УДК 519.2(075.8)
ББК 22.17я73
С47

Рецензенти:

д-р фіз.-мат. наук, проф. О.Г.Наконечний,
канд. фіз.-мат. наук, доц. В.А.Заславський

Рекомендовано до друку вченою радою факультету кібернетики (протокол № 2 від 16 жовтня 2006 року)

Слабоспицький О.С.

С47 Основи кореляційного аналізу даних: Навчальний посібник. – К.: ВПЦ "Київський університет", 2006. – 59 с.

ISBN 966-594-869-5

Висвітлено базові методи кореляційного аналізу, які дозволяють з'ясувати наявність статистичного зв'язку між змінними, що досліджуються. Розглянуто випадки аналізу кількісних, ординальних і номінальних змінних.

Для студентів, які вивчають дисципліну "Аналіз даних".

УДК 519.2(075.8)
ББК 22.17я73

ISBN 966-594-869-5

ВСТУП

Сучасні технології кожному надають доступ до неосяжних об'ємів різноманітної інформації. Але це не означає, що в подальшому отримані дані будуть кваліфіковано оброблені, а потім з результатів цієї обробки будуть зроблені правильні висновки.

Для розв'язання такого роду проблем можна скористатися математичними методами аналізу даних. Їх використанню прискореними темпами на практиці сприяла поява широкого кола відповідних програмних продуктів для сучасної обчислювальної техніки, які дозволяють обробляти інформацію незалежно від її природи. Це надало можливість ефективно розв'язувати різноманітні задачі в різних галузях, особливо в науці, бізнесі, медицині, фінансах, біології тощо.

Серед розділів аналізу даних, які найчастіше використовуються, можна назвати такі як: попередня обробка даних, кореляційний аналіз, дисперсійний аналіз, регресійний аналіз, коваріаційний аналіз, дискримінантний аналіз, кластерний аналіз, аналіз часових рядів.

Серед них чільне місце посідає кореляційний аналіз. Саме він дозволяє з'ясувати питання наявності істотного зв'язку між процесами (явищами, об'єктами), які досліджуються. Для цього, враховуючи тип даних, які обробляються, як правило, здійснюють такі кроки:

- обирають конкретну числову характеристику для парного чи множинного статистичного зв'язку;
- визначають оцінку значення характеристики, яка використовується;
- на основі оцінки значення характеристики парного чи множинного статистичного зв'язку приймають рішення, чи є істотним зв'язок між процесами (явищами, об'єктами), які досліджуються.

І тільки після ствердувальної відповіді про істотність статистичного зв'язку між змінними, що спостерігаються, має сенс переходити до наступного етапу – пошуку математичної моделі цього зв'язку засобами інших розділів аналізу даних.

У роботі висвітлено можливості, які надає кореляційний аналіз даних, при дослідженні наявності парних і множинних статистичних зв'язків. Окремо розглянуто випадки аналізу кількісних, ординальних і номінальних змінних.

Автор щиро вдячний студентам факультету кібернетики у сприянні покращенню даного посібника.

Усі зауваження та побажання щодо даної роботи будуть з вдячністю сприйняті автором. Їх можна надсилати електронною поштою (*sl@univ.kiev.ua*).

1. АНАЛІЗ НАЯВНОСТІ СТАТИСТИЧНОГО ЗВ'ЯЗКУ МІЖ КІЛЬКІСНИМИ ЗМІННИМИ

У цьому розділі увага буде зосереджена на дослідженні наявності статистичного зв'язку між кількісними змінними. Спочатку аналізуються скалярні залежна та незалежна змінні, потім випадок скалярної залежної змінної та вектора незалежних змінних. Із цією метою використовують відповідні парні та множинні характеристики статистичного зв'язку, які базуються на фундаментальній властивості функції регресії. Крім цього описується частинний коефіцієнт кореляції. У всіх випадках наводяться методики їх використання.

Як правило, характеристики статистичного зв'язку вводяться таким чином, щоб у разі рівності нулю, зв'язок був відсутній, а суттєвість зв'язку між змінними, що досліджуються, зростала зі збільшенням його відхилення від нуля. Таким чином, щоб з'ясувати істотність зв'язку за допомогою деякої характеристики статистичного зв'язку K, яка прийняла ненульове значення, достатньо перевірити гіпотезу про те, чи значимо відхиляється від нуля коефіцієнт K, тобто здійснити *перевірку* його *на значимість*. Це зводиться до перевірки гіпотези:

$$H_0 : K = 0,$$

з деяким рівнем значущості $\alpha > 0$.

Приклади задач кореляційного аналізу кількісних змінних:

- чи суттєво впливає рівень безробіття населення на рівень злочинності в країні?
- наскільки істотним є вплив на рівень життя населення основних показників економіки країни?
- чи залежить від оцінок, які отримав студент під час роботи протягом семестру, підсумковий результат у конкретного викладача з його предмету?
- чи істотним є зв'язок між температурою зовнішнього повітря та кількістю вжитих безалкогольних напоїв населенням?

1.1. Середньоквадратична апроксимація випадкової величини. Функція регресії та її властивості

Нехай на ймовірнісному просторі (Ω, F, P) розглядаються скалярна випадкова величина η і випадковий вектор $\vec{\xi} \in R^q$. Будемо інтерпретувати η, як залежну змінну, а $\vec{\xi}$ – як вектор незалежних змінних.

Припустимо, що $M\eta^2 < \infty$, тоді будуть існувати умовне математичне сподівання та умовна дисперсія випадкової величини η відносно події $\{\vec{\xi} = \vec{x}\}$, для яких у подальшому будуть використовуватися такі позначення:

$$f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x}), \qquad g(\vec{x}) = D(\eta / \vec{\xi} = \vec{x}).$$

Надалі буде корисне таке твердження.

Лема. Якщо η і $\vec{\xi}$ – випадкові величина та вектор, відповідно, а $M\eta^2 < \infty$, тоді для них справедливо:

$$D\eta = M\left\{\left(M\left(\eta / \vec{\xi}\right) - M\eta\right)^2\right\} + M\left\{M\left\{\left(\eta - M\left(\eta / \vec{\xi}\right)\right)^2 / \vec{\xi}\right\}\right\},$$

або скорочено $D\eta = Df(\vec{\xi}) + Mg(\vec{\xi})$.

З доведенням леми можна ознайомитися у додатку 2.

Означення. Нехай η і $\vec{\xi}$ – випадкові величина та вектор, відповідно, причому $M|\eta| < \infty$. Тоді *функцією регресії* η *на* $\vec{\xi}$ *(або функцією регресії* η *щодо* $\vec{\xi}$*)* називається функція

$$f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x}).$$

Функція регресії $f(\vec{x})$ має унікальну властивість, що підтверджує таке важливе твердження.

Теорема. Нехай $M\eta^2 < \infty$, Φ – множина борелівських функцій на R^q, тоді

$$f(\cdot) = \arg \min_{\varphi(\cdot) \in \Phi} M\left[\eta - \varphi\left(\vec{\xi}\right)\right]^2,$$

де $f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x})$.

Доведення цього фундаментального твердження викладено в додатку 2.

Наведена теорема дозволяє стверджувати, що найкращою в середньоквадратичному розумінні апроксимацією η на класі борелівських функцій від $\vec{\xi}$ є функція $f\left(\vec{\xi}\right)$, тобто за математичну модель можна взяти співвідношення

$$\eta = f\left(\vec{\xi}\right) + \varepsilon,$$

де ε – залишкова помилка апроксимації.

Для цієї моделі мають місце такі **властивості**:

1) $Mf\left(\vec{\xi}\right) = M\eta$, $M\varepsilon = 0$;
2) $f\left(\vec{\xi}\right)$ та ε – некорельовані;

3) $D\eta = Df(\vec{\xi}) + D\varepsilon$.

Зауваження. Перша властивість дозволяє запропонувати для останнього співвідношення ще одне представлення:

$$D\eta = Df(\vec{\xi}) + M\varepsilon^2.$$

Доведення. Скористаємося властивостями умовного математичного сподівання випадкової величини та доведемо послідовно кожну з властивостей:

1) $Mf(\vec{\xi}) = M\{M(\eta/\vec{\xi})\} = M\eta$. А це, у свою чергу, дозволяє стверджувати, що $M\varepsilon = M\{\eta - f(\vec{\xi})\} = M\eta - Mf(\vec{\xi}) = 0$;

2) оскільки, згідно з попередньою властивістю $M\varepsilon = 0$, то для доведення некорельованості достатньо впевнитися, що $M\left[f(\vec{\xi}) - Mf(\vec{\xi})\right]\varepsilon = 0$. Дійсно,

$$M\left[f(\vec{\xi}) - Mf(\vec{\xi})\right]\varepsilon = M\left[f(\vec{\xi}) - M\eta\right]\left[\eta - f(\vec{\xi})\right] =$$
$$= M\left\{M\left[\left(f(\vec{\xi}) - M\eta\right)\left(\eta - f(\vec{\xi})\right)/\vec{\xi}\right]\right\} =$$
$$= M\left\{\left(f(\vec{\xi}) - M\eta\right)M\left[\left(\eta - f(\vec{\xi})\right)/\vec{\xi}\right]\right\} = 0;$$

3) із доведеної в другій властивості некорельованості $f(\vec{\xi})$ та ε випливає: $D\eta = D\left(f(\vec{\xi}) + \varepsilon\right) = Df(\vec{\xi}) + D\varepsilon$.

Властивості доведено.

1.2. Індекс кореляції та його властивості. Коефіцієнт детермінації

Припустимо, що на ймовірнісному просторі (Ω, F, P) задано скалярну залежну змінну η і випадковий вектор незалежних змінних $\vec{\xi} \in R^q$. Введемо універсальну характеристику статистичного зв'язку для змінних η та $\vec{\xi}$. Будемо вважати, що $M\eta^2 < \infty$, тоді існуватиме функція регресії η на $\vec{\xi}$, а саме: $f(\vec{x}) = M(\eta/\vec{\xi} = \vec{x})$. А для випадкової величини η можна використовувати таку математичну модель: $\eta = f(\vec{\xi}) + \varepsilon$, де ε – залишкова помилка апроксимації. Причому результати з попереднього розділу дозволяють стверджувати, що мають місце такі властивості:

1) $Mf(\vec{\xi}) = M\eta$, $M\varepsilon = 0$;

2) $f(\vec{\xi})$ та ε – некорельовані;

3) $D\eta = Df(\vec{\xi}) + D\varepsilon$.

Після цих міркувань, враховуючи унікальну властивість функції регресії, доведену в теоремі (див. п. 1.1), цілком природним здається використання, як характеристики статистичного зв'язку для кількісних змінних η та $\vec{\xi}$, нижчевизначеного індексу кореляції.

Означення. Нехай η і $\vec{\xi}$ $\left(\vec{\xi} \in R^q\right)$ – випадкові величина та вектор, відповідно, причому $0 < M\eta^2 < \infty$. Тоді *(множинним) індексом кореляції* η *щодо* $\vec{\xi}$ називається величина

$$I_{\eta\vec{\xi}} = \sqrt{\frac{Df(\vec{\xi})}{D\eta}}.$$

Зауваження. Оскільки для дисперсії функції регресії справедливо $Df(\vec{\xi}) = D\eta - D\varepsilon$, звідси випливає, що для індексу кореляції $I_{\eta\vec{\xi}}$ можна отримати ще одне представлення:

$$I_{\eta\vec{\xi}} = \sqrt{1 - \frac{D\varepsilon}{D\eta}},$$

де $\varepsilon = \eta - f(\vec{\xi})$.

Властивості індексу кореляції:

1) $0 \le I_{\eta\vec{\xi}} \le 1$;

2) якщо $I_{\eta\vec{\xi}} = 0$, то відсутній вплив $\vec{\xi}$ на η;

3) якщо $I_{\eta\vec{\xi}} = 1$, то існує функціональний зв'язок між η та $\vec{\xi}$, а саме, з ймовірністю 1 справедливо $\eta = f(\vec{\xi})$.

(Тут і далі для випадкових величин/векторів рівності/нерівності вважаються справедливими з ймовірністю 1, якщо не наголошується на іншому.)

Доведення. Скористаємось властивостями умовного математичного сподівання та умовної дисперсії випадкової величини η відносно випадкового вектора $\vec{\xi}$ і послідовно доведемо потрібні властивості індексу кореляції:

1) згідно з вищенаведеними властивостями та враховуючи, що $D\varepsilon \ge 0$, отримаємо:

$$0 \le I_{\eta\vec{\xi}} = \sqrt{\frac{Df(\vec{\xi})}{D\eta}} = \sqrt{\frac{D\eta - D\varepsilon}{D\eta}} \le 1;$$

2) припустимо, що $I_{\eta\vec{\xi}} = 0$. Тоді

$$Df(\vec{\xi}) = 0 \Rightarrow f(\vec{\xi}) = Mf(\vec{\xi}) \Rightarrow f(\vec{\xi}) = M\eta = \text{const}.$$

Звідси випливає, що функція регресії, за допомогою якої апроксимується η, не залежить від значень свого аргументу – $\vec{\xi}$;

3) нехай $I_{\eta\vec{\xi}} = 1$, тоді

$$1 = I_{\eta\vec{\xi}} = \sqrt{1 - \frac{D\varepsilon}{D\eta}} \Rightarrow D\varepsilon = 0.$$

Остання умова еквівалентна наступній: $0 = D\varepsilon = M\left\{\eta - f\left(\vec{\xi}\right)\right\}^2$. А це, у свою чергу, дозволяє стверджувати, що з ймовірністю 1 має місце така функціональна залежність між η та $\vec{\xi}$: $\eta = f\left(\vec{\xi}\right)$.

Справедливість властивостей доведено.

Поряд з індексом кореляції $I_{\eta\vec{\xi}}$, використовується також характеристика, яку визначає таке визначення.

Означення. Нехай η і $\vec{\xi}$ $\left(\vec{\xi} \in R^q\right)$ – випадкові величина та вектор, відповідно, причому $0 < M\eta^2 < \infty$. Тоді *(множинним) коефіцієнтом детермінації* η *щодо* $\vec{\xi}$ називається величина

$$I^2_{\eta\vec{\xi}} = \frac{Df(\vec{\xi})}{D\eta}.$$

Зауваження 1. Коефіцієнт детермінації $I^2_{\eta\vec{\xi}}$ є більш привабливішим порівняно з індексом кореляції, тому що має більш прозору інтерпретацію, а саме: він вказує, яка частина дисперсії змінної η визначається варіацією функції регресії $f(\vec{\xi})$.

Зауваження 2. Аналогічно, для коефіцієнта детермінації також має інше представлення:

$$I^2_{\eta\vec{\xi}} = 1 - \frac{D\varepsilon}{D\eta}.$$

У підсумку, для кількісних змінних, випадкової величини η $(0 < M\eta^2 < \infty)$ і випадкового вектора $\vec{\xi}$ $\left(\vec{\xi} \in R^q\right)$, введено теоретичну характеристику статистичного зв'язку – індекс кореляції $I_{\eta\vec{\xi}}$ (коефіцієнт детермінації $I^2_{\eta\vec{\xi}}$).

1.3. Аналіз парних статистичних зв'язків кількісних змінних

При дослідженні зв'язку між двома скалярними змінними маємо справу з так званим *парним статистичним зв'язком*. Нехай на ймовірнісному просторі (Ω, F, P) розглядаються дві скалярні випадкові величини: η – залежна змінна, ξ – незалежна змінна. Потрібно з'ясувати, чи є істотним зв'язок з статистичної точки зору, між цими змінними з рівнем значущості $\alpha > 0$.

Якщо припустити, що $0 < M\eta^2 < \infty$, то можна скористатися в загальному випадку для аналізу статистичного зв'язку між змінними η та ξ універсальною характеристикою – (парним) індексом кореляції η щодо ξ:

$$I_{\eta\xi} = \sqrt{\frac{Df(\xi)}{D\eta}},$$

де $f(x) = M(\eta / \xi = x)$.

Для залежної змінної η можна використовувати представлення $\eta = f(\xi) + \varepsilon$, де ε – залишкова помилка апроксимації, для якої справедливі:

1) $Mf(\xi) = M\eta$, $M\varepsilon = 0$;
2) $f(\xi)$ та ε – некорельовані;
3) $D\eta = Df(\xi) + D\varepsilon$.

Остання властивість дозволяє використовувати також таке представлення для індексу кореляції η щодо ξ:

$$I_{\eta\xi} = \sqrt{1 - \frac{D\varepsilon}{D\eta}}.$$

Усі властивості індексу кореляції (див. розд. 1.2) залишаються в силі для $I_{\eta\xi}$:

1) $0 \le I_{\eta\xi} \le 1$;
2) якщо $I_{\eta\xi} = 0$, то відсутній вплив ξ на η;
3) якщо $I_{\eta\xi} = 1$, то існує функціональний зв'язок між η та ξ, а саме: з ймовірністю 1 справедливо $\eta = f(\xi)$.

Відповідно, (парний) коефіцієнт детермінації η щодо ξ обчислюється згідно з виразами

$$I^2_{\eta\xi} = \frac{Df(\xi)}{D\eta} = 1 - \frac{D\varepsilon}{D\eta}.$$

Таким чином, маємо для скалярних кількісних змінних η $(0 < M\eta^2 < \infty)$ і ξ теоретичну характеристику парного статистичного зв'язку – індекс кореляції $I_{\eta\xi}$ (коефіцієнт детермінації $I^2_{\eta\xi}$).

1.3.1. Вибіркові значення характеристики парного статистичного зв'язку в загальному випадку

Для практичного використання теоретичних показників кореляційного зв'язку потрібно вміти обчислювати їх вибіркові (емпіричні) значення, на основі отриманих спостережень над кількісними змінними, які досліджуються. Процедура побудови відповідної оцінки для індексу кореляції, у подальшому, буде ґрунтуватися на його теоретичному визначенні у вигляді

$$I_{\eta\xi} = \sqrt{\frac{Df(\xi)}{D\eta}} = \sqrt{1 - \frac{D(\eta - f(\xi))}{D\eta}}.$$

Обчислення конкретних емпіричних значень індексу кореляції (коефіцієнта детермінації) буде залежати в кожному випадку від умов проведення експерименту. Перейдемо до розгляду цих ситуацій.

1.3.1.1. Кореляційне відношення

Розглянемо спочатку випадок, коли проведено групування n спостережень над незалежною змінною ξ. Нехай отримано s інтервалів групування, причому виміри над залежною змінною η, які відповідають спостереженням над незалежною змінною з i-го інтервалу групування позначатимемо як $y_{i1}, y_{i2}, \ldots, y_{i\nu_i}$, $i = \overline{1,s}$. Тут ν_i – кількість вимірів, які потрапили в i-й інтервал групування $\left(n = \sum_{i=1}^{s} \nu_i\right)$.

Зауваження. У подальшому скрізь, як правило, в оцінках різних характеристик будемо опускати аргумент n, який вказує на об'єм вибірки, якщо кількість спостережень незмінна.

Скористаємося позначенням для загального середнього:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{s} \nu_i \bar{y}_{i\bullet},$$

де $\bar{y}_{i\bullet}$ – середнє на i-му інтервалі групування, яке обчислюється таким чином:

$$\bar{y}_{i\bullet} = \frac{1}{\nu_i}\sum_{j=1}^{\nu_i} y_{ij}, \quad i = \overline{1,s}.$$

Зауважимо, що крапка в $\bar{y}_{i\bullet}$ є позначенням сумування за тим індексом, замість якого вона фігурує. Зважаючи на те, що $D\eta$ можна оцінити згідно з

$$s^2_{\eta} = \frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{\nu_i}\left(y_{ij} - \bar{y}\right)^2,$$

а $Df(\xi)$ – відповідно, за формулою $s^2_f = \frac{1}{n}\sum_{i=1}^{s} \nu_i \left(\bar{y}_{i\bullet} - \bar{y}\right)^2$, отримаємо таку оцінку для індексу кореляції η щодо ξ:

$$\hat{\rho}_{\eta\xi} = \sqrt{\frac{s^2_f}{s^2_{\eta}}}.$$

Оцінку $\hat{\rho}_{\eta\xi}$ називають *парним кореляційним відношенням* η *щодо* ξ.

Методика використання $\hat{\rho}_{\eta\xi}$ є традиційною:

1) якщо $\hat{\rho}_{\eta\xi} = 0$, то вважаємо, що зв'язок між η і ξ відсутній;
2) якщо $\hat{\rho}_{\eta\xi} = 1$, то вважаємо, що зв'язок між η і ξ функціональний;
3) якщо $\begin{cases}\hat{\rho}_{\eta\xi} \neq 0\\ \hat{\rho}_{\eta\xi} \neq 1\end{cases}$, то істотність зв'язку з'ясовується шляхом перевірки на значимість індексу кореляції η щодо ξ:

$$H_0 : I_{\eta\xi} = 0,$$

з деяким рівнем значущості $\alpha > 0$. Якщо гіпотеза не справедлива, то статистичний зв'язок між η і ξ вважається істотним, інакше зв'язок вважається не істотним.

Для перевірки останньої гіпотези скористаємося тим, що при справедливості гіпотези H_0, можна вважати, що розподіл статистики

$$F = \frac{\hat{\rho}^2_{\eta\xi}}{1 - \hat{\rho}^2_{\eta\xi}} \cdot \frac{n-s}{s-1}$$

можна наблизити F-розподілом з $(s-1)$ та $(n-s)$ ступенями свободи. Тоді, з огляду на структуру статистики F, як критичну область для гіпотези H_0 потрібно взяти область великих значень, а відповідна область прийняття для нашої гіпотези матиме вигляд

$$F < F_{\alpha}(s-1, n-s),$$

де $F_{\alpha}(m,n)$ – $100\alpha\%$-на точка F-розподілу з m та n ступенями свободи.

1.3.1.2. Оцінка індексу кореляції

Розглянемо тепер ситуацію, коли спостереження відповідних змінних

$$\eta: y_1, y_2, \ldots, y_n, \qquad \xi: x_1, x_2, \ldots, x_n,$$

є доступними, а також є можливість припустити, що функція регресії η на ξ може бути наближена на деякому класі параметричних функцій $f(x,\lambda),\ \lambda \in R^p$.

Якщо отримана оцінка $\hat{\lambda}$ для вектора невідомих параметрів λ деяким відомим методом, то вибіркове значення $D\varepsilon$ можна обчислити таким чином:

$$\frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - f\left(x_i,\hat{\lambda}\right)\right)^2,$$

а *емпіричне (вибіркове) значення індексу кореляції* набуде вигляду

$$\hat{I}_{\eta\xi} = \sqrt{1 - \frac{\frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - f\left(x_i,\hat{\lambda}\right)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}},$$

де $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Для практичного використання оцінки $\hat{I}_{\eta\xi}$ не вистачає тільки процедури *перевірки* індексу кореляції *на значимість*, тобто гіпотези

$$H_0: I_{\eta\xi} = 0$$

з деяким рівнем значущості $\alpha > 0$. Для цього скористаємося тим фактом, що при справедливості гіпотези H_0 розподіл статистики

$$F = \frac{\hat{I}^2_{\eta\xi}}{1-\hat{I}^2_{\eta\xi}} \cdot \frac{n-p}{p-1}$$

можна апроксимувати F-розподілом з $(p-1)$ та $(n-p)$ ступенями свободи. Тоді область прийняття для гіпотези H_0 матиме вигляд:

$$F < F_\alpha(p-1, n-p),$$

де $F_\alpha(m,n)$ – $100\alpha\%$-на точка F-розподілу з m та n ступенями свободи.

У розглянутому випадку для індексу кореляції отримано формулу емпіричного значення та процедуру перевірки на значиме відхилення від нуля.

1.3.2. Коефіцієнт кореляції – характеристика парного статистичного зв'язку у нормальному випадку

Припустимо, що η і ξ – нормально розподілені випадкові величини, задані на ймовірнісному просторі (Ω, F, P), а саме:

$$\eta \sim N\left(m_\eta, \sigma^2_\eta\right),\ \xi \sim N\left(m_\xi, \sigma^2_\xi\right),\ \sigma^2_\eta, \sigma^2_\xi > 0.$$

Перш ніж проводити обчислення емпіричного значення характеристики парного статистичного зв'язку для нормальних змінних, проведемо аналіз теоретичної характеристики. Виявляється, що з'ясування питання про істотність зв'язку між змінними η та ξ з рівнем значущості $\alpha > 0$ спрощується. Дійсно, за цих припущень індекс кореляції $I_{\eta\xi}$ легко виражається через *(парний) коефіцієнт кореляції* $r_{\eta\xi}$ *випадкових величин* η *та* ξ, який обчислюється як:

$$r_{\eta\xi} = \frac{\operatorname{cov}(\eta,\xi)}{\sigma_\eta \sigma_\xi},$$

де $\operatorname{cov}(\eta,\xi) = M\left(\eta - m_\eta\right)\left(\xi - m_\xi\right)$ – *коефіцієнт коваріації випадкових величин* η *та* ξ.

У нормальному випадку для функцій $f(x) = M(\eta/\xi = x)$, $g(x) = D(\eta/\xi = x)$ справедливе твердження.

Теорема. Нехай пара (η,ξ) має гауссівський розподіл і $\sigma^2_\xi > 0$, тоді справедливо

$$f(x) = m_\eta + \frac{\operatorname{cov}(\eta,\xi)}{\sigma^2_\xi}\left(x - m_\xi\right), \quad g(x) = \sigma^2_\eta - \frac{\operatorname{cov}^2(\eta,\xi)}{\sigma^2_\xi}.$$

Якщо ж припустити, що $\sigma^2_\eta > 0$, то має місце

$$f(x) = m_\eta + r_{\eta\xi}\frac{\sigma_\eta}{\sigma_\xi}\left(x - m_\xi\right), \qquad g(x) = \sigma^2_\eta\left(1 - r^2_{\eta\xi}\right).$$

Остання теорема дозволяє переконатися, що має місце таке твердження.

Теорема. Нехай η і ξ – нормально розподілені випадкові величини: $\eta \sim N\left(m_\eta, \sigma^2_\eta\right),\ \xi \sim N\left(m_\xi, \sigma^2_\xi\right),\ \sigma^2_\eta, \sigma^2_\xi > 0$. Тоді індекс кореляції $I_{\eta\xi} = \left|r_{\eta\xi}\right|$, а відповідно, коефіцієнт детермінації $I^2_{\eta\xi} = r^2_{\eta\xi}$.

Доведення. Згідно з попередньою теоремою індекс кореляції $I_{\eta\xi}$ можна обчислити як:

$$I_{\eta\xi}=\sqrt{\frac{Df(\xi)}{D\eta}}=\sqrt{\frac{D\left(m_\eta+r_{\eta\xi}\frac{\sigma_\eta}{\sigma_\xi}\left(\xi-m_\xi\right)\right)}{\sigma_\eta^2}}=\sqrt{\frac{r_{\eta\xi}^2\frac{\sigma_\eta^2}{\sigma_\xi^2}D\left(\xi-m_\xi\right)}{\sigma_\eta^2}}=\left|r_{\eta\xi}\right|.$$

Тоді представлення для коефіцієнта детермінації у вигляді $I_{\eta\xi}^2=r_{\eta\xi}^2$ стає очевидним. Доведення завершено.

У результаті, у нормальному випадку як характеристику парного статистичного зв'язку можна використовувати не індекс кореляції $I_{\eta\xi}$, а звичайний коефіцієнт кореляції $r_{\eta\xi}$. Його **властивості** загальновідомі:

1) $-1\le r_{\eta\xi}\le 1$;
2) якщо $r_{\eta\xi}=0$, то η не залежить від ξ;
3) якщо $\left|r_{\eta\xi}\right|=1$, то з ймовірністю 1 існує лінійний зв'язок між η та ξ, а саме, справедливо

$$\eta=m_\eta+r_{\eta\xi}\frac{\sigma_\eta}{\sigma_\xi}\left(\xi-m_\xi\right);$$

4) $r_{\eta\xi}=r_{\xi\eta}$.

Доведення. Для перевірки їх справедливості скористаємося властивостями індексу кореляції:

1) дійсно, з першої властивості $I_{\eta\xi}$ випливає

$$1\ge I_{\eta\xi}=\left|r_{\eta\xi}\right|\Rightarrow -1\le r_{\eta\xi}\le 1;$$

2) оскільки $r_{\eta\xi}=0$, то $I_{\eta\xi}=0$. Згідно з другою властивістю індексу кореляції це означає, що η не залежить від ξ;

3) так як $\left|r_{\eta\xi}\right|=1$, то $I_{\eta\xi}=1$. Тоді третя властивість $I_{\eta\xi}$ дозволяє стверджувати, що з ймовірністю 1 справедливо $\eta=f\left(\xi\right)$. Але в нормальному випадку $f\left(x\right)=m_\eta+r_{\eta\xi}\frac{\sigma_\eta}{\sigma_\xi}\left(x-m_\xi\right)$. Властивість доведено;

4) очевидно.

Доведення завершено.

Зауваження. Вид функції регресії $f\left(x\right)$, який наведено в теоремі вказує на те, що зв'язок між η і ξ має монотонний характер, а знак коефіцієнта кореляції $r_{\eta\xi}$ конкретизує його, а саме: якщо $r_{\eta\xi}>0$ $\left(r_{\eta\xi}<0\right)$, то функція залежності буде зростаючою (спадною).

Перейдемо до практичного використання коефіцієнта кореляції $r_{\eta\xi}$ як характеристики парного статистичного зв'язку для скалярних кількісних змінних. Нехай спостереження над випадковими величинами

$$\eta: y_1, y_2, \ldots, y_n, \qquad \xi: x_1, x_2, \ldots, x_n$$

є доступними. Тоді, на основі цих вимірів *вибіркове (емпіричне) значення* $\hat{r}_{\eta\xi}$ *(парного) коефіцієнта кореляції* $r_{\eta\xi}$ обчислюється згідно з

$$\hat{r}_{\eta\xi}=\frac{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)\left(x_i-\bar{x}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}},$$

де $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$.

Методика аналізу статистичного зв'язку між змінними η та ξ така:

1) якщо $\hat{r}_{\eta\xi}=0$, то вважаємо, що зв'язок між η та ξ відсутній;
2) якщо $\left|\hat{r}_{\eta\xi}\right|=1$, то зв'язок між η та ξ функціональний (точніше лінійний) і має вигляд

$$\eta=m_\eta+r_{\eta\xi}\frac{\sigma_\eta}{\sigma_\xi}\left(\xi-m_\xi\right);$$

3) якщо $\left\{\begin{matrix}\hat{r}_{\eta\xi}\neq 0\\ \left|\hat{r}_{\eta\xi}\right|\neq 1\end{matrix}\right|$, то потрібно перевірити гіпотезу про те, чи суттєво відхиляється від нуля коефіцієнт кореляції $r_{\eta\xi}$ із статистичної точки зору, тобто здійснити *перевірку* його *на значимість.* Для цього потрібно перевірити гіпотезу:

$$H_0: r_{\eta\xi}=0,$$

з деяким рівнем значущості $\alpha>0$. І якщо гіпотеза виявиться справедливою, то вважається, що статистичний зв'язок між η та ξ є не істотним, у протилежному випадку – істотним.

Конкретизуємо процедуру перевірки останньої гіпотези. Скористаємося тим фактом, що при справедливості гіпотези H_0, статистика

$$t(n) = \frac{\sqrt{n-2}\ \hat{r}_{\eta\xi}}{\sqrt{1-\hat{r}^2_{\eta\xi}}}$$

має t-розподіл Стьюдента з $(n-2)$ ступенями свободи. Тоді логічно за критичну область цієї гіпотези взяти області набуття статистикою своїх екстремальних значень. А область прийняття гіпотези H_0 матиме представлення:

$$|t(n)| < t_{\frac{\alpha}{2}}(n-2),$$

де $t_\alpha(n)$ – $100\alpha\,\%$-на точка t-розподілу Стьюдента з n ступенями свободи.

Зауваження. Скрізь у подальшому, як правило, будемо вважати справедливим припущення про нормальність відповідних спостережень у разі необхідності перевірки гіпотез або побудови довірчих інтервалів, якщо не обумовлено інше.

Таким чином, у нормальному випадку за характеристику парного статистичного зв'язку можна використовувати звичайний коефіцієнт кореляції $r_{\eta\xi}$, який має прозору інтерпретацію й дозволяє стверджувати про наявність лінійного зв'язку, коли він набуває значення ± 1. У загальному випадку коефіцієнт кореляції вже не має такої яскравої інтерпретації і тому виникає потреба у зверненні до універсальної характеристики парного статистичного зв'язку – індексу кореляції $I_{\eta\xi}$.

1.4. Аналіз множинних статистичних зв'язків кількісних змінних

Будемо стверджувати, що маємо справу з *множинним статистичним зв'язком,* якщо досліджується зв'язок між більше, ніж двома скалярними змінними. До цього розглядалися зв'язки тільки між двома скалярними змінними, тобто аналізувалися парні статистичні зв'язки.

Нехай на ймовірнісному просторі (Ω, F, P) задані скалярна залежна змінна η і вектор незалежних змінних $\vec{\xi} \in R^q$ $\left(\vec{\xi}^T = (\xi_1, \xi_2, \ldots, \xi_q)\right)$. Проведемо дослідження питання про істотність зв'язку між змінними η та $\vec{\xi}$, з статистичної точки зору, з рівнем значущості $\alpha > 0$.

Припустимо, що $0 < M\eta^2 < \infty$. Тоді, якщо відсутня додаткова апріорна інформація, можна скористатися множинним індексом кореляції η щодо $\vec{\xi}$, який було введено раніше:

$$I_{\eta\vec{\xi}} = \sqrt{\frac{Df(\vec{\xi})}{D\eta}} = \sqrt{1 - \frac{D\left(\eta - f(\vec{\xi})\right)}{D\eta}},$$

де $f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x})$. Зауважимо, що залежна змінна η може бути представлена у вигляді $\eta = f\left(\vec{\xi}\right) + \varepsilon$, де ε – помилка моделі. Причому, як було доведено, мають місце властивості:

1) $Mf(\vec{\xi}) = M\eta,\ M\varepsilon = 0$;
2) $f\left(\vec{\xi}\right)$ та ε – некорельовані;
3) $D\eta = Df(\vec{\xi}) + D\varepsilon$.

А для індексу кореляції $I_{\eta\vec{\xi}}$, у свою чергу, справедливо:

1) $0 \le I_{\eta\vec{\xi}} \le 1$;
2) якщо $I_{\eta\vec{\xi}} = 0$, то відсутній вплив $\vec{\xi}$ на η;
3) якщо $I_{\eta\vec{\xi}} = 1$, то існує функціональний зв'язок між η та $\vec{\xi}$ у вигляді $\eta = f\left(\vec{\xi}\right)$.

У результаті, у загальному випадку можна використовувати для аналізу статистичного зв'язку між скалярною залежною змінною η і вектором незалежних змінних $\vec{\xi}$ теоретичну характеристику множинного статистичного зв'язку – індекс кореляції $I_{\eta\vec{\xi}}$.

1.4.1. Емпіричні значення характеристики множинного статистичного зв'язку в загальному випадку

Як і при аналізі парних зв'язків, обчислення вибіркового значення індексу кореляції $I_{\eta\vec{\xi}}$ залежатиме від конкретної ситуації під час обробки даних. Окремо розглянемо випадки обробки згрупованих спостережень і наявності можливості апроксимації функції регресії на деякому класі параметричних функцій.

1.4.1.1. Обробка згрупованих даних

Нехай отримано n спостережень над скалярною випадковою величиною η і випадковим вектором $\vec{\xi}$ $\left(\vec{\xi}^T = (\xi_1, \xi_2, \ldots, \xi_q)\right)$. Проведемо групування за вектором незалежних змінних $\vec{\xi}$. Якщо раніше у випадку однієї

незалежної змінної під час групування її область значень розбивалася на інтервали групування, то у векторному випадку область значень $\vec{\xi}$ розбивається на їх аналоги – гіперпаралелепіпеди групування. Якщо область значень змінної ξ_i розбилася на s_i $\left(i=\overline{1,q}\right)$ інтервалів, то отримуємо $s=\prod_{i=1}^{q} s_i$ гіперпаралелепіпедів групування за вектором незалежних змінних $\vec{\xi}$. Тоді спостереження над залежною змінною η, які відповідають вимірам над вектором незалежних змінних $\vec{\xi}$ з i-го гіперпаралелепіпеда групування будемо позначати як $y_{i1}, y_{i2}, \ldots, y_{i\nu_i}$, $i=\overline{1,s}$, де ν_i – кількість вимірів, які потрапили в i-й гіперпаралелепіпед групування $\left(n=\sum_{i=1}^{s}\nu_i\right)$.

У подальшому процедура визначення оцінки індексу кореляції буде подібною до випадку скалярної незалежної змінної. Введемо позначення для загального середнього

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{s}\nu_i\overline{y}_{i\bullet}\text{, де }\overline{y}_{i\bullet}=\frac{1}{\nu_i}\sum_{j=1}^{\nu_i}y_{ij},\ i=\overline{1,s}.$$

Тоді для підрахування оцінок $D\eta$ та $Df\left(\vec{\xi}\right)$ можна скористатися такими виразами відповідно:

$$s_\eta^2=\frac{1}{n}\sum_{i=1}^{s}\sum_{j=1}^{\nu_i}\left(y_{ij}-\overline{y}\right)^2,\ s_f^2=\frac{1}{n}\sum_{i=1}^{s}\nu_i\left(\overline{y}_{i\bullet}-\overline{y}\right)^2.$$

Це дає змогу обчислювати вибіркове значення індексу кореляції $I_{\eta\vec{\xi}}$ таким чином:

$$\hat{\rho}_{\eta\vec{\xi}}=\sqrt{\frac{s_f^2}{s_\eta^2}}.$$

Аналогічно до вже розглянутих парних кореляційних відношень, оцінку $\hat{\rho}_{\eta\vec{\xi}}$ будемо називати *множинним кореляційним відношенням* η *щодо* ξ.

Процедура використання статистики $\hat{\rho}_{\eta\vec{\xi}}$ є ідентичною до методики використання парного кореляційного відношення.

1.4.1.2. Використання наближення функції регресії

Припустимо, що функцію регресії η на $\vec{\xi}$ можна апроксимувати на класі параметричних функцій $f\left(\vec{x},\lambda\right)$, $\vec{x}\in R^q$, $\lambda\in R^p$. Нехай під час

аналізу можна використовувати такі спостереження над залежною змінною η і вектором незалежних змінних $\vec{\xi}$:

$$\eta: y_1, y_2, \ldots, y_n, \qquad \vec{\xi}: \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n.$$

Наявність оцінки $\hat{\lambda}$ для вектора параметрів λ дозволяє підрахувати *емпіричне (вибіркове) значення індексу кореляції* $I_{\eta\vec{\xi}}$ таким чином:

$$\hat{I}_{\eta\vec{\xi}}=\sqrt{1-\frac{\frac{1}{n-p}\sum_{i=1}^{n}\left(y_i-f\left(\vec{x}_i,\hat{\lambda}\right)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}},$$

де $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

Методика використання вибіркового значення $\hat{I}_{\eta\vec{\xi}}$ у множинному випадку не чим не відрізняється від відповідної процедури в парному випадку. *Перевірка* індексу кореляції *на значимість*

$$H_0: I_{\eta\vec{\xi}}=0,$$

з деяким рівнем значущості $\alpha>0$, здійснюється за допомогою статистики

$$F=\frac{\hat{I}^2_{\eta\vec{\xi}}}{1-\hat{I}^2_{\eta\vec{\xi}}}\cdot\frac{n-p}{p-1},$$

розподіл якої можна наблизити F-розподілом з $\left(p-1\right)$ та $\left(n-p\right)$ ступенями свободи. Це дозволяє область прийняття для гіпотези H_0 записати у вигляді

$$F<F_\alpha\left(p-1,n-p\right),$$

де $F_\alpha\left(m,n\right)$ – $100\alpha\,\%$-на точка F-розподілу з m та n ступенями свободи.

1.4.2. Обробка спостережень у нормальному випадку

Розглянемо на ймовірнісному просторі $\left(\Omega,F,P\right)$ нормально розподілені η та $\vec{\xi}^T=\left(\xi_1,\xi_2,\ldots,\xi_q\right)$ з параметрами:

$$M\eta=m_\eta,\ D\eta=\sigma_\eta^2;\ M\vec{\xi}=\vec{m}_{\vec{\xi}},\ M\left(\vec{\xi}-\vec{m}_{\vec{\xi}}\right)\left(\vec{\xi}-\vec{m}_{\vec{\xi}}\right)^T=\Sigma_{\vec{\xi}\vec{\xi}},$$

або скорочено $\eta\sim N\left(m_\eta,\sigma_\eta^2\right)$, $\vec{\xi}\sim N\left(\vec{m}_{\vec{\xi}},\Sigma_{\vec{\xi}\vec{\xi}}\right)$, $\sigma_\eta^2>0,\Sigma_{\vec{\xi}\vec{\xi}}>0$.

Для уніфікації позначень будемо у подальшому використовувати позначення ξ_0 для змінної η у разі потреби, тобто $\xi_0 \equiv \eta$. Аналіз наявності статистичного зв'язку між гауссівськими змінними η та $\vec{\xi}$ має свою специфіку. Дійсно, індекс кореляції $I_{\eta\vec{\xi}}$ за цих припущень виражається через множинний коефіцієнт кореляції $r_{\eta\vec{\xi}}$ випадкових змінних η та $\vec{\xi}$. А з'ясування наявності залежності для пари змінних (η, ξ_i), призводить до необхідності врахування впливу на них інших не врахованих змінних ξ_j, $j \neq i$. Останнє дозволяють зробити частинні коефіцієнти кореляції, які вводяться далі.

1.4.2.1. Частинний коефіцієнт кореляції та його властивості

Практика дослідження парного статистичного зв'язку між скалярними змінними ξ_i та ξ_j вказує, що звичайний коефіцієнт кореляції $r_{\xi_i\xi_j}$ (або скорочено r_{ij}) може неадекватно віддзеркалювати наявний зв'язок, коли існують інші змінні, наприклад ξ_l, $l \in \{0,1,2,\ldots,q\} \setminus \{i,j\}$, які мають суттєвий вплив як на ξ_i, так і ξ_j. Тому виникає бажання удосконалити звичайний коефіцієнт кореляції $r_{\xi_i\xi_j}$ таким чином, щоб модифікована характеристика парного статистичного зв'язку була не чутлива до впливу сторонніх змінних.

Для цього запропоновано підраховувати коефіцієнт кореляції для ξ_i та ξ_j не за їх сумісним розподілом, як це робилося раніше для парного коефіцієнта кореляції $r_{\xi_i\xi_j}$, а за їх умовним сумісним розподілом, де в умові сторонні змінні, наприклад ξ_l, $l \in \{0,1,2,\ldots,q\} \setminus \{i,j\}$, фіксуються на певному рівні.

Означення. *Частинним коефіцієнтом кореляції випадкових величин ξ_i та ξ_j при сторонніх змінних ξ_l, $l \in I$* називається числова характеристика, яка підрахована за формулою парного коефіцієнта кореляції для ξ_i та ξ_j, але за умовним сумісним розподілом випадкових величин ξ_i та ξ_j за умови, що змінні ξ_l, $l \in I$ набувають фіксованих значень, де I – множина індексів сторонніх змінних. Позначення – $r_{ij(I)}$.

Зауваження. Множина I не містить у собі індекси i та j.

У загальному випадку, така модифікована характеристика парного статистичного зв'язку для ξ_i та ξ_j буде залежати від постійних значень,

яких набувають сторонні змінні, і тому породжує певні труднощі в її використанні. Але з'ясувалося, що у нормальному випадку частинний коефіцієнт кореляції для випадкових величин ξ_i та ξ_j не залежить від фіксованих значень, яких набувають інші змінні ξ_l, $l \in I$, і тому знайшов своє широке застосування.

Якщо випадкові величини $\xi_0, \xi_1, \xi_2, \ldots, \xi_q$ мають невироджений сумісний гауссівський розподіл ($\xi_0 \equiv \eta$), а множина індексів сторонніх змінних $I = \{0,1,2,\ldots,q\} \setminus \{i,j\}$, то частинний коефіцієнт кореляції $r_{ij(I)}$ можна обчислити згідно з

$$r_{ij(I)} = -\frac{R_{ij}}{\sqrt{R_{ii}R_{jj}}},$$

де R_{kl} – алгебраїчне доповнення до елемента r_{kl} у матриці парних коефіцієнтів кореляції для змінних $\xi_0, \xi_1, \xi_2, \ldots, \xi_q$, а саме:

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0q} \\ r_{10} & 1 & r_{12} & \cdots & r_{1q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{q0} & r_{q1} & r_{q2} & \cdots & 1 \end{pmatrix}.$$

Зауваження. У позначенні $r_{ij(I)} = r_{ij(\{0,1,2,\ldots,q\}\setminus\{i,j\})}$ у дужках вказується множина індексів усіх сторонніх змінних, зафіксованих на певних рівнях.

Якщо підрахувати частинний коефіцієнт кореляції для ξ_0 та ξ_1, зафіксувавши в умові лише значення змінної ξ_2, то матимемо

$$r_{01(2)} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{\left(1-r_{02}^2\right)\left(1-r_{12}^2\right)}}.$$

Множину сторонніх змінних можна розширювати, послідовно додаючи їх по одній. Це дозволяє для обчислення частинного коефіцієнта кореляції $r_{01(2,3,\ldots,q)}$ запропонувати *рекурентну процедуру*

$$r_{01(2,3,\ldots,k)} = \frac{r_{01(2,3,\ldots,k-1)} - r_{0k(2,3,\ldots,k-1)}r_{1k(2,3,\ldots,k-1)}}{\sqrt{\left(1-r_{0k(2,3,\ldots,k-1)}^2\right)\left(1-r_{1k(2,3,\ldots,k-1)}^2\right)}},\ k = \overline{3,q}.$$

Зауваження. З огляду на означення $r_{ij(I)}$, частинного коефіцієнта кореляції випадкових величин ξ_i та ξ_j при сторонніх змінних ξ_l, $l \in I$, можна

стверджувати, що його властивості збігаються з властивостями парного коефіцієнта кореляції для випадкових величин ξ_i та ξ_j, тобто r_{ij}.

Для обчислення емпіричного значення частинного коефіцієнта кореляції $\hat{r}_{ij(I)}$, $I = \{0,1,2,\ldots,q\} \setminus \{i,j\}$ можна скористатися виразом

$$\hat{r}_{ij(I)} = -\frac{\hat{R}_{ij}}{\sqrt{\hat{R}_{ii}\hat{R}_{jj}}},$$

де $\hat{R}_{kl}$ – алгебраїчне доповнення до елемента $\hat{r}_{kl}$ у матриці вибіркових парних коефіцієнтів кореляції для змінних $\xi_0, \xi_1, \xi_2, \ldots, \xi_q$, а саме:

$$\hat{R} = \begin{pmatrix} 1 & \hat{r}_{01} & \hat{r}_{02} & \ldots & \hat{r}_{0q} \\ \hat{r}_{10} & 1 & \hat{r}_{12} & \ldots & \hat{r}_{1q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{r}_{q0} & \hat{r}_{q1} & \hat{r}_{q2} & \ldots & 1 \end{pmatrix}.$$

Рекурентний алгоритм для обчислення емпіричного значення частинного коефіцієнта кореляції $\hat{r}_{01(2,3,\ldots,q)}$ набуває вигляду

$$\hat{r}_{01(2,3,\ldots,k)} = \frac{\hat{r}_{01(2,3,\ldots,k-1)} - \hat{r}_{0k(2,3,\ldots,k-1)}\hat{r}_{1k(2,3,\ldots,k-1)}}{\sqrt{\left(1-\hat{r}^2_{0k(2,3,\ldots,k-1)}\right)\left(1-\hat{r}^2_{1k(2,3,\ldots,k-1)}\right)}}, \quad k = \overline{3,q},$$

з початковим значенням

$$\hat{r}_{01(2)} = \frac{\hat{r}_{01} - \hat{r}_{02}\hat{r}_{12}}{\sqrt{\left(1-\hat{r}^2_{02}\right)\left(1-\hat{r}^2_{12}\right)}}.$$

Нехай I – деяка множина з m індексів сторонніх змінних, які зафіксовані на деякому рівні при обчисленні частинного коефіцієнта кореляції $r_{ij(I)}$ для змінних ξ_i та ξ_j. (Очевидно, що множина I не містить у собі індекси i та j).

Методика аналізу статистичного зв'язку між змінними ξ_i та ξ_j за допомогою вибіркового значення частинного коефіцієнта кореляції $\hat{r}_{ij(I)}$ збігається з відповідною процедурою для значення парного коефіцієнта кореляції $\hat{r}_{ij}$ в усьому, крім етапу перевірки цієї характеристики на значиме відхилення від нуля, у який потрібно внести незначні корекції.

Дійсно, якщо $\begin{cases} \hat{r}_{ij(I)} \neq 0 \\ \left|\hat{r}_{ij(I)}\right| \neq 1 \end{cases}$, то для з'ясування суттєвості статистичного зв'язку між змінними ξ_i та ξ_j потрібно здійснити *перевірку на значимість* частинного коефіцієнта кореляції $r_{ij(I)}$, а саме, перевірити гіпотезу:

$$H_0 : r_{ij(I)} = 0$$

з деяким рівнем значущості $\alpha > 0$.

Виявляється, що при справедливості гіпотези H_0, статистика

$$t(n,m) = \frac{\sqrt{n-m-2}\ \hat{r}_{ij(I)}}{\sqrt{1-\hat{r}^2_{ij(I)}}}$$

має t-розподіл Стьюдента з $(n-m-2)$ ступенями свободи. Тоді область прийняття гіпотези H_0 набуває вигляду:

$$\left|t(n,m)\right| < t_{\frac{\alpha}{2}}(n-m-2),$$

де $t_\alpha(n)$ – 100α %-на точка t-розподілу Стьюдента з n ступенями свободи.

Якщо гіпотеза H_0 відхиляється, то вважається, що статистичний зв'язок між змінними ξ_i та ξ_j є істотним, у протилежному випадку – не істотним.

1.4.2.2. Множинний коефіцієнт кореляції та його властивості

Припустимо, що на ймовірнісному просторі (Ω, F, P) задані випадкова величина η і випадковий вектор $\vec{\xi}^T = (\xi_1, \xi_2, \ldots, \xi_q)$, які мають сумісний нормальний розподіл:

$$\eta \sim N\left(m_\eta, \sigma^2_\eta\right),\ \vec{\xi} \sim N\left(\vec{m}_{\vec{\xi}}, \Sigma_{\vec{\xi}\vec{\xi}}\right),\ \sigma^2_\eta > 0, \Sigma_{\vec{\xi}\vec{\xi}} > 0;$$

$$M\left(\eta - m_\eta\right)\left(\vec{\xi} - \vec{m}_{\vec{\xi}}\right)^T = \Sigma_{\eta\vec{\xi}}.$$

Означення. Нехай η і $\vec{\xi}$ випадкові величина та вектор відповідно. Тоді *множинним коефіцієнтом кореляції* η *і* $\vec{\xi}$ називається парний коефіцієнт кореляції випадкових величин η і лінійної функції від $\vec{\xi}$, яка є найкращою в середньоквадратичному розумінні апроксимацією для η. Позначення $r_{\eta\vec{\xi}}$.

Зауваження. У нормальному випадку такою лінійною функцією є функція регресії $f(\vec{x}) = M(\eta / \vec{\xi} = \vec{x})$, тому $r_{\eta\vec{\xi}} = r_{\eta f(\vec{\xi})}$.

Крім цього для функцій $f(\vec{x})$ і $g(\vec{x}) = D(\eta / \vec{\xi} = \vec{x})$ має місце твердження.

Теорема. Нехай пара $(\eta, \vec{\xi})$ має сумісний гауссівський розподіл, причому

$$\eta \sim N\left(m_\eta, \sigma_\eta^2\right),\ \vec{\xi} \sim N\left(\vec{m}_{\vec{\xi}}, \Sigma_{\vec{\xi}\vec{\xi}}\right),\ \sigma_\eta^2 > 0, \Sigma_{\vec{\xi}\vec{\xi}} > 0;$$

$$M\left(\eta - m_\eta\right)\left(\vec{\xi} - \vec{m}_{\vec{\xi}}\right)^T = \Sigma_{\eta\vec{\xi}}.$$

Тоді справедливо

$$f(\vec{x}) = m_\eta + \Sigma_{\eta\vec{\xi}}\Sigma_{\vec{\xi}\vec{\xi}}^{-1}\left(\vec{x} - m_{\vec{\xi}}\right), \qquad g(\vec{x}) = \sigma_\eta^2\left(1 - r_{\eta\vec{\xi}}^2\right).$$

Наслідок. Якщо додатково припустити, що $\xi_1, \xi_2, \ldots, \xi_q$ некорельовані, то

$$f(\vec{x}) = m_\eta + \sum_{i=1}^{q} r_{\eta\xi_i}\frac{\sigma_\eta}{\sigma_{\xi_i}}\left(x_i - m_{\xi_i}\right),$$

де $\vec{x}^T = \left(x_1, x_2, \ldots, x_q\right)$, $\sigma_{\xi_i}^2 = D\xi_i$, $i = \overline{1,q}$.

Наступне твердження дозволяє встановити зв'язок між множинним коефіцієнтом кореляції $r_{\eta\vec{\xi}}$ та індексом кореляції $I_{\eta\vec{\xi}}$.

Теорема. Припустимо, що η та $\vec{\xi}$ нормально розподілені

$$\eta \sim N\left(m_\eta, \sigma_\eta^2\right),\ \vec{\xi} \sim N\left(\vec{m}_{\vec{\xi}}, \Sigma_{\vec{\xi}\vec{\xi}}\right),\ \sigma_\eta^2 > 0, \Sigma_{\vec{\xi}\vec{\xi}} > 0;$$

$$M\left(\eta - m_\eta\right)\left(\vec{\xi} - \vec{m}_{\vec{\xi}}\right)^T = \Sigma_{\eta\vec{\xi}}.$$

Тоді індекс кореляції $I_{\eta\vec{\xi}} = \left|r_{\eta\vec{\xi}}\right|$, а коефіцієнт детермінації $I_{\eta\vec{\xi}}^2 = r_{\eta\vec{\xi}}^2$.

Зауваження. У гауссівському випадку для парного/множинного статистичного зв'язку індекс кореляції збігається з модулем, а коефіцієнт детермінації з квадратом парного/множинного коефіцієнта кореляції.

Доведення. Скористаємося попередньою теоремою для обчислення індексу кореляції $I_{\eta\vec{\xi}}$. Оскільки $D\eta = Df(\vec{\xi}) + Mg(\vec{\xi})$, можна стверджувати, що

$$I_{\eta\vec{\xi}} = \sqrt{\frac{Df(\vec{\xi})}{D\eta}} = \sqrt{\frac{\sigma_\eta^2 - Mg(\vec{\xi})}{\sigma_\eta^2}} = \sqrt{\frac{\sigma_\eta^2 - \sigma_\eta^2\left(1 - r_{\eta\vec{\xi}}^2\right)}{\sigma_\eta^2}} = \left|r_{\eta\vec{\xi}}\right|.$$

Що й треба було довести.

Таким чином, у нормальному випадку як характеристику множинного статистичного зв'язку можна використовувати замість індексу кореляції $I_{\eta\vec{\xi}}$ множинний коефіцієнт кореляції $r_{\eta\vec{\xi}}$.

Остання теорема дозволяє легко переконатись у справедливості **властивостей** $r_{\eta\vec{\xi}}$:

1) $-1 \le r_{\eta\vec{\xi}} \le 1$;
2) якщо $r_{\eta\vec{\xi}} = 0$, то η не залежить від $\vec{\xi}$;
3) якщо $\left|r_{\eta\vec{\xi}}\right| = 1$, то з ймовірністю 1 існує лінійний зв'язок між η та $\vec{\xi}$, а саме, справедливо $\eta = m_\eta + \Sigma_{\eta\vec{\xi}}\Sigma_{\vec{\xi}\vec{\xi}}^{-1}\left(\vec{\xi} - m_{\vec{\xi}}\right)$.

Для визначення коефіцієнта детермінації можна скористатися одним з таких підходів:

1) нехай випадкова величина η і випадковий вектор $\vec{\xi}^T = \left(\xi_1, \xi_2, \ldots, \xi_q\right)$ мають невироджений сумісний гауссівський розподіл. Для єдиності позначень покладемо $\xi_0 \equiv \eta$. Тоді

$$r_{\eta\vec{\xi}}^2 = 1 - \frac{\det(R)}{R_{00}},$$

де R_{kl} – алгебраїчне доповнення до елемента r_{kl} у матриці звичайних парних коефіцієнтів кореляції для змінних $\xi_0, \xi_1, \xi_2, \ldots, \xi_q$, а саме:

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0q} \\ r_{10} & 1 & r_{12} & \cdots & r_{1q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{q0} & r_{q1} & r_{q2} & \cdots & 1 \end{pmatrix};$$

2) коефіцієнт детермінації також можна визначити через частинні коефіцієнти кореляції таким чином:

$$r_{\eta\vec{\xi}}^2 = 1 - \left(1 - r_{01}^2\right)\left(1 - r_{02(\{1\})}^2\right)\left(1 - r_{03(\{1,2\})}^2\right)\cdots\left(1 - r_{0q(\{1,2,\ldots,q-1\})}^2\right).$$

Для використання на практиці введених характеристик множинного зв'язку потрібно вміти знаходити їх емпіричні значення. Припустимо, що доступні спостереження:

$$\eta: y_1, y_2, \ldots, y_n, \left(\xi_0 \equiv \eta\right), \qquad \vec{\xi}: \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n.$$

Тоді, враховуючи останні результати, для обчислення вибіркового значення коефіцієнта детермінації $\hat{r}_{\eta\vec{\xi}}^2$ можна використати вираз

$$\hat{r}^2_{\eta\vec{\xi}} = 1 - \frac{\det\left(\hat{R}\right)}{\hat{R}_{00}},$$

де $\hat{R}_{kl}$ – алгебраїчне доповнення до елемента $\hat{r}_{kl}$ у матриці вибіркових парних коефіцієнтів кореляції для змінних $\xi_0, \xi_1, \xi_2, \ldots, \xi_q$, а саме:

$$\hat{R} = \begin{pmatrix} 1 & \hat{r}_{01} & \hat{r}_{02} & \cdots & \hat{r}_{0q} \\ \hat{r}_{10} & 1 & \hat{r}_{12} & \cdots & \hat{r}_{1q} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{r}_{q0} & \hat{r}_{q1} & \hat{r}_{q2} & \cdots & 1 \end{pmatrix}.$$

З іншого боку, якщо доступні емпіричні значення для потрібних частинних коефіцієнтів кореляції, множинний коефіцієнт детермінації $\hat{r}^2_{\eta\vec{\xi}}$ можна підрахувати таким чином:

$$\hat{r}^2_{\eta\vec{\xi}} = 1 - \left(1 - \hat{r}^2_{01}\right)\left(1 - \hat{r}^2_{02(\{1\})}\right)\left(1 - \hat{r}^2_{03(\{1,2\})}\right)\ldots\left(1 - \hat{r}^2_{0q(\{1,2,\ldots,q-1\})}\right).$$

Методика аналізу множинного статистичного зв'язку між випадковою змінною η і випадковим вектором $\vec{\xi}$ за допомогою $\hat{r}_{\eta\vec{\xi}}$ така:

1) якщо $\hat{r}_{\eta\vec{\xi}} = 0$, то вважаємо, що зв'язок між η та $\vec{\xi}$ відсутній;
2) якщо $\left|\hat{r}_{\eta\vec{\xi}}\right| = 1$, то зв'язок між η та $\vec{\xi}$ є лінійний і має представлення: $\eta = m_\eta + \Sigma_{\eta\vec{\xi}}\Sigma^{-1}_{\vec{\xi}\vec{\xi}}\left(\vec{\xi} - m_{\vec{\xi}}\right)$;
3) якщо $\begin{cases} \hat{r}_{\eta\vec{\xi}} \neq 0 \\ \left|\hat{r}_{\eta\vec{\xi}}\right| \neq 1 \end{cases}$, то потрібно перевірити гіпотезу про те, чи суттєво відхиляється від нуля множинний коефіцієнт кореляції $r_{\eta\vec{\xi}}$ зі статистичної точки зору, а саме, зробити *перевірку* його *на значимість*, тобто потрібно перевірити гіпотезу:

$$H_0 : r_{\eta\vec{\xi}} = 0,$$

з деяким рівнем значущості $\alpha > 0$. І якщо гіпотеза виявиться справедливою, то вважається, що статистичний зв'язок між η та $\vec{\xi}$ $\left(\vec{\xi} \in R^q\right)$ є не істотним, у протилежному випадку – істотним.

Для перевірки останньої гіпотези скористаємося тим фактом, що при справедливості гіпотези H_0 розподіл статистики

$$F = \frac{\hat{r}^2_{\eta\vec{\xi}}}{1 - \hat{r}^2_{\eta\vec{\xi}}} \cdot \frac{n - q - 1}{q}$$

можна наблизити F-розподілом з q і $(n-q-1)$ ступенями свободи. А для обчислення самої статистики F потрібно вміти обраховувати вибіркове значення множинного коефіцієнта детермінації $\hat{r}^2_{\eta\vec{\xi}}$, але, згідно з вищенаведеними результатами, це не викликає труднощів. Останнє дозволяє область прийняття для гіпотези H_0 записати у вигляді

$$F < F_\alpha\left(q, n - q - 1\right),$$

де $F_\alpha(m,n)$ – 100α %-на точка F-розподілу з m та n ступенями свободи.

У підсумку, у гауссівському випадку як характеристику множинного статистичного зв'язку можна використовувати множинний коефіцієнт кореляції $r_{\eta\vec{\xi}}$, застосування якого на практиці є більш логічним, оскільки при набутті ним значень ± 1 можна зробити висновок про наявність лінійного зв'язку між η та $\vec{\xi}$. У загальному випадку множинний коефіцієнт кореляції не надає можливості в такій інтерпретації, тому буде необхідно звернутися до використання індексу кореляції $I_{\eta\vec{\xi}}$.

2. Кореляційний аналіз ординальних змінних

Розглянемо проблему аналізу наявності статистичного зв'язку між ординальними (порядковими) змінними. Введемо такі поняття, як ранг, ранжировка, таблиця рангів. Наведемо методику для дослідження парного зв'язку між ординальними змінними за допомогою рангових коефіцієнтів кореляції Спірмена та Кендела. Для аналізу множинного зв'язку використаємо коефіцієнт конкордації. Для всіх запропонованих характеристик статистичного зв'язку наведено основні властивості та процедури їх перевірки на значиме відхилення від нуля.

2.1. Ранги та таблиці рангів

Нехай необхідно провести аналіз ординальних змінних η, ξ_1, $\xi_2, \ldots, \xi_q$. У подальшому для змінної η (для єдиності позначень) будемо використовувати позначення ξ_0, тобто $\xi_0 \equiv \eta$. Їх дослідження про-

ведемо на основі доступних спостережень проявів цих змінних для деяких n об'єктів.

Спочатку проаналізуємо випадок, коли за кожною ординальною змінною (властивістю) ξ_i її прояви в різних об'єктів не однакові $\left(i=\overline{0,q}\right)$. Розглянемо прояви конкретної змінної ξ_i для n об'єктів. Оскільки в різних об'єктів прояв властивості ξ_i інший, то всі об'єкти можна впорядкувати в міру спадання прояву змінної ξ_i. Це дозволяє кожному k-му об'єкту зіставляти натуральне число $x_k^{(i)}$ – номер місця, яке $k-$й об'єкт зайняв у цій упорядкованій послідовності.

Означення. *Рангом k-го об'єкта за ординальною змінною ξ_i* називається натуральне число $x_k^{(i)}$, яке вказує номер місця, яке зайняв k-й об'єкт у послідовності всіх досліджуваних об'єктів, упорядкованій у міру спадання прояву властивості ξ_i, $k=\overline{1,n}$, $i=\overline{0,q}$.

Зрозуміло, що для якісного визначення рангів об'єктів потрібно залучати експертів відповідного профілю.

Означення. *Ранжировкою за ординальною змінною ξ_i* називається вектор-стовчик $x^{(i)}$, k-та компонента якого дорівнює рангу k-го об'єкта за ординальною змінною ξ_i, тобто $x^{(i)}=\left(x_1^{(i)},x_2^{(i)},\ldots,x_n^{(i)}\right)^T$.

У розглянутому випадку, ранжировки для змінних, які потрібно дослідити, будуть являти собою вектор-стовпчики, утворені внаслідок деяких перестановок компонент у вектор-стовпчику $(1,2,\ldots,n)^T$.

Означення. Ранжировка $y=(y_1,y_2,\ldots,y_n)^T$ *називається протилежною ранжировкою до ранжировки* $x=(x_1,x_2,\ldots,x_n)^T$, якщо справедливо $y_i=n+1-x_i$, $i=\overline{1,n}$.

Наприклад: протилежною ранжировкою до ранжировки $(5,1,4,2,3)^T$ є ранжировка $(1,5,2,4,3)^T$.

Розглянемо тепер більш загальний випадок, коли існує ординальна змінна для якої існує група об'єктів, яка складається з не менше, ніж двох членів, для яких прояв цієї ординальної змінної однаковий. Тоді маємо справу з *наявністю груп нерозрізнимих об'єктів за змінною,* яка аналізується. Узагальнимо для цього випадку поняття рангу.

Означення. *Зв'язаним рангом об'єкта з групи нерозрізнимих об'єктів за ординальною змінною,* називається середнє арифметичне номерів місць, які припали на цю групу об'єктів.

Наприклад, якщо на групу нерозрізнимих об'єктів за ординальною змінною, припали місця з номерами 3, 4, 5, 6, то кожному з цих об'єктів присвоюється зв'язаний ранг, який дорівнює

$$\frac{3+4+5+6}{4}=4,5.$$

Аналогічно обчислюється зв'язаний ранг для об'єктів з інших груп нерозрізнимих об'єктів. Приходимо до того, що в загальному випадку ранг, що віддзеркалює степінь прояву властивості, не завжди буде натуральним числом.

На основі рангів, які присвоєні об'єктам, за всіма ординальними змінними формується *таблиця рангів*, у якій стовпчик з номером i являє собою не що інше, як ранжировку $x^{(i)}=\left(x_1^{(i)},x_2^{(i)},\ldots,x_n^{(i)}\right)^T$ за ординальною змінною ξ_i, $i=\overline{0,q}$, а саме, таблицю такого вигляду:

№ об'єкта \ № змінної	0	1	2	⋯	i	⋯	q
1	$x_1^{(0)}$	x_1^{1}	$x_1^{(2)}$	⋯	$x_1^{(i)}$	⋯	$x_1^{(q)}$
2	$x_2^{(0)}$	$x_2^{(1)}$	$x_2^{(2)}$	⋯	$x_2^{(i)}$	⋯	$x_2^{(q)}$
⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮
k	$x_k^{(0)}$	$x_k^{(1)}$	$x_k^{(2)}$	⋯	$x_k^{(i)}$	⋯	$x_k^{(q)}$
⋮	⋮	⋮	⋮	⋯	⋮	⋯	⋮
n	$x_n^{(0)}$	$x_n^{(1)}$	$x_n^{(2)}$	⋯	$x_n^{(i)}$	⋯	$x_n^{(q)}$

Заповнена таблиця є основою при проведенні кореляційного аналізу ординальних змінних. Поява в ній значень зв'язаних рангів буде вимагати модифікації відповідних процедур для дослідження парних і множинних зв'язків.

З огляду на введені поняття, стає зрозумілим, чому розділ математики з досліджень істотності статистичного зв'язку між ординальними змінними називають також *аналізом рангових кореляцій.*

2.2. Аналіз парних рангових кореляцій

Розглянемо дві ординальні змінні. Нехай для них щодо кожного об'єкта відомий його ранг, тобто доступні відповідні ранжировки. Необхідно, спираючись на цю інформацію, з'ясувати істотність статистичного зв'язку між цими порядковими змінними. Пропонується використовувати рангові коефіцієнти кореляції. Розглянемо найбільш вживані з них.

2.2.1. Ранговий коефіцієнт кореляції Спірмена

Для аналізу статистичного зв'язку між двома ординальними змінними ξ_i та ξ_j з відомими для них ранжировками $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\right)^T$ та $x^{(j)} = \left(x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\right)^T$, відповідно, К. Спірмен запропонував оригінальну характеристику парного зв'язку.

Спочатку розглянемо ситуацію, коли групи об'єктів з однаковим проявом властивостей ξ_i та ξ_j відсутні.

Означення. *Ранговим коефіцієнтом кореляції Спірмена* ординальних змінних ξ_i та ξ_j з ранжировками $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\right)^T$ та $x^{(j)} = \left(x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\right)^T$, відповідно, називається характеристика, яка обчислюється таким чином:

$$\hat{\tau}_{ij}^{(S)} = 1 - \frac{\left\|x^{(i)} - x^{(j)}\right\|^2}{\frac{n^3 - n}{6}},$$

де $\left\|x^{(i)} - x^{(j)}\right\|^2 = \sum_{k=1}^{n}\left(x_k^{(i)} - x_k^{(j)}\right)^2$.

Цей показник має такі **властивості**:

1) $-1 \le \hat{\tau}_{ij}^{(S)} \le 1$;

2) якщо $\hat{\tau}_{ij}^{(S)} = 0$, то зв'язок між i-ою та j-ою змінними відсутній;

3) якщо $\hat{\tau}_{ij}^{(S)} = 1$, то ранжировки $x^{(i)}$ та $x^{(j)}$ рівні;

4) якщо $\hat{\tau}_{ij}^{(S)} = -1$, то ранжировки $x^{(i)}$ та $x^{(j)}$ протилежні;

5) $\hat{\tau}_{ij}^{(S)} = \hat{\tau}_{ji}^{(S)}$.

У загальному випадку, коли наявні групи нерозрізнимих об'єктів принаймні за однією зі змінних ξ_i чи ξ_j, використовується *модифікований ранговий коефіцієнт кореляції Спірмена*:

$$\hat{\tilde{\tau}}_{ij}^{(S)} = \frac{\frac{n^3 - n}{6} - \left\|x^{(i)} - x^{(j)}\right\|^2 - T^{(i)} - T^{(j)}}{\sqrt{\left[\frac{n^3 - n}{6} - 2T^{(i)}\right]\left[\frac{n^3 - n}{6} - 2T^{(j)}\right]}},$$

де $T^{(k)} = \frac{1}{12}\sum_{l=1}^{m^{(k)}}\left[\left(n_l^{(k)}\right)^3 - n_l^{(k)}\right]$; $m^{(k)}$ – кількість груп нерозрізнимих об'єктів за змінною ξ_k; $n_l^{(k)}$ – кількість об'єктів, які ввійшли в l-ту групу нерозрізнимих об'єктів за змінною ξ_k; $k = i, j$.

Зауваження. Коли відсутні групи нерозрізнимих об'єктів за змінними ξ_i та ξ_j, справедливим є $T^{(i)} = T^{(j)} = 0$, оскільки $m^{(i)} = m^{(j)} = n$, а $n_l^{(i)} = n_l^{(j)} = 1$, $l = \overline{1,n}$. Звідси випливає, що $\hat{\tilde{\tau}}_{ij}^{(S)}$ буде дорівнювати $\hat{\tau}_{ij}^{(S)}$.

Методика аналізу статистичного зв'язку між ординальними змінними ξ_i та ξ_j за допомогою рангового коефіцієнта кореляції Спірмена у випадках, коли він набуває одне зі значень -1, 0, 1 очевидна і випливає з його властивостей. А коли

$$\begin{cases} \hat{\tau}_{ij}^{(S)} \neq 0 \\ \left|\hat{\tau}_{ij}^{(S)}\right| \neq 1 \end{cases},$$

потрібно з'ясувати, чи суттєво відхиляється від нуля ранговий коефіцієнт кореляції Спірмена із статистичної точки зору, тобто здійснити *перевірку* його *на значимість*, а саме перевірити гіпотезу

$$H_0 : \tau_{ij}^{(S)} = 0$$

з деяким рівнем значущості $\alpha > 0$. Якщо гіпотеза буде відхилена, то вважається, що зв'язок між ординальними змінними ξ_i та ξ_j зі статистичної точки зору є істотним, у протилежному випадку – не суттєвим.

При $n=\overline{4,10}$ таку перевірку можна здійснити за допомогою спеціальних таблиць. А при $n\geq 11$ область прийняття гіпотези набуває вигляду

$$\left|\frac{\sqrt{n-2}\ \hat{\tau}_{ij}^{(S)}}{\sqrt{1-\left(\hat{\tau}_{ij}^{(S)}\right)^2}}\right| < t_{\frac{\alpha}{2}}(n-2),$$

де $t_\alpha(n)$ – 100α %-на точка t-розподілу Стьюдента з n ступенями свободи.

2.2.2. Ранговий коефіцієнт кореляції Кендела

Розглянемо ще одну характеристику, яку у свій час запропонував М. Кендел для аналізу статистичного зв'язку між двома ординальними змінними ξ_i та ξ_j з доступними для них ранжировками

$$x^{(i)}=\left(x_1^{(i)},x_2^{(i)},\ldots,x_n^{(i)}\right)^T \text{ та } x^{(j)}=\left(x_1^{(j)},x_2^{(j)},\ldots,x_n^{(j)}\right)^T.$$

Почнемо з випадку, коли відсутні групи об'єктів з однаковим проявом властивостей ξ_i та ξ_j.

Означення. *Ранговим коефіцієнтом кореляції Кендела* ординальних змінних ξ_i та ξ_j з ранжировками $x^{(i)}=\left(x_1^{(i)},x_2^{(i)},\ldots,x_n^{(i)}\right)^T$ і $x^{(j)}=\left(x_1^{(j)},x_2^{(j)},\ldots,x_n^{(j)}\right)^T$, відповідно, називається характеристика, що визначається таким чином:

$$\hat{\tau}_{ij}^{(K)}=1-\frac{\upsilon\left(x^{(i)},x^{(j)}\right)}{\frac{n(n-1)}{4}},$$

де $\upsilon\left(x^{(i)},x^{(j)}\right)$ – мінімальна кількість обмінів сусідніх компонент у ранжировці $x^{(i)}$, яка приводить її до ранжировки $x^{(j)}$.

Властивості рангового коефіцієнта кореляції Кендела збігаються з властивостями коефіцієнта Спірмена і мають вигляд:

1) $-1\leq\hat{\tau}_{ij}^{(K)}\leq 1$;

2) якщо $\hat{\tau}_{ij}^{(K)}=0$, то зв'язок між i-ою та j-ою змінними відсутній;

3) якщо $\hat{\tau}_{ij}^{(K)}=1$, то ранжировки $x^{(i)}$ та $x^{(j)}$ рівні;

4) якщо $\hat{\tau}_{ij}^{(K)}=-1$, то ранжировки $x^{(i)}$ та $x^{(j)}$ протилежні;

5) $\hat{\tau}_{ij}^{(K)}=\hat{\tau}_{ji}^{(K)}$.

За наявності груп нерозрізнимих об'єктів принаймні за однією зі змінних ξ_i чи ξ_j, ранговий коефіцієнт кореляції Кендела корегується й використовується у вигляді *модифікованого рангового коефіцієнта кореляції Кендела*:

$$\hat{\hat{\tau}}_{ij}^{(K)}=\frac{\hat{\tau}_{ij}^{(K)}-\frac{2\left(U^{(i)}+U^{(j)}\right)}{n(n-1)}}{\sqrt{\left(1-\frac{2U^{(i)}}{n(n-1)}\right)\left(1-\frac{2U^{(j)}}{n(n-1)}\right)}},$$

де $U^{(k)}=\frac{1}{2}\sum_{l=1}^{m^{(k)}}n_l^{(k)}\left(n_l^{(k)}-1\right)$; $m^{(k)}$ – кількість груп нерозрізнимих об'єктів за змінною ξ_k; $n_l^{(k)}$ – кількість об'єктів, які ввійшли в l-ту групу нерозрізнимих об'єктів за змінною ξ_k; $k=i,j$.

Зауваження. Для рангового коефіцієнта кореляції Кендела, коли відсутні групи нерозрізнимих об'єктів за змінними ξ_i та ξ_j, корегуючі величини $U^{(i)}=U^{(j)}=0$, оскільки $m^{(i)}=m^{(j)}=n$, а $n_l^{(i)}=n_l^{(j)}=1$, $l=\overline{1,n}$. У цій ситуації модифікований коефіцієнт $\hat{\hat{\tau}}_{ij}^{(K)}$ буде дорівнювати $\hat{\tau}_{ij}^{(K)}$.

Аналіз статистичного зв'язку між ординальними змінними ξ_i та ξ_j у випадках, коли ранговий коефіцієнт кореляції Кендела набуває одного із значень -1, 0, 1, випливає з вищенаведених його властивостей. А для прийняття рішення, коли

$$\begin{cases}\hat{\tau}_{ij}^{(K)}\neq 0\\ \left|\hat{\tau}_{ij}^{(K)}\right|\neq 1\end{cases},$$

потрібно звернутися до *перевірки на значимість* рангового коефіцієнта кореляції Кендела, тобто перевірити гіпотезу

$$H_0:\tau_{ij}^{(K)}=0$$

з деяким рівнем значущості $\alpha > 0$.

При $n = \overline{4,10}$ така перевірка здійснюється за допомогою спеціальних таблиць. А при $n \geq 11$ аналіз показав, що область прийняття гіпотези набуває такого представлення:

$$\left| 3\hat{\tau}_{ij}^{(K)} \sqrt{\frac{n(n-1)}{2(2n+5)}} \right| < u_{\frac{\alpha}{2}},$$

де u_β – 100β %-на точка нормального розподілу з параметрами 0 і 1.

Зв'язок між ординальними змінними ξ_i та ξ_j із статистичної точки зору вважається суттєвим при відхиленні гіпотези H_0, а в протилежному разі його слід сприймати як не істотний.

У підсумку, для аналізу парного статистичного зв'язку ординальних змінних отримали ще одну характеристику – ранговий коефіцієнт кореляції Кендела.

2.3. Коефіцієнт конкордації – характеристика множинних статистичних зв'язків для ординальних змінних

Перейдемо до аналізу наявності статистичного зв'язку між кількома ординальними змінними $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$, причому нехай для цих змінних доступні ранжировки $x^{(i_j)} = \left(x_1^{(i_j)}, x_2^{(i_j)}, \ldots, x_n^{(i_j)} \right)^T$, $j = \overline{1,m}$, $2 \leq m \leq q+1$, де n – кількість об'єктів, що досліджуються. Для розв'язання цієї проблеми М. Кендел запропонував використовувати спеціальний показник.

Розглянемо спочатку випадок, коли відсутні групи об'єктів з однаковим проявом у змінних $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$, $2 \leq m \leq q+1$. Введемо позначення $\vec{\zeta} = \left(\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m} \right)^T$.

Означення. *Коефіцієнтом конкордації* ординальних змінних $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$ з ранжировками $x^{(i_1)}, x^{(i_2)}, \ldots, x^{(i_m)}$ називається характеристика, що задається згідно з:

$$\hat{W}_{\vec{\zeta}} = \frac{\sum_{k=1}^{n} \left[\sum_{j=1}^{m} \left(x_k^{(i_j)} \right) - \frac{m(n+1)}{2} \right]^2}{\frac{m^2 \left(n^3 - n \right)}{12}},$$

де $x^{(i_j)} = \left(x_1^{(i_j)}, x_2^{(i_j)}, \ldots, x_n^{(i_j)} \right)^T$, $j = \overline{1,m}$, $2 \leq m \leq q+1$.

Для коефіцієнта конкордації мають місце такі **властивості**:

1) $0 \leq \hat{W}_{\vec{\zeta}} \leq 1$;

2) якщо $\hat{W}_{\vec{\zeta}} = 0$, то вважається, що зв'язок між змінними $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$ відсутній;

3) якщо $\hat{W}_{\vec{\zeta}} = 1$, то ранжировки $x^{(i_1)}, x^{(i_2)}, \ldots, x^{(i_m)}$ рівні;

4) зміна порядку розташування змінних у послідовності $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$ не змінює значення коефіцієнта конкордації.

Додаткової уваги потребує випадок наявності груп об'єктів з однаковим проявом у змінних $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$, $2 \leq m \leq q+1$. Після відповідної корекції, коефіцієнт конкордації можна використовувати у вигляді такого *модифікованого коефіцієнта конкордації*:

$$\hat{W}_{\vec{\zeta}} = \frac{\sum_{k=1}^{n} \left[\sum_{j=1}^{m} \left(x_k^{(i_j)} \right) - \frac{m(n+1)}{2} \right]^2}{\frac{m^2 \left(n^3 - n \right)}{12} - m \sum_{j=1}^{m} T^{(i_j)}},$$

де $T^{(i_j)} = \frac{1}{12} \sum_{l=1}^{m^{(i_j)}} \left[\left(n_l^{(i_j)} \right)^3 - n_l^{(i_j)} \right]$; $m^{(i_j)}$ – кількість груп нерозрізнимих об'єктів за змінною ξ_{i_j}; $n_l^{(i_j)}$ – кількість об'єктів, які ввійшли в l-ту групу нерозрізнимих об'єктів за змінною ξ_{i_j}; $j = \overline{1,m}$; $2 \leq m \leq q+1$.

Використання на практиці коефіцієнта конкордації, коли він набуває екстремальних значень 0 чи 1, з огляду на його властивості, не викликає труднощів. Зупинимося на ситуації, коли $0 < \hat{W}_{\vec{\zeta}} < 1$. Знову звертаємося до перевірки на значиме відхилення від нуля коефіцієнта, який використовується, тобто до перевірки гіпотези

$$H_0 : W_{\vec{\zeta}} = 0$$

з деяким рівнем значущості $\alpha > 0$.

При $n = \overline{3,7}$, $m = \overline{3,20}$ потрібна перевірка здійснюється за допомогою спеціальних таблиць. А при $n \geq 8$ можна скористатися тим фактом, що

розподіл статистики $m(n-1)\hat{W}_{\bar{\xi}}$ можна наблизити χ^2-розподілом з $(n-1)$ ступенями свободи за умови справедливості гіпотези H_0. У результаті область прийняття гіпотези набуває вигляду:

$$m(n-1)\hat{W}_{\bar{\xi}} < \chi^2_{\alpha}(n-1),$$

де $\chi^2_{\alpha}(n)$ – 100α %-на точка χ^2-розподілу з n ступенями свободи.

Зв'язок між ординальними змінними $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$ із статистичної точки зору будемо вважати суттєвим при відхиленні гіпотези H_0, а в протилежному випадку вважаємо не істотним.

3. Дослідження наявності зв'язку між номінальними змінними

Кореляційний аналіз номінальних (класифікаційних) змінних має свою специфіку, оскільки значення, яких вони набувають, не є впорядкованими. При роботі з двома змінними, які мають скінченні множини значень, на базі спостережень формується таблиця спряженості, яка є основою для побудови цілої низки характеристик статистичного зв'язку.

Деякі з них базуються на статистиці квадратичної спряженості, що й буде продемонстровано далі. З іншого боку, при побудові потрібного показника можна скористатися поняттям ентропії і на його основі побудувати інформаційну міру зв'язку. Аналіз цих характеристик дозволяє зробити висновок відносно істотності статистичного зв'язку між номінальними змінними, які досліджуються.

3.1. Таблиця спряженості

Розглянемо номінальні змінні η та ξ. Нехай множина значень у змінної η складається з r_1 градацій, а у змінної ξ – з r_2 градацій відповідно. Припустимо, що проведено всього n спостережень над деякими об'єктами. Вважатимемо, що у n_{ij} випадках η набула значень своєї i-ої градації, а ξ – j-ої градації. Тоді результати всіх спостережень можна занести до таблиці, у якій у комірку на перетині i-го рядка та j-го стовпчика заноситься значення n_{ij}, $i=\overline{1,r_1}$, $j=\overline{1,r_2}$. Побудована таким чином таблиця називається *таблицею спряженості* змінних η та ξ і матиме такий вигляд:

η \ ξ	1	2	…	j	…	r_2	
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1r_2}	$n_{1\bullet}$
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2r_2}	$n_{2\bullet}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ir_2}	$n_{i\bullet}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
r_1	$n_{r_1 1}$	$n_{r_1 2}$	…	$n_{r_1 j}$	…	$n_{r_1 r_2}$	$n_{r_1\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet r_2}$	n

де $n_{i\bullet}=\sum_{j=1}^{r_2} n_{ij}$, $n_{\bullet j}=\sum_{i=1}^{r_1} n_{ij}$, $n=\sum_{i=1}^{r_1} n_{i\bullet}=\sum_{j=1}^{r_2} n_{\bullet j}$. Причому, в останніх стовпчику та рядку таблиці підраховані суми значень по рядках і стовпчиках таблиці відповідно.

Таблиці спряженості можуть бути також використані при аналізі й інших класів змінних, які набувають скінченні множини значень, а саме: порядкових, дискретних кількісних зі скінченною множиною значень, неперервних кількісних для яких провели процедуру групування.

Використовуючи таблицю спряженості, можна конструювати показники, потрібні для дослідження наявності статистичного зв'язку між номінальними змінними.

3.2. Квадратична спряженість і характеристики парного зв'язку на її основі

Для подальшої побудови характеристик парного статистичного зв'язку введемо статистику *квадратичної спряженості* для змінних η та ξ:

$$\hat{\chi}^2_{\eta\xi}=\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\frac{\left(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}=n\left[\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\left(\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}}\right)-1\right].$$

Очевидно, що ця статистика набуває значень нуль тоді й тільки тоді, коли має місце

$$\frac{n_{ij}}{n}=\frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n},\ i=\overline{1,r_1},\ j=\overline{1,r_2}.$$

Але ця умова є не що інше, як умова незалежності змінних η та ξ, але записана у вибірковому представленні.

Останнє було використано різними авторами при побудові низки характеристик парного статистичного зв'язку на базі статистики квадратичної спряженості для змінних η та ξ, а саме:

1) *коефіцієнт спряженості – квадратний корінь з середньої квадратичної спряженості* $\hat{\varphi}_{\eta\xi} = \sqrt{\frac{\hat{\chi}^2_{\eta\xi}}{n}}$;

2) *коефіцієнт спряженості Пірсона* $\hat{P}_{\eta\xi} = \sqrt{\frac{\hat{\chi}^2_{\eta\xi}}{n+\hat{\chi}^2_{\eta\xi}}}$;

3) *коефіцієнт спряженості Чупрова* $\hat{T}_{\eta\xi} = \sqrt{\frac{\hat{\chi}^2_{\eta\xi}}{n\sqrt{(r_1-1)(r_2-1)}}}$;

4) *коефіцієнт спряженості Крамера* $\hat{C}_{\eta\xi} = \sqrt{\frac{\hat{\chi}^2_{\eta\xi}}{n\cdot\min(r_1-1, r_2-1)}}$.

Для цих характеристик справедливі такі **властивості**:

1) $\hat{\varphi}_{\eta\xi}, \hat{P}_{\eta\xi}, \hat{T}_{\eta\xi}, \hat{C}_{\eta\xi} \geq 0$;

2) $\hat{P}_{\eta\xi} < 1,\ \hat{C}_{\eta\xi} \leq 1$;

3) якщо $\begin{cases} \hat{\varphi}_{\eta\xi} = 0 \\ \hat{P}_{\eta\xi} = 0 \\ \hat{T}_{\eta\xi} = 0 \\ \hat{C}_{\eta\xi} = 0 \end{cases}$, то η не залежить від ξ.

При практичному використанні цих коефіцієнтів, коли вони набувають нульового значення, робимо висновок, що η не залежить від ξ, а якщо вони набувають значення відмінне від нуля, то потрібно звернутися до перевірки їх на значимість, але з огляду на їх структуру та залежність від квадратичної спряженості достатньо перевірити гіпотезу

$$H_0 : \chi^2_{\eta\xi} = 0$$

з деяким рівнем значущості $\alpha > 0$.

Враховуючи той факт, що розподіл статистики $\hat{\chi}^2_{\eta\xi}$ можна наблизити χ^2-розподілом з $(r_1-1)(r_2-1)$ ступенями свободи за умови справедливості H_0, то область прийняття гіпотези набуває вигляду:

$$\hat{\chi}^2_{\eta\xi} < \chi^2_{\alpha}\left((r_1-1)(r_2-1)\right),$$

де $\chi^2_{\alpha}(n)$ – 100α %-на точка χ^2-розподілу з n ступенями свободи.

Якщо остання нерівність буде справедлива, то будемо стверджувати, що η не залежить від ξ, у протилежному випадку йдеться про істотність зв'язку між η та ξ із статистичної точки зору.

3.3. Інформаційна міра парного статистичного зв'язку

Інший підхід до введення характеристики статистичного зв'язку базується на поняттях з теорії інформації. Нехай змінна η набуває своїх значень y_i з ймовірностями $p_\eta(y_i)$, змінна ξ – значень x_j з ймовірностями $p_\xi(x_j)$, а пара (η,ξ) набуває значень (y_i, x_j) з ймовірностями $p_{\eta\xi}(y_i, x_j)$, $i = \overline{1, r_1}$, $j = \overline{1, r_2}$.

Означення. *Ентропією* η називається величина, яка обчислюється як:

$$H_\eta = -\sum_{i=1}^{r_1} p_\eta(y_i)\ln\left(p_\eta(y_i)\right) = -M\ln\left(p_\eta(\eta)\right).$$

У свою чергу, ентропію ξ можна підрахувати згідно з:

$$H_\xi = -\sum_{j=1}^{r_2} p_\xi(x_j)\ln\left(p_\xi(x_j)\right) = -M\ln\left(p_\xi(\xi)\right).$$

Означення. *Ентропією* η *та* ξ називається значення, яке задається виразом

$$H_{\eta\xi} = -\sum_{i=1}^{r_1}\sum_{j=1}^{r_2} p_{\eta\xi}(y_i, x_j)\ln\left(p_{\eta\xi}(y_i, x_j)\right) = -M\ln\left(p_{\eta\xi}(\eta,\xi)\right).$$

Означення. *Інформаційна міра зв'язку* η *та* ξ визначається згідно з

$$N_{\eta\xi} = H_\eta + H_\xi - H_{\eta\xi}.$$

Основні **властивості** інформаційної міри зв'язку η та ξ:

1) $N_{\eta\xi} \geq 0$;

2) якщо $N_{\eta\xi} = 0$, то η не залежить від ξ;

3) $N_{\eta\eta} = H_\eta$.

Спираючись на доступні значення з таблиці спряженості, підрахуємо *вибіркове значення для інформаційної міри зв'язку* η *та* ξ. Оскільки

$$\hat{H}_\eta = -\sum_{i=1}^{r_1} \frac{n_{i\bullet}}{n} \ln\left(\frac{n_{i\bullet}}{n}\right),\ \hat{H}_\xi = -\sum_{j=1}^{r_2} \frac{n_{\bullet j}}{n} \ln\left(\frac{n_{\bullet j}}{n}\right),\ \hat{H}_{\eta\xi} = -\sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \frac{n_{ij}}{n} \ln\left(\frac{n_{ij}}{n}\right),$$

то

$$\hat{N}_{\eta\xi} = -\sum_{i=1}^{r_1} \frac{n_{i\bullet}}{n} \ln\left(\frac{n_{i\bullet}}{n}\right) - \sum_{j=1}^{r_2} \frac{n_{\bullet j}}{n} \ln\left(\frac{n_{\bullet j}}{n}\right) + \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} \frac{n_{ij}}{n} \ln\left(\frac{n_{ij}}{n}\right) =$$

$$= \frac{1}{n}\left\{ \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} n_{ij} \ln\left(n_{ij}\right) - \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} n_{ij} \ln(n) - \sum_{i=1}^{r_1} n_{i\bullet} \ln\left(n_{i\bullet}\right) + \sum_{i=1}^{r_1} n_{i\bullet} \ln(n) - \right.$$

$$\left. - \sum_{j=1}^{r_2} n_{\bullet j} \ln\left(n_{\bullet j}\right) + \sum_{j=1}^{r_2} n_{\bullet j} \ln(n) \right\} =$$

$$= \frac{1}{n}\left\{ \sum_{i=1}^{r_1}\sum_{j=1}^{r_2} n_{ij} \ln\left(n_{ij}\right) - \sum_{i=1}^{r_1} n_{i\bullet} \ln\left(n_{i\bullet}\right) - \sum_{i=1}^{r_2} n_{\bullet j} \ln\left(n_{\bullet j}\right) + n \ln(n) \right\}.$$

Коли $\hat{N}_{\eta\xi} = 0$, будемо приймати рішення про відсутність статистичного зв'язку між η та ξ, а у випадку $\hat{N}_{\eta\xi} > 0$, потрібно звернутися до *перевірки* інформаційної міри зв'язку η та ξ *на значимість*, тобто перевірити гіпотезу

$$H_0 : N_{\eta\xi} = 0$$

з деяким рівнем значущості $\alpha > 0$. Для цього розглянемо статистику

$$\hat{\hat{N}}_{\eta\xi} = 2n\hat{N}_{\eta\xi} - n_0 =$$

$$= 2\left[\sum_{i=1}^{r_1}\sum_{j=1}^{r_2} n_{ij} \ln\left(n_{ij}\right) - \sum_{i=1}^{r_1} n_{i\bullet} \ln\left(n_{i\bullet}\right) - \sum_{j=1}^{r_2} n_{\bullet j} \ln\left(n_{\bullet j}\right) + n \ln(n) \right] - n_0,$$

де n_0 – кількість нульових значень у таблиці спряженості.

Виявляється, що при справедливості цієї гіпотези розподіл статистики $\hat{\hat{N}}_{\eta\xi}$ можна наблизити χ^2-розподілом з $(r_1 - 1)(r_2 - 1)$ ступенями свободи. Тоді область прийняття гіпотези H_0 набуде вигляду

$$\hat{\hat{N}}_{\eta\xi} < \chi^2_\alpha\left((r_1 - 1)(r_2 - 1)\right),$$

де $\chi^2_\alpha(n)$ – 100α %-на точка χ^2-розподілу з n ступенями свободи.

Якщо остання нерівність не справедлива, то будемо говорити про істотність статистичного зв'язку між η та ξ, інакше вважаємо, що зв'язок між ними не є суттєвим.

ДОДАТОК 1

Умовні ймовірності та математичні сподівання. Основні властивості

Додаток містить довідкову інформацію необхідну при використанні умовних ймовірностей і математичних сподівань. Згадаємо основні необхідні поняття.

Означення. Нехай d_R є сукупністю усіх інтервалів виду $(a,b]$, $a,b \in R$. *Борелівською* σ*-алгеброю* σ_R *в* R називається найменша σ-алгебра, яка містить у собі систему d_R.

Зауваження 1. З означення борелівської σ-алгебри σ_R в R випливає, що в неї, крім інтервалів виду $(a,b]$, $a,b \in R$, також входять інтервали:

$$(a,b),\ [a,b),\ [a,b],\ (-\infty,b),\ (-\infty,b],\ (a,\infty),\ [a,\infty),\ a,b \in R,$$

порожня множина (a,a) і множини, які складаються з однієї точки $[a,a]$.

Зауваження 2. В означенні борелівської σ-алгебри σ_R в R замість інтервалів виду $(a,b]$ можна використовувати інтервали одного з видів:

$$(a,b),\ [a,b),\ [a,b],\ (-\infty,b),\ (-\infty,b],\ (a,\infty),\ [a,\infty),\ a,b \in R.$$

Означення. *Борелівською множиною в* R (*множиною вимірною за Борелем в* R) називається довільна множина з борелівської σ-алгебри σ_R в R.

Означення. *Борелівською функцією на* R *(функцією вимірною за Борелем на* R*)* називається дійсна функція задана на R, у якої прообраз довільної борелівської множини в R є борелівською множиною в R.

Аналогічно вводиться борелівська функція з областю визначення в просторі R^q, $q \in N$.

Означення. Нехай d_{R^q} є сукупністю всіх множин виду $\left\{(x_1, x_2, \ldots, x_q) \in R^q : x_i \in (a_i, b_i];\ a_i, b_i \in R, i = \overline{1,q}\right\}$. *Борелівською* σ*-алгеброю* σ_{R^q} *в* R^q називається найменша σ-алгебра, яка містить у собі систему d_{R^q}.

Зауваження 1. З означення борелівської σ-алгебри σ_{R^q} в R^q випливає, що в неї також входять множини виду

$$\left\{\left(x_1, x_2, \ldots, x_q\right) \in R^q : x_i \in B_i, i = \overline{1,q}\right\},$$

де $B_1, B_2, \ldots B_q$ – борелівські множини в R.

Зауваження 2. В означенні борелівської σ-алгебри σ_{R^q} в R^q замість множин

$$\left\{\left(x_1, x_2, \ldots, x_q\right) \in R^q : x_i \in \left(a_i, b_i\right]; a_i, b_i \in R, i = \overline{1,q}\right\}$$

можна використовувати множини виду

$$\left\{\left(x_1, x_2, \ldots, x_q\right) \in R^q : x_i \in B_i, i = \overline{1,q}\right\},$$

де $B_1, B_2, \ldots B_q$ – борелівські множини в R.

Означення. *Борелівською множиною в R^q (множиною вимірною за Борелем в R^q)* називається довільна множина з борелівської σ-алгебри σ_{R^q} в R^q.

Означення. *Борелівською функцією на R^q (функцією вимірною за Борелем на R^q)* називається дійсна функція, задана на R^q, у якої прообраз довільної борелівської множини в R є борелівською множиною в R^q.

Нехай на ймовірнісному просторі (Ω, F, P) задані дійсні випадкові величини $\eta = \eta(\omega)$ та $\xi = \xi(\omega)$, а G – деяка σ-алгебра ($G \subset F$).

Означення. *Випадкова величина ξ є вимірною відносно σ-алгебри G* ($G \subset F$), якщо прообраз довільної борелівської множини в R належить σ-алгебрі G, тобто $\forall A \in \sigma_R \ \ \xi^{-1}(A) \in G$.

Означення. *Умовним математичним сподіванням невід'ємної випадкової величини ξ відносно σ-алгебри G $(G \subset F)$* називається невід'ємна випадкова величина, яка позначається $M(\xi/G)$ або $M(\xi/G)(\omega)$ така, що

- $M(\xi/G)$ є вимірною відносно σ-алгебри G,
- $\forall A \in G$ справедливо $\int\limits_A \xi dP = \int\limits_A M(\xi/G) dP$.

Нескладно з'ясувати питання існування та єдиності $M(\xi/G)$. Дійсно, якщо випадкова величина ξ є інтегровною і $\xi \geq 0$ майже всюди, тоді на вимірному просторі (Ω, G) функція множин

$$\lambda(A) = \int\limits_A \xi dP, \ \forall A \in G$$

є мірою, яка абсолютно неперервна відносно ймовірнісної міри P (розглядається на вимірному просторі (Ω, G)). Очевидно, що міра λ буде скінченною та σ – адитивною.

Далі скористаємося відомим твердженням.

Теорема (Радона – Никодима). Нехай (Ω, σ_Ω) – вимірний простір, на σ-алгебрі σ_Ω задані скінченні σ-адитивні міри λ та μ. Якщо міра λ є абсолютно неперервною відносно міри μ, то існує σ_Ω-вимірна єдина (з точністю до μ-еквівалентності) така функція $\varphi(\cdot)$, задана на Ω, яка задовольняє вимоги

- $\varphi \in L_1(\Omega, \sigma_\Omega, \mu)$, (множина функцій інтегровних на Ω за мірою μ);
- $\varphi(\omega) \geq 0$ майже всюди на Ω відносно міри μ, що справедливо

$$\lambda(A) = \int\limits_A \varphi d\mu, \ \forall A \in \sigma_\Omega .$$

З теореми Радона – Никодима випливає, що існує невід'ємна інтегровна G-вимірна єдина (з точністю до P-еквівалентності) функція $M(\xi/G)$ задана на Ω така, що

$$\lambda(A) = \int\limits_A \xi dP = \int\limits_A M(\xi/G) dP \ \ \forall A \in G .$$

Таким чином, для невід'ємної випадкової величини ξ існує єдине (з точністю до P-еквівалентності) умовне математичне сподівання відносно σ-алгебри G $(G \subset F)$: $M(\xi/G)$.

Якщо ξ випадкова величина, то введемо позначення для таких випадкових величин $\xi^+ = \max(\xi, 0)$, $\xi^- = -\min(\xi, 0)$. Очевидно, що $\xi = \xi^+ - \xi^-$.

Означення. Умовним *математичним сподіванням довільної випадкової величини ξ відносно σ-алгебри G $(G \subset F)$* називається випадкова величина $M(\xi/G)$ (або $M(\xi/G)(\omega)$), яка визначається як:

$$M(\xi/G) = M(\xi^+/G) - M(\xi^-/G).$$

Зауваження. З означення випливає, що з ймовірністю 1 справедливо

$$\begin{cases} M\left(\xi^{+}/G\right)\geq 0, \\ M\left(\xi^{-}/G\right)\geq 0. \end{cases}$$

Означення. $M\left(\xi/G\right)$ *існує* (або *визначено*), якщо з ймовірністю 1 справедливо $\min\left(M\left(\xi^{+}/G\right),M\left(\xi^{-}/G\right)\right)<\infty$.

Означення. *Умовною ймовірністю події* $B\left(B\in F\right)$ *відносно* σ*-алгебри* $G\left(G\subset F\right)$, яка позначається $P\left(B/G\right)$ (або $P\left(B/G\right)(\omega)$), називається невід'ємна випадкова величина, що визначається згідно з

$$P\left(B/G\right)=M\left(I_B/G\right),$$

де I_B – *індикатор множини* B, тобто

$$I_B(\omega)=\begin{cases} 1, & \omega\in B, \\ 0, & \omega\notin B. \end{cases}$$

Можна використовувати й інше означення.

Означення. *Умовною ймовірністю події* $B\left(B\in F\right)$ *відносно* σ*-алгебри* $G\left(G\subset F\right)$ називається невід'ємна випадкова величина, яка позначається $P\left(B/G\right)$ (або $P\left(B/G\right)(\omega)$) така, що:

- $P\left(B/G\right)$ є вимірною відносно σ-алгебри G,
- $\forall A\in G$ справедливо $P\left(A\cap B\right)=\int\limits_A P\left(B/G\right)dP$.

Означення. σ*-алгебри* G_1,G_2 $(G_1,G_2\subset F)$ *незалежні,* якщо

$$P\left(A_1\cap A_2\right)=P\left(A_1\right)P\left(A_2\right),\ \forall A_1\in G_1, A_2\in G_2.$$

Означення. *Випадкові величини* η *та* ξ *незалежні,* якщо незалежні σ-алгебри, які породжені випадковими величинами η та ξ, а саме, σ-алгебри σ_η та σ_ξ.

Означення. η *не залежить від* σ*-алгебри* G $(G\subset F)$, якщо не залежні σ-алгебра G і σ-алгебра, яка породжена випадковою величиною η, тобто σ_η.

Основні властивості умовних математичних сподівань випадкових величин відносно σ-алгебр мають такий вигляд:

1) нехай $M|\eta|<\infty$, G – σ-алгебра $(G\subset F)$. Тоді існує $M\left(\eta/G\right)$;

2) нехай існує $M\left(\eta\right)$ і $Q=\{\varnothing,\Omega\}$ – тривіальна σ-алгебра. Тоді

$$M\left(\eta/Q\right)=M\left(\eta\right);$$

3) нехай існує $M\left(\eta\right)$ та η не залежить від σ-алгебри G $(G\subset F)$. Тоді $M\left(\eta/G\right)=M\left(\eta\right)$;

4) нехай η – вимірна відносно σ-алгебри G $(G\subset F)$, тоді $M\left(\eta/G\right)=\eta$. Тим паче $M\left(\eta/F\right)=\eta$;

5) нехай $M|\eta_1|<\infty$, $M|\eta_2|<\infty$, G – σ-алгебра $(G\subset F)$, $a,b\in R$, тоді $M\left(a\eta_1+b\eta_2/G\right)=aM\left(\eta_1/G\right)+bM\left(\eta_2/G\right)$;

6) нехай $M|\eta|<\infty$, $M|\xi\eta|<\infty$, ξ – вимірна відносно σ-алгебри G $(G\subset F)$, тоді $M\left\{\xi\eta/G\right\}=\xi M\left(\eta/G\right)$;

7) нехай $c\in R$, G – σ-алгебра $(G\subset F)$, тоді $M\left(c/G\right)=c$;

8) нехай $M|\eta|<\infty$, G_1,G_2 – σ-алгебри $(G_1\subset G_2\subset F)$, тоді

$$M\left\{M\left(\eta/G_2\right)/G_1\right\}=M\left(\eta/G_1\right),$$

$$M\left\{M\left(\eta/G_1\right)/G_2\right\}=M\left(\eta/G_1\right).$$

Означення. Нехай ξ – випадкова величина. Розглянемо сукупність множин $\left\{\omega:\xi(\omega)\in B\right\}$, $B\in\sigma_R$. Найменша σ-алгебра, що містить у собі ці множини, називається σ*-алгеброю, яка породжена випадковою величиною* ξ. Позначення – σ_ξ.

Аналогічне поняття вводиться для випадкового вектора.

Означення. Нехай випадковий вектор $\vec{\xi}\in R^q$. Розглянемо сукупність множин $\left\{\omega:\vec{\xi}(\omega)\in B\right\}$, $B\in\sigma_{R^q}$. Найменша σ-алгебра, що містить у собі ці множини, називається σ*-алгеброю, яка породжена випадковим вектором* $\vec{\xi}$. Позначення – $\sigma_{\vec{\xi}}$.

Умовні математичне сподівання та дисперсія випадкової величини η зручно вводити одразу відносно випадкового вектора $\vec{\xi}$ $\left(\vec{\xi}\in R^q,\ q\in N\right)$, а відносно випадкової величини отримуємо, як частинний випадок.

Означення. *Умовним математичним сподіванням випадкової величини* η *відносно випадкового вектора* $\vec{\xi}$ називається умовне математичне сподівання випадкової величини η відносно σ-алгебри, яка

породжена випадковим вектором $\vec{\xi}$, тобто відносно $\sigma_{\vec{\xi}}$. Позначення – $M\left(\eta/\vec{\xi}\right)$ (або $M\left(\eta/\vec{\xi}\right)(\omega)$).

Означення. *Умовною дисперсією випадкової величини* η *відносно випадкового вектора* $\vec{\xi}$ називається випадкова величина, яка позначається $D\left(\eta/\vec{\xi}\right)$ (або $D\left(\eta/\vec{\xi}\right)(\omega)$) і визначається як:

$$D\left(\eta/\vec{\xi}\right)=M\left\{\left[\eta-M\left(\eta/\vec{\xi}\right)\right]^2/\vec{\xi}\right\}.$$

Наслідок. Для умовної дисперсії випадкової величини η відносно випадкового вектора $\vec{\xi}$ справедливо

$$D\left(\eta/\vec{\xi}\right)=M\left\{\left[\eta-M\left(\eta/\vec{\xi}\right)\right]^2/\vec{\xi}\right\}=M\left(\eta^2/\vec{\xi}\right)-\left[M\left(\eta/\vec{\xi}\right)\right]^2.$$

Подібним чином вводиться відповідне поняття для умовної ймовірності події.

Означення. *Умовною ймовірністю події* B $\left(B\in F\right)$ *відносно випадкового вектора* $\vec{\xi}$ називається умовна ймовірність події B відносно σ-алгебри, яка породжена випадковим вектором $\vec{\xi}$. Позначення – $P\left(B/\vec{\xi}\right)$ (або $P\left(B/\vec{\xi}\right)(\omega)$).

У свою чергу, відповідні властивості умовних математичних сподівань випадкової величини відносно випадкового вектора $\vec{\xi}$ $\left(\vec{\xi}\in R^q\right)$ мають такий вигляд:

1) нехай $M|\eta|<\infty$, тоді існує $M\left(\eta/\vec{\xi}\right)$;
2) нехай $M|\eta_1|<\infty$, $M|\eta_2|<\infty$, $a,b\in R$, тоді
$$M\left(a\eta_1+b\eta_2/\vec{\xi}\right)=aM\left(\eta_1/\vec{\xi}\right)+bM\left(\eta_2/\vec{\xi}\right).$$
3) нехай $M|\eta|<\infty$, $M\left|\varphi\left(\vec{\xi}\right)\eta\right|<\infty$, $\varphi(\cdot)$ – борелівська функція на R^q, тоді $M\left\{\varphi(\vec{\xi})\eta/\vec{\xi}\right\}=\varphi(\vec{\xi})M\left(\eta/\vec{\xi}\right)$;
4) нехай $c\in R$, тоді $M\left(c/\vec{\xi}\right)=c$;
5) нехай $M|\eta|<\infty$, тоді $M\left\{M\left(\eta/\vec{\xi}\right)\right\}=M\eta$.

Теорема. Нехай на ймовірнісному просторі (Ω,F,P) задані дійсна випадкова величина η і випадковий вектор $\vec{\xi}\in R^q$, причому η є вимірною відносно σ-алгебри, яка породжена випадковим вектором $\vec{\xi}$, тобто відносно $\sigma_{\vec{\xi}}$. Тоді існує борелівська функція $\varphi(\cdot)$ на R^q така, що $\eta=\varphi\left(\vec{\xi}\right)$.

Зауваження. Зворотний результат очевидний. Нехай на ймовірнісному просторі (Ω,F,P) задані дійсна випадкова величина η і випадковий вектор $\vec{\xi}\in R^q$. Припустимо, що існує борелівська функція $\varphi(\cdot)$ на R^q така, що $\eta=\varphi\left(\vec{\xi}\right)$, тоді η є вимірною відносно σ-алгебри, яка породжена випадковим вектором $\vec{\xi}$, тобто відносно $\sigma_{\vec{\xi}}$.

Скористаємося останньою теоремою, аналізуючи $M\left(\eta/\vec{\xi}\right)$. Оскільки згідно з означенням $M\left(\eta/\vec{\xi}\right)$ є вимірною відносно σ-алгебри $\sigma_{\vec{\xi}}$, то існує борелівська функція $f(\cdot)$ на R^q така, що $M\left(\eta/\vec{\xi}\right)=f\left(\vec{\xi}\right)$. Останнє дозволяє сформулювати таке означення.

Означення. Для $M\left(\eta/\vec{\xi}\right)$ існує борелівська функція $f(\cdot)$ на R^q така, що $M\left(\eta/\vec{\xi}\right)=f\left(\vec{\xi}\right)$. *Умовне математичне сподівання випадкової величини* η *відносно події* $\left\{\vec{\xi}=\vec{x}\right\}$ *(або умовне математичне сподівання випадкової величини* η *за умови, що* $\vec{\xi}=\vec{x}$*) позначається* $M\left(\eta/\vec{\xi}=\vec{x}\right)$ і визначається таким чином: $M\left(\eta/\vec{\xi}=\vec{x}\right)=f\left(\vec{x}\right)$, $\vec{x}\in R^q$.

Зауваження 1. Функцію $f(\vec{x})=M(\eta/\vec{\xi}=\vec{x})$ називають *функцією регресії* η *на* $\vec{\xi}$ *(або функцією регресії* η *щодо* $\vec{\xi}$*).*

Зауваження 2. Легко бачити, що $Mf(\vec{\xi})=M\left\{M(\eta/\vec{\xi})\right\}=M\eta$.

Означення. *Умовна дисперсія випадкової величини* η *відносно події* $\left\{\vec{\xi}=\vec{x}\right\}$ *(або умовна дисперсія випадкової величини* η *за умови, що* $\vec{\xi}=\vec{x}$*)* позначається $D\left(\eta/\vec{\xi}=\vec{x}\right)$ і визначається таким чином:

$$D\left(\eta/\vec{\xi}=\vec{x}\right)=M\left\{\left[\eta-M\left(\eta/\vec{\xi}=\vec{x}\right)\right]^2/\vec{\xi}=\vec{x}\right\}.$$

Наслідок. Для умовної дисперсії випадкової величини η відносно події $\{\vec{\xi} = \vec{x}\}$ справедливо

$$D(\eta/\vec{\xi} = \vec{x}) = M\left\{\left[\eta - M(\eta/\vec{\xi} = \vec{x})\right]^2/\vec{\xi} = \vec{x}\right\} =$$

$$= M(\eta^2/\vec{\xi} = \vec{x}) - \left[M(\eta/\vec{\xi} = \vec{x})\right]^2.$$

Зауваження 1. Для нотації $D(\eta/\vec{\xi} = \vec{x})$ також буде використовуватися таке позначення: $g(\vec{x}) = D(\eta/\vec{\xi} = \vec{x})$.

Зауваження 2. Очевидно, що

$$g(\vec{\xi}) = M\left\{\left[\eta - M(\eta/\vec{\xi})\right]^2/\vec{\xi}\right\}.$$

ДОДАТОК 2

Основні властивості функції регресії

Лема. Якщо η та $\vec{\xi}$ – випадкові величина й вектор, відповідно, а $M\eta^2 < \infty$, тоді для них справедливо:

$$D\eta = M\left\{\left(M(\eta/\vec{\xi}) - M\eta\right)^2\right\} + M\left\{M\left\{\left(\eta - M(\eta/\vec{\xi})\right)^2/\vec{\xi}\right\}\right\},$$

або скорочено $D\eta = Df(\vec{\xi}) + Mg(\vec{\xi})$.

Доведення. Використовуючи властивості умовного математичного сподівання випадкових величин відносно випадкового вектора $\vec{\xi}$, дисперсію η можна розписати таким чином:

$$D\eta = M(\eta - M\eta)^2 = M\left\{M\left\{(\eta - M\eta)^2/\vec{\xi}\right\}\right\} =$$

$$= M\left\{M\left\{\left[\left(\eta - M(\eta/\vec{\xi})\right) + \left(M(\eta/\vec{\xi}) - M\eta\right)\right]^2/\vec{\xi}\right\}\right\} =$$

$$= M\left\{M\left\{\left[\left(\eta - M(\eta/\vec{\xi})\right)^2 + 2\left(\eta - M(\eta/\vec{\xi})\right)\left(M(\eta/\vec{\xi}) - M\eta\right) + \right.\right.$$

$$\left.\left. + \left(M(\eta/\vec{\xi}) - M\eta\right)^2\right]/\vec{\xi}\right\}\right\} = M\left\{M\left\{\left(\eta - M(\eta/\vec{\xi})\right)^2/\vec{\xi}\right\}\right\} +$$

$$+2M\left\{\left(M(\eta/\vec{\xi}) - M\eta\right)M\left\{\left(\eta - M(\eta/\vec{\xi})\right)/\vec{\xi}\right\}\right\} +$$

$$+M\left\{M\left\{\left(M(\eta/\vec{\xi}) - M\eta\right)^2/\vec{\xi}\right\}\right\} =$$

$$= M\left\{M\left\{\left(\eta - M(\eta/\vec{\xi})\right)^2/\vec{\xi}\right\}\right\} + M\left\{M\left\{\left(M(\eta/\vec{\xi}) - M\eta\right)^2/\vec{\xi}\right\}\right\} =$$

$$= M\left\{M\left\{\left(\eta - M(\eta/\vec{\xi})\right)^2/\vec{\xi}\right\}\right\} + M\left\{\left(M(\eta/\vec{\xi}) - M\eta\right)^2\right\} =$$

$$= Mg(\vec{\xi}) + Df(\vec{\xi}).$$

Що й треба було довести.

Теорема. Нехай $M\eta^2 < \infty$, Φ – множина борелівських функцій на R^q, тоді

$$f(\cdot) = \arg \min_{\varphi(\cdot)\in\Phi} M\left[\eta - \varphi(\vec{\xi})\right]^2,$$

де $f(x) = M(\eta/\vec{\xi} = \vec{x})$.

Доведення. Без втрати загальності будемо розглядати ті борелівські функції на R^q, для яких $M\varphi^2(\vec{\xi}) < \infty$. Скориставшись властивостями умовного математичного сподівання випадкової величини η відносно випадкового вектора $\vec{\xi}$ і врахувавши, що $M\left[f(\vec{\xi}) - \varphi(\vec{\xi})\right]^2 \geq 0$, легко бачити, що має місце ланцюжок перетворень:

$$M\left[\eta - \varphi(\vec{\xi})\right]^2 = M\left[\left(\eta - f(\vec{\xi})\right) + \left(f(\vec{\xi}) - \varphi(\vec{\xi})\right)\right]^2 =$$

$$= M\left[\eta - f(\vec{\xi})\right]^2 + 2M\left[\left(\eta - f(\vec{\xi})\right)\left(f(\vec{\xi}) - \varphi(\vec{\xi})\right)\right] +$$

$$+M\left[f(\vec{\xi}) - \varphi(\vec{\xi})\right]^2 \geq M\left[\eta - f(\vec{\xi})\right]^2 +$$

$$+2M\left\{M\left[\left(\eta - f(\vec{\xi})\right)\left(f(\vec{\xi}) - \varphi(\vec{\xi})\right)/\vec{\xi}\right]\right\} = M\left[\eta - f(\vec{\xi})\right]^2 +$$

$$+2M\left\{\left(f(\vec{\xi}) - \varphi(\vec{\xi})\right)M\left[\left(\eta - f(\vec{\xi})\right)/\vec{\xi}\right]\right\} = M\left[\eta - f(\vec{\xi})\right]^2.$$

В останньому переході скористалися тим, що $M\left[\left(\eta - f(\vec{\xi})\right)/\vec{\xi}\right] = 0$.

У результаті, маємо

$$M\left[\eta-\varphi\left(\vec{\xi}\right)\right]^2 \geq M\left[\eta-f\left(\vec{\xi}\right)\right]^2,$$

тобто отримано нижню межу для нашого функціоналу $M\left[\eta-\varphi\left(\vec{\xi}\right)\right]^2$, яка досягається на функції регресії $f(x)=M(\eta/\vec{\xi}=\vec{x})$.

Доведення завершено.

СПИСОК ЛІТЕРАТУРИ

1. *Айвазян С.А., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Исследование зависимостей. – М.: Финансы и статистика, 1985.
2. *Афифи А., Эйзен С.* Статистический анализ. Подход с использованием ЭВМ. – М.: Мир, 1982.
3. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1983.
4. *Кендалл М., Стьюарт А.* Статистические выводы и связи. – М.: Наука, Гл. редакция физ.-мат. лит-ры, 1973.
5. *Кендэл М.* Ранговые корреляции. – М.: Статистика, 1975.
6. *Крамер Г.* Математические методы статистики. – М.: Мир, 1975.
7. *Кульбак С.* Теория информации и статистика. – М.: Наука, 1967.
8. *Рао С.Р.* Линейные статистические методы и их применение. – М.: Наука, 1968.
9. *Ширяев А.Н.* Вероятность. – М.: Наука, Гл. редакция физ.-мат. литры, 1980.

Предметний покажчик

Покажчик позначень

З М І С Т