критеріїв щодо порівняння середніх значень і дисперсій, підтвердження можливості застосування експериментальних даних для моделювання. Окрім цього, невідповідність експериментального розподілу теоретичному нормальному може бути викликана наявністю у вибірці анормальних значень, неправильним вибором факторів, що впливають на параметр оптимізації, або їх рівнів варіювання, впливу неконтрольованих, або некерованих факторів, що призводить до появи не усунених залишків систематичних, методичних, або інструментальних похибок. Цей факт вимагає додаткового аналізу умов проведення експерименту і одержаних результатів.

Графічний статистичний аналіз експериментальних даних дозволяє візуально оцінити вигляд експериментального розподілу за формою полігону частот, або частостей, а також візуально порівняти графіки теоретичних і експериментальних диференціальних та інтегральних функцій розподілу.

Остаточний висновок щодо закону розподілу вибіркової сукупності можна зробити тільки шляхом оцінки його відповідності теоретичному за допомогою спеціальних критеріїв згоди.

Графічне порівняння емпіричного розподілу з теоретичним

Для графічної ілюстрації одержаних результатів будують графіки теоретичної та експериментальної інтегральної й диференціальної функцій нормального розподілу.

Для побудови інтегральних функцій розподілу:

- копіюють відсортовані за збільшенням результати вимірювань одного із показників (Л_р_№6) на новий аркуш, масиви класових інтервалів, відносних частот і відносних накопичених частот;
- у вільних комірках визначають, якщо не було встановлено раніше,

середнє вибіркове значення за допомогою вбудованої функції **СРЗНАЧ** і виправлене середнє квадратичне відхилення - за допомогою **СТАНДОТКЛОН;**

- у першу комірку вільного стовпчика, яка знаходиться в одному рядку з першим значенням скопійованих експериментальних даних, вводять формулу для підрахунку інтегральної функції нормального розподілу експериментальних величин: **=НОРМРАСП** (відносне посилання на адресу комірки з першим значенням вхідних даних; абсолютне посилання на адресу комірки зі значенням середнього; абсолютне посилання на адресу комірки зі значенням виправленого середнього квадратичного відхилення; **ИСТИНА**);
- копіюють формулу так, щоб були задіяні всі вхідні дані, тобто, до рядка, що містить останнє значення вхідних даних;
- виокремлюють діапазон комірок з вхідними даними та діапазон комірок зі значеннями інтегральної функції нормального розподілу, використовуючи клавішу **Ctrl**, якщо вони не сумісні;
- викликають **Мастер диаграмм** і будують графік, як це було описано в Л_р_ № 6;
- проводять відповідні редагування і форматування графіка, щоб він мав вигляд, поданий на рис.;
- для побудови експериментальної інтегральної функції розподілу натискують праву клавішу миші на вільному місці графіка теоретичної інтегральної функції;
- у контекстному меню, що відкриється, вибирають пункт **Исходные данные;**
- переходять на закладку **Ряд** і натискують кнопку **Добавить;**
- в полі **Название** вводять "експериментальна";
- в поле **Значення оси Х** вводять діапазон комірок зі значеннями класових інтервалів;
- в поле **Значення оси У** вводять діапазон комірок зі значеннями накопичених частостей;
- проводять редагування і форматування графіків відомими способами;

або:

- використовують побудований в Л_р_6 графік кумулятивної кривої (накопичених відносних частот) і до нього добавляють графік теоретичної інтегральної функції нормального розподілу.

Для побудови диференціальних функцій нормального розподілу виконують дії:

- у першу комірку вільного стовпчика вводять формулу для підрахунку диференціальної функції нормального розподілу експериментальних

величин, наприклад: = **НОРМРАСП** (адреса комірки з першим значенням експериментальних даних; адреса комірки зі значенням середнього; адреса комірки зі значенням виправленого середнього квадратичного відхилення; **ЛОЖЬ**), використовуючи абсолютні і відносні посилання на адреси комірок;

- натискують клавішу **Enter**;
- активізують комірку з формулою і за допомогою маркера авто заповнення розповсюджують формулу так, щоб вказана функція була визначена для всіх експериментальних даних;
- для побудови графіка теоретичної диференціальної функції нормального розподілу виокремлюють діапазон комірок з вхідними даними та діапазон комірок зі значеннями диференціальної функції;
- викликають **Мастер диаграмм** і аналогічним чином будують теоретичну диференціальну криву нормального розподілу;
- натискують праву клавішу миші на вільному місці діаграми;
- у контекстному меню, що відкриється, вибирають пункт **Исходные данные;**
- у діалоговому вікні, що відкриється, переходять на закладку **Ряд** і натискують кнопку **Добавить;**
- у полі **Название** вводять "експериментальна";
- у полі **Значення оси Х** вводять діапазон комірок зі значеннями класових інтервалів;
- в поле **Значення оси У** вводять діапазон комірок зі значеннями частостей;
- проводять редагування і форматування графіків;
- приклади побудованих графіків подано на рисунках.

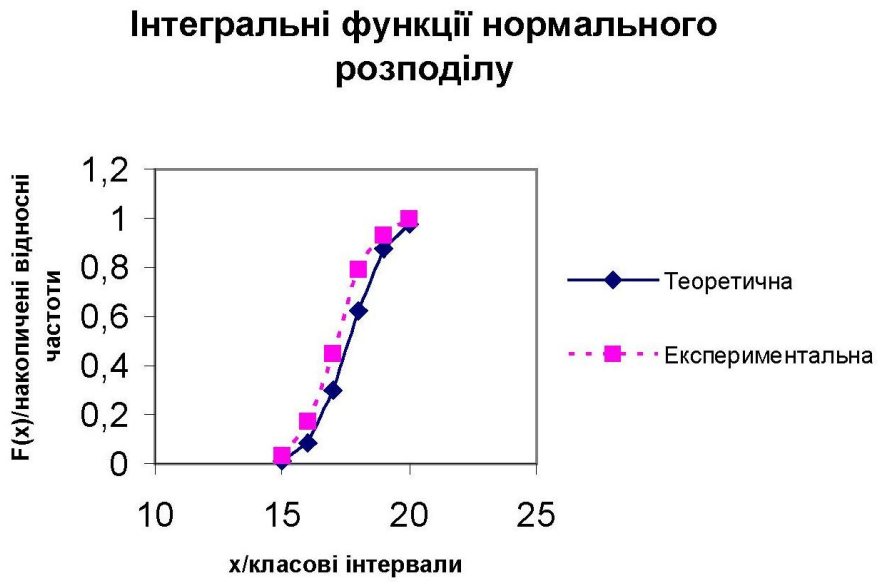
Програма виконання роботи

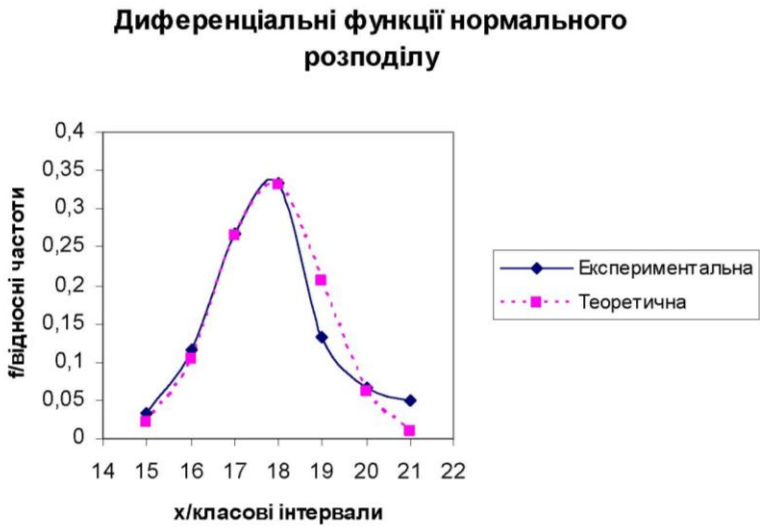
1. Відкрити документ, що містить Л_р_№3.
2. Скопіювати експериментальні дані, масив класових інтервалів, масиви частостей і накопичених частостей на новий аркуш, або в новий документ.
3. Побудувати теоретичні й експериментальні диференціальні та інтегральні функції нормального розподілу.
4. Провести редагування і форматування графіків згідно з рисунками.

Запитання для самоперевірки

1. Що називається інтегральною і диференціальною функціями нормального розподілу випадкових величин?

2. Яка вбудована функція Excel слугує для побудови теоретичних нормальних розподілів?
3. Як обчислити значення і побудувати графік інтегральної теоретичної функції нормального розподілу ?
4. Який експериментальний розподіл відповідає інтегральній функції?
5. Як обчислити значення і побудувати графік диференціальної теоретичної функції нормального розподілу засобами MS Excel?
6. Який експериментальний розподіл відповідає диференціальній функції?
7. Що є аргументами функції **НОРМРАСП**?
8. Про що свідчить подібність графіків теоретичного і експериментального розподілів випадкових величин?





Лабораторна робота № 5
Порівняння параметрів нормального розподілу

Мета роботи - отримання практичних навичок у використанні статистичних критеріїв для перевірки статистичних гіпотез, зокрема, критерію Фішера для перевірки однорідності дисперсій і критерію Ст'юдента для перевірки рівності середніх значень.

Теоретичні відомості

Поняття про статистичні гіпотези і критерії

Задачі порівняльного аналізу застосовуються у багатьох сферах наукової і професійної діяльності. Наприклад, при оцінці якості нової технології - показники, одержані при застосуванні нової технології, порівнюються з показниками, одержаними за традиційної технології. При оцінці ефективності очисних заходів - показники, одержані в результаті очисних заходів, порівнюються з нормативними тощо.

До задач порівняльного аналізу відносять: порівняння емпіричних законів розподілу, або емпіричного з теоретичним, а також параметрів цих розподілів.

При цьому висувається гіпотеза щодо результату, що передбачається. Наприклад, що показник нової технології перевищує показник класичної, або що в результаті проведення очисних заходів вміст шкідливих речовин не перевищує граничнодопустимого.

Статистичною гіпотезою називається припущення відносно виду невідомого розподілу або відносно значень параметрів емпіричного розподілу.

Нульовою (основною) гіпотезою називається висунута гіпотеза, яка позначається H_0. *Альтернативною (конкуруючою)* називається гіпотеза, яка суперечить основній (позначається H_1).

Якщо проводиться порівняння однієї вибірки, генеральний параметр якої z_1, з іншою вибіркою, генеральний параметр якої z_2, то основна гіпотеза формулюється так: генеральні параметри вибірок, що порівнюються, рівні, тобто різниця між вибірковими параметрами носить випадковий характер. У цьому разі основну гіпотезу записують у вигляді: $H_0: z_1 = z_2$.

Альтернативні гіпотези при цьому можуть мати один із наступних видів:

а) $H_1: z_1 > z_2$; *б)* $H_1: z_1 < z_2$; *в)* $H_1: z_1 \neq z_2$.

Гіпотези *"а)"* і *"б)"* називаються направленими, а гіпотеза *"в)"* - ненаправленою.

Перевірка гіпотези дозволяє зробити висновок щодо прийняття або протиріччя висунутої гіпотези емпіричним даним.

Статистичним критерієм називається спеціально напрацьована випадкова величина з відомою функцією розподілу, яка служить для перевірки основної гіпотези. Значення критерію, обчисленого за вибіркою, називається *"значенням критерію, що спостерігається"* (застосовується також назва *"розрахункове значення критерію").* Значення критерію, визначене за спеціальними формулами для вибраного рівня надійності і розрахованого числа степенів свободи, називається *критичним значенням критерію.*

Критичною областю називається множина значень критерію, при яких відхиляється основна гіпотеза, тобто множина критичних значень критерію.

Порівняння двох дисперсій

Порівняння дисперсій використовують для оцінки степені розсіювання двох випадкових величин, оцінки точності визначення певних показників тощо. Наприклад, при порівнянні двох технологій ремедиації нафтозабруднених середовищ кращою буде та, яка забезпечує, окрім мінімальних середніх значень показників забруднення, мінімальні відхилення окремих вимірювань цих показників від своїх середніх значень. Середнє квадратичне відхилення (корінь квадратний з дисперсії) використовується для оцінки точності визначення середніх значень показників, що є необхідною умовою при визначенні значущості різниці між ними. Порівняння дисперсій передує порівнянню середніх, оскільки від його результату залежить вибір інструменту аналізу для порівняння середніх значень.

Критерій Фішера (F- тест) для порівняння двох дисперсій

Для оцінки значущості різниці виправлених вибіркових дисперсій (оцінок дисперсій генеральних сукупностей) S^2_x та S^2_y (нехай $S^2_x > S^2_y$), розрахованих за двома вибірками із генеральних сукупностей X і Y, що мають розподіл, близький до нормального, використовується критерій Фішера *(F-*критерій).

Вимагається перевірити нульову гіпотезу H_0: $z_1 = z_2$.

Значення критерію Фішера, що спостерігається (розрахункове), обчислюють за формулою

$$F_{\text{розр}} = \frac{S_x^2}{S_y^2}. \qquad (31)$$

У чисельнику завжди має бути більша дисперсія.!!! Розрахункове значення критерію порівнюють з критичним, обчисленим з прийнятим рівнем значущості. Якщо розрахункове значення критерію перевищує

критичне, то нульова гіпотеза про рівність двох виправлених вибіркових дисперсій відхиляється. Тобто різниця між ними не є випадковою, отже дисперсії двох генеральних сукупностей неоднорідні. Якщо розрахункове значення критерію Фішера менше критичного, то нульова гіпотеза приймається, тобто дисперсії однорідні.

Порівняння дисперсій двох вибірок засобами MS Excel

Для порівняння дисперсій в MS Excel використовується засіб під назвою **Двухвыборочный F-тест для дисперсии.** Для його використання виконують таку послідовність дій: **Сервис=>Анализ данных=> Двухвыборочный F-тест для дисперсии.**

В діалоговому вікні **Двухвыборочный F-тест для дисперсии** вводять такі дані:

- У групі Входные данные у полі **Интервал переменной 1** вводять адресу інтервалу комірок, що містять дані першої вибірки (дисперсія якої мае бути більша), а в полі **Интервал переменной 2** вводять адресу інтервалу комірок, що містять дані другої вибірки;
- у полі Альфа встановлюють рівень значущості (за замовчуванням встановлено a =0,05);
- У групі **Параметры вывода** для виведення результатів обчислень на поточному робочому аркуші активізують перемикач Выходной **интервал** і вказують у полі справа від перемикача адресу комірки для виведення даних (верхню ліву);
- для виведення результатів обчислень на новий аркуш активізують перемикач **Новый рабочий лист;**
- для виведення результатів обчислень у новий файл активізують перемикач **Новая рабочая книга;**
- після встановлення всіх необхідних параметрів натискують кнопку ОК.

У результаті виконаних дій з'явиться таблиця, що містить обчислені середні значення, дисперсії, число степенів свободи для кожної вибірки (у рядку ***df),*** значення критерію Фішера, що спостерігається, (у рядку **F**), та інші. Якщо дисперсія першої змінної виявиться меншою за дисперсію другої змінної, то обчислення повторюють і в якості першої змінної використовують ту, дисперсія якої більше.

Для прийняття рішення порівнюють розрахункове значення критерію Фішера у рядку **F** даної таблиці, з критичним значенням розподілу Фішера $\mathbf{F_{крит}}$ із останнього рядка таблиці. Якщо $\mathbf{F>F_{крит}}$, то дисперсії, що порівнюються, неоднорідні (нульова гіпотеза про їх однорідність відхиляється). Якщо $\mathbf{F<F_{крит}}$, то дисперсії, що порівнюються, однорідні (нульова гіпотеза про їх однорідність приймається).

Приклад. Досліджували вплив важких металів на приріст калюсних клітин польовиці. Результати у відсотках від контрольного варіанту наведені у таблиці 1.

Вимагається встановити з рівнем надійності α=0,05, чи однорідні дисперсії двох вибіркових сукупностей. Результати розрахунків, виконаних у Ms Excel наведено в табл. 2.

Оскільки $F=3,549>F_{крит}=2,484$, то дисперсії мають статистично значущу відмінність, тобто вони неоднорідні. Нульова гіпотеза про рівність дисперсій відхиляється.

Таблиця 1

Приріст калюсних клітин польовиці

Вплив важких металів на приріст калюсних клітин польовиці, % від контролю		
№ вимірювань	Кадмій	Свинець
1	54	60
2	58	57
3	70	63
4	66	59
5	63	57
6	69	57
7	73	58
8	71	62
9	69	59
10	65	60
11	67	65
12	66	57
13	68	60
14	74	64
15	70	64

Таблиця 2

Результати перевірки однорідності дисперсій за критерієм Фішера

Двухвыборочный F-тест для дисперсии		
	Переменная 1	*Переменная 2*
Среднее	66,9136	60,2414
Дисперсия	27,32672583	7,698765971
Наблюдения	15	15
df	14	14
F	3,549494286	
P(F<=f) одностороннее	0,011977	
F критическое одностороннее	2,483723449	

Завдання для лабораторної роботи

Таблиця 4

Вплив важких металів на схожість рослин польовиці, %

Схожість рослин польовиці під дією важких металів, % від контролю									
1 Варіант		2 Варіант		3 Варіант		4 Варіант		5 Варіант	
Кадмій	Свинець	Кадмій	Свинець	Кадмій	Свинець	Кадмій	Свинець	Кадмій	Свинець
99	95	85	84	91	85	80	75	79	81
98	96	87	82	85	86	77	75	79	80
99	96	85	82	89	86	76	74	79	82
98	96	86	82	85	86	76	75	79	79
99	93	89	85	85	85	80	75	80	85
97	90	88	83	89	82	76	74	81	76
97	99	88	87	83	85	75	76	79	83
104	98	84	81	88	84	79	76	80	85
97	92	84	85	89	82	76	76	80	83
95	95	84	85	92	89	79	76	81	83
100	100	86	88	85	81	79	76	81	79
97	94	83	83	87	80	79	75	76	78
96	89	78	85	89	85	78	75	80	82
96	91	91	87	88	81	78	76	78	84

Програма виконання роботи

1. Завантажити табличний процесор MS Excel.

2. Вибрати один варіант даних із завдання (табл.4).
3. Перевірити однорідність дисперсій двох вибірок за критерієм Фішера.
4. Проаналізувати одержані результати.
5. Зберегти файл у власній папці. Надіслати на перевірку викладачу

Лабораторна робота №6
Перевірка гіпотези про рівність середніх

Критерій Ст'юдента (t-тест) для порівняння вибіркових середніх двох незалежних вибірок

Нехай із двох генеральних сукупностей *X* і *У*, що мають розподіл близький до нормального, взято по одній незалежній вибірці. Обчислені для цих вибірок середні значення $\bar{x}_в$ і $\bar{y}_в$, як правило, будуть відрізнятися. Оскільки вибірки є випадковими, то ця різниця може бути випадковою і генеральні середні можуть співпадати.

Потрібно перевірити нульову гіпотезу про рівність генеральних середніх: $H_o : \bar{x}_г = \bar{y}_г$. Гіпотеза про рівність генеральних середніх перевіряється шляхом порівняння вибіркових середніх значень. Значущість різниці між двома вибірковими середніми $\bar{x}_в$ і $\bar{y}_в$ визначається за допомогою критерію *Ст'юдента* (або *t*-критерію).

Значення критерію Ст'юдента, яке спостерігається, (розрахункове значення) $t_{розр}$ обчислюється за формулою

$$t_{\text{розр}} = \frac{|\bar{x}_в - \bar{y}_в|}{\sigma_{\bar{x}-\bar{y}}} \qquad (32)$$

де величина $\sigma_{\bar{x}_в-\bar{y}_в}$ називається *похибкою різниці двох середніх.* Вигляд для обчислення $\sigma_{\bar{x}_в-\bar{y}_в}$ залежить від обсягів вибірок і від того, чи припускаються рівними невідомі дисперсії генеральних сукупностей:

- якщо обсяги вибірок n_x і n_y приблизно одинакові і достатньо великі, тобто, якщо n_x>30 і n_y >30, то похибка різниці двох середніх визначається за формулою:

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \qquad (33)$$

де σ_x^2 і σ_y^2 - дисперсії вибірок із двох генеральних сукупностей *X* і *У*;

- якщо обсяги вибірок n_x і n_y малі, тобто n_x<30 і n_y<30, а дисперсії генеральних сукупностей невідомі, або припускаються рівними, то:

$$\sigma_{\bar{x}-\bar{y}} = \sqrt{\frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2} \times \frac{n_x+n_y}{n_x n_y}}, \qquad (34)$$

де s_x^2 і s_y^2 - виправлені дисперсії двох генеральних сукупностей *X* і *У*.

Тобто, вибір формули для визначення розрахункового значення критерію Ст'юдента залежить від результату перевірки однорідності дисперсії вибірок, що порівнюються. Тому застосуванню критерію Ст'юдента завжди передує перевірка однорідності дисперсій за критерієм Фішера.

Порівняння середніх двох незалежних вибірок засовами MS Excel

Для перевірки статистичної гіпотези про рівність середніх використовують **Пакет анализа.** Оскільки правила застосування критерію Ст'юдента залежать від результатів перевірки однорідності дисперсій генеральних сукупностей, із яких утворені дві незалежні вибірки, то для порівняння середніх цих вибірок, відповідно є два інструменти аналізу: **Двухвыборочный t-тест с одинаковыми дисперсиями** та **Двухвыборочный t-тест с различными дисперсиями.**

Для виклику будь-якого з цих інструментів аналізу виконують команду **Сервис=> Анализ данных,** а потім у діалоговому вікні, що відкриється, вибирають потрібне.

Вибір t-тесту здійснюють лише після того, як буде виконана перевірка гіпотези про рівність генеральних дисперсій двох вибірок і буде встановлено за допомогою критерію Фішера (F-тесту) значущість різниці вибіркових дисперсій. Якщо дисперсії вибірок виявилися рівними, тобто не мають статистично значущої різниці, то вибирають **Двухвыборочный t-тест с одинаковыми дисперсиями** і натискують кнопку ОК. Відкриється діалогове вікно з назвою вибраного інструменту аналізу. Порядок введення даних у цьому вікні такий же, як і у діалоговому вікні F- тесту. Після введення необхідних даних і натискання кнопки ОК з'явиться таблиця з аналогічною назвою.

У таблиці будуть міститися обчислені середні значення і дисперсії для кожної вибірки; обсяги цих вибірок (у рядку: **Наблюдения);** теоретичні генеральні дисперсії обох вибірок (у рядку: **Объединенная дисперсия);** число степенів свободи розподілу Ст'юдента (у рядку: ***df);*** значення критерію Ст'юдента, яке спостерігається (у рядку: **t-статистика);** критична точка розподілу Ст'юдента двосторонньої критичної області (у рядку: **t критическое двустороннее)** та інше.

Якщо після виконання F-тесту дисперсії вибірок виявляються різними, то вибирають **Двухвыборочный t-тест с различными дисперсиями.** Після натискання кнопки ОК відкриється діалогове вікно з аналогічною назвою.

Порядок заповнення полів у даному діалоговому вікні такий же, як і в попередньому випадку. Таблиця в t-тесті з різними дисперсіями практично співпадає з таблицею у t- тесті з однаковими дисперсіями за винятком того, що в останньому випадку відсутній рядок **Объединенная дисперсия.**

Порівнюють *за абсолютною величиною* значення критерію Ст'юдента, що знаходиться в рядку **t-статистика,** з критичною точкою розподілу Ст'юдента двосторонньої критичної області **t критическое двухстороннее.**

Приклад. За даними табл. 4 перевірялася однорідність дисперсій за критерієм Фішера. Виявилося, що дисперсії однорідні. Для перевірки середніх значень був застосований критерій Ст'юдента для вибірок з однорідними дисперсіями. Результати розрахунків наведені у табл. 3.

Таблиця 3

Результати перевірки середніх значень за критерієм Ст'юдента

Двухвыборочный t-тест с одинаковыми дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	10,35	8,45
Дисперсия	1,5536	0,9754
Наблюдения	12	12
Объединенная дисперсия	1,26454	
Гипотетическая разность средних	0	
df	22	
t-статистика	4,1386	
P(T<=t) одностороннее	0,0002	
t критическое одностороннее	1,7171	
P(T<=t) двухстороннее	0,0004	
t критическое двухстороннее	2,0738	

Оскільки $t_{статистика}=4,1386>t_{Крит.двухст.}=2,0738$, то між середніми, що порівнюються, є статистично значуща різниця, тобто середні двох вибірок відрізняються. Значення t-статистики завжди береться за модулем (за абсолютною величиною!).

Програма виконання роботи

1. Завантажити табличний процесор MS Excel.
2. Вибрати один варіант даних із завдання (табл.4).
3. Перевірити гіпотезу про рівність середніх значень двох незалежних вибірок і проаналізувати одержані результати.
4. Зберегти документ у власній папці і надіслати не перевірку викладачу.

Запитання для самоперевірки

1. Що називається статистичною гіпотезою?
2. Які бувають статистичні гіпотези?
3. Що називається статистичним критерієм?
4. Для чого використовується порівняння двох дисперсій?
5. За яких умов дисперсії вважають однорідними і за яких - ні?
6. Про що свідчить неоднорідність дисперсій?
7. За яким критерієм порівнюють дві дисперсії і який порядок його застосування?
8. Які засоби MS Excel використовують для порівняння дисперсій?
9. За якими показниками, виведеними у таблиці F-тесту MS Excel, визначають однорідність дисперсій?

10.За яким критерієм перевіряють гіпотезу про рівність середніх значень?

11.Від чого залежить вибір виду критерію Ст'юдента для порівняння середніх?

12.Які правила прийняття рішення при застосуванні t-тесту?

13.Як задати рівень надійності, за яким треба перевірити гіпотезу?

14.Який рівень надійності встановлений за замовчуванням у F- та t-тестах?

15.Як впливає значення рівня надійності на результати порівняння?

Лабораторна робота № 7
Однофакторний дисперсійний аналіз

Мета роботи: набуття навичок проведення однофакторного дисперсійного аналізу засобами Excel

Теоретичні відомості

Однофакторний дисперсійний аналіз

Метод досліджень, заснований на порівнянні дисперсій, називається *дисперсійним аналізом.*

Основна ідея дисперсійного аналізу полягає у порівнянні *«факторної дисперсії»* обумовленої впливом факторної ознаки, і *«залишкової дисперсії»* обумовленої впливом випадкових чинників. Якщо різниця між цими дисперсіями статистично значуща, то вплив факторної ознаки, що вивчається, на результативну ознаку також є статистично значущим.

Дисперсійний аналіз застосовують для оцінки впливу *кількісних* або *якісних* факторних ознак (факторів). *Кількісний фактор* - факторна ознака, яку можна виразити кількісно (можна виміряти), наприклад, кількість внесених добрив, температура і відносна вологість повітря тощо. Рівні варіювання кількісного фактора - конкретні значення, які цікавлять дослідника, наприклад, певна кількість внесених добрив, конкретні значення температури, або вологості повітря тощо.

Якісний фактор - факторна ознака, яку не можна виміряти, тобто яка приймає тільки якісні значення, наприклад, вид добрива, вид механічного засобу для обробки ґрунту тощо. Рівні варіювання якісного фактору - певні види добрив, певні види засобів тощо. Рівні варіювання якісного фактору не можуть приймати проміжних значень.

У MS Excel є засіб для проведення однофакторного дисперсійного аналізу. Для його використання виконують дії:

дослідні дані розміщують у таблиці, де: назви стовпчиків – рівні факторної ознаки; рядки - результати повторних вимірювань;

виконують дії: Анализ данных => Однофакторный дисперсионый анализ;

у поля однойменного діалогового вікна, що відкриється, вводять потрібне:

- у поле **Входной интервал** вводять прямокутний діапазон комірок з вихідними даними;
- серед трупи перемикачів **Группирование** відмічають по **столбцам**, оскільки дані, що обумовлені рівнями факторної ознаки, розміщені у стовпчиках;
- якщо у введений діапазон даних увійшли назви стовпчиків, то відмічають пункт **Метки в первой строке;**
- залишають рівень надійності **Альфа** без змін, якщо значення 0,05 задовольняє, або вводять інше значення у відповідне поле;
- у групі перемикачів **Параметры вывода** відмічають **Выходной интервал** і вводять адресу верхньої лівої комірки для виведення результатів, або вибирають інше;
- натискують кнопку **ОК.**

Приклад. Вивчали вплив якісного фактора на трьох рівнях варіювання (вилив трьох видів добрив) на кількість зерен в колосі ярої пшениці. Результати досліджень наведено в табл. 8.

Таблиця 8 - Результати досліджень впливу видів добрив на кількість зерен у колосі ярої пшениці

Номер досліду (повторності)	Вид. добрив		
	A	B	C
	Кількість зерен у колосі, шт.		
1	51	52	42
2	52	54	44
3	56	56	50
4	57	58	52

Після заповнення полів однойменного діалогового вікна виводяться дві таблиці. В табл. 9 представлені підсумкові розрахунки для рівнів факторної ознаки - для кожного із видів добрив, які трактуються як групові результати. Визначаються середні значення і дисперсії всередині груп (серед повторних вимірювань), обумовлені впливом випадкових чинників.

Таблиця 9 - Проміжні результати дисперсійного аналізу

Одно факторный дисперсионный анализ				
ИТОГИ				
Группы	Счет	Сумма	Среднее	Дисперсия
Столбец 1	4	216	54	8,6666
Столбец 2	4	220	55	6,6666
Столбец 3	4	188	47	22,6666

- **Столбец 1**, **Столбец 2**, **Столбец 3** - рівні варіювання якісної факторної ознаки - виду добрив;
- **Счет** - кількість повторностей для кожного рівня факторної ознаки;
- **Сумма** - загальна сума одержаних результатів;
- **Среднее** - середнє значення результативної ознаки для кожного рівня факторної ознаки;
- **Дисперсия** - дисперсія, обумовлена впливом випадкових чинників, тобто відмінностями між повторними вимірюваннями

.У табл. 10 наведені підсумкові результати дисперсійного аналізу.

Таблиця 10 - Підсумкові результати дисперсійного аналізу

Дисперсионный анализ						
Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	152	2	76	6	0,0220	4,2564
Внутри групп	114	9	12,6666			
Итого	266	11				

- **SS** - сума квадратів:
  - **Между группами** - міжгрупова сума квадратів (обумовлена впливом рівнів факторної ознаки, що вивчається): $\sum_{i=1}^{n}\left(\overline{X}_i - \overline{X}\right)^2$,

    де $\overline{X}$ - загальне середнє значення результативної ознаки; $\overline{X}_i$ - середнє значення результативної ознаки, обумовлене впливом *i*- го рівня факторної ознаки; i =1…n, де n- кількість рівнів факторної ознаки;
  - **Внутри групп** - у середині груп (випадковий вплив): $\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\overline{X}_i - \overline{X}\right)^2$, де X_{ij} - поточне значення результативної ознаки, вплив на яку вивчається; j = 1..m - поточне значення повторності; m - кількість повторних дослідів;
  - **Итого** - загальна $\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\overline{X}_i - \overline{X}\right)^2$, де m=4 - кількість дослідів;

n-3 - кількість рівнів факторної ознаки;

- **df** - число степенів свободи;
- **MS** - дисперсії (середнє квадратів: MS=SS/df);
- **F** розрахункове значення критерію Фішера;
- **P-Значение** - розрахункове значення мінімальної значущості;
- **F-критическое** - критичне значення критерію Фішера.

Оскільки $F>F_{крит}$, то між факторною дисперсією, обумовленою впливом рівнів факторної ознаки - виду добрив, і залишковою дисперсією, обумовленою впливом випадкових чинників, існує статистично значуща різниця. Отже, з рівнем надійності 0,05 можна стверджувати, що вплив видів добрив на кількість зерен у колосі ярої пшениці є статистично значущим.

Порядок виконання роботи

1. Завантажити табличний процесор MS Excel.
2. Ввести дані із завдання (один із варіантів табл. 11).
3. Визначити значущість впливу рівнів факторної ознаки, що досліджується, на результативну ознаку.
4. Провести аналіз одержаних результатів.
5. Зберегти документ в особистій папці.
6. Завершити роботу з Excel.

Завдання для лабораторної роботи

Таблиця 11 - Результати досліджень впливу виду технології вирощування на показники врожаю озимої пшениці

№ досліду (повторності)	Вид технології вирощування пшениці		
	1a	2b	3c
3 варіант	Густота стояння, шт./м^2		
1	397	410	437
2	399	412	432
3	401	415	436
4	389	409	429
2 варіант	Висота рослин, см		
1	90	98	110
2	91	99	108
3	93	97	109
4	92	100	115
3 варіант	Загальне кущення, шт.		
1	2,4	2,7	3,2
2	2,2	2,8	3,1
3	2,3	2,5	3,3
4	2,3	2,7	3,4
4 варіант	Продуктивне кущення, шт.		
1	2,0	2,5	3.1
2	2,1	2,6	3,0
3	2,2	2,4	3,2
4	2,1	2,4	3,0
5 варіант	Довжина колоса, см		
1	7,0	7,6	8,2
2	7,1	7,5	8,4
3	7,2	7,4	8,3
4	7,1	7,5	8,6
6 варіант	Кількість колосків, шт.		
1	15	19	22
2	16	Г __ І?	23
3	15	20	22
7 варіант	Кількість зерен в колосі, шт.		
1	35	39	42
2	37	38	43
3	34	41	45
4	36	40	44

8 варіант	Маса зерен з 10 рослин, г		
1	25	31	33
2	26	32	34
3	28	30	37
4	25	32	35
9 варіант	Маса соломи, г		
1	35	41	54
2	36	42	57
3	37	44	53
4	36	43	58
10 варіант	Маса 1000 зерен, г		
1	32	37	42
2	33	38	43
3	31	36	42
4	32	38	44

Запитання для самоперевірки

1. Які відмінності між якісними і кількісними факторними ознаками?
2. У чому полягає сутність дисперсійного аналізу?
3. Для чого застосовують дисперсійний аналіз?
4. Які особливості використання засобу MS Excel Од н о фа кто р и ы й дисперсионный анализ?
5. Які результати виводяться при використанні засобу Од н о ф а кто р НЫ й дисперсионный анализ?
6. На основі яких даних робиться висновок щодо статистичної значущості впливу рівнів факторної ознаки, що вивчається?
7. Чи залежать одержані результати від прийнятого рівня значущості?
8. Як задати потрібний рівень значущості при проведенні дисперсійного аналізу?
9. Як перевірити, чи існує статистично значуща різниця між впливом двох рівнів факторної ознаки?

Лабораторна робота № 8
Двофакторний дисперсійний аналіз

Мета роботи: набуття навичок проведення двофакторного дисперсійного аналізу засобами Excel

Теоретичні відомості

Двофакторний дисперсійний аналіз

Двофакторний дисперсійний аналіз застосовують для оцінки впливу як видів факторних ознак, так і їх рівнів. Причому оцінюють також, що із них має більший вплив: вид факторних ознак, чи рівні їх варіювання. Для проведення двофакторного дисперсійного аналізу застосовують засіб MS Excel з аналогічною назвою: **Анализ данных → Двухфакторный дисперсионный анализ без повторений**. Поля діалогового вікна заповнюються як і у випадку однофакторного дисперсійного аналізу.

Приклад. Досліджували вплив різних технологій вирощуваная (якісний фактор) і різних концентрацій добрив (кількісний фактор) и а показники врожаю ярої пшениці. Вимагається оцінити значущість виливу зазначених факторних ознак на результативну ознаку. Результати експеримент}/ наведено у табл. 12.

Таблиця 12 - Результати експериментальних досліджень впливу концентрації добрив і виду технологій вирощування на густоту стояння ярої пшениці

Густота стояння ярої пшениці, шт./м^2			
Варіант досліду	Технологія 1	Технологія 2	Технологія 3
$N_{30}P_{40}K_{15}$	407	412	395
$N_{6}OP_{8}OK_{3}O$	437	442	405
$N_{90}P_{120}K_{45}$	455	469	425

Результати розрахунків виводяться у вигляді таблиць 13 і 14.

Таблиця 13 - Результати проміжних розрахунків двофакторного дисперсійного аналізу

Двухфакторный дисперсионный анализ без повторений				
ИТОГИ	Счет	Сумма	Среднее	Дисперсія
Строка 1	3	1214	404,6666	76,3333
Строка 2	3	1284	428	403
Строка 3	3	1349	449,6666	505,3333
Столбец 1	3	1299	433	588
Столбец 2	3	1323	441	813
Столбец 3	3	1225	408,3333	233,3333

Таблиця 14 - Підсумкові результати двофакторного дисперсійного аналізу

Дисперсионный анализ						
Источник вариации	SS	*df*	MS	F	P-значение	F критич.
Строки	3038,88	2	1519,44	26,4506	0,0049	6,9442
Столбцы	1739,55	2	869,77	15,1411	0,0136	6,9442
Погрешность	229,77	4	57,44			
Итого	5008,22	8				

- **SS строки** - сума квадратів, обумовлена виливом рівнів факторної ознаки результати вимірювань якої розташовані у рядках (концентрації добрив);
- **SS столбцы** - сума квадратів, обумовлена впливом рівнів факторної ознаки результати вимірювань якої розташовані у стовпчиках (технології вирощування пшениці).
- **Погрешность** - залишкова сума квадратів, обумовлена випадковою похибкою;
- **df** - число степенів свободи;
- **MS** - середній квадрат (фактична дисперсія результативної ознаки);
- **F** - розрахункове значення критерію Фішера (значення критерію, що спостерігається),

P -значение - розрахункове значення мінімальної значущості;

F критическое - критичне значення критерію Фішера.

Згідно з даними, наведеними у табл. 12-14, у рядках розміщуються рівні кількісного фактора - концентрації добрив, у стовпчиках - рівні якісного фактора - технології вирощування. Порівнюється F значення з $F_{крит}$. Оскільки у першому випадку F=26,45 > $F_{крит}$=6,94, то концентрація добрив має статистично значущий вплив на вихідний параметр - густоту стояння ярої пшениці. У другому випадку F=15,14> $F_{крит}$=6,94, отже, вид технології вирощування також має статистично значущий вплив на вихідний параметр. Причому концентрація добрив (F=26,45) має більший вплив порівняно з видом технології вирощування (F=15,14).

Порядок виконання роботи

1. Завантажити табличний процесор MS Excel.
2. Ввести дані із завдання (один із варіантів табл. 15).
3. Визначити статистичну значущість впливу факторних ознак, що досліджуються, на результативну ознаку.
4. Встановити, яка із факторних ознак має більш сильний вилив?
5. Яку факторну ознаку можна віднести до якісної, а яку - до кількісної?
6. Оцінити вплив випадкових чинників на результативну ознаку.
7. Провести аналіз одержаних результатів.
8. Зберегти документ в особистій папці.
9. Завершити роботу з Excel .

Запитання для самопгревірки

1. У яких випадках використовують двофакторний дисперсійний аналіз?

2. Які результати виводяться при використанні засобу ***Двухфакторный дисперсионный анализ?***
3. На основі яких даних робиться висновок щодо статистичної значущості впливу факторних ознак, що вивчаються, на результативну ознаку?
4. Як оцінити, яка із факторних ознак має більш сильний вплив?
5. Як ураховується рівень надійності при застосуванні інструменту аналізу ***Двухфакторный дисперсионный анализ без повторений?***
6. Чи можна засобами дисперсійного аналізу вивчати вплив тільки кількісних, або тільки якісних факторних ознак?
7. Як оцінити вплив випадкових чинників на результативну ознаку?
8. Що розуміють під "впливом випадкових чинників" ?

Завдання для лабораторної роботи

Таблиця 15 - Результати досліджень виливу технології вирощування і сорту цибулі на якісні показники продукції

Вид технології вирощування	Сорт цибулі		
	Балстора	Сквирська	Полтавська
1 варіант	Вміст сухої речовини, %		
A	12,1	13,2	16,1
B	12,6	13,7	16,0
C	12,4	14,1	15,9
2 варіант	Загальний цукор, %		
A	6,9	8,4	7,7
B	7,0	8,6	7,6
C	7,1	8,9	7,8
3 варіант	Сахароза, %		
A	3,9	4,4	4,8
B	4,0	4,8	4,7
C	3,8	5,3	4,9
4 варіант	Вітамін С, %		
A	8,5	6,8	8,8
B	9,0	7,0	8,9
C	8,1	6,9	9,1
5 варіант	Нітрат, мг/кг		
A	40,7	44,3	38,4
B	40,8	43,7	38,9
C	40.6	45,0	37,9
6 варіант	Ефірні масла, мг/кг		

A	76,1	67,5	81,3
B	73,7	67,9	82,6
C	75,7	69,9	81,0
7 варіант	Вміст сухої речовини, %		
A	9,8	12,1	11,7
B	10,2	12,5	9,8
C	9,9	12,9	10,7
8 варіант	Загальний цукор, %		
A	7,5	8,4	9,5
B	7,3	8,9	9,1
C	7.9	8,2	9,3
9 варіант	Сахароза, %		
A	5,7	4.9	6,2
B	5,9	4,6	6,3
C	5,4	4,0	6,7
10 варіант	Вітамін С, %		
A	8,4	9,2	7,5
B	8,1	9,3	7,9
C	8,3	9,5	1,0

Лабораторна робота №9
Побудова двовимірної лінійної математичної моделі за методом найменших квадратів

Мета роботи: отримання практичних навичок побудови двовимірної лінійної математичної моделі за методом найменших квадратів засобами MS Excel

Теоретичні відомості

Оцінка наявності і тісноти лінійної залежності між двома змінними

Однією з найбільш розповсюджених задач наукових досліджень є задачі, пов’язані з вивченням статистичних залежностей між двома і більше випадковими величинами.

Залежність між двома змінними величинами, при якій значенню однієї змінної величини відповідає одне значення другої, називається функціональною. Функціональна залежність між двома величинами зустрічається дуже рідко. Найчастіше між величинами існує так звана

статистична залежність, при якій кожному значенню однієї із випадкових величин відповідає закон розподілу другої.

Статистична залежність випадкових величин, при якій досліджується вплив зміни однієї із величин на середнє значення другої називається *кореляційною.*

Кореляційну залежність *Y* від *X* виражає рівняння регресії *Y* на *X,* що задається у вигляді $\bar{y}(x) = \varphi(x)$.

Першою задачею кореляційного аналізу є з'ясування форми кореляційної залежності, тобто вигляду функції регресії $\varphi(x)$. Якщо функція $\varphi(x)$ – лінійна, то кореляцію називають лінійною. Інакше кореляція називається нелінійною.

Вищезгадана задача розв'язується шляхом побудови за вибірковими даними емпіричної функції регресії, яку підбирають так, щоб вона якомога краще відображала характерні особливості статистичних даних. Практично ця задача збігається з задачею підбору емпіричних формул за експериментальними даними і найчастіше розв'язується методом найменших квадратів.

Друга задача кореляційного аналізу полягає в тому, щоб оцінити, наскільки тісною (сильною) є кореляційна залежність між випадковими величинами.

Сила кореляційного зв'язку *Y* від *X* оцінюється за величиною розсіювання значень *Y* навколо умовного середнього $\bar{y}(x)$. Значне розсіювання свідчить про слабку залежність *Y* від *X* або навіть про відсутність такої залежності. Мале розсіювання говорить про існування сильної залежності *Y* від *X*.

Основними характеристиками, що описують силу зв'язку між складовими *X* і *Y* двовимірної випадкової величини (*X, Y)* є кореляційний момент (або коваріація),

$$\mu_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}, \qquad (1)$$

де $\bar{x}$, $\bar{y}$, $\overline{xy}$ - середні значення випадкових величин *x* та *y* та їх добутку;
і коефіцієнт кореляції

$$r_{x,y} = \frac{\mu_{xy}}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} \qquad (2)$$

де σ_x, σ_y – середні квадратичні відхилення величин *X* та *Y*.

Коефіцієнт парної кореляції є безрозмірною величиною, не залежить від вибору одиниць вимірювання величин, що спостерігаються і має такі властивості:

1. Коефіцієнт парної кореляції за модулем менше або дорівнює одиниці, $|r| \le 1$. А саме $-1 \le r \le 1$.

2. Якщо коефіцієнт парної кореляції дорівнює одиниці, то вибіркові значення ознак *X* та *У* зв'язані лінійною функціональною залежністю.

3. Якщо коефіцієнт парної кореляції дорівнює нулю, то вибіркові значення ознак *X* та *У* не зв'язані лінійною кореляційною залежністю.

4. При зростанні абсолютної величини коефіцієнта парної кореляції, лінійна кореляційна залежність між *X* та *У* стає більш тісною, а при зменшенні – зв'язок слабкий.

	0	зв'язок відсутній
0,1	0,29	слабкий
0,3	0,49	помірний
0,5	0,69	значний
0,7	0,89	високий
0,9	0,99	дуже високий
	1	зв'язок функціональний

Якщо коефіцієнт кореляції статистично значущий, то між змінними існує лінійний зв'язок. Якщо коефіцієнт кореляції більше нуля, то зв'язок між змінними прямий, якщо менше нуля, то зв'язок обернений. Якщо коефіцієнт кореляції статистично незначущий, або дорівнює нулю, то дві випадкові величини X і У не пов'язані лінійною залежністю. Але це не означає, що між випадковими величинами взагалі немає кореляційного зв'язку. Це означає, що між змінними може існувати нелінійний зв'язок.

Статистичну значущість коефіцієнта парної кореляції визначають за допомогою критерію Стьюдента. При цьому перевіряють нульову гіпотезу про рівність коефіцієнта кореляції нулю.

Розрахункове значення *t* – критерію, власне *t* - розрахункове $t_{розр}$ обчислюють за формулою:

$$t_{розр} = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}} \quad , \qquad (3)$$

де *r*- значення коефіцієнта парної кореляції;

n- кількість спостережень.

Розрахункове значення *t* критерію Стьюдента порівнюють з табличним (критичним) значенням, визначеним на деякому рівні значущості α (найчастіше α приймають рівним 0,05) та з *n*-2 степенями свободи. Якщо розрахункове значення *t* критерію більше табличного, то прийнята нульова

гіпотеза про рівність коефіцієнта кореляції нулю відхиляється, у протилежному випадку приймається.

Коефіцієнт кореляції, обчислений за даними вибірки, називається вибірковим і позначається $r_в$. Вибірковий коефіцієнт кореляції є оцінкою коефіцієнта кореляції генеральної сукупності. Коефіцієнт кореляції генеральної сукупності знаходиться в межах довірчого інтервалу.

Половину ширини довірчого інтервалу для коефіцієнта кореляції визначають за формулою:

$$\delta = \frac{t_{розр}(1-r_в^2)}{\sqrt{n}}, \qquad (4)$$

де n- число спостережень,

$r_в$ - вибірковий коефіцієнт кореляції;

$t_{розр}$ – розрахункове значення t- критерію Стьюдента.

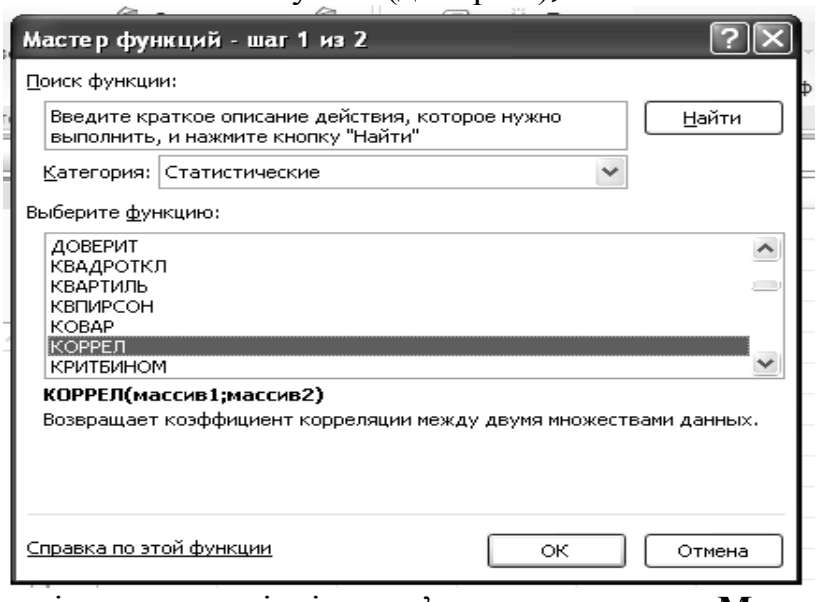
Межі коефіцієнту кореляції генеральної сукупності визначаються за допомогою подвійної нерівності

$$r_в - \delta \le r_в \le r_в + \delta \qquad (5)$$

Застосування засобів MS Excel для знаходження коефіцієнтів парної кореляції

Коефіцієнт кореляції знаходять за допомогою вбудованої функції **КОРРЕЛ.** Для цього виконують такі дії:

- виокремлюють комірку, в якій буде знаходитись результат обчислення $r_в$**;**
- на панелі інструментів в меню **Вставка→Функция** у полі **Категория** вибирають **Статистические**;
- у полі **Функция** обирають **КОРРЕЛ**;
- натискають кнопку **ОК** (див. рис.)**;**



- у діалоговому вікні, що з’явиться, у рядку **Массив 1** записують діапазон комірок однієї змінної, а у рядку **Массив 2** - діапазон

комірок другої, або виокремлюють діапазон комірок зі змінними безпосередньо у таблиці;

- натискають кнопку **ОК.**

Для визначення значущості отриманого коефіцієнта кореляції обчислюють значення *t* критерію Стьюдента – розрахункове та критичне (табличне).

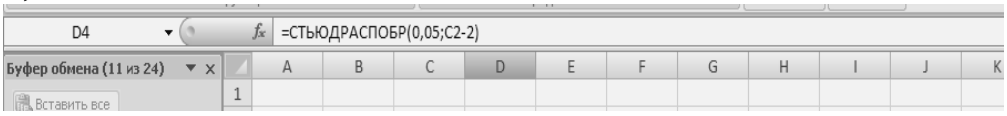
Табличне (критичне) значення критерію обчислюють за допомогою вбудованої функції MS Excel **СТЬЮДРАСПОБР** категорії **Статистические,** аргументами якої є рівень значущості (α=0,05) і число степенів свободи *n*-2.

Число ступенів свободи визначають, попередньо визначивши число спостережень. З цією метою використовують функцію **СЧЕТ** категорії **Статистические**. Для цього виконують такі дії:

- активізують комірку, в якій буде знаходитись результат;
- на панелі інструментів в меню **Вставка→Функція** у полі **Категория** вибирають **Статистические**;
- у полі **Функция** обирають **СЧЕТ**;
- у діалоговому вікні, що з'явиться виокремлюють діапазон комірок з даними;
- натискають кнопку **ОК**;
- у відповідній комірці з'явиться значення, що відповідає числу спостережень *n*.

Для знаходження значення табличного (критичного) значення *t* критерію Стьюдента виконують наступне:

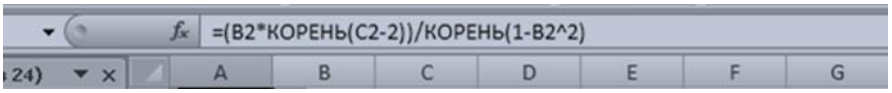
- активізують комірку, в якій буде знаходитись результат;
- на панелі інструментів в меню **Вставка→Функція** у полі **Категория** вибирають **Статистические**;
- у полі **Функция** обирають **СТЬЮДРАСПОБР**;
- у діалоговому вікні, що з'явиться у рядок **Вероятность** вводять число, що дорівнює рівню значущості 0,05;
- у рядок **Степень_свободы** вводять адресу комірки, в якій знаходиться значення числа спостережень і від нього віднімають 2;



- натискають кнопку **ОК**;
- у відповідній комірці з'явиться значення, що відповідає табличному (критичному) значенню *t* критерію Стьюдента.

Розрахункове значення *t* критерію Стьюдента обчислюють безпосередньо за формулою (3). Для цього виконують наступні дії:

- виокремлюють комірку, в якій буде знаходитись результат;
- з клавіатури вводять знак =;
- перемикають клавіатуру на латинську розкладку;
- вводять необхідну формулу у вигляді

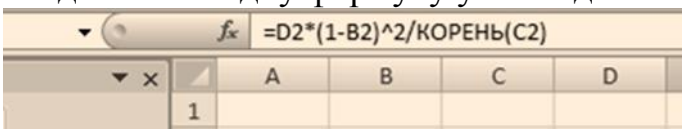


- натискують клавішу Enter;
- у відповідному вікні з’явиться результат.

Порівнюємо отримані розрахункове та критичне (табличне) значення *t* критерію Стьюдента. Робимо висновок про статистичну значущість отриманого коефіцієнту кореляції.

Якщо коефіцієнт кореляції виявився статистично значущим, то визначають величину довірчого інтервалу за формулою (4). Для цього виконують наступне:

- виокремлюють комірку, в якій буде знаходитись результат;
- з клавіатури вводять знак =;
- перемикають клавіатуру на латинську розкладку;
- вводять необхідну формулу у вигляді



- натискують клавішу Enter;
- у відповідному вікні з’явиться результат.

Визначають межі коефіцієнта кореляції за формулою (5). Для визначення нижньої межі: від значення коефіцієнту кореляції $r_в$ віднімають значення δ. Для визначення верхньої межі: до значення коефіцієнту кореляції $r_в$ додають значення δ.

Визначення параметрів і похибки моделі

Після встановлення наявності та оцінки тісноти лінійного зв’язку між змінними, переходять до побудови математичної моделі. Задача моделювання полягає в тому, щоб припускаючи лінійну залежність між змінними *X* і *Y* , одержати найкращу регресійну пряму у вигляді рівняння

$$Y = ax + b \qquad (1)$$

Для визначення параметрів *a* і *b* лінійної залежності

досліджують вираз *D*:

$$D = \sum_{i=1}^{n} \frac{(y_i - Y_i)^2}{n} = \sum_{i=1}^{n} (\frac{y_i - a \cdot x_i - b)^2}{n}, \qquad (2)$$

де Y_i і y_i - теоретичні (обчислені за формулою (1)) і дослідні значення. Вираз (2) являє собою дисперсію відхилень дослідних значень *y* відносно теоретичних значень *Y*. Коефіцієнти *a* і *b* потрібно підібрати таким чином, щоб вираз *D* (2) мав мінімальне значення. З математики відомо, що для цього потрібно прирівняти до нуля частинні похідні:

$$\frac{\partial D}{\partial a} = 0, \quad \frac{\partial D}{\partial b} = 0. \qquad (3)$$

Знайшовши похідні, одержимо систему двох рівнянь з двома невідомими *a* і *b*:

$$\begin{cases} b + a \cdot \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{n} \frac{y_i}{n} \\ b \cdot \sum_{i=1}^{n} \frac{x_i}{n} + a \cdot \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{n} x_i \cdot \frac{y_i}{n} \end{cases} \quad (4)$$

або

$$\begin{cases} b + a\bar{x} = \bar{y}, \\ b\bar{x} + a\overline{x^2} = \overline{xy}, \end{cases} \quad (5)$$

де

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n},\ \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n},\ \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}\ . \quad (6)$$

Система рівнянь (5) називається нормальною і у підручниках записується у вигляді

$$\begin{cases} b \cdot n + a \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b \cdot \sum_{i=1}^{n} x_i + a \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}, \quad (7)$$

яка виводиться з умови мінімуму виразу

$$L = \sum_{i=1}^{n} (y_i - Y_i)^2 \quad (8)$$

або множення системи (4) на n.

Знайдемо розв'язки системи (5) методом підстановки. З першого рівняння визначимо *b* і підставимо у друге, після чого визначимо *a*. В результаті одержимо:

$$b = \bar{y} - a\bar{x},$$
$$\bar{x} \cdot \bar{y} - a\bar{x}^2 + a\overline{x^2} = \bar{x} \cdot \bar{y},$$
$$a\left(\overline{x^2} - \bar{x}^2\right) = \overline{xy} - \overline{xy},$$
$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = R_{yx},$$
$$b = \bar{y} - \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \cdot \bar{x} = \bar{y} - R_{yx} \cdot \bar{x}.$$

Визначення параметрів двовимірної лінійної моделі за допомогою вбудованої функції ЛИНЕЙН

Функція **ЛИНЕЙН** категорії ***Статистические*** дозволяє за методом найменших квадратів визначити параметри лінійної моделі, яка найкращим чином апроксимує експериментальні дані. Функція повертає масив значень, який характеризує теоретичну прямолінійну залежність.

Синтаксис функції: ЛИНЕЙН (известные_значения_y; известные_значения_x; конст; статистика),

де ***известные значения*** **y** - діапазон комірок зі значеннями результативної ознаки;

известные значения x - діапазон комірок зі значеннями факторної ознаки;

конст - логічне значення, яке вказує на наявність у лінійної моделі вільного члена. Якщо ***конст*** має значення ***ИСТИНА,*** або значення не вказується, то модель матиме вигляд: Y=a+bX, якщо ***конст*** має значення ***ЛОЖЬ,*** то модель матиме вигляд Y=bX.

статистика - логічне значення, яке вказує, чи треба виводити додаткові статистичні показники. Якщо аргумент статистика має значення ***ИСТИНА,*** то функція ***ЛИНЕЙН*** повертає додаткову регресійну статистику; якщо аргумент статистика має значення ***ЛОЖЬ,*** то виводяться лише коефіцієнти рівняння.

Додаткова регресійна статистика включає:

se_1 - стандартне значення похибки оцінки коефіцієнта *b*;

se_2 - стандартне значення похибки оцінки коефіцієнта *a* - вільного члена;

r^2 - коефіцієнт детермінації;

se_y - стандартна похибка оцінки Y;

F - F-статистика - розрахункове значення критерію Фішера;

df~~ число степенів свободи;

$ss_{регр}$ - регресійна сума квадратів;

$ss_{зал}$, - залишкова сума квадратів.

m_1	a
se_1	se_2
r^2	se_v
F	df
$ss_{регр}$	$ss_{зал}$

Для застосування функції **ЛИНЕЙН** виконують дії:

- на вільному місці аркуша виокремлюють прямокутний діапазон комірок, який складається із двох стовпчиків і п'яти рядків для виведення результатів розрахунку;
- вводять знак "=" (дорівнює);
- **виконують дії: Вставка=> Функция :=>Статистические ЛИНЕЙН**;
  - заповнюють поля діалогового вікна, що відкриється:

- у поле ***известные значения*** **y** вводять діапазон комірок зі значеннями результативної ознаки;

- у поле ***известные значения*** **x** - діапазон комірок зі значеннями факторної ознаки;
- у поле ***конст*** вводять логічне значення ***ИСТИНА*** (або ***1);***
- у поле ***статистика*** - логічне значення ***ИСТИНА*** (або ***1);***
- натискують клавішу F2, щоб вивести масив значень;
- натискують одночасно клавіші Ctrl + Shift + Enter;
- після цього діапазон комірок 2x5 буде заповнено значеннями згідно наведеної таблички;
- якщо, в результаті виконання зазначених дій, буде виведене одне число, треба впевнитися, що необхідний діапазон комірок виокремлено, якщо - ні, то виокремити його;
- ще раз натискають клавішу F2, щоб вивести масив значень;
- натискують одночасно клавіші Ctrl + Shift + Enter.

У результаті виконання зазначених дій буде виведено масив значень, який містить параметри моделі і оцінку їх точності.

Перевірка правильності визначення параметрів моделі і побудова графічної залежності

Для перевірки правильності визначення параметрів рівняння, яке б описувало дану залежність, засобами Excel виконують дії:

будують точковий графік на основі двох рядів спостережень (x_i, y_i), між якими планується визначити функціональну залежність:

- виокремлюють ряди спостережень;
- **виконують дії Вставка => Диаграмма**;
- у діалоговому вікні, що відкриється, на закладці ***Стандартная*** вибирають ***Точечная,*** у ***полі Вид*** вибирають ***Точечная диаграмма позволяет сравнивать пары значений,*** натискують кнопку Далее;
- продовжують побудову графічної залежності, як описано в попередніх лабораторних роботах;

виокремлюють побудований графік (один раз клацають лівою клавішею миші по одній з точок графіка);

клацають правою клавішею миші по графіку, викликаючи контекстне меню;

у контекстному меню вибирають пункт **Добавить линию тренда**;

або виконують дії: меню ***Диаграммам /Добавить линию тренда;***

у діалоговому вікні**,** що відкриється, на закладці ***Тип*** вибирають запропоновану лінійну залежність, якою планується апроксимувати експериментальні точки;

▪ на закладці **Параметри** відмічають пункт **показывать уравнение на диаграме і поместить на диаграмму величину достоверности апроксимации**;

натискують кнопку **ОК**;

▪ на графіку з'явиться апроксимуюча лінія з рівнянням, за яким вона побудована.

За величиною коефіцієнта детермінації R^2 (квадрат коефіцієнта кореляції) оцінюють якість одержаної моделі. Чим більше коефіцієнт R^2 наближається до одиниці, тим краще математична залежність описує експериментальні дані. Якщо, наприклад, R^z=0,9144, то, у випадку лінійної залежності, 91,44% дисперсії результативної ознаки обумовлено впливом факторної ознаки, що досліджується, а решта - впливом випадкових чинників.

Для обчислення похибки моделі використовують вбудовану функцію ***СТОШУХ*** категорії ***Статистические,*** аргументами якої є діапазон комірок зі значеннями результативної ознаки $y_{i:}$ і діапазон комірок зі значеннями факторної ознаки x_t.

Програма виконання роботи

1. Завантажити табличний процесор Excel.
2. Використати дані двох стовпчиків із табл. до завдання.
3. Установити наявність та оцінити тісноту лінійного зв'язку між вибраними показниками за допомогою коефіцієнта кореляції.
4. Перевірити статистичну значущість одержаного коефіцієнта кореляції.
5. Визначити параметри моделі і додаткову регресійну статистику за допомогою функції ЛИНЕЙН.
6. Перевірити правильність розрахунків за допомогою засобу Добавить линию тренда.
7. Знайти похибку моделі за допомогою функції ***СТОШУХ.***
8. Порівняти одержані результати.
9. Зберегти документ в особистій папці.

Лабораторна робота №10
Виявлення наявності і оцінка тісноти статистичної залежності між змінними та робота з формулами масивів

Мета роботи: отримання практичних навичок проведення кореляційного аналізу засобами MS Excel, використання формул масивів для роботи з даними

Теоретичні відомості

Залежність випадкових величин, при якій кожному значенню однієї із них відповідає закон розподілу другої, тобто зміна однієї із величин спричиняє змінювання розподілу значення другої, називається *статистичною.* Статистична залежність випадкових величин, при якій досліджується вплив зміни однієї із величин на середнє значення другої називається *кореляційною.* Наприклад, кореляційною є залежність середніх значень гематологічних і біохімічних показників периферичної крові свиней від терміну, який пройшов від часу введення препарату, що на ці показники впливає.

Одним із показників, що дозволяє встановити наявність лінійної залежності між двома змінними і оцінити її тісноту, є *коефіцієнт парної кореляції.*

Коефіцієнт парної кореляції $r_{x,y}$ двох величин *X* і *Y* визначається за формулою:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \qquad (1)$$

де X_i, Y_i – змінні, зв'язок між якими вивчається;
$\bar{X}, \bar{Y}$ – середні значення змінних, що вивчаються;
n – кількість спостережень.

Модуль коефіцієнта парної кореляції не перевищує одиниці, тобто $|r_{x,y}| \le 1$, що еквівалентне подвійній нерівності: $-1 \le r_{x,y} \le 1$. Якщо коефіцієнт кореляції статистично значущий, то між змінними існує лінійний зв'язок. Якщо $r_{x,y} > 0$, то зв'язок між змінними прямий, якщо – $r_{x,y} < 0$, то зв'язок обернений. Якщо коефіцієнт кореляції статистично незначущий, або дорівнює нулю, то дві випадкові величини X і *Y* не мають лінійної залежні, але можуть мати нелінійну.

Статистичну значущість коефіцієнта парної кореляції визначають за допомогою критерію Стьюдента.

Розрахункове значення критерію $t_{розр}$ обчислюють за формулою:

$$t_{розр} = \frac{r_{x,y}\sqrt{(n-2)}}{\sqrt{1 - r_{x,y}^{\ 2}}}, \qquad (2)$$

де $r_{x,y}$ – значення коефіцієнта парної кореляції;
n – кількість спостережень.

Розрахункове абсолютне значення критерію Стьюдента порівнюють з табличним (критичним) значенням. Якщо розрахункове значення критерію більше табличного, то прийнята нульова гіпотеза про рівність коефіцієнта

кореляції нулю відхиляється, тобто величина коефіцієнта кореляції є статистично значущою.

Табличне (критичне) значення критерію обчислюється за допомогою вбудованої функції MS Excel **СТЬЮДРАСПОБР** категорії **Статистические,** аргументами якої є рівень значущості ($\alpha = 0{,}05$) і число ступенів свободи (*f*=(*n*-2)).

Вбудовані формули масивів MS Excel для роботи з даними

Для спрощення складних розрахунків використовують вбудовані формули масивів MS Excel, виконуючи дії: **Вставка ⇒Функція⇒ Математические**. За формулою масиву одночасно може виконуватися декілька видів обчислень, результатом яких може бути одне значення, або масив значень. Формула масиву обробляє декілька наборів значень, які називаються аргументами масиву. Кожний аргумент масиву повинен включати однакове число стовпчиків і рядків. Аргументи відділяються ";".

Д о формул масиву відносяться:

СУММКВРАЗН(массив_х;массив_у) **– сума квадратів різниці відповідних значень двох масивів.**

СУММПРОИЗВ(массив_х;массив_у) – **сума добутків відповідних елементів двох масивів.**

Вбудовані функції MS Excel для визначення кореляційних характеристик

Вбудовані функції MS Excel для визначення кореляційних характеристик згруповані в категорії **Статистические**.

Коефіцієнт кореляції обчислюється однією із двох вбудованих функції: **КОРРЕЛ** або **ПИРСОН** аргументами яких є діапазони комірок з вхідними даними.

Програма виконання роботи

1. Завантажити табличний процесор MS Excel.
2. Ввести дані із завдання (табл.10): значення одного із показників (Y) і стовпчика "Доба спостережень"(X).
3. Обчислити середнє значення рядів спостережень за допомогою відповідної вбудованої функції.
4. Доповнити таблицю такими стовпчиками:

Загальний білок, г/л (Y_i)	Доба спостережень (X_i)	Y_c	$Y_i - Y_c$	X_c	$X_i - X_c$

де Y_c, X_c – середні значення показників.

5. Заповнити таблицю значеннями, використовуючи відносні і абсолютні посилання на адреси комірок з даними.
6. За допомогою формул масивів **СУММПРОИЗВ** і **СУММКВРАЗН** обчислити складові формули (1).

7. Обчислити коефіцієнт парної кореляції за формулою (1), використовуючи обчислені значення складових.
8. Перевірити правильність розрахунків, використовуючи вбудовану функцію **КОРРЕЛ** або **ПИРСОН**.
9. Обчислити розрахункове значення критерію Стьюдента за формулою (2).
10. Обчислити критичне (табличне) значення критерію Стьюдента за допомогою вбудованої функції **СТЬЮДРАСПОБР**.
11. У вільну комірку ввести формулу для порівняння розрахункового і критичного значень критерію за допомогою вбудованої логічної функції **ЕСЛИ** так, щоб у випадку підтвердження значущості коефіцієнта кореляції було виведено надпис “Значущий!”, а в протилежному разі – ”Не значущий”.
12. Зробити висновки щодо наявності і тісноти лінійного зв’язку між змінними.

Завдання на лабораторну роботу

1. Залежність показників крові свиней від часу

Динаміка біохімічних показників сироватки крові свиней після застосування модифікованого ізотонічного розчину							
№ п.п.	Доба спостережень (X_i)	Назва показників (Y_i)					
		загальний білок, г/л	альбумін, %	глобулін, %	α-глобулін, %	β-глобулін, %	γ- глобулін, %
1	5	52,9	48,4	51,58	24,21	15,12	12,25
2	14	53,7	47,6	52,4	24,86	15,2	12,34
3	30	56,9	47,4	52,6	21,56	16,4	14,64
4	60	63,4	48,5	51,48	19,88	16,8	14,8
5	90	69,1	41,55	58,45	21,16	16,1	21,19
6	150	72,1	41,51	58,49	20,9	16,2	21,3
7	270	76	41,18	58,82	19,85	16,82	22,16

Запитання для самоперевірки

1. Який зв’язок між змінними називають кореляційним?
2. Які характеристики застосовують для встановлення наявності кореляційного зв'язку між двома випадковими величинами?
3. Про що свідчать величина і знак коефіцієнта парної кореляції?

4. Як перевіряється статистична значущість коефіцієнта парної кореляції?
5. Які вбудовані функції застосовуються для визначення коефіцієнта кореляції?
6. Для чого застосовують формули масивів?
7. Що використовують в якості аргументів формул масивів?

Список літератури