

СОЦИОЛОГИЯ

УДК 303.02

ОСОБЕННОСТИ ПРИМЕНЕНИЯ КОРРЕЛЯЦИОННОГО АНАЛИЗА В СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

В.И. Гостенина, К.С. Карандин

Брянский государственный университет имени академика И.Г. Петровского

В статье рассматривается понятие корреляционного анализа, его особенности и специфика применения при изучении социологическом изучении социальных проблем управления.

На обширной теоретической базе смоделированы разнообразные виды коэффициентов ранговой корреляции. На простых примерах показан алгоритм расчёта наиболее часто применяемых в социологии коэффициентов Спирмена и Пирсона. Выделены их достоинства и недостатки.

Приведен реальный пример использования коэффициента Спирмена при анализе данных, полученных членами научной лаборатории кафедры социологии и социальной работы Брянского государственного университета в ходе мониторинга информационного бюллетеня «Вестник. Труд и занятость» управления государственной службы по труду и занятости населения Брянской области.

Ключевые слова: корреляция, анализ, шкала Чертока, корреляционное поле, коэффициент Спирмена, коэффициент Юла, коэффициент Пирсона, множественный коэффициент, коэффициент Чупрова, коэффициент Крамера, мониторинг, социальное управление.

1. Теоретическое обоснование использования корреляционных расчетов в социологическом анализе

Разработка методики исследования осуществляется, исходя из логики научного поиска. В целом методика представляет комплекс теоретических и эмпирических методов, сочетая которые с наибольшей достоверностью прогнозируются социальные явления в области социального управления. Представленная авторская методика позволяет проводить исследования сложных и многомерных объектов.

Применяя инструменты математического анализа дается численная оценка явлений, независимо от их происхождения, социальной значимости и сложности принятия управленческого процесса [1,2].

В эмпирической социологии, пожалуй, самым распространенным видом анализа информации является – корреляционный анализ (от лат. *correlatio* «соотношение или взаимосвязь»). Корреляция или корреляционная зависимость – это некая зависимость двух или более случайных величин. Суть корреляционного анализа сводится к поиску связи между этими величинами. Особенность анализа заключается в том, что наличие связи характеризуется тем, насколько ярко она выражена. Показателем, отражающим тесноту двух коррелирующих величин, является определенный критерий, получивший название – коэффициент корреляции.

Коэффициент корреляции может принимать значения от +1 до -1, чем ближе значение к единице, тем более сильной считается связь между величинами. Нулевое или близкое к нему значение свидетельствует об отсутствии связи. Существует и более тонкая градация силы связи, она представлена шкалой Чертока (Таблица 1).

Корреляционный анализ широко применяется в гуманитарных науках [4]. В социологии, как правило, используется коэффициент ранговой корреляции Спирмена, коэффициент Юла, коэффициент Пирсона, множественный коэффициент корреляции, коэффициент Чупрова и Крамера.

Таблица 1 - Шкала Чертока

№	Коэффициент корреляции	Характеристика силы связи
1	$ r_s < 0,1$	Связь практически отсутствует
2	$0,1 < r_s < 0,3$	Слабая связь
3	$0,3 < r_s < 0,5$	Умеренная связь
4	$0,5 < r_s < 0,7$	Связь средней силы
5	$0,7 < r_s < 0,9$	Сильная связь
6	$0,9 < r_s < 1$	Очень сильная связь

Применение корреляционного анализа в социологии обусловлено определенными причинами. При изучении социальных проблем исследуемые вопросы могут включать большое количество влияющих факторов. Например, люди, имеющие высшее образование имеют более высокую зарплату, хотя среди них имеются люди с разной зарплатой. Корреляционно связаны, например уровень образования родителей и успеваемость детей, уровень заработной платы и удовлетворенность работой, квалификация сотрудника и производительность труда. Наличие корреляции может свидетельствовать о том, что одно из выделенных явлений-причина другого или же эти явления являются следствиями общих причин.

В социологии зачастую определенному значению одной переменной, может соответствовать ряд значений другой, то есть каждому x соответствует целый ряд значений y . Если изобразить на плоскости каждую пару значений этих величин, то наносимые точки расположатся в виде «облака». Это «облако» называется корреляционным полем [5]. Корреляционное поле является вспомогательным средством анализа выборочных данных, а расположение точек позволяет нам сделать предварительный вывод о форме и характере зависимости.

Однако стоит помнить, что корреляционная зависимость не может быть абсолютно полной и точной. Выявление корреляции не дает оснований нам утверждать о причинно – следственных связях явлений. Например, В. Шубкин обнаружил связь между зарплатой родителей и успеваемостью их детей. Ричардсон выявил связь между зарплатой учителей и потреблением вин в США. В первом примере мы можем объяснить корреляцию тем, что более образованные люди получают более высокую заработную плату. Во втором примере – тем, что в период с 1870 по 1910 года в Америке в целом наблюдался рост заработных плат (в том числе и учителей) что и вызвало увеличение потребления продуктов питания (в том числе и вина). То есть заработная плата учителей и потребление вина – это следствия общей причины, а не причина и следствия по отношению друг другу.

Любая социологическая проблема - это явление, зависящее от большого числа причин, которые действуют с разной силой. Когда мы занимаемся изучением влияния одной переменной на другую, то как правило выделяем один фактор, но на зависимую величину оказывают влияние и многие другие, чем и обуславливается характер корреляционной зависимости.

Еще одной причиной удобства использования корреляционного анализа в социологических исследованиях является его «универсальность». Мы можем применять его для переменных, которые относятся к различным шкалам. От вида шкалы зависит метод расчета коэффициента корреляции. Однако при всей простоте анализа не стоит забывать о корректной интерпретации полученных результатов. Зачастую провести корреляционный анализ не вызывает труда, проблемы возникают именно на этапе интерпретации.

2. Примеры расчета корреляционного анализа

На вымышленном примере разберем работу корреляционного анализа. Результаты Единого Государственного Экзамена по математике и физике (Таблица 2). Попробуем определить зависимость между результатами. Для определения корреляционной связи используем коэффициент Спирмена (r_s). Его формула представлена ниже.

$$r_s = 1 - \frac{6 \sum_i d_i^2}{(n-1)n(n+1)} \quad (1)$$

Таблица 2 - Результаты Единого Государственного Экзамена по математике и физике

Ученик	1	2	3	4	5	6	7	8	9	10
Результаты по математике	95	93	84	79	72	70	62	60	57	45
Результаты по физике	91	96	87	83	53	60	76	58	77	65

Технология расчета:

1. Проранжируем данные в порядке их убывания и найдем квадраты разности соответствующих рангов.
2. Для удобства эти действия лучше производить в таблице (Таблица 3).

Таблица 3 - Таблица рангов результатов ЕГЭ по математике и физике

Математика	1	2	3	4	5	6	7	8	9	10
Физика	2	1	3	4	10	8	6	9	5	7
d_i^2	1	1	0	0	25	4	1	1	16	9

3. Объем выборки (n) равен 10. Следовательно:

$$r_s = 1 - \frac{6 \sum_i d_i^2}{(n-1)n(n+1)} = 1 - \frac{6 \cdot (1+1+0+0+25+4+1+1+16+9)}{9 \cdot 10 \cdot 11} \approx 0,648 \quad (2)$$

Так как $r_s = 0,648$, мы можем сделать вывод о том, что имеет место прямая связь средней силы.

Применение коэффициента корреляции Спирмена удобно из-за его простоты расчета.

Однако следует помнить, что этот расчет может применяться не во всех случаях. Эффективность снижается, если достаточно велика разница между различными значениями кокой-либо из измеряемых величин. Так же не рекомендуется применять коэффициент ранговой корреляции Спирмена, если мы имеем дело с неравномерным распределением значений измеряемой величины.

В представленном выше примере целесообразно использовать именно коэффициент Спирмена, так как значения измеряемой величины изменяются без скачков, относительно равномерно.

Рассмотрим ещё один пример по применению коэффициента корреляции. Студентам было предложено пройти два простых теста. Было измерено среднее время решения заданий теста (в секундах). Попробуем ответить на вопрос: Есть ли взаимосвязь между временем

решения этих тестов? Обозначим среднее время решения теста №1 – переменной x , а среднее время решения теста №2 – y .

На этот раз используем формулу для расчёта коэффициента корреляции Пирсона (Представлена ниже). x_i – это значения, которые принимает переменная x , y_i – значения, принимаемые переменной y , \bar{x} , \bar{y} – средние по x и y соответственно.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 * \sum(y_i - \bar{y})^2}} \quad (3)$$

Исходная формула несколько неудобна, поэтому мы будем применять ее аналог, полученный с помощью преобразований. Представим исходные данные в таблице, с дополнительными столбцами, необходимыми для расчета по формуле (Таблица 4).

Таблица 4 - Исходные данные с дополнительными столбцами, необходимыми для расчета коэффициента корреляции Пирсона

№	x_i	y_i	$x_i * y_i$	x_i^2	y_i^2
1	20	18	360	400	324
2	31	6	186	961	36
3	34	18	612	1156	324
4	42	26	1092	1764	676
5	30	29	870	900	841
6	36	32	1152	1296	1024
7	38	19	722	1444	361
8	40	18	720	1600	324
9	45	35	1575	2025	1225
10	45	44	1980	2025	1936
11	25	11	275	625	121
12	36	27	972	1296	729
13	30	14	420	900	196
14	41	44	1804	1681	1936
15	43	46	1978	1849	2116
16	31	23	713	961	529
17	49	46	2254	2401	2116
18	40	24	960	1600	576
19	34	17	578	1156	289
20	48	27	1296	2304	729
Сумма	738	524	20519	28344	16408

Эмпирическую величину коэффициента корреляции рассчитываем по формуле:

$$r_{x \text{ эмп}} = \frac{20 * 20519 - 738 * 524}{\sqrt{(20 * 28344 - 738 * 738) * (20 * 16408 - 524 * 524)}} \approx 0,686 \quad (4)$$

Далее воспользуемся таблицей критических значений корреляции Пирсона и определим критические значения для полученного коэффициента. Так как в нашем случае $k = n - 2 = 20 - 18$, то $r_{кр}$:

$$0,44 \text{ для } p \leq 0,05$$

$$0,56 \text{ для } p \leq 0,01$$

Далее строим ось значимости (Рисунок 1).

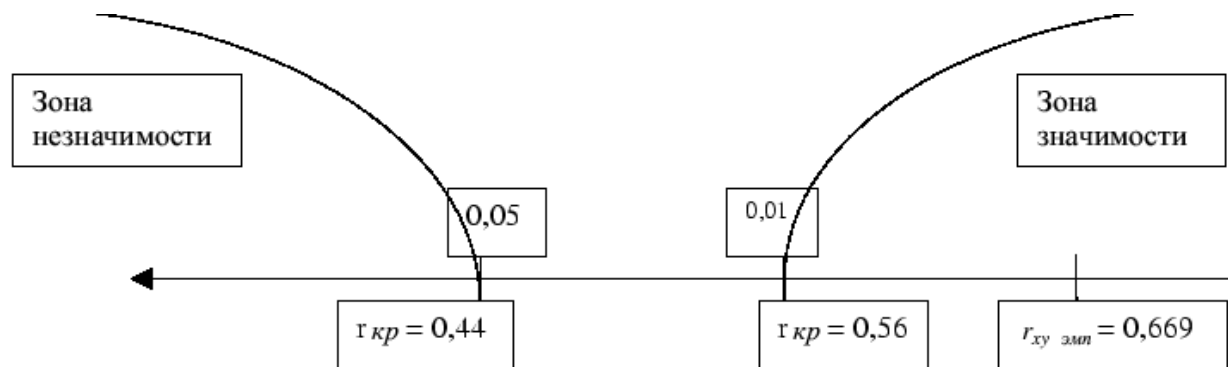


Рис.1. - Ось значимости

Так как величина корреляции попадает в зону значимости, H_0 - опровергается, принимается гипотеза H_1 . То есть, чем выше среднее время решения задач теста №1, тем выше среднее время решения задач теста №2 и наоборот.

Для применения коэффициента корреляции Пирсона, необходимо, чтобы соблюдался ряд условий.

Во-первых переменные которые подвергаются сравнению должны быть получены в интервальной шкале или шкале отношений.

Во-вторых распределения переменных x и y должны быть близки к нормальному.

В-третьих число варьирующих признаков в сравниваемых переменных x и y должно быть одинаковым.

3. Применение корреляционного анализа на практике

Для иллюстрации применения корреляционного анализа, за основу были взяты данные полученные в ходе мониторинга информационного бюллетеня «Вестник. Труд и занятость» управления государственной службы по труду и занятости населения Брянской области.

Цель мониторинга заключалась в создании информационных условий для формирования целостного представления о информационном бюллетене и его роли в информировании специалистов о состоянии рынка труда и занятости населения Брянской области.

Задачи, решаемые в ходе мониторинга, сводились к следующему:

а) Выработать комплекс показателей, обеспечивающих целостное представление о состоянии информационного бюллетеня и реализации его функций.

б) Систематизировать информации о состоянии и развитии информационного бюллетеня.

в) Обеспечить наглядное представление информации отражаемой в информационном бюллетене.

г) Обеспечить анализ и прогнозирование состояния и развития информационного бюллетеня, выработать управленческие рекомендации.

Всего было проанализировано 11 выпусков информационного бюллетеня (2015, 2016 годы).

Общее число публикаций: 70. Из них 15 публикаций касаются профессионально ориентационных материалов, 15 – посвящены проблемам охраны труда, 17 - социально трудовых отношений, 14 – по вопросам защиты от безработицы, 9 публикаций представляют информацию о положении на рынке труда.



Рис. 2. - Направление проблемы, освещаемой в бюллетене (Критерии)

Семь публикаций носят аналитический характер, 64 - информационный характер. Сорок публикаций являются актуальными для молодежи и школьников, 17 - для социальной незащищённых граждан, 27 - для безработных граждан, 4 - для пенсионеров, 28 - для работодателей, 17 - представляют интерес для работающего населения, 2 - для администрации университетов, 13 - для муниципальных органов власти, 10 - для профсоюзных организаций, 2 - для мигрантов, 2 - для женщин, находящихся в отпуске по уходу за ребенком.

Мониторинг позволил получить обширные данные по направленности публикационной активности для различных сегментов целевой аудитории. Это и натолкнуло на мысль о проведении корреляционного анализа.

Управлением государственной службы по труду и занятости населения Брянской области были предоставлены данные о содействии в поиске подходящей работы за 2015 год.

Проверке подверглась зависимость между количеством публикаций в информационном бюллетене (x) и количеством трудоустроенных (y). Все данные были объединены в таблицу (Таблица 5). Проверка осуществлялась с помощью коэффициента Спирмена (r_s).

Таблица 5 - Исходные данные для расчета коэффициента корреляции

	Молодежь и школьники	Социально незащищенные	Безработные граждане	Пенсионеры	Женщины, находящиеся в отпуске по уходу за ребенком
x	40	17	27	4	2
y	8 512	799	15 576	964	1 066

Таблица рангов для расчёта коэффициента корреляции представлена в таблице 6.

Таблица 6 - Таблица рангов

	Молодежь и школьники	Социально незащищенные	Безработные граждане	Пенсионеры	Женщины, находящиеся в отпуске по уходу за ребенком
х	5	3	4	2	1
у	4	1	5	2	3

Для удобства все расчёты осуществлялись в программе Microsoft Excel. Так как $r_s = 0,5$, мы можем сделать вывод о том, что связь между признаком у и фактором х умеренная и прямая.

Полученный результат можно считать удовлетворительным, мы можем говорить о наличии статистической взаимосвязи между переменными, однако проблема заключается в интерпретации значений этой взаимосвязи.

Таким образом корреляционный анализ – является удобным инструментом, позволяющим значительно расширить имеющиеся данные. Он способен сделать теоретическое мышление более эффективным. Однако применять его необходимо осмысленно. При всех явных плюсах корреляционный анализ не способен заменить специальные знания. Основным минусом методики является сложность интерпретации и поиск причинно-следственных связей.

Список литературы

1. Смердова С.Г. Применение статистических методов в испытаниях полимер-модифицированных битумов / С.Г. Смердова, Н.Н. Умарова, Л.В. Петухова, Г.И. Сибгатуллина // Вестник Казанского технологического университета. – 2010. – №5. – С. 259-266.
2. Абдулин И.А. Применение статистических методов для оценки технологического процесса производства кумулятивных зарядов / И.А. Абдуллин, О.И. Белобородова, Н.И. Лаптев, Е.Л. Москвичева, А.Д. Горяйнов // Вестник Казанского технологического университета. – 2010. – №12. – С. 477-482.
3. Ершов А.Н. Социология в современном мире / А.Н. Ершов А.Н., А.П. Кулапин // Вестник Казанского технологического университета. – 2012. – №2. – С. 118-121.
4. Андреева В.Р. Корреляционный анализ в социологических исследованиях / М. М. Андреева, В. Р. Волков // Вестник Казанского технологического университета. – 2013. – №7. – С. 271-274.
5. Гржибовский А.М. Анализ трех и более независимых групп количественных данных / А. М. Гржибовский // Экология человека. – 2008. – №3. – С. 50-58.
6. Гржибовский А.М. Корреляционный анализ данных с использованием программного обеспечения Statistica и SPSS / А. М. Гржибовский, С. В Иванов, М. А Горбатова // Наука и здравоохранение. – 2017. – №1. – С. 7-36.
7. Буров А.В. Корреляционный анализ взаимосвязей между индивидуально-психологическими особенностями детей и показателями адаптации к школьному обучению / А. В. Буров // Концепт. – 2014. – №7. – С. 101-105.
8. Bonett D. Sample size requirements for estimating Pearson, Kendall and Spearman correlations / D. Bonett, T. Wright // Psychometrica. – 2000. – Vol. 65. – P. 23–28.
9. Kraemer H. Correlation coefficients in medical research: from product moment correlation to the odds ratio / H. Kraemer // Statistical Methods in Medical Research. – 2006. – №6. – P. 525–545.

Сведения об авторе

В.И. Гостенина - доктор социологических наук, профессор, заведующий кафедрой социологии и социальной работы Брянского государственного университета имени академика И. г. Петровского, E-mail KafedraRandPR@yandex.ru

К.С. Карандин - магистр 1 курса направления подготовки «Социология управления» ФГБОУ ВО «Брянский государственный университет имени академика И.Г. Петровского», E-mail KafedraRandPR@yandex.ru

UDK 303.02

PECULIARITIES OF THE USE OF CORRELATION ANALYSIS IN SOCIOLOGICAL RESEARCH

V.I. Gostenina, K.S. Karandin

Bryansk state University named after academician I. G. Petrovsky

In the article, special attention is paid to the concept of correlation analysis, its features and specificity are examined in the study of sociological problems. The types of coefficients of rank correlation are disassembled.

On simple examples, the algorithm for calculating the Spearman and Pearson coefficients most frequently used in sociologists is analyzed. Their pluses and minuses are singled out.

A real example is given of using the Spearman coefficient, when analyzing the information received during the monitoring of the Bulletin. Labor and Employment "management of the state service for labor and employment of the population of the Bryansk region.

Keywords: correlation, analysis, Chertok scale, correlation field, Spearman coefficient, Yule coefficient, Pearson coefficient, multiple coefficient, Chuprov coefficient, Cramer coefficient, monitoring.

References

1. Smerdova S.G. *Primenenie statisticheskikh metodov v ispytaniyakh polimer-modifitsirovannykh bitumov* / S.G. Smerdova, N.N. Umarova, L.V. Petuhova, G.I. Sibgatullina // *Vestnik Kazanskogo tehnologicheskogo universiteta*. – 2010. – № 5. – S. 259-266.
2. Abdulin I.A. *Primenenie statisticheskikh metodov dlya otsenki tehnologicheskogo protsessa proizvodstva kumulyativnykh zaryadov* / I.A. Abdullin, O.I. Beloborodova, N.I. Laptev, E.L. Moskvicheva, A.D. Goryaynov // *Vestnik Kazanskogo tehnologicheskogo universiteta*. – 2010. – № 12. – S. 477-482.
3. Ershov A.N. *Sotsiologiya v sovremennom mire* / A.N. Ershov A.N., A.P. Kulapin // *Vestnik Kazanskogo tehnologicheskogo universiteta*. – 2012. – №2. – S. 118-121.
4. Andreeva V.R. *Korrelyatsionnyy analiz v sotsiologicheskikh issledovaniyakh* / M. M. Andreeva, V. R. Volkov // *Vestnik Kazanskogo tehnologicheskogo universiteta*. – 2013. – № 7. – S. 271-274.
5. Grzhibovskiy A.M. *Analiz trekh i bolee nezavisimyykh grupp kolichestvennykh dannykh* / A. M. Grzhibovskiy // *Ekologiya cheloveka*. – 2008. – № 3. – S. 50-58.
6. Grzhibovskiy A.M. *Korrelyatsionnyy analiz dannykh s ispolzovaniem programmnoho obespecheniya Statistica i SPSS* / A. M. Grzhibovskiy, S. V Ivanov, M. A Gorbatoва // *Nauka i zdravoohranenie*. – 2017. – № 1. – S. 7-36.
7. Burov A.V. *Korrelyatsionnyy analiz vzaimosvyazey mezhdru individualno-psihologicheskimi osobennostyami detey i pokazatelyami adaptatsii k shkolnomu obucheniyu* / A. V. Burov // *Kontsept*. – 2014. – № 7. – S. 101-105.
8. Bonett D. *Sample size requirements for estimating Pearson, Kendall and Spearman correlations* / D. Bonett, T. Wright // *Psychometrika*. — 2000. — Vol. 65. — P. 23—28.

9. Kraemer H. Correlation coefficients in medical research: from product moment correlation to the odds ratio / H. Kraemer // *Statistical Methods in Medical Research*.

Author`s information

Gostenina V.I. - Doctor of Sociology, Professor, Head of the Department of Sociology and Social Work of the Bryansk State University named after Academician I.G. Petrovsky.

K. S. Karandin - master student 1 course of the specialty «Sociology of management» E-mail KafedraRandPR@yandex.ru