ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

М.А. Харченко

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Учебное пособие для вузов

Издательско-полиграфический центр
Воронежского государственного университета
2008

Утверждено научно-методическим советом факультета философии и психологии 25 марта 2008 г., протокол № 1400-03

Рецензент Н.И. Вьюнова

Учебное пособие подготовлено на кафедре общей и социальной психологии факультета философии и психологии Воронежского государственного университета.

Рекомендовано для студентов 2-го курса очной и 4 курса очно-заочной форм обучения отделения психологии факультета философии и психологии ВГУ.

Для специальности: 030301 – Психология
ОПД.Ф.11

СОДЕРЖАНИЕ

I. СТАТИСТИЧЕСКИЕ МЕРЫ СВЯЗИ МЕЖДУ СЛУЧАЙНЫМИ ВЕЛИЧИНАМИ

§ 1. Понятие корреляции

Величины могут быть либо *независимыми*, либо связанными *функциональной* или *стохастической* (*вероятностной*) зависимостью.

Функциональная зависимость величин реализуется тогда, когда каждому значению одной величины (аргумента) соответствует определенное *значение* другой величины. Примером функциональной зависимости является длина окружности $l = 2\pi r$ в зависимости от ее радиуса r. Очевидно, для случайных величин такого соответствия нет, поэтому строгие функциональные зависимости встречаются лишь тогда, когда величины не подвержены действию случайных факторов.

В большинстве случаев между переменными существуют зависимости, при которых каждому значению одной величины (аргумента) соответствует не какое-то определенное значение другой величины, а множество ее возможных значений – определенное *распределение*. Такая зависимость называется *стохастической*, или *вероятностной*.

Например, с одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай. Случайные величины – количество внесенных удобрений и собранный урожай – связаны друг с другом стохастической зависимостью: вторая переменная подвержена влиянию целого ряда факторов помимо количества внесенных удобрений (количество осадков, температура воздуха и др.). Кроме того, измерение значений обеих переменных неизбежно сопровождается случайными ошибками.

Частным случаем вероятностной зависимости является ***корреляционная зависимость*** – *стохастическая зависимость между случайными величинами, при которой наблюдается функциональная зависимость между значениями одной величины и средними значениями другой величины*.

Вернемся к рассмотренному выше примеру. Связь между количеством удобрений и собранным урожаем корреляционная, потому что, как показывает опыт, *средний* урожай и количество внесенных в почву удобрений связаны друг с другом функциональной зависимостью.

Термин «корреляция» (от лат. *correlatio* – соотношение, связь, зависимость) появился в XIX в. благодаря работам английского математика *Карла Пирсона (Pearson)* (1857–1936) и английского антрополога и психолога *Френсиса Гальтона (Galton)* (1882–1911).

Изображенные на координатной плоскости точки (x_i, y_i), где x_i и y_i – значения первой и второй переменных, называются *корреляционным полем* (рис. 1*а*). Аналитическая функция, аппроксимирующая (приближенно описывающая) наблюдаемые эмпирические значения, называется *функцией регрессии* (рис. 1*б*).

Название функции (от лат. *regressio* – движение назад) дал Ф. Гальтон, который, изучая зависимость между ростом родителей и их детей, обнаружил явление регрессии к среднему: у детей, родившихся у очень высоких родителей, рост имел тенденцию быть ближе к средней величине.

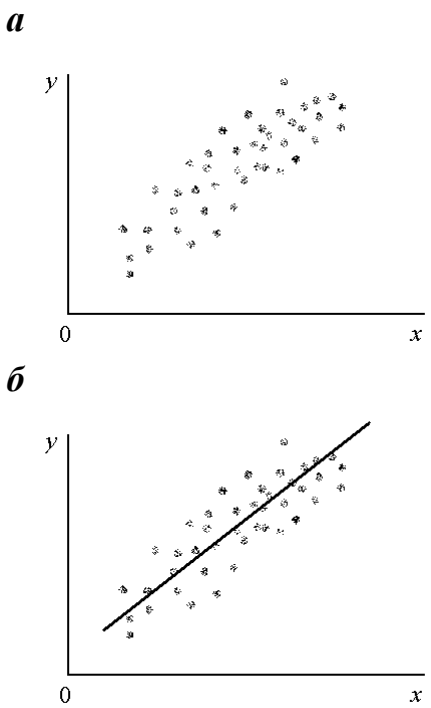


Рис. 1. Корреляционное поле (*а*) и функция регрессии (*б*)

Функция регрессии отражает тенденцию изменения одной величины под действием другой и строится таким образом, чтобы эмпирические точки корреляционного поля лежали как можно ближе к ней. Функция регрессии может быть линейной, параболической, гиперболической, логарифмической и др.

$$\begin{matrix} Y \\ \updownarrow \\ X \leftrightarrow \boldsymbol{Z} \end{matrix} \quad \not\Rightarrow X \leftrightarrow Y$$

1. **Наличие корреляционной зависимости между переменными не всегда означает наличие непосредственной связи этих величин друг с другом**: наблюдаемая связь часто существует благодаря другим переменным (не двум рассматриваемым), а изучаемые величины могут быть связаны между собой через латентные (скрытые от исследователя) переменные.

Примером подобного артефакта («искусственного» результата) является обнаруженная американскими психологами зависимость между уровнем интеллекта и уровнем дохода человека. Латентной переменной, обусловливающей эту корреляцию, является структура общества: подобное исследование, проведенное в современной России, дает иные результаты.

Другим примером является корреляция скорости опознавания изображения при его тахископическом (быстро пульсирующем) предъявлении и словарный запас человека (латентная переменная – общий интеллект испытуемого). Как видно, взаимосвязи переменных в психологии слишком сложны, чтобы их можно было объяснить единственной причиной.

2. Корреляционная связь, в отличие от функциональной, показывает лишь *тенденцию* изменения одной величины под действием другой, следовательно, на **основании корреляции можно утверждать лишь о степени связи между переменными, но не о существовании причинно-следственной зависимости между ними**. Другими словами, факт корре-

ляции переменных отнюдь не означает, что одна из них вызывает другую, однако дает возможность выдвинуть такую гипотезу. Например, между успеваемостью ребенка в начальных классах и возрастом, в котором он научился читать, имеется корреляционная зависимость. Однако из этого факта вовсе не следует причинная зависимость: можно встретить слабоуспевающего ребенка, который научился читать задолго до поступления в школу, и наоборот. Тем не менее, такая гипотеза не лишена оснований.

Рассмотрим другой пример. Имеется корреляционная зависимость между уровнем тревожности студентов и результатами их тестирования по окончанию курса «Математические методы в психологии». Этот факт можно объяснить, с одной стороны тем, что волнение, испытываемое частью студентов, могло привести к тому, что они хуже справились с тестовым заданием, а более спокойные студенты оказались в состоянии успешно проявить свои способности. Но разве не столь же правдоподобно считать, что сам тест является фактором, вызывающим беспокойство? Менее способные (как правило, более ленивые) студенты пугаются тестирования, а способные и ответственные не находят в нем ничего тревожного, кроме очередной проверки знаний. Причинно-следственную связь невозможно интерпретировать без экспериментальной проверки.

3. Иногда в психологических исследованиях устанавливается *случайная корреляция*, не обусловленная никакой причиной. Примером такой корреляции является связь между тревожностью и успеваемостью по английскому языку у школьников средних классов. Свидетельствует ли это о том, что повышенная тревожность заставляет учащегося усерднее трудиться? Вовсе нет. При проверке по шкале тревожности Тейлор девочки показывают более высокие показатели, чем мальчики. Известно также, что девочки в средних классах, как правило, имеют более высокие оценки по английскому языку по сравнению с мальчиками. Установить подобную связь отдельно для мальчиков и девочек еще не удалось никому.

§ 2. Свойства корреляционной связи

Корреляционные связи различаются по *тесноте (силе)* связи и *количеству признаков*.

По тесноте (силе) корреляционной связи принято выделять: 1) *функциональную*, 2) *тесную* (*сильную*), 3) *среднюю* (*умеренную*), 4) *слабую* и 5) *нулевую* (*отсутствующую*) виды связи.

По количеству признаков корреляция может быть *парной* (между двумя признаками) и *множественной* (между несколькими признаками).

Форма парной корреляции может быть *линейной*, описываемой линейной функцией регрессии, и *нелинейной* (*криволинейной*), описываемой нелинейными функциями (рис. 2). Нелинейные связи в психологических исследованиях встречаются чаще.

Линейная корреляция наблюдается, например, между личностной пластичностью и склонностью к смене социальных установок (рис. 2*а*).

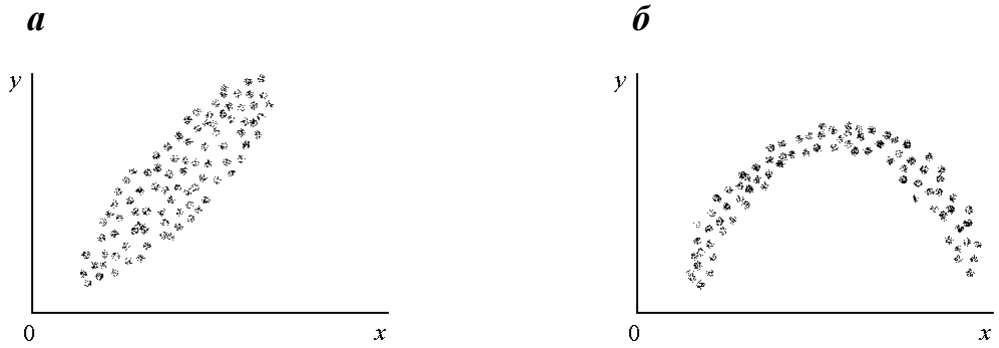


Рис. 2. Линейная (*а*) и нелинейная (*б*) корреляционные связи

Известным примером нелинейной корреляции является первый закон Йеркса–Додсона: по мере увеличения интенсивности мотивации качество деятельности изменяется по колоколообразной кривой: сначала повышается, а затем постепенно снижается (рис. 2*б*). Другим примером нелинейной связи является закон Хика: скорость переработки информации пропорциональна логарифму от числа альтернатив.

Заключение о форме парной корреляционной связи можно сделать, изобразив корреляционное поле на координатной плоскости.

Парная линейная корреляция, в свою очередь, может быть *положительной* (*«прямой»*) и *отрицательной* (*«обратной»*). При положительной корреляции при возрастании одного признака в среднем увеличивается другой, в случае же отрицательной корреляции при возрастании одного признака другой в среднем уменьшается.

Примеры: уровень личностной тревожности положительно коррелирует с риском заболеть язвой желудка, число детей в семье отрицательно коррелирует с показателем их интеллекта, возрастание громкости звука сопровождается ощущением повышения его тона, а количество ежедневно выкуриваемых сигарет отрицательно коррелирует с продолжительностью жизни.

Как видно из рисунков 2–6, в случае парной линейной корреляции корреляционное поле представляет собой эллипс. При этом чем теснее корреляционная связь, тем эллипс более сжат; в случае функциональной связи он преобразуется в прямую, а при отсутствии связи – в круг.

§ 3. Коэффициент детерминации и корреляционное отношение

Возможность косвенной оценки одних характеристик через другие связана с тем, что они оказываются зависимыми от ряда общих факторов.

Если один и тот же фактор F действует на обе переменные, то между ними в эмпирическом исследовании всегда обнаруживается корреляция. При этом наблюдаемый разброс переменных X и Y обусловлен не только действием общего фактора F, но и другими причинами, иррелевантными по отношению к нему. В результате каждому значению переменной X соответствует распределение переменной Y, а не определенное ее значение:

а ***б***

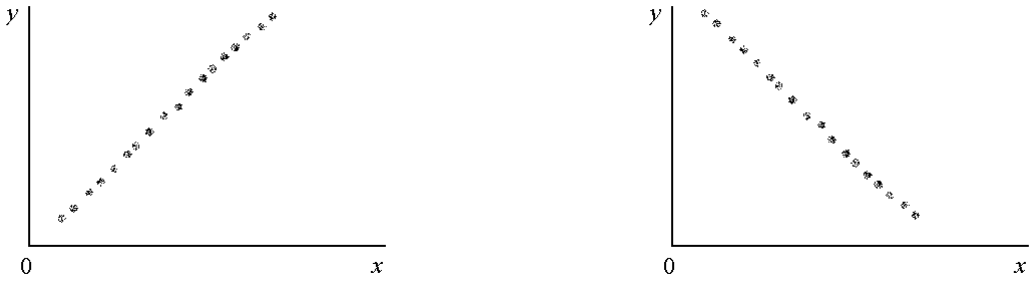


Рис. 3. Положительная (*а*) и отрицательная (*б*) линейные функциональные связи

а ***б***

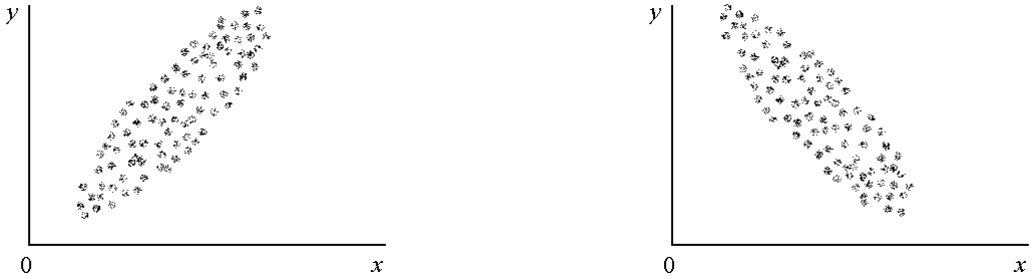


Рис. 4. Положительная (*а*) и отрицательная (*б*) тесные (сильные) линейные корреляционные связи

а ***б***

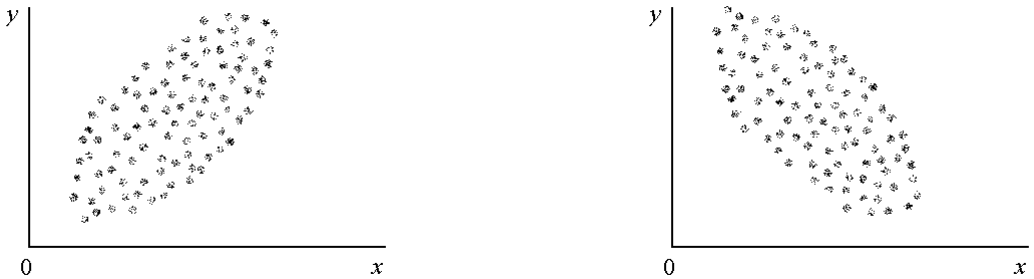


Рис. 5. Положительная (*а*) и отрицательная (*б*) средние (умеренные) линейные корреляционные связи

а ***б***

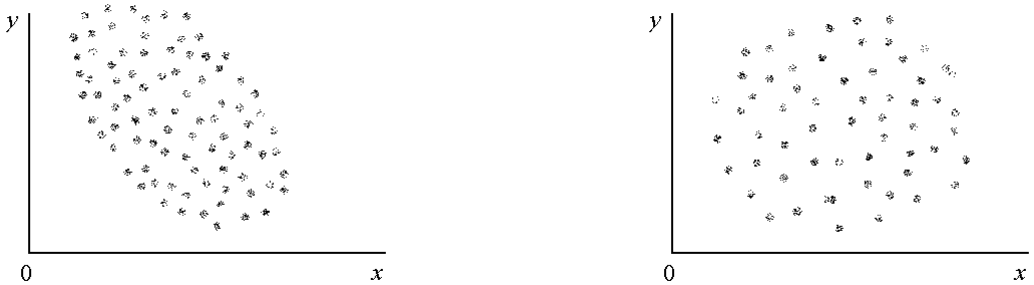


Рис. 6. Слабая (*а*) и нулевая (*б*) связи

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_{11}, y_{12}, y_{13} \ldots \end{pmatrix}, \begin{pmatrix} x_2 \\ y_{21}, y_{22}, y_{23} \ldots \end{pmatrix} \ldots$$

Другими словами, изменение переменной *Y* при изменении *X* может быть представлено в виде двух составляющих (релевантной общему фактору *F* и иррелевантной ему), а дисперсия измеряемой величины складывается из «факторной» σ_F^2 и «остаточной» σ_0^2 компонент:

$$\sigma^2 = \sigma_F^2 + \sigma_0^2 .$$

Факторная компонента дисперсии связана с зависимостью рассматриваемых величин и определяется действием общего фактора на обе переменные. **Если факторная составляющая дисперсии равна нулю, изучаемые величины являются независимыми**.

Остаточная компонента дисперсии обусловливает действие множества других факторов, не влияющих на обе переменные одновременно (индивидуальных различий испытуемых, ошибок измерения и др.). Остаточная дисперсия приводит к случайному разбросу значений y_i при фиксированном значении x_i, поэтому она называется также «случайной». **При нулевой случайной дисперсии наблюдается функциональная зависимость между величинами** (каждому значению одной величины соответствует единственное значение другой).

Таким образом, соотношение между факторной и случайной компонентами дисперсии может служить количественной мерой тесноты (силы) корреляционной связи между величинами: чем больше доля факторной компоненты в общей дисперсии, тем связь между величинами ближе к функциональной.

Отношение факторной дисперсии к полной называется ***коэффициентом детерминации***:

$$R = \frac{\sigma_F^2}{\sigma^2} = \frac{\sigma_F^2}{\sigma_F^2 + \sigma_0^2} .$$

Коэффициент детерминации является безразмерной неотрицательной величиной, изменяющейся от 0 до 1 (его часто выражают в процентах). Он показывает долю общей вариации одной переменной, обусловленной изменчивостью другой переменной.

Величина коэффициента детерминации не меняется при увеличении или уменьшении на одно и то же число или в одно и то же число раз всех значений переменных.

При отсутствии какой-либо связи между рассматриваемыми величинами ($\sigma_F^2 = 0$) коэффициент детерминации равен нулю, а в случае функциональной зависимости между ними ($\sigma_0^2 = 0$, то есть когда 100 % вариации первой переменной обусловлены изменчивостью второй переменной) коэффициент детерминации равен единице:

$$R = \frac{\sigma_F^2}{\sigma^2} = \frac{\sigma^2 - \sigma_0^2}{\sigma^2} = 1 - \frac{\sigma_0^2}{\sigma^2}.$$

Чем ближе коэффициент детерминации к единице, тем наблюдаемые значения теснее примыкают к линии регрессии, а уравнение регрессии лучше описывает зависимость переменных.

Квадратный корень из коэффициента детерминации называется ***корреляционным отношением***:

$$\eta = \sqrt{R} = \frac{\sigma_F}{\sigma} = \frac{\sigma_F}{\sqrt{\sigma_F^2 + \sigma_0^2}}.$$

Корреляционное отношение является универсальной количественной оценкой тесноты (силы) корреляционной связи, т.к. может быть применимо к корреляционной связи любой формы (как линейной, так и нелинейной). Оно, как и коэффициент детерминации, является безразмерной неотрицательной величиной, изменяющейся от 0 до 1.

Для независимых случайных величин $\eta = 0$, корреляционное поле в этом случае представляет собой круг. В случае функциональной связи корреляционное отношение равно единице, и все наблюдаемые значения располагаются строго на линии регрессии. В остальных случаях корреляционное отношение принимает значения, заключенные между нулем и единицей: $0 < \eta < 1$.

Отличные от нуля значения η являются *достаточным* условием установления корреляционной зависимости между исследуемыми признаками (корреляционной, а не причинно-следственной!). Чем ближе корреляционное отношение к единице, тем с бóльшим основанием можно считать, что изучаемые величины находятся в корреляционной зависимости:

$\eta = 0$	$\Leftrightarrow$	независимость величин
$\eta = 1$	$\Leftrightarrow$	функциональная зависимость
$\eta \neq 0$	$\Leftrightarrow$	корреляционная зависимость

Теснота (сила) связи между величинами измеряется величиной корреляционного отношения. С возрастанием η корреляционная связь становится более тесной:

- $\eta = 1$ – величины связаны функциональной зависимостью;
- $0{,}95 \le \eta < 1$ – связь очень сильная, практически функциональная;
- $0{,}75 \le \eta < 0{,}95$ – связь тесная (сильная);
- $0{,}5 \le \eta < 0{,}75$ – связь средняя (умеренная);
- $0{,}2 \le \eta < 0{,}5$ – связь слабая;
- $0 \le \eta < 0{,}2$ – практически нет связи.

Корреляционное отношение η позволяет установить лишь силу корреляционной связи; форму корреляционной зависимости можно определить только на основании графического метода.

§ 4. Коэффициенты ковариации и корреляции

Для оценки тесноты (силы) линейной связи служат коэффициенты *ковариации*[1] и *корреляции*.

Коэффициент ковариации представляет собой *математическое ожидание произведения отклонений величин от их мат. ожиданий*[2]:

$$\operatorname{cov}(X, Y) = M[(X - M(X))\,(Y - M(Y))].$$

Если рассматриваемые величины независимы, то коэффициент ковариации равен нулю:[3]

$$\operatorname{cov}(X, Y) = M[(X \cdot Y - M(X) \cdot Y - X \cdot M(Y) + M(X) \cdot M(Y)] =$$
$$= M(X) \cdot M(Y) - M(X) \cdot M(Y) - M(X) \cdot M(Y) + M(X) \cdot M(Y) = 0.$$

В случае же линейной связи между величинами коэффициент ковариации отличен от нуля.

Вследствие того, что значение коэффициента ковариации зависит от единиц измерения изучаемого признака, то его значение меняется при изменении масштаба измерительной шкалы. Например, в зависимости между показателем интеллекта и уровнем месячного дохода человека коэффициент ковариации изменится в 1000 раз, если величину дохода выразить не в рублях, а в тысячах рублей.

Очевидно, удобный показатель тесноты связи должен иметь стандартную систему единиц измерения, в которой данные по различным характеристикам были бы сравнимы между собой (или быть безразмерной величиной). Для получения такого показателя коэффициент ковариации необходимо нормировать – разделить на максимально возможное значение ковариации $\sigma_x \cdot \sigma_y$.

Нормированное значение коэффициента ковариации называется ***коэффициентом корреляции***:

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_x \cdot \sigma_y} = \frac{M\left[\left(X - M(X)\right) \cdot \left(Y - M(Y)\right)\right]}{\sigma_x \cdot \sigma_y},$$

где σ_x, σ_y – средние квадратические отклонения переменных X и Y; $M(X)$, $M(Y)$ – их математические ожидания.

Коэффициент корреляции представляет собой безразмерную величину, изменяющуюся в пределах от –1 до 1. Значение коэффициента корреляции выражает лишь долю от максимально возможной ковариации, в чем и состоит его преимущество перед коэффициентом ковариации.

Величина коэффициента корреляции не меняется при увеличении или уменьшении на одно и то же число или в одно и то же число раз всех значений переменных.

---

[1] Ковариация – от лат. *con* и *variare* – совместная изменчивость.

[2] Ковариация по каждому аргументу удовлетворяет свойствам математического ожидания:

$$\operatorname{cov}(C_1X, C_2Y) = C_1C_2 \cdot \operatorname{cov}(X, Y),$$

а ковариация переменной самой с собой представляет собой дисперсию величины:

$$\operatorname{cov}(X, X) = M[(X - M(X))\,(X - M(X))] = M[X - M(X)]^2 = D(X).$$

[3] Математическое ожидание раскрыто с учетом того, что X и Y – независимые случайные величины, а $M(X)$ и $M(Y)$ – константы.

При $\rho = \pm 1$ корреляционная связь представляет собой **линейную функциональную зависимость**[4], при этом все наблюдаемые значения располагаются на прямой (рис. 3). При $\rho = 0$ **линейная** корреляционная **связь отсутствует**, корреляционное поле представляет собой круг (рис. 6). Чем ближе значение $|\rho|$ к единице, тем с бóльшим основанием можно считать, что изучаемые величины находятся в линейной зависимости.

Для независимых случайных величин коэффициент корреляции равен нулю. Однако **равенство нулю коэффициента корреляции не всегда означает независимость случайных величин**: оно свидетельствует лишь об отсутствии линейной корреляционной зависимости между изучаемыми величинами, но не корреляционной зависимости вообще. Коэффициент корреляции может быть равен нулю в случае нелинейной связи. Итак,

независимость случайных величин	$\Rightarrow$ $\not\Leftarrow$	$\rho = 0$
$\rho \neq 0$	$\Rightarrow$ $\not\Leftarrow$	корреляционная зависимость

Направление линейной корреляционной **связи определяется знаком коэффициента корреляции**: для «прямой», положительной связи $\rho > 0$, для «обратной», отрицательной связи $\rho < 0$.

Теснота (сила) линейной **связи** между случайными величинами **определяется абсолютной величиной коэффициента корреляции**:

- $|\rho| = 1$ – величины связаны линейной функциональной зависимостью;
- $0{,}95 \le |\rho| < 1$ – связь очень сильная, практически функциональная;
- $0{,}75 \le |\rho| < 0{,}95$ – связь тесная (сильная);
- $0{,}5 \le |\rho| < 0{,}75$ – связь средняя (умеренная);
- $0{,}2 \le |\rho| < 0{,}5$ – связь слабая;
- $0 \le |\rho| < 0{,}2$ – практически нет связи.

Коэффициент корреляции связан с корреляционным отношением следующим образом: $0 \le |\rho| \le \eta \le 1$.

В случае линейной связи они равны: $|\rho| = \eta$, поэтому:

1) расхождение между коэффициентом корреляции и корреляционным отношением используется в качестве критерия линейности корреляционной зависимости;

2) в случае линейной связи коэффициент детерминации равен квадрату коэффициента корреляции.

---

[4] Если $Y = aX + b$ $(a \neq 0)$, то $\operatorname{cov}(X, Y) = \operatorname{cov}(X, aX + b) = M[(X - M(X))\,(aX + b - M(aX + b))] =$
$= M[(X - M(X))\,(aX + b - a \cdot M(X) - b)] = a \cdot M[(X - M(X))\,(X - M(X))] = a \cdot D(X)$;

$$\sigma_x \cdot \sigma_y = \sqrt{D(X)} \cdot \sqrt{D(aX+b)} = \sqrt{D(X) \cdot a^2 D(X)} = |a| \cdot D(X);$$

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{a \cdot D(X)}{|a| \cdot D(X)} = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}.$$

II. ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

§ 5. Коэффициент линейной корреляции Пирсона

Коэффициент линейной корреляции Пирсона используется для оценки тесноты (силы) связи между двумя переменными в случаях, если:

1) **рассматриваемая связь линейная**;

2) **обе переменные измерены в сильных шкалах** (реляционной или интервальной).

Коэффициент линейной корреляции Пирсона

$$r = \frac{\overline{\mathrm{cov}}(X,Y)}{s_x s_y}$$

представляет собой отношение выборочного коэффициента ковариации

$$\overline{\mathrm{cov}}(X,Y) = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

к произведению выборочных средних квадратических отклонений s_x, s_y:

$$s_x = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}, \qquad s_y = \sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2},$$

где x_i, y_i – числовые значения рассматриваемых переменных, n – объем выборки. После подстановки имеем:

$$\boxed{r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\cdot\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}}.$$

Величина выборочного коэффициента линейной корреляции Пирсона, как и генерального, изменяется в пределах от –1 до +1.

При **малом объеме выборки ($n < 100$)** значение коэффициента линейной корреляции Пирсона необходимо корректировать по формуле:

$$\boxed{r' = r \cdot \left[1 + \frac{1 - r^2}{2(n-3)}\right]}.$$

§ 6. Проверка гипотезы о значимости выборочного коэффициента линейной корреляции Пирсона

Выборочный коэффициент линейной корреляции Пирсона, как и все

выборочные характеристики, является случайной величиной и при повторении измерений может принимать другие значения. Поэтому для независимых случайных величин, для которых генеральный коэффициент корреляции ρ равен нулю, выборочный коэффициент r может заметно отличаться от нуля, и наоборот. В связи с этим *всегда* возникает важная практическая задача, заключающаяся в **проверке значимости** выборочного коэффициента корреляции.

Нулевая гипотеза h_0 заключается в отсутствии *линейной* корреляционной связи между исследуемыми переменными в генеральной совокупности: $\rho = 0$. Альтернативной гипотезой h_1 является утверждение о том, что генеральный коэффициент корреляции ρ отличен от нуля: $\rho \neq 0$.

Проверка нулевой гипотезы осуществляется по-разному, в зависимости от объема выборки.

1. Большой объем выборки ($n \geq 100$).

Проверка нулевой гипотезы осуществляется с помощью критерия Стьюдента и заключается в вычислении величины

$$| t | = \frac{| r |}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2},$$

которая затем сравнивается с критическими значениями $t_\alpha(df)$ для выбранного уровня значимости α и числа степеней свободы $df = n - 2$.

Если значение $| t |$ попадает в область допустимых значений, то есть если выполняется условие $| t | \leq t_{0,05}(n - 2)$, нулевая гипотеза $\rho = 0$ не отвергается. Считается, что в этом случае линейная связь между рассматриваемыми переменными отсутствует.

Выборочный коэффициент линейной корреляции Пирсона значимо (существенно) отличается от нуля, если эмпирическое значение $| t |$ попадает в критическую область критерия, то есть если $| t | > t_{0,01}(n - 2)$.

Для значимого коэффициента корреляции рассчитывается **доверительный интервал**, который с вероятностью $P = 1 - \alpha$ содержит неизвестный генеральный коэффициент корреляции ρ. Границы доверительного интервала находятся по формуле

$$\rho = r \pm t_\alpha (n - 2) \cdot \frac{1 - r^2}{\sqrt{n - 1}}.$$

2. Ограниченный объем выборки ($n < 100$).

Для проверки гипотезы об отсутствии корреляции между исследуемыми величинами используется преобразование Фишера

$$u = \frac{1}{2} \ln \frac{1 + r'}{1 - r'},$$

где r' – скорректированное значение выборочного коэффициента корреляции. Проверка нулевой гипотезы $\rho = 0$ заключается в вычислении значения u и сопоставления его с критическим

$$u_\alpha(n) = z_{1-\alpha/2}\,\frac{1}{\sqrt{n-3}},$$

где $z_{1-\alpha/2}$ – квантили нормированного распределения: $z_{1-\alpha/2} = 1{,}960$ для $\alpha = 0{,}05$ и $z_{1-\alpha/2} = 2{,}576$ для $\alpha = 0{,}01$.

Если эмпирическое значение u попадает в область допустимых значений, то есть если выполняется условие $|u| \le u_{\alpha n}$, нулевая гипотеза $\rho = 0$ не отвергается. Считается, что линейной корреляционной связи между рассматриваемыми величинами нет.

Корреляция считается значимой, если эмпирическое значение u попадает в критическую область: $|u| > u_\alpha(n)$.

Границы **доверительного интервала** для генерального коэффициента корреляции при ограниченном объеме выборки определяются как

$$r_1 < \rho < r_2,$$

где r_1 и r_2 находятся из выражения $u = \frac{1}{2}\ln\frac{1+r}{1-r}$ для $u_1 = u - u_\alpha(n)$ и $u_2 = u + u_\alpha(n)$:

$$r = \frac{e^{2u}-1}{e^{2u}+1}.$$

Пример II.1. По результатам измерения уровня ригидности (X) и времени решения креативной задачи (Y) у 16 испытуемых (табл. 2.1) требуется произвести оценку корреляционной связи. Переменная X измерена в интервальной шкале (в T-баллах), переменная Y – в реляционной (в секундах).

Таблица 2.1

Расчет коэффициента линейной корреляции Пирсона

Испытуемый	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1. Арбузов	39,3	15,0	-5,65	-1,25	7,0625	31,9225	1,5625
2. Веткина	33,3	13,0	-11,65	-3,25	37,8625	135,7225	10,5625
3. Дунайский	56,6	20,9	11,65	4,65	54,1725	135,7225	21,6225
4. Ёжикова	62,3	19,0	17,35	2,75	47,7125	301,0225	7,5625
5. Зубовских	31,1	13,6	-13,85	-2,65	36,7025	191,8225	7,0225
6. Карпова	36,7	15,0	-8,25	-1,25	10,3125	68,0625	1,5625
7. Лукин	52,9	17,1	7,95	0,85	6,7575	63,2025	0,7225
8. Мороз	32,9	13,5	-12,05	-2,75	33,1375	145,2025	7,5625
9. Носов	35,2	14,2	-9,75	-2,05	19,9875	95,0625	4,2025
10. Орлова	62,8	21,3	17,85	5,05	90,1425	318,6225	25,5025
11. Пригожин	34,2	13,5	-10,75	-2,75	29,5625	115,5625	7,5625
12. Русалин	58,1	17,0	13,15	0,75	9,8625	172,9225	0,5625
13. Сёмченко	29,3	13,0	-15,65	-3,25	50,8625	244,9225	10,5625
14. Ушаков	59,9	18,2	14,95	1,95	29,1525	223,5025	3,8025
15. Федулина	49,0	19,2	4,05	2,95	11,9475	16,4025	8,7025
16. Яблоков	45,6	16,5	0,65	0,25	0,1625	0,4225	0,0625
	719,2	**260,0**			**475,4000**	**2260,1000**	**119,1400**

Решение. В качестве оценки коэффициента корреляции можно использовать коэффициент корреляции Пирсона, т.к. обе переменные измерены в сильных шкалах.

1. Вычисление коэффициента линейной корреляции Пирсона.

Составляем расчетную таблицу (табл. 2.1). Выборочные средние арифметические значения $\bar{x}$ и $\bar{y}$ находим по результатам первого и второго столбцов:

$$\bar{x} = \frac{719{,}2}{16} = 44{,}95; \qquad \bar{y} = \frac{260{,}0}{16} = 16{,}25.$$

Выборочный коэффициент линейной корреляции есть отношение суммы пятого столбца к квадратному корню из произведения сумм шестого и седьмого столбцов:

$$r = \frac{475{,}40}{\sqrt{2260{,}10 \cdot 119{,}14}} = 0{,}916.$$

Ввиду оценки корреляции по выборке малого объема необходима поправка:

$$r' = 0{,}916 \cdot \left[1 + \frac{1 - 0{,}916^2}{2 \cdot (16 - 3)}\right] = 0{,}922.$$

2. Проверка значимости коэффициента корреляции.

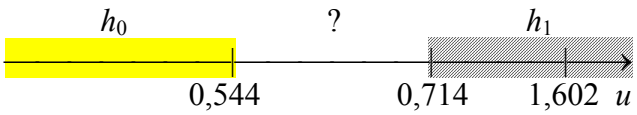
Нулевой гипотезой h_0 является предположение о том, что генеральный коэффициент корреляции равен нулю: $\rho = 0$, альтернативная гипотеза h_1 состоит в том, что генеральный коэффициент корреляции отличен от нуля: $\rho \neq 0$.

Для проверки нулевой гипотезы находим эмпирическое значение

$$u = \frac{1}{2} \ln \frac{1 + 0{,}922}{1 - 0{,}922} = 1{,}602,$$

которое сопоставляем с критическими значениями

$$u_{0,05} = \frac{1{,}96}{\sqrt{16 - 3}} = 0{,}544; \qquad u_{0,01} = \frac{2{,}576}{\sqrt{16 - 3}} = 0{,}714.$$



Эмпирическое значение $u = 1{,}602$ попадает в критическую область, что позволяет отвергнуть нулевую гипотезу. Коэффициент корреляции значимо отличается от нуля ($p < 0{,}01$).

Для построения 95 %-го доверительного интервала для генерального коэффициента корреляции находим $u_1 = 1{,}602 - 0{,}544 = 1{,}058$ и $u_2 = 1{,}602 + 0{,}544 = 2{,}146$.

Границы доверительного интервала находятся по формулам

$$r_1 = \frac{e^{2 \cdot 1{,}058} - 1}{e^{2 \cdot 1{,}058} + 1} = \frac{e^{2{,}116} - 1}{e^{2{,}116} + 1} = \frac{7{,}2979}{9{,}2979} = 0{,}785;$$

$$r_2 = \frac{e^{2 \cdot 2{,}146} - 1}{e^{2 \cdot 2{,}146} + 1} = \frac{e^{4{,}292} - 1}{e^{4{,}292} + 1} = \frac{72{,}1125}{74{,}1125} = 0{,}973.$$

Таким образом, результаты исследования свидетельствуют о наличии тесной ($| r | > 0{,}75$) прямой ($r > 0$) линейной корреляционной связи между уровнем ригидности испытуемого и временем решения им креативной задачи. Генеральный коэффициент корреляции с вероятностью 95 % лежит в интервале

$$0{,}785 < \rho < 0{,}973.$$

§ 7. Сравнение двух выборочных коэффициентов линейной корреляции Пирсона

Иногда необходимо сравнить два выборочных коэффициента линейной корреляции Пирсона с целью установления общности генерального коэффициента корреляции для двух рассматриваемых выборок.

Проверка нулевой гипотезы о незначимости различий между двумя коэффициентами линейной корреляции Пирсона h_0: $\rho_1 = \rho_2 = \rho$ при альтернативной гипотезе h_1: $\rho_1 \neq \rho_2$ заключается в вычислении величины

$$|z| = \frac{|u_1 - u_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}},$$

где u_1 и u_2 находятся для выборочных значений коэффициентов линейной корреляции Пирсона r_1 и r_2 по формуле

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

Вычисленное значение z сравнивают с квантилями нормального распределения $z_{1-\alpha/2}$ (1,960 для $\alpha = 0{,}05$ и 2,576 для $\alpha = 0{,}01$).

Нулевая гипотеза $\rho_1 = \rho_2$ не отвергается, если эмпирическое значение $|z|$ попадает в область допустимых значений: $|z| \leq z_{0{,}975} = 1{,}960$.

В случае попадания эмпирического значения $|z|$ в критическую область: $|z| > z_{0{,}995} = 2{,}576$ нулевая гипотеза $\rho_1 = \rho_2$ отвергается. В этом случае нельзя считать, что обе выборки взяты из общей генеральной совокупности и имеют один и тот же генеральный коэффициент корреляции.

Пример II.2. В примере II.1. по результатам обследования 16 испытуемых было получено значение коэффициента корреляции $r_1 = 0{,}922$ между уровнем ригидности и временем решения креативной задачи. При повторении исследования на выборке 100 человек значение коэффициента корреляции оказалось равным $r_2 = 0{,}880$. Требуется проверить гипотезу о незначимости различия между коэффициентами корреляции.

Решение. В примере II.1. для $r_1 = 0{,}922$ было получено значение $u_1 = 1{,}602$. Для $r_2 = 0{,}880$ имеем: $u_2 = 1{,}376$. Для проверки нулевой гипотезы h_0: $\rho_1 = \rho_2 = \rho$ вычисляем z

$$|z| = \frac{|1{,}602 - 1{,}376|}{\sqrt{\frac{1}{16 - 3} + \frac{1}{100 - 3}}} = 0{,}765 < 1{,}960,$$

которая попадает в область допустимых значений. Это значит, что оба коэффициента корреляции характеризуют выборки, взятые из общей генеральной совокупности.

III. РАНГОВАЯ КОРРЕЛЯЦИЯ

В психологии часто возникает потребность анализа связи между переменными, которые не могут быть измерены в интервальной или реляционных шкалах, но тем не менее поддаются упорядочению и могут быть проранжированы по степени убывания или возрастания признака. Для определения тесноты связи между признаками, измеренными в **порядковых шкалах**, применяются методы *ранговой корреляции*. К ним относятся: *коэффициенты ранговой корреляции Спирмена и Кендалла* (используются для определения тесноты связи между двумя величинами) и *коэффициент конкордации* (устанавливает статистическую связь между несколькими признаками). Использование коэффициента линейной корреляции Пирсона в случае, когда о законе распределения и о типе измерительной шкалы отсутствует сколько-нибудь надежная информация, может привести к существенным ошибкам.

Методы ранговой корреляции могут быть использованы для определения тесноты связи не только между количественными переменными, но и между качественными признаками при условии, что их значения можно упорядочить и проранжировать. Эти методы также могут быть использованы применительно к признакам, измеренным в интервальных и реляционных шкалах, однако их эффективность в этом случае всегда будет ниже.

§ 8. Коэффициент ранговой корреляции Спирмена

Каждая из двух совокупностей располагается в виде вариационного ряда с присвоением каждому члену ряда соответствующего порядкового номера (*ранга*), выраженного натуральным числом. Одинаковым значениям ряда присваивают среднее ранговое число.

Сравниваемые признаки можно ранжировать в любом направлении: как в сторону ухудшения качества (ранг 1 получает самый большой, быстрый, умный и т. д. испытуемый), так и наоборот. Главное, чтобы обе переменные были проранжированы одинаковым способом.

Коэффициент ранговой корреляции Спирмена находится по формуле

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n^3 - n},$$

где d_i – разность рангов для каждой i–пары из n наблюдений.

Если в вариационных рядах для X и Y встречаются члены ряда с одинаковыми ранговыми числами, то в формулу для коэффициента корреляции Спирмена необходимо внести поправки T_x и T_y на одинаковые ранги:

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{(n^3 - n) - \frac{1}{2}(T_x + T_y)}, \qquad T = \sum_{k=1}^{l} (t_k^3 - t_k).$$

Здесь l – число групп в вариационном ряду с одинаковыми ранговыми числами; t_k – число членов в каждой из l групп.

Ранговый коэффициент корреляции Спирмена, как и линейный, изменяется от –1 до +1, однако значение рангового коэффициента корреляции Спирмена всегда меньше значения коэффициента линейной корреляции Пирсона: $r_S < r$.

Проверка **гипотезы о значимости** коэффициента ранговой корреляции Спирмена проводится по-разному в зависимости от объема выборки.

1. Объем выборки больше 30 ($n > 30$).

Проверка нулевой гипотезы h_0: $\rho = 0$ при альтернативной h_1: $\rho \neq 0$ осуществляется с помощью критерия Стьюдента и заключается в вычислении величины

$$| t | = \frac{| r_S |}{\sqrt{1 - r_S^2}} \cdot \sqrt{n - 2},$$

имеющей распределение Стьюдента с $df = n - 2$ степенями свободы. Эмпирическое значение сравнивается с критическими значениями $t_\alpha(n - 2)$.

Нулевая гипотеза $\rho = 0$ не отвергается, если эмпирическое значение попадает в область допустимых значений:

$$| t | \le t_{0,05}(df), \quad df = n - 2.$$

Коэффициент ранговой корреляции Спирмена значимо отличается от нуля, если эмпирическое значение попадает в критическую область:

$$| t | > t_{0,01}(df), \quad df = n - 2.$$

2. Очень малый объем выборки ($n \le 30$).

Проверка нулевой гипотезы осуществляется путем сравнения вычисленного коэффициента r_S с критическими значениями $r_\alpha(n)$, взятым из статистических таблиц для выбранного уровня значимости α и числа пар наблюдений n (табл. 3.1).

Нулевая гипотеза $\rho = 0$ не отвергается, если эмпирическое значение попадает в область допустимых значений:

$$| r_S | \le r_{0,05}(n).$$

Коэффициент ранговой корреляции Спирмена значимо отличается от нуля, если вычисленное значение попадает в критическую область:

$$| r_S | > r_{0,01}(n).$$

Таблица 3.1

Критические значения коэффициента ранговой корреляции Спирмена

n	*α*		*n*	*α*		*n*	*α*	
	0,05	**0,01**		**0,05**	**0,01**		**0,05**	**0,01**
7	0,745	0,893	15	0,518	0,654	23	0,415	0,531
8	0,690	0,857	16	0,500	0,632	24	0,406	0,520
9	0,663	0,817	17	0,485	0,615	25	0,398	0,510
10	0,636	0,782	18	0,472	0,598	26	0,389	0,500
11	0,609	0,754	19	0,458	0,582	27	0,383	0,491
12	0,580	0,727	20	0,445	0,568	28	0,375	0,483
13	0,555	0,698	21	0,435	0,555	29	0,368	0,474
14	0,534	0,675	22	0,424	0,543	30	0,362	0,466

Пример III.1. В методике С.А. Будаси испытуемому предлагается проранжировать 20 качеств по степени желательности (ранг 20 присуждается самому желаемому качеству). Затем в другой колонке его просят проранжировать эти же качества по степени выраженности у него в данный момент (ранг 20 получает самое характе́рное качество). На основе расчета коэффициента ранговой корреляции Спирмена делается вывод об уровне самооценки испытуемого. Результаты испытуемого *С. О-ва* приведены в таблице 3.2. Требуется рассчитать коэффициент корреляции Спирмена между выраженностью качеств у обследуемого испытуемого в данный момент и его идеальным представлением.

Решение. Составляем расчетную таблицу, в которую заносим две ранговые последовательности (желаемую N и реальную N'), разности рангов d и d^2.

Таблица 3.2

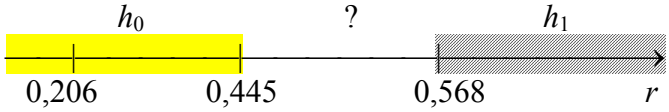
Расчет коэффициента ранговой корреляции Спирмена

Качества	*N*	*N'*	$d = N - N'$	d^2
уступчивость	14	15	-1	1
смелость	15	18	-3	9
вспыльчивость	2	16	-14	196
настойчивость	13	13	0	0
нервозность	1	7	-6	36
терпеливость	17	10	7	49
увлекаемость	12	20	-8	64
пассивность	8	2	6	36
холодность	10	19	-9	81
энтузиазм	9	17	-8	64
осторожность	16	4	12	144
капризность	3	1	2	4
медлительность	18	6	12	144
нерешительность	7	11	-4	16
энергичность	20	12	8	64
жизнерадостность	19	8	11	121
мнительность	4	3	1	1
упрямство	5	9	-4	16
беспечность	11	14	-3	9
застенчивость	6	5	1	1
				1056

Значение коэффициента корреляции Спирмена подсчитываем по формуле

$$r_S = 1 - \frac{6 \cdot 1056}{20^3 - 20} = 0{,}206 .$$

Вследствие малого n (меньше 30) гипотезу о значимости коэффициента корреляции проверяем с помощью статистических таблиц. Для $n = 20$ имеем (см. табл. 3.1):



Значение коэффициента корреляции $r_S = 0{,}206$ попадает в область допустимых значений, что не позволяет отвергнуть нулевую гипотезу. Коэффициент корреляции не отличается от нуля, что свидетельствует об отсутствии связи между выраженностью качеств у обследуемого испытуемого в данный момент и идеальным представлением.

§ 9. Коэффициент ранговой корреляции Кендалла

Коэффициент корреляции «тау» Кендалла имеет те же свойства, что и коэффициент Спирмена (изменяется от –1 до +1, для независимых случайных величин равен нулю), однако он считается более информативным.

Первым этапом расчета коэффициента «тау» Кендалла является ранжирование рядов переменных (одинаковым значениям ряда присваивают среднее ранговое число). Первая переменная должна быть упорядочена по возрастанию рангов.

Коэффициент корреляции Кендалла определяется по формуле

$$\tau = \frac{4 \cdot \sum_{i=1}^{n-1} R_i}{n(n-1)} - 1 .$$

Здесь n – объем выборки (число сопоставляемых пар); R_i – число рангов во втором вариационном ряду, бóльших, чем данное ранговое число и расположенных ниже него.

Проверка **гипотезы о значимости** коэффициента ранговой корреляции Кендалла заключается в сопоставлении вычисленного значения коэффициента «тау» по модулю с критическими значениями:

$$\tau_\alpha(n) = z_{1-\alpha/2} \sqrt{\frac{2(2n+5)}{9n(n+1)}} ,$$

где n – объем выборки, $z_{1-\alpha/2}$ – квантили нормированного нормального распределения ($z_{1-\alpha/2} = 1{,}960$ для $\alpha = 0{,}05$; $z_{1-\alpha/2} = 2{,}576$ для $\alpha = 0{,}01$).

Нулевая гипотеза $\tau = 0$ не отвергается, если значение коэффициента корреляции Кендалла (по модулю) попадает в область допустимых значений: $|\tau| \le \tau_{0{,}05}(n)$.

Корреляция считается значимой, если модуль коэффициента «тау» попадает в критическую область: $|\tau| > \tau_{0{,}01}(n)$.

Пример III.2. Требуется оценить корреляционную связь между скоростью чтения первоклассников и их усидчивостью. Скорость чтения первоклассников замерялась секундомером (слов/мин), усидчивость – с помощью экспертного оценивания по специально разработанной пятиточечной шкале: очень высокий (ОВ) – высокий (В) –

средний (С) – низкий (Н) – очень низкий (ОН) уровни. Учителю предъявлялись карточки с описанием уровня проявления усидчивости, которые он должен был соотнести с поведением каждого ученика. Результаты измерений приведены в таблице 3.3.

Решение. Скорость чтения первоклассников (переменная X) измерена в реляционной шкале, а их усидчивость (переменная Y) – в порядковой, поэтому определить связь между ними можно с помощью ранговой корреляции, проранжировав обе переменные. Решим задачу двумя способами, рассчитав коэффициенты ранговой корреляции Спирмена и Кендалла.

1. Расчет коэффициента корреляции Спирмена.

В таблицу с результатами (3.3) добавим четыре колонки: две для ранговых последовательностей первой и второй переменных, для разности рангов d и для d^2. Одинаковым значениям по второй переменной присваиваем средние ранговые значения.

Таблица 3.3

Расчет коэффициента ранговой корреляции Спирмена

Ф.И.	X	Y	R_x	R_y	$d = R_x - R_y$	d^2
1. Аня К.	12,0	Н	9	8,5	0,5	0,25
2. Боря Л.	18,8	ОН	6	10	-4	16
3. Вася Р.	11,0	В	10	2,5	7,5	56,25
4. Даша В.	29,0	С	2	5,5	-3,5	12,25
5. Зина С.	17,5	Н	7	8,5	-1,5	2,25
6. Игорь М.	23,4	ОВ	4	1	3	9
7. Катя Г.	35,6	С	1	5,5	-4,5	20,25
8. Лёня А.	15,4	С	8	5,5	2,5	6,25
9. Маша Д.	26,1	В	3	2,5	0,5	0,25
10. Яша Б.	20,7	С	5	5,5	-0,5	0,25
						123,00

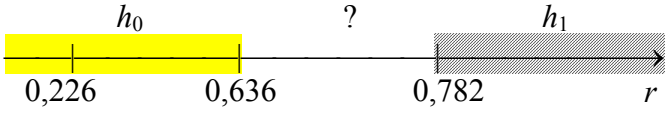
Коэффициент корреляции Спирмена рассчитывается по формуле с поправками на одинаковые значения. Вследствие того, что все значения X различны, значение поправки T_x равно нулю. По второй переменной наблюдаются три группы повторяющихся значений. В первой группе (Н) два повторения, во второй (С) – четыре, в третьей (В) – два. Значение поправки равно

$$T_y = (2^3 - 2) + (4^3 - 4) + (2^3 - 2) = 72.$$

Таким образом, коэффициент корреляции Спирмена равен

$$r_S = 1 - \frac{6 \cdot 123}{1000 - 10 - \frac{1}{2}(0 + 72)} = 0{,}226\,.$$

Критические значения находим в таблице 3.1 для десяти испытуемых:



Значение коэффициента корреляции Спирмена $r_S = 0{,}226$ попадает в область допустимых значений, что свидетельствует об отсутствии связи между скоростью чтения первоклассников и их усидчивостью.

2. Расчет коэффициента корреляции Кендалла.

Для расчета коэффициента корреляции Кендалла необходимо расчетную таблицу перегруппировать по возрастанию рангов первой переменной. Последний столбец (R) заполняется следующим образом: ниже Кати Г. ($R_y = 5{,}5$) имеется 3 ранга R_y, бóльших, чем у Кати (5,5); ниже Даши В. ($R_y = 5{,}5$) имеется 3 ранга R_y, бóльших, чем у Даши (5,5); ниже Маши Д. ($R_y = 2{,}5$) имеется 5 рангов R_y, бóльших, чем у Маши (2,5).

Таблица 3.4

Расчет коэффициента ранговой корреляции Кендалла

Ф.И.	*X*	*Y*	R_x	R_y	*R*
7. Катя Г.	35,6	С	1	5,5	3
4. Даша В.	29,0	С	2	5,5	3
9. Маша Д.	26,1	В	3	2,5	5
6. Игорь М.	23,4	ОВ	4	1	6
10. Яша Б.	20,7	С	5	5,5	3
2. Боря Л.	18,8	ОН	6	10	0
5. Зина С.	17,5	Н	7	8,5	0
8. Лёня А.	15,4	С	8	5,5	1
1. Аня К.	12,0	Н	9	8,5	0
3. Вася Р.	11,0	В	10	2,5	0
					21

Коэффициент корреляции Кендалла равен

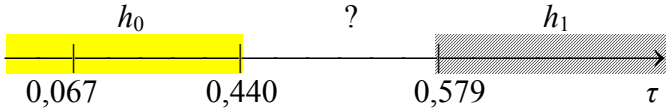
$$\tau = \frac{4 \cdot 21}{10 \cdot 9} - 1 = -0{,}067 .$$

Критические значения рассчитываем для $n = 10$ испытуемых:

$$\tau_{0,05} = 1{,}96 \cdot \sqrt{\frac{2 \cdot (2 \cdot 10 + 5)}{9 \cdot 10 \cdot 11}} = 0{,}440 ;$$

$$\tau_{0,01} = 2{,}576 \cdot \sqrt{\frac{2 \cdot (2 \cdot 10 + 5)}{9 \cdot 10 \cdot 11}} = 0{,}579 .$$

С критическими значениями сравнивается модуль коэффициента «тау»:



Значение модуля коэффициента «тау» Кендалла $|\tau| = 0{,}067$ попадает в область допустимых значений (как и в случае коэффициента корреляции Спирмена). Это еще раз подтверждает отсутствие связи между скоростью чтения исследуемых первоклассников и их усидчивостью.

§ 10. Коэффициент конкордации (согласованности) Кендалла

Коэффициент конкордации Кендалла используется в случае, когда совокупность объектов характеризуется несколькими последовательностями рангов, а исследователю необходимо установить статистическую связь между этими последовательностями. Такие задачи возникают, например, при анализе экспертных оценок: несколько экспертов ранжируют одних и тех же испытуемых по определенному качеству, а психологу для проведения углубленного анализа ситуации и принятия обоснованного решения требуется определить степень согласованности мнений группы экспертов.

Коэффициент конкордации Кендалла определяется по формуле

$$W = \frac{12 \cdot \sum_{i=1}^{n} D_i^2}{m^2 (n^3 - n)},$$

где n – число оцениваемых объектов (испытуемых), m – число ранговых последовательностей (количество экспертов), $D_i = d_i - \bar{d}$ – отклонение суммы рангов i-го объекта $d_i = \sum_{j=1}^{m} R_{ij}$ от средней суммы рангов всех объектов $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$. Средняя сумма рангов всех объектов может быть вычислена по формуле $\bar{d} = \frac{1}{2}m(n+1)$, которая используется для контроля.

Значения коэффициента конкордации, в отличие от коэффициента корреляции, заключены в интервале $0 \le W \le 1$. Коэффициент конкордации равен единице при полном совпадении всех ранговых последовательностей. Если мнения экспертов (ранговые последовательности) полностью противоположны, коэффициент конкордации равен нулю (коэффициент корреляции в этом случае будет равен –1).

При наличии одинаковых рангов у одного эксперта расчетная формула для коэффициента конкордации приобретает следующий вид

$$W = \frac{12 \cdot \sum_{i=1}^{n} D_i^2}{m^2(n^3 - n) - m\sum_{j=1}^{m} T_j}, \qquad T_j = \sum_{k=1}^{l} (t_k^3 - t_k).$$

В корректирующем члене для j-го эксперта через t_k обозначено число одинаковых значений в k-й группе (связке), l – число связок (групп с одинаковыми значениями) в ранговой последовательности j-го эксперта.

Проверка гипотезы об отсутствии связи. Проверка нулевой гипотезы h_0: $W = 0$ (мнения экспертов не согласуются друг с другом) при альтернативной h_1: $W \ne 0$ (мнения экспертов согласуются) при относительно большом количестве объектов $n \ge 7$ проводится с помощью критерия Пирсона «хи–квадрат». Эмпирическое значение

$$\chi^2 = m(n-1) \cdot W$$

сравнивается с критическими $\chi^2_\alpha(n-1)$, вычисленными для числа степеней свободы $df = n - 1$ и соответствующих уровней значимости α. Коэффициент конкордации значимо отличается от нуля, если эмпирическое значение попадает в критическую область: $\chi^2 > \chi^2_{0,01}(n-1)$.

Значимость коэффициента конкордации при малом количестве объектов n проверить сложно.

Пример III.3. Экспертная комиссия из 5 человек проранжировала 7 сочинений школьников – участников олимпиады по психологии (ранг 1 присваивался самой лучшей работе). Ранговые последовательности приведены в таблице 3.5. Требуется вычислить коэффициент конкордации.

Решение. В расчетную таблицу 3.5 заносим экспертные оценки, ранговые суммы d_i, отклонения D_i суммы рангов от средней $\bar{d}$ и D_i^2.

Таблица 3.5

Расчет коэффициента конкордации

Школьники (n)	*Эксперты (m)*					$d_i = \sum_{j=1}^{m} R_{ij}$	$D_i = d_i - \bar{d}$	D_i^2
	1	*2*	*3*	*4*	*5*			
1	1	1	2	1	3	8	-12	144
2	3	2	1	2	1	9	-11	121
3	4	5	7	4	5	25	5	25
4	2	3	5	6	4	20	0	0
5	6	6	6	3	2	23	3	9
6	7	4	4	5	6	26	6	36
7	5	7	3	7	7	29	9	81
						140		**416**

Средняя сумма рангов всех объектов равна $\bar{d} = \frac{140}{7} = 20{,}0$.

В качестве контроля используем выражение $\bar{d} = \frac{1}{2} m(n+1) = \frac{5 \cdot 8}{2} = 20$.

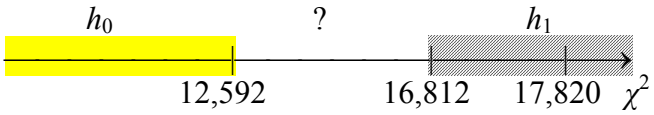
Коэффициент конкордации Кендалла определяется по формуле

$$W = \frac{12 \cdot 416}{5^2 (7^3 - 7)} = 0{,}594 .$$

Проверку нулевой гипотезы о том, что мнения экспертов не согласуются друг с другом (h_0: $W = 0$) проводим с помощью критерия Пирсона «хи–квадрат». Для этого вычисляем эмпирическое значение

$$\chi^2 = 5 \cdot 6 \cdot 0{,}594 = 17{,}820 ,$$

которое сравниваем с критическими значениями критерия «хи-квадрат» для числа степеней свободы $df = n - 1 = 6$:



Эмпирическое значение $\chi^2 = 17{,}820$ попадает в критическую область, что позволяет отвергнуть нулевую гипотезу. Коэффициент конкордации значимо отличается от нуля ($p < 0{,}01$), следовательно имеется достаточно тесная согласованность мнений экспертов относительно оцениваемых сочинений.

IV. БИСЕРИАЛЬНАЯ КОРРЕЛЯЦИЯ

Бисериальная[5] *корреляция* является методом корреляционного анализа между двумя переменными, одна из которых измерена в **дихотомической шкале**. Подобные задачи часто встречаются в психодиагностике.

Решение задач данного класса осуществляется с помощью бисериальных коэффициентов корреляции: *точечного бисериального коэффициента корреляции Пирсона* (если вторая переменная измерена в сильной шкале) и *рангово-бисериального коэффициента корреляции* (если вторая переменная измерена в порядковой шкале).

Дихотомические переменные, принимающие два значения (мужчина–женщина, верный ответ – неверный ответ и т.п.), можно обозначать любыми двумя знаками. Например, мужскую и женскую части популяции можно маркировать буквами (*М* и *Ж*), цифрами (0 и 1) или символами (♂ и ♀). Далее мы будем пользоваться цифровыми обозначениями.

§ 11. Точечный бисериальный коэффициент корреляции

Пусть переменная X измерена в сильной шкале, а переменная Y – в дихотомической. Точечный бисериальный коэффициент корреляции r_{pb} вычисляется по формуле[6]

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \cdot \sqrt{\frac{n_1 n_0}{n(n-1)}}.$$

Здесь $\bar{x}_1$ – среднее значение по X объектов со значением «единица» по Y;

$\bar{x}_0$ – среднее значение по X объектов со значением «ноль» по Y;

s_x – среднее квадратическое отклонение всех значений по X;

n_1 – число объектов «единица» по Y, n_0 – число объектов «ноль» по Y;

$n = n_1 + n_0$ – объем выборки.

Точечный бисериальный коэффициент корреляции можно рассчитать также с помощью других эквивалентных выражений:

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}}{s_x} \cdot \sqrt{\frac{n_1 n}{n_0(n-1)}}; \qquad r_{pb} = \frac{\bar{x} - \bar{x}_0}{s_x} \cdot \sqrt{\frac{n_0 n}{n_1(n-1)}}.$$

Здесь $\bar{x}$ – общее среднее значение по переменной X.

[5] Бисериальный от лат. *bis series* – два ряда, две серии.

[6] Обозначение «r_{pb}» от англ. *point bis series* – точечный бисериальный.

Точечный бисериальный коэффициент корреляции r_{pb} изменяется в пределах от –1 до +1. Его значение равно нулю в том случае, если переменные с единицей по Y имеют среднее по Y, равное среднему переменных с нулем по Y.

Проверка **гипотезы о значимости** точечного бисериального коэффициента корреляции заключается в проверке нулевой гипотезы h_0 о равенстве генерального коэффициента корреляции нулю: $\rho = 0$, которая осуществляется с помощью критерия Стьюдента. Эмпирическое значение

$$| t | = \frac{| r_{pb} |}{\sqrt{1 - r_{pb}^2}} \cdot \sqrt{n - 2}$$

сравнивается с критическими значениями $t_\alpha(df)$ для числа степеней свободы $df = n - 2$.

Если выполняется условие $| t | \le t_\alpha(df)$, нулевая гипотеза $\rho = 0$ не отвергается. Точечный биссериальный коэффициент корреляции значимо отличается от нуля, если эмпирическое значение $| t |$ попадает в критическую область, то есть если выполняется условие $| t | > t_\alpha(n - 2)$.

Достоверность связи, рассчитанной с помощью точечного бисериального коэффициента корреляции r_{pb}, можно определить также с помощью критерия χ^2 для числа степеней свободы $df = 2$.

Пример IV.1. У студентов железнодорожного института был измерен уровень потребности в достижении с помощью тест-опросника Ю.М. Орлова (результаты в T-баллах приведены в таблице 4.1). Требуется рассчитать бисериальную корреляцию между уровнем развития потребности в достижении и успеваемостью студентов.

Решение. Составляем расчетную таблицу, в которую заносим показатели потребности в достижении (X) и успеваемость (Y). В две последние колонки записываем результат разбиения выборки на две подвыборки по дихотомической переменной.

Таблица 4.1

Расчет точечного бисериального коэффициента корреляции

x_i, T	Y: *успеваемость*	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$Y = 1$	$Y = 0$
57,4	высокая (1)	5,16	26,5916	57,4	
61,9	высокая (1)	9,60	92,2287	61,9	
25,3	низкая (0)	-26,98	727,7756		25,3
79,6	высокая (1)	27,33	746,6679	79,6	
24,3	низкая (0)	-27,98	782,6510		24,3
45,2	низкая (0)	-7,07	50,0524		45,2
43,1	низкая (0)	-9,17	84,0462		43,1
73,4	высокая (1)	21,14	447,0984	73,4	
52,5	низкая (0)	0,23	0,0507		52,5
38,4	низкая (0)	-13,89	192,8180		38,4
66,5	высокая (1)	14,23	202,3570	66,5	
39,6	высокая (1)	-12,67	160,6499	39,6	
55,4	высокая (1)	3,13	9,7670	55,4	
49,4	низкая (0)	-2,87	8,2643		49,4
72,1	высокая (1)	19,83	393,0395	72,1	
784,1			**3924,0584**	**505,9**	**278,2**

По условию задачи $n_1 = 8$, $n_0 = 7$. Объем выборки $n = 15$, $df = 14$.

Находим средние значения по переменной X и среднее квадратическое отклонение s_x:

$$\bar{x} = \frac{784{,}1}{15} = 52{,}27; \qquad \bar{x}_1 = \frac{505{,}9}{8} = 63{,}24; \qquad \bar{x}_0 = \frac{278{,}2}{7} = 39{,}74;$$

$$s_x = c_n \cdot \sqrt{\frac{SS}{df}} = 1{,}018 \cdot \sqrt{\frac{3924{,}0584}{14}} = 17{,}043.$$

Значение точечного бисериального коэффициента корреляции можно вычислить по любой из трех формул:

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \cdot \sqrt{\frac{n_1 n_0}{n(n-1)}} = \frac{63{,}24 - 39{,}74}{17{,}043} \sqrt{\frac{8 \cdot 7}{15 \cdot 14}} = 0{,}712;$$

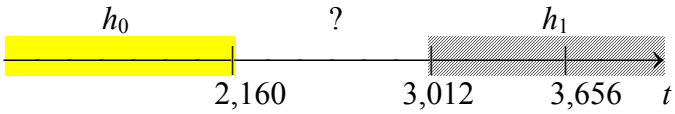
$$r_{pb} = \frac{\bar{x}_1 - \bar{x}}{s_x} \cdot \sqrt{\frac{n_1 n}{n_0 (n-1)}} = \frac{63{,}24 - 52{,}27}{17{,}043} \sqrt{\frac{8 \cdot 15}{7 \cdot 14}} = 0{,}712;$$

$$r_{pb} = \frac{\bar{x} - \bar{x}_0}{s_x} \cdot \sqrt{\frac{n_0 n}{n_1 (n-1)}} = \frac{52{,}27 - 39{,}74}{17{,}043} \sqrt{\frac{7 \cdot 15}{8 \cdot 14}} = 0{,}712.$$

Гипотезу о значимости точечного бисериального коэффициента корреляции проверяем с помощью критерия Стьюдента. Эмпирическое значение равно

$$|t| = \frac{0{,}712}{\sqrt{1 - 0{,}712^2}} \cdot \sqrt{13} = 3{,}656.$$

Критические значения критерия Стьюдента $t_\alpha(df)$ находим в статистических таблицах для числа степеней свободы $df = 13$:



Эмпирическое значение $|t| = 3{,}656$ попадает в критическую область, что позволяет отвергнуть нулевую гипотезу $\rho = 0$. Коэффициент корреляции значимо отличается от нуля ($p < 0{,}01$), следовательно имеется средняя[7] (умеренная) связь между уровнем развития потребности в достижении и успеваемостью исследуемых студентов.

§ 12. Рангово-бисериальный коэффициент корреляции

Рангово-бисериальный коэффициент корреляции, используемый в случаях, когда одна из переменных (X) представлена в порядковой шкале, а другая (Y) – в дихотомической, вычисляется по формуле

$$\boxed{r_{rb} = \frac{2}{n}\left(\bar{X}_1 - \bar{X}_0\right)}.$$

Здесь $\bar{X}_1$ – средний ранг объектов, имеющих единицу по Y; $\bar{X}_0$ – средний ранг объектов с нулем по Y, n – объем выборки.

Проверка **гипотезы о значимости** рангово-бисериального коэффициента корреляции осуществляется аналогично точечному биссериальному коэффициенту корреляции с помощью критерия Стьюдента с заменой в формулах r_{pb} на r_{rb}.

[7] Теснота связи определяется абсолютным значением коэффициента корреляции (см. § 4).

V. СОПРЯЖЕННОСТЬ

Стохастическая связь между качественными переменными **номинативной шкалы** называется *сопряженностью*. При корреляционном анализе дихотомических переменных используется *коэффициент контингенции Пирсона* (*φ-коэффициент*), при исследовании степени тесноты связи между качественными номинативными (но не дихотомическими) признаками – *коэффициенты сопряженности*.

§ 13. Коэффициент контингенции Пирсона (*φ*-коэффициент)

Дихотомическая корреляция используется при исследовании степени тесноты между переменными *x* и *y*, каждый из которых представлен в виде двух альтернатив («1» – «0»). Расчетная таблица *φ-коэффициента Пирсона* состоит из четырех ячеек:

Таблица 5.1

Тетрахорическая таблица

Переменные		***X***	
		«1»	«0»
Y	«1»	*a*	*b*
	«0»	*c*	*d*

Частоты *a*, *b*, *c*, *d* называются *тетрахорическими показателями*, их сумма равна объему выборки:

$$n = a + b + c + d.$$

Каждая из клеток соответствует частоте выбора определенной альтернативы того и другого признака. Например, частота *b* определяет количество случаев, в которых зарегистрировано «0» по переменной *X* и «1» по переменной *Y*. *φ*-коэффициент Пирсона определяется по формуле

$$\boxed{\varphi = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}}.$$

Его значения изменяются от –1 до +1. При $\varphi = 0$ признаки независимы. $\varphi = 1$ свидетельствует о положительной зависимости (всем X = «1» соответствует Y = «1»), при $\varphi = -1$ связь отрицательная.

Проверка **гипотезы о значимости связи** между исследуемыми переменными осуществляется с помощью критерия χ^2. Для этого эмпирическое значение

$$\chi^2 = \varphi^2 \cdot n,$$

где *n* – объем выборки, сравнивается с критическим $\chi_\alpha^2(1)$ для числа степе-

ней свободы $df = 1$ ($\chi_\alpha^2 = 3{,}841$ для $\alpha = 0{,}05$; $\chi_\alpha^2 = 6{,}635$ для $\alpha = 0{,}01$).

Если выполняется условие $\chi^2 \le \chi^2_{0,05}(1)$, нулевая гипотеза $\varphi = 0$ не отвергается. При $\chi^2 > \chi^2_{0,01}(1)$ нулевая гипотеза $\varphi = 0$ отвергается, и связь считается значимой.

Пример V.1. Изучается общественное мнение по очень важному поводу. Распределение ответов респондентов, мужчин и женщин, приведено в таблице 5.2. Требуется определить наличие связи между полом и определенным мнением.

Таблица 5.2

Тетрахорическая таблица

Переменные		***Мнение***	
		положительное	отрицательное
Пол	мужчины	59	41
	женщины	36	64

Решение. Нулевой гипотезой является предположение об отсутствии связи между рассматриваемыми переменными. Для ее проверки определяем значение φ-коэффициента Пирсона и критическое значение «хи-квадрат»:

$$\varphi = \frac{59 \cdot 64 - 36 \cdot 41}{\sqrt{(59+41)(41+64)(59+36)(36+64)}} = 0{,}23;$$

$$\chi^2 = 0{,}23^2 \cdot 200 = 10{,}580 > 6{,}635.$$

Вследствие того, что эмпирическое значение попало в критическую область ($\chi^2 > \chi^2_{0,01}(1)$), нулевая гипотеза $\varphi = 0$ отвергается, связь между полом и определенным мнением можно считать значимой ($p < 0{,}01$).

Учебное издание

Харченко Максим Андреевич

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Учебное пособие для вузов

Редактор Л.М. Носилова

Подписано в печать 27.06.08. Формат 60×84/16. Усл. печ. л. 1,9.
Тираж 100 экз. Заказ 1287.

Издательско-полиграфический центр
Воронежского государственного университета.
394000, г. Воронеж, пл. им. Ленина, 10. Тел. 208-298, 598-026 (факс)
http://www.ppc.vsu.ru; e-mail: pp_ctnter@ppc.vsu.ru

Отпечатано в типографии Издательско-полиграфического центра
Воронежского государственного университета.
394000, г. Воронеж, ул. Пушкинская, 3. Тел. 204-133.