Практикум до теми 2

Завдання 1. Виробник хімічної речовини досліджує її чистоту. Чистота речовини вимірюється в кожній партії.

Дані про вимірювання чистоти речовини в 20 послідовних партіях

Номер партії	Чистота, %	Номер партії	Чистота, %
1	0,81	11	0,81
2	0,82	12	0,83
3	0,81	13	0,81
4	0,82	14	0,82
5	0,82	15	0,81
6	0,83	16	0,85
7	0,81	17	0,83
8	0,80	18	0,87
9	0,81	19	0,86
10	0,82	20	0,84

Використовуючи наведені дані, визначити, чи перебуває процес у стані статистичного контролю.

Завдання 2. Досліджується концентрація активної добавки у рідині для чищення. У таблиці наведено результати 30 послідовних замірів концентрації.

Результати виміру концентрації активної добавки

Спостереження	Концентрація, г/л	Спостереження	Концентрація, г/л
1	60,4	16	99,9
2	69,5	17	59,3
3	78,4	18	60,0
4	72,8	19	74,7
5	78,2	20	75,8
6	78,7	21	76,6
7	56,9	22	68,4
8	78,4	23	83,1
9	79,6	24	61,1
10	100,8	25	54,9
11	99,6	26	69,1
12	64,9	27	67,5
13	75,5	28	69,2
14	70,4	29	87,2
15	68,1	30	73,0

На основі наведених даних створити контрольну карту окремих спостережень. Оцінити стабільність процесу.

Завдання 3. З певного виробничого процесу взято 20 вибірок обсягом 3 вироби кожна. Результати вимірів певної ознаки якості x для кожного виробу наведено в таблиці.

Результати 20 вибіркових спостережень з виробничого процесу

№ вибірки	X_1	X_2	X_3
1	70,204	70,263	69,270
2	69,982	71,257	69,738
3	70,558	73,019	69,794
4	68,993	71,871	79,400
5	70,064	72,973	70,935
6	70,291	73,090	72,224
7	71,401	74,323	71,930
8	70,048	74,539	70,534
9	69,028	74,444	69,836
10	69,892	74,247	68,808
11	70,152	72,979	70,559
12	71,006	71,824	69,288
13	70,196	74,612	68,740
14	70,477	74,368	68,322
15	69,510	75,109	68,713
16	67,744	76,569	68,973
17	67,607	75,959	69,508
18	68,168	76,005	68,808
19	69,979	73,206	69,931
20	68,227	72,692	69,763

На основі наведених даних розрахувати характеристики контрольної карти. Навести формули для центральної лінії, контрольних меж і техніку розрахунку.

Навести вибіркові значення ознаки на контрольну карту. Обґрунтувати наявність особливих причин (якщо вони є).

Завдання 4. З певного виробничого процесу взято 20 вибірок обсягом 4 вироби кожна. Результати вимірів певної ознаки якості *x* для кожного виробу наведено в таблиці.

Результати вибіркових спостережень з виробничого процесу

№ вибірки	X_1	X_2	X_3	X_4
1	70,291	72,224	70,566	73,090
2	71,401	71,930	70,311	74,323
3	70,048	70,534	69,762	74,539
4	69,028	69,836	69,552	74,444
5	69,892	68,808	70,884	74,247
6	70,152	70,559	71,593	72,979
7	71,006	69,288	70,242	71,824
8	70,196	68,740	70,863	74,612
9	70,477	68,322	69,895	74,368
10	69,510	68,713	70,244	75,109
11	67,744	68,973	69,716	76,569
12	67,607	69,508	68,914	75,959
13	68,168	68,808	69,216	76,005
14	69,979	69,931	68,431	73,206
15	68,227	69,763	67,616	72,692
16	68,497	69,541	67,542	72,251
17	67,113	69,889	69,136	70,386
18	67,993	71,243	69,905	70,519
19	68,113	69,701	70,515	71,005
20	69,142	71,135	70,234	71,542

На основі наведених даних визначити тип контрольної карти, яку доцільно застосовувати для оцінки стабільності процесу. Обґрунтувати вибір.

Розрахувати характеристики контрольної карти. Навести формули для центральної лінії, контрольних меж і техніку розрахунку.

Навести вибіркові значення ознаки на контрольну карту. Оцінити наявність особливих причин (якщо вони є).

Завдання 5. З виробничого процес, що досліджується у завд. 4, взято 20 вибірок обсягом 4 вироби кожна. Результати виміру певної ознаки якості *x* наведено в таблиці.

Результати вибіркових спостережень з виробничого процесу

№ вибірки	X_1	X_2	X_3	X_4
1	70,291	72,224	70,566	73,090
2	71,401	71,930	70,311	74,323
3	70,048	70,534	69,762	74,539
4	69,028	69,836	69,552	74,444
5	69,892	68,808	70,884	74,247
6	70,152	70,559	71,593	72,979
7	71,006	69,288	70,242	71,824
8	70,196	68,740	70,863	74,612
9	70,477	68,322	69,895	74,368
10	69,510	68,713	70,244	75,109
11	67,744	68,973	69,716	76,569
12	67,607	69,508	68,914	75,959
13	68,168	68,808	69,216	76,005
14	69,979	69,931	68,431	73,206
15	68,227	69,763	67,616	72,692
16	68,497	69,541	67,542	72,251
17	67,113	69,889	69,136	70,386
18	67,993	71,243	69,905	70,519
19	68,113	69,701	70,515	71,005
20	69,142	71,135	70,234	71,542

Навести вибіркові значення ознаки на контрольну карту, побудовану за результатами завд. 4. Оцінити стабільність процесу. Обґрунтувати наявність особливих причин (якщо вони є).

Завдання 6. Товщина шару фарби, нанесеної на покриття виробником комплектуючих для побутової техніки, є важливою ознакою якості. У таблиці наведено результати 15 вибірок спостережень обсягом 4 одиниці кожна.

Результати вибіркових спостережень за процесом покриття комплектуючих фарбою

№ вибірки	Товщина шару фарби			
	X_1	X_2	X_3	X_4

1	43,4	58,4	43,3	53,3
2	46,7	51,0	44,1	44,1
3	44,8	41,2	47,4	47,4
4	51,3	47,7	51,6	51,3
5	49,2	45,7	42,5	42,5
6	45,5	50,6	54,5	54,3
7	48,4	51,0	57,5	57,5
8	50,1	53,0	54,8	64,8
9	53,7	56,0	52,1	52,6
10	45,6	47,2	59,6	59,6
11	50,0	48,0	51,5	51,5
12	51,2	55,9	58,4	58,4
13	46,9	50,0	57,5	53,5
14	44,9	47,9	51,3	54,3
15	46,2	53,4	53,3	55,3

На основі наведених даних розробити форму для збирання даних ознаки якості. Передбачити можливість подальшого моніторингу процесу.

Розрахувати характеристики контрольної карти. Навести формули для центральної лінії, контрольних меж і техніку розрахунку.

Навести вибіркові значення ознаки на контрольну карту. Оцінити стабільність процесу. Обґрунтувати наявність особливих причин (якщо вони є).

Питання для самоконтролю

1. Які методи становлять основу проведення контролю якості на підприємстві?
2. Які проблеми якості аналізують за допомогою діаграм Парето?
3. Який порядок складання схеми Ісікави?
4. На які типи поділяють контрольні карти?
5. Що таке контрольні карти Шугарта? Які їх види вам відомі?
6. З якою метою застосовують контрольні карти на виробництві?
7. Як будується діаграма розсіювання?
8. Як застосовується метод гістограм в роботі підприємства?
9. Де використовують контрольні листки?
10. Як заповнюють діаграму послідовності дій?

Тема 3. Статистичний контроль якості продукції

Якість продукції тлумачиться як «сукупність характеристик продукції (процесу, послуги), які стосуються її здатності задовольняти встановлені і передбачені потреби».

На практиці неможливо забезпечити перевірку характеристик кожної одиниці готової для реалізації продукції. Тому роблять **вибірковий контроль** та статистичний аналіз результатів, так званий – **статистичний приймальний контроль** (СПК).

Під готовою продукцією можна розуміти сировину, матеріал, напівфабрикати, тобто будь-які результати діяльності на попередніх ланках

створення продукції, все те, що являє собою певну сукупність і відносно чого потрібно прийняти рішення про відповідність визначеним вимогам.

Стандарти статистичного приймального контролю товарів

Для успішного застосування статистичних методів контролю якості продукції необхідні стандарти, доступні широкому колу інженерно-технічних працівників. Стандарти та статистичний приймальний контроль забезпечують можливість об'єктивно порівнювати рівні якості партій однотипної продукції як у часі, так і по різних підприємствах. Сформулюємо деякі вимоги до стандартів по статистичному контролі.

Стандарт повинен містити досить велику кількість планів, що мають різні оперативні характеристики. Це дозволить вибирати плани контролю з урахуванням особливостей виробництва й вимог споживача до якості продукції. Бажано, щоб у стандарті були зазначені різні типи планів: одноступінчасті, двоступінчасті, плани послідовного контролю й т.д.

При приймальному контролі якості партій готові продукції, сировини й напівфабрикатів важливі не стільки результати контролю окремої партії, скільки результат контролю послідовності партій виробів. Тому в стандартах повинна бути представлена система правил, що вказує, який конкретно план контролю з безлічі наявних у стандарті варто використовувати для контролю контрольної партії. Бажано враховувати результати контролю попередніх партій. При розкладаннях технологічного процесу, коли на контроль надходять партії продукції з підвищеним утримуванням дефектних виробів, споживач може наполягати на використанні планів контролю, що забезпечують задане значення межі середнього рівня вихідної якості. У цьому зв'язку стандарт повинен містити правила переходу з нормального контролю на посилений і навпаки. У тих випадках, коли якість продукції досить висока, можна використовувати плани полегшеного контролю.

Таким чином, основними елементами стандартів по приймальному контролі є:

1) таблиці планів вибіркового контролю, використовувані в умовах нормального ходу виробництва, а також планів для посиленого контролю в умовах розладнань і для полегшеного контролю при досягненні високої якості;
2) правила вибору планів з урахуванням особливостей контролю;
3) правила переходу з нормального контролю на посилений або полегшений і зворотний перехід при нормальному ході виробництва;
4) методи обчислення наступних оцінок показників якості контрольного процесу.

У ряді сучасних стандартів плани посиленого й полегшеного контролю пов'язані з таблицями планів нормального контролю простим співвідношенням. Наприклад, при одноступінчастому контролі перехід від нормального до посиленого зводиться до зменшення приймального *числа з* без зміни *обсягу вибірки n*. Якщо ж *з=0*, то збільшується тільки обсяг вибірки.

У тих випадках, коли відхилені партії піддаються суцільному контролю, план вибирають із урахуванням середнього числа виробів $n_{ср}$, контрольованих у партії при нормальному ході виробництва. Розрахунок $n_{ср}$ провадять у припущенні, що частка дефектних виробів у партіях постійна й дорівнює $\bar{q}$.

Залежно від гарантій, забезпечуваних планами приймального контролю, існують різні методи побудови планів:

- установлюють значення ризику постачальника α і ризику споживача β і висувають вимогу, щоб оперативна характеристика $P(q)$ пройшла приблизно через дві точки: q_0, α і q_m, β, де q_0 і q_m – відповідно прийнятний і бракувальний рівні якості. Цей план можна назвати компромісним, тому що він забезпечує захист інтересів як споживача, так і постачальника;
- вибирають одну крапку оперативної характеристики й приймають одне або кілька додаткових незалежних умов (наприклад, q_L або мінімальний обсяг контролю $n_{ср}$ при заданому значені q).

Перша система планів статистичного приймального контролю, що знайшла широке застосування в промисловості, була розроблена Доджем і Ромінгом. Плани цієї системи передбачають суцільний контроль виробів із забракованих партій і заміну дефектних виробів придатними. Запропоновано таблиці одноступінчастих і двоступінчастих планів вибіркового контролю двох видів. Одні плани забезпечують ризик споживача $\beta=0{,}1$ при вісьмох значеннях бракувального рівня якості q_m (0,5; 1,0; 2,0; 3,0; 4,0; 5,0; 7,0; 10,0 %), а інші гарантують значення q_L межі середнього рівня вихідної якості. Крім того, плани Доджа-Ромінга забезпечують мінімальний обсяг контрольний операцій при середній частці дефектних виробів у продукції, що надходять на контроль m. Значення m визначається за результатами попереднього контролю. Умова мінімуму обсягу при $q_m = q_0$ є другим критерієм вибору конкурентного плану.

Для приклада розглянемо таблицю одноступінчастих планів вибіркового контролю, що гарантують ризик споживача $\beta=0{,}1$ при бракувальному рівні якості $q_m=0{,}01$ і обсягів партій від 1000 до 4000 виробів (табл. 3.1).

Таблиця 3.1

Одноступінчасті плани вибіркового контролю

Обсяг партій	Параметри плану при середньому рівні вхідної якості q, %								
	0 ÷ 0,010			0,011 ÷ 0,10			0,11 ÷ 0,2		
	n	$з$	q_L, %	n	$з$	q_L, %	n	$з$	q_L, %
1001 ÷ 2000	220	0	0,15	220	0	0,15	220	0	0,15
2001 ÷ 3000	220	0	0,15	375	1	0,20	505	2	0,23
3001 ÷ 4000	220	0	0,15	380	1	0,20	510	2	0,24

Поряд з параметрами планів n і $з$ у таблиці дані значення межі середнього рівня вихідної якості q_L, забезпечувані планами контролю.

У таблиці 3.2. як приклад приводяться одноступінчаті плани вибіркового контролю, що забезпечують значення $q_L=2\%$ при обсягів партії $N=801 \div 1000$ виробів. Поряд з параметрами планів n і $з$ дані значення бракувального рівня

якості q_m, що відповідають ризику постачальника $\beta = 0{,}1$.

Таблиця 3.2

Одноступінчасті плани вибіркового контролю

Обсяг партій	Параметри плану при середньому рівні вхідної якості q, %								
	0 ÷ 0,04			0,05 ÷ 0,40			0,41 ÷ 0,80		
	n	*з*	q_L, %	*n*	*з*	q_L, %	*n*	*з*	q_L, %
801 ÷ 1000	18	0	12,0	40	1	9,6	40	1	9,6

ГОСТ 18242-72 «Статистический приемочный контроль по альтернативному признаку. Планы контроля» містить плани одноступінчастого й двоступінчастого приймального контролю. В основу стандарту покладене поняття прийнятного рівня (ПРУК) q_0, що розглядається як максимально припустима споживачем частка дефектних виробів у партії, виготовленої при нормальному ході виробництва. Імовірність α забракувати партію із часток дефектних виробів, рівної q_0, для планів стандарту мала й зменшується в міру зростання обсягу вибірки.

Зрівняємо оперативні характеристики планів посиленого, нормального й полегшеного контролю партії з $N=1000$ виробів при прийнятному рівні якості $q_0=0{,}09\%$.

Відповідно до стандарту, нашим умовам відповідають плани: посилений ($n=120$, $з=2$), нормальний ($n=80$, $з=2$) і полегшений ($n=20$, $з=0$). На рис.3.1. приводяться оперативні характеристики цих планів.

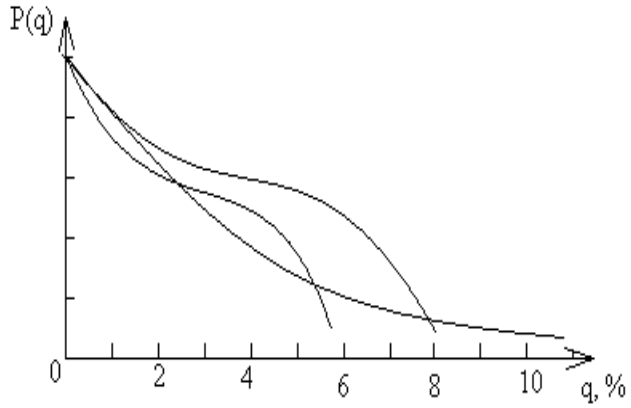


Рис. 3.1. Порівняння оперативних характеристик планів нормального, посиленого й полегшеного контролю партій виробів при N=1000 і q_0= 0,09%

Як бачимо, план посиленого контролю вимагає найбільших витрат на контроль ($n=120$), але забезпечує кращий захист інтересів споживача. Наприклад, при бракувальному рівні якості $q=5\%$ посилений план контролю забезпечує ризик споживача $p=0{,}05$, а нормальний план лише $\beta = 0{,}25$.

У рамках системи планів приймального контролю, передбачених ГОСТ 18242-72, двоступінчасті плани підібрані таким чином, що їхні оперативні характеристики близькі до одноступінчастого. Таким чином, при обраному значені прийнятого рівня якості q_0 одноступінчаті й двоступінчаті плани мають однакову ймовірність прийняття партії, але різняться по середньому числу виробів, контрольованих у партії. Середній обсяг вибірки для планів двоступінчастого контролю приблизно на 20% менше, ніж в одноступінчастих. Однак, як ми вже відзначали, двоступінчасті плани мають більше складну

організацію.

При контролі виробів по декількох ознаках стандартом рекомендується класифікувати дефекти на три класи: критичні, значні й малозначні – і для кожного класу вибирати свій план контролю.

В Україні застосовуються наступні стандарти по статистичному контролю:

ДСТУ 3021-95. Випробування та контроль якості продукції. Терміни та визначення.

ГОСТ 16504-81. Випробування і контроль якості продукції. Основні терміни і визначення.

ГОСТ 18321-73. Статистичний контроль якості. Методи випадкового відбору вибірок поштучної продукції.

ГОСТ 24297-87. Вхідний контроль продукції. Основні положення.

Інструкція. «Приймання продукції по кількості і якості».

Тести для перевірки засвоювання навчального матеріалу

1. Статистична оцінка рівня якості продукції здійснюється на стадії:
а) експлуатація або споживання продукції;
б) *виробництва продукції;*
в) маркетингу та вивчення ринку;
г) проектування і розроблення продукції.

2. Оцінювання рівня якості виготовлення продукції за показниками ефективності здійснюється на стадії:
а) експлуатації або споживання продукції;
б) *виробництва продукції;*
в) маркетингу та вивчення ринку;
г) проектування і розроблення продукції.

3. Оцінювання рівня виготовлення продукції за показниками ефективності здійснюється на стадії:
а) експлуатації або споживання продукції;
б) *виробництва продукції;*
в) маркетингу та вивчення ринку;
г) проектування і розроблення продукції.

4. Встановлення способу збирання і отримання інформації про рівень якості здійснюється на стадії:
а) *експлуатації або споживання продукції;*
б) виробництва продукції;
в) маркетингу та вивчення ринку;
г) проектування і розроблення продукції.

5. Статистичний контроль якості на підприємстві передбачає:
а) розподіл функцій і відповідальність за якість як між окремими

працівниками, так і цеховим керівником або майстром;

б) *використання контрольних карт;*

в) наявність відділу управління якістю у складі організаційної структури підприємства;

г) передбачає участь у роботах з якості всього персоналу підприємства.

Тема 4. Статистичні моделі (Моделювання якості)

Під моделюванням якості продукції розуміють «такі напрямки виконання функції загального управління, які визначають політику, цілі і відповідальність у сфері якості, а також здійснюють їх за допомогою таких засобів, як планування якості, оперативне управління якістю, забезпечення якості та поліпшення якості в межах системи якості» [1].

Основні принципи загального менеджменту якості (TQM) закладено у новій версії стандартів ISO 9001-2000. Відповідно до цих стандартів для ефективного досягнення результатів роботи організації слід управляти ресурсами і діяльністю як процесами. Система менеджменту якості орієнтована на безперервне вдосконалення і є метою організації. Цей підхід проілюстровано на рис. 4.1 [9].

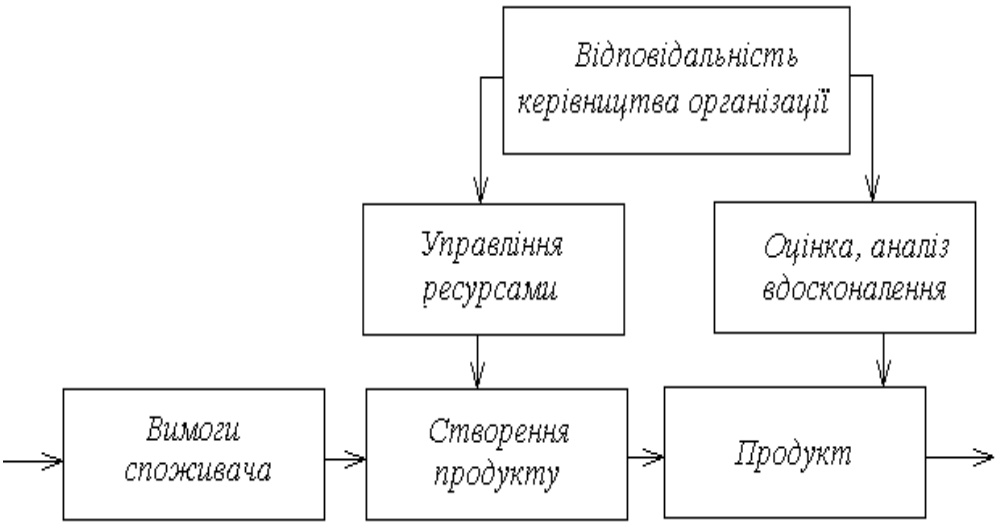


Рис. 4.1.

З цієї концептуальної моделі випливає висновок, що споживач відіграє основну роль у системі менеджменту якості, він формує вимоги до продукту, що створюється організацією, і визначає ступінь відповідності продукту своїм вимогам.

Побудова кореляційно-регресійних моделей

Для ефективного вдосконалення якості важливе значення має визначення наявності залежності певної ознаки якості від впливу різноманітних чинників, а також кількісна характеристика ступеня і виду цієї залежності.

Така залежність може бути **визначеною** (функціональною), але частіше на результативну ознаку різноманітні чинники впливають випадковим чином. В цьому випадку має місце стохастичний зв'язок, який визначається і

характеризується за допомогою статистичних методів.

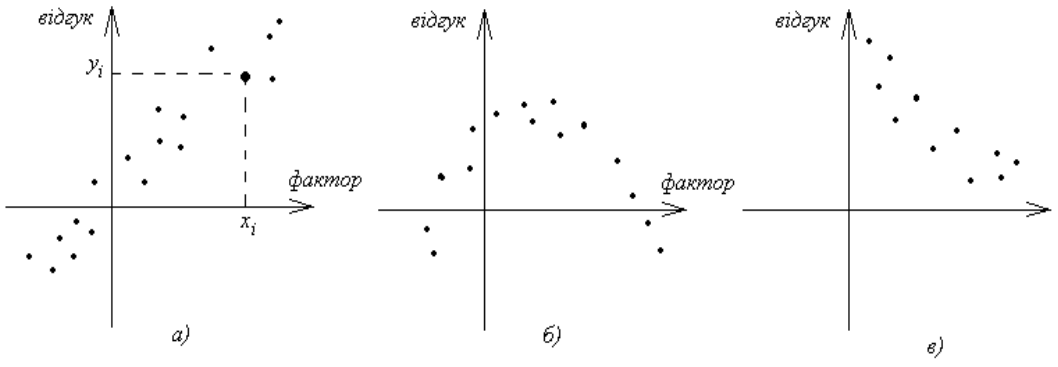
В математичній статистиці регресійний аналіз є одним з найбільш поширених методів обробки економічних даних при вивченні характеру залежностей між величинами. Термін «регресія» вперше був застосований для визначення залежності між двома величинами у останній треті XIX ст. англійським дослідником Френсісом Гальтоном. Він вивчав залежність зросту сина від зросту батька і встановив, що зріст сина завжди «регресує», тобто прагне повернутися до середньої величини.

В статистичній моделі припускається, що спостереження (або дані) є реалізацією (експериментом) деякої випадкової величини, що залежить від кількох інших не випадкових величин, значення яких відомі. Величини, або змінні, значення яких відомі і (або) визначаються умовами експерименту, називаються **факторами**. Величини, які вимірюються під час експерименту, називається **відгуками**. Наприклад, розглядається питання як якість продукції залежить від рівня освіти працівника, його статі, віку, місця проживання, сфери діяльності. Факторами в цій задачі будуть: рівень освіти (середня, середня спеціальна, вища), стать (чоловік чи жінка), вік (у роках), місце проживання (у місті чи сільській місцевості), галузь господарства (промисловість, сфера торгівлі і т. ін.). А відгуком буде якість кінцевого продукту.

В тому випадку, коли факторів два і більше задача називається **багатофакторною**, якщо ж фактор один – то експеримент називають **парним.**

Однофакторний регресійний аналіз

Найбільш просту ілюстрацію парних експериментів дає **графік**, або **діаграма розсіювання**. Будь яку пару «фактор-відгук» можна розглядати, як упорядковану пару чисел та відобразити її точкою на площині у прямокутній системі координат. По осі OX відкладають значення фактора x_i, $i=1\div n$, яке впливає на відгук, а по осі OY відкладають відповідні значення y_i, $i=1\div n$ відгуку. Множина цих точок створить графік розсіювання експерименту. Після цього вивчають взаємне розташування точок на графіку і за формою утвореної області точок роблять висновок про наявність залежності та її вид. Приклади діаграм наведено на рис. 4.2.



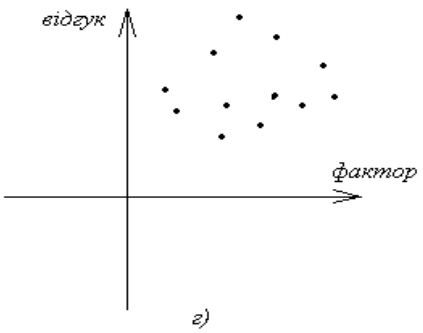


Рис. 4.2.

На діаграмі 4.2 а) видно, що із збільшенням фактору зростає і відгук, отже можемо припустити, що між ними існує пряма лінійна залежність. На діаграмі 4.2 б) можна припустити квадратичну залежність, а на 4.2 в) – експоненціальну. На діаграмі 4.2 г) залежність візуально не визначається. Таким чином, на діаграмах 4.2 а), 4.2 б), 4.2 в) точки $(x_i,\ y_i)$ близько розташовуються до деякої кривої, що описує аналітичну залежність $y = f(x)$. Така крива називається **лінією регресії** і може бути використана для цілей передбачення.

Нехай вибіркові спостереження складаються з n пар вимірювань $(x_i,\ y_i)$, $i = 1 \div n$. Мірою щільності зв'язку між фактором X і відгуком Y буде статистика

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \qquad (4.1)$$

Статистика r називається **коефіцієнтом кореляції.** Чисельник у формулі (4.1) є середній добуток відповідних відхилень фактора x та відгуку y від їх вибіркових середніх значень. Цю величину називають **вибірковою коваріацією** X та Y і позначають S_{xy}. Знаменник представляє собою добуток стандартних відхилень чисельних значень відповідно x та y, які зазвичай позначають відповідно S_x і S_y. Тоді формулу (4.1) можна представити у вигляді

$$r = \frac{S_{xy}}{S_x \cdot S_y} \qquad (4.2)$$

Величина коефіцієнта кореляції r не має одиниць вимірювання, вона не залежить від одиниць вимірювання X та Y та від вибору початку координат. Значення r належать інтервалу: $-1 \le r \le 1$ і що більше його абсолютне значення, то сильніший лінійний зв'язок між фактором і відгуком. Крім того коефіцієнт r вказує і на напрямок зв'язку: при позитивних значеннях r зв'язок прямий, при від'ємних – зв'язок зворотній. Обчислення r спрощується, якщо скористатись формулами:

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} \qquad (4.3)$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x} \qquad (4.4)$$

$$S_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y} \qquad (4.5)$$

Метод найменших квадратів (МНК) в лінійних і нелінійних моделях однофакторної регресії

Для визначення залежності відгуку від фактора будується **регресійна модель.** В моделі регресійного аналізу виходячи з вигляду діаграми розсіювання підбирають теоретичну **лінію регресії**, яка задається функцією $\hat{y} = f(x,\theta)$ від невідомого параметра θ. Ця функція називається **рівнянням регресії** або емпіричною формулою. Зазвичай функцію $\hat{y}$ обирають з числа простих за виглядом аналітичних функцій (табл. 4.1.) Зрозуміло, що строге співпадання значень $\hat{y}$ з експериментальними точками (x_i, y_i) спостерігається дуже рідко, але емпірична функція дозволяє прогнозувати значення реальної залежності для нетабличних значень x_i , тим самим «згладжувати» результати вимірювань величини відгуку. Треба зауважити, що множину точок (x_i, y_i) можна моделювати різними емпіричними формулами і обирати з них найкращу модель.

Таблиця 4.1

Вид регресії	№ п/п	Назва регресії	Рівняння регресії	Невідомий параметр θ
Лінійна	1	Лінійна	$\hat{y} = ax + b$	$\{a,b\}$
	2	Параболічна	$\hat{y} = ax^2 + bx + c$	$\{a,b,c\}$
	3	Степенева (геометрична)	$\hat{y} = ax^b$	$\{a,b\}$
	4	Гіперболічна	$\hat{y} = \frac{a}{x} + b$	$\{a,b\}$
	5	Показникова	$\hat{y} = a \cdot e^{bx}, a > 0$	$\{a,b\}$
	6	Логарифмічна	$\hat{y} = a \ln x + b$	$\{a,b\}$
	7	Дробово-лінійна	$\hat{y} = \frac{1}{ax + b}$	$\{a,b\}$
	8	Дробово-раціональна	$\hat{y} = \frac{x}{ax + b}$	$\{a,b\}$

Для оцінювання невідомого параметру рівняння регресії в математичній статистиці використовується декілька методів, наприклад метод максимальної вірогідності. Найбільш важливим серед всіх методів є метод найменших квадратів (МНК), який запропонував Лежандр у 1805 р. Основні досягнення цього методу пов’язані з іменами К. Гауса та О. Маркова. Цей метод відрізняється від методу

максимальної вірогідності, але має власні оптимальні властивості. Питання про доцільність використання МНК зводиться до питання про властивості оцінок найменших квадратів. Цей метод дає незсунені оцінки, дисперсія яких мінімальна у класі лінійних оцінок від спостережень.

Розглянемо задачу **лінійної регресії**. Припустимо, що є n вибіркових спостережень (x_i, y_i) і ми хочемо апроксимувати їх («підігнати») лінійним рівнянням регресії $\hat{y} = ax + b$, що задає пряму лінію. Основою МНК є вибір зі всієї множини прямих на площині такої, якій відповідає найменше значення суми квадратів відхилень від неї до точок графіка розсіювання, тобто

$$F = \sum_{i=1}^{n} \left[y_i - \hat{y}_i \right]^2 = \sum_{i=1}^{n} \left[y_i - \left(ax_i + b \right) \right]^2 \tag{4.6}$$

Величини x_i і y_i задачі фіксовані числа, невідомий параметр $\theta = \theta(a,b)$, тому функціонал F залежить від a і b:

$$F(a,b) = \sum_{i=1}^{n} \left[y_i - \hat{y}_i \right]^2 \to \min_{i=1 \div n} \tag{4.7}$$

Параметр a – **коефіцієнт регресії** показує на скільки одиниць у середньому зміниться y із зміною x на одиницю. Він має такі самі одиниці виміру, що і відгук. У разі прямого зв'язку a – величина додатня, а при зворотному – від'ємна. Параметр b – вільний член рівняння регресії, тобто це значення $\hat{y}$ при $x=0$.

Значення параметрів a і b, що мінімізують функціонал $F(a,b)$ (рис. 4.3), є рішенням системи рівнянь:

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases} \tag{4.8}$$

або

$$\begin{cases} \sum\limits_{i=1}^{n} y_i = a \sum\limits_{i=1}^{n} x_i + bn \\ \sum\limits_{i=1}^{n} x_i y_i = a \sum\limits_{i=1}^{n} x_i^2 + b \sum\limits_{i=1}^{n} x_i \end{cases} \tag{4.9}$$

Рівняння (4.9) називаються **нормальними рівняннями регресії**. Остаточні формули для обчислення значень параметрів a і b мають вигляд:

$$a = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i \right) \left(\sum\limits_{i=1}^{n} y_i \right)}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2} \tag{4.10}$$

$$b=\frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)\sum x_i y_i}{n\sum_{i=1}^{n} x_i^2-\left(\sum_{i=1}^{n} x_i\right)^2} \qquad (4.11)$$

Пряма $\hat{y}=ax+b$, де a та b є рішенням системи нормальних рівнянь (4.9) називається **регресією *Y* на *X*.** Аналогічно можна одержати рівняння **регресії *X* на *Y*:** $x=c\hat{y}+d$.

Взагалі, якщо треба завбачити значення y по даному значенню x, необхідно користуватися регресією Y на X; для знаходження x по даному y – регресією X на Y.

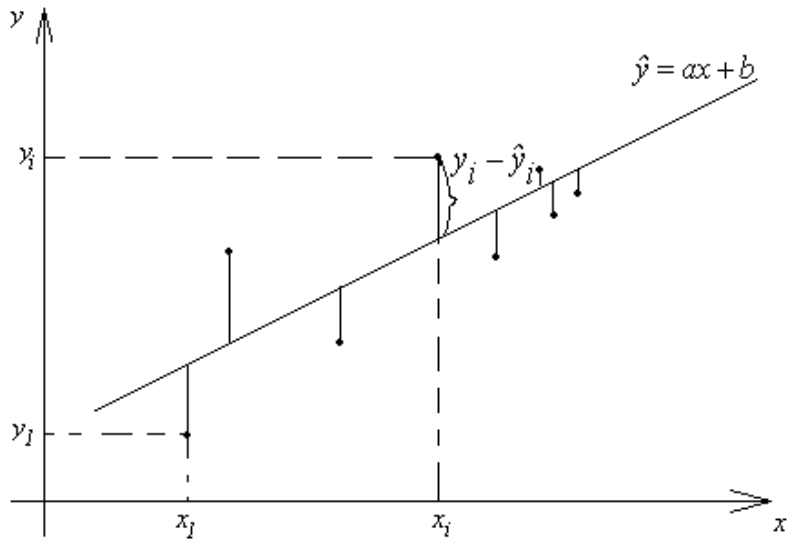


Рис.4.3. Лінійна модель регресії

Рівняння прямої регресії може бути записано й за допомогою коефіцієнту кореляції. Регресія Y на X:

$$y'=x'r, \qquad (4.12)$$

регресія X на Y:

$$x'=y'r$$

де x' та y' - стандартизовані змінні:

$$x'=\frac{x-\bar{x}}{S_x},\quad y'=\frac{y-\bar{y}}{S_y} \qquad (4.13)$$

МНК узагальнюється й на випадок, коли лінія регресії парного експерименту довільна, тобто не є прямою.

Параболічна регресія

Рівняння регресії має вигляд (табл. 4.1):

$$\hat{y}=ax^2+bx+c.$$

Треба підібрати параметр $\theta=\{a,b,c\}$ цієї моделі мінімізуючи функціонал

$$F(a,b,c)=\sum_{i=1}^{n}\left[y_i-\left(ax_i^2+bx_i+c\right)\right]^2.$$

Необхідна умова мінімуму функції $F(a,b,c)$:

$$\frac{\partial F}{\partial a}=0,\quad \frac{\partial F}{\partial b}=0,\quad \frac{\partial F}{\partial c}, \tag{4.14}$$

приймає вигляд:

$$\begin{cases} a\sum\limits_{i=1}^{n}x_i^4+b\sum\limits_{i=1}^{n}x_i^3+c\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}x_i^2y_i \\ a\sum\limits_{i=1}^{n}x_i^3+b\sum\limits_{i=1}^{n}x_i^2+c\sum\limits_{i=1}^{n}x_i=\sum\limits_{i=1}^{n}x_iy_i \\ a\sum\limits_{i=1}^{n}x_i^2+b\sum\limits_{i=1}^{n}x_i+c\cdot n=\sum\limits_{i=1}^{n}y_i \end{cases}. \tag{4.15}$$

Геометрична регресія

В цьому випадку рівняння регресії моделюється функцією

$$\hat{y}=ax^m. \tag{4.16}$$

Треба оцінити невідомий параметр $\theta=\{a,m\}$.

Нехай $x_i>0$ і $y_i>0$. Прологарифмуємо рівняння (4.16):

$$\ln\hat{y}=\ln a+m\ln x. \tag{4.17}$$

Введемо нову змінну $u=\ln x$, тоді з (4.17) матимемо:

$$\ln\hat{y}=\ln a+mu=g(u).$$

Позначимо:

$$m=A,\quad \ln a=B \tag{4.18}$$

Тоді рівняння (4.17) приймає вигляд:

$$\hat{g}(u,A,B)=Au+B, \tag{4.19}$$

Тобто задача звелась до задачі лінійної регресії. Практично для знаходження теоретичної функції (4.16) необхідно зробити наступне:

- за таблицю вибіркових значень (x_i, y_i) скласти нову таблицю, в якій прологарифмувати значення x_i і y_i;
- по новій таблиці знайти параметри A і B для функції $\hat{g}(u,A,B)$ за допомогою формул (4.10) і (4.11);
- використавши позначення (4.18), знайти значення невідомих параметрів a і m і підставити їх у (4.16).

Приклад 4.1.

Побудувати регресійну модель залежності відгуку від фактору для множини експериментальних даних, наведених в таблиці 4.2:

Таблиця 4.2.

x	1,1	1,7	2,4	3,0	3,7	4,5	5,1	5,8
y	0,3	0,6	1,1	1,7	2,3	3,0	3,8	4,6

Діаграма розсіювання представлена на рис. 4.4.

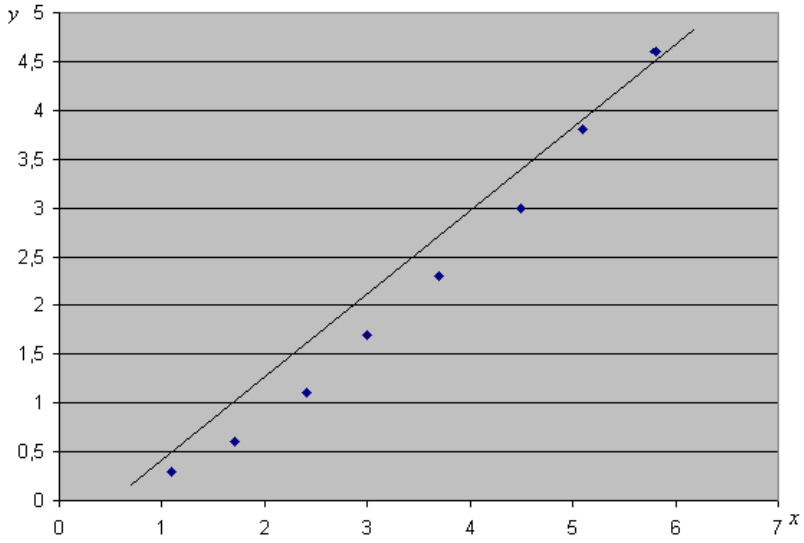


Рис. 4.4 Діаграма розсіювання для значень (x_i, y_i) табл.4.2

Для порівняння якості моделювання розглянемо паралельно два типа регресії: пряму $\hat{y}=ax+b$ і геометричну $\hat{y}=ax^m$. Після знаходження значень параметрів $\theta_1=\{a,m\}$ і $\theta_2=\{c,m\}$ можна порівняти якість регресійних моделей.

Значення параметрів *a* та *b* лінійної регресії знаходяться за формулами (4.10) і (4.11) з використанням обчислень в табл. 4.3.:

Таблиця 4.3

№	x	y	xy	x^2
1	1.1	0,3	0,33	1,21
2	1,7	0,6	1,02	2,89
3	2,4	1,1	2,64	5,76
4	3,0	1,7	5,10	9,00
5	3,7	2,3	8,51	13,69
6	4,5	3,0	13,5	20,25
7	5,1	3,8	19,38	26,01
8	5,8	4,6	26,68	33,64
Σ	27,3	17,4	77,16	112,45

Розв'язавши систему рівнянь, маємо:

$$a=0{,}921,\ \ b=-0{,}968.$$

Таким чином

$$\hat{y}=0{,}921x-0{,}968 \qquad (4.20)$$

Тепер знайдемо параметри *c* і *m* геометричної регресії. Для цього складаємо нову таблицю (табл. 4.4) з логарифмів значень *x* і *y*. Позначимо значення нових змінних відповідно *u* і *z*, тобто

$$u=\ln x,\ \ z=\ln y.$$

Таблиця 4.4

№	u	z	uz	u^2
1	0,095	-1,204	-0,114	0,009
2	0,531	-0,511	-0,271	0,282
3	0,875	0,095	0,083	0,766

4	1,099	0,531	0,584	1,208
5	1,308	0,833	1,090	1,711
6	1,504	1,099	1,653	2,262
7	1,629	1,335	2,175	2,654
8	1,758	1,526	2,683	3,091
Σ	8,799	3,704	7,883	11,983

Тепер шукаємо $A=m$ і $B=\ln c$ для лінійної регресії (4.19) за допомогою табл. 4.4. Отримуємо

$$A=1{,}656\,,\ B=-1{,}359\,.$$

Тому

$$m=1{,}656,\ \ c=e^{(-1{,}359)}=0{,}257\,.$$

Таким чином геометрична регресія має вигляд:

$$\hat{y}=0{,}257x^{1{,}656} \qquad (4.21)$$

Для порівняння якості моделей (4.20) і (4.21) у сенсі МНК обчислимо суму квадратів відхилень

$$\varepsilon=y_i-\hat{y}_i\,.$$

Для лінійної регресії $\sum\limits_{i=1}^{8}\varepsilon_1^2=0{,}2012$, а для геометричної – $\sum\limits_{i=1}^{8}\varepsilon_2^2=0{,}0425$.

Таким чином степенева регресійна модель краща за лінійну.

Множинний регресійний аналіз

У практичному аналізі якості здебільшого використовують методи багатофакторної регресії, тобто досліджується зв'язок між відгуком *y* і факторами $x_1,x_2,\ldots,x_n$ за допомогою функції регресії, що має невідомі параметри. Рівняння такої моделі можна записати у вигляді

$$\hat{y}=f(x_1,x_2,\ldots,x_p;\beta_1,\beta_2,\ldots,\beta_k)+e,$$

де $\beta_1,\beta_2,\ldots,\beta_k$ - невідомі параметри;

e – помилка апроксимації *У* за допомогою функції регресії.

Зокрема, якщо $k=p+1$ і

$$f(x_1,x_2,\ldots,x_p;\beta_0,\beta_1,\ldots,\beta_p)=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p,$$

модель називається **<u>лінійною множинною</u>**. Тоді

$$\hat{y}=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+e \qquad (4.22)$$

В цьому рівнянні деякі факторні змінні можуть бути функціями інших змінних або залежати одна від іншої. Наприклад,

$$\hat{y}=\beta_0+\beta_1\sin z_1+\beta_2\cos z_1+e$$

є модель множинної лінійної регресії з $x_1=\sin z_1$ і $x_2=\cos z_1$. В частковому випадку, якщо $x_i=x^i, i=1\div p$ отримується модель **<u>поліноміальної регресії</u>**.

$$\hat{y}=\beta_0+\beta_1x+\beta_2x^2+\ldots+\beta_px^p+e\,.$$

Нарешті, треба розуміти, що слово «лінійна» розуміє лінійність відносно параметрів, але не відносно факторних змінних. Так

$$\hat{y} = \beta_0 + sin(\beta x_1) + \beta_2 x_2$$

не є лінійною функцією параметрів.

В цьому розділі будемо розглядати модель множинної лінійної регресії у вигляді (4.22). В цій моделі коефіцієнти регресії $\beta_1, \beta_2, ..., \beta_n$ показують, на скільки одиниць зміниться відгук $\hat{y}$ у разі зміни відповідної факторної ознаки на одиницю за умови, що значення інших факторних ознак *x*, які входять до моделі, є фіксованими.

Параметри моделі оцінюються за вибіркою об'єму *n*, що отримана таким чином: фіксуємо факторні змінні $x_1, x_2, ..., x_p$ і спостерігаємо значення відгуку y_1; знову фіксуємо факторні змінні $x_1, x_2, ..., x_p$ і спостерігаємо значення відгуку y_2 і т.д. Таким чином отримуємо вибірку з *n* спостережень $(y_1; x_{11}; x_{21}, ..., x_{p1}), ..., (y_n; x_{1n}, ..., x_{pn})$. Для моделі множинної лінійної регресії маємо

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + e_i, \quad i = 1 \div n \qquad (4.23)$$

де $\beta_1, \beta_2, ..., \beta_k$ - невідомі параметри;

$e_1, ..., e_n$ - незалежні випадкові помилки, які розподілені за законом

$N(0, \sigma^2)$.

Крім того, будемо вважати, що $x_{1i}, x_{2i}, ..., x_{pi}, \ i = 1 \div n$ є фіксовані значення незалежних змінних $x_1, ..., x_p$.

МНК – оцінки $b_0, b_1, ... b_p$ параметрів $\beta_0, \beta_1, ..., \beta_p$ мінімізують суму квадратів відхилень.

$$F = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_p x_{pi})^2 \qquad (4.24)$$

Зазвичай їх називають **частинними коефіцієнтами регресії**. Іноді оцінку b_0 називають **вільним членом, константою або зсуненням по *y***. Оцінка рівняння множинної лінійної регресії або **площина найменших квадратів** може бути записаною у вигляді

$$\hat{y} = b_0 + b_1 x_1 + ... + b_p x_p. \qquad (4.25)$$

Модель регресії у загальному вигляді в матричній формі може бути записана як:

$$Y = X\beta + e,$$

де β - матриця невідомих параметрів розмірності $(p+1) \cdot 1$;

Y – матриця з *n* спостережень розмірності *n+1*;

e – матриця з *n* помилок розмірності *n+1*;

X^T – так звана **матриця плану** розмірності $n \cdot (p+1)$.

$$X^T = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{pmatrix}$$

Рівняння (4.23) можуть бути записані у вигляді

$$У = X^T\beta + e,$$

а вираз (4.24) приймає вид

$$F = (У - X^T\beta)^T(У - X^T\beta).$$

Матриця МНК – оцінок $B = (b_0, b_1, \dots b_p)^T$ невідомих параметрів знаходиться як розв'язок системи **нормальних рівнянь**

$$X \cdot X^T \cdot \beta = XУ$$

і має вигляд

$$B = (X \cdot X^T)^{-1}(X \cdot У) \qquad (4.26)$$

Для отримання оцінок коефіцієнтів багатофакторної моделі регресії доцільно використовувати пакети статистичних програм (ПСП). Наприклад, використання функції пакета EXCEL «ЛИНЕЙН», наведено в [16].

Приклад 4.2. Експериментально вивчалось октанове число бензину, що містить різні концентрації двох домішок А і В. Нехай y – октанове число, x_1 - відсоток домішки А і x_2 - відсоток домішки В. Вважається, що ефекти домішок А і В складаються, так що для опису залежності відгуку y від факторів x_1 і x_2 можна використовувати множинну лінійну регресію $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Кожна з двох факторних змінних приймала одне з чотирьох фіксованих значень, а значення y визначалось для кожної комбінації значень x_1 і x_2. Значення, що аналізувались, наведені в таблиці 4.5:

Таблиця 4.5

№ п/п	x_1	x_2	У	№ п/п	x_1	x_2	У
1	2	2	96,3	9	4	2	96,2
2	2	3	95,7	10	4	3	100,1
3	2	4	99,9	11	4	4	103,2
4	2	5	99,4	12	4	5	104,2
5	3	2	95,1	13	5	2	97,8
6	3	3	97,8	14	5	3	102,2
7	3	4	99,3	15	5	4	104,7
8	3	5	104,9	16	5	5	108,8

Побудувати оцінку множинної лінійної регресії.

Розв'язання. За допомогою програми множинної регресії з ПСП отримані МНК - оцінки $b_0 = 84{,}553, b_1 = 1{,}833, b_2 = 2{,}683$. Таким чином, площина найменших квадратів має рівняння:

$$\hat{y} = 84{,}553 + 1{,}833x_1 + 2{,}683x_2 .$$

У вихідних даних програм множинної лінійної регресії зазвичай містяться ще декілька величин. Наприклад – **залишкова сума квадратів** (або **помилок**) SF_R, що представляє собою значення F (4.24), яке отримується при підстановці МНК - оцінок замість невідомих параметрів, тобто

$$SF_R = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - \ldots - b_p x_{pi})^2 \qquad (4.27)$$

Якщо цю величину поділити на **число степенів свободи залишків або помилок**

$$\upsilon_R = n - p - 1, \qquad (4.28)$$

то отримується незсунена оцінка дисперсії помилок, яка називається **залишковим середнім квадратом помилки** і позначається

$$MF_R = \frac{SF_R}{\upsilon_R} \qquad (4.29)$$

Квадратний корінь з MF_R називається **стандартною похибкою оцінки**

$$S = \sqrt{MF_R} \qquad (4.30)$$

Для приклада 4.2 вихідні дані мають вигляд: $SF_R = 25{,}182$; $\upsilon_R = 16 - 2 - 1 = 13$; $MF_R = 1{,}937$; $S = 1{,}392$.

Зауваження. У регресійній моделі (4.22) коефіцієнт β_i вимірює степінь зміни відгуку y в залежності від фактора x_i, коли значення інших факторів фіксовані. На практиці часто виникає ситуація, коли факторні змінні вимірюються в різних одиницях (на відміну від ситуації приклада 4.2) і коефіцієнти не можливо порівняти за величиною. Ця проблема вирішується завдяки застосування **стандартизованих факторних змінних**

$$z_j = \frac{x_j}{S_j}, \quad j = 1 \div p, \qquad (4.31)$$

де

$$S_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ji} - \overline{x_j}\right)^2 .$$

Тоді модель множинної лінійної регресії в термінах z_j буде задаватись рівняннями, аналогічними (4.23):

$$y_i = \gamma_0 + \gamma_1 z_{1i} + \ldots + \gamma_p z_{pi} + e_i, \; i = 1 \div n, \qquad (4.32)$$

де γ_k - невідомі параметри, $k = 0 \div p$;

e_i - незалежні випадкові помилки, які розподілені за законом $N(0, \sigma^2)$.

Перевага стандартизації (4.31) полягає в тому, що $\gamma_1, \ldots, \gamma_p$ вимірюють степінь зміни в одній шкалі, що дозволяє робити висновки про вплив факторних змінних на відгук (згадаємо наведений на початку розділу приклад аналізу якості продукції в залежності від таких факторів як рівень освіти (x_1), вік (x_2), стать (x_3), місце проживання (x_4) і т.д. працівника).

Після знаходження оцінок невідомих параметрів (4.25) визначають їх значущість. Для цього при заданому рівні значущості α найчастіше використовують t- критерій Стьюдента для двосторонньої області з υ_R степенями вільності і обчислюють статистичну t:

$$H_0 : b_j = 0; \qquad H_1 : b_j \neq 0; \qquad t = \frac{|b_j|}{\sqrt{MF_R \cdot c_{jj}}},$$

де c_{jj} - елемент головної діагоналі оберненої матриці $\left(X \cdot X^T\right)^{-1}$.

Критичне значення $t_{кр.}$ знаходять за таблицями критичних точок t-розподілу Стьюдента для υ_R степенів вільності при заданому рівні значущості α (додаток 2). Якщо гіпотеза $H_0 : b_j = 0$ приймається для всіх j, то на цьому регресійний аналіз закінчується. Для значущих факторів рівняння регресії будують інтервальні оцінки коефіцієнтів. Довірчий інтервал для β_j:

$$b_j - t_{кр}\sqrt{MF_R \cdot c_{jj}} \leq \beta_j \leq b_j + t_{кр}\sqrt{MF_R \cdot c_{jj}} \qquad (4.32)$$

Деякі програми з ПСП друкують значення статистики t для кожного коефіцієнта b_j. Іноді це значення називають величиною ***t* -включення.**

Зауважимо, що гіпотеза $H_0 : b_j = 0, j = 1 \div p$ може розглядатися як гіпотеза про те, що факторна змінна x_j не покращує передбачення відгуку *y* в порівнянні з передбаченням, що отримане за допомогою регресії *y* за *(p-1)* іншими змінними. Коефіцієнт b_0 можна вважати передбаченим значенням відгуку *y* при $x_1 = \ldots = x_p = 0$.

Важливим моментом побудови шуканої залежності (4.22) є відбір факторів x_j, що суттєво впливають на відгук *y*. Відомо достатньо методів відбору. Умовно ці методи можна поділити на два класи: **формальні** (семантичні) і **змістовні.**

Формальні методи базують на ідеї перебору різних рівнянь до моменту досягнення деякого критерію (наприклад описаного вище t - критерію Стьюдента), що при заданому рівні значущості α характеризує значущість вкладу змінної *y* регресію.

Змістовні методи виконують досягнення цілей моделювання, при цьому розрізняють:

- **фізичні моделі**, що описують особливості аналізованих процесів. Побудова таких моделей рідкий випадок, так як принципово неможливо врахувати всі причинно-наслідкові зв'язки та їх взаємодію;
- **моделі для управління процесом**. Пропонується можливим для будь-якого y_i знайти такі x_{ij} (керуючі впливи), що задавши їх у моделі отримуємо значення y_i, що вимагається;
- **модель для передбачення**. Дає можливість за відомими факторами x_{ij} визначати відгук y_i, що прогнозується.

Спостереження за впливом факторів на відгук можна уявити у вигляді

множини точок деякого фазового простору. Тоді фізичні моделі, моделі управління і передбачення – це можливі проекції відгуку на різні площини. Тому на практиці ці моделі не співпадають.

Зазвичай на практиці більш необхідні моделі, що описують змістовні впливи факторів. Спроба поєднання формальних і змістовних критеріїв – це типова багатокритеріальна задача, розв'язок якої зазвичай не є однозначним.

Степінь лінійного зв'язку між факторами і відгуком і факторами визначається за допомогою часткових і множинних коефіцієнтів кореляції (Нелінійні зв'язки не визначаються. В цьому випадку для оцінки використовують кореляційне відношення).

Оцінкою парного коефіцієнта кореляції є статистика

$$r_{jk} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{nS_xS_y}, \quad j,k = 1 \div p.$$

Оцінкою часткового коефіцієнта кореляції l-го порядку $(l = p-2)$ є **вибірковий частковий коефіцієнт кореляції l-го порядку**. Він обчислюється на базі кореляційної матриці, що складена з вибіркових парних коефіцієнтів кореляції:

$$Q_p = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ & & r_{jk} & \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}, \qquad (4.33)$$

за формулою

$$r_{jk\cdot 1,2,\dots,p} = \frac{q_{jk}}{\left(q_{jj}\cdot q_{kk}\right)^{1/2}} \qquad (4.34)$$

де q_{jk}, q_{jj}, q_{kk} - алгебраїчні доповнення до відповідних елементів матриці (4.33).

Зв'язок одного з факторів (наприклад x_j) з усіма іншими можна оцінити за допомогою **вибіркового множинного коефіцієнта кореляції**

$$R_{j\cdot 1,2,\dots,p} = \sqrt{1 - \frac{\left|Q_p\right|}{q_{jj}}}, \qquad (4.35)$$

де $\left|Q_p\right|$ - визначник кореляційної матриці (4.33),

q_{jj} - алгебраїчне доповнення до елемента r_{jj}.

Квадрат вибіркового множинного коефіцієнта кореляції

$$D = R^2_{j\cdot 1,2,\dots,p}$$

називається **вибірковим множинним коефіцієнтом детермінації**. Коефіцієнти $R_{j\cdot 1,2,\dots,p}$ і D - величини додатні і приймають значення від 0 до 1.

Розглянемо частковий випадок двофакторної (x_1, x_2) моделі

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e \qquad (4.36)$$

За міру лінійного зв'язку між відгуком y і факторами x_1, x_2 використовують статистику множинний коефіцієнт кореляції

$$R_{y\cdot 1,2} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2 \cdot r_{y1} \cdot r_{y2} \cdot r_{12}}{1 - r_{12}^2}} \qquad (4.37)$$

де r_{ij} - парні коефіцієнти кореляції.

Можна розглядати також $R_{1\cdot y,2}$ і $R_{2\cdot y,1}$. Якщо $R_{y\cdot 1,2} = 1$, то відгук y однозначно визначається функціональною залежністю (4.36).

Для встановлення впливу фактору x_1 (або x_2) на зміну відгуку y використовують частковий коефіцієнт кореляції

$$r_{yx_1\cdot x_2} = \frac{r_{y1} - r_{y2} \cdot r_{12}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{y2}^2\right)}} \qquad (4.38)$$

Аналогічно визначається і $r_{yx_2\cdot x_1}$.

Приклад 4.3. Задана кореляційна матриця вибіркових парних коефіцієнтів кореляції $Q_3 = \begin{pmatrix} 1 & -0{,}85 & -0{,}62 & -0{,}21 \\ -0{,}85 & 1 & -0{,}53 & 0{,}34 \\ -0{,}62 & 0{,}53 & 1 & 0{,}46 \\ -0{,}21 & 0{,}34 & 0{,}46 & 1 \end{pmatrix}$.

Обчислити оцінки часткових і множинних коефіцієнтів кореляції.

Розв'язання. З формули (4.34) у випадку трьох факторів маємо $r_{12\cdot 3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$. Тому $r_{12\cdot 3} = \frac{0{,}85 - (-0{,}62) \cdot (-0{,}53)}{\sqrt{1 - (-0{,}62)^2}\sqrt{1 - (-0{,}53)^2}} = -0{,}78$.

Аналогічно $r_{23\cdot 1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{\left(1 - r_{21}^2\right)\left(1 - r_{31}^2\right)}}$ і $r_{23\cdot 1} = \frac{-0{,}53 - (-0{,}85) \cdot (-0{,}62)}{\sqrt{1 - (-0{,}85)^2}\sqrt{1 - (-0{,}62)^2}} = 0{,}007$

Обчислимо далі оцінку множинного коефіцієнта кореляції і детермінації першого фактору x_1 з усіма іншими x_2 і x_3. Будемо користуватися формулою (4.35) для випадку $p = 3$: $R_{1\cdot 2,3} = \sqrt{1 - \frac{|Q_3|}{q_{11}}}$.

Визначник кореляційної матриці $|Q_3| = 0.11$. Для елемента r_{11} алгебраїчне доповнення $q_{11} = (-1)^{1+1}\begin{vmatrix} 1 & -0.53 & 0.34 \\ 0.53 & 1 & 0.46 \\ 0.34 & 0.46 & 1 \end{vmatrix} = 0.56$.

Звідки $R_{1\cdot 2,3} = 0.89,\ D = R_{1\cdot 2,3}^2 = 0.8.$

Задачі до теми 4

4.1. На підприємстві А вивчається залежність ознаки якості продукції y від чинника x. Результати спостережень представлені в таблиці:

№ одиниці продукції	1	2	3	4	5	6	7	8	9	10
Значення чинника	5	9	14	17	23	31	35	42	46	50
Значення якості	9,14	9,50	9,33	9,34	9,31	9,26	9,22	9,30	9,15	9,08

а) Зобразіть графік розсіювання для даного парного експерименту;
б) чи існує залежність між фактором і відгуком? Якщо це так, то дайте словесний опис цієї залежності;
в) якщо чинник має значення 40, то чому приблизно буде дорівнювати ознака якості продукції?
г) обчисліть коефіцієнт кореляції r.

4.2. Намалюйте графік розсіювання та обчисліть коефіцієнт кореляції r для кожної з наступних множин спостережень. Порівняйте коефіцієнти кореляції та графіки розсіювання.

x_i	1	2	3	4	5
y_i	1	2	3	4	5

а)

x_i	1	2	3	4	5
y_i	2	1	3	3	6

б)

x_i	1	2	3	4	5
y_i	4	3	1	2	0

в)

x_i	1	2	3	4	5
y_i	1	-1	1	-1	1

г)

4.3. Задана множина даних:

x_i	1	2	3	4	5
y_i	2	1	3	4	6

а) Намалюйте графік розсіювання;
б) обчисліть коефіцієнт кореляції r;
в) за допомогою лінійного перетворення $\begin{cases} x' = 2x+3 \\ y' = 3y-1 \end{cases}$ складіть множину «умовних» даних;
г) намалюйте графік розсіювання та обчисліть r для множини «умовних» даних;
д) порівняйте графіки розсіювання та значення r у пунктах а), б) і г).

4.4. Нехай над змінними x і y зроблені лінійні перетворення: $\begin{cases} x' = ax+b \\ y' = cx+d \end{cases}$

Покажіть, що $r_{xy} = r_{x'y'}$.

4.5. Залежність якості продукції підприємства від чинника задається даними:

x \ y	1	2	3	4	5	m_x
1	2					2
2	1	1	2			4
3		2	2	1		5
4				2	2	4
m_y	3	3	4	3	2	15

Знайдіть коефіцієнт лінійної кореляції між x і y.

4.6. Значення ознак ξ та η у членів деякої сукупності даються наступними кореляційними таблицями:

I.

ξ \ η	0	1	2	3	4	5	6	7	m_ξ
25	2	1							3
35		5	3						8
45			4	2	4				10
55					2	3	1	5	11
65							6	2	8
m_η	2	6	7	2	6	3	7	7	40

II.

ξ \ η	15-20	20-25	25-30	30-35	35-40	40-45	m_ξ
210-220			1	1			2
220-230		1	4	3	2		10
230-240	2	7	8	9	7	3	36
240-250		3	4	3	3		13
250-260			3	2	2		7
260-270			2	2			4
m_η	2	11	22	20	14	3	72

Знайти у кожному випадку коефіцієнт кореляції та написати рівняння прямих регресії ξ на η та η на ξ.

4.7. Керівництво банку отримало результати тестування (у балах) 10 банківських службовців. Перший тест перевіряє пам'ять (ξ), другий – здатність до логічного мислення (η). Оцінка за тест по перевірці пам’яті (ξ): 5, 8, 7, 10, 4, 7, 9, 6, 8, 6. Оцінка за тест по перевірці здатності до логічного мислення (η): 7, 9, 6, 9, 6, 7, 10, 7, 6, 8.

а) Намалюйте графік розсіювання;

б) знайдіть коефіцієнт кореляції між ξ та η;

в) методом найменших квадратів знайдіть регресію ξ на η та регресію η на ξ;

г) намалюйте лінії регресії на графіку розсіювання. Які координати точки перетину ліній регресії.

4.8. В банку N встановлені такі відсоткові ставки за кредити, що видаються в

залежності від кількості днів, на які видаються кредити:

Кількість діб	0^*	4	10	15	21	29	36	51	max
Відсоткова ставка за кредит в %	66,7	71,0	76,3	80,6	85,7	92,9	99,4	113,6	max

Примітка: 0^* означає, що кредит видається на ніч; max – це максимальна тривалість відстрочки виплат кредиту – 68 діб (при цьому % - ставка складає 125,1 %).

Вважається, що відсоткова ставка за виданий кредит (η) лінійно залежить від кількості діб (ξ), по закінченню яких необхідно з відсотками повернути банківський кредит, знайти параметри залежності $\eta = a\xi + b$ за МНК.

4.9. Вартість випуску фірмою промислових товарів (η) за семирічну (роки ξ) характеризується такими даними:

ξ	1	2	3	4	5	6	7
η млн. грв.	0,5	0,5	1,5	3,5	6,5	10,5	15,5

Вирівняти залежність η від ξ по параболі $\eta = a\xi^2 + b\xi + c$.

4.10. Фізичний обсяг грошового обороту (η) комерційного банку (у цінах за кредит, що зіставляються) за 12 років (ξ) наводиться у наступній таблиці:

ξ	1	2	3	4	5	6	7	8	9	10	11	12
η у %	100	113	121	148	183	194	219	260	277	304	338	352

Вирівняти за МНК ці дані по прямій $\eta = a\xi + b$.

4.11. Нехай залежність прибутку фірми (у млн. грн.) по роках характеризується такими даними:

Роки ξ	1989	1990	1991	1992	1993	1994
Прибуток η	-2	-3	-3	-1	3	7

Вважаючи, що $\eta = a\xi^2 + b\xi + c$, знайти параметри цієї залежності, користуючись МНК.

4.12. Дослідження залежності тривалості (t) часу обслуговування електронних рахунків клієнтів від кількості (η) електронних рахунків однакового ступеню складності дало такі результати:

η	2	3	4	5	6	7	8	9	10
t (сек)	12	35	75	130	210	315	445	600	800

Вважаючи, що $t = an^{\alpha}$, знайти за МНК параметри a і α.

4.13. Експериментально вивчались терміни y схоплення бетонної суміші в залежності від концентрації двох домішок x_1, x_2. Припускається, що ефекти домішок додаються, так що для опису залежності відгуку y від факторів x_1, x_2 використовується модель множинної лінійної регресії $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$.

За допомогою програми множинної регресії з ПСП отримати МНК – оцінки коефіцієнтів регресії і стандартну похибку оцінки, якщо результати експериментів представлені в таблиці.

№ н/п	x_1(%)	x_2(%)	y(год)
1	1,1	0,11	2,5
2	1,2	0,12	2,4
3	1,3	0,13	2,3
4	1,4	0,14	2,2
5	1,5	0,15	2,1
6	1,6	0,16	2,0
7	1,7	0,17	2,1
8	1,8	0,18	2,2
9	1,9	0,19	2,3
10	2,0	0,2	2,5

4.14. Експериментально оцінювалась товщина карбонізаційного шару бетону y в залежності від віку бетону (x_1), вологості середи (x_2) і пористості бетону (x_3). Обчислити оцінки часткових і множинних коефіцієнтів кореляції, якщо результати експериментів пока0зані в таблиці

№ н/п	x_1(роки)	x_2(%)	x_3	y(мм)
1	0,5	100	1,0	0
2	1,0	95	1,2	0
3	1,5	80	1,4	1,2
4	2,0	85	1,6	1,6
5	2,5	80	1,8	2,0
6	3,0	75	2,0	2,4
7	3,5	70	2,2	2,8
8	4,0	65	2,4	3,2
9	5,0	60	2,6	3,6
10	5,5	50	2,8	5,0

Практикум до теми 4

Завдання №1.

За результатами перевірки якості продукції підприємства виявлено залежність її від певного фактору.

№ вар-ту	Спостереження										x_0
1	x	10	15	20	25	30	35	40			32
	y	14	12	10	8,5	7	5,5	4,5			
2	x	74	88	102	116	130	144	158	172		100
	y	325	355	385	415	445	475	505	535		
3	x	46	38	34	32	28	24	22	18	16	12	30
	y	47,9	52,9	54	55,5	62,5	64,2	70,3	75	77	81,6	
4	x	21	32	45	55	57	64	77	80	90	96	60
	y	0,39	0,55	0,64	0,76	0,81	0,91	1,05	1,05	1,16	1,22	

5	x	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	1,2
	y	0,96	0,97	0,98	0,97	0,96	0,97	0,95	0,94	0,92	
6	x	0	4	10	15	21	29	36	51	68	32
	y	66,7	71	76,3	80,6	85,7	92,9	99,4	114	125,1	
7	x	19,1	25,0	30,1	36,0	40,0	45,1	50,0			38
	y	76,30	77,80	79,75	80,80	82,35	83,9	85,10			
8	x	8,1	16,1	21,8	43,9	65,8	87,6	96,5			50
	y	0,330	0,271	0,242	0,183	0,158	0,142	0,138			
9	x	125	175	225	275	325	375	425	475		300
	y	3,25	2,25	1,25	0,75	0,75	0,75	0,25	0,25		
10	x	10	30	50	70	90	110	130	150		80
	y	-40	-20	-20	0	20	40	80	100		

На основі наведених даних:

- побудувати діаграму розсіювання;
- оцініть щільність зв'язку за допомогою коефіцієнта *r;*
- побудувати лінійну і геометричну регресійну модель і порівняти їх;
- знайдіть прогноз якості продукції від значення x_0 фактора, за умови лінійної моделі.

Відповіді та вказівки

4.1. $r = -0{,}89$.

4.2. а) $r = +1$, б) $r = \frac{5}{7}$, в) $r = -0{,}9$, г) $r = 0{,}2$.

4.5. $r = 0{,}834$.

4.6. I. $r = 0{,}9,\ \eta - 3{,}9 = \frac{0{,}17}{\xi - 48{,}2},\ \xi - 48{,}2 = \frac{4{,}8}{\eta - 3{,}9}$.

II. $r = 0{,}015,\ \eta - 30{,}4 = \frac{0{,}0089}{\xi - 238{,}5},\ \xi - 238{,}5 = \frac{0{,}034}{\eta - 30{,}4}$.

4.7. б) $r = 0{,}6366$, в) $\frac{\eta - 7{,}5}{1{,}36} = 0{,}64\left(\frac{\xi - 7}{1{,}73}\right)$, друга лінія регресії отримується очевидним чином з першої.

4.8. $\eta = 0{,}87\xi + 57{,}5$.

4.9. $\eta = 0{,}5\xi^2 - 1{,}5\xi + 1{,}5$.

4.10. $\eta = \frac{27{,}315}{\xi - 7} + 233{,}172$.

Вказівка. Змінити початок відліку, взявши за 0 відліку 7-й рік.

4.11. $\eta = \xi^{*^2} + \xi^* - 3$, де $\xi^* = \xi - 1991$.

Вказівка. В якості 0 відліку (базового року) взяти 1991 рік.

4.12. $t = 2{,}001 \cdot n$.

Вказівка. Вихідну таблицю слід представити у логарифмічній шкалі, тобто у вигляді

$\ln \eta$	
$\ln t$	

Тоді, прологарифмувавши ліву і праву частини функціонального співвідношення, одержуємо $\ln t = \ln a + \alpha \ln n$, де $\ln t$, $\ln a$, $\ln \alpha$ можуть бути знайдені за МНК.

ДОДАТКИ

Додаток 1

Словник уживаних термінів

α **(рівень значущості)** у тому випадку, коли висунута нульова гіпотеза є правильною і вона відхиляється, припускається помилка I виду. Імовірність зробити помилку I виду позначається як грецькою літерою α і називається рівнем значущості.

Варіація – коливання ознаки, мінливість величини ознаки одиниць, що входять до складу сукупності. Існує декілька статистичних характеристик коливання ознаки: варіаційний розмах, середнє лінійне відхилення, середнє квадратичне відхилення, дисперсія, коефіцієнт.

Вибіркова сукупність – сукупність випадково відібраних об'єктів з генеральної сукупності.

Випадкова величина – величина, яка набуває залежно від деяких випадкових обставин одне зі значень x_1, x_2, …, x_n, що мають певні імовірності p_1, p_2, **…,** p_n.

Випадкова змінна – це будь-яка змінна, значення якої не може бути визначено.

Генеральна сукупність – сукупність об'єктів, з яких відбирається вибірка.

Гіпотеза – наукове припущення, що висувається для пояснення якогось явища та потребує перевірки досвідом і теоретичного обґрунтування для того, щоб стати науковою теорією.

Дисперсія – середній квадрат відхилення варіант x від середнього арифметичного значення $\bar{x}$.

Кількісні зміни – скалярне вимірювання у визначеній шкалі ступеня проявлення досліджуваних властивостей об'єкта.

Контрольна карта – карта, на якій для наочності відображення стану процесу відмічають значення відповідної вибіркової характеристики суміжних вибірок у часовій послідовності.

Кореляція – систематичний і обумовлений зв'язок між двома явищами чи процесами.

Кореляційна матриця – матриця, елементами якої є коефіцієнти кореляції між парами випадкових величин.

Моніторинг – 1) постійне спостереження за будь-яким процесом з метою виявлення його відповідності бажаному результату або первинними припущенням; 2) спостереження, оцінка і прогноз стану навколишнього середовища у зв'язку з господарською діяльністю людини.

Якість – сукупність характеристик об'єкта, які стосуються його здатності задовольнити встановлені й передбачені потреби.

Якість продукції – сукупність характеристик продукції (процесу, послуги), які стосуються її здатності задовольняти встановлені і передбачені потреби.