Лабораторне заняття 5
КОРЕЛЯЦІЙНИЙ АНАЛІЗ ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ

Мета: лабораторне заняття передбачає засвоєння методів графічного (діаграми розсіювання) та математичного (коефіцієнт кореляції Пірсона та ранговий коефіцієнт кореляці Спірмена) визначення кореляційного зв'язку із застосуванням табличного процесору MS Excel 2010.

Обладнання та матеріали: ПЕОМ із встановленою ОС Microsoft Windows XP або Windows Seven, процесор таблиць Microsoft Excel 2007 або 2010, Інтернет браузер, методичні вказівки.

План:

1. Представлення кореляції у графічному вигляді та визначення та її типу.
2. Визначення параметричного коефіцієнту кореляції Пірсона.
   2.1. Обчислення довірчої імовірності коефіцієнту кореляції Пірсона.
3. Визначення непараметричного коефіцієнту рангової кореляції Спірмена.
   3.1. Обчислення довірчої імовірності коефіцієнту кореляції Спірмена.

Теоретичні відомості.

При вирішенні багатьох наукових та практичних завдань досліднику часто доводиться проводити статистичний аналіз зв'язків між факторними та результативними ознаками статистичної сукупності (причинно-наслідковий зв'язок) або визначати залежності паралельних змін декількох ознак цієї сукупності від якої-небудь третьої величини (від їх загальної причини). Для вивчення особливостей зв'язків подібного роду, визначення їх розмірів та напрямків, а також оцінки їх достовірності використовують методи визначення кореляційного зв'язку. Є два види проявів кількісних зв'язків між ознаками: а) функціональний зв'язок; б) кореляційний зв'язок.

Функціональний зв'язок – це така форма співвідношень між двома ознаками, коли кожному значенню однієї з них строго відповідає лише певне значення іншої (наприклад площа кола залежить від його радіусу). Функціональний зв'язок є характерним для фізико-математичних процесів.

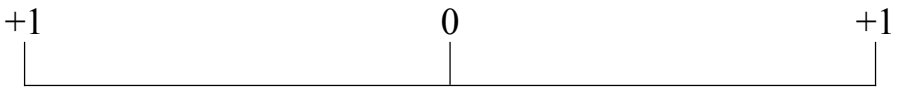
Кореляційний зв'язок – такий зв'язок, при якому кожному визначеному значенню однієї ознаки відповідає кілька значень іншої, взаємозв'язаної з нею ознаки. Інакше кажучи – це зв'язок, де вплив окремих факторів проявляється лише як тенденція (у середньому) при масовому спостереженні фактичних даних. Прикладами кореляції є: зв'язок між ростом і масою тіла людини, зв'язок між температурою тіла і частотою пульсу та ін. Кореляційний зв'язок характерний для медико-біологічних процесів.

Величина, що характеризує напрямок і силу зв'язку між ознаками це коефіцієнт кореляції, який одним числом дає уявлення про напрямок та силу зв'язків між ознаками (явищами); межі його коливань знаходяться в діапазоні

від 0 до ± 1.

Напрямок кореляційного зв'язку може бути прямим або зворотнім, а його сила поділяється на декілька градацій:

- сильна: ± 0,7 до ± 1
- середня: ± 0,3 до ± 0,699
- слабка: 0 до ± 0,299



Сильна позитивна кореляція | Кореляція відсутня | Сильна негативна кореляція

Найчастіше для визначення коефіцієнта кореляції застосовують метод квадратів (метод Пірсона) та ранговий метод (метод Спірмена).

При визначенні коефіцієнта кореляції важливо враховувати те, що:

- вимірювання зв'язку можливе лише в якісно однорідних сукупностях (наприклад, вимірювання зв'язку між ростом і вагою в сукупностях, однорідних за статтю та віком);
- розрахунок може проводитися з використанням як абсолютних так і похідних величин;
- для розрахунку коефіцієнта кореляції використовуються негруповані варіаційні ряди (це вимога стосується тільки обчислення коефіцієнта кореляції за методом квадратів);
- кількість спостережень менше 30.

Метод рангової кореляції (метод Спірмена) застосовують тоді, коли:

а) нема необхідності в точному встановленні сил зв'язку, а досить орієнтовних даних;

б) ознаки представлені не тільки кількісними, але й атрибутивними значеннями;

в) ряди розподілів ознак мають відкриті варіанти (наприклад, вік тварини до 1 року).

Метод квадратів (метод Пірсона) застосовують у випадку коли потрібно точне встановлення сили зв'язку між ознаками, які мають лише кількісне вираження.

Хід виконання роботи.

Завдання 1. У окуня озера Біле були виміряні довжина голови (*x*) та грудного плавця (*y*):

X	**66**	**61**	**67**	**73**	**51**	**59**	**48**	**47**	**58**	**44**	**41**	**54**	**52**	**47**	**51**	**45**
Y	**38**	**31**	**36**	**43**	**29**	**33**	**28**	**25**	**36**	**26**	**21**	**30**	**28**	**27**	**28**	**26**

Визначити кореляцію між X та Y у графічному та математичному вигляді.

У MS Excel для визначення типу та ступені кореляційного зв'язку в графічній та математичній формі потрібно виконати наступний й алгоритм дій:

☑ *Створення таблиці з даними.*

☑ *Побудова діаграми розсіювання*

☑ *Попередня оцінка ступеня та спрямованості кореляції за виглядом діаграми*

☑ *Формулювання H_0 та H_A гіпотез.*

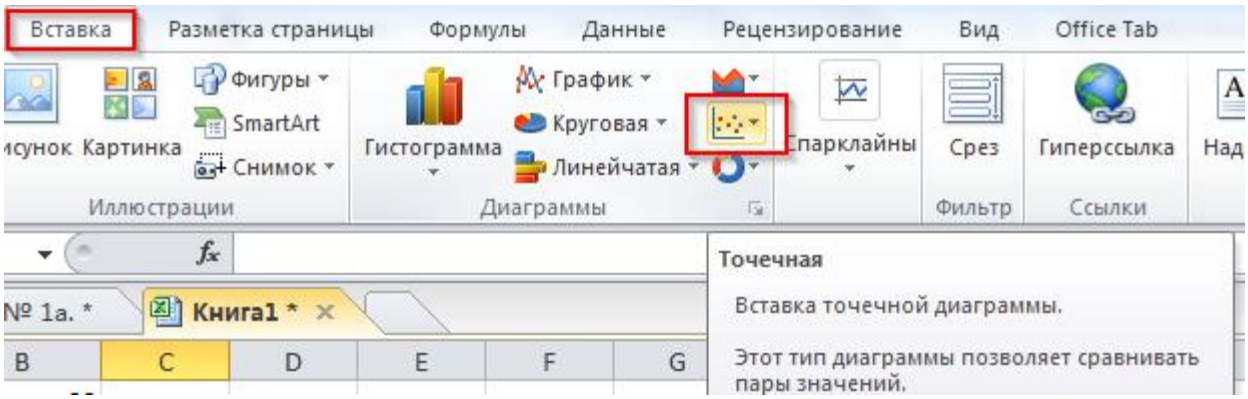
☑ *Обрахування обраного коефіцієнту кореляції.*

☑ *Обрахування довірчої ймовірності обрахованого коефіцієнту кореляції.*

☑ *Відхилення чи прийняття H_0 гіпотези на прийнятому рівні статистичної значущості за результатами інтерпретації результатів статистичного тесту.*

1.1. Для визначення кореляції у графічному вигляді використовують діаграми розсіювання. Для цього створіть таблицю із данними. У комірки A1 та B1 уведіть назви змінних «Довжина голови, см» та «Довжина грудного плавця, см», а у комірки A2:B17 – значення відповідних змінних.

1.2. Виділіть діапазон комірок A1:B17, та оберіть на стрічці Excel вкладку «**Вставка**», панель інструментів «**Диаграммы**», тип діаграми «**Точечная**».

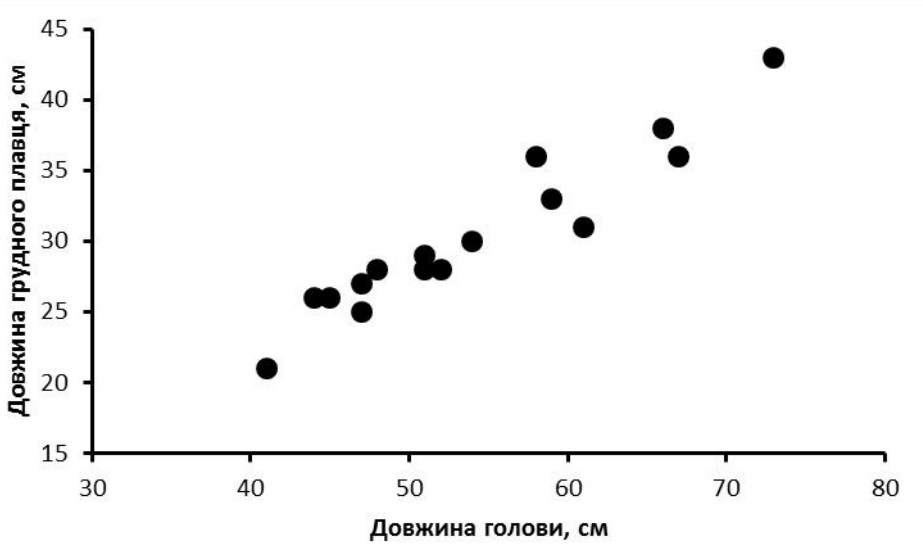


Відредагуйте отриманий графік, змінивши мінімальне значення вертикальної осі з «**авто**» на «**фиксированное**» із значенням, що дорівнює 15, та мінімальне значення горизонтальної осі з «**авто**» на «**фиксированное**» із значенням, що дорівнює 30.

1.3. Видаліть з графіку додаткові лінії, заголовок та легенду, додайте підписи до осей, змініть форму маркерів на круглу а колір – на чорний.

Щоб змінити параметри будь-якої осі графіку клацніть по ньому ПКМ та у контекстному меню оберіть пункт «Формат осі…».
Щоб змінити вигляд маркерів, клацніть по будь-якому з них ПКМ та у контекстному меню оберіть пункт «Формат ряда даннях…»

У результаті буде побудована точкова діаграма (діаграма розсіювання), на якій у графічній формі буде відображено наявність або відсутність кореляційного зв’язку.



1.4. **Інтерпретація результатів**. У даному випадку дійсно, спостерігається позитивна лінійна кореляція між довжиною голови та довжиною грудного плавця у окуня озера Біле, оскільки точки вишукались зліва направо та знизу догори. Отже, можна зробити висновок, що чим більша довжина голови у окуня, тим більша довжина його грудного плавця.

Тип кореляції можна визначити у залежності від виду діаграми розсіювання. Якщо точки вишиковуються в уявну лінію знизу догори (починаючи з лівого боку діаграми), то наявна позитивна лінійна кореляція. У випадку коли точки вишиковуються з гори донизу – то це свідчить про наявність негативної лінійної кореляції. Якщо ж кореляції між показниками немає, точки майже рівномірно розсіюються по усій площі діаграми.

1.5. Визначте коефіцієнт кореляції для даної вибірково сукупності. Для цього використайте функцію PEARSON(массив1;массив2), де массив1 – набір даних незалежної змінної, а массив2 – набір даних залежної змінної. У комірку A19 уведіть загальноприйняте позначення коефіцієнту кореляції (малу латинську літеру r) та знак «дорівнює», а у комірку B19 – формулу для обчислення коефіцієнту кореляції =PEARSON(A2:A18;B2:B18).

17	45	26
18	73	43
19	r=	0,96498775
20		

У результаті буде обрахований коефіцієнт кореляції, який у нашому

випадку буде дорівнювати 0,96. Отже, дані графічної оцінки кореляції підтверджено, і остаточний висновок проведеного кореляційного аналізу наступний: «***Між довжиною голови та довжиною грудного плавця окуня озера Біла наявна сильна позитивна кореляція***».

Коефіцієнти кореляції, як міри зв'язку між випадковими величинами, також є величинами випадковими та носять імовірнісний характер. Статистичні висновки про кореляційний зв'язок між величинами роблять не з генерального коефіцієнта кореляції ρ (значення якого є зазвичай невідомим), а за його вибірковим аналогом – r. Оскільки коефіцієнти кореляції r розраховується за значеннями змінних, які випадково потрапили у вибірку з генеральної сукупності, то й статистика r є величиною випадковою, яка потребує статистичної оцінки. Як правило, перевіряють нульову гіпотезу про відсутність кореляційного зв'язку між змінними в генеральної сукупності, тобто H_0: ρ = 0. Достовірність (довірча ймовірність) коефіцієнтів кореляції залежить від прийнятого рівня значущості α і обсягу вибірки n.

1.6. Для визначення довірчої імовірності коефіцієнта кореляції Пірсона спочатку потрібно впевнитись, що кореляція має лінійний характер. З побудованої у завданні 1 діаграми розсіювання добре видно, що кореляція лінійна. Далі, у комірці B20 розрахуйте емпіричний критерій $t_{ст}$ за допомогою формули =B19*КОРЕНЬ((СЧЁТ(A2:A18)-2)/(1-B19^2)) та отримайте значення $t_{ст}$. У комірку B21 уведіть рівень статистичної значущості (α), що дорівнює 0,05. Отримайте двобічне критичне значення t-критерію Стьюдента ($t_{кр}$) за допомогою функції =СТЬЮДРАСПОБР(). Для цього у комірку B22 уведіть вираз =СТЬЮДРАСПОБР(B21;СЧЁТ(B2:B18)-2).

19	r=	0,96498775
20	$t_{ст}$=	14,2487741
21	α=	0,05
22	$t_{кр}$=	2,13144955

1.7. Інтерпретація результатів та прийняття рішення. Оскільки $t_{ст} > t_{кр}$ (14,25 > 2,13), нульова гіпотеза відхиляється, тобто **із 95% довірчою імовірністю** ми можемо стверджувати що «***Між довжиною голови та довжиною грудного плавця окуня озера Біле наявна сильна позитивна кореляція»***

Завдання 2. Під час моніторингу для проекту MoorLIFE еколог зібрав дані про два два види рослин – верес (*Calluna vulgaris*) та чорницю (*Vaccinium myrtillus* L.) з району вересової пустки, яку було відновлено у 2003 році.

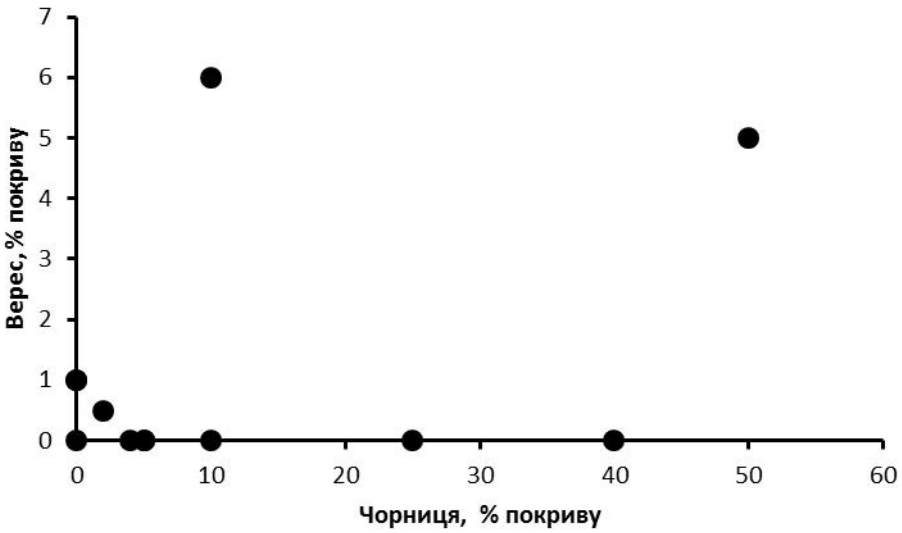
Ділянка	Чорниця, % покриву	Верес, % покриву
1	**5**	**0**
2	**40**	**0**
3	**50**	**5**
4	**5**	**0**
5	**10**	**0**
6	**25**	**0**
7	**0**	**1**
8	**4**	**0**
9	**0**	**0**
10	**0**	**1**
11	**10**	**6**
12	**2**	**0,5**

Чорниця (*Vaccinium myrtillus)*

Верес (*Calluna vulgaris)*

Оцінити наявність та значущість кореляції (із 99% довірчою ймовірністю, $\alpha = 0{,}01$) між зростанням цих двох видів рослин.

2.1. Перенесіть дані до електронної таблиці. У комірки A1, B1 та C1 уведіть назви змінних, а у комірки A2:C13 – варіанти. Для того, щоб зробити попередню оцінку наявності та характеру кореляційного зв'язку між змінними побудуйте діаграму розсіювання.

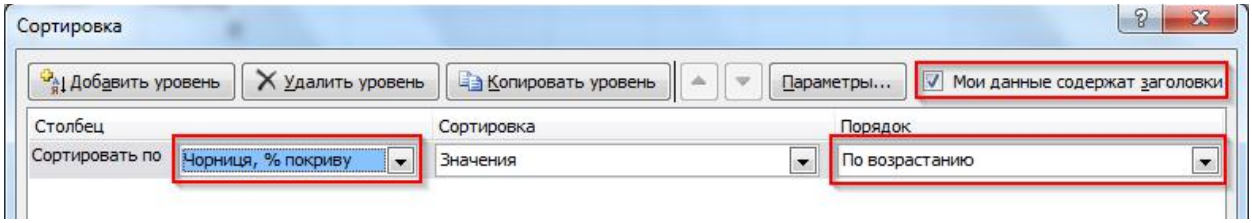


З діаграми видно, деякі точки вишиковуються у пряму (ліворуч знизу, а праворуч догори), що може вказувати на наявність слабкої негативної лінійної кореляції між двома видами рослин.

2.2. Наступний крок – розрахунок коефіцієнту кореляції. Оскільки дані виражені у відсотках, то розумним вибором буде використання рангового коефіцієнту кореляції Спірмена. Для цього спочатку потрібно проранжувати усі дані спостережень **<u>окремо для кожного виду рослин</u>**. На новому аркуші створіть таблицю за наступним зразком:

	A	B	C	D	E	F	G
1	Ділянка	Ранг	Чорниця, % покриву	Ділянка	Ранг	Верес, % покриву	
2	1		5	1		0	
3	2		40	2		0	
4	3		50	3		5	
5	4		5	4		0	
6	5		10	5		0	

2.3. Відсортуйте дані у таблиці за збільшенням за змінною «Чорниця, % покриву». Виділіть комірки A1:C13 та оберіть на стрічці вкладку «**Данные**», панель інструментів «**Сортировка и фильтр**», кнопка «**Сортировать**». У діалоговому вікні встановіть необхідні параметри.



Після того, як ваші дані будуть відсортовані, у комірках B2:B13 за допомогою автозаповнення проранжуйте варіанти, починаючи із 1 із кроком 1. Знову відсортуйте дані (коміки A1:C13) але за змінною «Ділянка».

Подібну послідовність дій виконайте для частини таблиці, що містить дані про площу покриття вересової пустки для вересу. Таким чином ви отримаєте ранжовані дані для кожного виду рослин, окремо на кожній ділянці.

2.4. Створіть таблицю зі значеннями рангів за наведеним зразком:

	A	B	C	D	E
1	Ранг				
2	чорниця	верес			
3	6	1			
4	11	2			
5	12	11			
6	7	3			

Коефіцієнт рангів r_s розраховується за формулою:

$$r_s = 1 - \frac{6 \times \sum_{i=1}^{n} (x_i - y_i)^2}{n \times (n^2 - 1)}$$

де: n – обсяг сукупності об’єктів; $(x_i - y_i)$ – різниця рангів i-го об’єкта. Коефіцієнт r_s приймає значення в інтервалі від -1 до +1.

2.5. Виконайте наступну послідовність дій. У комірки C1 та D1 уведіть текст «(x-y)» і «$(x-y)^2$» відповідно, а у комірках C2:C14 та D2:D14 розрахуйте відповідні значення. Для цього у комірку C2 уведіть формулу =A2-B2, а у комірку D2 – формулу =C2^2.

Додайте новий стовпчик A, та у комірку A15 уведіть текст «Суми», а у комірках B15:E15 обрахуйте суму усіх значень у відповідних стовпчиках за допомогою формули СУММ().

У комірку A16 уведіть текст «r_s=», а у комірку B16 – формулу для обчислення коефіцієнта кореляції Спірмена: =1-((6*E15)/СЧЁТ(B3:B14) *((СЧЁТ(B3:B14)^2)-1))). Отримане значення r_s дорівнює приблизно -0,15, що свідчить про наявність слабкої негативної кореляції між обома видами рослин.

2.6. Далі обрахуйте довірчу ймовірність для визначеного коефіцієнту кореляції. Для цього у комірку B17 уведіть формулу для обчислення t-критерію =B15/КОРЕНЬ(ABS(1-B15^2)/(СЧЁТ(B3:B14)-2)), а у комірку B18 – формулу для обчислення критичного значення t-критерію =СТЬЮДРАСПОБР(0,01/2;СЧЁТ(B3:B14)-2).

15	**Суми**	**78**	**78**	**0**	**328**
16	**r_s =**	**-0,14685**			
17	**$t_{емп}$ =**	**3,162538**			
18	**$t_{кр}$ =**	**3,581406**			

2.7. Отриманні значення $t_{емп}$ = 3,16 та $t_{кр}$ = 3,58 не дозволяють нам відхилити нульову гіпотезу на прийнятому рівні статистичної значущості (α = 0,01), оскільки $t_{емп} < t_{кр}$. Отже, остаточний висновок буде наступним: «**За результатами кореляційного аналізу площі покриття вересової пустки рослинами верес (*Calluna vulgaris*) та чорниця (*Vaccinium myrtillus* L.) зв'язку між зростанням цих рослин на визначений ділянці не виявлено**».

Завдання для самостійного виконання.

1. Для 10 півників леггорнів 15-денного віку були отримані наступні дані про масу їх тіла x (в г) і масу гребеня y (в мг):

x	**83**	**72**	**69**	**90**	**90**	**95**	**95**	**91**	**75**	**70**
y	**56**	**42**	**18**	**84**	**56**	**107**	**90**	**68**	**31**	**48**

Дайте відповідь на питання, чи наявний кореляційний зв'язок між масою тіла та масою гребеня півників, для чого побудуйте діаграму розсіювання та розрахуйте коефіцієнт кореляції Пірсона та його довірчу ймовірність при рівні статистичної значущості $\alpha = 0,05$.

2. При відновленні деяких земельних ділянок, переданих до заповідного фонду, екологами проводився моніторинг процесів сукцесії. У таблиці подано інформацію про домінуючи види рослин на одній із ділянок, умовно поділену на 26 квадратів.

Ідентифікаційний код квадрату	Deschampsia flexuosa, % покриття	Ранг 1	Agrostis spp, % покриття	Ранг 2
JP001	**30**		**15**	
JP003	**35**		**1**	
JP004	**15**		**5**	
JP005	**15**		**10**	
JP006	**15**		**2**	
JP007	**10**		**10**	
JP007b	**3**		**3**	
JP008	**5**		**1**	
JP009	**10**		**5**	
JP010	**25**		**0,25**	
JP011	**35**		**30**	
JP012	**50**		**20**	
JP013	**50**		**10**	
JP014	**30**		**15**	
JP015	**40**		**15**	
JP016	**50**		**5**	
JP017	**30**		**4**	
JP018	**15**		**10**	
JP019	**40**		**7**	
JP020b	**35**		**15**	
JP021	**30**		**30**	
JP022	**50**		**5**	
JP023	**15**		**10**	
JP024	**40**		**10**	
JP025	**40**		**12**	
JP026	**10**		**4**	

Побудуйте діаграму розсіювання. Проаналізуйте припущення про наявність або відстуність кореляційного зв'язку між зростанням цих двох видів рослин на визначеній ділянці. Визначте довірчу ймовірність для коефіцієнта кореляції ($\alpha = 0,05$).

Контрольні питання.

1. Дайте визначення поняттю «кореляція».
2. Поясніть, у чому полягає різниця між кореляційною і функціональною залежностями?
3. Розкрийте сутність позитивної і негативної кореляції. Які між ними відмінності?
4. Чому дорівнює коефіцієнт кореляції при повному кореляційному зв'язку?
5. Проаналізуйте, чим відрізняється r від ρ?