Федеральное агентство по образованию

Санкт-Петербургский Государственный Политехнический Университет

МАТЕМАТИКА В ПОЛИТЕХНИЧЕСКОМ УНИВЕРСИТЕТЕ

Ю.Д. Максимов

МАТЕМАТИКА

НОВЫЕ МЕТОДЫ ДЕТЕРМИНАЦИОННОГО И КОРРЕЛЯЦИОННОГО АНАЛИЗА

Санкт-Петербург
Издательство политехнического университета
2007

УДК 519.2
ББК 22.16я 73
М 171

Рецензент – действительный член Международной академии наук высшей школы, заслуженный деятель науки и техники Российской Федерации, доктор технических наук, профессор Петербургского университета путей сообщения *В.Г. Дегтярев*.

Максимов Ю.Д.
Математика. Новые методы детерминационного и корреляционного анализа. СПб.: Изд-во Политехнического университета, 2007.252 с. (Математика в политехническом университете).

В книге дается изложение новых методов детерминационного и корреляционного анализа, основанных на новых коэффициентах детерминации и корреляции, описывающих парную связь случайных величин. 8 типов рассматриваемых коэффициентов детерминации обладают важным свойством: равенство коэффициента нулю обеспечивает независимость случайных величин. Этим свойством не обладает обычный линейный коэффициент корреляции. 4 коэффициента детерминации могут применяться для описания парных связей количественных, качественных и смешанных признаков. Каждый коэффициент детерминации путем введения параметра порождает спектр аналогичных коэффициентов. Дается история вопроса. 2-я глава целиком посвящена теории максимального коэффициента корреляции, которая создана в середине 20-го столетия и неизвестна широкой математической общественности.

Предназначена для математиков, инженеров, исследователей в области экономики, социологии, медицины и всех тех, кто занимается изучением связей в Природе и обществе.

Табл. 60 Илл. 14 Библиогр.: 42 назв.

Печатается по решению редакционно-издательского совета Санкт-Петербургского государственного политехнического университета.

ISBN 5-7422-1573-8

Что же ты ищешь, мальчик-бродяга,
В этой забытой богом земле?
Синяя птица манит куда-то.
Что ты увидел в таинственной мгле?

Фрагмент песни.

Предисловие

Задача предлагаемой вниманию книги – изучение парной зависимости между случайными величинами с помощью новых числовых характеристик. В этой области среди всего многообразия вероятностно-статистических тем наметился прорыв в область монополии линейного коэффициента корреляции, составляющего основу корреляционного анализа, построенного К. Пирсоном и его последователями в начале 20-го века. Недостатком линейного коэффициента корреляции, как меры величины зависимости, в общем случае является невозможность гарантировать независимость случайных величин при равенстве коэффициента корреляции нулю. Именно поэтому предпринимаются попытки найти другие простые коэффициенты связи, лишенные указанного недостатка. Первый успех в этом направлении был достигнут в 40-е годы 20-го века, когда молодой тогда ученый О.В. Сарманов, ученик академика С.Н. Бернштейна, разработал теорию максимального коэффициента корреляции, обладающего теми же основными свойствами, что и линейный коэффициент корреляции, но гарантирующий независимость случайных величин при равенстве коэффициента нулю. Теория – достаточно сложная, поэтому, видимо, и неизвестная широкой математической аудитории.

В настоящей книге предлагаются, как инструменты исследования, несколько спектров коэффициентов связи между случайными величинами. Они построены совсем на других идеях и названы коэффициентами детерминации, а не корреляции, так как нормированы более узкими границами изменения: от нуля до единицы, но зато все они обладают нужным свойством: гарантируют независимость случайных величин при равенстве коэффициента нулю. Верхняя граница – единица, как правило, достижима и указывает в этом случае на функциональную зависимость случайных величин, в общем случае – нелинейную. Об этом подробнее – во введении и в основном тексте книги. Тогда предлагаемые коэффициенты могут служить полноправными мерами связи, а не тесноты связи, как стали теперь говорить, используя линейный коэффициент корреляции. В книге 13 глав. При изложении глав, где вводятся и изучаются все новые коэффициенты связи, используется математический аппарат в объеме втузовского курса математики. При чтении параграфов, посвященных дискретным случайным величинам, используются знания в еще меньшем объеме, которые обеспечиваются курсами математики для

гуманитариев. Учитывая эти обстоятельства, можно полагать, что книга доступна широким кругам исследователей – математиков, инженеров, научных работников – технического, экономического, медицинского, гуманитарного профилей, то есть всем тем, кто исследует связи между явлениями, событиями, признаками. Книга – актуальна, так как установление связи – это задача любого закона, применяемого для прогнозирования. Первые две главы книги посвящены изучению коэффициентов связи, ранее известных, - линейному, максимальному и ряду других. Сравнение с ними новых коэффициентов – важная задача для уяснения роли и тех и других. Кроме того, новые коэффициенты связаны с некоторыми старыми и конструктивно. 9-я глава посвящена приложениям новых коэффициентов пока к анализу регрессии. В дальнейшем предполагаются приложения к корреляционной и спектральной теориям случайных процессов, к анализу явлений социальной сферы. 10-я глава посвящена изложению общего оптимизационного метода построения согласованных числовых характеристик положения и рассеяния и их статистических оценок. Этот метод дает единый подход к построению многих известных характеристик, которые ранее воспринимались изолированно. На первый взгляд эта глава стоит в стороне от вопросов исследования связей, однако величина рассеяния влияет на величину связи. При отсутствии рассеяния имеет место детерминизм и все коэффициенты связи должны быть равны нулю. Единый подход позволяет построить спектр пар характеристик – так же, как и для коэффициентов связи. На идее согласования характеристик положения и рассеяния построены комбинированные коэффициенты детерминации в главах 7и 10. 12-я глава содержит таблицы.

Это таблица 1 коэффициентов детерминации и корреляции. В таблицу помещены 26 коэффициентов (не все), указаны их названия, конструкции, комментарии. Это целесообразно, так как их очень много, нужны рациональные обозначения, названия, классификация. Таблица 2 содержит оригинальную таблицу нормирующих коэффициентов для ликвидации смещения оценок положения и рассеяния в случае нормального закона.

Даны также таблица обозначений и названий. В главе 13 рассматриваются историко-философские аспекты зависимости.

Для каждого коэффициента, числовой характеристики приведены примеры, иногда требующие применения приближенных численных методов. Пакеты программ типа

Mathcad не применялись, что позволяет легко оценить объем и трудности вычислений. В дальнейшем для вычислений могут быть составлены специальные программы.

В книге принята автономная система нумерации. Нумерация параграфов – двухпозиционная, привязанная к главам. Номера формул, таблиц, примеров, теорем – в каждом параграфе свои. Для удобства пользования имеются в конце книги предметно-именной указатель и список литературы.

Автор Максимов Юрий Дмитриевич – профессор кафедры высшей математики Санкт-Петербургского государственного политехнического университета.

Санкт - Петербургский государственный политехнический университет. 2007г.

Сущность любого закона
Природы – описание зависимости.
Сначала же эту зависимость нужно
установить и как-либо измерить.

Введение.

Эта книга – о корреляции, детерминации, регрессии, то есть о тех областях науки, где исследуется зависимость между случайными величинами, которые математически описывают связи между явлениями и признаками. Закономерно, что эти понятия и термины сначала появились в прикладных областях знания, а не в математике. Познакомимся сначала с некоторой историей вопросов, рассматриваемых в книге, что позволит лучше уяснить их сущность и прикладные возможности.

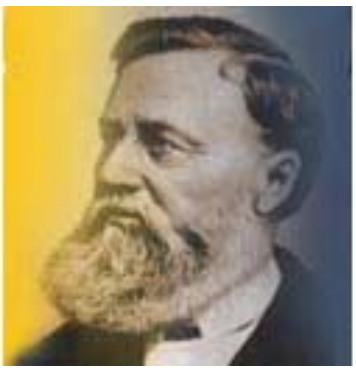
1°. Корреляция. Понятие «корреляция» ввел в науку французский палеонтолог профессор Жорж Кювье (1769 – 1832) в 1806 году, когда занимался палеонтологией и сравнительной анатомией.

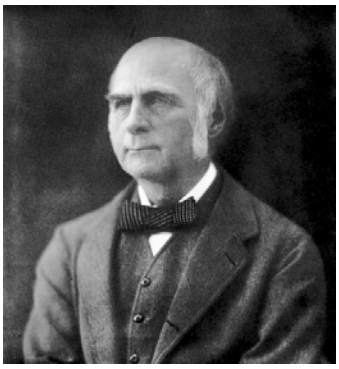
Кювье Жорж

Слово «корреляция» происходит от позднелатинского слова «correlation», что означает соответствие, соотношение. В отличие от слова «relation» это не просто «отношение, связь», а «как бы связь», то есть связь, но не в привычной в то время детерминистской, функциональной форме. Кювье заметил связь органов живого организма между собой, связь строения животного с образом его жизни, связь видов животных и растений с определенным временем их жизни и много других связей. Это позволило ему сформулировать общие принципы «корреляции органов» и «функциональной корреляции», относящиеся к сравнительной анатомии и палеонтологии. Эти принципы были применены для определения возраста земных слоев по ископаемым остаткам и наоборот, реконструкции вида ископаемых форм животных по ископаемым остаткам. Эти принципы также были основанием для выполненной им классификации видов живых организмов, ископаемых и ныне живущих. Для лучшего понимания смысла корреляции по Кювье можно привести следующий интересный эпизод из жизни самого Кювье [8]. В дни университетского праздника студенты решили подшутить над профессором Кювье. Они вырядили одного из студентов в козлиную шкуру с рогами и копытами и подсадили его в окно спальни Кювье. Ряженый загремел копытами и завопил: «Я тебя съем!». Кювье проснулся. Увидел силуэт с рогами и копытами и спокойно отвечал: «Если у тебя рога и копыта, то по закону корреляции ты травоядное, и съесть меня не можешь. А за то, что не знаешь закона корреляции, получишь двойку!».

Корреляция от Кювье и до Ф.Гальтона понималась как связь, описываемая качественно. Она не детерминированная связь, а статистическая, но наблюдавшаяся в то время в узком круге явлений.

2. Регрессия. Понятие регрессии связано с именем Ф. Гальтона. ФренсисГальтон (1822 – 1911), знаменитый английский антрополог, биолог. психолог и метеоролог, понял, что ***корреляция, иначе – корреляционная зависимость – это связь в среднем между любыми случайными величинами***. Корреляционный анализ (термин также принадлежит Гальтону)





Гальтон Френсис

занимается измерением величины (тесноты, как теперь выражаются,) корреляции с помощью числовых показателей, коэффициентов. Гальтон же ввел в науку и понятие регрессии (1885 г.), тесно связанной с корреляцией. Регрессия выражает корреляционную зависимость между случайными величинами функционально в среднем. Так же, как и Кювье, Гальтон открыл регрессионную зависимость, наблюдая живую природу. Изучая размеры семян бобов, он заметил, что наследники семян не проявляли тенденции к воспроизведению размеров своих родителей, а, напротив, всегда были ближе к середине, чем они (под серединой имеется в виду среднее арифметическое). А именно: семена были меньше, чем их родители, если родители были велики, и больше, если родители были малы. Эту закономерность Гальтон назвал регрессией, так как наблюдалось изменение в противоположную сторону. Дальнейшие наблюдения показали, что в среднем сыновняя регрессия к середине прямо пропорциональна отклонению родителей от нее, что позволяет описать ее функционально в виде линейной функции (например, $y = \overline{y} + \frac{2}{3}\left(x - \overline{x}\right)$). Та же регрессионная закономерность описывается Гальтоном в результате наблюдений над ростом 930 взрослых детей и 205 их родителей. Термин «регрессия» прижился и сейчас им называют функциональную зависимость в среднем между любыми случайными величинами, а не только между теми, которые изучал Гальтон. Большой вклад в науку внес Гальтон при изучении вопросов передачи количественных и качественных признаков по наследству. Они изучались Гальтоном с помощью численных расчетов на основе концепции корреляции. До сих пор мы пользуемся собранной Гальтоном статистикой из области демографии, наследственности, социологии с примерами численных расчетов корреляции. Гальтон являлся большим сторонником эволюционной теории Ч. Дарвина (Дарвин был двоюродным братом Гальтона). В этой области также

ярко проявляются корреляционные связи. Являясь президентом Британского королевского научного общества (академии наук), он, конечно, немало способствовал популяризации эволюционного учения (дарвинизма). Отметим также, что Гальтон является основоположником науки, называемой евгеника, основанной также и на корреляционном принципе. Евгеника (от греч. Eugenes – хорошего рода) – это учение о наследственном здоровье человека и путях улучшения его наследственных свойств
(книга: Ф. Гальтон. Наследственность таланта, его законы и последствия. 1869). Не случайно, что основоположники современной математической статистики, крупнейшие ученые-статистики Френсис Гальтон, Карл Пирсон, Рональд Фишер в конце своей деятельности стали преподавать в университетах евгенику и биологию, так как там нашли и истоки статистики и лучшие ее приложения.

3º. Формула вычисления линейного коэффициента корреляции. Формулу, которой мы теперь пользуемся для (линейного) коэффициента корреляции (см. далее §1.1), явно выписал К Пирсон в 1896 г., но вплотную к ней подошел в 1846 г., не сделав еще одного шага **Огюст Браве (Bravais Auguste,1811 – 1863).** К этому времени был уже известен метод наименьших квадратов, созданный независимо работами французского математика Адриена Лежандра (1805г.), американского математика Роберта Эдрейна (1808г.), немецкого математика Карла Гаусса (1809г.). Была осмыслена случайная природа ошибок измерения и сформулирован нормальный закон (Гаусса) их распределения. Развивая теорию ошибок, Браве аналитически и геометрически нашел прямую регрессии, угловой коэффициент которой, как выяснил позднее К. Пирсон, выражается через коэффициент корреляции. Биография Браве свидетельствует о многогранности его таланта и полна курьезных метаморфоз. Морской офицер и участник научных экспедиций в Алжир и на Мон Блан (1832-1840); профессор астрономии (Лион, с 1841-1845.), ботаник, метеоролог, физик-теоретик, кристаллограф (именем Браве названы открытые им кристаллические решетки), член Парижской Академии Наук (1854), профессор l’Ecole Polytechnique (1845-1856).

Браве Огюст

4º. Карл Пирсон (1857 – 1936), английский математик-статистик, биолог, философ, основоположник знаменитого журнала биометрика превратил концепцию корреляции в математическую, статистическую теорию.

На основе теоретических исследований, аналогичных тем, что проводил О. Браве, К. Пирсон выписал формулу для линейного коэффициента корреляции (**product-moment correlation coefficient) :**

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\frac{x_i - \overline{x}}{s_X}\frac{y_i - \overline{y}}{s_y}.$$

Здесь $\overline{x}, \overline{y}$ – выборочные средние, s_X, s_Y – выборочные средние квадратические отклонения случайных величин X, Y; n – объем выборки.

Эта формула оказалась очень плодотворной для исследования зависимостей, на ней построены целые теории – корреляционный анализ в теории вероятностей, корреляционная теория случайных процессов и др.

К.Пирсону принадлежит теорема об асимптотическом законе распределения статистики теста хи-квадрат (теория проверки гипотез о законе распределения случайной величины).

Пирсон Карл

Со времени К. Пирсона в корреляционном анализе кроме линейного коэффициента корреляции применяется большое число других коэффициентов, которые называются и коэффициентами корреляции с указанием имен их создателей и коэффициентами с другими названиями и свойствами, призванные измерять величину (тесноту) связи (детерминации, ассоциации, контингенции и т. д.). Некоторые из этих коэффициентов принадлежат и Пирсону.

5°.Рональд Фишер (1890, Лондон – 1962, Аделаида, Австралия), английский генетик, математик-статистик внес огромный вклад в теорию вероятностей и математическую статистику. Достаточно сказать, что он является родоначальником дисперсионного анализа и вместе с К.Пирсоном заложил основы теории проверки статистических гипотез. Р. Фишер ввел понятие достаточной статистики, создал метод максимального правдоподобия построения статистических оценок и теорию фидуциальных доверительных интервалов. Хотя в настоящее время используются в основном доверительные интервалы по Ю. Нейману, тем не менее интерес к доверительным интервалам по Р. Фишеру не пропал. Много внимания им уделялось и корреляционному анализу. До сих пор применяются приемы Фишера проверки значимости связи в случае нормального закона.

Р. Фишера по праву можно считать одним из основоположников современной математической статистики. Член Лондонского королевского общества (1929). Окончил колледж в Кембридже (1912). Работал статистиком в «Меркантайл энд дженерал инвестмент компани» (1913–15). В 1919–33 работал в отделе статистики Ротемстедской экспериментальной (сельскохозяйственной) станции. В 1933–43 профессор евгеники Лондонского университета. В 1943–57

профессор генетики Кембриджского университета, в 1956–59 руководил одним из его колледжей.

Фишер Рональд Айлмер

6º. Юл Джордж Одни (Yule George Udny. Scotland, 1871 – 1951, Cambridge, England), английский статистик, профессор Кембриджского университета, президент Королевского статистического общества, ученик К. Пирсона.

Работал в области теорий регрессии и корреляции. Его коэффициент контингенции для событий (§1.3.) применяется в данной книге для построения контингенциального коэффициента детерминации для случайных величин (гл.5).

Юл Георг Одни

7º. Бернштейн Сергей Натанович (!880 – 1968), российский математик, академик АН СССР (1929), иностранный член Парижской академии наук (1955) работал в Харьковском университете (1907 – 1933), Ленинградских университете и политехническом институте (1933 – 1941), математическом институте АН СССР (1935 – 1968), Известен выдающимися достижениями во многих областях математики, в том числе в теории вероятностей и математической статистике. Исследовал коэффициент корреляции между событиями (§1.2.), применяемый здесь в качестве ядра для построения более сложных коэффициентов детерминации и корреляции.

Бернштейн Сергей Натанович

8º. Сарманов Олег Васильевич (1916 – 1977), ученик С.Н. Бернштейна, профессор, работал на математико-механическом факультете Ленинградского университета (1946 – 1955), затем в Москве в математическом институте АН СССР. Разработал теорию максимального коэффициента корреляции (гл.2), применил ее к исследованию стационарных марковских процессов [34]. Максимальный коэффициент корреляции – более совершенный, но аналитически гораздо более сложный инструмент для измерения зависимости между случайными величинами. Равенство нулю этого коэффициента обеспечивает независимость случайных величин, чего нельзя сказать о линейном коэффициенте корреляции Пирсона. Автор этой книги – ученик О.В. Сарманова.

Слева направо: Линник Юрий Владимирович,
Сарманов Олег Васильевич, Максимов Юрий Дмитриевич
1949 г.

К коэффициентам связи целесообразно предъявить ряд требований. Перечислим их.

1. Нормированность: $0 \le |k| \le 1$ или $0 \le k \le 1$.
2. Достижимость граничных значений 0 и 1.
3. Сопоставимость граничных значений коэффициента k с предельными случаями связи между случайными величинами: равенство нулю коэффициента k означает независимость случайных величин X, Y, а равенство модуля коэффициента k единице означает функциональную зависимость между X и Y.

Не все эти свойства имеют место у известных и применяемых на практике коэффициентов. Наиболее жестким требованием является третье о независимости случайных величин при равенстве коэффициента нулю. Чтобы это требование выполнялось, необходимо применять в конструкции коэффициента необходимые и достаточные условия независимости случайных величин.

Для дискретных случайных величин эти условия имеют вид

$$p_{ij} = p_{i\cdot} p_{\cdot j} \text{ для любых } i, j = 1, 2, \ldots \quad (1)$$

Здесь $p_{ij} = P\left(X = x_i, Y = y_j\right) \quad i, j = 1, 2, \ldots$ – закон распределения двумерной случайной величины $\left(X, Y\right)$;

$p_{i\cdot} = P\left(X = x_i\right);\quad p_{\cdot j} = P\left(Y = y_j\right);\quad i, j = 1,2,\ldots -$ законы распределения компонент двумерной случайной величины $\left(X, Y\right)$.

Для непрерывных случайных величин условия независимости, выраженные через плотности, имеют вид

$$f_{XY}\left(x, y\right) = f_X\left(x\right) f_Y\left(y\right) \text{ для любых } x, y\,. \quad (2)$$

Здесь $f_{XY}\left(x, y\right)-$ плотность вероятности двумерной случайной величины $\left(X, Y\right)$,

$f_X\left(x\right), f_Y\left(y\right)-$ плотности вероятности компонент двумерной случайной величины.

Не все применяемые коэффициенты связи удовлетворяют всем перечисленным свойствам. Известные в науке парные коэффициенты корреляции линейный, А.А.Чупрова, К.Пирсона и ряд других не обладают в полной мере важным свойством 3: при равенстве коэффициента нулю не гарантируется независимость случайных величин. Этим свойством обладает максимальный коэффициент корреляции О.В.Сарманова, являющийся для непрерывного симметричного двумерного распределения величиной, обратной минимальному собственному числу некоторого интегрального оператора, сложный для понимания и вычисления.

Здесь предлагаются вниманию коэффициенты детерминации, обладающие указанным свойством, лежащие в достижимых пределах от 0 до 1. Их несколько, отдельно для дискретных и непрерывных случайных величин. Перечислим их: ассоциативный, контингенциальный, комбинированные, предельный, дефектологический. Всем им даны краткие названия. Для удобства пользования и обозрения они помещены в таблицу 1 приложения (гл.12). Коэффициенты, предназначенные для дискретных случайных величин, могут быть применены для статистических исследований. В этом случае вместо вероятностей используются относительные частоты.

В основе конструкции каждого коэффициента детерминации лежит некоторая функция с определенными свойствами, называемая ядром. Это ядро усредняется с помощью двойного интеграла и плотности распределения или с помощью двойной суммы и вероятностей значений случайной величины. В конструкции коэффициента детерминации применяется модуль ядра. Если снять знак модуля, то коэффициент детерминации становится коэффициентом корреляции, но он уже не гарантирует независимость случайных величин при равенстве коэффициента нулю, так как это равенство нулю может произойти не вследствие независимости случайных величин, а за счет взаимного уничтожения положительных и отрицательных элементов под знаком суммы или интеграла. В конструкции коэффициентов детерминации для дискретных

случайных величин (кроме «комби») не используются значения случайных величин, используются только вероятности или относительные частоты, поэтому они годятся для исследования, как количественных признаков, так и для исследования качественных и смешанных признаков. Таким свойством не обладают упомянутые выше другие коэффициенты корреляции.

Заметим, что каждый коэффициент детерминации порождает спектр коэффициентов детерминации, если, кроме усреднения по вероятностям или с помощью плотности, применять степенное усреднение с помощью параметра. Эта операция степенного усреднения применена практически для всех коэффициентов детерминации.

Переход от дискретных и непрерывных случайных величин к случайным величинам общего вида требует введения интеграла Лебега-Стилтьеса. В некоторых параграфах этот переход сделан. Эти части параграфов можно пропустить при чтении без ущерба для понимания основного материала.

В главе 2, где излагается теория максимального коэффициента корреляции О.В. Сарманова, используется теория линейных однородных интегральных уравнений Фредгольма и теория Гильберта-Шмидта. Для читателей, незнакомых с этими теориями, соответствующие вопросы можно пропустить и воспользоваться только выводами и окончательными формулами.

Вообще, все известные коэффициенты для исследования связи между случайными величинами (признаками) можно разделить на несколько групп. Отметим из них 5: коэффициенты, использующие числовые значения случайных величин и вероятности этих значений, коэффициенты между качественными признаками, коэффициенты между смешанными признаками, ранговые коэффициенты корреляции, робастные (устойчивые к засорению выборки) коэффициенты. 5 ранее перечисленных коэффициентов детерминации среди указанных ― достаточно универсальны.

Корреляционный и регрессионный анализы возникли вне математики при изучении проблем естественных наук – биологии, палеонтологии, психологии, медицины, астрономии и других наук, они и применяются теперь как в этих науках, так и в статистических исследованиях практически во всех областях знания – в экономике, социологии, технике и т. д. Так, например, прежде, чем выпустить лекарство в практику, нужно убедиться в значимой связи его применения и результата лечения, в социологии важно знать величину связи между многими социальными факторами – уровнем безработицы и уровнем преступности, принадлежностью семьи к определенной социальной группе и количеством детей в семье и т. д. Корреляционный анализ, как правило, предшествует регрессионному анализу. Логика исследования такова, что сначала устанавливается значимая связь между признаками, позволяющая отобрать существенные признаки для описания явления, а затем строится функциональная зависимость между ними в среднем, то есть регрессия.

В книге рассматриваются только парные коэффициенты детерминации и корреляции, однако они могут и должны будут перенесены в область изучения связи многих случайных величин – в дальнейшем будут созданы частные и множественный коэффициенты. Большая область их применения – теория случайных процессов – детерминационная и спектральная теории. Нужно также найти асимптотические законы распределения коэффициентов. Новые результаты автора, изложенные в книге, частично опубликованы в статьях [16 – 23].

Область применения математики не имеет
границ, кроме границ самого знания.
С.Н. Бернштейн.

Глава 1. Известные коэффициенты связи между случайными величинами, используемые в учебной литературе и научных исследованиях

Прежде, чем рассматривать новые коэффициенты связи, естественно для полноты изложения привести минимальные сведения об известных коэффициентах. Будет видна разница в их свойствах и возможностях приложений. Кроме того, часть старых коэффициентов является базой для построения новых.

Старые коэффициенты связи имеют большую историю: с начала 19-го века – с Кювье и Браве, о чем говорилось во введении. Первым и основным в ряду коэффициентов является линейный коэффициент корреляции (К.Пирсона). С его рассмотрения и начинается изложение главы.

§1.1. Линейный коэффициент корреляции между двумя случайными величинами.

Название коэффициента – линейный объясняется его тесной связью с линейной регрессией. Когда нет опасности спутать его с другими коэффициентами, приставка «линейный» снимается для краткости.

Корреляционный момент двух случайных величин определяется формулой

$$K_{XY} = M\left[\left(X - m_X\right)\left(Y - m_Y\right)\right] \qquad (1)$$

Определение 1. Коэффициентом корреляции ρ_{XY} двух случайных величин

X и Y называется отношение их корреляционного момента к произведению их средних квадратических отклонений:

$$\rho_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y}. \qquad (2)$$

Очевидно, что $\rho_{XY} = \rho_{YX}$.

Корреляционный момент и коэффициент корреляции – это числовые характеристики двумерной случайной величины, причем ρ_{XY} – безразмерная числовая характеристика. Как увидим далее, они характеризуют зависимость между X и Y.

Свойства корреляционного момента и коэффициента корреляции.

Свойство 1. Коэффициент корреляции по модулю не превосходит единицы:

$$|\rho_{XY}| \le 1 \qquad (3)$$

По неравенству Коши – Буняковского :

$$|K_{XY}| = |\mathsf{M}[(X - m_X)(Y - m_Y)]| \le \sqrt{\mathsf{M}[(X - m_X)^2] \cdot \mathsf{M}[(Y - m_Y)^2]} = \sigma_X \sigma_Y .$$

Таким образом,

$$|\rho_{XY}| = \left|\frac{K_{XY}}{\sigma_X \sigma_Y}\right| \le \frac{\sigma_X \sigma_Y}{\sigma_X \sigma_Y} = 1 .$$

Свойство 2. Для независимых случайных величин X, Y корреляционный момент, а следовательно, и коэффициент корреляции равны нулю.

Известна формула $K_{XY} = \mathsf{M}[XY] - m_X m_Y$. Далее, по теореме о математическом ожидании произведения двух независимых случайных величин $\mathsf{M}[XY] = \mathsf{M}[X] \cdot \mathsf{M}[Y] = m_X m_Y$. Отсюда $K_{XY} = m_X m_Y - m_X m_Y = 0$.

Замечание 1. Обратное предложение неверно, т.е. существуют зависимые случайные величины, для которых корреляционный момент равен нулю. Из рассмотрения формулы (1) очевидно, что K_{XY} будет равен нулю всякий раз, когда распределение двумерной случайной величины (X, Y) обладает симметрией относительно какой-либо из прямых $x = m_X$ или $y = m_Y$, так как в этом случае симметричные элементы интеграла или суммы взаимно уничтожатся. Воспользуемся этим соображением для построения примера.

Пример 1. Рассмотрим две функционально зависимые случайные величины X и $Y = X^2$. Пусть X распределена нормально с $m_X = 0$ и $\sigma_X = 1$:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} .$$

Покажем, что $K_{XY} = 0$.

$$\mathsf{M}[XY] = \mathsf{M}[X^3] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^3 e^{-x^2/2} dx = 0$$ как интеграл от нечетной

функции по симметричному промежутку. Далее $K_{XY} = \mathrm{M}[XY] - m_X m_Y = 0$.

Свойство 2 и сделанное замечание позволяют ввести новое понятие.

Определение 1. Две случайные величины X, Y называются некоррелированными, если их корреляционный момент равен нулю.

Если $K_{XY} \neq 0$, то говорят, что X, Y коррелируют между собой.

Замечание 2. Из свойства 2 и замечания 1 заключаем:

1) Если случайные величины независимы, то они и некоррелированы.

2) Если случайные величины некоррелированы, то отсюда не следует, что они независимы. Построенный выше пример показывает, что существуют зависимые некоррелированные случайные величины.

Замечание 3. Если в общем случае из некоррелированности случайных величин не следует их независимость, то в частном случае, когда случайные величины X, Y образуют двумерную нормальную случайную величину (X, Y), из некоррелированности X, Y следует их независимость.

Следствие из теоремы 2. Если $K_{XY} \neq 0$, то случайные величины X, Y зависимы.

Действительно, если бы они были независимыми, то по свойству 2 $K_{XY} = 0$, что противоречит условию.

Замечание 4. Соотношение между независимыми и некоррелированными случайными величинами может быть представлено диаграммой:

независимые с.в.	зависимые с.в.
некоррелированные с.в.	коррелированные с.в.

Свойство 3. Для случайных величин X, $Y = aX + b$, связанных линейной зависимостью, коэффициент корреляции по модулю равен единице. При этом $\rho_{XY} = +1$ при $a > 0$ и $\rho_{XY} = -1$ при $a < 0$.

$$D_Y = \mathrm{M}[(Y - m_Y)^2] = \mathrm{M}[(Y - \mathrm{M}[aX + b])^2] = \mathrm{M}[(aX + b - am_X - b)^2] =$$

$$= \mathrm{M}[a^2 (X - m_X)^2] = a^2 D_X = a^2 \sigma_X^2.$$

Отсюда $\sigma_Y = |a| \sigma_X$. Далее,

$$K_{XY} = \mathrm{M}[(X - m_X)(Y - m_Y)] = \mathrm{M}[(X - m_X)(aX + b - (am_X + b))] =$$

$$= \mathrm{M}[a(X - m_X)^2] = aD_X = a\sigma_X^2$$

Тогда

$$\rho_{XY} = \frac{K_{XY}}{\sigma_X \sigma_Y} = \frac{a\sigma_X^2}{\sigma_X |a| \sigma_X} = \frac{a}{|a|} = \begin{cases} +1 & \text{при} \quad a > 0, \\ -1 & \text{при} \quad a < 0. \end{cases}$$

Рассмотрим теперь одно свойство дисперсии, которое потребуется для установления очередного свойства коэффициента корреляции.

Теорема. Если дисперсия некоторой случайной величины X равна нулю, то с вероятностью единица случайная величина постоянна:

$$P\{X = \mathrm{const}\} = 1. \qquad (4)$$

Применим неравенство Чебышёва:

$$P\left\{\left|X - m_X\right| \geq \varepsilon\right\} \leq \frac{D_X}{\varepsilon^2} \quad \text{для} \quad \forall \varepsilon > 0. \tag{5}$$

По условию $D_X = 0$, поэтому

$$P\left\{\left|X - m_X\right| \geq \varepsilon\right\} = 0 \quad \text{для} \quad \forall \varepsilon > 0 \tag{6}$$

Событие $X = m_X$ эквивалентно событию $\left|X - m_X\right| = 0$. Рассмотрим противоположное событие $\left|X - m_X\right| \neq 0$. Выберем последовательность положительных чисел $\varepsilon_1 > \varepsilon_2 > \ldots > \varepsilon_n > \ldots$, сходящуюся к нулю. Заметим, что событие $\left|X - m_X\right| \neq 0$ происходит тогда и только тогда, когда происходит одно и только одно из попарно несовместных событий $\left|X - m_X\right| > \varepsilon_1$, $\varepsilon_n \leq \left|X - m_X\right| < \varepsilon_{n-1}$, $n = 2, 3, \ldots$. Это означает, что событие $\left|X - m_X\right| \neq 0$ равно сумме указанных событий:

$$\left\{\left|X - m_X\right| \neq 0\right\} = \left\{\left|X - m_X\right| > \varepsilon_1\right\} + \sum_{n=2}^{\infty}\left\{\varepsilon_n \leq \left|X - m_X\right| < \varepsilon_{n-1}\right\}.$$

По аксиоме сложения вероятностей получаем

$$P\left\{\left|X - m_X\right| \neq 0\right\} = P\left\{\left|X - m_X\right| > \varepsilon_1\right\} + \sum_{n=2}^{\infty} P\left\{\varepsilon_n \leq \left|X - m_X\right| < \varepsilon_{n-1}\right\}.$$

Все слагаемые справа в этом равенстве равны нулю в силу (6). Тогда $P\left\{\left|X - m_X\right| \neq 0\right\} = 0$. Отсюда вероятность противоположного события равна единице, т.е. имеет место равенство (1): $P\left\{\left|X - m_X\right| = 0\right\} = P\left\{X = m_X\right\} = 1$, где роль указанной константы выполняет математическое ожидание m_X.

Пусть случайная величина X имеет математическое ожидание m_X и конечную дисперсию σ_X^2, отличную от нуля. Тогда, как известно, случайная величина

$$X' = \frac{X - m_X}{\sigma_X} \tag{7}$$

является центрированной и нормированной.

На основе свойств математического ожидания и дисперсии проверяется, что

$$\mathsf{M}[X'] = 0; \quad \mathsf{D}[X'] = 1 \tag{8}$$

Пусть $X' = \frac{X - m_X}{\sigma_X}$, $Y' = \frac{Y - m_Y}{\sigma_Y}$ – две центрированные и нормированные случайные величины. Рассмотрим их коэффициент корреляции $\rho_{X'Y'}$.

$$\rho_{X'Y'} = \frac{\mathsf{M}[(X' - m_{X'})(Y' - m_{Y'})]}{\sigma_{X'}\sigma_{Y'}} = \mathsf{M}[X'Y'] = \frac{\mathsf{M}[(X - m_X)(Y - m_Y)]}{\sigma_X\sigma_Y} = \rho_{XY}.$$

Итак,

$$\rho_{XY} = \rho_{X'Y'} = \mathsf{M}[X'Y']. \tag{9}$$

Свойство 5. Если коэффициент корреляции ρ_{XY} случайных величин X, Y по модулю равен единице: $|\rho_{XY}|=1$, то с вероятностью единица между X и Y существует линейная зависимость $Y=aX+b$, причем $a>0$ при $\rho_{XY}=1$ и $a<0$ при $\rho_{XY}=-1$:

$$P\{Y=aX+b\}=1. \qquad (10)$$

Пусть $\rho_{XY}=1$. Тогда для центрированных и нормированных случайных величин $X'=\dfrac{X-m_X}{\sigma_X}$, $Y'=\dfrac{Y-m_Y}{\sigma_Y}$ имеем по формулам (8), (9):

$$\mathsf{M}[(X'-Y')^2]=\mathsf{M}[X'^2]-2\mathsf{M}[X'Y']+\mathsf{M}[Y'^2]=\mathsf{D}[X'^2]-2\rho_{XY}+\mathsf{D}[Y']=1-2\cdot1+1=0$$

.

Далее,

$$\mathsf{M}[X'-Y']=\mathsf{M}[X']-\mathsf{M}[Y']=0;\quad \mathsf{D}[X'-Y']=\mathsf{M}[(X'-Y')^2]=0.$$

По доказанной теореме случайная величина, имеющая дисперсию и математическое ожидание, равные нулю, сама равна нулю с вероятностью единица:

$$P\{X'-Y'=\mathsf{M}[X'-Y']=0\}=1,$$

т.е. $Y'=X'$ с вероятностью единица. Это означает, что с вероятностью единица имеют место равенства

$$\frac{Y-m_Y}{\sigma_Y}=\frac{X-m_X}{\sigma_X}, \qquad Y=\frac{\sigma_Y}{\sigma_X}X+\frac{m_Y}{\sigma_Y}-\frac{m_X}{\sigma_X}=aX+b \text{,где} \qquad a=\frac{\sigma_Y}{\sigma_X}>0,$$

$$b=\frac{m_Y}{\sigma_Y}-\frac{m_X}{\sigma_X}.$$

Аналогично рассматривается случай, когда $\rho_{XY}=-1$. В этом случае нужно изучать величину $\mathsf{M}[(X'+Y')^2]$.

§ 1.2. Коэффициент корреляции между двумя событиями

Коэффициент корреляции между событиями имеет довольно длинную историю, начиная с К. Пирсона [18,19,20]. Его исследованием занимались М. Дж. Кендалл [11] и С.Н. Бернштейн [1,2]. В настоящей книге продолжено изучение его свойств. Увидим, что он выполняет роль ядра для образования коэффициентов детерминации и корреляции между случайными величинами. Многие свойства этих коэффициентов опираются на свойства коэффициента корреляции между событиями.

Приступим к изучению самого коэффициента и его свойств.

Характеристиками зависимости между двумя событиями можно считать условные вероятности $P(A/B)=P(AB)/P(B)$ и $P(B/A)=P(AB)/P(A)$. Их неудобство, однако, состоит в том, что каждая из этих характеристик

несимметрична по отношению к событиям A и B. Желательно иметь такую меру связи, сконструированную из трех вероятностей $P(AB)$, $P(A)$, $P(B)$, которая, во-первых, была бы симметричной по отношению к A и B, а во-вторых, имела бы удобные свойства, характеризующие зависимость или независимость A и B. Такой мерой связи является коэффициент корреляции между событиями A и B:

$$\rho_{AB} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}. \qquad (1)$$

свойства коэффициента корреляции.

Свойство 1. $\rho_{AB} = 0$ тогда и только тогда, когда события A и B независимы.

Следует из формулы $P(AB) = P(A)P(B)$, являющейся необходимым и достаточным условием независимости событий A и B.

Свойство 2. $-1 \le \rho_{AB} \le 1$.

Определение коэффициента корреляции между событиями по формуле (1) находится в согласии с общим определением линейного коэффициента корреляции между случайными величинами (§1.1.). Коэффициент корреляции между событиями есть коэффициент корреляции между индикаторами событий. Индикатором события называется переменная (случайная величина), определяемая по формуле

$$I_A = \begin{cases} 1, \text{ если событие } A \text{ произошло,} \\ 0, \text{ если событие } A \text{ не произошло.} \end{cases}$$

(2)

$$M[I_A] = 0 \cdot P(\overline{A}) + 1 \cdot P(A) = P(A).$$

Аналогично

$$M[I_B] = P(B);\ M[I_A^2] = P(A);\ M[I_B^2] = P(B);\ M[I_A I_B] = P(A)P(B).$$

$$D[I_A] = M[I_A^2] - M^2[I_A] = P(A) - P^2(A) = P(A)(1 - P(A)) = P(A)P(\overline{A}).$$

Аналогично $D[I_B] = P(B)P(\overline{B})$. Далее

$$K(I_A I_B) = M[I_A I_B] - M[I_A]M[I_B] =$$

$$= P(AB) - P(A)P(B).\ \rho_{AB} = \frac{K(I_A I_B)}{\sqrt{D[I_A]D[I_B]}} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(B)P(\overline{A})P(\overline{B})}}.$$

В §1.1. рассматриваемое свойство доказано в общем случае.

Свойство 3. $\rho_{AB} = 1$ тогда и только тогда, когда $P(A) = P(B) = P(AB)$, т.е. $A = B$ (совпадают) с точностью до множества элементарных событий меры нуль. (Применяется аксиоматическая схема случайных событий).

Аналогично, $\rho_{AB} = -1$ тогда и только тогда, когда $A = \overline{B}$ с точностью до множества элементарных событий меры нуль.

Пусть $P(A) = P(B) = P(AB)$. Тогда

$$\rho_{AB} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = \frac{P(A) - P^2(A)}{P(A)P(\overline{A})} = \frac{P(A)\big(1 - P(A)\big)}{P(A)P(\overline{A})} = 1.$$

Обратно, пусть $\rho_{AB} = 1$. Докажем сначала , что

$$\rho_{\overline{A}B} = -\rho_{AB} \tag{3}$$

Действительно, $B = BI = B(A + \overline{A}) = AB + \overline{A}B$. Отсюда $P(B) = P(AB) + P(\overline{A}B)$, так как AB и $\overline{A}B$ – несовместные события. Тогда $P(\overline{A}B) = P(B) - P(AB)$. Далее,

$$\rho_{\overline{A}B} = \frac{P(\overline{A}B) - P(\overline{A})P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = \frac{P(B) - P(AB) - P(\overline{A})P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} =$$

$$= \frac{P(B)\big(1 - P(\overline{A})\big) - P(AB)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = \frac{P(A)P(B) - P(AB)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = -\rho_{AB}.$$

Теперь преобразуем формулу для $\rho_{\overline{A}B}$, учитывая, что $P(\overline{A}B) = P(B/\overline{A})P(\overline{A})$:

$$\rho_{\overline{A}B} = \frac{P(B/\overline{A})P(\overline{A}) - P(\overline{A})P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = \frac{P(\overline{A})\big(P(B/\overline{A}) - P(B)\big)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}.$$

Так как $\rho_{AB} = 1$, то $\rho_{\overline{A}B} = -1$, поэтому $-\sqrt{\frac{P(A)P(B)P(\overline{B})}{P(\overline{A})}} = P(B/\overline{A}) - P(B)$.

Отсюда $P(B/\overline{A}) = P(B)\left(1 - \sqrt{\frac{P(A)P(\overline{B})}{P(\overline{A})P(B)}}\right)$. Так как $P(B/\overline{A}) \geq 0$ и $P(B) \geq 0$, то получаем неравенство $1 - \sqrt{\frac{P(A)P(\overline{B})}{P(\overline{A})P(B)}} \geq 0$, откуда $\sqrt{P(\overline{A})P(B)} \geq \sqrt{P(A)P(\overline{B})}$. Это неравенство равносильно следующим:

$\big(1 - P(A)\big)P(B) \geq P(A)\big(1 - P(B)\big)$;

$P(B) \geq P(A)$.

Меняя ролями A и B, найдем, что $P(A) \leq P(B)$, откуда $P(A) = P(B)$. Далее, учитывая этот результат, получаем $\rho_{AB} = 1 = \frac{P(AB) - P^2(A)}{P(A)P(\overline{A})}$. Отсюда

$P(A)\big(1 - P(A)\big) = P(AB) - P^2(A)$;

$P(AB) = P(A)$.

Итак, $P(A) = P(B) = P(AB)$.

Замечание 1. Из формулы (1) следует, что если $\rho_{AB} < 0$, то $P(A/B) < P(A)$;

если $\rho_{AB} > 0$, то $P(AB)/P(B) = P(A/B) > P(A)$.

Рассмотрим классический пример из области демографии и теории наследственности.

Пример 1. При переписи населения Англии и Уэльса в 1891 году получены следующие данные:

темноглазые отцы и темноглазые сыновья составляли	5,0 %;
темноглазые отцы и светлоглазые сыновья составляли	7,9 %;
светлоглазые отцы и темноглазые сыновья составляли	8,9 %;
светлоглазые отцы и светлоглазые сыновья составляли	78,2 %.

Требуется охарактеризовать зависимость между цветом глаз у отца и сына в данной группе обследованных.

Введем события: A – отец имеет темный цвет глаз, B – сын имеет темный цвет глаз. Тогда по условию задачи

$P(AB) = 0,050;\quad P(A\overline{B}) = 0,079;\quad P(\overline{A}B) = 0,089;\quad P(\overline{A}\,\overline{B}) = 0,782$.

Требуемую зависимость можно охарактеризовать четырьмя условными вероятностями $P(B/A)$, $P(B/\overline{A})$, $P(A/B)$, $P(A/\overline{B})$, а также четырьмя коэффициентами корреляции.

Так же, как формула (3), доказывается более общая формула, связывающая четыре

коэффициента корреляции

$$\rho_{AB} = \rho_{\overline{A}\,\overline{B}} = -\rho_{\overline{A}B} = -\rho_{A\overline{B}} \qquad (4)$$

Поэтому достаточно найти ρ_{AB}. Далее, так как $A = AB + A\overline{B}$; $B = AB + \overline{A}B$, то

$P(A) = P(AB) + P(A\overline{B}) = 0,050 + 0,079 = 0,129$;

$P(B) = P(AB) + P(\overline{A}B) = 0,050 + 0,089 = 0,139$.

Далее, $P(\overline{A}) = 1 - P(A) = 0,871$; $P(\overline{B}) = 0,861$. Отсюда

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{0,050}{0,129} = 0,388;$$

$$P(B/\overline{A}) = \frac{P(\overline{A}B)}{P(\overline{A})} = \frac{0,089}{0,871} = 0,102;$$

$$P(\overline{B}/\overline{A}) = \frac{P(\overline{A}\,\overline{B})}{P(\overline{A})} = \frac{0,782}{0,861} = 0,908;$$

$$P(\overline{B}/A) = \frac{P(A\overline{B})}{P(A)} = \frac{0,079}{0,129} = 0,612;$$

$$\rho_{AB} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}} = \frac{0,050 - 0,129 \cdot 0,139}{\sqrt{0,129 \cdot 0,871 \cdot 0,139 \cdot 0,861}} = 0,277.$$

Таким образом, связь между цветом глаз отца и сына характеризуется в целом модулем коэффициента корреляции, равным 0,277. Все условные вероятности

находятся в промежутке [0,102; 0,908]. Связь достаточно заметная.

§ 1.3. Коэффициент контингенции для исследования связи между двумя событиями

Наряду с коэффициентом корреляции существует еще одна мера связи между двумя событиями, восходящая к ученику К. Пирсона Д. Юлу [41, 42, 1900г], называемая коэффициентом контингенции (сопряженности, связи). Продолжим ее изучение на основе результатов Д. Юла и М. Кендалла [11, с.723]. Коэффициент контингенции далее будет применен в измененном виде как ядро более сложного коэффициента связи, названного контингенциальным коэффициентом детерминации между двумя случайными величинами. Его свойства будут опираться на свойства коэффициента контингенции.

Рассмотрим случай корреляционной таблицы 2x2. Ее можно записать в виде табл.1.

Таблица 1. Корреляционная таблица 2x2 между двумя признаками.

Признаки $X, Y \rightarrow$ $\downarrow$	B	$\overline{B}$	Σ
A	a	b	$a+b$
$\overline{A}$	c	d	$c+d$
Σ	$a+c$	$b+d$	$n=a+b+c+d$

Здесь a, b, c, d – частоты, с которыми n выборочных элементов классифицируются по признакам A и B.

Коэффициент контингенции определяется формулой [11,с. 723], [8, с. 292]

$$k_k = \frac{ad - bc}{ad + bc}$$

(1)

Для изучения свойств коэффициента контингенции запишем таблицу 1 и формулу (1) в другом виде (таблица 2).

Таблица 2. Корреляционная таблица связи между событиями 2x2.

События	B	$\overline{B}$	Σ
A	$P(AB)$	$P(A\overline{B})$	$P(A) = P(AB) + P(A\overline{B})$

$\overline{A}$	$P\left(\overline{A}B\right)$	$P\left(\overline{A}\overline{B}\right)$	$P\left(\overline{A}\right)=P\left(\overline{A}B\right)+P\left(\overline{A}\overline{B}\right)$
Σ	$P(B)=$ $=P(AB)+P\left(\overline{A}B\right)$	$P\left(\overline{B}\right)=$ $=P\left(A\overline{B}\right)+P\left(\overline{A}\overline{B}\right)$	1

На основе таблицы 2 коэффициент контингенции k_k представим в виде

$$k_k=\frac{P(AB)P\left(\overline{A}\overline{B}\right)-P\left(A\overline{B}\right)P\left(\overline{A}B\right)}{P(AB)P\left(\overline{A}\overline{B}\right)+P\left(A\overline{B}\right)P\left(\overline{A}B\right)}$$

(2)

Преобразуем числитель формулы (2). С помощью таблицы 2 находим

$A=AI=A\left(B+\overline{B}\right)=AB+A\overline{B}$. Тогда $P(A)=P(AB)+P\left(A\overline{B}\right)$. Отсюда

$P\left(A\overline{B}\right)=P(A)-P(AB)$; Аналогично $P\left(\overline{A}B\right)=P(B)-P(AB)$;

$P\left(\overline{A}\overline{B}\right)=P\left(\overline{A}\right)-P\left(\overline{A}B\right)=1-P(A)-P(B)+P(AB)$. Тогда

$P(AB)P\left(\overline{A}\overline{B}\right)-P\left(A\overline{B}\right)P\left(\overline{A}B\right)=$

$=P(AB)\left[1-P(A)-P(B)+(AB)\right]-\left[P(A)-P(AB)\right]\left[P(B)-P(AB)\right]=$

$=P(AB)-P(AB)P(A)-P(AB)P(B)+P^2(AB)-$

$-P(A)P(B)+P(AB)(A)+P(AB)P(B)-P^2(AB)$

$=P(AB)-P(A)P(B)$.

Аналогично преобразуется знаменатель формулы (2).

$P(AB)P\left(\overline{A}\overline{B}\right)+P\left(A\overline{B}\right)P\left(\overline{A}B\right)=$

$=P(AB)+2P^2(AB)-2P(AB)P(A)-2P(AB)P(B)+P(A)P(B)$.

Тогда коэффициент контингенции записывается в виде

$$k_k=\frac{P(AB)-P(A)P(B)}{P(AB)+2P^2(AB)-2P(AB)P(A)-2P(AB)P(B)+P(A)P(B)}.$$

(3)

Свойства коэффициента контингенции.

1. Нормированность: $-1\le k_k\le 1$.

Это свойство следует непосредственно из формулы (1):

$$\left|k_k\right|=\frac{|ad-bc|}{ad+bc}\le\frac{ad+bc}{ad+bc}=1.$$

2. Признак независимости событий. Для того чтобы события A, B были независимыми, необходимо и достаточно выполнение равенства $k_k = 0$.

Действительно, необходимым и достаточным условием независимости событий A, B является выполнение равенства $P(AB) - P(A)P(B) = 0$, которое выполняется тогда и только тогда, когда числитель в формуле (3) равен нулю.

3. Достижение верхней границы. Если $A \subset B$ или $B \subset A$, то $k_k = 1$.

Если $A \subset B$, то в этом случае $P(AB) = P(A);\ P(\overline{A}\overline{B}) = P(\overline{B});\ P(A\overline{B}) = 0;$

$$k_k = \frac{P(A)P(\overline{B})}{P(A)P(\overline{B})} = 1.$$

Если $B \subset A$, то $P(AB) = P(B);\ P(\overline{A}\overline{B}) = P(\overline{A});\ P(\overline{A}B) = 0;$

$$k_k = \frac{P(B)P(\overline{A})}{P(B)P(\overline{A})} = 1.$$

4. Достижение нижней границы. Если события A, B несовместны или события $\overline{A}, \overline{B}$ несовместны, то $k_k = -1$.

Действительно, если A, B несовместны, то $P(AB) = 0;\ P(A\overline{B}) = P(A);$ $P(\overline{A}B) = P(B)$.

Тогда $k_k = -\dfrac{P(A)P(B)}{P(A)P(B)} = -1$. Аналогично, если $\overline{A}, \overline{B}$ несовместны, то $P(\overline{A}\overline{B}) = 0;$

$P(A\overline{B}) = P(\overline{B});\ P(\overline{A}B) = P(\overline{A})$. Тогда $k_k = -\dfrac{P(\overline{A})P(\overline{B})}{P(\overline{A})P(\overline{B})} = -1$.

5. Если $k_k = 1$, то $B \subset A$ или $A \subset B$ с точностью до множества элементарных событий меры нуль. В частности, если имеют место оба соотношения $B \subset A;\ A \subset B$ совместно, то $A = B$ с точностью до множества элементарных событий меры нуль.

Действительно, из равенства

$$\frac{P(AB)P(\overline{A}\overline{B}) - P(A\overline{B})P(\overline{A}B)}{P(AB)P(\overline{A}\overline{B}) + P(A\overline{B})P(\overline{A}B)} = 1$$ получаем

$P(AB)P(\overline{A}\overline{B})-P(A\overline{B})P(\overline{A}B)=P(AB)P(\overline{A}\overline{B})+P(A\overline{B})P(\overline{A}B)$; Тогда $P(A\overline{B})P(\overline{A}B)=0$.

Отсюда $P(A\overline{B})=0$ или $P(A\overline{B})=0$.

Если $P(A\overline{B})=0$, то $A\subset B$ с точностью до множества элементарных событий меры нуль.

Если $P(\overline{A}B)=0$, то $B\subset A$ с точностью до множества элементарных событий меры нуль.

6. Если $k_k=-1$, то A,B несовместны или $\overline{A},\overline{B}$ несовместны с точностью до множества элементарных событий меры нуль.

Действительно, из равенства $\dfrac{P(AB)P(\overline{A}\overline{B})-P(A\overline{B})P(\overline{A}B)}{P(AB)P(\overline{A}\overline{B})+P(A\overline{B})P(\overline{A}B)}=-1$

находим

$$P(AB)P(\overline{A}\overline{B})-P(A\overline{B})P(\overline{A}B)=-P(AB)P(\overline{A}\overline{B})-P(A\overline{B})P(\overline{A}B);$$

Тогда $P(AB)P(\overline{A}\overline{B})=0$. Отсюда $P(AB)=0$ или $P(\overline{A}\overline{B})=0$.

Если $P(AB)=0$, то A,B несовместны с точностью до множества элементарных событий меры нуль. Если $P(\overline{A}\overline{B})=0$, то $\overline{A},\overline{B}$ несовместны с точностью до множества элементарных событий меры нуль.

7. Знаменатель коэффициента контингенции в ноль не обращается.

Действительно, предположим противное, что $P(AB)P(\overline{A}\overline{B})+P(A\overline{B})P(\overline{A}B)=0.$

Тогда в силу неотрицательности слагаемых оба слагаемых равны нулю. Отсюда следует, что имеет место один из случаев

$$1)\begin{cases}P(AB)=0\\P(A\overline{B})=0\end{cases}\quad 2)\begin{cases}P(AB)=0\\P(\overline{A}B)=0\end{cases}\quad 3)\begin{cases}P(\overline{A}\overline{B})=0\\P(A\overline{B})=0\end{cases}\quad 4)\begin{cases}P(\overline{A}\overline{B})=0\\P(\overline{A}B)=0\end{cases}.$$

Рассмотрим 1-й случай. Из первого равенства следует, что события A,B несовместны. Тогда событие $A\subset\overline{B}$. Из второго равенства следует, что события $A,\overline{B}$ несовместны. Это противоречит тому, что событие A влечет событие $\overline{B}$. Первый случай невозможен. Аналогично доказывается , что и остальные случаи невозможны. Итак, знаменатель коэффициента контингенции в ноль не обращается.

Замечание. Свойства 5 и 6 означают жесткую зависимость между событиями A,B.

§1.4. Коэффициент ассоциации для исследования двух качественных признаков.

Коэффициент ассоциации определяется формулой [11,с. 723], [8, с. 292]

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

(1)

Этот коэффициент так же, как и коэффициент контингенции k_k (§ 1.3) применяется для исследования связи между качественными признаками, на основе корреляционной таблицы 2×2 (таблица 1):

Таблица 1. Корреляционная таблица 2×2 между двумя признаками.

Признаки $X, Y \rightarrow$ $\downarrow$	B	$\overline{B}$	Σ
A	a	b	$a+b$
$\overline{A}$	c	d	$c+d$
Σ	$a+c$	$b+d$	$n=a+b+c+d$

Здесь a, b, c, d – частоты, с которыми n выборочных элементов классифицируются по признакам A и B.

Этот коэффициент имеет свойства, аналогичные свойствам коэффициента контингенции k_k .

Пример 1.[8, с. 291]. При социологическом обследовании 1000 жителей города было поставлено 2 вопроса: А – считаете ли вы, что ваши доходы обеспечивают основные потребности? В – удовлетворяет ли вас деятельность мэра города? Распределение ответов представлено в следующей таблице (табл.2)

Таблица2. Данные к примеру 1.

признаки	B -да	$\overline{B}$ -нет	Σ
A -да	170	80	250
$\overline{A}$ -нет	230	520	750
Σ	400	600	1000

По данным таблицы 2 имеем:

$$k_a = \frac{170 \cdot 520 - 80 \cdot 230}{\sqrt{250 \cdot 750 \cdot 400 \cdot 600}} = 0,330; \quad k_k = \frac{170 \cdot 520 - 80 \cdot 230}{170 \cdot 520 + 80 \cdot 230} = 0,655.$$

В экономической практике [29,с. 203] принято считать связь установленной, если k_a >0,5 и k_k >0,3.

В примере – связь заметная.

Простыми преобразованиями непосредственно проверяется, что модуль коэффициента ассоциации k_a равен ассоциативному коэффициенту детерминации as (§ 3.2), если вероятности событий для вычисления as мы заменим относительными частотами. Действительно, положим $n = a + b + c + d; r = \sqrt{(a+b)(b+d)(a+c)(c+d)}.$

Тогда

$$as = \frac{\left|\frac{a}{n} - \frac{a+b}{n}\frac{a+c}{n}\right|}{\frac{r}{n^2}}\frac{a}{n} + \frac{\left|\frac{b}{n} - \frac{a+b}{n}\frac{b+d}{n}\right|}{\frac{r}{n^2}}\frac{b}{n} +$$

$$+ \frac{\left|\frac{c}{n} - \frac{c+d}{n}\frac{a+c}{n}\right|}{\frac{r}{n^2}}\frac{c}{n} + \frac{\left|\frac{d}{n} - \frac{c+d}{n}\frac{b+d}{n}\right|}{\frac{r}{n^2}}\frac{d}{n} =$$

$$= \frac{a}{nr}\left|an - (a+b)(a+c)\right| + \frac{b}{nr}\left|bn - (a+b)(b+d)\right| +$$

$$+ \frac{c}{nr}\left|cn - (c+d)(a+c)\right| + \frac{d}{nr}\left|dn - (c+d)(b+d)\right| =$$

$$= \frac{\left|ad - bc\right|(a+b+c+d)}{nr} = \frac{\left|ad - bc\right|}{r}.$$

На основе полученной формулы можем утверждать, что $as = 0{,}330.$

Свойства ассоциативного коэффициента детерминации as в § 4.1 исследованы подробно. В частности они имеют место и для модуля коэффициента ассоциации $\left|k_a\right|$.

§1.5. Бисериальный коэффициент корреляции К. Пирсона для исследования смешанных признаков

Бисериальный коэффициент корреляции k_b (К. Пирсон), [11, с. 411] применяется для оценивания величины связи между качественным признаком, имеющим две градации и количественным признаком с любым числом градаций. Предполагается, что распределение по количественному признаку нормально. Коэффициент k_b определяется формулой

$$k_b = \frac{\left|\overline{y}_2 - \overline{y}_1\right|}{\sigma_y}\frac{pq}{z_k}. \quad (1)$$

Здесь $\overline{y}_1, \overline{y}_2 -$ средние в группах качественного признака по переменной y;

$\sigma_y -$ среднее квадратическое отклонение переменной y;

$p -$ доля первой группы;

$q -$ доля второй группы;

z_k находится по формулам

$$\Phi(k) = 1 - p = q;\ z_k = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}k^2\right) = \varphi(k);\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Пример1. Зависимость успеха в сдаче приемного экзамена в Лондонский университет от возраста абитуриента в 1908 – 1909 гг. (К. Пирсон), [11, с. 415]. Исходные данные о 6156 абитуриентах представлены в таблице 1.

Таблица 1. Данные к примеру 1.

Возраст абитуриента Y	16 – 18 лет среднее 17 лет	19 – 21 среднее 20лет	22 – 30 среднее 25 лет	Свыше 30 Среднее 33 года	Σ
Сдавшие, y_1	1774	383	214	40	2411
Не сдавшие, y_2	2411	814	439	81	3745
Σ	4185	1197	653	121	6156

$$\overline{y}_1 = \frac{17 \cdot 1774 + 20 \cdot 383 + 25 \cdot 214 + 33 \cdot 40}{2411} = 18,452;$$

$$\overline{y}_2 = \frac{17 \cdot 2411 + 20 \cdot 814 + 25 \cdot 439 + 33 \cdot 81}{3745} = 18,936;$$

$$\overline{y} = \frac{17 \cdot 4185 + 20 \cdot 1197 + 25 \cdot 653 + 33 \cdot 121}{6156} = 18,746$$

$$\sigma_y^2 = \frac{(17 - 18,75)^2 4185 + (20 - 18,75)^2 1197 + (25 - 18,75)^2 653 + (33 - 18,75)^2 121}{6156};$$

$$\sigma_y^2 = 10,520671; \quad \sigma_y = 3,244; \quad p = \frac{2411}{6156} = 0,392; \quad q = \frac{3745}{6156} = 0,608;$$

Решаем уравнение $F(k) = q = 0,608.$ Получаем

$$k = 0,275;\ z_k = \varphi(k) = \varphi(0,275) = 0,384.$$

$$k_b = \frac{|18,936 - 18,452| \cdot 0,392 \cdot 0,608}{3,244 \cdot 0,384} = 0.0925 \approx 0,093.$$

Оцененная корреляция между возрастом и успехом невелика.

Оценим эту корреляцию между возрастом и успехом абитуриента с помощью ассоциативного коэффициента детерминации as (§ 4.1). Для удобства вычислений умножим числитель и знаменатель в формуле (1) на $n^2 = 6156^2$. Получаем

$$as = \frac{|1774 \cdot 6156 - 2411 \cdot 4185|}{\sqrt{2411 \cdot 3745 \cdot 4185 \cdot 1971}} \cdot \frac{1774}{6156} + \frac{|383 \cdot 6156 - 2411 \cdot 1197|}{\sqrt{2411 \cdot 3745 \cdot 1197 \cdot 4959}} \frac{383}{6156} +$$

$$+ \frac{|214 \cdot 6156 - 2411 \cdot 653|}{\sqrt{2411 \cdot 3745 \cdot 653 \cdot 5503}} \frac{214}{6156} + \frac{|40 \cdot 6156 - 2411 \cdot 121|}{\sqrt{2411 \cdot 3745 \cdot 121 \cdot 6035}} \cdot \frac{40}{6156} +$$

$$+ \frac{|2411 \cdot 6156 - 3745 \cdot 4185|}{\sqrt{3745 \cdot 2411 \cdot 4185 \cdot 1971}} \cdot \frac{2411}{6156} + \frac{|814 \cdot 6156 - 3745 \cdot 1197|}{\sqrt{3745 \cdot 2411 \cdot 1197 \cdot 4959}} \cdot \frac{814}{6156} +$$

$$+ \frac{|439 \cdot 6156 - 3745 \cdot 653|}{\sqrt{3745 \cdot 2411 \cdot 653 \cdot 5503}} \cdot \frac{439}{6156} + \frac{|81 \cdot 6156 - 3745 \cdot 121|}{\sqrt{3745 \cdot 2411 \cdot 121 \cdot 6035}} \cdot \frac{81}{6156};$$

$$\rho_1 = 0,02774 + 0,01420 + 0,00157 + 0,00016 + 0,03770 + 0,00954 + 0,00322 + 0,00023.$$

$$as = 0,09436 \approx 0,094.$$

Этот результат практически совпадает с результатом для коэффициента k_b.

Для измерения величины связи между двумя качественными признаками с любым числом групп по каждому признаку применяются коэффициенты (взаимной сопряженности) k_p К.Пирсона и k_{ch} А.А.Чупрова (§ 1.6).

§1.6. Коэффициенты К.Пирсона и А.А. Чупрова для исследования качественных признаков

Коэффициент k_p К.Пирсона вычисляется по формулам

$$k_p = \sqrt{\frac{\varphi^2}{1+\varphi^2}} \qquad (1)$$

$$\varphi^2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}^2}{n_{ix} n_{jy}} - 1 \qquad (2)$$

Здесь n_{ix} – частоты по признаку X; n_{jy} – частоты по признаку Y; n_{ij} – частоты по обоим признакам (X, Y);

k_1 – число значений (групп) первого признака X;

k_2 – число значений (групп) второго признака Y;

φ^2 определяется как сумма отношений квадратов частот n_{ij} каждой клетки корреляционной таблицы к произведению итоговых частот n_{ix}, n_{jy} соответствующего столбца и строки минус единица.

Коэффициент k_{ch} А.А.Чупрова вычисляется по формуле

$$k_{ch} = \sqrt{\frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}}}$$

(3)

Здесь обозначения – те же, что и для коэффициента К.Пирсона.

Пример 1. Распределение новорожденных в ФРГ по религиозной принадлежности отца и матери в 1993 г.
(Источник: $Statistisches\ Jahrbuch\ für\ die\ BRD. - 1995, c.74.$ [8, с. 294]).
Статистические данные представлены в следующей таблице 1 (тыс.чел.).

Таблица 1. распределения новорожденных в ФРГ по религиозной принадлежности отца и матери.

Религия отца Y $\rightarrow$ Религия матери X $\downarrow$	Евангелическая	Римско-католическая	Прочие христиане	Другие религии	Неверующие и не указавшие	Σ
Евангелическая	146,1	57,6	1,1	0,5	8,8	214,1
Римско-католическая	57,3	195,9	1,1	0,7	5,2	260,2
Прочие христиане	1,3	1,4	10,5	0,1	0,3	13,6
Другие религии	1,8	2,0	0,1	62,8	1,1	67,8
Неверующие и не указавшие	29,1	16,1	0,7	0.8	77,7	124,4
Σ	235,6	273,0	13,5	64,9	93.1	680,1

Расчеты по этой таблице дают $k_p = 0,825;\quad k_{ch} = 0,7308$; [8.с. 295].

Вычислим для этого примера выборочный ассоциативный коэффициент детерминации as .

$\hat{as} = \sum_i \sum_j \frac{\left|\hat{p}_{ij} - \hat{p}_{i.}\hat{p}_{.j}\right|}{\sqrt{\hat{p}_{i.}\left(1 - \hat{p}_{i.}\right)\hat{p}_{.j}\left(1 - \hat{p}_{.j}\right)}}\hat{p}_{ij}$. Для удобства вычислений умножим

числитель и знаменатель в формуле на $n^2 = (680,1)^2$.Получаем

$\hat{as}$ =

$$\frac{\left|146,1\cdot 680,1-214,1\cdot 235,6\right|}{\sqrt{214,1\cdot 466,0\cdot 235,6\cdot 444,5}}\cdot\frac{146,1}{680,1}+\frac{\left|57,6\cdot 680,1-214,1\cdot 273,0\right|}{\sqrt{214,1\cdot 466,0\cdot 273,0\cdot 407,1}}\cdot\frac{57,6}{680,1}+$$

$$+\frac{\left|1,1\cdot 680,1-214,1\cdot 13,5\right|}{\sqrt{214,1\cdot 466,0\cdot 13,5\cdot 666,6}}\cdot\frac{1,1}{680,1}+\frac{\left|0,5\cdot 680,1-214,1\cdot 64,9\right|}{\sqrt{214,1\cdot 466,0\cdot 64,9\cdot 615,2}}\cdot\frac{0,5}{680,1}+$$

$$+\frac{\left|8,8\cdot 680,1-214,1\cdot 93,1\right|}{\sqrt{214,1\cdot 466,0\cdot 93,1\cdot 587,0}}\cdot\frac{8,8}{680,1}+\frac{\left|57,3\cdot 680,1-260,2\cdot 235,6\right|}{\sqrt{260,2\cdot 419,9\cdot 235,6\cdot 444,5}}\cdot\frac{57,3}{680,1}+$$

$$+\frac{\left|195,9\cdot 680,1-260,2\cdot 273,0\right|}{\sqrt{260,2\cdot 419,9\cdot 273,0\cdot 407,1}}\cdot\frac{195,9}{680,1}+\frac{\left|1,1\cdot 680,1-260,2\cdot 13,5\right|}{\sqrt{260,2\cdot 419,9\cdot 13,5\cdot 666,6}}\cdot\frac{1,1}{680,1}+$$

$$+\frac{\left|0,7\cdot 680,1-260,2\cdot 64,9\right|}{\sqrt{260,2\cdot 419,9\cdot 64,9\cdot 615,2}}\cdot\frac{0,7}{680,1}+\frac{\left|5,2\cdot 680,1-260,2\cdot 93,1\right|}{\sqrt{260,2\cdot 419,9\cdot 93,1\cdot 587,0}}\cdot\frac{5,2}{680,1}+$$

$$+\frac{\left|1,3\cdot 680,1-13,6\cdot 235,6\right|}{\sqrt{13,6\cdot 666,5\cdot 235,6\cdot 444,5}}\cdot\frac{1,3}{680,1}+\frac{\left|1,4\cdot 680,1-13,6\cdot 273,0\right|}{\sqrt{13,6\cdot 666,5\cdot 273,0\cdot 407,1}}\cdot\frac{1,4}{680,1}+$$

$$+\frac{\left|10,5\cdot 680,1-13,6\cdot 13,5\right|}{\sqrt{13,6\cdot 666,5\cdot 13,5\cdot 666,6}}\cdot\frac{10,5}{680,1}+\frac{\left|0,1\cdot 680,1-13,6\cdot 64,9\right|}{\sqrt{13,6\cdot 666,5\cdot 64,9\cdot 615,2}}\cdot\frac{0,1}{680,1}+$$

$$+\frac{\left|0,3\cdot 680,1-13,6\cdot 93,1\right|}{\sqrt{13,6\cdot 666,5\cdot 93,1\cdot 587,0}}\cdot\frac{0,3}{680,1}+\frac{\left|1,8\cdot 680,1-67,8\cdot 235,6\right|}{\sqrt{67,8\cdot 612,3\cdot 235,6\cdot 444,5}}\cdot\frac{1,8}{680,1}+$$

$$+\frac{\left|2,0\cdot 680,1-67,8\cdot 273,0\right|}{\sqrt{67,8\cdot 612,3\cdot 273,0\cdot 407,1}}\cdot\frac{2,0}{680,1}+\frac{\left|0,1\cdot 680,1-67,8\cdot 13,5\right|}{\sqrt{67,8\cdot 612,3\cdot 13,5\cdot 666,6}}\cdot\frac{0,1}{680,1}+$$

$$+\frac{\left|62,8\cdot 680,1-67,8\cdot 64,9\right|}{\sqrt{67,8\cdot 612,3\cdot 64,9\cdot 615,2}}\cdot\frac{62,8}{680,1}+\frac{\left|1,1\cdot 680,1-67,8\cdot 93,1\right|}{\sqrt{67,8\cdot 612,3\cdot 93,1\cdot 587,0}}\cdot\frac{1,1}{680,1}+$$

$$+\frac{\left|29,1\cdot 680,1-124,4\cdot 235,6\right|}{\sqrt{124,4\cdot 555,7\cdot 235,6\cdot 444,5}}\cdot\frac{29,1}{680,1}+\frac{\left|16,1\cdot 680,1-124,4\cdot 273,0\right|}{\sqrt{124,4\cdot 555,7\cdot 273,0\cdot 407,1}}\cdot\frac{16,1}{680,1}+$$

$$+\frac{\left|0,7\cdot 680,1-124,4\cdot 13,5\right|}{\sqrt{124,4\cdot 555,7\cdot 13,5\cdot 666,6}}\cdot\frac{0,7}{680,1}+\frac{\left|0,8\cdot 680,1-124,4\cdot 64,9\right|}{\sqrt{124,4\cdot 555,7\cdot 64,9\cdot 615,2}}\cdot\frac{0,8}{680,1}+$$

$$+\frac{\left|77,7\cdot 680,1-124,4\cdot 93,1\right|}{\sqrt{124,4\cdot 555,7\cdot 93,1\cdot 587,0}}\cdot\frac{77,7}{680,1}.$$

$\hat{as}$ =0,102812+0,015503+0,000116+0,000158+0,002444+0,017591+0,162582+
+0,000143+0,000256+0,002047+0,000144+0,000176+0,011893+0,000006+
+0,000021+0,000592+0,000742+0,000006+0,086890+0,000189+0,004786+
+0,006215+0,000050+0,000169+0,076697=0,301605+0,016050+0,174996=0,492651
.

Итак, все три показателя дают результаты $k_p = 0,825;\quad k_{ch} = 0,731;\quad \hat{as} = 0,493$, хотя и не близкие, но указывающие на значительную связь между религиозной принадлежностью новобрачных в ФРГ. Наиболее жестким показателем является $\hat{as}$.

§1.7. Коэффициент М.М. Юзбашева для исследования двух качественных признаков

Коэффициент детерминации М.М. Юзбашева $[8, c.293, 297]$определяется формулой

$$\eta^2 = \frac{\left| \sum_{i=1}^{k} n_{ii} - \frac{1}{n} \sum_{i=1}^{k} n_{i\cdot} n_{\cdot i} \right|}{n - \frac{1}{n} \sum_{i=1}^{k} n_{i\cdot} n_{\cdot i}}.$$

(1)

Здесь $n = \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}$ – сумма всех частот корреляционной таблицы;

$n_{i\cdot} = \sum_{j=1}^{k} n_{ij};\quad n_{\cdot j} = \sum_{i=1}^{k} n_{ij};\quad i, j = 1,...,k;\ k$ – число групп по каждому из двух признаков.

n_{ii} – диагональные частоты корреляционной таблицы.

Отметим свойства коэффициента η^2 .

1. $\eta^2 \le 1$, так как числитель не превосходит знаменателя.

2. $\eta^2 = 0$, если числитель равен нулю. Выясним когда это может быть. Для этого запишем формулу для η^2 в другой форме. Разделим числитель и знаменатель на n . Получаем

$$\eta^2 = \frac{\left| \sum_{i=1}^{k} \frac{n_{ii}}{n} - \sum_{i=1}^{k} \frac{n_{i\cdot}}{n} \frac{n_{\cdot i}}{n} \right|}{1 - \sum_{i=1}^{k} \frac{n_{i\cdot}}{n} \frac{n_{\cdot i}}{n}}.$$

(2)

Если $\frac{n_{ii}}{n} = \frac{n_{i\cdot}}{n} \frac{n_{\cdot i}}{n}$ для всех $i = 1,...,k$, то $\eta^2 = 0$. Этот факт означает независимость принятия значений переменными X, Y на главной диагонали корреляционной таблицы.

3. $\eta^2 = 1$, если все частоты сосредоточены на главной диагонали: $\sum_{i=1}^{k} n_{ii} = n$ (предельный случай).

4. Если $\eta^2 = 0$, то это не означает независимости случайных величин X, Y так как равенства

$\frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n} \frac{n_{\cdot j}}{n}$ могут не выполняться для всех $i, j = 1,...,k.$

Пример 1. Вычисляем коэффициент М.М. Юзбашева по данным табл. 1, §1.6.

$n = 680,1;$

$$\frac{1}{n} \sum_{i=1}^{5} n_{i\cdot} n_{\cdot i} = \frac{1}{680,1} (214,1 \cdot 235,6 + 260,2 \cdot 273 + 13,6 \cdot 13,5 +$$

$$67,8 \cdot 64,9 + 124,4 \cdot 93,1) = 202,17.$$

$$\sum_{i=1}^{5} n_{ii} = 146,1 + 195,9 + 10,5 + 62,8 + 77,7 = 493,0.$$

$$\eta^2 = \frac{493 - 202,17}{680,1 - 202,17} = 0,6085; \quad \eta = 0,780.$$

Таким образом, за счет предпочтения браков между лицами одинаковых религий на главную диагональ таблицы «собралось» 60,85% возможных родительских пар сверх равномерного распределения. Итак все способы измерения показали, что влияние религии на формирование супружеских пар в ФРГ в 1993 году было значительное.

Напомним коэффициенты связи, примененные для решения этой же задачи в предыдущем параграфе: $k_p = 0,825; \quad k_{ch} = 0,731; \quad \hat{as} = 0,493.$

Итак, все способы измерения показали, что влияние религии на формирование супружеских пар в ФРГ в 1993 году было значительное.

Глава 2. Максимальные коэффициенты корреляции О.В.Сарманова

Приводится достаточно полная теория максимальных коэффициентов корреляции.

§ 2.1. Максимальный коэффициент корреляции для непрерывных случайных величин. Симметричный случай.

1°. История вопроса. Линейный парный коэффициент корреляции ρ, как числовая характеристика связи случайных величин, и ранее со времен К. Пирсона (1857 – 1936) и теперь занимает преимущественное, по сути даже монопольное положение, в теории вероятностей, теории случайных процессов, математической статистике. Первый прорыв монополии произошел в сороковые годы прошлого, 20-го века, когда Ленинградский математик О.В. Сарманов (1916 – 1977) в своей докторской диссертации (1948) построил теорию максимального коэффициента корреляции. Этот коэффициент отличается от обычного линейного коэффициента корреляции тем, что обращается в нуль тогда и только тогда, когда случайные величины X, Y независимы. Теория сложная, основана на теории интегральных уравнений и функциональном анализе. К сожалению, она не получила широкого развития и не попала в учебники, видимо, из-за своей относительной сложности и отсутствия в то время эффективных электронных вычислительных средств. Все началось с решения задачи выпрямления криволинейной регрессии, которую в конце тридцатых (довоенных) годов прошлого века поставил академик Сергей Натанович Бернштейн (1880 – 1968). О.В. Сарманов был его учеником. Решив задачу, он, как побочный продукт, создал теорию максимального коэффициента корреляции. Об этой теории и пойдет речь в этом параграфе.

2°. Постановка задачи. По определению теоретическое уравнение регрессии y на x есть уравнение вида $y = M_x Y = m_Y(x)$, где $M_x Y -$ условное математическое ожидание случайной величины Y при заданном значении x случайной величины X. В непрерывном случае оно задается формулой

$$M_x Y = \int_{-\infty}^{+\infty} y f(y/x) dy = \int_{-\infty}^{+\infty} y \frac{f_{XY}(x,y)}{f_X(x)} dy = m_Y(x).$$

(1)

Аналогично

$$M_y X = \int_{-\infty}^{+\infty} x f(x/y)\,dx = \int_{-\infty}^{+\infty} x \frac{f_{XY}(x,y)}{f_X(x)}\,dx = m_X(y) -$$

(2)

условное математическое ожидание X при заданном значении y случайной величины Y.Теоретическое уравнение регрессии x на y записывается в виде $x = m_X(y)$. В общем случае эти уравнения – нелинейные.

Здесь

$$f(y/x) = \frac{f_{XY}(x,y)}{f_X(x)} -$$

(3)

условная плотность случайной величины Y при заданном значении x случайной

величины X. Аналогично

$$f(x/y) = \frac{f_{XY}(x,y)}{f_Y(y)} -$$

(4)

условная плотность X при заданном значении y случайной величины Y.

Регрессия считается линейной тогда и только тогда, когда обе функции регрессии являются линейными. Известно, что в случае двумерного нормального закона регрессия линейна.

Зависимость между случайными величинами в случае линейности регрессий называется линейной корреляцией. Термин «Регрессия» обычно понимается двояко. С одной стороны –это вероятностная зависимость, выраженная уравнением регрессии, с другой стороны – это само уравнение регрессии.

Задача, поставленная академиком С.Н. Бернштейном, формулируется следующим образом.

В случае нелинейной регрессии между X, Y требуется найти все функции

$$\xi = \varphi(X); \eta = \psi(Y)$$

(5)

такие, что между случайными величинами ξ, η регрессия прямолинейна.

Частная задача состоит в отыскании дифференцируемых строго монотонных функций

$$\xi = \varphi(x), \eta = \psi(y).$$

(6)

Не ограничивая общности можно ограничиться строго возрастающими функциями.

В этом случае существуют однозначные, строго возрастающие, дифференцируемые обратные функции

$$x = \varphi^{-1}(\xi);\ y = \psi^{-1}(\eta).$$

(7)

3°. Решение задачи в симметричном непрерывном случае. В симметричном случае двумерная плотность симметрична: $f_{XY}(x,y) = f_{XY}(y,x)$. Тогда по симметрии $f_X(x)$ и $f_Y(y)$ – одинаковые функции: $f_X(x) = f_Y(x) = f(x)$. Будем также обозначать $f_{XY}(x,y) = f(x,y)$.

В этом симметричном случае функции, решающие задачу выпрямления регрессии, также одинаковы:

$$\varphi(x) = \psi(x).$$

(8)

Задача поиска функции φ сводится к решению так называемого корреляционного интегрального уравнения:

$$\varphi(x) = \lambda \int_a^b \varphi(y) \frac{f(x,y)}{f(x)} dy.$$

(9)

Пределы интеграла a, b, в частности, могут быть и бесконечными.

Здесь λ – числовой параметр, которому далее будет дано определенное толкование.

Это уравнение (9) выражает факт, означающий, что случайная величина $\varphi(Y)$ имеет условное математическое ожидание $\frac{1}{\lambda}\varphi(x)$. Таким образом, функцией регрессии будет $\frac{1}{\lambda}\varphi(x)$. По симметрии случайная величина $\varphi(X)$ имеет условное математическое ожидание $\frac{1}{\lambda}\varphi(y)$. Если положить

$$\eta = \varphi(Y);\quad \xi = \varphi(X),$$

(10)

то получаем, что

$$M_\xi \eta = \frac{1}{\lambda}\xi;\quad M_\eta \xi = \frac{1}{\lambda}\eta.$$

(11)

Это означает, что между новыми случайными величинами ξ,η корреляция линейная, так как регрессионные уравнения (11) – линейные.

Заметим, что случайные величины $\varphi(X)$ и $\varphi(Y)$ всегда можно нормировать.

Пусть среднее квадратическое отклонение случайной величины $\varphi(X)$ равно σ. Перейдем в уравнении (7) к функциям $\tilde{\varphi}(x)=\dfrac{\varphi(x)}{\sigma}$,разделив обе части уравнения (9) на σ. Получаем

$$\tilde{\varphi}(x)=\lambda\int_a^b\tilde{\varphi}(y)\frac{f(x,y)}{f(x)}dy.$$

Чтобы не осложнять записей, нормированные случайные величины снова будем обозначать

$\varphi(X),\varphi(Y)$. Как известно, при нормировании случайных величин их коэффициент корреляции не меняется.

По традиции в теории интегральных уравнений параметр λ ставится перед интегралом, а не перед функцией φ слева в уравнении (9), как это принято в функциональном анализе. Там функциональное уравнение записывается в виде: $A\varphi=\lambda\varphi$. В нашем случае под оператором A понимается линейный интегральный оператор, выраженный интегралом в формуле (9) справа. В теории интегральных уравнений разработан метод симметризации некоторых интегральных уравнений [30,37]. В нашем случае симметричного распределения уравнение (9) симметризуется. Запишем его в виде

$$\varphi(x)\sqrt{f(x)}=\lambda\int_a^b\varphi(y)\sqrt{f(y)}\frac{f(x,y)}{\sqrt{f(x)}\sqrt{f(y)}}dy.$$

(12)

Полагаем

$$\omega(x)=\varphi(x)\sqrt{f(x)};\quad K(x,y)=\frac{f(x,y)}{\sqrt{f(x)f(y)}}.$$

(13)

Тогда уравнение (12) можно записать в виде

$$\omega(x)=\lambda\int_a^b\omega(y)K(x,y)dy.$$

(14)

Получили уравнение с симметрическим ядром. По классификации – это однородное уравнение Фредгольма 2-го рода (Фредгольм Эрих Ивар, 1866 – 1927, Стокгольмский университет).

Из теории Гильберта-Шмидта (Гильберт Давид, 1862 – 1943; Шмидт Эрхард, 1876 – 1959) интегральных уравнений с симметрическим вещественным ядром известны следующие результаты.

Ищутся ненулевые решения, выражающиеся через собственные функции. Предполагается, что

$$\int_a^b \int_a^b K^2(x,y)dxdy = k^2 < +\infty.$$

(15)

Решения ищутся в классе L^2 функций с интегрируемым квадратом. Уравнение имеет спектр характеристических (собственных) чисел и собственных функций:

$$|\lambda_0| \le |\lambda_1| \le |\lambda_2| \le ... \le |\lambda_n| \le ...$$

(16)

$$\omega_0(x), \omega_1(x), \omega_2(x), ..., \omega_n(x), ...$$

(17)

Каждое собственное число считается столько раз, какова его кратность (кратность конечна).

$$|\lambda_n| \xrightarrow[n\to\infty]{} \infty$$

(18)

Собственные функции $\omega_k(x)$ ортогональны:

$$\int_a^b \omega_k(x)\omega_n(x)dx = 0; \quad k \ne n.$$

(19)

Можно считать собственные функции ортонормальными.

Ядро разлагается в билинейный ряд Фурье

$$K(x,y) = \sum_{i=0}^{\infty} \frac{\omega_i(x)\omega_i(y)}{\lambda_i},$$

(20)

сходящийся к $K(x,y)$ в среднем:

$$\int_a^b \int_a^b \left[K(x,y) - \sum_{i=0}^{n} \frac{\omega_i(x)\omega_i(y)}{\lambda_i} \right]^2 \xrightarrow[n\to\infty]{} 0.$$

(21)

Известно, что в линейном уравнении регрессии y на x, записанном в виде $y = ax + b$, угловой коэффициент имеет вид $a = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$. Если случайные

величины X,Y нормированы, то есть $\sigma_X = \sigma_Y = 1$, то угловой коэффициент регрессии – коэффициент регрессии равен коэффициенту корреляции $a = \rho_{XY}$. В нашем случае нормирования собственных функций это будет иметь место.

Пересчитаем условную плотность $\frac{f(x,y)}{f(x)}$ к новым переменным $\xi = \varphi(x), \eta = \varphi(y)$.

$$\frac{f(x,y)}{f(x)} = \frac{f\left[\varphi^{-1}(\xi),\varphi^{-1}(\eta)\right]\left[\varphi^{-1}(\xi)\right]'_{\xi}\left[\varphi^{-1}(\eta)\right]'_{\eta}}{f\left[\varphi^{-1}(\xi)\right]\left[\varphi^{-1}(\xi)\right]'_{\xi}} =$$

$$\frac{f\left[\varphi^{-1}(\xi),\varphi^{-1}(\eta)\right]}{f\left[\varphi^{-1}(\xi)\right]}\left[\varphi^{-1}(\eta)\right]'_{\eta}.$$

Будем считать случайные величины ξ,η нормированными. Перейдем от переменных x, y к переменным ξ,η по формулам

$$\xi = \varphi(x), \quad \eta = \varphi(y),$$

где $\varphi(x)$ - собственная функция рассматриваемого ядра, и подставим эти новые переменные в интегральное уравнение (9). Получаем

$$\frac{1}{\lambda}\xi = \int_{a_1}^{b_1} \eta \frac{f\left[\varphi^{-1}(\xi),\varphi^{-1}(\eta)\right]}{f\left[\varphi^{-1}(\xi)\right]}\left[\varphi^{-1}(\eta)\right]'_{\eta} d\eta.$$

(22)

Это уравнение регрессии η на ξ – линейное. Запишем его в виде

$$\eta = \frac{1}{\lambda}\xi$$

Здесь ξ,η – неслучайные переменные. Угловой коэффициент в линейном уравнении регрессии при нормированных случайных величинах есть коэффициент корреляции
между ними, поэтому коэффициент корреляции между случайными величинами ξ,η равен

$$\rho = \frac{1}{\lambda}.$$

(23)

Покажем, что уравнение (9) имеет собственное число $\lambda_0 = 1$ и собственную функцию $\varphi_0(x) \equiv 1$. Действительно, подставим их в (9). Получаем $1 = 1 \cdot \int_a^b 1 \cdot \frac{f(x,y)}{f(x)} dy.$

Это равенство верно, так как является нормировкой условной плотности. Тогда спектры собственных чисел и собственных функций уравнения (12), (14) имеют вид:

$$\lambda_0 = 1, |\lambda_1| \le |\lambda_2| \le \ldots$$

(24)

$$1 \cdot \sqrt{f(x)}, \varphi_1(x)\sqrt{f(x)}, \varphi_2(x)\sqrt{f(x)}, \ldots$$

(25)

Определение 1. Максимальным коэффициентом корреляции случайных величин X, Y называется число, обратное наименьшему по модулю характеристическому числу λ_1 корреляционного интегрального уравнения (9), или, что то же, уравнения (12):

$$\rho^* = \frac{1}{\lambda_1}.$$

(26)

Роль максимального коэффициента корреляции в характеристике связи случайных величин определяется следующей теоремой.

Теорема 1. Для независимости случайных величин X, Y необходимо и достаточно обращение в нуль их максимального коэффициента корреляции:

$$\rho^* = 0.$$

(27)

Необходимость.

Пусть X, Y независимы. Тогда $\varphi_1(X) = \xi, \varphi_1(Y) = \eta$ также независимы. Линейный коэффициент корреляции между ними равен нулю по независимости ξ, η. С другой стороны он определяется по формуле (23). В силу формулы (26) и по определению он есть максимальный коэффициент корреляции между X, Y.

Достаточность.

Пусть $\rho^* = \frac{1}{\lambda_1} = 0$. Отсюда следует, что $\lambda_1 = \infty$. Тогда в спектре характеристических чисел ядра $K(x,y) = \frac{f(x,y)}{\sqrt{f(x)f(y)}}$ остается только $\lambda_0 = 1$, а в спектре собственных функций – только функция $\omega_0(x) = 1 \cdot \sqrt{f(x)}$. В разложении (20) ядра в билинейный ряд по собственным функциям будет только одно слагаемое $\frac{\omega_0(x)\omega_0(y)}{\lambda_0} = \sqrt{f(x)}\sqrt{f(y)}$.

Запишем условие сходимости (21) билинейного ряда к ядру в среднем квадратическом:

$\lim\limits_{n\to\infty} \int\limits_a^b \int\limits_a^b \left[K(x,y) - \sqrt{f(x)}\sqrt{f(y)} \right]^2 dxdy = 0$. Отсюда следует $K(x,y) = \sqrt{f(x)}\sqrt{f(y)}$ почти везде. Подробнее: $\frac{f(x,y)}{\sqrt{f(x)}\sqrt{f(y)}} = \sqrt{f(x)}\sqrt{f(y)}$. Тогда окончательно находим $f(x,y) = f(x)f(y)$ почти везде. Это и означает независимость случайных величин X, Y. Доказательство закончено.

Теорема 2. Если корреляция между случайными величинами X, Y – линейная, то обычный линейный коэффициент корреляции ρ между X, Y совпадает с максимальным коэффициентом корреляции ρ^*. В частности, если (X,Y) – двумерная нормальная случайная величина, то корреляция между X, Y – линейная, а потому и здесь линейный и максимальный коэффициенты совпадают.

Доказательство. В случае линейной корреляции ядро $K(x,y)$ в качестве первой собственной функции имеет линейную функцию (и, следовательно, строго монотонную функцию) $(x-m)/\sigma$, где m – математическое ожидание, а σ^2 – дисперсия X. Если в спектре ядра корреляционного интегрального уравнения имеется строго монотонная функция, то она принадлежит всегда первому собственному числу [24,33], поэтому линейный коэффициент корреляции между X, Y, равный коэффициенту корреляции между функциями $(X-m)/\sigma$ и $(Y-m)/\sigma$, совпадает с максимальным коэффициентом корреляции.

Замечание 1. Название «Максимальный коэффициент корреляции» объясняется двояко. С одной стороны, $\rho^* = \frac{1}{\lambda_1}$ является наибольшим из возможных отношений $\frac{1}{\lambda_n}; n = 1, 2, \ldots$

С другой стороны, в теории интегральных уравнений известно экстремальное свойство собственных чисел [30]. Максимум модуля функционала

$$I = \int_a^b \int_a^b \varphi(x)\varphi(y) f(x,y)\, dxdy$$

(28)

при условиях

$$\int_a^b \varphi(x) f(x)\, dx = 0; \quad \int_a^b \varphi^2(x) f(x)\, dx = 1$$

(29)

достигается при $\varphi(x) = \varphi_1(x)$ и равен $1/\lambda_1$.

3°. Обзор методов решения интегрального уравнения и вычисления его собственных чисел.

Вычисление максимального коэффициента ρ^* – задача гораздо более сложная, чем вычисление линейного коэффициента корреляции ρ или коэффициентов детерминации, рассматриваемых в этой книге. Дело в том, что максимальный коэффициент корреляции задан неявно интегральным уравнением, которое нужно решать, выбрав соответствующий алгоритм. Нужно решать также задачу поиска минимального по модулю собственного числа ядра уравнения. Линейный же коэффициент корреляции и рассматриваемые в книге коэффициенты детерминации определены явно через интеграл. Способов решения интегральных уравнений разработано много [10], [30], [37]. Многие из них пригодны и для решения задачи приближенного или точного вычисления максимального коэффициента корреляции.

1) Случай вырожденного ядра

$$K(x,t) = \sum_{i=1}^{n} \sigma_i(x) q_i(t).$$

(30)

В этом случае решение интегрального уравнения сводится к решению системы n линейных алгебраических уравнений с n неизвестными, а характеристическое уравнение для нахождения характеристических чисел будет алгебраическим степени n.

2) Замена ядра общего вида близким к нему вырожденным ядром. В этом случае решения уравнения и характеристические числа находятся приближенно,

[10]. В качестве вырожденного ядра может быть взят, например, отрезок ряда Тейлора.

3) Метод конечных сумм. В этом случае интеграл в интегральном уравнении заменяется конечной суммой с помощью какой-либо квадратурной формулы, например, Симпсона. Интегральное уравнение заменяется квадратной системой алгебраических уравнений.

4) Метод моментов. В этом случае решение ищется в виде линейной комбинации линейно независимых опорных функций. Коэффициенты линейной комбинации находятся из условия ортогональности невязки к опорным функциям линейной комбинации/

5) Метод последовательных приближений для нахождения максимального коэффициента корреляции. Метод изложен в работе [31,33] О.В. Сарманова.

Суть его следующая. В качестве нулевого приближения берем любую функцию $r_0(x)$, такую, что случайная величина $r_0(X)$ имеет конечную дисперсию. Если первый момент

$C_0 = \int_a^b r_0(x) f(x)\,dx \neq 0$, то $r_0(x)$ удобнее заменить на $r_0(x) - C_0$, поэтому, не ограничивая общности, будем считать $C_0 = 0$. Положим

$r_k(x) = \int_a^b r_{k-1}(y) \frac{f(x,y)}{f(x)}\,dy; \quad k = 1,2,...$ Тогда

$$C_1 \varphi_1(x) = \lim_{k\to\infty} r_k(x) \lambda_1^k;$$

(31)

$$C_1 = \int_a^b \varphi_1(x) r_0(x) f(x)\,dx.$$

(32)

Если k достаточно велико, то

$$\rho^* = \frac{1}{\lambda_1} \approx \frac{r_k(x)}{r_{k-1}(x)}.$$

(33)

Действительно, из (31) находим

$C_1 \varphi_1(x) \approx r_{k-1}(x) \lambda_1^{k-1}; \quad C_1 \varphi_1(x) \approx r_k(x) \lambda_1^k$. Делим первое равенство на второе.

Получаем равенство (33).

4°. Примеры.

Пример 1. Вычисление максимального коэффициента корреляции для двумерного непрерывного распределения в квадрате $D=\{(x;y):0\le x\le 1;0\le y\le 1\}$, заданного симметричной плотностью

$$f(x,y)=\begin{cases} x+y; & (x;y)\in D \\ 0;(x;y)\notin D \end{cases}.$$

В этом случае $f(x)=\int_0^1 (x+y)dy=x+\frac{1}{2};\quad f(y)=y+\frac{1}{2}.$

Интегральное уравнение (9) в этом случае имеет вид

$\varphi(x)=\lambda\int_0^1 \varphi(y)\frac{x+y}{x+\frac{1}{2}}dy.$ Это – уравнение с вырожденным ядром. Запишем его иначе

$$\varphi(x)=\lambda\left[\frac{x}{x+\frac{1}{2}}\int_0^1\varphi(y)dy+\frac{1}{x+\frac{1}{2}}\int_0^1 y\varphi(y)dy\right].$$ Положим

$s_1=\int_0^1\varphi(y)dy;\quad s_2=\int_0^1 y\varphi(y)dy.$ Тогда приходим к равенству

$$\varphi(x)=\lambda\left(\frac{x}{x+\frac{1}{2}}s_1+\frac{1}{x+\frac{1}{2}}s_2\right).$$

(34)

Интегрируем это равенство по x в пределах от 0 до 1, а затем умножаем на x и снова интегрируем по x в тех же пределах. Получаем систему алгебраических уравнений

$$\begin{cases} s_1=\lambda s_1\int_0^1\frac{x}{x+0,5}dx+\lambda s_2\int_0^1\frac{dx}{x+0,5}; \\ s_2=\lambda s_1\int_0^1\frac{x^2}{x+0,5}dx+\lambda s_2\int_0^1\frac{x}{x+0,5}dx. \end{cases}$$

Вычислим полученные интегралы: $a_{11}=a_{22}=\int_0^1\frac{x}{x+0,5}dx=1-\frac{1}{2}\ln 3;$

$a_{12}=\int\limits_0^1\frac{dx}{x+0,5}=\ln 3;\quad a_{21}=\int\limits_0^1\frac{x^2}{x+0,5}dx=\frac{1}{4}\ln 3.$ Запишем ранее полученную систему в общем виде:

$\begin{cases}s_1=\lambda s_1a_{11}+\lambda s_2a_{12}\\ s_2=\lambda s_1a_{21}+\lambda s_2a_{22}\end{cases}$. Далее получаем $\begin{cases}s_1=\lambda s_1\left(1-\frac{1}{2}\ln 3\right)+\lambda s_2\ln 3\\ s_2=\lambda s_1\left(\frac{1}{4}\ln 3\right)+\lambda s_2\left(1-\frac{1}{2}\ln 3\right)\end{cases}$.

Запишем эту однородную систему в стандартной форме.

$$\begin{cases}s_1\left[\lambda\left(1-\frac{1}{2}\ln 3\right)-1\right]+\lambda s_2\ln 3=0\\ s_1\lambda\frac{1}{4}\ln 3+s_2\left[\lambda\left(1-\frac{1}{2}\ln 3\right)-1\right]=0\end{cases}.$$

(35)

Сами решения нам не нужны, а нужны только собственные числа матрицы коэффициентов. Вычислим определитель системы (35).

$$\begin{vmatrix}\lambda\left(1-\frac{1}{2}\ln 3\right)-1 & \lambda\ln 3\\ \frac{1}{4}\lambda\ln 3 & \lambda\left(1-\frac{1}{2}\ln 3\right)-1\end{vmatrix}=\left[\lambda\left(1-\frac{1}{2}\ln 3\right)-1\right]^2-\frac{1}{4}\lambda^2\left(\ln 3\right)^2=$$

$=\lambda^2\left(1-\ln 3\right)-2\lambda\left(1-\frac{1}{2}\ln 3\right)+1.$ Приравнивая определитель к нулю, получаем характеристическое уравнение:

$$\lambda^2\left(1-\ln 3\right)-2\lambda\left(1-\frac{1}{2}\ln 3\right)+1=0.$$

(36)

Его корни равны $\lambda_0=1;\quad \lambda_1=\frac{1}{1-\ln 3}$. Согласно построенной теории

$$\rho^*=\frac{1}{\lambda_1}=1-\ln 3\approx -0,0986.$$

(37)

Пример 2. Вычислим максимальный коэффициент корреляции по формуле (33).для того же распределения, что и в примере 1.

Возьмем $\overline{r}_0\left(x\right)=x$. Тогда $C_0=\int\limits_0^1\overline{r}_0\left(x\right)f\left(x\right)dx=\int\limits_0^1x\left(x+\frac{1}{2}\right)dx=\frac{7}{12}.$

В качестве нулевого приближения возьмем $r_0(x) = x - \frac{7}{12}$. Тогда

$$r_1(x) = \int_0^1 r_0(y)\frac{f(x,y)}{f(x)}dy = \int_0^1\left(y - \frac{7}{12}\right)\frac{x+y}{x+\frac{1}{2}}dy = -\frac{1}{12}\frac{2x-1}{2x+1};$$

$$r_2(x) = \int_0^1 r_1(y)\frac{f(x,y)}{f(x)}dy = -\frac{1}{12}\int_0^1\frac{2y-1}{2y+1}\frac{x+y}{x+\frac{1}{2}}dy = \frac{\ln 3-1}{12}\frac{2x-1}{2x+1};$$

$$r_3(x) = \int_0^1 r_2(y)\frac{f(x,y)}{f(x)}dy = \frac{\ln 3-1}{12}\int_0^1\frac{2y-1}{2y+1}\frac{x+y}{x+\frac{1}{2}}dy = -\frac{(\ln 3-1)^2}{12}\frac{2x-1}{2x+1};$$

По индукции находим

$r_k(x) = (-1)^k\frac{(\ln 3-1)^{k-1}}{12}\frac{2x-1}{2x+1}$. Далее получаем

$\frac{r_k(x)}{r_{k-1}(x)} = -(\ln 3-1) = 1-\ln 3 \approx -0,0986.$ Отсюда

$$\rho^* = \lim_{k\to\infty}\frac{r_k(x)}{r_{k-1}(x)} = 1-\ln 3 \approx -0,0986.$$

этот результат совпадает с тем, что получен предыдущим методом.

Пример 3. Для распределения из примера 1 вычислим линейный коэффициент корреляции ρ для сравнения с максимальным коэффициентом корреляции ρ^*.

$\rho = \frac{K_{XY}}{\sigma_X\sigma_Y}$. Последовательно вычисляем:

$$m_X = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \frac{7}{12};\quad m_Y = \frac{7}{12};\ \alpha_2 = MX^2 = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \frac{5}{12};$$

$$D_X = \alpha_2 - m_X^2 = \frac{5}{12} - \frac{49}{144} = \frac{11}{144};\ D_Y = \frac{11}{144};$$

$$\alpha_{11} = M[XY] = \int_0^1\int_0^1 xy(x+y)dxdy = \frac{1}{3};$$

$$K_{XY} = \alpha_{11} - m_X m_Y = \frac{1}{3} - \frac{49}{144} = -\frac{1}{144};$$

$\rho = \frac{-1/144}{11/144} = -\frac{1}{11}$. Итак, $\rho = -\frac{1}{11} \approx -0,0909$. Получили значение близкое к ρ^*.

Для этого же распределения приведем также значение дефектологического коэффициента детерминации, вычисленное далее в §8.3: $def = \frac{1}{24} \approx 0,0417.$

Сводка коэффициентов связи для примера 1:

$$\rho^* = -0,0986; \quad \rho = -0,0909; \quad def = 0,0417.$$

§ 2.2. Максимальный коэффициент корреляции для дискретных случайных величин. Симметричный случай.

Для дискретных случайных величин максимальный коэффициент корреляции определяется аналогично его определению в непрерывном случае.

Пусть вероятностная зависимость между дискретными случайными величинами X, Y определяется симметричной матрицей вероятностей

$$\left(p_{ij}\right), \quad i, j = 1,2,...$$

(1)

Здесь

$$0 \le p_{ij} = p_{ji} = P\left(X = x_i, Y = y_j\right); \qquad \sum_i \sum_j p_{ij} = 1.$$

(2)

$$p_i = \sum_{j=1}^{n} p_{ij} = P\left(X = x_i\right) = P\left(Y = y_i\right); \quad i = 1,2,...$$

(3)

Образуем матрицу

$$\mathbf{P} = \left(\frac{p_{ij}}{\sqrt{p_i p_j}} \right).$$

(4)

Пусть $\mu_1 = \lambda_1^{-1} -$ первое (наибольшее) собственное число матрицы **P**.

Определение. Максимальным коэффициентом корреляции между случайными величинами X, Y называется число

$$\rho^* = \frac{1}{\lambda_1} = \mu_1 \quad .$$

(5)

Максимальный коэффициент корреляции в дискретном случае обладает теми же свойствами, что и в непрерывном. Сначала докажем лемму С.Н. Бернштейна [1, с. 330].

Лемма.

Если регрессия y на x линейна, то есть $M_x Y = ax + b$, то имеет место формула
для углового коэффициента регрессии

$$a = \rho \frac{\sigma_Y}{\sigma_X}$$

(6)

Здесь $\rho = \rho_{XY}$ – линейный коэффициент корреляции между случайными величинами X, Y, σ_X, σ_Y – средние квадратические отклонения X, Y.

Доказательство.

Без ограничения общности будем считать, что случайные величины X, Y центрированы, то есть $MX = MY = 0$. При этой операции линейный коэффициент корреляции ρ не меняется. Пусть

$$M_x Y = \varphi(x)$$

(7)

есть условное математическое ожидание y на x, то есть функция регрессии. Пусть далее

$\psi(x)$ – произвольная однозначная функция; $p_i = P(X = x_i)$; $i = 1, 2, \ldots$

Условную вероятность $P(Y = y_k / X = x_i)$ обозначим для краткости символом $p_{k/i}$:

$$P(Y = y_k / X = x_i) = p_{k/i}$$

(8)

Тогда

$$\varphi(x_i) = M_{x_i} Y = \sum_k p_{k/i} y_k .$$

(9)

$$M[\psi(X) \cdot Y] = \sum_i \sum_k P(X = x_i) P(Y = y_k / X = x_i) \psi(x_i) \cdot y_k =$$

$$\sum_i \sum_k p_i p_{k/i} \psi(x_i) y_k .$$

Далее суммируем по i и по k раздельно. Получаем

$$M[\psi(X) \cdot Y] = \sum_i p_i \psi(x_i) \sum_k p_{k/i} y_k = \sum_i p_i \psi(x_i) \cdot \varphi(x_i) =$$

$$M[\psi(X) \varphi(X)].$$

Итак,

$$M[\psi(X) \cdot Y] = M[\psi(X) \varphi(X)].$$

(10)

Полагаем, в частности, в формуле (10) $\psi(x)=1$. Тогда имеем

$$M[Y]=M[\varphi(X)]=0,$$

(11)

так как $Y-$центрированная случайная величина. Пусть теперь $\varphi(x)=ax+b-$регрессия линейная. Тогда

$M[\varphi(X)]=M[aX+b]=aM[X]+b=0+b=b=0,$ так как $M[X]=0$ и $M[\varphi(x)]=0$ по (11). Итак, $b=0$. Таким образом $\varphi(x)=ax.$

Положим теперь в формуле (10) $\varphi(x)=ax;\quad \psi(x)=x$. Тогда получаем

$M[XY]=K_{XY}=\rho_{XY}\sigma_X\sigma_Y=M[XaX]=aM[X^2]=a\sigma_X^2.$ Отсюда заключаем, что

$\rho_{XY}\sigma_X\sigma_Y=a\sigma_X^2.$ далее $a=\rho_{XY}\dfrac{\sigma_Y}{\sigma_X}=\rho\dfrac{\sigma_Y}{\sigma_X}.$ Лемма доказана.

Теорема 1. Для того, чтобы дискретные случайные величины X,Y были независимыми, необходимо и достаточно обращение в нуль их максимального коэффициента корреляции.

Доказательство.

Необходимость. Пусть случайные величины X,Y независимы. Тогда, как известно, их линейный коэффициент корреляции равен нулю: $\rho=0.$ Рассмотрим теперь условное математическое ожидание случайной величины Y при заданном значении x_i случайной величины X:

$M_{x_i}Y=\sum_j p_{k/i}y_k=\sum_j y_j\dfrac{p_{ij}}{p_i},$ здесь $p_{ij}=P(X=x_i,Y=y_j)$, и перейдем к системе алгебраических уравнений

$$\sum_{i=1}^{n} y_j\frac{p_{ij}}{p_i}=\mu x_i;\quad i,j=1,2,...,n.$$

(12)

Заметим, что в силу симметрии $y_j=x_j;\quad j=1,2,...,n.$

Система – квадратная. Симметризуем ее:

$$\sum_{j=1}^{n} y_j\sqrt{p_j}\frac{p_{ij}}{\sqrt{p_ip_j}}=\mu x_i\sqrt{p_i}$$

(13)

и введем новые величины, положив $x_i\sqrt{p_i}=\xi_i;\quad i=1,2,...,n.$ Тогда получим

$$\sum_{j=1}^{n} \xi_j \frac{p_{ij}}{\sqrt{p_i p_j}} = \mu \xi_i; \quad i = 1,2,...,n.$$

(14)

Здесь μ можем рассматривать как собственное число линейного оператора A, представленного симметричной матрицей $\left(\frac{p_{ij}}{\sqrt{p_i p_j}} \right)$; $i, j = 1,2,...,n.$

С другой стороны система равенств (12) представляет собой линейную регрессию дискретных случайных величин X, Y в симметричном случае. Угловой коэффициент линейной регрессии a с одной стороны по лемме равен $a = \rho \frac{\sigma_Y}{\sigma_X}$,с другой стороны $a = \mu$.

Отсюда $\mu = \rho \frac{\sigma_Y}{\sigma_X} = 0$, так как $\rho = 0$. Итак любое собственное число в этом случае равно нулю, а потому и максимальный коэффициент корреляции, являющийся наибольшим по модулю собственным числом, тоже равен нулю. Необходимость доказана.

Достаточность.

Пусть $\rho^* = 0.$ Докажем, что в этом случае случайные величины X, Y независимы.

1). Рассмотрим квадратичную форму

$$H = \sum_i \sum_j \frac{p_{ij}}{\sqrt{p_i p_j}} \xi_i \xi_j .$$

(15)

Квадратичную форму H приводим к сумме квадратов с помощью соответствующего ортогонального преобразования. При этом коэффициент у каждого квадрата является собственным числом матрицы коэффициентов $\left(\frac{p_{ij}}{\sqrt{p_i p_j}} \right)$ квадратичной формы H. Так как $\rho^* = 0$ и ρ^* является наибольшим по модулю из собственных чисел матрицы коэффициентов квадратичной формы H, то все собственные числа этой матрицы равны нулю. Поэтому форма H, приведенная к сумме квадратов, будет тождественно равна нулю: $H \equiv 0$.

Получаем тождество

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{p_{ij}}{\sqrt{p_i p_j}}\xi_i\xi_j \equiv 0\,.$$

(16)

2). Пусть $P\left(X = x_i\right) = p_i\,;\; m = M\left[X\right];\; \sigma = \sigma\left[X\right].$

Положим

$$\xi_i = \frac{x_i - m}{\sigma}\sqrt{p_i}\,;\quad i = 1,2,...,n.$$

(17)

Отметим, что величины (17) удовлетворяют двум условиям:

$$\sum_{i=1}^{n}\xi_i\sqrt{p_i} = \sum_{i=1}^{n}\frac{x_i - m}{\sigma}\sqrt{p_i} = \frac{1}{\sigma}\left(\sum_{i=1}^{n}x_i p_i - m\sum_{i=1}^{n}p_i\right) = \frac{1}{\sigma}\left(m - m\right) = 0;$$

$$\sum_{i=1}^{n}\xi_i^2 = \sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2 p_i = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2m\sum_{i=1}^{n}x_i p_i + m^2\sum_{i=1}^{n}p_i\right) =$$

$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i^2 p_i - m^2\right) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

Подставим эти величины (17) в тождество (16). Получаем $\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{p_{ij}}{\sqrt{p_i p_j}}\frac{x_i - m}{\sigma}\sqrt{p_i}\frac{x_j - m}{\sigma}\sqrt{p_j} \equiv 0$. Отсюда далее находим

$$\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ij}\left[x_i x_j + m^2 - m\left(x_i + x_j\right)\right] \equiv 0.$$

(18)

Выполним некоторые преобразования.

$$m\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i + x_j\right)p_{ij} = m\sum_{i=1}^{n}x_i\sum_{j=1}^{n}p_{ij} + m\sum_{j=1}^{n}y_j\sum_{i=1}^{n}p_{ij} =$$

$$m\left(\sum_{i=1}^{n}x_i p_i + \sum_{j=1}^{n}x_j p_j\right) = = m\left(m + m\right) = 2m^2.$$ Преобразуем далее равенство (18).

$$\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ij}x_i x_j + m^2\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ij} - 2m^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}p_{ij}x_i x_j - m^2 =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}p_{ij}x_i x_j\left(\sum_{i=1}^{n}x_i p_i\right)^2 =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}x_i x_j - \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j x_i x_j = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(p_{ij} - p_i p_j\right)x_i x_j \equiv 0.$$

В силу произвольности чисел $x_i \quad (i = 1,2,...,n)$ получаем $p_{ij} - p_i p_j = 0; \quad i, j = 1,2,...,n.$

Равенства

$p_{ij} = p_i p_j; \quad i, j = 1,2,...,n$ и означают независимость дискретных случайных величин X, Y. Достаточность доказана. Теорема доказана полностью.

Пример 1. Триномиальное распределение определяется формулой

$$p_{ij} = \frac{n!}{i!\,j!(n-i-j)!} p_1^i p_2^j (1 - p_1 - p_2)^{n-i-j};$$

$i, j = 0.1,...,n; \quad 0 < p_1 < 1; \quad 0 < p_2 < 1; \quad p_1 + p_2 < 1; \quad i + j \le n.$

Рассмотрим случай $n = 2; \quad p_1 = p_2 = 1/4.$ Построим таблицу распределения (табл.1).

Таблица 1 триномиального распределения при $n = 2.$

$X \setminus Y \to$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00} = 1/4$	$p_{01} = 1/4$	$p_{02} = 1/16$	$p_{0\cdot} = 9/16$
1	$p_{10} = 1/4$	$p_{11} = 1/8$	$p_{12} = 0$	$p_{1\cdot} = 3/8$
2	$p_{20} = 1/16$	$p_{21} = 0$	$p_{22} = 0$	$p_{2\cdot} = 1/16$
$p_{\cdot j}$	$p_{\cdot 0} = 9/16$	$p_{\cdot 1} = 3/8$	$p_{\cdot 2} = 1/16$	1

Линейный коэффициент корреляции вычисляется по формуле

$$\rho = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}} = -\frac{1}{3} \;;[3, cmp.142].$$

Вычислим максимальный коэффициент корреляции ρ^*. Для этого из матрицы вероятностей $\left(p_{ij}\right)$ построим симметричную матрицу $P = \left(\frac{p_{ij}}{\sqrt{p_i p_j}}\right)$ для вычисления максимального коэффициента корреляции. Здесь $p_i = p_{i\cdot}; p_j = p_{\cdot j}$.

$$P=\begin{pmatrix} 4/9 & 4/3\sqrt{6} & 1/3 \\ 4/3\sqrt{6} & 1/3 & 0 \\ 1/3 & 0 & 0 \end{pmatrix}.$$

(19)

На основе этой матрицы составляем характеристическое уравнение.

$$\begin{vmatrix} \mu-\frac{4}{9} & \frac{4}{3\sqrt{6}} & \frac{1}{3} \\ \frac{4}{3\sqrt{6}} & \mu-\frac{1}{3} & 0 \\ \frac{1}{3} & 0 & \mu \end{vmatrix}=0.$$

Раскладываем определитель по последней строке и записываем уравнение в стандартной форме.

$$\mu^3-\frac{7}{9}\mu^2-\frac{7}{27}\mu+\frac{1}{27}=0.$$

(20)

Характеристический многочлен легко раскладывается на множители:

$$\left(\mu-1\right)\left(\mu+\frac{1}{3}\right)\left(\mu-\frac{1}{9}\right)=0.$$

Выписываем 3 собственных числа матрицы P:

$$\mu_0=1;\quad \mu_1=-\frac{1}{3};\quad \mu_2=\frac{1}{9}.$$

Наибольшее по модулю собственное число $\mu_1=-\frac{1}{3}$ ($\mu_0=1$ во внимание не принимается)

и есть максимальный коэффициент корреляции: $\rho^*=-\frac{1}{3}$. Он совпадает с линейным коэффициентом корреляции.

§ 2.3. Максимальный коэффициент детерминации для непрерывных случайных величин. Несимметричный случай.

1º. Сведение несимметричного случая к двум симметричным.

Пусть $f_{XY}(x,y)-$ плотность распределения двумерной случайной величины (X,Y), определяющая зависимость между случайными величинами X,Y в прямоугольной области

$D=\{(x;y): a\le x\le b; a_1\le y\le b_1\}$, которая может быть и бесконечной. Пусть далее

$$f_X(x)=\int_{a_1}^{b_1} f_{XY}(x,y)\,dy;\quad f_Y(y)=\int_a^b f_{XY}(x,y)dx-$$

(1)

плотности распределения компонент X,Y двумерной случайной величины (X,Y).

Предполагаем, что квадрат ядра

$$K(x,y)=\frac{f_{XY}(x,y)}{\sqrt{f_X(x)f_Y(y)}}$$

(2)

интегрируем по обеим переменным.

Несимметричная плотность $f_{XY}(x,y)$ определяет две симметричные плотности [32,33]

$$f_1(x,y)=\int_{a_1}^{b_1}\frac{f_{XY}(x,t)f_{XY}(y,t)}{f_Y(t)}dt;\quad f_2(x,y)=\int_a^b\frac{f_{XY}(t,x)f_{XY}(t,y)}{f_X(t)}dt.$$

(3)

Первая функция $f_1(x,y)$ определена по первому аргументу исходной плотности $f_{XY}(x,y)$, то есть в квадрате $a\le x\le b;\quad a\le y\le b$. Вторая функция определена по второму аргументу плотности $f_{XY}(x,y)$, то есть в квадрате $a_1\le x\le b_1; a_1\le y\le b_1$.

Функции (3) удовлетворяют требованию нормировки плотности. Проверим это, например, для первой из них.

$$\int_a^b\int_a^b f_1(x,y)\,dxdy=\int_a^b dx\int_a^b dy\int_{a_1}^{b_1}\frac{f_{XY}(x,t)f_{XY}(y,t)}{f_Y(t)}dt=$$

$$\int_a^b dy\int_{a_1}^{b_1}\frac{f_{XY}(y,t)}{f_Y(t)}dt\int_a^b f_{XY}(x,t)dx=$$

$$=\int_a^b dy\int_{a_1}^{b_1}\frac{f_{XY}(y,t)}{f_Y(t)}f_Y(t)\,dt=\int_a^b dy\int_{a_1}^{b_1}f_{XY}(y,t)dt=\int_a^b f_X(y)\,dy=1.$$

Предполагаем, что смену порядка интегрирования произвести можно.

Через плотности $f_1(x,y), f_2(x,y)$ определяются два симметричных ядра

$$K_1(x,y)=\frac{f_1(x,y)}{\sqrt{f_X(x)f_X(y)}};\quad K_2(x,y)=\frac{f_2(x,y)}{\sqrt{f_Y(x)f_Y(y)}}.$$

(4)

Первое ядро определено в квадрате $a\le x\le b; a\le y\le b$.

Второе ядро определено в квадрате $a_1\le x\le b_1; a_1\le y\le b_1$.

Ядра (4) положительны и имеют одинаковые спектры собственных чисел

$$1<\lambda_1^2\le\lambda_2^2\le\ldots\le\lambda_k^2\le\ldots$$

(5)

и, вообще говоря, различные спектры собственных функций

$$\{\varphi_i(x)\},\quad \{\psi_i(x)\};\quad i=1,2,...$$

(6)

Согласно теории Гильберта-Шмидта спектр ядра $K(x,y)$ собственных чисел и собственных функций имеет вид

$$\lambda_0=1,\lambda_1,\lambda_2,...,\lambda_k,...$$

(7)

$$\sqrt{f_X(x)},\sqrt{f_Y(y)};\qquad \sqrt{f_X(x)}\varphi_k(x),\sqrt{f_k(y)}\psi_k(y);\quad k=1,2,...$$

(8)

Билинейное разложение

$$\sqrt{f_X(x)}\sqrt{f_Y(y)}+\sum_{k=1}^{\infty}\frac{\varphi_k(x)\sqrt{f_X(x)}\psi_k(y)\sqrt{f_Y(y)}}{\lambda_k}$$

(9)

сходится в среднем к ядру $K(x,y)$ в области D.

2º. Определение максимального коэффициента корреляции в несимметричном непрерывном случае.

Определение максимального коэффициента корреляции ρ^* в несимметричном случае аналогично его определению в случае симметричном:

Это есть число, определяемое по формуле

$$\rho^* = \frac{1}{\lambda_1}.$$

(10)

Здесь $|\lambda_1|$ есть корень из первого собственного числа $\mu_1 = \lambda_1^2$ симметричных ядер $K_1(x,y), K_2(x,y)$, определяемых формулой (4). Знак λ_1, следовательно и ρ^*, определяется из соображений, лежащих вне рассматриваемой теории. Данная теория позволяет найти только $|\lambda_1| = \sqrt{\lambda_1^2}$, а потому $|\rho^*|$, который является по сути дела коэффициентом детерминации между случайными величинами X, Y.

3º. Приближенное вычисление квадрата максимального коэффициента корреляции.

$(\rho^*)^2$ вычисляется приближенно с помощью следующих последовательных приближений функций $r_{2k+1}(x)$ и $r_{2k+2}(y)$. В качестве «нулевого приближения» $r_0(y)$ можно взять любую функцию, такую, что $r_0(Y)$ имеет конечную дисперсию. Не ограничивая общности, будем считать, что математическое ожидание $r_0(Y)$ равно нулю.

Положим

$$r_{2k+1}(x) = \int_{a_1}^{b_1} r_{2k}(y) \frac{f_{XY}(x,y)}{f_X(x)} dy;$$

(11)

$$r_{2k+2}(y) = \int_{a}^{b} r_{2k+1}(y) \frac{f_{XY}(x,y)}{f_Y(y)} dx; \qquad k = 0,1,2,...$$

(12)

Сходимость процесса последовательных приближений доказана в работе О.В. Сарманова [32,33]. Первая пара собственных функций с точностью до нормирующих множителей q_1, g_1 определяется равенствами

$$q_1 \psi_1(y) = \lim_{k\to\infty} r_{2k}(y) \lambda_1^{2k};$$

(13)

$$g_1 \varphi_1(x) = \lim_{k\to\infty} r_{2k+1}(x) \lambda_1^{2k}.$$

(14)

Если k достаточно велико то

$$\left(\rho^*\right)^2 = \frac{1}{\lambda_1^2} \approx \frac{r_{2k}(y)}{r_{2k-2}(y)} \approx \frac{r_{2k+1}(x)}{r_{2k-1}(x)}.$$

(15)

4º. Свойства максимального коэффициента корреляции.

Свойство 1. Для независимости случайных величин X, Y необходимо и достаточно обращение в нуль их максимального коэффициента корреляции.

Свойство 2. Если корреляция между случайными величинами X, Y линейная, то их линейный коэффициент корреляции ρ совпадает с максимальным коэффициентом корреляции ρ^*.

Доказательство этих свойств аналогично доказательству свойств в симметричном случае
([32, 33]).

Пример 1. Вычисление максимального коэффициента детерминации $\left|\rho^*\right|$ в случае несимметричного непрерывного двумерного распределения, заданного плотностью

$$f_{XY}(x,y) = \begin{cases} \frac{1}{4}(x+2y); & (x;y) \in D \\ 0; & (x;y) \notin D \end{cases}.$$

(16)

D – прямоугольник, заданный неравенствами $0 \le x \le 2; \quad 0 \le y \le 1.$

Для решения задачи последовательно находим

$$f_X(x) = \int_0^1 f_{XY}(x,y)dy = \int_0^1 \frac{1}{4}(x+2y)\,dy = \frac{1}{4}(x+1); \qquad 0 \le x \le 2.$$

(17)

$$f_Y(y) = \int_0^2 f_{XY}(x.y)dx = \int_0^2 \frac{1}{4}(x+2y)\,dx = \frac{1}{2}(2y+1); \qquad 0 \le y \le 1.$$

(18)

Образуем две симметричные плотности $f_1(x,y)$ и $f_2(x,y)$.

$$f_1(x,y) = \int_{a_1}^{b_1} \frac{f_{XY}(x,t)\,f_{XY}(y,t)}{f_Y(t)}dt = \int_0^1 \frac{\frac{1}{4}(x+2t)\frac{1}{4}(y+2t)}{\frac{1}{2}(1+2t)}dt =$$

$$= \frac{1}{8}\int_0^1 \frac{(2t+x)(2t+y)}{2t+1}dt = \frac{1}{16}\left[2x+2y+(x-1)(y-1)ln3\right].$$

$$f_1(x,y)\frac{1}{16}\left[2x+2y+(x-1)(y-1)\ln 3\right]; \qquad 0 \le x \le 2; 0 \le y \le 2.$$

(19)

$$f_2(x,y) = \int_a^b \frac{f_{XY}(t,x)f_{XY}(t,y)}{f_X(t)}dt = \int_0^2 \frac{\frac{1}{4}(t+2x)(t+2y)}{\frac{1}{4}(t+1)}dt =$$

$$\frac{1}{4}\int_0^2 \frac{(t+2x)(t+2y)}{t+1}dt = = x+y+\frac{1}{4}(2x-1)(2y-1)\ln 3.$$

$$f_2(x,y) = x+y+\frac{1}{4}(2x-1)(2y-1)\ln 3; \qquad 0 \le x \le 1; 0 \le y \le 1.$$

(20)

Из этих плотностей образуем два симметричных ядра (4):

$$K_1(x,y) = \frac{1}{4}\frac{2x+2y+(x-1)(y-1)\ln 3}{\sqrt{(x+1)(y+1)}}; \qquad 0 \le x \le 2; 0 \le y \le 2.$$

(21)

$$K_2(x,y) = \frac{1}{2}\frac{4x+4y+(2x-1)(2y-1)\ln 3}{\sqrt{(2x+1)(2y+1)}}; \qquad 0 \le x \le 1; 0 \le y \le 1.$$

(22)

Оба ядра вырожденные, поэтому корреляционное интегральное уравнение для отыскания собственных чисел сводится к системе алгебраических уравнений. Достаточно рассмотреть одно ядро, так как оба уравнения приводят к одним и тем же собственным числам. Рассмотрим первое ядро как более простое, Составим для него интегральное уравнение:

$$\omega(x) = \mu\int_{a_1}^{b_1} \omega(y)K_1(x,y)dy;$$

(23)

$$\omega(x) = \mu\int_0^2 \omega(y)\frac{1}{4}\frac{2x+2y+(x-1)(y-1)\ln 3}{\sqrt{(x+1)(y+1)}}dy.$$

(24)

Запишем его в виде

$$\omega(x)=\mu\left[\frac{2x-\ln 3(x-1)}{4\sqrt{x+1}}\int_0^2\omega(y)\frac{dy}{\sqrt{y+1}}+\frac{2+\ln 3(x-1)}{4\sqrt{x+1}}\int_0^2\omega(y)\frac{y}{\sqrt{y+1}}dy\right].$$

Положим

$$s_1=\int_0^2\omega(y)\frac{dy}{\sqrt{y+1}};\quad s_2=\int_0^2\omega(y)\frac{y}{\sqrt{y+1}}dy.$$

(25)

Тогда уравнение принимает вид

$$\omega(x)=\mu\left[\frac{2x-\ln 3(x-1)}{4\sqrt{x+1}}s_1+\frac{2+\ln 3(x-1)}{4\sqrt{x+1}}s_2\right].$$

(26)

Умножаем уравнение (26) на функцию $1/\sqrt{x+1}$ и интегрируем полученное равенство по промежутку [0;2]:

$$s_1=\mu\left[s_1\int_0^2\frac{2x-\ln 3(x-1)}{4(x+1)}dx+s_2\int_0^2\frac{2+\ln 3(x-1)}{4(x+1)}dx\right].$$

(27)

Теперь умножаем уравнение (26) на функцию $x/\sqrt{x+1}$ и интегрируем полученное равенство по промежутку [0;2]:

$$s_2=\mu\left[s_1\int_0^2\frac{2x^2-\ln 3x(x-1)}{4(x+1)}dx+s_2\int_0^2\frac{2x+\ln 3x(x-1)}{4(x+1)}dx\right].$$

(28)

Вычисляем интегралы.

$$\int_0^2\frac{2x-\ln 3(x-1)}{4(x+1)}dx=\frac{2-2\ln 3+\ln^2 3}{2};\quad\int_0^2\frac{2+\ln 3(x-1)}{4(x+1)}dx=\frac{(2-\ln 3)\ln 3}{2};$$

$$\int_0^2\frac{2x^2-\ln 3x(x-1)}{4(x+1)}dx=\frac{(2-\ln 3)\ln 3}{2};$$

$$\int_0^2\frac{2x+\ln 3x(x-1)}{4(x+1)}dx=\frac{2-2\ln 3+\ln^2 3}{2}.$$

Получаем систему алгебраических уравнений с неизвестными s_1 s_2

$$\begin{cases} s_1 = \mu s_1 \dfrac{2-2\ln 3+\ln^2 3}{2} + \mu s_2 \dfrac{(2-\ln 3)\ln 3}{2} \\ s_2 = \mu s_1 \dfrac{(2-\ln 3)\ln 3}{2} + \mu s_2 \dfrac{2-2\ln 3+\ln^2 3}{2} \end{cases}.$$

$$\begin{cases} s_1 = \mu s_1 \dfrac{2-2\ln 3+\ln^2 3}{2} + \mu s_2 \dfrac{(2-\ln 3)\ln 3}{2} \\ s_2 = \mu s_1 \dfrac{(2-\ln 3)\ln 3}{2} + \mu s_2 \dfrac{2-2\ln 3+\ln^2 3}{2} \end{cases}.$$

Запишем эту систему в стандартной форме:

$$\begin{cases} s_1\left(\mu \dfrac{2-2\ln 3+\ln^2 3}{2} - 1\right) + s_2 \mu \dfrac{(2-\ln 3)\ln 3}{2} = 0 \\ s_1 \mu \dfrac{(2-\ln 3)\ln 3}{2} + s_2\left(\mu \dfrac{2-2\ln 3+\ln^2 3}{2} - 1\right) = 0 \end{cases}.$$

(29)

Для того, чтобы однородная система имела ненулевые решения, необходимо и достаточно обращение в нуль определителя системы:

$$\begin{vmatrix} \mu \dfrac{2-2\ln 3+\ln^2 3}{2} - 1 & \mu \dfrac{(2-\ln 3)\ln 3}{2} \\ \mu \dfrac{(2-\ln 3)\ln 3}{2} & \mu \dfrac{2-2\ln 3+\ln^2 3}{2} - 1 \end{vmatrix} = 0.$$

Это есть характеристическое уравнение системы относительно неизвестной величины μ.

Вычисляя определитель, запишем его в виде

$$\left(\mu \frac{2-2\ln 3+\ln^2 3}{2} - 1\right)^2 - \left(\mu \frac{(2-\ln 3)\ln 3}{2}\right)^2 = 0.$$

Раскладываем разность квадратов на множители и находим корни квадратного уравнения:

$$\mu_0 = 1; \quad \mu_1 = \frac{1}{(1-\ln 3)^2}.$$

Отсюда $\left(\rho^*\right)^2 = \dfrac{1}{\mu_1} = (1-\ln 3)^2; \ \left|\rho^*\right| = \ln 3 - 1.$ Таким образом,

максимальный коэффициент детерминации

$$\left|\rho^*\right| = \ln 3 - 1 \approx 0,0986 .$$

(30)

Из сопоставления этого результата с аналогичным симметричным случаем, можем утверждать, что $\rho^* < 0$. Тогда максимальный коэффициент корреляции

$$\rho^* = 1 - \ln 3 \approx -0,0986.$$

(31)

Пример 2. Для сравнения с максимальным коэффициентом корреляции (31), вычислим линейный коэффициент корреляции для этого же распределения из примера 1.

Используя найденные там результаты, получаем далее

$$m_X = \int_0^2 x f_X(x)\,dx = \int_0^2 x\frac{1}{4}(x+1)dx = \frac{7}{6};\ m_Y = \int_0^1 y\frac{1}{2}(1+2y)\,dy = \frac{7}{12};$$

$$\alpha_{2X} = \int_0^2 x^2\frac{1}{4}(x+1)\,dx = \frac{5}{3};\ \alpha_{2Y} = \int_0^1 y^2\frac{1}{2}(1+2y)\,dy = \frac{5}{12};$$

$$D_X = \alpha_{2X} - m_X^2 = \frac{5}{3} - \frac{49}{36} = \frac{11}{36};\ \sigma_X = \frac{\sqrt{11}}{6};\ D_Y = \frac{5}{12} - \frac{49}{144};\ \sigma_Y = \frac{\sqrt{11}}{12};$$

$$M[XY] = \int_0^2 dx\int_0^1 xy\frac{1}{4}(x+2y)\,dy = \frac{2}{3};$$

$$K_{XY} = M[XY] - m_X m_Y = \frac{2}{3} - \frac{7}{6}\frac{7}{12} = -\frac{1}{72};$$

$\rho = \dfrac{K_{XY}}{\sigma_X\sigma_Y} = \dfrac{-1/72}{\frac{\sqrt{11}}{6}\frac{\sqrt{11}}{12}} = -\dfrac{1}{11}$. Итак, линейный коэффициент корреляции равен

$$\rho = -\frac{1}{11} = -0,0909.$$

(32)

Значения (31) и (32) близкие, в том числе и по знаку.

Пример 3. Для сравнения вычислим дефектологический коэффициент детерминации для распределения из примеров 1и 2, (§ 8.3).

$$F_{XY}(x,y) = \frac{1}{4}\int_0^x du\int_0^y (u+2v)dv = \frac{1}{8}\left(x^2y + 2xy^2\right);$$

$$F_X(x) = F_{XY}(x,1) = \frac{1}{8}\left(x^2+2x\right);\ \ F_Y(y) = F_{XY}(2,y) = \frac{1}{2}\left(y^2+y\right);$$

$$def = 6\int_0^2 dx \int_0^1 \left| F_{XY}(x,y) - F_X(x) F_Y(y) \right| f_{XY}(x,y)\, dy =$$

$$= 6\int_0^2 dx \int_0^1 \left| \frac{1}{8}\left(x^2 y + 2xy^2\right) - \frac{1}{8}\left(x^2 + 2x\right)\frac{1}{2}\left(y^2 + y\right) \right| \frac{1}{4}(x + 2y)\, dy;$$

$$def = \frac{3}{32}\int_0^2 dx \int_0^1 \left| 2\left(x^2 y + 2xy^2\right) - \left(x^2 + 2x\right)\left(y^2 + y\right) \right| (x + 2y)\, dy =$$

$$= \frac{3}{32}\int_0^2 dx \int_0^1 \left| x + 2y - xy - 2 \right| xy(x + 2y)\, dy =$$

$$\frac{3}{32}\int_0^2 dx \int_0^1 \left| (x-2)(1-y) \right| xy(x + 2y)\, dy =$$

$$= \frac{3}{32}\int_0^2 dx \int_0^1 (2 - x)(1 - y)\, xy(x + 2y)\, dy =$$

$$= \frac{3}{32}\left[\int_0^2 \left(2x^2 - x^3\right) dx \int_0^1 \left(y - y^2\right) dy + 2\int_0^2 \left(2x - x^2\right) dx \int_0^1 \left(y^2 - y^3\right) dy \right] = \frac{1}{24}.$$

$$def = \frac{1}{24} \approx 0,0417.$$

(33)

Для сравнения записываем все три характеристики связи вместе:

$$\rho^* = -0,0986; \quad \rho = -0,0909; \quad def = 0,0417.$$

Все 3 характеристики указывают на слабую связь, дублируя друг друга, хотя каждая в меру своей жесткости.

§ 2.4. Максимальный коэффициент детерминации для дискретных случайных величин. Несимметричный случай.

Пусть вероятностная зависимость между двумя дискретными случайными величинами определяется прямоугольной матрицей вероятностей

$$\left\{ p_{ij} \right\}, \quad i = 1,2,...,m; \quad j = 1,2,...,n.$$

(1)

Здесь

$$p_{ij} = P\left(X = x_i, Y = y_j\right); \quad \sum_{i,j} p_{ij} = 1;$$

(2)

$$p_{i\cdot} = \sum_{j=1}^{n} p_{ij} = P\left(X = x_i\right); \qquad p_{\cdot j} = \sum_{i=1}^{m} p_{ij} = P\left(Y = y_j\right).$$

(3)

С помощью матрицы вероятностей (1) образуем две квадратные симметричные матрицы

$$P_1 = \left\{ \frac{p_{ij}^{(1)}}{\sqrt{p_{i\cdot} p_{j\cdot}}} \right\}; \qquad i, j = 1, 2..., m;$$

(4)

$$P_2 = \left\{ \frac{p_{ij}^{(2)}}{\sqrt{p_{\cdot i} p_{\cdot j}}} \right\}; \quad i, j = 1, 2, ..., n,$$

(5)

где

$$p_{ij}^{(1)} = \sum_{k=1}^{n} \frac{p_{ik} p_{jk}}{p_{\cdot k}}; \qquad p_{ij}^{(2)} = \sum_{k=1}^{m} \frac{p_{ki} p_{kj}}{p_{k\cdot}}.$$

(6)

Обе симметричные матрицы P_1 и P_2 имеют конечный спектр собственных чисел, в общем случае разный, так как матрицы имеют разный порядок: $\lambda_0 = 1, \lambda_1^2, \ldots$ Однако, первое собственное число λ_1^2 у обеих матриц – одно и то же.

Определение. Максимальным коэффициентом детерминации между случайными величинами X, Y называется число, определяемое формулой

$$\rho^* = \left|\lambda_1\right|.$$

(7)

Здесь $\left|\lambda_1\right| = \sqrt{\lambda_1^2} = \sqrt{\mu_1}; \quad (\lambda_1^2 = \mu_1)$ – корень из первого (наибольшего, отличного от нуля и единицы) собственного числа μ_1 матрицы P_1 (или P_2).

Замечание. Изложенная теория не дает возможности найти знак λ_1, а, следовательно, и знак максимального коэффициента корреляции ρ^*. С помощью рассмотренной теории определяется только коэффициент детерминации $\left|\rho^*\right|$. Знак ρ^* определяется из других соображений, лежащих вне данной теории.

Процесс последовательных приближений для нахождения максимального коэффициента детерминации.

Пусть $\overline{r_0} = \left(y_1^{(0)}, y_2^{(0)}, ..., y_n^{(0)} \right)$ – произвольный вектор такой, что $\sum_{j=1}^{n} y_j^{(0)} p_{\cdot j} = 0$.

Положим

$$x_i^{(2k+1)} = \sum_{j=1}^{n} \frac{p_{ij}}{p_{i\cdot}} y_j^{(2k)}; \qquad i = 1,2,...,m; \quad k = 0,1,2,...$$

(8)

$$y_j^{(2k+2)} = \sum_{i=1}^{m} \frac{p_{ij}}{p_{\cdot j}} x_i^{(2k+1)}; \qquad j = 1,2,...,n; \quad k = 0,1,2,...$$

(9)

Если k достаточно велико, то

$$\left(\rho^* \right)^2 \approx \frac{y_j^{(2k)}}{y_j^{(2k-2)}} \approx \frac{x_i^{(2k+1)}}{x_i^{(2k-1)}} \qquad \text{при всех } i, j.$$

(10)

Сходимость процесса приближений доказана в работе О.В. Сарманова [32,33].
Роль максимального коэффициента корреляции (детерминации) здесь – та же, что и в симметричном случае.

Пример 1. Вычисление максимального коэффициента детерминации для дискретного несимметричного распределения, представленного следующей таблицей 1.

Таблица 1. Несимметричное дискретное распределение вероятностей к примеру 1 для вычисления максимального коэффициента детерминации.

$X \downarrow Y \rightarrow$	1	2	3	$p_{i\cdot}$
1	$p_{11} = 2/12$	$p_{12} = 1/12$	$p_{13} = 1/12$	$4/12$
2	$p_{21} = 4/12$	$p_{22} = 3/12$	$p_{23} = 1/12$	$8/12$
$p_{\cdot j}$	$6/12$	$4/12$	$2/12$	1

По формулам (6) несимметричное распределение, определенное таблицей 1, преобразуем в два симметричных дискретных распределения, представленных таблицами 2 и 3.

Таблица 2 размера 2×2 симметричного дискретного распределения вероятностей, полученная из таблицы 1 по формулам (6).

$X_1 \downarrow Y_1 \rightarrow$	$y_1^{(1)}$	$y_2^{(1)}$	$p_{i\cdot}^{(1)}$
$x_1^{(1)}$	17/144	31/144	48/144
$x_2^{(1)}$	31/144	65/144	96/144
$p_{\cdot j}^{(1)}$	48/144	96/144	1

Таблица 3 размера 3×3 симметричного дискретного распределения вероятностей, полученная из таблицы 1 по формулам (6).

$X_2 \downarrow Y \rightarrow$	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	$p_{i\cdot}^{(2)}$
$x_1^{(2)}$	24/96	16/96	8/96	48/96
$x_2^{(2)}$	16/96	11/96	5/96	32/96
$x_3^{(2)}$	8/96	5/96	3/96	16/96
$p_{\cdot j}^{(2)}$	48/96	32/96	16/96	1

Из таблиц 2 и 3 по формулам (4) и (5) образуем две квадратные симметричные матрицы P_1, P_2.

$$P_1 = \begin{pmatrix} \frac{17}{48} & \frac{31}{48\sqrt{2}} \\ \frac{31}{48\sqrt{2}} & \frac{65}{96} \end{pmatrix}; \qquad P_2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{11}{32} & \frac{5}{16\sqrt{2}} \\ \frac{1}{2\sqrt{3}} & \frac{5}{16\sqrt{2}} & \frac{3}{16} \end{pmatrix}. \tag{11}$$

Из этих матриц P_1, P_2 образуются характеристические уравнения для отыскания их характеристических чисел.

Рассмотрим сначала матрицу P_1 и составим для нее характеристическое уравнение:

$$\begin{vmatrix} \frac{17}{48}-\mu & \frac{31}{48\sqrt{2}} \\ \frac{31}{48\sqrt{2}} & \frac{65}{96}-\mu \end{vmatrix} = 0.$$

(12)

Вычисляя определитель, приведем характеристическое уравнение к виду

$$\mu^2 - \frac{33}{32}\mu + \frac{1}{32} = (\mu - 1)\left(\mu - \frac{1}{32}\right) = 0.$$

(13)

Выписываем корни уравнения

$$\mu_1 = \frac{1}{32}; \quad \mu_2 = 1.$$

(14)

Применяется наименьший из этих корней $\mu_1 = 1/32$. Из него образуется максимальный коэффициент детерминации:

$$\left|\rho^*\right| = \sqrt{\mu_1} = \sqrt{\lambda_1^2} = \left|\lambda_1\right| = \frac{1}{\sqrt{32}} = \frac{1}{4\sqrt{2}} \approx 0{,}177.$$

(15)

Для контроля рассмотрим теперь матрицу P_2. Составим для нее характеристическое уравнение. Оно должно привести к тому же корню μ_1 и к тому же коэффициенту детерминации $\left|\rho^*\right|$. Характеристическое уравнение для матрицы P_2 имеет вид:

$$\begin{vmatrix} \frac{1}{2}-\mu & \frac{1}{\sqrt{6}} & \frac{1}{2\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{11}{32}-\mu & \frac{5}{16\sqrt{2}} \\ \frac{1}{2\sqrt{3}} & \frac{5}{16\sqrt{2}} & \frac{3}{16}-\mu \end{vmatrix} = 0.$$

(16)

Вычисляем определитель и записываем уравнение в канонической форме:

$$\mu^3 - \frac{33}{32}\mu^2 + \frac{1}{32}\mu = \mu(\mu - 1)\left(\mu - \frac{1}{32}\right) = 0.$$

(17)

Выписываем корни уравнения – характеристические числа:

$$\mu_1 = \frac{1}{32}; \quad \mu_2 = 1; \quad \mu_3 = 0.$$

(18)

Корни 0 и 1 во внимание не принимаются. Корень μ_1 – тот же, что и для уравнения (13).

С помощью матрицы P_2 приходим к тому же значению для максимального коэффициента детерминации.

Чтобы показать как работает метод последовательных приближений, вычислим максимальный коэффициент детерминации рассматриваемого распределения этим методом по формулам (8), (9), (10).

В качестве нулевого (начального) вектора возьмем вектор

$\overline{r}_0(y) = \left(y_1^{(0)}, ..., y_n^{(0)}\right) = \left(-\frac{1}{6}; \frac{1}{8}; \frac{1}{4}\right)$. Для этого вектора выполнено требование

$\sum_{j=1}^{n} y_j^{(0)} p_{\cdot j} = \frac{1}{12}\left[6\left(-\frac{1}{6}\right) + 4\frac{1}{8} + 2\frac{1}{4}\right] = 0.$ Тогда

$$x_1^{(1)} = \sum_{j=1}^{3} \frac{p_{1j}}{p_{1\cdot}} y_j^{(0)} = \frac{2/12}{4/12}\left(-\frac{1}{6}\right) + \frac{1/12}{4/12}\frac{1}{8} + \frac{1/12}{4/12}\frac{1}{4} = \frac{1}{96};$$

$$x_2^{(1)} = \sum_{j=1}^{3} \frac{p_{2j}}{p_{2\cdot}} y_j^{(0)} = \frac{4/12}{8/12}\left(-\frac{1}{6}\right) + \frac{3/12}{8/12}\frac{1}{8} + \frac{1/12}{8/12}\frac{1}{4} = -\frac{1}{192};$$

$$y_1^{(2)} = \sum_{i=1}^{2} \frac{p_{i1}}{p_{\cdot 1}} x_i^{(1)} = \frac{2/12}{6/12}\frac{1}{96} + \frac{4/12}{6/12}\left(-\frac{1}{192}\right) = 0;$$

$$y_2^{(2)} = \sum_{i=1}^{2} \frac{p_{i2}}{p_{\cdot 2}} x_i^{(1)} = \frac{1/12}{4/12}\frac{1}{96} + \frac{3/12}{4/2}\left(-\frac{1}{192}\right) = -\frac{1}{8 \cdot 96};$$

$$y_3^{(2)} = \sum_{i=1}^{2} \frac{p_{i3}}{p_{\cdot 3}} x_i^{(1)} = \frac{1/12}{2/12}\frac{1}{96} + \frac{1/12}{2/12}\left(-\frac{1}{192}\right) = \frac{1}{4 \cdot 96};$$

Аналогично вычисляем

$$x_1^{(3)} = \frac{1}{32 \cdot 96}; \qquad x_2^{(3)} = -\frac{1}{64 \cdot 96}; \qquad y_1^{(4)} = 0; \qquad y_2^{(4)} = -\frac{1}{256 \cdot 96};$$

$$y_3^{(4)} = \frac{1}{128 \cdot 96}.$$

Для вычисления максимального коэффициента детерминации возьмем, например, третью координату текущего вектора:

$\left(\rho^*\right)^2 \approx \dfrac{y_3^{(4)}}{y_3^{(2)}} = \dfrac{\dfrac{1}{128\cdot 96}}{\dfrac{1}{4\cdot 96}} = \dfrac{1}{32}$; Отсюда $\left|\rho^*\right| = \dfrac{1}{\sqrt{32}} \approx 0{,}177$. Получили то же значение максимального коэффициента детерминации, что и ранее.

Так как все познается в сравнении, то вычислим для рассматриваемого распределения, определенного таблицей 1, и некоторые другие коэффициенты детерминации и корреляции.

Пример 2. Вычисление линейного коэффициента корреляции ρ для распределения, определенного таблицей 1.

Последовательно вычисляем нужные величины.

Математические ожидания случайных величин:

$$m_X = 1\cdot\frac{4}{12} + 2\cdot\frac{8}{12} = \frac{20}{12} = \frac{5}{3}; \quad m_Y = 1\cdot\frac{6}{12} + 2\cdot\frac{4}{12} + 3\cdot\frac{2}{12} = \frac{20}{12} = \frac{5}{3}.$$

Вторые начальные моменты:

$$\alpha_{2X} = 1\cdot\frac{4}{12} + 4\frac{8}{12} = 3; \quad \alpha_{2Y} = 1\cdot\frac{6}{12} + 4\cdot\frac{4}{12} + 9\cdot\frac{2}{12} = \frac{40}{12} = \frac{10}{3}.$$

Дисперсии:

$$D_X = \alpha_{2X} - m_X^2 = 3 - \frac{25}{9} = \frac{2}{9}; \quad D_Y = \alpha_{2Y} - m_Y^2 = \frac{10}{3} - \frac{25}{9} = \frac{5}{9}.$$

Средние квадратические отклонения:

$$\sigma_X = \sqrt{D_X} = \frac{\sqrt{2}}{3}; \quad \sigma_Y = \sqrt{D_Y} = \frac{\sqrt{5}}{3}.$$

Второй смешанный начальный момент:

$$\alpha_{11} = 1\cdot 1\cdot\frac{2}{12} + 1\cdot 2\frac{1}{12} + 1\cdot 3\frac{1}{12} + 2\cdot 1\cdot\frac{4}{12} + 2\cdot 2\frac{3}{12} + 2\cdot 3\frac{1}{12} = \frac{33}{12} = \frac{11}{4}.$$

Ковариация:

$$K_{XY} = \alpha_{11} - m_X m_Y = \frac{11}{4} - \frac{25}{9} = -\frac{1}{36}.$$

Линейный коэффициент корреляции:

$$\rho = \frac{K_{XY}}{\sigma_X\sigma_Y} = \frac{-1/36}{\dfrac{\sqrt{2}}{3}\dfrac{\sqrt{5}}{3}} = -\frac{1}{4\sqrt{10}} \approx -0{,}0791 \approx -0{,}079.$$

Пример 3. Вычисление ассоциативного коэффициента детерминации as и ассоциативного коэффициента корреляции as_c (§ 4.2):

$$as=\sum_{i,j}\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}p_{ij}=\frac{\left|\frac{2}{12}-\frac{4}{12}\frac{6}{12}\right|}{\sqrt{\frac{4}{12}\frac{8}{12}\frac{6}{12}\frac{6}{12}}}\frac{2}{12}+$$

$$\frac{\left|\frac{1}{12}-\frac{4}{12}\frac{4}{12}\right|}{\sqrt{\frac{4}{12}\frac{8}{12}\frac{4}{12}\frac{8}{12}}}\frac{1}{12}+$$

$$+\frac{\left|\frac{1}{12}-\frac{4}{12}\frac{2}{12}\right|}{\sqrt{\frac{4}{12}\frac{8}{12}\frac{2}{12}\frac{10}{12}}}\frac{1}{12}+\frac{\left|\frac{4}{12}-\frac{8}{12}\frac{6}{12}\right|}{\sqrt{\frac{8}{12}\frac{4}{12}\frac{6}{12}\frac{6}{12}}}\frac{4}{12}+\frac{\left|\frac{3}{12}-\frac{8}{12}\frac{4}{12}\right|}{\sqrt{\frac{4}{12}\frac{8}{12}\frac{4}{12}\frac{8}{12}}}\frac{3}{12}+$$

$$\frac{\left|\frac{1}{12}-\frac{8}{12}\frac{2}{12}\right|}{\sqrt{\frac{8}{12}\frac{4}{12}\frac{2}{12}\frac{10}{12}}}\frac{1}{12}=$$

$$=\frac{1}{12}\left(0+\frac{1}{8}+\frac{1}{2\sqrt{10}}+0+\frac{3}{8}+\frac{1}{2\sqrt{10}}\right)=0,068.$$

$$as=0,068.$$

(19)

Ассоциативный коэффициент корреляции (в формуле для ассоциативного коэффициента детерминации снимаем знак модуля):

$$as_c=\frac{1}{12}\left(-\frac{1}{8}+\frac{1}{2\sqrt{10}}+\frac{3}{8}-\frac{1}{2\sqrt{10}}\right)=\frac{1}{48}\approx 0,021.$$

$$as_c=0,021.$$

(20)

Пример 4. Вычисление контингенциального коэффициента детерминации co и контингенциального коэффициента корреляции co_c (§ 5.4.) для распределения, определенного таблицей 1. Используем результаты вычислений в примере 3.

$$co=\sum_{i,j}\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}p_{ij}=$$

$$= 0 + \frac{\left|-4/12^2\right|}{\frac{1}{12}\left(1 + 2\frac{1}{12} - 2\frac{4}{12} - 2\frac{4}{12}\right) + \frac{4}{12}\frac{4}{12}}\frac{1}{12} +$$

$$\frac{4/12^2}{\frac{1}{12}\left(1 + 2\frac{1}{12} - 2\frac{4}{12} - 2\frac{2}{12}\right) + \frac{4}{12}\frac{2}{12}}\frac{1}{12} +$$

$$+0 + \frac{4/12^2}{\frac{3}{12}\left(1 + 2\frac{3}{12} - 2\frac{8}{12} - 2\frac{4}{12}\right) + \frac{8}{12}\frac{4}{12}}\frac{3}{12} +$$

$$\frac{\left|-4/12^2\right|}{\frac{1}{12}\left(1 + 2\frac{1}{12} - 2\frac{8}{12} - 2\frac{2}{12}\right) + \frac{8}{12}\frac{2}{12}}\frac{1}{12} =$$

$$= \frac{1}{12}\left(\frac{2}{7} + \frac{2}{5} + \frac{6}{7} + \frac{2}{5}\right) = \frac{17}{105} \approx 0{,}162.$$

$$co = 0{,}162.$$

(21)

Контингенциальный коэффициент корреляции (снимаем знак модуля в формуле для контингенциального коэффициента детерминации). Используем результаты предыдущих вычислений.

$$co_c = \frac{1}{12}\left(-\frac{2}{7} + \frac{2}{5} + \frac{6}{7} - \frac{2}{5}\right) = \frac{1}{21} \approx 0{,}048.$$

$$co_c = 0{,}048.$$

(22)

Для распределения, представленного таблицей 1, вычислены 3 коэффициента детерминации и 3 коэффициента корреляции. Для сравнения сведем их вместе.

Коэффициенты детерминации: $\rho^* = 0{,}177; \quad as = 0{,}068; \quad co = 0{,}162$

(максимальный, ассоциативный, контингенциальный).

Коэффициенты корреляции: $\rho = -0{,}079; \quad as_c = 0{,}021; \quad co_c = 0{,}048$

(линейный, ассоциативный, контингенциальный).

Глава 3. Новые коэффициенты детерминации и корреляции для исследования зависимости между двумя случайными величинами и связи между количественными, качественными и смешанными признаками

В предыдущих главах были достаточно обстоятельно рассмотрены ранее известные коэффициенты связи между случайными величинами (признаками). В этой главе вводятся новые, ранее неизвестные, коэффициенты связи между двумя случайными величинами. Этих коэффициентов много, фактически бесконечное количество. Для их образования формулируются общие принципы. Все они обладают общим важным свойством: равенство нулю коэффициента связи обеспечивает независимость изучаемых случайных величин. Кроме того, они имеют достаточно простую конструкцию, что обеспечивает простоту применения.

§ 3.1. Общие принципы конструирования коэффициентов связи и их применения. Спектр коэффициентов.

Коэффициент связи для случайных величин образуется из элементарных коэффициентов связи для событий, состоящих в принятии этими случайными величинами отдельных значений. Элементарные парные коэффициенты связи событий, например, рассмотренные в §§ 1.2., 1.3., генерируют функциональные ядра K_{ij} или $K(x,y)$, из которых с помощью операторов суммирования и интегрирования строятся коэффициенты связи случайных величин с целью усреднения элементарных коэффициентов связи отдельных значений.

Для дискретных случайных величин коэффициент связи δ_λ, называемый коэффициентом детерминации, конструируется по формуле

$$\delta_\lambda = \left[\sum_i \sum_j \left| K_{ij} \right|^\lambda p_{ij} \right]^{1/\lambda} . \qquad (1)$$

Так как новых коэффициентов связи создано много, то для них применяются обозначения символами, позволяющими узнать и запомнить эти коэффициенты по первым буквам названия ядра. Например, для ассоциативного ядра соответствующий коэффициент обозначается as_λ. Для контингенциального ядра соответствующий коэффициент обозначается co_λ и т.д.

В формуле (1): $p_{ij} = P\left(X = x_i, Y = y_j \right)$; $i, j = 1, 2, ...$ – закон распределения двумерной дискретной случайной величины $\left(X, Y \right)$. Кроме усреднения по

вероятностям в формуле (1) применяется степенное усреднение с помощью вещественного параметра λ , $\lambda > 0$ (степень усреднения).

Ассоциативное ядро имеет вид

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}.$$

(2)

Контингенциальное ядро выражается формулой

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1-2p_{ij}+2p_{i\cdot}+2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}.$$

(3)

Для непрерывных случайных величин коэффициент детерминации конструируется по формуле

$$\delta_\lambda = \left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left|K\left(x,y\right)\right|^\lambda f_{XY}\left(x,y\right)dxdy\right]^{1/\lambda}.$$

(4)

Ассоциативное ядро в непрерывном случае имеет вид

$$K\left(x,y\right) = \frac{F_{XY}\left(x,y\right)-F_X\left(x\right)F_Y\left(y\right)}{\sqrt{F_X\left(x\right)\left(1-F_X\left(x\right)\right)F_Y\left(y\right)\left(1-F_Y\left(y\right)\right)}}$$

(5)

Аналогичное выражение в непрерывном случае имеет и контингенциальное ядро.

Для случайных величин общего вида коэффициент детерминации конструируется по формуле

$$\delta_\lambda = \left[M\left|K\left(X,Y\right)\right|^\lambda\right]^{1/\lambda}.$$

(6)

Формула (6) в общем случае выражается интегралом Лебега-Стилтьеса по вероятностной мере P.

$$\delta_\lambda = \left[\iint_D\left|K\left(x,y\right)\right|^\lambda dP\right]^{1/\lambda}.$$

(7)

Для ассоциативного ядра эти коэффициенты по-прежнему обозначаются as_λ, а для контингенциального ядра они обозначаются co_λ.

Здесь $F_{XY}(x,y)-$ функция распределения двумерной случайной величины (X,Y), $f_{XY}(x,y)-$ее плотность вероятности, M – оператор математического ожидания.

Эти коэффициенты связи δ_λ целесообразно назвать коэффициентами детерминации, так как их основное назначение фиксировать наличие или отсутствие связи между случайными величинами или признаками, принимающими числовые значения или имеющими качественные градации. Свойства ядер K определяют и свойства коэффициентов детерминации и корреляции, в частности:

1) $0 \le \delta_\lambda \le 1$;

2) необходимым и достаточным условием независимости случайных величин является равенство нулю коэффициента детерминации. По величине коэффициента можно судить о величине зависимости и сравнивать по этому признаку зависимости между собой.

Из теории неравенств [40] известно, что $\rho_{\lambda_1} < \rho_{\lambda_2}$ при $\lambda_1 < \lambda_2$. Для практики интересны случаи $\lambda = 1,2.$ Может применяться также среднее геометрическое элементарных коэффициентов связи при их отличии от нуля. Среднее геометрическое является предельным случаем среднего степенного при $\lambda \to 0$.

Формулы (1) и (4) дают несколько спектров коэффициентов детерминации. Наряду с ними можно применять и другие, используя другие приемы усреднения. Отметим некоторые.

Для конкретной выборки построим вариационный ряд из значений модулей ядер $K_{ij} = K\left(p_{ij}\right)$, где $p_{ij}-$относительные частоты двумерной выборки. Вариационный ряд должен строится с учетом вероятностей p_{ij} появления этих ядер как элементарных коэффициентов корреляции событий $X = x_i; Y = y_j$. Например, если вероятности выражены десятичными дробями с двумя цифрами после запятой, то умножаем их все на число 100 и превращаем в веса, с которыми модули ядер входят в вариационный ряд, то есть повторяем в вариационном ряде модуль ядра столько раз, каков его вес.

Можно поступить иначе. Выписываем вариационный ряд из модулей ядер вместе с вероятностями их появления $|K|_{(1)}, p_1; |K|_{(2)}, p_2; ...; |K|_{(l)}, p_l$. Пусть, например, требуется найти медиану. Последовательно суммируем вероятности $p_1 + p_2 + ...$ до тех пор пока не наберем в сумме 0,5. Пусть $p_1 + ... + p_{i-1} < 0,5$, а $p_1 + ... + p_{i-1} + p_i \ge 0,5.$ Тогда $|K|_{(i)} = med.$

На основе вариационного ряда можно построить следующие числовые характерстики:

1. $med\left|K\left(p_{ij}\right)\right|;\quad i,j=1,2,...-$ выборочная медиана – вариант среднего элементарных коэффициентов связи.
2. $\left|K\right|_{\min},\left|K\right|_{\max}-$крайние значения элементарных коэффициентов связи.
3. $t_R=\frac{1}{2}\left(\left|K\right|_{\min}+\left|K\right|_{\max}\right)-$ полусумма крайних значений элементарных коэффициентов связи.
4. $t_q=\frac{1}{2}\left(\left|K\right|_{1/4}+\left|K\right|_{3/4}\right)-$ полусумма квартилей элементарных коэффициентов связи.

Если в формулах (1),(4) снять знак модуля и положить $\lambda=1$, то коэффициенты детерминации переходят в коэффициенты корреляции

$$\delta_c=\left[\sum_i\sum_j K\left(p_{ij}\right)p_{ij}\right] \quad ; \tag{7}$$

$$\delta_c=\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}K\left(x,y\right)f_{XY}\left(x,y\right)dxdy\right]. \tag{8}$$

Эти коэффициенты корреляции обладают свойствами, аналогичными свойствам линейного коэффициента корреляции (§1.1), в частности, $-1\le\delta_c\le1$. В отличие от коэффициентов детерминации равенство нулю коэффициента корреляции не обеспечивает независимость исследуемых случайных величин.

§ 3.2. Общие требования к коэффициентам связи. Классификация связей

Новые коэффициенты связи позволяют проводить исследование связи случайных величин системно, привлекая новые и старые коэффициенты.

Коэффициент корреляции δ_c может иметь любой знак и может быть равным нулю, поэтому для конкретных коэффициентов корреляции целесообразно выделить 3 класса связей:

1) отрицательная – при $\delta_c<0$; в среднем события $X=x_i$ и $Y=y_J\ \left(i,j=1,2,...\right)$ совместно появляются реже, чем в случае независимости случайных величин X,Y.

2) уравновешенная – при $\delta_c = 0$; в среднем события $X = x_i$ и $Y = y_J \left(i, j = 1,2,...\right)$ совместно появляются с теми же вероятностями, что и в случае независимости случайных величин X, Y.

3) положительная – при $\delta_c > 0$. В среднем события $X = x_i$ и $Y = y_J \left(i, j = 1,2,...\right)$ совместно появляются чаще, чем в случае независимости случайных величин X, Y.

Величина коэффициента детерминации δ_λ указывает на силу связи, а знак коэффициента корреляции ρ указывает на принадлежность связи одному из классов.

Для исследования связи можно параллельно привлекать и линейный коэффициент корреляции ρ, который учитывает значения случайных величин, чего не делает коэффициент детерминации δ_λ. По его величине аналогично можно выделить еще 3 класса связей и сопоставить их с ранее выделенными.

Ко всем коэффициентам корреляции или детерминации ρ предъявляются следующие требования:

1. Нормированность: $|\rho| \le 1$.
2. Достижимость крайних пределов изменения $|\rho|$: $|\rho| = 0$; $|\rho| = 1$.
3. Взаимно однозначная связь крайних значений $|\rho| = 0$ с независимостью случайных величин X, Y и $|\rho| = 1$ - с их полной зависимостью, например, вида $Y = X$.

Не все коэффициенты ρ, известные в науке и применяющиеся на практике, удовлетворяют всем перечисленным требованиям. 1-е требование обычно выполняется, а 2-е,3-е – не всегда. Особенно трудно удовлетворить 3-му требованию.

Для того, чтобы третье требование выполнялось, требуется в конструкции коэффициента использовать необходимые и достаточные условия независимости случайных величин.

Для произвольных случайных величин они имеют вид

$$F_{XY}\left(x, y\right) - F_X\left(x\right) F_Y\left(y\right) = 0; \quad \forall x, y.$$

(9)

В формуле (9): $F_{XY}\left(x, y\right)-$ функция распределения двумерной случайной величины

$\left(X, Y\right)$, а $F_X\left(x\right), F_Y\left(y\right)-$ функции распределения ее компонент X, Y.

Замечание. В дальнейшем для упрощения записей все коэффициенты детерминации δ_λ при $\lambda = 1$ будем обозначать просто δ без знака λ. Например, as, co и т. д.

Принципиально различных новых коэффициентов будет представлено в следующих главах 8 видов: ассоциативный, контингенциальный, комбинированный ассоциативный, комбинированный контингенциальный, комбинированный мед-ас, комбинированный мед-конт, предельный, дефектологический.

Комбинированные коэффициенты применяются для исследования числовых признаков, остальные - и для числовых, и для качественных, и для смешанных. Рассмотрены случаи дискретных и непрерывных распределений. Приведенные многочисленные примеры позволяют сопоставить величины этих коэффициентов между собой и с известными ранее коэффициентами, в частности с линейным и максимальным.

Глава 4. Ассоциативный коэффициент детерминации

Ассоциативный коэффициент детерминации – первый из восьми новых коэффициентов, представленных в этой книге.

§ 4.1. Ассоциативный коэффициент детерминации для любых, в частности непрерывных, случайных величин.

В § 3.1. приведена формула, выражающая общую конструкцию коэффициентов детерминации

$$\delta_\lambda = \left[M \left| K\left(X,Y\right) \right|^\lambda \right]^{1/\lambda}$$

(1)

В соответствии с этой формулой предлагается конструкция парных коэффициентов детерминации с помощью ассоциативного ядра

$$K = \frac{P\left(AB\right) - P\left(A\right)P\left(B\right)}{\sqrt{P\left(A\right)\left(1-P\left(A\right)\right)P\left(B\right)\left(1-P\left(B\right)\right)}},$$

рассматриваемого как коэффициент корреляции между событиями A, B (§1.2, §1.4).

В качестве вероятностей событий здесь взяты функции распределения случайных величин. Соответствующий коэффициент детерминации назван ассоциативным и обозначен символом as_λ

$$as_\lambda = \left[M\left(\frac{\left| F_{XY}(X,Y) - F_X(X)F_Y(Y) \right|}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_Y(Y)\left(1-F_Y(Y)\right)}} \right)^\lambda \right]^{1/\lambda}; \qquad \lambda > 0.$$

(2)

При $\lambda = 1$ этот коэффициент обозначается просто as.

Здесь символ M – оператор математического ожидания. Эта формула выражает степенное среднее с помощью операции математического ожидания.

Если двумерная случайная величина – непрерывная с плотностью вероятности $f_{XY}(x,y)$, то формула (2) подробнее может быть записана в виде

$$as_\lambda = \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{\left| F_{XY}(x,y) - F_X(x)F_Y(y) \right|}{\sqrt{F_X(x)\left(1-F_X(x)\right)F_Y(y)\left(1-F_Y(y)\right)}} \right)^\lambda f_{XY}(x,y)\,dxdy \right]^{1/\lambda};$$

(3) $\lambda > 0.$

Множители знаменателя подкоренного выражения в формуле (3) могут обращаться в нуль. Тогда и числитель дроби под интегралом равен нулю, так как сама дробь является модулем линейного коэффициента корреляции между индикаторами событий $X < x;\quad Y < y$ и потому ограничена, в силу того, что находится в пределах промежутка $[0;1]$. В соответствующих точках дробь имеет устранимый разрыв. Пусть, например, $F_X(x_0) = 0$. Тогда $(X < x_0) = \varnothing;$

$$F_{XY}(x_0, y) = P(X < x_0, Y < y) = P(\varnothing, Y < y) = P(\varnothing) = 0.$$

Все разрывы предполагаются устраненными.

Для практики интересны случаи $\lambda = 1$ и $\lambda = 2$. По сравнению со случаем $\lambda = 1$ случай $\lambda = 2$ удобней тем, что снимается знак модуля и выражение под знаком интеграла становится аналитическим.

Если двумерная случайная величина – дискретная с законом распределения $P(X = x_i; Y = y_j) = p_{ij};\quad (i, j = 1,2,...)$, то общая формула (2) для ассоциативного коэффициента детерминации принимает вид

$$as_\lambda = \left[\sum_i \sum_j \left(\frac{\left| F_{XY}(x_i, y_j) - F_X(x_i)F_Y(y_j) \right|}{\sqrt{F_X(x_i)\left(1-F_X(x_i)\right)F_Y(y_j)\left(1-F_Y(y_j)\right)}} \right)^\lambda p_{ij} \right]^{1/\lambda}.$$

(4)

Знаменатель дроби в этой формуле может обращаться в нуль. При этом и числитель обращается в нуль, поэтому требуется доопределение того, как понимается дробь в случае неопределенности. Рассмотрим возможные случаи.

$F_X(x_1) = P(X < x_1) = P(\varnothing) = 0.$ При этом

$$F_{XY}(x_1, y_j) = P(X < x_1, Y < y_j) = P(\varnothing, Y < y_j) = 0; \quad j = 1,2,...$$

Пусть $j = 1$. События $\varnothing$ и $\varnothing$ по аналогии с любыми одинаковыми событиями A и A следует считать полностью зависимыми. Линейный коэффициент корреляции между ними равен 1. Этот коэффициент корреляции и есть та дробь в формуле (4), о которой идет речь. Пусть теперь $j = 2,3,...$ В этом случае событие $Y < y_j$ отлично от невозможного. Рассмотрим условную

вероятность $P(\varnothing / Y < y_j) = P(\varnothing, Y < y_j) / P(Y < y_j) = 0.$ Это означает, что события $\varnothing$ и $Y < y_j$ являются независимыми. Линейный коэффициент корреляции между ними равен нулю. Так как он есть дробь в формуле (4), то это и является основанием того, чтобы эту дробь принять равной нулю. Аналогично решается вопрос о величине дроби, когда возникает неопределенность в случае $F_Y(y_0) = 0$.

Свойства коэффициентов детерминации.

Свойство 1. Если случайные величины X, Y независимы, то $as_\lambda = 0$.

Это свойство вытекает из того, что в случае независимости выполняются равенства

$$F_{XY}(x, y) - F_X(x) F_Y(y) = 0; \quad \forall x, y,$$

а поэтому числитель в формуле (2) равен нулю.

Свойство 2. Если $as_\lambda = 0$, то случайные величины X, Y независимы.

Для простоты ограничимся рассмотрением непрерывных случайных величин. Тогда заметим, что подынтегральная функция в формуле (3) неотрицательна. Если интеграл от неотрицательной функции равен нулю, то подынтегральная функция равна нулю почти везде. Тогда числитель равен нулю и, поэтому равенства (1) выполняются почти везде, что и означает независимость X, Y с вероятностью единица.

Свойство 3.

$$0 \le as_\lambda \le 1.$$

(5)

Для доказательства заметим, что в формуле (3) выражение

$$\rho = \frac{F_{XY}(x,y) - F_X(x)F_Y(y)}{\sqrt{F_X(x)\left(1-F_X(x)\right)F_Y(y)\left(1-F_Y(y)\right)}}$$

(6)

является линейным коэффициентом корреляции ρ между индикаторами событий $X < x;\ \ Y < y$. Для линейного коэффициента корреляции ρ справедливы неравенства $-1 \le \rho \le 1$. Отсюда $0 \le |\rho|^{\lambda} \le 1$. Умножаем все части этих неравенств на плотность $f_{XY}(x,y)$ и интегрируя по всей плоскости, получаем неравенства $0 \le M|\rho|^{\lambda} \le 1$. Возводя все части этих неравенств в степень $1/\lambda$,приходим к равносильным неравенствам (5).

Свойство 4. Если $Y = X$, то $as_{\lambda} = 1$. (Полная зависимость).

В этом случае двумерный закон распределения становится одномерным. Из формулы (2) получаем

$$as_{\lambda} = \left[M\left(\frac{F_X(X) - F_X(X)F_X(X)}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_X(X)\left(1-F_X(X)\right)}} \right)^{\lambda} \right]^{1/\lambda} =$$

$$= \left[M\left(\frac{F_X(X)\left(1-F_X(X)\right)}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_X(X)\left(1-F_X(X)\right)}} \right)^{\lambda} \right]^{1/\lambda} = [M(1)]^{1/\lambda} = 1,$$

так как $F_{XY}(x,y) = P(X < x, X < x) = P(X < x) = F_X(x).$

Свойство 5.Если случайные величины X, Y связаны строго возрастающей функциональной зависимостью $Y = \varphi(X)$,то $as_{\lambda} = 1$. Предполагается, что функция φ определена для всех значений случайной величины X.

Для доказательства заметим, что в этом случае двумерный закон распределения

становится одномерным. Если случайная величина X приняла значение x, то случайная величина $\varphi(X)$ приняла значение $\varphi(x)$. События $X < x$ и $\varphi(X) < \varphi(x)$ эквивалентны в силу строгого возрастания функции φ. Тогда вероятности этих событий равны: $P(X < x) = P\left(\varphi(X) < \varphi(x)\right)$. Отсюда

следует, что $F_X(x) = F_Y(y)$, где $y = \varphi(x)$;

$$F_{XY}(x,y) = P\left(X < x, \varphi(X) < \varphi(x)\right) = P\left(X < x, X < x\right) = P\left(X < x\right) = F_X(x).$$

Тогда

$$as_\lambda = \left[M\left(\frac{\left| F_{XY}(X,Y) - F_X(X)F_Y(Y) \right|}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_Y(Y)\left(1-F_Y(Y)\right)}} \right)^\lambda \right]^{1/\lambda} =$$

$$= \left[M\left(\frac{F_X(X) - F_X(X)F_X(X)}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_X(X)\left(1-F_X(X)\right)}} \right)^\lambda \right]^{1/\lambda} =$$

$$= \left[M\left(\frac{F_X(X)\left(1-F_X(X)\right)}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_X(X)\left(1-F_X(X)\right)}} \right)^\lambda \right]^{1/\lambda} = [M(1)]^{1/\lambda} = 1.$$

Свойство 6. Если $as_\lambda = 1$, то случайная величина Y является строго возрастающей функцией $Y = \varphi(X)$ случайной величины X с вероятностью единица.

Для доказательства воспользуемся формулой (3), в которой соответствующее математическое ожидание равно 1:

$$M\left[\frac{\left| F_{XY}(X,Y) - F_X(X)F_Y(Y) \right|}{\sqrt{F_X(X)\left(1-F_X(X)\right)F_Y(Y)\left(1-F_Y(Y)\right)}} \right]^\lambda = 1.$$

(7)

Для краткости обозначим функцию, стоящую в квадратных скобках под знаком модуля через $K(x,y)$, Тогда формула (7) запишется короче в виде

$$M\left[\left| K(X,Y) \right| \right]^\lambda = 1.$$

(8)

В общем случае математическое ожидание в формуле (8) выражается двукратным интегралом Лебега-Стилтьеса по вероятностной мере:

$$\int_{R_2} \left|K\left(x,y\right)\right|^{\lambda} dP = 1.$$

(9)

Здесь R_2 – плоскость изменения переменных x, y .Интеграл в формуле (9) берется по той области $D \subset R_2$, в которой определена вероятностная мера P.

Подынтегральная функция удовлетворяет неравенствам $0 \le \left|K\left(x,y\right)\right|^{\lambda} \le 1$, так как $K\left(x,y\right)$ есть модуль линейного коэффициента корреляции между индикаторами событий $X < x, Y < y$, который заключен в промежутке [0;1]. Заметим также, что $\int_D dP = 1$. Разобьем область D на две части: $D = D_1 \cup D_2$. В D_1 имеет неравенство $\left|K\left(x,y\right)\right| < 1$, а в D_2 - равенство $\left|K\left(x,y\right)\right| = 1$. Тогда интеграл в формуле (9) разобьется на два интеграла

$$\int_{D_1} \left|K\left(x,y\right)\right|^{\lambda} dP + \int_{D_2} \left|K\left(x,y\right)\right|^{\lambda} dP = 1.$$

(10)

Так как во втором интеграле $\left|K\left(x,y\right)\right| = 1$, то формула (10) принимает вид

$$\int_{D_1} \left[K\left(x,y\right)\right]^{\lambda} dP + \mu\left(D_2\right) = 1.$$

(11)

здесь $\mu\left(D_2\right)$ – вероятностная мера множества D_2.

Интегрируем неравенство $\left|K\left(x,y\right)\right|^{\lambda} < 1$ по множеству D_1. Получаем

$\int_{D_1} \left|K\left(x,y\right)\right|^{\lambda} dP < \int_{D_1} dP = \mu\left(D_1\right)$. Таким образом,

$\int_{D_1} \left|K\left(x,y\right)\right|^{\lambda} dP + \mu\left(D_2\right) < \mu\left[D_1\right] + \mu\left(D_2\right) = 1$. Этот результат означает, что при $\mu\left(D_1\right) > 0$ равенство (11) нарушается, следовательно $\mu\left(D_1\right) = 0$. Итак, $\left|K\left(x,y\right)\right|^{\lambda} < 1$, а потому $\left|K\left(x,y\right)\right| < 1$ только с вероятностью 0. Соответственно

$|K(x,y)|=1$ с вероятностью 1. (12)

Проанализируем равенство (12). Для краткости обозначений введем события $A=(X<x);\ B=(Y<y)$. Тогда формула (12) запишется в виде

$$|K(x,y)|=\frac{|F_{XY}(X,Y)-F_X(X)F_Y(Y)|}{\sqrt{F_X(X)(1-F_X(X))F_Y(Y)(1-F_Y(Y))}}=$$

$$=\frac{|P(AB)-P(A)P(B)|}{\sqrt{P(A)(1-P(A))P(B)(1-P(B))}}=1.$$

(13)

В § 1.2. доказано, что при выполнении равенства (13) либо $B=A$, либо $B=\overline{A}$.

Рассмотрим оба случая.

1) Если $B=A$, то события $Y<y;\ \ X<x$ эквивалентны при любом x и любом y. Зафиксируем y. Тогда $F_X(x)=P(X<x)=P(Y<y)=const$, что невозможно, так как $F_X(x)-$ возрастающая функция. Это означает, что распределение Y не является независимым, а определяется распределением X, то есть $Y=\varphi(X)-$ является функцией от X. Эта функция – строго монотонная, так как в противном случае обратная функция будет неоднозначной и события $X<x$ и $\varphi(X)<\varphi(x)$ не будут эквивалентны. Функция $\varphi(x)$ не может быть строго убывающей, так как иначе неравенство $X<x$ эквивалентно неравенству $\varphi(X)>\varphi(x)$. Тогда

$$F_X(x)=P(X<x)=P(\varphi(X)>\varphi(x))=P(Y>y)=1-P(Y<y)=$$
$$1-F_Y(y).$$

Это означает, что функция $F_X(x)$ возрастает, чего быть не может.

2) Пусть теперь $B=\overline{A}$. Тогда события $Y<y$ и $X\geq x$ эквивалентны, следовательно

$P(X\geq x)=1-F_X(x)=P(Y<y)=F_Y(y)$. Это означает, что функция $F_Y(y)$ убывает, так как функция $1-F_X(x)$ убывает. Для функции

распределения убывание невозможно. Второй случай отпадает. Оба возможных случая проанализированы. Свойство 6 доказано.

Пример 1. Двумерная случайная величина (X,Y) распределена равномерно в треугольнике Δ с вершинами O(0;0), B(1;0), C(0;1). Вычислим коэффициент детерминации as ; $(\lambda = 1)$.

Двумерная плотность вероятности в этом случае определяется формулами

$f_{XY}(x,y) = 2$ при $(x;y) \in \Delta$; $f_{XY}(x,y) = 0$ при $(x;y) \notin \Delta$.

Из этих формул следует, что двумерная функция распределения $F_{XY}(x,y) = 2xy$ при $(x,y) \in \Delta$. Одномерные плотности распределения компонент X,Y соответственно выражаются формулами $f_X(x) = 2(1-x)$; $f_Y(y) = 2(1-y)$; $0 \le x \le 1; 0 \le y \le 1$.

Соответствующие функции распределения компонент выражаются формулами $F_X(x) = 2x - x^2$; $F_Y(y) = 2y - y^2$; $0 \le x \le 1; 0 \le y \le 1$. Тогда по формуле (3) находим

$$as = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{\left|F_{XY}(x,y) - F_X(x)F_Y(y)\right|}{\sqrt{F_X(x)\left(1-F_X(x)\right)F_Y(y)(1-)F_Y(y)}}\right)f_{XY}(x,y)\,dxdy =$$

$$= 2\int_0^1 dx \int_0^{1-x} \frac{(2+xy-2x-2y)\sqrt{xy}}{(1-x)(1-y)\sqrt{(2-x)(2-y)}}dy = 2\int_0^1 \varphi(x)dx.$$

Здесь

$$\varphi(x) = \frac{4\sqrt{x}}{(1-x)\sqrt{2-x}}\,arctg\sqrt{\frac{1-x}{1+x}} - \frac{x\sqrt{x}}{(1-x)\sqrt{2-x}}\ln\frac{\sqrt{1+x}+\sqrt{1-x}}{\sqrt{1+x}-\sqrt{1-x}} -$$

$$-\sqrt{x(2-x)}\sqrt{\frac{1+x}{1-x}}.$$

Составим таблицу значений функции $\varphi(x)$ с шагом $h = 1/12$.

x	$\varphi(x)$	x	$\varphi(x)$
0	0	$7/12$	0,12942
$1/12$	0,18199	$8/12$	0,09909

$2/12$	0,21203	$9/12$	0,06862
$3/12$	0,21465	$10/12$	0,03974
$4/12$	0,20325	$11/12$	0,01493
$5/12$	0,18336	1	0
$6/12$	0,15805	–	–

С помощью формулы Симпсона численного интегрирования вычисляем интеграл

$\int_0^1 \varphi(x)dx = \frac{h}{3}\left[y_0 + y_{12} + 4\left(y_1 + y_3 + ... + y_{11}\right) + 2\left(y_2 + y_4 + ... + y_{10}\right)\right]$. Здесь

$y_k = \varphi(x_k);\quad k = 0,1,...,12.\ x_k = kh$. Тогда $\int_0^1 \varphi(x)dx = 0,12767.$ Отсюда

$$as = 2\int_0^1 \varphi(x)dx = 2 \cdot 0,12767 = 0,25535.$$

Итак, ассоциативный коэффициент детерминации равен $as = 0,255$.

Пример 2.

Полезно сравнить этот коэффициент as в примере 1 с линейным коэффициентом корреляции ρ, который в этом случае вычисляется легко.

$$\rho = \frac{K_{XY}}{\sigma_X \sigma_Y};\ \ K_{XY} = \alpha_{11} - m_X m_Y;\ \ \alpha_{11} = M\left[XY\right];\ \ \sigma_X = \sqrt{D_X};\sigma_Y = \sqrt{D_Y};$$

$$D_X = \alpha_{2X} - m_X^2;\ \ \alpha_{2X} = M\left[X^2\right].$$

Используем результаты вычислений из примера 1.

$m_X = \int_0^1 x2(1-x)dx = \frac{1}{3};$ по симметрии $m_Y = \frac{1}{3}$.

$$\alpha_{2X} = \int_0^1 x^2 2(1-x)\,dx = \frac{1}{6};\quad D_X = \alpha_2 - m_X^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18};$$

$\sigma_X = \sqrt{1/18} = \dfrac{1}{3\sqrt{2}}$; по симметрии $\sigma_Y = \dfrac{1}{3\sqrt{2}}$.

$\alpha_{11} = 2\int\limits_0^1 dx \int\limits_0^{1-x} xydy = \dfrac{1}{12}$: $K_{XY} = \dfrac{1}{12} - \dfrac{1}{3}\dfrac{1}{3} = -\dfrac{1}{36}$; $\rho = \dfrac{-1/36}{1/18} = -\dfrac{1}{2}$.

Итак, $\rho = -0,5$.

Сравнивая $|\rho| = 0,5$ и $as = 0,255$, заключаем, что ассоциативный коэффициент детерминации оценивает величину связи в этом случае более жестко, чем модуль линейного коэффициента корреляции.

§ 4.2. Ассоциативные коэффициенты детерминации и корреляции для дискретных случайных величин. Примеры.

В соответствии с общей схемой построения коэффициентов детерминации и корреляции, развитой в § 3.1. здесь в качестве ядра используется линейный парный коэффициент корреляции между событиями (ассоциативное ядро)

$$K\left(p_{ij}\right) = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}.$$

Пусть закон распределения двумерной случайной величины $(X;Y)$ задан формулой $P\left(X = x_i; Y = y_j\right) = p_{i,j}; \quad i = 1,2,...,m; j = 1,2,...,k;$ (m,k могут быть бесконечными). Пусть далее $p_{i\cdot} = \sum_j p_{ij}; \; p_{\cdot j} = \sum_i p_{ij}$.

Формулы $P\left(X = x_i\right) = p_{i\cdot}; P\left(Y = y_j\right) = p_{ij}$ определяют законы распределения компонент двумерной случайной величины.

Ассоциативный коэффициент детерминации для дискретных случайных величин определяется формулой

$$as_\lambda = \left[\sum_i \sum_j \left(\frac{\left|p_{ij} - p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}\right)^\lambda p_{ij}\right]^{1/\lambda}. \qquad (1)$$

Параметр λ - степень среднего; $0 < \lambda < +\infty$. Для практики интересны случаи $\lambda = 1$ и $\lambda = 2$. Если $\lambda = 1$. то коэффициент обозначается проще: as.

Из теории неравенств для средних [40] известно, что $as_\lambda < as_\mu$ при $\lambda < \mu$.

Свойства коэффициента as_λ.

Свойства коэффициента детерминации as_λ определяются структурой множителей каждого слагаемого в формуле (1) . Введем события $A_i = \left(X = x_i\right); B_j = \left(Y = y_j\right); i, j = 1,2,...$ Тогда

$p_{i\cdot} = P\left(X = x_i\right); p_{\cdot j} = P\left(Y = y_j\right).$

$$\frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}} = \frac{P\left(A_iB_j\right) - P\left(A_i\right)P\left(B_j\right)}{\sqrt{P\left(A_i\right)\left(1-P\left(A_i\right)\right)P\left(B_j\right)\left(1-P\left(B_j\right)\right)}} = \rho_{ij}$$

(2)

есть коэффициент корреляции между событиями A_i, B_j, который определяется как линейный коэффициент корреляции между индикаторами I_{A_i}, I_{B_j} этих событий (§ 1.2.).

Из теории вероятностей известно, что линейный коэффициент корреляции нормирован неравенствами $-1 \le \rho_{ij} \le 1$. Кроме того, по структуре числителя формулы (2) видно, что он равен нулю тогда и только тогда, когда события A_i, B_j независимы, так как равенство $P\left(A_iB_j\right) - P\left(A_i\right)P\left(AB_j\right) = 0$ является необходимым и достаточным условием независимости событий A_i, B_j. Все степени модулей коэффициентов корреляции ρ_{ij} далее усредняются с помощью вероятностей p_{ij} на основе линейной выпуклой комбинации (1).

Таким образом, as_λ является средне степенным модуля линейного парного коэффициента корреляции ρ_{ij} между отдельными значениями или отдельными градациями признаков дискретных случайных величин X, Y.

$$as_\lambda = \left(\sum_i\sum_j \left|\rho_{ij}\right|^\lambda p_{ij}\right)^{1/\lambda}.$$

(3)

Отметим и докажем свойства ассоциативного коэффициента детерминации as_λ.

Свойство 1. $0 \le as_\lambda \le 1$**.**

Действительно, $\left(\rho_\lambda\right)^\lambda = \sum_i\sum_j \left|\rho_{ij}\right|^\lambda p_{ij} \le \sum_i\sum_j 1\cdot p_{ij} = 1.$

Свойство 2. Если дискретные случайные величины X, Y независимы, то $as_\lambda = 0$.

Действительно, равенства

$$p_{ij} - p_{i\cdot} p_{\cdot j} = 0; \forall i, j$$

(4)

являются необходимым и достаточным условием независимости X, Y. Тогда все числители в формуле (1) равны 0. Следовательно, $as_\lambda = 0$.

Свойство 3. Если $as_\lambda = 0$, то X, Y независимы.

Действительно, сумма неотрицательных слагаемых равна нулю тогда и только тогда, когда все слагаемые равны нулю. Поэтому все числители в формуле (1) равны нулю. Следовательно, выполняются равенства (4).

Свойство 4. Если $Y = X$ (полная зависимость, по терминологии Д. Кендалла), то $as_\lambda = 1$.

Действительно, в этом случае $p_{ij} = P\left(X = x_i; X = x_j\right) = \begin{cases} 0; & i \neq j \\ p_{i\cdot}; & i = j \end{cases}$;

$p_{\cdot j} = p_{i\cdot}$. Тогда

$$p_{ij} - p_{i\cdot} p_{\cdot j} = p_{i\cdot} - \left(p_{i\cdot}\right)^2 = p_{i\cdot}\left(1 - p_{i\cdot}\right);$$

$$\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)} = p_{i\cdot}\left(1 - p_{i\cdot}\right).$$

$$\rho_\lambda = \sum_i \sum_j 1 \cdot p_{ij} = \sum_i p_{i\cdot} = 1.$$

Свойство 5. Если $Y = \varphi\left(X\right)$, где $\varphi\left(x\right)$ - функция, определенная для всех значений X, выполняющая взаимно однозначное отображение (в частности, строго монотонная), то $as_\lambda = 1$.

Действительно, в этом случае

$$p_{\cdot i} = P\left(Y = y_i\right) = P\left(\varphi\left(X\right) = y_i\right) = P\left(X = x_i\right) = p_{i\cdot};$$

$$p_{ij} = P\left(X = x_i; \varphi\left(X\right) = y_j\right) = P\left(X = x_i; X = x_j\right) = \begin{cases} 0; & i = j \\ p_{i\cdot}; & i = j \end{cases}.$$

Здесь $y_i = \varphi\left(x_i\right)$. Приходим к случаю, равносильному $Y = X$.

Свойство 6. Если $as_\lambda = 1$, то случайная величина Y является однозначной функцией случайной величины X ,имеющей однозначную обратную функцию.

Действительно, в этом случае все множители $\dfrac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}$ в слагаемых формулы (1) равны 1, так как все они неотрицательны и не превосходят 1, а другие множители p_{ij} положительны и в сумме равны 1. Из теории коэффициента корреляции между событиями (§ 1.2.) известно, что в этом случае или $B_j = A_i$, или $B_j = \overline{A}_i$. В любом случае значения случайной величины Y однозначно определяются значениями случайной величины X, так как $P\left(B_j\right) = P\left(Y = y_j\right) = p_{\cdot j} = P\left(A_i\right) = P\left(X = x_i\right) = p_{i\cdot}$ или $P\left(B_j\right) = 1 - p_{i\cdot}$.

Свойство 7. Коэффициент детерминации as_λ не изменяется, если случайные величины X, Y заменить функциями $\varphi\left(X\right), \psi\left(Y\right)$, где $\varphi\left(x\right), \psi\left(y\right)$ – однозначные функции своих переменных с однозначными обратными функциями.

Действительно, это следует из того, что $p_{i\cdot} = P\left(X = x_i\right) = P\left(\varphi\left(X\right) = \varphi\left(x_i\right)\right)$; $p_{\cdot j} = P\left(Y = y_j\right) = P\left(\psi\left(Y\right) = \psi\left(y_j\right)\right)$; $p_{ij} = P\left(X = x_i, Y = y_j\right) = P\left(\varphi\left(X\right) = \varphi\left(x_i\right), \psi\left(Y\right) = \psi\left(y_j\right)\right)$.

Приложения к математической статистике.

Статистическую оценку $\hat{\rho}_\lambda$ ассоциативного коэффициента детерминации ρ_λ получим, заменив в формуле (1) вероятности $p_{ij}, p_{i\cdot}, p_{\cdot j}$ соответствующими относительными частотами событий $\hat{p}_{ij} = \dfrac{n_{ij}}{n}$, $\hat{p}_{i\cdot} = \dfrac{n_{i\cdot}}{n}$, $\hat{p}_{\cdot j} = \dfrac{\hat{n}_{\cdot j}}{n}$:

$$\hat{as}_\lambda = \left[\sum_i \sum_j \left(\frac{\left|\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j}\right|}{\sqrt{\hat{p}_{i\cdot}\left(1-\hat{p}_{i\cdot}\right)\hat{p}_{\cdot j}\left(1-\hat{p}_{\cdot j}\right)}}\right)^\lambda \hat{p}_{ij}\right]^{1/\lambda}. \qquad (5)$$

Эта оценка $\hat{as}_\lambda$ является состоятельной оценкой as_λ, так как относительные частоты стремятся по вероятности к вероятностям соответствующих событий.

Пример 1. Триномиальное распределение определяется формулой

$$p_{ij} = \frac{n!}{i!\,j!(n-i-j)!} p_1^i p_2^j (1-p_1-p_2)^{n-i-j};$$

$$i, j = 0.1,...,n; \quad 0 < p_1 < 1; \quad 0 < p_2 < 1; \quad p_1 + p_2 < 1; \quad i + j \le n.$$

Рассмотрим случай $n = 2; \quad p_1 = p_2 = 1/4.$ Построим таблицу распределения (табл.1).

Таблица 1 триномиального распределения при $n = 2.$

$X \setminus Y \to$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00} = 1/4$	$p_{01} = 1/4$	$p_{02} = 1/16$	$p_{0\cdot} = 9/16$
1	$p_{10} = 1/4$	$p_{11} = 1/8$	$p_{12} = 0$	$p_{1\cdot} = 3/8$
2	$p_{20} = 1/16$	$p_{21} = 0$	$p_{22} = 0$	$p_{2\cdot} = 1/16$
$p_{\cdot j}$	$p_{\cdot 0} = 9/16$	$p_{\cdot 1} = 3/8$	$p_{\cdot 2} = 1/16$	1

Линейный коэффициент корреляции вычисляется по формуле

$$\rho = -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}} = -\frac{1}{3}$$; [3, с. 142].

Вычисления as на основе построенной таблицы дают

$$as = \frac{\left|\frac{1}{4} - \frac{9}{16}\frac{9}{16}\right|}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{9}{16}\frac{7}{16}}}\frac{1}{4} + \frac{\left|\frac{1}{4} - \frac{9}{16}\frac{3}{8}\right|}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{3}{8}\frac{5}{8}}}\frac{1}{4} + \frac{\left|\frac{1}{16} - \frac{9}{16}\frac{1}{16}\right|}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{1}{16}\frac{15}{16}}}\frac{1}{16} + \frac{\left|\frac{1}{4} - \frac{3}{8}\frac{9}{16}\right|}{\sqrt{\frac{3}{8}\frac{5}{8}\frac{9}{16}\frac{7}{16}}}\frac{1}{4} + \frac{\left|\frac{1}{8} - \frac{3}{8}\frac{3}{8}\right|}{\sqrt{\frac{3}{8}\frac{5}{8}\frac{3}{8}\frac{5}{8}}}\frac{1}{8} +$$

$$+ \frac{\left|\frac{1}{16} - \frac{1}{16}\frac{9}{16}\right|}{\sqrt{\frac{1}{16}\frac{15}{16}\frac{9}{16}\frac{7}{16}}}\frac{1}{16} = \frac{191}{2520} + \frac{9}{8\sqrt{105}} = 0{,}18558 \approx 0{,}186.$$

Из этого примера видим, что ассоциативный коэффициент детерминации as оценивает величину связи между случайными величинами более жестко, чем модуль линейного коэффициента корреляции $|\rho|$: $\frac{|\rho|}{\rho_1} = \frac{0{,}333}{0{,}186} \approx 1{,}8.$

Пример 2.

Пользуясь результатами примера 1, вычислим ассоциативный коэффициент корреляции

$$as_c = -\frac{191}{2520} + \frac{9}{8\sqrt{105}} = -0,0758 + 0,1098 = 0,034.$$

Этот результат означает, что в среднем связь между значениями случайных величин близка к варианту, когда случайные величины независимы. Здесь отрицательные и положительные элементарные коэффициенты корреляции практически уравновешивают друг друга.

Пример 3.

С помощью этих же результатов из примера 1 найдем медианный коэффициент детерминации as_m.

Построим вариационный ряд элементарных коэффициентов детерминации с учетом их вероятностей из таблицы распределения.

0,067 (вер. 0,125), 0,163 (вер. 0,5), 0,228 (вер. 0,125), 0,270 (вер. 0,25), 1 (вер.0).

Из анализа ряда следует, что $as_m = 0,163$. Этот показатель близок к показателю $as = 0,186.$

Глава 5. Контингенциальные коэффицинты детерминации и корреляции

Контингенциальный коэффициент детерминации – второй из восьми новых коэффициентов, представленных в этой книге.

§ 5.1. Контингенциальный коэффициент детерминации для дискретных случайных величин, «конти»

Применяем общую схему построения коэффициентов детерминации, развитую в § 3.1.

В этом случае используется ядро, генерируемое коэффициентом связи Юла между событиями (коэффициентом контингенции):

$$k = \frac{P(AB)\left(\overline{A}\overline{B}\right) - P\left(A\overline{B}\right)P\left(\overline{A}B\right)}{P(AB)\left(\overline{A}\overline{B}\right) + P\left(A\overline{B}\right)P\left(\overline{A}B\right)}.$$

(1)

Здесь он записывается в несколько измененном виде

$$K\left(p_{ij}\right)=\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}.$$

(2)

Контингенциальный коэффициент детерминации для дискретных случайных величин конструируется из ядра (2) по формуле

$$co_{\lambda}=\left[\sum_{i}\sum_{j}\left(\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}\right)^{\lambda}p_{ij}\right]^{1/\lambda}.$$

(3)

Здесь $P\left(X=x_i,Y=y_j\right)=p_{ij};\quad i,j=1,2,...$ – закон распределения двумерной дискретной случайной величины $\left(X,Y\right)$; $P\left(X=x_i\right)=p_{i\cdot};\quad P\left(Y=y_j\right)=p_{\cdot j};\quad i,j=1,2,...$ – законы распределения компонент двумерной случайной величины $\left(X,Y\right)$.

$$p_{i\cdot}=\sum_{j}p_{ij};\quad p_{j\cdot}=\sum_{i}p_{ij}.$$

Параметр λ - степень среднего; $0<\lambda<+\infty$. Для практики интересны случаи $\lambda=1$ и $\lambda=2$. При $\lambda=1$ контингенциальный коэффициент обозначается проще: co.

Из теории неравенств для средних [32] известно, что $co_{\lambda}<co_{\mu}$ при $\lambda<\mu$.

Свойства коэффициента детерминации co_{λ}.

Свойства коэффициента детерминации co_{λ} определяются структурой ядра, то есть множителей каждого слагаемого в формуле (3). Введем события $A_i=\left(X=x_i\right);B_j=\left(Y=y_j\right);i,j=1,2,...$ Тогда

$$p_{i\cdot}=P\left(X=x_i\right)=P\left(A_i\right);p_{\cdot j}=P\left(Y=y_j\right)=P\left(B_j\right);\quad p_{ij}=P\left(A_iB_j\right).$$

Введем величину

$$\tau_{ij}=\frac{P\left(A_iB_j\right)-P\left(A_i\right)P\left(B_j\right)}{P\left(A_iB_j\right)\left[1+2P\left(A_iB_j\right)-2P\left(A_i\right)-2P\left(B_j\right)\right]+P\left(A_i\right)P\left(B_j\right)}.$$

(4)

Это коэффициент контингенции между событиями A_i,B_j, исследованный в §1.3. Тогда

$$co_\lambda = \left(\sum_i \sum_j \left|\tau_{ij}\right|^\lambda p_{ij} \right)^{1/\lambda} \qquad (5)$$

есть средне степенное коэффициентов контингенции τ_{ij} между отдельными значениями, градациями признаков X, Y. Опираемся на свойства коэффициента контингенции (§1.3.).

Свойство 1. Нормированность: $0 \le co_\lambda \le 1$.

Действительно,

$\left(co_\lambda\right)^\lambda = \sum_i \sum_j \left|\tau_{ij}\right|^\lambda p_{ij} \le \sum_i \sum_j 1 \cdot p_{ij} = 1$, так как $0 \le \tau_{ij} \le 1$; $\sum_i \sum_j p_{ij} = 1$.

Свойство 2. Признак независимости случайных величин: $co_\lambda = 0$.

Для того, чтобы случайные величины X, Y были независимы, необходимо и достаточно равенство нулю коэффициента co_λ.

Пусть X, Y независимы. Тогда выполняется необходимое и достаточное условие независимости дискретных случайных величин

$$p_{ij} - p_{i\cdot} p_{\cdot j} = 0. \qquad i, j = 1, 2, \ldots \qquad (6)$$

Тогда все числители слагаемых в суммах формулы (3) равны нулю, поэтому $co_\lambda = 0$.

Обратно, пусть $co_\lambda = 0$. Тогда в силу неотрицательности все слагаемые в сумме (3) равны нулю. Отсюда следует выполнение равенств (6), что означает независимость случайных величин X, Y.

Свойство 3. Достижимость верхней границы. Если $Y = X$, то $co_\lambda = 1$.

Действительно, в этом случае в формуле (4) $P\left(A_i B_j\right) = p_{ij} = 0$ при $i \ne j$; $P\left(A_i B_j\right) = p_{ij} = P\left(A_i\right) = p_i$ при $i = j$.

Тогда $\tau_{ij} = \dfrac{P\left(A_i B_j\right) - P\left(A_i\right) P\left(A_j\right)}{P\left(A_i B_j\right)\left[1 + 2P\left(A_i B_j\right) - 2P\left(A_i\right) - 2P\left(A_j\right)\right] + P\left(A_i\right) P\left(A_j\right)} =$

$= \dfrac{1 - P\left(A_j\right)}{1 - P\left(A_j\right)} = 1$. По формуле (5) получаем $\tau_\lambda = \left[\sum_i 1 \cdot p_i \right]^{1/\lambda} = 1$.

Свойство 4. Если $Y=\varphi(X)$, где $\varphi(x)$ - функция, определенная для всех значений X, выполняющая взаимно однозначное отображение (в частности, строго монотонная), то $co_\lambda = 1$.

Действительно, в этом случае $p_{\cdot i} = P(Y=y_i) = P(\varphi(X)=y_i) = P(X=x_i) = p_{i\cdot}$;

Здесь $y_i = \varphi(x_i)$.

$$p_{ij} = P(X=x_i;\varphi(X)=y_j) = P(X=x_i;X=x_j) = \begin{cases} 0; & i \neq j \\ p_{i\cdot}; & i=j \end{cases}.$$

Приходим к случаю, равносильному $Y=X$.

Свойство 5.

Если $co_\lambda = 1$, то случайные величины X,Y связаны функционально-логической зависимостью (с вероятностью 1).

Для доказательства рассмотрим все возможные случаи. Пусть $co_\lambda = 1$. Тогда все слагаемые в сумме формулы (3) равны 1, так как $0 \le |\tau_{ij}| \le 1$ и $\sum_i \sum_j p_{ij} = 1$. Это в свою очередь означает, что $|\tau_{ij}| = 1$ для любых i, j. Если, хотя бы для одной пары значений i, j имело место неравенство $|\tau_{ij}| < 1$, то тогда $co_\lambda < 1$. Рассмотрим возможные случаи для значений τ_{ij}.

1. Пусть $\tau_{ij} = 1$. Следовательно,

$$\tau_{ij} = \frac{P(A_iB_j) - P(A_i)P(A_j)}{P(A_iB_j)\left[1 + 2P(A_iB_j) - 2P(A_i) - 2P(A_j)\right] + P(A_i)P(A_j)} = 1.$$

Отсюда

$$P(A_iB_j) - P(A_i)P(B_j) =$$
$$= P(A_iB_j) + P^2(A_iB_j) - 2P(A_iB_j)P(A_i) - 2P(A_iB_j)P(B_j) + P(A_i)P(B_j).$$

После преобразований получаем

$P(A_iB_j)\left[P(A_iB_j) - P(A_i)\right] - P(B_j)\left[P(A_iB_j) - P(A_i)\right] = 0$. Далее имеем

$\left[P(A_iB_j) - P(A_i)\right]\left[P(A_iB_j) - P(B_j)\right] = 0$. Возможны случаи

1.1. $P(A_iB_j) - P(A_i) = 0; \quad P(A_iB_j) = P(A_i)$. Отсюда следует, что $A_i \subset B_j$.

Для рассматриваемых значений i, j это означает, что событие $X = x_i$ влечет событие
$Y = y_j$, то есть величина y_j является функцией значения x_i.

Зафиксируем $i = i_0;\quad j = j_0$. Для различных значений j мы имеем два равенства $p_{i_0 j_0} = p_{i_0 \cdot};\ \sum_j p_{i_0 j} = p_{i_0 \cdot}$. Последнее равенство есть формула согласованности. Отсюда $\sum_j p_{i_0 j} = p_{i_0 j_0}$. Следовательно, $\sum_{j;j \neq j_0} p_{i_0 j} = 0$. Из этого равенства заключаем, что $p_{i_0 j} = 0$ при всех значениях $j;\quad j \neq j_0$.

В таблице рассматриваемого дискретного распределения в строке с номером i_0, задающей вероятности значения $X = x_{i_0}$, принимаемого совместно со значениями $Y = y_j$ при различных значениях j, все вероятности равны нулю кроме $p_{i_0 j_0}$.

Вероятность $p_{i_0 j_0}$ не может равняться нулю. Иначе вся строка с номером i_0 будет состоять из нулей. Это приведет к тому, что вероятность $P\left(X = x_{i_0}\right) = P\left(A_{i_0}\right) = p_{i_0 \cdot} = \sum_j p_{i_0 j} = 0$, чего быть не может, так как предполагается, что событие $X = x_{i_0}$ реализуемо с вероятностью, отличной от нуля.

Формула (4) при этом дает $\tau_{i_0 j} = \dfrac{0-0}{0+0} = \dfrac{0}{0}$. Это противоречит тому, что по предположению
$\tau_{i_0 j} = 1$. Итак, строки из нулей не может быть. Пункт 1.1. рассмотрен.

1.2. Пусть теперь $P\left(A_i B_j\right) - P\left(B_j\right) = 0$. Пункт рассматривается аналогично. Получаем, что $B_j \subset A_i$. В таблице рассматриваемого дискретного распределения столбец с номером j будет содержать одну клетку с вероятностью $p_{i_0 j_0}$, отличной от нуля; в остальных клетках стоят нули. Значение x_{i_0} определяется значением y_{j_0}, то есть является функцией этого значения. Столбец из одних нулей невозможен.

Из рассмотрения пунктов 1.1 и 1.2 следует, что таблица распределения должна быть квадратной; иначе будут строки или столбцы, состоящие из нулей. За счет перенумерации значений случайных величин X, Y можно считать, что в таблице распределения числа, отличные от нуля, стоят на главной диагонали, а в остальных клетках стоят нули.

Приведем пример распределения, для которого $co = 1$.

Таблица 1 дискретного распределения с коэффициентом детерминации $co = 1$.

$Y \rightarrow$ $X \downarrow$	y_1	y_2	y_3	$p_{i\cdot}$
x_1	$\frac{1}{3}$	0	0	$\frac{1}{3}$
x_2	0	$\frac{1}{3}$	0	$\frac{1}{3}$
x_3	0	0	$\frac{1}{3}$	$\frac{1}{3}$
$p_{\cdot j}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

Для этой таблицы

$$\tau_{11} = \frac{\frac{1}{3} - \frac{1}{3}\frac{1}{3}}{\frac{1}{3}\left(1 + 2\frac{1}{3} - 2\frac{1}{3} - 2\frac{1}{3}\right) + \frac{1}{3}\frac{1}{3}} = \frac{1 - \frac{1}{3}}{1 - \frac{1}{3}} = 1. \qquad \tau_{12} = \frac{0 - \frac{1}{3}\frac{1}{3}}{0 + \frac{1}{3}\frac{1}{3}} = -1.$$

.

Аналогично

$$\tau_{22} = \tau_{33} = 1;\ \tau_{13} = \tau_{21} = \tau_{23} = \tau_{31} = \tau_{32} = -1.$$

Тогда

$$co = \sum_{i=1}^{3}\sum_{j=1}^{3} \tau_{ij} p_{ij} = 3 \cdot 1 \cdot \frac{1}{3} + 0 = 1.$$

Таблица 2 дискретного распределения с коэффициентом детерминации $co = 1$.

$Y \rightarrow$ $X \downarrow$	y_1	y_2	y_3	$p_{\cdot j}$
x_1	0	$\frac{1}{3}$	0	$\frac{1}{3}$
x_2	0	0	$\frac{1}{3}$	$\frac{1}{3}$
x_3	$\frac{1}{3}$	0	0	$\frac{1}{3}$
$p_{i\cdot}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1

Таблица 2 путем перенумерации значений случайных величин приводится к табл. 1.

Случай 1 исчерпан.

2. Пусть теперь $\tau_{ij} = -1$. Тогда имеем равенство

$$\tau_{ij} = \frac{P\left(A_iB_j\right) - P\left(A_i\right)P\left(A_j\right)}{P\left(A_iB_j\right)\left[1 + 2P\left(A_iB_j\right) - 2P\left(A_i\right) - 2P\left(A_j\right)\right] + P\left(A_i\right)P\left(A_j\right)} = -1.$$

Отсюда

$$P\left(A_iB_j\right) - P\left(A_i\right)P\left(B_j\right) =$$

$$= -P\left(A_iB_j\right) - P^2\left(A_iB_j\right) + 2P\left(A_iB_j\right)P\left(A_i\right) + 2P\left(A_iB_j\right)P\left(B_j\right) - P\left(A_i\right)P\left(B_j\right).$$

Далее получаем

$P\left(A_iB_j\right)\left[1 + P\left(A_iB_j\right) - P\left(A_i\right) - P\left(B_j\right)\right] = 0.$ Рассмотрим возможные случаи.

2.1. $P\left(A_iB_j\right) = p_{ij} = 0.$ Это равенство означает, что события $A_i = \left(X = x_i\right)$ и $B_j = \left(Y = y_j\right)$ несовместны. Несовместность - это тоже вариант функционально-логической зависимости между значениями $x_i; y_j$ случайных величин X, Y. Примеры таблиц 1,2 показывают, что рассматриваемый случай реализуется.

2.2. Пусть теперь выполняется равенство $1 + P\left(A_iB_j\right) - P\left(A_i\right) - P\left(B_j\right) = 0.$

Отсюда

$$P\left(A_i\right) + P\left(B_j\right) - P\left(A_iB_j\right) = 1.$$

(7)

Так как $P\left(A_i + B_j\right) = P\left(A_i\right) + P\left(B_j\right) - P\left(A_i B_j\right)$, то получаем

$$P\left(A_i + B_j\right) = 1.$$

Это означает, что $A_i + B_j = I$ (достоверное событие). Это есть логическая связь между событиями A_i и B_j, то есть связь между возможностями принятия значений x_i, y_j случайными величинами X, Y. Происходит достоверно или событие $A_i \overline{B}_j$ или событие $\overline{A}_i B_j$ или событие $A_i B_j$. Таблица распределения двумерной случайной величины $\left(X, Y\right)$ имеет следующую структуру: строка с номером i и столбец с номером j содержат вероятности, отличные от нуля, удовлетворяющие равенству (7), а в остальных клетках таблицы стоят нули. Пример такого распределения вероятностей представлен в таблице 3.

В этой таблице вероятности событий $P\left(A_2\right) = P\left(X = x_2\right)$ и $P\left(B_2\right) = P\left(Y = y_2\right)$ удовлетворяют условию (7):

$$P\left(A_2\right) + P\left(B_2\right) - P\left(A_2 B_2\right) = \frac{5}{8} + \frac{5}{8} - \frac{2}{8} = 1.$$

Таблица 3 распределения с коэффициентом детерминации $co = 1$.

$Y \rightarrow$ $X \downarrow$	y_1	y_2	y_3	$p_{i\cdot}$
x_1	0	$\frac{2}{8}$	0	$\frac{2}{8}$
x_2	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{5}{8}$
x_3	0	$\frac{1}{8}$	0	$\frac{1}{8}$
$p_{\cdot j}$	$\frac{2}{8}$	$\frac{5}{8}$	$\frac{1}{8}$	1

Вычислим все элементарные коэффициенты связи.

$$\tau_{11}=\frac{0-\frac{2}{8}\frac{2}{8}}{0+\frac{2}{8}\frac{2}{8}}=-1;\ \tau_{12}=\frac{\frac{2}{8}-\frac{2}{8}\frac{5}{8}}{\frac{2}{8}\left(1+2\frac{2}{8}-2\frac{2}{8}-2\frac{5}{8}\right)+\frac{2}{8}\frac{5}{8}}=1;\ \tau_{13}=\tau_{11}=-1;$$

$$\tau_{21}=1;\ \tau_{22}=\frac{\frac{2}{8}-\frac{5}{8}\frac{5}{8}}{\frac{2}{8}\left(1+2\frac{2}{8}-2\frac{5}{8}-2\frac{5}{8}\right)+\frac{5}{8}\frac{5}{8}}=-1;$$

$$\tau_{23}=\frac{\frac{1}{8}-\frac{5}{8}\frac{1}{8}}{\frac{1}{8}\left(1+2\frac{1}{8}-2\frac{5}{8}-2\frac{1}{8}\right)+\frac{5}{8}\frac{1}{8}}=1;$$

$$\tau_{31}=-1;\ \tau_{32}=1;\ \tau_{33}=-1.$$

Из найденных элементарных коэффициентов связи составим коэффициент детерминации co.

$$co=\sum_i\sum_j|\tau_{ij}|p_{ij}=$$

$$=|-1|\cdot 0+1\cdot\frac{2}{8}+|-1|\cdot 0+1\cdot\frac{2}{8}+|-1|\frac{2}{8}+1\cdot\frac{1}{8}+|-1|\cdot 0+1\cdot\frac{1}{8}+|-1|\cdot 0=1.$$

С помощью этих же элементарных коэффициентов связи вычислим также коэффициент корреляции co_c.

$$co_c=(-1)\cdot 0+1\cdot\frac{2}{8}+(-1)\cdot 0+1\cdot\frac{2}{8}+(-1)\frac{2}{8}+1\cdot\frac{1}{8}+(-1)\cdot 0+1\cdot\frac{1}{8}+(-1)\cdot 0=\frac{1}{2}.$$

Коэффициент детерминации co показывает, что связь между исследуемыми случайными величинами – функционально-логическая, максимальная. Коэффициент корреляции указывает на положительность связи по предложенной классификации, так как $co_c>0$.

Итак, если $co_\lambda=1$, то случайные величины X,Y находятся в функционально-логической зависимости с вероятностью 1. Перечислим эти виды связей для отдельных значений случайных величин.

1. $A_i\subset B_j$, то есть событие $A_i=(X=x_i)$ влечет событие $B_j=(Y=y_j)$ или $B_j\subset A_i$.
2. $A_iB_j=\varnothing$ - события A_i и B_j несовместны.

3. $A_i + B_j = I$ – сумма событий A_i и B_j есть достоверное событие.

§ 5.2. Контингенциальный коэффициент детерминации для любых, в частности непрерывных, случайных величин

В соответствии с общей схемой, изложенной в § 3.1. образуем коэффициент детерминации для любых случайных величин на базе ядра контингенции. Такой коэффициент назовем контингенциальным коэффициентом детерминации, коротко «конти ».

Этот коэффициент имеет следующую конструкцию

$$co_\lambda = \left[M\left(\frac{\left| F_{XY}(X,Y) - F_X(X)F_Y(Y) \right|}{F_{XY}(X,Y)\left(1 + 2F_{XY}(X,Y) - 2F_X(X) - 2F_Y(Y)\right) + F_X(X)F_Y(Y)} \right)^\lambda \right]^{1/\lambda} \tag{1}$$

Здесь символ M означает оператор математического ожидания; $F_{XY}(x,y)$ – функция распределения случайной двумерной величины (X,Y); $F_X(X) = F_{XY}(x,+\infty)$;

$F_Y(y) = F_{XY}(+\infty, y)$ – функции распределения компонент X, Y двумерной случайной величины (X,Y); λ – параметр степенного усреднения; $\lambda > 0$. Для практики интересны случаи $\lambda = 1$ и $\lambda = 2$. В случае $\lambda = 1$ коэффициент co_λ будем обозначать просто co.

Формулу (1) можно записать короче, Применив в записи контингенциальное ядро $K(x,y)$.

Ядро $K(x,y)$ имеет вид

$$K(x,y) = \frac{F_{XY}(x,y) - F_X(x)F_Y(y)}{F_{XY}(x,y)\left(1 + 2F_{XY}(x,y) - 2F_X(x) - 2F_Y(y)\right) + F_X(x)F_Y(y)}. \tag{2}$$

Тогда формула (1) запишется в виде

$$co_\lambda = \left[\iint_D \left| K(x,y) \right|^\lambda dP \right]^{1/\lambda}. \tag{3}$$

Здесь D – область значений случайной двумерной величины (X,Y); интеграл понимается в смысле Лебега-Стилтьеса по вероятностной мере P.

Для непрерывных случайных величин с двумерной плотностью $f_{XY}(x,y)$ формулу (1) можно записать в виде

$$co_\lambda = \left(\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left| K(x,y) \right|^\lambda f_{XY}(x,y)\, dxdy \right)^{1/\lambda},$$

(4)

Знаменатель в формуле (2) обращается в нуль только тогда, когда и числитель равен нулю.

В противном случае ядро бы обращалось в бесконечность, чего быть не может, так как ядро является коэффициентом контингенции, для которого в § 1.3 установлены границы промежутка [0;1]. В случае равенства нулю функций распределения, стоящих в числителе и знаменателе формулы (2), требуется дополнительное соглашение о значении дроби для устранения разрывов.

Для дискретных случайных величин с законом распределения $P(X = x_i; Y = y_j) = p_{ij}; \quad i, j = 1,2,...$

формула (1) может быть записана в виде

$$co_\lambda = \left(\sum_i \sum_j \left| K(x_i, y_j) \right|^\lambda p_{ij} \right)^{1/\lambda},$$

(5)

где ядро $K(x,y)$ определяется формулой (2).

Свойства коэффициента co_λ аналогичны свойствам коэффициентов co_λ и as_λ, рассмотренных ранее. Сформулируем и докажем эти свойства.

Свойство 1. (Необходимое и достаточное условие независимости случайных величин).

Для того, чтобы случайные величины X, Y были независимы, необходимо и достаточно равенство нулю коэффициента детерминации co_λ.

Для доказательства используем необходимое и достаточное условие независимости случайных величин X, Y, выраженное через их функции распределения:

$$F_{XY}(x,y) - F_X(x) F_Y(y) = 0; \quad \forall x, y.$$

(6)

Пусть случайные величины X,Y независимы. Тогда выполняется условие (6), поэтому величина, стоящая под знаком математического ожидания в формуле (1) тождественно равна нулю. Математическое ожидание нуля равно нулю.
Обратно. Пусть $co_\lambda = 0$. Тогда

$$M\left(\frac{\left|F_{XY}(x,y) - F_X(x)F_Y(y)\right|}{F_{XY}(x,y)\left(1 + 2F_{XY}(x,y) - 2F_X(x) - 2F_Y(y)\right) + F_X(x)F_Y(y)} \right) = 0.$$

Математическое ожидание является интегралом, в общем случае Лебега-Стилтьеса. Из теории интеграла следует, что, если интеграл от неотрицательной функции, каковой является функция, стоящая под знаком математического ожидания, равен нулю, то и сама функция равна нулю с вероятностью единица. Тогда числитель функции равен нулю. Выполняется условие (6). Случайные величины X,Y – независимы с вероятностью единица.

Свойство 2. (Свойство нормировки): $0 \le co_\lambda \le 1$.

Действительно, ядро $K(x,y)$ в формуле (3) является коэффициентом контингенции событий $A = (X < x); B = (Y < y)$. В § 1.3 доказано, что модуль коэффициента контингенции заключен в промежутке [0;1]. Тогда $0 \le \left|K(x,y)\right|^\lambda \le 1$. Математическое ожидание $M\left(\left|K(X,Y)\right|^\lambda\right)$ как интеграл по вероятностной мере от функции $\left|K(x,y)\right|^\lambda$ удовлетворяет этим же неравенствам: $0 \le M\left(\left|K(X,Y)\right|^\lambda\right) \le 1$. Отсюда $0 \le co_\lambda \le 1$, так как

$$co_\lambda = \left[M\left(\left|K(X,Y)\right|^\lambda\right)\right]^{1/\lambda} \quad . \tag{7}$$

Свойство 3. (достижение верхней границы).
Если $Y = X$, то $co_\lambda = 1$.

Действительно, в этом случае $F_{XY}(x,y) = P(X < x, X < y) = P(X < x) = F_X(x)$ при $x \le y$. Аналогично $F_{XY}(x,y) = F_Y(y) = F_X(y)$ при $y < x$. Тогда

$$K(x,y) = \frac{F_X(x) - F_X(x)F_X(y)}{F_X(x)\left(1 + 2F_X(x) - 2F_X(x) - 2F_X(y)\right) + F_X(x)F_X(y)} = 1 \quad \text{при}$$

$x \le y$.

Аналогично $K(x,y)=1$ при $y<x$. Тогда $M\left(\left|K(X,Y)\right|^{\lambda}\right)=1$ и $co_{\lambda}=1$.

Свойство 4. Если случайная величина Y является строго возрастающей функцией случайной величины X: $Y=\varphi(X)$, то $co_{\lambda}=1$.

Действительно, в этом случае $y=\varphi(x)$ и события $X<x$ и $(Y<y)=(\varphi(X)<\varphi(x))$ являются эквивалентными в силу эквивалентности неравенств $X<x$ и $\varphi(X)<\varphi(x)$.

Тогда эти события имеют равные вероятности: $P(X<x)=F_X(x)=P(Y<y)=F_Y(y)$.

Далее, $F_{XY}(x,y)=P(X<x,Y<y)=P(X<x,\varphi(X)<\varphi(x))=$

$=P(X<x,X<x)=P(X<x)=F_X(x)$. Отсюда, как и в свойстве 3, мы получаем

$$K(x,y)=\frac{F_X(x)-F_X(x)F_X(x)}{F_X(x)(1+2F_X(x)-2F_X(x)-2F_X(x))+F_X(x)F_X(x)}=$$

$$=\frac{F_X(x)-F^2{}_X(x)}{F_X(x)-F^2{}_X(x)}=1.$$ Поэтому $M\left[\left|K(X,Y)\right|^{\lambda}\right]=1$ и $co_{\lambda}=1$.

Свойство 5. Если $co_{\lambda}=1$ то распределение двумерной случайной величины (X,Y) является вырожденным. В этом случае либо $F_{XY}(x,y)=F_X(x)$, либо $F_{XY}(x,y)=F_Y(y)$.

Доказательство проведем в два этапа.

5.1. Сначала докажем, что в этом случае $\left|K(x,y)\right|=1$ тождественно, то есть для любых x,y из области значений случайных величин X,Y.

Это доказательство во многом повторяет то доказательство, которое было проведено в § 4.1. для свойства 6 в случае ассоциативного ядра.

Воспользуемся формулой (7), в которой соответствующее математическое ожидание равно 1:

$$M\left[\left|K(X,Y)\right|\right]^{\lambda}=1.$$

(8)

В общем случае математическое ожидание в формуле (8) выражается двукратным интегралом Лебега-Стилтьеса по вероятностной мере:

$$\int_{R_2} \left|K\left(x,y\right)\right|^{\lambda} dP = 1.$$

(9)

Здесь R_2 – плоскость изменения переменных x, y .Интеграл в формуле (9) берется по той области $D \in R_2$, в которой определена вероятностная мера P.

Подынтегральная функция удовлетворяет неравенствам $0 \le \left|K\left(x,y\right)\right|^{\lambda} \le 1$, так как $K\left(x,y\right)$ есть коэффициент контингенции событий $A = \left(X < x\right)$ и $B = \left(Y < y\right)$, который заключен в промежутке [–1;1].

Заметим также, что $\int_D dP = 1$. Разобьем область D на две части: $D = D_1 \cup D_2$. В D_1 имеет неравенство $\left|K\left(x,y\right)\right| < 1$, а в D_2 - равенство $\left|K\left(x,y\right)\right| = 1$. Тогда интеграл в формуле (9) разобьется на два интеграла

$$\int_{D_1} \left|K\left(x,y\right)\right|^{\lambda} dP + \int_{D_2} \left|K\left(x,y\right)\right|^{\lambda} dP = 1.$$

(10)

Так как во втором интеграле $\left|K\left(x,y\right)\right| = 1$, то формула (10) принимает вид

$$\int_{D_1} \left[K\left(x,y\right)\right]^{\lambda} dP + \mu\left(D_2\right) = 1.$$

(11)

здесь $\mu\left(D_2\right)$ – вероятностная мера множества D_2.

Интегрируем неравенство $\left|K\left(x,y\right)\right|^{\lambda} < 1$ по множеству D_1. Получаем

$\int_{D_1} \left|K\left(x,y\right)\right|^{\lambda} dP < \int_{D_1} dP = \mu\left(D_1\right)$. Таким образом,

$\int_{D_1} \left|K\left(x,y\right)\right|^{\lambda} dP + \mu\left(D_2\right) < \mu\left[D_1\right] + \mu\left(D_2\right) = 1$. Этот результат означает, что при $\mu\left(D_1\right) > 0$

равенство (11) нарушается, следовательно $\mu(D_1)=0$. Итак, $|K(x,y)|^{\lambda}<1$ с вероятностью ноль, а потому $|K(x,y)|<1$ только с вероятностью 0. Соответственно

$$|K(x,y)|=1 \quad \text{с вероятностью 1.} \tag{12}$$

Проанализируем равенство (12). Для краткости обозначений положим $A=(X<x);B=(Y<y)$. Тогда получим

$$\frac{P(AB)-P(A)P(B)}{P(AB)(1+2P(AB)-2P(A)-2P(B))+P(A)P(B)}=\pm 1. \tag{13}$$

5.2. Рассмотрим сначала случай, когда в формуле (13) стоит знак плюс. Тогда получаем

$$P(AB)-P(A)P(B)=$$
$$=P(AB)+2P^2(AB)-2P(AB)P(A)-2P(AB)P(B)+P(A)P(B).$$

Отсюда $P^2(AB)-P(AB)P(A)-P(AB)P(B)+P(A)P(B)=0.$

Далее получаем

$$P(AB)(P(AB)-P(A))-P(B)(P(AB)-P(A))=0;$$

$$[P(AB)-P(A)][P(AB)-P(B)]=0. \tag{14}$$

Приравняем нулю первый множитель, Тогда

$$P(AB)=P(A). \tag{15}$$

Это означает, что

$$A\subset B, \tag{16}$$

то есть событие $X<x$ влечет событие $Y<y$ для любых значений x,y.

Отсюда следует, что

$$F_X(x)=P(X<x)\le P(Y<y)=F_Y(y). \tag{17}$$

Переменная y не может быть независимой переменной. В противном случае фиксируем значение y такое, что $F_Y(y)<1$. Получаем $F_X(x)<1$. Это противоречит равенству $F_X(+\infty)=1$. Противоречие доказывает, что y зависит от x.

Запишем равенство (15) подробнее:

$$F_{XY}(x,y)=F_X(x)$$

(18)

Это равенство означает, что при любых значениях аргументов x,y функция распределения двумерной случайной величины (X,Y) равна функции распределения своей компоненты X, то есть двумерное распределение является вырожденным – не зависит от переменной y. Примеры таких распределений ранее указаны. Это случаи, когда $Y=X$ или Y является строго возрастающей функцией от X. Есть ли другие случаи – вопрос пока остается открытым.

Вариант, когда $P(AB)=P(B)$, то есть $F_{XY}(x,y)=F_Y(y)$, рассматривается аналогично.

5.3. Рассмотрим теперь случай, когда в формуле (13) стоит знак минус. Тогда

$$\frac{P(AB)-P(A)P(B)}{P(AB)\big(1+2P(AB)-2P(A)-2P(B)\big)+P(A)P(B)}=-1.$$ Отсюда

$$P(AB)-P(A)P(B)=$$
$$=-P(AB)\big(1+2P(AB)-2P(A)-2P(B)\big)-P(A)P(B);$$

$$P(AB)\big[1+P(AB)-P(A)-P(B)\big]=0.$$

(19)

Приравняем поочередно оба множителя в формуле (19) нулю.

Пусть $P(AB)=0$. Тогда $F_{XY}(x,y)=0$ тождественно, то есть для любых x,y. Но $F_{XY}(+\infty,+\infty)=1$. Это равенство противоречит тому, что функция распределения тождественно равна нулю. Случай отпадает.

Пусть теперь $1+P(AB)-P(A)-P(B)=0.$ Тогда $P(AB)=P(A)+P(B)-1.$

Запишем это равенство через функции распределения. $F_{XY}(x,y)=F_X(x)+F_Y(y)-1.$

По свойствам функций распределения $F_{XY}(-\infty,-\infty)=0; F_X(-\infty)=0; F_Y(-\infty)=0.$

При этих значениях аргументов предыдущее равенство становится неверным, так как слева стоит нуль, а справа минус единица. Случай отпадает.

Свойство 5 доказано.

§ 5.3.(3.7) Пример вычисления контингенциальных коэффициентов детерминации и корреляции для непрерывных случайных величин

Расчетные формулы для контингенциального коэффициента детерминации в случае непрерывного распределения приведены в § 5.2 – формулы (2), (4):

$$co_\lambda=\left(\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left|K(x,y)\right|^\lambda f_{XY}(x,y)\,dxdy\right)^{1/\lambda}$$

(1)

$$K(x,y)=\frac{F_{XY}(x,y)-F_X(x)F_Y(y)}{F_{XY}(x,y)\left(1+2F_{XY}(x,y)-2F_X(x)-2F_Y(y)\right)+F_X(x)F_Y(y)}.$$

(2)

Применим эти формулы при $\lambda=1$ для вычисления коэффициента co в конкретном примере.

Пример 1. Рассмотрим равномерное распределение в треугольнике с вершинами O(0;0), A(1;0), B(0;1).

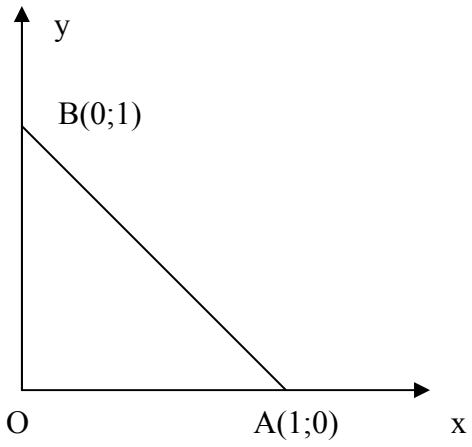


Рис. 1. Треугольник OAB.

1. Уравнение линии AB: $y = 1 - x$. Обозначим треугольник OAB символом Δ. Плотность вероятности двумерной случайной величины (X,Y) задается формулами: $f_{XY}(x,y) = 2$ при $(x;y) \in \Delta$; $f_{XY}(x,y) = 0$ при $(x;y) \in \Delta$. Плотности компонент :

$f_X(x) = 2(1-x); f_Y(y) = 2(1-y)$ при $(x;y) \in \Delta$.

Функции распределения: $F_{XY}(x,y) = 2xy$

$F_X(x) = 2x - x^2; F_Y(y) = 2y - y^2$

при $(x,y) \in \Delta$.

Для этого распределения контингенциальное ядро $K(x,y)$ равно

$$K(x,y) = \frac{F_{XY}(x,y) - F_X(x)F_Y(y)}{F_{XY}(x,y)\left[1 + 2F_{XY}(x,y) - 2F_X(x) - 2F_Y(y)\right] + F_X(x)F_Y(y)} =$$

$$= \frac{2xy - \left(2x - x^2\right)\left(2y - y^2\right)}{2xy\left[1 + 4xy - 2\left(2x - x^2\right) - 2\left(2y - y^2\right)\right] + \left(2x - x^2\right)\left(2y - y^2\right)} =$$

$$= \frac{-2 + 2x + 2y - xy}{4x^2 + 4y^2 + 9xy - 10x - 10y + 6}.$$

2. Исследуем знак числителя и знаменателя ядра .

Сначала исследуем знак числителя $u = -2 + 2x + 2y - xy.$ Ищем наибольшее и наименьшее значения этой функции в треугольнике Δ.

Стационарная точка: $u'_x = 2 - y = 0; u'_y = 2 - x = 0.$ Стационарная точка $(2,2)$ лежит вне $\Delta.$ На границе: $x = 0 \Rightarrow u = -2 + 2y = 2(y-1) \le 0.$

Аналогично: $y = 0 \Rightarrow u = 2(x-1) \le 0.$

На границе AB $y = 1 - x$:

$u(x,1-x) = -2 + 2x + (2-x)(1-x) = x(x-1) \le 0$.

Итак, $u \le 0$ в Δ и обращается в нуль только в точках $A(1;0), B(0;1).$

Теперь исследуем знак знаменателя $v = 4x^2 + 9xy + 4y^2 - 10x - 10y + 6.$

По свойствам контингенциального ядра его знаменатель неотрицателен, так как является суммой двух вероятностей. Однако нам нужно установить не только его знак, но и его нули.

Определяем стационарную точку: $v'_x = 8x + 9y - 10 = 0; v'_y = 9x + 8y - 10 = 0.$

$x = y = 10/17$; Точка (10/17;10/17) лежит вне Δ. Исследуем границу. При $y = 0$

получаем $v = 4x^2 - 10x + 6 = 4(1-x)\left(\frac{3}{2} - x\right) \geq 0.$ В пределах треугольника Δ обращается в нуль только в точке (1;0). Аналогично при $x = 0$ получаем $v = 4(1-y)\left(\frac{3}{2} - y\right) \geq 0.$

В пределах треугольника Δ равенство нулю только в точке (0;1). При $y = 1 - x$ получаем в Δ:

$$v = 4x^2 + 9x(1-x) + 4(1-x)^2 - 10x - 10(1-x) + 6 = x(1-x) \geq 0.$$

Возвращаемся к ядру $K(x,y)$. Из предыдущего исследования следует, что

$$|K(x,y)| = \frac{xy - 2x - 2y + 2}{4x^2 + 9xy + 4y^2 - 10x - 10y + 6}.$$

(3)

Найдем значения этой функции в угловых точках треугольника, раскрыв неопределенность.

На прямой $y = 1 - x$ имеем: числитель $u(x, 1-x) = x(1-x)$; знаменатель $v(x, 1-x) = x(1-x)$. Тогда $|K(x,y)| = \frac{x(1-x)}{x(1-x)} = 1.$ Следовательно и в угловых точках $A(1;0), B(0;1)$ будет $|K(x,y)|$ равен 1. В точке $O(0;0)$ $|K(0;0)| = 1/3.$

Итак, во всем треугольнике Δ справедливы неравенства

$$-1 \leq K(x,y) \leq -\frac{1}{3};$$

(4)

$$\frac{1}{3} \leq |K(x,y)| \leq 1.$$

(5)

Используя формулу (3), составим коэффициент детерминации co $(\lambda = 1)$:

$$co = \int_0^1 dx \int_0^{1-x} |K(x,y)| f_{XY}(x,y)\, dy =$$

$$\int_0^1 dx \int_0^{1-x} \frac{xy - 2x - 2y + 2}{4x^2 + 9xy + 4y^2 - 10x - 10y + 6} 2dy. \quad (6)$$

Вычисление интеграла (6) связано с громоздкими вычислениями. Внешний интеграл по аргументу x не выражается в элементарных функциях

(неберущийся), поэтому целесообразно для отыскания двойного интеграла применить численные методы. Один из численных методов основан на аппроксимации подынтегральной функции более простой функцией, например, многочленом. Модуль ядра меняется очень плавно в узком диапазоне (5), поэтому можно ограничиться интерполяционным многочленом невысокой степени – линейным или квадратическим. Результат усреднения модуля ядра будет лежать в тех же пределах (5): $\frac{1}{3} \le co \le 1$.

3. Интерполяционный многочлен $z = Ax + By + C$, построенный по трем угловым точкам $O(0;0), A(1;0), B(0;1)$ имеет вид $z = -\frac{1}{3}(2x + 2y + 1)$. Таким образом,

$$\left|K(x,y)\right| \approx \frac{1}{3}(2x + 2y + 1).$$

(7)

$$co \approx \frac{2}{3}\int_0^1 dx \int_0^{1-x} (2x + 2y + 1)dy = \frac{7}{9} \approx 0,78.$$

(8)

4. Построим теперь квадратический интерполяционный многочлен

$z = Ax^2 + Bxy + Cy^2 + Dx + Ey + F$ по шести точкам границы треугольника Δ

$O(0;0), z = -\frac{1}{3}; F\left(\frac{1}{2};0\right), z = -\frac{1}{2}; \quad B(1;0), z = -1; D\left(\frac{1}{2};\frac{1}{2}\right), z = -1;$

$C(0;1), z = -1; E\left(0;\frac{1}{2}\right), z = -\frac{1}{2}.$

Интерполяционный многочлен имеет вид $z = -\frac{1}{3}(2x^2 + 4xy + 2y^2 + 1)$, поэтому

$$\left|K(x,y)\right| \approx \frac{1}{3}(2x^2 + 4xy + 2y^2 + 1).$$

(9)

Тогда

$$co \approx \frac{2}{3}\int_0^1 dx \int_0^{1-x} (2x^2 + 4xy + 2y^2 + 1)dy = \frac{2}{3} \approx 0,67.$$

(10)

5. Выполняем кусочно-линейную интерполяцию ядра $K(x,y)$. Для этого разобьем треугольник Δ на 16 малых треугольников Δ_k прямыми, параллельными сторонам треугольника Δ : $x=\frac{1}{4};\quad x=\frac{1}{2};\quad x=\frac{3}{4};$

$y=\frac{1}{4};\quad y=\frac{1}{2};\quad y=\frac{3}{4};$

$x+y=\frac{1}{4};\quad x+y=\frac{1}{2};\quad x+y=\frac{3}{4}.$

В каждом треугольнике Δ_k аппроксимируем ядро $K(x,y)$ линейным интерполяционным многочленом $z=A_kx+B_ky+C_k=P_k(x,y)$, вычисляем интеграл

$\iint\limits_{\Delta_k}\left|P_k(x,y)\right|2dxdy=\iint\limits_{\Delta_k}\left|A_kx+B_ky+C_k\right|dxdy=I_k$, а затем все интегралы складываем:

$$co\approx 2\sum_{k=1}^{16}I_k.$$

(11)

Для вычислений интегралов составим таблицу.

В ней x_k, y_k – вершины треугольников Δ_k. $z_k=K(x_k,y_k)$.

Таблица1. Вычисление интегралов I_k и коэффициента co по формуле (11).

k	x_k	y_k	z_k	A_k	B_k	C_k	I_k
1	0 0 1/4	1 3/4 3/4	−1 −2/3 −1	−4/3	−4/3	1/3	1/36 = 0,027778
2	1/4 1/4 1/2	3/4 1/2 1/2	−1 −5/7 −1	−8/7	−8/7	1/7	1/672 = 0,001488
3	1/2 1/2 3/4	1/2 1/4 1/4	−1 −5/7 −1	симмет	рично	к=2	1/672 = 0,001488
4	3/4	1/4 0	−1 −2/3	симмет	рично	к=1	1/36 = 0,027778

	3/4 1	0	−1				
5	0 1/4 1/4	3/4 1/2 3/4	−2/3 −5/7 −1	−4/3	−8/7	4/21	25/1008 = 0,024802
6	1/4 1/2 1/2	1/2 1/4 1/2	−5/7 −5/7 −1	−8/7	−8/7	1/7	17/672 = 0,025298
7	1/2 3/4 3/4	1/4 0 1/4	−5/7 −2/3 −1	симмет	рично	к=5	25/1008 = 0,024802
8	0 0 1/4	3/4 1/2 1/2	−2/3 −1/2 −5/7	−6/7	−2/3	−1/6	−79/4032 = 0,019593
9	1/4 1/4 1/2	1/2 1/4 1/4	−5/7 −17/33 −5/7	−184/231	−184/231	−9/77	0,021965
10	1/2 1/2 3/4	1/4 0 0	−5/7 −1/2 −2/3	симмет	рично	к=8	−79/4032 = 0,019593
11	0 1/4 1/4	1/2 1/2 1/2	−1/2 −17/33 −5/7	−0,857143	−0,796537	−0,101732	0,018015
	x_k	y_k	z_k	A_k	B_k	C_k	I_k
12	1/4 1/2 1/2	1/4 0 1/4	−17/33 −1/2 −5/7	симмет	рично	к=11	0,018015
13	0 0 1/4	1/4 1/2	−2/5 −1/2	−76/165	−2/5	−3/10	0,014741

		1/4	−17/33				
14	1/4 1/4 1/2	1/4 0 0	−17/33 −2/5 −1/2	симмет	рично	к=13	0,014741
15	0 1/4 1/4	1/4 0 1/4	−2/5 −2/5 −17/38	−76/165	−76/165	−47/165	0,013699
16	0 0 1/4	0 1/4 0	−1/3 −2/5 −2/5	−4/15	−4/15	−1/3	0,011806
Σ	-	-	-	-	-	-	0,285602

Используя суммарный результат вычислений из последней строки и последнего столбца таблицы 1, находим значение коэффициента детерминации

$$co = 2\iint\limits_{\Delta} \left|K\left(x,y\right)\right| dxdy \approx 2 \cdot 0,285602 = 0,571204 \approx 0,57.$$

(12)

Этот результат является более точным, чем результаты, полученные при линейной и квадратической интерполяции. Они были соответственно равны 0,78 и 0,67.

Для сравнения приведем здесь также для этого распределения значение линейного коэффициента корреляции: $\rho = -0,5.$

Если снять знак модуля подынтегральной функции в формуле (12), то получим значение коэффициента корреляции $co_c \approx -0,57$ на базе контингенциального ядра для случайных величин X, Y, связанных рассматриваемым двумерным распределением. Знак минус появился потому, что в треугольнике Δ ядро $K\left(x,y\right)$ отрицательно. Знак минус здесь так же, как и для линейного коэффициента корреляции, показывает убывание в среднем значений Y при возрастании X.

§ 5.4. Пример вычисления контингенциального коэффициента детерминации для дискретных случайных величин

Напомним, что этот коэффициент вычисляется по формуле

$$co_\lambda = \left[\sum_i \sum_j \left| K_{ij} \right|^\lambda p_{ij} \right]^{1/\lambda} .$$

(1)

Здесь $p_{ij} = P\left(X = x_i, Y = y_j \right);\ p_{i\cdot} = P\left(X = x_i \right);\ p_{\cdot j} = P\left(Y = y_j \right);$

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j} \right) + p_{i\cdot} p_{\cdot j}}.$$

(2)

Рассмотрим случай $\lambda = 1$. Тогда

$$co = \sum_i \sum_j \left| K_{ij} \right| p_{ij}.$$

(3)

Рассмотрим случай равномерного дискретного распределения для треугольной таблицы вида

Таблица 1. Дискретное двумерное равномерное распределение в треугольнике.

$Y \rightarrow$ $X \downarrow$	1	2	3	4	5	6	7	$p_{i\cdot}$
1	p_{11} 1/16	0	0	0	0	0	0	1/16
2	p_{21} 1/16	p_{22} 1/16	p_{23} 1/16	0	0	0	0	3/16
3	p_{31} 1/16	p_{32} 1/16	p_{33} 1/16	p_{34} 1/16	p_{35} 1/16	0	0	5/16
4	p_{41} 1/16	p_{42} 1/16	p_{43} 1/16	p_{44} 1/16	p_{45} 1/16	p_{46} 1/16	p_{47} 1/16	7/16
$p_{\cdot j}$	4/16	3/16	3/16	2/16	2/16	1/16	1/16	1

Таблица 1, хотя и не является в точности треугольной, названа так потому, что может служить примером дискретизации равномерного непрерывного распределения в треугольнике, которое рассмотрено в предыдущем § 5.3. Коэффициент детерминации в этом дискретном случае должен быть близким тому, который вычислен в § 5.3.

Вычисления по формуле (2) дают

$$K_{11}=\frac{\frac{1}{16}-\frac{1}{16}\frac{4}{16}}{\frac{1}{16}\left(1+2\frac{1}{16}-2\frac{1}{16}-2\frac{4}{16}\right)+\frac{1}{16}\frac{4}{16}}=1;$$

$$K_{21}=\frac{\frac{1}{16}-\frac{3}{16}\frac{4}{16}}{\frac{1}{16}\left(1+2\frac{1}{16}-2\frac{3}{16}-2\frac{4}{16}\right)+\frac{3}{16}\frac{4}{16}}=\frac{1}{4};$$

и так далее получаем

$$K_{22}=\frac{7}{15};K_{23}=\frac{7}{15};K_{31}=-\frac{1}{5};\ K_{32}=\frac{1}{17};K_{33}=\frac{1}{17};K_{34}=\frac{3}{7};K_{35}=\frac{3}{7};$$

$$K_{41}=-\frac{1}{2};K_{42}=-\frac{5}{19};K_{43}=-\frac{5}{19};K_{44}=\frac{1}{6};K_{45}=\frac{1}{6};K_{46}=1;K_{47}=1.$$

Суммируя все результаты по формуле (3), получаем

$$co=6,71777/16=0,41986\approx 0,42.$$

(4)

Напомним, что для аналогичного непрерывного равномерного распределения коэффициент детерминации равнялся $co=0,57$.

Глава 6. Предельный коэффициент детерминации

Контингенциальное ядро можно выразить через плотность распределения, если произвести некоторые преобразования, связанные с переходом к пределу. Этому и посвящаются следующие параграфы.

§ 6.1. Коэффициенты детерминации и корреляции с предельным ядром контингенции для непрерывных распределений.

1º. Построение коэффициента детерминации с предельным контингенциальным ядром.

Отправляемся от ядра контингенции, применяемого для образования коэффициента детерминации двух дискретных случайных величин X,Y

$$K_{ij}=\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}. \qquad (1)$$

Здесь $p_{ij}=P\left(X=x_i,Y=y_j\right); p_{i\cdot}=P\left(X=x_i\right); p_{\cdot j}=P\left(Y=y_j\right).$

Пусть двумерное непрерывное распределение определяется плотностью $f_{XY}\left(x,y\right)$ с хорошими аналитическими свойствами.

Для простоты рассмотрим прямоугольную область $D=\left[a\le x\le b;c\le y\le d\right]$ распределения вероятностей двумерной случайной величины $\left(X,Y\right)$. Разобьем ее на прямоугольники $D_{ij}=\left[a_i\le x\le b_i;c_j\le y\le d_j\right],\ i=1,...,m;j=1,...,n.$сеткой прямых, параллельных координатным осям с шагом Δx вдоль одной оси и шагом Δy вдоль другой оси. Плотности распределения компонент X,Y двумерной случайной величины $\left(X,Y\right)$ строятся по формулам $f_X\left(x\right)=\int\limits_c^d f_{XY}\left(x,y\right)dy;\quad f_Y\left(y\right)=\int\limits_a^b f_{XY}\left(x,y\right)dx$. Перейдем от непрерывного распределения к дискретному, сосредоточив вероятностные массы, распределенные в прямоугольниках D_{ij}, в отдельных точках. Положим

$p_{ij}=\iint\limits_{D_{ij}} f_{XY}\left(x,y\right)dxdy=f_{XY}\left(x_i,y_j\right)\Delta x\Delta y$ по теореме о среднем для интеграла.

$$p_{i\cdot} = \sum_{j=1}^{n} p_{ij} = \sum_{j=1}^{n} \int_{a_i}^{b_i} dx \int_{c_j}^{d_j} f_{XY}(x,y)dy = \int_{a_i}^{b_i} \left[\sum_{j=1}^{n} \int_{c_j}^{d_j} f_{XY}(x,y)\,dy \right] dx =$$

$$\int_{a_i}^{b_i} \left[\int_{c}^{d} f_{XY}(x,y)\,dy \right] dx =$$

$$= \int_{a_i}^{b_i} f_X(x)\,dx = f_X(x_i')\Delta x \text{ по теореме о среднем для интеграла; } x_i' \in [a_i; b_i].$$

Аналогично $p_{\cdot j} = \sum_{i=1}^{m} p_{ij} = f_Y(y')\Delta y; \quad y_j' \in [c_j; d_j]$.

Имея непрерывное распределение, построили двумерное дискретное распределение вероятностей:

$$p_{ij} = P(X' = x_i; Y' = y_j); \quad i = 1,...,m; j = 1,...,n.$$

и два одномерных дискретных распределения компонент X', Y' двумерной дискретной случайной величины (X', Y'):

$$p_{i\cdot} = \sum_{j=1}^{n} p_{ij} = (X' = x_i); \quad p_{\cdot j} = \sum_{i=1}^{m} p_{ij} = P(Y' = y_j).$$

Составим для построенного дискретного распределения контингенциальное ядро

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij}(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}) + p_{i\cdot}p_{\cdot j}} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij}u_{ij} + p_{i\cdot}p_{\cdot j}}.$$

Здесь для сокращения записи положено $u_{ij} = 1 + 2p_{ij} - 2p_{i\cdot}p_{\cdot j}$. Заменим в нем вероятности дискретного распределения выражениями через плотности. Получим

$$K_{ij} = \frac{f_{XY}(x_i, y_j)\Delta x\Delta y - f_X(x_i')\Delta x f_Y(y_j')\Delta y}{f_{XY}(x_i, y_j)\Delta x\Delta y u_{ij} + f_X(x_i')\Delta x f_Y(y_j')\Delta y}. \qquad (2)$$

Здесь

$$u_{ij} = 1 + 2f_{XY}(x_i, y_j)\Delta x\Delta y - 2f_X(x_i')\Delta x - 2f_Y(y_j'). \qquad (3)$$

Сократим дробь (2) на произведение $\Delta x\Delta y$ и перейдем к пределу при $\lambda = \max\{\Delta x, \Delta y\} \to 0$. Так как при этом $u_{ij} \to 1$ и точки (x_i, y_j) и (x_i', y_j') сходятся к общей точке (x, y), то в пределе получим

$$K(x,y) = \frac{f_{XY}(x,y) - f_X(x)f_Y(y)}{f_{XY}(x,y) + f_X(x)f_Y(y)}. \qquad (4)$$

Полученную функцию (4) назовем предельным контингенциальным ядром. Оно станет основой для построения предельного контингенциального коэффициента детерминации для непрерывных распределений. Этот коэффициент строится по формуле

$$l_\lambda = \left[\iint\limits_D \left| K(x,y) \right|^\lambda f_{XY}(x,y)\,dxdy \right]^{1/\lambda} =$$

$$= \left[\iint\limits_D \left(\frac{\left| f_{XY}(x,y) - f_X(x) f_Y(y) \right|}{f_{XY}(x,y) + f_X(x) f_Y(y)} \right)^\lambda f_{XY}(x,y)\,dxdy \right]^{1/\lambda}. \qquad (5)$$

Здесь D – область значений случайной величины (X,Y); $(\lambda > 0)$.

Коэффициент детерминации l_λ степени λ предназначен для измерения величины связи между непрерывными случайными величинами X,Y. В случае $\lambda = 1$ этот коэффициент будем обозначать просто l. Если снять знак модуля при $\lambda = 1$, то получим коэффициент корреляции

$$l_c = \iint\limits_D \frac{f_{XY}(x,y) - f_X(x) f_Y(y)}{f_{XY}(x,y) + f_X(x) f_Y(y)} f_{XY}(x,y)\,dxdy. \qquad (6)$$

Коэффициент корреляции также измеряет величину связи между случайными величинами X,Y, но в другом смысле, он дает среднюю разность между положительными и отрицательными значениями ядра. Коэффициент корреляции представляется менее информативным в части величины зависимости между случайными величинами, чем коэффициент детерминации.

2º.Свойства коэффициента детерминации с предельным контингенциальным ядром.

1. Коэффициенты детерминации и корреляции l_λ и l_c соответственно нормированы условиями

$$0 \le l_\lambda \le 1; \quad -1 \le l_c \le 1. \qquad (7)$$

Действительно, из вида ядра $K(x,y)$ в формуле (4) заключаем, что $\left| K(x,y) \right| \le 1$. Тогда справедливо и неравенство $\left| K(x,y) \right|^\lambda \le 1; \quad (\lambda > 0)$. Интегрирование этого неравенства с весом, равным плотности $f_{XY}(x,y)$, дает первое неравенство (7). Отсюда следует и второе неравенство (7).

2. Коэффициент детерминации l_λ равен нулю тогда и только тогда, когда непрерывные случайные величины независимы (с вероятностью единица).

Доказательство.

2.1. Пусть непрерывные случайные величины независимы. Тогда согласно необходимому и достаточному условию их независимости справедливо равенство

$$f_{XY}(x,y) - f_X(x) f_Y(y) = 0; \quad \forall x, y. \tag{7}$$

Тогда интеграл в формуле (5) равен нулю, поэтому $l_\lambda = 0$.

2.2. Пусть $l_\lambda = 0$. Это означает, что интеграл в формуле (5) от неотрицательной функции равен нулю. Согласно общим свойствам интеграла Римана подынтегральная функция равна нулю почти везде, то есть с точностью до множества нулевой меры Лебега. Тогда $f_{XY}(x,y) - f_X(x) f_Y(y) = 0$ с точностью до множества меры нуль. Это означает, что случайные величины X, Y независимы с вероятностью единица.

3. Если случайные величины X, Y равны, то есть $Y = X$ тождественно (полная зависимость), то точной верхней границей коэффициента детерминации l_λ является единица.

Доказательство.

Пусть $Y = X$. Это означает, что двумерное распределение – вырожденное, является одномерным. Тогда функция распределения двумерной случайной величины представляется в виде

$$F_{XY}(x,y) = P(X < x, X < y) = \begin{cases} P(X < x) \, при & x \le y \\ P(Y < y) \, при & y < x \end{cases}.$$

Тогда $f_{XY}(x,y) = \dfrac{\partial^2 F_{XY}(x,y)}{\partial x \partial y} = 0$ вне прямой $y = x$. На прямой $y = x$ имеем $F_{XY}(x,x) = F_X(x)$. Поэтому $F'_X(x) = f_X(x)$ является одномерной плотностью на прямой $y = x$. Итак, в этом случае $f_{XY}(x,y) = f_X(x)$ (или $f_Y(y)$).

Предельное контингенциальное ядро принимает вид

$$K(x,y) = \frac{f_{XY}(x,y) - f_X(x) f_Y(y)}{f_{XY}(x,y) + f_X(x) f_Y(y)} = \frac{f_X(x) - f^2{}_X(x)}{f_X(x) + f^2{}_X(x)} = \frac{1 - f_X(x)}{1 + f_X(x)}. \tag{7}$$

Из выражения

$$\left|K(x,y)\right| = \frac{\left|1 - f_X(x)\right|}{1 + f_X(x)} \tag{8}$$

видим, что

$$0 \le \left|K(x,y)\right| \le 1. \tag{9}$$

$\left|K\left(x,y\right)\right|=1$ имеет место при $f_X\left(x\right)\equiv 0$, что для случая непрерывного распределения невозможно. Однако это равенство возможно как предельное, то есть возможно, чтобы $l_\lambda \to 1$. В этом случае

$$l_\lambda = \left[\int\limits_{-\infty}^{+\infty}\left(\frac{\left|1-f_X\left(x\right)\right|}{1+f_X\left(x\right)}\right)^\lambda f_X\left(x\right)dx\right]^{1/\lambda}. \qquad (10)$$

Рассмотрим для простоты случай $\lambda = 1$. Тогда

$$l = l_1 = \int\limits_{-\infty}^{+\infty}\frac{\left|1-f_X\left(x\right)\right|}{1+f_X\left(x\right)}f_X\left(x\right)dx\ . \qquad (11)$$

Нужно подобрать плотность, которая просто и эффективно концентрирует распределение около одной точки с помощью одного параметра. Такой плотностью может быть, например, плотность показательного распределения: $f_X\left(x\right)=\mu e^{-\mu x}$ при $x\ge 0;$ $\left(\mu>0\right);$ $f_X\left(x\right)=0$ при $x<0$. Вычислим для нее коэффициент детерминации η_1 по формуле (11).

$l=\int\limits_0^{+\infty}\frac{\left|1-\mu e^{-\mu x}\right|}{1+\mu e^{-\mu x}}\mu e^{-\mu x}dx=-\int\limits_0^{+\infty}\frac{\left|1-\mu e^{-\mu x}\right|}{1+\mu e^{-\mu x}}de^{-\mu x}$. Подстановка $z=e^{-\mu x}$. Тогда

$l=\int\limits_o^1\frac{\left|1-\mu z\right|}{1+\mu z}dz.$ Для снятия знака модуля заметим, что $1-\mu z=0$ при $z=\frac{1}{\mu}.$

Тогда $1-\mu z\ge 0$ при $0\le z\le\frac{1}{\mu};$ $1-\mu z<0$ при $\frac{1}{\mu}<z<1;\left(\mu>1\right).$

Далее получаем

$$l=\int\limits_0^{1/\mu}\frac{1-\mu z}{1+\mu z}dz-\int\limits_{1/\mu}^1\frac{1-\mu z}{1+\mu z}dz=-\int\limits_0^{1/\mu}\frac{\left(\mu z+1\right)-2}{\mu z+1}dz+\int\limits_{1/\mu}^1\frac{\left(\mu z+1\right)-2}{\mu z+1}dz=$$

$$=-\int\limits_0^{1/\mu}\left(1-\frac{2}{\mu z+1}\right)dz+\int\limits_{1/\mu}^1\left(1-\frac{2}{\mu z+1}\right)dz=$$

$=-\frac{1}{\mu}+\frac{2}{\mu}\ln\left(\mu z+1\right)\Big|_0^{1/\mu}+1-\frac{1}{\mu}-\frac{2}{\mu}\ln\left(\mu z+1\right)\Big|_{1/\mu}^1$. Отсюда следует, что

$$l=1-\frac{2}{\mu}+\frac{4\ln 2}{\mu}-\frac{2}{\mu}\ln\left(\mu+1\right)\underset{\mu\to+\infty}{\to}1.$$

Этот результат доказывает, что при $Y = X$ можно подобрать распределение с коэффициентом детерминации l, сколь угодно близким к единице.

4. Распределение, для которого $l_\lambda = 1$ не существует.

Доказательство.

Пусть $l_\lambda = 1$. Тогда интеграл (5) равен 1. Так же, как в §§ 4.1, 5.2 докажем, что $\left|K(x,y)\right| = 1$ для любых значений x, y с вероятностью 1.

Отсюда следует, что $K(x,y) = \pm 1; \quad \forall x, y.$

4.1. Рассмотрим 1-й случай, когда $K(x,y) = -1$ на некотором множестве D положительной меры. Тогда $f_{XY}(x,y) - f_X(x) f_Y(y) = -f_{XY}(x,y) - f_X(x) f_Y(y).$ Отсюда $f_{XY}(x,y) = 0$ для любых x, y на множестве D значений случайной величины (X,Y), что исключается, так как $\iint\limits_D f_{XY}(x,y)\,dxdy = 1.$

4.2. Пусть теперь $K(x,y) = 1; \quad \forall x, y.$ Тогда $f_{XY}(x,y) - f_X(x) f_Y(y) = f_{XY}(x,y) + f_X(x) f_Y(y).$ Отсюда $f_X(x) f_Y(y) = 0$ везде в области D с вероятностью 1. Любой из этих случаев также исключается, так как $\int\limits_{-\infty}^{+\infty} f_X(x)\,dx = 1; \quad \int\limits_{-\infty}^{+\infty} f_Y(y)\,dy = 1.$

§ 6.2. Пример вычисления коэффициента детерминации с предельным ядром для непрерывных случайных величин.

Для вычисления коэффициента детерминации с предельным ядром при $\lambda = 1$ применяем формулу

$$l = \iint\limits_D \frac{\left|f_{XY}(x,y) - f_X(x) f_Y(y)\right|}{f_{XY}(x,y) + f_X(x) f_Y(y)} f_{XY}(x,y)\,dxdy\,.$$

Рассмотрим непрерывное распределение в треугольнике Δ с вершинами в точках O(0;0), A(1;0), B(0;1), (Рис. 1). Это распределение рассматривалось ранее в § 5.3 для вычисления другого коэффициента детерминации с ядром контингенции. Здесь его удобно рассмотреть еще раз в сравнительных целях. Были ранее получены плотности вероятности

$f_{XY}(x,y) = 2; \quad f_X(x) = 2(1-x); \quad f_Y(y) = 2(1-y); \quad 0 \le x \le 1; \quad 0 \le y \le 1 - x.$

Тогда

$$K(x,y)=\frac{2-2(1-x)2(1-y)}{2+2(1-x)2(1-y)}=\frac{2x+2y-2xy-1}{2xy-2x-2y+3}. \qquad (1)$$

$$l=\int_0^1 dx\int_0^{1-x}\frac{|2x+2y-2xy-1|}{2xy-2x-2y+3}2dxdy. \qquad (2)$$

Чтобы снять знак модуля под знаком интеграла, исследуем знак ядра $K(x,y)$. Знаменатель положителен, так как является суммой плотностей, а числитель может иметь различные знаки. Исследуем знак числителя. Сначала приравняем его нулю: $2x+2y-2xy-1=0.$ Это уравнение кривой внутри треугольника Δ, которая является гиперболой. Действительно, запишем уравнение в виде $2y(1-x)=1-2x;$ Отсюда $y=\frac{x-0,5}{x-1}$. Это дробно-линейная функция, описывающая гиперболу с асимптотами $x=1$ и $y=1$. Гипербола проходит через точки $M\left(\frac{1}{2};0\right);N\left(0;\frac{1}{2}\right)$ и разделяет треугольник Δ на две части. В части, примыкающей к началу координат, ядро $K(x,y)<0$, в остальной части треугольника $K(x,y)>0.$ Запишем теперь формулу (2) подробнее

$$l=2\int_0^{1/2}dx\int_0^{\frac{x-0,5}{x-1}}\frac{2xy+1-2x-2y}{2xy-2x-2y+3}dy+2\int_0^{1/2}dx\int_{\frac{x-0,5}{x-1}}^{1-x}\frac{2x+2y-2xy-1}{2xy-2x-2y+3}dy+$$

$$+2\int_{1/2}^{1}dx\int_0^{1-x}\frac{2x+2y-2xy-1}{2xy-2x-2y+3}dy=2(I_1+I_2+I_3).$$

Здесь через I_1,I_2,I_3 для краткости по порядку написания обозначены интегралы суммы для l. Вычислим каждый из них в отдельности.

$$I_1=\int_0^{1/2}dx\int_0^{\frac{x-0,5}{x-1}}\frac{2xy+1-2x-2y}{2xy-2x-2y+3}dy=\int_0^{1/2}dx\int_0^{\frac{x-0,5}{x-1}}\left(1-\frac{1}{(x-1)}\frac{2(x-1)}{2y(x-1)-2x+3}\right)dy=$$

$$=\int_0^{1/2}\left(\frac{x-0,5}{x-1}-\frac{1}{x-1}\ln(2y(x-1)-2x+3)\Big|_0^{\frac{x-0,5}{x-1}}\right)dx=$$

$$= \int_0^{1/2} \left(\frac{x-0,5}{x-1} - \frac{1}{x-1} \ln\left(2y(x-1) - 2x + 3\right) \Big|_0^{\frac{x-0,5}{x-1}} \right) dx =$$

$$= \int_0^{1/2} \left(1 + \frac{0,5}{x-1} - \frac{\ln 2}{x-1} + \frac{\ln(3-2x)}{x-1} \right) dx =$$

$$= 0,5 + (0,5 - \ln 2) \ln|x-1| \Big|_0^{0,5} + \int_0^{1/2} \frac{\ln(3-2x)}{x-1} dx =$$

$$= 0,5 + (\ln 2 - 0,5) \ln 2 - \int_0^{1/2} \frac{\ln(3-2x)}{1-x} dx .$$

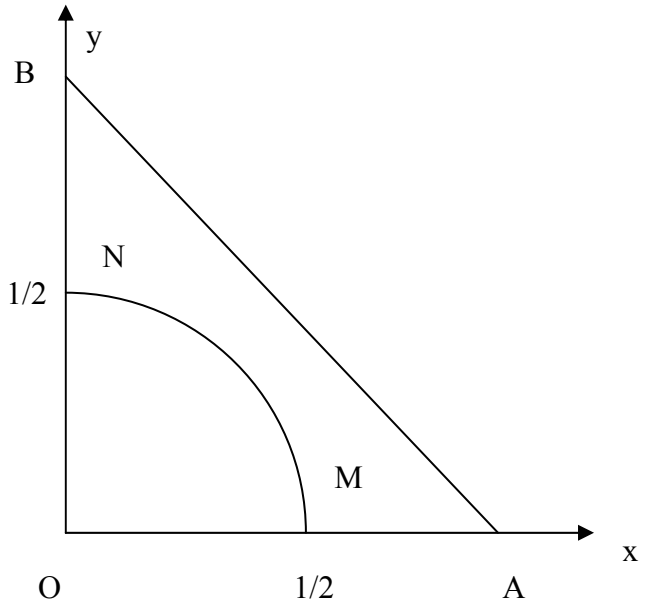


Рис. 1. Треугольник Δ OAB, в котором равномерно распределена случайная величина.

Итак, получено

$$I_1 = 0,5 + (\ln 2 - 0,5) \ln 2 - \int_0^{1/2} \frac{\ln(3-2x)}{1-x} dx. \quad . \qquad (3)$$

Далее вычисляем интеграл I_2.

$$I_2=\int\limits_0^{1/2}dx\int\limits_{\frac{x-0,5}{x-1}}^{1-x}\frac{2x+2y-2xy-1}{2xy-2x-2y+3}dx=\int\limits_0^{1/2}dx\int\limits_{\frac{x-0,5}{x-1}}^{1-x}\left(-1+\frac{1}{x-1}\frac{2(x-1)}{2y(x-1)-2x+3}\right)dy=$$

$$=\int\limits_0^{1/2}\left(-1+x+\frac{x-0,5}{x-1}+\frac{1}{x-1}\ln\left(2y(x-1)-2x+3\right)\Bigg|_{\frac{x-0,5}{x-1}}^{1-x}\right)dx=$$

$$=\int\limits_0^{1/2}\left(x+\frac{0,5}{x-1}+\frac{1}{x-1}\ln\left(-2(x-1)^2-2x+3\right)-\frac{1}{x-1}\ln\left(2(x-0,5)-2x+3\right)\right)dx=$$

$$=0,5+0,5\ln(1-x)\Big|_0^{0,5}+\int\limits_0^{1/2}\left(\frac{1}{x-1}\ln\left(-2x^2+2x+1\right)-\frac{1}{x-1}\ln 2\right)dx=$$

$$=0,5+0,5\ln 0,5-\ln 2\ln|x-1|\Big|_0^{0,5}+\int\limits_0^{1/2}\frac{\ln\left(1+2x-2x^2\right)}{x-1}dx=$$

$$=0,5-0,5\ln 2+(\ln 2)^2+\int\limits_0^{1/2}\frac{\ln\left(1+2x-2x^2\right)}{x-1}dx.$$

Итак, интеграл I_2 выразился формулой

$$I_2=0,5-0,5\ln 2+(\ln 2)^2+\int\limits_0^{1/2}\frac{\ln\left(1+2x-2x^2\right)}{x-1}dx. \qquad (4)$$

Далее вычисляем интеграл I_3 с теми же начальными выкладками, что и для I_2.

$$I_3=\int\limits_{1/2}^{1}dx\int\limits_0^{1-x}\frac{2x+2y-2xy-1}{2xy-2x-2y+3}dy=$$

$$\int\limits_{1/2}^{1}\left[\left(-y+\frac{1}{x-1}\ln\left(2y(x-1)-2x+3\right)\right)\Big|_0^{1-x}\right]dx=$$

$$=\int\limits_{1/2}^{1}\left(-1+x+\frac{1}{x-1}\ln\left(-2(x-1)^2-2x+3\right)-\frac{1}{x-1}\ln(3-2x)\right)dx=$$

$$=-\frac{1}{8}+\int\limits_{1/2}^{1}\frac{\ln\left(-2x^2+2x+1\right)}{x-1}dx-\int\limits_{1/2}^{1}\frac{\ln(3-2x)}{x-1}dx=-\frac{1}{8}-\int\limits_{1/2}^{1}\frac{1}{1-x}\ln\frac{1+2x-2x^2}{3-2x}dx.$$

Итак, интеграл I_3 выразился формулой

$$I_3 = -\frac{1}{8} - \int_{1/2}^{1} \frac{1}{1-x} \ln \frac{1+2x-2x^2}{3-2x} dx . \qquad (5)$$

Запишем теперь выражение для l на основе результатов (3),(4),(5).

$$l = 2\left(I_1 + I_2 + I_3\right) = 1+2\ln^2 2 - \ln 2 + 1 - \ln 2 + 2\ln^2 2 - 0{,}125 - 2\int_0^{1/2} \frac{\ln\left(3-2x\right)}{1-x} dx -$$

$$-2\int_0^{1/2} \frac{\ln\left(1+2x-2x^2\right)}{1-x} dx - 2\int_{1/2}^{1} \frac{\ln\left(1+2x-2x^2\right)}{1-x} dx + 2\int_{1/2}^{1} \frac{\ln\left(3-2x\right)}{1-x} dx =$$

$$=1{,}875+4\ln^2 2 - 2\ln 2 - 2\int_0^{1} \frac{\ln\left(1+2x-2x^2\right)}{1-x} dx - 2\int_0^{1} \frac{\ln\left(3-2x\right)}{1-x} dx +$$

$$4\int_{1/2}^{1} \frac{\ln\left(3-2x\right)}{1-x} dx =$$

$$= 2{,}410518 - 2\int_0^{1} \frac{\ln\left[\left(1+2x-2x^2\right)\left(3-2x\right)\right]}{1-x} dx + 4\int_{1/2}^{1} \frac{\ln\left(3-2x\right)}{1-x} dx .$$

Итак, получаем

$$l = 2{,}410518 - 2\int_0^{1} \frac{\ln\left[\left(1+2x-2x^2\right)\left(3-2x\right)\right]}{1-x} dx + 4\int_{1/2}^{1} \frac{\ln\left(3-2x\right)}{1-x} dx . \qquad (6)$$

Неберущиеся интегралы в формуле (6) вычислим приближенно с помощью метода Симпсона. Заметим, что $x=1$ является устранимой особой точкой для подынтегральных функций. Разрыв устраняем, полагая при $x=1$

$$\frac{\ln\left[\left(1+2x-2x^2\right)\left(3-2x\right)\right]}{1-x} = 4 ; \qquad \frac{\ln\left(3-2x\right)}{1-2x} = 2 . \qquad (7)$$

(Предельные значения функций при $x \to 1-0$).

Так как высокая точность вычислений не нужна (важен лишь метод), то берем небольшое число точек деления промежутков интегрирования: $n=8$.

Составим таблицы значений подынтегральных функций. Положим

$$u = \frac{\ln\left[\left(1+2x-2x^2\right)\left(3-2x\right)\right]}{1-x}; \quad v = \frac{\ln\left(3-2x\right)}{1-x} . \qquad (8)$$

Таблица 1. Значения функции u.

x	0	0,125	0,25	0,375	0,5
u	1,098612	1,382202	1,646326	1,912547	2,197225
x	0,625	0,75	0,875	1	–
u	2,517407	2,895675	3,367754	4	–

Вычисление интеграла $\int_0^1 u(x)dx$ по формуле Симпсона.

$$\int_0^1 u(x)dx \approx \frac{\Delta x}{3}\left[y_0 + y_8 + 4(y_1 + y_3 + y_5 + y_7) + 2(y_2 + y_4 + y_6)\right].$$

Вычислим сначала отдельные части суммы.

$$y_0 + y_8 = 1,098612 + 4 = 5,098612;$$

$$y_1 + y_3 + y_5 + y_7 = 1,382202 + 1,912547 + 2,517407 + 3,367754 = 9,179910;$$

$$y_2 + y_4 + y_6 = 1,646326 + 2,197225 + 2,895675 = 6,739226;$$

Тогда

$$\int_0^1 u(x)dx \approx \frac{0,125}{3}(5,098612 + 4 \cdot 9,179910 + 2 \cdot 6,739226) =$$

$$= \frac{0,125}{3} 55,296704 = 2,304029.$$

Таблица 2. Значения функции v.

x	0,5	0,5625	0,625	0,6875	0,75
v	1,386294	1,436820	1,492309	1,553625	1,621860
x	0,8125	0,875	0,9375	1	–
v	1,698420	1,785148	1,884529	2	–

Вычисление интеграла $\int_{1/2}^1 v(x)dx$ по формуле Симпсона.

$$y_0 + y_8 = 1,386294 + 2 = 3,386294;$$

$$y_1 + y_3 + y_5 + y_7 = 1,436820 + 1,553625 + 1,698420 + 1,884529 = 6,573394;$$

$$y_2 + y_4 + y_6 = 1,492309 + 1,621860 + 1,785148 = 4,899317;$$

Тогда

$$\int_{1/2}^{1} v(x)dx \approx \frac{0,0625}{3}\left(3,386294 + 4 \cdot 6,573394 + 2 \cdot 4,899317\right) =$$

$$= \frac{0,0625}{3}\, 33,066164 = 0,688878.$$

Далее

$$l = 2,410518 - 2\int_{0}^{1} u(x)dx + 4\int_{1/2}^{1} v(x)dx = 2,410518 - 2 \cdot 2,304029 + 4 \cdot 0,688878 =$$

$$= 0,557972 \approx 0,558.$$

Итак, коэффициент детерминации с предельным ядром для данного непрерывного распределения равен

$$l = 0,558. \qquad (9)$$

Полезно сравнить полученное значение коэффициента детерминации со значениями других коэффициентов. Это корректно, так как они вычислены для одного и того же распределения.

Модуль линейного коэффициента корреляции $|\rho| = 0,5$.

Коэффициент детерминации с ассоциативным ядром $as = 0,255.$

Коэффициент детерминации с контингенциальным ядром $co = 0,571.$

Коэффициент детерминации с контингенциальным ядром для дискретного распределения, аппроксимирующего рассматриваемое непрерывное распределение $co = 0,420$.

Хотя некоторые значения коэффициентов вычислены с невысокой точностью, тем не менее значения – близкие. Наиболее жестко оценивает связь коэффициент as.

§ 6.3. Коэффициент детерминации с предельным ядром для дискретных распределений.

Исходим из конструкции предельного ядра для непрерывного распределения

$$K(x,y) = \frac{f_{XY}(x,y) - f_X(x) f_Y(y)}{f_{XY}(x,y) + f_X(x) f_Y(y)}. \qquad (1)$$

Здесь $f_{XY}(x,y); f_X(x); f_Y(y)$ – плотности распределения двумерной случайной величины (X,Y) и ее компонент $X, Y.$

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy; \;\; f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx.$$

Формально, основываясь на аналогии образования конструкций числовых характеристик дискретных и непрерывных распределений, образуем ядро (назовем его по аналогии конструкции также предельным) для дискретного распределения

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij} + p_{i\cdot}p_{\cdot j}}. \qquad (2)$$

Здесь

$$p_{ij} = P\left(X = x_i, Y = y_j\right); \qquad (3)$$

$$p_{i\cdot} = \sum_i p_{ij} = P\left(X = x_i\right);\ p_{\cdot j} = \sum_j p_{ij} = P\left(Y = y_j\right);\ i, j = 1,2,... \qquad (4)$$

Это ядро (2) можно считать аппроксимирующим ядро (1) для соответствующего непрерывного распределения. Действительно, пусть D – область значений двумерной случайной величины $\left(X, Y\right)$. Для простоты будем считать эту область прямоугольником. Разобьем ее на элементарные прямоугольники $D_{ij} = \left[a_i \le x \le b_i; c_j \le y \le d_j\right]$ системой прямых, параллельных координатным осям, с соответственно равным шагом Δx и Δy. $D = \bigcup_{i,j} D_{ij}$. Пусть

$$\iint\limits_{D_{ij}} f_{XY}\left(x, y\right)dxdy = f_{XY}\left(x_i, y_j\right)\Delta x\Delta y = p_{ij};\ \left(x_i, y_j\right) \in D_{ij}.\ \ i, j = 1,2,... \qquad (5)$$

Здесь применена теорема о среднем для двойного интеграла. Пусть далее

$$\int\limits_{a_i}^{b_i} f_X\left(x\right)dx = f_X\left(x_i'\right)\Delta x;\ \int\limits_{c_j}^{d_j} f_Y\left(y\right)dy = f_Y\left(y_j'\right)\Delta y;\ \left(x_i' y_j'\right) \in D_{ij}.$$

Применены теоремы о среднем для определенного интеграла. Будем предполагать плотности распределения непрерывными функциями. Тогда

$f_X\left(x_i'\right) = f_X\left(x_i\right) + \alpha_i;$ $\quad f_Y\left(y_j'\right) = f_Y\left(y_j\right) + \beta_j;$ $\quad \alpha_i \to 0; \beta_j \to 0$ при $\Delta x \to 0, \Delta y \to 0$.

$$p_{i\cdot} = \sum_j p_{ij} = \sum_j \int\limits_{a_i}^{b_i}\int\limits_{c_j}^{d_j} f_{XY}\left(x, y\right)dxdy = \int\limits_{a_i}^{b_i}\int\limits_{c}^{d} f_{XY}\left(x, y\right)dxdy = \int\limits_{a_i}^{b_i} f_X\left(x\right)dx = f_X\left(x_i'\right)\Delta x.$$

Аналогично $p_{\cdot j} = f_Y\left(y_j'\right)\Delta y$. Тогда

$$p_{i\cdot} = f_X\left(x_i\right)\Delta x + \alpha_i\Delta x;\ \ p_{\cdot j} = f_Y\left(y_j\right)\Delta y + \beta_j\Delta y;$$

$$K\left(x_i, y_j\right) = \frac{f_{XY}\left(x_i, y_j\right)\Delta x\Delta y - f_X\left(x_i\right)\Delta x f_Y\left(y_j\right)\Delta y}{f_{XY}\left(x_i, y_j\right)\Delta x\Delta y + f_X\left(x_i\right)\Delta x f_Y\left(y_j\right)\Delta y} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j} - \varepsilon_{ij}}{p_{ij} + p_{i\cdot}p_{\cdot j} + \varepsilon_{ij}}.$$

Здесь $\varepsilon_{ij} = \alpha_i \beta_j \Delta x \Delta y - p_{i.} \beta_j \Delta y - p_{.j} \alpha_i \Delta x \to 0$ при $\Delta x \to 0, \Delta y \to 0$.

Основываясь на ядре (2), построим коэффициент детерминации для случая дискретного распределения

$$l_\lambda = \left[\sum_i \sum_j \left(\frac{\left| p_{ij} - p_{i.} p_{.j} \right|}{p_{ij} + p_{i.} p_{.j}} \right)^\lambda p_{ij} \right]^{1/\lambda}; \quad (\lambda > 0). \qquad (6)$$

Свойства коэффициента детерминации l_λ.

1. Нормированность. $|l_\lambda| \le 1$.

Из конструкции ядра K_{ij} следует, что $|K_{ij}| \le 1$. Тогда $|K_{ij}|^\lambda \le 1$, $(\lambda > 0)$, а потому среднее весовое (весами являются вероятности p_{ij})также удовлетворяет этому неравенству:

$\sum_i \sum_j |K_{ij}|^\lambda p_{ij} \le 1$. Отсюда следует, что $\left[\sum_i \sum_j |K_{ij}|^\lambda \right]^{1/\lambda} \le 1$.

2. Коэффициент детерминации l_λ равен нулю тогда и только тогда, когда случайные величины X, Y независимы.

2.1. Пусть случайные величины X, Y независимы. Тогда выполняется равенство

$$p_{ij} - p_{i.} p_{.j} = 0; \quad \forall ij = 1, 2, ... \qquad (7)$$

Поэтому все слагаемые в сумме $\sum_i \sum_j |K_{ij}|^\lambda$ равны нулю и, следовательно $l_\lambda = 0$.

2.2. Пусть $l_\lambda = 0$. Тогда все слагаемые в сумме формулы (6) равны нулю, так как все слагаемые неотрицательны. Тогда выполняются условия (7), что означает независимость случайных величин X, Y, ибо условия (7) являются необходимыми и достаточными для независимости этих величин.

3. Если случайные величины X, Y равны (полная зависимость), то точной верхней границей значений l_λ является 1.

Доказательство.

Пусть $Y = X$. Это означает, что двумерное распределение – вырожденное, является одномерным. Тогда

$$p_{ij} = P\left(X = x_i, X = x_j \right) = \begin{cases} 0; & i \ne j \\ P\left(X = x_i \right) = p_{i.}; & i = j \end{cases}.$$

В этом случае

$$K_{ij} = K_{ii} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij} + p_{i\cdot}p_{\cdot j}} = \frac{p_{i\cdot} - p_{i\cdot}^2}{p_{i\cdot} + p_{i\cdot}^2} = \frac{1 - p_{i\cdot}}{1 + p_{i\cdot}}; \quad (i = j). \tag{8}$$

Тогда при $\lambda = 1$ получаем

$$l = \sum_i \frac{1 - p_{i\cdot}}{1 + p_{i\cdot}} p_{i\cdot} \ . \tag{9}$$

Положим, например, в этом выражении $p_{i\cdot} = \frac{1}{n}; \quad i = 1, 2, ..., n.$ Тогда

$$l = \sum_{i=1}^{n} \frac{1 - \frac{1}{n}}{1 + \frac{1}{n}} \frac{1}{n} = n \frac{1 - \frac{1}{n}}{1 + \frac{1}{n}} \frac{1}{n} = \frac{1 - \frac{1}{n}}{1 + \frac{1}{n}} \underset{n \to \infty}{\to} 1.$$

Замечание.

Вырожденное одномерное распределение, описывающее случай $Y = X$, является предельным для двумерного. Чтобы показать это, проварьируем все вероятности p_{ij}, определяющие двумерное распределение. Для простоты рассмотрим предыдущий пример, когда $p_{ii} = \frac{1}{n}; \quad p_{ij} = 0$ при $i \neq j; \quad i = 1, 2, ..., n.$ Рассмотрим бесконечно малую величину $\alpha \square \frac{1}{n^2}$. Проварьируем предыдущие вероятности, положив $p_{ii}^* = \frac{1}{n} - \alpha; \quad p_{ij} = \frac{\alpha}{n-1}$ при $i \neq j$. Имеем следующую таблицу распределения

Таблица 1.Двумерное дискретное распределение, для которого $l = l_1 \to 1$.

$X \downarrow \quad Y \to$	1	2	…	n	$p_{i\cdot}$
1	$\frac{1}{n} - \alpha$	$\frac{\alpha}{n-1}$	…	$\frac{\alpha}{n-1}$	$\frac{1}{n}$
2	$\frac{\alpha}{n-1}$	$\frac{1}{n} - \alpha$	…	$\frac{\alpha}{n-1}$	$\frac{1}{n}$
…	…	…	…	…	…
n	$\frac{\alpha}{n-1}$	$\frac{\alpha}{n-1}$	…	$\frac{1}{n} - \alpha$	$\frac{1}{n}$
$p_{\cdot j}$	$1/n$	$1/n$	…	$1/n$	1

Для этого распределения получаем

$$l=\frac{\left(\frac{1}{n}-\alpha\right)-\frac{1}{n^2}}{\left(\frac{1}{n}-\alpha\right)+\frac{1}{n^2}}\left(\frac{1}{n}-\alpha\right)n+\frac{\left|\frac{\alpha}{n-1}-\frac{1}{n^2}\right|}{\frac{\alpha}{n-1}+\frac{1}{n^2}}\frac{\alpha}{n-1}n(n-1)=$$

$$=\frac{1-\alpha n-\frac{1}{n}}{1-\alpha n+\frac{1}{n}}(1-\alpha n)+\frac{\left|\alpha-\frac{n-1}{n^2}\right|}{\alpha+\frac{n-1}{n^2}}\alpha n\underset{n\to\infty}{\to}1\ ,\ \text{так как}\ \alpha n\to 0,\quad \frac{\left|\alpha-\frac{n-1}{n^2}\right|}{\alpha+\frac{n-1}{n^2}}\to 1\ \text{при}$$

$n\to\infty$.

4. Дискретное распределение, для которого $l_\lambda=1$, не существует.

Действительно, пусть $l_\lambda=1$. Тогда так же, как и в § 3.4 докажем, что $\left|K_{ij}\right|=1$ для любых $i,j,$ то есть

$$\left|K_{ij}\right|=\frac{\left|p_{ij}-p_{i.}p_{.j}\right|}{p_{ij}+p_{i.}p_{.j}}=1.\ \ (\forall i,j). \tag{9}$$

Рассмотрим два возможных случая.

4.1. $K_{ij}=1$. Тогда $p_{ij}-p_{i.}p_{.j}=p_{ij}+p_{i.}p_{.j}$. Отсюда $p_{i.}p_{.j}=0$. Получаем, что $p_{i.}=0,\ \ p_{.j}=0$; $(\forall i,j)$. Если предположить, что для какого-либо значения i вероятность $p_{i.}\neq 0$, то тогда для любого значения j должно выполняться равенство $p_{.j}=0$. Иначе не выполняется равенство $p_{i.}p_{.j}=0$. Но тогда не выполняется условие $\sum_j p_{.j}=1$.

4.2. Пусть теперь $K_{ij}=-1$. Из (9) следует, что $p_{ij}-p_{i.}p_{.j}=-p_{ij}-p_{i.}p_{.j}$. Отсюда получаем $p_{ij}=0;(\forall i,j)$, но тогда не выполняется условие $\sum_i\sum_j p_{ij}=1$.

Итак, распределение с указанными свойствами не существует.

§ 6.4. Примеры вычисления коэффициентов детерминации и корреляции для дискретных распределений.

Для вычисления предельного коэффициента детерминации при $\lambda=1$ применяем формулу

$$l = \sum_i \sum_j \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij} .$$

Пример 1. Вычислим коэффициент детерминации l для триномиального распределения, рассмотренного в § 3.3 Для сравнения возьмем те же значения параметров:

$$p_1 = p_2 = \frac{1}{4}; n = 2.$$

Триномиальное распределение определяется формулой

$$p_{ij} = \frac{n!}{i!\, j! \left(n - i - j\right)!} p_1^i p_2^j \left(1 - p_1 - p_2\right)^{n-i-j};$$

$$i, j = 0.1, ..., n; \quad 0 < p_1 < 1; \quad 0 < p_2 < 1; \quad p_1 + p_2 < 1; \quad i + j \le n.$$

Таблица 1 триномиального распределения при $n = 2.$

$X \setminus Y \rightarrow$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00} = 1/4$	$p_{01} = 1/4$	$p_{02} = 1/16$	$p_{0\cdot} = 9/16$
1	$p_{10} = 1/4$	$p_{11} = 1/8$	$p_{12} = 0$	$p_{1\cdot} = 3/8$
2	$p_{20} = 1/16$	$p_{21} = 0$	$p_{22} = 0$	$p_{2\cdot} = 1/16$
$p_{\cdot j}$	$p_{\cdot 0} = 9/16$	$p_{\cdot 1} = 3/8$	$p_{\cdot 2} = 1/16$	1

Вычисления по формуле

$$l = \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij} . \qquad (1)$$

приводятся ниже

$$l = \frac{\left|\frac{1}{4} - \frac{9}{16}\frac{9}{16}\right|}{\frac{1}{4} + \frac{9}{16}\frac{9}{16}}\frac{1}{4} + \frac{\left|\frac{1}{4} - \frac{9}{16}\frac{3}{8}\right|}{\frac{1}{4} + \frac{9}{16}\frac{3}{8}}\frac{1}{4} + \frac{\left|\frac{1}{16} - \frac{9}{16}\frac{1}{16}\right|}{\frac{1}{16} + \frac{9}{16}\frac{1}{16}}\frac{1}{16} + \frac{\left|\frac{1}{4} - \frac{3}{8}\frac{9}{16}\right|}{\frac{1}{4} + \frac{3}{8}\frac{9}{16}}\frac{1}{4} + \frac{\left|\frac{1}{8} - \frac{3}{8}\frac{3}{8}\right|}{\frac{1}{8} + \frac{3}{8}\frac{3}{8}}\frac{1}{8} +$$

$$+ \frac{\left|\frac{1}{16} - \frac{1}{16}\frac{9}{16}\right|}{\frac{1}{16} + \frac{1}{16}\frac{9}{16}}\frac{1}{16} = \frac{17}{145 \cdot 4} + \frac{5}{59 \cdot 4} + \frac{7}{25 \cdot 16} + \frac{5}{59 \cdot 4} + \frac{1}{17 \cdot 8} + \frac{7}{25 \cdot 16} =$$

=0,029310+0,021186+0,017500+0,021186+0,007353+0.017500=0,114035 ≈ 0,114.

Для сравнения для этого же распределения:

Модуль линейного коэффициента корреляции $|\rho| = 0,333.$

Ассоциативный коэффициент детерминации $as = 0,186.$

Контингенциальный коэффициент детерминации $co = 0,430$; (вычисления ниже).

Пример 2. Вычислим контингенциальный коэффициент детерминации co для триномиального распределения при тех же значениях параметров, что и в примере 1.

Вычисления проводятся по формуле

$$co = \sum_{i=0}^{n}\sum_{j=0}^{n} \frac{\left|p_{ij} - p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot}p_{\cdot j}} p_{ij}. \qquad (2)$$

$$co = \frac{\left|\frac{1}{4} - \frac{9}{16}\frac{9}{16}\right|}{\frac{1}{4}\left(1 + 2\frac{1}{4} - 2\frac{9}{16} - 2\frac{9}{16}\right) + \frac{9}{16}\frac{9}{16}}\frac{1}{4} + \frac{\left|\frac{1}{4} - \frac{9}{16}\frac{3}{8}\right|}{\frac{1}{4}\left(1 + 2\frac{1}{4} - 2\frac{9}{16} - 2\frac{3}{8}\right) + \frac{9}{16}\frac{3}{8}}\frac{1}{4} +$$

$$+ \frac{\left|\frac{1}{16} - \frac{9}{16}\frac{1}{16}\right|}{\frac{1}{16}\left(1 + 2\frac{1}{16} - 2\frac{9}{16} - 2\frac{1}{16}\right) + \frac{9}{16}\frac{1}{16}}\frac{1}{16} + \frac{\left|\frac{1}{4} - \frac{3}{8}\frac{9}{16}\right|}{\frac{1}{4}\left(1 + 2\frac{1}{4} - 2\frac{3}{8} - 2\frac{9}{16}\right) + \frac{3}{8}\frac{9}{16}}\frac{1}{4} +$$

$$+ \frac{\left|\frac{1}{8} - \frac{3}{8}\frac{3}{8}\right|}{\frac{1}{8}\left(1 + 2\frac{1}{8} - 2\frac{3}{8} - 2\frac{3}{8}\right) + \frac{3}{8}\frac{3}{8}}\frac{1}{8} + \frac{\left|\frac{1}{16} - \frac{1}{16}\frac{9}{16}\right|}{\frac{1}{16}\left(1 + 2\frac{1}{16} - 2\frac{1}{16} - 2\frac{9}{16}\right) + \frac{1}{16}\frac{9}{16}}\frac{1}{16} =$$

$$= \frac{17}{33\cdot 4} + \frac{5}{18\cdot 4} + \frac{7}{7\cdot 16} + \frac{5}{15\cdot 4} + \frac{1}{13\cdot 8} + \frac{7}{7\cdot 16} =$$

=0,128788+0,083333+0,062500+0,083333+0,009615+0,062500= 0,430069 ≈ 0.430.

Пример 3. Вычисление предельного коэффициента детерминации для дискретного распределения, определяемого таблицей 2,(§ 5.3).

Вычислим значения предельного ядра для этого распределения по формуле

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij} - p_{i\cdot}p_{\cdot j}}. \qquad (3)$$

Таблица 2. Дискретное двумерное равномерное распределение в треугольнике.

$Y \to$ $X \downarrow$	1	2	3	4	5	6	7	$p_{i\cdot}$
1	p_{11} 1/16	0	0	0	0	0	0	1/16
2	p_{21} 1/16	p_{22} 1/16	p_{23} 1/16	0	0	0	0	3/16
3	p_{31} 1/16	p_{32} 1/16	p_{33} 1/16	p_{34} 1/16	p_{35} 1/16	0	0	5/16
4	p_{41} 1/16	p_{42} 1/16	p_{43} 1/16	p_{44} 1/16	p_{45} 1/16	p_{46} 1/16	p_{47} 1/16	7/16
$p_{\cdot j}$	4/16	3/16	3/16	2/16	2/16	1/16	1/16	1

$$K_{11} = \frac{\frac{1}{16} - \frac{1}{16}\frac{4}{16}}{\frac{1}{16} + \frac{1}{16}\frac{4}{16}} = \frac{3}{5};\ K_{21} = \frac{\frac{1}{16} - \frac{3}{16}\frac{4}{16}}{\frac{1}{16} + \frac{3}{16}\frac{4}{16}} = \frac{1}{7};\ K_{22} = K_{23} = \frac{\frac{1}{16} - \frac{3}{16}\frac{3}{16}}{\frac{1}{16} + \frac{3}{16}\frac{3}{16}} = \frac{7}{25};$$

$$K_{31} = \frac{\frac{1}{16} - \frac{4}{16}\frac{5}{16}}{\frac{1}{16} + \frac{4}{16}\frac{5}{16}} = -\frac{1}{9};\ K_{32} = K_{33} = \frac{\frac{1}{16} - \frac{3}{16}\frac{5}{16}}{\frac{1}{16} + \frac{3}{16}\frac{5}{16}} = \frac{1}{31};$$

$$K_{34} = K_{35} = \frac{\frac{1}{16} - \frac{2}{16}\frac{5}{16}}{\frac{1}{16} + \frac{2}{16}\frac{5}{16}} = \frac{3}{13};\ K_{41} = \frac{\frac{1}{16} - \frac{4}{16}\frac{7}{16}}{\frac{1}{16} + \frac{4}{16}\frac{7}{16}} = -\frac{3}{11};$$

$$K_{42} = K_{43} = \frac{\frac{1}{16} - \frac{3}{16}\frac{7}{16}}{\frac{1}{16} + \frac{3}{16}\frac{7}{16}} = -\frac{5}{37};\ K_{44} = K_{45} = \frac{\frac{1}{16} - \frac{2}{16}\frac{7}{16}}{\frac{1}{16} + \frac{2}{16}\frac{7}{16}} = \frac{1}{15};$$

$$K_{46} = K_{47} = \frac{\frac{1}{16} - \frac{1}{16}\frac{7}{16}}{\frac{1}{16} + \frac{1}{16}\frac{7}{16}} = \frac{9}{23}.$$

Используя эти значения ядра, вычислим коэффициент детерминации по формуле

$$l = \sum_i \sum_j \left| K_{ij} \right| p_{ij}. \qquad (4)$$

$$l = \frac{1}{16}\left(\frac{3}{5} + \frac{1}{7} + 2\frac{7}{25} + \frac{1}{9} + 2\frac{1}{31} + 2\frac{3}{13} + \frac{3}{11} + 2\frac{5}{37} + 2\frac{1}{15} + 2\frac{9}{23} \right) =$$

=(0,600000+0,142857+0,560000+0,111111+0,064516+0,461538+0,272727+0,270270+
+0,133333+0,782609)/16=3,398961/16=0,212435.

Итак, $l = 0,212$.

Для сравнения приводим значения других коэффициентов, вычисленные ранее в § 5.4 для этого же дискретного и аналогичного непрерывного распределений.

$co = 0,42$ (контингенциальный коэффициент детерминации для дискретных распределений).

$l = 0,57$ (предельный коэффициент детерминации для непрерывных распределений).

Пример 4. Вычисление предельного коэффициента детерминации для равномерного дискретного треугольного распределения в таблице 8×8, (таблица 3).

Таблица 3. Треугольное дискретное равномерное распределение, (таблица 8×8).

$X \downarrow Y \rightarrow$	1	2	3	4	5	6	7	8	$p_{i\cdot}$
1	1/36								1/36
2	1/36	1/36							2/36
3	1/36	1/36	1/36						3/36
4	1/36	1/36	1/36	1/36					4/36
5	1/36	1/36	1/36	1/36	1/36				5/36
6	1/36	1/36	1/36	1/36	1/36	1/36			6/36
7	1/36	1/36	1/36	1/36	1/36	1/36	1/36		7/36
8	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36	8/36
$p_{\cdot j}$	8/36	7/36	6/36	5/36	4/36	3/36	2/36	1/36	1

Вычисляем коэффициент l, по формулам (3), (4).

$$K_{11} = \frac{\frac{1}{36} - \frac{1}{36}\frac{8}{36}}{\frac{1}{36} + \frac{1}{36}\frac{8}{36}} = \frac{28}{44};\ K_{21} = \frac{\frac{1}{36} - \frac{2}{36}\frac{8}{36}}{\frac{1}{36} + \frac{2}{36}\frac{8}{36}} = \frac{20}{52};\ K_{22} = \frac{\frac{1}{36} - \frac{2}{36}\frac{7}{36}}{\frac{1}{36} + \frac{2}{36}\frac{7}{36}} = \frac{22}{50};$$

$$K_{31}=\frac{\frac{1}{36}-\frac{3}{36}\frac{8}{36}}{\frac{1}{36}+\frac{3}{36}\frac{8}{36}}=\frac{12}{60};\ K_{32}=\frac{\frac{1}{36}-\frac{3}{36}\frac{7}{36}}{\frac{1}{36}+\frac{3}{36}\frac{7}{36}}=\frac{15}{57};\ K_{33}=\frac{\frac{1}{36}-\frac{3}{36}\frac{6}{36}}{\frac{1}{36}+\frac{3}{36}\frac{6}{36}}=\frac{18}{54};$$

$$K_{41}=\frac{\frac{1}{36}-\frac{4}{36}\frac{8}{36}}{\frac{1}{36}+\frac{4}{36}\frac{8}{36}}=\frac{4}{68};\ K_{42}=\frac{\frac{1}{36}-\frac{4}{36}\frac{7}{36}}{\frac{1}{36}+\frac{4}{36}\frac{7}{36}}=\frac{8}{64};\ K_{43}=\frac{\frac{1}{36}-\frac{4}{36}\frac{6}{36}}{\frac{1}{36}+\frac{4}{36}\frac{6}{36}}=\frac{12}{60};$$

$$K_{44}=\frac{\frac{1}{36}-\frac{4}{36}\frac{5}{36}}{\frac{1}{36}+\frac{4}{36}\frac{5}{36}}=\frac{16}{56};\ K_{51}=\frac{\frac{1}{36}-\frac{5}{36}\frac{8}{36}}{\frac{1}{36}+\frac{5}{36}\frac{8}{36}}=-\frac{4}{76};\ K_{52}=\frac{\frac{1}{36}-\frac{5}{36}\frac{7}{36}}{\frac{1}{36}+\frac{5}{36}\frac{7}{36}}=\frac{1}{71};$$

$$K_{53}=\frac{\frac{1}{36}-\frac{5}{36}\frac{6}{36}}{\frac{1}{36}+\frac{5}{36}\frac{6}{36}}=\frac{6}{66};\ K_{54}=\frac{\frac{1}{36}-\frac{5}{36}\frac{5}{36}}{\frac{1}{36}+\frac{5}{36}\frac{5}{36}}=\frac{11}{61};\ K_{55}=\frac{\frac{1}{36}-\frac{5}{36}\frac{4}{36}}{\frac{1}{36}+\frac{5}{36}\frac{4}{36}}=\frac{16}{56};$$

$$K_{61}=\frac{\frac{1}{36}-\frac{6}{36}\frac{8}{36}}{\frac{1}{36}+\frac{6}{36}\frac{8}{36}}=-\frac{12}{84};\ K_{62}=\frac{\frac{1}{36}-\frac{6}{36}\frac{7}{36}}{\frac{1}{36}+\frac{6}{36}\frac{7}{36}}=-\frac{6}{78};\ K_{63}=\frac{\frac{1}{36}-\frac{6}{36}\frac{6}{36}}{\frac{1}{36}+\frac{6}{36}\frac{6}{36}}=0;$$

$$K_{64}=\frac{\frac{1}{36}-\frac{6}{36}\frac{5}{36}}{\frac{1}{36}+\frac{6}{36}\frac{5}{36}}=\frac{6}{66};\ K_{65}=\frac{\frac{1}{36}-\frac{6}{36}\frac{4}{36}}{\frac{1}{36}+\frac{6}{36}\frac{4}{36}}=\frac{12}{60};\ K_{66}=\frac{\frac{1}{36}-\frac{6}{36}\frac{3}{36}}{\frac{1}{36}+\frac{6}{36}\frac{3}{36}}=\frac{18}{54};$$

$$K_{71}=\frac{\frac{1}{36}-\frac{7}{36}\frac{8}{36}}{\frac{1}{36}+\frac{7}{36}\frac{8}{36}}=-\frac{20}{92};\ K_{72}=\frac{\frac{1}{36}-\frac{7}{36}\frac{7}{36}}{\frac{1}{36}+\frac{7}{36}\frac{7}{36}}=-\frac{13}{85};\ K_{73}=\frac{\frac{1}{36}-\frac{7}{36}\frac{6}{36}}{\frac{1}{36}+\frac{7}{36}\frac{6}{36}}=-\frac{6}{78};$$

$$K_{74}=\frac{\frac{1}{36}-\frac{7}{36}\frac{5}{36}}{\frac{1}{36}+\frac{7}{36}\frac{5}{36}}=\frac{1}{71};\ K_{75}=\frac{\frac{1}{36}-\frac{7}{36}\frac{4}{36}}{\frac{1}{36}+\frac{7}{36}\frac{4}{36}}=\frac{8}{64};\ K_{76}=\frac{\frac{1}{36}-\frac{7}{36}\frac{3}{36}}{\frac{1}{36}+\frac{7}{36}\frac{3}{36}}=\frac{15}{57};$$

$$K_{77}=\frac{\frac{1}{36}-\frac{7}{36}\frac{2}{36}}{\frac{1}{36}+\frac{7}{36}\frac{2}{36}}=\frac{22}{50};\ K_{81}=\frac{\frac{1}{36}-\frac{8}{36}\frac{8}{36}}{\frac{1}{36}+\frac{8}{36}\frac{8}{36}}=-\frac{28}{100};\ K_{82}=\frac{\frac{1}{36}-\frac{8}{36}\frac{7}{36}}{\frac{1}{36}+\frac{8}{36}\frac{7}{36}}=-\frac{20}{92};$$

$$K_{83}=\frac{\frac{1}{36}-\frac{8}{36}\frac{6}{36}}{\frac{1}{36}+\frac{8}{36}\frac{6}{36}}=-\frac{12}{84};\ K_{84}=\frac{\frac{1}{36}-\frac{8}{36}\frac{5}{36}}{\frac{1}{36}+\frac{8}{36}\frac{5}{36}}=-\frac{4}{76};\ K_{85}=\frac{\frac{1}{36}-\frac{8}{36}\frac{4}{36}}{\frac{1}{36}+\frac{8}{36}\frac{4}{36}}=\frac{4}{68};$$

$$K_{86}=\frac{\frac{1}{36}-\frac{8}{36}\frac{3}{36}}{\frac{1}{36}+\frac{8}{36}\frac{3}{36}}=\frac{12}{60};\ K_{87}=\frac{\frac{1}{36}-\frac{8}{36}\frac{2}{36}}{\frac{1}{36}+\frac{8}{36}\frac{2}{36}}=\frac{20}{52};\ K_{88}=\frac{\frac{1}{36}-\frac{8}{36}\frac{1}{36}}{\frac{1}{36}+\frac{8}{36}\frac{1}{36}}=\frac{28}{44};$$

$l=$ (0,636364+0,384615+0,440000+0,200000+0,263158+0,333333+0,058824+0,125000+
+0,200000+0,285714+0,052632+0,014085+0,090909+0,180328+0,285714+0,142857+0,076
923+
+0+0,090909+0,200000+0,333333+0,217391+0,152941+0,076923+0,014085+0,125000+
+0,263158+o,440000+0,280000+0,217391+0,142857+0,052632+0,058824+0,20000+
+0,384615+0,636364)/36=7,656879/36=0,212691.

Полезно сравнить этот результат с результатом вычисления этого коэффициента l=0,212435 в предыдущем примере 3 с вдвое меньшим числом клеток таблицы. Небольшая разница лишь в четвертом знаке после запятой.

Пример 5. Распределение новорожденных в ФРГ по религиозной принадлежности отца и матери в 1993 г.

(Источник: *Statistisches Jahrbuch für die BRD.* $-1995, c.74.\ [8, c.294]$).

Статистические данные представлены в следующей таблице 4 (тыс.чел.).

Для распределения, представленного в этой таблице, ранее в § 1.6 были указаны и вычислены следующие коэффициенты связи:

Ассоциативный коэффициент детерминации $as=0,493;$

Коэффициент К. Пирсона $k_p=0,825;$

Коэффициент А.А. Чупрова $k_{ch}=0,731.$

Вычислим для этого распределения контингенциальный коэффициент детерминации co по формулам:

$$co=\sum_i\sum_j\left|K_{ij}\right|p_{ij};\quad K_{ij}=\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}.\qquad(5)$$

Таблица 4. распределения новорожденных в ФРГ по религиозной принадлежности отца и матери.

Религия отца Y → Религия матери X ↓	Евангелическая	Римско-католическая	Прочие христиане	Другие религии	Неверующие и не указавшие	Σ
Евангелическая	146,1	57,6	1,1	0,5	8,8	214,1
Римско-католическая	57,3	195,9	1,1	0,7	5,2	260,2
Прочие христиане	1,3	1,4	10,5	0,1	0,3	13,6
Другие религии	1,8	2,0	0,1	62,8	1,1	67,8
Неверующие и не указавшие	29,1	29,1	0,7	0.8	77,7	124,4
Σ	235,6	273,0	13,5	64,9	93.1	680,1

Формулу (5) для ядра переделаем, заменив вероятности относительными частотами. Получим

$$K_{ij} = \frac{\frac{n_{ij}}{n} - \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}}{\frac{n_{ij}}{n}\left(1 + 2\frac{n_{ij}}{n} - 2\frac{n_{i\cdot}}{n} - 2\frac{n_{\cdot j}}{n}\right) + \frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n}} = \frac{n_{ij}n - n_{i\cdot}n_{\cdot j}}{n_{ij}\left(n + 2n_{ij} - 2n_{i\cdot} - 2n_{\cdot j}\right) + n_{i\cdot}n_{\cdot j}}. \quad (6)$$

Здесь $n = 680{,}1$, а остальные частоты равны числам из соответствующих клеток таблицы 4.

$$K_{11} = \frac{146{,}1 \cdot 680{,}1 - 214{,}1 \cdot 235{,}6}{146{,}1\left(680{,}1 + 2 \cdot 146{,}1 - 2 \cdot 214{,}1 - 2 \cdot 235{,}6\right) + 214{,}1 \cdot 235{,}6} = 0{,}800762;$$

$$K_{12} = \frac{57{,}6 \cdot 680{,}1 - 214{,}1 \cdot 273{,}0}{57{,}6\left(680{,}1 + 2 \cdot 57{,}6 - 2 \cdot 214{,}1 - 2 \cdot 273{,}0\right) + 214{,}1 \cdot 273{,}0} = -0{,}400367;$$

$$K_{13} = \frac{1{,}1 \cdot 680{,}1 - 214{,}1 \cdot 13{,}5}{1{,}1\left(680{,}1 + 2 \cdot 1{,}1 - 2 \cdot 214{,}1 - 2 \cdot 13{,}5\right) + 214{,}1 \cdot 13{,}5} = -0{,}682207;$$

$$K_{14} = \frac{0,5 \cdot 680,1 - 214,1 \cdot 64,9}{0,5\left(680,1 + 2 \cdot 0,5 - 2 \cdot 214,1 - 2 \cdot 64,9\right) + 214,1 \cdot 64,9} = -0,971225;$$

$$K_{15} = \frac{8,8 \cdot 680,1 - 214,1 \cdot 93,1}{8,8\left(680,1 + 2 \cdot 8,8 - 2 \cdot 214,1 - 2 \cdot 93,1\right) + 214,1 \cdot 93,1} = -0,674925;$$

$$K_{21} = \frac{57,3 \cdot 680,1 - 260,2 \cdot 235,6}{57,3\left(680,1 + 2 \cdot 57,3 - 2 \cdot 260,2 - 2 \cdot 235,6\right) + 260,2 \cdot 235,6} = -0,446483;$$

$$K_{22} = \frac{195,9 \cdot 680,1 - 260,2 \cdot 273,0}{195,9\left(680,1 + 2 \cdot 195,9 - 2 \cdot 260,2 - 2 \cdot 273,0\right) + 260,2 \cdot 273,0} = 0,875519;$$

$$K_{23} = \frac{1,1 \cdot 680,1 - 260,2 \cdot 13,5}{1,1\left(680,1 + 2 \cdot 1,1 - 2 \cdot 260,2 - 2 \cdot 13,5\right) + 260,2 \cdot 13,5} = -0,757920;$$

$$K_{24} = \frac{0,7 \cdot 680,1 - 260,2 \cdot 64,9}{0,7\left(680,1 + 2 \cdot 0,7 - 2 \cdot 260,2 - 2 \cdot 64,9\right) + 260,2 \cdot 64,9} = -0,970549;$$

$$K_{25} = \frac{5,2 \cdot 680,1 - 260,2 \cdot 93,1}{5,2\left(680,1 + 2 \cdot 5,2 - 2 \cdot 260,2 - 2 \cdot 93,1\right) + 260,2 \cdot 93,1} = -0,854579;$$

$$K_{31} = \frac{1,3 \cdot 680,1 - 13,6 \cdot 235,6}{1,3\left(680,1 + 2 \cdot 1,3 - 2 \cdot 13,6 - 2 \cdot 235,6\right) + 13,6 \cdot 235,6} = -0,673693;$$

$$K_{32} = \frac{1,4 \cdot 680,1 - 13,6 \cdot 273,0}{1,4\left(680,1 + 2 \cdot 1,4 - 2 \cdot 13,6 - 2 \cdot 273,0\right) + 13,6 \cdot 273,0} = -0,714017;$$

$$K_{33} = \frac{10,5 \cdot 680,1 - 13,6 \cdot 13,5}{10,5\left(680,1 + 2 \cdot 10,5 - 2 \cdot 13,6 - 2 \cdot 13,5\right) + 13,6 \cdot 13,5} = 0,997720;$$

$$K_{34} = \frac{0,1 \cdot 680,1 - 13,6 \cdot 64,9}{0,1\left(680,1 + 2 \cdot 0,1 - 2 \cdot 13,6 - 2 \cdot 64,9\right) + 13,6 \cdot 64,9} = -0,871290;$$

$$K_{35} = \frac{0,3 \cdot 680,1 - 13,6 \cdot 93,1}{0,3\left(680,1 + 2 \cdot 0,3 - 2 \cdot 13,6 - 2 \cdot 93,1\right) + 13,6 \cdot 93,1} = -0,755239;$$

$$K_{41} = \frac{1,8 \cdot 680,1 - 67,8 \cdot 235,6}{1,8\left(680,1 + 2 \cdot 1,8 - 2 \cdot 67,8 - 2 \cdot 235,6\right) + 67,8 \cdot 235,6} = -0,889716;$$

$$K_{42} = \frac{2,0 \cdot 680,1 - 67,8 \cdot 273,0}{2,0\left(680,1 + 2 \cdot 2,0 - 2 \cdot 67,8 - 2 \cdot 273,0\right) + 67,8 \cdot 273,0} = -0,926274;$$

$$K_{43}=\frac{0,1\cdot 680,1-67,8\cdot 13,5}{0,1\left(680,1+2\cdot 0,1-2\cdot 67,8-2\cdot 13,5\right)+67,8\cdot 13,5}=-0,876141;$$

$$K_{44}=\frac{62,8\cdot 680,1-67,8\cdot 64,9}{62,8\left(680,1+2\cdot 62,8-2\cdot 67,8-2\cdot 64,9\right)+67,8\cdot 64,9}=0,999452;$$

$$K_{45}=\frac{1,1\cdot 680,1-67,8\cdot 93,1}{1,1\left(680,1+2\cdot 1,1-2\cdot 67,8-2\cdot 93,1\right)+67,8\cdot 93,1}=-0,829378;$$

$$K_{51}=\frac{29,1\cdot 680,1-124,4\cdot 235,6}{29,1\left(680,1+2\cdot 29,1-2\cdot 124,4-2\cdot 235,6\right)+124,4\cdot 235,6}=-0,318946;$$

$$K_{52}=\frac{16,1\cdot 680,1-124,4\cdot 273,0}{16,1\left(680,1+2\cdot 16,1-2\cdot 124,4-2\cdot 273,0\right)+124,4\cdot 273,0}=-0,705164;$$

$$K_{53}=\frac{0,7\cdot 680,1-124,4\cdot 13,5}{0,7\left(680,1+2\cdot 0,7-2\cdot 124,4-2\cdot 13,5\right)+124,4\cdot 13,5}=-0,612884;$$

$$K_{54}=\frac{0,8\cdot 680,1-124,4\cdot 64,9}{0,8\left(680,1+2\cdot 0,8-2\cdot 124,4-2\cdot 64,9\right)+124,4\cdot 64,9}=-0,905417;$$

$$K_{55}=\frac{77,7\cdot 680,1-124,4\cdot 93,1}{77,7\left(680,1+2\cdot 77,7-2\cdot 124,4-2\cdot 93,1\right)+124,4\cdot 93,1}=0,966315;$$

Запишем все значения ядра в таблицу 5.

Таблица 5.Значения контингенциального ядра K_{ij} и вероятностей p_{ij} к примеру 4.

$j\rightarrow$ $i\downarrow$	1	2	3	4	5
1	0,800762 0,214821	–0,400367 0,084693	–0,682207 0,001617	–0,971225 0,000735	–0,674925; 0,012939
2	–0,446483; 0,084252	0,875519; 0,288046	–0,757920; 0,001617	–0,970549; 0,001029	–0,854579; 0,007646
3	–0,673693; 0,001911	–0,714017 0,002059	0,997720; 0,015439	–0,871290; 0,000147	–0,755239; 0,000441
4	–0,889716; 0,002647	–0,926274; 0,002941	–0,876141; 0,000147	0,999452; 0,092339	–0,829378; 0,001617
5	–0,318946; 0,042788	–0,705164; 0,042788	–0,612884; 0,001029	–0,905417; 0,001176	0,966315; 0,114248

С помощью таблицы 5 вычисляем коэффициент co по первой формуле (5):

$co\ \tau_1 = 0{,}172020+0{,}033908+0{,}001103+0{,}000714+0{,}008733+$

$+0{,}037617+0{,}252190+0{,}001226+0{,}000999+0{,}006534+$

$+0{,}001287+0{,}001470+0{,}015404+0{,}001281+0{,}000333+$

$+0{,}002355+0{,}002724+0{,}000129+0{,}092288+0{,}001341+$

$+0{,}013647+0{,}030173+0{,}000631+0{,}001065+0{,}110400=0{,}789572.$

Итак, $co = 0,789572 \approx 0,790.$

С помощью таблицы 5 вычислим также контингенциальный коэффициент корреляции co_c, учтя, что положительные значения ядер в таблице 5 стоят на главной диагонали.

$$co_c = \sum_i \sum_j K_{ij} p_{ij} = -0,789572 + 2 \cdot 0,172020 + 2 \cdot 0,252190 + 2 \cdot 0,015404 +$$

$$+2 \cdot 0,092288 + 2 \cdot 0,110400 = 0,495032 \approx 0,495.$$

Пример 6. Вычисление предельного коэффициента детерминации l по данным таблицы 5 (пример 5) по формулам (1), (3), (4):

$$l = \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij}; \qquad K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij} - p_{i\cdot} p_{\cdot j}}; \qquad l = \sum_i \sum_j \left| K_{ij} \right| p_{ij}.$$

Сначала вычислим ядра (3).

$$K_{11} = \frac{146,1 \cdot 680,1 - 214,1 \cdot 235,6}{146,1 \cdot 680,1 + 214,1 \cdot 235,6} = \frac{99362.61 - 50441,96}{99362.61 + 50441,96} = \frac{48920.65}{149804.57} =$$

$0,326563$

$$K_{12} = \frac{57,6 \cdot 680,1 - 214,1 \cdot 273,0}{57,6 \cdot 680,1 + 214,1 \cdot 273,0} = \frac{-19275,54}{97623,06} = -0,197449$$

$$K_{13} = \frac{1,1 \cdot 680,1 - 214,1 \cdot 13,5}{1,1 \cdot 680,1 + 214,1 \cdot 13,5} = \frac{748,11 - 2890,35}{748,11 + 2890,35} = \frac{-2142,24}{3638,46} = -0,588777$$

$$K_{14} = \frac{0,5 \cdot 680,1 - 214,1 \cdot 64,9}{0,5 \cdot 680,1 + 214,1 \cdot 64,9} = \frac{340,05 - 13895,09}{340,05 + 13895,09} = \frac{-13555,04}{14235,14} = -0,952224$$

$$K_{15} = \frac{8,8 \cdot 680,1 - 214,1 \cdot 93,1}{8,8 \cdot 680,1 + 214,1 \cdot 93,1} = \frac{5984,88 - 19932,71}{5984,88 + 19932,71} = -\frac{13947,83}{25917,59} = -0,538161$$

$$K_{21}=\frac{57,3\cdot 680,1-260,2\cdot 235,6}{57,3\cdot 680,1+260,2\cdot 235,6}=\frac{38969,73-61303,12}{38969,73+61303,12}=\frac{-22333,39}{100272,85}=$$
$$-0,222726$$

$$K_{22}=\frac{195,9\cdot 680,1-260,2\cdot 273,0}{195,9\cdot 680,1+260,2\cdot 273,0}=\frac{133231,59-71034,6}{133231,59+71034,6}=\frac{62196,99}{204266,19}=$$
$$0,304490$$

$$K_{23}=\frac{1,1\cdot 680,1-260,2\cdot 13,5}{1,1\cdot 680,1+260,2\cdot 13,5}=\frac{748,11-3512,7}{748,11+3512,7}=\frac{-2764,59}{4260,81}=-0,648841$$

$$K_{24}=\frac{0,7\cdot 680,1-260,2\cdot 64,9}{0,7\cdot 680,1+260,2\cdot 64,9}=\frac{476,07-16886,98}{476,07+16886,98}=\frac{-16410,91}{17363,05}=-0,945163$$

$$K_{25}=\frac{5,2\cdot 680,1-260,2\cdot 93,1}{5,2\cdot 680,1+260,2\cdot 93,1}=\frac{3536,52-24224,62}{3536,52+24224,62}=\frac{-20688,1}{27761,14}=-0,745218$$

$$K_{31}=\frac{1,3\cdot 680,1-13,6\cdot 235,6}{1,3\cdot 680,1+13,6\cdot 235,6}=\frac{884,13-3204,16}{884,13+3204,16}=\frac{-2320,03}{4088,29}=-0,567482$$

$$K_{32}=\frac{1,4\cdot 680,1-13,6\cdot 273,0}{1,4\cdot 680,1+13,6\cdot 273,0}=\frac{-2760,66}{4664,94}=-0,591789$$

$$K_{33}=\frac{10,5\cdot 680,1-13,6\cdot 13,5}{10,5\cdot 680,1+13,6\cdot 13,5}=\frac{6958,8}{7323,3}=0,950227$$

$$K_{34}=\frac{0,1\cdot 680,1-13,6\cdot 64,9}{0,1\cdot 680,1+13,6\cdot 64,9}=\frac{-814,63}{950,65}=-0,856919$$

$$K_{35}=\frac{0,3\cdot 680,1-13,6\cdot 93,1}{0,3\cdot 680,1+13,6\cdot 93,1}=\frac{-1062,13}{1470,19}=-0,722444$$

$$K_{41}=\frac{1,8\cdot 680,1-67,8\cdot 235,6}{1,8\cdot 680,1+67,8\cdot 235,6}=\frac{-14749,5}{17197,86}=-0,857636$$

$$K_{42}=\frac{2,0\cdot 680,1-67,8\cdot 273,0}{2,0\cdot 680,1+67,8\cdot 273,0}=\frac{-17149,2}{19869,6}=-0,863087$$

$$K_{43}=\frac{0,1\cdot 680,1-67,8\cdot 13,5}{0,1\cdot 680,1+67,8\cdot 13,5}=\frac{-847,29}{983,31}=-0,861671$$

$$K_{44}=\frac{62,8\cdot 680,1-67,8\cdot 64,9}{62,8\cdot 680,1+67,8\cdot 64,9}=\frac{38310,06}{47110,5}=0,813196$$

$$K_{45}=\frac{1{,}1\cdot 680{,}1-67{,}8\cdot 93{,}1}{1{,}1\cdot 680{,}1+67{,}8\cdot 93{,}1}=\frac{-5564{,}07}{7060{,}29}=-0{,}788080$$

$$K_{51}=\frac{29{,}1\cdot 680{,}1-124{,}4\cdot 235{,}6}{29{,}1\cdot 680{,}1+124{,}4\cdot 235{,}6}=\frac{-9517{,}73}{49099{,}55}=-0{,}193846$$

$$K_{52}=\frac{16{,}1\cdot 680{,}1-124{,}4\cdot 273{,}0}{16{,}1\cdot 680{,}1+124{,}4\cdot 273{,}0}=\frac{-23011{,}59}{44910{,}81}=-0{,}512384$$

$$K_{53}=\frac{0{,}7\cdot 680{,}1-124{,}4\cdot 13{,}5}{0{,}7\cdot 680{,}1+124{,}4\cdot 13{,}5}=\frac{-1203{,}33}{2155{,}47}=-0{,}558268$$

$$K_{54}=\frac{0{,}8\cdot 680{,}1-124{,}4\cdot 64{,}9}{0{,}8\cdot 680{,}1+124{,}4\cdot 64{,}9}=\frac{-7529{,}48}{8617{,}64}=-0{,}873729$$

$$K_{55}=\frac{77{,}7\cdot 680{,}1-124{,}4\cdot 93{,}1}{77{,}7\cdot 680{,}1+124{,}4\cdot 93{,}1}=\frac{41262{,}13}{64425{,}41}=0{,}640464$$

Значения ядер и вероятности запишем в таблицу 6.

Таблица 6. Значения ядер и вероятности для вычисления предельного коэффициента детерминации по формуле (4).

$j\rightarrow$ $i\downarrow$	1	2	3	4	5
1	0,326563 0,214821	–0,197449 0 ,084693	–0,588777 0,001617	–0,952224 0,000735	–0,538161 0,012939
2	–0,222726 0,084252	0,304490 0,288046	–0,648841 0,001617	–0,945163 0,001029	–0,745218 0,007646
3	– 0,567482 0,001911	– 0,591789 0,002059	0,950227 0,015439	– 0,856919 0,000147	– 0,722444 0,000441
4	– 0,857636 0,002647	– 0,863087 0,002941	– 0,861671 0,000147	0,813196 0,092339	– 0,788080 0.001617
5	– 0,193846 0,042788	– 0,512384 0,023673	– 0,558268 0,001029	– 0,873729 0,001176	0,640464 0,114248

На основе таблицы (6) вычисляем предельный коэффициент детерминации по формуле (4):

0,070153+0,016723+0,000952+0,000700+0,006963+0,018765+0,087707+0,001049+

+0,000973+0,005698+0,001084+0,001218+0,014671+0,000126+0,000319+0,002270+

+0,002538+0,000127+0,075090+0,001274+0,008294+0,012130+0,000574+0,001028+

+0,073172=0,402646.

$l \approx 0,403.$

Глава 7. Комбинированные коэффициенты детерминации для дискретных случайных величин

В предыдущих параграфах рассматривались коэффициенты детерминации для дискретных случайных величин, которые регистрируют величину зависимости только по вероятностям ($as_\lambda, co_\lambda, l_\lambda$), что является их определенным достоинством, так как применимы для количественных, качественных и смешанных признаков. Линейный коэффициент корреляции ρ применяется для исследования зависимости только количественных признаков, так как в его конструкции участвуют значения случайных величин. Однако величина зависимости определяется как вероятностями значений, так и самими значениями случайных величин, поэтому представляется целесообразным рассмотреть комбинированные измерители величины связи и по значениям и по вероятностям значений случайных величин.

Внимательный анализ показывает, что представляет интерес комбинация линейного коэффициента корреляции с ассоциативным или контингенциальным коэффициентами детерминации as_λ и co_λ. Это объясняется тем, что ядра указанных коэффициентов нормированы неравенствами $-1 \le K_{ij} \le 1$ с достижимыми границами и обращаются в нуль только в случае независимости случайных величин. Ограничимся случаем $\lambda = 1.$

§ 7.1.(3.12) Комбинированные коэффициенты детерминации дискретных случайных величин: комби-ас и комби-конт.

Коэффициент детерминации комби-ас определяется формулой

$$com_a = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left| x_i - m_X \right| \left| y_j - m_Y \right| \left| K_{ij} \right| p_{ij} \quad . \tag{1}$$

Здесь

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}}; \tag{2}$$

$p_{ij} = P\left(X = x_i, Y = y_j\right); \quad p_{i\cdot} = P\left(X = x_i\right); \quad p_{\cdot j} = P\left(Y = y_j\right); \quad i, j = 1, 2, \ldots$;

m_X, m_Y – математические ожидания, σ_X, σ_Y – средние квадратические отклонения случайных величин X, Y.

Коэффициент детерминации комби-конт определяется формулой

$$com_c = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left| x_i - m_X \right| \left| y_j - m_Y \right| \left| K_{ij} \right| p_{ij} . \tag{3}$$

Здесь

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} . \tag{4}$$

Свойства коэффициентов $com_a; com_c$**.**

Так как свойства $com_a; com_c$ одинаковы, то доказательство свойств проведем одновременно для обоих коэффициентов, обозначив их единым символом com**.**

1. Нормировка: $0 \le com \le 1$.

Доказательство. Используем свойство ядер $\left| K_{ij} \right| \le 1$ для любых i, j. Тогда получаем неравенство $com \le \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left| x - m_X \right| \left| y - m_Y \right| p_{ij}$.

Далее применяем неравенство Коши-Буняковского. Получаем

$$\frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left[\left| x_i - m_X \right| \sqrt{p_{ij}} \right] \left[\left| y_j - m_Y \right| \sqrt{p_{ij}} \right] \le$$

$$\le \frac{1}{\sigma_X \sigma_Y} \sqrt{\sum_i \sum_j \left(x_i - m_X \right)^2 p_{ij}} \sqrt{\sum_i \sum_j \left(y_j - m_Y \right)^2 p_{ij}} =$$

$$= \frac{1}{\sigma_X \sigma_Y} \sqrt{\sum_i \left(x_i - m_X \right)^2 \sum_j p_{ij}} \sqrt{\sum_j \left(y_j - m_Y \right)^2 \sum_i p_{ij}} =$$

$$= \frac{1}{\sigma_X \sigma_Y} \sqrt{\sum_i \left(x_i - m_X\right)^2 p_{i\cdot}} \sqrt{\sum_j \left(y_j - m_Y\right)^2 p_{\cdot j}} = \frac{1}{\sigma_X \sigma_Y} \sigma_X \sigma_Y = 1.$$

Применены формулы согласованности $\sum_j p_{ij} = p_{i\cdot}; \sum_i p_{ij} = p_{\cdot j}.$

Итак, получили: $com \le 1$. Неравенство $com \ge 0$ очевидно, так как все слагаемые под знаком суммы неотрицательны.

2. Коэффициент детерминации com обращается в нуль тогда и только тогда, когда случайные величины X, Y независимы.

Доказательство. В составе ядер их числитель равен разности $p_{ij} - p_{i\cdot} p_{\cdot j}$. Он обращается в нуль при всех i, j тогда и только тогда, когда случайные величины X, Y независимы.

Отсюда следует справедливость свойства 2.

3. Если между случайными величинами X, Y имеется линейная зависимость: $Y = aX + b$, то коэффициент детерминации $com = 1$.

Доказательство. В этом случае

$$p_{\cdot j} = P\left(Y = y_j\right) = P\left(aX + b = ax_j + b\right) = P\left(X = x_j\right) = p_{j\cdot}.$$

$$p_{ij} = P\left(X = x_i, Y = y_j\right) = P\left(X = x_i, aX + b = ax_j + b\right) = P\left(X = x_i, X = x_j\right) =$$

$$= \begin{cases} p_{i\cdot}; & i = j \\ 0; & i \ne j \end{cases}.$$

Тогда контингенциальное ядро (4) принимает вид

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} = \frac{p_{i\cdot} - p_{i\cdot} p_{j\cdot}}{p_{i\cdot}\left(1 + 2p_{i\cdot} - 2p_{i\cdot} - 2p_{j\cdot}\right) + p_{i\cdot} p_{j\cdot}} =$$

$$= \frac{1 - p_{j\cdot}}{1 - p_{j\cdot}} = \frac{1 - p_{i\cdot}}{1 - p_{i\cdot}} = 1 \text{ при } i = j;$$

$$K_{ij} = \frac{-p_{i\cdot} p_{j\cdot}}{p_{i\cdot} p_{j\cdot}} = -1 \text{ при } i \ne j.$$

Аналогично и для ассоциативного ядра получаем

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}} = \frac{p_{i\cdot} - p_{i\cdot} p_{i\cdot}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{i\cdot}\left(1 - p_{i\cdot}\right)}} = \frac{p_{i\cdot}\left(1 - p_{i\cdot}\right)}{p_{i\cdot}\left(1 - p_{i\cdot}\right)} = 1 \text{ при}$$

$i = j$;

$$K_{ij}=\frac{-p_{i\cdot}p_{j\cdot}}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{j\cdot}\left(1-p_{j\cdot}\right)}}=\frac{-\sqrt{p_{i\cdot}p_{j\cdot}}}{\sqrt{\left(1-p_{i\cdot}\right)\left(1-p_{j\cdot}\right)}}\ \text{при}\ \ i\neq j\,.$$

В суммах (1) и (3) останутся слагаемые только при $i=j$, так как $p_{ij}=0$ при $i\neq j$.

Тогда

$$com=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j\left|x_i-m_X\right|\left|y_j-m_Y\right|p_{ij}=$$

$$=\frac{1}{\sigma_X\left|a\right|\sigma_X}\sum_i\left|x_i-m_X\right|\left|ax_i+b-am_X-b\right|p_{i\cdot}=$$

$$=\frac{1}{\left|a\right|\sigma_X^2}\sum_i\left|a\right|\left(x_i-m_X\right)^2p_{i\cdot}=\frac{\left|a\right|\sigma_X^2}{\left|a\right|\sigma_X^2}=1\,.$$

Здесь использован тот факт, что

$$\sigma_Y=\sqrt{\sum_i\left(ax_i+b-am_X-b\right)^2p_{i\cdot}}=\left|a\right|\sqrt{\sum_i\left(x_i-m_X\right)^2p_{i\cdot}}=\left|a\right|\sigma_X\,.$$

4. Если коэффициент детерминации com (com_a или com_c) равен 1, то случайные величины линейно зависимы.

Пусть

$$com=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j\left|x_i-m_X\right|\left|y_j-m_Y\right|\left|K_{ij}\right|p_{ij}=1.\qquad(5)$$

Тогда объединенный множитель при вероятности p_{ij} в каждом слагаемом суммы формулы (5) равен 1:

$$A_{ij}=\frac{\left|x_i-m_X\right|}{\sigma_X}\frac{\left|y_j-m_Y\right|}{\sigma_Y}\left|K_{ij}\right|=1;\quad\forall i,j.\qquad(6)$$

Действительно, предположим противное, что для какой-либо пары $\left(l,n\right)$ имеет место неравенство (для простоты для одной пары)

$$A_{\ln}=\frac{\left|x_l-m_X\right|}{\sigma_X}\frac{\left|y_n-m_Y\right|}{\sigma_Y}\left|K_{\ln}\right|<1\,.\qquad(7)$$

(Неравенство противоположного смысла быть не может, так как $com\le1$).

Рассмотрим систему соотношений

$$\begin{cases}A_{ij}=1;\quad i,j=1,2,...;i\neq l,j\neq n;\\ A_{\ln}<1\end{cases}.\qquad(8)$$

Умножим каждое соотношение системы (8) на соответствующую вероятность p_{ij} и просуммируем по всем i, j. Получаем

$\sum_{i\neq l}\sum_{j\neq n} A_{ij}p_{ij} + A_{\ln}p_{\ln} < \sum_{i\neq l}\sum_{j\neq n} p_{ij} + p_{\ln} = 1$. Получаем: $com < 1$. Это неравенство противоречит исходному равенству (5). Противоречие доказывает, что $A_{ij} = 1; \forall i, j.$

Докажем теперь, что $\left|K_{ij}\right| = 1; \forall i, j.$

Доказываем от противного, пусть для какой-либо пары индексов $i = l, j = n$ имеет место неравенство

$$\left|K_{\ln}\right| < 1. \tag{9}$$

(Для простоты для одной пары). Заметим, что $\left|K_{\ln}\right| > 1$ быть не может, так как имеет место общее неравенство $\left|K_{ij}\right| \le 1$. Тогда в силу доказанного равенства (6) имеем неравенство

$$B_{\ln} = \frac{\left|x_l - m_X\right|}{\sigma_X}\frac{\left|y_n - m_n\right|}{\sigma_Y} > 1\ . \tag{10}$$

Запишем систему соотношений

$$\begin{cases} B_{ij} = \dfrac{\left|x_i - m_X\right|}{\sigma_X}\dfrac{\left|y_j - m_Y\right|}{\sigma_Y} = 1; \quad \left(i, j = 1,2,...; i \neq l, j \neq n\right) \\ B_{\ln} > 1 \end{cases}. \tag{11}$$

Умножим каждое соотношение (11) на соответствующую вероятность p_{ij} и просуммируем по всем $i, j.$ Получаем

$\sum_{i\neq l}\sum_{j\neq n} B_{ij}p_{ij} + B_{\ln}p_{\ln} > \sum_{i\neq l}\sum_{j\neq n} p_{ij} + p_{\ln} = 1$. Таким образом, имеем

$$\sum_i\sum_j \frac{\left|x_i - m_X\right|}{\sigma_X}\frac{\left|y_j - m_Y\right|}{\sigma_Y}p_{ij} > 1. \tag{12}$$

С другой стороны, в силу неравенства Коши-Буняковского, получаем другое соотношение:

$$\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j\left(\left|x_i - m_X\right|\sqrt{p_{ij}}\right)\left(\left|y_j - m_Y\right|\sqrt{p_{ij}}\right) \le$$

$$\le \frac{1}{\sigma_X\sigma_Y}\sqrt{\sum_i\left(x_i - m_X\right)^2\sum_j p_{ij}}\sqrt{\sum_j\left(y_j - m_Y\right)^2\sum_i p_{ij}} =$$

$$= \frac{1}{\sigma_X \sigma_Y} \sqrt{\sum_i \left(x_i - m_X\right)^2 p_{i\cdot}} \sqrt{\sum_j \left(y_j - m_Y\right)^2 p_{\cdot j}} = \frac{1}{\sigma_X \sigma_Y} \sigma_X \sigma_Y = 1.$$

Итак,

$$\sum_i \sum_j \frac{\left|x_i - m_X\right|}{\sigma_X} \frac{\left|y_j - m_Y\right|}{\sigma_Y} p_{ij} \le 1. \tag{13}$$

Это неравенство противоречит неравенству (12). Противоречие доказывает, что

$$\left|K_{ij}\right| = 1; \quad \forall i, j. \tag{14}$$

При этом одновременно имеют место равенства

$$B_{ij} = \frac{\left|x_i - m_X\right|}{\sigma_X} \frac{\left|y_j - m_Y\right|}{\sigma_Y} = 1; \quad \forall i, j. \tag{15}$$

Докажем теперь, что случайные величины X, Y линейно зависимы.

Для этого умножим каждое равенство (15) на соответствующую вероятность p_{ij} и просуммируем по всем i, j. Получаем

$$\sum_i \sum_j B_{ij} p_{ij} = \sum_i \sum_j p_{ij} = 1. \tag{16}$$

Наряду со случайными величинами X, Y рассмотрим центрированные и нормированные случайные величины

$$X' = \frac{X - m_X}{\sigma_X}, \quad Y' = \frac{Y - m_Y}{\sigma_Y} \ . \tag{17}$$

$$MX' = MY' = 0; \quad DX' = DY' = 1. \tag{18}$$

Отсюда

$$M\left(X'\right)^2 = DX' = 1; \ M\left(Y'\right)^2 = DY' = 1. \tag{19}$$

Далее, равенство (16) можно записать короче в виде

$$M\left(\left|X'\right|\left|Y'\right|\right) = 1. \tag{20}$$

Рассмотрим

$$M\left[\left(\left|X'\right| - \left|Y'\right|\right)^2\right] = M\left(X'\right)^2 - 2M\left(\left|X'\right|\left|Y'\right|\right) + M\left(Y'\right)^2 = 1 - 2 \cdot 1 + 1 = 0.$$

Это равенство запишем подробнее

$\sum_i \sum_j \left(\frac{|x_i - m_X|}{\sigma_X} - \frac{|y_j - m_Y|}{\sigma_Y} \right)^2 p_{ij} = 0.$ Так как все слагаемые неотрицательны, то равенство нулю суммы означает равенство нулю каждого слагаемого. Вероятности $p_{ij} \neq 0$, поэтому имеют место равенства

$\frac{|x_i - m_X|}{\sigma_X} - \frac{|y_j - m_Y|}{\sigma_Y} = 0; \quad \forall i, j.$ Отсюда

$$\frac{|y_j - m_Y|}{\sigma_Y} = \frac{|x_i - m_X|}{\sigma_X}; \quad \forall i, j. \tag{21}$$

Зафиксируем в этом равенстве индекс j и будем менять индекс i. Получим серию противоречивых равенств. Равенства (21) возможны только при $j = i$. Это означает, что двумерное распределение в этом случае вырождается в одномерное и равенства (21) принимают вид

$$\frac{|y_i - m_Y|}{\sigma_Y} = \frac{|x_i - m_X|}{\sigma_X}; \quad i = 1, 2, \ldots \tag{22}$$

Отсюда

$y_i - m_Y = \pm \frac{\sigma_Y}{\sigma_X}(x_i - m_X);$ и далее

$$y_i = m_Y \pm \frac{\sigma_Y}{\sigma_X}(x_i - m_X); \quad i = 1, 2, \ldots \tag{23}$$

Равенства (23) означают, что случайная величина Y является линейной функцией случайной величины X:

$$Y = m_Y - \frac{\sigma_Y}{\sigma_X} m_X + \frac{\sigma_Y}{\sigma_X} X = a_1 X + b_1 \text{ или } Y = m_Y + \frac{\sigma_Y}{\sigma_X} m_X - \frac{\sigma_Y}{\sigma_X} X = a_2 X + b_2 . \tag{24}$$

Свойство доказано.

Примеры вычисления детерминационных коэффициентов комби-ас и комби-конт приведены в следующем параграфе.

§ 7.2. Примеры вычисления коэффициентов комби-ас и комби-конт.

Сравнение величин коэффициентов детерминации для триномиального распределения.

Пример1. Двумерная случайная величина (X,Y) распределена по триномиальному закону, который определяется формулой

$$p_{ij}=\frac{n!}{i!\,j!(n-i-j)!}p_1^i p_2^j(1-p_1-p_2)^{n-i-j}; \qquad (1)$$

$$i,j=0.1,...,n;\quad 0<p_1<1;\quad 0<p_2<1;\quad p_1+p_2<1;\quad i+j\le n.$$

Рассмотрим случай $n=2;\quad p_1=p_2=1/4$. Построим таблицу распределения (табл.1).

Таблица 1 триномиального распределения при $n=2$.

$X \setminus Y \to$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00}=1/4$	$p_{01}=1/4$	$p_{02}=1/16$	$p_{0\cdot}=9/16$
1	$p_{10}=1/4$	$p_{11}=1/8$	$p_{12}=0$	$p_{1\cdot}=3/8$
2	$p_{20}=1/16$	$p_{21}=0$	$p_{22}=0$	$p_{2\cdot}=1/16$
$p_{\cdot j}$	$p_{\cdot 0}=9/16$	$p_{\cdot 1}=3/8$	$p_{\cdot 2}=1/16$	1

С помощью этой таблицы вычислим коэффициент детерминации комби-ас по формулам

$$com_a=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j|x_i-m_X||y_j-m_Y||K_{ij}|p_{ij}; \qquad (2)$$

$$K_{ij}=\frac{p_{ij}-p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}(1-p_{i\cdot})p_{\cdot j}(1-p_{\cdot j})}}; \qquad (3)$$

$p_{ij}=P(X=x_i,Y=y_j);\quad p_{i\cdot}=P(X=x_i);\quad p_{\cdot j}=P(Y=y_j);\quad i,j=1,2,...$

m_X,m_Y – математические ожидания, σ_X,σ_Y – средние квадратические отклонения случайных величин X,Y. Для рассматриваемого случая

$$m_X=np_1=2\cdot\frac{1}{4}=\frac{1}{2};\ m_Y=np_2=2\cdot\frac{1}{4}=\frac{1}{2};$$

$$\sigma_X=\sqrt{np_1(1-p_1)}=\sqrt{2\frac{1}{4}\frac{3}{4}}=\frac{\sqrt{6}}{4}=\sigma_Y.$$

Предварительно создадим таблицу значений ядер K_{ij}, (аналогичный пример в § 4.3.)

Таблица 2. Значения ассоциативных ядер K_{ij} к примеру 1.

$i\downarrow\ j\rightarrow$	0	1	2
0	– 0,269841	0,162650	0,227710
1	0,162650	– 0,066667	–
2	0,227710	–	–

$$com_a = \frac{8}{3}\left|0-\frac{1}{2}\right|\left|0-\frac{1}{2}\right|0,269841\cdot\frac{1}{4}+\frac{8}{3}\left|0-\frac{1}{2}\right|\left|1-\frac{1}{2}\right|0,162650\cdot\frac{1}{4}+$$

$$+\frac{8}{3}\left|0-\frac{1}{2}\right|\left|2-\frac{1}{2}\right|0,227710\cdot\frac{1}{16}+\frac{8}{3}\left|1-\frac{1}{2}\right|\left|0-\frac{1}{2}\right|0,162650\cdot\frac{1}{4}+$$

$$+\frac{8}{3}\left|1-\frac{1}{2}\right|\left|1-\frac{1}{2}\right|0,066667\cdot\frac{1}{8}+\frac{8}{3}\left|2-\frac{1}{2}\right|\left|0-\frac{1}{2}\right|0,227710\cdot\frac{1}{16}=$$

=(1/6)0,269841+(1/6)0,162650+(1/8)0,227710+(1/6)0,162650+(1/12)0,066667+(1/8)0,22771
0=

=0,044974+0,027108+0,028464+0,027108+0,005556+0,028464=0,161674$\approx 0,162$. Итак,

$$com_a = 0,162.$$

Пример 2. Вычисление коэффициента детерминации комби-конт для триномиального распределения.

Применим ту же формулу (2), но с другим ядром – контингенциальным ядром:

$$K_{ij} = \frac{p_{ij} - p_{i.}p_{.j}}{p_{ij}\left(1+2p_{ij}-2p_{i.}-2p_{.j}\right)+p_{i.}p_{.j}}; \qquad (4)$$

Для начала создадим таблицу значений ядра (4), (аналогичный пример 2 в § 6.4).

Таблица 3. Значения контингенциального ядра K_{ij} из формулы (4) примера 2.

$i\downarrow\ j\rightarrow$	0	1	2
0	0,117241	0,084746	0,280000
1	0,084746	0,058824	–
2	0,280000	–	–

$$com_c = \frac{8}{3}\left|0-\frac{1}{2}\right|\left|0-\frac{1}{2}\right|0,117241\cdot\frac{1}{4}+\frac{8}{3}\left|0-\frac{1}{2}\right|\left|1-\frac{1}{2}\right|0,084746\cdot\frac{1}{4}+$$

$$+\frac{8}{3}\left|0-\frac{1}{2}\right|\left|2-\frac{1}{2}\right|0,280000\cdot\frac{1}{16}+\frac{8}{3}\left|1-\frac{1}{2}\right|\left|0-\frac{1}{2}\right|0,084746\cdot\frac{1}{4}+$$

$$+\frac{8}{3}\left|1-\frac{1}{2}\right|\left|1-\frac{1}{2}\right|0,058824\cdot\frac{1}{8}+\frac{8}{3}\left|2-\frac{1}{2}\right|\left|0-\frac{1}{2}\right|0,280000\cdot\frac{1}{16}=$$

=(1/6)0,117241+(1/6)0,084746+(1/8)0,280000+(1/6)0,084746+(1/12)0,058824+(1/8)0,280000=

=0,019540+0,014124+0,035000+0,014124+0,004902+0,035000=0,122690 ≈ 0,123.

Итак,

$$com_c = 0,123.$$

Для сравнения приведем для этого распределения величины всех вычисленных ранее коэффициентов детерминации.

Модуль линейного коэффициента корреляции $|\rho| = 0,333$;

Ассоциативный коэффициент детерминации $as = 0,186$;

Контингенциальный коэффициент детерминации $co = 0,430$;

Комбинированный коэффициент детерминации комби-ас $com_a = 0,162$;

Комбинированный коэффициент детерминации комби-конт $com_c = 0,123$;

Предельный коэффициент детерминации $l = 0,114$.

Ранжируем их по величине:

$$l,\quad com_c,\quad com_a,\quad as,\quad |\rho|,\quad co. \qquad (5)$$

Эту ранжировку в дальнейшем сопоставим с ранжировками для других распределений.

§ 7.3 Случай равенства нулю линейного коэффициента корреляции. Сравнение с коэффициентами детерминации.

Пример 1.

Рассмотрим двумерное дискретное распределение, в котором случайная величина X имеет симметричное равномерное распределение, а $Y = |X|$.

Закон распределения X зададим формулой $P(X = k) = 1/4;\ k = \pm 1, \pm 2.$

Тогда таблица распределения двумерной случайной величины (X, Y) примет следующий вид.

Таблица 1. Распределение двумерной случайной величины $(X, |X|)$ в примере 1.

$X \downarrow Y \rightarrow$	1	2	$p_{i\cdot}$
-2	0	1/4	1/4
-1	1/4	0	1/4
1	1/4	0	1/4

2	0	1/4	1/4
$p_{\cdot j}$	1/2	1/2	1

Для этого распределения находим:

$$m_X = 0;\ D_X = M\left[X^2\right] = \frac{1}{4}\left[(-2)^2 + (-1)^2 + 1^2 + 2^2\right] = \frac{5}{2};\quad \sigma_X = \sqrt{D_X} = \sqrt{\frac{5}{2}};$$

$$m_Y = \frac{1}{2}(1+2) = \frac{3}{2};\ D_Y = M\left[Y^2\right] - m_Y^2 = \frac{1}{2}\left(1+2^2\right) - \left(\frac{3}{2}\right)^2 = \frac{5}{2} - \frac{9}{4} = \frac{1}{4};\ \sigma_Y = \frac{1}{2}.$$

Вычислим для этого распределения 6 ранее изученных коэффициентов связи.

1. Линейный коэффициент корреляции ρ.

$$\rho = \frac{1}{\sigma_X \sigma_Y}\sum_i\sum_j (x_i - m_X)(y_j - m_Y)\,p_{ij} =$$

$$= 2\sqrt{\frac{2}{5}}\left[(-2)\left(2-\frac{3}{2}\right) + (-1)\left(1-\frac{3}{2}\right) + 1\cdot\left(1-\frac{3}{2}\right) + 2\left(2-\frac{3}{2}\right)\right]\frac{1}{4} =$$

$$= \frac{1}{2}\sqrt{\frac{2}{5}}\left(-1+\frac{1}{2}-\frac{1}{2}+1\right) = 0.$$

2. Ассоциативный коэффициент детерминации as.

$$as = \sum_i\sum_j \frac{\left|p_{ij} - p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}(1-p_{i\cdot})\,p_{\cdot j}(1-p_{\cdot j})}}\,p_{ij} = \frac{1}{4}\frac{\frac{1}{4}-\frac{1}{4}\frac{1}{2}}{\sqrt{\frac{1}{4}\frac{3}{4}\frac{1}{2}\frac{1}{2}}}4 = \frac{\frac{1}{8}}{\frac{1}{4}\frac{1}{2}\sqrt{3}} = \frac{1}{\sqrt{3}} = 0{,}577.$$

3. Контингенциальный коэффициент детерминации co.

$$co = \sum_i\sum_j \frac{\left|p_{ij} - p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot}p_{\cdot j}}\,p_{ij} =$$

$$= \frac{1}{4}\left[\frac{\frac{1}{4}-\frac{1}{4}\frac{1}{2}}{\frac{1}{4}\left(1+2\frac{1}{4}-2\frac{1}{4}-2\frac{1}{2}\right)+\frac{1}{4}\frac{1}{2}}\right]4 = \frac{1/8}{1/8} = 1.$$

4. Предельный коэффициент детерминации l.

$$l=\sum_i\sum_j\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{p_{ij}+p_{i\cdot}p_{\cdot j}}p_{ij}=\frac{1}{4}\frac{\frac{1}{4}-\frac{1}{4}\frac{1}{2}}{\frac{1}{4}+\frac{1}{4}\frac{1}{2}}4=\frac{1/8}{3/8}=\frac{1}{3}=0{,}333.$$

5. Коэффициент детерминации комби-ас com_a.

Используем результаты вычисления в пункте 2.

$$com_a=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j\left|x_i-m_X\right|\left|y_j-m_Y\right|\frac{\left|p_{ij}-p_{i\cdot}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}p_{ij}=$$

$$=\frac{1}{4}\frac{1}{\sqrt{3}}\left(\left|-2\right|\left|2-\frac{3}{2}\right|+\left|-1\right|\left|1-\frac{3}{2}\right|+1\cdot\left|1-\frac{3}{2}\right|+2\left|2-\frac{3}{2}\right|\right)=\frac{3}{4\sqrt{3}}=\frac{\sqrt{3}}{4}=0{,}433.$$

6. Коэффициент детерминации комби-конт com_c.

Используем результаты вычисления в пунктах 3,5.

$$com_c=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j\left|x_i-m_X\right|\left|y_j-m_Y\right|\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}p_{ij}=$$

$$=3\cdot1\cdot\frac{1}{4}=\frac{3}{4}=0{,}75.$$

Ранжируем величины всех вычисленных коэффициентов:

$$\rho=0,\quad l=0{,}333,\quad com_a=0{,}433,\quad as=0{,}577,\quad com_c=0{,}750,\quad co=1.\ (1)$$

Видим, что лучше всего реагирует на связь случайных величин $Y=\left|X\right|$ контингенциальный коэффициент детерминации co.

Пример 2. Рассматриваем двумерное дискретное распределение, в котором случайная величина X имеет симметричное равномерное распределение примера 1, а $Y=X^2$.

$P\left(X=k\right)=1/4;\quad k=\pm1,\pm2.$ Составим таблицу распределения.

Таблица 2. Распределение двумерной случайной величины $\left(X,X^2\right)$ в примере 2.

$X\downarrow Y\rightarrow$	1	4	$p_{\cdot j}$
– 2	0	1/4	1/4
– 1	1/4	0	1/4
1	1/4	0	1/4
2	0	1/4	1/4
$p_{\cdot j}$	1/2	1/2	1

Распределение вероятностей в таблице 2 – такое же, как в таблице 1. Отсюда следует, что коэффициенты детерминации, не зависящие от значений случайных величин, – те же, что и в примере 1: $as = 0,577;\quad co = 1;\quad l = 0,333$. Линейный коэффициент корреляции $\rho = 0$ в силу симметричности распределения по X. Различия будут для коэффициентов комби. Для их вычисления нужно сначала найти моменты первых двух порядков.

$$m_X = 0;\ DX = M\left[X^2\right] = \frac{1}{4}\left[(-2)^2 + (-1)^2 + 1^2 + 2^2\right] = \frac{5}{2};\ \sigma_X = \sqrt{\frac{5}{2}};$$

$$m_Y = \frac{1}{2}(1+4) = \frac{5}{2};$$

$$DY = M\left[Y^2\right] - m_Y^2 = \frac{1}{2}\left(1^2 + 4^2\right) - \left(\frac{5}{2}\right)^2 = \frac{17}{2} - \frac{25}{4} = \frac{9}{4};\ \sigma_Y = \frac{3}{2}.$$

- Коэффициент детерминации комби-ас com_a

$$com_a = \frac{1}{\sigma_X \sigma_Y}\sum_i\sum_j |x_i - m_X|\,|y_j - m_Y|\,\frac{\left|p_{ij} - p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}(1-p_{i\cdot})\,p_{\cdot j}(1-p_{\cdot j})}}\,p_{ij} =$$

$$= \frac{2}{3}\sqrt{\frac{2}{5}}\,\frac{1}{\sqrt{3}}\,\frac{1}{4}\left(|-2|\left|4-\frac{5}{2}\right| + |-1|\left|1-\frac{5}{2}\right| + 1\cdot\left|1-\frac{5}{2}\right| + 2\cdot\left|4-\frac{5}{2}\right|\right) = \frac{\sqrt{2}}{6\sqrt{15}}\left(2\frac{3}{2}2 + \frac{3}{2}2\right) =$$

$$= \frac{3}{2}\sqrt{\frac{2}{15}} = 0,548.$$

- Коэффициент детерминации комби-конт com_c.

$$com_c = \frac{1}{\sigma_X \sigma_Y}\sum_i\sum_j |x_i - m_X|\,|y_j - m_Y|\,\frac{\left|p_{ij} - p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot}p_{\cdot j}}\,p_{ij} =$$

$$= \frac{2}{3}\sqrt{\frac{2}{5}}\cdot 1\cdot\frac{1}{4}9 = \frac{3}{2}\sqrt{\frac{2}{5}} = 0,949.$$

Ранжируем все 6 вычисленных коэффициентов:

$$\rho = 0;\quad l = 0,333;\quad com_a = 0,548;\quad as = 0,577;\quad com_c = 0,949;\quad co = 1 \quad .(2)$$

Сравним эту ранжировку (2) с ранжировкой (1) из примера 1. Несмотря на некоторое различие в цифрах, порядок следования коэффициентов сохранился.

Пример 3. Рассматриваем двумерное дискретное распределение, в котором случайная величина X имеет симметричное равномерное распределение примера 1, а $Y = X^3$.

$P(X = k) = 1/4; \quad k = \pm 1, \pm 2.$ Составим таблицу распределения.

Таблица 3. . Распределение двумерной случайной величины $\left(X, X^3\right)$ в примере 3.

$X \downarrow Y \rightarrow$	– 8	– 1	1	8	$p_{i\cdot}$
– 2	1/4	0	0	0	1/4
– 1	0	1/4	0	0	1/4
1	0	0	1/4		1/4
2	0	0	0	1/4	1/4
$p_{\cdot j}$	1/4	1/4	1/4	1/4	1

Для этого распределения вычислим 6 ранее введенных коэффициентов детерминации.

Сначала вычислим моменты первых двух порядков случайных величин.

$m_X = m_Y = 0$ по симметрии распределений относительно нуля.

$$D_X = M\left[X^2\right] = \frac{1}{4}(4+1+1+4) = \frac{5}{2}; \;\; \sigma_X = \sqrt{\frac{5}{2}};$$

$$D_Y = M\left[Y^2\right] = \frac{1}{4}(64+1+1+64) = \frac{65}{2}; \;\; \sigma_Y = \sqrt{\frac{65}{2}}.$$

1. Линейный коэффициент корреляции ρ.

$$\rho = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j x_i y_j p_{ij} = \sqrt{\frac{2}{5}} \sqrt{\frac{2}{65}} \left[(-2)(-8) + (-1)(-1) + 1 \cdot 1 + 2 \cdot 8\right] \frac{1}{4} =$$

$$= \frac{2}{5\sqrt{13}} 34 \frac{1}{4} = \frac{17}{5\sqrt{13}} = 0{,}943.$$

2. Ассоциативный коэффициент детерминации as.

$$as = \sum_i \sum_j \frac{\left|p_{ij} - p_{i\cdot} p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}} p_{ij} = 4 \frac{1}{4} \frac{\frac{1}{4} - \frac{1}{4}\frac{1}{4}}{\sqrt{\frac{1}{4}\frac{3}{4}\frac{1}{4}\frac{3}{4}}} = \frac{\frac{3}{16}}{\frac{1}{4}\frac{3}{4}} = 1.$$

3. Контингенциальный коэффициент детерминации co.

$$co = \sum_i \sum_j \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij} =$$

$$= 4 \frac{\frac{1}{4} - \frac{1}{4}\frac{1}{4}}{\frac{1}{4}\left(1 + 2\frac{1}{4} - 2\frac{1}{4} - 2\frac{1}{4}\right) + \frac{1}{4}\frac{1}{4}} \frac{1}{4} = \frac{\frac{3}{16}}{\frac{1}{4}\frac{1}{2} + \frac{1}{16}} = 1.$$

4. Предельный коэффициент детерминации l_1.

$$l = \sum_i \sum_j \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij} = 4\frac{1}{4}\frac{\frac{1}{4} - \frac{1}{16}}{\frac{1}{4} + \frac{1}{16}} = \frac{3}{5} = 0,6.$$

5. Коэффициент детерминации комби-ас com_a.

$$com_a = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left| x_i - m_X \right| \left| y_j - m_Y \right| \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}} p_{ij} =$$

$$= \frac{2}{5\sqrt{13}} \frac{1}{4} 34 \cdot 1 = \frac{17}{5\sqrt{13}} = 0,943.$$

6. Коэффициент детерминации комби-конт com_c.

$$com_c = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left| x_i - m_X \right| \left| y_j - m_Y \right| \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij} =$$

$$= \frac{17}{5\sqrt{13}} \cdot 1 = 0,943.$$

Ранжируем все 6 вычисленных коэффициентов.

$$l = 0,6; \quad \rho = com_a = com_c = 0,943; \quad as = co = 1.$$

Эта ранжировка существенно отличается от ранжировок в примерах 1 и 2. По- прежнему лучше других отражает зависимость коэффициент co.

Пример 4. Рассматриваем двумерное дискретное распределение, в котором случайная величина X имеет симметричное равномерное распределение примера 1, $Y = \left(-1\right)^k \left| X \right|$.

Здесь k – номер значения x_k случайной величины X: $x_1 = -2; x_2 = -1; x_3 = 1; x_4 = 2$.

Составим таблицу распределения.

Таблица 4. Распределение двумерной случайной величины $\left(X, \left(-1\right)^k \left| X \right| \right)$ в прим. 4.

$X \downarrow Y \rightarrow$	-2	-1	1	2	$p_{\cdot j}$
–2	1/4	0	0	0	1/4
–1	0	0	1/4	0	1/4
1	0	1/4	0	0	1/4
2	0	0	0	1/4	1/4
$p_{i\cdot}$	1/4	1/4	1/4	1/4	1

Распределение, представленное таблицей 4, не является симметричным и дает пример зависимости, не являющейся линейной, поэтому линейный коэффициент корреляции не равен нулю и не равен единице. Представляет интерес сравнить для этого распределения значения всех коэффициентов связи, рассмотренных в предыдущих примерах.

Сначала вычислим первые два момента распределения, нужные для дальнейших вычислений.

$$m_X = m_Y = 0; \quad D_X = D_Y = \frac{1}{4}\left[(-2)^2 + (-1)^2 + 1^2 + 2^2\right] = \frac{5}{2};$$

$$K_{XY} = \frac{1}{4}\left[(-2)(-2) + (-1)\cdot 1 + 2\cdot 2\right] = \frac{3}{2};$$

1. $\rho = \dfrac{K_{XY}}{\sigma_X \sigma_Y} = \dfrac{3/2}{5/2} = \dfrac{3}{5} = 0,6.$

2. $as = \dfrac{1}{4}4\dfrac{\frac{1}{4} - \frac{1}{4}\frac{1}{4}}{\sqrt{\frac{1}{4}\frac{3}{4}\frac{1}{4}\frac{3}{4}}} = \dfrac{3/16}{3/16} = 1.$

3. $co = \dfrac{1}{4}4\dfrac{\frac{1}{4} - \frac{1}{4}\frac{1}{4}}{\frac{1}{4}\left(1 + 2\frac{1}{4} - 2\frac{1}{4} - 2\frac{1}{4}\right) + \frac{1}{4}\frac{1}{4}} = \dfrac{3/16}{3/16} = 1.$

4. $l = \dfrac{1}{4}4\dfrac{3/16}{5/16} = \dfrac{3}{5} = 0,6.$

5. $com_a = com_c = \dfrac{1}{4}\dfrac{2}{5}\left(|-2||-2| + |-1|\cdot 1 + 1\cdot|-1| + 2\cdot 2\right) = \dfrac{10}{4}\dfrac{2}{5} = 1.$

Ранжируем все коэффициенты: $\rho = l = 0,6; \quad as = co = com_a = com_c = 1.$

Последние 4 коэффициента хорошо реагируют на рассматриваемую зависимость.

Пример 5. Рассматриваем дискретную двумерную случайную величину (X,Y), компоненты которой независимы и распределены симметрично и равномерно согласно следующей таблице распределения 5:

Таблица 5. Двумерное распределение с независимыми компонентами, распределенными симметрично и равномерно.

$X \downarrow Y \rightarrow$	–1	0	1	$p_{i\cdot}$
–2	1/12	1/12	1/12	1/4
–1	1/12	1/12	1/12	1/4
1	1/12	1/12	1/12	1/4
2	1/12	1/12	1/12	1/4
$p_{\cdot j}$	1/3	1/3	1/3	1

Для этого распределения все 6 рассматриваемых коэффициентов связи равны нулю: $\rho = as = co = l = com_a = com_c = 0.$ Чтобы разрушить независимость, варьируем вероятности p_{ij}, оставляя распределение симметричным и равномерным. Для этого прибавляем или отнимаем от вероятностей первых двух столбцов малое число 1/24=0,042. Получаем новую таблицу распределения 6:

Таблица 6. Двумерное распределение со слабо зависимыми компонентами, распределенными симметрично и равномерно.

$X \downarrow Y \rightarrow$	1	0	1	$p_{i\cdot}$
2	1/24	3/24	1/12	1/4
1	3/24	1/24	1/12	1/4
1	3/24	1/24	1/12	1/4
2	1/24	3/24	1/12	1/4
$p_{\cdot j}$	1/3	1/3	1/3	1

Для распределения, представленного таблицей 6, вычисляем все 6 ранее рассмотренных коэффициентов связи.

Предварительно вычисляем моменты первых двух порядков, нужные для дальнейших вычислений.

$m_X = m_Y = 0$ по симметрии распределения.

$$D_X = M\left[X^2\right] = \frac{1}{4}(4+1+1+4) = \frac{5}{2}; \quad \sigma_X = \sqrt{\frac{5}{2}};$$

$$D_Y = M\left[Y^2\right] = \frac{1}{3}(1+0+1) = \frac{2}{3}; \quad \sigma_Y = \sqrt{\frac{2}{3}};$$

$K_{XY} = 0$ по симметрии распределения.

1. Линейный коэффициент корреляции $\rho = \dfrac{K_{XY}}{\sigma_X \sigma_Y} = 0.$

2. Ассоциативный коэффициент детерминации

$$as = \frac{1}{24}\left(\frac{\left|\frac{1}{24} - \frac{1}{4}\frac{1}{3}\right|}{\sqrt{\frac{1}{4}\frac{3}{4}\frac{1}{3}\frac{2}{3}}} 4 + \frac{\left|\frac{3}{24} - \frac{1}{4}\frac{1}{3}\right|}{\sqrt{\frac{1}{4}\frac{3}{4}\frac{1}{3}\frac{2}{3}}} 4\cdot 3 + \frac{\left|\frac{1}{12} - \frac{1}{4}\frac{1}{3}\right|}{\sqrt{\frac{1}{4}\frac{3}{4}\frac{1}{3}\frac{2}{3}}} 4\cdot 2 \right) =$$

$$\frac{4}{24}\left(\frac{1/24}{\sqrt{6}/12} + \frac{3/24}{\sqrt{6}/12} \right) =$$

$$= \frac{1}{3\sqrt{6}} = 0{,}136.$$

3. Контингенциальный коэффициент детерминации

$$co = \frac{1}{24} 4 \left[\frac{\left|\frac{1}{24} - \frac{1}{4}\frac{1}{3}\right|}{\frac{1}{24}\left(1 + 2\frac{1}{24} - 2\frac{1}{4} - 2\frac{1}{3}\right) + \frac{1}{12}} + \frac{\left|\frac{3}{24} - \frac{1}{4}\frac{1}{3}\right| 3}{\frac{3}{24}\left(1 + 2\frac{3}{24} - 2\frac{1}{4} - 2\frac{1}{3}\right) + \frac{1}{12}} \right] =$$

$$= \frac{1}{6}\left(\frac{12}{23} + \frac{36}{27} \right) = \frac{192}{23\cdot 27} = 0{,}309.$$

4. Предельный коэффициент детерминации

$$l = \frac{1}{24} 4 \left(\frac{\left|\frac{1}{24} - \frac{1}{12}\right|}{\frac{1}{24} + \frac{1}{12}} + \frac{\left|\frac{3}{24} - \frac{1}{12}\right|}{\frac{3}{24} + \frac{1}{12}} 3 \right) = \frac{1}{6}\left(\frac{1}{3} + \frac{3}{5} \right) = \frac{7}{45} = 0{,}156.$$

5. Комбинированный коэффициент детерминации комби-ас.

Используем результаты вычислений пункта 2.

$$com_a = \sqrt{\frac{3}{2}\frac{2}{5}}\frac{1}{24}\frac{1/24}{\sqrt{6}/12}\left(2\cdot 1+1\cdot 1\cdot 3+1\cdot 1\cdot 3+2\cdot 1\right)=\sqrt{\frac{3}{5}}\frac{1}{24}\frac{1}{2\sqrt{6}}10=\frac{5}{24\sqrt{10}}.$$

$com_a = 0,066.$

6. Комбинированный коэффициент детерминации комби-конт.

Используем результаты вычислений пункта 3.

$$com_c = \sqrt{\frac{3}{2}\frac{2}{5}}\frac{1}{24}\left(2\cdot 1\frac{12}{23}+1\cdot 1\frac{36}{27}+1\cdot 1\frac{36}{27}+2\cdot 1\frac{12}{23}\right)=\sqrt{\frac{3}{5}}\frac{1}{2}\left(\frac{4}{23}+\frac{2}{9}\right).$$

$$com_c = \sqrt{\frac{3}{5}}\frac{41}{207}=0,153.$$

Ранжируем все вычисленные коэффициенты связи от меньшего к большему:

$\rho = 0;\quad com_a = 0,066;\quad as = 0,136;\quad com_c = 0,153;\quad l = 0,156;\quad co = 0,309.$

Линейный коэффициент корреляции никак не отреагировал на введение слабой зависимости, так как распределение осталось симметричным. Все остальные коэффициенты, как и полагается, – отреагировали; из них наиболее существенно отреагировал контингенциальный коэффициент детерминации «Конт».

Глава 8. Дефектологический коэффициент детерминации

Из 8 новых коэффициентов детерминации дефектологический коэффициент является наиболее простым по структуре.

§ 8.1. Дефект независимости событий и его свойства.

Известно необходимое и достаточное условие независимости двух событий A, B:

$$P(AB) - P(A)P(B) = 0. \qquad (1)$$

Это условие может быть положено в основу конструкции коэффициента

$$\delta_{AB} = \left|P(AB) - P(A)P(B)\right|, \qquad (2)$$

который целесообразно назвать дефектом независимости событий A и B (коэффициентом дефектности, дефектом). Этот коэффициент как множитель под знаками суммы или интеграла входит во все коэффициенты детерминации и определяет их важнейшее

свойство: указывать на независимость случайных величин, если коэффициент равен нулю. Дефект δ_{AB} ограничен снизу нулем и сверху единицей: $0 \le \delta_{AB} \le P(AB) \le 1.$ Указанная верхняя граница – неточная, но указывает, что можно найти точную границу и поэтому, нормировав δ_{AB} числовым коэффициентом, можно применять нормированный дефект $\overline{\delta}_{AB}$ как меру зависимости событий. Числовой нормирующий коэффициент подберем так, что нормированный дефект будет заключен между точными границами 0 и 1:

$$0 \le \overline{\delta}_{AB} \le 1.$$

(3)

Точную верхнюю границу дефекта найдем, исследовав его свойства.

В дальнейшем для краткости дефект δ_{AB} будет иногда обозначаться просто δ.

Свойства дефекта.

1. Дефект δ_{AB} событий A, B равен нулю тогда и только тогда, когда события A, B независимы.

Действительно, равенство нулю δ_{AB} происходит тогда и только тогда, когда вполняется условие независимости событий (1).

2. Наибольшее значение дефекта $\delta = \delta_{AB}$ равно 1/4 .

Оно достигается только в двух случаях:

2.1. $A = B = AB; \quad P(A) = P(B) = P(AB) = 1/2.$

2.2. A и B несовместны; $P(AB) = 0; \quad P(A) = P(B) = 1/2.$

Доказательство.

1º. Сначала применим интуитивные, эвристические соображения о том, что полная-зависимость, то есть, когда $A = B$, должна отвечать наибольшему значению δ.

Если $A = B,$ то $AB = A.$ Отсюда следует, что $P(A) = P(B) = P(AB);$ $\delta = \left| P(A) - P^2(A) \right|.$ Положим $P(A) = x.$ Тогда $\delta = \left| x - x^2 \right| = x - x^2 \ge 0,$ так как $0 \le x \le 1.$ Ищем наибольшее и наименьшее значения этой функции на промежутке $[0;1].$ $\delta(0) = \delta(1) = 0$ – наименьшее значение.

$\delta' = 1 - 2x = 0 \quad \Rightarrow \quad x = 1/2; \ \delta\left(\frac{1}{2}\right) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$ – наибольшее значение.

2º. Применим теперь эвристические соображения другого рода.

Модуль разности двух чисел принимает наибольшее значение, когда одно слагаемое равно нулю, а второе имеет наибольшее возможное значение.

Если $P(A)=0$ то $A=\varnothing$ – невозможное событие, но тогда $P(AB)=P(A\varnothing)=P(\varnothing)=0$. Поэтому $\delta_{AB}=0.$ Это – наименьшее значение.

Если $P(AB)=0$, то события A,B несовместны и $\delta_{AB}=\delta=P(A)P(B).$

Положим $P(A)=x,\quad P(B)=y.$ Тогда $\delta=xy$.

Переменные x,y удовлетворяют неравенствам:

$$0\le x\le 1;\ 0\le y\le 1;\ x+y\le 1 \qquad (4)$$

Последнее неравенство вытекает из формулы теории вероятностей для вероятности суммы двух любых событий:
$P(A+B)=P(A)+P(B)-P(AB)=P(A)+P(B)\le 1,$
так как $P(A+B)\le 1$. Неравенства (4) на плоскости xOy определяют замкнутый треугольник, в котором ищем наибольшее значение функции $\delta(x,y)=\delta$.

Находим стационарную точку: $\delta'_x=y=0;\quad \delta'_y=x=0.$ Стационарная точка $O(0;0)$ находится на границе и в ней $\delta=0$. Это – наименьшее значение $\delta.$

Исследуем границу. На прямых $x=0$ и $y=0$ функция δ принимает наименьшее значение. Исследуем гипотенузу $x+y=1$. Имеем $y=1-x$;

$\delta(x,1-x)=u(x)=x(1-x)=x-x^2;\ u'(x)=1-2x=0.$ Отсюда

$x=1/2;\quad y=1-x=1/2;\quad \delta=xy=\frac{1}{2}\frac{1}{2}=\frac{1}{4}.$ Итак, наибольшее значение $\delta=\frac{1}{4}.$

3°.Приведенные в пунктах 1°, 2° эвристические соображения дают возможность высказать только предположения о том, что наибольшее значение δ равно 1/4, приведем теперь полное доказательство. Положим

$$\delta=|z-xy| \qquad (5)$$

где $x=P(A);\quad y=P(B);\quad z=P(AB).$

Эти три переменные связаны соотношениями

1) $x\ge 0;\quad y\ge 0;$

2) $z\ge 0;$

3) $x\le 1;\quad y\le 1;\quad z\le 1;$ (6)

4) $z\le x;\quad z\le y$

5) $x+y-z\le 1.$

Последнее неравенство следует из формулы
$P(A+B)=P(A)+P(B)-P(AB)\le 1,$

так как $P(A+B)\le 1$.

Неравенства 4) следуют из формул $AB\subset A;\quad P(AB)\le P(A).$

Эти неравенства (6) в системе координат $Oxyz$ определяют треугольную пирамиду OABC, лежащую внутри единичного куба: $0\le x\le 1;\quad 0\le y\le 1;\quad 0\le z\le 1$. Рис. 1.

Эта пирамида имеет 4 вершины $O(0;0;0), A(1;0;0), B(0;1;0), C(1;1;1);$

4 грани:

1) $OAB:\ z=0;$ 2) $OAC:\ z=y;$ 3) $OBC:\ z=x;$ 4) $ABC:\ x+y-z=1;$

6 ребер:

1) $AB:\ \begin{cases} z=0 \\ x+y=1 \end{cases};$ 2) $OA:\ \begin{cases} y=0 \\ z=0 \end{cases};$ 3) $OB:\ \begin{cases} x=0 \\ z=0 \end{cases};$ 4) $OC:\ \begin{cases} z=x \\ z=y \end{cases};$

5) $AC:\ \begin{cases} x=1 \\ z=y \end{cases};$ 6) $BC:\ \begin{cases} y=1 \\ z=x \end{cases};$

Пирамида

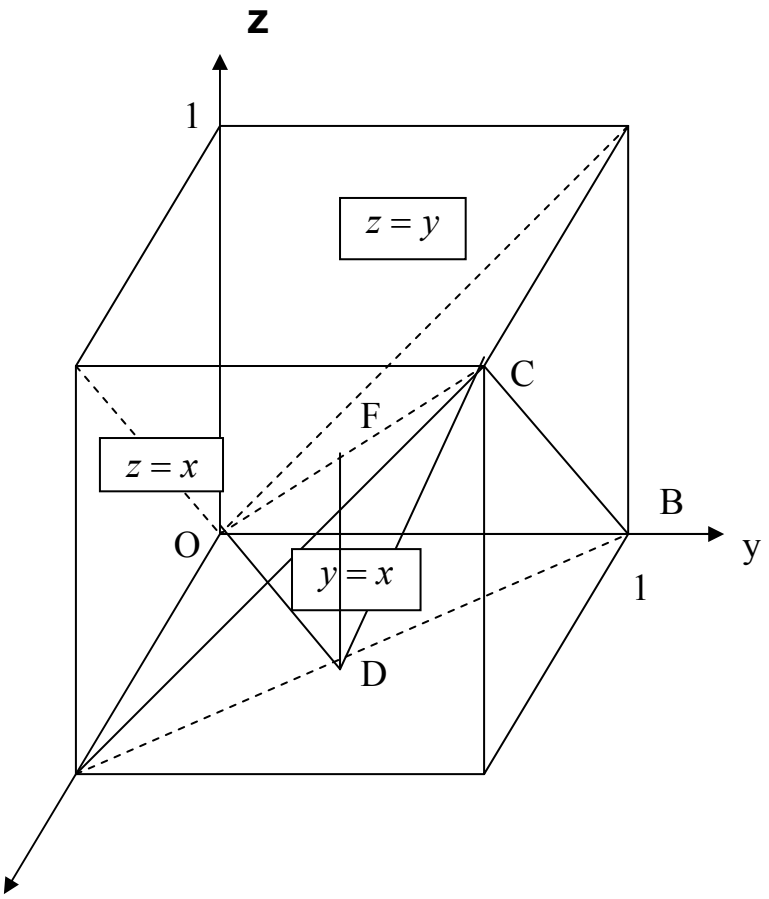


Рис. 1. Пирамида OABC, лежащая внутри единичного куба.

Вместо отыскания наибольшего значения δ можно искать наибольшее значение двух аналитических функций $\delta_1 = z - xy$ и $\delta_2 = xy - z$, отличающихся только знаком. У этих функций нет стационарных точек, так как

$$\frac{\partial \delta_{1,2}}{\partial z} = \pm 1 \neq 0; \quad \frac{\partial \delta_{1,2}}{\partial x} = \pm x; \quad \frac{\partial \delta_{1,2}}{\partial y} = \pm y.$$

Поэтому наибольшее значение $\delta_1, \delta_2, \delta$ нужно искать на границе пирамиды OABC.

Последовательно перебираем всю границу пирамиды.

1). Грань OAB: $z = 0$. Тогда $\delta = |-xy| = xy$. Так как $x + y - z \le 1$, то $x + y \le 1$. Итак, ищем наибольшее значение функции $\delta = xy$ в треугольнике, ограниченном прямыми

$x = 0; y = 0; x + y = 1$ плоскости xOy. Эта задача решена в пункте 2°. Найдено наибольшее значение $\delta = 1/4$ при $x = y = 1/2$. При этом $z = 0$. Это значение δ принимается в точке D(1/2;1/2;0) плоскости xOy (Рис. 1).

2)Грань OAC: $z = y$. Тогда $\delta = |z - xy| = |y - xy| = y(1 - x) = y - xy = v(x, y)$. Ищем наибольшее значение этой функции в треугольнике OAD, ограниченном линиями $x = 0; \quad y = 0; \quad y = x$ плоскости xOy. В этот треугольник проектируется грань OAC.

Стационарная точка: $v'_x = -y = 0; y = 0; \quad v'_y = 1 - x = 0; x = 1;$ A(1;0). В этой точке $v = 0$. Это – наименьшее значение δ. Далее исследуем границу треугольника OAD.

Сторона OA: $y = 0$. Тогда $\delta = v(x, 0) = 0$. Это наименьшее значение δ.

Сторона OD: $y = x$.Тогда $v = x - x^2; \quad 0 \le x \le \frac{1}{2}. \quad v(0) = 0; \quad v\left(\frac{1}{2}\right) = \frac{1}{4}.$ Эти значения были получены ранее. Критическая точка: $v' = 1 - 2x = 0 \Rightarrow x = \frac{1}{2};$ Это значение x отвечает точке D, исследованной ранее. На грани OAC, которую мы ис-

следуем, точке D соответствует точка F(1/2;1/2;1/2), так как для нее $x = y = z = 1/2$.

Сторона AD: $y = 1 - x;$ Тогда

$v = (x, 1-x) = 1 - x - x(1-x) = 1 - 2x + x^2 = (1-x)^2 = w(x);$

$\frac{1}{2} \le x \le 1;$ Эта функция монотонно убывает от $w\left(\frac{1}{2}\right) = \frac{1}{4}$ до $w(1) = 0.$ Эти значения функции w, а потому и δ,были получены раньше. Значение $w\left(\frac{1}{2}\right) = \frac{1}{4}$ на грани OAC соответствует точке F.

3)Грань OBC исследуется аналогично, однако, вследствие симметрии вхождения переменных x, y в выражение для функции δ, можно утверждать, что δ принимает те же значения, что и на грани OAC.

4)Грань ABC: $z = x + y - 1$. Тогда $\delta = |z - xy|; \quad \delta_{1,2} = \pm(x + y - 1 - xy);$

$0 \le x \le 1; \quad 0 \le y \le 1.$

$\frac{\partial \delta_{1,2}}{\partial x} = \pm(1-y) = 0 \quad \Rightarrow y = 1; \qquad \frac{\partial \delta_{1,2}}{\partial y} = \pm(1-x) = 0 \qquad \Rightarrow x = 1;$

$z = x + y - 1 = 1.$

Стационарной точкой функций $\delta_{1,2}$ является точка C(1;1;1). Эта точка является также вершиной пирамиды OABC. В точке C дефект δ равен нулю. Исследуем границу треугольника ABE, в который проектируется грань ABC для функции $\delta = |x + y - 1 - xy|.$

Сторона AB: $y = 1 - x.$ Тогда

$\delta = |x + y - 1 - xy| = |x + 1 - x - 1 - x(1-x)| = x(1-x).$ Эта функция исследовалась в пункте 2. Ее наибольшее значение принимается при $x = \frac{1}{2}.$ Тогда

$y = 1 - x = \frac{1}{2};$

$z = x + y - 1 = 0.$

Получили знакомую точку D(1/2;1/2;0).

Сторона AE: $x = 1.$ Тогда $\delta = |1 + y - 1 - y| = 0.$

Сторона BE: $y = 1.$ По симметрии роли переменных x, y имеем $\delta = 0.$

Итак, исследована внутренность и граница пирамиды OABC. Доказали, что наибольшим значением дефекта δ является 1/4. Это значение принимается только в точках
D(1/2;1/2;0) и F(1/2;1/2;1/2).

3. Если дефект $\delta_{AB} = \frac{1}{4}$, то либо $A = B$, либо события A, B несовместны.

Доказательство. Основываемся на том, что функция трех переменных $\delta = |z - xy|$, построенная в пункте 2, определена в пирамиде OABC, рис. 1, и в этой пирамиде значение $\delta = 1/4$ принимает только в двух точках D(1/2;1/2;0) и F(1/2;1/2;1/2).

Точка D соответствует случаю, когда события A, B равны.

Точка F соответствует случаю, когда события A, B несовместны.

Исследование свойств дефекта δ завершено.

Знание свойств дефекта δ позволяет построить нормированный дефект

$$\overline{\delta} = 4\delta \tag{7}$$

и утверждать, что

$$0 \le \overline{\delta} \le 1. \tag{8}$$

Границы точные, то есть достигаются.

Из сравнения дефекта δ_{AB} с коэффициентом корреляции между событиями

$$\rho_{AB} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$$

получаем формулу связи между ними

$$|\rho_{AB}| = \frac{|\delta_{AB}|}{\sqrt{P(A)\left(1-P(A)\right)P(B)\left(1-P(B)\right)}}. \tag{9}$$

Отсюда

$$|\delta_{AB}| = |\rho_{AB}|\sqrt{P(A)\left(1-P(A)\right)P(B)\left(1-P(B)\right)}\ . \tag{10}$$

Основываясь на этой формуле (10), дадим другое более короткое доказательство неравенства

$$|\delta_{AB}| \le \frac{1}{4}. \tag{11}$$

Доказательство свойств дефекта на основе формулы (10).

1. Оценим сверху оба множителя в формуле (10). Первый множитель есть модуль линейного корреляции между индикаторами событий A, B, а потому $0 \le |\rho_{AB}| \le 1$,

(см. § 1.2).

Оценим сверху теперь второй множитель в формуле (10). Пусть $P(A)=x; P(B)=y$. Тогда

$P(A)(1-P(A))=x(1-x)\le \frac{1}{2}\frac{1}{2}=\frac{1}{4}$. (Вершина параболы лежит посредине между точками пересечения ею оси абсцисс, то есть в точке $x=\frac{1}{2}$). По той же причине $P(B)(1-P(B))\le \frac{1}{4}$. Тогда

$$\sqrt{P(A)(1-P(A))P(B)(1-P(B))}\le \frac{1}{4}. \qquad (12)$$

Далее получаем $\delta_{AB}\le 1\cdot\frac{1}{4}=\frac{1}{4}$.

2. Дефект δ_{AB} обращается в ноль тогда и только тогда, когда события A,B независимы.

Это свойство следует из того, что в ноль обращается коэффициент корреляции ρ_{AB} только в случае независимости событий A,B (см. §1.2).

3. Если $\delta_{AB}=\frac{1}{4}$, то события A,B или равны, при этом $P(A)=\frac{1}{2}$, или несовместны, при этом $P(A)=P(B)=\frac{1}{2}$.

Действительно, если $\delta_{AB}=\frac{1}{4}$, то на основании формул (10), (11), (12) заключаем, что $|\rho_{AB}|=1$. В этом случае или $A=B$, или $B=\overline{A}$ (см. §1.2). Далее, если $\delta_{AB}=\frac{1}{4}$, то $\sqrt{P(A)(1-P(A))P(B)(1-P(B))}=\frac{1}{4}$. Как мы выяснили в пункте 2 в этом случае $P(A)=P(B)=\frac{1}{2}$. Этим завершается доказательство свойства 3.

Пример 1. Рассмотрим дискретное двумерное триномиальное распределение, определяемое формулой

$$p_{ij}=\frac{n!}{i!\,j!(n-i-j)!}p_1^i p_2^j(1-p_1-p_2)^{n-i-j};$$

$$i,j=0.1,...,n;\quad 0<p_1<1;\quad 0<p_2<1;\quad p_1+p_2<1;\quad i+j\le n.$$

Рассмотрим случай $n=2;\quad p_1=p_2=1/4.$ Построим таблицу распределения (табл.1).

Таблица 1 триномиального распределения при $n=2.$

$X \setminus Y \rightarrow$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00}=1/4$	$p_{01}=1/4$	$p_{02}=1/16$	$p_{0\cdot}=9/16$
1	$p_{10}=1/4$	$p_{11}=1/8$	$p_{12}=0$	$p_{1\cdot}=3/8$
2	$p_{20}=1/16$	$p_{21}=0$	$p_{22}=0$	$p_{2\cdot}=1/16$
$p_{\cdot j}$	$p_{\cdot 0}=9/16$	$p_{\cdot 1}=3/8$	$p_{\cdot 2}=1/16$	1

Линейный коэффициент корреляции вычисляется по формуле

$$\rho=-\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}}=-\frac{1}{3}$$;[3,с.142].

Введем события

$C_{ij}=(X=i,Y=j);\quad A_i=(X=i);\quad B_j=(Y=j);\ (i,j=0,1,2).$

Их вероятности: $P(C_{ij})=p_{ij};\quad P(A_i)=p_{i\cdot};\quad P(B_j)=p_{\cdot j}.$

Вычислим нормированные дефекты для всех пар событий A_i,B_j и сравним их с модулем линейного коэффициента корреляции. Пусть $\overline{\delta}_{ij}$ – нормированный дефект пары событий

$A_i,B_j.$

$$\overline{\delta}_{ij}=4\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right| \qquad (13)$$

Тогда на основании таблицы 1 получаем

$$\overline{\delta}_{oo}=4\left|\frac{1}{4}-\frac{9}{16}\frac{9}{16}\right|=\frac{17}{64};\ \overline{\delta}_{01}=4\left|\frac{1}{4}-\frac{9}{16}\frac{3}{8}\right|=\frac{10}{64};\ \overline{\delta}_{02}=4\left|\frac{1}{16}-\frac{9}{16}\frac{1}{16}\right|=\frac{7}{64};$$

$$\overline{\delta}_{10}=4\left|\frac{1}{4}-\frac{3}{8}\frac{9}{16}\right|=\frac{10}{64};\ \overline{\delta}_{11}=4\left|\frac{1}{8}-\frac{3}{8}\frac{3}{8}\right|=\frac{4}{64};\ \overline{\delta}_{12}=4\left|0-\frac{3}{8}\frac{1}{16}\right|=\frac{6}{64};$$

$$\overline{\delta}_{20}=4\left|\frac{1}{16}-\frac{1}{16}\frac{9}{16}\right|=\frac{7}{64};\ \overline{\delta}_{21}=4\left|0-\frac{1}{16}\frac{3}{8}\right|=\frac{6}{64};\ \overline{\delta}_{22}=4\left|0-\frac{1}{16}\frac{1}{16}\right|=\frac{1}{64}.$$

Составим вариационный ряд из этих дефектов для сравнения:

$\overline{\delta}_{22}=1/64=0{,}016;\ \overline{\delta}_{11}=4/64=0{,}063;\ \overline{\delta}_{12}=6/64=0{,}094;\ \overline{\delta}_{21}=\ 6/64=0{,}094;$

$\overline{\delta}_{02}=7/64=0{,}109;$ (14)

$\overline{\delta}_{20}=7/64=0{,}109;\ \overline{\delta}_{01}=10/64=0{,}156;\ \overline{\delta}_{10}=10/64=0{,}156;\ \overline{\delta}_{00}=17/64=0{,}266.$

Медианный дефект (пятый слева) равен 7/64=0,109.

Все значения нормированного дефекта меньше модуля линейного коэффициента корреляции. Интересно сравнить значения дефекта (14) с соответствующими значениями модуля коэффициента корреляции, вычисленными по формуле (9).

$$|\rho_{00}|=\frac{\delta_{00}}{\sqrt{p_{0\cdot}(1-p_{0\cdot})\,p_{\cdot 0}(1-p_{\cdot 0})}}=\frac{17/256}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{9}{16}\frac{7}{16}}}=\frac{17\cdot}{7\cdot 9}=\frac{17}{63}=0{,}270;$$

$$|\rho_{01}|=\frac{\delta_{01}}{\sqrt{p_{0\cdot}(1-p_{0\cdot})\,p_{\cdot 1}(1-p_{\cdot 1})}}=\frac{10/256}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{3}{8}\frac{5}{8}}}=\frac{5}{\sqrt{945}}=0{,}163;$$

$$|\rho_{02}|=\frac{\delta_{02}}{\sqrt{p_{0\cdot}(1-p_{0\cdot})\,p_{\cdot 2}(1-p_{\cdot 2})}}=\frac{7/256}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{1}{16}\frac{15}{16}}}=\frac{7}{\sqrt{945}}=0{,}228;$$

$$|\rho_{10}|=\frac{\delta_{10}}{\sqrt{p_{1\cdot}(1-p_{1\cdot})\,p_{\cdot 0}(1-p_{\cdot 0})}}=\frac{10/256}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{3}{8}\frac{5}{8}}}=\frac{5}{\sqrt{945}}=0{,}163;$$

$$|\rho_{11}|=\frac{\delta_{11}}{\sqrt{p_{1\cdot}(1-p_{1\cdot})\,p_{\cdot 1}(1-p_{\cdot 1})}}=\frac{4/256}{\sqrt{\frac{3}{8}\frac{5}{8}\frac{3}{8}\frac{5}{8}}}=\frac{1}{15}=0{,}067;$$

$$|\rho_{12}|=\frac{\delta_{12}}{\sqrt{p_{1\cdot}(1-p_{1\cdot})\,p_{\cdot 2}(1-p_{\cdot 2})}}=\frac{6/256}{\sqrt{\frac{3}{8}\frac{5}{8}\frac{1}{16}\frac{15}{16}}}=\frac{3}{15}=0{,}2;$$

$$|\rho_{20}|=\frac{\delta_{20}}{\sqrt{p_{2\cdot}(1-p_{2\cdot})\,p_{\cdot 0}(1-p_{\cdot 0})}}=\frac{7/256}{\sqrt{\frac{9}{16}\frac{7}{16}\frac{1}{16}\frac{15}{16}}}=\frac{7}{\sqrt{945}}=0{,}228;$$

$$|\rho_{21}|=\frac{\delta_{21}}{\sqrt{p_{2\cdot}(1-p_{2\cdot})\,p_{\cdot 1}(1-p_{\cdot 1})}}=\frac{6/256}{\sqrt{\frac{1}{16}\frac{15}{16}\frac{3}{8}\frac{5}{8}}}=\frac{3}{15}=0{,}2;$$

$$|\rho_{22}| = \frac{\delta_{22}}{\sqrt{p_{2\cdot}(1-p_{2\cdot})\,p_{\cdot 2}(1-p_{\cdot 2})}} = \frac{1/256}{\sqrt{\frac{1}{16}\frac{15}{16}\frac{1}{16}\frac{15}{16}}} = \frac{1}{15} = 0{,}067;$$

Составим вариационный ряд из полученных значений:

$|\rho_{11}| = 0{,}067;\ |\rho_{22}| = 0{,}067;\ |\rho_{01}| = 0{,}163;\ |\rho_{10}| = 0{,}163;\ |\rho_{12}| = 0{,}2;\ |\rho_{21}| = 0{,}2;$

$|\rho_{02}| = 0{,}228;\ |\rho_{20}| = 0{,}228;\ |\rho_{00}| = 0{,}270.$

Результаты вычислений дефекта и модуля коэффициента корреляции сведем в табл. 2.

Таблица 2. Значения дефекта и модуля коэффициента корреляции событий, определяемых триномиальным распределением при n=2.

$\bar{\delta}_{00}$	$\bar{\delta}_{01}$	$\bar{\delta}_{02}$	$\bar{\delta}_{10}$	$\bar{\delta}_{11}$	$\bar{\delta}_{12}$	$\bar{\delta}_{20}$	$\bar{\delta}_{21}$	$\bar{\delta}_{22}$	медиана
0,266	0,156	0,109	0,156	0,063	0,094	0,109	0,094	0,016	0,109
$\lvert\rho_{00}\rvert$	$\lvert\rho_{01}\rvert$	$\lvert\rho_{02}\rvert$	$\lvert\rho_{10}\rvert$	$\lvert\rho_{11}\rvert$	$\lvert\rho_{12}\rvert$	$\lvert\rho_{20}\rvert$	$\lvert\rho_{21}\rvert$	$\lvert\rho_{22}\rvert$	медиана
0,270	0,163	0,228	0,163	0,067	0,200	0,228	0,200	0,067	0,200

Из этой таблицы следует, что значения дефекта примерно в 2 раза меньше, чем значения модуля коэффициента корреляции.

Пример 2. Вычисляем дефект событий, определяемых равномерным дискретным распределением в случае, когда $Y = X$, при этом

$$p_{ij} = P\left(X = x_i, X = x_j\right) = \begin{cases} 0; & i \neq j \\ P\left(X = x_i\right) = \dfrac{1}{n}; & i = j \end{cases}; \qquad (15)$$

$$p_{i\cdot} = p_{\cdot j} = \frac{1}{n};\ \ i, j = 1, 2, ..., n;$$

Формулы (15) определяют следующую таблицу 3 распределения. Приведем ее .

Таблица 3. Равномерное дискретное распределение при условии, что $Y = X$.

$X \downarrow Y \rightarrow$	1	2	…	n – 1	n	$p_{i\cdot}$
1	1/n	0	…	0	0	1/n
2	0	1/n	…	0	0	1/n
…	…	…	…	…	…	…
n - 1	0	0	…	1/n	0	1/n
n	0	0	…	0	1/n	1/n
$p_{\cdot j}$	1/n	1/n	…	1/n	1/n	1

$$\overline{\delta}_{ij} = 4\left|p_{ij} - p_{i\cdot}^2\right|;\ \overline{\delta}_{ij} = 4\left(\frac{1}{n} - \frac{1}{n^2}\right) = 4\frac{n-1}{n^2} \text{ при } i = j;$$

$$\overline{\delta}_{ij} = 4\left|0 - \frac{1}{n^2}\right| = \frac{4}{n^2} \text{ при } i \neq j.$$

При $n = 2$ получаем $\overline{\delta}_{11} = \overline{\delta}_{22} = 4\frac{n-1}{n^2} = 1;\quad \overline{\delta}_{12} = \overline{\delta}_{21} = \frac{4}{n^2} = 1;\ \overline{\delta}_{ij} \xrightarrow[n\to\infty]{} 0$.

Для сравнения вычислим модуль коэффициента корреляции между этими событиями.

$$\left|\rho_{ij}\right| = \frac{\delta_{ij}}{\sqrt{\frac{1}{n}\frac{n-1}{n}\frac{1}{n}\frac{n-1}{n}}} = \delta_{ij}\frac{n^2}{n-1};\ \left|\rho_{ij}\right| = \frac{n-1}{n^2}\frac{n^2}{n-1} = 1 \text{ при } i = j;$$

$$\left|\rho_{ij}\right| = \frac{1}{n^2}\frac{n^2}{n-1} = \frac{1}{n-1} \xrightarrow[n\to\infty]{} 0 \text{ при } i \neq j.$$

Между $\overline{\delta}_{ij}$ и $\left|\rho_{ij}\right|$ имеются определенные различия в характеристике величины связи между событиями при рассматриваемом распределении.

§ 8.2. Коэффициент детерминации с дефектологическим ядром для дискретных распределений

Коэффициент детерминации с дефектологическим ядром (дефектологический коэффициент детерминации) строится на основе дефекта событий, который изучен в предыдущем параграфе. В этом параграфе рассматривается случай дискретного распределения, а в следующем будет рассмотрен общий случай.

Коэффициент детерминации с дефектологическим ядром для двумерного дискретного распределения строится по формуле

$$def = 6\sum_{i=1}^{m}\sum_{j=1}^{n}\left|F_{ij} - F_{i\cdot}F_{\cdot j}\right|p_{ij} \qquad (1)$$

Здесь

$$p_{ij} = P\left(X = x_i, Y = y_j\right);\ F_{ij} = F_{XY}\left(x_i, y_j\right);\ F_{i\cdot} = F_X\left(x_i\right);\ F_{\cdot j} = F_Y\left(y_j\right); \qquad (2)$$

$F_{XY}\left(x,y\right) = P\left(X < x, Y < y\right);\ F_X\left(x\right) = P\left(X < x\right);\ F_Y\left(y\right) = P\left(Y < y\right)$ – функции распределения.

Значения функций распределения (2) будем вычислять по формулам

$$F_{i\cdot} = \sum_{k=1}^{i} p_{k\cdot}; \;\; F_{\cdot j} = \sum_{l=1}^{j} p_{\cdot l}; \;\; F_{ij} = \sum_{k=1}^{i}\sum_{l=1}^{j} p_{kl}, \qquad (3)$$

которые кумулируют (накапливают) вероятности отдельных значений случайных величин.

Коэффициент 6 в формуле (1) подобран из условия нормировки коэффициента детерминации

$$0 \le def \le 1 \qquad (4)$$

и основан на величине максимума суммы в формуле (1). Максимум находится из следующих соображений. Предполагается, что он соответствует максимальной зависимости

между случайными величинами, которая имеет место при равенстве случайных величин:

$Y = X$. В этом случае

$$p_{ij} = P\left(X = x_i, X = x_j\right) = \begin{cases} p_{i\cdot}; i = j \\ 0; \quad i \ne j \end{cases};$$

$F_{ii} = \sum_{k=1}^{i} p_{kk} = \sum_{k=1}^{i} p_{k\cdot}$; $F_{i\cdot} = F_{\cdot i}$; Положим

$$F_{ii} = F_{i\cdot} = F_{\cdot i} = F_i; \;\; F_0 = 0. \qquad (5)$$

Тогда

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left|F_{ij} - F_{i\cdot}F_{\cdot j}\right|p_{ij} = \sum_{i=1}^{m}\left|F_i - F_i^2\right|p_i = \sum_{i=1}^{m} F_i\left(1-F_i\right)\left(F_i - F_{i-1}\right) = \sum_{i=1}^{m} F_i\left(1-F_i\right)\Delta F_i.$$

Далее находим

$$\sum_{i=1}^{m} F_i\left(1-F_i\right)\Delta F_i \xrightarrow[m\to\infty]{} \int_0^1 F\left(1-F\right)dF = \left(\frac{F^2}{2} - \frac{F^3}{3}\right)\Bigg|_0^1 = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

Итак, предполагается, что

$$\max \sum_{i=1}^{m}\sum_{j=1}^{n}\left|F_{ij} - F_{i\cdot}F_{\cdot j}\right| = \frac{1}{6}. \qquad (6)$$

Свойства дефектологического коэффициента детерминации.

1. $def = 0$ тогда и только тогда, когда случайные величины X, Y независимы.

Свойство следует из того, что необходимым и достаточным условием независимости дискретных случайных величин является равенство

$F_{ij} - F_{i\cdot}F_{\cdot j} = 0$ для любых $i, j = 1, 2, \ldots$

2. $$0 \le def \le 1. \qquad (4)$$

Для доказательства заметим, что левая граница очевидна, так как коэффициент детерминации неотрицателен, а правая граница следует из формулы (6) .

3. Верхняя граница в неравенстве (4) точная для случая, когда $Y = X$, что подтверждается примером 1.

Замечание. Формула (6) получена с помощью нестрогих эвристических рассуждений. Строгое доказательство пока неизвестно.

Пример 1. Рассмотрим двумерное распределение при $Y = X$, определенное табл. 1.

Вычисляем дефектологический коэффициент детерминации def .

Таблица 1. Диагональное равномерное распределение к примеру 1.

$X \downarrow Y \rightarrow$	1	2	…	n	$p_{i\cdot}$
1	$1/n$	0	…	0	$1/n$
2	0	$1/n$	…	0	$1/n$
…	…	…	…	…	…
n	0	0	…	$1/n$	$1/n$
$p_{\cdot j}$	$1/n$	$1/n$	…	$1/n$	1

Непосредственно по таблице 1 находим

$$def =$$

$$= 6\left[\left(\frac{1}{n}-\frac{1}{n^2}\right)\frac{1}{n}+\left(\frac{2}{n}-\frac{4}{n^2}\right)\frac{1}{n}+\left(\frac{3}{n}-\frac{9}{n^2}\right)\frac{1}{n}+\ldots+\left(\frac{n-1}{n}-\left(\frac{n-1}{n}\right)^2\right)\frac{1}{n}+(1-1)\frac{1}{n}\right]=$$

$$= \frac{6}{n^2}\left[\left(1+2+3+\ldots+(n-1)\right)-\frac{1}{n}\left(1^2+2^2+3^2+\ldots+(n-1)^2\right)\right]=$$

$$= \frac{6}{n^2}\left[\frac{1+(n-1)}{2}(n-1)-\frac{1}{n}\frac{(n-1)n\left(2(n-1)+1\right)}{6}\right]=\frac{n^2-1}{n^2}\xrightarrow[n\to\infty]{}1.$$

Пример2. Рассмотрим триномиальное распределение при $n = 2; p_1 = p_2 = 1/2$ из примера 1, § 3.15. Там была приведена таблица распределения, здесь таблица 1.

Таблица 2. триномиального распределения при $n = 2.$

$X \setminus Y \rightarrow$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00} = 1/4$	$p_{01} = 1/4$	$p_{02} = 1/16$	$p_{0\cdot} = 9/16$
1	$p_{10} = 1/4$	$p_{11} = 1/8$	$p_{12} = 0$	$p_{1\cdot} = 3/8$

2	$p_{20}=1/16$	$p_{21}=0$	$p_{22}=0$	$p_{2\cdot}=1/16$
$p_{\cdot j}$	$p_{\cdot 0}=9/16$	$p_{\cdot 1}=3/8$	$p_{\cdot 2}=1/16$	1

Непосредственно по таблице 2 находим

$$def=6\left[\left|\frac{1}{4}-\frac{9}{16}\frac{9}{16}\right|\frac{1}{4}+\left|\frac{1}{2}-\frac{9}{16}\frac{15}{16}\right|\frac{1}{4}+\left|\frac{9}{16}-\frac{9}{16}\cdot 1\right|\frac{1}{16}+\left|\frac{1}{2}-\frac{15}{16}\frac{9}{16}\right|\frac{1}{4}\right]+$$

$$+6\left[\left|\frac{7}{8}-\frac{15}{16}\frac{15}{16}\right|\frac{1}{8}+0+0+0+\left|\frac{9}{16}-1\cdot\frac{9}{16}\right|\frac{1}{16}\right]=\frac{6}{16^2\cdot 4}\left(17+7+7+\frac{1}{2}\right)=$$

$=\frac{189}{16^2\cdot 4}=\frac{189}{1024}\approx 0{,}185$. Таким образом,

$$def=0{,}185. \qquad (7)$$

Учитывая, что под знаком модуля все, отличные от нуля числа, отрицательны, вычисляем дефектологический коэффициент корреляции

$$def_c=-0{,}185. \qquad (8)$$

Для сравнения линейный коэффициент корреляции для этого же распределения (таблица 1)

$\rho=-1/3\approx -0{,}333$;

ассоциативный коэффициент детерминации (§ 7.2) $as=0{,}186$.

Пример 3. Вычисляем дефектологический коэффициент детерминации для распределения, определенного таблицей 3.

Таблица 3. Дискретное несимметричное распределение к примеру 3.

$X\downarrow Y\rightarrow$	1	2	3	$p_{i\cdot}$
1	2/12	1/12	1/12	4/12
2	4/12	3/12	1/12	8/12
$p_{\cdot j}$	6/12	4/12	2/12	1

Для этой таблицы 3 непосредственно получаем

$F_{1\cdot}=4/12;\ F_{2\cdot}=1;\ F_{\cdot 1}=6/12;F_{\cdot 2}=10/12;F_{\cdot 3}=1;$

$F_{11}=2/12;F_{12}=3/2;F_{13}=4/12;\ F_{21}=6/12;F_{22}=10/12;F_{23}=1.$

$def=6\left[\left|F_{11}-F_{1\cdot}F_{\cdot 1}\right|p_{11}+\left|F_{12}-F_{1\cdot}F_{\cdot 2}\right|p_{12}+\left|F_{13}-F_{1\cdot}F_{\cdot 3}\right|p_{13}\right]+$

$+6\left[\left|F_{21}-F_{2\cdot}F_{\cdot 1}\right|p_{21}+\left|F_{22}-F_{2\cdot}F_{\cdot 2}\right|p_{22}+\left|F_{23}-F_{2\cdot}F_{\cdot 3}\right|p_{23}\right];$

$$def = 6\left[0 + \left|-\frac{4}{12}\right|\frac{1}{12} + 0 + 0 + 0 + 0\right] = \frac{1}{6} \approx 0{,}167.$$

§ 8.3. Дефектологический коэффициент детерминации для непрерывных распределений.

Конструкцию и свойства дефекта событий положим в основу определения дефектологического коэффициента детерминации общего вида:

$$def = C \cdot M\left[\left|F_{XY}(X,Y) - F_X(X)F_Y(Y)\right|\right] \quad . \tag{1}$$

Для непрерывных распределений эта формула записывается в виде

$$def = C\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left|F_{XY}(x,y) - F_X(x)F_Y(y)\right| f_{XY}(x,y)\,dxdy\,. \tag{2}$$

Константу C нужно подобрать из условия нормировки коэффициента детерминации: $def \le 1$. Из свойств дефекта (§ 3.15) следует, что $\left|F_{XY}(x,y) - F_X(x)F_Y(y)\right| \le \frac{1}{4}$.

Отсюда вытекает, что можно положить $C = 4$.

Однако, есть соображения, что константу C можно увеличить, чтобы верхняя граница def была достижимой. Эти соображения следующие.

Полная зависимость между случайными величинами, то есть, когда $Y = X$, должна отвечать максимальному значению коэффициента детерминации. Тогда

$$F_{XY}(x,y) = P(X < x, X < y) = \begin{cases} P(X < x) = F_X(x); & x \le y \\ P(X < y) = F_X(y); & y < x \end{cases}.$$

В этом случае закон распределения является вырожденным: двумерная плотность распределения равна нулю вне прямой $y = x$ и равна $f_X(x)$ на прямой $y = x$:

$$\frac{\partial^2 f_{XY}(x,y)}{\partial x \partial y} = \begin{cases} 0; & y \ne x \\ f_X(x); & y = x \end{cases}.$$

Формулы (1), (2) в случае $Y = X$ для непрерывных случайных величин записываются в виде

$$def = C\int_{-\infty}^{+\infty}\left|F_X(x) - F_X^2(x)\right| dF_X(x)\,. \tag{3}$$

Так как выражение под знаком модуля неотрицательно, то знак модуля можно снять. Формула (3) принимает вид

$$def = C\int_{-\infty}^{+\infty}\left(F_X(x) - F_X^2(x)\right) dF_X(x)\,. \tag{4}$$

В интеграле (4) выполним подстановку $F_X(x) = z$. Тогда получаем

$$def = C\int_0^1 \left(z - z^2\right)dz = C\left(\frac{z^2}{2} - \frac{z^3}{3}\right)\Bigg|_0^1 = C\left(\frac{1}{2} - \frac{1}{3}\right) = C\frac{1}{6}.$$

Для выполнения равенства $def = 1$ достаточно положить $C = 6$.

Тогда формула (2) принимает вид

$$def = 6\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left|F_{XY}(x,y) - F_X(x)F_Y(y)\right| f_{XY}(x,y)\,dxdy, \qquad (5)$$

так как предположительно выполняется свойство

$$\max_f \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left|F_{XY}(x,y) - F_X(x)F_Y(y)\right| f_{XY}(x,y)dxdy = \frac{1}{6}. \qquad (6)$$

Свойства дефектологического коэффициента детерминации.

1.Коэффициент def равен нулю тогда и только тогда, когда случайные величины X, Y независимы .

Доказательство. Необходимое и достаточное условие независимости случайных величин

X, Y может быть записано в виде

$$F_{XY}(x,y) - F_X(x)F_Y(y) = 0, \ \forall x, y \text{ (с вероятностью 1).} \qquad (7)$$

При его выполнении подынтегральная функция в формуле (5) равна нулю, а потому $def = 0$.

Обратно, если $def = 0$, то вследствие неотрицательности подынтегральной функции она равна нулю везде кроме, может быть, множества меры нуль, то есть с вероятностью 1. Это означает выполнение равенства (7), то есть независимость X, Y

с вероятностью 1.

2. $$def \le 1. \qquad (8)$$

Это свойство предположительное, так как свойство (6) сформулировано лишь на основе эвристических рассуждений.

3. Верхняя граница в неравенстве (8) достигается при $Y = X$.

Это свойство доказано выше при проведении эвристических рассуждений для формулировки свойства, выраженного формулой (6).

Пример 1. Вычисление дефектологического коэффициента детерминации def для двумерной случайной величины (X, Y), распределенной равномерно в треугольнике с вершинами $O(0;0), A(1;0), B(0;1)$.

Указанное распределение применялось для вычисления ассоциативного и контингенциального коэффициентов детерминации (§ 4.1.) Используем полученные результаты.

Двумерная плотность вероятности в этом случае определяется формулами

$f_{XY}(x,y)=2$ при $(x;y)\in\Delta$; $f_{XY}(x,y)=0$ при $(x;y)\notin\Delta$.

Из этих формул следует, что двумерная функция распределения $F_{XY}(x,y)=2xy$ при $(x,y)\in\Delta$. Одномерные плотности распределения компонент X,Y соответственно выражаются формулами $f_X(x)=2(1-x);\quad f_Y(y)=2(1-y);\ 0\le x\le 1;0\le y\le 1$.

Соответствующие функции распределения компонент выражаются формулами $F_X(x)=2x-x^2;\quad F_Y(y)=2y-y^2;\ 0\le x\le 1;0\le y\le 1$. Тогда дефект δ выражается формулой

$$\delta=\left|F_{XY}(x,y)-F_X(x)F_Y(y)\right|=\left|2xy-\left(2x-x^2\right)\left(2y-y^2\right)\right|=xy\left|2x+2y-xy-2\right|.$$

Функция $u=2x+2y-xy-2$ в треугольнике Δ имеет наименьшее значение, равное (-2), а наибольшее значение, равное 0 , поэтому

$\delta=\left|F_{XY}(x,y)-F_X(x)F_Y(y)\right|=(2+xy-2x-2y)xy$. Отсюда

$$def=12\int_0^1\int_0^{1-x}(2+xy-2x-2y)xydxdy=12\int_0^1 xdx\int_0^{1-x}\left(2y+xy^2-2xy-2y^2\right)dy=$$

$$=4\int_0^1 x(1-x)^2\left(1-x^2\right)dx=\frac{4}{15}=0,267.$$ Итак,

$def=0,267.$

Приведем для сравнения величины других коэффициентов детерминации.

$as=0,255-$ ассоциативный коэффициент детерминации.

$co=0,57-$ контингенциальный коэффициент детерминации.

Линейный коэффициент корреляции $\rho=-0,5.$

Пример 2. Вычисление дефектологического коэффициента детерминации для двумерной случайной величины (X,Y), имеющей в квадрате

$D=\{(x;y):0\le x\le 1;0\le y\le 1\}$ симметричную плотность

$$f_{XY}(x,y)=\begin{cases}x+y;\quad (x;y)\in D\\ 0;\quad (x;y)\notin D\end{cases}.$$

В этом случае $f_X(x)=\int_0^1 f_{XY}(x,y)dy=\int_0^1(x+y)dy=x+\frac{1}{2};$

$F_X(x)=\int_0^x f_X(t)dt=\int_0^x\left(t+\frac{1}{2}\right)dt=\frac{1}{2}\left(x^2+x\right)$. По симметрии $F_Y(y)=\frac{1}{2}\left(y^2+y\right)$.

Далее

$$F_{XY}(x,y)=\int_0^x\int_0^y(u+v)dudv=\frac{1}{2}\left(x^2y+xy^2\right);$$

$$K(x,y)=F_{XY}(x,y)-F_X(x)F_Y(y)=\frac{1}{2}\left(x^2y-xy^2\right)-\frac{1}{2}\left(x^2+x\right)\frac{1}{2}\left(y^2+y\right)=$$

$$=\frac{1}{4}xy\left(x+y-xy-1\right)=-\frac{1}{4}xy(1-x)(1-y).$$ Отсюда

$$def=6\int_0^1\int_0^1\left|K(x,y)\right|f_{XY}(x,y)\,dxdy=6\int_0^1\int_0^1\frac{1}{4}xy(1-x)(1-y)(x+y)\,dxdy=$$

$$=2\frac{3}{2}\int_0^1\int_0^1xy(1-x)(1-y)xdxdy=3\int_0^1x^2(1-x)\,dx\int_0^1y(1-y)dy=3\left(\frac{1}{3}-\frac{1}{4}\right)\left(\frac{1}{2}-\frac{1}{3}\right)=$$

$$=3\frac{1}{12}\frac{1}{6}=\frac{1}{24}\approx 0,0417.$$

Глава 9. Приложения коэффициентов детерминации и корреляции

§ 9.1. Приложение коэффициентов детерминации и корреляции к исследованию криволинейной регрессии

Для исследования линейной регрессии удобен линейный коэффициент корреляции ρ, Так как основные характеристики регрессии – коэффициенты регрессии, угол между прямыми регрессии «y на x» и «x на y», остаточная дисперсия и другие очень просто через него выражаются. И вообще линейный коэффициент корреляции очень хорошо характеризует величину вероятностной зависимости между случайными величинами, описываемую линейной регрессией. Это равно касается, как теоретической, так и эмпирической регрессионной зависимости.

Если же регрессия является нелинейной, то линейный коэффициент корреляции характеризует величину зависимости уже не столь хорошо. Во-первых, равенство нулю коэффициента корреляции вовсе не означает в этом случае отсутствие зависимости.
Во- вторых, нет простых формул, выражающих основные характеристики регрессии через коэффициент корреляции.

Для исследования величины корреляционной (вероятностной) зависимости в случае нелинейной регрессии удобно применить коэффициенты детерминации, рассматриваемые в этой книге – коэффициенты ассоциации, контингенции, предельный и комби. Когда эмпирические точки имеют разброс относительно тренда, то эти коэффициенты реагируют на величину разброса (чем разброс больше, тем коэффициенты меньше). Следовательно, эти коэффициенты, характеризуя величину корреляционной зависимости, характеризуют и рассеяние случайных величин относительно тренда. Целесообразно использовать не один коэффициент, а несколько и коэффициенты корреляции в том числе для системного анализа. Коэффициенты корреляции отличаются от коэффициентов детерминации тем, что снят знак модуля у функций, стоящих под знаком суммы или интеграла, поэтому они заключены в пределах промежутка [–1;1]. У коэффициентов детерминации «комби» знак модуля можно снимать различно, поэтому и коэффициентов корреляции в этом случае может быть несколько.
В дальнейшем будем рассматривать коэффициенты детерминации и корреляции только для дискретных случайных величин. В этом случае вероятности, из которых построены ядра коэффициентов, заменяем на относительные частоты для экспериментальных точек. Эти экспериментальные точки группируются около графика функции регрессии, который будем именовать трендом. Тренд можно построить различными способами. Наиболее распространенный способ – метод наименьших квадратов. Наша задача сейчас состоит не в построении тренда, а в изучении поведения коэффициентов связи при различном расположении экспериментальных точек относительно тренда. Поэтому будем исходить из готового тренда, не вдаваясь в способ его построения. Целесообразно взять самый простой криволинейный тренд для простоты вычислений и для простоты осмысления результатов.

В качестве тренда выберем параболу $y = x^2$. Рассмотрим 3 случая группировки экспериментальных точек относительно тренда или какой-либо точки:

1. Экспериментальные точки лежат точно на параболе, то есть в этом случае случайные величины связаны точной квадратической зависимостью $Y = X^2$.

2. Экспериментальные точки не лежат на параболе, а лишь группируются около параболы с некоторым разбросом.

3. Экспериментальные точки образуют круговое облако, не выражая тенденции к какой-либо зависимости в среднем, то есть группируются около какой-либо точки.

Наиболее часто встречающийся случай построения регрессии – случай однократных наблюдений для каждого значения x, то есть повторные наблюдения отсутствуют. Тогда относительная частота каждого наблюдения равна $1/n$ – одна и та же (n – общее число наблюдений), хотя в общем случае мы не имеем дело с равномерным распределением. Выход – в группировке, то есть в построении двумерного группированного статистического ряда. Для этого мы заключим область группирования точек в прямоугольник и разобьем его на $5^2 = 25$ равных по величине клеток. Пусть n – число наблюдений двумерной случайной величины $(X;Y)$. Число n наблюдений для простоты вычислений возьмем небольшое: $n = 20$. Хотя это число явно небольшое, однако достаточное для образования относительных частот и выявления закономерностей. Размеры прямоугольника возьмем 10 и 100; $x \in [0;10]$, $y \in [0;100]$.

Приступим к вычислению коэффициентов связи в трех объявленных выше случаях. Это будет выполнено в параграфах 9.2 – 9.4.

§ 9.2. Примеры вычисления коэффициентов детерминации и корреляции, когда экспериментальные точки лежат точно на квадратичной-параболе

Случай, когда экспериментальные точки лежат точно на параболе, является вырожденным. Он означает, что величины X, Y связаны функциональной зависимостью $Y = X^2$. Будем считать, что переменная X принимает значения с равным шагом $h = 0,5$ на промежутке $[0;10]$. Получаем следующую последовательность экспериментальных точек:

$(0,5;0,25),(1;1),(1,5;2,25),(2;4),(2,5;6,25),(3;9),(3,5;12,25),(4;16),$
$(4,5;20,25),(5;25),(5,5;30,25),(6;36),(6,5;42,25),(7;49),$
$(7,5;56,25),(8;64),(8,5;72,25),(9;81),(9,5;90,25),(10;100)$.

Распределим эти точки по клеткам корреляционной таблицы 1. Таблица построена так, что клетки, в которых стоят относительные частоты, выявляют контур рассматриваемой параболы $y = x^2$.

Пример 1. По данным таблицы 1 вычисляем линейный коэффициент корреляции ρ.

Предварительно вычисляем моменты первых двух порядков.

$$m_X = \sum_i x_i p_{i\cdot} = \frac{1}{20}(1 \cdot 3 + 3 \cdot 4 + 5 \cdot 4 + 7 \cdot 4 + 9 \cdot 5) = 5,4;$$

$$m_Y = \sum_j y_j p_{\cdot j} = \frac{1}{20}\left(10\cdot 8 + 30\cdot 4 + 50\cdot 3 + 70\cdot 2 + 90\cdot 3\right) = 38;$$

$$MX^2 = \sum_i x_i^2 p_{i\cdot} = \frac{1}{20}\left(1\cdot 3 + 9\cdot 4 + 25\cdot 4 + 49\cdot 4 + 81\cdot 5\right) = 37;$$

$$D_X = MX^2 - m_X^2 = 37 - \left(5,4\right)^2 = 7,84; \quad \sigma_X = \sqrt{D_X} = 2,8;$$

$$MY^2 \sum y_j^2 p_{\cdot j} = \frac{1}{20}\left(100\cdot 8 + 900\cdot 4 + 2500\cdot 3 + 4900\cdot 2 + 8100\cdot 3\right) = 2300;$$

$$DY = MY^2 - m_Y^2 = 2300 - \left(38\right)^2 856; \quad \sigma_Y = \sqrt{DY} = 29,257478;$$

$$M\left[XY\right] =$$

$$= \frac{1}{20}\left(1\cdot 10\cdot 3 + 3\cdot 10\cdot 4 + 5\cdot 10\cdot 1 + 5\cdot 30\cdot 3 + 7\cdot 30\cdot 1 + 7\cdot 50\cdot 3 + 9\cdot 70\cdot 2 + 9\cdot 90\cdot 3\right) =$$

$$= 280; \quad K_{XY} = M\left[XY\right] - m_X m_Y = 280 - 5,4\cdot 38 = 74,8;$$

$$\rho = \frac{K_{XY}}{\sigma_X \sigma_Y} = \frac{74,8}{2,8\cdot 29,257} = 0,913 \qquad (1)$$

Таблица 1. Группированный статистический ряд значений двумерной случайной величины $\left(X, X^2\right)$.

$p_{\cdot j}$	3/20	4/20	4/20	4/20	5/20	1
$[80;100]$ $y_5 = 90$					3/20	3/20
$[60;80)$ $y_4 = 70$					2/20	2/20
$[40;60)$ $y_3 = 50$				3/20		3/20
$[20;40)$ $y_2 = 30$			3/20	1/20		4/20

$[0;20)$ $y_1 = 10$	3/20	4/20	1/20			8/20
Промежутки. Средние точки. $y \uparrow x \rightarrow$	$[0;2)$ $x_1 = 1$	$[2;4)$ $x_2 = 3$	$[4;6)$ $x_3 = 5$	$[6;8)$ $x_4 = 7$	$[8;10]$ $x_5 = 9$	$p_{i\cdot}$

Пример 2. По данным таблицы 1 вычисляем ассоциативный коэффициент детерминации.

$$as = as_1 = \sum_i \sum_j \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}} p_{ij} =$$

$$= \frac{\left| \frac{3}{20} - \frac{8}{20}\frac{3}{20} \right|}{\sqrt{\frac{8}{20}\frac{12}{20}\frac{3}{20}\frac{17}{20}}} \frac{3}{20} + \frac{\left| \frac{4}{20} - \frac{8}{20} \right|}{\sqrt{\frac{8}{20}\frac{12}{20}\frac{4}{20}\frac{16}{20}}} \frac{4}{20} +$$

$$+ \frac{\left| \frac{1}{20} - \frac{8}{20}\frac{4}{20} \right|}{\sqrt{\frac{8}{20}\frac{12}{20}\frac{4}{20}\frac{16}{20}}} \frac{1}{20} + \frac{\left| \frac{3}{20} - \frac{4}{20}\frac{4}{20} \right|}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{4}{20}\frac{16}{20}}} \frac{3}{20} + \frac{\left| \frac{1}{20} - \frac{4}{20}\frac{4}{20} \right|}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{4}{20}\frac{16}{20}}} \frac{1}{20} + \frac{\left| \frac{3}{20} - \frac{3}{20}\frac{4}{20} \right|}{\sqrt{\frac{3}{20}\frac{17}{20}\frac{4}{20}\frac{16}{20}}} \frac{3}{20} +$$

$$+ \frac{\left| -\frac{2}{20}\frac{5}{20} \right|}{\sqrt{\frac{2}{20}\frac{18}{20}\frac{5}{20}\frac{15}{20}}} \frac{2}{20} + \frac{\left| \frac{3}{20} - \frac{3}{20}\frac{5}{20} \right|}{\sqrt{\frac{3}{20}\frac{17}{20}\frac{5}{20}\frac{15}{20}}} \frac{3}{20} = \frac{1}{20}[1{,}543487 + 2{,}449490 + 0{,}153093 +$$

$2{,}962500 + + 0{,}062500 + 2{,}520504 + 1{,}154701 + 2{,}182821] = 0{,}606455 \approx 0{,}606.$

$$as = 0{,}606. \qquad (2)$$

Пример 3. Вычисляем контингенциальный коэффициент детерминации $co = co_1$.

$$co = \sum_i \sum_j \frac{\left| p_{ij} - p_{i\cdot} p_{\cdot j} \right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij}.$$

$$co = \frac{\left(\frac{3}{20} - \frac{8}{20}\frac{3}{20}\right)\frac{3}{20}}{\frac{3}{20}\left(1 + 2\frac{3}{20} - 2\frac{8}{20} - 2\frac{3}{20}\right) + \frac{8}{20}\frac{3}{20}} + \frac{\left(\frac{4}{20} - \frac{8}{20}\frac{4}{20}\right)\frac{4}{20}}{\frac{4}{20}\left(1 + 2\frac{4}{20} - 2\frac{8}{20} - 2\frac{4}{20}\right) + \frac{8}{20}\frac{4}{20}} +$$

$$+ \frac{\left|\frac{1}{20} - \frac{8}{20}\frac{4}{20}\right|\frac{1}{20}}{\frac{1}{20}\left(1 + 2\frac{1}{20} - 2\frac{8}{20} - 2\frac{4}{20}\right) + \frac{8}{20}\frac{4}{20}} + \frac{\left(\frac{3}{20} - \frac{4}{20}\frac{4}{20}\right)\frac{3}{20}}{\frac{3}{20}\left(1 + 2\frac{3}{20} - 2\frac{4}{20} - 2\frac{4}{20}\right) + \frac{4}{20}\frac{4}{20}} +$$

$$+ \frac{\left(\frac{1}{20} - \frac{4}{20}\frac{4}{20}\right)\frac{1}{20}}{\frac{1}{20}\left(1 + 2\frac{1}{20} - 2\frac{4}{20} - 2\frac{4}{20}\right) + \frac{4}{20}\frac{4}{20}} + \frac{\left(\frac{3}{20} - \frac{3}{20}\frac{4}{20}\right)\frac{3}{20}}{\frac{3}{20}\left(1 + 2\frac{3}{20} - 2\frac{3}{20} - 2\frac{4}{20}\right) + \frac{3}{20}\frac{4}{20}} +$$

$$+ \frac{\left(\frac{2}{20} - \frac{2}{20}\frac{5}{20}\right)\frac{2}{20}}{\frac{2}{20}\left(1 + 2\frac{2}{20} - 2\frac{2}{20} - 2\frac{2}{20}\right) + \frac{2}{20}\frac{5}{20}} + \frac{\left(\frac{3}{20} - \frac{3}{20}\frac{5}{20}\right)\frac{3}{20}}{\frac{3}{20}\left(1 + 2\frac{3}{20} - 2\frac{3}{20} - 2\frac{5}{20}\right) + \frac{3}{20}\frac{5}{20}} =$$

$$= \frac{1}{20}\left(3 + 4{,}8 + 0{,}4 + 3{,}869565 + 0{,}181818 + 3 + 2 + 3\right) = 0{,}963.$$

$$co = 0{,}963. \qquad (3)$$

Пример 4. Вычисляем комбинированный коэффициент детерминации комби-конт. Используем результаты вычислений в примерах 1,3.

$$com_c = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j |x_i - m_X| |y_j - m_Y| \frac{\left|p_{ij} - p_{i\cdot} p_{\cdot j}\right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij}.$$

$$com_c = \frac{1}{20 \cdot 2{,}8 \cdot 29{,}257478}\left(|1 - 5{,}4||10 - 38|3 + |3 - 5{,}4||10 - 38|4{,}8 + \right.$$

$$+ |5 - 5{,}4||10 - 38|0{,}4 + |5 - 5{,}4||30 - 38|2{,}869565 + |7 - 5{,}4||30 - 38|0{,}181818 +$$

$$\left. + |7 - 5{,}4||50 - 38|3 + |9 - 5{,}4||70 - 38|2 + |9 - 5{,}4||90 - 38|3\right) = 0{,}950764 \approx 0{,}951.$$

$$com_c = 0{,}951. \qquad (4)$$

Пример 5. Вычисляем комбинированный коэффициент детерминации комби-ас. Используем результаты вычислений в примерах 2, 4.

$$com_a = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j |x_i - m_X| |y_j - m_Y| \frac{|p_{ij} - p_{i\cdot} p_{\cdot j}|}{\sqrt{p_{i\cdot}(1-p_{i\cdot}) p_{\cdot j}(1-p_{\cdot j})}} p_{ij} .$$

$$com_a = \frac{1}{20 \cdot 2{,}8 \cdot 29{,}257478} \big(4{,}4 \cdot 28 \cdot 1{,}543487 + 2{,}4 \cdot 28 \cdot 2{,}449490 +$$

$$+0{,}4 \cdot 28 \cdot 0{,}153093 + 0{,}4 \cdot 8 \cdot 2{,}062500 + 1{,}6 \cdot 8 \cdot 0{,}062500 + 1{,}6 \cdot 12 \cdot 2{,}520504 +$$

$$+3{,}6 \cdot 32 \cdot 1{,}154701 + 3{,}6 \cdot 52 \cdot 2{,}182821 \big) = 0{,}582218 \approx 0{,}582.$$

$$com_a = 0{,}582. \qquad (5)$$

Пример 6. Вычисляем предельный коэффициент детерминации $l = l_1$.

$$l = \sum_i \sum_j \frac{|p_{ij} - p_{i\cdot} p_{\cdot j}|}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij}.$$

$$l = \frac{1}{20} \left(\frac{60-24}{60+24} 3 + \frac{80-32}{80+32} 4 + \frac{|20-32|}{20+32} \cdot 1 + \frac{60-16}{60+16} 3 + \frac{20-16}{20+16} \cdot 1 + \frac{60-12}{60+12} 3 + \right.$$

$$\left. + \frac{40-10}{40+10} 2 + \frac{60-15}{60+15} 3 \right) = 0{,}503936 \approx 0{,}504.$$

$$l = 0{,}504. \qquad (6)$$

§ 9.3. Примеры вычисления коэффициентов детерминации и корреляции, когда экспериментальные точки имеют разброс относительно тренда: квадратичной параболы

Рассмотрим следующую последовательность из 20 экспериментальных точек**,** имеющих некоторый разброс относительно тренда – параболы $y = x^2$:

(0,5;2), (1;0,5), (1,5;4), (2;1), (2,5;3), (3;19), (3,5;9), (4;21), (4,5;15), (5;10), (5,5;42), (6;30), (5,5;62), (7;35), (7,5;70), (8;50), (8,5;85), (9;70), (9,5;95), 10;85).

Составим из этих точек группированный статистический ряд, представленный таблицей 1.

Этот ряд аналогичен тому, что построен в § 9.2. (табл.1). Общие вопросы применения коэффициентов связи к исследованию криволинейной регрессии изложены в § 9.1..

Заложенный разброс точек относительно тренда $y = x^2$ в приводимой ниже таблице 1 виден.

Пример 1. Вычисляем линейный коэффициент корреляции ρ по данным таблицы 1.

Предварительно вычисляем моменты первых двух порядков.

$$m_X = 1\cdot\frac{3}{20} + 3\frac{4}{20} + 5\frac{4}{20} + 7\frac{4}{20} + 9\frac{5}{20} = 5,4.$$

$$m_Y = \frac{1}{20}\left(10\cdot 9 + 30\cdot 3 + 50\cdot 2 + 70\cdot 3 + 90\cdot 3\right) = 38.$$

$$MX^2 = \frac{1}{20}\left(1\cdot 3 + 9\cdot 4 + 25\cdot 4 + 49\cdot 4 + 81\cdot 5\right) = 37.$$

$$D_X = MX^2 - m_X^2 = 37 - \left(5,4\right)^2 = 7,84;$$

$$\sigma_X = \sqrt{D_X} = \sqrt{7,84} = 2,8.$$

$$MY^2 = \frac{1}{20}\left(100\cdot 9 + 900\cdot 3 + 2500\cdot 2 + 4900\cdot 3 + 8100\cdot 3\right) = 2380.$$

$$D_Y = MY^2 - m_Y^2 = 2380 - \left(38\right)^2 = 936.$$

$$\sigma_Y = \sqrt{D_Y} = \sqrt{936} = 30,594117.$$

$$M\left[XY\right] = \frac{1}{20}\left(1\cdot 10\cdot 3 + 3\cdot 10\cdot 4 + 5\cdot 10\cdot 2 + 5\cdot 30\cdot 1 + 5\cdot 50\cdot 1 + 7\cdot 30\cdot 2 + 7\cdot 70\cdot 2 + \right.$$

$$\left. + 9\cdot 50\cdot 1 + 9\cdot 70\cdot 1 + 9\cdot 90\cdot 3\right) = 278.$$

$$K_{XY} = M\left[XY\right] - m_X m_Y = 278 - 5,4\cdot 38 = 72,8;$$

Таблица 1. Группированный статистический ряд экспериментальных точек с разбросом относительно тренда $y = x^2$.

$p_{\cdot j}$	3/20	4/20	4/20	4/20	5/20	1
[80;10] $y_5 = 90$					3/20	3/20
[60;80) $y_4 = 70$				2/20	1/20	3/20
[40;60) $y_3 = 50$			1/20		1/20	2/20
[20;40) $y_2 = 30$			1/20	2/20		3/20
[0;20) $y_1 = 10$	3/20	4/20	2/20			9/20
Промежутки. Средние	[0;2)	[2;4)	[4;6)	[6;8)	[8;10)	$p_{i\cdot}$

точки. $y\uparrow\ x\rightarrow$	$x_1=1$	$x_2=3$	$x_3=5$	$x_4=7$	$x_5=9$	

$$\rho=\frac{K_{XY}}{\sigma_X\sigma_Y}=\frac{72{,}8}{2{,}8\cdot 30{,}594117}=0{,}849837\approx 0{,}850.$$

$$\rho=0{,}850. \qquad (1)$$

Пример 2. Вычисляем ассоциативный коэффициент детерминации $as=as_1$.

Используем данные таблицы 1.

$$as=\sum_i\sum_j\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}p_{ij}=\frac{1}{20}\left(\frac{\left(\frac{3}{20}-\frac{3}{20}\frac{9}{20}\right)3}{\sqrt{\frac{3}{20}\frac{17}{20}\frac{9}{20}\frac{11}{20}}}+\frac{\left(\frac{4}{20}-\frac{4}{20}\frac{9}{20}\right)4}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{9}{20}\frac{11}{20}}}+\right.$$

$$+\frac{\left(\frac{2}{20}-\frac{4}{20}\frac{9}{20}\right)2}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{9}{20}\frac{11}{20}}}+\frac{\left(\frac{1}{20}-\frac{4}{20}\frac{3}{20}\right)\cdot 1}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{3}{20}\frac{17}{20}}}+\frac{\left(\frac{1}{20}-\frac{4}{20}\frac{2}{20}\right)\cdot 1}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{2}{20}\frac{18}{20}}}+\frac{\left(\frac{2}{20}-\frac{4}{20}\frac{3}{20}\right)2}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{3}{20}\frac{17}{20}}}+$$

$$\left.+\frac{\left(\frac{2}{20}-\frac{4}{20}\frac{3}{20}\right)2}{\sqrt{\frac{4}{20}\frac{16}{20}\frac{3}{20}\frac{17}{20}}}+\frac{\left(\frac{1}{20}-\frac{5}{20}\frac{2}{29}\right)\cdot 1}{\sqrt{\frac{5}{20}\frac{15}{20}\frac{2}{29}\frac{18}{20}}}+\frac{\left(\frac{1}{20}-\frac{5}{20}\frac{3}{20}\right)\cdot 1}{\sqrt{\frac{5}{20}\frac{15}{20}\frac{3}{20}\frac{17}{20}}}+\frac{\left(\frac{3}{20}-\frac{5}{20}\frac{3}{20}\right)3}{\sqrt{\frac{5}{20}\frac{15}{20}\frac{3}{20}\frac{17}{20}}}\right)=$$

$$=\frac{1}{20}(1{,}393261+2{,}211083+0{,}100504+0{,}140028+0{,}250000+0{,}980196+$$

$$+0{,}980196+0{,}192450+0{,}046676+1{,}260252)=0{,}377732\approx 0{,}378.$$

$$as=0{,}378. \qquad (2)$$

Пример 3. Вычисляем контингенциальный коэффициент детерминации $co=co_1$ по данным таблицы 1.

$$co=\sum_i\sum_j\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}p_{ij}=$$

$$=\frac{1}{20}\left[\frac{\left(\frac{3}{20}-\frac{3}{20}\frac{9}{20}\right)3}{\frac{3}{20}\left(1+2\frac{3}{20}-2\frac{3}{20}-2\frac{9}{20}\right)+\frac{3}{20}\frac{9}{20}}+\frac{\left(\frac{4}{20}-\frac{4}{20}\frac{9}{20}\right)4}{\frac{4}{20}\left(1+2\frac{4}{20}-2\frac{4}{20}-2\frac{9}{20}\right)+\frac{4}{20}\frac{9}{20}}+\right.$$

$$+\frac{\left(\frac{2}{20}-\frac{4}{20}\frac{9}{20}\right)2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{4}{20}-2\frac{9}{20}\right)+\frac{4}{20}\frac{9}{20}}+\frac{\left(\frac{1}{20}-\frac{4}{20}\frac{3}{20}\right)\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{4}{20}-2\frac{3}{20}\right)+\frac{4}{20}\frac{3}{20}}+$$

$$+\frac{\left(\frac{1}{20}-\frac{4}{20}\frac{2}{20}\right)\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{4}{20}-2\frac{2}{20}\right)+\frac{4}{20}\frac{2}{20}}+\frac{\left(\frac{2}{20}-\frac{4}{20}\frac{3}{20}\right)2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{4}{20}-2\frac{3}{20}\right)+\frac{4}{20}\frac{3}{20}}+$$

$$+\frac{\left(\frac{2}{20}-\frac{3}{20}\frac{4}{20}\right)2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{3}{20}-2\frac{4}{20}\right)+\frac{3}{20}\frac{4}{20}}+\frac{\left(\frac{1}{20}-\frac{5}{20}\frac{2}{20}\right)\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{5}{20}-2\frac{2}{20}\right)+\frac{5}{20}\frac{2}{20}}+$$

$$\left.+\frac{\left(\frac{1}{20}-\frac{5}{20}\frac{3}{20}\right)\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{5}{20}-2\frac{3}{20}\right)+\frac{5}{20}\frac{3}{20}}+\frac{\left(\frac{3}{20}-\frac{5}{20}\frac{3}{20}\right)3}{\frac{3}{20}\left(1+2\frac{3}{20}-2\frac{5}{20}-2\frac{3}{20}\right)}\right]=$$

$$=\frac{1}{20}(3+4+1/4+2/5+2/3+7/4+7/4+5/9+5/21+3)=0,780516\approx 0,781.$$

$$co=0,781. \qquad (3)$$

Пример 4. Вычисляем комбинированный коэффициент детерминации комби-ас по данным таблицы 1.

Воспользуемся результатами вычислений в примерах 1,2.

$$com_a=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j\left|x_i-m_X\right|\left|y_j-m_Y\right|\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}p_{ij}=$$

$$=\frac{1}{20\cdot 2,8\cdot 30,594117}(4,4\cdot 28\cdot 1,393261+\ 2,4\cdot 28\cdot 2,211083+0,4\cdot 28\cdot 0,100504+$$

$$+0,4\cdot 8\cdot 0,140028+0,4\cdot 12\cdot 0,25+1,6\cdot 32\cdot 0,980196+\ 4,4\cdot 12\cdot 0,192450+$$

$$+4,4\cdot 32\cdot 0.046676+4,4\cdot 52\cdot 1,260252)=0,403217\approx 0,403.$$

$$com_a = 0,403. \qquad (4)$$

Пример 5. Вычисляем по данным таблицы 1 комбинированный коэффициент детерминации

Комби-конт. Используем результаты вычислений в примерах 1,3.

$$com_c = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j |x_i - m_X| |y_j - m_Y| \frac{|p_{ij} - p_{i\cdot} p_{\cdot j}|}{p_{ij}(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}) + p_{i\cdot} p_{\cdot j}} p_{ij} =$$

$$= \frac{1}{20 \cdot 2,8 \cdot 30,594117} \left(4,4 \cdot 28 \cdot 3 + 2,4 \cdot 28 \cdot 4 + 0,4 \cdot 28 \cdot \frac{1}{4} + \right.$$

$$+0,4 \cdot 8 \cdot \frac{2}{5} + 0,4 \cdot 12 \cdot \frac{2}{3} +$$

$$\left. +1,6 \cdot 8 \cdot \frac{7}{4} + 1,6 \cdot 32 \cdot \frac{7}{4} + 4,4 \cdot 12 \cdot \frac{5}{9} + 4,4 \cdot 32 \cdot \frac{5}{21} + 4,4 \cdot 52 \cdot 3 \right) = 0,870229 \approx 0,870.$$

$$com_c = 0,870. \qquad (5)$$

Пример 6. Вычисляем по данным таблицы 1 предельный коэффициент детерминации $l = l_1$.

$$l = \sum_i \sum_j \frac{|p_{ij} - p_{i\cdot} p_{\cdot j}|}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij} = \frac{1}{20} \left(\frac{60-27}{60+27} 3 + \frac{80-36}{80+36} 4 + \frac{40-36}{40+36} 2 + \frac{20-12}{20+12} \cdot 1 + \right.$$

$$\left. + \frac{20-8}{20+8} \cdot 1 + \frac{40-12}{40+12} 2 + \frac{40-12}{40+12} 2 + \frac{20-10}{20+10} \cdot 1 + \frac{20-15}{20+15} \cdot 1 + \frac{60-15}{60+15} 3 \right) =$$

$=0,361035 \approx 0,361.$

$$l = 0,381.$$

(6)

§ 4.4. Примеры вычисления коэффициентов детерминации и корреляции при круговом облаке экспериментальных точек.

Рассмотрим следующую последовательность экспериментальных точек, которые образуют круговое облако. Это означает, что при возрастании абсцисс точек их ординаты в среднем не возрастают и не убывают:

(2,5;30), (2,2;43), (3,5;55), (2,4;70), (3,4;75), (4,5;32), (5,5;45), (4,4; 46), (4,6;50), (5,3;55), (5,5;58), (4,2;60), (5,4;65), (5;90), (7;35), (6,5;52), (7,2;45), (7,5;55), (6,4;73), (9;50).

Распределим эти точки по клеткам корреляционной таблицы 1. Таблицу по форме построим такую же, как в параграфах 9.2 и 9.3 Общие вопросы применения коэффициентов связи к исследованию криволинейной регрессии изложены в § 9.1. Относительные частоты, стоящие в клетках таблицы, соответствуют количеству экспериментальных

точек, находящихся в той части плоскости, которая накрывается этой клеткой. Облако точек круговой формы в таблице наблюдается.

Пример 1. По данным таблицы 1 вычисляем линейный коэффициент корреляции ρ.

Сначала вычисляем моменты первых двух порядков.

Таблица 1. Группированный статистический ряд экспериментальных точек, составляющих круговое облако.

$p_{i\cdot}$	0	5/20	9/20	5/20	1/20	**1**
[80;100) $y_5 = 90$			1/20			1/20
[60;80) $y_4 = 70$		2/20	2/20	1/20		5/20
[40;60) $y_3 = 50$		2/20	4/20	3/20	1/20	10/20
[20;40) $y_2 = 30$		1/20	2/20	1/20		4/20
[0;20) $y_1 = 10$						0
Промежутки. Средние точки. $y \uparrow x \rightarrow$	[0;2) $x_1 = 1$	[2;4) $x_2 = 3$	[4;6) $x_3 = 5$	[6;8) $x_4 = 7$	[8;10] $x_5 = 9$	$p_{\cdot j}$

$$m_X = \frac{1}{20}\left(3\cdot 5 + 5\cdot 9 + 7\cdot 5 + 9\cdot 1\right) = 5,2;$$

$$m_Y = \frac{1}{20}\left(30\cdot 4 + 50\cdot 10 + 70\cdot 5 + 90\cdot 1\right) = 53;$$

$$MX^2 = \frac{1}{20}\left(9\cdot 5 + 25\cdot 9 + 49\cdot 5 + 81\cdot 1\right) = 29,8;$$

$$D_X = MX^2 - m_X^2 = 29,8 - (5,2)^2 = 2,76;\ \ \sigma_X = \sqrt{D_X} = \sqrt{2,76} = 1,661325;$$

$$MY^2 = \frac{1}{20}\left(900\cdot 4 + 2500\cdot 10 + 4900\cdot 5 + 8100\cdot 1\right) = 3060;$$

$$D_Y = MY^2 - m_Y^2 = 3060 - -2805 = 255;\ \ \sigma_Y = \sqrt{D_Y} = \sqrt{255} = 15,968719;$$

$$M\left[XY\right] = \frac{1}{20}\left(90\cdot 1 + 150\cdot 2 + 210\cdot 2 + 150\cdot 2 + 250\cdot 4 + 350\cdot 2 + 450\cdot 1 + \right.$$

$$+210\cdot 1+350\cdot 3+490\cdot 1+450\cdot 1)=273;$$

$$K_{XY}=M\left[XY\right]-m_X m_Y=273-5,2\cdot 53=-2,6;$$

$$\rho=\frac{K_{XY}}{\sigma_X\sigma_Y}=\frac{-2,6}{1,661325\cdot 15,968719}=-0,09800\approx -0,100.$$

$$\rho=-0,1. \qquad (1)$$

Пример 2. Вычисляем ассоциативный коэффициент детерминации as по данным таблицы 1.

$$as=\sum_i\sum_j\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}p_{ij}=$$

$$=\frac{1}{20}\left[0+\frac{\left|\frac{2}{20}-\frac{5}{20}\frac{10}{20}\right|2}{\sqrt{\frac{5}{20}\frac{15}{20}\frac{10}{20}\frac{10}{20}}}+\frac{\left(\frac{2}{20}-\frac{5}{20}\frac{5}{20}\right)2}{\sqrt{\frac{5}{20}\frac{5}{20}\frac{5}{20}\frac{5}{20}}}+\right.$$

$$+\frac{\left(\frac{2}{20}-\frac{9}{20}\frac{4}{20}\right)2}{\sqrt{\frac{9}{20}\frac{11}{20}\frac{4}{20}\frac{16}{20}}}+\frac{\left|\frac{4}{20}-\frac{9}{20}\frac{10}{20}\right|4}{\sqrt{\frac{9}{20}\frac{11}{20}\frac{10}{20}\frac{10}{20}}}+\frac{\left|\frac{2}{20}-\frac{9}{20}\frac{5}{20}\right|2}{\sqrt{\frac{9}{20}\frac{11}{20}\frac{5}{20}\frac{15}{20}}}+\frac{\left(\frac{1}{20}-\frac{9}{20}\frac{1}{20}\right)\cdot 1}{\sqrt{\frac{9}{20}\frac{11}{20}\frac{1}{20}\frac{19}{20}}}+0+$$

$$\left.+\frac{\left(\frac{3}{20}-\frac{5}{20}\frac{10}{20}\right)3}{\sqrt{\frac{5}{20}\frac{15}{20}\frac{10}{20}\frac{10}{20}}}+\frac{\left|\frac{1}{20}-\frac{5}{20}\frac{5}{20}\right|\cdot 1}{\sqrt{\frac{5}{20}\frac{5}{20}\frac{5}{20}\frac{5}{20}}}+\frac{\left(\frac{1}{20}-\frac{1}{20}\frac{10}{20}\right)\cdot 1}{\sqrt{\frac{1}{20}\frac{19}{20}\frac{10}{20}\frac{10}{20}}}\right]=\frac{1}{20}\left[0+0,230940+0,4+\right.$$

$$+0,100504+0,402015+0,116052+0,253629+0,346410+0,066667+0,229416]=$$

$$=0,107282\approx 0,107.$$

$$as=0,107. \qquad (2)$$

Пример 3. Вычисляем контингенциальный коэффициент детерминации co по данным таблицы 1.

$$co=\sum_i\sum_j\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}p_{ij}=\frac{1}{20}\left[0+\right.$$

$$+\frac{\left|\frac{2}{20}-\frac{5}{20}\frac{10}{20}\right|2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{5}{20}-2\frac{10}{20}\right)+\frac{5}{20}\frac{10}{20}}+\frac{\left(\frac{2}{20}-\frac{5}{20}\frac{5}{20}\right)2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{5}{20}-2\frac{5}{20}\right)+\frac{5}{20}\frac{5}{20}}+$$

$$\frac{\left(\frac{2}{20}-\frac{9}{20}\frac{4}{20}\right)2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{9}{20}-2\frac{4}{20}\right)+\frac{9}{20}\frac{4}{20}}+\frac{\left|\frac{4}{20}-\frac{9}{20}\frac{10}{20}\right|4}{\frac{4}{20}\left(1+2\frac{4}{20}-2\frac{9}{20}-2\frac{10}{20}\right)+\frac{9}{20}\frac{10}{20}}+$$

$$\frac{\left|\frac{2}{20}-\frac{9}{20}\frac{5}{20}\right|2}{\frac{2}{20}\left(1+2\frac{2}{20}-2\frac{9}{20}-2\frac{5}{20}\right)+\frac{9}{20}\frac{5}{20}}+\frac{\left(\frac{1}{20}-\frac{9}{20}\frac{1}{20}\right)\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{9}{20}-2\frac{1}{20}\right)+\frac{9}{20}\frac{1}{20}}+0+$$

$$+\frac{\left(\frac{3}{20}-\frac{5}{20}\frac{10}{20}\right)3}{\frac{3}{20}\left(1+2\frac{3}{20}-2\frac{5}{20}-2\frac{10}{20}\right)+\frac{5}{20}\frac{10}{20}}+\frac{\left|\frac{1}{20}-\frac{5}{20}\frac{5}{20}\right|\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{5}{20}-2\frac{5}{20}\right)+\frac{5}{20}\frac{5}{20}}+$$

$$+\frac{\left(\frac{1}{20}-\frac{1}{20}\frac{10}{20}\right)\cdot 1}{\frac{1}{20}\left(1+2\frac{1}{20}-2\frac{1}{20}-2\frac{10}{20}\right)+\frac{1}{20}\frac{10}{20}}\Bigg]=$$

$$=\frac{1}{20}\left(0+\frac{10}{19}+\frac{10}{11}+\frac{1}{4}+\frac{20}{17}+\frac{10}{37}+1+0+\frac{15}{19}+\frac{5}{27}+1\right)=0{,}183581\approx 0{,}184.$$

$$co=0{,}184. \qquad (3)$$

Пример 4. Вычисляем комбинированный коэффициент детерминации комби-ас по данным таблицы 1. Используем результаты вычислений в примерах 1,2.

$$com_a=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j|x_i-m_X|\left|y_j-m_Y\right|\frac{\left|p_{ij}-p_{i\cdot}p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}p_{ij}=$$

$$=\frac{1}{20\sigma_X\sigma_Y}\Big[|3-5{,}2||30-53|\cdot 0+|3-5{,}2||50-53|\frac{2}{\sqrt{75}}+|3-5{,}2||70-53|\frac{2}{5}+$$

$$+|5-5{,}2||30-53|\frac{1}{3\sqrt{11}}+|5-5{,}2||50-53|\frac{4}{3\sqrt{11}}+|5-5{,}2||7-53|\frac{2}{3\sqrt{33}}+$$

$$+|5-5,2||90-53|\frac{11}{3\sqrt{11\cdot 19}}+0+|7-5,2||50-53|\frac{3}{5\sqrt{3}}+$$

$$+|7-5,2||70-53|\frac{1}{15}+|9-5,2||50-53|\frac{1}{\sqrt{19}}\Big]=$$

$$=\frac{1}{20\cdot 1,661325\cdot 15,968719}\Big[0+2,2\cdot 3\frac{2}{\sqrt{75}}+2,2\cdot 17\frac{2}{5}+0,2\cdot 23\frac{1}{3\sqrt{11}}+$$

$$0,2\cdot 3\frac{4}{3\sqrt{11}}+$$

$$+0,2\cdot 17\frac{2}{3\sqrt{33}}+0,2\cdot 37\frac{11}{3\sqrt{11\cdot 19}}+0+1,8\cdot 3\frac{3}{5\sqrt{3}}+1,8\cdot 17\frac{1}{15}+3,8\cdot 3\frac{1}{\sqrt{19}}\Big]=$$

$=0,048974\approx 0,049.$

$$com_a=0,049. \qquad (4)$$

Пример 5. Вычисляем комбинированный коэффициент детерминации комби-конт по данным таблицы 1. Используем результаты вычислений из примеров 1,3,4.

$$com_c=\frac{1}{\sigma_X\sigma_Y}\sum_i\sum_j|x_i-m_X||y_j-m_Y|\frac{|p_{ij}-p_{i\cdot}p_{\cdot j}|}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}p_{ij}=$$

$$=\frac{1}{20\cdot 1,661325\cdot 15,968719}\Big(0+2,2\cdot 3\frac{10}{19}+2,2\cdot 17\frac{10}{11}+0,2\cdot 23\frac{1}{4}+0,2\cdot 3\frac{20}{17}+$$

$$0,2\cdot 17\frac{10}{37}+0,2\cdot 37\cdot 1+0+1,8\cdot 3\frac{15}{19}+1,8\cdot 17\frac{5}{27}+3,8\cdot 3\cdot 1\Big)=0,130004\approx 0,130.$$

$$com_c=0,130. \qquad (5)$$

Пример 6. Вычисляем предельный коэффициент детерминации по данным таблицы 1.

$$l=\sum_i\sum_j\frac{|p_{ij}-p_{i\cdot}p_{\cdot j}|}{p_{ij}+p_{i\cdot}p_{\cdot j}}p_{ij}=\frac{1}{20}\Big(0+\frac{|40-50|}{40+50}2+\frac{40-25}{40+25}2+\frac{40-36}{40+36}2+$$

$$+\frac{|80-90|}{80+90}4+\frac{|40-45|}{40+45}2+\frac{20-9}{20+9}\cdot 1+0+\frac{60-50}{60+50}3+\frac{|20-25|}{20+25}\cdot 1+\frac{20-10}{20+10}\cdot 1\Big)=$$

$=0,111922\approx 0,112.$

$$l=0,112. \qquad (6)$$

Пример 7. Вычисление ассоциативного коэффициента корреляции по данным табл. 1.

Коэффициент корреляции отличается от коэффициента детерминации тем, что в формуле для этого коэффициента снят знак модуля:

$$\rho_{as} = \sum_i \sum_j \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}}. \qquad (7)$$

Используем результаты вычислений в примере 2.

$$\rho_{as} = \frac{1}{20}(0–0{,}230940+0{,}4+0{,}100504–0{,}402015–$$

$$–0{,}116052+0{,}253629+0+0{,}346410–0{,}066667+0{,}229416)=0{,}025941\approx 0{,}026.$$

$$\rho_{as} = 0{,}026. \qquad (8)$$

Пример 8. Вычисляем контингенциальный коэффициент корреляции ρ_{co}, используя результаты вычислений в примере 3.

$$\rho_{co} = \sum_i \sum_j \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij}. \qquad (9)$$

$$\rho_{co} = \frac{1}{20}\left(0 - \frac{10}{19} + \frac{10}{11} + \frac{1}{4} - \frac{20}{17} - \frac{10}{37} + 1 + 0 + \frac{15}{19} - \frac{5}{27} + 1\right) =$$

$$= \frac{1}{20}(2+0{,}263158+0{,}909091+0{,}25–1{,}176471–0{,}270270–0{,}185185)=$$

$$=0{,}089516\approx 0{,}090.$$

$$\rho_{co} = 0{,}090. \qquad (10)$$

Пример 9. Вычисляем комбинированный коэффициент корреляции ρ_{cas}, используя результаты вычислений в примере 4.

$$\rho_{cas} = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left(x_i - m_X\right)\left(y_j - m_Y\right) \frac{\left|p_{ij} - p_{i\cdot} p_{\cdot j}\right|}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}} p_{ij}. \qquad (11)$$

$$\rho_{cas} = \frac{1}{20 \cdot 1{,}661325 \cdot 15{,}968719}\left[0 + 2{,}2 \cdot 3\frac{2}{\sqrt{75}} - 2{,}2 \cdot 17\frac{2}{5} + 0{,}2 \cdot 23\frac{1}{3\sqrt{11}} + \right.$$

$$+0{,}2 \cdot 3\frac{4}{3\sqrt{11}} –0{,}2 \cdot 17\frac{2}{3\sqrt{33}} - 0{,}2 \cdot 37\frac{11}{3\sqrt{11 \cdot 19}} + 0 –$$

$$\left. –1{,}8 \cdot 3\frac{3}{5\sqrt{3}} + 1{,}8 \cdot 17\frac{1}{15} - 3{,}8 \cdot 3\frac{1}{\sqrt{19}}\right] = \frac{4{,}267731 - 21717382}{530{,}584642} = –0{,}032888 \approx$$

$$\approx –0{,}033.$$

$$\rho_{cas} = –0{,}033. \qquad (12)$$

Пример10. Вычисляем комбинированный контингенциальный коэффициент корреляции

$$\rho_{comc} = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left(x_i - m_X\right)\left(y_j - m_Y\right) \frac{\left|p_{ij} - p_{i\cdot} p_{\cdot j}\right|}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij} \,. \quad (13)$$

Используем результаты вычислений в примере 5.

$$\rho_{comc} = \frac{1}{20 \cdot 1{,}661325 \cdot 15{,}968719}\left(0 + 2{,}2 \cdot 3\frac{10}{19} - 2{,}2 \cdot 17\frac{10}{11} +\right.$$

$$+0{,}2 \cdot 23\frac{1}{4} + 0{,}2 \cdot 3\frac{20}{17} -$$

$$\left. -0{,}2 \cdot 17\frac{10}{37} - 0{,}2 \cdot 37 \cdot 1 + 0 - 1{,}8 \cdot 3\frac{15}{19} + 1{,}8 \cdot 17\frac{5}{27} - 3{,}8 \cdot 3 \cdot 1\right) =$$

$$= \frac{1}{530{,}584642}(0+3{,}473684-34+1{,}15+0{,}705882-\ 0{,}918919 - 7{,}4+0 - 4{,}263158+$$

$$+5{,}666667 - 11{,}4) = \frac{10{,}996233 - 57{,}982077}{530{,}584642} = -\,0{,}088555 \approx -\,0{,}089.$$

$$\rho_{comc} = -0{,}089. \quad (14)$$

В следующем § 9.5 анализируются результаты вычисления коэффициентов связи, полученные в §§ 9.2 – 9.4. Дается их сводка.

§ 9.5. Сводка и анализ результатов вычислений коэффициентов связи в §§ 9.2 – 9.4.

Сводка результатов вычислений из 22 примеров.

1. Коэффициенты связи при строгой функциональной зависимости $Y = X^2$.

$$\rho = 0{,}913; \quad as = 0{,}606; \quad co = 0{,}963;$$

$$com_a = 0{,}582; \quad com_c = 0{,}951; \quad l = 0{,}504.$$

2. Коэффициенты связи при разбросе точек относительно тренда $y = x^2$.

$$\rho = 0{,}850; \quad as = 0{,}378; \quad co = 0{,}801;$$

$$com_a = 0{,}403; \quad com_c = 0{,}871; \quad l = 0{,}361.$$

3. Коэффициенты связи при круговом разбросе точек.

$$\rho = -0{,}100; \quad as = 0{,}107; \quad co = 0{,}184;$$

$$com_a = 0{,}049; \quad com_c = 0{,}130; \quad l = 0{,}112;$$

4. Коэффициенты корреляции при круговом разбросе экспериментальных точек.

$$\rho_{as} = -0,036; \quad \rho_{co} = 0,090; \quad \rho_{cas} = -0,033; \quad \rho_{comc} = -0,089.$$

Анализ результатов.

1. В §§ 9.2 – 9.4 на 22 примерах практически показано как вычисляются новые коэффициенты связи – детерминации и корреляции в случаях, когда в массивах экспериментальных точек для построения регрессии нет повторных наблюдений – для каждой абсциссы одна ордината. Построение группированного статистического ряда и вычисления проводились вручную с помощью микрокалькулятора. Эти вычисления не сложнее, чем для обычного линейного коэффициента корреляции. В дальнейшем для их вычисления могут быть составлены программы и привлечены пакеты программ – excel, mathcad и другие.

2. Все 5 коэффициентов детерминации – as, co, com_a, com_c, l монотонно уменьшаются при переходе от случая 1 строгой нелинейной зависимости к случаю 3 практической независимости в среднем. Каждый коэффициент меняется по своему, но отражает тенденцию уменьшения силы зависимости при увеличении разброса экспериментальных точек относительно тренда или центра рассеяния $\left(m_X; m_Y\right)$.

3. Коэффициент детерминации co в случае строгой зависимости $y = x^2$ должен равняться 1.Здесь он равен $0{,}963$. Отступление от 1 объясняется группировкой.

4. Наиболее близким по характеру изменения к линейному коэффициенту корреляции ρ представляется контингенциальный коэффициент детерминации co, однако он лучше регистрирует зависимость, так как равенство нулю этого коэффициента означает независимость случайных величин, чего не скажешь о коэффициенте ρ.

5. В качестве измерителя силы связи можно употреблять любой из коэффициентов детерминации. Все познается в сравнении с типичными случаями из практики. Однако, чтобы избежать ошибок и более основательно судить о величине зависимости случайных величин лучше исследовать зависимость системно с помощью нескольких коэффициентов детерминации и корреляции.

6. Все коэффициенты корреляции обозначены символом ρ с индексами или без них. Их целесообразно применять системно с коэффициентами детерминации в случае, близком к независимости случайных величин.

7.Для коэффициентов детерминации корректно применять понятие «Сила зависимости»
вместо «Теснота зависимости», которое применяется к линейному коэффициенту корреляции ρ в силу их свойств.

§ 9.6. Выборочные коэффициенты детерминации и корреляции в математической статистике

Введенные в главе 3 коэффициенты детерминации и корреляции – ассоциативный, контингенциальный, предельный, «комби», дефектологический могут применяться в математической статистике, и уже применялись, путем замены вероятностей на относительные частоты, а теоретических функций распределения – на эмпирические функции распределения. Полученные таким образом характеристики будут оценками соответствующих теоретических коэффициентов связи. Они будут состоятельными оценками, так как относительные частоты являются состоятельными оценками вероятностей, а эмпирические функции распределения являются состоятельными оценками теоретических функций распределения. Исследования по этому вопросу приведены в работе Г. Крамера [12].

Среди выборочных оценок можно выделить группу робастных оценок, устойчивых к выбросам. Эти оценки можно построить по следующей схеме.

Рассмотрим оценку $\hat{\delta}$ общего вида

$$\hat{\delta} = \sum_i \sum_j \left| K\left(x_i, y_j\right) \right| \hat{p}_{ij} . \qquad (1)$$

Модуль ядра $K\left(x_i, y_j\right)$ можно рассматривать как коэффициент детерминации между отдельными значениями x_i, y_j случайных величин. Из значений модуля можно составить вариационный ряд и построить из его элементов робастные характеристики – медиану med и полусумму квартилей t_q по известным правилам математической статистики. Эти характеристики действительно робастные, так как не содержат выбросов. Выбросы группируются на концах вариационного ряда, а медиана – средний элемент или полусумма средних элементов. Квартили – элементы, на четверть отстоящие от краев, и поэтому тоже не содержат выбросов. Полученные таким образом характеристики можно назвать соответственно медианными или квартильными. Например, медианный ассоциативный коэффициент детерминации. В § 3.1 такой коэффициент был отмечен, а в § 4.2 - вычислен в примере 3.

Глава 10. Общий оптимизационный метод получения спектра числовых характеристик положения и рассеяния и их оценок для непрерывных распределений

§ 10.1. Сущность оптимизационного метода получения согласованного спектра числовых характеристик положения и рассеяния.

1°. В различных отраслях прикладного знания применяется большое количество числовых характеристик положения и рассеяния случайных величин (с.в.). Все они поразному характеризуют изучаемые распределения и дополняют друг друга, образуя систему. Эта система числовых характеристик в теории вероятностей порождает соответствующую систему выборочных оценок этих характеристик в математической статистике. Каждая характеристика рассеяния имеет смысл по отношению к определенному центру рассеяния. Они попарно связаны.

Для того, чтобы определить подход к построению общего метода получения такого рода характеристик, рассмотрим частную задачу о минимальном свойстве математического ожидания m_X:

Утверждение. Математическое ожидание m_X минимизирует среднее значение квадрата отклонения с.в. X от любой точки η вещественной оси, т.е.

$\mathsf{M}\left[(X-m_X)^2\right] \le \mathsf{M}\left[(X-\eta)^2\right]$.

Применяя свойства математического ожидания, последовательно получаем

$$\mathsf{M}\left[(X-\eta)^2\right] = \mathsf{M}\left[\left((X-m_X)+(m_X-\eta)\right)^2\right] =$$

$$= \mathsf{M}\left[(X-m_X)^2\right] + 2(m_X-\eta)\mathsf{M}\left[X-m_X\right] + (m_X-\eta)^2 = \mathsf{M}\left[(X-m_X)^2\right] + (m_X-\eta)^2,$$

так как $\mathsf{M}\left[X-m_X\right] = m_X - m_X = 0$. В силу того, что $(m_X-\eta)^2 \ge 0$, получаем: $\mathsf{M}\left[(X-\eta)^2\right] \ge \mathsf{M}\left[(X-m_X)^2\right]$. Таким образом,

$$\sigma_X^2 = \mathsf{M}\left[(X-m_X)^2\right] = \min_{\eta} \mathsf{M}\left[(X-\eta)^2\right]. \qquad (1)$$

2°. Основываясь на минимальном свойстве m_X, рассмотрим с.в. X с плотностью $f(x) = F'(x)$. Для определения величины рассеяния с.в. X в общем случае выбирается так называемая функция потерь на ее рассеяние. В качестве такой функции мы выбираем степенную $|u|^\lambda$ $(\lambda \ge 1)$ как наиболее простую среди возможных. Величину

рассеяния с.в. X относительно точки η, $\eta \in \mathbf{R}$, выражаем через функцию потерь по формуле

$$J_\lambda(\eta) = \int_{-\infty}^{\infty} |x-\eta|^\lambda f(x)\,dx\,. \qquad (2)$$

Интеграл в формуле (2) предполагается сходящимся при любом $\eta \in \mathbf{R}$ и рассматриваемом фиксированном λ ($\lambda \geq 1$).

Множество распределений, для которых интеграл в формуле (2) сходится, заведомо не пустое, так как содержит равномерное, показательное, нормальное и другие распределения.

Положим далее

$$\sigma_\lambda = \min_\eta \left[\int_{-\infty}^{\infty} |x-\eta|^\lambda f(x)\,dx \right]^{1/\lambda} \qquad (3)$$

Заметим, что при фиксированном λ минимум степени в формуле (2) достигается при тех же значениях η, что и минимум основания степени.

Теорема 1. $\min_\eta J_\lambda(\eta)$ при фиксированном λ существует (при сделанном предположении о сходимости интеграла).

Пусть (a,b) – интервал, в котором плотность $f(x)$ непрерывна и положительна, $f(x)$ предполагается кусочно-непрерывной, следовательно, такой интервал существует. Тогда

$$J_\lambda(\eta) = \int_{-\infty}^{\infty} |x-\eta|^\lambda f(x)\,dx \geq \int_a^b |x-\eta|^\lambda f(x)\,dx = |\xi-\eta|^\lambda f(\xi)(b-a), \quad a<\xi<b,$$

по теореме о среднем. Отсюда следует, что $J_\lambda(\eta) \to +\infty$ при $\eta \to \pm\infty$.

Далее покажем, что $J_\lambda(\eta)$ непрерывна в любой точке $\eta \in \mathbf{R}$.

Пусть A, B – произвольные числа, $A < \eta < B$. Тогда

$$J_\lambda(\eta) = \int_{-\infty}^{A} |x-\eta|^\lambda f(x)\,dx + \int_A^B |x-\eta|^\lambda f(x)\,dx + \int_B^{\infty} |x-\eta|^\lambda f(x)\,dx\,. \qquad (4)$$

Средний интеграл является собственным и в силу непрерывности подынтегральной функции относительно переменной η также является непрерывной функцией по η. Два крайних несобственных интеграла равномерно сходятся относительно η по признаку Вейерштрасса, так как мажорируются сходящимися интегралами, не зависящими от η: $\int_{-\infty}^{A} (B-x)^\lambda f(x)\,dx$ и $\int_B^{\infty} (x-A)^\lambda f(x)\,dx$ соответственно.

Кроме того, подынтегральные функции в интегралах (4) непрерывны по η. Отсюда следует, что крайние интегралы в (4) являются непрерывными функциями относитель-

но переменной η. Итак, все слагаемые в формуле (4) непрерывны по η, следовательно, непрерывна и их сумма $J_\lambda(\eta)$ в рассматриваемом промежутке (A, B). Так как числа A, B – любые, то это означает непрерывность $J_\lambda(\eta)$ на всей оси. Доказанные два свойства функции $J_\lambda(\eta)$ обеспечивают существование ее глобального минимума $\min\limits_{\eta} J_\lambda(\eta)$.

В одном частном случае можно указать точку глобального минимума.

Теорема 2. Если распределение симметрично относительно точки m, то при $\lambda \geq 1$ в точке m функция $J_\lambda(\eta)$ имеет глобальный минимум.

Без ущерба общности можно считать, что $m = 0$, так как иначе в интеграле (2) выполняем подстановку $x - m = z$ и вводим новый параметр $\eta_1 = \eta - m$. Итак, будем предполагать, что плотность распределения $f(x)$ – четная функция.

Докажем сначала, что функция $J_\lambda(\eta)$ также четная. Для этого преобразуем интеграл (3) к другому виду:

$$J_\lambda(\eta) = \int\limits_{-\infty}^{\infty} |x - \eta|^\lambda f(x)\,dx = \int\limits_{-\infty}^{0} |x - \eta|^\lambda f(x)\,dx + \int\limits_{0}^{\infty} |x - \eta|^\lambda f(x)\,dx .$$

В первом интеграле делаем подстановку $x = -z$:

$$\int\limits_{-\infty}^{0} |x - \eta|^\lambda f(x)\,dx = -\int\limits_{+\infty}^{0} |-z - \eta|^\lambda f(-z)\,dz = \int\limits_{0}^{+\infty} |z + \eta|^\lambda f(z)\,dz .$$

Следовательно,

$$J_\lambda(\eta) = \int\limits_{0}^{\infty} \left[|x - \eta|^\lambda + |x + \eta|^\lambda \right] f(x)\,dx .$$

Из полученного представления функции $J_\lambda(\eta)$ и видна ее четность. Это позволяет ограничиться рассмотрением значений $\eta \geq 0$.

Пусть $\lambda \geq 1$ фиксировано. Рассмотрим функцию

$$\varphi(\eta) = J_\lambda(\eta) - J_\lambda(0) = \int\limits_{0}^{\infty} \left[|x - \eta|^\lambda + (x + \eta)^\lambda - 2x^\lambda \right] f(x)\,dx .$$

Покажем, что $\varphi(\eta) \geq 0$. Обозначим подынтегральную функцию

$$v(\eta, x) = \left[|x - \eta|^\lambda + (x + \eta)^\lambda - 2x^\lambda \right] f(x) .$$

Заметим, что $v(0, x) = 0$. Докажем, что при любом фиксированном x функция $v(\eta, x)$ возрастает в широком смысле.

Пусть $0 \leq \eta \leq x$. Тогда $v(\eta, x) = \left[(x - \eta)^\lambda + (x + \eta)^\lambda - 2x^\lambda \right] f(x)$;

$$\frac{\partial}{\partial \eta} v(\eta, x) = \lambda f(x) \left[(x + \eta)^{\lambda - 1} - (x - \eta)^{\lambda - 1} \right] \geq 0 \quad \text{при} \quad \lambda \geq 1 .$$

Пусть $\eta \ge x$. Тогда $v(\eta, x) = \left[(\eta - x)^{\lambda} + (\eta + x)^{\lambda} - 2x^{\lambda}\right] f(x)$;

$\dfrac{\partial}{\partial \eta} v(\eta, x) = \lambda f(x)\left[(\eta - x)^{\lambda-1} + (\eta + x)^{\lambda-1}\right] \ge 0$ при $\lambda \ge 1$.

Таким образом, при любом фиксированном $x \ge 0$ имеем $\dfrac{\partial v(\eta, x)}{\partial \eta} \ge 0$. Отсюда следует, что $v(\eta, x)$ возрастает в широком смысле при любом фиксированном x. Тогда $v(\eta, x) \ge v(0, x) = 0$. Интеграл по x от неотрицательной функции – неотрицателен, следовательно, $\varphi(\eta) = J_{\lambda}(\eta) - J_{\lambda}(0) \ge 0$. Отсюда $J_{\lambda}(\eta) \ge J_{\lambda}(0)$.

3°. Пусть η_{λ} – любая точка, в которой достигается $\min\limits_{\eta} J_{\lambda}(\eta)$. (Если эта точка не единственна, то для ее выбора можно предъявить дополнительные условия, например, соображения симметрии.)

Определение 1. При заданной функции потерь $|u|^{\lambda}$ за меру рассеяния с.в. X принимается величина

$$\sigma_{\lambda} = \left[J_{\lambda}(\eta_{\lambda})\right]^{1/\lambda}. \qquad (5)$$

Точка η_{λ} называется центром рассеяния с.в. X, согласованным с выбранной мерой рассеяния.

Определение 2. Множество пар $\{(\eta_{\lambda}, \sigma_{\lambda})\}$ при всевозможных λ $(\lambda \ge 1)$ называется спектром согласованных числовых характеристик положения и рассеяния.

Для практики представляют интерес три случая: $\lambda = 1,\ 2,\ \infty$.

Случай $\lambda = 2$ уже рассмотрен в п.1°. Он дает согласованную пару (m_X, σ_X). Рассмотрим случай $\lambda = 1$.

$$J_1(\eta) = \int_{-\infty}^{\infty} |x - \eta| f(x)\,dx = \int_{-\infty}^{\eta} (\eta - x) f(x)\,dx + \int_{\eta}^{\infty} (x - \eta) f(x)\,dx =$$

$$= \eta \int_{-\infty}^{\eta} f(x)\,dx - \int_{-\infty}^{\eta} x f(x)\,dx + \int_{\eta}^{\infty} x f(x)\,dx - \eta \int_{\eta}^{\infty} f(x)\,dx =$$

$$= \eta F_X(\eta) - \int_{-\infty}^{\eta} x f(x)\,dx + \int_{\eta}^{\infty} x f(x)\,dx - \eta\left[1 - F_X(\eta)\right].$$

Отсюда:

$J_1'(\eta) = F_X(\eta) + \eta f_X(\eta) - \eta f_X(\eta) - \eta f_X(\eta) - 1 + F_X(\eta) + \eta f_X(\eta) = 2F_X(\eta) - 1 = 0$;

$F_X(\eta) = 1/2$.

Корнем этого уравнения является медиана $\eta_1 = \mathrm{Me}X$. Тогда

$$\sigma_1 = \int_{-\infty}^{\infty} |x - \mathrm{Me}| f(x)\,dx = \delta$$

– среднее абсолютное отклонение. Получаем вторую согласованную пару (Me, δ).

Замечание. Остальные случаи конечных значений λ сложны для рассмотрения, так как приводят к сложным уравнениям, которые нужно решать численно, учитывая вид распределения. Так, например, при $\lambda = 4$ для нахождения η_4 получаем уравнение $\eta_4^3 - 3\alpha_1\eta_4^2 + 3\alpha_2\eta_4 - \alpha_3 = 0$, где $\alpha_k = \mathsf{M}\left[X^k\right]$, $k = 1, 2, 3$. Решая это уравнение для показательного распределения с плотностью $f(x) = \alpha e^{-\alpha x}$ при $x \geq 0$ и $f(x) = 0$ при $x < 0$, найдем $\eta_4 = 1.5962/\alpha$. Соответственно находим $\sigma_4 = 1.5962/\alpha$.

4°. $\lambda = \infty$. Этот случай рассматривается как предельный при $\lambda \to \infty$. Величину рассеяния в этом случае относительно точки η можно выразить формулой

$$H(\eta) = \lim_{\lambda\to\infty}\left[\int_a^b |x - \eta|^{\lambda} f(x)\,dx\right]^{1/\lambda}. \qquad (6)$$

для усеченного конечного промежутка $[a, b]$, на котором изучается рассеяние случайной величины.

Естественно, что точки a, b должны быть согласованы с распределением. Полагаем их равными симметричным p-квантилям: $a = F_X^{-1}(p) = \zeta_p$; $b = F_X^{-1}(1-p) = \zeta_{1-p}$; $0 < p < 1/2$. Фактически рассматриваемое распределение с плотностью $f(x)$ мы аппроксимируем усеченным распределением на промежутке $[a, b]$ с плотностью $cf(x)$, где c – нормирующий множитель. Вне промежутка $[a, b]$ плотность полагается равной нулю.

Рассматриваемый случай охватывает те распределения (Коши, Стьюдента и другие), для которых интеграл (3) расходится при всех $\lambda \geq 1$ или при некоторых λ.

Теорема 3.

$$H(\eta) = \begin{cases} b - \eta, & -\infty < \eta \leq \dfrac{(a+b)}{2}; \\ \eta - a, & \dfrac{(a+b)}{2} < \eta < +\infty. \end{cases} \qquad (7)$$

Рассмотрим случай $-\infty < \eta \leq (a+b)/2$. Тогда $|x - \eta| \leq |b - \eta|$ при $a \leq x \leq b$ (рис.1).

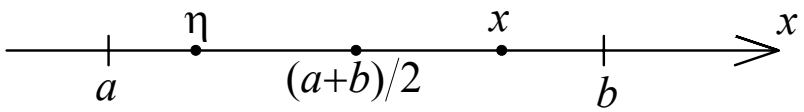


Рис. 1. Иллюстрация неравенства $|x - \eta| \leq |b - \eta|$.

Поэтому $\int_a^b |x - \eta|^{\lambda} f(x)\,dx \leq |b - \eta|^{\lambda} \int_a^b f(x)\,dx \leq |b - \eta|^{\lambda} \int_{-\infty}^{\infty} f(x)\,dx = |b - \eta|^{\lambda}$. Таким образом,

$$\int_a^b |x-\eta|^{\lambda} f(x)\,dx \le |b-\eta|^{\lambda}. \qquad (8)$$

Оценим этот же интеграл снизу. Пусть $\varepsilon > 0$ – достаточно малое. Получаем:

$$\int_a^b |x-\eta|^{\lambda} f(x)\,dx \ge \int_{b-\varepsilon}^b |x-\eta|^{\lambda} f(x)\,dx \ge |b-\varepsilon-\eta|^{\lambda} \int_{b-\varepsilon}^b f(x)\,dx. \qquad (9)$$

На основе неравенств (8) и (9) получаем:

$$|b-\varepsilon-\eta| \left[\int_{b-\varepsilon}^b f(x)\,dx\right]^{1/\lambda} \le \left[\int_a^b |x-\eta|^{\lambda} f(x)\,dx\right]^{1/\lambda} \le |b-\eta|.$$

Перейдем в этих неравенствах к пределу при $\lambda \to +\infty$ и учтем, что $\left[\int_{b-\varepsilon}^b f(x)\,dx\right]^{1/\lambda} \to 1$, предполагая, что $f(x) > 0$ на отрезке $[b-\varepsilon, b]$. Получаем

$|b-\varepsilon-\eta| \le H(\eta) \le |b-\eta|$.

Устремим в этих неравенствах ε к нулю. В пределе находим: $|b-\eta| \le H(\eta) \le |b-\eta|$.

Отсюда: $H(\eta) = |b-\eta| = b-\eta$.

Аналогично рассматривается и второй случай, когда $(a+b)/2 < \eta < +\infty$.

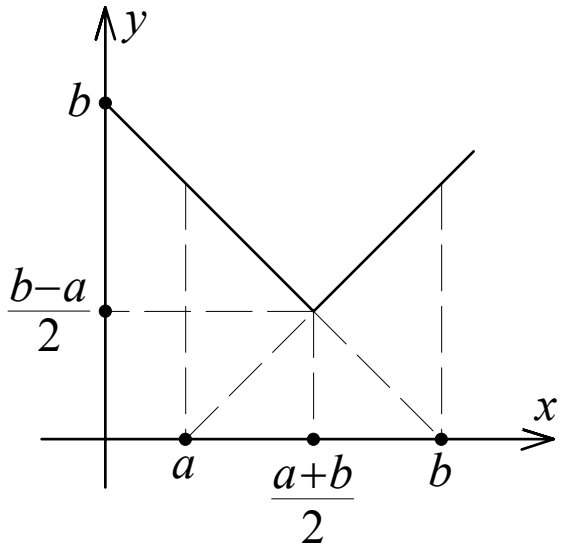


Рис. 2. График кусочно-линейной функции $y = H(\eta)$.

График функции (7) показан на рис.2. Минимум кусочно-линейной функции $H(\eta)$ равен $(b-a)/2$. Он достигается в точке $\eta = (a+b)/2$.

Определение 3. За меру рассеяния с.в. X при степенной функции потерь $|u|^{\lambda}$ в случае $\lambda = \infty$ для промежутка $[\zeta_p, \zeta_{1-p}]$ между симметричными p-квантилями, $0 < p < 1/2$, принимается величина

$$\sigma_{\infty}(p) = \min_{\eta} H(\eta). \qquad (10)$$

За центр рассеяния принимается точка $\eta_{\infty}(p)$, в которой этот минимум достигается.

$$\eta_{\infty}(p) = \frac{\zeta_p + \zeta_{1-p}}{2} \qquad (11)$$

– полусумма симметричных p-квантилей,

$$\sigma_{\infty}(p) = \frac{\zeta_{1-p} - \zeta_p}{2} \qquad (12)$$

– семиинтерквантильная широта.

Эти числовые характеристики удобны в тех случаях, когда для распределения не существуют δ и σ, например, для распределения Коши.

Случай $p = 0.25$ отметим особо. Величина

$$\sigma_\infty(0.25) = \frac{\zeta_{3/4} - \zeta_{1/4}}{2} = w \qquad (13)$$

называется вероятным отклонением (эта характеристика рассеяния принята в теории ошибок и во внешней баллистике).

Вместо него как характеристика рассеяния часто используется величина $2w$ – разность квартилей, называемая интерквартильной широтой:

$$Q = \zeta_{3/4} - \zeta_{1/4}. \qquad (14)$$

Соответствующая характеристика положения – полусумма квартилей

$$\theta_Q = \frac{\zeta_{3/4} + \zeta_{1/4}}{2}. \qquad (15)$$

Итак, случай $\lambda = \infty$ при $p = 1/4$ приводит к следующей паре согласованых числовых характеристик положения и рассеяния: (θ_Q, Q).

5°. Примеры.

Пример 1. Нормальное распределение.

$N(m,\sigma)$; $F_X(x) = \Phi\left(\dfrac{x-m}{\sigma}\right)$.

В этом случае в силу симметрии $m_X = \mathrm{Me} = \theta_Q = m$; $\sigma_X = \sigma$.

$\delta = \int\limits_{-\infty}^{\infty} |x-m| \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{(x-m)^2}{2\sigma^2}\right] dx = \dfrac{2\sigma}{\sqrt{2\pi}} \int\limits_0^\infty z e^{-z^2/2} dz = \dfrac{\sigma\sqrt{2}}{\sqrt{\pi}} \approx 0.7979\sigma$. Здесь произведена замена переменной: $x = m + \sigma z$.

$w = (\zeta_{3/4} - \zeta_{1/4})/2 = \sigma\Phi^{-1}(0.75) \approx 0.6745\sigma$.

$Q = 2w \approx 1.349\sigma$.

Пример 2. Показательное распределение.

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{при } x \ge 0, \\ 0 & \text{при } x \ge 0. \end{cases}$$

$m_X = 1/\lambda$.

Решая уравнение $F_X(\zeta_p) = p$, найдем $\zeta_p = -\dfrac{\ln(1-p)}{\lambda}$. Отсюда

$\mathrm{Me} = \zeta_{1/2} = \dfrac{\ln 2}{\lambda} \approx \dfrac{0.693}{\lambda}$; $\theta_Q = \dfrac{\zeta_{3/4} + \zeta_{1/4}}{2} = -\dfrac{1}{2\lambda}\left(\ln\dfrac{1}{4} + \ln\dfrac{3}{4}\right) \approx \dfrac{0.837}{\lambda}$.

$\sigma_X = \dfrac{1}{\lambda}$; $\delta = \int\limits_0^\infty \left| x - \dfrac{\ln 2}{\lambda} \right| \lambda e^{-\lambda x} dx = \dfrac{\ln 2}{\lambda} \approx \dfrac{0.693}{\lambda}$.

$$w = \frac{1}{2\lambda}\left(-\ln\frac{1}{4} + \ln\frac{3}{4}\right) = \frac{\ln 3}{2\lambda} \approx \frac{0.549}{\lambda};\quad Q = \frac{\ln 3}{\lambda} \approx \frac{1.099}{\lambda}.$$

Пример 3. Равномерное распределение.

$$f_X(x) = \begin{cases} 1/(b-a), & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

В силу симметрии распределения $m_X = \mathrm{Me} = \theta_Q = (a+b)/2$.

$$\sigma_X = \frac{b-a}{2\sqrt{3}};\quad \delta = \int_a^b \left| x - \frac{a+b}{2} \right| \frac{1}{b-a} dx = \frac{b-a}{4}.$$

$$w = \frac{b-a}{4};\quad Q = 2w = \frac{b-a}{2}.$$

Пример 4. Распределение Коши.

$$f_X(x) = \frac{a}{\pi}\frac{1}{a^2 + x^2};\quad F_X(x) = \frac{1}{2} + \frac{1}{\pi}\mathrm{arctg}\frac{x}{a}.$$

В этом случае m_X, σ_X, δ не существуют.

$\mathrm{Me} = \theta_Q = 0$ – центр симметрии.

$$w = a \cdot \mathrm{tg}\frac{\pi}{4} = a;\quad Q = 2w = 2a.$$

§ 10.2. Системный метод точечного оценивания числовых характеристик положения и рассеяния

В § 10.1 с помощью общего метода введены три числовых характеристики положения и три числовых характеристики рассеяния генерального распределения. Напомним их. Математическое ожидание m_X, медиана $\mathrm{Me}X$, полусумма квартилей t_Q являются различного вида средними значениями рассматриваемой случайной величины, а также точками вещественной оси, относительно которых измеряются отклонения значений этой случайной величины и, следовательно, характеризуется ее рассеяние.

Среднее квадратическое отклонение σ_X, среднее абсолютное отклонение δ_X, интерквартильная широта $Q = \xi_{3/4} - \xi_{1/4}$ являются характеристиками рассеяния. Они вместе с характеристиками положения образуют согласованные пары (m, σ), (Me, δ), (t_Q, Q).

Для построения статистических оценок указанных числовых характеристик применяется так называемый метод аналогии, иначе – подстановки, который заключается в следующем. Вместо случайной величины X рассматривается аппроксимирующая ее с.в. X^*, значения которой – элементы выборки, принимаемые равновероятно, т.е. с вероятностью $1/n$.

Числовые характеристики этой дискретной с.в. X^* и есть оценки аналогичных генеральных числовых характеристик. Так появляются согласованные пары $(\overline{x},\sigma)$, (med,d), (t_q,q). К ним добавляется пара $t_R=\dfrac{x_{\max}-x_{\min}}{2}$, $R=x_{\max}-x_{\min}$, для которой нет генерального аналога. Эта пара возникает естественным образом из пары $\left(\dfrac{z_p+z_{1-p}}{2},z_p-z_{1-p}\right)$, являющейся статистическим аналогом пары генеральных характеристик $\left(\dfrac{\zeta_p+\zeta_{1-p}}{2},\zeta_p-\zeta_{1-p}\right)$ – полусуммы и разности квантилей порядка *p*. Действительно, $z_p=x_{([np]+1)}=x_{\max}$ при *p* достаточно близком к 1 и *np* не целом. Соответственно, $z_{1-p}=x_{\min}$.

Все эти выборочные характеристики дополняют и взаимно контролируют друг друга, страхуя от ошибок, аномалий, выбросов, так как обладают различными свойствами.

В этом смысл системного подхода в оценивании. Особенно целесообразен системный подход при исследовании симметричных распределений, когда с естественным центром симметрии совпадают рассмотренные характеристики положения $m=\mathsf{Me}=\theta_Q$. В этом случае все указанные оценки генеральных характеристик положения являются оценками *m*, а все оценки генеральных характеристик рассеяния могут рассматриваться как оценки σ.

Для выполнения этой их роли каждую смещенную оценку *b* для устранения смещения нужно разделить на соответствующий нормирующий коэффициент $k_b(n)=\mathsf{M}[b]$, зависящий от объема выборки *n*.

Для симметричного распределения выборочные характеристики

$$\overline{x},\ \mathsf{med},\ t_q,\ t_R \qquad (1)$$

являются несмещенными оценками математического ожидания *m*, а оценки

$$s,\ d,\ q,\ R \qquad (2)$$

относительно σ смещены.

Соответствующие нормирующие коэффициенты

$k_s(n),\ k_d(n),\ k_q(n),\ k_R(n)$

для нормального распределения при $2\le n\le 20$ для четных *n* содержатся в таблице 2 (Гл. 12). Тогда получаем следующие несмещенные оценки σ:

$$s'=s/k_s(n),\quad d^*=d/k_d(n),\quad q^*=q/k_q(n),\quad R^*=R/k_q(n). \qquad (3)$$

Все восемь отмеченных оценок являются состоятельными. Для оценок (1) это было доказано в [3, раздел4, гл. 1, §7 и гл. 2, §2]. В случае нормального распределения для оценки *s* имеются представления $\mathsf{M}s=\sigma+u_n$; $\mathsf{D}s=v_n$; $u_n\to 0$, $v_n\to 0$ при $n\to\infty$ [14, с. 387]. Тогда состоятельность *s*, а, следовательно, и s', следует из теоремы 1, §2.

Состоятельность d как оценки σ следует из того, что она является оценкой максимального правдоподобия [3, раздел 4, гл. 1,§4). Состоятельность q доказана в [3, раздел 4, гл. 1, , §5, пример 2]. Состоятельность R^* доказана в [14, с. 412]. Состоятельность d^* как оценки σ доказана ниже в § 10.3.

У всех оценок (1), (2) различные дисперсии, и следовательно, различная относительная эффективность.

В таблице 1 представлены дисперсии оценок центра наиболее важного в инженерных вопросах нормального распределения с $\sigma = 1$ для объемов выборки n от 2 до 20. На основе этой таблицы заключаем, что при $n > 4$

$$\mathsf{D}\bar{x} < \mathsf{D}t_q < \mathsf{D}\,\mathrm{med} < \mathsf{D}t_R . \qquad (4)$$

Указанное ранжирование оценок по величине дисперсии определяет и их эффективность. Относительная эффективность оценок (1) в сравнении с $\bar{x}$ представлена в таблице 3. Относительная эффективность падает с ростом n, довольно значительно для выборочной медианы med и полусуммы крайних элементов t_R. Для полусуммы квартилей t_q она остается сравнительно высокой: $0.82 - 0.84$, что и говорит о целесообразности применения этой оценки при $4 \le n \le 20$. Отмеченное падение относительных эффективностей оценок ограничивает и диапазон их применения малым объемом n выборки от 4 до 20 элементов.

В таблице 2 представлены дисперсии четырех несмещенных оценок σ для случая нормального распределения. На основе этой таблицы заключаем, что при $n > 6$

$$\mathsf{D}s' < \mathsf{D}d^* < \mathsf{D}R^* < \mathsf{D}q^* . \qquad (5)$$

С помощью таблицы 2 составлена таблица 4 относительных эффективностей оценок (3) в сравнении с s'. Из их анализа делаем вывод, что применение оценок R^*, q^* при $n > 20$ вряд ли целесообразно из-за низкой относительной эффективности.

Оценки med, t_q, t_R, q^*, R^* принято называть "**быстрыми**", так как при очень малом объеме выборки ($n \le 10$) они дают результат при незначительных вычислениях, и в то же время они достаточно эффективны.

Отметим, наконец, что оценки med, t_q, q^* являются устойчивыми к появлению выбросов и аномальных данных. Оценка d^*, хотя и реагирует на такие данные, но меньше, чем $\bar{x}$, s', R^*, так как в своем составе содержит медиану. Подобные оценки называются **робастными**.

Резюмируя, скажем, что в работе с малыми выборками при $n \le 20$ целесообразен системный метод точечного оценивания, особенно в автоматическом режиме, когда применяются все 8 оценок (1), (3) или их часть. Если оценки положения или соответственно оценки рассеяния – взаимно близкие, то это свидетельствует об их близости к оцениваемой числовой характеристике. Значительные расхождения в полученных числах

свидетельствуют о неоднородности выборки, о наличии аномальных наблюдений. Аномальные наблюдения следует подвергнуть дополнительному анализу.

Таблица 1. Дисперсии оценок центра нормального распределения с $\sigma = 1$. (Составлена Максимовым Ю.Д.)

Объем выборки n	2	4	6	8	10	12	14	16	18	20
$D\bar{x}$	0.5000	0.2500	0.1667	0.1250	0.1000	0.0833	0.0714	0.0625	0.0556	0.0500
Dt_q	—	0.2982	0.1928	0.1513	0.1190	0.1015	0.0860	0.0764	0.0674	0.0613
Dmed	0.5000	0.2982	0.2148	0.1682	0.1384	0.1175	0.1022	0.0904	0.0810	0.0734
Dt_R	0.5000	0.2982	0.2361	0.2049	0.1855	0.1721	0.1622	0.1544	0.1482	0.1430

Таблица 2. Дисперсии несмещенных оценок σ для случая нормального распределения. (Составлена Максимовым Ю.Д.)

Объем выборки n	2	4	6	8	10	12	14	16	18	20
$D\left[s'/\sigma\right]$	0.5708	0.1781	0.1045	0.0738	0.0570	0.0464	0.0392	0.0339	0.0298	0.0267
$D\left[d^*/\sigma\right]$	0.5708	0.1952	0.1157	0.0822	0.0637	0.0521	0.0440	0.0381	0.0336	0.0301
$D\left[R^*/\sigma\right]$	0.5708	0.1826	0.1120	0.0829	0.0671	0.0571	0.0502	0.0451	0.0412	0.0381
$D\left[q^*/\sigma\right]$	—	0.1826	0.2108	0.1213	0.1301	0.0899	0.0941	0.0713	0.0737	0.0589

Таблица 3. Относительные эффективности оценок центра нормального распределения (в сравнении с $\bar{x}$).

Объем выборки n	2	4	6	8	10	12	14	16	18	20
$D\bar{x}/Dt_q$	—	0.8384	0.8646	0.8262	0.8403	0.8207	0.8302	0.8181	0.8249	0.8157
$D\bar{x}/$Dmed	1.0000	0.8384	0.7761	0.7432	0.7225	0.7089	0.6986	0.6914	0.6864	0.6812
$D\bar{x}/Dt_R$	1.0000	0.8384	0.7061	0.6101	0.5391	0.4840	0.4402	0.4048	0.3752	0.3497

Таблица 4. Относительные эффективности несмещенных оценок σ в случае нормального распределения (в сравнении с s').

Объем выборки n	2	4	6	8	10	12	14	16	18	20
Ds'/Dd^*	1.0000	0.9124	0.9032	0.8976	0.8946	0.8919	0.8896	0.8887	0.8870	0.8863
Ds'/DR^*	1.0000	0.9754	0.9330	0.8900	0.8499	0.8136	0.7809	0.7513	0.7246	0.7002
Ds'/Dq^*	—	0.9754	0.4957	0.6082	0.4382	0.5167	0.4164	0.4754	0.4050	0.4522

Пример. Имеются 3 выборки по 10 элементов из нормальной генеральной совокупности с параметрами $m = 0$ и $\sigma = 1$ ([3], первые 3 строки таблицы XVI приложения, содержащей нормально распределенные случайные числа). По выборкам построены 3 вариационных ряда:

n	1	2	3	4	5	6	7	8	9	10
1	–0.957	–0.323	–0.288	–0.068	0.137	0.241	0.296	0.464	1.298	2.455
2	–2.526	–1.558	–1.190	–0.531	–0.194	0.022	0.060	0.187	0.525	0.543
3	–1.865	–0.963	–0.853	–0.634	–0.354	0.697	0.785	0.926	1.375	1.486

Требуется для каждой выборки построить 4 оценки m и 4 оценки σ (1), (3), произвести анализ их близости к оцениваемым характеристикам, сопоставить результаты с рядами ранжированных дисперсий (4), (5).

Вычисляем оценки (1), (3) и заносим их в следующую таблицу:

n	$\overline{x}$	med	t_q	t_R	s'	d^*	q^*	R^*
1	0.33	0.19	0.09	0.75	0.93	0.85	0.57	1.11
2	–0.47	–0.09	–0.50	–0.99	0.97	0.99	1.05	1.00
3	0.06	0.17	0.04	–0.19	1.11	1.34	1.36	1.09

Каждая из четырех оценок положения действительно колеблется около нуля, допуская значительную вариабельность, особенно для оценки t_R .

Каждая из четырех оценок рассеяния в трехкратном повторении колеблется около 1. Вариабельность колебаний незначительна.

Имеющаяся объединенная выборка в 30 элементов – ни малая, ни большая. Для ее обработки поэтому не применим описанный выше метод обработки малых выборок. С другой стороны, к ней не применимы асимптотические методы обработки больших выборок. Выход – в разделении этой средней выборки на 3 малых, а затем усреднение результатов. Средние арифметические по каждой из восьми оценок представлены ниже:

$\bar{x}_{\text{ср}}$	$\text{med}_{\text{ср}}$	$t_{q\,\text{ср}}$	$t_{R\,\text{ср}}$	$s'_{\text{ср}}$	$d^*_{\text{ср}}$	$q^*_{\text{ср}}$	$R^*_{\text{ср}}$
–0.03	0.09	–0.12	–0.14	1.00	1.06	0.99	1.07

Усредненные оценки близки и взаимно, и к оцениваемым характеристикам соответственно: $m = 0$ и $\sigma = 1$.

Произведем ранжировку оценок положения по величине модуля:

$$\left|\bar{x}_{\text{ср}}\right| < \left|\text{med}_{\text{ср}}\right| < \left|t_{q\,\text{ср}}\right| < \left|t_{R\,\text{ср}}\right|.$$

Эта ранжировка близка к теоретической (4). Ранжировка оценок рассеяния производится по величине модуля разности оценки и единицы:

$$\left|s'_{\text{ср}} - 1\right| < \left|q^*_{\text{ср}} - 1\right| < \left|d^*_{\text{ср}} - 1\right| < \left|R^*_{\text{ср}} - 1\right|.$$

Она несколько отличается от теоретической (5). Отличия объясняются малым числом (3) усредненных вариантов.

§ 10.3. Состоятельность нормированного среднего абсолютного отклонения d^* как оценки среднего квадратического отклонения σ для нормального генерального распределения

Наряду со средним абсолютным отклонением от выборочной медианы *med*

$$d = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - med\right| \qquad (1)$$

рассмотрим среднее абсолютное отклонение от генерального математического ожидания m

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - m\right| \qquad (2)$$

и их нормированные варианты

$$d^* = \frac{d}{M\left[d/\sigma\right]} = \frac{\sigma d}{M\left[\bar{d}\right]}; \quad \bar{d}^* = \frac{\bar{d}}{M\left[\bar{d}/\sigma\right]} = \frac{\sigma d}{M\left[\bar{d}\right]}, \qquad (3)$$

которые являются несмещенными оценками σ.

Докажем состоятельность d^* и $\bar{d}^*$ как оценок σ в случае нормальности генерального распределения.

1°. Вычислим сначала $M\bar{d}$ для нормального закона

$$M\bar{d} = M\left[\frac{1}{n}\sum_{i=1}^{n}|x_i - m|\right] = \frac{1}{n}\sum_{i=1}^{n} M|x_i - m| = \frac{1}{n} nM|X - m| = M|X - m| =$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}|x-m|e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \left[\begin{matrix}(x-m)/\sigma = t\\ dx = \sigma dt\end{matrix}\right] = \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|t|e^{-\frac{t^2}{2}}dt = -\frac{2\sigma}{\sqrt{2\pi}}\int_{0}^{\infty}de^{-\frac{t^2}{2}} = \sigma\sqrt{\frac{2}{\pi}}.$$

Итак,

$$M\bar{d} = \sigma\sqrt{2/\pi}. \qquad (4)$$

Далее, аналогично с помощью той же подстановки и по частям получаем:

$$M(x-m)^2 = \frac{1}{\sigma\sqrt{2\pi}}\int(x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}}dx = \sigma^2.$$

Используя эти два интеграла, находим

$$D\bar{d} = \frac{1}{n^2}\sum D|x_i - m| = \frac{1}{n}D|X-m| = \frac{1}{n}\left[M(X-m)^2 - \left(M|X-m|\right)^2\right] =$$

$$= \frac{1}{n}\left(\sigma^2 - \sigma^2\frac{2}{\pi}\right) = \frac{\sigma^2}{n}\left(1 - \frac{2}{\pi}\right) \xrightarrow[n\to\infty]{} 0.$$

2°. Применим неравенство Чебышева :

$$P\left(\left|\bar{d} - M\bar{d}\right| \ge \varepsilon\right) \le \frac{D\bar{d}}{\varepsilon^2}; \quad \forall\varepsilon > 0.$$

Этот результат означает, что

$$\bar{d} \xrightarrow[n\to\infty]{P} M\bar{d} = \sigma\sqrt{2\pi}. \qquad (5)$$

Тогда

$$\bar{d}^* = \sigma\frac{\bar{d}}{M\bar{d}} \xrightarrow[n\to\infty]{P} \sigma\frac{M\bar{d}}{M\bar{d}} = \sigma.$$

Этот результат доказывает состоятельность $\bar{d}^*$ как оценки σ.

3°. Рассмотрим теперь среднее абсолютное отклонение от выборочной медианы

$$d=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-med\right|=\frac{1}{n}\sum\left|\left(x_i-m\right)+\left(m-med\right)\right|\le\frac{1}{n}\sum_{i=1}^{n}\left(\left|x_i-m\right|+\left|m-med\right|\right)=$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-m\right|+\left|m-med\right|=\overline{d}+\left|m-med\right|.$$

Аналогично: $d\ge\overline{d}-\left|m-med\right|$. Итак,

$$\overline{d}-\left|m-med\right|\le d\le\overline{d}+\left|m-med\right|. \qquad (6)$$

Известно [3,с.428] , что выборочная медиана является состоятельной оценкой генеральной медианы: $med\xrightarrow[n\to\infty]{P}Me=m.$ Тогда $\left|m-med\right|\xrightarrow[n\to\infty]{P}0.$ Учитывая формулу (5), из неравенств (6) получаем, что

$$d\xrightarrow[n\to\infty]{P}\sigma\sqrt{\frac{2}{\pi}}. \qquad (7)$$

Здесь и в предыдущем пункте использованы известные из математического анализа теоремы о пределах, которые распространяются и на предел по вероятности.

4°. Если $X\ge0$, то и $MX\ge0.$

Тогда при $X\ge Y$ имеем $X-Y\ge0;\quad M\left[X-Y\right]\ge0;\ MX\ge MY.$

Применив ко всем членам неравенств (6) оператор математического ожидания, получаем:

$$M\overline{d}-M\left|m-med\right|\le Md\le M\overline{d}+M\left|m-med\right|. \qquad (8)$$

Покажем, что $M\left|m-med\right|\xrightarrow[n\to\infty]{}0.$ Для этого воспользуемся асимптотическим распределением выборочной медианы для нормального генерального распределения [3, с.428]. Асимптотически она распределена нормально по закону $N\left(m,\sigma\sqrt{\frac{\pi}{2n}}\right).$

Тогда при больших n имеем:

$$M\left|m-med\right|=\frac{\sqrt{n}}{\pi\sigma}\int_{-\infty}^{\infty}\left|x\right|e^{-\frac{nx^2}{\pi\sigma^2}}dx=-\frac{\sigma}{\sqrt{n}}\int_{0}^{\infty}de^{-\frac{nx^2}{\pi\sigma^2}}=\frac{\sigma}{\sqrt{n}}\xrightarrow[n\to\infty]{}0.$$

На основе полученного результата, формул (4) и (8) заключаем, что

$$Md\xrightarrow[n\to\infty]{}\sigma\sqrt{\frac{2}{\pi}}. \qquad (9)$$

5°. Рассмотрим теперь d^{*}.

$$d^* = \sigma \frac{d}{Md} \xrightarrow[n\to\infty]{P} \sigma \frac{\sigma\sqrt{2/\pi}}{\sigma\sqrt{2/\pi}} = \sigma.$$

Итак,

$$d^* \xrightarrow[n\to\infty]{P} \sigma, \qquad (10)$$

что и означает состоятельность d^* как оценки σ.

Замечание. В процессе доказательства для оценивания σ введена еще одна выборочная характеристика $\overline{d}^*$. Она менее эффективна, чем d^*, и ее применение на практике затруднительно, так как в реальных ситуациях m обычно неизвестно.

Глава 11. Комбинированные коэффициенты детерминации «мед-ас» и «мед-конт» на основе медианы

В главе 10 исследована роль медианы как центра распределения, согласованного с мерой рассеяния, называемой средним абсолютным отклонением. В главе 11 изучаются комбинированные коэффициенты детерминации, отличающиеся от тех, которые рассмотрены в главе 7, тем, что математические ожидания заменены медианами, а средние квадратические отклонения заменены средними абсолютными отклонениями.

§11.1. Комбинированные коэффициенты детерминации «мед-ас» и «мед-конт» для дискретных случайных величин.

Комбинированный коэффициент детерминации сконструирован на основе линейного коэффициента корреляции К. Пирсона заменой математических ожиданий на медианы и внесением в его структуру ассоциативного или контингенциального ядра, как гена, улучшающего свойства коэффициента.

Комбинированный коэффициент детерминации на основе медианы и ассоциативного ядра для дискретных случайных величин определяется формулой

$$mas = \frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \sqrt{\left|x_i - me_X\right|\left|y_j - me_Y\right|}\left|K_{ij}\right| p_{ij} . \qquad (1)$$

Здесь K_{ij} – ассоциативное ядро, определяемое формулой

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}} . \qquad (2)$$

$p_{ij} = P\left(X = x_i, Y = y_j\right); \quad p_{i\cdot} = P\left(X = x_i\right); \quad p_{\cdot j} = P\left(Y = y_j\right).$

$me_X -$ медиана случайной величины X, $me_Y -$ медиана случайной величины Y.

$d_X = \sum_i \left|x_i - me_X\right| p_{i\cdot}$ – среднее абсолютное отклонение случайной величины X,

$d_Y = \sum_j \left|y_j - me_Y\right| p_{\cdot j}$ – среднее абсолютное отклонение случайной величины Y.

Для дискретной случайной величины X медиана me_X, если не совпадает со значением случайной величины, определяется неоднозначно. Поэтому условимся вычислять ее на основе пропорционального деления промежутка, в котором она находится между значением x_i с вероятностью p_i и значением x_{i+1} с вероятностью p_{i+1}:

$$me_X = \frac{x_i p_i + x_{i+1} p_{i+1}}{p_i + p_{i+1}} .$$

Аналогично определяется комбинированный коэффициент детерминации для дискретных случайных величин на основе медианы и контингенциального ядра

$$mco = \frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \sqrt{\left|x_i - me_X\right|\left|y_j - me_Y\right|}\left|K_{ij}\right| p_{ij} \ . \qquad (3)$$

Здесь K_{ij} – контингенциальное ядро, определяемое формулой

$$K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} \qquad (4)$$

Свойства коэффициентов *mas* и *mco*.

Так как свойства *mas* и *mco*. одинаковы, то доказательство этих свойств проведем одновременно для обоих коэффициентов, обозначив их единым символом δ.

1. Нормировка: $0 \le \delta \le 1$.

Доказательство. Используем свойство ядер $\left|K_{ij}\right| \le 1$ для любых i, j. Тогда получаем неравенство $\delta \le \frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \sqrt{\left|x_i - me_X\right|\left|y_j - me_Y\right|} p_{ij}$.

Далее применяем неравенство Коши-Буняковского. Получаем

$$\frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \left[\sqrt{\left|x_i - me_X\right| p_{ij}}\right]\left[\sqrt{\left|y_j - me_Y\right| p_{ij}}\right] \le$$

$$\le \frac{1}{\sqrt{d_X d_Y}} \sqrt{\sum_i \sum_j \left|x_i - me_X\right| p_{ij}} \sqrt{\sum_i \sum_j \left|y_j - me_Y\right| p_{ij}} =$$

$$= \frac{1}{\sqrt{d_X d_Y}} \sqrt{\sum_i \left|x_i - me_X\right| \sum_j p_{ij}} \sqrt{\sum_j \left|y_j - me_Y\right| \sum_i p_{ij}} =$$

$$= \frac{1}{\sqrt{d_X d_Y}} \sqrt{\sum_i \left|x_i - me_X\right| p_{i\cdot}} \sqrt{\sum_j \left|y_j - me_Y\right| p_{\cdot j}} = \frac{1}{\sqrt{d_X d_Y}} \sqrt{d_X d_Y} = 1.$$

Применены формулы согласованности $\sum_j p_{ij} = p_{i\cdot}; \sum_i p_{ij} = p_{\cdot j}$.

Итак, получили: $\delta \le 1$. Неравенство $\delta \ge 0$ очевидно, так как все слагаемые под знаком суммы неотрицательны.

2. Коэффициент детерминации δ обращается в нуль тогда и только тогда, когда случайные величины X, Y независимы.

Доказательство. В составе ядер их числитель равен разности $p_{ij} - p_{i\cdot} p_{\cdot j}$. Он обращается в нуль при всех i, j тогда и только тогда, когда случайные величины X, Y независимы.

Отсюда следует справедливость свойства 2.

3. Если между случайными величинами X, Y имеется линейная зависимость: $Y = aX + b$, то коэффициент детерминации $\delta = 1$.

Доказательство. В этом случае

$$p_{\cdot j} = P\left(Y = y_j\right) = P\left(aX + b = ax_j + b\right) = P\left(X = x_j\right) = p_{j\cdot}.$$

$$p_{ij} = P\left(X = x_i, Y = y_j\right) = P\left(X = x_i, aX + b = ax_j + b\right) = P\left(X = x_i, X = x_j\right) =$$

$$= \begin{cases} p_{i\cdot}; & i = j \\ 0; & i \neq j \end{cases}.$$

Тогда контингенциальное ядро (4) принимает вид

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot}p_{\cdot j}} = \frac{p_{i\cdot} - p_{i\cdot}p_{j\cdot}}{p_{i\cdot}\left(1 + 2p_{i\cdot} - 2p_{i\cdot} - 2p_{j\cdot}\right) + p_{i\cdot}p_{j\cdot}} =$$

$$= \frac{1 - p_{j\cdot}}{1 - p_{j\cdot}} = \frac{1 - p_{i\cdot}}{1 - p_{i\cdot}} = 1 \text{ при } i = j;$$

$$K_{ij} = \frac{-p_{i\cdot}p_{j\cdot}}{p_{i\cdot}p_{j\cdot}} = -1 \text{ при } i \neq j.$$

Аналогично и для ассоциативного ядра получаем

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right)p_{\cdot j}\left(1 - p_{\cdot j}\right)}} = \frac{p_{i\cdot} - p_{i\cdot}p_{i\cdot}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right)p_{i\cdot}\left(1 - p_{i\cdot}\right)}} = \frac{p_{i\cdot}\left(1 - p_{i\cdot}\right)}{p_{i\cdot}\left(1 - p_{i\cdot}\right)} = 1 \text{ при}$$

$i = j$;

$$K_{ij} = \frac{-p_{i\cdot}p_{j\cdot}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right)p_{j\cdot}\left(1 - p_{j\cdot}\right)}} = \frac{-\sqrt{p_{i\cdot}p_{j\cdot}}}{\sqrt{\left(1 - p_{i\cdot}\right)\left(1 - p_{j\cdot}\right)}} \text{ при } i \neq j.$$

В суммах (1) и (3) останутся слагаемые только при $i = j$, так как $p_{ij} = 0$ при $i \neq j$.

Тогда

$$\delta = \frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \sqrt{\left|x_i - me_X\right|\left|y_j - me_Y\right|}p_{ij} =$$

$$= \frac{1}{\sqrt{d_X \left|a\right| d_X}} \sum_i \sqrt{\left|x_i - me_X\right|\left|ax_i + b - ame_X - b\right|}p_{i\cdot} =$$

$$= \frac{1}{\sqrt{|a|}d_X}\sum_i \sqrt{|a|}\,|x_i - me_X|p_{i\cdot} = \frac{\sqrt{|a|}d_X}{\sqrt{|a|}d_X} = 1.$$

Здесь использован тот факт, что

$$d_Y = \sum_i |ax_i + b - ame_X - b|\,p_{i\cdot} = |a|\sum_i |x_i - me_X|\,p_{i\cdot} = |a|d_X.$$

4. Если коэффициент детерминации δ (mas или mco) равен 1, то случайные величины линейно зависимы.

Доказательство.

Доказательство разобьем на несколько этапов.

4.1. Пусть

$$mco = \frac{1}{\sqrt{d_X d_Y}}\sum_i\sum_j \sqrt{|x_i - me_X||y_j - me_Y|}\,|K_{ij}|\,p_{ij} = 1. \qquad (5)$$

Тогда объединенный множитель при вероятности p_{ij} в каждом слагаемом суммы формулы (5) равен 1:

$$A_{ij} = \frac{\sqrt{|x_i - me_X|}}{\sqrt{d_X}}\frac{\sqrt{|y_j - me_Y|}}{\sqrt{d_Y}}|K_{ij}| = 1; \quad \forall i, j. \qquad (6)$$

Действительно, предположим противное, что для какой-либо пары (l, n) имеет место неравенство (для простоты для одной пары)

$$A_{\ln} = \frac{\sqrt{|x_l - me_X|}}{\sqrt{d_X}}\frac{\sqrt{|y_n - me_Y|}}{\sqrt{d_Y}}|K_{\ln}| < 1. \qquad (7)$$

(Неравенство противоположного смысла быть не может, так как $com \le 1$).

Рассмотрим систему соотношений

$$\begin{cases} A_{ij} = 1; \quad i, j = 1, 2, ...; i \ne l, j \ne n; \\ A_{\ln} < 1 \end{cases}. \qquad (8)$$

Умножим каждое соотношение системы (8) на соответствующую вероятность p_{ij} и просуммируем по всем i, j. Получаем

$\sum_{i\ne l}\sum_{j\ne n} A_{ij}p_{ij} + A_{\ln}p_{\ln} < \sum_{i\ne l}\sum_{j\ne n} p_{ij} + p_{\ln} = 1$. Получаем: $mco < 1$. Это неравенство противоречит исходному равенству (5). Противоречие доказывает, что $A_{ij} = 1; \forall i, j.$

4.2. Докажем теперь, что $|K_{ij}| = 1; \forall i, j.$

Доказываем от противного, пусть для какой-либо пары индексов $i = l, j = n$ имеет место неравенство

$$\left|K_{\ln}\right| < 1. \tag{9}$$

(Для простоты для одной пары). Заметим, что $\left|K_{\ln}\right| > 1$ быть не может, так как имеет место общее неравенство $\left|K_{ij}\right| \le 1$. Тогда в силу доказанного равенства (6) имеем неравенство

$$B_{\ln} = \frac{\sqrt{\left|x_l - me_X\right|}}{\sqrt{d_X}} \frac{\sqrt{\left|y_n - me_n\right|}}{\sqrt{d_Y}} > 1 \tag{10}.$$

Запишем систему соотношений

$$\begin{cases} B_{ij} = \dfrac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} \dfrac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} = 1; \quad \left(i, j = 1,2,...; i \ne l, j \ne n\right) \\ B_{\ln} > 1 \end{cases}. \tag{11}$$

Умножим каждое соотношение (11) на соответствующую вероятность p_{ij} и просуммируем по всем i, j. Получаем

$\sum\limits_{i \ne l} \sum\limits_{j \ne n} B_{ij} p_{ij} + B_{\ln} p_{\ln} > \sum\limits_{i \ne l} \sum\limits_{j \ne n} p_{ij} + p_{\ln} = 1$. Таким образом, имеем

$$\sum_i \sum_j \frac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} \frac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} p_{ij} > 1. \tag{12}$$

С другой стороны, в силу неравенства Коши-Буняковского, получаем другое соотношение:

$$\frac{1}{\sqrt{d_X}\sqrt{d_Y}} \sum_i \sum_j \left(\sqrt{\left|x_i - me_X\right|}\sqrt{p_{ij}}\right)\left(\sqrt{\left|y_j - me_Y\right|}\sqrt{p_{ij}}\right) \le$$

$$\le \frac{1}{\sqrt{d_X}\sqrt{d_Y}} \sqrt{\sum_i \left|x_i - m_X\right| \sum_j p_{ij}} \sqrt{\sum_j \left|y_j - me_Y\right| \sum_i p_{ij}} =$$

$$= \frac{1}{\sqrt{d_X}\sqrt{d_Y}} \sqrt{\sum_i \left|x_i - me_X\right| p_{i\cdot}} \sqrt{\sum_j \left|y_j - me_Y\right| p_{\cdot j}} = \frac{1}{\sqrt{d_X}\sqrt{d_Y}} \sqrt{d_X}\sqrt{d_Y} = 1.$$

Итак,

$$\sum_i \sum_j \frac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} \frac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} p_{ij} \le 1. \tag{13}$$

Это неравенство противоречит неравенству (12). Противоречие доказывает, что

$$\left|K_{ij}\right| = 1; \quad \forall i, j. \tag{14}$$

При этом одновременно имеют место равенства

$$B_{ij} = \frac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} \frac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} = 1; \quad \forall i, j. \tag{15}$$

4.3. Докажем теперь, что случайные величины X, Y линейно зависимы.

Для этого умножим каждое равенство (15) на соответствующую вероятность p_{ij} и просуммируем по всем i, j. Получаем

$$\sum_i \sum_j B_{ij} p_{ij} = \sum_i \sum_j p_{ij} = 1. \tag{16}$$

Наряду со случайными величинами X, Y рассмотрим центрированные и нормированные случайные величины

$$X' = \frac{\sqrt{\left|X - m_X\right|}}{\sqrt{d_X}}, \quad Y' = \frac{\sqrt{\left|Y - m_Y\right|}}{\sqrt{d_Y}} \tag{17}.$$

Далее получаем

$M\left(X'\right)^2 = \frac{1}{d_X} \sum_i \left|x_i - me_X\right| p_{i\cdot} = \frac{d_X}{d_X} = 1$. Аналогично $M\left(Y'\right)^2 = 1$. Итак,

$$M\left(X'\right)^2 = 1; \quad M\left(Y'\right)^2 = 1. \tag{18}$$

$$M\left(X'Y'\right) = \sum_i \sum_j x'_i y'_j p_{ij} = \sum_i \sum_j \frac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} \frac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} p_{ij} = 1. \tag{19}$$

Рассмотрим

$$M\left[\left(X' - Y'\right)^2\right] = M\left(X'\right)^2 - 2M\left(X'Y'\right) + M\left(Y'\right)^2 = 1 - 2 \cdot 1 + 1 = 0.$$

Это равенство запишем подробнее

$$\sum_i \sum_j \left(\frac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} - \frac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} \right)^2 p_{ij} = 0. \tag{20}$$

Так как все слагаемые неотрицательны, то равенство нулю суммы означает равенство нулю каждого слагаемого. Вероятности $p_{ij} \neq 0$, поэтому имеют место равенства

$$\frac{\sqrt{\left|x_i - me_X\right|}}{\sqrt{d_X}} - \frac{\sqrt{\left|y_j - me_Y\right|}}{\sqrt{d_Y}} = 0; \quad \forall i, j.$$ Отсюда

$$\frac{|y_j - me_Y|}{d_Y} = \frac{|x_i - me_X|}{d_X}; \quad \forall i, j. \tag{21}$$

Зафиксируем в этом равенстве индекс j и будем менять индекс i. Получим серию противоречивых равенств. Равенства (21) возможны только при $j = i$. Это означает, что двумерное распределение в этом случае вырождается в одномерное и равенства (21) принимают вид

$$\frac{|y_i - me_Y|}{d_Y} = \frac{|x_i - me_X|}{d_X}; \quad i = 1, 2, \ldots \tag{22}$$

Отсюда

$$y_i - me_Y = \pm \frac{d_Y}{d_X}(x_i - me_X); \quad \text{и далее}$$

$$y_i = me_Y \pm \frac{d_Y}{d_X}(x_i - me_X); \quad i = 1, 2, \ldots \tag{23}$$

Равенства (23) означают, что случайная величина Y является линейной функцией случайной величины X:

$$Y = me_Y - \frac{d_Y}{d_X} me_X + \frac{\sigma_Y}{\sigma_X} X = a_1 X + b_1 \qquad \text{или}$$

$$Y = me_Y + \frac{d_Y}{d_X} me_X - \frac{d_Y}{d_X} X = a_2 X + b_2 .$$

Свойство доказано.

Примеры вычисления детерминационных коэффициентов мед-ас и мед-конт приведены в следующем параграфе.

§ 11.2. Примеры вычисления детерминационных коэффициентов «мед-ас» и «мед-конт» для дискретных случайных величин

Пример 1. Вычисление коэффициента мед-ас для триномиального распределения.

Таблица 1 триномиального распределения при $n = 2;\ p_1 = p_2 = 0,25$.

$X \setminus Y$	0	1	2	$p_{i\cdot}$
0	$p_{00} = 1/4$	$p_{01} = 1/4$	$p_{02} = 1/16$	$p_{0\cdot} = 9/16$
1	$p_{10} = 1/4$	$p_{11} = 1/8$	$p_{12} = 0$	$p_{1\cdot} = 3/8$
2	$p_{20} = 1/16$	$p_{21} = 0$	$p_{22} = 0$	$p_{2\cdot} = 1/16$
$p_{\cdot j}$	$p_{\cdot 0} = 9/16$	$p_{\cdot 1} = 3/8$	$p_{\cdot 2} = 1/16$	1

Применяем формулу

$$mas = \frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \sqrt{|x_i - me_X||y_j - me_Y|}\,|K_{ij}|\,p_{ij}\,. \qquad (1)$$

Здесь K_{ij} – ассоциативное ядро, определяемое формулой

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}(1-p_{i\cdot})\,p_{\cdot j}(1-p_{\cdot j})}}\,. \qquad (2)$$

Предварительно создадим таблицу 2 значений ядер K_{ij} и вычислим me_X, d_X.

Таблица 2. Значения ассоциативных ядер K_{ij} к примеру 1.

$i\downarrow\ j\rightarrow$	0	1	2
0	– 0,269841	0,162650	0,227710
1	0,162650	– 0,066667	–
2	0,227710	–	–

Медиану me_X находим из условия положения ее между точками $x_1 = 0$ и $x_2 = 1$ пропорционально вероятностям $p_1 = 9/16$ и $p_2 = 3/8$: $(me_X - x_1)\,p_1 = (x_2 - me_X)\,p_2$.

Отсюда $me_X = \dfrac{x_1 p_1 + x_2 p_2}{p_1 + p_2} = \dfrac{0\cdot\frac{9}{16} + 1\cdot\frac{6}{16}}{\frac{9}{16}+\frac{6}{16}} = \dfrac{2}{5} = me_Y$. Тогда

$$d_X = d_Y = \left|0-\frac{2}{5}\right|\frac{9}{16} + \left|1-\frac{2}{5}\right|\frac{6}{16} + \left|2-\frac{2}{5}\right|\frac{1}{16} = \frac{11}{20} = 0{,}55\,.$$

Далее вычисляем

$$mas = \frac{20}{11}\left[\sqrt{\left|0-\frac{2}{5}\right|\left|0-\frac{2}{5}\right|}\,0{,}269841\frac{1}{4} + \sqrt{\left|0-\frac{2}{5}\right|\left|1-\frac{2}{5}\right|}\,0{,}162650\frac{1}{4} + \sqrt{\left|0-\frac{2}{5}\right|\left|2-\frac{2}{5}\right|}\,0{,}227710\frac{1}{16} + \right.$$

$$\left. + \sqrt{\left|1-\frac{2}{5}\right|\left|0-\frac{2}{5}\right|}\,0{,}162650\frac{1}{4} + \sqrt{\left|1-\frac{2}{5}\right|\left|1-\frac{2}{5}\right|}\,0{,}066667\frac{1}{8} + \sqrt{\left|2-\frac{2}{5}\right|\left|0-\frac{2}{5}\right|}\,0{,}227710\frac{1}{16} + 0\right] =$$

=0,172. Итак,

$$mas = 0{,}172. \qquad (3)$$

Пример 2. Вычисление коэффициента детерминации *mco* для триномиального распределения, представленного таблицей 1.

Применяем формулу

$$mco = \frac{1}{\sqrt{d_X d_Y}} \sum_i \sum_j \sqrt{|x_i - me_X||y_j - me_Y|}\,|K_{ij}|\,p_{ij} \qquad (4)$$

Здесь K_{ij} – контингенциальное ядро, определяемое формулой

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}} \qquad (5)$$

Сначала создадим таблицу 3 значений контингенциального ядра (5).

Таблица 3. Значения контингенциального ядра K_{ij} из формулы (5) .

$i\downarrow\ j\rightarrow$	0	1	2
0	0,117241	0,084746	0,280000
1	0,084746	0,058824	–
2	0,280000	–	–

$$mco = \frac{20}{11}\left[\sqrt{\left|0-\frac{2}{5}\right|\left|0-\frac{2}{5}\right|}\,0{,}117241\frac{1}{4}+\sqrt{\left|0-\frac{2}{5}\right|\left|1-\frac{2}{5}\right|}\,0{,}084746\frac{1}{4}+\sqrt{\left|0-\frac{2}{5}\right|\left|2-\frac{2}{5}\right|}\,0{,}280000\frac{1}{16}+\right.$$

$$\left.+\sqrt{\left|1-\frac{2}{5}\right|\left|0-\frac{2}{5}\right|}\,0{,}084746\frac{1}{4}+\sqrt{\left|1-\frac{2}{5}\right|\left|1-\frac{2}{5}\right|}\,0{,}058824\frac{1}{8}+\sqrt{\left|2-\frac{2}{5}\right|\left|0-\frac{2}{5}\right|}\,0{,}280000\frac{1}{16}+0\right]=$$

=0,118. Итак,

$$mco = 0{,}118. \qquad (6)$$

Для сравнения в § 7.2 вычислены для этого же распределения комбинированные коэффициенты детерминации $com_a = 0{,}162;\quad com_c = 0{,}123.$

Значения близкие.

§ 11.3. Сравнение величин различных коэффициентов детерминации для триномиального распределения

После рассмотрения всех коэффициентов детерминации полезно сравнить их величины для одного и того же применяемого на практике распределения, например, триномиального. Используем ранее полученные результаты, сведя их вместе.

Пример1. Двумерная случайная величина $\left(X,Y\right)$ распределена по триномиальному закону, который определяется формулой

$$p_{ij} = \frac{n!}{i!\,j!\left(n-i-j\right)!}p_1^i p_2^j\left(1-p_1-p_2\right)^{n-i-j}; \qquad (1)$$

$$i,j = 0.1,...,n;\quad 0<p_1<1;\quad 0<p_2<1;\quad p_1+p_2<1;\quad i+j\le n.$$

Рассмотрим случай $n=2;\quad p_1=p_2=1/4.$ Построим таблицу распределения (табл.1).

Таблица 1 триномиального распределения при $n = 2.$

$X \setminus Y \rightarrow$ $\downarrow$	0	1	2	$p_{i\cdot}$
0	$p_{00} = 1/4$	$p_{01} = 1/4$	$p_{02} = 1/16$	$p_{0\cdot} = 9/16$
1	$p_{10} = 1/4$	$p_{11} = 1/8$	$p_{12} = 0$	$p_{1\cdot} = 3/8$
2	$p_{20} = 1/16$	$p_{21} = 0$	$p_{22} = 0$	$p_{2\cdot} = 1/16$
$p_{\cdot j}$	$p_{\cdot 0} = 9/16$	$p_{\cdot 1} = 3/8$	$p_{\cdot 2} = 1/16$	1

Для этого распределения ранее в предыдущих параграфах было получено:

1. Модуль линейного коэффициента корреляции $|\rho| = 0{,}333.$
2. Ассоциативный коэффициент детерминации $as = 0{,}186.$
3. Контингенциальный коэффициент детерминации $co = 0{,}430.$
4. Комбинированный коэффициент детерминации комби-ас $com_a = 0{,}162.$
5. Комбинированный коэффициент детерминации комби-конт $com_c = 0{,}123.$
6. Предельный коэффициент детерминации $l = 0{,}114.$
7. Дефектологический коэффициент детерминации $def = 0{,}185.$
8. Комбинированный коэффициент детерминации мед-ас $mas = 0{,}172.$
9. Комбинированный коэффициент детерминации мед-конт $mco = 0{,}118.$

Ранжируем эти коэффициенты в порядке возрастания:

$$l,\quad mco,\quad com_c,\quad com_a,\quad mas,\quad def,\quad as,\quad |\rho|,\quad co.$$

Взаимные отношения этих коэффициентов связи следует учитывать в работе.

§ 11.4. Комбинированные коэффициенты детерминации для непрерывных случайных величин

В главе 7 комбинированные коэффициенты для непрерывных случайных величин не рассматривались, поэтому здесь рассмотрим комбинированные коэффициенты для непрерывных распределений как на основе математического ожидания, так и на основе медианы.

Комбинированный коэффициент детерминации комби-ас для непрерывной случайной величины (X, Y) с плотностью $f_{XY}(x, y)$ и функцией распределения $F_{XY}(x, y)$ определяется формулой

$$com_a = \frac{1}{\sigma_X \sigma_Y} \iint\limits_D |x - m_X||y - m_Y||K(x,y)| f_{XY}(x,y)\,dxdy \ . \qquad (1)$$

Здесь

$$K(x,y) = \frac{F_{XY}(x,y) - F_X(x)F_Y(Y)}{\sqrt{F_X(x)(1-F_X(x))F_Y(Y)(1-F_Y(Y))}}; \qquad (2)$$

m_X, m_Y – математические ожидания, σ_X, σ_Y – средние квадратические отклонения компонент двумерной случайной величины (X,Y). D – область ее значений.

Аналогично, комбинированный коэффициент детерминации комби-конт для непрерывной случайной величины (X,Y) определяется формулой

$$com_c = \frac{1}{\sigma_X \sigma_Y} \iint\limits_d |x - m_X||y - m_Y||K(x,y)| f_{XY}(x,y)\,dxdy \ . \qquad (3)$$

Здесь

$$K(x,y) = \frac{F_{XY}(x,y) - F_X(x)F_Y(y)}{F_{XY}(x,y)(1 + 2F_{XY}(x,y) - 2F_X(x) - 2F_Y(y)) + F_X(x)F_Y(y)} \ . \qquad (4)$$

Комбинированные коэффициенты детерминации мед-ас и мед-конт (на основе медианы) определяются формулой

$$m(as,co) = \frac{1}{\sqrt{d_X d_Y}} \iint\limits_D \sqrt{|x - me_X||y - me_Y|}\,|K(x,y)| f_{XY}(x,y)\,dxdy \ . \qquad (5)$$

Здесь

$K(x,y)$ определяется формулой (2) для мед-ас и формулой (4) для мед-конт. me_X, me_Y – медианы, d_X, d_Y – средние абсолютные отклонения компонент двумерной случайной величины (X,Y). Для строго возрастающей функции распределения медианы определяются однозначно. В противном случае требуется дополнительное соглашение для их вычисления, подобное тому, что сделано в § 11.1. для дискретного случая (промежуток постоянства делится пропорционально соседним вероятностным массам). Свойства этих коэффициентов детерминации – такие же, как и в дискретном случае. Там они рассмотрены подробно. Примеры вычисления этих коэффициентов для конкретных распределений приводят к неберущимся интегралам, что представляет определенные трудности. Здесь эти примеры не рассматриваются.

Замечание. Для комбинированного коэффициента детерминации в случае дискретного распределения возможен случай, когда случайная величина (X,Y) принимает значения центра рассеяния $(m_X; m_Y)$ или $(me_X; me_Y)$. Равенство нулю комбинированного коэффициента детерминации приводит к равенству нулю всех слагаемых под знаком двойной суммы ((3),§ 11.1). В точке $(me_X; me_Y)$ соответствующее слагаемое равно нулю, а ядро K_{ij}, казалось бы, может и не равняться нулю. Тогда не обеспечивается независимость случайных величин X, Y. Однако, применяя формулы согласованности для двумерного распределения, удается доказать, что и в этом случае условия независимости случайных величин выполняются.

Глава 12.Таблицы

Глава 12 содержит 3 вспомогательные таблицы и список названий и обозначений.

§12.1. Сводные таблицы коэффициентов детерминации и корреляции

Большое количество коэффициентов детерминации и корреляции, рассмотренных в этой книге, заставляет свести их воедино в таблицу для сопоставления и выбора нужных для исследования на основе конструкции и свойств коэффициентов.

Все коэффициенты детерминации δ обладают следующими общими свойствами:

1. $0 \le \delta \le 1$. Границы достижимы.
2. Для независимости случайных величин X, Y необходимо и достаточно выполнение равенства $\delta = 0$.

Все коэффициенты корреляции κ обладают следующими общими свойствами:

1. $-1 \le \kappa \le 1$. Границы достижимы.
2. Если случайные величины X, Y независимы, то $\kappa = 0$.
3. Равенство нулю коэффициента корреляции в общем случае не обеспечивает независимость случайных величин X, Y. Исключение составляет максимальный коэффициент корреляции.

Таблица 1. Коэффициенты детерминации.

№	Название. Краткое название. Конструкция. Комментарии.
1	Ассоциативный коэффициент детерминации для дискретных случайных величин, «Ас» $as = \sum_{i,j} \frac{\left\lvert p_{ij} - p_{i\cdot} p_{\cdot j} \right\rvert}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right) p_{\cdot j}\left(1-p_{\cdot j}\right)}} p_{ij}$. $p_{ij} = P\left(X = x_i, Y = y_j\right);\ p_{i\cdot} = P\left(X = x_i\right); p_{\cdot j} = P\left(Y = y_j\right)$. Случайные величины могут выражать числовые, качественные и смешанные признаки.
2	Ассоциативный коэффициент детерминации для непрерывных случайных величин, «Ас» $as = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\left\lvert F_{XY}(x,y) - F_X(x) F_Y(y) \right\rvert f_{XY}(x,y)\,dxdy}{\sqrt{F_X(x)\left(1-F_X(x)\right) F_Y(y)\left(1-F_Y(y)\right)}}$. $as = 1 \Rightarrow X, Y$ связаны взаимно однозначной функциональной зависимо-

	стью.
№	Название. Конструкция. Краткое название. Комментарии.
3	Контингенциальный коэффициент детерминации для дискретных случайных величин, «Конти» $co = \sum_{i,j} \frac{\left\lvert p_{ij} - p_{i\cdot} p_{\cdot j} \right\rvert}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}} p_{ij}$ $p_{ij} = P\left(X = x_i, Y = y_j\right);\ p_{i\cdot} = P\left(X = x_i\right);\ p_{\cdot j} = P\left(Y = y_j\right).$ Случайные величины могут выражать числовые, качественные и смешанные признаки.
4	Контингенциальный коэффициент детерминации для непрерывных случайных величин, «Конти» $co = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left\lvert K\left(x,y\right)\right\rvert f_{XY}\left(x,y\right)dxdy;$ $K\left(x,y\right) =$ $= \frac{F_{XY}\left(x,y\right) - F_X\left(x\right)F_Y\left(y\right)}{F_{XY}\left(x,y\right)\left(1 + 2F_{XY}\left(x,y\right) - 2F_X\left(x\right) - 2F_Y\left(y\right)\right) + F_X\left(x\right)F_Y\left(y\right)}.$
5	Предельный коэффициент детерминации для дискретных случайных величин, «Предельный» $l = \sum_{i,j} \frac{\left\lvert p_{ij} - p_{i\cdot} p_{\cdot j} \right\rvert}{p_{ij} + p_{i\cdot} p_{\cdot j}} p_{ij}.$ Верхняя граница $l = 1$ не достижима, но является точной. Простая конструкция. Применяется для исследования числовых, качественных и смешанных признаков.
6	Предельный коэффициент детерминации для непрерывных случайных величин, «Предельный»

	$l=\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\left\|K\left(x,y\right)\right\|f_{XY}\left(x,y\right)dxdy;$ $K\left(x,y\right)=\dfrac{f_{XY}\left(x,y\right)-f_X\left(x\right)f_Y\left(y\right)}{f_{XY}\left(x,y\right)+f_X\left(x\right)f_Y\left(y\right)}.$ Верхняя граница $l=1$ не достижима, но является точной. Простая конструкция, выраженная через плотности.
№	Название. Конструкция. Краткое название. Комментарии.
7	Комбинированный коэффициент детерминации для дискретных случайных величин комби-ас на основе математических ожиданий и средних квадратических отклонений. $com_a=\dfrac{1}{\sigma_X\sigma_Y}\sum_{i,j}\left\|x_i-m_X\right\|\left\|y_j-m_Y\right\|\left\|K_{ij}\right\|p_{ij};$ $K_{ij}=\dfrac{p_{ij}-p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}.$ $com_a=1\Rightarrow X,Y$ связаны линейной зависимостью. Явно учитывает и значения и вероятности этих значений случайных величин.
8	Комбинированный коэффициент детерминации для дискретных случайных величин комби-конт на основе математических ожиданий и средних квадратических отклонений. $com_c=\dfrac{1}{\sigma_X\sigma_Y}\sum_{i,j}\left\|x_i-m_X\right\|\left\|y_j-m_Y\right\|\left\|K_{ij}\right\|p_{ij};$ $K_{ij}=\dfrac{p_{ij}-p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1+2p_{ij}-2p_{i\cdot}-2p_{\cdot j}\right)+p_{i\cdot}p_{\cdot j}}.$ $com_c=1\Rightarrow X,Y$ связаны линейной зависимостью. Явно учитывает и значения и вероятности этих значений случайных величин.
9	Комбинированный коэффициент детерминации для дискретных случайных величин

	мед-ас на основе медиан, средних абсолютных отклонений и ассоциативного ядра. $mas = \frac{1}{d_X d_Y} \sum_{i,j} \left\| x_i - me_X \right\| \left\| y_j - me_Y \right\| \left\| K_{ij} \right\| p_{ij}$; $K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1 - p_{i\cdot}\right) p_{\cdot j}\left(1 - p_{\cdot j}\right)}}$. $mas = 1 \Rightarrow X, Y$ связаны линейной зависимостью. Явно учитывает и значения и вероятности этих значений случайных величин.
10	Комбинированный коэффициент детерминации для дискретных случайных величин мед-конт на основе медиан, средних абсолютных отклонений и контингенциального ядра. $mco = \frac{1}{d_X d_Y} \sum_{i,j} \left\| x_i - me_X \right\| \left\| y_j - me_Y \right\| \left\| K_{ij} \right\| p_{ij}$; $K_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot} p_{\cdot j}}$. $mco = 1 \Rightarrow X, Y$ связаны линейной зависимостью. Явно учитывает и значения и вероятности этих значений случайных величин.
№	Название. Конструкция. Краткое название. Комментарии.
11	Дефектологический коэффициент детерминации для дискретных случайных величин, «Дефект» $def = 6 \sum_{i=1}^{m} \sum_{j=1}^{n} \left\| F_{ij} - F_{i\cdot} F_{\cdot j} \right\| p_{ij}$. Здесь $F_{ij} = \sum_{k=1}^{i} \sum_{l=1}^{j} p_{kl}$; $F_{i\cdot} = \sum_{k=1}^{i} p_{k\cdot}$; $F_{\cdot j} = \sum_{l=1}^{j} p_{\cdot l}$; $p_{ij} = P\left(X = x_i, Y = y_j\right)$. Верхняя граница $def = 1$ точная. Простая конструкция. Применяется для исследования числовых, качественных и смешанных признаков.
12	Дефектологический коэффициент детерминации для непрерывных случайных, «Дефект» величин $def = 6 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left\| F_{XY}\left(x, y\right) - F_X\left(x\right) F_Y\left(y\right) \right\| f_{XY}\left(x, y\right) dxdy$

	Верхняя граница $def = 1$ достижима. Простая конструкция. Применяется для исследования числовых, качественных и смешанных признаков.
13	Максимальный коэффициент детерминации для дискретных случайных величин. Несимметричный случай. «Макси». $\rho^* = \lvert\lambda_1\rvert$; $\lambda_1^2 = \mu_1$ – корень из первого (наименьшего, отличного от нуля и единицы) собственного числа матрицы $P_1 = \left(\frac{p_{ij}^{(1)}}{\sqrt{p_{i\cdot} p_{j\cdot}}} \right)$; $i, j = 1, 2..., m;$ $\quad p_{ij}^{(1)} = \sum_{k=1}^{n} \frac{p_{ik} p_{jk}}{p_{\cdot k}}$; $p_{ij} = P\left(X = x_i, Y = y_j \right)$; $\quad i = 1, 2, ..., m;$ $\quad j = 1, 2, ..., n.$
14	Максимальный коэффициент детерминации для непрерывных случайных величин. Несимметричный случай. «Макси». $\rho^* = \frac{1}{\lvert\lambda_1\rvert}$; $\lambda_1^2 = \mu_1$ – первое (наименьшее) собственное число симметричного ядра $K_1(x, y) = \frac{f_1(x, y)}{\sqrt{f_X(x) f_X(y)}}$ линейного однородного интегрального уравнения $\omega(x) = \lambda \int_a^b \omega(y) K_1(x, y) dy$; $f_1(x, y) = \int_{a_1}^{b_1} \frac{f_{XY}(x, t) f_{XY}(y, t)}{f_Y(t)} dt$; $f_{XY}(x, y)$ – плотность несимметричного распределения в прямоугольнике $a \le x \le b; a_1 \le y \le b_1$.

Таблица 2. Коэффициенты корреляции.

№	Название. Конструкция. Краткое название. Комментарии.
1	Линейный коэффициент корреляции. Общее определение. «Коэффициент корреляции». $\rho = \frac{K_{XY}}{\sigma_X \sigma_Y}$; $K_{XY} = M\left[\left(X - m_X \right) \left(Y - m_Y \right) \right]$; $\sigma_X = \sqrt{M\left[\left(X - m_X \right)^2 \right]}$;

	$$\sigma_Y = \sqrt{M\left[\left(X - m_X\right)^2\right]}.$$ $\rho = 0$ не обеспечивает независимость случайных величин X, Y. Границы изменения ρ достижимы.
2	Линейный коэффициент корреляции для дискретных случайных величин. $$\rho = \frac{K_{XY}}{\sigma_X \sigma_Y};\ K_{XY} = \sum_{i,j}\left(x_i - m_X\right)\left(y_j - m_Y\right)p_{ij};$$ $$\sigma_X = \sqrt{\sum_i \left(x_i - m_X\right)^2 p_{i\cdot}};\ \sigma_Y = \sqrt{\sum_j \left(y_j - m_j\right)^2 p_{\cdot j}}$$
3	Линейный коэффициент корреляции для непрерывных случайных величин. $$\rho = \frac{K_{XY}}{\sigma_X \sigma_Y};\ K_{XY} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(x - m_X\right)\left(y - m_Y\right)f_{XY}\left(x, y\right)dxdy;$$ $$\sigma_X = \sqrt{D_X};\sigma_Y = \sqrt{D_Y};\ D_X = \int_{-\infty}^{+\infty}\left(x - m_X\right)^2 f_X\left(x\right)dx;$$ $$D_Y = \int_{-\infty}^{+\infty}\left(y - m_Y\right)^2 f_Y\left(y\right)dy.$$
4	Максимальный коэффициент корреляции для дискретных случайных величин. Симметричный случай. «Макси». $\rho^* = \mu_1$ – наибольшее по модулю собственное число симметричной матрицы $\left(\frac{p_{ij}}{\sqrt{p_{i\cdot}p_{\cdot j}}}\right)$.
№	Название. Конструкция. Краткое название. Комментарии.
5	Максимальный коэффициент корреляции для непрерывных случайных величин. Симметричный случай. «Макси».

	$\rho^* = \frac{1}{\lambda_1}$; λ_1 – наименьшее по модулю собственное число ядра $K(x,y) = \frac{f_{XY}(x,y)}{\sqrt{f_X(x) f_Y(y)}}$ линейного однородного интегрального уравнения $\omega(x) = \lambda \int_a^b \omega(y) K(x,y)\, dy.$ Симметричная плотность определена в квадрате $a \le x \le b; a \le y \le b$. $\rho^* = 0 \Leftrightarrow$ случайные величины X, Y независимы.
6	Коэффициент корреляции между событиями A, B. «Корсоб». $\rho_{AB} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)(1-P(A))P(B)(1-P(B))}}$. По определению ρ_{AB} есть линейный коэффициент корреляции между индикаторами событий A, B. $\rho_{AB} = 0 \Leftrightarrow$ случайные события A, B независимы.
7	Ассоциативный коэффициент корреляции для дискретных случайных величин «Аскор». $as = \sum_{i,j} \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot}(1-p_{i\cdot}) p_{\cdot j}(1-p_{\cdot j})}} p_{ij};$ $p_{ij} = P(X = x_i, Y = y_j);\ p_{i\cdot} = P(X = x_i); p_{\cdot j} = P(Y = y_j).$
8	Ассоциативный коэффициент корреляции для непрерывных случайных величин. «Аскор». $as = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{F_{XY}(x,y) - F_X(x) F_Y(y)}{\sqrt{F_X(x)(1-F_X(x)) F_Y(y)(1-F_Y(y))}} f_{XY}(x,y)\, dxdy;$ $F_{XY}(x,y) = P(X < x, Y < y); F_X(x) = P(X < x); F_Y(y) = P(Y < y)$, $\forall x, y.$

№	Название. Конструкция. Краткое название. Комментарии.
9	Контингенциальный коэффициент корреляции для дискретных случайных величин. «Контикор». $co_c = \sum_{i,j} \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij}\left(1 + 2p_{ij} - 2p_{i\cdot} - 2p_{\cdot j}\right) + p_{i\cdot}p_{\cdot j}} p_{ij}$ $p_{ij} = P\left(X = x_i, Y = y_j\right);\ p_{i\cdot} = P\left(X = x_i\right); p_{\cdot j} = P\left(Y = y_j\right).$
10	Контингенциальный коэффициент корреляции для непрерывных случайных величин. «Контикор». $co = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K(x,y) f_{XY}(x,y)\,dxdy;$ $K(x,y) =$ $= \frac{F_{XY}(x,y) - F_X(x)F_Y(y)}{F_{XY}(x,y)\left(1 + 2F_{XY}(x,y) - 2F_X(x) - 2F_Y(y)\right) + F_X(x)F_Y(y)}.$
11	Предельный коэффициент корреляции для дискретных случайных величин, «Предкор» $l_c = \sum_{i,j} \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{p_{ij} + p_{i\cdot}p_{\cdot j}} p_{ij}.$
12	Предельный коэффициент корреляции для непрерывных случайных величин, «Предкор» $co_c = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K(x,y) f_{XY}(x,y)\,dxdy;$ $K(x,y) = \frac{f_{XY}(x,y) - f_X(x)f_Y(y)}{f_{XY}(x,y) + f_X(x)f_Y(y)}.$

13	Дефектологический коэффициент корреляции для дискретных случайных величин, «Дефектокор» $def = 6\sum_i \sum_j \left(F_{ij} - F_{i\cdot}F_{\cdot j}\right) p_{ij}.$ Здесь $F_{ij} = \sum_{k=1}^{i}\sum_{l=1}^{j} p_{kl};\ F_{i\cdot} = \sum_{k=1}^{i} p_{k\cdot}; F_{\cdot j} = \sum_{l=1}^{j} p_{\cdot l};\ p_{ij} = P\left(X = x_i, Y = y_j\right).$

	Простая конструкция. Применяется для исследования числовых, качественных и смешанных признаков.
№	Название. Конструкция. Краткое название. Комментарии.
14	Дефектологический коэффициент корреляции для непрерывных случайных величин. «Дефектокор» $def = 6\int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}\left(F_{XY}(x,y) - F_X(x)F_Y(y)\right)f_{XY}(x,y)\,dxdy$. Простая конструкция. Применяется для исследования числовых, качественных и смешанных признаков.

§ 12.2. Таблица нормирующих коэффициентов для устранения смещения оценок положения и рассеяния при нормальном законе распределения

Таблица. Нормирующие коэффициенты для устранения смещения оценок среднего квадратического отклонения σ в случае нормального распределения.

n	2	4	6	8	10	12	14	16	18	20
$k_s(n)$	0.7979	0.9213	0.9515	0.9650	0.9727	0.9776	0.9810	0.9835	0.9854	0.9869
$k_d(n)$	0.5642	0.6632	0.7035	0.7253	0.7390	0.7482	0.7550	0.7602	0.7642	0.7675
$k_R(n)$	1.128	2.059	2.534	2.847	3.078	3.258	3.407	3.532	3.640	3.735
$k_q(n)$	—	2.059	1.284	1.704	1.312	1.586	1.324	1.526	1.330	1.491

Составлена Максимовым Ю.Д. Рассматриваются четыре несмещенные оценки σ.

1. Нормированное выборочное среднее квадратическое отклонение $s' = s/k_s(n)$.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad k_s(n) = \sqrt{\frac{2}{n}}\cdot\Gamma\left(\frac{n}{2}\right)\Big/\Gamma\left(\frac{n-1}{2}\right); \quad \mathsf{M}[s'] = \sigma.$$

2. Нормированное выборочное среднее абсолютное отклонение $d^* = d/k_d(n)$.

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mathsf{med}|; \quad k_d(n) = \mathsf{M}[d/\sigma]; \quad \mathsf{M}[d^*] = \sigma.$$

3. Нормированный размах $R^* = R/k_R(n)$.

$$R = x_{(n)} - x_{(1)} = x_{\max} - x_{\min}; \quad k_R(n) = \mathsf{M}\left[R/\sigma\right]; \quad \mathsf{M}[R^*] = \sigma.$$

4. Нормированная выборочная интерквартильная широта $q^* = q/k_q(n)$.

$$q = z_{3/4} - z_{1/4}; \quad z_{1/4} = x_{(i)}; \quad z_{3/4} = x_{(n-i+1)};$$

$$i = \begin{cases} \left[n/4\right]+1 & \text{при } n/4 \text{ дробном}; \\ n/4 & \text{при } n/4 \text{ целом}; \end{cases} \qquad \mathsf{M}[q^*] = \sigma.$$

Оценки s', d^*, R^*, q^* – несмещенные.

§ 12.3. Обозначения и названия числовых характеристик

1. δ_λ – коэффициент связи степени усреднения λ.
2. δ – коэффициент связи при $\lambda = 1$.
3. as_λ – ассоциативный коэффициент детерминации степени λ (для дискретных случайных величин и общего вида).
4. as – ассоциативный коэффициент детерминации при $\lambda = 1$, «ас» (для дискретных случайных величин и общего вида).
5. co_λ – контингенциальный коэффициент детерминации степени λ (для дискретных случайных величин и общего вида).
6. co – контингенциальный коэффициент детерминации при $\lambda = 1$,«конти»(для дискретных случайных величин и общего вида).
7. l – предельный коэффициент детерминации (для дискретных и непрерывных случайных величин).
8. com_a – комбинированный коэффициент детерминации с ассоциативным ядром, «комби-ас» (для дискретных случайных величин).
9. com_c – комбинированный коэффициент детерминации с контингенциальным ядром, «комби-конт»
(для дискретных случайных величин).
10. def – дефектологический коэффициент детерминации (для дискретных и непрерывных случайных величин).
11. ρ – линейный коэффициент корреляции (К Пирсона).
12. ρ^* – максимальный коэффициент корреляции (О.В.Сарманова).
13. ρ_{AB} – коэффициент корреляции между событиями A, B, «корсоб».
14. as_c – ассоциативный коэффициент корреляции, «аскор» (для дискретных и непрерывных случайных величин).
15. co_c – контингенциальный коэффициент корреляции, «контикор», (для дискретных и непрерывных случайных величин).
16. l_c – предельный коэффициент корреляции «предкор», (для дискретных и непрерывных случайных величин).

17. def_c – дефектологический коэффициент корреляции «дефектокор», (для дискретных и непрерывных случайных величин).
18. as_m – ассоциативный медианный коэффициент детерминации для дискретных случайных величин.
19. co_m – контингенциальный медианный коэффициент детерминации для дискретных случайных
величин.
20. mas – комбинированный коэффициент детерминации на основе медианы и ассоциативного ядра.
21. mco – комбинированный коэффициент детерминации на основе медианы и контингенциального ядра.
22. m_X – математическое ожидание случайной величины X.
23. $\overline{x}$ – выборочное среднее.
24. me – генеральная медиана.
25. med – выборочная медиана.
26. t_q – полусумма выборочных квартилей.
27. t_R – полусумма крайних элементов вариационного ряда выборки.
28. σ – генеральное среднее квадратическое отклонение.
29. s – выборочное среднее квадратическое отклонение.
30. s' – нормированное выборочное среднее квадратическое отклонение.
31. d – выборочное среднее абсолютное отклонение.
32. d^* – нормированное среднее абсолютное отклонение.
33 d_X – генеральное среднее абсолютное отклонение случайной величины X.
34. q – выборочная интерквартильная широта.
35. q^* – нормированная выборочная интерквартильная широта.
36. R – размах выборки.
37. R^* – нормированный размах выборки.

Глава 13. Краткая история и философия зависимости

Для лучшего понимания места, ценности и назначения мер зависимости, изучаемых в этой книге, целесообразно обратиться к историческим и философско-методологическим корням этой темы.

§ 13.1. Краткая история корреляции и регрессии

Корреляция и регрессия применяются для измерения связей между явлениями, признаками, процессами, случайными величинами. Так как все в мире связано, то обнаружение и измерение величины этих связей позволяет обоснованно прогнозировать явления. В настоящей книге и в более ранних статьях опубликованы новые коэффициенты связи, более совершенные, чем известные. Они являются закономерным развитием старых

идей, а не плодом чистого мышления. Выделим основные этапы развития понятий корреляции и регрессии, которые тесно связаны и развивались совместно.

1. 1806 г. Жорж Кювье (1769 – 1832), французский палеонтолог ввел в науку в 1806 г. понятие и слово «Корреляция», когда занимался исследованиями в области палеонтологии и сравнительной анатомии. Слово «корреляция» происходит от позднелатинского слова «correlation», что означает соответствие, соотношение. В отличие от слова «relation» это не просто «отношение, связь», а «как бы связь», то есть связь, но не в привычной в то время детерминистской, функциональной форме. Кювье заметил связь органов живого организма между собой, связь строения животного с образом его жизни, связь видов животных и растений с определенным временем их жизни и много других связей. Это позволило ему сформулировать общие принципы «корреляции органов» и «функциональной корреляции».
2. 1805 – 1809 г.г. Метод наименьших квадратов предопределил появление формулы для вычисления величины связи между признаками явления – формулы линейного коэффициента корреляции. Этот метод появился при решении практических задач астрономии, геодезии, связанных с минимизацией вычислительных ошибок при обработке результатов измерений. Независимо, хотя спор о приоритетах существует, метод открыли и сформулировали три математика в близкие времена: француз Адриен Лежандр (1752 – 1833) в 1805 г., американец Роберт Эдрейн в 1808 г., немец Карл Фридрих Гаусс (1777 – 1855) в 1809 г.
3. 1812.г. Пьер Симон Лаплас (1749 – 1827) – французский математик, физик, астроном опубликовал аналитическую теорию вероятностей, развил теорию ошибок и с вероятностных позиций обосновал метод наименьших квадратов. Его работы позволили в дальнейшем вопросы корреляции и регрессии поставить на твердую вероятностную основу.
4. 1846.г. Огюст Браве (1811 – 1863) – французский астроном, физик, кристаллограф, ботаник, метеоролог, развивая вероятностную теорию ошибок, с помощью метода наименьших квадратов аналитически и геометрически нашел прямую, которая впоследствии Ф. Гальтоном была названа прямой регрессии. Угловой коэффициент этой прямой выражается через коэффициент корреляции, что позже показал

К. Пирсон.

5. 1885 г. Френсис Гальтон (1822 – 1911) – английский антрополог, биолог, психолог,

метеоролог, президент Британского королевского научного общества ввел понятие «Регрессия», изучая размеры детей и их родителей (на бобах и людях). Он установил, что размеры наследников в среднем были ближе к середине, чем размеры родителей, что и означает регрессию (термин прижился, им теперь называют любую функцию, описывающую изменение в среднем). Им же в 1888 г. введен и

термин «Корреляционный анализ». Сотрудничая с К. Пирсоном, Ф. Гальтон произвел численные расчеты корреляции из области демографии и социологии на основе собранной им многочисленной статистики. Ф. Гальтон пришел к регрессии эмпирически.

6 1896 г. Карл Пирсон (1857 – 1936), английский математик-статистик, биолог, философ, основоположник знаменитого журнала «Биометрика» дал аналитическое выражение линейному коэффициенту корреляции (1896 г., Pearson product-moment correlation coefficient) и ряду других коэффициентов связи, превратил концепцию корреляции в математическую, статистическую теорию. Благодаря Чарльзу Дарвину, К. Пирсон познакомился с Ф. Гальтоном и его идеями, став последователем, соратником и продолжателем его работ, не только в области статистики. В 1892 году К. Пирсон по рекомендации Альберта Ейнштейна издал свой знаменитый философский труд «Грамматика науки» (рус. пер. в 1911г.), где, в частности, анализирует вопросы случайности и вероятности.

7. 1900 – 1912 г.г. Джордж Одни Юл (1871 – 1951), английский статистик, профессор Кембриджского университета, президент Королевского статистического общества, ученик К. Пирсона. Работал в области теорий регрессии и корреляции. Его коэффициент контингенции для событий применяется автором в настоящей монографии для построения контингенциального коэффициента детерминации случайных величин.

8. 1915 – 1928 г.г. Рональд Айлмер Фишер (1890 – 1962), английский математик-статистик, генетик, продолжая работы К. Пирсона, получил и исследовал распределение выборочного линейного коэффициента корреляции и его многомерных модификаций – частного и множественного. Благодаря другим его статистическим работам, Р. Фишер считается одним из основоположников современной математической статистики.

9. 1930 – 1933 г.г. Сергей Натанович Бернштейн (1880 – 1960), академик исследовал коэффициент корреляции между событиями, введенный К. Пирсоном, создал теорию нормальной корреляции. Коэффициент корреляции между событиями автором книги используется как ядро для построения ассоциативных коэффициентов корреляции и детерминации для случайных величин.

10. 1946 – 1958 г.г. Олег Васильевич Сарманов (1916 – 1977), ученик С.Н. Бернштейна, профессор создал теорию максимального коэффициента корреляции – более совершенный, но аналитически гораздо более сложный инструмент измерения зависимости. Равенство нулю этого коэффициента обеспечивает независимость случайных величин, чего нельзя сказать о линейном коэффициенте корреляции.

11. 1998 – 2006 г.г. Юрий Дмитриевич Максимов, ученик О.В.Сарманова, профессор разработал теорию серии коэффициентов детерминации, более простых по структуре, чем коэффициент О.В. Сарманова, но обеспечивающих независимость случайных величин при равенстве коэффициента нулю.

В качестве резюме выделим 3 периода истории корреляции и регрессии.

1806 – 1896 г.г.: от Кювье до Пирсона – период зарождения и осмысления.

1896 – 1946 г.г.: от Пирсона до Сарманова – период становления и развития математической теории линейного коэффициента корреляции.

1946 – … от Сарманова до настоящего времени - период создания и развития новых коэффициентов, обладающих свойством: их равенство нулю является необходимым и достаточным условием независимости случайных величин.

§ 13.2. **Философские аспекты зависимости**

Природа, изучаемая наукой, безгранична во всех смыслах. Мир, в котором мы живем, материален, а потому един. Это единство и означает бесконечность в пространстве, во времени, в движении, в свойствах. Эта бесконечность и актуальная и потенциальная. Актуальная потому, что она есть в любой части мира – свойства любой части неисчерпаемы и любая часть дробима до бесконечности. Бесконечность мира, с другой стороны, – потенциальная, так как познается ограниченными частями, но процесс познания неограничен ничем, ни временем, ни количеством свойств.

Сила человека – в знаниях. Познавая Природу, человек предвидит события, развитие явлений, движение, изменение ее частей. Знания нужно добывать, создавать. Перефразируя высказывание знаменитого биолога-селекционера Ивана Владимировича Мичурина, скажем: «Мы не можем ждать милостей от Природы, изучать ее – наша задача». Именно, изучать и на основе знаний брать «милости» у Природы, а не вымаливать у Бога. Древние имели много богов по причине недостаточности знаний. Иудаизм, Христианство, Ислам сделали большой шаг вперед в понимании Мира, перейдя к единому Богу. Наука этот шаг поставила на четкую прочную ступень материализма, воплотив знания в законы.

Законы природы и общества – это модели, приближенно описывающие Природу на каком-либо языке (определение академика Никиты Моисеева).

Известное положение идеализма о том, что «Разум дает законы Природе» надо понимать соответствующим образом (независимо от того как это положение понима-

лось в том или ином идеалистическом учении). У Природы на самом деле нет законов. Ни на каких скрижалях они не записаны. Природа существует сама по себе. Ее элементы взаимодействуют и связаны материально, временем, движением, энергетически. Природа так устроена, что ее элементы (звезды, частицы микромира, их совокупности и др.) имеют схожие структуры и свойства, повторяемость. Все это можно заметить. Человек познает Природу и это познание формулирует в виде законов. В этом смысле Разум дает законы, но не Природе, а человеку – законы, приближенно описывающие существование Природы. Под Разумом здесь следует понимать науку, созданную всем человечеством.

Природные зависимости, подмеченные Разумом, формулируются в виде законов, которые и принято называть законами науки или Природы.

Приведем на эту тему высказывание Карла Пирсона из его философского труда «Грамматика науки». Англичанин Карл Пирсон выражается со свойственной ему определенностью: "Законы науки - гораздо больше продукты человеческого ума, чем факты внешнего мира" ("The Grammar of Science", 2nd ed., p. 36<<*108>>). Цитируется по книге В.И.Ленина «Материализм и эмпириокритицизм», гл.3, где В.И.Ленин подвергает критике концепцию идеализма, в том числе и К. Пирсона, о первичности сознания и вторичности материи.

В.И.Ленин ранее там же, гл. 2 пишет «Итак, человеческое мышление по природе своей способно давать и дает нам абсолютную истину, которая складывается из суммы относительных истин. Каждая ступень в развитии науки прибавляет новые зерна в эту сумму абсолютной истины, но пределы истины каждого научного положения относительны, будучи то раздвигаемы, то суживаемы дальнейшим ростом знания. "Абсолютную истину, - говорит И. Дицген в "Экскурсиях", - мы можем видеть, слышать, обонять, осязать, несомненно также *познавать,* но она не входит целиком (geht nicht auf) в познание" (S. 195). "Само собою разумеется, что картина не исчерпывает предмета, что художник остается позади своей модели... Как может картина "совпадать" с моделью? Приблизительно, да" (197). "Мы можем лишь относительно (релятивно) познавать природу и части ее; ибо всякая часть, хотя она является лишь относительной частью природы, имеет все же природу абсолютного, природу природного целого самого по себе (des Naturganzen an sich), не исчерпываемого познанием...»

Человечество познавало окружающий Мир постепенно все более точно и полно. Так для древних Солнце было богом. Птолемей дал материалистическую объективную геоцентрическую систему Мира, которая долгое время удовлетворяла людей и науку. Со временем Николай Коперник и его последователи разрушили старые представления и дали людям и науке новую гелиоцентрическую картину Мира. Эта новая система обросла далее законами Иоганна Кеплера, Исаака Ньютона. Новая математическая модель механики Ньютона и сейчас работает хорошо, но в решении ряда задач она дает

сбои. Применяется модель Альберта Эйнштейна. С развитием наших знаний о Природе и эта Эйнштейновская модель будет уточняться.

Любой научный закон описывает зависимости между элементами Природы. Эти законы могут быть сформулированы словесно на обычном разговорном языке. Например, законы диалектики, логики, биологии (Законы об условных и безусловных рефлексах). Законы Природы могут быть сформулированы также на специальных языках – химическом, физическом, математическом и других. Прежде, чем зависимости, закладываемые в закон, будут сформулированы, нужно их заметить, осмыслить, создав «щель» (ограниченное множество элементов) в Природе, то есть абстрагировавшись от всего остального бесконечного Мира. Изучать Природу приходится в пределах этой «щели». На большее человек не способен в силу своей ограниченности.

Нас сейчас интересуют математические модели Природы, то есть законы Природы, записанные с помощью математических понятий и математической символики. Обычная последовательность действий при построении модели такова. Сначала наблюдения, опыты, сбор статистических данных в любом виде, не обязательно в числовом. Далее формулировка зависимости в различной форме:

1. Описание причин и следствий.
2. Число или система чисел.
3. Функция.
4. График.
5. Уравнение или система уравнений.
6. Алгоритм.
7. Теоретико-множественные структуры.
8. Распределения значений случайных величин.
9. Случайные процессы и поля.
10. Другие.

На третьей ступени построения модели производится логический анализ сформулированных зависимостей, в частности, решение уравнений. На четвертой ступени происходит сопоставление следствий модели с практикой. Если есть расхождения, то модель либо совершенствуется, либо заменяется другой.

Законы Природы, выраженные математически, обычно формулируют в двух формах – детерминистской и стохастической (иначе, вероятностной, статистической).

В детерминистски выраженной закономерности задание значений независимых переменных и параметров дает строго определенные значения функций (их может быть много, например, при решении системы уравнений). Таковы классические законы физики: закон Ома в теории электричества, законы Ньютона в механике.

В стохастически выраженной закономерности задание значений независимых переменных дает лишь вероятности принятия зависимыми переменными определенного множества значений или описывается изменение процесса в среднем. Такие законы

есть в физике (статистические закономерности газа, микрочастиц), экономике, социологии, медицине, инженерных вопросах. Например, можно лишь с определенной вероятностью утверждать сколько будет завтра пожаров, аварий, рождений в городе N. Экстраполируя кривую регрессии на один шаг вперед, можно указать значение исследуемого процесса в среднем через этот шаг. Так прогнозируется развитие производства, науки, экономики, социальной обстановки.

В 1965 г. профессор Калифорнийского университета иранского происхождения Лотфи Заде (Lotfy A. Zadeh) дал миру еще один способ описания зависимостей с помощью теории нечетких множеств (Fuzzy sets). По нашему представлению эта теория впоследствии может соединиться со статистической теорией, ибо откуда брать числовые характеристики нечетких множеств, как не из статистических наблюдений. Теория нечетких множеств сейчас активно развивается. Создает приемы описания нечетких суждений и зависимостей.

При стохастическом описании зависимости неизбежно возникает вопрос об измерении силы, величины зависимости. Для этого нужна мера, выраженная числом. Если зависимость слабая, то ее в практике можно и не учитывать, игнорировать, а рассматривать в изучаемом явлении лишь сильно зависимые факторы и отклики. Для них пытаться найти функциональные зависимости в среднем.

Независимых величин, явлений, событий в Природе нет. В силу единства Мира, его бесконечности всякие два природные элемента (предметы, величины, события, явления) связаны бесчисленным количеством взаимодействий. Нам открыта лишь небольшая «щель» в Природу, в пределах которой рассматриваемые элементы кажутся независимыми. Математическая абстракция независимости событий, случайных величин, полей удобна, так как упрощает описание слабо зависимых элементов Природы.

Силу взаимодействия нужно уметь измерять. Зависимость между двумя событиями A, B можно измерить условными вероятностями $P(A/B)$ и $P(B/A)$, но эти вероятности отражают зависимость несимметрично относительно событий. Симметричной мерой является коэффициент корреляции между событиями, рассмотренный в этой книге в § 1.2. и исследованный С.Н. Бернштейном:

$$\rho_{AB} = \frac{P(AB) - P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}.$$

Этот коэффициент берет свое начало от коэффициента ассоциации К.Пирсона. Может быть выведен оттуда (§§1.2,1.4.)

Карлом Пирсоном в 1896 г. создан линейный коэффициент корреляции – Pearson-product-moment correlation coefficient для измерения величины парной связи между случайными величинами. Он сыграл в науке важную роль, вызвав к жизни много теорий. Однако у него есть конструктивный недостаток. Равенство нулю линейного коэффициента корреляции не обеспечивает независимости случайных величин. Чтобы коэффи-

циент связи был способен выполнять свою роль, он должен удовлетворять ряду требований:

1. Измерять связь в среднем.
2. Равенство нулю коэффициента означает независимость случайных величин.
3. Равенство единице коэффициента по модулю означает жесткую (в частности, линейную) зависимость случайных величин.
4. Модуль коэффициента изменяется в пределах от нуля до единицы.

Коэффициенты связи, удовлетворяющие всем перечисленным требованиям структурно достаточны и могут выполнять роль измерителя связи. Этих коэффициентов связи в предлагаемой читателю книге много. Каждый из этих коэффициентов хорош в соответствующей ситуации. Некоторые из них могут конкурировать друг с другом. Для этих коэффициентов термин «Коэффициент величины (силы) связи» корректен. Нами эти коэффициенты названы коэффициентами детерминации. Для линейного коэффициента связи стихийно установился другой термин «Коэффициент тесноты связи», а не величины в силу указанного выше недостатка. Так как в книге указан общий метод образования коэффициентов связи, то в дальнейшем появятся и другие коэффициенты связи, удовлетворяющие перечисленным требованиям. Среди них нужно будет выбирать лучшие в конкретной ситуации и сопоставлять друг с другом.

Коэффициенты детерминации, введенные в этой книге, измеряют связь между двумя случайными величинами в среднем. Они могут применяться на практике при решении следующих задач.

1. Для отбора существенных (сильно связанных) факторов исследуемого явления.
2. Для построения функциональной модели регрессии между основными факторами.
3. Для построения многомерной матрицы детерминационных коэффициентов между компонентами многомерной случайной величины.
4. Для построения многомерного распределения из нескольких связанных между собой одномерных.
5. Для построения множественных и частных коэффициентов детерминации группы случайных величин.

Измеряемая парная связь между случайными величинами может не быть причинной. Дело в том, что на эти величины действует еще бесконечное число других случайных величин, которые мы не учитываем, но они влияют на исследуемые и поэтому не позволяют построить детерминистскую функциональную модель. Возможно лишь описание связи в среднем или вероятностное с указанием вероятностей множества значений случайных величин. Например, урожаи на двух полях причинно не связаны, но вероятностная связь есть, так как есть общий фактор – погода.

Случайность при описании природных явлений исключить нельзя принципиально, так как человек не может учесть бесконечное число влияющих факторов, а учитывает лишь некоторые, ему доступные.

При описании общей закономерности для повторяющихся явлений случайность также учитывать приходится, так как при каждом повторении явление протекает несколько иначе, чем ранее, поэтому общая закономерность прослеживается лишь в среднем. Что для человека случайно, для Природы – нет. Внимательный анализ показывает, что только так все и должно было произойти.

Случайность, хаос наблюдаются и в явлениях, которые изначально описывались детерминистской моделью дифференциальных уравнений. В условиях неустойчивости эти явления иногда начинают себя вести непредсказуемо. Такой детерминированный хаос также заставляет обратиться к возможностям стохастического описания.

Подводя итог, скажем, что описание зависимости числом (коэффициентом детерминации или корреляции) – только начало более глубокого исследования случайного явления с помощью более сложных моделей.

2º. Базисные понятия математики – основа построения математических моделей.

Любой закон Природы, как мы уже отмечали, выражает природную зависимость и призван служить средством прогноза для Человека. Математически выраженные закономерности обладают лучшей информативностью, чем закономерности, дающие лишь качественную картину явления. Это математические модели природных явлений. Математические модели конкретных явлений, таких, например, как распространение тепла, работа ядерного реактора, строятся с помощью абстрактных математических моделей, таких как дифференциальное исчисление, теория вероятностей и многих других. Эти абстрактные математические модели являются теориями, построенными на базисных математических понятиях. Все понятия математики так или иначе отражают действительность. Многие возникли из решения практических задач. Из них базисные определяют математические теории.

Следует различать базисные понятия 1-го, 2-го и более высоких уровней, которые последовательно обобщают друг друга. Базисные понятия 1-го уровня отличаются меньшей общностью. Это, например, геометрический вектор, производная функции одной переменной. Группа базисных понятий разного уровня и разных модификации, связанных общим содержанием, образуют обобщающее базисное понятие с общим названием. Так, например, понятие «Предел» следует считать обобщающим базисным понятием, так как оно объединяет такие понятия разного уровня и разных модификаций, как предел функции, предел последовательности – числовой, функциональной, матричной, предел интегральной суммы, предел поточечный, равномерный, равностепенный, по вероятности, в среднем и т. д. Определения всех этих пределов могут быть сформулированы на языке «эпсилон-дельта» или на языке окрестностей, что и обнаруживает общность во всех этих понятиях.

Важнейшие обобщающие базисные понятия целесообразно перечислить, начиная со школьных: число, алгебраическое выражение, функция, уравнение, множество, геометрическая фигура, длина, площадь, объем, геометрический вектор, координаты точки и вектора, производная, первообразная, определенный интеграл, высказывание – это школьные базисные математические понятия.

К ним добавляются вузовские: матрица, определитель, пространство, арифметический вектор, предел, непрерывность, дифференциал, интеграл, дифференциальное уравнение, поле, ряд, вероятностно-статистические понятия; их отметим позднее.

Для выделения базисных понятий нужны определенные принципы. Иначе процесс сужения всего множества понятий до базисных будет с сильным разбросом.

К базисному понятию можно предъявить следующие требования:

1. Оно достаточно абстрактно охватывает множество частных случаев.
2. Наряду с другими понятиями образует теорию или несколько теорий.
3. Понятие необходимо, то есть без него построить теорию нельзя.

Возьмем, к примеру, такое базисное понятие, как предел, Оно имеет множество модификаций: предел числовой последовательности, предел функции одной или нескольких переменных, предел слева, справа, верхний, нижний, в конечной точке, на бесконечности, предел по мере, в среднем, равномерный и т.д. На понятии предела построено несколько теорий.

В каждое из перечисленных базисных понятий вкладывается определенное содержание, которое раскрывается в базисных понятиях 1-го уровня, более конкретных.

Мировая практика отдает предпочтение индуктивному методу введения математических понятий: от физических задач – к понятию в его простейшей форме и, далее к более усложненным формам.

Так, в школе учащимся на примере задачи о скорости движения тела дается понятие производной функции одной переменной, а далее в вузе оно развивается. Вводятся понятия производной по направлению, частной производной, производной векторной функции и т. д.

Таким образом, базисное понятие производной для учащихся представляется в виде конкретных модификаций. Для одних оно заканчивается понятием частной производной, для других – производной матрицы, тензора, для третьих – производной оператора. Интуитивно ощущается общее во всех этих определениях: скорость изменения, способ определения производной через предел и др.

Естественно, что самого общего базисного понятия дать нельзя, ибо наука развивается и переходит ко все более общим представлениям.

Базисные понятия – это как бы каркас, на котором строится все здание теории. Они определяют теорию и, следовательно, основные задачи теории, а те, в свою очередь, определяют базисные методы решения основных задач. Очень часто базисный метод представляет собой применение системы формул, так как теория доводится до

уровня исчисления. Примером является дифференциальное исчисление функций одной переменной, в котором выводится таблица формул для производных. С помощью этой таблицы можно найти производную любой элементарной функции.

Базисными понятиями в теории вероятностей являются понятия события, вероятности, случайной величины. На них построена теория вероятностей. Естественно, в теории вероятностей используются и другие базисные понятии меньшего уровня – числовая характеристика случайной величины, закон распределения и другие. Базисными понятиями в математической статистике являются: выборка, статистическая оценка, статистическая гипотеза и ряд понятий меньшего уровня, с помощью которых строятся различные статистические теории.

Детерминационная и корреляционная теории являются частями и теории вероятностей и математической статистики. Они призваны исследовать зависимость между событиями, случайными величинами, а, следовательно, природными и общественными явлениями. Базисные понятия здесь – коэффициенты детерминации и корреляции и их статистические оценки. Величина зависимости в этих теориях исследуется с помощью указанных числовых характеристик и их оценок. Теории еще далеки до завершения и на роль исчислений пока не тянут, но в будущем это несомненно произойдет.

Предметно-именной указатель

А

Б

В

Г

Д

Ж –

З

И

К

– – – предельное 112, 113

Библиографический список

1. Бернштейн С.Н. Теория вероятностей. М. Л.: ГТТИ, 1934, 412 с. 2-е изд.
2. Бернштейн С.Н. Теория вероятностей. М.-Л.: ОГИЗ., 1946, 556 с. 4-е изд.
3. Вероятностные разделы математики / под ред. Максимова Ю.Д.. СПб.: изд. «Иван Федоров», 2001, с. 592.
4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 1, изд. 2-е. / Пер. с англ. М.: Финансы и статистика, 1986, с 368.
5. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 2, изд. 2-е. / Пер. с англ. М.: Финансы и статистика, 1987, с. 352.
6. Елисеева И.И. Статистические методы измерения связей. Л. Изд. Ленинградского университета, 1982.
7. Елисеева И.И., Рукавишников В.О. Логика прикладного статистического анализа. М.: Финансы и статистика, 1982.
8. Елисеева И.И., Юзбашев М.М. Общая теория статистики. М.: Финансы и статистика,2001, 480 с.
9. Канаев И.И. Френсис Гальтон. Л.: Наука, 1972, С. 100.
10. Канторович Л.В. и Крылов В,И. Приближенные методы высшего анализа. М.-Л.: Физматгиз, 1962, 708 с.
11. Кендалл М.Дж., Стьюарт А.. Статистические выводы и связи. М.: Наука. Пер. с англ.,1973, 900 с.
12. Козлов В.Н., Максимов Ю.Д., Хватов Ю.А. Базис дисциплины «Высшая математика». Учебное пособие. С.-П.: Изд. СПбГТУ, 1995, с.76.
13. Козлов В.Н., Максимов Ю.Д., Хватов Ю.А. Математика. Структурированная программа (базис). Учебное пособие. С-П.: Изд. СПбГТУ, 2001,с.56.
14. Крамер Г. Математические методы статистики. М.: Мир, 1975, 648с.
15. Ленин В.И. Материализм и эмпириокритицизм, ПСС, т.18.
16. Максимов Ю.Д. Статистический коэффициент ассоциации между двумя признаками как выборочный коэффициент корреляции между индикаторами двух событий. Сб. Трудов V междун. н.-м. конф. «Высокие интеллектуальные технологии образования и науки». Материалы н.-м. программы «Университеты России». Направл.3; 30 – 31 янв. 1998. СПб. Изд. СПбГТУ; с. 146 – 147. Статья.
17. Максимов Ю.Д. Системный метод точечного оценивания числовых характери стик положения и рассеяния распределений. Журн. «Заводская лаборатория». Диагностика материалов. №1, 1999, с. 56 – 61.
18. Максимов Ю.Д. Новые коэффициенты детерминации для исследования зависимостей случайных явлений. Материалы X Всероссийской конф. по проблемам науки и высшей школы. «Фундаментальные исследования в технических университетах».СПб.: Изд. СПбГПУ, 2006, с. 109 – 111. Статья.
19. Максимов Ю.Д. Спектр новых коэффициентов детерминации для анализа парной зависимости случайных величин. Сб. «Математика в вузе» Труды XIX межд. научно-метод. конф. Сентябрь 2006, г. Псков. С.132 – 134. Статья.
20. Максимов Ю.Д. Контингенциальный коэффициент детерминации для исследова ния случайных явлений. Материалы XIV Международной научно-методической

конференции «Высокие интеллектуальные технологии и инновации в образовании и науке». СПб. 14 -15 февраля 2007 г. Изд. СПбГПУ, 2007 г., с.228 – 229. Статья.
21. Максимов Ю.Д. Краткая история теорий корреляции и регрессии. Материалы XIV Международной научно-методической конференции «Высокие интеллектуальные технологии и инновации в образовании и науке». СПб. 14 -15 февраля 2007 г. Изд. СПбГПУ, 2007 г., с.225 – 228. Статья.
22. Максимов Ю.Д. Комбинированный коэффициент детерминации с ассоциативным ядром для дискретных случайных величин. Материалы XI Всероссийской конф. по проблемам науки и высшей школы. «Фундаментальные исследования в технических университетах».СПб.: Изд. СПбГПУ, 2007, с. – . Статья.
23. Максимов Ю.Д**.** Коэффициенты детерминации и корреляции с предельным ядром для непрерывных распределений. Материалы XI Всероссийской конф. по проблемам науки и высшей школы. «Фундаментальные исследования в техниче ских университетах».СПб.: Изд. СПбГПУ, 2007, с. – . Статья.
24. Номоконов М.К. О простоте второго характеристического числа корреляционых интегральных уравнений. ДАН СССР, 1950, т. 72, № 6, с. 1021 – 1024.
25. Пирсон К. Karl Pearson's Early Statistical Papers, Cambridge Univ. Press, London, 1948.
26. Пирсон К. (Pearson K. 1904). On the theory of contingency and its relation to association and normal correlation, Draper's Co. Memoirs, Biometric Series, No. 1, London.
27. Пирсон К. (Pearson K. 1909). On the a new method for determining correlation between a measured character A and a character B , of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. Biometrica 7, 96.
28. Пирсон К. Грамматика науки. СПб., 1911.
29. Практикум по теории статистики. Уч. пособие. / под ред. Шмойловой Р..А. М.: Финансы и статистика, 2001, 416 с.
30. Привалов И.И. Интегральные уравнения. М.-Л.: Объед. научно-техн. изд., 1935, 248 с.
31. Сарманов О.В. Максимальный коэффициент корреляции (симметричный случай). ДАН СССР, 1958, т.120, №4, с.715 – 718.
32. Сарманов О.В. Максимальный коэффициент корреляции (несимметричный случай). ДАН СССР, 1958, т.121, №1, с.52 – 55.
33. Сарманов О.В., Сарманов И.О. Основные типы корреляции, применяемые в гидрологии. М.: Наука, 1983, 200 с.
34. Сарманов О.В. Исследование стационарных марковских процессов методом разложения по собственным функциям. М.: изд. АН СССР. Труды математического института имени В.А. Стеклова, т.60, 1961, с.238 – 261.
35. Сарманов О.В. О монотонных решениях корреляционных интегральных урав нений. Дан СССР, 1946, т.53, № 9, с.781 – 784.
36. Сборник задач по математике для втузов, т.4. Методы оптимизации. Уравнения в частных производных. Интегральные уравнения. /под ред. А.В. Ефимова. М.: Наука, 1990, 304 с.
37. Смирнов В.И. Курс высшей математики, т.4. М.: Гос. Техтеоретиздат, 1953, 804с.

38. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов / Пер. с нем. М.: Финансы и статистика, 1983.

39. Фишер Р. Статистические методы для исследователей. Пер. с англ. М.: Госстатиздат, 1958, с. 268.
40. Харди Г.Г., Литтльвуд Д.Е., Полиа Г. Неравенства. М.: Гос. изд. иностр. литер., 1948, 456 с.
41. Юл (Yule G. U. 1900). On the association of attributes in statistics. Phil. Trans**. A, 194**, 257.
42. Юл (Yule G. U. 1912). On the methods of measuring association between two attributes, J. Roy. Statist. Soc. **75**, 579.

Оглавление

МАТЕМАТИКА
В ПОЛИТЕХНИЧЕСКОМ УНИВЕРСИТЕТЕ

Максимов Юрий Дмитриевич

МАТЕМАТИКА

НОВЫЕ МЕТОДЫ ДЕТЕРМИНАЦИОННОГО И КОРРЕЛЯЦИОННОГО АНАЛИЗА

Дизайн обложки Т.М. Ивановой

Директор Издательства Политехнического университета А.В. Иванов

Лицензия ЛР № 020593 от 07.08.97

Налоговая льгота – Общероссийский классификатор продукции

ОК 005-93, Т. 2; 95 3005 – научная и производственная литература

Подписано в печать 06. 08. 2007. Формат 60×90/16.
Усл. печ. л. 15,75. Уч.-изд. л. 15,75. Тираж 150. Заказ 287.

Отпечатано с готового оригинал-макета, предоставленного автором,
в типографии издательства политехнического университета.
195251, Санкт-Петербург, Политехническая, 29.

Дополнение

§ Лемма о внутренней связи условий независимости компонент двумерной дискретной случайной величины

Коэффициенты детерминации комби (комби-ас, комби-конт, комби-мед) должны обладать свойством: их равенство нулю означает независимость компонент X,Y двумерной случайной величины (X,Y). Это свойство обеспечивается выполнением равенств

$$p_{ij} - p_{i\cdot}p_{\cdot j} = 0; \quad \forall i,j. \tag{1}$$

Выражения в левой части этих равенств входят в структуру коэффициентов, например, коэффициент комби-ас имеет вид

$$com_a = \frac{1}{\sigma_X \sigma_Y} \sum_i \sum_j \left|x_i - m_X\right| \left|y_j - m_Y\right| \left|K_{ij}\right| p_{ij} \,. \tag{2}$$

Здесь

$$K_{ij} = \frac{p_{ij} - p_{i\cdot}p_{\cdot j}}{\sqrt{p_{i\cdot}\left(1-p_{i\cdot}\right)p_{\cdot j}\left(1-p_{\cdot j}\right)}}; \tag{3}$$

$p_{ij} = P\left(X = x_i, Y = y_j\right); \quad p_{i\cdot} = P\left(X = x_i\right); \quad p_{\cdot j} = P\left(Y = y_j\right); \quad i,j = 1,2,\ldots$;

m_X, m_Y – математические ожидания, σ_X, σ_Y – средние квадратические отклонения случайных величин X,Y.

Однако, если при некотором $i = k$ будет $x_k = m_X$ или при некотором $j = l$ будет $y_l = m_Y$, то абстрактно возможно невыполнение равенства (1) при $i = k$ или $j = l$. Тогда нельзя гарантировать независимость случайных величин X, Y.

Подробное исследование случаев $x_k = m_X$ и $y_l = m_Y$ при фиксированных k и l показывает, что равенства (1) выполняются и при этих значениях индексов, если при всех других они выполнялись, то есть независимость случайных величин X, Y гарантируется и в этих случаях, если $com = 0$. Под символом com понимается любой из комбинированных коэффициентов детерминации. Доказательство высказанного утверждения содержится в следующей лемме.

Лемма.

Пусть для вероятностей p_{ij} двумерного дискретного распределения с законом распределения

$$P\left(X = x_i, Y = y_j\right) = p_{ij} \tag{4}$$

справедливы равенства

$$p_{ij} = p_{i\cdot} p_{\cdot j} \tag{5}$$

для $$i = 1,2,...,m; \left(i \neq k\right);\ \ j = 1,2,...,n; \left(j \neq l\right) \tag{6}$$

(m, n могут быть и бесконечными).

Тогда равенства (5) имеют место и для значений $i = k$ и $j = l$.

Вероятности $p_{i\cdot}, p_{\cdot j}$ распределения компонент X, Y находятся по формулам согласованности

$$p_{i\cdot} = \sum_j p_{ij}, \quad p_{\cdot j} = \sum_j p_{ij}\,. \tag{7}$$

Все рассматриваемые вероятности помещены в таблицу 1 распределения.

Таблица1. Распределение двумерной дискретной случайной величины.

$\downarrow X \setminus Y \rightarrow$	y_1	…	y_j	…	y_l	…	y_n	$p_{\lambda\square}$
x_1	p_{11}	…	p_{1j}	…	p_{1l}	…	p_{1n}	$p_{1\square}$
…	…	…	…	…	…	…	…	…
x_i	p_{i1}	…	p_{ij}	…	p_{il}	…	p_{in}	$p_{i\square}$
…	…	…	…	…	…	…	…	…
x_k	p_{k1}	…	p_{kj}	…	p_{kl}	…	p_{kn}	$p_{k\square}$
…	…	…	…	…	…	…	…	…
x_m	p_{m1}	…	p_{mj}	…	p_{ml}	…	p_{mn}	$p_{m\square}$
$p_{\square\nu}$	$p_{\square 1}$	…	$p_{\square j}$	…	$p_{\square l}$	…	$p_{\square n}$	1

Доказательство.

Рассмотрим любую вероятность p_{il} при $i=1,2,...,m; i\neq k$ столбца с номером l таблицы 1 распределения. Из формул согласованности (7) получаем

$$p_{il}=p_{i\square}-\sum_{\substack{j=1\\ j\neq l}}^{n}p_{ij}=p_{i\square}-\sum_{\substack{j=1\\ j\neq l}}^{n}p_{i\square}p_{\square j}=p_{i\square}\left(1-\sum_{\substack{j=1\\ j\neq l}}^{n}p_{\square j}\right)=p_{i\square}p_{\square l}\ .$$

Этот результат означает, что равенства (2) выполняются и для всех $i=1,2,...,m; i\neq k$ и $j=l$. Аналогично получаем при $j=1,2,...,n;\ j\neq l$ (вероятности строки с номером k таблицы 1 распределения):

$$p_{kj}=p_{\square j}-\sum_{\substack{j=1\\ j\neq l}}^{m}p_{ij}=p_{\square j}-\sum_{\substack{j=1\\ j\neq l}}^{m}p_{i\square}p_{\square j}=p_{\square j}\left(1-\sum_{\substack{j=1\\ j\neq l}}^{m}p_{i\square}\right)=p_{\square j}p_{k\square}\ .$$

Это означает, что равенства (2) выполняются при всех $j=1,2,...,n;\ j\neq l$ и $i=k$.

Таким образом все вероятности p_{ij} обладают мультипликативным свойством (5) в столбце с номером l и в строке с номером k таблицы распределения 1, кроме p_{kl}.

Покажем, что и для вероятности p_{kl} свойство (5) выполняется.

$$p_{kl}=p_{k\square}-\sum_{\substack{j=1\\ j\neq l}}^{n}p_{kj}=p_{k\square}-\sum_{\substack{j=1\\ j\neq l}}^{n}p_{k\square}p_{\square j}=p_{k\square}\left(1-\sum_{\substack{j=1\\ j\neq l}}^{n}p_{\square j}\right)=p_{k\square}p_{\square l}\ .$$

Итак, все вероятности без исключения обладают мультипликативным свойством (5). Лемма доказана.

Сравнение коэффициентов ассоциации и контингенции

Коэффициенты ассоциации и контингенции строятся на основе корреляционной таблицы размера 2x2 (таблица 1).

Таблица 1. Корреляционная таблица 2x2 между двумя признаками.

Признаки $X,Y\rightarrow$ $\downarrow$	B	$\overline{B}$	Σ
A	a	b	$a+b$
$\overline{A}$	c	d	$c+d$

Σ	$a+c$	$b+d$	$n=a+b+c+d$

Здесь a,b,c,d – частоты, с которыми n выборочных элементов классифицируются по признакам A и B.

Коэффициент контингенции определяется формулой

$$k_k=\frac{ad-bc}{ad+bc}=\frac{\frac{a}{n}\frac{d}{n}-\frac{b}{n}\frac{c}{n}}{\frac{a}{n}\frac{d}{n}+\frac{b}{n}\frac{c}{n}}. \qquad (1)$$

Коэффициент ассоциации определяется формулой

$$k_a=\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}=\frac{\frac{a}{n}-\frac{a+b}{n}\frac{a+c}{n}}{\sqrt{\frac{a+b}{n}\frac{c+d}{n}\frac{a+c}{n}\frac{b+d}{n}}}=\frac{an-(a+b)(a+c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

У дробей, через которые выражаются коэффициенты k_a и k_k числители равны. Действительно,

$an-(a+b)(a+c)=a(a+b+c+d)-(a+b)(a+c)=ad-bc$.

Сравним квадраты знаменателей этих дробей

$(a+b)(c+d)(a+c)(b+d)=\left[(ac+bc)+(ad+bc)\right]\left[(ad+bc)+(ab+cd)\right]=$

$=(ac+bc)+\left[(ad+bc)+(ab+cd)\right]+(ad+bc)(ab+cd)+(ad+bc)^2\geq(ad+bc)^2$.

Из этого неравенства для знаменателей дробей следует, что

$$k_a\leq k_k.$$

Знак равенства достигается при $b=c=0$. В этом случае $k_a=k_k=1$.

Сравнение l и co

Ядро предельного коэффициента имеет вид

$$l=\frac{P(AB)-P(A)P(B)}{P(AB)+P(A)P(B)}$$

Для случая корреляционной таблицы 2x2 (таблица 1) это ядро принимает вид

$$l=\frac{\frac{a}{n}-\frac{a+b}{n}\frac{a+c}{n}}{\frac{a}{n}+\frac{a+b}{n}\frac{b+c}{n}}=\frac{an-(a+b)(a+c)}{an+(a+b)(a+c)}=\frac{ad-bc}{an+(a+b)(a+c)}.$$

Числители дробей, через которые выражаются l и co одинаковы. Сравним знаменатели.

Знаменатель $l=an+(a+b)(a+c)=a(a+b+c+d)+(a+b)(a+c)=$

$ad+bc+$ положительные слагаемые $\geq ad+bc=$знаменатель co. Отсюда следует, что $l\leq co$. Знак равенства достигается при $a=0$.В этом случае ядра l и co равны (-1). Итак,

$$l\leq co.$$