УДК 519.8:681.3

АНАЛИЗ ПОРЯДКОВЫХ ДАННЫХ

© 2008 г. **А. М. Гржибовский**
Национальный институт общественного здоровья, г. Осло, Норвегия

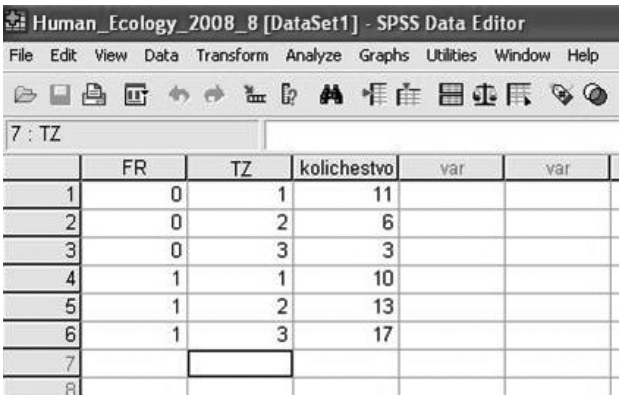
В статье рассматривается сравнение порядковых (ранговых, ординальных) данных с помощью гамма (γ)-критерия Гудмана – Краскела, d-критерия Сомера, критериев тау-b (τ-b) и тау-c (τ-c) Кендалла, а также критерия хи-квадрат (χ2) для линейного тренда с помощью пакета статистических программ SPSS. Изложенный материал дает общие сведения о статистических критериях, применяемых для проверки гипотез о наличии взаимосвязи между порядковыми переменными и призван вызвать интерес читателей журнала к прочтению специализированной литературы перед началом работы над будущими публикациями.
Ключевые слова: порядковые данные, анализ, SPSS.

В одном из предыдущих выпусков журнала были рассмотрены статистические критерии, предназначенные для проверки гипотез о наличии взаимосвязи между номинальными переменными [1]. Так как данные, измеряемые на порядковой (ранговой, ординальной) шкале, также относятся к качественным данным, критерии, которые можно применять для анализа многопольных таблиц номинальных данных, применимы и для порядковых. Однако порядковые данные несут больше информации, чем номинальные (известна направленность), поэтому применение критериев, предназначенных для сравнения номинальных данных, приведет к потере (точнее, к неиспользованию) части информации. Так, при анализе данных когортного исследования гипотезу о наличии статистической взаимосвязи между уровнем физической активности (низкий, средний, высокий) и вероятностью возникновением артериальной гипертезии через 10 лет наблюдения (есть/нет) можно проверить с помощью уже знакомого нам критерия χ^2 Пирсона, а величину эффекта оценить с помощью критерия V Крамера. Однако использование информации о направленности одной из переменных позволит сформулировать гипотезу более четко, например, изменяется ли вероятность возникновения артериальной гипертензии при увеличении физической активности.

В данной статье будут рассмотрены только те критерии для сравнения ординальных переменных, которые можно применить с помощью пакета статистических программ SPSS. Сразу определимся, что все из рассматриваемых в ней критериев предназначены для оценки величины эффекта и только критерий хи-квадрат (χ^2) для линейного тренда – для проверки гипотезы о наличии статистической взаимосвязи между переменными. Кроме того, гамма (γ)-критерий Гудмана – Краскела, критерии тау-b (τ-b) и тау-c (τ-c) Кендалла являются симметричными мерами взаимосвязи между переменными, то есть при их расчете не имеет значения, какая из переменных является зависимой, а какая независимой. В то же время d-критерий Сомера является асимметричным и его значения зависят от того, какая из переменных является зависимой.

Рассмотрим пример исследования, в которое были включены 60 человек с определенным заболеванием, тяжесть которого (переменная TZ) можно разделить на 3 степени (легкая или 1, средней тяжести или 2 и тяжелая или 3). Кроме того, собиралась информация о наличии или отсутствии у участников исследования определенного фактора риска (переменная FR). Таким образом, одна из переменных является порядковой (тяжесть заболевания, TZ), а другая дихотомической (фактор риска, FR). Исходные данные для примера сохранены в файле

Human_Ecology_2008_8.sav, который можно «скачать» с сайта журнала «Экология человека»: http://www.nsmu.ru/nauka_sgmu/rio/eco_human/. Обратите внимание, что если во всех прошлых выпусках журнала мы рассматривали только файлы, в которых каждая строчка соответствовала одному участнику исследования, то в файле Human_Ecology_2008_8.sav данные представлены в сгруппированном виде, как показано на рис. 1.



	FR	TZ	kolichestvo	var	var
1	0	1	11		
2	0	2	6		
3	0	3	3		
4	1	1	10		
5	1	2	13		
6	1	3	17		
7					
8					

Рис. 1. Окно «SPSS Data Editor» для файла Human_Ecology_2008_8.sav

Из рис. 1 видно, что из 60 участников исследования у 11 не было фактора риска (FR = 0) и заболевание было легкой степени (TZ = 1), у 6 участников не было фактора риска (FR = 0), а заболевание было средней степени тяжести (TZ = 2), и так далее до последней строки, из которой следует, что у 17 человек был выявлен фактор риска (FR = 1) и тяжелая степень изучаемого заболевания (TZ = 3). Переменная «kolichestvo» показывает абсолютное число участников исследования с характеристиками, обозначаемыми двумя предыдущими переменными. Такой способ введения данных в SPSS удобен, когда необходимо быстро провести анализ не очень сложных таблиц. В данном случае речь идет о шестиполльной таблице. Для введения данных в SPSS обычным способом пришлось бы использовать 60 строк – по одной строке на каждого участника исследования, что значительно увеличило бы время работы.

Однако, несмотря на удобство введения данных вышеописанным образом, программе необходимо будет «объяснить», что обозначает каждая из переменных, иначе SPSS будет считать, что в исследовании принимало участие всего 6 человек (так как всего 6 строк), для которых имеются данные по трем признакам (в файле три переменные), а это на самом деле не так. Для того чтобы SPSS могла произвести верные расчеты, нам необходимо провести «взвешивание» данных. Для этого на панели инструментов нужно выбрать меню «Data» (четвертое слева на рис. 1), затем в выскакивающем меню выбрать нижнюю строчку «Weight Cases», после чего появится диалоговое окно «Weight Cases» (рис. 2). В левом окне его следует выделить ту переменную, которая показывает количество наблюдений (в данном случае «kolichestvo» и перенести ее в правое окно, обозначенное «Frequency Variable». Теперь SPSS «знает», что каждая строка представляет собой столько участников исследования, сколько обозначено переменной «kolichestvo».

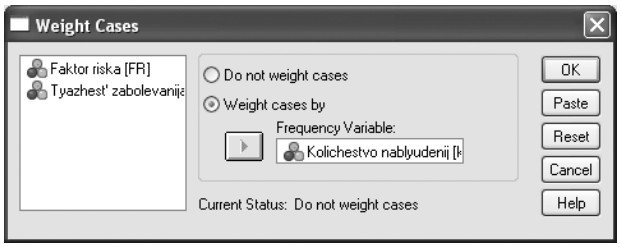


Рис. 2. Диалоговое окно «Weight Cases»

После этого можно проводить уже знакомую по предыдущим выпускам журнала процедуру кросстабуляции (Crosstabs), выбрав меню «Analyze» (шестое слева на рис. 1), затем «Descriptive Statistics», а затем «Crosstabs», после чего в одноименном диалоговом окне следует выбрать независимую (фактор риска, FR) и зависимую (тяжесть заболевания, TZ) переменную (рис. 3).

Для выбора статистических критериев, с помощью которых будем проводить анализ данных, выберем меню «Statistics», в нем отметим критерии, предназначенные для анализа порядковых данных, которые сгруппированы в правом столбце окна: «Gamma», «Somer's d», «Kendall's tau-b», «Kendall's tau-c» (рис. 4). Кроме того, отметим «Chi-square» в левом верхнем углу для проверки гипотезы об отсутствии взаимосвязи между переменными с помощью критерия χ^2 Пирсона и χ^2 для тренда, после чего, нажав на «Continue», вернемся в диалоговое окно «Crosstabs» (см. рис. 3) и запустим анализ путем нажатия на «OK».

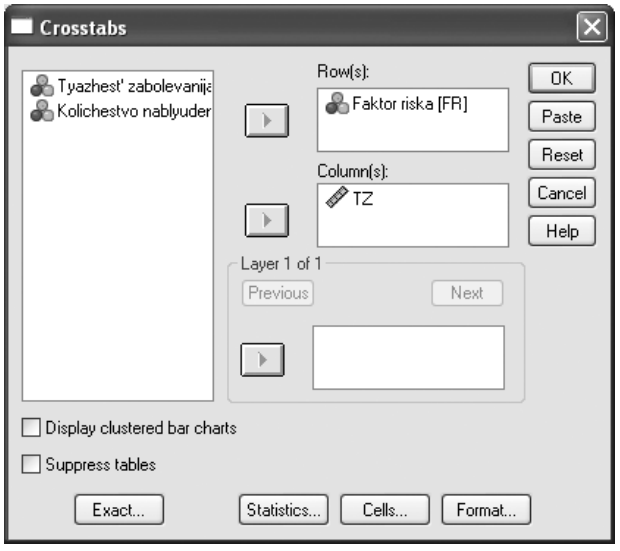


Рис. 3. Диалоговое окно «Crosstabs»

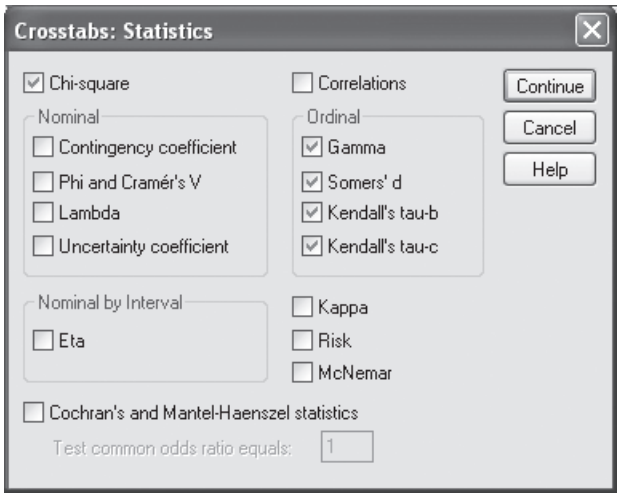


Рис. 4. Диалоговое окно «Crosstabs: Statistics»

Первая таблица результатов (рис. 5) выглядит знакомо и сообщает, что в исследовании всего (столбец Total) принимали участие 60 человек и что пропущенных значений (столбец Missing) ни для одной из изучаемых переменных обнаружено не было.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Faktor riska * Tyazhest' zabolevanija	60	100,0%	0	,0%	60	100,0%

Рис. 5. Данные об общем количестве значений и о количестве пропущенных значений переменных

Во второй таблице результатов (таблица сопряженности) представлены абсолютные значения (Count) и доли (%) для каждой из комбинаций признаков (рис. 6). Всего в исследовании было 60 человек, из которых 66,7 % (40 из 60) подвергались воздействию фактора риска (ряд FR+), а 33,3 % (20 из 60) — нет (ряд FR−). В то же время в изучаемой выборке 35 % (21 из 60) имели легкую степень заболевания, 31,7 % (19 из 60) — средней степени тяжести и 33,3 % (20 из 60) — тяжелую степень заболевания.

Faktor riska * Tyazhest' zabolevanija Crosstabulation

			Tyazhest' zabolevanija			
			1	2	3	Total
Faktor riska	FR-	Count	11	6	3	20
		% within Faktor riska	55,0%	30,0%	15,0%	100,0%
		% within Tyazhest' zabolevanija	52,4%	31,6%	15,0%	33,3%
		% of Total	18,3%	10,0%	5,0%	33,3%
	FR+	Count	10	13	17	40
		% within Faktor riska	25,0%	32,5%	42,5%	100,0%
		% within Tyazhest' zabolevanija	47,6%	68,4%	85,0%	66,7%
		% of Total	16,7%	21,7%	28,3%	66,7%
Total		Count	21	19	20	60
		% within Faktor riska	35,0%	31,7%	33,3%	100,0%
		% within Tyazhest' zabolevanija	100,0%	100,0%	100,0%	100,0%
		% of Total	35,0%	31,7%	33,3%	100,0%

Рис. 6. Таблица сопряженности

В следующей таблице результатов (рис. 7) представлены также уже известные из предыдущих публикаций (1) результаты применения критерия χ^2 Пирсона (Pearson Chi Square) и отношение правдоподобия (Likelihood ratio), называемое также критерием χ^2 Пирсона с поправкой на правдоподобие. Дополнительно к этим знакомым критериям в таблице присутствует строка «Linear-by-Linear Association», которая показывает результаты применения критерия χ^2 для линейного тренда. SPSS рассчитывает этот критерий во всех случаях, когда пользователь запрашивает критерий χ^2, однако следует помнить, что этот критерий наиболее целесообразно использовать в ситуациях, когда обе переменные являются порядковыми. Дихотомическую переменную всегда можно представить как порядковую. Номинальную же переменную представлять в виде порядковой было бы ошибкой. В даном случае независимая переменная является дихотомической, а зависимая — порядковой. Из данной таблицы следует, что существует взаимосвязь между фактором риска и степенью тяжести заболевания ($\chi^2(2) = 6{,}480$, $p = 0{,}039$), причем зависимость носит линейный характер ($\chi^2_{\text{для тренда}} = 6{,}346$, $p = 0{,}012$), однако окончательный вывод можно будет сделать только после обсуждения наличия в данном исследовании систематических ошибок и смешивающих факторов (конфаундеров).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,480[a]	2	,039
Likelihood Ratio	6,710	2	,035
Linear-by-Linear Association	6,346	1	,012
N of Valid Cases	60		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 6,33.

Рис. 7. Результаты применения критерия χ^2 Пирсона, отношения правдоподобия и теста для линейного тренда

«Вручную» критерий χ^2 для линейного тренда рассчитывается по формуле:

$$M^2 = r^2 \cdot (N-1),$$

где M^2 — значение критерия χ^2 для тренда (для 1 степени свободы), N — объем выборки, а r — значение коэффициента корреляции Пирсона между изучаемыми переменными. Интересно отметить, что критерий χ^2 для линейного тренда менее чувствителен к ситуациям, когда ожидаемое количество наблюдений в некоторых ячейках мало, то есть при наличии малого числа наблюдений надежность этого критерия может быть выше надежности классического критерия χ^2 Пирсона. Подробно расчет критерия χ^2 для линейного тренда с обсуждением плюсов и минусов метода описывается в [2].

Дополнительно к тому, что дает SPSS в таблице результатов на рис. 7, можно также оценить вероятность отклонения существующей зависимости от линейной. Для этого из общего значения χ^2 Пирсона можно вычесть значение χ^2 для линейного тренда и

полученную разность (6,480 – 6,346 = 0,134) сравнить с критическим значением χ^2 для оставшегося числа степеней свободы (2 – 1 = 1). Для 1 степени свободы критическое значение χ^2 = 3,841 (для уровня значимости 5 %), а полученная для данного примера разность составляет всего 0,134, что не позволяет отвергнуть нулевую гипотезу об отсутствии отклонения данных от линейной статистической взаимосвязи.

Оставшиеся две таблицы результатов показывают меры взаимосвязи между изучаемыми переменными (рис. 8 и 9).

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	,303	,108	2,742	,006
		Faktor riska Dependent	,253	,092	2,742	,006
		Tyazhest' zabolevanija Dependent	,379	,134	2,742	,006

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Рис. 8. Результаты применения асимметричного d-критерия Сомера

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	,309	,110	2,742	,006
	Kendall's tau-c	,337	,123	2,742	,006
	Gamma	,540	,171	2,742	,006
	Spearman Correlation	,328	,117	2,645	,010[c]
Interval by Interval	Pearson's R	,328	,117	2,644	,011[c]
N of Valid Cases		60			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Рис. 9. Результаты применения симметричных мер взаимосвязи порядковых переменных

Перед тем как перейти к разбору результатов применения критериев, предназначенных для ранговых переменных, остановимся на основных принципах, лежащих в основе расчетов. Поскольку для порядковых переменных характерна направленость, в основе всех критериев расчет количества нарушений порядка. Все пары данных можно классифицировать как конкордантные, дискордантные и связанные либо по зависимой, либо по независимой переменной. Конкордантные пары также называют проверсиями, а дискордантные – инверсиями. Для начала обозначим каждую из ячеек таблицы сопряженности (рис. 6) относительно ее положения по рядам (R) и столбцам (C).

Для начала рассмотрим ячейки R1C1 (первый ряд, первый столбец) и R2C2 (второй ряд, второй столбец). Для ячейки R2C2 номер как ряда, так и столбца выше, чем таковые для ячейки R1C1, что позволяет считать эту пару ячеек конкордантной (проверсия). Всего в ячейке R1C1 11 наблюдений, а в ячейке R2C2 – 13, значит, в ячейках R1C1 и R2C2 содержится информация о 13 x 10 = 130 конкордантных парах индивидуальных наблюдений.

Теперь рассмотрим ячейки R1C2 (первый ряд, второй столбец) и R2C1 (второй ряд, первый столбец). В данном случае порядковый номер ряда для ячейки R1C2 меньше, чем для ячейки R2C1, а порядковый номер столбца – больше, что позволяет считать эту пару ячеек дискордантной (инверсия). По аналогии с предыдущим примером, поскольку в ячейке R1C2 6 наблюдений, а в ячейке R2C1 – 10, общее число конкордантных пар индивидуальных наблюдений для этих ячеек составит 6 x 10 = 60 наблюдений.

Помимо конкордантных и дискордантных пар существуют так называемые связанные или сцепленные пары, которые в англоязычной литературе обозначаются как ties или tied ranks. Посмотрим на ячейки R2C2 (второй ряд, второй столбец) и R2C3 (второй ряд, третий столбец). Для этой пары ячеек ряд остается постоянным, хотя номер столбца для второй ячейки выше. Такие ячейки называются связанными по ряду (так как ряд один и тот же несмотря на принадлежность ячеек к разным столбцам). Для ячеек R2C2 и R2C3 количество индивидуальных пар наблюдений составит 13 x 17 = 221. Примером пар ячеек, связанных по столбцам, может служить пара R2C2, R1C2, так как ячейки принадлежат к одному столбцу, но к разным рядам. Общее количество связанных по столбцу наблюдений для этой пары ячеек составит 3 x 17 = 51.

Количество возможных комбинаций между ячейками n = 0,5N(N – 1), где N – количество ячеек. Поскольку в нашей таблице 6 ячеек, всего возможно 15 комбинаций, которые представлены в таблице. Расчеты количества конкордантных, дискордантных пар наблюдений, и т. д. пригодятся позже при расчете статистических критериев.

Все комбинации ячеек таблицы сопряженности (см. рис. 6) **для рассматриваемого в тексте примера**

Пары ячеек		Количество пар индивидуальных наблюдений	
		Расчет	Всего
Конкордантные пары			
R1C1	R2C2	11x13	143
R1C1	R2C3	11x17	187
R1C2	R2C3	6x17	102
Всего:			432
Дискордантные пары			
R2C1	R1C2	10x6	60
R2C1	R1C3	10x3	30
R2C2	R1C3	13x3	39
Всего:			129
Пары, связанные по рядам			
R1C1	R1C2	11x6	66
R1C1	R1C3	11x3	33
R1C2	R1C3	6x3	18
R2C1	R2C2	10x13	130
R2C1	R2C3	10x17	170
R2C2	R2C3	13x17	221
Всего:			638
Пары, связанные по столбцам			
R1C1	R2C1	11x10	110
R1C2	R2C2	6x13	78
R1C3	R2C3	3x17	51
Всего:			239

Таким образом, всего для данного примера существует 432 конкордантных, 129 дискордантных и 877 связанных пар наблюдений, из которых 638 связаны по рядам, а 239 — по столбцам.

Гамма-критерий Гудмена — Краскела (Goodman-Kruskal's gamma) основан на сравнении количества конкордантных и дискордантных пар. Расчет критерия может производиться вручную по формуле:

$$\gamma = \frac{C - D}{C + D},$$

где C — количество конкордантных, а D — количество дискордантных пар. Для нашего примера:

$$\gamma = \frac{432 - 129}{432 + 129} = \frac{303}{561} = 0{,}540,$$

что соответствуют результату, выданному SPSS и представленному на рис. 9 в строке «Gamma». Помимо абсолютного значения критерия SPSS рассчитывает страндартную ошибку и асимптотический уровень значимости. Для данного примера нулевую гипотезу о том, что значение γ не отличается от нуля, можно отклонить для критического уровня значимости 5 %, так как достигнутый уровень значимости составил всего 0,006 (см. рис. 9). Критерий γ может варьировать от −1 до 1, причем 1 означает полную прямо пропорциональную взаимосвязь между переменными, −1 — полную обратную взаимосвязь между переменными, а 0 — полное отсутствие какой-либо связи между изучаемыми признаками. Чем ближе значение критерия к 1 или −1, тем сильнее взаимосвязь. Гамма — симметричный критерий и не зависит от того, какая из переменных является зависимой. Технически γ показывает, насколько больше в исследуемой выборке конкордантных пар, чем дискордантных, относительно общего числа конкордантных и дискордантных пар. При этом полностью игнорируются связанные пары наблюдений (см. формулу). Можно также интерпретировать γ-критерий как пропорциональное уменьшение ошибки прогнозирования одной переменной при наличии информации о другой. Так, полученное значение 0,54 можно интерпретировать следующим образом: наличие информации о факторе риска может уменьшить ошибку предсказания степени тяжести заболевания на 54 %. Следует повторить, что при расчете γ-критерия совершенно не используется информация о связанных парах наблюдений, то есть для рассматриваемого примера при расчете γ-критерия используются данные только для 561 пары и не используется информация о 877 парах наблюдений.

Из критериев, предназначенных для сравнения порядковых данных с учетом связанных пар наблюдений, SPSS позволяет рассчитывать критерии τ-b и τ-c Кендалла (Kendall's tau-b, tau-c). Оба критерия могут принимать значения в том же диапазоне, что и γ-критерий Гудмена — Краскела. Оба τ-критерия Кендалла также показывают силу взаимосвязи между переменными.

Критерий τ-b чаще всего применяется для таблиц 2 x 2, однако возможно его применение и для многопольных таблиц. Как и γ-критерий, он показывает разность между количеством конкордантных и дискордантных пар, но с делением на геометрическое среднее количества пар, связанных по рядам, и количества пар, связанных по столбцам, что можно представить в виде:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + X) \cdot (C + D + Y)}},$$

где X — количество пар наблюдений, связанных по рядам, Y — количество пар наблюдений, связанных по столбцам, а C и D — количество конкордантных и дискордантных пар. Для нашего примера формула для расчета τ-b Кендалла принимает вид:

$$\tau_b = \frac{432 - 129}{\sqrt{(432 + 129 + 638) \cdot (432 + 129 + 239)}} = \frac{303}{979} = 0{,}309$$

Результат аналогичен рассчитанному SPSS и представленному на рис. 9 в строке Kendall's tau-b. Кроме того, SPSS рассчитывает значение стандартной ошибки и уровень значимости, последний из которых равен в данном случае 0,006, что говорит о том, что гипотезу о равенстве критерия нулю можно отвергнуть, используя традиционный 5 % критический уровень значимости. Данный критерий лучше использовать только для квадратных таблиц, то есть для таблиц, в которых количество рядов равно количеству столбцов. Критерий τ-b является симметричным критерием, для которого неважно, какая из переменных является зависимой. Интерпретировать абсолютное значение критерия достаточно сложно, но следует помнить, что он, как и γ-критерий, может принимать значения от −1 до 1 и показывает силу взаимосвязи между переменными.

Критерий τ-c Кендалла, называемый иногда еще критерием τ-c Стюарта (Stuart's tau-c) или критерием Кендалла — Стюарта, осуществляет коррекцию на общее количество рядов и столбцов в таблице сопряженности и использует общее количество наблюдений, а не только конкордантных и дискордантных пар, как γ–критерий. Кроме того, в отличие от критерия τ-b он может применяться не только для четырехпольных или квадратных таблиц, что обеспечивает его более широкое использование. Расчет критерия τ-c Кендалла производится по формуле:

$$\tau_c = \frac{2m(C - D)}{N^2 (m - 1)},$$

где m — меньшее значение количества рядов или столбцов (в нашем примере таблица имеет 2 ряда и 3 столбца, значит, m = 2), N — объем выборки, C

и D – количество конкордантных и дискордантных пар. Для рассматриваемого примера

$$\tau_c = \frac{2 \cdot 2(432-129)}{60^2(2-1)} = \frac{1212}{3600} = 0{,}337,$$

что соответствует значению, рассчитанному SPSS (см. рис. 9, строка Kendall's tau-c). Кроме того, SPSS рассчитывает значение стандартной ошибки и уровень значимости, последний из которых равен в данном случае 0,006, что говорит о том, что гипотезу о равенстве критерия нулю можно отвергнуть, используя традиционный 5 % критический уровень значимости. Из проблем, связанных с применением этого критерия, следует отметить, что его значение в большой степени зависит от размеров таблицы (количества рядов и столбцов), то есть от степени категоризации данных исследователем, что является поводом для критики критерия.

Коэффициенты корреляции Пирсона (строка Pearson's R) и Спирмена (строка Spearman Correlation), также присутствующие в результатах, представленных на рис. 9, предназначены прежде всего для оценки взаимосвязи между количественными перемеными и будут рассматриваться в следующем выпуске журнала.

На рис. 8 представлены результаты применения единственного предлагаемого SPSS асимметричного, или направленного, критерия для оценки взаимосвязи между двумя порядковыми переменными. Асимметричность критерия подразумевает, что значение его обусловлено тем, какая из переменных является зависимой, а какая независимой, что не имело значения для всех критериев, рассмотренных выше.

Критерий d Сомера (Somer's d) учитывает только связанные пары данных по столбцам, если в столбцах записана зависимая переменная. Для нашего примера зависимой переменной является тяжесть заболевания (переменная TZ), которая была занесена нами в столбцы, а независимая переменная (фактор риска, FR) – в ряды (см. рис. 6). Критерий d Сомера рассчитывается по формуле:

$$d = \frac{C - D}{C + D + Y},$$

где C и D – количество конкордантных и дискордантных пар, а Y – количество пар, связаных по зависимой переменной. Для нашего примера значение критерия d Сомера будет

$$d = \frac{432 - 129}{432 + 129 + 239} = 0{,}379,$$

что не противоречит расчетам SPSS. В таблице, представленной на рис. 8, следует смотреть только на нижнюю строку «Tyazhest' zabolevanija Dependent», так как SPSS автоматически всегда рассчитывает данный критерий, представляя поочередно то одну, то другую переменную в виде зависимой, поэтому исследователю надо четко помнть, какая гипотеза проверяется в исследовании. Помимо двух асимметричных мер критерия d Сомера SPSS также рассчитывает симметричное значение d (верхняя строка, Symmetric) путем усреднения количества связанных пар наблюдений по рядам и столбцам и показывает силу взаимосвязи независимо от того, какая из переменных является зависимой. Неудивительно, что это значение всегда находится между двумя асимметричными значениями d. Значение d Сомера показывает разность между вероятностью того, что случайно выбранная пара наблюдений конкордантна, и вероятностью того, что эта пара дискордантна, при условии, что наблюдения не связаны по независимой переменной. Возможные значения d Сомера варьируют от –1 до 1 (от полной прямо пропорциональной до полной обратно пропорциональной взаимосвязи), а 0 обозначает полную независимость переменных друг от друга. В зависимости от количества связанных пар значение d Сомера в большинстве случаев будет несколько меньше, чем значение γ-критерия Гудмена – Краскела [2].

Подробную информацию о способах сравнения порядковых данных можно получить из [3–6] и специализированной литературы по анализу качественных данных [2, 8, 9]. Подробно о расчетах ошибки и уровня значимости для каждого из рассмотренных критериев можно прочитать в [7].

В следующем выпуске будут рассмотрены меры взаимосвязи между непрерывными переменными.

Список литературы

1. *Гржибовский А. М.* Анализ номинальных данных (независимые наблюдения) / А. М. Гржибовский // Экология человека. – 2008. – № 6. – С. 58–68.

2. *Agresti A.* Introduction to categorical data analysis / A. Agresti. – N. Y. : John Wiley and Sons, 2007. – 400 p.

3. *Goodman L.* Measures of Association for cross-classification I / L. Goodman, W. H. Kruskal // Journal of the American Statistical Association. – 1954. – Vol. 49. – P. 732–764.

4. *Goodman L.* Measures of Association for cross-classification II / L. Goodman, W. H. Kruskal // Journal of the American Statistical Association. – 1959. – Vol. 54. – p. 123 – 163.

5. *Goodman L.* Measures of Association for cross-classification III / L. Goodman, W. H. Kruskal // Journal of the American Statistical Association. – 1963. – Vol. 58. – p. 310 – 364.

6. *Goodman L.* Measures of Association for cross-classification IV / L. Goodman, W. H. Kruskal // Journal of the American Statistical Association. – 1972. – Vol. 67. – P. 415–421.

7. *Liebetrau A. M.* Measures of Association / A. M. Liebetrau. – Newbery Park : Sage Publications, 1983. – 100 p.

8. *Hildebrand D. K.* Analysis of ordinal data / D. K. Hildebrand, J. D. Laing, H. Rosenthal. – Newbery Park : SAGE University Paper, 1977. – 80p.

9. *Siegel S.* Nonparametric Statistics for the Behavioural Sciences / S. Siegel, N. J. Castellan. – N. Y. : McGraw-Hill, 1988. – 399 p.

ANALYSIS OF ORDINAL DATA

A. M. Grjibovski

National Institute of Public Health, Oslo, Norway

The article describes analysis of ordinal data by chi-squared test for linear trend, symmetric measures of the effect size such as Goodman-Kruskal's gamma, Kendall's tau-b and tau-c and asymmetric Somer's d-statistic. The use of all abovementioned tests in SPSS software is also presented. The paper provides only general introduction to the methods of analysis of ordinal data. The readers are encouraged to consult statistical literature prior to analysing own data and preparing manuscripts.

Key words: ordinal data, analysis, SPSS.

Контактная информация:

Гржибовский Андрей Мечиславович — старший советник Национального института общественного здоровья, г. Осло, Норвегия

Адрес: Nasjonalt folkehelseinstitutt, Pb 4404 Nydalen, 0403 Oslo, Norway

Тел.: +47 21076392, +47 45268913; e-mail: angr@fhi.no

Статья поступила 21.07.2008 г.

Уважаемые читатели, редакция приносит извинения за допущенные в статьях Практикума в № 5 и 6 ошибки. Просим учесть приведенные ниже исправления.

Номер 5

С. 57. Нижняя и верхняя границы доверительного интервала поменяны местами.

Следует читать:

от $p - 1.96\sqrt{\frac{p(1-p)}{N}}$ до $p + 1.96\sqrt{\frac{p(1-p)}{N}}$

Номер 6

С. 59. Правая колонка, сразу под первой формулой.

Следует читать:

i — номер ряда (строки, от 1 до r), j — номер столбца (от 1 до c.

Oij— фактическое количество наблюдений в ячейке ij,

Eij— ожидаемое число наблюдений в ячейке ij.

С. 60. Левая колонка. Должна быть формула:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(|O_{ij} - E_{ij}| - 0{,}5)^2}{E_{ij}}$$

С. 64. Левая колонка, вторая строка сверху. Разность рисков рассчитывается по формуле A(A+B) — C(C+D).

С. 67. Таблица 3

Различия между относительным риском и отношением шансов (объяснения в тексте)

Ситуация	Вероятность исхода для тех, у кого **нет фактора** риска	Вероятность исхода для тех, у кого есть фактор риска	Относительный риск	Отношение шансов
1	0,05, или 5 %	0,1, или 10 %	2,0	2,1
2	0,45, или 45 %	0,90, или 90 %	2,0	11,0