Кореляція

Лінійна регресія

- Кореляцією називають взаємозв'язок між середніми показниками сукупностей, а метод оцінки тісноти взаємозв’язку між середніми показниками досліджуваних сукупностей має назву кореляційного аналізу.
- Кореляція – це така залежність, коли будь-якому значенню однієї змінної величини може відповідати декілька різноманітних значень іншої змінної.
- Кореляція – взаємозв’язок між ознаками, що полягає в зміні середнього значення однієї з них залежно від зміни іншої.

Форма кореляційного зв'язку

- Під *формою кореляційного зв’язку* розуміємо тип аналітичного рівняння, що виражає залежність між досліджуваними ознаками. Розрізняють дві форми зв’язку: лінійну і нелінійну (криволінійну). Лінійна виражається рівнянням прямої лінії, нелінійна – рівнянням кривих ліній: гіперболи, параболи, степеневої, показникової тощо.

- За напрямом зв’язки бувають прямими і оберненими.
- Кореляцію і регресію називають *простою*, якщо досліджується зв'язок між двома ознаками,
- *множинною*, коли досліджується залежність між трьома і більшою кількістю ознак.

Коефіцієнт кореляції

- Ми розглянемо метод оцінки тісноти взаємозв’язку між двома явищами, який ґрунтується на визначенні так званого коефіцієнта кореляції.

Коефіцієнт кореляції

- є середнім арифметичним значенням добутку нормованих відхилень за двома досліджуваними ознаками

$$r = \overline{t_x t_y} = \frac{\sum t_x t_y}{n} \qquad r = \frac{\sum (x - M_x)(y - M_y)}{\sqrt{\sum (x - M_x)^2 \sum (y - M_y)^2}}$$

- Значення коефіцієнта кореляції лежить у межах від +1 до –1.
- $-1 \le r \le +1$.
- Чим ближче значення коефіцієнта кореляції до 1, тим тісніший зв'язок між досліджуваними явищами. Коли коефіцієнт кореляції наближається до 0, то кореляція між досліджуваними ознаками дуже мала, або її немає зовсім. Отже, абсолютна величина характеризує ступінь тісноти зв’язку.

Градації тісноти зв’язку:

- $0,7 \leq |r| < 1$ – сильна кореляція (тісний зв’язок);
- $0,5 \leq |r| < 0,7$ – середня кореляція (середньої тісноти зв’язок);
- $0 < |r| < 0,5$ – слабка кореляція (мала залежність або відсутня залежність).

Напрямленість коефіцієнта кореляції

- Якщо коефіцієнт кореляції позитивний, то досліджувані ознаки характеризуються позитивною кореляцією, тобто збільшення однієї ознаки веде до збільшення іншої. Наприклад, при збільшенні росту в середньому збільшується вага.
- Якщо коефіцієнт кореляції від'ємний, то існує обернена залежність між показниками, а досліджувані ознаки характеризуються негативною кореляцією, тобто при збільшенні одного показника – інший зменшується. Залежність між імовірністю захворювання дітей на дитячі інфекційні хвороби та їх віком існує обернена залежність: чим старша дитина, тим менша ймовірність захворювання.

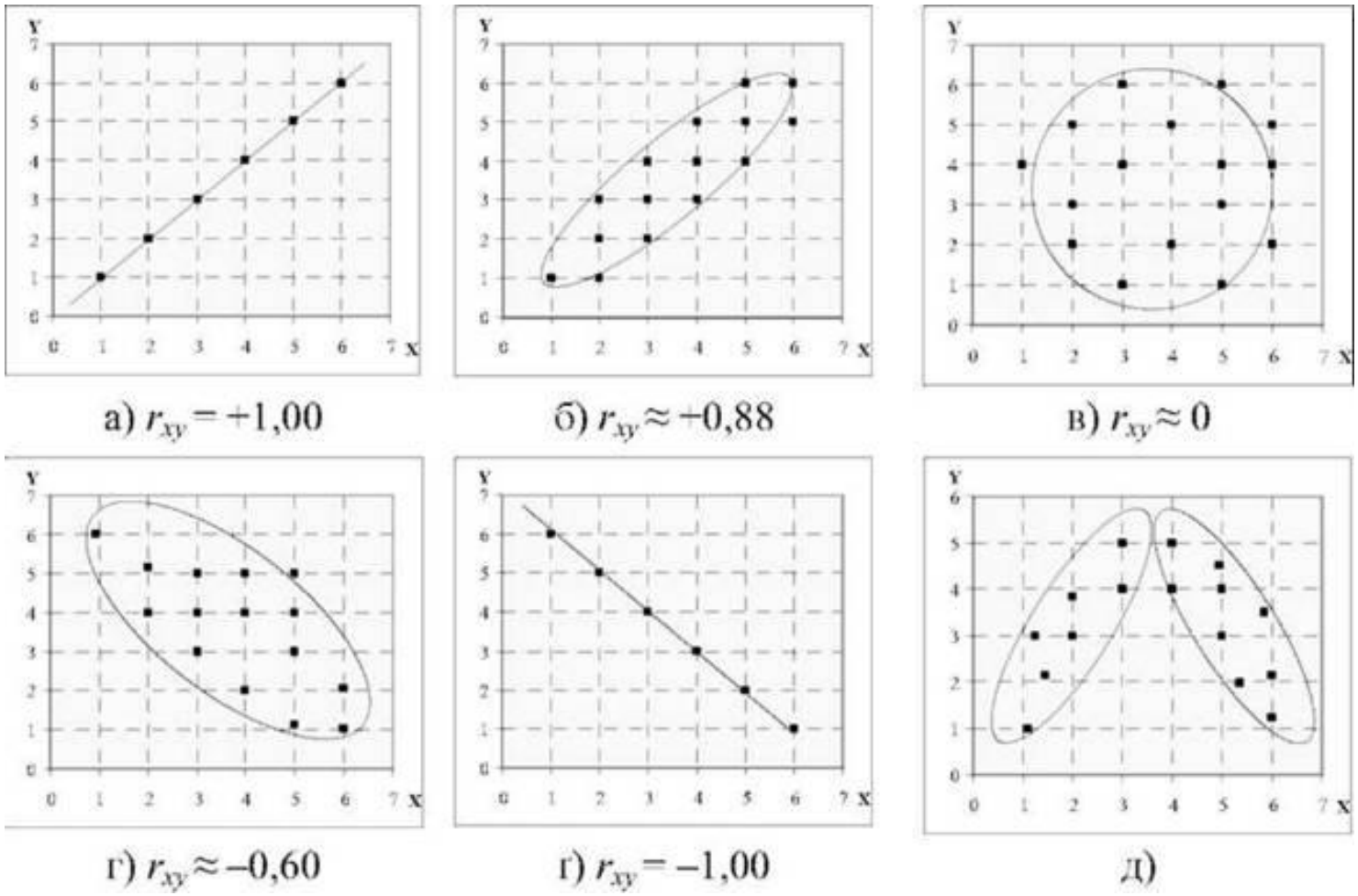
Кореляційні зв'язки

- Кореляційні зв'язки можна вивчати на якісному рівні з діаграм розсіяння емпіричних значень змінних X і Y і відповідним чином їх інтерпретувати. Так, наприклад, якщо підвищення рівня однієї змінною супроводжується підвищенням рівня іншої, то йдеться про *позитивну* кореляцію або прямий зв'язок.

Кореляційні зв'язки

- Якщо ж зростання однієї змінної супроводжується зниженням значень іншої, то маємо справу з негативною кореляцією або зворотним зв'язком. Нульовою називається кореляція за відсутності зв'язку змінних. Проте нульова загальна кореляція може свідчити лише про відсутність *лінійної* залежності, а не взагалі про відсутність будь якого *статистичного* зв'язку .

а) строга позитивна кореляція; б) сильна позитивна кореляція; в) нульова кореляція; г) помірна негативна кореляція; ґ) строга негативна кореляція; д) нелінійна кореляція



Достовірність кореляції.

- Достовірність кореляційного зв’язку безпосередньо пов’язана з кількістю проведених досліджень, тобто з обсягом сукупності n. Сильні кореляційні зв’язки можна з високою вірогідністю довести на малому обсязі експериментального матеріалу. Зате слабкі взаємовпливи в природі можна виявити тільки на основі великого обсягу досліджень.

- Імовірність статистичної істотності будь-якого показника, що характеризується нормальним розподілом, можна оцінити, визначивши коефіцієнт Стьюдента. Але в зв'язку з тим, що коефіцієнт кореляції не підлягає законові нормального розподілу, для встановлення ступеня вірогідності треба перевести коефіцієнт кореляції r у такий показник z, який підлягає закону нормального розподілу.

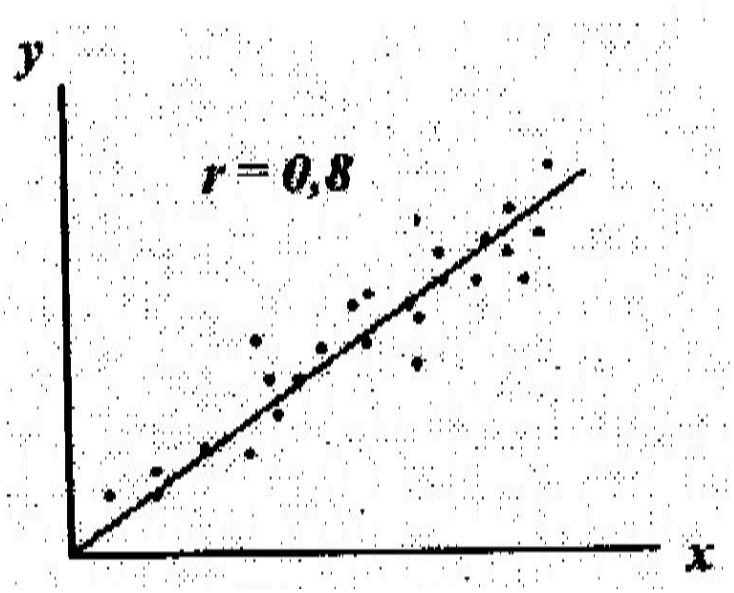
$$z = \frac{1}{2} \ln \frac{r+1}{r-1}$$

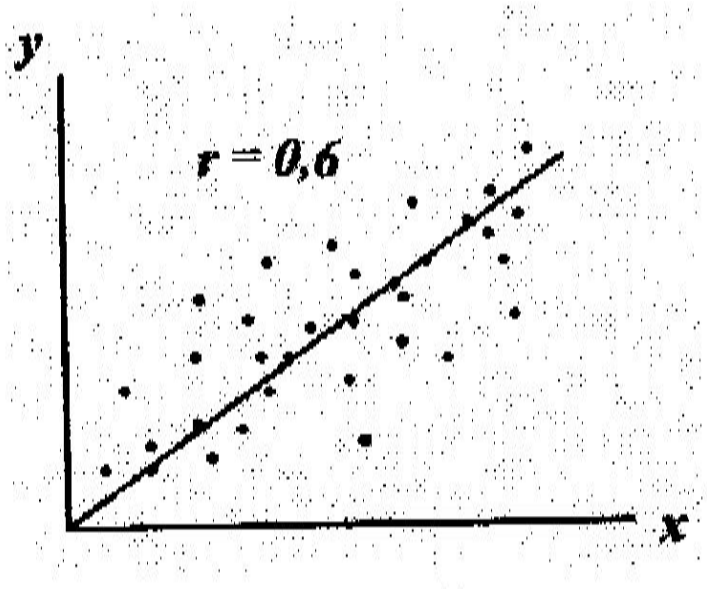
Рівняння лінійної регресії

- Під лінійною кореляційною залежністю між двома ознаками розуміють таку залежність, яка має лінійний характер і виражається рівнянням прямої лінії
- $y = a + bx,$
- де a і b – відповідні коефіцієнти.

- Лінійна регресія – це така залежність, коли рівномірні зміни аргументу x викликають одинакові зміни функції *y*.
- Чим більший кореляційний зв’язок, тим тісніше точки зосереджені навколо прямої лінії регресії.

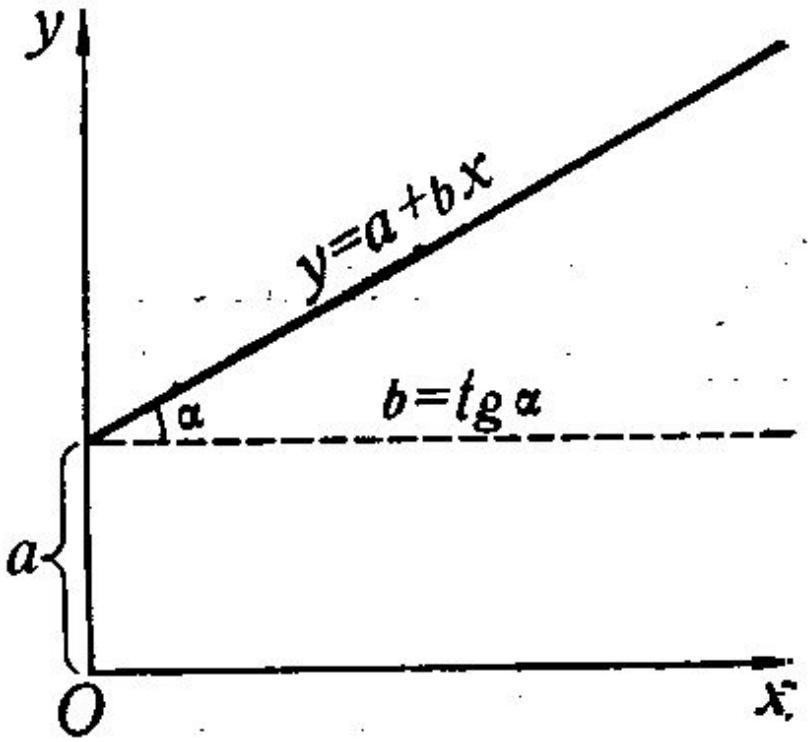
Лінія регресії та залежність від коефіцієнта кореляції.





- Вільний член рівняння a – це відрізок від початку координат до точки перетину лінії з віссю ординат,
- a b – тангенс кута нахилу лінії до осі абсцис.

Графічне зображення рівняння прямої лінії $y = a + bx$.



- Виведення рівняння лінійної регресії полягає в тому, щоб встановити, на скільки одиниць змінюється одна ознака (наприклад у), якщо друга ознака (х) змінюється на одиницю. Цю умову можна записати у вигляді такої лінійної пропорції, коли обидві ознаки х та у задані як відхилення від середніх арифметичних значень Мх і Му:

$$\frac{x - M_x}{y - M_y} = \frac{1}{b}$$

Рівняння регресії

виведене з даної пропорції, набуває такого вигляду:

$$y = M_y + b(x - M_x)$$

- У цьому рівнянні b є так званим коефіцієнтом регресії, який показує , на скільки одиниць зміниться ознака у, якщо ознака х зміниться на одиницю.

Коефіцієнт регресії.

- Коли вивчають регресію між двома ознаками, то слід вказати, яка ознака змінюється фіксованими, одиничними кроками, а зміна якої при цьому досліджується. Як правило, ознаку з фіксованими змінами позначають символом х, а ознаку, зміни якої вивчають, – символом у. Тоді говорять про регресію у по х.

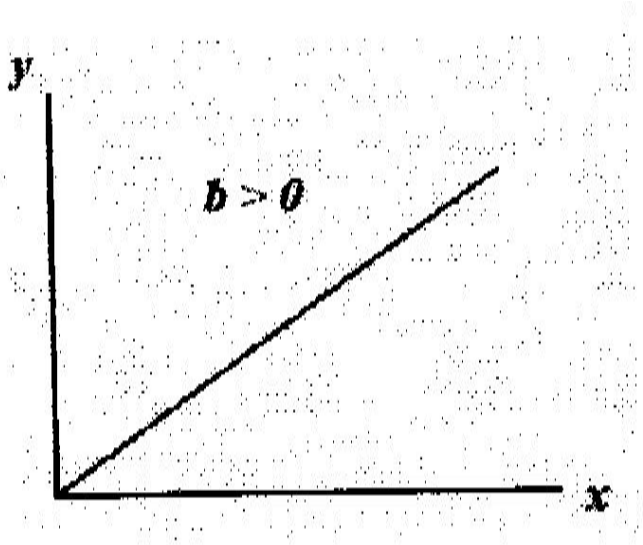
- При позитивному зв’язку між ознаками лінія регресії утворює гострий кут з віссю абсцис, коефіцієнт регресії b >0. При негативному зв’язку лінія регресії утворює тупий кут з віссю абсцис, коефіцієнт регресії b< 0.

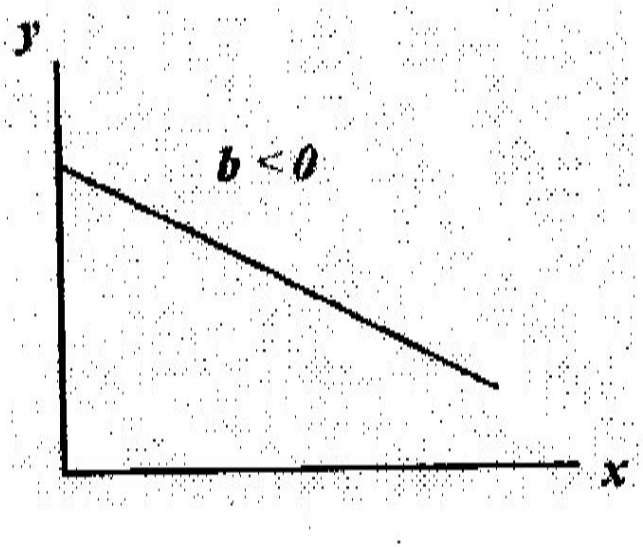
Коефіцієнт регресії

$$b_{y/x} = \frac{\Sigma(x - M_x)(y - M_y)}{\Sigma(x - M_x)^2}$$

$$b_{y/x} = r\frac{\sigma_y}{\sigma_x}$$

Напрямок нахилу лінії регресії





Емпірична та теоретична лінії регресії

- Емпірична лінія регресії є ламаною лінією, бо на неї впливають випадкові фактори статистичної природи. Теоретична лінія регресії загладжує цю ламану лінію до прямої, що проходить на найменшій відстані між експериментальними точками.

Криволінійна регресія

- Якщо зв’язок між досліджуваними явищами суттєво відрізняється від лінійної, то коефіцієнт кореляції непридатний для визначення міри зв’язку. Він може вказати на відсутність взаємозв’язку, там де простежується сильна криволінійна залежність. При нелінійному кореляційному зв’язку рівномірним змінам однієї ознаки відповідають в середньому нерівномірні, які підлягають відповідній закономірності змін другої ознаки. Зовнішнім проявом нелінійної регресії є те, що емпіричні лінії регресії на графіку мають вигляд кривих різної конфігурації. Тому необхідний новий показник, який би встановив степінь криволінійної залежності.

Кореляційне відношення (η)

- визначають як лінійну, так і нелінійну залежність. В першому випадку η = r , але чим сильніша виражена нелінійність зв’язку, тим більше значення кореляційного відношення переважає величину коефіцієнта кореляції r. Кореляційне відношення є кількісною мірою спряженості ознак при будь-якій формі зв’язку між ними. Він є двосторонньою мірою спряженості ознак, отже, говорять про кореляційне відношення *y* по *x* ηy/x і кореляційне відношення x по y ηx/y

Кореляційне відношення обчислюють за формулою

$$\eta_{y/x} = \frac{\overline{\sigma_y}}{\sigma_y}$$

$$\sigma_y = \sqrt{\frac{\Sigma f_y (\upsilon_y - M_y)^2}{n-1}}$$

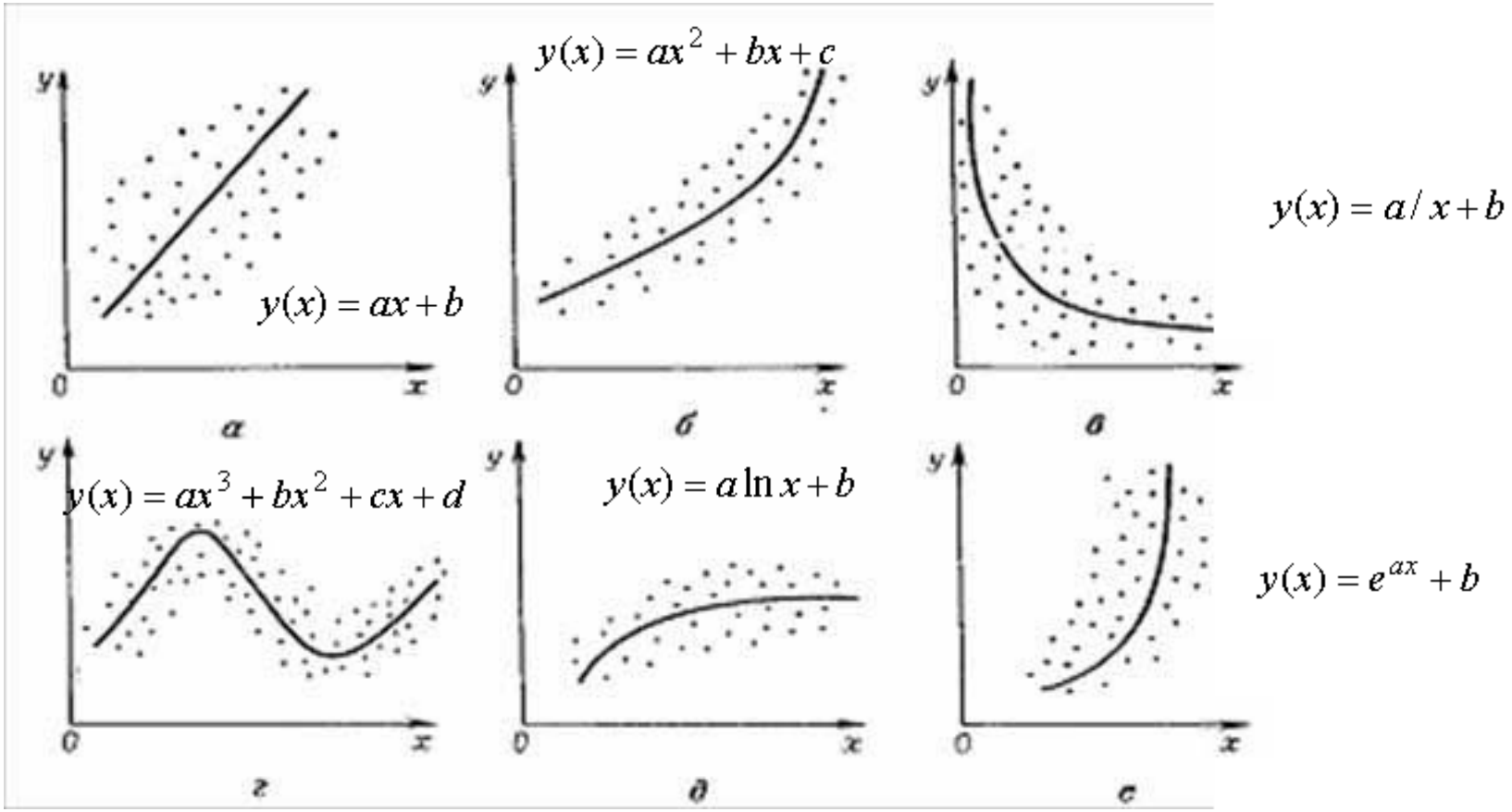
$$\overline{\sigma_y} = \sqrt{\frac{\Sigma f_x (\overline{y - M_y})^2}{n-1}}$$

$$F = \frac{\eta^2 - r^2}{1-\eta^2} \cdot \frac{\gamma_2}{\gamma_1}$$

- Доказ лінійності зв'язку полягає в тому, щоб дослідити, чи існує статистично істотна різниця між показниками будь-якого зв'язку - кореляційним відношенням і показником лінійного зв'язку . Якщо ця різниця статистично неістотна, то гіпотеза про лінійність кореляційного зв'язку приймається. В протилежному випадку гіпотезу про лінійність зв'язку треба відхилити.

Кореляційний та регресійний аналізи з використанням засобів Excel

- Для оцінювання парного кореляційного зв'язку між показниками можна використати інструмент **Кореляция** з пакету «Аналіз даних» або статистичну функцію **КОРРЕЛ.** У першому випадку дістанемо таблицю парних коефіцієнтів кореляції для кількох показників одночасно (але без зворотного зв'язку з вхідними даними), у другому випадку можемо виконати обчислення лише для двох масивів.
- При проведенні кореляційно-регресійного аналізу можна застосовувати також додаткові статистичні функції для оцінювання параметрів моделі та залежності між показниками:
- **НАКЛОН** – визначає коефіцієнт *b* у рівнянні $y = a + bx$,
- **ОТРЕЗОК** – визначає коефіцієнт *a* у рівнянні $y = a + bx$,
- **ЛИНЕЙН** – вводяться масиви y та x та обчислюються коефіцієнти *b* і *a;*
- **ПИРСОН –** визначає коефіцієнт кореляції у межах -1 до +1;
- **КОВАР –** визначає коефіцієнти коваріації, а також середні попарні добутки відхилень.



Кореляційні поля й гіпотетичні рівняння регресії :

а – лінійне ;

б – квадратичне $y(x) = ax^2 + bx + c$

в – гіперболічне $y(x) = a/x + b$

г) поліноміальне $y(x) = ax^3 + bx^2 + cx + d$

д – логарифмічне $y(x) = a\ln x + b$

е – експоненціальне $y(x) = e^{ax} + b$