ЛЕКЦІЯ 6
КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Як часто нам доводиться чути висловлювання про те, що одне явище корелює з іншим? «Ціна на нафту корелює з курсами валют», «Біль у м’язах після тренування не корелює з гіпертрофією м’язових волокон», «Результати проведених досліджень добре корелюють…». Складається враження, що поняття «кореляція» стало широко використовуватись не лише в науці, а й у повсякденному житті. Можна сказати, що кореляція виражає міру лінійної залежності між двома явищами. При описанні двовимірних випадкових величин буває недостатньо таких характеристик, як математичне очікування, дисперсія, середньоквадратичне відхилення. Тому часто використовуються ще дві важливі характеристики – *коваріація* та *кореляція*.

Наприклад, діти, які часто дивляться по телевізору бойовики, менше читають. Діти, які більше читають, краще навчаються. Не так-то просто визначити, де тут причини, а де наслідки, та про це мова не йдеться. Ми можемо лише, висунути гіпотезу про наявність зв’язку, підкріпити його цифрами. Якщо зв'язок дійсно є, говорять, що між двома випадковими величинами є кореляція. Якщо збільшення однієї випадкової величини пов’язано зі збільшенням іншої, кореляція називається *прямою*. Наприклад кількість прочитаних за рік сторінок та середній бал (успішність). Якщо ж навпаки збільшення однієї величини пов’язане зі зменшенням іншої, говорять про *обернену кореляцію*. Наприклад, кількість бойовиків та прочитаних сторінок.

Як відомо, задачі *дисперсійного аналізу* передбачають дослідження питання залежності результату експеримента від зміни деяких факторів без виявлення конкретного вигляду цієї залежності (що вже є задачею регресійного аналізу). Тобто *регресійний аналіз* полягає у визначенні аналітичних залежностей зв’язку, в якому зміна результативної ознаки обумовлюється впливом однієї або кількох факторних ознак, а множина всіх

інших факторів застосовується як постійні (або усереднені) величини. Як правило, такі залежності будуються за допомогою МНК (лекція 5).

У свою чергу, *кореляційний аналіз* передбачає вимірювання параметрів рівняння, що виражає зв'язок середніх значень залежної змінної зі значеннями незалежної змінної (залежність середніх величин результативної ознаки від значень одного або декількох факторних ознак), а також вимірювання тісноти зв'язку двох (або більшої кількості) ознак між собою.

Термін *«кореляція»* вперше застосував французький палеонтолог Ж. Кюв'є, який вивів закон кореляції частин та органів тварин (цей закон дозволяє відтворювати за знайденими частинами тіла вигляд всієї тварини). У статистику вказаний термін ввів у 1886 р. англійський біолог і статистик Ф. Гальтон (не просто зв'язок – relation, а *ніби* зв'язок – *co*-relation). Проте точну формулу для підрахунку коефіцієнта кореляції розробив його учень – математик і біолог К. Пірсон.

Кореляція у прямому перекладі означає «співвідношення». Якщо зміна однієї змінної супроводжується зміною іншої, то можна говорити про кореляцію цих змінних. *Наявність кореляції двох змінних не передбачає причинно-наслідкової залежності між ними, але дає можливість висунути таку гіпотезу. Відсутність же кореляції дозволяє відкинути гіпотезу про причинно-наслідковий зв'язок змінних.*

6.1. Змішаний другий момент при розрахунку похибок непрямих вимірювань

Припустимо, що для знаходження функції $q(x,y)$ ми вимірюємо дві величини x та y кілька разів і отримуємо N пар даних $(x_1,y_1),\ldots,(x_N,y_N)$. За результатами N вимірювань $x_1,\ldots,x_N$ ми можемо обчислити середнє $\bar{x}$ та стандартне відхилення σ_x; аналогічно за даними $y_1,\ldots,y_N$ ми можемо обчислити $\bar{y}$ та σ_y. Далі, використовуючи N пар результатів вимірювань, ми

розраховуємо N значень величини, що нас цікавить: $q_i = q(x_i, y_i)$, $(i = \overline{1,N})$. Отримавши $q_1,\ldots,q_N$, ми зможемо обчислити їх середнє $\bar{q}$, яке, за нашими припущеннями, дає найкращу оцінку для q, та стандартне відхилення σ_q, що є мірою випадкової похибки у значеннях q_i.

Припускатимемо, що всі наші похибки малі і, отже, всі числа x_1, ..., x_N близькі до $\bar{x}$, а всі y_1, ..., y_N – до $\bar{y}$. В цьому випадку справедливе наближення

$$q_i = q(x_i, y_i) \approx q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}). \qquad (6.1)$$

У виразі (6.1) часткові похідні $\frac{\partial q}{\partial x}$ та $\frac{\partial q}{\partial y}$ обчислюються у точці $x = \bar{x}$, $y = \bar{y}$, і, отже, вони однакові для всіх $i = \overline{1,N}$. В рамках цього наближення середнє набуває вигляду $\bar{q} = \frac{1}{N}\sum_{i=1}^{N} q_i = \frac{1}{N}\sum_{i=1}^{N}\left[q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}) \right]$.

Тут другий та третій доданки дорівнюють нулю. (Так, з визначення середнього $\bar{x}$ виходить, що $\sum(x_i - \bar{x}) = 0$.) Таким чином, ми отримуємо простий результат

$$\bar{q} = q(\bar{x}, \bar{y}), \qquad (6.2)$$

тобто щоб знайти середнє $\bar{q}$, ми просто повинні обчислити значення функції $q(x,y)$ в точці $x = \bar{x}$ та $y = \bar{y}$.

Стандартне відхилення для N значень $q_1,\ldots,q_N$ є $\sigma_q^2 = \frac{1}{N}\sum(q_i - \bar{q})^2$. Підставляючи в цей вираз (6.1) та (6.2), ми знаходимо, що

$$\sigma_q^2 = \frac{1}{N}\sum\left[\frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}) \right]^2 = \left(\frac{\partial q}{\partial x}\right)^2 \frac{1}{N}\sum(x_i - \bar{x})^2 + \left(\frac{\partial q}{\partial y}\right)^2 \frac{1}{N}\sum(y_i - \bar{y})^2 +$$

$$+ 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y}). \qquad (6.3)$$

Суми у перших двох членах – це визначення стандартних відхилень σ_x та σ_y. Остання сума називається ***змішаним другим моментом x та y*** (а ще ***коваріацією***) і позначається як $\mathrm{cov}(x, y)$ або

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}). \qquad (6.4)$$

З цим визначенням вираз (6.3) для стандартного відхилення σ_q набуває вигляду

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy}. \qquad (6.5)$$

Цей вираз визначає стандартне відхилення σ_q незалежно від того, чи є незалежними вимірювання x та y і чи нормально вони розподілені.

Розглянемо основні властивості змішаного другого моменту (коваріації).

1. *Коваріація випадкової величини з самою собою є її дисперсією*: $\mathrm{cov}(x,x) = D(x)$ або $\sigma_{xx} = \sigma_x^2 = D(x)$.

2. *Коваріація симетрична*: $\mathrm{cov}(x,y) = \mathrm{cov}(y,x)$ або $\sigma_{xy} = \sigma_{yx}$.

3. *Постійний множник можна виносити за знак коваріації*: $\mathrm{cov}(ax,y) = \mathrm{cov}(x,ay) = a\cdot\mathrm{cov}(x,y)$.

4. *Коваріація не зміниться, якщо до однієї з випадкових величин (або до двох одразу) додати постійну величину*:

$$\mathrm{cov}(x+a,y) = \mathrm{cov}(x,y+a) = \mathrm{cov}(x+a,y+a) = \mathrm{cov}(x,y).$$

5. $\mathrm{cov}(ax+b,cy=d) = ac\cdot\mathrm{cov}(x,y)$.

6. *Якщо вимірювання x та y незалежні, то легко бачити, що після багатьох вимірювань змішаний другий момент (коваріація) σ_{xy} повинен прямувати до нуля*: $\mathrm{cov}(x,y) = 0$ (або $\sigma_{xy} = 0$). Незалежно від значення y_i величина $x_i - \bar{x}$ з рівною імовірністю може бути як додатною, так і від'ємною. Таким чином, після багатьох вимірювань додатні й від'ємні члени у (6.4) повинні приблизно компенсуватися, і для нескінченного числа вимірювань множник $\frac{1}{N}$ у (6.4) забезпечить рівність σ_{xy} нулю. (Після скінченного числа вимірювань величина σ_{xy} не буде точно дорівнюватись нулю, але повинна бути *малою,* якщо помилки в *x* та *y* дійсно є незалежними й випадковими). При σ_{xy}=0 вираз для σ_q зводиться до

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2, \qquad (6.6)$$

знайомого результату для незалежних і випадкових похибок.

7. *Якщо вимірювання x та y не є незалежними, то змішаний другий момент (коваріація)* σ_{xy} *не повинен дорівнюватися нулю*: $\mathrm{cov}(x, y) \neq 0$ (або $\sigma_{xy} \neq 0$). Наприклад, переоцінка x завжди тягне за собою переоцінку y і навпаки. В цьому випадку числа $(x_i - \bar{x})$ та $(y_i - \bar{y})$ завжди матимуть один знак (плюс або мінус), а їх добуток буде завжди додатним. Оскільки всі члени в сумі (6.4) додатні, то σ_{xy} не повинна зникати навіть тоді, коли ми проводимо нескінченно багато вимірювань.

Коли змішаний другий момент $\sigma_{xy} \neq 0$ (навіть для нескінченно великого числа вимірювань), ми говоримо, що помилки в x та y є ***корельованими***. В цьому випадку похибка σ_q в $q(x, y)$, що визначається (6.5), – це не одне й те ж, що ми отримали б згідно з формулою (6.6) для незалежних і випадкових помилок.

8. За допомогою формули (6.5) ми можем отримати верхню границю для σ_q, яка завжди справедлива. *Змішаний другий момент (коваріація)* σ_{xy} *задовольняє так званій* ***нерівності Шварца***

$$\left|\mathrm{cov}(x, y)\right| \leq \sigma_x \sigma_y \qquad \text{або} \qquad \left|\sigma_{xy}\right| \leq \sigma_x \sigma_y. \qquad (6.7)$$

Нерівність (6.7) можна записати у такому вигляді: $\left|\mathrm{cov}(x, y)\right| \leq \sqrt{D(x)D(y)}$ або $\left|\sigma_{xy}\right| \leq \sqrt{D(x)D(y)}$.

Якщо підставити (6.7) у вираз (6.5) для похибки σ_q, то отримаємо, що $\sigma_q^2 \leq \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_x^2 + 2\left|\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right|\sigma_x \sigma_y = \left[\left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y\right]^2$, тобто

$$\sigma_q \leq \left|\frac{\partial q}{\partial x}\right|\sigma_x + \left|\frac{\partial q}{\partial y}\right|\sigma_y. \qquad (6.8)$$

6.2. Коефіцієнт лінійної кореляції

Поняття змішаного другого моменту (коваріації) σ_{xy} дозволяє відповісти на питання, про те, наскільки добре набір результатів вимірювань $(x_1, y_1),\ldots,(x_N, y_N)$ для двох змінних підтверджує гіпотезу про лінійну залежність x та y.

Припустимо, що ми отримали N пар виміряних значень $(x_1, y_1),\ldots,(x_N, y_N)$ двох змінних, які, за нашими очікуваннями, повинні бути по'вязані лінійною залежністю вигляду $y = A + Bx$.

Важливо, що $x_1,\ldots,x_N$ у даному випадку є не результатами вимірювань лише однієї величини, а результатами вимірювань N різних значень однієї змінної (наприклад, N різних висот, з яких кидали камінь). Те ж саме відноситься й до $y_1,\ldots,y_N$.

За допомогою МНК ми можемо знайти значення A та B для лінії, яка найкращим чином апроксимує точки $(x_1, y_1),\ldots,(x_N, y_N)$. Якщо у нас є надійні оцінки похибок вимірювань, то ми можемо бачити, чи дійсно виміряні точки розміщуються розумно близько до лінії (порівняно з відомими похибками). Якщо це так, то вимірювання підтверджують наше припущення, що x та y пов'язані лінійно.

На жаль, у багатьох експериментах важко визначити надійні оцінки похибок заздалегідь, і тому ми повинні використати вихідні дані, щоб визначити, чи пов'язані дві змінні лінійно.

Розглянемо приклад експерименту, коли неможливо визначити величину похибок заздалегідь. Уявімо собі, що професор, що бажає переконати своїх студентів у тому, що виконання домашніх завдань допоможе їм добре скласти іспити, збирає відомості про їх оцінки за домашнє завдання та за іспит і виводить їх на графік розкидів, як наведено на рис.6.1. На цьому графіку оцінки за домашнє завдання відкладені по горизонтальній осі, а за іспит – по вертикальній (оцінки наведені за 100-бальною системою). Кожна точка (x_i, y_i)

показує оцінку одного студента за домашнє завдання x_i та за іспит y_i. Професор сподівається показати, що високі оцінки за іспит *корелюють* з високими оцінками за домашнє завдання і навпаки (і його графік розкидів підтверджує, що це приблизно так). (В цьому прикладі експерименту немає жодних похибок у точках; дві оцінки кожного студента відомі точно). Похибка буде скоріше у мірі, якою корельовано оцінки, і саме це повинно бути визначено з даних.)

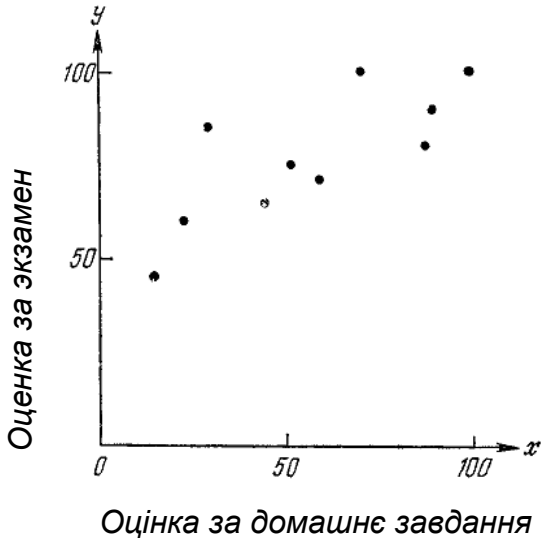

Рисунок 6.1 – Графік розкидів

Дві змінні x та y можуть бути пов'язані й складнішою залежністю, ніж простий лінійний зв'язок вигляду $y = A + Bx$. Наприклад, багато фізичних законів призводить до квадратичної залежності типу $y=A+Bx+Cx^2$. Тим не менш ми обмежимо наш розгляд випадком простішої задачі, коли потрібно визначити, чи підтверджує даний набір точок гіпотезу про *лінійний зв'язок* $y = A + Bx$.

Міра, якою набір точок $(x_1, y_1),\ldots,(x_N, y_N)$ підтверджує лінійну залежність між x та y, вимірюється ***коефіцієнтом лінійної кореляції*** (***коефіцієнтом кореляції***)

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{6.9}$$

де змішаний другий момент σ_{xy} та стандартні відхилення σ_x та σ_y визначаються так, як я раніше, формулами (6.4) і $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$. Підставляючи ці

визначення у (6.9), можна переписати вираз для ***коефіцієнта кореляції Пірсона*** у вигляді

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}. \qquad (6.10)$$

Число r показує, наскільки добре точки (x_i, y_i) *апроксимуються прямою.*

Наведемо основні властивості коефіцієнту кореляції.

1. *Коефіцієнт кореляції r приймає значення між -1 та 1*. Якщо *r* близьке до ±1, точки лежать поблизу деякої прямої; якщо *r* близьке до 0, точки не корельовані та або незначно, або взагалі не групуються біля прямої.

2. Якщо дві змінні *x* та *y* такі, що при нескінченно великому числі вимірювань їх змішаний другий момент дорівнює нулю (і, отже, *r=0*), то ми говоримо, що змінні ***не корельовані***. Якщо після скінченного числа вимірювань коефіцієнт кореляції *r* малий, це буде підтвердженням гіпотези про те, що *x* та *y* не корельовані.

3. Можна показати, що $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - N\bar{x}\bar{y}$. Тоді вираз для коефіцієнта кореляції (6.10) можна записати у вигляді

$$r = \frac{\sum x_i y_i - N\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - N\bar{x}^2\right)\left(\sum y_i^2 - N\bar{y}^2\right)}}. \qquad (6.11)$$

Кореляційним аналізом називають оцінювання коефіцієнту кореляції за даними спостережень.

Розглянемо приклад з оцінками за іспит та домашнє завдання (рис.6.1). Оцінки наведені також у табл.6.1.

Таблиця 6.1 – Оцінки студентів

Студент, *i*	1	2	3	4	5	6	7	8	9	10
Домашнє завдання, x_i	90	60	45	100	15	23	52	30	71	88
Іспит, y_i	90	71	65	100	45	60	75	85	100	80

Простий розрахунок показує, що коефіцієнт кореляції для цих десяти пар оцінок дорівнює $r = 0.7807 \approx 0.8$. Тоді можна зробити висновок, що це

значення «розумно близьке» до 1, і тому можна сказати, що оскільки наявна висока кореляція між оцінками за домашнє завдання і за іспит, то важливо виконувати домашнє завдання.

Якщо б коефіцієнт кореляції r вийшов близьким до нуля, це означало б, що оцінки за домашнє завдання аж ніяк не пов'язані з оцінками за іспит.

Якщо б виявилось, що величина r близька до -1, це означало б, що оцінки за домашнє завдання і за іспит підпадають під від'ємну кореляцію, тобто що студенти, які добре виконують домашнє завдання, погано складають іспити.

6.3. Достовірність коефіцієнту кореляції

Достовірність коефіцієнту кореляції визначається шляхом його порівняння з обчислювальною середньою помилкою (похибкою).

Середня помилка коефіцієнту кореляції визначається формулою

$$S_r = \pm \frac{1 - r^2}{\sqrt{N - 1}}, \tag{6.12}$$

де r – коефіцієнт кореляції, N – число спостережень.

Коефіцієнт кореляції вважається ***достовірним***, якщо втричі перевищує свою середню помилку:

$$\left| \frac{r}{S_r} \right| \geq 3. \tag{6.13}$$

В іншому випадку необхідно збільшити кількість спостережень (вимірювань).

6.4. Кількісний критерій значущості

Припустимо, що дві змінні *x* та *y* в дійсності не є корельованими, тобто для нескінченно великого числа вимірювань коефіцієнт кореляції дорівнював би нулю. Після скінченного числа вимірювань дуже малоймовірно, щоб *r* точно дорівнював би нулю (у випадку скінченного числа вимірювань використовується термін «***вибірковий коефіцієнт кореляції***»). Виявляється, можна обчислити імовірність того, що *r* буде не меншим за будь-яке задане значення.

Позначимо через $P_N\left(|r| \geq r_0\right)$ імовірність того, що *N* вимірювань двох некорельованих змінних *x* та *y* приведуть до значення коефіцієнта *r*, не меншого, ніж будь-яке значення r_0 (оскільки кореляція означає, що *r* близьке до +1, або -1, то ми розглядаємо імовірність отримання ***абсолютного значення*** $|r| \geq r_0$.) Наприклад, можна розрахувати імовірність $P_N\left(|r| \geq 0.8\right)$ того, що після *N* вимірювань некорельованих змінних *x* та *y* коефіцієнт кореляції буде, принаймні, не меншим, ніж отримане у розглянутому прикладі значення 0.8.

Результати таких обчислень наведені у табл.А.1 Додатку А. Лівий стовпець таблиці показує число експериментальних точок *N*. (В нашому прикладі *N=10*). Числа в кожному наступному стовпці – імовірності того, що *N* вимірювань двох некорельованих змінних дадуть коефіцієнт *r*, який, принаймні, не менший, ніж верхнє число у стовпці. Наприклад, як ми бачимо, імовірність того, що десять некорельованих точок дадуть $|r| \geq 0.8$, є невеликою і складає лише 0.5%. Отже, можна сказати, що дуже неймовірно, щоб некорельовані оцінки дали для коефіцієнта кореляції значення $|r|$, більше або рівне величині 0.8. Іншими словами, дуже імовірно, що оцінки за домашнє завдання і за іспит дійсно є корельованими.

Отримавши значення імовірності, можна дати найповнішу можливу відповідь на питання про те, наскільки добре N пар значень (x_i, y_i) підтверджують лінійний зв'язок між x та y. За виміряними точками можна спочатку обчислити значення коефіцієнта кореляції r_0. Потім знайти імовірність $P_N(|r| \geq |r_0|)$ того, що N некорельованих точок дадуть для коефіцієнта значення не менше, ніж отриманий коефіцієнт r_0. Якщо ця імовірність «достатньо мала», то ми можемо зробити висновок, що *дуже* ***неймовірно****, щоб x та y були не корельовані, і, отже, дуже* ***імовірно****, що вони дійсно корельовані.*

Ми ще повинні обрати значення імовірності, яке розглядатимемо як «достатньо мале». Розповсюджений вибір полягає в тому, щоб розглядати кореляцію r_0, що спостерігається, як ***значущу***, якщо імовірність отримання коефіцієнта r, такого що $|r| \geq |r_0|$, для некорельованих змінних менше 5%. Кореляцію іноді називають ***високозначущою***, якщо відповідна імовірність менше 1%. *Який би вибір ми не зробили, ми не отримаємо точно визначеної відповіді, які дані корельовані, а які ні. Замість цього в нас є кількісна міра, що показує, наскільки неімовірним є припущення про те, що вони не корельовані*.

Приклад. Припустимо, що ми вимірюємо три пари значень (x_i, y_i) і визначаємо, що коефіцієнт кореляції дорівнює 0.7 (або -0.7). Чи підтверджує це значення гіпотезу про те, що x та y пов'язані лінійно?

Звертаючись до табл.А.1 Додатку А, ми бачимо, що навіть якщо змінні x та y взагалі не корельовані, то імовірність отримання $|r| \geq 0.7$ при N=3 складає 51%. Іншими словами, цілком можливо, що x та y не корельовані; таким чином, у нас немає надійного доказу кореляції. Дійсно, у випадку лише трьох вимірювань було б дуже складно отримати переконливе підтвердження кореляції. Навіть значення коефіцієнта 0.9 є недостатнім для ствердження кореляції, оскільки імовірність отримання $|r| \geq 0.9$ у випадку трьох вимірювань некорельованих змінних дорівнює 29%.

Якщо б ми знайшли значення коефіцієнта 0.7 за шістьма вимірюваннями, то ситуація була б дещо краща, але все ще недостатньо хорошою. З N=6 імовірність отримання $|r| \geq 0.7$ для некорельованих змінних дорівнює 12%. Це число не таке вже й мале, щоб виключити можливість того, що x та y не корельовані.

З іншого боку, якщо б ми отримали $|r| \geq 0.7$ після 20 вимірювань, то у нас було б вагоме підтвердження кореляції, оскільки при N=20 імовірність отримання $|r| \geq 0.7$ для двох некорельованих змінних дорівнює лише 0.1%. За будь-якими критеріями це дуже неправдоподібно, і ми могли б впевнено стверджувати, що кореляція виявлена. Зокрема, ця кореляція могла бути названа високозначущою, оскільки відповідна імовірність менше 1%.

Контрольні питання

1. Змішаний другий момент при розрахунку похибок непрямих вимірювань.
2. Коефіцієнт лінійної кореляції та його основні властивості.
3. Достовірність коефіцієнту кореляції.
4. Кількісний критерій значущості.

Додаток А

Імовірності коефіцієнтів кореляції

Міра, якою N точок $(x_1, y_1),\ldots,(x_N, y_N)$ апроксимуються прямою, визначається коефіцієнтом лінійної кореляції (коефіцієнтом кореляції Пірсона) $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$, який завжди лежить в інтервалі $-1 \le r \le 1$. Значення r, близькі до ±1, означають дуже високий міру лінійної кореляції; значення, близькі до 0, вказують на слабку кореляцію або на її відсутність.

Кількісна міра апроксимації може бути отримана за допомогою табл.А.1. Для будь-якого визначеного r_0 $P_N(|r| \ge |r_0|)$ є імовірність того, що результати N вимірювань двох некорельованих змінних матимуть коефіцієнт кореляції r, не менший, ніж r_0. Таким чином, якщо ми отримаємо коефіцієнт r_0, для якого імовірність $P_N(|r| \ge |r_0|)$ мала, то неймовірно, щоб наші змінні були некорельованими, тобто кореляція існує. Зокрема, якщо $P_N(|r| \ge |r_0|) \le 5\%$, кореляція називається ***значущою***, якщо ця імовірність менше 1%, то кореляція називається ***високозначущою***.

Наприклад, імовірність того, що результати 20 вимірювань (N=20) двох некорельованих змінних дадуть $|r| \ge 0.5$, визначається таблицею у 2.5%. Таким чином, якщо б результати 20 вимірювань дали r=0.5, то в нас був би значущий доказ лінійної кореляції між двома змінними.

Таблиця А.1 – Виражена у відсотках імовірність $P_N\left(|r| \geq |r_0|\right)$

(Прочерки вказують на імовірності, менші за 0.05%.)

	r_0										
N	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
3	100	94	87	81	74	67	59	51	41	29	0
4	100	90	80	70	60	50	40	30	20	10	0
5	100	87	75	62	50	39	28	19	10	3.7	0
6	100	85	70	56	43	31	21	12	5.6	1.4	0
7	100	83	67	51	37	25	15	8.0	3.1	0.6	0
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0
9	100	80	61	43	29	17	8.8	3.6	1.0	0.1	0
10	100	78	58	40	25	14	6.7	2.4	0.5	–	0
11	100	77	56	37	22	12	5.1	1.6	0.3	–	0
12	100	76	53	34	20	9.8	3.9	1.1	0.2	–	0
13	100	75	51	32	18	8.2	3.0	0.8	0.1	–	0
14	100	73	49	30	16	6.9	2.3	0.5	0.1	–	0
15	100	72	47	28	14	5.8	1.8	0.4	–	–	0
16	100	71	46	26	12	4.9	1.4	0.3	–	–	0
17	100	70	44	24	11	4.1	1.1	0.2	–	–	0
18	100	69	43	23	10	3.5	0.8	0.1	–	–	0
19	100	68	41	21	9.0	2.9	0.7	0.1	–	–	0
20	100	67	40	20	8.1	2.5	0.5	0.1	–	–	0
25	100	63	34	15	4.8	1.1	0.2	–	–	–	0
30	100	60	29	11	2.9	0.5	–	–	–	–	0
35	100	57	25	8.0	1.7	0.2	–	–	–	–	0
40	100	54	22	6.0	1.1	0.1	–	–	–	–	0
45	100	51	19	4.5	0.6	–	–	–	–	–	0
		0	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**
50		100	73	49	30	16	8.0	3.4	1.3	0.4	0.1
60		100	70	45	25	13	5.4	2.0	0.6	0.2	–
70		100	68	41	22	9.7	3.7	1.2	0.3	0.1	–
80		100	66	38	18	7.5	2.5	0.7	0.1	–	–
90		100	64	35	16	5.9	1.7	0.4	0.1	–	–
100		100	62	32	14	4.6	1.2	0.2	–	–	–