А. А. ЧУПРОВ

ОСНОВНЫЕ ПРОБЛЕМЫ ТЕОРИИ КОРРЕЛЯЦИИ

О СТАТИСТИЧЕСКОМ ИССЛЕДОВАНИИ СВЯЗИ МЕЖДУ ЯВЛЕНИЯМИ

Издание М. и С. Сабашниковых
1926

БИБЛИОТЕКА
КОММУНИСТИЧЕСКОЙ
АКАДЕМИИ
1076
26

Ленинградский Гублит № 3005/11 л. — 2000 экз.

Государственная Академическая Типография, 9 линия 12

ПРЕДИСЛОВИЕ.

Весною 1924 года я был приглашен прочесть небольшой курс теории корреляции в страховом семинаре при Университете в Кристиании. Курс этот положен в основу настоящего труда. Он ставит себе задачи, несколько отличные от тех, каким обычно служит изложение теории корреляции в статистических руководствах. Я не преследую цели научить применению на практике приемов статистического исследования корреляционной связи и почти не касаюсь калькуляционной техники; исходя из убеждения, что в результате стремительного роста теории корреляции за истекшее тридцатилетие назрела потребность в систематическом пересмотре основных понятий и исходных предпосылок, в более точной постановке проблем и в критическом обзоре приемов их решения, я предлагаю попытку свести в отвечающее таким, именно, запросам слитное целое результаты выполненной в этой области конструктивной работы путем укоренения их в математической теории вероятностей. Строго выдержанный вероятностный подход открывает возможность подвергнуть все отдельные учения планомерной шлифовке, необходимой в целях обеспечения прочности и стройности возводимых построений.

Особое значение я придаю уяснению того, что можно назвать вероятностным априори корреляционных измерений. Лишь отчетливое и выдержанное разграни-

чение априорных искомых и эмпирических данных статистического исследования корреляционной связи способно устранить ту сбивчивость, которая столь затрудняет порою усвоение работ английской школы. Отсюда вытекает избранный мною порядок изложения: я начинаю не со счетной обработки эмпирического материала, а отправляюсь от анализа априорных, в смысле теории вероятностей, величин и их взаимоотношений и через обзор системы случайных значений переменной перехожу к рассмотрению путей оценки априорных искомых по эмпирическим данным. Примыкая при этом в общем к работам английской школы, я перестраиваю их, где требуется, на априорный лад и как бы перевожу их на иной математический язык.

Изложение я ограничиваю проблемами, возникающими при рассмотрении приемов исследования корреляционной — или стохастической (вероятностной), как я предпочитаю называть ее, — связи между двумя случайными переменными. Вопросов, к которым приводит исследование стохастической связи между тремя и более случайными переменными, я касаюсь лишь мимоходом; систематическое их рассмотрение должно составить предмет отдельного труда. Равным образом не останавливаюсь я на вопросах, к постановке которых мы приходим, если одна из переменных носит характер случайной, другая же представляется неслучайною. К изучению связи между случайными и неслучайными переменными применяют обычно те же приемы, которые теория корреляции предлагает для изучения стохастической связи между случайными переменными. Подобный перенос методов исследования, построенных в иных целях, требует большой осторожности. Даже в тех случаях, где он может быть оправдан, внутренний смысл получаемых результатов претерпевает существенные изменения. Теоретическое освещение этой группы проблем, выдвигающейся за последнее время на видную очередь, находится еще в зачатках. Прежде чем явится возможность включить их в сводный и сжатый систематический обзор, необходима интенсивная монографическая разработка. В настоящем труде я вынужден оставить их в стороне, но хотел бы отметить, что последовательное развитие тех общих точек зрения, которые положены в его основу, позволяет и сюда внести систему и порядок.

Основной замысел книги требовал возможно малой ее загрузки техническими деталями не только статистического, но и математического характера. В виду этого я избегаю приводить в тексте подробные выкладки. Читатель, обладающий некоторой математической сноровкой, без труда восполнит опускаемые звенья цепи рассуждений. Для тех же, кто мало привычен к самостоятельным алгебраическим вычислениям, в особом приложении сведено в возможно простой форме то, что может быть надобно для вывода рассматриваемых в тексте формул.

В целях упрощения логического и математического анализа я исхожу от рассмотрения прерывных законов распределения и связи. Это позволяет не считаться с рядом осложнений, отклоняющих внимание от логического существа проблем, а вместе с тем, создается возможность облечь математические рассуждения в формы элементарной алгебры и тем сделать доступным усвоение их хода для статистиков, математическая подготовка которых ограничена курсом средней школы. Дифференциальным и интегральным исчислениями я почти не пользуюсь. Есть, однако, построения, которые невозможно изложить, не предполагая некоторого с ними знакомства: таково учение о так наз. „нормальной корреляции“. Тут читатель, незнакомый с основами исчисления бесконечно-малых, схватит лишь результаты, не будучи в состоянии проследить хода выкладок.

На логических проблемах, связанных с понятием вероятности, я не останавливаюсь. Лично я примыкаю к тому направлению, которое ведет свое начало от Курно и Криса (см. мои „Очерки по теории статистики“). Но математическое изложение я стремлюсь облечь в формы, дозволяющие уместить его без особого напряжения в рамки иных логических воззрений.

Статистическими приемами изучения связи между явлениями пользуются ныне разнообразнейшие научные дисциплины. Исследование по теории статистики может в силу этого по разному намечать круг читателей, на который оно в первую очередь рассчитано. Приспособлять изложение к определенному кругу читателей побуждает ближайшим образом то соображение, что различия в характере

эмпирического материала, с которым имеют дело статистики-обществоведы и статистики-естествоиспытатели, осязательно сказывается на практике применения общих по замыслу методов работы. То обстоятельство, что я не останавливаюсь в настоящем труде подробнее на технической стороне статистического исследования стохастической связи, избавляет от необходимости выбирать между читателями естествоиспытателями и обществоведами. Рассматриваемые ниже вопросы являются общими и тем, и другим.

Я совсем не касаюсь в настоящем труде проблемы подгонки уравнения корреляционной поверхности к имеющимся эмпирическим данным. Отчасти это обусловлено тем, чте методы решения задачи не поддаются пересказу в легко доступных математических формах. Да и разработана проблема не так полно, как вопрос о кривых распределения значений одной случайной переменной; притом на практике вычисление уравнений поверхностей корреляции пока почти не применяется. Решающее значение имело для меня, однако, то соображение, что познавательная ценность такой обработки вызывает живейшие споры. Некоторые представители так наз. „математической“ статистики склонны усматривать в уравнениях кривых распределения и корреляционных поверхностей высшее достижение, доступное статистику. С другой стороны, есть немало статистиков, видящих в подобных вычислениях научно бесплодную забаву любителей играть формулами и математическими знаками. По общему строю настоящего труда логическая оценка результатов, получаемых путем подгонки уравнений корреляционных поверхностей, должна была бы стоять в центре изложения. Но логика этой проблемы не имеет ближайшего касательства к теории корреляции, как таковой: ей место в общей теории статистики, — тем более, что на примере кривых распределения логические вопросы поддаются уяснению несравненно легче и полнее, как в силу их большей разработанности, так, в особенности, в силу того, что лишь по отношению к кривым распределения мы располагаем возможностью опереть общие соображения на более или менее достаточный по охвату материал конкретных данных, подвергнутых счетной обработке.

Обзор литературы, заключающий книгу, не преследует цели дать исчерпывающую библиографию. Его задача — предложить читателю указания, к чему следует обратиться в первую очередь для дальнейшего углубления в заинтересовавшие его вопросы и для ознакомления с их литературой.

Глубокою признательностью обязан я Н. С. Четверикову, без дружеской помощи и неустанных забот которого русское издание моей книги не могло бы увидеть свет.

А. Чупров.

Рим, 21 ноября 1925 г.

ГЛАВА ПЕРВАЯ.

„Математическая" теория корреляции и „элементарные" приемы уловления связи между явлениями.

§ I.

В ряду многообразных задач, которым служит статистическое знание, раскрытие связей между интересующими нас явлениями выдвигается на одно из самых видных мест по своему научному и жизненному значению. Выработке приемов уловления связи статистическая научная мысль издавна уделяет напряженное внимание. Широкое применение статистических способов исследования в области естествознания, которым знаменуются конец девятнадцатого, начало двадцатого века, сопровождалось особым обострением интереса к тем своеобразным зависимостям между явлениями, с которыми имеет дело статистик. Трудами, главнейше, английских биологов и математиков, смыкающихся вокруг Карла Пирсона, набор приемов, которыми статистик располагает для их изучения, начал обогащаться с быстротою, дотоле неведанною. Благодаря участию математиков работа направилась при этом в несколько иное русло, нежели в трудах статистиков-обществоведов, стремящихся решать те же задачи без помощи столь совершенного математического вооружения. В результате новая система „математических" приемов изучения связи, получившая наименование „теория корреляции", стала яблоком раздора между статистиками. Статистики-математики смотрят пренебрежительно свысока на грубые и недодуманные, с их точки зрения, способы работы, которыми пользуются статистики-не-математики. Последние также не остаются в долгу и отвергают „математические" приемы, как научно бесплодную забаву любителей играть в цифры и в математические знаки. Вражда между статистиками-математиками и статистиками-не-математиками, заходящая в некоторых странах (например, в Германии) очень далеко, служит в настоящее время серьезною помехою для успешного роста статистики. Между тем она лишена реальных оснований, и именно в области приемов изучения связей между явлениями, где она проявляется особенно резко, для нее всего менее оправдания: исторически и логически новые „математические" приемы работы тесно примыкают к методам, изобретенным статистиками-обществоведами, представляя их естественное развитие и утончение. Корни тех идей, которые развернулись на наших глазах в современную теорию корреляции, мы находим в трудах пред-

ставителей не-математической статистики. Для того, чтоб обнаружить эти корни, нет даже надобности в глубоких изысканиях. Достаточно окинуть беглым, но непредубежденным взором те способы, какими статистики издавна исследуют связи между явлениями. Подобный смык теории корреляции с приемами работы статистиков-не-математиков имеет значение и для уяснения основ теории корреляции и для устранения того недоверчиво пугливого к ней отношения, которое встречается так часто среди статистиков-не-математиков. Попытаемся в виду этого осуществить спайку в ряде показательных в этом смысле точек. Такая спайка даст нам опору для дальнейших „математических“ построений.

§ II.

Перед нами так называемая *корреляционная таблица*: сведенная в удобо-обозримую форму совокупность чисел, отмечающих, сколько раз в поле зрения наблюдателя повторились всевозможные сочетания различных значений тех двух признаков, связь между которыми подвергается изучению. Излюбленною иллюстрацией служит таблица частостей, с какими за рассматриваемый период времени в пределах избранной для изучения социальной группы женихи разных возрастов вступали в брак с разного возраста невестами. Для наших целей выгоднее, однако, не осложнять задачи конкретными чертами такой специальной проблемы, как возрастный подбор брачущихся. Положим в основу наших формально методологических размышлений данные выполненного мною эксперимента, на подробностях которого нам нет надобности останавливаться; для нас достаточно отметить, что оба рассматриваемых признака — X и Y — способны принимать лишь целые значения от 0 до 18; за этою оговоркою мы можем воображать себе под нашим X и Y что угодно (см. таблицу 1):

Таблица 1.

		Y																		
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
X	0	1	—	1	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—
	1	—	—	3	—	—	3	—	1	—	—	—	—	—	—	—	—	—	—	—
	2	—	2	1	1	—	—	1	—	4	1	1	1	—	—	—	—	—	—	—
	3	—	—	—	—	3	2	1	2	—	—	2	1	—	—	—	—	—	—	—
	4	—	—	—	—	1	—	2	—	1	1	1	—	—	—	—	—	—	—	—
	5	—	—	1	1	1	3	1	—	1	1	1	1	2	—	—	—	—	—	—
	6	—	—	1	1	1	—	1	1	—	2	1	—	—	2	—	—	—	—	—
	7	—	2	2	2	—	—	2	—	—	1	—	3	1	—	1	1	—	—	—
	8	—	—	1	1	1	1	3	—	—	—	4	1	1	—	3	—	—	1	—
	9	—	1	—	—	—	1	3	4	3	2	3	4	4	1	1	1	1	—	—
	10	—	—	2	—	—	1	1	4	4	—	1	2	6	—	2	—	2	—	—
	11	—	—	1	—	—	1	1	1	1	2	—	3	3	2	—	1	—	—	—
	12	—	—	—	1	—	1	1	1	1	2	—	3	1	—	3	—	2	—	1
	13	—	—	—	—	—	1	—	1	1	—	2	—	1	1	—	1	1	1	—
	14	—	—	—	—	—	1	1	2	1	2	1	2	—	—	—	—	1	—	—
	15	—	—	—	—	—	—	—	—	1	1	1	1	—	1	1	—	—	—	—
	16	—	—	—	—	—	—	—	2	—	—	—	1	4	2	—	—	—	—	—
	17	—	—	—	—	—	—	—	—	—	—	2	1	—	—	1	—	—	—	—
	18	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	1	1

Что же говорят нам о связи между X и Y числа, сведенные в корреляционную таблицу? Каким путем можем мы выяснить по ним, есть ли связь между X и Y или обе переменные представляются взаимно независимыми? Если всматриваться в отдельные числа, то открываешь самые разнообразные соотношения между значениями X и соответствующими им значениями Y: бо́льшим значениям X соотвествуют то бо́льшие, то меньшие значения Y. В целом довольно явственно проступает, правда, известный порядок, — некоторый рост значений Y с ростом значений X. Но порядок этот заметен только в целом, да и то легко может быть оспариваем указанием на исключения, — и исключения, порою очень резко бьющие в глаза: так, в одном случае мы видим, что X равняется 0, а Y равняется 6; в то же время наблюдается и такой случай, где X равняется 9, а Y равняется 1. Всматриваясь в единичные числа повторений, сведенные в нашу таблицу, не выносишь целостного и неоспоримого впечатления. Чтоб притти к более уверенному выводу, необходимо подвергнуть эти числа дальнейшей обработке, — привести их к такому виду, чтоб наличие или отсутствие связи поступало с большею четкостью.

К этой цели можно итти разными путями.

А) Основной прием не-математиков заключается в том, что все значения переменной Y, отвечающие определенным значениям переменной X, объединяются в одну группу, а затем подвергаются рассмотрению их средние, расположенные в ряд в порядке возрастания значений X. Если с ростом X средние значения Y правильно возрастают или правильно убывают, то отсюда мы заключаем о наличии связи, прямой или обратной, между X и Y. Если же средние значения Y обнаруживают неправильные колебания, то делается вывод: связи нет или, в более осторожной формулировке, связи не заметно.

В нашем случае мы получаем при такой обработке исходного нашего материала таблицу 2:

Таблица 2.

Значение X .	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Среднее значение Y . .	2,7	4,0	6,25	6,6	7,2	7,1	7,6	7,1	9,1	9,4	9,6	9,75	10,8	11,3	9,0	10,8	11,0	11,25	16

Мы видим, что с ростом X заметно нарастают в целом и средние значения Y. Но и тут опять-таки лишь в целом. Наблюдаются и здесь исключения: значению 7 переменной X отвечает, например, меньшее среднее значение Y, нежели значению 6. Применяя этот прием, — на языке современной теории корреляции мы именуем его „рассмотрение эмпирической линии регрессии переменной Y по переменной X“, — мы сталкиваемся, таким образом, снова с тою же трудностью, какую надеялись преодолеть при его посредстве: чем руководствоваться при решении, колеблются ли средние значения Y неправильно или обнаруживают достаточно явственный уклон вниз или вверх? Ответ дать не трудно, если во всех без исключения случаях последующие значения больше или меньше

непосредственно им предшествующих. Но часто не получается такой целостной картины, а средние значения Y, как в нашем примере, то нарастают, то убывают. Математик, вычисляющий уравнение линии регрессии при помощи метода наименьших квадратов, располагает своими приемами для решения вопроса, в какой мере наблюдающиеся зигзаги эмпирической линии регрессии могут почитаться за случайные скачки, примиримые с допущением, что линия регрессии имеет в сущности вид прямой, параллельной оси X. Не-математик стоит перед тою же задачей, но вынужден искать иных путей ее решения.

Естественно напрашивается мысль подвергнуть ряд средних значений Y повторной сходной обработке и свести их в более крупные группы, так, чтоб в ряде вновь полученных средних уже обнаружилось совершенно правильное нарастание или убывание с ростом X. Если мы разобьем наш ряд (таблица 2) на три части, отнеся к первой значения X от 0 до 5, ко второй — значения X от 6 до 12, а к третьей — значения X от 13 до 18, то получим для Y значения средних: 5,6; 9,05; 11,6. С ростом X правильно возрастают без единого исключения и значения Y; между обеими переменными существует прямая связь. Картину правильного нарастания без единого исключения мы получим и в том случае, если разобьем ряды таблицы 2 не на три, а на пять групп (см. таблицу 3):

Таблица 3.

Среднее значение X	1,5	5,5	9	12,5	16,5
Среднее значение Y	4,9	7,2	9,35	10,2	12,4

Но если мы попытаемся увеличить число групп до девяти, то целостной картины вполне неуклонного нарастания значений Y уже не получим (см. таблицу 4):

Таблица 4.

Среднее значение X.	0,5	2,5	4,5	6,5	9	11,5	13,5	15,5	17,5
Среднее значение Y.	3,3	6,4	7,1	7,4	9,35	10,3	10,15	10,9	13,6

Ясно, что путем такого объединения в крупные группы мы всегда можем получить желанную картину неуклонного движения ряда Y в одном направлении: если мы ограничим число групп двумя, то среднее значение Y для второй из них будет или больше или меньше среднего значения Y для первой или равно ему; вывод, наблюдается ли связь и является ли она прямою или обратною, не представит тут затруднений. Но ясно в то же время, что заключение наше не будет обладать и большим весом, так как слишком значительна будет вероятность того, что подмечаемая нами картина носит случайный характер. Чем больше групп мы в состоянии нарезать, не наталкиваясь ни на одно исключение, тем прочнее

вывод, что найденная связь или подмеченное отсутствие связи не случайны, а свидетельствуют о действительных взаимоотношениях между изучаемыми признаками. Прибегающий к такому приему исследования статистик-не-математик стремится в виду этого разбить свой материал на части так, чтоб выдержанное нарастание или убывание наблюдалось при возможно значительном числе групп. Внимательно вглядываясь в первоначальную линию регрессии (таблица 2), можно обычно при некотором старании подогнать границы образуемых групп так, чтоб уклоны линии регрессии в нежелательную сторону сглаживались в возможно тесных границах. Какое широкое поле открывается тут произволу и даже прямой подтасовке, распространяться нет особой надобности: нередко случается, что ловкою обработкою одного и того же материала можно выжать из него при помощи этого приема прямо противоположные заключения. Чтоб ввести произвол в границы, рекомендуется соблюдать известные правила при сводке материала в группы: образовывать группы, например, так, чтоб они охватывали равные по возможности числа единичных значений, или разбивать скалу значений X на равные по возможности интервалы.

Основная мысль рассматриваемого нами приема воплощается, таким образом, на практике в очень разнообразные формы. Но все это разнообразие направлено к преодолению одной и той же трудности: к выпутыванию из под покрова случайных колебаний смутно проглядывающей через него действительной связи между X и Y. К преодолению той же трудности на основе тех же исходных представлений направлены и усилия математиков, вычисляющих уравнения линий регрессии. Вся разница между математиками и не-математиками заключается в том, что первые с большей систематичностью продумывают исходные положения и строже обосновывают зиждущиеся на них выводы.

В) Другой излюбленный прием статистиков-не-математиков состоит в рассмотрении отклонений обеих переменных от их средних. В основе лежит то соображение, что при полной взаимной независимости X и Y отклонения X в одну какую-либо сторону от средней должны сопровождаться приблизительно равно часто отклонениями переменной Y от ее средней в ту же и в обратную сторону. Если же замечается явственный перевес совпадений или несовпадений знаков отклонений, то это свидетельствует о наличии прямой или обратной связи. Чтоб ближе вникнуть в ход рассуждений, характерных для этой группы приемов, рассмотрим пример, заимствуемый мною из курса статистики проф. Г. Яна[1]) (см. таблицу 5).

Речь идет о связи высоты заработной платы в разных провинциях Норвегии в 1910 и в 1915 гг. Перед нами два ряда цифр: средняя заработная плата поденщика на своих харчах в 18 провинциях Норвегии в 1910 году — ряд X — и средняя заработная плата поденщика на своих харчах в тех же провинциях в 1915 году — ряд Y. Ряд X расположен мною в порядке убывания заработной платы в 1910 году. Первые 8 провинций стояли в 1910 году по уровню заработной платы выше обще-норвежской средней; последние 10 провинций ниже ее. Лишь в одной из первых восьми провинций заработная плата

1) G. Jahn, Statistikkens teknik og metode, p. 224 (Kristiania, 1920).

Таблица 5.

Заработная плата поденщика на своих харчах (в эре).

Провинции	1910 год.	1915 год.	Отклонения от средней	
			1910 год	1915 год.
Finmark	350	445	+ 85	+ 89
Telemark	289	390	+ 24	+ 34
Troms	286	417	+ 21	+ 61
Rogaland	281	396	+ 16	+ 40
Vestfold	280	374	+ 15	+ 18
Vestagder	276	375	+ 11	+ 19
Nordland	272	362	+ 7	+ 6
Buskerud	267	350	+ 2	— 6
Austagder	264	359	— 1	+ 3
Hedmark	263	347	— 2	— 9
N. Trondelag	259	339	— 6	— 17
Hordaland	255	359	— 10	+ 3
Ostfold	249	335	— 16	— 21
S. Trondelag	240	316	— 25	— 40
Akershus	236	317	— 29	— 39
Sogn og Fjordane	233	317	— 32	— 39
More	232	344	— 33	— 12
Opland	231	312	— 34	— 44
Средняя	265	356		

в 1915 году была ниже обще-норвежской средней для 1915 года, в остальных семи она и в 1915 году держалась выше среднего уровня для всей страны. Из десяти последних провинций с заработною платою в 1910 году ниже обще-норвежского среднего уровня лишь в двух заработная плата превышала в 1915 году средний уровень для всей страны, в восьми провинциях она и в 1915 году держалась ниже обще-норвежской средней. Отклонения от обще-норвежских средних совпадают по знаку в 15 случаях из 18; лишь в 3 случаях отклонение X в одну сторону сопровождается отклонением Y в сторону противоположную. Явственно поступает, таким образом, прямая связь между изучаемыми нами величинами. Чтоб дать ей возможно более осязательное выражение, мы можем, по примеру Фехнера[1]), вычислить сводный показатель, разделяя разность чисел совпадений и несовпадений знаков на сумму их, — т. е. на общее число членов в сопоставляемых рядах.

Этотъ Фехнеровский показатель, как нетрудно сообразить, не может быть по абсолютной величине больше 1. Он равняется + 1, если все отклонения совпадают по знаку, т.-е., если между рассматриваемыми величинами наблюдается прямая связь, выраженная с максимальною ясностью. Показатель равняется — 1,

1) G. Th. Fechner, Kollektivmasslehre, s. 382 ff (Leipzig, 1897).

если все отклонения не совпадают по знаку, если каждое отклонение X в одну сторону сопровождается отклонением Y в противоположную сторону, т.-е., если между X и Y наблюдается обратная связь, выраженная с максимальною четкостью. Наконец, показатель равняется нулю, если числа совпадений и несовпадений равны между собою и между величинами X и Y не наблюдается ни прямой, ни обратной связи. Численное значение показателя дает, таким образом, яркое представление о том, что мы стремимся схватить в картине расположения сравниваемых нами рядов. Обозначая этот Фехнеровский показатель через i, находим в нашем примере при ясно выраженной. как мы уже видели, прямой связи

$$i = \frac{15 - 3}{15 + 3} = \frac{2}{3} = + 0{,}67.$$

Знаками отклонений не исчерпывается, однако, то, что можно извлечь из имеющегося в нашем распоряжении цифрового материала для суждения о наличии или отсутствии связи между X и Y. Связь, если она существует, должна находить выражение не только в совпадении или несовпадении знаков (связь прямая и связь обратная), но также и в размерах сопутствующих друг другу отклонений. При полной взаимной независимости не только знак, но и величина отклонения X от средней не будут сказываться на отклонениях Y: большие и небольшие отклонения переменной X от ее среднего уровня будут равно часто сопровождаться и большими, и малыми отклонениями переменной Y от ее среднего уровня как в ту же, так и в противоположную сторону. В случае прямой связи должен, напротив, наблюдаться не только перевес совпадений знаков над несовпадениями, но также сравнительно частое совпадение больших отклонений X с большими же отклонениями Y и малых отклонений X с малыми отклонениями Y. В случае связи обратной большим отклонениям X должны сравнительно часто сопутствовать большие же отклонения Y в противоположную сторону, а малые отклонения X должны сравнительно часто сопровождаться малыми отклонениями Y в противоположную сторону. В случае независимости алгебраическая сумма произведений отклонений X и Y, взятых с соответственными знаками, должна быть недалека от нуля. В случае прямой связи она должна иметь значительную положительную величину; в случае связи обратной она должна иметь более или менее заметное отрицательное значение. В нашем примере сумма положительных произведений оказывается равною 16550 и очень сильно перевешивает сумму отрицательных произведений, равную всего лишь 45. И этот прием исследования подтверждает, следовательно, наличие резко выраженной прямой связи между высотою заработной платы в различных частях Норвегии в 1910 и в 1915 годах. Можно и эти выкладки дополнить, по предложению того же Фехнера, вычислением сводного показателя, вполне аналогичного i. Разделим алгебраическую сумму всех произведений (в нашем случае разность между 16550 и 45) на сумму всех произведений, взятых с положительным знаком (в нашем случае на сумму 16550 и 45). Этот новый показатель по своей абсолютной величине также не может превышать 1 и равняется + 1 при максимально четкой прямой связи, — 1 при максимально ясной обратной связи и 0 при отсутствии сколько-нибудь ясно вы-

раженной прямой или обратной связи. Обозначая его через I, находим для нашего примера:

$$I = \frac{16550 - 45}{16550 + 45} = + 0{,}99.$$

Цифровой материал, которым мы располагаем, позволяет нам пойти дальше. Мы только-что вели речь о том, что при наличии связи большие отклонения одной переменной должны сравнительно часто сопровождаться большими же отклонениями другой, а малые отклонения совпадать с малыми. Что, собственно, имеем мы в виду, говоря о больших, о малых отклонениях? Чем должна измеряться величина отклонений? Общая теория статистики рекомендует, как известно, в качестве меры колебаний переменной около ее средней величины так называемое среднее квадратическое отклонение (см. ниже глава третья, § III). Представляется целесообразным остановиться на этой мере и в данном случае. Тогда мы будем, действительно, сопоставлять друг с другом отклонения, которые строго соответствуют друг другу по своему относительному размаху. Если одна из переменных обнаруживает большие колебания, а другая колеблется очень мало, то сравнительно небольшое отклонение второй должно, с точки зрения взаимного соответствия размахов, расцениваться выше, чем много бо́льшее по своей абсолютной величине отклонение первой. Выражая все отклонения в квадратических ошибках, — мы с этим и считаемся, — приводим все отклонения к сравнимому виду. По отношению к таким сравнимым отклонениям мы можем повторить те рассуждения, из которых только-что исходили, когда рассматривали абсолютные отклонения. Если переменные X и Y взаимно независимы, то алгебраическая сумма произведений соразмеренных друг с другом отклонений должна быть близка к нулю. Она должна быть тем больше по абсолютной величине, чем яснее выражена связь, и должна быть положительною, если связь прямая, — отрицательною, если связь обратная. Обозначим буквою r алгебраическую сумму произведений таких приведенных к сравнимому виду отклонений. Подобно показателям i и I, показатель r (см. ниже глава пятая § V) не может превышать по абсолютной величине 1. Он равняется $+1$, если между сопоставляемыми рядами наблюдается максимально ясная прямая связь; он равняется -1, если наблюдается максимально ясная обратная связь; он обращается в 0, если не наблюдается сколько-нибудь заметной прямой или обратной связи. В нашем примере мы находим

$$r = + 0{,}93.$$

Новый прием исследования подтверждает, как и следовало ожидать, полученный ранее вывод, что между заработной платой разных частей Норвегии в 1910 и в 1915 гг. наблюдается резко выраженная прямая связь.

Показатель r, к которому мы пришли путем постепенного развития мысли, лежащей в основе наипростейшего из приемов уловления связи путем сопоставления отклонений X и Y от их средних, играет очень видную роль в современной теории корреляции: это так называемый коэффициент корреляции. Мы видим, таким образом, что вычисление коэффициентов корреляции непосред-

ственно примыкает к приемам работы статистиков-не-математиков. Тут, как и по отношению к линиям регрессии, мы находим зачатки „математических“ приемов в способах работы, давно придуманных и постоянно применяемых статистиками-не-математиками. Но лишь в системе идей современной теории корреляции получают эти зачатки свое естественное и полное развитие. Лишь на ее почве удается дать законченное обоснование приемам исследования и этого типа.

C) Большим расположением статистиков-не-математиков пользуется издавна прием нумерации членов сопоставляемых рядов. Все члены обоих рядов нумеруются, причем номером 1 обозначается наибольший, номером 2—следующий по величине и т. д. Ряд X располагается затем в порядке восходящих номеров от наибольшего значения X к наименьшему. Наличие или отсутствие связи должно найти себе при этом выражение в том, как расположатся соответственные номера

Таблица 6.

Заработная плата поденщика на своих харчах.

Провинции.	1910 год.	1915 год	Разность номеров: d.	Квадрат разности: d^2.
Finmark	1	1	0	0
Telemark	2	4	— 2	4
Troms	3	2	+ 1	1
Rogaland	4	3	+ 1	1
Vestfold	5	6	— 1	1
Vestagder	6	5	+ 1	1
Nordland	7	7	0	0
Buskerud	8	10	— 2	4
Austagder	9	$8\frac{1}{2}$	+ $\frac{1}{2}$	$\frac{1}{4}$
Hedmark	10	11	— 1	1
N. Trondelag	11	13	— 2	4
Hordaland	12	$8\frac{1}{2}$	+ $3\frac{1}{2}$	$12\frac{1}{4}$
Ostfold	13	14	— 1	1
S. Trondelag	14	17	— 3	9
Akershus	15	$15\frac{1}{2}$	— $\frac{1}{2}$	$\frac{1}{4}$
Sogn og Fjordane	16	$15\frac{1}{2}$	+ $\frac{1}{2}$	$\frac{1}{4}$
More	17	12	+ 5	25
Opland	18	18	0	0

в ряде Y. При вполне ясно выраженной прямой связи между X и Y номера ряда Y будут идти в правильно возрастающем порядке, начиная с номера 1 и кончая наивысшим номером. При вполне ясно выраженной обратной связи номера ряда Y будут итти в правильно убывающем порядке: номеру 1 ряда X будет соответствовать наивысший номер ряда Y, последнему члену ряда X будет соответствовать номер 1 ряда Y. Наконец, при отсутствии связи в расположении номеров ряда Y не будет заметно никакого порядка, ни систематического нарастания, ни систематического убывания. Случаи совершенно ясно выраженной прямой или обратной связи встречаются, конечно, не часто. Так, например, (см. таблицу 6) в рассматриваемой нами иллюстрации те провинции Норвегии, которые стояли в 1910 году на первом и на последнем месте по высоте заработной платы остались на тех же местах и в 1915 году: осталась на

своем месте и провинция, занимавшая в 1910 году седьмое место; провинции, занимавшие в 1910 году 15 и 16 место, характеризуются в 1915 году одинаковой заработной платой и приходятся как раз на 15 и 16 место. Наличие такого числа совпадений при 18 номерах в ряде является веским свидетельством в пользу прямой связи. При помощи теории вероятностей можно бы учесть, сколь мало правдоподобно, чтоб такая картина установилась при отсутствии прямой связи между X и Y. Статистики-не-математики прибегают, однако, к иным способам сделать более наглядным, к какой из трех возможностей — прямая связь, обратная связь, отсутствие связи — приближается действительное расположение номеров ряда Y. Задачу эту и в такой постановке можно решать при помощи приемов, описанных выше под А) и В). Но при таком подходе можно итти к ее решению и своими путями.

Можно, например, разбить ряды на несколько частей и посмотреть, держатся ли номера ряда Y в пределах тех частей, где им надлежало бы стоять в предположении прямой или обратной связи. Разобьем провинции Норвегии на три группы, отнеся к первой группе шесть первых номеров по высоте заработной платы в 1910 году, ко второй номера 7 по 12 по высоте заработной платы в 1910 году и к третьей те шесть провинций, которые занимали в 1910 году наинизшие места. Ни одна из первых шести провинций (см. таблицу 6) не выходит и в 1915 году за пределы шестого места. Отдельные номера ряда Y стоят, правда, не все точно на тех местах, где им надлежало бы стоять при вполне ясно выраженной прямой связи, но ни один из них не перебегает в другие группы. Между второю и третьею группою замечается по одной перебежке: номер 12 в ряде Y перескочил в третью группу, номер 13 перешел в среднюю группу. Но остальные десять номеров остаются прикованными к пределам своей группы. В целом тенденция номеров ряда Y держаться рамок расположения, предуказываемого прямой связью с X, проступает, как видим, с очень большою отчетливостью.

Общему впечатлению от расположения номеров ряда Y можно придать напрашивающееся само собою численное выражение при посредстве особого показателя. Вычислим разности между соответствующими друг другу номерами X и Y. Если связь — вполне ясно выраженная прямая, все разности будут равны нулю. Чем сильнее разности отклоняются от нуля, тем менее ясно выражена прямая связь между X и Y. На величине разностей можно, следовательно, обосновать меру, аналогичную показателям i, I и r. Нужно лишь подыскать удобный сводный показатель для характеристики ряда разностей. Так как алгебраическая сумма разностей тожественно равна нулю при всяком расположении ряда Y, то сводный показатель приходится опереть или на абсолютные величины разностей, или на их квадраты. Остановимся на квадратах разностей. При вполне ясно выраженной прямой связи все разности равны нулю, равна нулю и сумма их квадратов. В случае вполне ясно выраженной обратной связи, при общем числе номеров n, ряд разностей начинается с $1 — n$, затем следует $2 — (n — 1)$ и т. д. Что касается суммы квадратов разностей, то она в таких условиях равняется

$$\frac{n\,(n^2 — 1)}{3}.$$

Наконец, в случае взаимной независимости X и Y — т.-е. в предположении, что ряд Y не обнаруживает ни нарастания, ни убывания, — сумма квадратов разностей получается равною

$$\frac{n(n^2-1)}{6}.$$

Таким образом, если мы, следуя примеру Пирсона, обозначим через ρ величину

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)},$$

то показатель ρ будет принимать значение $+1$ при вполне ясной прямой связи между X и Y, он будет принимать значение -1 при вполне ясной обратной связи и будет обращаться в 0 при взаимной независимости. Чем больше ρ по своей абсолютной величине, тем яснее выражена связь; знак же ρ указывает, наблюдается ли связь прямая или обратная.

В нашем примере заработной платы в 18 провинциях Норвегии (см. таблицу 6) сумма квадратов разностей равняется 65. Показатель ρ получается равным $+0{,}93$. Он оказывается случайно в точности равным полученному нами раньше коэффициенту корреляции. Это случайность, так как в общем случае показатель ρ отличен от коэффициента корреляции r, и взаимоотношения между ними носят весьма сложный характер. Лишь при известных допущениях — например, для случая так называемой „нормальной корреляции“ (см. ниже глава четвертая, § V) — зависимость может быть сведена, как показал Пирсон, к более простым формам. Эта задача отходит уже, однако, неоспоримо, к области ведения „математической“ статистики. Мы снова, таким образом, развивая исходные мысли статистиков-не-математиков, перешли незаметно в область приемов „математических“, — поставили проблему, рассматриваемую в современной теории корреляции под обозначением „корреляция рангов“, и даже вывели основную формулу теории корреляции рангов — формулу Спирмана-Пирсона. И опять-таки никаких скачков, никакого резкого расхождения между тем, что относится к приемам статистиков-не-математиков и приемами статистиков-математиков не наблюдается и тут. Нет между ним принципиального разрыва: нет пропасти, чрез которую было бы трудно перебросить мост. Приемы, величаемые математическими, и здесь отличаются от нематематических лишь тем, что с большей последовательностью развивают те же основные мысли и облекают их в более общие и более точные формы. Математическими, по существу, можно с равным правом признать и те, и другие: с точки зрения „математичности“ разницы между ними не больше, чем между арифметическим вычислением $8.2 = (5+3)(5-3) = 25 - 9 = 16$ и алгебраической выкладкой $(a+b)(a-b) = a^2 - b^2$.

§ III.

Статистики-не-математики подготовили почву для статистиков-математиков не только тем, что наметили в основных чертах ряд приемов уловления той своеобразной связи между явлениями, с которою имеет дело статистическое исследо-

вание. Не меньшее значение имеет то обстоятельство, что благодаря их трудам самое представление об этой связи начало постепенно принимать более отчетливые очертания и вместе с тем создалась возможность планомерного построения приемов ее изучения. Первоначально вопрос шел лишь о том, замечается ли между изучаемыми явлениями связь, или нет. Самое большее, спрашивалось еще, прямая связь или обратная. Затем начали, однако, подмечать, что подвергаемые изучению ряды чисел отличаются друг от друга не только тем, что связь, прямая или обратная, проступает иногда явственнее, иногда менее явственно. Накопляющийся опыт обнаружил, что самая связь не всегда остается одною и тою же, что порою она носит характер более тесной, порою менее тесной. Случайные колебания сопоставляемых рядов всегда перекрывают в известной степени существующую между переменными связь. Связь обнаруживается с тем большею явственностью, чем менее значительны случайные колебания по сравнению с теми изменениями в величине соответствующих друг другу значений X и Y, которые обусловлены наличием связи между X и Y. Первоначально все помыслы были направлены к тому, чтоб искусною обработкою обоих рядов ослабить относительный вес случайных колебаний и тем открыть возможность для связи, буде таковая существует, проступить на их фоне с большею отчетливостью. К этой цели непосредственно и клонились разнообразные приемы уловления связи, изобретавшиеся статистиками-не-математиками. По мере применения этих приемов на практике начало, однако, выясняться, что успех в высокой мере зависит от самого характера уловляемой связи, что, с этой точки зрения, связи представляют в разных случаях весьма резкие различия: встречаются связи такие, которые рельефно проступают даже при малых числах наблюдений и сравнительно больших размерах случайных колебаний; в других же случаях связь становится сколько-нибудь приметною только после того, как размах случайных колебаний будет сведен к минимуму. В итоге таких наблюдений статистическая мысль начала по новому истолковывать результаты, получаемые путем сопоставления рядов. Стало складываться представление о большей или меньшей тесноте, как свойстве самой связи, отличном от большей или меньшей ясности, с какою связь проступает в рассматриваемых рядах. Это знаменовало собою первый решительный шаг на пути к той рационализации приемов статистического изучения связей между явлениями, которая воплотилась в современную „математическую“ теорию корреляции.

ГЛАВА ВТОРАЯ.

Объект и задачи статистического изучения связей между явлениями. Неразрывная причинная связь и корреляция.

1) Представление о бóльшей или меньшей тесноте, как свойстве связи, присущем ей независимо от того, с какою степенью ясности она прорывается в изучаемом статистическом материале сквозь слой перекрывающих ее случайных колебаний, лежит в основе теории корреляции. С этого центрального понятия нам и надо начать анализ системы построений, охватываемых обозначением „теория корреляции“. На первый взгляд представление о связи, более или менее тесной, стоит в непримиримом противоречии с коренным нашим представлением о причинной связи. Причинная связь мыслится, как связь неразрывная: если A причина A', то всегда и везде, где имеет место A, неизменно следует A', и всегда и везде, где наблюдается A', неизменно предшествует A. О бóльшей или меньшей тесноте связи нет при этом речи: или A является причиною A' и неразрывно с A' связано или A не причина A' и ни в какой связи с A' не стоит: tertium non datur, — ничего промежуточного между неразрывной связью и отсутствием связи, казалось бы, быть не может. Откуда же появляются в нашем поле зрения те более или менее тесные связи, с которыми имеет дело статистик?

На сходное противоречие наталкивается исследователь и в тех случаях, когда он ведет работу, не прибегая к статистике. Понятие неразрывной связи исключает, на первый взгляд, все количественные соотношения между причиною и следствием, не сводящиеся к прямой пропорциональности. Если A причина A', то всегда за A следует A', и нигде не встречается A' без того, чтоб предшествовало A. Отсюда, казалось бы, вытекает, что если осуществляются $2A$ (т.-е. $A + A$), то должны следовать $2A'$ ($A' + A'$) и что за kA должны следовать kA'. На деле, между тем, труд естествоиспытателей, работающих в области так называемых точных наук над изучением причинных соотношений между явлениями, поглощается изысканием математических формул законов связей в бóльшей, пожалуй мере, нежели простым констатированием наличия или отсутствия связи. Чем может быть объяснено это противоречие?

Не трудно убедиться, что противоречие здесь только кажущееся. При работе в неизведанной еще области мы никогда не можем поручиться, что измеряются количественные значения причины и ее непосредственного следствия;

измерению постоянно подвергаются величины, с ними лишь так или иначе сопряженные, и нет способа ни оберечь себя от этого, ни удостовериться до исследования, как в этом отношении обстоит дело. Вес, например, завися от массы, устанавливается прямо пропорциональным объему, если рассматриваются тела из одного и того же совершенно однородного вещества. Но допустим, что, имея дело с медными кубиками, мы начнем сопоставлять с весом не объем кубов, а длину ребра. Связь получится вполне определенная, так как длина ребра определяет объем куба. Но закон связи будет уже иной: с ростом ребра вдвое объем, а с ним вместе и вес, возрастают не вдвое, а в восемь (2^3) раз. Предположим, с другой стороны, что измеряется не длина ребра, а площадь одной из сторон куба. Тогда сопоставление с весом даст закон связи еще более сложного вида: с ростом площади в четыре раза вес возрастает в восьмеро, с ростом площади в девять раз вес возрастает в 27 раз. А кто подскажет нам до начала работы, что измерять следует не длины и не площади, а именно объемы, если неизвестно, как что с чем связано в подлежащей исследованию области. Пример этот с ясностью свидетельствуют, что лишь счастливая, но мало вероятная случайность может навести нас при выборе величин, подвергаемых измерению, на такие, которые стоят как раз в отношении прямой пропорциональности. Вероятнее, что мы натолкнемся на функциональную зависимость иного вида. Ибо на один подход к проблеме, дающий в результате измерений картину прямой пропорциональности, обычно имеется на лицо множество таких, при которых прямой пропорциональности не получается. Правда, мы редко идем при исследовании совсем в слепую. Как правило, кое-что о подлежащих изучению явлениях нам уже известно. Это в некоторой мере повышает вероятность удачи при выборе. Но все же лишь как редкое исключение осведомлены мы при начале исследования достаточно, чтоб с уверенностью нацеливаться как раз на величины, стоящие в отношении прямой пропорциональности друг к другу. Отсутствие прямой пропорциональности не может, следовательно, почитаться доказательством отсутствия связи. „Законы“ связи могут облекаться в любые математические формы. Этим и объясняется, что среди законов, которые открывает физик, химик и т. д., прямая пропорциональность — эта простейшая форма функциональной зависимости — занимает столь скромное сравнительно место.

Сходным образом может быть объясняемо появление в поле зрения исследователя связей, не неразрывных, а лишь более или менее тесных. На практике научной работы мы постоянно встречаемся с такими не неразрывными связями: их мы и разумеем, когда ведем речь о корреляции. Существует неоспоримая связь наследственности между свойствами родителей и свойствами потомства. Существует неоспоримая связь между барометрическим давлением и высотою над уровнем моря. Но если начать рассматривать отдельные измерения, на основе которых устанавливается наличие связи, то взору открывается весьма пестрая картина: случается, что у отца великана родится малорослый сын; случается и наоборот, что сын большого роста, а отец карлик; единичные показания барометра в любом месте колеблются в таких широких границах, что в записях метеорологической станции, лежащей на изрядной высоте, можно встретить давления, превышающие некоторые из тех, какие значатся в записях наблюдений

на уровне моря. Как примирить это с допущением неразрывной связи между причиной и следствием?

Предположим, что причина A неразрывно связана со следствием A', что причина B неразрывно связана с следствием B' и т. д. Предположим, далее, что исследователь интересуется взаимоотношениями не между A и A', B и B', а между сравнительно сложными явлениями X и Y, слагающимися из причин A, B и др. и следствий A', B' и т. д. в разных сочетаниях. Возможны тут четыре случая. Если X слагается из A и B, а Y слагается из A' и B', то и сложные явления X и Y будут связаны неразрывно. Но если X слагается из A и B, а Y включает в себя A', B' и, кроме того, C', то зависимость будет носить уже иной характер: X будет всегда предшествовать Y, но не всегда за X будет следовать Y: могут за X следовать и иные сложные явления, включающие в себя A' и B', но отличные от Y в том отношении, что вместо C' в них приходит какое-нибудь иное следствие, — например, D'. Напротив, в том случае, если X включает „лишние“ составные части, которым нет соответствующих в составе Y, то за X всегда будет следовать Y, но у Y будут наблюдаться и иные предшествующие, кроме X. Наконец, как четвертая возможность, мыслим случай, когда X слагается из A и B, а Y слагается из A' и C'; в таких условиях за причиною X будут следовать различные следствия, кроме Y, а следствию Y будут предшествовать различные причины, кроме X. Если приводятся в связь явления, где причина сложнее следствия, то возникает множественность причин. Если следствие сложнее причины, то наблюдается множественность следствий. Если же как причина, так и следствие, имея некоторые элементы, друг другу соответствующие, дополняются сверх того элементами, друг с другом не связанными причинно, то между подобными сложными явлениями устанавливаются взаимоотношения, характеризуемые одновременно и множественностью причин, и множественностью следствий.

В точном соответствии с этими простейшими схемами складываются причинные соотношения в тех примерах, на которых обычно иллюстрируются положения теории вероятностей. Из закрытой урны, в которой лежит один белый и один черный шар, производится извлечение. Какой шар вынется? Ответ: или белый, или черный. Перед нами множественность следствий: рассматриваемый комплекс причин не определяет действия вполне; он оставляет возможными два разных следствия. Бросается шестигранная кость; наверху может оказаться любая из граней куба; у причины не одно, а шесть возможных следствий. Из колоды карт вынимается одна карта; возможных следствий 52.

В чем здесь, собственно, дело? Откуда появляется такая неопределенность следствий? Почему рассматриваемые причины оставляют возможными несколько разных следствий? Корень множественности, как легко убедиться, именно в том, что сравнительно простую причину мы пытаемся привести в связь с более сложным следствием. Из закрытой урны, содержащей шары белые и черные, вынимается один шар. Причина не определяет следствия во всей его полноте: может вынуться и белый шар, и черный. Но дополним в надлежащей мере группу причин, — укажем, как располагаются шары в урне и как движется рука, и неопределенность отпадет: если в глубине урны лежит шар белый, а рука схватывает шар, лежащий

в глубине урны, то вынется, именно, шар белый, а не черный. С таким пополненным комплексом причин следствие стоит в связи, не допускающей уже никакой множественности. С другой стороны, и тот комплекс причин, который оставляет неопределенным, вынется ли шар белый, или черный, имеет свое следствие, которое определяет всецело: именно, вынется из урны шар, а не куб, которого в урне нет, и притом шар белый или черный, а не красный и не синий. Возможность свести все к связям неразрывным всегда есть. В принципе всегда можно устранить множественность причин, упрощая причины или усложняя следствие, а множественность следствий, индивидуализируя причину или обобщая следствия. Часто нас интересуют, однако, не такие неразрывные, но ничего нам в данном случае не говорящие связи, а именно взаимоотношения между выхватываемыми нами частичными комплексами причин и следствий, хотя они и представляются не неразрывными, а лишь более или менее тесными. На практике научного исследования мы и не располагаем, как правило, возможностью выдержанно проводить линию: не приводи в связь того, что в связи не стоит! не вводи в причину обстоятельств, не сказывающихся на следствии! не вводи в следствие элементов, независимых от рассматриваемой причины! Мы в состоянии отшлифовывать понятия, которыми оперируем, по заданному образцу, если строим их планомерно, складывая в более сложные группы простейшие элементы, законы причинной связи между которыми нам уже в точности известны. Если мы знаем, что A причина A', а B причина B', то ничто, разумеется, не мешает подобрать к явлению Y, слагающемуся из $A' + B'$, такую причину $A + B$, которая отвечает требованию неразрывности связи. Но при работе в неисследованной еще области мы не располагаем таким знанием связей между изучаемыми явлениями. Те понятия, которыми мы оперируем, приступая к исследованию, понемногу обтесываются и подвергаются рационализации по мере того, как подвигается вперед работа. Когда же мы ставим вопрос о взаимоотношениях между причиною X и следствием Y, мы не вправе уповать, что X и Y окажутся либо неразрывно связанными, либо вполне независимыми друг от друга. И здесь, как при попытке схватить закон функциональной связи, лишь счастливая, но мало вероятная случайность может навести нас на одну из двух предельных возможностей: на полное соответствие строения X и Y, обусловливающее неразрывность связи, или же на полное отсутствие в составе X и Y связанных между собою причинно частей. Как правило, следует ожидать, что в составе X и Y будут части, соответствующие друг другу, но, помимо них, будут на лицо и части, в связи не стоящие. В силу этого, приступая к работе, мы должны быть готовы к тому, что связь между X и Y, хотя и окажется на лицо, но будет не неразрывною, а лишь более или менее тесною.

2) Бо́льшая или меньшая теснота связи может обусловливаться различиями строения X и Y. Чем меньше относительный вес соответствующих друг другу частей в их составе, тем слабее связь: если X слагается из $A + B + C$, а Y из $A' + B' + D'$, то связь теснее, нежели в том случае, когда X слагается из $A + B + C$, в состав же Y входят $A' + D' + E'$; между $A' + B' + C'$ и $A' + B' + D'$ больше общего, чем между $A' + B' + C'$ и $A' + D' + E'$. Наследственная связь между свойствами отца и свойствами сына теснее, нежели между свойствами деда и свойствами внука. Длина правой и длина левой руки у

одного человека теснее связаны между собою, нежели длина рук у двух братьев: ибо среди условий, сказывающихся на длине руки, больше общих для обеих сопоставляемых величин в первом случае, когда и наследственность, и все условия развития, начиная с утробной жизни, одни и те же, нежели во втором, когда в условиях зачатия и роста могут быть на лицо существенные различия. Еще менее тесною представляется связь между длиною рук у двоюродных братьев, так как в этом случае и часть факторов наследственности оказывается неодинаковою, и условия жизненной обстановки расходятся резче.

Различиями строения дело, однако, не исчерпывается. При одном и том же структурном соотношении между X и Y связь может быть и более, и менее тесною в зависимости от того, как складывается фактически в поле наблюдения исследователя картина колебаний тех частей X и Y, которые не стоят друг с другом в связи. Между объемом шара и его весом устанавливается вполне определенная связь, если изучению подвергаются шары из одного и того же вполне однородного вещества. Но объем шара перестает определять его вес, если не все шары из одного вещества. Чем разнообразнее при этом состав шаров, тем менее тесною окажется связь между объемом и весом: если все шары каменные, то объем теснее определяет вес, нежели в том случае, когда к каменным шарам примешиваются деревянные. Связь между X и Y может, таким образом, в одних условиях наблюдения быть более тесною, в иных менее тесною. Мыслимо даже, что стечение обстоятельств в поле наблюдения будет симулировать несуществующую в действительности связь или перекрывать до неузнаваемости связь, действительно имеющую силу. Цвет шара не оказывает влияния на вес: если посеребрить золотой шарик, он не станет от этого легче. Но если исследованию подвергаются в случайном подборе шары из разных металлов, то непосредственные измерения установят, что желтые шары, при равенстве объема, тяжелее белых. С другой стороны, если случайный подбор материала бросит в поле наблюдения шары каменные малого радиуса и несколько бо́льшие деревянные, то измерения покажут нам, вместо прямой связи между весом и объемом, связь обратную. Наряду со связями причинными — связь объема шара и его веса — исследователь постоянно встречает на своем пути связи беспричинные, „встречи“ во времени и в пространстве независимых друг от друга явлений, — связь веса шара с цветом. Статистические данные устанавливают, например, наличие связи между опустошительностью сельских пожаров и участием пожарной трубы в тушении пожара: для пожаров, которые заливались трубою, опустошительность заметно выше. Но заключать отсюда, что наиболее верное средство борьбы с пожарами, это — взять трубы все и сжечь, было бы не более осмотрительно, нежели обосновывать на статистических сопоставлениях практический совет, делать желтые предметы легче путем перекраски их в белый цвет. Подмеченная статистиками связь обусловлена в действительности тем, что пожарные трубы имеются преимущественно в более крупных селениях и редко привлекаются к тушению пожаров, истребляющих небольшое число строений.

Мы видим, таким образом, насколько задача изучения взаимной зависимости между явлениями сложнее в действительности, нежели представляется тем, кто исходит из допущения, что либо X должен быть связан неразрывно с Y,

либо между X и Y не может быть никакой связи. Задача исследования не исчерпывается решением вопроса, есть ли связь. Если связь есть, необходимо выяснить, что́ это за связь. Должен быть установлен ее „закон“, далеко не всегда сводящийся, как мы видели, к прямой пропорциональности, а могущий принимать форму функциональной зависимости любого вида. Должна быть так или иначе охарактеризована степень тесноты связи. И, наконец, должен быть освещен характер связи: должен быть учтен вес тех „беспричинных встреч“, которые способны порождать видимость связи между X и Y, перестающей сковывать их, как только мы выходим за пределы первоначального поля наблюдения. Задача истолкования подмеченной связи представляется нередко и наиважнейшей, и наитруднейшей. Обрабатывая сходным образом эмпирические данные, которые имеют совершенно одинаковый вид, исследователь приходит в разных случаях к выводам, глубоко различным по их внутреннему смыслу. Один исследователь производит барометрические наблюдения на разных высотах над уровнем моря. Другой исследователь измеряет барометрическое давление на разных расстояниях от берега моря. И тот, и другой получают в результате своих измерений по два ряда чисел, выглядящих совершенно сходно. И тот, и другой применяют к обработке полученных рядов одни и те же приемы. Допустим, что они приходят к одной и той же математической формуле закона связи. При всем сходстве выполняемой ими исследовательской работы, внутреннее содержание и логическая ценность того, что ими устанавливается, коренным образом различны: один открывает закон связи атмосферического давления с высотою; другой же лишь описывает в своеобразной форме рельеф местности, где шла работа, — начни он отступать от той же точки у берега моря вглубь страны в несколько ином направлении, получился бы совершенно иной „закон“.

Правильное истолкование подмечаемой связи представляется особенно существенным, когда статистическое знание привлекается к обоснованию жизненно важных решений и практических мероприятий. Тут знание связей, остающихся без истолкования или неверно истолковываемых, часто хуже полного незнания. Недостаточное внимание к этому обстоятельству является одним из злейших статистических преступлений. Здесь и корень наиопаснейших для статистики нападок на нее. Статистика обнаруживает явственную связь между характером родовспомогательной помощи и долей мертворождений: роды, к которым привлекается врач, дают больше мертворожденных. Плохую услугу оказали бы, однако, людям те благодетели человечества, которые склонили бы рожениц ссылаясь на авторитет статистики, не обращаться к врачу при трудных родах. Статистические данные раскрывают для России яркую обратную связь между высотою горимости и урожаем: чем выше урожай, тем ниже горимость; в годы же плохих урожаев Россия выгорает сильнее. С. И. Устинов, обнаруживший эту интересную зависимость, полагал, что „причина кроется в значении... урожая, как главного экономического фактора народной жизни страны“. Если бы это было так, то мерами к поднятию сельскохозяйственной техники можно бы надеяться понизить горимость: повышая народно-хозяйственный доход от земледелия, агрономическая помощь крестьянам сулила бы вместе с тем сократить потери, которые народное имущество терпит у нас от пожаров. Более вероятным представляется,

однако, что подмеченная С. И. Устиновым связь является лишь отражением неодинакового влияния засухи на урожай хлебов и на горимость зданий. В таком случае меры агрономической помощи крестьянству улучшат сборы, но едва ли скажутся на пожарах. Исследование влияния метеорологических условий на урожай, выполненное Ц.С.У. под руководством Н. С. Четверикова, обнаружило для некоторых местностей любопытную обратную связь между урожаем озимых и количеством осадков, выпадающих за несколько недель до посева. Объяснения можно искать или во влиянии влаги на качество дозревающих семян, или же в воздействии осадков на почву. Если бы дальнейшие статистические и агрономические изыскания подтвердили первое объяснение, то заботою о снабжении крестьян лучшими семенами в годы с большим количеством осадков в критический период можно было бы поднять урожаи. Если же правильно второе объяснение, помыслы об улучшении урожаев должны быть направлены совсем в иную сторону.

3) Необходимость считаться с возможностью, что связи между изучаемыми явлениями окажутся не неразрывными, а лишь более или менее тесными, приводит не только к расширению круга задач исследователя, но и к перестройке самых приемов исследования. Те способы уловления связи, которые намечаются логиками под именем методов индукции, здесь явно неприменимы. Они покоятся всецело на допущении, что связь между X и Y может быть лишь неразрывною. К выводу, что следствие Y связано с данною именно причиною X, они идут путем исключения причин Z, U, W, подбирая такие случаи, где либо Z, или U, или W на лицо, а нет Y среди следствий, либо Y на лицо, а нет Z или U, или W среди причин. Аргументация теряет силу, если связь между Y и X может быть и не неразрывною. Для уловления таких не неразрывных, а лишь более или менее тесных связей, индуктивные методы разницы, сходства, сопутствующих изменений непригодны. Для этих целей должны быть изысканы иные способы работы.

Ясно, притом, что новых приемов не построишь на отрицании, — на том свойстве подлежащих уловлению связей, что они не неразрывны. Зацепкою для исследователя могут служить лишь положительные свойства этих не неразрывных связей. В первую очередь приходит, естественно, на мысль то их свойство, что они могут быть более или менее тесными. Измерение тесноты связи выдвигается в силу этого на центральное место в теории приемов, направленных на изучение не неразрывных связей. В тех наипростейших условиях, которые воспроизводят структурную схему, рассмотренную выше, практичною характеристикою тесноты связи может служить математическая вероятность (см. мои „Очерки по теории статистики“, Очерк третий). В более общем случае не представляется удобным использовать в этих целях непосредственно математическую вероятность. Необходимы более сложные приемы измерения тесноты связи. Представляется возможным, как увидим ниже, использовать и иные свойства не неразрывных связей, менее бросающиеся в глаза. Логическою основоюприемов изучения таких не неразрывных связей, охватываемых общим понятием „корреляции“, является понятие стохастической связи между случайными переменными, опирающееся на понятие математической вероятности и облекающее в точные математические формы намеченное выше представление о не неразрывной связи. С анализа понятий „случайная переменная“ и „стохастическая связь“ мы начнем обзор основ теории корреляции.

ГЛАВА ТРЕТЬЯ.

Стохастическая связь и функциональная зависимость.

§ 1.

Краеугольным камнем, на котором зиждется здание математической теории корреляции, является понятие стохастической связи между случайными переменными.

Исходя из традиционного определения математической вероятности как отношения числа благоприятных равновозможных статочностей ко всему числу равновозможных несовместных статочностей, будем называть *случайной переменной k-ого порядка* величину, способную принимать k разных значений с определенными вероятностями. Совокупность возможных значений случайной переменной и присущих им вероятностей будем называть: *закон распределения значений случайной переменной.* Число очков, выпадающих на подбрасываемой кости, представляется случайною переменною шестого порядка, так как может принимать шесть разных значений 1, 2, 3, 4, 5, 6 с равными вероятностями в 1/6.

Понятие случайной переменной может быть подведено под общее понятие переменной величины как genus proximum, причем наличие закона распределения выступает как differentia specifica. Однозначное число является переменною, способною принимать десять разных значений от 0 до 9. Оно приобретает характер случайной переменной десятого порядка, если каждому из этих десяти значений присваивается определенная вероятность. Закон распределения складывается при этом по разному в зависимости от условий, определяющих вероятности отдельных значений. Допустим, что условия задачи приводят к равенству всех десяти вероятностей; тогда закон распределения будет сводиться к десяти возможным значениям 0, 1, ... , 9 и к десяти их вероятностям в 1/10. Число белых шаров в группе двадцати шаров — переменная, способная принимать 21 различное значение от 0 до 20. Оно приобретает характер случайной переменной 21-го порядка, если добавляется, что двадцать шаров извлекаются из закрытой урны, содержащей в равных числах белые и не белые шары, и что вынимаемый шар бросается назад в урну, прежде чем производится следующее извлечение: ибо в таких условиях каждому числу белых шаров среди двадцати вынувшихся шаров соответствует определенная вероятность, вычисляемая по

известным формулам теории вероятностей. Число белых шаров в группе двадцати шаров приобретает характер случайной переменной 21-го порядка с иным законом распределения значений, если дело идет о двадцати извлечениях из закрытой урны, содержащей шары белые и не белые в иной пропорции, как равно и в том случае, когда числа белых и не белых шаров в урне равны между собою, но вынимаемый шар не бросается обратно в урну до производства следующего извлечения.

Пример извлечений из урны хорошо иллюстрирует не только самое понятие случайной переменной, но и его значение для хода научной работы, как она складывается на практике. Исследователь часто имеет дело с материалом, добываемым способами, которые более или менее близко уподобляются извлечению шаров из закрытой урны. При выборочных исследованиях постановка отбора изучаемых экземпляров нередко заботливо воспроизводит технику извлечения шаров из закрытой урны с возвращением вынутого шара обратно в урну или без возвращения. Естествоиспытатель извлекает из глубины морской небольшие пробы планктона и пытается по взятым образцам морской фауны мелких организмов составить представление о содержании той необъятной „урны“, из которой они почерпнуты в слепую. Врач подсчитывает под микроскопом кровяные шарики в капле разведенного во много раз образца крови пациента, стремясь установить по найденным числам существенные для диагноза свойства крови больного. Числа красных и белых кровяных шариков в поле зрения микроскопа отвечают определению понятия случайной переменной с тою же строгостью, как числа белых и не белых шаров в примере извлечений из закрытой урны.

Случаями, когда исследователь преднамеренно прибегает к взятию образцов, не исчерпывается значение понятия случайной переменной для научной работы. Нередко приходится ставить работу на почву рассмотрения образцов и помимо воли. Ботаник находит новый цветок, срывает его и подсчитывает число лепестков. Устанавливаемое им число носит совершенно тот же характер, как если бы он вынул билет из закрытой урны с большим количеством билетов, на которых написаны разные числа, и прочел начертанное на нем число. Подсчет лепестков на втором экземпляре цветка может дать то же число, но может привести и к иному результату. Отдельные экземпляры одного и того же цветка не всегда ведь имеют одно и то же число лепестков. У сирени преобладают цветки с четырьмя лепестками, но попадается порою и „счастье“ о пяти лепестках, попадаются и цветки с тремя лепестками. Проф. Шарлье[1]) подсчитал числа лепестков у 321 экземпляра Trientalis europaea из окрестностей Лунда и нашел почти у половины по шести лепестков; более одной трети экземпляров имели по пяти лепестков, более $^1/_{12}$ общего числа экземпляров имели по семи лепестков, а у двух цветков оказалось даже по 9 лепестков. Н. С. Четвериков[2]) подсчитал числа цветков ландышей на отдельных стеблях для 1158 экземпляров, собранных им весною 1911 года в Богородском уезде, Московской губ.: в большинстве случаев на стебле было 6 цветков, но часто встречались стебли,

1) См. E Czuber. Die statistischen Forschungsmethoden, S. 115—116 (Wien, 1921).

2) См Л. К. Лахтин. Кривые распределения и построение для них интерполяционных формул по способам Пирсона и Брунса, стр. 114—115 (Москва, 1922).

на которых было по пяти или по семи цветков, а попадались и такие, где было всего три или целых десять и даже одиннадцать цветков. Таким образом, если бы даже наш ботаник открывший новый цветок, произвел подсчеты не для одного экземпляра, а для многих сотен, он тем не менее имел бы дело с извлечениями из неисчерпаемой урны билетов с написанными на них номерами, — он располагал бы пробами многочисленными, но все же оставался бы при рассмотрении проб.

Можно пойти и далее. Допустим, что так или иначе нашему ботанику удалось бы пересчитать числа лепестков на всех без исключения экземплярах изучаемого им растения, цветущих в данное время на поверхности земного шара. Его исчерпывающее в пространстве исследование все же имело бы дело лишь с обращиками, но с обращиками в ином смысле: в смысле сменяющихся поколений. Строго говоря, во всех подобных случаях исследователь изучает лишь пробы, — на подобие того, как врач при счете кровяных шариков под микроскопом рассматривает лишь обращик пробы крови пациента.

Последние примеры приводят нас к углубленному пониманию значения понятия случайной переменной для научной работы. Случайные переменные могут не только играть служебную роль. Они могут представлять и самостоятельный интерес, — быть сами по себе предметом исследования, привлекать к себе внимание, как таковые. Число лепестков Trientalis europaea имеет характер случайной переменной, принимающей всего чаще значение 6, но способной принимать и иные значения, вероятности которых убывают по мере отклонения их от наиболее частого значения. Рост новобранца может принимать любые значения в довольно широких пределах с вероятностями, которые поразительно правильно убывают по мере отклонения роста от его наивероятнейшей величины, оставаясь все время равными для равных отклонений вверх и вниз. Такого рода случайные переменные входят весьма заметною составною частью в круг объектов изучения разных научных дисциплин. На вопросе, чем может объясняться их происхождение, я не буду останавливаться, так как это слишком отвлекло бы нас от наших прямых задач. Для наших целей достаточно убедиться, что случайная переменная не представляется праздным измышлением математиков, забавляющихся игрою в ни к чему ненужные построения, — что научное исследование наталкивается на нее чуть не на каждом шагу, при чем порою случайная переменная имеет лишь служебное значение, иногда же вступает как непосредственный предмет научного интереса.

Двойная роль случайной переменной в научном обиходе — то обстоятельство, что порою она рассматривается как переходный этап к конечной цели исследования, а порою сама является конечною целью научных изысканий, — имеет чрезвычайно существенное значение для теории статистики. Представляется, в виду этого, целесообразным рассмотреть еще несколько примеров, освещающих эту двойственность с новой стороны.

Одним из путей, которыми случайная переменная проникает в поле зрения исследователя, является сопряженное с погрешностями измерение. Допустим, что мы забыли начатки геометрии и видим в вопросе о сумме углов треугольника задачу, подлежащую решению не при помощи умозрения, а посредством действительных промеров. Для ряда треугольников вымеряются углы, для каждого

из них подсчитывается сумма всех трех углов. Истинная величина суммы равняется, как мы знаем из геометрии, неизменно во всех случаях 180 градусам. Но сумма углов треугольника, устанавливаемых путем измерения, отнюдь не представляется постоянною величиною, неизменно равною 180 градусам: значения отдельных сумм обнаруживают колебания, — то превышают 180 градусов, то не доходят до 180 градусов. Сумма, получаемая в результате измерений, носит характер величины переменной, способной принимать разные значения, и принимать их с определенными вероятностями; если измерения производятся с должной тщательностью одним и тем же способом, — она является случайною переменною в смысле данного нами определения понятия.

Предположим далее, что измеряется рост Аполлона Бельведерского. То, что интересует исследователя, не представляется тут переменною величиною; это вполне определенная постоянная величина, остающаяся все время неизменною. Но повторные измерения не будут давать неизменно одной и той же величины, точно совпадающей с истинным ростом Аполлона. Вследствие неизбежных погрешностей измерений мы будем получать величину, отклоняющуюся от искомой то больше, то меньше, то в одну, то в другую сторону. Рост, который дается измерением, носит характер случайной переменной, при чем закон распределения значений определяется техникой измерения и степенью ловкости исследователя. Чтобы добраться до искомой величины истинного роста приходится подвергать изучению эту случайную переменную. Лишь путем целесоответственной обработки ряда ее значений можем мы составить суждение об истинном росте Аполлона. Но случайная переменная не является здесь предметом исследования. И она, и закон распределения ее значений нужны лишь ради того, чтоб оценить возможно точно величину той постоянной, которая одна только и интересует нас, и чтоб иметь возможность прикинуть степень надежности нашей оценки.

Сходным образом в случае, если какой-либо молодой человек в день своего совершеннолетия пожелает увековечить свой рост в этот памятный момент жизни и произведет с помощью приятеля ряд измерений, то рост, устанавливаемый единичными измерениями, будет принимать то одно, то другое значение и будет носить характер случайной переменной. Но случайная переменная не будет и тут предметом исследования и непосредственного интереса; она нужна лишь для того, чтобы от нее перейти к интересующей друзей постоянной величине, — к росту одного из них в день, когда ему исполняется 21 год.

Предположим, с другой стороны, что дело идет не о росте одного какого-либо молодого человека в день его совершеннолетия, а о росте норвежца, англичанина, француза на 21-ом году жизни и что ради этого производится ряд измерений роста новобранцев. Тут уже не только величина, получаемая в результате измерения, но и величина, подвергаемая измерению, представляется случайною переменною. Истинный рост новобранца норвежца не остается одним и тем же для всех норвежцев; он колеблется в довольно широких пределах и распределение его значений следует вполне определенному закону — так называемому закону Гаусса-Лапласа. Истинный рост новобранца носит характер случайной переменной в точном смысле нашего определения. Вместе с тем и рост, устанавливаемый измерением, носит характер случайной переменной, закон

распределения значений которой определяется тут уже не только техникою измерений и ловкостью измерителей, как ранее, но вместе с тем и законом распределения значений истинного роста. Одна из этих случайных переменных — рост, устанавливаемый измерением, — играет и тут служебную роль, но искомым является уже не численное значение некоторой постоянной величины, а вторая из случайных переменных — истинный рост — и закон распределения ее значений.

§ II.

1) Понятие стохастической связи, занимающее центральное место в теории корреляции, опирается на понятие случайной переменной. Оно должно быть строго отграничиваемо от понятия функциональной зависимости. Если, по определении значения X, Y сохраняет характер случайной переменной, способной принимать разные значения с определенными вероятностями, то Y стоит в *стохастической связи* с X. Если же, по определении значения X, переменная Y перестает быть случайною переменною, а становится способна принимать лишь одно определенное значение или несколько разных значений, которым не присущи определенные вероятности, то Y находится в функциональной зависимости от X. Сумма чисел, выбрасываемых на двух разного цвета костях, стоит в стохастической связи с числом, выпадающим на белой и красной кости, ибо при любом значении числа очков на одной из костей, она может принимать с равными вероятностями в $\frac{1}{6}$ шесть разных значений, смотря по тому, что выпадет на другой кости: если на одной кости выпадает 3, то сумма может равняться и 4, и 5, и т. д. до 9; если на первой кости выпадает 5, сумма может равняться 6, 7 и т. д. до 11. Напротив, число очков, выпадающих на кости, и квадрат этого числа находятся в функциональной зависимости, так как при данном числе выпавших очков квадрат этого числа величина не переменная, а строго определенная: если $X = 4$, а $Y = X^2$, то $Y = 16$. В функциональной зависимости от числа очков, выпадающих на кости, находится и корень квадратный из этого числа; если выпадает на кости 4, то корень квадратный из 4 может быть, правда, равен и $+2$ и -2, но ни тому, ни другому значению не присуща определенная вероятность; какое из двух возможных значений придается корню квадратному, решается не случаем, а определяется соображениями, к случаю и к вероятности не имеющими касательства.

Если X случайная переменная, а Y находится в функциональной зависимости от X, то Y также представляется случайною переменною, поскольку рассматривается вне зависимости от X. Если X означает число очков, выпадающих на кости, а $Y = X^2$, то Y может принимать шесть разных значений 1, 4, 9, 16, 25 и 36 с равными вероятностями в $\frac{1}{6}$. Но при данном значении X, Y перестает быть случайною переменною величиною, а получает характер постоянной величины: при X, равном 3, Y равняется 9; при X, равном 5, Y равняется 25.

Понятие стохастической связи может быть распространено на любое число переменных. Если по установлении значений X, Y, Z и т. д. переменная T сохраняет характер случайной переменной и может с определенными вероятностями принимать некоторое число разных значений, то T стоит в стохасти-

ческой связи с переменными X, Y. Z и т. д. Если же переменная T по установлении значений X, Y, Z утрачивает способность принимать разные значения с определенными вероятностями, то переменная T находится в функциональной зависимости от X, Y, Z. Пусть, например, T обозначает сумму чисел, выпадающих на трех разного цвета костях; пусть X обозначает число, выпадающее на белой кости; пусть Y обозначает число, выпадающее на красной кости. Переменная T стоит в стохастической связи с переменными X и Y, так как и при данных числах очков, выпавших на этих двух костях, T может с равными вероятностями в $^1/_6$ принимать шесть разных значений, смотря по тому, выпадет ли на третьей кости 1, 2, 3, 4, 5 или 6.

Весьма существенным представляется то обстоятельство, что переменная T может стоять в стохастической связи с рядом переменных X, Y, Z в отдельности и находиться в функциональной зависимости от этих переменных, вместе взятых. Пусть T означает сумму чисел, выпадающих на красной и на белой кости; T стоит, как мы видели, в стохастической связи и с числом, выпадающим на красной кости, и с числом, выпадающим на белой кости. Но с совокупностью обоих чисел T уже не стоит в стохастической связи; при данных значениях того и другого числа оказывается однозначно определенною и величина T: если на красной кости выпадает 1, а на белой кости выпадает 6, то T равняется 7 и никаких иных значений, кроме семи, принимать не может, — от X и Y, вместе взятых, T находится в функциональной зависимости.

Такого рода взаимоотношения между тремя переменными поучительно сопоставить с случаем функциональной зависимости переменной T от двух взаимно независимых не случайных переменных X и Y. Температура кипения раствора поваренной соли в воде находится в функциональной зависимости от концентрации раствора и от давления на поверхность: данными значениями X и Y однозначно определяется значение T. Наблюдается также функциональная зависимость между температурою кипения и давлением при неизменной концентрации раствора и между температурою кипения и концентрацией раствора при неизменном давлении. Но между температурою кипения и давлением при остающейся неопределенною концентрации раствора нет никакой уловимой зависимости. На вопрос, при какой температуре закипает раствор соли в воде под давлением в одну атмосферу, нет разумного ответа. Вопрос бессмыслен, если не добавляется указаний касательно концентрации раствора. Такие дополнительные указания могут быть даны в виде точного численного определения насыщенности раствора. Тогда мы возвращаемся к функциональной зависимости T от обоих определяющих ее факторов — от концентрации раствора и от давления. Но мыслимо также, что указания даются в форме, при которой концентрация раствора получает характер случайной переменной с определенным законом распределения значений. Пусть, например, спрашивается, при какой температуре закипает под давлением в одну атмосферу раствор соли, взятый на удачу из имеющегося в лаборатории запаса. Такой вопрос лишен научно ценного содержания, но не бессмыслен по существу, если только в лаборатории действительно запасены растворы разной концентрации. Ответ на него должен гласить: температура кипения представляется случайною переменною с таким-то, возможно точно

указываемым, законом распределения значений. И далеко не во всех условиях такая постановка вопроса представляется лишенною научно ценного содержания. На практике научного исследования нередко приходится иметь дело с такими случайными позаимствованиями на удачу из „лабораторного“ запаса окружающей нас природы. С случайными переменными такого происхождения мы встречаемся порою даже там, где в этом нелегко дать себе отчет. Атомный вес свинца всякий химик укажет, не колеблясь, с большой точностью. Между тем мы знаем ныне, что свинец, атомный вес которого нам называет химик, представляется смесью веществ с разными атомными весами. Атомный вес свинца, приводимый в руководствах по химии, является в сущности случайною переменною того именно типа, как температура кипения раствора соли, заимствованного на удачу из лабораторного запаса. Разница лишь в том, что разные составные части обычного свинца встречаются в земной коре в почти неизменной пропорции, в силу чего атомный вес смеси и получает видимость величины постоянной. Успехи знания за последние десятилетия выяснили, что с подобными случайными переменными, преображенными в лже-постоянные величины, научные изыскания имеют дело несравненно чаще, чем думалось еще недавно.

2) Стохастически связанные переменные проникают в поле зрения исследователя теми же путями, как и единичные случйные переменные. Как и последние, они могут являться переходным этапом к конечной цели, не представляя непосредственного интереса для исследователя сами по себе, но могут также составлять конечную цель изысканий.

Переменные не-случайные, находящиеся в функциональной зависимости, преображаются в случайные переменные, стоящие в стохастической связи, при сопряженных с погрешностями измерениях. Температура кипения чистой воды и давление на поверхность находятся в функциональной зависимости: данному давлению отвечает определенная температура; данной температуре соответствует определенное давление. Но между результатами измерений температуры кипения и соответствующих им давлений наблюдается не функциональная зависимость, а стохастическая связь: если произвести достаточно измерений, то окажется, что одно и то же измеренное давление сопровождается то более высокой, то более низкой температурой кипения, а одной и той же установленной измерением температуре отвечает далеко не всегда одно и то же давление.

При исследовании зависимости температуры кипения воды от давления производство измерений сходно сказывается на обеих переменных: устанавливаемые измерением значения оказываются сопряженными с погрешностями и для давлений, и для температур. Бывает и так, что лишь для одной из переменных исследователь не располагает точными значениями, для другой же ему известны с полной точностью ее истинные значения. Допустим снова, что нами забыты начатки геометрии, и предположим, что требуется экспериментальным путем открыть, в какой зависимости находится сумма углов многоугольника от числа сторон. Перемеряются углы для большого числа треугольников, четырехугольников, пятиугольников и т. д. Для каждого измеренного многоугольника подсчитывается сумма углов. Суммы эти для каждого рода многоугольников не будут неизменно равны тому числу градусов, какое выводится для них путем умозрения, а будут

более или менее колебаться около истинного значения. Напротив, числа сторон могут быть при некотором внимании установлены вполне точно. В этом примере истинным значениям одной переменной противостоят совокупности обладающих определенными вероятностями значений другой переменной. В этом весьма существенное отличие от связи между температурою кипения воды и давлением. Но общим обоим примерам представляется, что и тут, и там исследователя интересует лишь функциональная зависимость между не случайными переменными, хотя ему и приходится подвергать рассмотрению переменные, стоящие в стохастической связи.

Предположим, с другой стороны, что изучению подлежит зависимость между ростом новобранцев и объемом груди или между ростом отцов и ростом сыновей. Тут не только между данными, полученными в результате измерений, наблюдается стохастическая связь, но и между величинами, подвергаемыми измерению, существует не функциональная зависимость, а стохастическая связь: одному и тому же росту соответствует то больший, то меньший объем груди не только в записях наших измерений, но и в действительности; среди измеряемых лиц встречаются высокорослые со впалой грудью, встречаются и карапузы с отлично развитою грудною клеткою. Сыновья одного и того же отца часто бывают очень разного роста. Если бы мы даже вместо сопряженных с ошибками измерений могли располагать вполне точными истинными значениями переменных, мы и тогда не имели бы перед собою функциональной зависимости.

Мы видим, таким образом, что при исследовании переменных, стоящих в стохастической связи, задачи могут ставиться по разному. Иногда стохастическая связь является лишь покровом, за которым скрыта интересующая нас функциональная зависимость. Иногда же изучению подлежит стохастическая связь, как таковая. К преследуемой цели исследования должны приспособляться и приемы работы. Если дело идет о том, чтоб посредством изучения стохастической связи добраться до закона функциональной зависимости между интересующими исследователя величинами, то исследователь прибегает к приемам, аналогичным тем, какие по отношению к одной переменной носят название „сглаживания“. Если же ставится задача научного освещения стохастической связи между случайными переменными, то в ход идут приемы, относящиеся к области ведения теории корреляции. На вопросе, что следует, собственно, разуметь под научным освещением стохастической связи, мы остановимся подробно в следующей главе. Теперь наметим лишь вкратце главные черты задачи с тем, чтоб до более подробного рассмотрения отдельных приемов вернуться еще раз к принципиальному вопросу о взаимоотношениях между приемами, уподобленными нами сглаживанию, и теми, которые изучаются теорией корреляции.

§ III.

Выяснение задач научного освещения стохастической связи между случайными переменными должно исходить из существа понятий случайной переменной и стохастической связи.

По отношению к одной случайной переменной, взятой отдельно, все, что так или иначе способно характеризовать ее, может почитаться известным, если

дан закон распределения ее значений. Все прочее может быть выведено из закона распределения путем — не всегда, правда, легких — математических выкладок. Целью исследования может, в силу этого, почитаться нахождение совокупности чисел, которые образуют закон распределения: всех возможных значений переменной и присущих им вероятностей. В такой форме знания наши касательно переменной представляются, однако, при всей их полноте, слишком громоздкими и трудно обозримыми. Чтобы использовать их в научной и в жизненной работе, требуется придать им более сжатое выражение. За исключением тех случаев, когда законы распределения носят самый несложный характер, лишь вычисление сводных характеристик позволяет выжать из них в практичной форме то содержание, которое представляет интерес для исследователя.

Среди сводных характеристик закона распределения значений одной случайной переменной наибольшую ценность представляют, как известно, математическое ожидание переменной и ее среднее квадратическое отклонение (средняя квадратическая ошибка). *Математическим ожиданием* случайной переменной называют среднюю всех возможных значений переменной, взятых с весом, равным их вероятностям. Пусть случайная переменная X способна принимать k разных значений $x_1, x_2, \ldots, x_k$ с вероятностями $p_1, p_2, \ldots, p_k$; условимся обозначать математическое ожидание переменной буквою E, помещаемой перед обозначением переменной; математическое ожидание X может быть в таком случае определено формулою

$$EX = p_1 x_1 + p_2 x_2 + \cdots + p_k x_k = \sum_i p_i x_i.$$

Средняя квадратическая ошибка X, обозначаемая обычно греческою буквою σ с припискою справа внизу обозначения переменной, определяется формулою

$$\sigma_X = \sqrt{E\{X - EX\}^2} = \sqrt{\sum_i p_i [x_i - EX]^2}.$$

Вторую степень средней квадратической ошибки X я буду называть *рассеяние* переменной X.

Математическое ожидание X характеризует тот средний уровень, кругом которого группируются разные значения случайной переменной. Средняя квадратическая ошибка и рассеяние X характеризуют размах колебаний отдельных значений кругом их среднего уровня, определяемого математическим ожиданием X. Рассеяние, отличное от нуля, представляется одним из характерных свойств случайной переменной. Рассеяние может равняться нулю, лишь если все значения переменной тожественны, т.-е. если перед нами не случайная переменная, а величина постоянная.

Чтобы пояснить понятия математического ожидания и рассеяния на примерах, с которыми нам часто придется иметь дело, предположим, что X означает число очков, выпадающих на кости. Так как все шесть возможных значений X равновероятны, то

$$EX = \frac{1}{6}[1 + 2 + 3 + 4 + 5 + 6] = 3{,}5.$$

$$\sigma_X^2 = \frac{1}{6}\left[\frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4}\right] = \frac{35}{12}.$$

Предположим далее, что T означает сумму чисел, выпадающих на двух костях; сходным путем находим

$$ET = 7, \quad \sigma_T^2 = \frac{35}{6}.$$

На сводных характеристиках асимметрии и иных более тонких черт закона распределения для наших целей нет надобности останавливаться.

§ IV.

1) Сходно намечаются и задачи изучения нескольких случайных переменных, стоящих в стохастической связи. Существо стохастической связи между двумя случайными переменными заключается в том, что возможные значения одной переменной встречаются в сочетании с разными значениями другой переменной и что каждому сочетанию присуща определенная вероятность. Совокупность различных сочетаний возможных значений переменных и присущих этим сочетаниям вероятностей будем называть: *закон связи между переменными.* Понятие закона связи легко распространить на любое число случайных переменных, стоящих в стохастической связи.

Если закон связи известен, то все, что так или иначе способно характеризовать взаимоотношения между переменными, может почитаться известным. Все прочее может быть выведено из закона связи путем математических выкладок. Целью исследования может и тут почитаться нахождение совокупности чисел, которые образуют закон связи. Но, как и в случае закона распределения значений одной случайной переменной, знание закона связи должно быть дополнено вычислением сводных его характеристик, облекающих в сжатые и выразительные формы те или иные его черты, важные для целей исследования.

2) Для того, чтоб разобраться в системе приемов, применяемых для характеристики законов связи, необходимо положить в основу точные определения ряда понятий, тесно примыкающих к понятию стохастической связи.

Будем называть *условным законом распределения значений* Y для данного значения X совокупность значений, которые переменная Y способна принимать, если X получает данное значение, и присущих им при этом предположении вероятностей. Математическое ожидание переменной Y, ее рассеяние и т. д., вычисляемые на основе условного закона распределения, будем называть *условным математическим ожиданием*, *условным рассеянием* и т. д. в отличие от *полного математического ожидания* Y, *полного рассеяния* и т. д., вычисляемых по закону распределения значений Y безотносительно к значению, получаемому переменною X; будем обозначать их обычными буквами с добавкою

наверху справа в скобках указания на то значение X, которое рассматривается, как данное: $E^{(i)}Y$ — условное математическое ожидание Y для значения x_i переменной X; $\sigma_Y^{(i)}$ — условная квадратическая ошибка Y для того же значения X.

Выражая в аналитической форме функциональную зависимость условного математического ожидания переменной Y от переменной X, мы получаем так называемое *уравнение регрессии* Y по X: изображая функциональную зависимость условного математического ожидания Y от X графически в прямоугольных координатах, находим *линию регрессии* Y по X. Выражая в аналитической или графической форме зависимость условного рассеяния или условной средней квадратической ошибки Y от X, находим *скедастическое уравнение* или *скедастическую линию* (от греческого глагола σκεδάννυμι, я рассеиваю). Если условное рассеяние Y остается неизменным при всех значениях X, то связь Y с X называется *гомоскедастичною*. В противном случае она именуется *гетероскедастичною*.

Поясним на примере понятия линии регрессии и скедастической линии. Пусть X обозначает число очков, выпадающих на белой кости; пусть Y обозначает число очков, выпадающих на красной кости, а T обозначает сумму X и Y. Обозначая, согласно указанному выше, через $E^{(i)}T$ условное математическое ожидание T, отвечающее значению x_i переменной X, и через $E^{(j)}X$ условное математическое ожидание X, отвечающее значению t_j переменной T, находим, что $E^{(i)}T$ слагается из соответственного значения X и математического ожидания Y, равного неизменно 3,5. Уравнение регрессии Y по X имеет вид: $E^{(i)}T = 3{,}5 + x_i$, — регрессия T по X прямолинейна. Не трудно найти и уравнение регрессии X по T. Возможные значения T лежат в пределах между 2 и 12. Значение 2 переменная T может получить лишь, если на обеих костях выпадает 1; переменная X может в этом случае иметь лишь одно единственное значение 1, которому и равняется условное математическое ожидание X: $E^{(2)}X$ равняется, следовательно, половине соответственного значения T. Значение 3 переменная T может получить, если $X = 1$, а $Y = 2$, или если $X = 2$, а $Y = 1$; обе возможности равновероятны; условное математическое ожидание X равно, следовательно, $\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 2 = 1{,}5$; и в этом случае оно равняется половине соответственного значения T. Сходным образом убеждаемся, что и при всех иных значениях T условное математическое ожидание X равняется половине соответственного значения T. Уравнение регрессии X по T имеет, следовательно, вид: $E^{(j)}X = \frac{1}{2}t_j$. Таким образом, и регрессия X по T прямолинейна.

Условное рассеяние T, отвечающее данному значению X, определяется колебаниями переменной Y и остается при всех значениях X неизменно равным $[\sigma_T^{(i)}]^2 = \frac{35}{12}$: связь T с X гомоскедастична. Что касается условного рассеяния X, то оно равно нулю, если T принимает значения 2 или 12. Если T принимает значение 3, то X может с равными вероятностями получать значения 1 и 2; отклонения от условного математического ожидания, равного 1,5, равняются

$+1/2$ и $-1/2$; условное рассеяние получается равным $\frac{1}{2}\cdot\frac{1}{4}+\frac{1}{2}\cdot\frac{1}{4}=\frac{1}{4}$, условная средняя квадратическая ошибка равняется $1/2$. Тому же равняется условная средняя квадратическая ошибка X, отвечающая значению 11 переменной T, так как при этом X может получать с равными вероятностями значения 5 и 6, отклоняющиеся на $1/2$ от условного математического ожидания X. Сходным образом находим для условного рассеяния X, отвечающего значениям 3 и 10 переменной T, величину $2/3$ и т. д. Зависимость X от T представляется, следовательно, гетероскедастичною; скедастическое уравнение имеет вид

$$\sigma_X^{(t)}=\sqrt{\frac{(5-|t_j-7|)(7-|t_j-7|)}{12}},$$

при чем $|t_j-7|$ означает абсолютную величину разности t_j-7.

3) Если условный закон распределения значений Y остается неизменным, какое бы из числа своих возможных значений ни принимала переменная X, то переменная Y *стохастически независима* от X. Не трудно показать (см. ниже глава четвертая, § I, 2), что в случае стохастической независимости Y от X переменная X представляется в свою очередь также стохастически независимою от Y, т.-е. что условный закон распределения значений X остается неизменным, какое бы из числа своих возможных значений ни принимала переменная Y: две случайных переменных могут быть независимы лишь взаимно. Если условный закон распределения значений Y изменяется в каком бы то ни было отношении с переходом от одних значений X к другим, то переменная Y не является стохастически независимою от X.

Понятие стохастической независимости — один из основных устоев теории корреляции. При исследовании стохастической связи всегда возникает предварительный вопрос, не представляются ли переменные взаимно независимыми. Если переменные X и Y взаимно независимы, то нет нужды подвергать особому рассмотрению закон их связи: он вполне определяется тогда законами распределения значений переменных. Если же переменные не представляются стохастически независимыми, то по законам распределения их значений нет возможности установить закон их связи: закон связи требует тогда самостоятельного изучения.

Понятие стохастической независимости может быть конструируемо и иначе. В литературе теории кореляции имеется целый ряд определений, более или менее расходящихся. Некоторые из этих разнообразных понятий независимости представляют самостоятельную научную ценность и вполне целесообразно удерживать их в научном обиходе. Необходимым условием является лишь, чтоб различные понятия независимости отчетливо различались друг от друга и покрывались разными терминами. Определение, предложенное выше, наилучше приспособлено к тому, чтоб служить основою теории корреляции. Оно является наиболее строгим изо всех формальных определений независимости. Если переменные стохастически независимы в этом смысле, то они представляются независимыми и в смысле всех иных формальных определений понятия. При этом, с каким бы из понятий ни иметь дело, следует твердо держать в памяти, что речь идет о независимости

в формально математическом смысле этого слова, как terminus technicus теории корреляции; к обычному представлению о независимости в смысле отсутствия причинного сцепления эти понятия независимости стоят в довольно своеобразных отношениях. Пусть, например, имеется шесть закрытых урн, содержащих шары белые и черные; урны перенумерованы в порядке чисел 1, 2, 3, 4, 5 и 6. Бросанием кости решается, из какой урны производить извлечения: если выпадает на кости 3, то шары вынимаются из урны № 3 и т. д. Число вынимаемых из урны шаров n остается неизменным, из какой бы урны ни производились извлечения. Относительное число вынувшихся белых шаров, могущее принимать $n+1$ разных значений — $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$ — и число, выпадающее на кости, представляются стохастически связанными случайными переменными, если доля белых шаров в разных урнах не одна и та же, и стохастически независимыми переменными, если доля белых шаров во всех урнах одна и та же. Причинный же механизм, определяющий ход явлений, не обнаруживает столь глубоких различий при этих двух предположениях. Своеобразие конструкции понятия независимости, с которым имеет дело теория корреляции, должно быть всегда внимательно учитываемо при истолковании результатов статистического исследования связи между явлениями. Иначе возникает риск запутаться в заключениях, далеко выходящих за пределы компетенции статистика.

Из прочих понятий независимости, используемых в теории статистики, заслуживает быть отмеченным первоначальное Пирсоновское определение понятия коррелированности переменных. Пирсон называет переменную Y *коррелированною* с X, если условное математическое ожидание Y не остается неизменным при всех значениях X. Если же условное математическое ожидание Y сохраняет одно и то же значение при всех значениях X, то переменная Y признается *некоррелированною* с X. Некоррелированность Y с X находит себе, следственно, выражение в том, что линия регрессии Y по X, изображающая графически зависимость условного математического ожидания Y по X, имеет вид прямой, параллельной оси X; если линия регрессии Y по X отлична от прямой, параллельной оси X, то переменная Y коррелирована с X.

Понятия коррелированности и некоррелированности имеют большую ценность для теории статистики, и мы будем постоянно пользоваться ими. От понятия стохастической независимости понятие некоррелированности отличается глубоко. Если переменная Y стохастически независима от X, то она и некоррелирована с X. Но из того, что переменная Y некоррелирована с X, нельзя заключать, что она стохастически независима от X: условное математическое ожидание Y может оставаться неизменным при всех значениях X, но условный закон распределения значений Y испытывать иные изменения при переходе от одних значений X к другим, — связь Y с X может, например, быть гетероскедастичною. Предположим, что в тех шести урнах, о которых только-что была речь, содержатся шары белые и черные в одном и том же числе, но что число извлечений из урны не остается неизменным, а пропорционально числу очков, выпадающих на кости: если выпадает 1, то производится n извлечений; если выпадает 2, то производится $2n$ извлечений и т. д. Доля вынувшихся белых шаров представляется

в таких условиях некоррелированною с числом очков на кости, так как условное математическое ожидание ее, определяемое долею белых шаров в урне, остается одним и тем же, сколько бы извлечений из урны ни производилось. Но условная средняя квадратическая ошибка доли вынувшихся белых шаров не остается неизменною: завися от числа извлечений, она имеет в два раза меньшие размеры, если выпадает 4, чем в случае, если выпадает 1. Доля вынувшихся белых шаров, будучи некоррелированною с числом очков, не представляется, таким образом, от него стохастически независимою.

Весьма существенно отметить, что в отличие от стохастической независимости, которая всегда взаимна, некоррелированность Y с X не обусловливает собою некоррелированности X с Y. Регрессия Y по X может иметь вид прямой, параллельной оси X, а в то же время регрессия X по Y иметь вид сколько угодно отличный от прямой, параллельной оси Y. Чтоб дать себе вполне ясный отчет в этом чрезвычайно важном свойстве стохастической связи, рассмотрим такой пример.

Из закрытой урны, содержащей поровну белых и черных шаров, производится t серий извлечений, при чем вынутый шар бросается в урну прежде, чем производится следующее извлечение. Число производимых извлечений определяется, в свою очередь, случаем: из другой закрытой урны, содержащей ряд билетов с написанными на них разными числами, вынимается один билет, и числу, стоящему на нем, приравнивается число извлечений; вынутый билет опускается обратно в урну прежде, чем приступают к установлению числа извлечений для следующей серии; вероятность каждого значения числа извлечений остается, таким образом, все время неизменною. Обозначим через n_i общее число шаров, извлекаемых в i-ой серии извлечений, и через w_i долю белых шаров среди них, и рассмотрим стохастическую связь между N и W в предположении, что числа, значащиеся на билетах, колеблются в широких границах, так что некоторые серии включают очень незначительное число извлечений, другие же, напротив, весьма большое. Начнем со связи переменной W с переменною N. Условное математическое ожидание W остается при данной постановке эксперимента неизменным при всяком N, так как W определяется однозначно долею белых шаров в урне: переменная W некоррелирована с N, и регрессия W по N имеет вид прямой параллельной оси N. Что касается условных законов распределения значений N, отвечающих разным значениям W, то они характеризуются линией регрессии совсем иного вида: значениям W, весьма близким к нулю, с одной стороны, и значениям W, весьма близким к 1, с другой, будут соответствовать преимущественно малые числа извлечений; напротив, значениям W, близким к $^1/_2$, будут отвечать сравнительно большие числа извлечений. Линия регрессии N по W будет иметь вид кривой, поднимающейся всего выше в точке 0,5 и более или менее круто спадающей по обе стороны этой точки. В известном руководстве Юла [1]) можно найти изображение линий регрессии W по N и N по W для конкретного примера связи между долей мальчиков среди новорожденных в различных районах Англии

[1]) G.U. Yule. An introduction to the theory of statistics p. 176 (London, 1911); см E. Czuber. Die statistischen Forschungsmethoden, S. 142.

и общим числом рождений в этих районах: вид их близко соответствует тому представлению, к которому мы пришли на основании общих соображений.

§ V.

Вернемся теперь к выяснению различий между строем мысли естествоиспытателя, стремящегося установить закон функциональной зависимости между изучаемыми величинами, и подходом к задаче статистика, интерес которого сосредоточен на стохастической связи.

В уравнении регрессии Y по X находит выражение функциональная зависимость условного математического ожидания Y от X. В уравнении регрессии X по Y находит выражение функциональная зависимость условного математического ожидания X от Y. Формально уравнения регрессии вполне соответствуют тем формулам функциональной зависимости между переменными, с которыми привык иметь дело естествоиспытатель. При известных условиях уравнение регрессии может совпадать и по содержанию с формулою закона функциональной зависимости, найти которую стремится естествоиспытатель. Предположим, что при экспериментальном исследовании связи между суммою углов многоугольника и числом его сторон ошибки измерений строго следуют закону Гаусса-Лапласа. Получаемые в результате измерений суммы углов треугольников, четыреугольников и т. д. представляются в таких условиях случайными переменными, математические ожидания которых равны соответственным истинным значениям сумм: 180 градусам для треугольника, 360 градусам для четыреугольника и т. д. Функциональная зависимость между величиною математического ожидания суммы углов, отвечающей данному числу сторон, и числом сторон — или линия регрессии суммы углов по числу сторон — совпадает с искомым законом: сумма углов n-угольника равна $(n-2).180°$.

При всем сходстве между уравнением регрессии и тою формулою закона функциональной зависимости, которую обычно ищет естествоиспытатель, между ними существуют, однако, глубокие различия. Закон функциональной зависимости всегда обратим: если дано выражение Y в виде функции X, то представляется в принципе всегда возможным перейти путем математических выкладок к выражению X, как функции Y: если $Y=X^2$, то $X=\pm\sqrt{Y}$; если $Y=aX+b$, то $X=\frac{1}{a}Y-\frac{b}{a}$. Уравнение же регрессии Y по X и уравнение регрессии X по Y не могут быть получены одно из другого. Регрессия суммы чисел, выпадающих на двух костях, по числу, выпадающему на одной из костей, линейна, как мы видели (см. выше § IV, 2)) и уравнение ее имеет вид: $E^{(x)}T=3.5+x_i$. Уравнение регрессии одного из чисел по сумме обоих чисел также линейно, но имеет вид $E^{(t)}X=\frac{1}{2}t_j$. Никакими математическими ухищрениями нельзя вывести одно уравнение из другого. Оба уравнения приходится устанавливать совершенно самостоятельно по закону связи. Само по себе это и не удивительно, так как эти два уравнения связывают совсем разные величины: не Y с X и X с Y, а условное математическое ожидание Y с X и условное матема-

тическое ожидание X с Y. Они также далеки друг от друга, как уравнение, связывающее X с Y, от уравнения, связывающего Z с T. Но мысль исследователя, переходящего к статистике от нестатистических приемов работы в области естествознания, не может оторваться от привычного представления о функциональной зависимости между X и Y, о котором, как ему чудится, говорят и уравнения регрессии. Своеобразие не обратимых взаимоотношений между уравнениями регрессии воспринимается им как нечто несуразное. Он не может преодолеть чувства, что дело здесь в несовершенстве обычных способов статистического подхода к изучению стохастической связи, — в недодуманности, которую надо преодолеть путем вычисления единого уравнения регрессии, связующего функционально X и Y и позволяющего с равным успехом выражать X, как функцию Y и Y, как функцию X. Подобные стремления свидетельствуют о неясном понимании подлинного характера стохастической связи. Они представляются, однако, законными, поскольку не выходят из должных границ.

Где эти границы проходят, явствует из предшествующего изложения. Они намечаются двойственностью роли, выпадающей на долю понятия стохастической связи в обиходе научной работы. Если стохастическая связь носит характер лишь оболочки, скрывающей в себе ядро функциональной зависимости между истинными значениями исследуемых величин, то исследователь с полным правом не удовлетворяется знанием уравнений регрессии. Ему нужна формула функциональной зависимости между X и Y; условные математические ожидания X и Y его нисколько не интересуют сами по себе; формула, связующая условные математические ожидания одной переменной с значениями другой переменной, ценна для него лишь постольку, поскольку может дать опору для перехода к закону функциональной зависимости между истинными значениями X и Y. Допустим, что ищется закон связи между температурою кипения воды и давлением; на вычислении уравнений регрессии, связующих условное математическое ожидание температуры с давлением и условное математическое ожидание давления с температурою, исследователь остановиться не может, так как они выражают не то, что ему требуется знать. Задача его остается не решенною, если ему не удается продвинуться дальше уравнений регрессии. Лишь в случае, если он сумеет путем надлежащей обработки величин, поставляемых его измерениями, добраться до функциональной зависимости, скрытой за стохастической связью, может он признать, что достиг своей цели. Какими путями следует итти к этой цели, вопрос особый и весьма сложный; систематическая разработка этой проблемы начата сравнительно недавно и далека еще от завершения. Останавливаться на приемах решения этой задачи нам нет надобности. Для нас достаточно отметить, что поиски единого уравнения, выражающего закон функциональной зависимости между X и Y, представляются во всех подобных случаях вполне законными и что уравнение это не имеет ничего общего с уравнениями регрессии.

Иначе обстоит дело, если конечным объектом изучения являются стохастически связанные случайные переменные. Если изучаемые величины X и Y являются подлинными случайными переменными, стоящими в подлинной стохастической связи, то между X и Y нет и не может быть функциональной зависимости: данным значениям одной из переменных не соответствуют определенные

значения другой; если X получает какое-либо из числа своих возможных значений, то переменная Y остается, как и до того, способною принимать с определенными вероятностями целый ряд разных значений, и ни одно из них не может быть рассматриваемо как то значение, которое соответствует данному значению X. Уравнение, выражающее функциональную зависимость между X и Y, представлялось в ранее рассмотренных условиях искомым, которое трудно было найти: тут его не приходится и искать. К характеристике взаимоотношений между X и Y можно итти тут разными путями: закон связи исчерпывающе описывает их; уравнения регрессии и иные методы, на которых мы остановимся в следующей главе, воспроизводят в сжатой и практической форме разные их свойства, представляющие интерес для исследователя. Но к числу разнообразных способов характеристики стохастической связи между X и Y формула закона функциональной зависимости между X и Y не принадлежит. Такой способ характеристики чужд по существу строю стохастической связи.

Мы можем теперь подвести итог. Условное математическое ожидание Y может совпадать с истинным значением находящейся от X в функциональной зависимости величины, которою интересуется исследователь; уравнение регрессии Y по X передает тут непосредственно искомый закон функциональной зависимости между изучаемыми переменными. С другой стороны, возможно, что исследователь интересуется функциональной зависимостью между X и Y, но ни уравнение регрессии Y по X, ни уравнение регрессии X по Y не передают ее непосредственно; в таких условиях лишь путем дальнейшей обработки данных может исследователь подняться до того, что он, собственно, ищет. И, наконец, встречаются случаи, когда ни о какой функциональной зависимости речи не идет, когда взаимоотношения складываются так, что характеристика их при посредстве формулы функциональной зависимости оказывается по существу неприложимою. Тут уравнения регрессии представляются законченными, никакого усовершенствования не требующими и не допускающими характеристиками известных сторон связи между величинами, составляющими действительный предмет изучения. Эти глубокие различия в целях исследования должны быть внимательно учитываемы при выборе методов работы и при истолковании полученных результатов. Исследователь должен быть всегда в состоянии дать себе отчет в том, что он собственно стремится установить путем счетной обработки своих статистических данных. Теория корреляции, изложение основ которой составляет нашу задачу, ставит своей целью свод и рационализацию приемов исследования, пригодных для изучения подлинной стохастической связи между переменными. Систематический обзор важнейших результатов теоретической работы в этом направлении и составит содержание последующих глав.

ГЛАВА ЧЕТВЕРТАЯ.

Априорный закон связи и система его сводных характеристик.

§ 1.

1) Своеобразие стохастической связи между переменными определяет строй приемов ее исследования. Способы, пригодные для изучения функциональной зависимости, не пригодны для изучения стохастической связи. Подобно тому, как рыбак подбирает снасть по рыбе, которую стремится поймать, так точно и научный исследователь должен заботливо подгонять свое оборудование к особенностям связи, подлежащей уловлению.

Мы ставим себе задачею дать систематический обзор основных методов изучения стохастической связи между двумя случайными переменными. Как выяснить, стоят ли случайные переменные X и Y в стохастической связи, или являются стохастически независимыми? Как выявить при наличии связи с наибольшею полнотою и рельефностью ее характерные черты, представляющие интерес для исследователя? При поисках ответа на эти вопросы почти неизбежно обратиться за поддержкою к математике. Без сложных выкладок можно обойтись, но было бы трудно вести изложение, не прибегая к языку алгебраических формул: придерживаясь исключительно словесных формулировок, мы были бы вынуждены лавировать между двумя крайностями, — либо оставлять формулировки чрезмерно расплывчатыми, либо придавать им запутанный и неудобопонятный вид. Представляется более экономным начать с усвоения ряда обозначений и основных для нас математических предложений. Предварительная затрата внимания окупается затем стократно.

2) Предположим, что даны две случайных переменных X и Y. Обозначим через $p_{i.}$ вероятность, что переменная X принимает одно определенное из ряда ее k возможных значений, — именно, значение x_i; обозначим через $p_{.j}$ вероятность, что переменная Y принимает одно определенное из ряда ее l возможных значений, — именно, значение y_j; обозначим через p_{ij} вероятность, что переменная X принимает значение x_i, а переменная Y одновременно принимает значение y_j. Введем, кроме того, обозначения: $p_{.j}^{(i)}$ для условной вероятности значения y_j переменной Y в предположении, что X получает значение x_i, и $p_{i.}^{(j)}$ для условной вероятности значения x_i переменной X в предположении, что переменная Y получает значение y_j. Как известно, вероятность совпадения двух взаимно не независимых событий равняется произведению вероятности одного из них на

условную вероятность другого в предположении, что первое имело место; следовательно,

$$p_i\, p_j^{(i)} = p_{ij} = p_j\, p_i^{(j)}.$$

Согласно данному выше определению (см. глава третья, § IV, 3)) переменная Y представляется стохастически независимою от X, если условный закон распределения значений Y остается неизменным при всех значениях X, т.-е. если условные вероятности разных значений Y сохраняют одну и ту же величину при всех значениях X. На языке наших алгебраических обозначений стохастическая независимость переменной Y от переменной X определяется, следовательно, условиями:

$$p_j^{(i)} = p_j \qquad \text{при } i = 1, 2, \ldots, k \qquad \text{и } j = 1, 2, \ldots, l.$$

Если переменная Y стохастически независима от X, то имеет силу эта система условий. Обратно: если

$$p_j^{(i)} = p_j \qquad \text{при } i = 1, 2, \ldots, k \qquad \text{и } j = 1, 2, \ldots, l,$$

то переменная Y стохастически независима от X, так как условный закон распределения значений Y остается неизменным при всех значениях X.

Положим $p_j^{(i)} = p_j$ в тожестве $p_i\, p_j^{(i)} = p_j\, p_i^{(j)}$. Сокращая на p_j, находим: $p_i = p_i^{(j)}$. Таким образом, если $p_j^{(i)} = p_j$ при всех значениях i и j, то и $p_i^{(j)} = p_i$ при всех значениях j и i. Стохастическая независимость Y от X обусловливает собою стохастическую независимость X от Y. Переменные X и Y могут быть лишь взаимно независимыми (см. выше глава третья, § IV, 3)).

Сходным путем убеждаемся, что при взаимной независимости переменных $p_{ij} = p_i\, p_j$ при всех значениях i и j.

Обратно: если $p_{ij} = p_i\, p_j$ при всех значениях i и j, то переменные X и Y взаимно независимы. Равенство величин p_{ij} и $p_i\, p_j$ при всех значениях i и j представляется не только необходимым, но и достаточным условием для того, чтоб переменные X и Y были взаимно независимы.

§ II.

1) В случае взаимной независимости переменных X и Y все разности $p_{ij} - p_i p_j$ равняются, как мы видели, нулю. Если хоть одна из разностей отлична от нуля, то переменные не могут быть стохастически независимыми. Чем значительнее разности, тем резче уклоняются взаимоотношения между переменными от стохастической независимости. Таким образом, в величине разностей $p_{ij} - p_i p_j$ находит себе выражение весьма существенное свойство стохастической связи между X и Y. В силу этого, на рассмотрении системы разностей $p_{ij} - p_i p_j$ зиждется одна из основных групп приемов изучения стохастической связи между случайными переменными.

Если переменная X и переменная Y могут принимать лишь по два разных

значения, то все четыре разности тожественно равны друг другу по абсолютной величине. Обозначая абсолютную величину их греческою буквою $\hat{c}$, находим:

$$\hat{c} = p_{11} - p_{1/} p_{/1} = p_{22} - p_{2/} p_{/2} = -[p_{12} - p_{1/} p_{/2}] = -[p_{21} - p_{2/} p_{/1}].$$

Величина $\hat{c}$ представляется в таких условиях практичною сводною характеристикою тесноты связи между переменными. Подставляя в выражение для $\hat{c}$ значения

$$p_{1/} = p_{11} + p_{12}, \quad p_{2/} = p_{21} + p_{22} \text{ и т. д.}$$

и замечая, что

$$p_{11} + p_{12} + p_{21} + p_{22} = 1,$$

находим:

$$\hat{c} = p_{11} p_{22} - p_{12} p_{21}.$$

В таком симметричном виде величина $\hat{c}$ особенно охотно используется для построения мер тесноты стохастической связи между переменными в случае, если обе переменных способны принимать лишь по два разных значения.

Если переменные способны принимать более, чем по два разных значения, то разности $p_{ij} - p_{i/} p_{/j}$ могут и не быть равными. Одна какая-либо из них не может тут уже служить характеристикою тесноты связи: сводные характеристики тесноты связи должны использовать всю совокупность разностей. В виду того, что алгебраическая сумма разностей тожественно равна нулю, представляется наиболее естественным положить в основу построения сводной характеристики вторые степени разностей. На этой основе могут быть построены различные меры тесноты связи. Наибольшее значение имеет среди них *показатель взаимной сопряженности* переменных (Mean Square Contingency), введенный Пирсоном, обозначаемый обычно греческою буквою φ и определяемый соотношением

$$\varphi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{[p_{ij} - p_{i/} p_{/j}]^2}{p_{i/} p_{/j}} = \sum_{i} \sum_{j} \frac{p^2_{ij}}{p_{i/} p_{/j}} - 1.$$

Если переменные взаимно независимы и разности $p_{ij} - p_{i/} p_{/j}$ все равны нулю, то обращается в нуль и показатель взаимной сопряженности. Обратно: φ^2 может равняться нулю лишь, если все разности $p_{ij} - p_{i/} p_{/j}$ равны нулю; равенство нулю показателя взаимной сопряженности свидетельствует, следовательно, о стохастической независимости переменных. Если Y находится в однозначной функциональной зависимости от X и каждому значению X соответствует одно лишь значение Y, то все вероятности p_{ij}, в которых j неравно i, равняются нулю, а вероятность p_{ii} совпадает с $p_{i/}$; подставляя эти значения в формулу для φ^2, находим $\varphi^2 = k - 1$ (k обозначает при этом число возможных значений X и Y).

Коэффициент $\frac{1}{\sqrt{(k-1)(l-1)}} \varphi^2$ обращается, таким образом, в нуль в случае взаимной независимости переменных — и только в этом случае! — и при-

нимает значение 1 в случае однозначной функциональной зависимости Y от X, какой бы вид ни имела при этом функциональная зависимость. В границах между 0 и 1 и держится численное значение $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$, приближаясь тем более к верхнему пределу, чем ближе условные законы распределения значений Y к такому виду, при котором каждому данному значению X отвечает лишь одно значение Y, — другими словами, чем ближе взаимоотношения между переменными к однозначной функциональной зависимости Y от X. В коэффициенте $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$ мы располагаем мерою, отчетливо характеризующею тесноту стохастической связи между переменными при любом виде закона связи.

Если обе переменных способны принимать лишь по два разных значения, то коэффициент $\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2$ легко приводится к виду

$$\frac{1}{\sqrt{(k-1)(l-1)}}\varphi^2 = \varphi^2 = \frac{\delta^2}{p_1 p_2 p_1 p_2}.$$

2) Все меры, использующие величины δ или φ, обладают одним общим свойством; они привлекают лишь вероятности разных значений X и Y и их различных сочетаний, опуская самые значения На величине δ и φ совсем не сказываются ни различия в размерах значений X и Y, ни различия в размахе их колебаний, если только все вероятности p_1, p_1, p_{ij} сохраняют свою величину. Эта особенность делает меры тесноты стохастической связи, опирающиеся на значения δ и φ, пригодными для изучения связи и в таких случаях, которые не подходят непосредственно под данное нами выше определение понятия стохастической связи (см. глава третья, § II, 1)). Раз численные значения X и Y не входят в формулы, то для пользования этими формулами нет надобности и знать их. А раз не надо знать их, то нет надобности и измерять их. А раз не надо измерять их, то вполне безразлично, допускают ли они вообще измерение, или нет; для применения этих приемов работы не требуется, чтобы разные значения X и Y были численно измеримы: достаточно, чтоб они были различимы,— чтоб можно было считать их и их разные сочетания. Признаки, связь между которыми подвергается изучению, могут быть не количественными, а качественными; возможности вычисления показателя взаимной сопряженности не затрогивает и это. В статистике брачности теснота связи между вероисповеданиями жениха и невесты может быть измеряема посредством показателя взаимной сопряженности с тем же удобством, как теснота связи между возрастом или ростом жениха и возрастом или ростом невесты. Случайно переменною величиною в смысле нашего определения вероисповедание, конечно, не является, так как оно вообще не является величиною, будь то случайно, будь то неслучайно переменною. Но оно может быть рассматриваемо как переменный признак, и мы можем среди качественных переменных признаков так же различать случайно переменные от неслучайно переменных, как и среди количественных. Если различным проявлениям какого-либо признака, не допускающим точного измерения

или количественно совсем несравнимым, присущи определенные вероятности, то мы имеем дело с *случайно переменным качественным признаком*. Сходным образом должно быть расширено и понятие стохастической связи: два переменных признака стоят в стохастической связи, если по определении значения одного из них другой сохраняет характер случайно переменного признака, способного с определенными вероятностями облекаться в разные формы. Признак же, который утрачивает способность облекаться с определенными вероятностями в разные формы после того, как другой признак получает определенное значение, находится от последнего в зависимости, напоминающей функциональную. Примером случайно переменного качественного признака может служить цвет шара, вынимаемого из закрытой урны. Предположим, что шары, находящиеся в закрытой урне, различаются не только цветом, но также какими-либо пометками; тогда между цветом шаров и типом пометки на них будет наблюдаться стохастическая связь, при чем оба связанных признака носят характер чисто качественных. Предположим, что шары различаются, кроме цвета, помеченными на них номерами; тогда перед нами будет стохастическая связь между признаками, из которых один качественный, а другой количественный.

Примыкая к обобщенным понятиям случайно переменного признака и стохастической связи между случайно переменными признаками, мы можем отмеченное выше свойство приемов измерения тесноты стохастической связи, опирающихся на вычисление величины $\hat{c}$ или $\hat{\varphi}$, характеризовать, как приложимость их к изучению стохастической связи между случайно переменными признаками, не только количественными, но и качественными. Это свойство их представляется весьма ценным. Приемы, к рассмотрению которых мы теперь переходим, приложимы непосредственно лишь к изучению связи между признаками, допускающими достаточно точное измерение. Для изучения связи между качественными признаками они пригодны лишь в случае, если обе переменных способны принимать только по два разных значения (см. ниже § VII). Для того, чтоб пользоваться ими, когда признаки не носят количественного характера, приходится качественные признаки искусственно „квантифицировать", — располагать их разные проявления в ряд так, чтоб порядок членов ряда можно было с некоторым правом отожествлять с нарастанием или убыванием, допускающим не совсем произвольную количественную разметку.

§ III.

1) Другая группа приемов исследования стохастической связи между двумя случайными переменными исходит от рассмотрения условных законов распределения значений. Если переменная Y стохастически независима от X, то условные законы распределения значений Y остаются неизменными при всех возможных значениях X (см. глава третья, § IV, 3)). Наличие связи находит себе выражение в том, что те или иные сводные характеристики условных законов распределения значений Y изменяются определенным образом с переходом от одних значений X к другим. В том, как именно они изменяются, и выявляется своеобразие закона связи.

В целях систематического обзора этой группы приемов исследования стохастической связи введем ряд новых обозначений. Положим:

$$m_{f/g} = EX^f Y^g = \sum_i \sum_j p_{i/j} x_i^f y_j^g,$$

$$\mu_{f/g} = E[X - m_{1/0}]^f [Y - m_{0/1}]^g = \sum_i \sum_j p_{i/j} [x_i - m_{1/0}]^f [y_j - m_{0/1}]^g.$$

$$r_{f/g} = \frac{\mu_{f/g}}{\mu_{2/0}^{1/2f} \mu_{0/2}^{1/2g}}.$$

Полагая здесь $g = 0$, находим сводные характеристики закона распределения значений X. Полагая $f = 0$, находим сводные характеристики закона распределения значений Y. В частности, $m_{1/0} = EX$ обозначает математическое ожидание X, $\mu_{2/0} = E[X - m_{1/0}]^2 = \sigma_X^2$ обозначает рассеяние X, $m_{0/1} = EY$ обозначает математическое ожидание Y, $\mu_{0/2} = \sigma_Y^2$ обозначает рассеяние Y.

Параметрами m, μ и r мы будем все время пользоваться для характеристики законов связи. По данному закону связи все они определяются однозначно. Обратно: закон связи представляется однозначно определимым, если даны параметры m в целесоответственном подборе и в потребном числе, или, если наряду со значениями $m_{1/0}$ и $m_{0/1}$, даны параметры μ, как равно, если наряду со значениями $m_{1/0}$, $\mu_{2/0}$, $m_{0/1}$ и $\mu_{0/2}$ даны параметры r в потребном числе. На доказательстве этих положений я не буду останавливаться. Для случая прерывных законов связи оно не представляет особой трудности. В случае непрерывных законов связи доказательство математически весьма сложно. Проблема эта, известная в математической литературе под обозначением problème des moments, слишком далеко отвлекла бы нас от наших прямых задач. Для наших ближайших целей она к тому же не имеет особого значения.

Параметры типа r обладают тем ценным свойством, что носят характер отвлеченных чисел; это делает их весьма удобными для сравнительной характеристики законов связи. Первый в ряду параметров r — именно $r_{1/1}$ — именуется обычно *коэффициентом корреляции* и определяется, как

$$r_{1/1} = \frac{\mu_{1/1}}{\mu_{2/0}^{1/2} \mu_{0/2}^{1/2}} = \frac{\mu_{1/1}}{\sigma_X \sigma_Y}.$$

Не трудно убедиться, что коэффициент корреляции по своей абсолютной величине не может быть больше 1. Замечая, что математическое ожидание переменной, не принимающей отрицательных значений, не может быть меньше нуля, находим

$$E\left[\frac{X - m_{1/0}}{\sigma_X} - r_{1,1}\frac{Y - m_{0/1}}{\sigma_Y}\right]^2 \geqslant 0.$$

Но

$$E\left[\frac{X-m_{1/0}}{\sigma_X}-r_{1\,1}\frac{Y-m_{0\,1}}{\sigma_Y}\right]^2=E\frac{[X-m_{1/0}]^2}{\sigma_X^2}-2r_{1\,1}E\frac{[X-m_{1\,0}]\,[Y-m_{0\,1}]}{\sigma_X\sigma_Y}+$$

$$+r_{1\,1}^2E\frac{[Y-m_{0/1}]^2}{\sigma_Y^2}=1-r_{1,1}^2.$$

Следовательно,

$$1-r_{1/1}^2\geqslant 0,\quad r_{1\,1}^2\leqslant 1.$$

Параметры, характеризующие условные законы распределения значений Y, будем отличать пометкою на верху справа в скобках указания на то значение X, которое принимается за данное: $m_{\,1}^{(i)}$ обозначает условное математическое ожидание Y в предположении, что X принимает значение x_i; $\mu_{\,2}^{(i)}$ обозначает соответственное условное рассеяние Y. Согласно определениям,

$$m_{\,1}^{(i)}=\sum_j p_{j/i}\,y_j,\qquad \mu_{\,2}^{(i)}=\sum_j p_{j/i}\,[y_j-m_{\,1}^{(i)}]^2.$$

Сходным образом находим для условного математического ожидания X и условного рассеяния X, отвечающих значению y_j переменной Y,

$$m_{1/j}^{(j)}=\sum_i p_{i/j}\,x_i,\qquad \mu_{2/j}^{(j)}=\sum_i p_{i/j}\,[x_i-m_{1/j}^{(j)}]^2.$$

Между параметрами, характеризующими условные законы распределения, и сводными характеристиками законов распределения значений переменных, рассматриваемых безотносительно одна к другой, существуют многообразные соотношения, из коих отметим лишь тожества, связывающие полные математические ожидания переменных со средними арифметическими их условных математических ожиданий и тожества, связывающие полные рассеяния со средними арифметическими условных рассеяний:

$$\sum_i p_{i/}\,m_{\,1}^{(i)}=m_{0/1},\qquad \sum_i p_{i/}\,\mu_{\,2}^{(i)}=\mu_{0/2}-\sum_i p_{i/}\,[m_{\,1}^{(i)}-m_{0/1}]^2,$$

$$\sum_j p_{j/}\,m_{1/}^{(j)}=m_{1/0},\qquad \sum_j p_{j/}\,\mu_{2/}^{(j)}=\mu_{2/0}-\sum_j p_{j/}\,[m_{\,1}^{(j)}-m_{1/0}]^2.$$

В случае взаимной независимости переменных имеем при любых значениях h

$$m_{h/}^{(j)}=m_{h/0}\quad\text{при}\quad j=1,\,2,\,3,\cdots,\,l$$

$$m_{/h}^{(i)}=m_{0/h}\quad\text{при}\quad i=1,\,2,\,3,\cdots,k.$$

и, с другой стороны, при любых значениях f и g

$$m_{f,g} = m_{f,0} m_{0,g}, \quad \mu_{f,g} = \mu_{f,0} \mu_{0,g}, \quad r_{f,g} = r_{f,0} r_{0,g}.$$

Обратно: в случае, если

$$m_{f,g} = m_{f,0} m_{0,g} \quad \text{или} \quad \mu_{f,g} = \mu_{f,0} \mu_{0,g} \quad \text{или} \quad r_{f,g} = r_{f,0} r_{0,g}$$

при любых или хотя бы лишь при любых целых положительных значениях f и g то переменные взаимно независимы.

2) В ряду сводных характеристик условных законов распределения значений переменных интерес исследователя останавливается, естественно, в первую очередь на условных математических ожиданиях. Выражая в аналитической форме функциональную зависимость условного математического ожидания одной переменной от значений другой переменной, мы получаем так называемое уравнение регрессии первой переменной по второй: $m_{1,}^{(i)} = f(x_i)$ — уравнение регрессии Y по X; $m_{1,}^{(j)} = F(y_j)$ — уравнение регрессии X по Y. Слово „регрессия“ имело первоначально определенный смысл в такой связи, но затем начало употребляться распространительно, и первоначальный смысл стушевался; в настоящее время оно носит характер искусственно построенного условного термина, с которым уже не приходиться связывать каких-либо корнесловных ассоциаций.

А) Предположим, что линия регрессии Y по X имеет вид параболы с уравнением

$$(I) \qquad m_{1}^{(i)} = a_{,0} + a_{,1} x_i + a_2 x_i^2 + \cdots + a_{,f} x_i^f.$$

Коэффициенты этого уравнения легко могут быть выражены через посредство параметров m, μ и r. Умножим обе стороны на $p_{i,} x_i^h$, оставляя h пока неопределенным, и суммируем по i. Так как

$$\sum_i p_{i,} x_i^h m_{,1}^{(i)} = m_{h,1},$$

находим:

$$m_{h,1} = a_{,0} m_{h,0} + a_{,1} m_{h+1,0} + \cdots + a_{,f} m_{h+f,0}.$$

Полагая здесь h равным 0, 1, 2, 3, ... и т. д. до f, получаем $f+1$ линейных уравнений, из которых коэффициенты a находятся без труда. Подставляя их в уравнение (I), получаем условия, которым должны отвечать параметры m, дабы регрессия Y по X имела вид параболы f-ой степени.

Пусть, например, регрессия Y по X линейна и уравнение ее имеет вид

$$m_{,1}^{(i)} = a_{,0} + a_{,1} x_i.$$

Замечая, что $m_{0,0}$ тождественно равняется 1, находим

$$m_{0,1} = a_{,0} + a_{,1} m_{1,0}, \qquad m_{1,1} = a_{,0} m_{1,0} + a_{,1} m_{2,0}$$

и отсюда

$$a_1 = \frac{m_{11} - m_{10}m_{01}}{m_{20} - m_{10}^2}. \qquad a_0 = \frac{m_{20}m_{01} - m_{10}m_{11}}{m_{20} - m_{10}^2} = m_{01} - m_{10}a_1.$$

Уравнению линейной регрессии Y по X может быть, следовательно, придан вид:

$$m_1^{(i)} - m_{01} = \frac{m_{11} - m_{10}m_{01}}{m_{20} - m_{10}^2}[x_i - m_{10}].$$

Условие же, которому должны удовлетворять параметры m, дабы регрессия Y по X была линейною, сводится к требованию, чтоб

$$\frac{m_{h1} - m_{h0}m_{01}}{m_{h+10} - m_{h0}m_{10}} = \frac{m_{11} - m_{10}m_{01}}{m_{20} - m_{10}^2}$$

при всяких целых положительных значениях h.

В) Для того, чтоб выразить коэффициенты уравнения регрессии чрез параметры μ, представляется более удобным придать уравнению регрессии несколько иной вид: выразить в аналитической форме не функциональную зависимость условного математического ожидания Y от X, а функциональную зависимость разности $m_1^{(i)} - m_{01}$ от разности $x_i - m_{10}$. Предполагая, что линия регрессии Y по X имеет вид параболы f-ой степени с уравнением

$$m_1^{(i)} - m_{01} = b_0 + b_1[x_i - m_{10}] + b_2[x_i - m_{10}]^2 + \cdots +$$
$$+ b_f[x_i - m_{10}]^f.$$

умножим обе стороны на $p_i[x_i - m_{10}]^h$ и суммируем по i. Замечая, что

$$\sum_i p_i[x_i - m_{10}]^h[m_1^{(i)} - m_{01}] = \mu_{h1},$$

находим:

$$\text{(II)} \qquad \mu_{h1} = b_0\mu_{h0} + b_1\mu_{h+10} + \cdots + b_f\mu_{h+f0}.$$

Полагая здесь h равным 0, 1, 2, 3, ... и т. д. до f, получаем вновь систему $f+1$ линейных уравнений, позволяющих выразить коэффициенты b через параметры μ. Подставляя коэффициенты b назад в уравнение (II), устанавливаем условия, которым должны отвечать параметры μ, дабы регрессия Y по X имела вид параболы f-ой степени.

В случае линейной регрессии с уравнением

$$m_1^{(i)} - m_{01} = b_0 + b_1[x_i - m_{10}]$$

находим:

$$b_{/0} = 0, \qquad b_{/1} = \frac{\mu_{1/1}}{\mu_{2/0}}.$$

Уравнение линейной регрессии Y по X может быть, следовательно, облечено в форму:

$$m_{/1}^{(i)} - m_{0/1} = \frac{\mu_{1/1}}{\mu_{2/0}}\left[x_i - m_{1/0}\right].$$

Коэффициент $b_{/1} = \frac{\mu_{1/1}}{\mu_{2,0}} = \frac{\sigma_Y}{\sigma_X} r_{1/1}$ в уравнении линейной регрессии Y по X обычно именуется *коэффициентом регрессии* Y по X, — обозначение неудачное, так как коэффициентами регрессии могли бы именоваться с равным правом все коэффициенты b в уравнениях регрессии любого вида. Однако, такое словоупотребление прочно укоренилось, и когда идет речь о коэффициенте регрессии Y по X, всегда разумеют именно коэффициент $b_{/1} = \frac{\sigma_Y}{\sigma_X} r_{1/1}$ в линейном уравнении регрессии.

Условия, коим должны отвечать параметры μ, дабы регрессия Y по X была линейною, имеют более простой вид, чем условия, связывающие параметры m. Они сводятся к требованию, чтоб $\frac{\mu_{h/1}}{\mu_{h+1/0}} = \frac{\mu_{1/1}}{\mu_{2/0}}$ при любых целых положительных значениях h.

Сходным образом находим для уравнения линейной регрессии X по Y:

$$m_{1/}^{(i)} - m_{1/0} = b_{1/}\left[y_j - m_{0/1}\right] = \frac{\mu_{1/1}}{\mu_{0/2}}\left[y_j - m_{0\,1}\right].$$

и для условия, которому должны отвечать параметры μ, дабы регрессия X по Y была линейною $\frac{\mu_{1/h}}{\mu_{0/h+1}} = \frac{\mu_{1/1}}{\mu_{0/2}}$ при любых целых положительных значениях h.

Если линейны обе регрессии, то одновременно $b_{/1} = \frac{\sigma_Y}{\sigma_X} r_{1/1}$ и $b_{1/} = \frac{\sigma_X}{\sigma_Y} r_{1/1}$. Отсюда находим: $b_{1/}\, b_{/1} = r_{1/1}^2$: в случае, если линейна и регрессия Y по X, и регрессия X по Y, произведение коэффициентов регрессии $b_{1/}$ и $b_{/1}$ тождественно равно квадрату коэффициента корреляции или, другими словами, коэффициент корреляции тождественно равен средней геометрической из коэффициентов регрессии.

С) Подставляя в уравнение линейной регрессии Y по X $b_{/1} = \frac{\sigma_1}{\sigma_X} r_{1/1}$, приводим уравнение к виду:

$$m_{/1}^{(i)} - m_{0/1} = \frac{\sigma_Y}{\sigma_X} r_{1/1}\left[x_i - m_{/01}\right] \quad \text{или} \quad \frac{m_{\,1}^{(i)} - m_{0\,1}}{\sigma_Y} = r_{1/1} \; \frac{x_i - m_{1/0}}{\sigma_X}.$$

Перейдем теперь к так называемым *нормальным координатам*, полагая

$$\frac{m_{\,1}^{(i)} - m_{0/1}}{\sigma_1} = \mathfrak{M}_{/1}^{(i)}, \quad \frac{x_i - m_{1/0}}{\sigma_X} = \mathfrak{x}_i.$$

Уравнение линейной регрессии Y по X примет тогда вид, особенно удобный для алгебраических выкладок:

$$\mathfrak{M}_{/1}^{(i)} = r_{1/1}\,\mathfrak{X}_i.$$

Отсюда получаем условия, которым должны отвечать параметры r, дабы регрессия Y по X была линейною:

$$r_{h/1} = r_{1/1}\,r_{h-1/0} \qquad \text{при } h = 2,\ 3,\ 4\ldots.$$

Переход к нормальным координатам сводится к измерению разностей, входящих в уравнение регрессии вида (II), при посредстве соответственных квадратических ошибок. Выгода его заключается в том, что все коэффициенты в уравнении регрессии в нормальных координатах принимают вид отвлеченных чисел, зависящих лишь от параметров r. Предположим снова, что линия регрессии имеет вид параболы f-ой степени. Уравнение ее в нормальных координатах имеет вид:

$$\mathfrak{M}_{/1}^{(i)} = c_{/0} + c_1\,\mathfrak{X}_i + c_2\,\mathfrak{X}_i^2 + \cdots + c_f\,\mathfrak{X}_i^f.$$

Отсюда, как ранее, находим

(III) $$r_{h/1} = c_0\,r_{h/0} + c_1\,r_{h+1/0} + \cdots + c_f\,r_{h+f/0}.$$

Придавая h значения 0, 1, 2, 3, $\cdots$, и т. д. до f, получаем систему $f+1$ линейных уравнений, позволяющих выразить коэффициенты c через параметры r. Подставляя полученные для коэффициентов выражения в уравнение (III), устанавливаем условия, которым должны отвечать параметры r, дабы регрессия Y по X имела вид параболы f-ой степени. В дополнение к приведенным выше формулам для случая линейной регрессии Y по X, рассмотрим также соответственные формулы для случая, когда регрессия Y по X имеет вид параболы второй степени. Уравнение регрессии приводится в этом случае в нормальных координатах к виду

$$\mathfrak{M}_{/1}^{(i)} = -\frac{r_{2/1} - r_{3/0}\,r_{1/1}}{r_{4/0} - r_{3/0}^2 - 1} + \left\{r_{1/1} - \frac{r_{3/0}\,[r_{2/1} - r_{3/0}\,r_{1/1}]}{r_{4/0} - r_{3/0}^2 - 1}\right\}\mathfrak{X}_i +$$

$$+ \frac{r_{2/1} - r_{3/0}\,r_{1/1}}{r_{4/0} - r_{3/0}^2 - 1}\,\mathfrak{X}_i^2.$$

Существенно заметить, что член, свободный от $\mathfrak{X}_i$, обращающийся в нуль в уравнении регрессии линейной, не пропадет в уравнении регрессии параболической, и что при $r_{1/1} = 0$ уравнение сводится к

$$\mathfrak{M}_{/1}^{(i)} = \frac{r_{2/1}}{r_{4/0} - r_{3/0}^2 - 1}\,\{\mathfrak{X}_i^2 - r_{3/0}\,\mathfrak{X}_i - 1\}.$$

На этих весьма важных отличиях от линейной регрессии мы еще будем иметь случай остановиться подробнее.

Условия, которым должны отвечать параметры r, дабы регрессия Y по X имела вид параболы второй степени, могут быть сведены к требованию, чтоб

$$\frac{r_{h1} - r_{11} r_{h+10}}{r_{h+20} - r_{30} r_{h+10} - r_{h0}} = \frac{r_{21} - r_{11} r_{30}}{r_{40} - r_{30}^2 - 1} \quad \text{при } h = 3, 4, 5, \cdots$$

3) Если переменная Y некоррелирована с X (см. выше глава третья, § IV, 3)), т.-е. если условное математическое ожидание Y остается неизменным при всех значениях X, то уравнение регрессии Y по X принимает вид

$$m_1^{(1)} - m_{01} = 0 \quad \text{или} \quad \mathfrak{M}_1^{(1)} = 0.$$

Для того, чтоб переменная Y была некоррелирована с X, параметры m μ и r должны удовлетворять условиям:

$$m_{h1} = m_{h0} m_{01}, \quad \mu_{h1} = 0, \quad r_{h1} = 0 \quad \text{при } h = 1, 2, 3, \cdots$$

Положим $h = 1$ в равенстве $r_{h1} = 0$: мы находим: $r_{11} = 0$. Некоррелированность переменной Y с X обусловливает равенство нулю коэффициента корреляции между X и Y. Равным образом и некоррелированность X с Y приводит к $r_{11} = 0$. Но обратное соотношение не имеет силы: из того, что коэффициент корреляции между X и Y равняется нулю, не следует, что переменная Y некоррелирована с X или переменная X некоррелирована с Y. Лишь при наличии линейной регрессии равенство коэффициента корреляции нулю обусловливает собою некоррелированность. Если регрессия Y по X линейна, то подставляя $r_{11} = 0$ в уравнение $\mathfrak{M}_1^{(1)} = r_{11} \mathfrak{X}_1$, находим $\mathfrak{M}_1^{(1)} = 0$. Если же регрессия нелинейна, то из $r_{11} = 0$ не следует, чтоб было $\mathfrak{M}_1^{(1)} = 0$. Мы имели случай убедиться в этом на примере линии регрессии, имеющей вид параболы второй степени (см. выше § III, 2), C)).

Существенно, далее, отметить, что из $\mu_{h1} = 0$ при $h = 1, 2, 3, \cdots$, не следует, чтобы μ_{1h} равнялось нулю при $h = 2, 3, \cdots$: некоррелированность Y с X не обусловливает собою некоррелированности X с Y.

Если переменные взаимно независимы, то они вместе с тем и некоррелированы. Это следует непосредственно из определений понятий и может быть легко выведено из установленных выше условий взаимной независимости и некоррелированности: замечая, что $\mu_{10} = \mu_{01} = 0$, находим из условия взаимной независимости, — $\mu_{fg} = \mu_{f0} \mu_{0g}$ при любых целых положительных f и g, — что $\mu_{f1} = 0$ и $\mu_{1g} = 0$ при любых целых положительных f и g.

Обратное заключение не имеет и здесь силы: некоррелированность не обусловливает собою взаимной независимости даже в том случае, когда некоррелирована как переменная Y с X, так и переменная X с Y. Из $\mu_{h1} = 0$ и $\mu_{1h} = 0$ при $h = 1, 2, 3, \cdots$ не вытекает, чтобы μ_{fg} равнялось произведению $\mu_{f0} \mu_{0g}$ при f и g, отличных от 1.

4) Допустим, что линия регрессии Y по X имеет вид параболы p-ой степени, и поставим задачею подобрать прямую, которая наилучше подходила бы к истинной линии регрессии Y по X — наилучше в том смысле, чтоб сумма квадратов разностей между условными математическими ожиданиями Y, вычисленными по уравнению прямой, и истинными условными математическими ожиданиями Y была наименьшею.

Придавая уравнению искомой прямой вид

$$M_1^{(i)} = A_0 + A_1 x_i,$$

мы должны так выбрать коэффициенты A_0 и A_1, чтобы сумма

$$\sum_i p_i \, [m_1^{(i)} - M_1^{(i)}]^2 = \sum_i p_i \, [m_1^{(i)} - A_0 - A_1 x_i]^2$$

получала возможно малое значение. Принимая во внимание, что

$$\sum_i p_i = 1, \quad \sum_i p_i \, x_i = m_{1,0}, \quad \sum_i p_i \, m_1^{(i)} = m_{0\,1},$$

$$\sum_i p_i \, x_i \, m_1^{(i)} = m_{1\,1}, \quad \sum_i p_i \, x_i^2 = m_{2\,0},$$

приходим к уравнениям

$$m_{0\,1} - A_0 - A_1 m_{1\,0} = 0, \quad m_{1\,1} - A_0 m_{1,0} - A_1 m_{2,0} = 0,$$

из которых находим

$$A_1 = \frac{m_{1\,1} - m_{1,0} m_{0\,1}}{m_{2\,0} - m_{1,0}^2} = \frac{\mu_{1\,1}}{\mu_{2,0}}, \quad A_0 = \frac{m_{2,0} m_{0\,1} - m_{1,0} m_{1,1}}{m_{2\,0} - m_{1,0}^2} = m_{0/1} - m_{1\,0} A_1.$$

Для A_0 и $A_{/1}$ получаются, таким образом, те же значения, как в случае, когда истинная линия регрессии Y по X прямая (см. выше § III, 2), A)). Прямая с уравнением

$$m_1^{(i)} - m_{0\,1} = \frac{\mu_{1/1}}{\mu_{2/0}} [x_i - m_{1,0}]$$

или, в нормальных координатах, $\mathfrak{M}_{/1}^{(i)} = r_{1/1} \mathfrak{X}_i$ изображает истинную линию регрессии, если регрессия действительно прямолинейная, и наилучше подходит к истинной линии регрессии, если регрессия в действительности криволинейная. Это свойство прямой $\mathfrak{M}_{/1}^{(i)} = r_{1/1} \mathfrak{X}_i$ представляется для статистика в высокой мере ценным. Оно позволяет получить известное представление о регрессии при любом виде закона связи путем вычисления уравнения прямой $\mathfrak{M}_1^{(i)} = r_{1/1} \mathfrak{X}_i$.

5) В уравнении регрессии находит себе выражение одно из тех свойств стохастической связи между переменными, которые представляют наибольший интерес для исследователя. Оно дает картину изменений среднего уровня значений одной из случайных переменных с ростом значений другой переменной. Его содержание в случае стохастической связи подобно тому, какое имеет формула закона функциональной зависимости между переменными неслучайными. Зная формулу, связывающую Y с X функционально, мы в состоянии предвычислять значения Y для данных значений X, не обращаясь к непосредственному их измерению. Сходным образом уравнение регрессии Y по X облегчает предвычисление ожидаемого значения Y по данным значениям X. Оно, правда, не позволяет судить с уверенностью о том, какое значение переменная Y примет в каждом отдельном случае. При стохастической связи Y с X переменная Y остается случайною переменною и после того, как X получает определенное значение; отдельные же значения случайной переменной группируются кругом ее математического ожидания с размахом, характеризуемым размерами средней квадратической ошибки. Знание уравнения регрессии позволяет нам, таким образом, предвычислять ожидаемые значения Y по данным значениям X лишь с тою степенью надежности, какая соответствует размерам условных рассеяний Y. Но, как мы видели (см. выше § III. 1)), средний размер условного рассеяния связан с полным рассеянием Y формулою

$$\sum_i p_i \mu_2^{(i)} = \mu_{0\,2} - \sum_i p_i [m_1^{(i)} - m_{0\,1}]^2.$$

Лишь при некоррелированности переменной Y с X

$$\sum_i p_i \mu_2^{(i)}$$

может равняться $\mu_{0,2}$; если же переменная Y коррелирована с X, то разности $m_1^{(i)} - m_{0\,1}$ не обращаются в нуль, и средний размер условного рассеяния Y всегда меньше полного рассеяния Y. Не располагая знанием уравнения регрессии Y по X, мы в состоянии оценивать ожидаемое значение переменной Y по ее полному математическому ожиданию со степенью надежности, намечаемою величиною $\mu_{0\,2}$. Если же мы судим об ожидаемом значении Y, опираясь на знание уравнения регрессии, то средняя степень надежности наших предвычислений определяется размерами

$$\sum_i p_i \mu_2^{(i)}.$$

При некоррелированности Y с X мы ничего не выигрываем, привлекая уравнение регрессии. Но если переменная Y коррелирована с X, то уравнение регрессии Y по X позволяет предучитывать ожидаемое значение Y с меньшим риском ошибки, чем на основании одного лишь закона распределения значений Y.

6) Так же, как условные математические ожидания, могут быть изучаемы и иные сводные характеристики условных законов распределения значений, — условное рассеяние или условная средняя квадратическая ошибка, различные меры асимметрии условных законов распределения и т. д. Приемы их изучения имеют крупный теоретический интерес, — в частности для построения методов исследования стохастической связи между более, чем двумя случайными переменными. Я не буду тем не менее останавливаться на них с подробностью, чтоб не загромождать изложения. Обходя их, мы не много теряем, так как принципиально нового эти приемы в теорию корреляции не вносят, на практике же применяются редко.

§ IV.

1) Третью группу приемов изучения стохастической связи между случайными переменными образуют сводные характеристики закона связи, подобные показателю взаимной сопряженности (см. выше § II, 1)), в том отношении, что имеют задачей обрисовать одним числом те или иные существенные для исследователя черты закона связи, но отличающиеся от показателя взаимной сопряженности тем, что используют в этих целях не только вероятности разных значений переменных, но и самые значения. Мы остановимся с некоторой подробностью на двух таких мерах: на коэффициенте корреляции и на корреляционном отношении. Вычисление коэффициента корреляции является ныне излюбленным приемом измерения тесноты стохастической связи между случайными переменными. Представляется, однако, более практичным начать с рассмотрения корреляционного отношения, так как ценность коэффициента корреляции для статистика покоится отчасти на том, что при некоторых условиях коэффициент корреляции способен служить заменою корреляционного отношения.

2) Мы уже видели (см. выше § III, 5)), сколь ценною представляется для сжатой характеристики стохастической связи между случайными переменными величина среднего условного рассеяния

$$\overline{\quad} \qquad \sum_i p_i \mu_2^{(i)}.$$

К ней примыкает мера тесноты стохастической связи, построенная Пирсоном и названная им корреляционным отношением. Обозначая, по примеру Пирсона, корреляционное отношение греческою буквою η, определим *корреляционное отношение* Y к X формулою

$$\eta_{Y\,X}^2 = 1 - \frac{1}{\mu_{0\,2}} \sum_i p_i \mu_2^{(i)}.$$

Корреляционное отношение является, таким образом, дополнением до единицы частного от деления среднего условного рассеяния на полное рассеяние. Замечая, что (см. выше § III, 1)),

$$\sum_i p_i \mu_2^{(i)} = \mu_{0\,2} - \sum_i p_i [m_1^{(i)} - m_{0\,1}]^2,$$

находим

$$\eta_{Y\,X}^{\,2} = \frac{1}{\sigma_{0\,2}} \sum_{i} p_i \, [m_1^{(i)} - m_{0\,1}]^2$$

или в нормальных координатах (см. выше § III, 2, C)),

$$\eta_{Y\,X}^{\,2} = \sum_{i} p_i \, [\mathfrak{M}_1^{(i)}]^2.$$

Существо стохастической связи Y с X заключается в том, что переменная Y сохраняет характер случайной переменной и после того, как X получает определенное значение, другими словами, в том, что условные рассеяния Y отличны от нуля. При наличии функциональной зависимости Y от X все условные рассеяния Y обращаются в нуль, и корреляционное отношение Y к X принимает значение 1. Обратно: из равенства корреляционного отношения Y к X единице следует, что все условные рассеяния Y равны нулю, т.-е. что Y находится в однозначной функциональной зависимости от X. С другой стороны, корреляционное отношение Y к X принимает значение 0, если среднее условное рассеяние Y равно полному рассеянию Y, т.-е. в случае некоррелированности Y с X. Обратно, из равенства корреляционного отношения нулю следует, что среднее условное рассеяние Y равно полному рассеянию Y, а это, в свою очередь, свидетельствует о некоррелированности Y с X. Так как условное рассеяние не может принимать отрицательных значений, то корреляционное отношение не может быть больше 1. Его численная величина заключена в пределах 0 и 1. Корреляционное отношение Y к X тем больше, чем ближе взаимоотношения между Y и X уподобляются функциональной зависимости, чем меньше в среднем размах случайных колебаний Y при данных значениях X, — словом, чем теснее стохастическая связь между Y и X. Если корреляционное отношение Y к X равно 0, то переменная Y некоррелирована с X, и знание значения, полученного переменною X, не повышает степени надежности нашей оценки ожидаемого значения Y по сравнению с тою, к какой мы приходим, не считаясь с переменной X, на основе одного лишь знания закона распределения значений Y. Если корреляционное отношение Y к X равно 1, то переменная Y находится в однозначной функциональной зависимости от X и, зная полученное переменною X значение, мы по формуле, связующей Y с X, с полной точностью предустанавливаем значение Y. При значениях корреляционного отношения Y к X, промежуточных между 0 и 1, степень надежности наших предвычислений повышается, если мы имеем возможность опереться на уравнение регрессии Y по X и на знание значения полученного переменною X. Чем ближе корреляционное отношение к 1, тем заметнее при этом наш выигрыш. Корреляционное отношение даст, таким образом, весьма ценную характеристику стохастической связи, очень существенно дополняющую те представления касательно закона связи, какие даются уравнением регрессии. Каков бы ни был вид уравнения регрессии и каков бы ни был вообще вид закона связи, численные значения корреляционного отношения всегда сохраняют один и тот же смысл. В этом крайне важное его преимущество перед

коэффициентом корреляции, численные значения которого лишь при линейной регрессии имеют тот смысл, какой им обычно присваивается (см. ниже § IV, 3)).

При пользовании корреляционным отношением, как мерою тесноты связи, существенно не упускать из виду, что корреляционное отношение не является мерою абсолютной тесноты связи: численное значение корреляционного отношения Y к X свидетельствует лишь о том, насколько убывает в среднем размах колебаний переменной Y после того, как переменная X получает определенное значение; оно ничего не говорит о том, каков размах колебаний Y при данных значениях X. В частности, значение 0 корреляционного отношения Y к X свидетельствует лишь об отсутствии коррелированности Y с X; равенство корреляционного отношения Y к X нулю означает, что определение значения X не сказывается на размахе колебаний Y, но не дает ни малейшего права на заключение, велик или мал этот размах. Для суждения об абсолютной тесноте связи Y с X необходимо привлекать величину

$$\sum_i p_i \mu_2^{(i)} = \mu_{0\,2} [1 - \eta_{1|X}^2].$$

При вычислении корреляционного отношения по закону связи между переменными можно исходить как из формулы

$$\eta_{1|X}^2 = 1 - \frac{1}{\mu_{0\,2}} \sum_i p_i \mu_2^{(i)},$$

так и из формулы

$$\eta_{1|X}^2 = \frac{1}{\mu_{0,2}} \sum_i p_i [\mathfrak{M}_1^{(i)}]^2.$$

Подставляя в последнюю формулу значение $\mathfrak{M}_1^{(i)}$ из уравнения регрессии, мы без особого труда выражаем корреляционное отношение чрез коэффициенты уравнения регрессии и, через их посредство, чрез параметры r. Так, в случае, если регрессия Y по X имеет вид параболы второй степени (см. выше § III, 2) C)), находим, после некоторых преобразований,

$$\eta_{Y|X}^2 = r_{1/1}^2 + \frac{[r_{2/1} - r_{3/0} r_{1/1}]^2}{r_{4/0} - r_{3/0}^2 - 1}.$$

3) В случае прямолинейной регрессии Y по X с уравнением

$$\mathfrak{M}_1^{(i)} = r_{1/1} \mathfrak{X}_i,$$

замечая, что

$$\sum_i p_i \mathfrak{X}_i^2 = 1,$$

находим

$$\eta_{Y|X}^2 = r_{1\,1}^2.$$

При прямолинейной регрессии Y по X корреляционное отношение Y к X и коэффициент корреляции между X и Y оказываются, таким образом, по своей абсолютной величине тожественно равными.

Если регрессия Y по X не прямолинейная, то r_{11}^2 всегда меньше $\eta_{Y/X}^2$. Доказать это всего проще следующим путем. Обозначим, как ранее (см. выше § III, 4)), чрез $M_1^{(i)}$ величину условного математического ожидания Y, находимую по уравнению прямой, наилучше подгоняемой к истинной линии регрессии Y по X. Сохраняя обычное обозначение $m_1^{(i)}$ для истинного условного математического ожидания Y и замечая, что

$$M_1^{(i)} = m_{01} + \frac{\mu_{11}}{\mu_{20}}[x_i - m_{10}],$$

находим

$$\sum_i p_i [m_1^{(i)} - M_1^{(i)}]^2 = \mu_{02}[\eta_{Y/X}^2 - r_{11}^2]$$

и отсюда

$$\eta_{Y/X}^2 - r_{11}^2 = \frac{1}{\mu_{02}} \sum_i p_i [m_1^{(i)} - M_1^{(i)}]^2.$$

Разность $\eta_{Y/X}^2 - r_{11}^2$ не может быть, следовательно, отрицательною. Она может обращаться в 0 лишь, если все величины $m_1^{(i)}$ совпадают с соответственными величинами $M_1^{(i)}$, т.-е. если истинная линия регрессии Y по X прямая с уравнением

$$m_1^{(i)} = m_{01} + \frac{\mu_{11}}{\mu_{20}}[x_i - m_{10}].$$

При линейной регрессии Y по X численные значения коэффициента корреляции могут быть, таким образом, истолковываемы совершенно так, как истолковываются значения корреляционного отношения Y к X: равенство коэффициента корреляции единице свидетельствует о наличии линейной функциональной зависимости Y от X; равенство коэффициента корреляции нулю свидетельствует о некоррелированности Y с X; чем больше коэффициент корреляции по абсолютной величине, тем меньшим случайным колебаниям подвержена переменная Y после того, как X получает определенное значение, тем теснее связь между Y и X, тем ближе взаимоотношения между Y и X подходят к линейной функциональной зависимости.

Но при криволинейной регрессии Y по X численные значения коэффициента корреляции уже не могут быть так истолковываемы. При криволинейной регрессии Y по X коэффициент корреляции остается всегда по абсолютной величине меньше корреляционного отношения Y к X. При функциональной, но не линейной зависимости Y от X коэффициент корреляции принимает значение не 1, а отстает от 1 больше или меньше, смотря по виду функциональной зависимости. Равным образом и равенство коэффициента корреляции нулю не служит при криволинейной регрессии порукою в том, что переменная Y некоррелирована с X. Оба предельных значения коэффициента корреляции имеют при непрямолинейной

регрессии не тот смысл, какой присущ им при регрессии прямолинейной. Если не представляется доподлинно известным, что регрессия прямолинейная, то численные значения коэффициента корреляции надлежит истолковывать с осторожностью. Во избежание ошибок лучше сопровождать или заменять в таких случаях вычисление коэффициентов корреляции вычислением корреляционных отношений, сохраняющих один и тот же смысл при всяком виде закона связи. Корреляционное отношение представляется более совершенною мерою тесноты стохастической связи между случайными переменными, нежели коэффициент корреляции. Коэффициент корреляции является вполне надежною мерою тесноты связи лишь тогда, когда он заступает место корреляционного отношения, будучи ему численно равен. Во всех же остальных случаях мы более или менее недооцениваем тесноту связи, измеряя ее при помощи коэффициента корреляции.

Интерес, который представляет вычисление коэффициента корреляции, не исчерпывается, однако, тем, что при известных условиях коэффициент корреляции может служить хорошею мерою тесноты связи. Мы видели (см. выше § III, 4)), что и в случае криволинейной регрессии прямая с уравнением, в нормальных координатах,

$$\mathfrak{M}_1^{(x)} = r_{11}\mathfrak{X}_1$$

может быть рассматриваема, как приближенное воспроизведение истинной линии регрессии Y по X. В качестве такового, она всегда представляет известный интерес для исследователя. Прямая эта указывает, нарастает или убывает в общем и целом условное математическое ожидание Y с ростом X, и дает представление о среднем темпе нарастания и убывания. Уравнение же этой прямой определяется, в свою очередь, численною величиною коэффициента корреляции. Если коэффициент корреляции положителен, то условное математическое ожидание Y нарастает в среднем с ростом X и нарастает тем быстрее, чем больше коэффициент корреляции; если коэффициент корреляции отрицателен, то условное математическое ожидание Y в среднем убывает с ростом X и убывает тем быстрее, чем ближе коэффициент корреляции к -1. Величина коэффициента корреляции определяет и уравнение прямой, наилучше передающей истинное уравнение регрессии X по Y. Численное значение коэффициента корреляции дает, таким образом, и в случае криволинейной регрессии в сжатой форме довольно содержательное представление о стохастической связи между переменными, не смотря на то, что измеряет тесноту связи не вполне совершенно.

В случае прямолинейной регрессии численные значения коэффициента корреляции допускают иногда особенно богатое содержанием истолкование. Предположим, что подбрасывается m белых и n красных костей; обозначим через U сумму очков, выпадающих на красных костях, через W сумму очков на белых костях и положим $X = U + W$. Оставим белые кости лежать, как они выпали, а красные кости подберем и снова бросим: обозначим через T сумму очков, выпадающих на красных костях при вторичном подбрасывании, и положим $Y = W + T$. Если мы вычислим по определяющей его формуле коэффициент корреляции между X и Y, то найдем его равным отношению числа белых костей

ко всему числу подбрасываемых костей: $r_{1,1} = \frac{m}{m+n}$. Численное значение коэффициента корреляции позволяет тут судить о доле костей, оставляемых без перебрасывания. Примеру этому может быть придана и более общая форма: если переменные X и Y могут быть представлены в виде сумм взаимно независимых слагаемых, следующих одному и тому же закону распределения значений, то коэффициент корреляции между X и Y равняется отношению числа общих обеим суммам слагаемых к геометрической средней из чисел слагаемых в каждой из сумм:

$$r_{1,1} = \frac{m}{\sqrt{(m+n)(m+l)}}.$$

при чем m обозначает число общих слагаемых, $m+n$ — число слагаемых в составе X, $m+l$ — число слагаемых в составе Y.

§ V

1) Особое значение получает коэффициент корреляции в случае, если стохастическая связь между переменными имеет вид так называемой нормальной корреляции. *Нормально коррелированными* называются переменные X и Y, если они способны принимать все значения между $-\infty$ и $+\infty$, при чем вероятность совпадения значения переменной X, лежащего между X и $X+dX$, с значением переменной Y, лежащим между Y и $Y+dY$, равняется

$$\frac{1}{2\pi\sqrt{\mu_{2/0}\,\mu_{0/2} - \mu_{1/1}^{\,2}}} \times$$

$$\times\, e^{-\frac{\mu_{0\cdot 2}[X-m_{1/0}]^2 - 2\mu_{1/1}[X-m_{1/0}][Y-m_{0/1}] + \mu_{2/0}[Y-m_{0/1}]^2}{2[\mu_{2/0}\,\mu_{0/2} - \mu_{1/1}]^2}} dX\,dY.$$

Переходя к нормальным координатам (см. выше § III, 2), C))

$$\mathfrak{X} = \frac{X-m_{1/0}}{\sqrt{\mu_{2/0}}},\ \mathfrak{Y} = \frac{Y-m_{0,1}}{\sqrt{\mu_{0/2}}}.$$

приводим вероятность совпадения значения $\mathfrak{X}$, лежащего между $\mathfrak{X}$ и $\mathfrak{X}+d\mathfrak{X}$ с значением $\mathfrak{Y}$, лежащим между $\mathfrak{Y}$ и $\mathfrak{Y}+d\mathfrak{Y}$, к виду

$$\frac{1}{2\pi\sqrt{1-r_{1,1}^{\,2}}}\, e^{-\frac{\mathfrak{X}^2 - 2r_{1/1}\mathfrak{X}\mathfrak{Y} + \mathfrak{Y})^2}{2[1-r_{1/1}^{\,2}]}} d\mathfrak{X}\,d\mathfrak{Y}.$$

Всматриваясь в это выражение, убеждаемся, что при нормальной корреляции между X и Y закон связи между переменными определяется однозначно

величиною коэффициента корреляции r_{11} вкупе со значениями математических ожиданий переменных и их рассеяний, характеризующими законы распределения значений обеих переменных в отдельности и необходимыми для определения системы нормальных координат. Коэффициент корреляции является в силу этого при нормальной корреляции как бы ключем ко всему, что может нас интересовать во взаимоотношениях между X и Y: зная величину коэффициента корреляции r_{11}, мы располагаем знанием закона связи между X и Y и можем получить из него все, что требуется знать, путем математической дедукции.

Математические выкладки сводятся при этом к задачам из области интегрального исчисления и не представляют трудности в случае двух переменных. Я не буду приводить их в подробностях. Но на главных результатах нам следует остановиться, дабы освоиться с понятием нормальной корреляции, играющим весьма видную роль в современной теории статистики.

2) При нормальной корреляции закон связи определяется величиною коэффициента корреляции r_{11}. Все сводные характеристики закона связи могут быть, следовательно, выражены через посредство r_{11}.

Для высших параметров типа r находим:

$$r_{31} = r_{13} = 3r_{11}, \qquad r_{22} = 1 + 2r_{11}^2, \qquad r_{10} = r_{01} = 3,$$

$$r_{51} = r_{15} = 15r_{11}, \qquad r_{42} = r_{24} = 3[1 + 4r_{11}^2], \qquad r_{33} = 3r_{11}[3 + 2r_{11}^2],$$

$$r_{60} = r_{06} = 15,$$

$$r_{f\,2h+1-f} = 0,$$

$$r_{2f\,2h} = 1.3.5\cdots(2f-1).1.3.5\cdots(2h-1)\left\{1 + \sum_{t=1} \frac{2^{2t} f^t h^t}{1.2.3\cdots(2t)} r_{11}^{2t}\right\},$$

$$r_{2f+1\,2h+1} =$$

$$= 1.3.5\cdots(2f+1).1.3.5\cdots(2h+1) r_{11} \sum_{t=0} \frac{2^{2t} f^t . h^t .}{1.2.3\cdots(2t+1)} r_{11}^{2t},$$

при чем

$$z^t = z(z-1)(z-2)\cdots(z-t+1).$$

Закон связи имеет симметричный вид. Распределение значений как X, так и Y следует закону Гаусса - Лапласа. Регрессия Y по X и X по Y линейны. Корреляционные отношения Y к X и X к Y тожественно равны, в силу этого, коэффициенту корреляции и равны между собою. Коэффициент корреляции представляется при нормальной корреляции между переменными вполне совершенною мерою тесноты связи. При $r_{11} = 0$ находим

$$r_{2f+1\,2h+1} = 0 = r_{2f+1\,0} r_{0\,2h+1},$$

$$r_{2f\,2h} = 1.3.5\cdots(2f-1).1.3.5\cdots(2h-1) = r_{2f\,0} r_{0\,2h};$$

равенство коэффициента корреляции нулю свидетельствует о взаимной независимости переменных (см. выше § III, 1)).

Все условные законы распределения значений переменной X и переменной Y имеют вид закона Гаусса-Лапласа. Связь Y с X и связь X с Y гомоскедастичны: условное рассеяние X остается при всех значениях Y равным

$$\mu_2^{(y)} = \mu_{20}[1 - r_{11}^2].$$

Условное рассеяние Y остается при всех значениях X равным

$$\mu_2^{(x)} = \mu_{02}[1 - r_{11}^2].$$

Представляет интерес отметить, что показатель взаимной сопряженности переменных (см. выше § II, 1)) связан при нормальной корреляции между переменными с коэффициентом корреляции соотношением:

$$\varphi^2 = \frac{r_{11}^2}{1 - r_{11}^2}.$$

При нормальной корреляции можно, следовательно, итти к определению величины коэффициента корреляции путем вычисления показателя взаимной сопряженности, переходя затем к коэффициенту корреляции по формуле

$$r = \pm \sqrt{\frac{\varphi^2}{1 + \varphi^2}}.$$

3) В современной теории статистики нормальная корреляция занимает совершенно особое место: ей уделяется несравненно больше внимания, нежели иным законам стохастической связи между случайными переменными. Объясняется это отчасти исторически. Теория корреляции выросла на почве изучения нормальной корреляции; в своих началах она еще теснее примыкала к допущению, что стохастическая связь между изучаемыми переменными имеет вид нормальной корреляции, нежели теория приемов изучения одной случайной переменной к предположению, что распределение значений переменной следует закону Гаусса-Лапласа. Основные понятия теории корреляции — коэффициент корреляции, уравнение регрессии и т. д. — выработаны были на рассмотрении нормальной корреляции, и печать такого их происхождения долго оставалась лежать на них. Лишь к концу XIX века начали, например, статистики с большею отчетливостью выделять в коэффициенте корреляции те свойства, которые присущи ему при нормальной корреляции, от тех, какими он обладает в случае линейной регрессии при любом законе связи, и от тех, какие присущи ему всегда. В настоящее время благодаря преимущественно трудам Пирсона и Юла мы видим в нормальной корреляции лишь одну из равновидностей общего понятия закона связи между случайными переменными. Но еще недавно трудно было мыслить с некоторою ясностью стохастическую связь между случайными переменными иначе как в формах

нормальной корреляции. Это и по сейчас сказывается на изложении теории корреляции в большинстве руководств тем, что случай нормальной корреляции выдвигается на первый план.

Надо, однако, добавить, что привилегированное положение нормальной корреляции в учебниках статистики не сводится всецело к историческим пережиткам. Есть и внутренние основания к тому. Большое значение имеет прежде всего то обстоятельство, что в предположении нормальной корреляции математический анализ складывается проще и элегантнее, чем в других случаях. При нормальной корреляции из общих формул выпадают все иные параметры, кроме r_{11}, так как все они могут быть выражены через r_{11}. В силу этого формулы принимают более простой вид, существенно облегчающий применение их на практике. В связи отчасти со сравнительной простотой математической трактовки, а отчасти с тем, что изучению случая нормальной корреляции издавна уделялось наибольшее внимание, теория нормальной корреляции представляется, далее, несравненно лучше разработанною, нежели теория приемов изучения стохастической связи для законов связи иного вида. Ряд основных задач, на которых мы остановимся ниже в главе шестой, представляется более или менее удовлетворительно решенным в предположении нормальной корреляции, тогда как для иных законов связи решения едва только намечаются. Благодаря этому теория нормальной корреляции может быть ныне излагаема с полнотою и законченностью, которые пока еще недостижимы при изложении приемов изучения стохастической связи между случайными переменными в ее более общих формах.

Что касается фактического распространения взаимоотношений между случайными переменными, отвечающих понятию нормальной корреляции, то в настоящее время статистики не склонны признавать, что нормальная корреляция является общим правилом или, хотя бы, что она встречается часто. Но случаи, где закон связи не слишком резко отходит от нормальной корреляции, не так редки, и зачастую представляется возможным для упрощения работы использовать формулы, относящиеся к случаю нормальной корреляции, без ощутительного ущерба для точности результатов. Чрезвычайно существенным представляется и тут, как при использовании закона Гаусса-Лапласа для исследования одной случайной переменной, то обстоятельство, что закон связи между средними арифметическими случайных значений X и Y стремится с ростом числа испытаний принять вид нормальной корреляции при почти любом виде закона связи между переменными, если только отдельные испытания представляются взаимно независимыми. При законах связи между X и Y, не слишком уклоняющихся от нормального вида, не требуется даже большого числа испытаний, дабы закон связи между средними был близок к нормальному. На практике это позволяет при изучении стохастической связи между средними почти всегда исходить из допущения, что связь имеет вид нормальной корреляции.

§ VI.

Вид линии регрессии Y по X и теснота связи Y с X ни в коей мере не определяют друг друга. Если Y находится в функциональной зависимости от X,

то все условные рассеяния Y равны нулю, корреляционное отношение Y к X равно единице, а линия регрессии Y по X совпадает с линией, передающей графически закон функциональной зависимости Y от X, и может являть, как и последняя, вид и прямой линии, и самой замысловатой кривой. По виду линии регрессии Y по X нельзя ни в малой мере судить о тесноте связи Y с X.

Равным образом нельзя судить о тесноте связи X с Y по виду линии регрессии X по Y. Но по совокупности обеих линий регрессии представляется возможным составить известное представление о тесноте связи между переменными.

В случае функциональной зависимости между переменными уравнение, выражающее Y как функцию X, определяет собою уравнение, выражающее X как функцию Y. Отправляясь от вида одной из линий, мы всегда можем установить, какой вид должна иметь другая, если переменные находятся в функциональной зависимости. Если действительная линия регрессии второй переменной по первой имеет иной вид, то предположение, что переменные находятся в функциональной зависимости, может быть с уверенностью отвергнуто: переменные стоят в стохастической более или менее тесной связи.

Пусть, например, регрессия Y по X имеет вид прямой с уравнением

$$m_1^{(i)} = a_0 + a_1 x_i.$$

При наличии функциональной зависимости между переменными $m_{j1}^{(i)}$ совпадает с значением Y, отвечающим значению x_i переменной X, а $m_{1j}^{(j)}$ совпадает с значением X, отвечающим значению y_j переменной Y. Уравнение регрессии Y по X может быть, следовательно, в предположении функциональной зависимости начертано в виде

$$Y = a_{j0} + a_{j1} X.$$

Определяя из этого уравнения X как функцию Y, находим

$$X = -\frac{a_{j0}}{a_{j1}} + \frac{1}{a_{j1}} Y.$$

Если переменные находятся в функциональной зависимости, то уравнение регрессии X по Y должно, таким образом, иметь вид

$$m_1^{(j)} = -\frac{a_{j0}}{a_{j1}} + \frac{1}{a_{j1}} y_j.$$

Если линия регрессии X по Y не имеет вида прямой или если регрессия X по Y прямолинейна, но в уравнении регрессии X по Y

$$m_{1j}^{(j)} = a_0 + a_1 y_j$$

коэффициент a_1 не равен $\frac{1}{a_1}$, то переменные X и Y не могут находиться в функциональной зависимости.

При помощи сходных рассуждений вопрос о допустимости предположения, что переменные находятся в функциональной зависимости, решается и в тех случаях, когда ни одна из линий регрессии не имеет вида прямой.

Если обе линии регрессии прямые, мы можем составить по их уравнениям и вполне точное представление о тесноте связи между переменными. Предположим, что уравнения регрессии имеют вид

$$m_1^{(x)} = a_0 + a_1 x, \qquad m_1^{(y)} = a_0 + a_1 y.$$

Произведение коэффициентов a_1 и a_1 равняется, как мы знаем (см. выше § III, 2)), квадрату коэффициента корреляции: $a_1\, a_1 = r_{11}^2$. Раз коэффициенты a_1 и a_1 известны, мы можем, таким образом, вычислить величину коэффициента корреляции между X и Y. А так как в случае, когда обе линии регрессии прямые, квадрат коэффициента корреляции тожественно равен корреляционным отношениям Y к X и X к Y (см. выше § IV, 3)), то мы располагаем в произведении коэффициентов a_1 и a_1 мерою тесноты связи, которая признана нами за наиболее совершенную.

В случае, если действительные линии регрессии Y по X и X по Y не имеют вида прямых, но известны уравнения прямых, наилучше подходящих к действительным линиям регрессии (см. выше § III, 4)), мы также располагает возможностью установить значение коэффициента корреляции, так как произведение коэффициентов A_1 и A_1 в уравнениях таких прямых равняется r_{11}^2. Но тут уже мы не имеем права на переход от коэффициента корреляции к корреляционным отношениям. Если действительная линия регрессии Y по X не имеем вида прямой, квадрат коэффициента корреляции не равняется корреляционному отношению Y к X, а всегда меньше его, как мы видели (см. выше § IV, 3)). В таких условиях численное значение коэффициента корреляции, полученное нами путем перемножения коэффициентов A_1 и A_1, не даст нам уже точного представления о действительной тесноте связи между переменными. Оценивая тесноту связи по величине коэффициента корреляции, мы тут преуменьшаем ее, — более или менее, смотря по виду действительных линий регрессии. Коэффициент корреляции остается тут меньше единицы даже, если переменные находятся в функциональной зависимости.

§ VII.

Наш обзор тех приемов, которые применяются для сводной характеристики стохастической связи между двумя случайными переменными, не претендует на исчерпывающую полноту. Он ограничен изложением основных идей, на которых покоится современная теория методов статистического изучения стохастической связи. Приспособление общих идей к тому, как ставятся проблемы в отдельных областях применения статистического метода, и к тому, как складывается статистический материал, которым располагает исследователь, выходит из рамок наших задач. В дополнение к тому, что было сказано выше при изложении теории показателя взаимной сопряженности (см. выше § II, 2)), я позволю себе лишь

отметить, что, хотя корреляционное отношение и коэффициент корреляции не представляются в общем случае непосредственно пригодными для измерения тесноты стохастической связи между неколичественными признаками, но при условии, что оба случайно переменных качественных признака способны принимать лишь по два разных значения, теснота стохастической связи между ними допускает измерение при посредстве корреляционного отношения и коэффициента корреляции.

Предположим сперва, что X и Y имеют характер случайно переменных величин, способных принимать лишь по два точно измеримых значения, и вычислим коэффициент корреляции между ними по обычной формуле

$$r_{1\,1} = \frac{\mu_{1\,1}}{\sigma_X\,\sigma_1}.$$

Полагая, как раньше (см. выше § II, 1)), $\delta = p_{1\,1}\,p_{2\,2} - p_{1\,2}\,p_{2\,1}$, находим без труда:

$$\mu_{1\,1} = \delta\,[x_1 - x_2]\,[y_1 - y_2],\ \mu_{2\,0} = \sigma_X^2 = p_{1\prime}\,p_{2\prime}\,[x_1 - x_2]^2,$$

$$\mu_{0\,2} = \sigma_Y = p_{\prime 1}\,p_{\prime 2}\,[y_1 - y_2]^2.$$

Численные значения переменных X и Y появляются, таким образом, и в числителе, и в знаменателе выражения для $r_{1\,1}$ в виде произведения

$$[x_1 - x_2]\,[y_1 - y_2]$$

и могут быть сокращены, после чего коэффициент корреляции принимает вид

$$r_{1,1} = \frac{\delta}{\sqrt{p_{1\prime}\,p_{2\prime}\,p_{\prime 1}\,p_{\prime 2}}}.$$

Выражение для коэффициента корреляции между X и Y не содержит, таким образом, численных значений переменных. Коэффициент корреляции не меняется при каких угодно изменениях численных значений X и Y, если только остаются неизменными вероятности этих значений и разных их сочетаний. Нет, следовательно, надобности знать численные значения для того, чтоб вычислить коэффициент корреляции. Нет надобности их и измерять. Не важно даже, допускают ли они вообще измерение, или изучаемые случайно переменные признаки представляются не количественными, а качественными.

Значение коэффициента корреляции, к которому мы пришли в предположении, что X и Y способны принимать только по два разных значения, тождественно тому, которое мы ранее получили в этом предположении для показателя взаимной сопряженности переменных (см. выше § II, 1)). Вычисление корреля-

ционного отношения приводит к тому же выражению. Таким образом, в случае, если обе переменных способны принимать лишь по два разных значения, величина

$$\frac{\delta}{\sqrt{p_1' p_2 \, p_1 p_2'}}$$

может быть рассматриваема и как корреляционное отношение одной из переменных к другой, и как коэффициент корреляции между ними, и как показатель их взаимной сопряженности. В связи с общим расположением статистиков к коэффициенту корреляции к ней чаще всего подходят, как к коэффициенту корреляции между X и Y.

ГЛАВА ПЯТАЯ.

Сводные характеристики совокупности эмпирических данных.

§ I.

Закон связи и совокупность его сводных характеристик дают нам полное и точное представление о стохастической связи между случайными переменными. Ближайшею задачею изучения стохастической связи и является получение возможно отчетливых представлений о законе связи и о величине тех его сводных характеристик, которые имеют наибольшее значение для целей данного исследования. Лишь справившись так или иначе с этой задачей, — путем приемов „элементарных“ или „математических“ — исследователь получает возможность подойти к увенчанию своего труда тем выяснением действительного смысла наблюдаемой связи, которое сообщает полученным им результатам подлинную ценность (см. выше глава вторая, 2) и ниже глава восьмая § II).

Лишь в редких случаях располагает исследователь на практике возможностью составить отчетливое и надежное представление о законе связи путем таких соображений „по существу“, какими руководствуется теория вероятностей при иллюстрации общих априорных положений примерами из области так называемых случайных игр. Как правило, все, что известно о переменных и их связи, исчерпывается знанием некоторого числа пар сопряженных случайных значений переменных. По этим случайным значениям переменных исследователь вынужден судить о тех сводных характеристиках закона связи, которые его интересуют. Изыскание наивыгоднейших приемов подхода к оценке „априорных“ величин по эмпирическим данным и составляет главный предмет забот теории корреляции.

Прежде чем обратиться к систематическому обзору выработанных теорией приемов решения этой задачи, необходимо остановиться подробнее на самом эмпирическом материале и подвергнуть рассмотрению те приемы, при помощи которых он облекается в формы, наиболее удобные для дальнейшей обработки.

§ II.

1) Исходные эмпирические данные имеют вид некоторого числа пар сопряженных случайных значений X и Y. Обозначим через N общее число пар; через n_{ij} — число таких пар, где переменная X имеет значение x_i, а перемен-

ная Y имеет значение y_j; через n_i — число пар, где переменная X имеет значение x_i, какое бы при этом ни имела значение переменная Y; через n_j — число пар, где переменная Y имеет значение y_j, какое бы при этом ни имела значение переменная X. Обозначая через k все число разных значений, получаемых переменною X, и через l все число разных значений, получаемых переменною Y, имеем тожественно:

$$n_i = \sum_{j=1}^{l} n_{ij}, \quad n_j = \sum_{i=1}^{k} n_{ij}.$$

$$N = \sum_{i=1}^{k} n_{ij} = \sum_{j=1}^{l} n_j = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}.$$

Совокупность чисел n_{ij}, расположенных в удобообозримой форме по образцу, приведенному выше (см. глава первая, § II), именуется *корреляционною таблицею*. Каждая строка и каждый столбец корреляционной таблицы именуются *рядом*, при чем ряд характеризуется значением той переменной, которая остается неизменною для всех членов ряда: совокупность чисел n_{ij} при неизменном i образует ряд x_i; совокупность чисел n_{ij} при неизменном j образует ряд y_j.

Совокупность чисел n_i дает распределение численностей значений переменной X. Совокупность чисел n_j дает распределение численностей значений переменной Y. Совокупность чисел n_{ij} при неизменном i даст распределение численностей значений Y для ряда x_i. Совокупность чисел n_{ij} при неизменном j дает распределение численностей значений X для ряда y_j.

Положим:

$$\frac{n_{ij}}{N} = p_{ij}', \quad \frac{n_i}{N} = p_i', \quad \frac{n_j}{N} = p_j',$$

$$\frac{n_{ij}}{n_j} = \frac{p_{ij}'}{p_j'} = p_i^{(j)'}, \quad \frac{n_{ij}}{n_i} = \frac{p_{ij}'}{p_i'} = p_j^{(i)'}.$$

Совокупность чисел p_i' дает распределение частостей значений X; совокупность чисел $p_i^{(j)'}$ дает распределение частостей значений X для ряда y_j. Совокупность чисел p_j' дает распределение частостей значений Y; совокупность чисел $p_j^{(i)'}$ дает распределение частостей значений Y для ряда x_i.

Помечая все пары сопряженных значений X и Y последовательными номерами, будем обозначать через $x^{[f]'}$ и $y^{[f]'}$ сопряженные значения переменных f-ой пары. Обозначим через x_0' среднюю арифметическую всех значений X, через $x_0^{(j)'}$ — среднюю арифметическую значений X для ряда y_j, через y_0' — среднюю арифметическую всех значений Y, через $y_0^{(i)'}$ — среднюю арифметическую значений Y для ряда x_i. Согласно определениям, имеем тожественно:

$$x_0' = \frac{1}{N} \sum_{f=1}^{N} x^{[f]'} = \frac{1}{N} \sum_{i=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} p_{ij}' x_i,$$

$$y_0' = \frac{1}{N}\sum_{f=1}^{N} x^{[f]'} = \frac{1}{N}\sum_{j=1}^{l} n_j y_j = \sum_{j=1}^{l} p_j' y_j,$$

$$x_0^{(j)'} = \frac{1}{n_j}\sum_{i=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} p_i^{(j)'} x_i,$$

$$y_0^{(i)'} = \frac{1}{n_i}\sum_{j=1}^{l} n_{ij} y_j = \sum_{j=1}^{l} p_j^{(i)'} y_j.$$

Замечая, что

$$\sum_{j=1}^{l} n_{ij} x_0^{(j)'} = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} n_i x_i,$$

находим

$$x_0' = \sum_{j=1}^{l} \frac{n_j}{N} x_0^{(j)'} = \sum_{j=1}^{l} p_j' x_0^{(j)'}.$$

Сходным образом получаем

$$y_0' = \sum_{i=1}^{k} p_i' y_0^{(i)'}.$$

2) Подобно тому, как закон связи между X и Y может почитаться данным, если известен ряд целесоответственно подобранных параметров $m_{f/g}$ (см. глава четвертая § III, 1)), так же точно и совокупность чисел p_{ij}' может быть дана в виде ряда величин $m_{f/g}'$, определяемых соотношениями

$$m_{f/g}' = \sum_i \sum_j p_{ij}' x_i^f y_j^g.$$

При $g = 0$ мы получаем ряд величин $m_{f/0}'$, воспроизводящий распределение частостей переменной X. При $f = 0$ получаем ряд величин $m_{0/g}'$, воспроизводящий распределение частостей переменной Y. В частности величина

$$m_{1/0}' = \sum_i \sum_j p_{ij}' x_i = \sum_i p_{i/}' x_i = x_0'$$

равняется средней арифметической всех значений X, а величина

$$m_{0/1}' = \sum_j p_{/j}' y_j = y_0'$$

равняется средней арифметической всех значений Y.

Для передачи совокупности чисел $p_{i,j}'$ могут быть также привлекаемы величины $\mu_{f,g}'$ и $r_{f,g}$, определяемые соотношениями

$$\mu_{f,g}' = \sum_i \sum_j p_{i,j}' [x_i - m_{1,0}']^f [y_j - m_{0,1}']^g$$

$$r_{f,g} = \frac{\mu_{f,g}'}{[\mu_{2,0}']^{\frac{1}{2}f} [\mu_{0,2}']^{\frac{1}{2}g}}.$$

И здесь при $g = 0$ мы получаем ряд величин $\mu_{f,0}'$ и $r_{f,0}'$, характеризующих распределение частостей переменной X, а при $f = 0$ ряд величин $\mu_{0,g}'$ и $r_{0,g}'$, характеризующих распределение частостей переменной Y. В частности величина

$$\mu_{0,2}' = \sum_i p_{i,j}' [x_i - m_{1,0}']^2 = \frac{1}{N} \sum_i n_{i,j} [x_i - x_0']^2 = m_{2,0}' - [m_{1,0}']^2$$

дает *эмпирическое рассеяние* X, а величина

$$\mu_{0,2}' = \sum_j p_{i,j}' [y_j - m_{0,1}']^2 = \frac{1}{N} \sum_j n_{i,j} [y_j - y_0']_1 = m_{0,2}' - [m_{0,1}']^2$$

дает эмпирическое рассеяние Y.

Первую из величин ряда $r_{f,g}'$ именно,

$$r_{1,1}' = \frac{\mu_{1,1}'}{\sqrt{\mu_{2,0}' \mu_{0,2}'}} = \frac{\sum_i \sum_j p_{i,j}' [x_i - m_{1,0}'] [y_j - m_{0,1}']}{\sqrt{\{\sum_i p_{i,j}' [x_i - m_{1,0}']^2\} \{\sum_j p_{i,j}' [y_j - m_{0,1}']^2\}}} =$$

$$= \frac{\sum_{i=1}^{N} [x^{[i]'} - x_0'] [y^{[i]'} - y_0']}{\sqrt{\{\sum_{i=1}^{N} [x^{[i]'} - x_0']^2\} \{\sum_{i=1}^{N} [y^{[i]'} - y_0']^2\}}}$$

будем называть: *эмпирический коэффициент корреляции.*

Замечая, что

$$\mu_{1,1}' = m_{1,1}' - m_{1,0}' m_{0,1}',$$

можем придать эмпирическому коэффициенту корреляции также вид

$$r_{1,1}' = \frac{m_{1,1}' - m_{1,0}' m_{0,1}'}{\sqrt{\{m_{2,0}' - [m_{1,0}']^2\} \{m_{0,2}' - [m_{0,1}']^2\}}}.$$

3) Обозначая, как ранее (см. глава четвертая, § III, 1)), пометкою справа наверху в скобках тот ряд, к которому относится рассматриваемая величина, положим:

$$m_g^{(j')} = \sum_j p_j^{(j')} y_j^g, \qquad m_i^{(i')} = \sum_i p_{ii}^{(i')} x_i^r,$$

$$\mu_{jg}^{(j')} = \sum_j p_j^{(j')} [y_j - m_1^{(j')}]^g, \quad \mu_i^{(i')} = \sum_i p_{ii}^{(i')} [x_i - m_1^{(i')}]^r.$$

Из многочисленных тожеств, связывающих сводные характеристики отдельных рядов со сводными характеристиками всей совокупности значений переменных, отметим лишь соотношения между средними арифметическими и между рассеяниями:

$$m_{1,0}' = \sum_j p_j' m_1^{(j')}, \quad m_{0,1}' = \sum_i p_i' m_1^{(i')},$$

$$\mu_{2,0}' = \sum_j p_j' \mu_2^{(j')} + \sum_j p_j' [m_1^{(j')} - m_{1,0}']^2,$$

$$\mu_{0,2}' = \sum_i p_i' \mu_2^{(i')} + \sum_i p_i' [m_1^{(i')} - m_{0,1}']^2$$

§ III.

Нанесем на график в прямоугольных координатах значения $m_1^{(i')}$, отвечающие разным значениям переменной X. Последовательность намечаемых ими точек будем называть: *эмпирическая ломаная регрессии* Y по X.

По отношению к эмпирической ломаной регрессии может быть формально поставлена задача, аналогичная той, которую мы рассматривали по отношению к априорной линии регрессии (см. глава четвертая, § III, 2)), — именно: в предположении, что точки ломаной лежат на плавной кривой, имеющей вид параболы f-ой степени, выразить коэффициенты уравнения параболы чрез посредство величин m', μ' и r'. Математическое решение этой задачи сходно с тем, к какому мы пришли выше. Задача эта лишена, однако, статистического интереса. Если сопряженные значения X и Y складываются случайно так, что все точки ломаной попадают на сравнительно несложную плавную кривую, статистик не в состоянии сделать из подобного случайного совпадения ценных выводов, так как знает, что достаточно было бы увеличить число сопряженных пар значений X и Y и видимость расположения точек на несложной плавной кривой исчезла бы.

В плоскости построений, имеющих дело с эмпирическими случайными данными, интерес статистика сосредоточивается на уравнении такой прямой линии или сравнительно несложной кривой, которые наилучше подходят к совокупности точек, независимо от того, располагаются ли эти точки случайно вполне точно

на подгоняемой кривой или только беспорядочно группируются вокруг нее. Предположим, что требуется подогнать прямую линию так, чтоб взвешенная средняя квадратов разностей между истинными средними рядов x_i и величинами, получаемыми по уравнению прямой, была возможно мала. Напишем уравнение искомой прямой в виде

$$M_{1}^{(x)} = A_{0}' + A_{1}' x_i.$$

Условие, чтоб сумма

$$\sum_i p_i' [m_1^{(x)} - M_1^{(x)}]^2 = \sum_i p_i' [m_1^{(x)} - A_0' - A_1' x_i]^2$$

принимала возможно малое значение, приводит к уравнениям

$$m_{0,1}' - A_0' - m_{1,0}' A_1' = 0, \quad m_{1,1}' - m_{1,0}' A_0' - m_{2,0}' A_1' = 0,$$

из которых находим

$$A_1' = \frac{m_{1,1}' - m_{1,0}' m_{0,1}'}{m_{2,0}' - [m_{1,0}']^2} = \sqrt{\frac{\mu_{0,2}'}{\mu_{2,0}'}}\, r_{1,1}', \quad A_0' = m_{0,1}' - A_1' m_{1,0}'.$$

Уравнению подгоняемой прямой может быть, следовательно, придан вид

$$M_{1}^{(x)} - m_{0,1}' = r_{1,1}' \sqrt{\frac{\mu_{0,2}'}{\mu_{2,0}'}}\, [x_i - m_{1,0}'].$$

Коэффициент при $[x_i - m_{1,0}']$ в уравнении прямой, наилучше передающей эмпирическую ломаную регрессии Y по X, будем называть: *эмпирический коэффициент регрессии* Y по X. Обозначая его по аналогии с априорным коэффициентом регрессии (см. глава четвертая, § III, 2), B)) через b_{1}', имеем согласно определению,

$$b_{1}' = r_{1,1}' \sqrt{\frac{\mu_{0,2}'}{\mu_{2,0}'}}.$$

Сходным образом находим для уравнения прямой, наилучше передающей ломаную регрессии X по Y

$$M_{1}^{(y)} - m_{1,0}' = r_{1,1}' \sqrt{\frac{\mu_{2,0}'}{\mu_{0,2}'}}\, [y_i - m_{0,1}']$$

и для эмпирического коэффициента регрессии X по Y

$$b_{1}' = r_{1,1}' \sqrt{\frac{\mu_{2,0}'}{\mu_{0,2}'}}.$$

Эмпирический коэффициент корреляции равняется, следовательно, геометрической средней из эмпирических коэффициентов регрессии Y по X и X по Y:

$$r_{1,1}' = \pm \sqrt{b_{1}' b_{1}'}.$$

§ IV.

Положим (ср. глава четвертая, § IV))

$$[\eta_{1/X}']^2 = 1 - \frac{1}{\mu_{0/2}'} \sum_i p_{i/}' \mu_{/2}^{(i)'}$$

и будем называть величину $\eta_{1/X}'$ *эмпирическое корреляционное отношение* Y к X.

Ни одна из величин $\mu_{/2}^{(i)'}$ не может быть отрицательною; не может быть отрицательною и взвешенная средняя их

$$\sum_i p_{i/}' \mu_{/2}^{(i)'}.$$

Отсюда следует, что $[\eta_{1/X}']^2$ не может быть больше 1.

Замечая, что (см. выше § II, 3))

$$\sum_i p_{i/}' \mu_{/2}^{(i)'} = \mu_{0/2}' - \sum_i p_{i/}' [m_{/1}^{(i)'} - m_{0/1}']^2,$$

преобразуем соотношение, служащее для определения понятия эмпирического корреляционного отношения, в тожество

$$[\eta_{1/X}']^2 = \frac{1}{\mu_{0/2}'} \sum_i p_{i/}' [m_{/1}^{(i)'} - m_{0/1}']^2.$$

Всматриваясь в это тожество, убеждаемся, что $[\eta_{1/X}']^2$, представляя сумму квадратов, не может быть меньше нуля. Таким образом, эмпирическое корреляционное отношение лежит между 0 и 1.

Эмпирическое корреляционное отношение Y к X принимает значение 1, если все величины $\mu_{/2}^{(i)'}$ обращаются в нуль; необходимым и достаточным для этого условием представляется, чтоб все значения Y, соответствующие каждому данному значению X, были равны между собою. Допустим, что число пар сопряженных значений X и Y равно числу наблюденных различных значений X. Тогда каждому значению X соответствует одно лишь значение Y; все величины $\mu_{/2}^{(i)'}$ оказываются равными нулю, и эмпирическое корреляционное отношение Y к X тожественно обращается в 1. При таком виде материала равенство эмпирического корреляционного отношения единице ничего не говорит нам о стохастической связи между Y и X.

Эмпирическое корреляционное отношение принимает значение 0, если

$$\mu_{0/2}' = \sum_i p_{i/}' \mu_{/2}^{(i)'}$$

и

$$\sum_i p_{i/} [m_{/1}^{(i)'} - m_{0/1}']^2 = 0,$$

что возможно лишь в том случае, если все величины $m_1^{(i)'}$ случайно в точности равны друг другу. Следует памятовать, что равенство всех величин $m_1^{(i)'}$ друг другу не является еще доказательством, что величины $m_1^{(i)}$ также равны между собою и что переменная Y некоррелирована с X (ср. глава четвертая, § IV, 2)): равенство величин $m_1^{(i)'}$ друг другу может получиться и в результате случайных отклонений значений $m_1^{(i)'}$ от соответственных значений $m_1^{(i)}$ при больших или меньших расхождениях между величинами $m_1^{(i)}$.

Замечая что (см. выше § III)

$$\sum_i p_i' [m_1^{(i)'} - M_1^{(i)'}]^2 = \mu_{02}' [\eta_{1|x}']^2 - \mu_{02}' [r_{11}']^2,$$

находим:

$$[\eta_{1|x}']^2 - [r_{11}']^2 = \frac{1}{\mu_{02}'} \sum_i p_i' [m_1^{(i)'} - M_1^{(i)'}]^2$$

и убеждаемся, что эмпирический коэффициент корреляции по своей абсолютной величине не может превосходить эмпирического корреляционного отношения. Он достигает размеров эмпирического корреляционного отношения лишь в том случае, если величины $m_1^{(i)'}$ располагаются в точности на прямой линии, уравнение которой имеет вид

$$M_1^{(i)} - m_{01}' = r_{11}' \sqrt{\frac{\mu_{02}'}{\mu_{20}'}} [x_i - m_{10}'].$$

Следует отметить и тут, что расположение всех точек эмпирической ломаной регрессии Y по X в точности на некоторой прямой не является еще доказательством прямолинейности регрессии Y по X. Не исключается, что истинное уравнение регрессии Y по X, связывающее величины $m_1^{(i)}$ с значениями X, имеет и иной вид, а лишь случайные отклонения величин $m_1^{(i)'}$ от соответственных величин $m_1^{(i)}$ приводят к видимости прямолинейной регрессии. Равенство разности $\eta_{Y|X}^2 - r_{11}^2$ нулю представляется доказательством, что регрессия Y по X прямолинейна. Равенство разности $[\eta_{1|x}']^2 - [r_{11}']^2$ нулю делает лишь предположение, что истинная линия регрессии Y по X имеет вид прямой, более или менее правдоподобным.

§ V.

Эмпирический коэффициент корреляции, определяемый соотношением,

$$r_{11}' = \frac{\mu_{11}'}{\sqrt{\mu_{20}' \mu_{02}'}},$$

не может быть по абсолютной величине больше эмпирического корреляционного отношения Y к X или эмпирического корреляционного отношения X к Y

(см. выше § IV). А так как эмпирическое корреляционное отношение не может в свою очередь превышать 1, то численное значение эмпирического коэффициента корреляции заключено в границы $+1$ и -1.

К доказательству, что эмпирический коэффициент корреляции лежит в границах между $+1$ и -1, можно итти, и не привлекая эмпирических корреляционных отношений. Замечая, что

$$\sum_i \sum_j p_{ij}' \left[\frac{x_i - m_{10}'}{\sqrt{\mu_{20}'}} - \frac{y_j - m_{01}'}{\sqrt{\mu_{02}'}} \right]^2,$$

как сумма квадратов, не может быть величиною отрицательною, находим:

$$\sum_i \sum_j p_{ij}' \left[\frac{x_i - m_{10}'}{\sqrt{\mu_{20}'}} - \frac{y_j - m_{01}'}{\sqrt{\mu_{02}'}} \right]^2 =$$

$$= \sum_i \sum_j p_{ij}' \left\{ \frac{[x_i - m_{10}']^2}{\mu_{20}'} - \frac{2[x_i - m_{10}'][y_i - m_{01}']}{\sqrt{\mu_{20}'}\sqrt{\mu_{02}'}} + \frac{[y_i - m_{01}']^2}{\mu_{02}'} \right\} =$$

$$= 1 - 2r_{11}' + 1 \geqslant 0$$

и отсюда $r_{11}' \leqslant 1$

С другой стороны,

$$\sum_i \sum_j p_{ij}' \left[\frac{x_i - m_{10}'}{\sqrt{\mu_{20}'}} + \frac{y_j - m_{01}'}{\sqrt{\mu_{02}'}} \right]^2 = 1 + 2r_{11}' + 1 \geqslant 0$$

и отсюда $r_{11}' \geqslant -1$.

Всматриваясь в соотношение, приводящее к заключению, что $r_{11}' \leqslant 1$, убеждаемся, что эмпирический коэффициент корреляции равняется точно $+1$ лишь, если все без исключения разности

$$\frac{x_i - m_{10}'}{\sqrt{\mu_{20}'}} - \frac{y_j - m_{01}'}{\sqrt{\mu_{02}'}}$$

равны нулю, т.-е. если каждому значению X отвечает одно лишь значение Y и притом все сопряженные отклонения X и Y от их средних арифметических строго пропорциональны друг другу и имеют один и тот же знак. Сходным образом убеждаемся, что r_{11}' может в точности равняться -1 лишь, если каждому значению X отвечает одно лишь значение Y и притом все сопряженные отклонения переменных X и Y от их средних арифметических строго пропорциональны друг другу, но имеют противоположные знаки. Равенство эмпирического коэффициента корреляции ± 1 свидельствует, таким образом, о том, что установленные наблюдением значения Y и X находятся в линейной функциональной зависимости друг от друга. Из этого опять-таки не вытекает с несомненностью, что и переменные X и Y стоят в линейной функциональной зависимости. Не вытекает

даже, что регрессия Y по X прямолинейна. Не исключено, что наблюдаемая видимость линейной функциональной зависимости обязана своим происхождением случайным отклонениям эмпирических значений Y от соответственных величин $m_1^{(i)}$.

Представляется существенным отметить, что эмпирический коэффициент корреляции способен принимать неопределенные значения. Если все значения X или все значения Y случайно оказываются равными друг другу, то как μ_{11}', так и одно из стоящих в знаменателе r_{11}' эмпирических рассеяний — μ_{20}' или μ_{02}' — обращаются в нуль, и для эмпирического коэффициента корреляции мы находим:

$$r_{11}' = \frac{0}{0}.$$

При небольших числах наблюдений неопределенные значения эмпирического коэффициента корреляции могут встречаться на практике нередко.

§ VI.

Положим

$$[\varphi']^2 = \sum_i \sum_j \frac{[p_{ij}' - p_i' p_j']^2}{p_i' p_j'} = \sum_i \sum_j \frac{\left[n_{ij} - \frac{1}{N} n_i' n_j\right]^2}{n_i' n_j}.$$

Величину φ' будем называть: *эмпирический показатель взаимной сопряженности переменных* (см. глава четвертая § II, 1)).

Замечая, что

$$\sum_i \sum_j n_{ij} = N, \quad \sum_i \sum_j n_i n_j = \left\{\sum_i n_i\right\}\left\{\sum_j n_j\right\} = N^2.$$

находим:

$$[\varphi']^2 = \sum_i \sum_j \frac{n_{ij}^2}{n_i' n_j'} - 1.$$

Эмпирический показатель взаимной сопряженности переменных принимает значение 0, если все разности $p_{ij}' - p_i' p_j'$ равны нулю. Из этого не следует, однако, с несомненностью, что все разности $p_{ij} - p_i p_j$ также равны нулю и что переменные X и Y взаимно независимы. Видимость взаимной независимости может и здесь сводиться к игре случайностей. Если все значения Y, соответствующие одному какому-либо значению X, равны между собою, то $p_{ij}' = 0$ при $j \neq i$, а $p_{ii}' = p_i'$. Для $[\varphi']^2$ находим, подставляя эти значения, величину $k - 1$. Такое значение эмпирического показателя взаимной сопряженности сообщает известное правдоподобие предположению, что Y находится в функциональной зависимости от X. Но и здесь необходимы те же оговорки, как в случае эмпирического корреляционного отношения (см. выше § IV).

§ VII.

Предположим, что обе переменные X и Y принимают лишь по два разных значения и положим (см. выше глава четвертая § II, 1))

$$p_{1\,1}'\,p_{2\,2}' - p_{1\,2}'\,p_{2\,1}' = \delta'.$$

Исходя из тожеств, связывающих величины

$$p_{i\,j}',\ p_{i\,.}',\ p_{.\,j}',$$

находим без труда

$$\delta' = p_{1\,1}' - p_{1\,.}'\,p_{.\,1}' = p_{2\,2}' - p_{2\,.}'\,p_{.\,2}' = -[p_{1\,2}' - p_{1\,.}'\,p_{.\,2}'] = -[p_{2\,1}' - p_{2\,.}'\,p_{.\,1}'].$$

Подставляя эти значения разностей $p_{1\,1}' - p_{1\,.}'\,p_{.\,1}'$ и пр. в формулу, определяющую понятие эмпирического показателя взаимной сопряженности (см. выше § VI), находим

$$[\varphi']^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{[\delta']^2}{p_{i\,.}'\,p_{.\,j}'} = \frac{|\delta'|^2}{p_{1\,.}'\,p_{2\,.}'\,p_{.\,1}'\,p_{.\,2}'}.$$

К той же величине

$$\frac{|\delta'|^2}{p_{1\,.}'\,p_{2\,.}'\,p_{.\,1}'\,p_{.\,2}'}$$

сводятся в случае, если обе переменные получают только по два разных значения, квадрат эмпирического коэффициента корреляции и эмпирические корреляционные отношения.

ГЛАВА ШЕСТАЯ.

Оценка априорных искомых по эмпирическим данным.

§ 1.

Мы познакомились (глава третья и глава четвертая) с основными чертами тех приемов, которые теория статистики рекомендует для сжатой характеристики стохастической связи между двумя случайными переменными. Мы рассмотрели (глава пятая) те формы, в какие исследователь облекает имеющиеся в его распоряжении эмпирические данные. Мы видели, что́ статистик должен стремиться установить при исследовании стохастической связи; мы видели, что́ ему известно, — из чего он обычно может исходить. Как правило, он не располагает знанием закона связи, которое позволило бы вычислить нужные ему значения сводных характеристик связи по определяющим их формулам без обращения к данным опыта. Обычно в его распоряжении имеется лишь запись случайных сопряженных значений, полученных переменными при ряде производившихся над ними „испытаний“. Возникает, таким образом, вопрос, какими путями может исследователь от имеющихся данных — от этих случайных значений переменных — перейти к искомым — к численным значениям тех сводных характеристик закона связи, которые в сжатой и удобной для использования форме выражают важные для исследования свойства взаимоотношений между изучаемыми переменными.

Задача перехода от эмпирических данных к априорным искомым научно точно была впервые поставлена Яковом Бернулли. Ее решение покоится на законе больших чисел, перекидывающем мост от эмпирических данных статистического учета к лежащим в их основе априорным величинам, недоступным, как правило, для непосредственного измерения. Основная мысль закона больших чисел может облекаться в весьма разнообразные формы. При наших дальнейших рассуждениях всего удобнее исходить из формулировки, которая всего проще может быть обоснована на знаменитом неравенстве Чебышева: предположим, что над случайною переменною производится N взаимно независимых испытаний; вероятность, что разность между среднею арифметическою случайных значений переменной при отдельных испытаниях и математическим ожиданием переменной не превзойдет данной, произвольно малой величины возрастает с ростом N и при

достаточно большом числе испытаний может быть сделана сколь угодно близкою к 1; можно, следовательно, при большом числе испытаний с весьма высокою степенью уверенности принимать, что численные значения эмпирической средней и лежащего в ее основе математического ожидания переменной близки друг к другу. Я не воспроизвожу математического вывода теоремы Бернулли и неравенства Чебышева, — их можно найти в каждом руководстве теории вероятностей. Не буду останавливаться и на сложных логических проблемах, поднимаемых законом больших чисел, дабы не отвлекаться от нашей ближайшей задачи — от использования приведенного выше положения для построения приемов оценки сводных характеристик законов связи по эмпирическим данным, имеющимся в распоряжении исследователя. Логический анализ закона больших чисел представляет, бесспорно, первостепенную важность для осмысления основ теории корреляции. Но к теории корреляции, как таковой, он не имеет ближайшего касательства: это — вопрос общей теории статистики, и загромождать его разбором изложение начал теории корреляции не представляется целесообразным.

§ II.

1) К оценке априорных величин по случайным значениям переменных можно итти разными путями.

Наиболее естественным представляется исходить при оценке численного значения априорной величины U от такой функции случайных значений переменной, математическое ожидание которой в точности равняется U. Численное значение, получаемое этой функцией — обозначим ее через U' — при подстановке найденных при опыте данных, может быть рассматриваемо как предположительное значение U. В каждом отдельном случае численное значение U' может уклоняться более или менее от интересующего нас истинного значения U, при чем предположительный размах таких случайных расхождений намечается рассеянием U'. Но в среднем эти случайные погрешности перекрывают друг друга, и общий уровень наших оценок будет всего ближе к истине, если мы выдержанно будем придерживаться такого порядка составления суждений об интересующих нас априорных величинах. Предположим, что из закрытой урны производится N извлечений, при чем белый шар появляется n раз; предположительное значение неизвестной вероятности вынуть белый шар из этой урны — p — может быть приравнено отношению $\frac{n}{N}$ как в том случае, если вынутый шар возвращается в урну, прежде чем вынимается следующий, — случай, который имеется в виду теоремою Бернулли, — так и при условии, что вынутые шары не возвращаются в урну до конца опыта. Ибо нетрудно доказать, что и в том, и в другом предположении математическое ожидание $\frac{n}{N}$ равняется p при любом числе извлечений. Повторим эксперимент извлечения N шаров из данной урны t раз, возвращая все время вынутый шар в урну, прежде чем извлекается следующий, или откладывая все время вынутые шары, пока не будет произведено N извлечений, а затем возвращая их все в урну, прежде чем приступить к следующей

серии N извлечений. Обозначим через n_1 число белых шаров, вынувшихся при первой серии N извлечений; через n_2 — число белых шаров, вынувшихся при второй серии извлечений, и т. д.; через n_t — число белых шаров, вынувшихся при последней серии извлечений. Согласно формулированному выше положению, разность между

$$\frac{1}{t}\left(\frac{n_1}{N}+\frac{n_2}{N}+\cdots+\frac{n_t}{N}\right)$$

и искомой вероятностью p будет при достаточно большом числе серий t весьма мало отлична от нуля. Принимая за предположительное значение p отношение числа вынувшихся белых шаров ко всему числу извлечений — $\frac{n}{N}$, — мы в каждом отдельном случае рискуем сделать ошибку, — можем преувеличить, можем и преуменьшить величину p. Но в конечном счете наши оценки неизвестных вероятностей по установленным наблюдением частостям будут при таком способе подхода наилучше попадать в цель.

Чтоб составить представление о том, как велик риск ошибки, с которою сопряжены единичные оценки, необходимо вычислить среднее квадратическое отклонение частости $\frac{n}{N}$. Если вынутый шар бросается назад в урну, прежде чем производится следующее извлечение, то, как известно из теории вероятностей,

$$\sigma_{\frac{n}{N}}=\sqrt{E\left[\frac{n}{N}-p\right]^2}=\sqrt{\frac{p(1-p)}{N}}.$$

Если шары в урну не возвращаются, то

$$\sigma_{\frac{n}{N}}=\sqrt{\frac{A-N}{A-1}\,\frac{p(1-p)}{N}},$$

при чем A обозначает все число шаров в урне. При такой постановке, когда вынутые шары в урну не возвращаются, оценки оказываются, таким образом, более надежными, и степень надежности их возрастает с ростом N быстрее, чем в том случае, когда вынутый шар бросается обратно в урну до производства следующего извлечения. При N хотя бы и малом, но равном A, оценка становится тут вполне надежною: вынув все шары из урны, мы по доле белых среди них, разумеется, можем с полной уверенностью судить о вероятности p.

2) На-ряду с этим подходом к оценке численных значений априорных величин по данным опыта приходится нередко прибегать и к более грубому способу. По отношению к функции U', численное значение которой приравнивается искомому значению априорной величины U, удовлетворяются требованием, чтоб математическое ожидание U' стремилось к U с ростом числа наблюдений, не выдвигая более строгого требования, чтобы математическое ожидание U' в точности равнялось U при всяком конечном числе испытаний. Такой способ оценки

опирается на соображение, что при сколько-нибудь большом числе испытаний погрешность не может быть велика, так как при бесконечно большом числе испытаний математическое ожидание U' совпадает с U: этот более грубый способ как бы нащупывает приближенно предположительное значение U, и приближение тем выше, чем больше число испытаний.

Обычно при таком подходе U' — функция случайных значений переменной, от которой отправляется оценка, — получается путем замены математических вероятностей статистическими частостями в формуле, определяющей U через посредство априорного закона распределения или связи. При этом исходят из мысли, что при большом числе испытаний разность между функцией математических вероятностей и тою же функцией статистических частостей не может быть велика, так как мало расходятся входящие в формулы частости и вероятности. Так, например, случайное значение эмпирического показателя взаимной сопряженности (см. глава пятая, § VI)

$$\varphi'^2 = \sum_i \sum_j \frac{[p_{ij}' - p_i' p_j']^2}{p_i' p_j'}$$

принимается за предположительное значение априорного показателя взаимной сопряженности (см. глава четвертая, § II, 1))

$$\varphi^2 = \sum_i \sum_j \frac{[p_{ij} - p_i p_j]^2}{p_i p_j}.$$

Сходным образом, случайное значение эмпирического коэффициента корреляции $r_{1/1}'$ принимается за предположительное значение априорного коэффициента корреляции $r_{1/1}$, случайное значение эмпирического корреляционного отношения $[\eta_{1/x}']^2$ принимается за предположительное значение априорного корреляционного отношения $\eta_{1/x}{}^2$ и т. д.

Невыгода подобного подхода заключается в том, что мы уже не вправе рассчитывать при повторных оценках на полное совпадение средней арифметической полученных таким путем предположительных значений с истинным значением искомой априорной величины. Сверх неизбывной случайной погрешности, с которой остается сопряженным и наиболее совершенный способ оценки, тут привходит еще систематическая погрешность, — бо́льшая или меньшая, положительная или отрицательная, смотря по обстоятельствам. Пусть, например, требуется оценить значение квадрата вероятности вынуть белый шар из урны по данным произведенных из урны N извлечений в предположении, что извлекаемый шар возвращается обратно в урну до производства следующего извлечения; приравнивая предположительное значение p^2 второй степени отношения числа вынувшихся белых шаров ко всему числу извлечений — $\left(\frac{n}{N}\right)^2$, — мы переоцениваем искомую величину, так как математическое ожидание $\left(\frac{n}{N}\right)^2$ равняется

не p^2, а $p^2 + \frac{1}{N} p(1-p)$. Пусть, с другой стороны, мы желаем составить по данным опыта представление о размерах произведения $p(1-p)$: приравнивая его предположительное значение произведению

$$\frac{n}{N}\left(1-\frac{n}{N}\right),$$

мы недооцениваем искомую величину, так как

$$E\frac{n}{N}\left(1-\frac{n}{N}\right) = E\frac{n}{N} - E\left(\frac{n}{N}\right)^2 = \frac{N-1}{N} p(1-p).$$

Но и в том, и в другом случае систематическая погрешность оценки тем меньше, чем больше число испытаний N; при N бесконечно большом систематическая погрешность сводится к нулю, ибо

$$\lim_{N\to\infty}\left[p^2 + \frac{1}{N} p(1-p)\right] = p^2, \quad \lim_{N\to\infty}\left[\frac{N-1}{N} p(1-p)\right] = p(1-p);$$

при конечном же, но достаточно большом N систематическая погрешность и тут, и там весьма невелика. Что и в других случаях применения этого метода оценки априорных величин по данным опыта систематическая погрешность, как правило, стремится к нулю с ростом числа испытаний, было доказано в общей форме Г. Больманом[1]), которому удалось подвести научный фундамент под способы работы, применявшиеся статистиками дотоле по большей части без должной критики. Остается все же желательным при использовании этого приема располагать возможностью составить достаточно точное представление о порядке величины систематической погрешности при небесконечном числе испытаний и, в особенности, о ее знаке. В этом направлении ведется ныне энергичная работа. Первый шаг на пути оценки интересующих статистика априорных величин по данным опыта состоял почти неизменно в замене математических вероятностей статистическими частостями по указанным выше образцам. Лишь позднее начали принимать во внимание, что это может быть сопряжено с систематическою погрешностью, и стали предпринимать попытки прикинуть ее размеры. Результаты исследования систематической погрешности получились в разных случаях различные. Иногда удавалось обнаружить, что систематической погрешности при таком способе оценки и не делается, так как математическое ожидание построенной по этому шаблону функции частостей в точности равняется искомой функции вероятностей не только при бесконечно большом, но и при любом конечном числе испытаний. В некоторых случаях удалось установить точно размеры систематической погрешности и устранить ее затем при посредстве целесообразной

1) G. Bohlmann. Formulierung und Begrundung zweier Hilfssatze der mathematischen Statistik (Math. Annalen. Bd 74; 1913)

перестройки функции U': так, например, если мы при оценке численного значения произведения $p(1-p)$ будем исходить не из величины произведения

$$\frac{n}{N}\left(1-\frac{n}{N}\right),$$

а из величины

$$\frac{N}{N-1}\,\frac{n}{N}\left(1-\frac{n}{N}\right)=\frac{n(N-n)}{N(N-1)},$$

то систематическая погрешность окажется устраненною, так как математическое ожидание

$$\frac{n(N-n)}{N(N-1)}$$

в точности равняется $p(1-p)$ при всяком N. В очень многих случаях оказалось, однако, невозможным найти точные размеры систематической погрешности, и приходится удовлетворяться примерною прикидкою ее величины. Если функция U' не имеет вида целой рациональной функции значений X и Y, то, как правило, не удается пойти далее приближенного вычисления ее математического ожидания, а вместе с тем и систематической погрешности оценки величины U по случайному значению U'. А так как основные сводные характеристики закона связи — корреляционное отношение, коэффициент корреляции, показатель взаимной сопряженности и др. — имеют вид частного, то в области теории корреляции мы в большинстве случаев оказываемся вынужденными иметь дело с такими приближенными вычислениями систематических погрешностей, как равно, надо заметить, и средних квадратических ошибок.

§ III.

1) На основе общих начал, которыми статистик руководствуется при оценке априорных величин по данным опыта, теория корреляции должна выработать приемы оценки тех сводных характеристик априорных законов связи, с которыми мы познакомились в четвертой главе. Не задаваясь и здесь целью достичь исчерпывающей полноты обзора, я попытаюсь свести возникающие при этом задачи к известной системе, наметить на ряде примеров основные линии решений, указать главные трудности, на которые наталкивается исследование, и разобрать важнейшие приемы, при помощи которых эти трудности преодолеваются. При наших рассуждениях мы будем исходить из предположения, что закон связи между переменными остается неизменным при всех производимых над ними испытаниях, а отдельные испытания все взаимно независимы, так что вероятность сочетания значения x_i переменной X и значения y_j переменной Y при любом из испытаний остается одною и тою же, какие бы значения ни принимали переменные X и Y при прочих испытаниях.

2) Начнем с оценки априорной величины δ, к которой примыкают, как мы видели (см. глава четвертая, §§ II, VII) сводные характеристики связи в случае, если обе переменные способны принимать лишь по два разных значения.

Исходя из соотношения

$$\delta = p_{1/1}\,p_{2/2} - p_{1/2}\,p_{2/1},$$

как определения априорной величины δ, заменим здесь математические вероятности соответственными статистическими частостями и обозначим получаемую при этом функцию частостей через δ' (см. глава пятая, § VII):

$$\delta' = p_{1/1}'\,p_{2/2}' - p_{1/2}'\,p_{2/1}' = \frac{1}{N^2}\,[n_{1/1}\,n_{2/2} - n_{1/2}\,n_{2/1}].$$

Требуется установить, в какой мере численное значение δ' пригодно для подхода к оценке значения δ, и выяснить, нельзя ли предложить более совершенный способ оценки.

Замечая, что

$$E n_{1/1}\,n_{2/2} = N(N-1)\,p_{1/1}\,p_{2/2}, \quad E n_{1/2}\,n_{2/1} = N(N-1)\,p_{1/2}\,p_{2/1},$$

находим:

$$E\delta' = \frac{N-1}{N}\,\delta.$$

Принимая численное значение δ' за предположительное значение δ, мы систематически недооцениваем величину δ. Систематическая погрешность — величина порядка $1/N$; при сколько-нибудь значительном числе испытаний она не велика, и ею можно без особого ущерба пренебречь. Нетрудно, однако, ее и устранить: достаточно для этого приравнивать предположительное значение δ не численному значению δ', а численному значению $\frac{N}{N-1}\,\delta'$, т.-е. увеличивать δ' на $1/N-1$ ее величины; ибо $E\,\frac{N}{N-1}\,\delta'$ в точности равняется δ при любом числе испытаний.

Чтоб составить представление о степени надежности, с которою можно судить о величине δ по выхваченному на удачу случайному значению δ', необходимо найти среднюю квадратическую ошибку $\frac{N}{N-1}\,\delta'$. Путем несложных алгебраических выкладок получаем:

$$\sigma^2_{\frac{N}{N-1}\delta'} = E\left[\frac{N}{N-1}\,\delta' - \delta\right]^2 =$$

$$= \frac{1}{N-1}\left\{\left[p_{1/1}\,p_{2/2}\,(p_{1/1}+p_{2/2}) + p_{1/2}\,p_{2/1}\,(p_{1/2}+p_{2/1}) - 4\delta^2\right] + \right.$$

$$\left. + \frac{1}{N}\,[p_{1/2}+p_{2/1}-p_{1/1}-p_{2/2}+6\delta]\,\delta\right\}.$$

Если априорная величина δ равна нулю и переменные взаимно независимы, то нулю равняется и математическое ожидание $\frac{N}{N-1}\delta'$; рассеяние же $\frac{N}{N-1}\delta'$ сводится при этом к

$$\frac{1}{N-1} p_1' p_2 p_1' p_2,$$

так как, при

$$\delta = 0, \quad p_{1,1} = p_1' p_1', \quad p_{1'2} = p_{1'} p_{'2} \text{ и т. д.}$$

Предположим, что $\frac{N}{N-1}\delta'$ равняется нулю. Из этого не следует еще с несомненностью, чтоб и априорная величина δ точно равнялась нулю и переменные были взаимно независимы. Не исключается, что δ' случайно принимает значение 0, хотя переменные и не представляются стохастически независимыми, и величина δ более или менее отличается от нуля. Равным образом и из того, что $\frac{N}{N-1}\delta'$ отлична от нуля, не следует еще с несомненностью, что отлична от нуля и априорная величина δ и переменные не представляются взаимно независимыми. И тут может быть замешана игра случайностей. Степень правдоподобия вывода в обоих случаях определяется величиною средней квадратической ошибки $\frac{N}{N-1}\delta'$ и возрастает, как явствует из приведенной выше формулы, с ростом числа испытаний пропорционально корню квадратному из N.

Для того, чтоб составить более полное представление о законе распределения значений $\frac{N}{N-1}\delta'$, надо вычислить математические ожидания высших степеней $\frac{N}{N-1}\delta'$ или разности $\frac{N}{N-1}\delta' - \delta$. Выкладки становятся все более сложными по мере повышения степени, но непреодолимых трудностей не представляют и могут быть выполнены в том же порядке, как вычисление математического ожидания и средней квадратической ошибки $\frac{N}{N-1}\delta'$. Не выписывая громоздких формул, отмечу лишь, что при конечном N распределение значений $\frac{N}{N-1}\delta'$ не отвечает закону Гаусса-Лапласа, — в частности не симметрично, — но что с ростом числа испытаний вид распределения приближается к тому, какой соответствует закону Гаусса-Лапласа, и при достаточно большом N может быть без особого риска уподоблен ему.

3) Проблема оценки априорного значения δ по данным опыта решается сравнительно легко. На трудности совсем иного порядка наталкиваемся мы при попытке выработать приемы оценки величины

$$\frac{\delta}{\sqrt{p_1' p_2' p_1' p_2'}} = r_{1'1},$$

выступающей, как мы видели (см. глава четвертая, § VII) в качестве коэффициента корреляции между переменными в случае, если обе переменные способны

принимать лишь по два разных значения. Если мы произведем обычную замену вероятностей частостями в выражении для r_{11}, то для получаемой функции частостей (см. глава пятая, § VII)

$$r_{1/1}' = \frac{\delta'}{\sqrt{p_1' p_2' p_1' p_2'}}$$

при современном состоянии разработки вопроса оказывается невозможным указать точно математическое ожидание. Как упомянуто выше (см. § II, 2)), установить точно математическое ожидание функции, имеющей вид частного, удается вообще лишь в очень редких случаях. При оценке значения

$$\frac{\delta'}{\sqrt{p_1' p_2' p_1' p_2'}}$$

по данным опыта мы оказываемся вынуждены довольствоваться приближенными результатами, на которых я остановлюсь ниже (см. § IV, 3), A) и § IV, 5)) при рассмотрении общей проблемы оценки априорного коэффициента корреляции между переменными по численной величине эмпирического коэффициента корреляции между их случайными значениями.

§ IV.

1) Трудности вычисления математических ожиданий функций, имеющих вид частного, занимают в теории приемов оценки сводных характеристик законов связи по данным опыта столь видное место, что представляется необходимым уделить особое внимание этой проблеме.

Найти точное значение математического ожидания частного двух переменных, стоящих в стохастической связи, удалось доселе лишь в очень немногих случаях. Первым примером было, повидимому, вычисление математического ожидания Лексисовского коэффициента дисперсии. В настоящее время можно указать несколько примеров и из области теории корреляции.

A) Согласно определению понятия математического ожидания

$$E\frac{X}{Y} = \sum_i \sum_j p_{ij} \frac{x_i}{y_j}.$$

Если закон связи между X и Y предполагается известным, то вероятности p_{ij} могут почитаться за данные, и задача сводится к тому, чтоб, подставив в формулу $E\frac{X}{Y}$ значение p_{ij}, привести двойную сумму к более удобообозримому виду, выполнив указываемые ею суммирования. Пусть, например, требуется найти математическое ожидание числа случаев, когда одновременно переменная X принимает значение x_i, а переменная Y принимает значение y_j, ко всему числу случаев, когда переменная X принимает значение x_i. т.-е. в наших обозначениях

$$E\frac{n_{ij}}{n_i} \quad \text{или} \quad E\frac{p_{ij}'}{p_{i'}}.$$

Вероятность, что n_{ij} примет значение h, равняется, как известно из теории вероятностей,

$$\binom{N}{h} p_{ij}^{\,h} (1 - p_{ij})^{N-h}.$$

Условное математическое ожидание n_{ij} в предположении, что n_{ij} равняется h, должно быть положено равным $h \frac{p_{ij}}{p_{i'}}$, ибо h испытаний, с которыми мы имеем дело в этом предположении, распределяются между l группами — $n_{i1}, n_{i2}, \cdots, n_{il}$ —, при чем вероятности попасть в каждую из этих групп — $p_{i1}, p_{i2}, \cdots, p_{il}$ — дают в сумме $p_{i'}$. Математическое ожидание частного равняется, следовательно,

$$E \frac{n_{ij}}{n_{i'}} = \sum_h \binom{N}{h} p_{ij}^{\,h} (1 - p_{ij})^{N-h} \frac{1}{h} h \frac{p_{ij}}{p_{ij}} =$$

$$= \frac{p_{ij}}{p_{i'}} \sum_h \binom{N}{h} p_{ij}^{\,h} (1 - p_{ij})^{N-h} = \frac{p_{ij}}{p_{i'}}.$$

В этом случае выкладки не представляют трудности вследствие того, что h в числителе и h в знаменателе стоящего за знаком суммы выражения могут быть сокращены. Но предположим, что требуется найти математическое ожидание $\left[\frac{n_{ij}}{n_{i'}}\right]^2$. Замечая, что условное математическое ожидание $\left[\frac{n_{ij}}{n_{i'}}\right]^2$ в предположении, что n_{ij} принимает значение h, получается равным

$$h^2 \left[\frac{p_{ij}}{p_{ij}}\right]^2 + h \frac{p_{ij}}{p_{ij}} \left[1 - \frac{p_{ij}}{p_{ij}}\right],$$

находим, выполняя выкладки в том же порядке,

$$E\left[\frac{n_{ij}}{n_{i'}}\right]^2 = \left[\frac{p_{ij}}{p_{i'}}\right]^2 + \frac{p_{ij}}{p_{ij}} \left[1 - \frac{p_{ij}}{p_{ij}}\right] \sum_h \binom{N}{h} p_{ij}^{\,h} (1 - p_{ij})^{N-h} \frac{1}{h}.$$

Мы попадаем, таким образом, в тупик, так как не представляется возможным выполнить суммировку

$$\sum_h \binom{N}{h} p_{ij}^{\,h} (1 - p_{ij})^{N-h} \frac{1}{h}.$$

С препятствием подобного рода и встречается в большинстве случаев вычисление математического ожидания частного. Иногда удается свести двойную сумму к простой, но не представляется возможным выполнить второе суммирование. Такой результат является известным успехом, хотя и не дает полного решения задачи. Но если выкладки ограничиваются подстановкою значения p_{ij} без дальнейшего преобразования двойной суммы, то о решении поставленной

задачи не может быть, конечно, и речи, хотя бы вывод общей формулы для вероятности p_{ij} и представлялся сам по себе ценным достижением.

B) В некоторых случаях к цели, недостижимой прямым путем, удается притти путем обходным. Следующим образом можно, например, доказать, что математическое ожидание эмпирического коэффициента корреляции в случае взаимной независимости переменных в точности равняется нулю, а рассеяние его представляется в точности равным $1/(N-1)$.

Исходя из определения эмпирического коэффициента корреляции (см. глава пятая, § II, 2))

$$r_{11}' = \frac{\sum_{f=1}^{N} [x^{[f]'} - x_0'] [y^{[f]'} - y_0']}{\sqrt{\left\{\sum_{f=1}^{N} [x^{[f]'} - x_0']^2\right\} \left\{\sum_{f=1}^{N} [y^{[f]'} - y_0']^2\right\}}}.$$

введем для облегчения начертания формул сокращенные обозначения:

$$\sqrt{\left\{\sum_{f=1}^{N} [x^{[f]'} - x_0']^2\right\}} = \Sigma_1, \qquad \sqrt{\left\{\sum_{f=1}^{N} [y^{[f]'} - y_0']^2\right\}} = \Sigma_2.$$

В предположении, что закон связи остается неизменным при всех испытаниях, а испытания взаимно независимы, находим

$$E\frac{x^{[f]'}}{\Sigma_1} = E\frac{x^{[\alpha]'}}{\Sigma_1} = E\frac{x_0'}{\Sigma_1}$$

и отсюда

$$E\frac{x^{[f]'} - x_0'}{\Sigma_1} = 0.$$

Сходным образом получаем

$$E\frac{y^{[f]'} - y_0'}{\Sigma_2} = 0.$$

Если переменные X и Y стохастически независимы, то

$$E\frac{[x^{[f]'} - x_0'] [y^{[f]'} - y_0']}{\Sigma_1 \Sigma_2} = \left\{E\frac{x^{[f]'} - x_0'}{\Sigma_1}\right\} \left\{E\frac{y^{[f]'} - y_0'}{\Sigma_2}\right\}.$$

Следовательно

$$Er_{1/1}' = E\frac{\sum_f [x^{[f]'} - x_0'] [y^{[f]'} - y_0']}{\Sigma_1 \Sigma_2} = \sum_f E\frac{[x^{[f]'} - x_0'] [y^{[f]'} - y_0']}{\Sigma_1 \Sigma_2} =$$

$$= \sum_f \left\{E\frac{x^{[f]'} - x_0'}{\Sigma_1}\right\} \left\{E\frac{y^{[f]'} - y_0'}{\Sigma_2}\right\} = 0.$$

Таким образом, в случае стохастической независимости переменных X и Y математическое ожидание $r_{1/1}'$ совпадает с истинным значением априорного коэффициента корреляции между переменными, которое, как мы знаем (см. глава четвертая, § III, 3)), также равнятся нулю при взаимной независимости переменных.

Отметим, что приведенный вывод опирается на допущение стохастической независимости в строгом смысле данного выше определения (см. глава третья, § IV, 3)), а не на предположения, что переменные некоррелированы или что $r_{1/1} = 0$, представляющиеся необходимыми, но не достаточными условиями стохастической независимости (см. глава четвертая, § III, 3)). Если $r_{1/1} = 0$, но переменные не представляются взаимно независимыми, то нельзя принимать, что

$$E\frac{[x^{[J]\prime} - x_0'][y^{[J]\prime} - y_0']}{\Sigma_1 \Sigma_2} = \left\{E\frac{x^{[J]\prime} - x_0'}{\Sigma_1}\right\}\left\{E\frac{y^{[J]\prime} - y_0'}{\Sigma_2}\right\}.$$

Как мы увидим ниже (см. § IV, 3), А)), математическое ожидание $r_{1\,1}'$ при $r_{1/1} = 0$ может быть и равно 0, и больше 0, и меньше 0.

Сходным путем может быть точно вычислено и рассеяние $r_{1/1}'$ в предположении, что переменные X и Y стохастически независимы. Замечая, что при сделанных нами допущениях

$$E\frac{[x^{[J]\prime} - x_0']^2}{\Sigma_1^2} = \frac{1}{N}E\frac{\sum_{J=1}^{N}[x^{[J]\prime} - x_0']^2}{\Sigma_1^2} = \frac{1}{N}E\frac{\Sigma_1^2}{\Sigma_1^2} = \frac{1}{N}$$

и

$$E\frac{[y^{[J]\prime} - y_0']^2}{\Sigma_2^2} = \frac{1}{N},$$

находим путем несложных выкладок

$$E[r_{1/1}']^2 = \frac{1}{N-1}.$$

А так как $Er_{1/1}' = 0$, то

$$\sigma^2 r_{1/1}' = E[r_{1/1}' - Er_{1/1}']^2 = E[r_{1/1}']^2 = \frac{1}{N-1}.$$

Полученные нами соотношения

$$Er_{1/1}' = 0, \qquad E[r_{1/1}']^2 = \frac{1}{N-1}$$

представляются при сделанных допущениях — неизменный закон связи, взаимно независимые испытания, стохастически независимые переменные X и Y — строго точными при любом виде законов распределения значений переменных X и Y.

Что касается математических ожиданий высших степеней $r_{1/1}'$, то их нельзя вычислить точно и при сделанных нами допущениях. При вычислении математического ожидания $[r_{1/1}']^4$ мы приходим, например, к выражениям вида

$$E\frac{\sum_J [x^{[J]\prime} - x_0']^4}{\Sigma_1^4} \quad \text{и} \quad E\frac{\sum_J [y^{[J]\prime} - y_0']^4}{\Sigma_2^4}.$$

Если законы распределения значений X и Y остаются неопределенными, суммировки эти не могут быть выполнены. Величина $E[r_{1,1}']^4$ зависит, таким образом, от вида законов распределения значений X и Y в случае, если переменные X и Y стохастически независимы.

2) В тех случаях, когда ни прямым путем, ни обходным не удается добраться до точного значения искомого математического ожидания, не остается иного выхода, как пытаться установить значение $E\frac{X}{Y}$ приближенно. Сама собою напрашивается мысль представить математическое ожидание $E\frac{X}{Y}$ в виде суммы членов, расположенных по восходящим степеням $1/N$, дабы хоть для достаточно больших чисел испытаний иметь возможность составить представление о его величине. Мысль эта может облекаться в разные формы. Всего чаще прибегают к следующему приему, получившему разработку преимущественно в трудах представителей английской школы.

A) Пусть требуется найти математическое ожидание

$$\left[\frac{p_{ij}'}{p_{i'}'}\right]^2 = \left[\frac{n_{ij}}{n_{i_1}}\right]^2$$

(см. выше § IV, 1), A)).

Введем для упрощения начертания формул сокращенные обозначения

$$p_{ij}' - p_{ij} = dp_{ij}', \qquad p_{i'}' - p_{i'} = dp_{i'}'$$

и разложим $\left[\frac{p_{ij}'}{p_{i'}'}\right]^2$ в ряд по степеням dp_{ij}' и $dp_{i'}'$:

$$\left[\frac{p_{ij}'}{p_{i'}'}\right]^2 = \frac{[p_{ij}+dp_{ij}']^2}{[p_{i'}+dp_{i'}']^2} = \frac{p_{ij}^2}{p_{i'}^2}\left[1+\frac{dp_{ij}'}{p_{ij}}\right]^2\left[1+\frac{dp_{i'}'}{p_{i'}}\right]^{-2} =$$

$$= \frac{p_{ij}^2}{p_{i'}^2}\left[1+\frac{2dp_{ij}'}{p_{ij}}+\frac{(dp_{ij}')^2}{p_{ij}^2}\right]\left[1-\frac{2dp_{i'}'}{p_{i'}}+\frac{3(dp_{i'}')^2}{p_{i'}^2}-\cdots\right] =$$

$$= \frac{p_{ij}^2}{p_{i'}^2}\left[1+\frac{2dp_{ij}'}{p_{ij}}-\frac{2dp_{i'}'}{p_{i'}}+\frac{(dp_{ij}')^2}{p_{ij}^2}-\frac{4dp_{ij}'\,dp_{i'}'}{p_{ij}\,p_{i'}}+\frac{3(dp_{i'}')^2}{p_{i'}^2}+\cdots\right].$$

Переходя к математическим ожиданиям, находим отсюда:

$$E\left[\frac{p_{ij}'}{p_{i'}'}\right]^2 = \frac{p_{ij}^2}{p_{i'}^2}\left\{1+E\left[\frac{2dp_{ij}'}{p_{ij}}-\frac{2dp_{i'}'}{p_{i'}}\right]+\right.$$

$$\left.+E\left[\frac{(dp_{ij}')^2}{p_{ij}^2}-\frac{4dp_{ij}'\,dp_{i'}'}{p_{ij}\,p_{i'}}+\frac{3(dp_{i'}')^2}{p_{i'}^2}\right]+\cdots\right\}$$

Выражения

$$E\{(dp_{ij}')^h(dp_{i'}')^{2s-1-h}\} \qquad \text{и} \qquad E\{(dp_{ij}')^h(dp_{i'}')^{2s-h}\}$$

не содержат членов порядка ниже $(1/N)^s$. Останавливая разложение на членах вида $(dp_{ij}')^h(dp_{i'}')^{2s-h}$, мы получим, следовательно, искомое математическое ожидание $E\left[\frac{n_{ij}}{n_{i'}}\right]^2$ с точностью до членов порядка $(1/N)^s$ включительно. Если удовольствоваться приближением до членов порядка $1/N$, то ряд может быть оборван там, где он остановлен в выписанной выше формуле. Если бы требовалось найти также члены порядка $(1/N)^2$, то надо было бы прихватить также члены, в которые dp_{ij}' и $dp_{i'}'$ входят в третьих и в четвертых степенях и т. д. Вычисление математических ожиданий отдельных членов правой стороны равенства не представляет чрезмерных трудностей. С точностью до $(1/N)^2$ получаем, например, после некоторых сокращений

$$E\left[\frac{n_{ij}}{n_{i'}}\right]^2 = \left[\frac{p_{ij}}{p_{i'}}\right]^2\left\{1 + \frac{1}{Np_{i'}}\frac{p_{i'} - p_{ij}}{p_{ij}} + \frac{1}{N^2p_{i'}^2}\frac{(p_{i'} - p_{ij})(1 - p_{i'})}{p_{ij}} + \cdots\right\}.$$

B) Замечая, что

$$\frac{n_{ij}^2}{n_{i'}n_{j'}} = \frac{p_{ij}^2}{p_{i'}p_{j'}}\left[1 + \frac{dp_{ij}'}{p_{ij}}\right]^2\left[1 + \frac{dp_{i'}'}{p_{i'}}\right]^{-1}\left[1 + \frac{dp_{j'}'}{p_{j'}}\right]^{-1},$$

находим сходным путем

$$E\frac{n_{ij}^2}{n_{i'}n_{j'}} = \frac{p_{ij}^2}{p_{i'}p_{j'}}\left\{1 + \frac{1}{Np_{i'}p_{j'}}\frac{(p_{i'} - p_{ij})(p_{j'} - p_{ij})}{p_{ij}} + \right.$$
$$\left. + \frac{1}{N^2p_{i'}^2p_{j'}^2}\frac{(p_{i'} - p_{ij})(p_{j'} - p_{ij})(2p_{ij} - p_{i'}p_{j'})}{p_{ij}} + \cdots\right\}.$$

C) Опираясь на последнюю формулу, мы можем составить представление о систематической погрешности оценки априорного показателя взаимной сопряженности между переменными (см. глава четвертая, § II, 1)) по случайному значению эмпирического показателя взаимной сопряженности (см. глава пятая, § VI). Подставляя полученное выше выражение для

$$E\frac{n_{ij}^2}{n_{i'}n_{j'}}$$

в формулу, определяющую величину эмпирического показателя взаимной сопряженности, находим с тем же приближением

$$E[\varphi']^2 = \varphi^2 + \frac{1}{N}\left\{\sum_i\sum_j \frac{p_{ij}[p_{i'} - p_{ij}][p_{j'} - p_{ij}]}{p_{i'}^2p_{j'}^2}\right\} +$$
$$+ \frac{1}{N^2}\left\{\sum_i\sum_j \frac{p_{ij}(p_{i'} - p_{ij})(p_{j'} - p_{ij})(2p_{ij} - p_{i'}p_{j'})}{p_{i'}^3p_{j'}^3}\right\} + \cdots$$

Мы убеждаемся, таким образом, что математическое ожидание $[\varphi']^2$ всегда больше φ^2: величина показателя взаимной сопряженности между переменными систематически переоценивается, если за предположительное значение φ^2 принимается численное значение $[\varphi']^2$, — в отличие от величины δ, которая, как мы видели (см. выше § III, 2) систематически недооценивается, если ее предположительное значение приравнивается δ'. Устранить погрешность тем путем, как в случае δ', мы в случае $[\varphi']^2$ не в состоянии, так как точное значение $E[\varphi']^2$ неизвестно, а полученное разложение в ряд не дает возможности притти к заключению, каким образом следовало бы видоизменить $[\varphi']^2$, дабы математическое ожидание видоизмененной функции в точности равнялось φ^2 при всяком числе испытаний.

Если переменные X и Y стохастически независимы, то $\varphi^2 = 0$. Для математического же ожидания $[\varphi']^2$ находим при стохастической независимости X и Y значение

$$E[\varphi']^2 = \frac{(k-1)(l-1)}{N} + \frac{(k-1)(l-1)}{N^2} + \cdots$$

Представляется весьма правдоподобным, что математическое ожидание $[\varphi']^2$ при стохастической независимости X и Y в точности равняется

$$\frac{(k-1)(l-1)}{N-1};$$

мне не удалось еще, однако, найти доказательство этого предположения. Если бы гипотеза подтвердилась, то, вычетая

$$\frac{(k-1)(l-1)}{N-1}$$

из $[\varphi']^2$, мы могли бы получить функцию случайных значений X и Y, математическое ожидание которой при взаимной независимости переменных в точности равнялось бы нулю при всяком числе испытаний.

Сходным образом может быть приближенно вычислено рассеяние $[\varphi']^2$. В общем случае закона связи любого вида находим в первом приближении

$$\sigma^2_{[\varphi']^2} = E\{[\varphi']^2 - E[\varphi']^2\}^2 =$$

$$= \frac{1}{N}\Big\{4\sum_i\sum_j \frac{p_{ij}^3}{p_i^2 p_j^2} - 3\sum_i\left[\frac{1}{p_i^3}\Big(\sum_j \frac{p_{ij}^2}{p_j}\Big)^2\right] - 3\sum_j\left[\frac{1}{p_j^3}\Big(\sum_i \frac{p_{ij}^2}{p_i}\Big)^2\right] +$$

$$+ 2\sum_i\sum_j\left[\frac{p_{ij}}{p_i^2 p_j^2}\Big(\sum_{f=1}^{k} \frac{p_{fj}^2}{p_f}\Big)\Big(\sum_{g=1}^{l} \frac{p_{ig}^2}{p_g}\Big)\right]\Big\} + \cdots$$

Если переменные X и Y стохастически незаьисимы, то член порядка $1/N$ в разложении рассеяния $[\varphi']^2$ по степеням $1/N$ обращается в нуль. Рассеяние

$[\varphi']^2$ представляется, таким образом, при взаимной независимости переменных величиною порядка не ниже $(1/N)^2$, а средняя квадратическая ошибка $[\varphi']^2$ величиною порядка не ниже $1/N$.

D) При помощи сходных приемов могут быть приближенно вычисляемы математические ожидания всех тех функций случайных значений переменных, которые получаются путем замены вероятностей частостями в формулах, служащих для определения сводных характеристик априорного закона связи. Математические выкладки, по большей части, весьма утомительны, — в особенности, если не довольствоваться приближением до членов порядка $1/N$, — но не представляют непреодолимых трудностей и при вычислениях, идущих до членов более высокого порядка. Необходимо лишь твердо держать в памяти, что математическое ожидание четных степеней разностей dp_{ij}', dp_i' и т. д. всегда содержит члены того же порядка, как и математическое ожидание непосредственно предшествующей нечетной степени. Для того, чтоб повысить степень точности конечного результата, необходимо всегда делать сразу два шага в направлении удлинения исходного разложения, — добавлять члены не только следующей нечетной степени разностей dp_{ij}', dp_i' и т. д., но также и следующие за ними члены четной степени. При недостаточном внимании к этому правилу в формулы вкрадываются грубые ошибки.

3) Вместо того, чтоб исходить из разложения в ряд по степеням разностей dp_{ij}', dp_i' и т. д., можно также отправляться от разложения в ряд по степеням разностей тех или иных функций вероятностей и соответственных функций частостей. Если математические ожидания удерживаемых степеней разностей таких функций представляются известными, то такою заменою можно иногда осязательно облегчить выкладки. Необходимо лишь так выбирать функции вероятностей, чтоб математические ожидания высших степеней разностей представлялись величинами малыми по сравнению с математическими ожиданиями низших степеней.

A) Положим (ср. глава пятая, § II, 2)):

$$\mu_{f/g}'' = \sum_i \sum_j p_{ij}' [x_i - m_{1/0}]^f [y_i - m_{0/1}]^g, \qquad d\mu_{f/g}'' = \mu_{f/g}'' - \mu_{f/g}.$$

Замечая, что

$$\mu_{1/0} = \mu_{0/1} = 0,$$

находим:

$$m_{1/1}' - m_{1/0}' m_{0/1}' = \mu_{1/1}'' - \mu_{1/0}'' \mu_{0/1}'' = \mu_{1/1} + d\mu_{1/1}'' - d\mu_{1/0}'' d\mu_{0/1}''$$

$$m_{2/0}' - [m_{1/0}']^2 = \mu_{2/0}'' - [\mu_{1/0}'']^2 = \mu_{2/0} + d\mu_{2/0}'' - (d\mu_{1/0}'')^2$$

$$m_{0/2}' - [m_{0/1}']^2 = \mu_{0/2}'' - [\mu_{0/1}'']^2 = \mu_{0/2} + d\mu_{0/2}'' - (d\mu_{0/1}'')^2.$$

Мы можем, следовательно, представить эмпирический коэффициент корреляции (см. глава пятая, § II, 2)) в виде

$$r_{1/1}' = \frac{\mu_{1/1}'' - \mu_{1/0}''\mu_{0\,1}''}{\sqrt{[\mu_{2/0}'' - (\mu_{1/0}'')^2][\mu_{0/2}'' - (\mu_{0\,1}'')^2]}} =$$

$$= r_{1/1}\frac{\left[1 + \frac{d\mu_{1/1}''}{\mu_{1/1}} - \frac{d\mu_{1/0}''d\mu_{0\,1}''}{\mu_{1/1}}\right]}{\left\{\left[1 + \frac{d\mu_{2/0}''}{\mu_{2/0}} - \frac{(d\mu_{1/0}'')^2}{\mu_{2\,0}}\right]\left[1 + \frac{d\mu_{0/2}''}{\mu_{0/2}} - \frac{(d\mu_{0\,1}'')^2}{\mu_{0\,2}}\right]\right\}^{1/2}}$$

Разлагая в ряд по степеням разностей

$$d\mu_{1/1}'',\ d\mu_{2/0}'',\ d\mu_{0\,2}'',\ d\mu_{1,0}'',\ d\mu_{0\,1}'',$$

находим:

$$r_{1/1}' = r_{1,1}\left[1 + \frac{d\mu_{1/1}''}{\mu_{1/1}} - \frac{d\mu_{1\,0}''\,d\mu_{0/1}''}{\mu_{1/1}}\right] \times$$

$$\left[1 - \frac{1}{2}\left(\frac{d\mu_{2/0}'' - (d\mu_{1/0}'')^2}{\mu_{2,0}}\right) + \frac{3}{8}\left(\frac{d\mu_{2/0}'' - (d\mu_{1\,0}'')^2}{\mu_{2,0}}\right)^2 + \cdots\right] \times$$

$$\left[1 - \frac{1}{2}\left(\frac{d\mu_{0/2}'' - (d\mu_{0/1}'')^2}{\mu_{0/2}}\right) + \frac{3}{8}\left(\frac{d\mu_{0/2}'' - (d\mu_{0\,1}'')^2}{\mu_{0\,2}}\right)^2 + \cdots\right] =$$

$$= r_{1/1}\left\{1 + \left[\frac{d\mu_{1/1}''}{\mu_{1/1}} - \frac{1}{2}\frac{d\mu_{2/0}''}{\mu_{2/0}} - \frac{1}{2}\frac{d\mu_{0/2}''}{\mu_{0/2}}\right] + \left[\frac{1}{2}\frac{(d\mu_{1/0}'')^2}{\mu_{2/0}} + \frac{1}{2}\frac{(d\mu_{0/1}'')^2}{\mu_{0/2}} + \right.\right.$$

$$+ \frac{3}{8}\frac{(d\mu_{2/0}'')^2}{\mu_{2/0}^2} + \frac{3}{8}\frac{(d\mu_{0/2}'')^2}{\mu_{0/2}^2} + \frac{1}{4}\frac{d\mu_{2/0}''\,d\mu_{0/2}''}{\mu_{2/0}\,\mu_{0/2}} - \frac{1}{2}\frac{d\mu_{1/1}''\,d\mu_{0\,2}''}{\mu_{1/1}\,\mu_{0\,2}} -$$

$$\left.\left. - \frac{1}{2}\frac{d\mu_{1/1}''\,d\mu_{2/0}''}{\mu_{1/1}\,\mu_{2/0}} - \frac{d\mu_{1\,0}''\,d\mu_{0/1}''}{\mu_{1/1}}\right] + \cdots\right\}.$$

Математическое ожидание $[d\mu_{1/1}'']^{2s-1}$, $[d\mu_{2/0}'']^{2s-1}$ и т. п. и математическое ожидание $[d\mu_{1/1}'']^{2s}$, $[d\mu_{2/0}'']^{2s}$ и т. п. не содержат членов порядка ниже $(1/N)^s$. Переходя к математическим ожиданиям, мы находим, следовательно, величину $Er_{1/1}'$ с точностью до членов порядка $(1/N)^s$, если останавливаем разложение на членах вида $[d\mu_{1/1}'']^{2s}$, $[d\mu_{2/0}'']^{2s}$, $[d\mu_{1/1}'']^h\,[d\mu_{2/0}'']^{2s-h}$ и т. д. и подставляем значения $E[d\mu_{1/1}'']^{2s}$, $E[d\mu_{1/1}'']^h\,[d\mu_{2\,0}'']^{2s-h}$ и т. д., вычисление которых не представляет непреодолимых трудностей.

Выполняя выкладки, я нахожу с точностью до членов порядка $\left(\frac{1}{N}\right)^2$

$$Er_{1/1}' = r_{1/1} + \frac{1}{N}\left\{\frac{1}{4}r_{2/2}r_{1/1} + \frac{3}{8}r_{1/1}[r_{4/0} + r_{0/4}] - \frac{1}{2}[r_{3\,1} + r_{1/3}]\right\} +$$

$$+ \frac{1}{N^2}\left\{\frac{1}{4}r_{3/3} - \frac{5}{16}r_{1/1}[r_{6/0} + r_{0/6}] - \frac{3}{16}r_{1/1}[r_{4/2} + r_{2/4}] + \frac{3}{8}[r_{5/1} + r_{1/5}] -\right.$$

$$- \frac{1}{4}r_{2/2}r_{1/1} + \frac{1}{2}[r_{3/1} + r_{1/3}] + \frac{15}{32}r_{1/1}r_{2/2}[r_{4/0} + r_{0/4}] - \frac{3}{8}r_{1/1}[r_{4/0} + r_{0/4}] +$$

$$\left. + \frac{9}{64}r_{1/1}r_{4/0}\,r_{0/4} + \frac{9}{32}r_{2/2}^2 r_{1/1} - \frac{15}{16}[r_{3/1}r_{4/0} + r_{1,3}r_{0/4}] - \frac{3}{16}r_{3/1}r_{0/4} + r_{1/3}r_{4/0}] -\right.$$

$$-\frac{3}{8}r_{2/2}[r_{3/1}+r_{1/3}]+\frac{105}{128}r_{1/1}[r_{4/0}^2+r_{0/4}^2]-\frac{1}{4}r_{3/0}r_{0/3}+\frac{15}{8}r_{1/1}[r_{3/0}^2+r_{0/3}^2]-$$

$$-\frac{9}{4}[r_{2/1}r_{3/0}+r_{1/2}r_{0/3}]-\frac{5}{4}r_{2/1}r_{1/2}+\frac{3}{8}r_{1/1}[r_{2/1}^2+r_{1/2}^2]+$$

$$+\frac{3}{4}r_{1/1}[r_{2/1}r_{0/3}+r_{1/2}r_{3/0}]-\frac{1}{2}r_{1/1}+\frac{1}{2}r_{1/1}^3\Big\}+\cdots.$$

Формула эта имеет силу при любом законе связи между X и Y. Если закон связи таков, что регрессия Y по Y и регрессия X по Y прямолинейны, то (см. глава четвертая, § III, 2), C)) $r_{h/1}=r_{1/1}\,r_{h+1/0}$ и $r_{1/h}=r_{1/1}\,r_{0/h+1}$; для математического ожидания эмпирического коэффициента корреляции находим в этом случае

$$Er_{1/1}'=r_{1/1}+\frac{1}{N}r_{1/1}\left\{\frac{1}{4}r_{2/2}-\frac{1}{8}[r_{4/0}+r_{0/4}]\right\}+$$

$$+\frac{1}{N^2}\left\{\frac{1}{4}r_{3/3}+\frac{1}{16}r_{1/1}[r_{6/0}+r_{0/6}]-\frac{3}{16}r_{1/1}[r_{4/2}+r_{2/4}]-\frac{1}{4}r_{1/1}r_{2/2}+\right.$$

$$+\frac{9}{32}r_{1/1}r_{2/2}^2+r_{1/1}[r_{4/0}+r_{0/4}]\left[\frac{1}{8}+\frac{3}{32}r_{2/2}\right]-\frac{15}{128}r_{1/1}[r_{4/0}+r_{0/4}]^2-$$

$$\left.-[1-r_{1/1}^2]\left[\frac{1}{2}r_{1/1}+\frac{1}{4}r_{3/0}r_{0/3}+\frac{3}{8}r_{1/1}(r_{3/0}^2+r_{0/3}^2)\right]\right\}+\cdots.$$

При $r=0$ получаем, каков бы ни был при этом закон связи,

$$Er_{1/1}'=-\frac{1}{2N}[r_{3/1}+r_{1/3}]+\frac{1}{N^2}\left\{\frac{1}{4}r_{3/3}+\frac{3}{8}[r_{5/1}+r_{1/5}]+\frac{1}{2}[r_{3/1}+r_{1/3}]-\right.$$

$$-\frac{15}{16}[r_{3/1}r_{4/0}+r_{1/3}r_{0/4}]-\frac{3}{16}[r_{3/1}r_{0/4}+r_{1/3}r_{4/0}]-\frac{3}{8}r_{2,2}[r_{3/1}+r_{1/3}]-\frac{1}{4}r_{3/0}r_{0/3}-$$

$$\left.-\frac{9}{4}[r_{2/1}r_{3/0}+r_{1/2}r_{0/3}]-\frac{5}{4}r_{2/1}r_{1/2}\right\}+\cdots.$$

Математическое ожидание $r_{1/1}'$ представляется, таким образом, при $r=0$ в общем случае отличным от нуля и может принимать как положительные, так и отрицательные значения в зависимости от знака суммы $[r_{3/1}+r_{1/3}]$.

Если переменные X и Y стохастически независимы, то (см. глава четвертая, § III, 1)) $r_{g/f}=r_{g/0}\,r_{0/f}$ при всяких g и f, причем $r_{1/0}=r_{0/1}=0$. Общая формула дает в границах нашего приближения $Er_{1/1}'=0$ в соответствии с найденным выше (см. § IV, 1), B)) точным результатом.

В случае нормальной корреляции между X и Y находим, подставляя выражения параметров $r_{f/g}$ через коэффициент корреляции $r_{1/1}$ (см. глава четвертая, § V, 2)):

$$Er_{1/1}'=r_{1/1}\left\{1-\frac{1-r_{1/1}^2}{2N}-\frac{3[1-r_{1/1}^2][1+3r_{1/1}^2]}{8N^2}+\cdots\right\}.$$

Величина коэффициента корреляции между X и Y систематически недооценивается в случае нормальной корреляции, если за предположительное значение коэффициента корреляции между переменными $r_{1/1}$ принимается величина эмпирического коэффициента корреляции между случайными значениями переменных $r_{1/1}'$. Но при иных видах закона связи величина коэффициента корреляции между переменными может и преувеличиваться, если предположительное значение $r_{1/1}$ приравнивается $r_{1/1}'$.

Устранить систематическую погрешность оценки значения $r_{1/1}$ по значению $r_{1/1}'$ путем преобразования функции $r_{1/1}'$ мы так же не в состоянии, как устранить систематическую погрешность оценки показателя взаимной сопряженности φ^2 по случайному значению $[\varphi']^2$. По сравнению с оценкою показателя взаимной сопряженности мы находимся при оценке коэффициента корреляции все же в лучшем положении в силу того, что хоть для случая стохастической независимости переменных располагаем точным значением $Er_{1/1}'$ и знаем, что оно в точности равняется $r_{1/1} = 0$ при всяком числе испытаний.

Сходным путем находим среднюю квадратическую ошибку $r_{1/1}'$. В первом приближении

$$\sigma^2 r_{1/1}' = E\{r_{1/1}' - Er_{1/1}'\}^2 =$$

$$= \frac{1}{N}\left\{r_{2/2}\left[1 + \frac{1}{2}r_{1/1}^2\right] - r_{1/1}[r_{3/1} + r_{1/3}] + \frac{1}{4}r_{1/1}^2[r_{4/0} + r_{0/4}]\right\} + \cdots.$$

То же значение имеет в первом приближении и математическое ожидание второй степени разности $r_{1/1}' - r_{1/1}$:

$$E\{r_{1/1}' - r_{1/1}\}^2 =$$

$$= \frac{1}{N}\left\{r_{2/2}\left[1 + \frac{1}{2}r_{1/1}^2\right] - r_{1/1}[r_{3/1} + r_{1/3}] + \frac{1}{4}r_{1/1}^2[r_{4/0} + r_{0/4}]\right\} + \cdots.$$

При нормальной корреляции находим с тою же степенью приближения

$$E\{r_{1/1}' - Er_{1/1}'\}^2 = E\{r_{1/1}' - r_{1/1}\}^2 = \frac{[1 - r_{1/1}^2]^2}{N} + \cdots.$$

В случае прямолинейной регрессии Y по X имеем;

$$\sigma^2 r_{1/1}' = \frac{1}{N}\left\{r_{2/2}\left[1 + \frac{1}{2}r_{1/1}^2\right] - \frac{3}{4}r_{1/1}^2 r_{4/0} - r_{1/1}r_{1/3} + \frac{1}{4}r_{1/1}^2 r_{0/4}\right\} + \cdots.$$

Если прямолинейна также регрессия X по Y, то

$$\sigma^2 r_{1/1}' = \frac{1}{N}\left\{r_{2/2}\left[1 + \frac{1}{2}r_{1/1}^2\right] - \frac{3}{4}r_{1/1}^2[r_{4/0} + r_{0/4}]\right\} + \cdots.$$

При $r_{1\,1} = 0$ получаем

$$\sigma^2_{r_{1/1}'} = E\,[r_{1\,1}']^2 = \frac{r_{2\,2}}{N} + \cdots$$

Если переменные стохастически независимы, то $r_{2\,2} = r_{2/0} r_{0/2} = 1$. Из последней формулы находим, подставляя $r_{2\,2} = 1$, в первом приближении

$$\sigma^2_{r'_{1\,1}} = \frac{1}{N},$$

тогда как точное значение рассеяния $r_{1/1}'$ равняется в случае взаимной независимости переменных (см. выше § IV, 1), B)) $\frac{1}{N-1}$.

B) Применим тот же метод к исследованию вопроса о нахождении предположительных значений коэффициентов уравнения регрессии. Ограничивая задачу исследованием прямой, наилучше подходящей к действительной линии регрессии (см. глава четвертая. § III, 4) и глава пятая, § III), попытаемся выяснить, в какой мере эмпирическая прямая

$$m_{/1}^{(i)'} = \frac{m_{2\,0}'\, m_{0/1}' - m_{1\,1}'\, m_{1\,0}'}{m_{2/0}' - [m_{1/0}']^2} + \frac{m_{1,1}' - m_{1/0}'\, m_{0/1}'}{m_{2/0}' - [m_{1/0}']^2}\, x_i = a_{/0}' + a_{/1}'\, x_i$$

может почитаться удовлетворительным приближением к интересующей нас априорной прямой

$$m_{/1}^{(i)} = \frac{m_{2/0}\, m_{0/1} - m_{1/1}\, m_{1/0}}{m_{2/0} - m_{1/0}^2} + \frac{m_{1/1} - m_{1/0}\, m_{0/1}}{m_{2/0} - m_{1/0}^2}\, x_i = a_{/0} - a_{/1}\, x_i.$$

Другими словами, в какой мере коэффициенты $a_{/0}'$ и $a_{/1}'$ могут рассматриваться, как предположительные значения коэффициентов $a_{/0}$ и $a_{/1}$?

Я не выписываю разложений в ряд, так как выкладки складываются подобно тем, какие мы рассмотрели на примере эмпирического коэффициента корреляции, и не дают ничего особо поучительного. С приближением до членов порядка $1/N$ находим:

$$Ea_{/1}' = a_{/1} + \frac{1}{N}\sqrt{\frac{\mu_{0/2}}{\mu_{2,0}}}\,[r_{4/0} r_{1/1} - r_{3/1}] + \cdots.$$

$$Ea_{/0}' = a_{/0} + \frac{1}{N}\sqrt{\frac{\mu_{0/2}}{\mu_{2\,0}}}\,\{[r_{3/0} r_{1/1} - r_{2/1}] - m_{1/0}[r_{4/0} r_{1/1} - r_{3/1}]\} + \cdots.$$

Эти соотношения имеют силу при любом виде закона связи между X и Y и при любом виде истинной линии регрессии Y по X. Если истинная линия регрессии Y по X прямая, то (см. глава четвертая, § III, 2), C)) $r_{h/1} = r_{1/1}\, r_{h+1/0}$, и, в границах нашего приближения, $Ea_{/1}' = a_{/1}$, $Ea_{/0}' = a_{/0}$. Представляется правдоподобным, что при прямолинейной регрессии Y по X математическое ожи-

дание a_1' в точности равно a_1, а математическое ожидание a_0' в точности равно a_0. Доказать это, мне, однако пока не удается.

Для рассеяний a_1' и a_0' находим тем же путем:

$$\sigma^2 a_1' = \frac{1}{N}\frac{\mu_{0,2}}{\mu_{2,0}}\left[r_{2,2} + r_{1,0}r_{1,1} - 2r_{3,1}r_{1,1}\right] + \cdots.$$

$$\sigma^2 a_0' = \frac{1}{N}\frac{\mu_{0,2}}{\mu_{2,0}}\left\{\mu_{2,0}\left[1 - r_{1,1}^2\right] - 2m_{1,0}\sqrt{\mu_{2,0}}\left[r_{1,2} - 2r_{2,1}r_{1,1} + r_{3,0}r_{1,1}^2\right] + \right.$$

$$\left. + m_{1,0}\left[r_{2,2} + r_{1,0}r_{1,1}^2 - 2r_{3,1}r_{1,1}\right]\right\} + \cdots.$$

Эти соотношения имеют силу при любом виде закона связи. Если закон связи таков, что истинная регрессия Y по X прямолинейна, то

$$\sigma^2 a_1' = \frac{1}{N}\frac{\mu_{0,2}}{\mu_{2,0}}\left[r_{2,2} - r_{1,0}r_{1,1}^2\right] + \cdots$$

$$\sigma^2 a_0' = \frac{1}{N}\frac{\mu_{0,2}}{\mu_{2,0}}\left\{\mu_{2,0}\left[1 - r_{1,1}^2\right] - \right.$$

$$\left. - 2m_{1,0}\sqrt{\mu_{2,0}}\left[r_{1,2} - r_{3,0}r_{1,1}^2\right] + m_{1,0}^2\left[r_{2,2} - r_{1,0}r_{1,1}^2\right]\right\} + \cdots.$$

Если регрессия Y по X прямолинейна, а связь Y с X гомоскедастична, то

$$\sigma^2 a_1' = \frac{1}{N}\frac{\mu_{0,2}}{\mu_{2,0}}\left[1 - r_{1,1}^2\right] + \cdots.$$

$$\sigma^2 a_0' = \frac{1}{N}\frac{\mu_{0,2}}{\mu_{2,0}}\left[1 - r_{1,1}^2\right]\left[\mu_{2,0} + m_{1,0}^2\right] + \cdots.$$

или, в первом приближении,

$$\sigma^2 a_0' = \left[\mu_{2,0} + m_{1,0}^2\right]\sigma^2 a_1' = m_{2,0}\,\sigma^2 a_1'.$$

Последние формулы имеют силу и в случае нормальной корреляции, так как при нормальной корреляции регрессия прямолинейна, а связь гомоскедастична (см. глава четвертая § V, 2)).

C) Исходя из определений априорного и эмпирического корреляционного отношений (см. глава четвертая, § IV, 2) и глава пятая § IV)

$$\eta_{Y/X}^2 = \frac{\sum_i p_i\left[m_1^{(i)} - m_{0,1}\right]^2}{\sum_j p_j\left[y_j - m_{0,1}\right]^2}, \qquad \left[\eta_{Y/X}'\right]^2 = \frac{\sum_i p_i'\left[m_1^{(i)\prime} - m_{0,1}'\right]^2}{\sum_j p_j'\left[y_j - m_{0,1}'\right]^2},$$

находим при помощи аналогичных разложений в ряд с точностью до членов порядка $1/N$

$$E[\eta_{Y/X}']^2=\eta_{X/1}{}^2+\frac{1}{N}\left\{[r_{0/4}+1]\,\eta_{Y/X}{}^2-1+\frac{1}{\mu_{0/2}}\sum_i \mu_{i2}{}^{(i)}-\right.$$

$$-\frac{1}{\mu_{0/2}{}^2}\left[\sum_i p_{i/}(m_{i1}{}^{(i)}-m_{0/1})^4+5\sum_i p_{i/}(m_{i1}{}^{(i)}-m_{0/1})^2\,\mu_{i2}{}^{(i)}+\right.$$

$$\left.\left.+2\sum_i p_{i/}(m_{i1}{}^{(i)}-m_{0/1})\,\mu_{i3}{}^{(i)}\right]\right\}+\cdots$$

$$\sigma^2[\eta_{Y/X}']^2=\frac{1}{N}\left\{r_{0/4}\,\eta_{Y/X}{}^4+\frac{1}{\mu_{0/2}{}^2}\left[(1-2\eta_{Y/X}{}^2)\sum_i p_i\,(m_{i1}{}^{(i)}-m_{0/1})^4+\right.\right.$$

$$\left.\left.+2\,(2-5\eta_{Y/X}{}^2)\sum_i p_{i/}(m_{i1}{}^{(i)}-m_{0/1})^2\mu_{i2}{}^{(i)}-4\eta_{1/X}{}^2\sum_i p_i\,(m_{i1}{}^{(i)}-m_{0/1})\,\mu_{i3}{}^{(i)}\right]\right\}+\cdots$$

Эти соотношения имеют силу, каков бы ни был закон связи между X и Y. Если закон связи таков, что линия регрессии Y по X имеет вид прямой, находим

$$E[\eta_{Y/X}']^2=r_{1/1}{}^2+\frac{1}{N}\left\{\frac{1}{\mu_{0/2}}\sum_i \mu_{i2}{}^{(i)}-1+r_{1/1}{}^2+r_{1/1}{}^2\,r_{0/4}+r_{1/1}{}^2\,r_{2/2}-2r_{1/1}\,r_{1/3}\right\}+\cdots$$

$$\sigma^2[\eta_{Y/X}']^2=\frac{1}{N}r_{1/1}{}^2\left\{2\,[r_{1/1}{}^2+2]\,r_{2/2}+r_{1/1}{}^2\,r_{0/4}-3r_{1/1}{}^2\,r_{1/0}-4r_{1/1}\,r_{1/3}\right\}+\cdots.$$

Если регрессия Y по X прямолинейна, а связь Y с X гомоскедастична, то

$$E[\eta_{Y/X}']^2=r_{1/1}{}^2+\frac{1}{N}\left\{[k-1]\,[1-r_{1/1}{}^2]+r_{1/1}{}^2\,r_{0/4}+r_{1/1}{}^2\,r_{2/2}-2r_{1/1}\,r_{1/3}\right\}+\cdots.$$

Если переменная Y не коррелирована с X, то

$$E[\eta_{Y/X}']^2=\frac{1}{N}\left[\frac{1}{\mu_{0/2}}\sum_i \mu_{i2}{}^{(i)}-1\right]+\cdots.$$

Что касается рассеяния $[\eta_{Y/X}']^2$, то в разложении его по степеням $1/N$ член порядка $1/N$ обращается в нуль в случае некоррелированности Y с X. При некоррелированности Y с X квадратическая ошибка $[\eta_{Y/X}']^2$ представляется, таким образом, величиною порядка не ниже $1/N$.

Если переменная Y некоррелирована с X и кроме того связь Y с X гомоскедастична, то

$$E[\eta_{Y/X}']^2=\frac{k-1}{N}+\cdots.$$

В случае нормальной корреляции находим

$$E[\eta_{Y/X}']^2 = r_{1/1}^2 + \frac{1}{N}[1 - r_{1/1}^2][k - 1 - 2r_{1/1}^2] + \cdots$$

$$\sigma^2[\eta_{1/X}']^2 = \frac{4}{N} r_{1/1}^2 [1 - r_{1/1}^2]^2 + \cdots .$$

D) Мы видели (см. глава четвертая, § IV, 3)), что равенство $\eta_{1/X}^2 = r_{1/1}^2$ представляется необходимым и достаточным условием прямолинейности регрессии Y по X. Величина разности $\eta_{1/X}^2 - r_{1/1}^2$, обозначаемой обычно греческою буквою $\zeta_{1/X}^2$, может, следовательно, служить критерием при решении, прямолинейна ли регрессия Y по X, или нет: если $\zeta_{Y/X}^2$ равняется нулю, регрессия Y по X не может не быть прямолинейною; если $\zeta_{Y/X}^2$ отлична от нуля, регрессия Y по X не может быть прямолинейною.

Оценка величины $\zeta_{1/X}^2$ по данным опыта исходит от случайного значения разности между соответственными эмпирическими величинами

$$[\zeta_{Y/X}']^2 = [\eta_{Y/X}']^2 - [r_{1/1}']^2.$$

При помощи тех же приемов, как выше, находим

$$E[r_{1/1}']^2 = r_{1/1}^2 + \frac{1}{N}\{r_{2/2}[1 + r_{1/1}^2] + r_{1/1}^2[r_{4/0} + r_{0/4}] - 2r_{1/1}[r_{3/1} + r_{1/3}]\} + \cdots .$$

$$\sigma^2[r_{1/1}']^2 = \frac{1}{N} r_{1/1}^2 \{2[r_{1/1}^2 + 2] r_{2/2} + r_{1/1}^2[r_{4/0} + r_{0/4}] - 4r_{1/1}[r_{3/1} + r_{1/3}]\} + \cdots .$$

Эти соотношения имеют силу, каков бы ни был закон связи между X и Y. Если закон связи таков, что регрессия Y по X прямолинейна, то

$$E[r_{1/1}']^2 = r_{1/1}^2 + \frac{1}{N}\{r_{2/2}[1 + r_{1/1}^2] + r_{1/1}^2 r_{0/4} - r_{1/1}^2 r_{4/0} - 2r_{1/1} r_{1/3}\} + \cdots$$

$$\sigma^2[r_{1/1}']^2 = \frac{1}{N} r_{1/1}^2 \{2[r_{1/1}^2 + 2] r_{2/2} + r_{1/1}^2 r_{0/4} - 3r_{1/1}^2 r_{4/0} - 4r_{1/1} r_{1/3}\} + \cdots .$$

Если прямолинейна также регрессия X по Y, то

$$E[r_{1/1}']^2 = r_{1/1}^2 + \frac{1}{N}\{r_{2/2}[1 + r_{1/1}^2] - r_{1/1}^2[r_{4/0} + r_{0/4}]\} + \cdots$$

$$\sigma^2[r_{1/1}'] = \frac{1}{N} r_{1/1}^2 \{2[r_{1/1}^2 + 2] r_{2/2} - 3r_{1/1}^2[r_{4/0} + r_{0/4}]\} + \cdots .$$

При $r_{1/1} = 0$ находим, каков бы ни был при этом вид регрессии Y по X:

$$E[r_{1/1}']^2 = \frac{1}{N} r_{2/2} + \cdots .$$

В разложении же рассеяния $[r_{1/1}']^2$ по степеням $1/N$ член порядка $1/N$ обращается в нуль, если $r_{11}=0$. В частности, если переменная Y некоррелирована с X, то

$$E[r_{1/1}']^2=\frac{1}{N}r_{2'2}+\cdots,$$

а средняя квадратическая ошибка $[r_{1/1}']^2$ оказывается величиною порядка не ниже $1/N$.

При стохастической независимости переменных $r_{2/2}=1$, и для математического ожидания $[r_{11}']^2$ разложение в ряд по степеням $1/N$ дает

$$E[r_{11}']^2=\frac{1}{N}+\cdots.$$

Точное значение $E[r_{1/1}']^2$, как мы видели (см. выше § IV, 1), В)), равняется $\frac{1}{N-1}$.

В случае нормальной корреляции между переменными получаем

$$E[r_{11}']^2=r_{1/1}^2+\frac{1}{N}[1-r_{11}^2][1-2r_{11}^2]+\cdots$$

$$\sigma^2[r_{1/1}']^2=\frac{4}{N}r_{11}^2[1-r_{11}^2]^2+\cdots.$$

Сопоставляя формулы, полученные для математического ожидания и для рассеяния $[r_{1/1}']^2$, с формулами, найденными для математического ожидания и для рассеяния $[\eta_{Y/X}']^2$, мы видим, что при равенстве $\eta_{Y/X}^2$ и $r_{1/1}^2$ средняя квадратическая ошибка оценки значения $r_{1/1}^2$ по случайному значению $[\eta_{Y/X}']^2$ и по случайному значению $[r_{1/1}']^2$ имеет одну и ту же величину, но систематическая погрешность оценки не одинакова. В случае нормальной корреляции разность между систематическими погрешностями равняется, например,

$$\frac{1}{N}[k-2][1-r_{1/1}^2]+\cdots$$

Разность между систематическими погрешностями оценки величины $r_{1/1}^2$ по случайным значениям $[\eta_{Y/X}']^2$ и $[r_{1/1}']^2$ дает нам в случае прямолинейной регрессии Y по X систематическую погрешность оценки величины $\zeta_{Y/X}^2$, равной при этом нулю, по случайному значению $[\zeta_{Y/X}']^2$. Из приведенных выше формул находим в предположении, что регрессия Y по X прямолинейная,

$$E[\zeta_{1/X}']^2-\zeta_{Y/X}^2=E[\zeta_{Y/X}']^2=E[\eta_{1/X}']^2-E[r_{1/1}']^2=$$

$$=\frac{1}{N}\left\{\frac{1}{\mu_{0,2}}\sum_i\mu_{i2}^{(i)}-[1-r_{1/1}^2]-r_{2,2}+r_{1/1}^2r_{4,0}\right\}+\cdots.$$

Если регрессия Y по X прямолинейна, а связь Y с X гомоскедастична, то

$$E[\zeta_{1/X}']^2 = \frac{1}{N}\{[k-1][1-r_{11}^2] - r_{22} + r_{11}^2 r_{40}\} + \cdots .$$

В случае нормальной корреляции получаем, как указано выше,

$$E[\zeta_{1/X}']^2 = \frac{1}{N}[k-2][1-r_{11}^2] + \cdots$$

Что касается рассеяния $[\zeta_{\eta/X}']^2$, то в общем случае получаем для него путем довольно утомительных выкладок

$$\sigma^2[\zeta_{1/X}']^2 = \frac{1}{N}\Big\{[\eta_{1/X}^2 - r_{11}^2]^2 r_{04} + [\eta_{1/X}^2 - r_{11}^2]\,2r_{1/1}[2r_{13} - r_{11}r_{22}] +$$

$$+ 4r_{11}^2[r_{22} - r_{11}r_{31}] + r_{11}^4 r_{40} + [1 - 2(\eta_{1/X}^2 - r_{11}^2)]\frac{1}{\mu_{02}^2}\sum_i p_i'[m_1^{(i)} - m_{01}]^4 +$$

$$+ [4 - 10(\eta_{1/X}^2 - r_{11}^2)]\frac{1}{\mu_{02}^2}\sum_i p_i'[m_1^{(i)} - m_{01}]^2\mu_2^{(i)} -$$

$$- 4[\eta_{1/X} - r_{11}^2]\frac{1}{\mu_{02}^2}\sum_i p_i'[m_1^{(i)} - m_{01}]\mu_3^{(i)} +$$

$$+ 2r_{1/1}^2\frac{1}{\mu_{20}\mu_{02}}\sum_i p_i'[m_1^{(i)} - m_{01}]^2[x_i - m_{10}]^2 -$$

$$- 4r_{1/1}\frac{1}{\sqrt{\mu_{20}\mu_{02}^3}}\sum_i p_i'[m_1^{(i)} - m_{01}]^3[x_i - m_{1/0}] -$$

$$- 8r_{1/1}\frac{1}{\sqrt{\mu_{20}\mu_{02}^3}}\sum_i p_i'[m_1^{(i)} - m_{01}][x_i - m_{1/0}]\mu_2^{(i)}\Big\} + \cdots .$$

Если регрессия Y по X прямолинейна, то член порядка $1/N$ в разложении рассеяния $[\zeta_{\eta/X}']^2$ по степеням $1/N$ обращается в нуль. Средняя квадратическая ошибка разности $[\eta_{\eta/X}']^2 - [r_{1/1}']^2$ представляется, следовательно, в случае, если $\eta_{1/X}^2 = r_{1/1}^2$ и регрессия Y по X прямолинейна, величиною порядка не ниже $1/N$.

Е) Разложение в ряд по восходящим степеням разностей $dm_{1/0}'$, $d\mu_{1/0}''$ и т. д. предполагает, что математические ожидания высших степеней этих разностей представляются величинами малыми по сравнению с математическими ожиданиями степеней менее высоких. При достаточно большом числе испытаний предположение это может быть обосновано в общем виде для функций эмпирических значений переменных, которые могут быть, в свою очередь, разложены в ряд по восходящим степеням разностей dp_{ij}', dp_i' и т. д.

Пусть функция Z' представляется в виде суммы членов

$$Z' = C + d_1 + d_2 + d_3 + d_4 + \cdots ,$$

причем C обозначает член, не содержащий разностей $dp_{1,}'$, $dp_{1,}'$ и т. д., d_1 — обозначает сумму членов, содержащих эти разности в первой степени, d_2 — сумму членов, содержащих их во второй степени и т. д. Переходя к математическим ожиданиям, находим:

$$EZ' = C + Ed_2 + E[d_3 + d_4] + \cdots, \quad dZ' = Z' - EZ = D + \frac{1}{N}K,$$

где

$$D = d_1 + d_2 + \cdots, \quad \text{а} \quad \frac{1}{N}K = -\{Ed_2 + E[d_3 + d_4] + \cdots\}.$$

Отсюда получаем.

$$E\{dZ'\}^{2h} = ED^{2h} + \binom{2h}{1}\frac{1}{N}KED^{2h-1} + \cdots + \frac{1}{N^{2h}}K^{2h},$$

$$E\{dZ'\}^{2h+1} = ED^{2h+1} + \binom{2h+1}{1}\frac{1}{N}KED^{2h} + \cdots + \frac{1}{N^{2h+1}}K^{2h+1}$$

и убеждаемся, что $E\{dZ'\}^{2h}$ не содержит членов порядка ниже $(1/N)^h$, а $E\{dZ'\}^{2h+1}$ не содержит членов порядка ниже $(1/N)^{h+1}$. Чтоб получить математическое ожидание U' с точностью до членов порядка $(1/N)^h$ включительно, необходимо и достаточно, таким образом, удерживать в разложении U' по степеням разностей $dm_{1\,0}'$, $d\mu_{1\,0}''$ и т. д. члены вида $E\{dZ'\}^{2h}$ и т. п. Правило это должно быть строжайше соблюдаемо. Недостаточное внимание к нему внесло в некоторые из английских исследований недавнего времени весьма прискорбные ошибки. Стремясь повысить точность формул, при вычислении которых ряд ранее обрывался на членах, содержащих разности во второй степени, исследователи начали принимать во внимание члены, содержащие разности в третьей степени, упуская из виду, что четвертая степень разности дает члены того же порядка $(1/N)^2$, как и третья. В результате степень точности полученных формул становилась обманчивою, и вместо уточнения формул получалось порою их искажение.

4) Разложение в ряд по степеням $1/N$ предполагает, что члены, содержащие $1/N$ в высших степенях настолько малы по сравнению с теми, которые содержат $1/N$ в степенях менее высоких, что совокупность их, следующая за последним удерживаемым членом разложения, может быть без ущерба отброшена. Доказать, что это действительно так, можно в строгости лишь путем исследования остаточного члена и сходимости ряда. Так далеко статистики пока не идут в своем математическом анализе. По большей части исследование ограничивается вычислением членов порядка $1/N$. В некоторых случаях — сравнительно очень еще не многих — вычислены также члены порядка $(1/N)^2$. Для статистических исследований остается здесь открытым необъятно широкое поле, разработка которого представляет благодарную задачу для начинающих статистиков: упражнение своих не вполне еще окрепших сил они могут тут с выгодою для себя и с пользою для науки объединить с получением результатов, представляющихся объективно ценными.

Как бы, однако, далеко ни шло усовершенствование метода разложения в ряд по степеням $1/N$, требование, чтоб число испытаний было достаточно большим, неизбежно остается conditio sine qua non его приложимости. Если N невелико, метод этот не способен дать сколько-нибудь надежных результатов. Нередко границы его приложимости намечаются еще уже. Разложение в ряд получается в действительности не по степеням $1/N$, а по степеням $\frac{1}{Np_i}$ или $\frac{1}{Np_{i'}}$ и т. д., и условие приложимости метода сводится уже к требованию, чтобы были достаточно велики не только общее число испытаний, но также числа случаев повторений разных значений переменных и разных комбинаций этих значений. Если вероятности p_{ij} и т. д. не слишком малы, то требование это почти покрывается требованием, чтоб было велико общее число испытаний. Но при малых значениях вероятностей p_{ij} и т. д. разница может быть очень заметна. Если же вероятность p_{ij} представляется малою величиною такого порядка, что произведение Np_{ij} стремится с ростом N к конечному пределу — случай так называемого закона малых чисел, — то метод разложения в ряд по степеням $1/N$ становится совсем неприложимым. Во всех подобных случаях приходится обращаться к иным приемам исследования, чтоб составить представление о размерах априорных величин по данным опыта.

Поскольку задача сводится к вычислению математического ожидания частного, можно нередко с успехом использовать следующий прием. Пусть функция U, математическое ожидание которой нас интересует, представляется в виде частного двух целых рациональных функций Z' и W', вычисление математических ожиданий разных степеней которых не встречается с чрезмерными трудностями. Замечая, что при любом значении постоянной C мы имеем тожественно

$$\frac{1}{W'} = \frac{1}{C} - \frac{W'-C}{CW'} = \frac{1}{C} - \frac{W'-C}{C^2} + \frac{(W'-C)^2}{C^2 W'},$$

находим

$$E\frac{Z'}{W'} = \frac{EZ'}{C} - \frac{EZ'(W'-C)}{C^2} + \frac{1}{C^2} E\frac{Z'(W'-C)^2}{W'}.$$

Предположим, что частное $\frac{Z'}{W'}$ остается при всех своих возможных значениях положительным, — допущение, которое оправдывается во многих случаях. Тогда и $\frac{Z'(W'-C)^2}{W'}$ не может принимать отрицательных значений, и вместе с тем $E\frac{Z'(W'-C)^2}{W'}$ не может быть меньше нуля. Следовательно,

$$(\text{I}) \qquad E\frac{Z'}{W'} > \frac{1}{C} EZ' - \frac{1}{C^2} EZ'(W'-C).$$

Пусть, с другой стороны, частное $\frac{Z'}{W'}$ не принимает значений, превосходящих некоторую постоянную K^2, например, значений, больших единицы. Тогда

$$E\frac{Z'}{W'} < \frac{1}{C}EZ' - \frac{1}{C^2}Z'(W'-C) + \frac{K^2}{C^2}E(W'-C)^2$$

или, при $K^2 = 1$,

$$\text{(II)} \qquad E\frac{Z'}{W'} < \frac{1}{C}EZ' - \frac{1}{C^2}EZ'(W'-C) + \frac{1}{C^2}E(W'-C)^2.$$

Мы получаем, таким образом, два неравенства, намечающих высший и низший предел для искомого значения $E\frac{Z'}{W'}$. Из этих пределов значение $E\frac{Z'}{W'}$ не выходит, каково бы ни было число испытаний N. Входящей в неравенства константе C можно придавать разные значения.

Полагая $C = EW'$, находим:

$$E\frac{Z'}{W'} > \frac{EZ'}{EW'} - \frac{1}{[EW']^2}EZ'[W'-EW']$$

$$E\frac{Z'}{W'} < \frac{EZ'}{EW'} - \frac{1}{[EW']^2}EZ'[W'-EW'] + \frac{1}{[EW']^2}E[W'-EW']^2.$$

Величина $E\frac{Z'}{W'}$ оказывается, таким образом, включенною в пределы, разность между которыми равняется $\frac{1}{[EW']^2}\sigma_{W'}{}^2$ и представляется величиною порядка $1/N$. Если пределы эти недостаточно тесны, то повторением подстановки $\frac{1}{W'} = \frac{1}{EW'} - \frac{W'-EW'}{W'EW'}$ еще раз или еще несколько раз можно их съузить. Мы приходим к неравенствам

$$\text{(I')} \qquad E\frac{Z'}{W'} > \frac{EZ'}{EW'} + \sum_{h=1}^{2t-1}\frac{(-1)^h}{[EW']^{h+1}}EZ'[w'-EW']^h$$

$$\text{(II')} \qquad E\frac{Z'}{W'} < \frac{EZ'}{EW'} + \sum_{h=1}^{2t-1}\frac{(-1)^h}{[EW']^{h+1}}EZ'[W'-EW']^h +$$

$$+ \frac{1}{[EW']^{2t}}E[W'-EW']^{2t},$$

разность между которыми равняется $\frac{1}{[EW']^{2t}}E[W'-EW']^{2t}$ и представляется величиною порядка $(1/N)^t$.

Величине C можно придавать также в каждом из неравенств свое значение, выбирая его так, чтоб получить в первом из неравенств возможно высокий низший предел для $E\frac{Z'}{W'}$, а во втором из неравенств — возможно низкий высший предел. Неравенство (I) принимает при этом вид

$$E\frac{Z'}{W'} > \frac{[EZ']^2}{EZ'W'}.$$

Неравенство (II) приводится к

$$E\frac{Z'}{W'} < 1 - \frac{[EW' - EZ']^2}{E(W')^2 - EZ'W'}.$$

Располагая правые части неравенств (I') и (II') по возрастающим степеням $1/N$, мы получаем $E\frac{Z'}{W'}$ в составе членов порядка до $(1/N)^{t-1}$ включительно в том же виде, как и при помощи описанных выше методов, располагая вместе с тем возможностью оценить размеры погрешности, сопряженной с отбрасыванием членов порядка $(1/N)^t$ и выше. Это теоретическое преимущество использования неравенств покупается, однако, ценою еще бо́льшей громоздкости выкладок. Поскольку интерес сосредоточивается на нахождении членов порядка до $(1/N)^{t-1}$ включительно в разложении $E\frac{Z'}{W'}$ по восходящим степеням $1/N$, представляется более практичным итти путем разложения $\frac{Z'}{W'}$ по восходящим степеням разностей $dp_{i/}'$, $dm_{1/0}'$ и т. д.

5) Мы видели (см. глава четвертая, § II, 1) и § VII), что в случае если обе переменные могут принимать лишь по два разных значения, то величина

$$\frac{\delta}{\sqrt{p_1 p_2 p_{/1} p_{/2}}}$$

может быть рассматриваема и как коэффициент корреляции между переменными, и как показатель их взаимной сопряженности и как корреляционное отношение одной из них к другой. Равным образом и соответственное эмпирическое выражение

$$\frac{\delta'}{\sqrt{p_{1/}' p_{2/}' p_{/1}' p_{/2}'}}$$

может быть получено и из общей формулы для $r_{1/1}'$, и из формулы для φ', и из формулы для $\eta_{Y/X}'$ (см. глава пятая, § VII).

Мы можем, следовательно, при вычислении математического ожидания и рассеяния

$$\frac{\delta'}{\sqrt{p_{1/}' p_{2/}' p_{/1}' p_{/2}'}}$$

итти не только прямым путем применения одного из описанных выше методов, но также отправляться от общих формул для Er_{11}', $\sigma_{r_{1/1}}'$ и т. д. Достаточно лишь подставить в общие формулы те значения m, μ и r, какие получаются в предположении, что обе переменные способны принимать лишь по два разных значения. Ограничиваясь теми величинами μ и r, какие входят в нужные для нас формулы, находим без труда

$$\mu_{2/0} = p_{1/} p_2 \,[x_1 - x_2]^2, \quad \mu_{0,2} = p_{\,1} p_{/2}\,[y_1 - y_2]^2,$$

$$r_{1'1} = \frac{\delta}{\sqrt{p_{1/} p_{2'} p_{\,1} p_{/2}}}.$$

$$r_{4,0} = \frac{1}{p_{1,}\, p_{2'}} - 3, \quad r_{0'1} = \frac{1}{p_{'1}\, p_{/2}} - 3.$$

$$r_{3/1} = r_{4,0}\, r_{1/1}; \quad r_{1,3} = r_{0\,4}\, r_{1,1},$$

$$r_{2/2} = \frac{(p_{1/} - p_{2'})\,(p_{/1} - p_{/2})\,\delta}{p_{1,}\, p_{2'} p_{/1} p_{\,2}} + 1 = r_{1/1}\,\frac{(p_{1/} - p_{2/})\,(p_{/1} - p_{/2})}{\sqrt{p_{1'} p_{2/} p_{'1} p_{'2}}} + 1.$$

По подстановке в общие формулы для $Er_{1/1}'$ и $\sigma_{r_{1/1}'}$ (см. выше § IV, 3), A)) получаем

$$E\frac{\delta'}{\sqrt{p_{1/}'\, p_{2/}'\, p_{/1}'\, p_{/2}'}} = \frac{\delta}{\sqrt{p_{1/} p_{2/} p_{/1} p_{/2}}} \times$$

$$\left\{1 + \frac{1}{N}\left[1 + \frac{(p_{1/} - p_{2/})\,(p_{/1} - p_{/2})\,\delta - \frac{1}{2}\,(p_{1,}\, p_{2,} + p_{/1}\, p_{/2})}{4\, p_{1/} p_{2'} p_{/1} p_{/2}}\right] + \cdots\right\};$$

$$\sigma^2_{r_{1/1}'} = \frac{1}{N}\left\{\left[1 + \frac{1}{2}\, r_{1/1}^{\,2}\right]\left[1 + \frac{(p_{1/} - p_{2/})\,(p_{/1} - p_{/2})\,\delta}{p_{1/} p_{2/} p_{/1} p_{/2}}\right] - \right.$$

$$\left. - \frac{3}{4}\, r_{1/1}^{\,2}\left[\frac{p_{1/} p_{2'} + p_{'1} p_{/2}}{p_{1/} p_{2/} p_{/1} p_{/2}} - 6\right]\right\} + \cdots.$$

Если $\delta = 0$ и переменные взаимно независимы, то $r_{1'1} = 0$ и $Er_{1/1}' = 0$: систематическая погрешность сводится, в пределах нашего приближения, к нулю; как мы видели (см. выше § IV, 1), B)), она и точно получается равною нулю. Что касается средней квадратической ошибки $r_{1/1}'$, то она в границах нашего приближения при $\delta = 0$ оказывается равною $\frac{1}{\sqrt{N}}$. Точное ее значение, как мы нашли выше (см. § IV, 1), B)), равняется $\frac{1}{\sqrt{N-1}}$.

В пределах нашего приближения систематическая погрешность обращается в нуль также, если все вероятности $p_{1/}$, $p_{2/}$, $p_{/1}$, $p_{/2}$ равны друг другу и равны $1/_2$. Средняя квадратическая ошибка принимает при этом в первом приближении значение

$$\sqrt{\frac{1 - r_{1'1}^{\,2}}{N}} = \sqrt{\frac{1 - 16\delta^2}{N}}.$$

Если $p_1 = p_2' = \frac{1}{2}$, но разность $p_1 - p_2$ отлична от нуля, то систематическая погрешность оказывается отрицательною. Она может быть и положительною, например, в случае, если $\delta > 0$, а обе разности $p_1 - p_2$ и $p_1 - p_2$ положительны и достаточно велики.

Если предположительное значение

$$\frac{\delta}{\sqrt{p_1 p_2 p_1 p_2}}$$

приравнивается случайному значению

$$\frac{\delta'}{\sqrt{p_1' p_2' p_1' p_2'}},$$

то искомая величина может, смотря по виду закона связи, и преувеличиваться и преуменьшаться, а при некоторых условиях такой подход к оценке оказывается свободным от систематической погрешности.

§ V.

Сопоставленные выше формулы для математических ожиданий тех функций случайных значений переменных, которые служат обычно для оценки важнейших сводных характеристик закона связи между переменными, свидетельствуют, что почти во всех случаях оценка сопряжена с известною систематическою погрешностью: иногда мы систематически преувеличиваем искомую величину, иногда систематически преуменьшаем ее; сравнительно редко систематическая погрешность сводится к нулю, и степень надежности нашей оценки определяется исключительно размерами средней квадратической ошибки.

С точки зрения практической важен, разумеется, не столько самый факт наличия некоторой систематической погрешности, сколько размеры ее: если систематическая погрешность незначительна по сравнению с размахом неизбежных случайных колебаний, считаться с нею на практике нет нужды. Заключения, к которым приводит теоретический анализ систематических погрешностей представляются сравнительно благоприятными с этой точки зрения. С ростом числа испытаний систематические погрешности быстро убывают. Лишь при небольших числах испытаний они достигают значительного размера. Но при небольших числах испытаний велик и размах случайных колебаний. При этом систематическая погрешность представляется по большей части величиною порядка $1/N$, тогда как средняя квадратическая ошибка, как правило, — величина порядка $\sqrt{1/N}$: при достаточно большом числе испытаний систематическая погрешность составляет, следовательно, лишь небольшую долю средней квадратической ошибки. В предположении нормальной корреляции получаем, например, для коэффициента корреляции при небольших числах испытаний довольно заметную систематическую погрешность: для того, чтоб можно было с уверенностью почитать *второй* десятичный знак коэффициента корреляции свободным от систематической погреш-

ности, число испытаний должно превышать два десятка. Но средняя квадратическая ошибка при столь малых числах испытаний лишь при очень значительном априорном коэффициенте корреляции не затрогивает *первого* десятичного знака. В первом приближении средняя квадратическая ошибка коэффициента корреляции при нормальной корреляции равняется

$$\frac{1 - r_{1/1}{}^2}{\sqrt{N}},$$

тогда как систематическая погрешность равна

$$\frac{r_{11}\,[1 - r_{11}{}^2]}{2N};$$

отношение систематической погрешности к средней квадратической ошибке равняется

$$\frac{r_{11}}{2\sqrt{N}},$$

при 16 испытаниях оно составляет лишь одну восьмую априорного коэффициента корреляции.

Теоретическое исследование приводит, таким образом, к успокоительному выводу, что, не считаясь на практике при вычислении коэффициентов корреляции с систематическою погрешностью, мы не наносим осязательного ущерба надежности наших выводов. Столь же благоприятно обстоит дело, как правило, и по отношению к иным сводным характеристикам закона связи. Есть, однако, и исключения из этого правила. Систематическая погрешность оценки величины показателя взаимной сопряженности равняется в первом приближении при взаимной независимости переменных (см. выше § IV, 2), C))

$$\frac{(k-1)\,(l-1)}{N};$$

в общей же формуле для рассеяния $[\varphi']^2$ член порядка $1/N$ обращается при стохастической независимости переменных в нуль, так что средняя квадратическая ошибка $[\varphi']^2$ является величиною того же порядка $1/N$, как и систематическая погрешность. Сходным образом систематическая погрешность оценки априорного корреляционного отношения Y к X и средняя квадратическая ошибка эмпирического корреляционного отношения оказываются величинами одного и того же порядка $1/N$ в случае, если переменная Y некоррелирована с X. Систематическая погрешность оценки величины $r_{1/1}{}^2$ по значению $[r_{1/1}']^2$ и средняя квадратическая ошибка $[r_{1/1}']^2$ оказываются также величинами одного порядка $1/N$, если переменные стохастически независимы. При интерпретации малых значений эмпирического корреляционного отношения и эмпирического показателя взаимной сопряженности необходимо, следовательно, соблюдать сугубую осторожность: необхо-

димо не только учитывать размеры средней квадратической ошибки, но и считаться также с размерами систематической погрешности. Равным образом нельзя игнорировать систематическую погрешность при заключении по малой величине разности $[\zeta_{\eta/x}']^2$ о прямолинейности регрессии Y по X.

§ VI.

1) Большая или меньшая теснота стохастической связи Y с X находит себе наиболее явственное выражение в том, насколько сокращается в среднем размах случайных колебаний, которым переменная Y остается подверженною при данных значениях X, по сравнению с полным размахом случайных колебаний Y. Мы судим о тесноте связи Y с X по отношению среднего условного рассеяния $Y — \sum p_i \mu_{i2}^{(i)}$ к полному рассеянию $Y — \mu_{0,2}$ — оценивая априорные величины $\mu_{i2}^{(i)}$ на основании опыта по эмпирическим значениям величин $\mu_{i2}^{(i)'}$. В известной мере представляется, однако, возможным подойти к суждению о тесноте связи на основании данных опыта и минуя величины $\mu_{i2}^{(i)'}$ и $\mu_{i2}^{(i)}$.

Предположим, что число производимых испытаний не превышает числа возможных значений переменной X и что X случайно принимает при всех испытаниях разные значения. Каждому наблюденному значению X соответствует лишь одно значение Y. Ни одно из эмпирических условных рассеяний Y не может быть установлено. Остаются неуловимыми и все априорные рассеяния Y. Можно ли притти к каким-либо заключениям относительно связи Y с X на основании такого ряда наблюдений?

Ясно, с одной стороны, что Y может находиться в функциональной зависимости от X: как бы причудливо ни расположились точки, всегда могут быть без труда указаны такие законы функциональной зависимости, которые дают для установленных наблюдением значений X установленные наблюдением значения Y. Ясно, с другой стороны, что не представляется исключенною и возможность наличия стохастической связи между Y и X: мыслимо, что Y остается случайною переменною и по определении значений X, сохраняя и при данных значениях X способность принимать разные значения с определенными вероятностями; мыслимо, что случайные значения, полученные переменною Y при отдельных испытаниях, более или менее отклоняются от истинных условных математических ожиданий Y, которые в свою очередь располагаются по линии регрессии, более или менее близкой к ломаной размещения точек, установленных наблюдением.

Некоторые ограничения налагаются все же на мыслимый закон связи между переменными тем расположением точек, какое устанавливается наблюдением: при данном размещении точек возможный вид линии регрессии и возможная теснота связи до некоторой степени обусловливают друг друга. Предположение некоррелированности Y с X может, например, делать неправдоподобною значительную тесноту связи, если точки группируются достаточно явственно при меньших значениях X не на той же высоте, как при бо́льших. Равным образом предположение линейной регрессии или регрессии, имеющей вид параболы второй

степени и т. д. могут при тех или иных группировках точек оказываться примиримыми лишь с допущением, что связь не представляется тесною. Порою сопряженные значения X и Y могут настолько сузить рамки сколько-нибудь правдоподобного выбора вида линии регрессии и степени тесноты связи, что получается с достаточною убедительностью вполне определенный вывод. Если, например, при не слишком малом числе испытаний все значения Y оказываются вполне точно пропорциональными сопряженным значениям X, то даже при таком характере материала, когда каждому значению X отвечает лишь одно значение Y, все предположения, кроме допущения функциональной линейной зависимости, представляются в высокой мере неправдоподобными и могут быть отброшены с тем бо́льшей уверенностью, чем больше число испытаний.

Равным образом оказывается исследователь в состоянии притти к более определенным заключениям касательно тесноты связи, если располагает возможностью опереться на определенные предположения касательно вида линии регрессии. О виде же линии регрессии он в состоянии с большей определенностью судить при подобном характере материала, если имеет возможность исходить из достаточно определенных допущений касательно тесноты связи. При обработке измерений, сопряженных с случайными погрешностями, задача обычно ставится именно так, что степень точности измерений почитается более или менее надежно известною и ищется уравнение линии регрессии, с достаточным правдоподобием подходящей к данному расположению точек при данной тесноте связи.

2) Вид линии регрессии Y по X и теснота связи Y с X ни в коей мере, как мы знаем, не обусловливают друг друга (см. глава четвертая, § VI): по совокупности априорных величин $m_{j1}^{(i)}$ мы не можем судить о тесноте связи Y с X; ряд $m_{j1}^{(a)}$ дает нам лишь вполне точное представление о регрессии Y по X. Напротив, совокупность эмпирических величин $m_{j1}^{(i)}$ не дает нам вполне надежного представления об истинном виде регрессии Y по X, но позволяет в некоторой мере судить и о регрессии Y по X, и о тесноте связи Y с X. Задача ставится вполне сходно с тою, которую мы только-что рассматривали: так же, как и единичные значения Y, их средние арифметические $m_{j1}^{(i)'}$ группируются около истинной линии регрессии Y по X тем теснее, чем теснее связь Y с X; при данном ряде значений $m_{j1}^{(i)'}$ не всякие предположения касательно истинного вида линии регрессии Y по X и касательно действительной тесноты связи Y с X представляются совместимыми друг с другом с достаточной степенью правдоподобия. Отсюда, как и выше, вытекает возможность, отбрасывая мало правдоподобные предположения, выработать более или менее надежное, смотря по материалу, суждение касательно вида линии регрессии и тесноты связи. Определенность заключений о тесноте связи повышается, если исследователь располагает более или менее определенными сведениями касательно характера регрессии Y по X. Заключения о виде регрессии принимают, в свою очередь, бо́льшую определенность, если исследователь располагает возможностью исходить из более или менее точных представлений о тесноте связи.

3) На более прочную почву становятся заключения о стохастической связи между переменными, если в основу их полагается не только установленный наблюдением ряд величин $m_{j1}^{(i)'}$, но одновременно также ряд величин $m_{1j}^{(j)'}$. Как мы

видели (см. глава четвертая, § VI), по уравнениям обеих линий регрессии представляется всегда возможным составить некоторое представление о тесноте связи между переменными. Располагая же знанием эмпирических значений $m_1^{(y)'}$ и $m_1^{(x)'}$, мы можем с известной степенью надежности судить о виде обеих линий регрессии. Мы можем вычислить коэффициенты b_1' и b_1' в уравнениях прямых линий, наилучше подходящих к ряду величин $m_1^{(y)'}$ и к ряду величин $m_1^{(x)'}$ (см. глава пятая, § III) и по произведению их, тожественно равному $[r_{1/1}']^2$, можем судить о величине априорного коэффициента корреляции $r_{1\,1}$ между переменными X и Y. Величина же априорного коэффициента корреляции может служить для характеристики тесноты связи, — с оговорками, на которых мы неоднократно останавливались.

Надо, однако, заметить, что ни размеры систематических погрешностей оценки априорных величин по данным опыта, ни размеры средних квадратических ошибок не могут быть найдены по значениям $m_1^{(y)'}$ и $m_1^{(x)'}$ и что суждения об истинном виде линий регрессии, опирающиеся на рассмотрение всей совокупности сопряженных пар значений X и Y, допускают бо́льшую определенность. Даже по сравнению с использованием обеих эмпирических ломаных регрессий для обоснования суждения о связи между переменными представляется все же предпочтительным располагать исходным материалом, сведенным в корреляционную таблицу: лишь планомерная обработка первичных измерений позволяет извлечь из них все то, что данные, доставленные опытом, способны сказать нам о связи между изучаемыми переменными.

§ VII.

Ограничив задачи курса рамками теории корреляции между двумя переменными, я отмечу лишь в заключение, что при изучении стохастической связи между тремя и более случайными переменными основные понятия и основные проблемы остаются те же самые. Однако, привходят и новые понятия, возникают некоторые новые задачи. Кроме условных законов распределения значений, приходится иметь дело с условными законами связи — с условными коэффициентами корреляции, с условными корреляционными отношениями и т. д. На-ряду с линиями регрессии появляются в поле зрения поверхности регрессии и т. п. Чрезвычайно большое значение для статистического изучения стохастической связи имеют понятия кратных и частичных коэффициентов корреляции и корреляционных отношений, не имеющие аналогии в мерах, которые применяются при исследовании связи между двумя случайными переменными. Совокупность всех этих новых проблем, возникающих, когда число стохастически связанных случайных переменных достигает трех, представляет высокий теоретический и практический интерес, но требует так много места для того, чтоб быть изложенною с достаточной точностью и ясностью, что представляется более целесообразным выделить этот круг тем в особое исследование.

ГЛАВА СЕДЬМАЯ.

Стохастические предпосылки числовой характеристики связи.

§ I.

Приемы оценки сводных характеристик априорных законов связи, описанные в главе шестой, покоятся на предположении, что закон связи остается неизменным при всех испытаниях, а испытания представляются взаимно независимыми (см. глава шестая, § III, 1)). Если закон связи не остается неизменным или между испытаниями существует стохастическая связь, то приведенные в главе шестой формулы становятся неприложимыми.

Допустим, что закон связи остается неизменным, но отдельные испытания находятся в такой связи друг с другом, какая отвечает схеме извлечений из урны „без возвращения“. Пусть в урне находится A билетов, помеченных каждый двумя числами; одно из чисел, написанное черными чернилами, дает значение переменной X; другое число, написанное красными чернилами, дает значение переменной Y. Из урны производится N извлечений. Если бы извлечения производились в таком порядке, что вынутый билет опускался обратно в урну до производства следующего извлечения, то испытания представлялись бы взаимно независимыми, и к оценке сводных характеристик закона связи между числами на билетах в урне по числам на вынувшихся билетах можно было бы итти путями, указанными в главе шестой; в частности, математическое ожидание эмпирического коэффициента корреляции и его рассеяние определялись бы формулами параграфа IV, 3), A):

$$Er_{1/1}' = r_{1/1}' + \frac{1}{N}\left\{\frac{1}{4}r_{2/2}r_{1/1} + \frac{3}{8}r_{1/1}[r_{4/0} + r_{0/4}] - \frac{1}{2}[r_{3/1} + r_{1/3}]\right\} + \cdots$$

$$\sigma^2 r_{1/1}' = \frac{1}{N}\left\{r_{2/2}\left[1 + \frac{1}{2}r_{1/1}^2\right] - r_{1/1}[r_{3/1} + r_{1/3}] + \frac{1}{4}r_{1/1}^2[r_{4/0} + r_{0/4}]\right\} + \cdots.$$

Но если вынутые билеты не возвращаются в урну, то математическое ожидание и рассеяние эмпирического коэффициента корреляции между числами на

вынувшихся билетах уже не могут быть найдены по этим формулам, а должны быть вычислены самостоятельно.

К вычислению их можно итти теми же путями, как в случае взаимно независимых испытаний. Так же, как в главе шестой, мы начинаем с разложения $r_{1\,1}'$ в ряд по степеням разностей $dp_{1/}'$, $dp_{/}'$ и т. д. или разностей $dm_{1\,0}'$, $d\mu_{1\,0}''$ и т. д.: так же, как там, переходим затем к математическим ожиданиям. Но математические ожидания разных степеней разностей $dp_{1/}'$, $dp_{/}'$, $dm_{1,0}'$, $d\mu_{1\,0}''$ и т. д. имеют при наличии связи между испытаниями не те значения, как при взаимной независимости испытаний. Подставляя значения математических ожиданий разных степеней разностей, какие получаются, если связь между испытаниями отвечает схеме невозвращаемого билета, находим:

$$Er_{1\,1}' = r_{1/1} + \frac{A-N}{A-1}\frac{1}{N}\left\{\frac{1}{4}r_{2/2}r_{1/1} + \frac{3}{8}r_{1\,1}[r_{4\,0}+r_{0/4}] - \frac{1}{2}[r_{3\,1}+r_{1\,3}]\right\} + \cdots$$

$$\sigma^2 r_{1/1}' = \frac{A-N}{A-1}\frac{1}{N}\left\{r_{2\,2}\left[1+\frac{1}{2}r_{1/1}^{\,2}\right] - r_{1/1}[r_{3/1}+r_{1\,3}] + \frac{1}{4}r_{1\,1}^{\,2}[r_{1/0}+r_{0/4}]\right\} + \cdots.$$

Мы видим, что систематическая погрешность оценки, определяемая разностью $Er_{1\,1}' - r_{1/1}$, и рассеяние $r_{1/1}'$ уменьшаются при наличии подобной связи между испытаниями по сравнению со случаем взаимно независимых испытаний в первом приближении в $\frac{A-N}{A-1}$ раз: математическое ожидание $r_{1/1}'$ ближе подходит тут к искомой априорной величине $r_{1/1}$, сокращается и размах случайных колебаний, — схема невозвращаемого билета с большею меткостью нацеливает $r_{1/1}'$ в точку, лежащую ближе к той, куда желательно попасть.

Предположим, с другой стороны, что из урны производятся извлечения в таком порядке: вынутый билет опускается в урну, прежде чем производится следующее извлечение, а вместе с ним бросается в урну еще один билет, на котором черными и красными чернилами написаны те же числа, как на вынувшемся билете, — „извлечение с прикладыванием“, буду я обозначать такую постановку эксперимента. При наличии такой связи между испытаниями к оценке сводных характеристик априорного закона связи между числами на билетах в урне по числам на вынувшихся билетах нельзя итти ни тем, ни другим из описанных выше путей. Разложение $r_{1/1}'$ в ряд по степеням разностей $dp_{1/}'$, $dp_{/}'$ и т. д. или разностей $dm_{1/0}'$, $d\mu_{1/0}''$ и т. д. сохраняет и тут свой прежний вид, но математические ожидания разных степеней разностей складываются снова по иному. Подставляя их, находим

$$Er_{1/1}' = r_{1/1} + \frac{A+N}{A+1}\frac{1}{N}\left\{\frac{1}{4}r_{2/2}r_{1|1} + \frac{3}{8}r_{1/1}[r_{4/0}+r_{0/4}] - \frac{1}{2}[r_{3\,1}+r_{1/3}]\right\} + \cdots$$

$$\sigma^2 r_{1/1}' = \frac{A+N}{A+1}\frac{1}{N}\left\{r_{2/2}\left[1+\frac{1}{2}r_{1/1}^{\,2}\right] - r_{1|1}[r_{3/1}+r_{1/3}] + \frac{1}{4}r_{1/1}^{\,2}[r_{4/0}+r_{0/4}]\right\} + \cdots.$$

В отличие от схемы невозвращаемого билета схема прикладываемого билета повышает и систематическую погрешность оценки и рассеяние $r_{1/1}'$ по сравнению

со случаем взаимно независимых испытаний. Если испытания связаны так, как отвечает схеме прикладываемого билета, то, приравнивая предположительное значение r_{11} тому случайному значению эмпирического коэффициента корреляции, какое мы находим по числам на вынувшихся билетах, мы с меньшею меткостью нацеливаемся в точку, лежащую дальше от той, куда нам желательно попасть.

Если внести иные предположения касательно связи между испытаниями или предположить, что закон связи изменяется определенным образом от испытания к испытанию, то необходимо будет формулы, на которые опирается оценка априорного коэффициента корреляции по данным опыта, подвергнуть новым преобразованиям.

Равным образом и формулы, служащие для оценки всех иных сводных характеристик закона связи по данным опыта, складываются существенно различно в зависимости от того, представляются ли испытания взаимно независимыми или связаны так или иначе, и остается ли закон связи между переменными неизменным или изменяется так или иначе от испытания к испытанию.

§ II.

Невозможность итти всегда одним и тем же путем от данных опыта к оценке сводных характеристик стохастической связи между переменными ставит перед статистиком вопрос об основаниях, какими следует руководствоваться при выборе применяемых при оценке формул. Тот же вопрос возникает, когда статистик имеет дело с одной случайной переменной и стремится составить по данным опыта представление о сводных характеристиках закона распределения ее значений. Если мы по рассеянию красных номеров на билетах, вынувшихся из урны, желаем судить о рассеянии красных номеров на билетах в урне, то при взаимной независимости испытаний — схема возвращаемого билета — мы исходим из соотношения

$$E\mu_{0/2}' = \frac{N-1}{N}\mu_{0/2}.$$

Если же испытания находятся в связи, отвечающей схеме невозвращаемого билета, то вместо этой формулы мы должны опираться на иную:

$$E\mu_{0/2}' = \frac{A}{A-1}\,\frac{N-1}{N}\mu_{0/2}.$$

А при связи, отвечающей схеме прикладываемого билета, мы исходим из формулы:

$$E\mu_{0/2}' = \frac{A}{A+1}\,\frac{N-1}{N}\mu_{0/2}.$$

Чтоб оценить размеры априорного рассеяния недостаточно, таким образом, знать, какие значения получены переменною при ряде произведенных над нею испытаний. Необходимо, сверх того, располагать некоторыми сведениями касательно постановки опыта: иначе нет возможности обоснованно решить, преувеличивается или преуменьшается априорное рассеяние, когда его предположительное значение приравнивается рассеянию вынувшихся из урны чисел.

Чтоб выяснить, из каких стохастических предпосылок надлежит исходить статистик, имеющий дело с одной случайной переменной, обычно прибегает к вычислению коэффициента дисперсии. Прием этот был, как известно, введен в употребление Лексисом. В первоначальной постановке задачи у Лексиса границы его приложимости были очерчены сравнительно узко условием, чтоб рассматриваемый статистический ряд имел внешнюю форму частостей (или известных исследователю функций частостей). Борткевичем Лексисовское построение было затем распространено на общий случай любого закона распределения значений переменной. Сущность приема сводится к тому, что конструируется особый критерий, — коэффициент дисперсии, — получающий значение 1, если закон распределения остается неизменным, а испытания представляются взаимно независимыми, — если устойчивость „нормальна", по терминологии Лексиса. Если выкладки приводят к значению коэффициента дисперсии, сильнее отклоняющемуся от единицы, нежели примиримо с размерами его средней квадратической ошибки, то приходится признать, что стохастические предпосылки не отвечают понятию нормальной устойчивости, т. е. или закон распределения значений не остается неизменным, или испытания так или иначе связаны, или имеет место и то и другое уклонение от условий нормальной устойчивости. В случае же, если коэффициент дисперсии получается достаточно близкий к единице, можно, с некоторыми оговорками, принимать, что устойчивость нормальна, что закон распределения значений неизменен и испытания взаимно независимы.

Тот же прием может быть привлечен к выяснению характера стохастических предпосылок и в случае двух или более стохастически связанных случайных переменных. Условимся называть стохастическую связь между переменными *нормально устойчивою*, если соблюдены требования неизменности закона связи от испытания к испытанию и взаимной независимости испытаний. Не трудно и для этого случая построить коэффициент дисперсии, вполне аналогичный и по структуре, и по тому смыслу, какой применяется при исследовании законов распределения значений одной случайной переменной. Если коэффициент дисперсии сильно отклоняется от 1, стохастические предпосылки нормальной устойчивости должны почитаться не соответствующими действительности: либо закон связи не остается неизменным, либо испытания находятся в связи друг с другом, либо и закон связи меняется и испытания не представляются взаимно независимыми. Если коэффициент дисперсии достаточно близок к 1, то — с теми же оговорками, как в случае одной переменной — можно принимать, что стохастические предпосылки отвечают понятию нормальной устойчивости, т.-е., что закон связи остается неизменным от испытания к испытанию, а испытания взаимно независимы.

§ III.

Сохраняя наши обычные обозначения и замечая, что при неизменном законе связи и взаимно независимых испытаниях

$$E\left[x^{[J]'} - m_{1/0}\right]^c \left[y^{[J]'} - m_{0/1}\right]^d = \mu_{c/d},$$

$$E\left[x^{[J]'} - m_{1/0}\right]^c \left[y^{[g]'} - m_{0/1}\right]^d = \left\{E\left[x^{[J]'} - m_{1/0}\right]^c\right\}\left\{E\left[y^{[g]'} - m_{0/1}\right]^d\right\} = \mu_{c/0}\,\mu_{0/d},$$

находим без труда

$$E\frac{1}{N}\sum_{f=1}^{N}[x^{[f]'}-m_{1/0}]\,[y^{[f]'}-m_{0\,1}]=\mu_{1\,1},$$

$$E\frac{1}{N-1}\sum_{f=1}^{N}[x^{[f]'}-x_0']\,[y^{[f]'}-y_0']=\mu_{1/1}.$$

Предположим, что N испытаний, произведенных над переменными, разбиваются на r серий по n испытаний в каждой серии. Обозначим через $x_0^{[h]'}$ и $y_0^{[h]'}$ средние арифметические случайных значений, полученных переменными x и y при испытаниях h-ой серии.

Рассматривая все испытания, как одно целое, имеем:

$$E\frac{1}{rn-1}\sum_{f=1}^{rn}[x^{[f]'}-x_0']\,[y^{[f]'}-y_0']=\mu_{1/1}.$$

С другой стороны,

$$E\frac{1}{r-1}\sum_{h=1}^{r}[x_0^{[h]'}-x_0']\,[y_0^{[h]'}-y_0']=\frac{1}{n}\mu_{1/1}.$$

Положим:

$$Q=\frac{\frac{n}{r-1}\sum_{h=1}^{r}[x_0^{[h]'}-x_0']\,[y_0^{[h]'}-y_0']}{\frac{1}{rn-1}\sum_{f=1}^{rn}[x^{[f]'}-x_0']\,[y^{[f]'}-y_0']}.$$

Нетрудно доказать тем же путем, как для случая одной переменной, что $EQ=1$, если соблюдены условия нормальной устойчивости, т. е. если закон связи остается неизменным, а испытания взаимно независимы.

В случае, если между переменными X и Y существует такого рода связь, что значение, принимаемое Y, всегда совпадает со значением X, построенный нами коэффициент устойчивости обращается в обычный коэффициент устойчивости для одной переменной.

§ IV.

Если устойчивость рассматриваемого ряда пар сопряженных случайных значений переменных X и Y может быть признана нормальною, то закон связи остается неизменным от испытания к испытанию; остаются неизменными и его сводные характеристики. Оценивая по данным опыта их предположительные значения, статистик получает, следовательно, величины, имеющие вполне определенный смысл численных значений вполне определенных мер. Если устойчивость не представляется нормальною, то меняется не только способ подхода к оценке

априорных величин, но также и внутренний смысл тех результатов, к которым приводят вычисления. Необходимо различать при этом две возможности. Устойчивость может не быть нормальною вследствие того, что испытания не представляются взаимно независимыми, хотя закон связи и остается неизменным от испытания к испытанию; примером может служить схема невозвращаемого билета или схема прикладываемого билета. В таких условиях отсутствие нормальной устойчивости вносит трудности в процесс оценки априорных величин по данным опыта, так как приходится подыскивать формулы, соответствующие характеру связи между испытаниями. Но если эти трудности удается преодолеть, то смысл получаемых предположительных значений остается прежним: они остаются предположительными значениями вполне определенных сводных характеристик неизменного закона связи, — численные значения эмпирического коэффициента корреляции между числами на вынувшихся из урны билетах одинаково отображают величину коэффициента корреляции между числами на билетах в урне и при извлечениях с возвращением, и при извлечениях без возвращения, и при извлечениях с прикладыванием; разница лишь в размерах систематических погрешностей и средних квадратических ошибок. Напротив, если устойчивость оказывается ненормальною вследствие того, что закон связи меняется от испытания к испытанию, то получаемые нами предположительные значения априорных величин не характеризуют уже какой-либо один определенный закон связи, а приурочиваются к нелегко уловимой совокупности разнствующих между собою законов связи. Ход выкладок остается во всех случаях один и тот же; эмпирический коэффициент корреляции вычисляется, например, всегда по одной формуле. Но выигрыш в нашем понимании связи между изучаемыми явлениями складывается, несмотря на это, существенно по разному. Те представления, которые обычно связываются с численными значениями сводных характеристик законов распределения и законов связи, — рассеяний, коэффициентов корреляции, корреляционных отношений и т. д. — могут почитаться строго законными лишь в условиях нормальной устойчивости. Иначе то, что ими характеризуется, принимает более или менее расплывчатые очертания. Такая зависимость внутреннего содержания результатов счетной обработки статистического материала от стохастических предпосылок придает выяснению характера стохастических предпосылок еще бо́льшую важность, нежели зависимость от них систематических погрешностей и средних квадратических ошибок наших оценок априорных величин по данным опыта. Статистик всегда должен стремиться к тому, чтоб выработать возможно ясное представление о стохастических предпосылках, вдумываясь внимательно в условия, в которых протекают изучаемые явления, и привлекая на помощь, когда это требуется, вычисление коэффициентов дисперсии. При изучении стохастической связи между случайными переменными задача эта имеет не меньшую важность, чем в случае, когда обработке подвергаются эмпирические численные значения одной случайной переменной.

ГЛАВА ВОСЬМАЯ.

Задачи и выгоды числовой характеристики связи.

§ 1.

1) Что выигрывает исследователь, вычисляя корреляционные отношения, коэффициенты корреляции и т. д.? В чем преимущества так называемых „математических“ приемов исследования стохастической связи между переменными пред „элементарными“ приемами статистиков-не-математиков?

Прежде всего: в бо́льшей определенности опирающихся на них суждений. И тот исследователь, который не прибегает к вычислению коэффициентов корреляции и прочих сводных характеристик закона связи, не обходится без суждений о бо́льшей или меньшей тесноте стохастической связи, о характере регрессии и т. п. Но суждения его страдают расплывчатостью и носят ярко выраженную субъективную окраску. Не производя выкладок, можно подметить, что некоторые из рядов обнаруживают заметный „параллелизм“ колебаний, тогда как другие ряды сильно расходятся. Лица, обладающие острым глазомером, способны оценивать степень параллелизма двух рядов на глаз с довольно значительной точностью. В моем семинаре процветал одно время своеобразный статистический спорт: некоторые из участников семинара упражнялись в глазомерной оценке коэффициентов корреляции между рядами по графическому изображению: те, у кого были хорошие задатки, дошли до почти безошибочной оценки первого десятичного знака коэффициента корреляции даже в тех случаях, когда картина представлялась довольно сбивчивою. Но даже когда глазомерные суждения близко отвечают истине, они не допускают ни обоснования, ни проверки. Возможность облечь составление суждения в формы, обеспечивающие получение тожественного результата всеми исследователями, имеет и тут те же выгоды, как и в иных подобных случаях. Нет, ведь, необходимости производить точные вычисления и средней арифметической ряда. Человек, набивший руку, скажет без вычислений, что средний уровень одного ряда выше, другого — ниже, а при некоторой сноровке оценит даже довольно верно и разницу уровней. Сходным образом размах колебаний ряда кругом его средней может быть схвачен глазом без вычисления

средних квадратических ошибок или иных мер устойчивости ряда: внимательно вглядываясь в движение чисел, мы можем составить более или менее отчетливое представление о том, где колебания больше, где меньше. Но все подобные утверждения неизбежно остаются недоказуемыми даже в тех случаях, когда лицо, высказывающее суждение, твердо убеждено в своей правоте. Сплошь и рядом два исследователя, всматривающиеся в одни и те же числа, будут приходить к противоречивым заключениям, и каждый из них будет стоять на своем, утверждая, что его оценка правильна, иная же оценка субъективно окрашена и не соответствует действительной картине. Решить, кто прав, можно лишь при помощи „математических“ приемов. Американский статистик Митчель [1]) в известном исследовании о хозяйственных циклах перерабатывает весьма обширный статистический материал, при чем не пользуется „математическими“ приемами, а опирает свои суждения о размахе колебаний рядов и о степени тесноты связи между ними на рассмотрение графиков. Другой американский статистик, Кинг [2]) дал себе труд проверить суждения, к которым приходит Митчель, путем вычисления средних квадратических ошибок и коэффициентов корреляции. В большинстве случаев Митчель оказался прав, но встречаются и промахи: при этом ошибки в оценке сравнительной тесноты связи делаются чаще, чем ошибки в оценке сравнительного размаха колебаний.

2) Существенным недостатком „элементарных“ приемов работы представляется затруднительность дать себе отчет в том, какая именно из мер тесноты связи лежит в основе глазомерного суждения. Статистик-не-математик, всматриваясь в ряды чисел, приходит к известным заключениям относительно того, где теснота связи выше, где ниже. Ни сам он, ни критик, пытающийся проверить его суждения, не в состоянии с уверенностью определить, имеются ли при этом в виду те свойства связи, которые измеряются величиною коэффициента корреляции, или те, для измерения которых служит корреляционное отношение, или еще какие-либо. Статистик же математик отчетливо выделяет разные меры и разные свойства связи, ими измеряемые, и из сопоставления численных значений разных мер почерпает ценные заключения о характере взаимоотношений между изучаемыми явлениями. Если, например, численные значения коэффициента корреляции и корреляционного отношения расходятся более, чем примиримо с размерами средней квадратической ошибки их разности, то исследователь знает, что регрессия не может быть прямолинейною. В тех случаях, когда численные значения избранных мер допускают наглядное истолкование, установление их представляется само по себе ценным научным достижением. Меры тесноты стохастической связи обычно конструируются так, чтоб определенное значение — по большей части, значение 1 — отвечало предельно тесной связи, а другое определенное значение — по большей части, значение 0 — отвечало независимости переменных. Скалу промежуточных значений стараются градуировать так, чтоб ее ступени возможно близко передавали нарастание тесноты связи. Надо, правда,

---

1) W. C. Mitchell. Business Cycles (Memoirs of the University of California, vol. 3); 1913.
2) B. W. King. A study of Mitchell's inquiries into prices (Quarterly Journal of Economics, vol. 31; 1917)

сказать, что представления изобретателей новых мер о том, что именно при этом градуируется, не всегда отличаются полной ясностью, а нередко остаются смутными и понятия независимости и предельно тесной связи, которыми они оперируют. Но даже и в тех случаях, когда представляется сомнительным, чтоб градусы скалы были действительно строго пропорциональны нарастанию измеряемого свойства связи, вычисления все же сохраняют по большей части тот смысл, что бо́льшим значениям меры соответствует высшая степень интенсивности измеряемого свойства. Что подобного рода измерения не лишены научного интереса, подтверждает практика научной работы во всех областях знания. Что, собственно, увеличивается ровно вдвое, когда температура поднимается с 5 градусов Цельсия до 10? Или, быть может, что-либо изменяется при этом в отношении 41 к 50, к которому приводит измерение в градусах Фаренгейта?

3) Независимо от возможности наглядного истолкования численных значений сводных характеристик законов распределения и связи знание их представляет ту выгоду, что облегчает оценку мыслимого размаха случайных колебаний. Уже при решении вопроса, есть ли какая-либо связь между переменными, или они представляются стохастически независимыми, возможность привлечь к рассмотрению численные значения целесообразно избранных мер дает ценную точку опоры. Сопряженные значения, принимаемые переменными при производимых над ними испытаниях, могут случайно и при взаимной независимости переменных являть видимость некоторой связи. Статистик-математик устанавливает в таких случах разность между численным значением избранной меры, вычисленным по данным опыта, и тем ее численным значением, которое отвечает предположению стохастической независимости, и сопоставляет величину разности с величиною ее средней квадратической ошибки. Таким путем он в состоянии выяснить, представляется ли наблюдаемая связь настолько ярко выраженною, чтоб можно было с достаточной уверенностью считать ее несовместимою с допущением взаимной независимости переменных, или же наблюдаемый параллелизм численных значений переменных с достаточным вероятием может быть сводим к случайным колебаниям чисел. Для статистика-не-математика такой подход к решению вопроса закрыт. Он вынужден искать окольных путей, прибегать к сложным построениям, требующим не малой затраты труда и не малой изобретательности и все же редко приводящим к вполне убедительным заключениям. С теми же трудностями приходится иметь дело при решении вопроса, не сводятся ли видимые различия в тесноте связи к воздействию случайностей, будучи не настолько значительными, чтоб представляться неправдоподобными при отсутствии действительной разницы в тесноте связи между рассматриваемыми переменными. И тут математическим приемам свойственны те же преимущества по сравнению с „элементарными“ методами исследования.

4) Сопоставим, далее, элементарные приемы исследования, равнозначные вычислению эмпирической ломаной регрессии (см. глава первая, § II, A)), с совокупностью соответственных математических приемов. Если вид регрессии не слишком запутан, статистик-не-математик, бесспорно, приходит также к более или менее правильным заключениям не только касательно того, замечается ли в целом нарастание или убывание значений одной переменной с ростом другой,

но и касательно среднего темпа нарастания или убывания. Однако, и понятия, которыми он оперирует, и — в еще бо́льшей мере — представления его о предпосылках, лежащих в основе его суждений, остаются смутными и недодуманными, а его количественные суждения страдают неточностью и субъективизмом. Кроме того, для него представляет почти непреодолимые трудности вопрос о том, где лежит грань между подлинными извивами линии регрессии и зыбью чисто случайных колебаний, отклоняющих ее в тех или иных точках вверх и вниз, но не нарушающих общего характера регрессии. Статистик-математик располагает, напротив, возможностью поставить и тут решение на сравнительно прочную основу учета средних квадратических ошибок. Он в состоянии прикинуть, например, с какою степенью правдоподобия можно принимать, что истинная линия регрессии — прямая, а наблюдающиеся зигзаги идут на счет случайных колебаний. Уравнение регрессии позволяет ему предвычислять для данных значений X условные математические ожидания Y; величина же корреляционного отношения Y к X дает представление о степени надежности, с которою можно строить предположения касательно значений, которые примет Y, на основе знания величины условных математических ожиданий Y. Если исследование охватывает несколько переменных, то по величине корреляционного отношения T к X, к Y, к Z и т. д. исследователь в состоянии судить, какие из рассматриваемых факторов налагают печать на значения T с бо́льшею силою; вычисляя же корреляционное отношение T к X, Y, Z, вместе взятым, он устанавливает, в какой мере на значениях T сказывается влияние факторов, им еще не принятых во внимание. Если корреляционное отношение T к X, Y, Z может быть принято равным единице, то он заключает, что от совокупности переменных X, Y, Z переменная T находится в функциональной зависимости и что, следовательно, им открыты все те факторы, которые сказываются на величине Y. Если корреляционное отношение T к X, Y, Z не равно единице в точности, но близко к единице, он убеждается, что напал на след тех факторов, которые оказывают на T наибольшее воздействие. Для многих областей научной работы такое одновременное вычисление уравнений регрессии и корреляционных отношений представляет наиболее рациональный подход к решению задачи. В частности проблема прогноза, опирающегося на знание стохастических связей между явлениями, — например, проблема предучета движений конъюнктуры — лишь таким путем может быть выведена на высоту научных изысканий из низин обывательского полу-инстинктивного гадания.

§ II.

1) С особою ясностью превосходство „математических“ приемов над „элементарными“ обнаруживается, если дело не ограничивается вычислением разрозненных единичных характеристик связи, а производятся систематические измерения по обдуманному плану. Статистик-не-математик, полагающийся в своих суждениях на глазомер, редко способен пойти далее утверждения, что в некоторых случаях наблюдается прямая связь, а в других связь обратная. Иные различия в характере связи ускользают от него. Если же производятся планомерные измерения, то результаты их могут быть подвергнуты дальнейшей научной обра-

ботке. Вычисляются значения избранной меры для рядов статистических чисел, устанавливаемых наблюдением в различных условиях: расхождения или совпадения их приводят к заключениям, которые проливают свет на существо взаимоотношений между изучаемыми явлениями и позволяют глубже вникнуть в истинный характер связи. Пусть, например, ряд вычисленных коэффициентов корреляции обнаруживает закономерную картину распределения в пространстве или во времени: крупные значения коэффициентов корреляции приурочиваются на карте к отчетливо очерченным границам, так что могут быть намечены зоны более и менее тесной стохастической связи между изучаемыми явлениями в форме картограмм или, по образцу метеорологов, в виде линий, подобных изотермам и изобарам; хронологические ряды коэффициентов корреляции обнаруживают правильное движение, нарастание или убывание и т. п. Каждая подобная закономерность ставит исследователю новую загадку, требует напряжения мысли, а порою и кропотливых дополнительных изысканий для ответа. Но затраты труда на разгадку окупаются сторицею. В своей совокупности такие загадки и отгадки дают незаменимую точку опоры для выполнения наиболее ответственной задачи статистического исследования стохастической связи — для выяснения действительного смысла подмеченной связи между явлениями (см. глава вторая, 2)).

Присмотримся ближе на ряде иллюстраций, заимствуемых мною из работы М. М. Виноградовой „Потребление водки в России и урожай“ [1]), к тому, как могут быть используемы систематические измерения тесноты связи в целях обогащения нашего понимания действительного характера взаимоотношений между изучаемыми явлениями.

2) Вопрос о влиянии урожая на потребление водки издавна привлекает к себе внимание русских статистиков. До сравнительно недавнего времени господствовало единодушное убеждение, что высота урожая налагает ясную печать на уровень потребления водки в России: чем обильнее сбор хлебов, тем больше выпивается водки; в годы плохих урожаев потребление водки сжимается. Связующим звеном является, по общему представлению, уровень хозяйственного благосостояния населения: хороший урожай поднимает покупательную способность земледельцев, сказывается и на положении промышленности повышенным спросом деревни на ее изделия; располагая в годы хороших урожаев большими средствами, население больше уделяет и на покупку спиртных напитков. Этим прочно укоренившимся у нас воззрениям В. К. Дмитриев [2]) противопоставил диаметрально противоположный взгляд на условия, определяющие уровень потребления водки в России. „Решающим моментом, определяющим у нас уровень потребления в стране алкоголя, является не „Господин Урожай“, а уж если употреблять образное выражение, „Господин Капитал“... как бы при этом ни складывались прочие обстоятельства, в том числе и результаты урожая“. При этом, по мнению Дмитриева, колебания промышленности действуют на потребление алко-

---

[1]) М. М. Виноградова. Потребление водки в России и урожай (Труды студентов экономического отделения Петроградского Политехнического Института Императора Петра Великого, № 17). Петроград, 1916.

[2]) В. К. Дмитриев, Критические исследования о потреблении алкоголя в России. Москва, 1911.

голя не через посредство влияния ее на благосостояние населения: промышленный подъем вызывает рост фабрично-заводского и городского пролетариата за счет крестьян-земледельцев; наоборот, всякая остановка в поступательном движении промышленности влечет за собою уменьшение, если не абсолютной, то по крайней мере относительной численности промышленного пролетариата. Потребитель индустриально-городского типа пьет в среднем больше, чем крестьянин, еще не порвавший связи с землей. Поэтому колебания среднего душевого потребления водки в стране идут параллельно изменению относительной численности группы капиталистического пролетариата, и, следовательно, отражают на себе состояние промышленности в каждый данный момент. Объясняется же высота потребления алкоголя представителями этой группы не тем, что они располагают большими средствами, а тем, что в их бюджете расход на алкоголь является одною из важнейших расходных статей. В условиях жизни индустриального пролетариата водка делается почти предметом первой необходимости, так как она дает средство заглушить те тягостные ощущения, которые влечет за собою переутомление, плохое питание, неуверенность в завтрашнем дне и т. п.

Дмитриев настолько убежден в правильности своего построения, что мнение, будто наблюдается тесная связь между урожаем и потреблением водки, относит к области „легенд“: он не находит никакого соответствия между их колебаниями за рассматриваемые им годы. „После тщательного сопоставления данных о душевом потреблении спиртных напитков за почти сорокалетний период с данными об изменении урожаев“, он приходит к выводу, что „за весь рассматриваемый период не было ни одного случая повышения или понижения потребления алкоголя, которое бы с уверенностью (или хотя бы с большой степенью вероятности) могло быть отнесено к следствиям колебаний урожаев“. Отвергает Дмитриев и распространенный взгляд, будто осенний максимум потребления водки, особенно ясно выраженный в губерниях чисто земледельческих, объясняется обилием в это время свободных средств у крестьян. По мнению Дмитриева, дело тут в переутомлении после тяжелой для земледельца страдной поры. Он вообще полагает, что „значительная часть колебаний душевого потребления, которая обыкновенно относится на счет изменения покупательных средств населения под влиянием тех или иных экономических факторов, на самом деле является результатом прямого воздействия этих факторов на психическое настроение потребителей“. Выдающиеся урожаи или неурожаи, подобно экстренным сложениям недоимок, объявлению войны и заключению мира, действуют так или иначе на настроение потребителей, увеличивая или уменьшая их влечение к алкоголю.

Сопоставляемые Дмитриевым ряды — погодные данные о сборе хлебов и о потреблении водки в 50 губерниях Европейской России за время до введения казенной продажи питей — действительно не обнаруживают связи между колебаниями урожая и движением потребления водки. Примененный им прием исследования вызывает, однако, существенные возражения. Потребление водки в течение календарного года стоит под действием двух урожаев: на потреблении последних месяцев года может сказываться урожай того же года, на потреблении же с января по осень, если и сказывается урожай, то, разумеется, лишь урожай прошлого года. Если урожаи смежных лет не сходятся по высоте, то действие

их будет взаимно нейтрализоваться, и ряд чисел, выражающих потребление водки, не будет обнаруживать ясного параллелизма с высотою урожая, даже если связь между потреблением водки и урожаем представляется на самом деле очень тесною. Сходным образом может сказываться на результатах сопоставлений и объединение в одно целое таких разнородных слагаемых, как разные части обширной территории 50 губерний Европейской России. Годы хороших урожаев в одних губерниях бывают нередко годами плохих урожаев в других; при этом вес, с каким входят разные части Европейской России в общую сумму сбора хлебов и в общий итог потребления водки, различен. В результате опять-таки в огульных сопоставлениях легко может стушевываться до неприметности даже сильная связь, существующая в действительности между урожаем и потреблением водки в отдельных, более однородных, частях этого пестрого целого. Чтоб решить путем статистических сопоставлений, сказывается ли урожай на потреблении водки, необходимо сопоставлять данные по менее обширным, возможно однородным районам и брать при этом потребление водки, по возможности, за такие периоды времени, которые стоят под действием одного урожая или, если и испытывают перекрещивающее влияние урожая смежного года, то в возможно ослабленной степени. Статистические материалы за время казенной продажи питей дают возможность осуществить такие сопоставления: за эти годы имеются помесячные данные о потреблении водки по отдельным губерниям (отчасти, даже по отдельным уездам). Есть, следовательно, возможность проследить, как сказывается урожай на потреблении водки отдельных месяцев; есть и возможность подойти к задаче путем вычисления потребления за нарочито образуемые урожайные или сельско-хозяйственные годы. Само по себе такое усовершенствование метода исследования может и не быть сопряжено с переходом от „элементарных“ к „математическим“ приемам работы. Систематические измерения тесноты связи между потреблением водки в отдельных губерниях за разные месяцы со сбором разных хлебов в них позволяют, однако, с такою определенностью разъяснить поднятые Дмитриевым спорные вопросы, какая едва ли могла бы быть достигнута иначе. Результаты исследования Виноградовой рельефно иллюстрируют поучительность анализа закономерностей, наблюдающихся в рядах сводных характеристик связи, вычисляемых по тщательно продуманному плану.

Исследование Виноградовой охватывает 19 губерний. Уже географическое распределение полученных для них коэффициентов корреляции между размерами урожая и потреблением водки за сельско-хозяйственный год обнаруживает весьма характерные правильности: коэффициент корреляции достигает максимальной высоты в Самарской губернии, где он почти равняется 1 ($r = +0{,}98$); полоса очень высоких коэффициентов корреляции включает, кроме Самарской, также губернии Оренбургскую, Уфимскую, Херсонскую, Екатеринославскую, Полтавскую, Киевскую, Подольскую. С нею смыкается полоса значительных, но не столь высоких коэффициентов корреляции: губернии Пермская, Харьковская, Черниговская, Волынская. При дальнейшем продвижении на север, мы переходим к коэффициентам корреляции сравнительно незначительным и притом то положительным, то отрицательным. Лишь Олонецкая губерния выделяется высоким коэффициентом корреляции, равным $+0{,}7$; объясняется это, повидимому, осо-

бымп условиями периода 1904—07 гг. (годы войны и революционного движения). Еще более выразительную картину обнаруживают коэффициенты корреляции между урожаем и потреблением водки за отдельные месяцы года по тем же губерниям. Полоса высоких коэффициентов корреляции характеризуется тем, что в ней тесная связь между урожаем и потреблением водки наблюдается в течение всего сельско-хозяйственного года: в Самарской губернии, например, ни один из коэффициентов корреляции за время с октября по сентябрь следующего года не спускается ниже + 0,5. В губерниях, где потребление водки за сельско-хозяйственный год не так тесно связано с урожаем, выделяются при расчленении года на отдельные месяцы более или менее длительные периоды, в течение которых связь потребления водки с урожаем характеризуется высокими коэффициентами корреляции. Даже в тех губерниях, где весь сельско-хозяйственный год в целом не обнаруживает заметной связи или дает отрицательный коэффициент корреляции, при переходе к отдельным месяцам подмечается для некоторых из осенних месяцев довольно тесная прямая связь. Высота урожая сказывается, таким образом, в различных местностях по разному на потреблении водки: местами урожай оказывает длительное влияние, отражаясь на потреблении водки в течение целого года вплоть до следующего урожая; местами он налагает печать на потребление водки в течение более или менее длинного ряда месяцев, следующих за сбором хлебов; местами действие его, сказываясь в начале сельско-хозяйственного года, быстро затем ослабевает. Заметны, кроме того, и различия во времени, с которого урожай начинает оказывать влияние — в Самарской и в Херсонской губерниях август и сентябрь дают уже значительные коэффициенты корреляции с урожаем того же года, тогда как в Харьковской и в Полтавской губерниях коэффициент корреляции достигает тех же размеров лишь к ноябрю. Нигде при этом не наблюдается особенно тесной связи в те месяцы, когда происходит уборка урожая; повсюду максимальные значения коэффициентов корреляции получаются для месяцев более поздних: так, например, в Самарской губернии коэффициент корреляции не спускается с октября по апрель следующего года ниже $^4/_5$, в июле же, августе и сентябре он держится около $^1/_2$. Эта закономерность имеет решающее значение для оценки гипотезы Дмитриева, связывающей рост потребления водки в урожайные годы с обостренным переутомлением во время уборки: ясно, что дело не в этом, раз влияние урожая сказывается с особою силою не во время уборки, а через несколько недель и даже месяцев. Отпадает и попытка свести влияние урожая к непосредственному воздействию радостной картины богатого сбора на настроение: ясно, что пьют больше водки не от избытка радостных чувств, раз особенное усиление потребления замечается не в те месяцы, когда радость ощущается всего острее, а лишь много спустя, когда впечатления успевают притупиться.

Напротив, господствовавшее до Дмитриева понимание связи между урожаем и потреблением водки, сводившее ее к повышенному хозяйственному благополучию населения в годы хороших урожаев, находит себе в картине рассматриваемых нами коэффициентов корреляции полное подтверждение. Все отмеченные различия в тесноте связи прекрасно согласуются с ним. Урожай не повсюду играет в хозяйстве одну и ту же роль и не везде реализуется в одно и то же время.

В Самарской губернии сбор хлебов почти всецело определяет собою доходы населения, и урожай притом рано выносится на рынок, так что и деньги, выручаемые за него, сравнительно рано поступают в распоряжение земледельца. Напротив, на севере сбор хлебов даже в урожайные годы не составляет единственного — местами не составляет даже главного — источника доходов; к нему присоединяются промыслы, заработки от лесных работ и т. д. Здесь воздействие высоты урожая не может долго тянуться и, кроме того, легко перебивается перекрещивающим его воздействием колебаний дохода из иных источников.

Исследование М. М. Виноградовой еще прочнее обосновывает это объяснение связи потребления водки с урожаем путем дальнейшего анализа имеющихся в распоряжении данных. Если дело, действительно, в колебаниях имущественной обеспеченности, то воздействие на потребление водки, должно, очевидно исходить не столько от количества собранных хлебов, сколько от ценности сбора, при чем связь потребления водки с размерами дохода от сельского хозяйства должна быть особенно сильна в те месяцы, когда урожай поступает в данной местности на рынок. Чтоб уловить отражение этих моментов, М. М. Виноградова вычисляет по ряду губерний коэффициенты корреляции между потреблением водки за отдельные месяцы и ценностью урожая тех культур, которые играют там более видную роль в денежных поступлениях крестьянского хозяйства. Результаты, к которым она приходит, складываются в чрезвычайно отчетливую картину. Неизменно связь оказывается наиболее тесною с ценностью сбора тех продуктов, которые в данной местности в данном месяце обычно выносятся крестьянами на рынок. Так, например, в Смоленской губернии в августе наблюдается некоторая связь потребления водки с ценностью урожая ржи, но уже в сентябре рожь перестает сказываться на потреблении; в сентябре замечается сравнительно тесная связь с ценностью льняного и конопляного семени, к которым в октябре, ноябре присоединяются овес и картофель, а в декабре льняное волокно и пенька.

Богатство статистических данных о потреблении водки за время казенной продажи питей позволяет внести в эту, и так уже весьма яркую картину, любопытные добавочные штрихи. Представляется, между прочим, возможным подвергнуть изучению сбыт водки в посуде различного объема: в четвертях, в бутылках, в полубутылках и в мелкой посуде. Теснота связи с урожаем оказывается далеко не одинаковою. По Уфимской губернии, например, коэффициент корреляции между сбытом водки в четвертях за календарный год и урожаем предшествующего года получается равным $+0{,}8$; для сбыта водки в бутылках коэффициент корреляции равняется $+0{,}7$; для полубутылок коэффициент корреляции опускается до $+0{,}6$ и, наконец, для мелкой посуды падает до $+0{,}4$. Сходная картина наблюдается и по другим губерниям. Объяснения следует, несомненно, искать, как указывает М. М. Виноградова, в том обстоятельстве, что размер посуды свидетельствует „о характере удовлетворяемой потребности. „Привычно-регулярный потребитель“... прибегает главным образом к мелкой посуде; напротив к большим праздникам, к свадьбам водка закупается в более крупных размерах“; „потребление в крупной посуде относится главным образом к случаям бытового или обрядового алкоголизма“. Тесная связь между потреблением водки в крупной посуде и урожаем указывает, таким образом, на важное посредствующее

звено в связи урожая с потреблением водки: в годы хороших урожаев население, располагая бо́льшими свободными средствами, пышнее справляет праздники, чаще играет свадьбы и не скупится на свадьбах на угощение. Связь потребления водки с числом свадеб, с одной стороны, и числа свадеб с урожаем, с другой, с большой отчетливостью обнаруживается и путем непосредственного сопоставления соответственных рядов чисел, — параллелизм резко бросается в глаза.

К некоторым из полученных Виноградовою выводов можно было бы притти и при помощи „элементарных“ приемов исследования. Во многих случаях связь так тесна, что одного взгляда на график, достаточно, дабы ее подметить. Но те характерные различия в силе связи, которые схватываются при помощи измерения ее тесноты, ускользнут от исследователя, пользующегося „элементарными“ приемами. Между сбытом водки в крупной и в мелкой посуде, между потреблением разных месяцев в разных губерниях, между количеством и ценностью сбора, между ценностью сбора разных продуктов и потреблением водки в разные месяцы для него не будет осязательной разницы. Везде он будет констатировать наличие связи и останавливаться на этом. Лишь „математические“ приемы вносят краску и жизнь в бесцветный эскиз, дальше которого не пошел бы статистик, отказывающийся от их использования. Лишь планомерные промеры дают возможность вникнуть глубже в самое существо стохастической связи между изучаемыми явлениями.

Отсюда, разумеется, не следует, что достаточно начать вычислять по формулам ряды коэффициентов корреляции, корреляционных отношений и т. д.. и выполняемое исследование наполнится жизненным содержанием, заиграет красками и приведет к научно-ценным выводам. Теория корреляции предоставляет в распоряжение исследователя богатый набор усовершенствованных рабочих инструментов. Используя их умело, статистик может извлечь из своего материала не мало такого, что иначе от него ускользнуло бы. Но не характер инструмента сообщает подлинную ценность произведению труда, а мастерство того, кто держит инструмент в руках. Несколько беглых черт, нанесенных рукою гения на обтрепанный клочек бумаги, лучше запечатлеют образ, нежели иной старательно выписанный роскошнейшими красками портрет. Для того, чтоб методы, выработанные теорией корреляции, вели к углубленному пониманию изучаемых явлений, исследователь, их применяющий, должен сам стоять на высоте задачи. Он должен не только владеть инструментом, но также владеть материалом и владеть предметом своего исследования. Он должен быть способен применяться в технике статистической работы к преследуемым целям и к имеющимся возможностям. Шаблонно же механическое использование готовых рецептов, хотя бы даже опирающихся на самые точные формулы и на самые тонкие математические соображения, ведет не к умножению наших знаний ценою больших, но оправдывающих себя затрат труда, а к бесплодному расточению сил и к нагромождению цифрового материала, мало продвигающего вперед наше понимание изучаемых явлений.

ПРИЛОЖЕНИЕ.

Глава первая.

§ II, C). Обозначая через d_i разность i-ого номера в ряду X и соответствующего ему номера в ряду Y и замечая, что

$$\sum_{X=1}^{n} X = \frac{n(n+1)}{2}, \qquad \sum_{X=1}^{n} X^2 = \frac{n(n+1)(2n+1)}{6};$$

находим в случае, если номера ряда Y располагаются в правильно убывающем порядке:

$$\sum_{i=1}^{n} d_i^2 = [1-n]^2 + [2-(n-1)]^2 + \cdots + [h-[(n-h+1)]^2 + \cdots +$$

$$+ [(n-1)-2]^2 + [n-1]^2 = \sum_{h=1}^{n} [h-(n-h+1)]^2 =$$

$$= \sum_{h=1}^{n} [2h-(n+1)]^2 = 4\sum_{h=1}^{n} h^2 - 4(n+1)\sum_{h=1}^{n} h + n(n+1)^2 =$$

$$= \frac{n(n^2-1)}{3}.$$

Если между расположением номеров ряда Y и расположением номеров ряда X нет никакой связи, то с каждым номером в ряду X может оказаться сопряженным любой из n номеров ряда Y с равною вероятностью в $\frac{1}{n}$. Рассмотрим в ряду X член, получающий номер i. Ожидаемое значение квадрата

разности между номером i в ряду X и соответственным номером в ряду Y равняется

$$d_i^2 = \frac{1}{n}[i-1]^2 + \frac{1}{n}[i-2]^2 + \cdots + \frac{1}{n}[i-(i-1)]^2 + \frac{1}{n}[i-i]^2 +$$

$$+ \frac{1}{n}[i-(i+1)]^2 + \cdots + \frac{1}{n}[i-(n-1)]^2 + \frac{1}{n}[i-n]^2 = \frac{1}{n}\sum_{h=1}^{i-1} h^2 + \frac{1}{n}\sum_{h=1}^{n-i} h^2 =$$

$$= \frac{1}{n}\left\{\frac{(i-1)\,i\,(2i-1)}{6} = \frac{(n-i)(n-i+1)(2n-2i+1)}{6}\right\} =$$

$$= \frac{1}{6}[2n^2 + 3n + 1] - (n+1)i + i^2.$$

Суммируя теперь по i от $i = 1$ до $i = n$, находим для случая взаимной независимости рядов

$$\sum_{i=1}^{n} d_i^2 = \frac{n(2n^2+3n+1)}{6} - \frac{n(n+1)^2}{2} + \frac{n(n+1)(2n+1)}{6} = \frac{n(n^2-1)}{6}.$$

Глава четвертая.

§ I, 2). В дополнение к приведенным в тексте соотношениям следует заметить следующие тожества, на которые приходится постоянно опираться при выкладках:

$$p_{i/} = \sum_{j=1}^{l} p_{ij} \qquad p_{/j} = \sum_{i=1}^{k} p_{ij}$$

$$1 = \sum_{i} p_{i/} = \sum_{j} p_{/j} = \sum_{i}\sum_{j} p_{ij}$$

$$1 = \sum_{i} p_{i/}^{(j)} = \sum_{j} p_{/j}^{(i)}$$

§ II, 1). Принимая во внимание, что, при $k = l = 2$,

$$p_{1/} = p_{1/1} + p_{1/2},\ p_{1/} + p_{2/} = 1 \text{ и т. д.},$$

находим без труда:

$$\delta = p_{1/1} - p_{1/}p_{/1} = p_{1/} - p_{1/2} - p_{1/}[1 - p_{/2}] = -[p_{1/2} - p_{1/}p_{/2}] = p_{2/2} - p_{2/}p_{/2} =$$

$$= -[p_{2/1} - p_{2/}p_{/1}].$$

Принимая, с другой стороны, во внимание, что

$$p_{1/1} + p_{1/2} + p_{2/1} + p_{2/2} = 1,$$

получаем:

$$\delta = p_{1/1} - p_{1/}\,p_{/1} = p_{1/1}[p_{1/1} + p_{1/2} + p_{2/1} + p_{2/2}] - [p_{1/1} + p_{1/2}]\,[p_{1/1} + p_{2/1}] =$$
$$= p_{1/1}\,p_{2/2} - p_{1/2}\,p_{2/1}.$$

Подставляя найденные значения в общую формулу для показателя взаимной сопряженности, находим:

$$\varphi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{[p_{i/j} - p_{i/}\,p_{/j}]^2}{p_{i/}\,p_{/j}} = \delta^2\left\{\frac{1}{p_{1/}\,p_{/1}} + \frac{1}{p_{1/}\,p_{/2}} + \frac{1}{p_{2/}\,p_{/1}} + \frac{1}{p_{2/}\,p_{/2}}\right\} =$$
$$= \frac{\delta^2}{p_{1/}\,p_{2/}\,p_{/1}\,p_{/2}}.$$

§ III, 1). Замечая, что

$$m_{/1}^{(i)} = \sum_j p_{j}^{(i)}\,y_j,$$

находим:

$$\sum_i p_{i/}\,m_{/1}^{(i)} = \sum_i \sum_j p_{i/}\,p_{j}^{(i)}\,y_j = \sum_i \sum_j p_{ij}\,y_j = \sum_j p_{j}\,y_j = m_{0/1}.$$

Принимая во внимание, что

$$\sum_i \sum_j p_{ij}\,[y_j - m_{0/1}]\,[m_{/1}^{(i)} - m_{0/1}] = \sum_i \{p_{i/}[m_{/1}^{(i)} - m_{0/1}]\sum_j p_{j}^{(i)}[y_j - m_{0/1}]\} =$$
$$= \sum_i p_{i/}\,[m_{/1}^{(i)} - m_{0/1}]^2,$$

получаем

$$\sum_i p_{i/}\,\mu_{/2}^{(i)} = \sum_i \sum_j p_{i/}\,p_{j}^{(i)}[y_j - m_{/1}^{(i)}]^2 = \sum_i \sum_j p_{ij}[(y_j - m_{0/1}) - (m_{/1}^{(i)} - m_{0/1})]^2 =$$
$$= \sum_i \sum_j p_{ij}\,[y_j - m_{0/1}]^2 - \sum_i p_{i/}[m_{/1}^{(i)} - m_{0/1}]^2 = \mu_{0/2} - \sum_i p_{i/}\,[m_{/1}^{(i)} - m_{0/1}]^2.$$

§ III, 1). Принимая во внимание, что

$$m_{f/g} = \sum_i \sum_j p_{ij}\,x_i^f\,y_j^g, \quad \text{а} \quad m_{f/0}\,m_{0/g} = \sum_i \sum_j p_{i/}\,p_{j}\,x_i^f\,y_j^g,$$

находим:

$$m_{f/g} - m_{f/0}\,m_{0/g} = \sum_i \sum_j [p_{ij} - p_{i/}\,p_{j}]\,x_i^f\,y_j^g.$$

Таким образом, если

$$m_{f,g} - m_{f,0}\, m_{0,g} = 0,$$

то

$$\sum_i \sum_j [p_{ij} - p_{i/} p_{/j}]\, x_i^f\, y_j^g = 0.$$

Последнее соотношение может иметь силу при любых целых положительных значениях f и g лишь, если все разности $p_{ij} - p_{i/} p_{/j}$ равны нулю, т.-е. если переменные взаимно независимы.

§ III, 2), A).

$$\sum_i p_{i/}\, x_i^h\, m_{/1}^{(i)} = \sum_i p_{i/}\, x_i^h \sum_j p_{/j}^{(i)}\, y_j = \sum_i \sum_j p_{ij}\, x_i^h\, y_j = m_{h,1}.$$

§ III, 2), B). Замечая, что

$$\sum_j p_{/j}^{(i)} = 1$$

и что, следовательно,

$$m_{/1}^{(i)} - m_{0/1} = \sum_j p_{/j}^{(i)}\, y_j - \sum_j p_{/j}^{(i)}\, m_{0,1} = \sum_j p_{/j}^{(i)} [y_j - m_{0,1}],$$

находим:

$$\sum_i p_{i/} [x_i - m_{1/0}]^h [m_{/1}^{(i)} - m_{0\,1}] = \sum_i \sum_j p_{ij} [x_i - m_{1/0}]^h [y_j - m_{0/1}] = \mu_{h/1}.$$

§ III, 2), C). Замечая, что

$$\mathfrak{X}_i = \frac{x_i - m_{1/0}}{\sigma_X}, \quad \mathfrak{M}_{/1}^{(i)} = \frac{m_{/1}^{(i)} - m_{0/1}}{\sigma_Y},$$

находим

$$\sum_i p_{i/}\, \mathfrak{X}_i^h = \frac{1}{\sigma_X^h} \sum_i p_{i/} [x_i - m_{1/0}]^h = \frac{\mu_{h/0}}{\sigma_X^h} = r_{h/0},$$

$$\sum_i p_{i/}\, \mathfrak{M}_{/1}^{(i)}\, \mathfrak{X}_i^h = \frac{1}{\sigma_X^h \sigma_Y} \sum_i p_{i/} [x_i - m_{1/0}]^h [m_{/1}^{(i)} - m_{0/1}] = \frac{\mu_{h/1}}{\sigma_X^h \sigma_1} = r_{h/1}.$$

Принимая во внимание, что

$$r_{0/0} = 1, \; r_{1/0} = r_{0/1} = 0, \; r_{2/0} = r_{0/2} = 1,$$

приходим в случае параболы второй степени к уравнениям

$$0 = c_{/0} + c_{/2}$$
$$r_{1/1} = c_{/1} + r_{3/0}\, c_{/2}$$
$$r_{2/1} = c_{/0} + r_{3/0}\, c_{/1} + r_{4/0}\, c_{/2},$$

из которых без труда находятся значения коэффициентов $c_{/0}$, $c_{/1}$, $c_{/2}$.

§ III, 4). Приравнивая нулю первые производные

$$\sum_i p_{i/} [m_{/1}^{(i)} - A_{/0} - A_{/1} x_i]^2$$

по $A_{/0}$ и по $A_{/1}$, приходим к уравнениям

$$\sum_i p_{i/} [m_{/1}^{(i)} - A_{/0} - A_{/1} x_i] = 0, \quad \sum_i p_{i/} x_i [m_{/1}^{(i)} - A_{/0} - A_{/1} x_i] = 0,$$

которые приводятся к

$$m_{0/1} - A_{/0} - A_{/1} m_{1/0} = 0, \quad m_{1/1} - A_{/0} m_{1/0} - A_{/1} m_{2/0} = 0.$$

§ IV, 2). В случае, если линия регрессии Y по X имеет вид параболы второй степени с уравнением в нормальных координатах

$$\mathfrak{M}_{/1}^{(i)} = r_{1/1} \mathfrak{X}_i + c_{/2} [\mathfrak{X}_i^2 - r_{3/0} \mathfrak{X}_i - 1],$$

находим:

$$[\mathfrak{M}_{/1}^{(i)}]^2 = r_{1/1}^2 \mathfrak{X}_i^2 + 2 r_{1/1} c_{/2} \mathfrak{X}_i [\mathfrak{X}_i^2 - r_{3/0} \mathfrak{X}_i - 1] + c_{/2}^2 [\mathfrak{X}_i^2 - r_{3/0} \mathfrak{X}_i - 1]^2.$$

Принимая во внимание, что

$$\sum_i p_{i/} \mathfrak{X}_i [\mathfrak{X}_i^2 - r_{3/0} \mathfrak{X}_i - 1] = r_{3/0} - r_{3/0} = 0,$$

$$\sum_i p_{i/} [\mathfrak{X}_i^2 - r_{3/0} \mathfrak{X}_i - 1]^2 = \sum_i p_{i/} \{\mathfrak{X}_i^4 - 2 r_{3/0} \mathfrak{X}_i^3 + [r_{3/0}^2 - 2] \mathfrak{X}_i^2 + 2 r_{3/0} \mathfrak{X}_i + 1\} =$$
$$= r_{4/0} - r_{3/0}^2 - 1$$

и что

$$c_{/2} = \frac{r_{2/1} - r_{3/0}\, r_{1/1}}{r_{4/0} - r_{3\,0}^2 - 1},$$

получаем:

$$\eta_{Y/X}^2 = \sum_i p_{i/} [\mathfrak{M}_{/1}^{(i)}]^2 = r_{1/1}^2 + c_{/2}^2 [r_{4/0} - r_{3/0}^2 - 1] = r_{1/1}^2 + \frac{[r_{2/1} - r_{3/0}\, r_{1/1}]^2}{r_{4/0} - r_{3/0}^2 - 1}.$$

§ IV, 3).

$$\sum_i p_{i/}[m_{/1}^{(i)} - M_{/1}^{(i)}]^2 = \sum_i p_{i/}\left[(m_{/1}^{(i)} - m_{0/1}) - \frac{\mu_{1/1}}{\mu_{2/0}}(x_i - m_{1/0})\right]^2 =$$

$$= \sum_i p_{i/}[m_{/1}^{(i)} - m_{0/1}]^2 - 2\frac{\mu_{1/1}}{\mu_{2/0}}\sum_i p_{i/}[x_i - m_{1/0}][m_{/1}^{(i)} - m_{0/1}] +$$

$$+ \frac{\mu_{1/1}^2}{\mu_{2/0}^2}\sum_i p_{i/}[x_i - m_{1/0}]^2 = \mu_{0/2}\,\eta_{Y/X}^2 - \mu_{0/2}\,r_{1/1}^2.$$

§ IV, 3). Предположим, что

$$X = W_1 + W_2 + \cdots + W_m + U_1 + U_2 + \cdots + U_n,$$
$$Y = W_1 + W_2 + \cdots + W_m + T_1 + T_2 + \cdots + T_l,$$

причем

$$W_1,\ W_2,\ \cdots,\ W_m;\ U_1,\ U_2,\ \cdots,\ U_n;\ T_1,\ T_2,\ \cdots,\ T_l$$

взаимно независимые случайные переменные, следующие одному и тому же закону распределения значений. Обозначая через m_1 математическое ожидание этих переменных и через μ_2 их рассеяние и замечая, что

$$E[W_i - m_1][W_j - m_1] = 0, \qquad E[U_i - m_1][U_j - m_1] = 0,$$
$$E[T_i - m_1][T_j - m_1] = 0 \text{ при } j \neq i;$$
$$E[W_i - m_1][U_j - m_1] = 0, \qquad E[W_i - m_1][T_j - m_1] = 0,$$
$$E[U_i - m_1][T_j - m_1] = 0 \text{ при всяком } j,$$

находим:

$$X - EX = \sum_{i=1}^{m}[W_i - m_1] + \sum_{j=1}^{n}[U_j - m_1],$$

$$Y - EY = \sum_{i=1}^{m}[W_i - m_1] + \sum_{j=1}^{l}[T_j - m_1],$$

$$\sigma_X^2 = E[X - EX]^2 = \sum_{i=1}^{m}E[W_i - m_1]^2 + 2\sum_i\sum_j E[W_i - m_1][U_j - m_1] +$$

$$+ \sum_{j=1}^{n}E[U_j - m_1]^2 = [m + n]\,\mu_2,$$

$$\sigma_Y^2 = [m + l]\,\mu_2,$$

$$E\{[X - EX][Y - EY]\} = \sum_{i=1}^{m}E[W_i - m_1]^2 = m\,\mu_2,$$

$$r_{1/1} = \frac{E\{[X - EX][Y - EY]\}}{\sigma_X\,\sigma_Y} = \frac{m}{\sqrt{[m + n][m + l]}}.$$

§ V, 2). Параметрам

$$r_{f/h} = \frac{1}{2\pi\sqrt{1-r^2_{1/1}}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{\mathfrak{X}^2 - 2r_{1/1}\mathfrak{X}\mathfrak{Y} + \mathfrak{Y}^2}{2[1-r^2_{1/1}]}} \mathfrak{X}^f \mathfrak{Y}^h \, d\mathfrak{X}\, d\mathfrak{Y}$$

могут быть придаваемы разные формы. Приведенные в тексте выражения для $r_{2f/2h}$ и $r_{2f+1/2h+1}$ даны в статье K. Pearson and A. W. Young. On the product-moments of various orders of the normal correlation surface of two variates (Biometrika, vol. XII).

Наиболее прямой путь вычисления $r_{f/h}$ сводится к подстановке $\mathfrak{X} = r_{1/1}\mathfrak{Y} + \mathfrak{Z}$ и к разложению $[r_{1/1}\mathfrak{Y} + \mathfrak{Z}]^f$ по степеням $\mathfrak{Y}$ и $\mathfrak{Z}$, после чего двойной интеграл приводится к произведению двух простых, легко вычисляемых по формуле

$$\int_{-\infty}^{+\infty} t^{2n} e^{-ct^2} dt = \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^n}\sqrt{\frac{\pi}{c^{2n+1}}}.$$

Замечая, что

$$r_{2f+1/1} = 1\cdot 3\cdot 5\cdots(2f+1)\, r_{1/1} = r_{2f+2/0}\, r_{1/1},$$

убеждаемся, что регрессия Y по X прямолинейная.

Замечая, что $r_{2f/0} = 1\cdot 3\cdot 5\cdots(2f-1)$, убеждаемся, что распределение значений X следует закону Лапласа-Гаусса.

Придавая переменной Y определенное значение, без труда убеждаемся, что условный закон распределения значений X имеет вид закона Лапласа-Гаусса, причем условное рассеяние X остается при всех значениях Y равным

$$[1 - r_{1/1}{}^2]\,\mu_{2/0}.$$

Соотношение

$$\varphi^2 = \frac{r_{1/1}{}^2}{1 - r_{1/1}{}^2}$$

указано Пирсоном, — см. K. Pearson, On the theory of contingency and its relation to association and normal correlation, p. 7-8 (Drapers' Company Research Memoirs, Biometric Series, I, 1904).

§ VII. В случае, если обе переменных могут принимать лишь по два разных значения, находим:

$$EX = p_{1/}x_1 + p_{2/}x_2, \quad x_1 - EX = p_{2/}[x_1 - x_2], \quad x_2 - EX = -p_{1/}[x_1 - x_2],$$

$$EY = p_{/1}y_1 + p_{/2}y_2, \quad y_1 - EY = p_{/2}[y_1 - y_2], \quad y_2 - EY = -p_{/1}[y_1 - y_2],$$

$$\mu_{2/0} = p_{1/}[x_1 - EX]^2 + p_{2/}[x_2 - EX]^2 = p_{1/}p_{2/}[x_1 - x_2]^2,$$

$$\mu_{0/2} = p_{/1} p_{/2} [y_1 - y_2]^2,$$

$$\mu_{1/1} = EXY - [EX][EY] = (p_{1/1} - p_{1/} p_{/1}) x_1 y_1 + (p_{1/2} - p_{1/} p_{/2}) x_1 y_2 +$$
$$+ (p_{2/1} - p_{2/} p_{/1}) x_2 y_1 + (p_{2/2} - p_{2/} p_{/2}) x_2 y_2 = \delta [x_1 y_1 - x_1 y_2 - x_2 y_1 + x_2 y_2] =$$
$$= \delta [x_1 - x_2][y_1 - y_2],$$

$$m_{/1}^{(1)} = \frac{1}{p_{1/}} [p_{1/1} y_1 + p_{1/2} y_2], \quad m_{/1}^{(2)} = \frac{1}{p_{2/}} [p_{2/1} y_1 + p_{2/2} y_2].$$

$$m_{0/1} = p_{/1} y_1 + p_{/2} y_2,$$

$$m_{/1}^{(1)} - m_{0/1} = \frac{1}{p_{1/}} \{[p_{1/1} - p_{1/} p_{/1}] y_1 + [p_{1/2} - p_{1/} p_{/2}] y_2\} = \frac{1}{p_{1/}} \delta [y_1 - y_2],$$

$$m_{/1}^{(2)} - m_{0/1} = \frac{1}{p_{2/}} \delta [y_1 - y_2],$$

$$\eta_{1/x}^2 = \frac{1}{\mu_{0/2}} \sum_{i=1}^{2} p_{i/} \left[m_{/1}^{(i)} - m_{0/1}\right]^2 = \frac{\delta^2 [y_1 - y_2]^2 \left[\frac{1}{p_{1/}} + \frac{1}{p_{2/}}\right]}{p_{/1} p_{/2} [y_1 - y_2]^2} = \frac{\delta^2}{p_{1/} p_{2/} p_{/1} p_{/2}}.$$

Глава пятая.

Формулы главы пятой так тесно примыкают к формулам главы четвертой. что вывод их не нуждается в пояснениях: чтоб получить их, достаточно следовать линиям вывода соответствующих априорных формул.

Глава шестая.

§ II, 1). Если вынутый шар возвращается в урну до производства следующего извлечения, то вероятность P_x вынуть при N извлечениях x белых шаров. равняется, как известно,

$$P_x = \frac{N(N-1)\cdots(N-x+1)}{1 \cdot 2 \cdots (x-1)\, x} p^x (1-p)^{N-x},$$

причем

$$\sum_{x=1}^{N} P_x = [p + (1-p)]^N = 1.$$

Математическое ожидание числа белых шаров, появляющихся при N извлечениях, проще всего может быть получено следующим путем:

$$En = \sum_{x=0}^{N} \frac{N(N-1)\cdots(N-x+1)}{1 \cdot 2 \cdots (x-1)\, x} p^x (1-p)^{N-x} x =$$
$$= \sum_{x=1}^{N} \frac{N(N-1)\cdots(N-x+1)}{1 \cdot 2 \cdots (x-1)} p^x (1-p)^{N-x} =$$
$$= Np \sum_{x=1}^{N} \frac{(N-1)\cdots(N-x+1)}{1 \cdot 2 \cdots (x-1)} p^{x-1} (1-p)^{N-x};$$

заменяя x через $t+1$, находим отсюда:

$$En = Np \sum_{t=0}^{N-1} \frac{(N-1)\cdots(N-t)}{1\cdot 2\cdots t} p^t (1-p)^{N-1-t} = Np\,[p+(1-p)]^{N-1} = Np.$$

Если вынутый шар не возвращается в урну, то вероятность вынуть x белых шаров при N извлечениях, если число белых шаров в урне равно $B = Ap$, а число черных шаров равно $C = A(1-p)$, равняется, как известно,

$$P_x = \frac{N(N-1)\cdots(N-x+1)}{1\cdot 2\cdots(x-1)\,x} \frac{B(B-1)\cdots(B-x+1)\,C(C-1)\cdots(C-N+x+1)}{A(A-1)\cdots(A-N+1)};$$

причем

$$\sum_{x=0}^{N} P_x = 1.$$

Тем же путем, как выше, находим:

$$En = \sum_{x=0}^{N} P_x x =$$

$$= \frac{NB}{A} \sum_{x=1}^{N} \frac{(N-1)\cdots(N-x+1)}{1\cdot 2\cdots(x-1)} \frac{(B-1)\cdots(B-x+1)\,C(C-1)\cdots(C-N+x+1)}{(A-1)\cdots(A-N+1)} =$$

$$= Np \sum_{t=0}^{N-1} \frac{(N-1)\cdots(N-t)}{1\cdot 2\cdots t} \frac{(B-1)\cdots(B-t)\,C(C-1)\cdots(C-N+t+2)}{(A-1)\cdots(A-N+1)} = Np;$$

ибо сумма, на которую умножается Np, равна 1, как сумма вероятностей всех возможных значений числа белых шаров, появляющихся при $N-1$ извлечениях из урны, содержащей $B-1$ белых и C черных шаров.

Замечая, что $x^2 = x(x-1)+x$, получаем при помощи того же приема в случае шара возвращаемого

$$En^2 = \sum_{x=0}^{N} P_x x^2 = N(N-1)p^2 + Np = N^2p^2 + Np(1-p)$$

и отсюда

$$\sigma_{\frac{n}{N}}^2 = E\left[\frac{n}{N} - E\frac{n}{N}\right]^2 = E\frac{n^2}{N^2} - p^2 = \frac{1}{N}p(1-p),$$

а в случае шара невозвращаемого

$$En^2 = N(N-1)\frac{B(B-1)}{A(A-1)} + N\frac{B}{A} = N^2\frac{B(B-1)}{A(A-1)} + N\frac{B}{A}\left[1 - \frac{B-1}{A-1}\right],$$

$$\sigma_{\frac{n}{N}}^2 = E\frac{n^2}{N^2} - \frac{B^2}{A^2} = \frac{B(A-B)}{A^2(A-1)}\left[\frac{A}{N} - 1\right] = \frac{A-N}{A-1}\frac{1}{N}p(1-p).$$

§ III, 2). Обозначая через P_h вероятность, что $n_{1/1}$ принимает значение h и через $E^{(h)} n_{2/2}$ условное математическое ожидание $n_{2/2}$ в предположении, что $n_{1/1} = h$, находим:

$$P_h = \frac{N(N-1)\cdots(N-h+1)}{1\cdot 2\cdots h} p_{1/1}{}^h (1-p_{1/1})^{N-h}.$$

$$En_{1/1} n_{2/2} = \sum_h P_h h\, En_{2/2}^{(h)}.$$

Замечая, что при $n_{1/1} = h$ на остальные три группы приходится $N-h$ испытаний, причем вероятность сочетания значений x_2 и y_2 получает значение

$$\frac{p_{2/2}}{p_{2/2}+p_{1/2}+p_{2/1}} = \frac{p_{2/2}}{1-p_{1/1}},$$

находим:

$$En_{2/2}^{(h)} = (N-h)\frac{p_{2/2}}{1-p_{1/1}},$$

$$En_{1/1}\, n_{2/2} = \sum_h P_h h (N-h)\frac{p_{2/2}}{1-p_{1/1}} = \frac{p_{2/2}}{1-p_{1/1}}\{N\sum_h P_h h - \sum_h P_h h^2\} =$$

$$= \frac{p_{2/2}}{1-p_{1/1}}\{N^2 p_{1/1} - N^2 p_{1/1}^2 - N p_{1/1}(1-p_{1/1})\} = N(N-1)p_{1/1}p_{2/2}.$$

Сходным путем, полагая

$$N(N-1)\cdots(N-t+1) = N^{t\cdot}$$

находим:

$$En_{1/1} n_{2/2} n_{1/2} n_{2/1} = N^{4\cdot} p_{1/1} p_{2/2} p_{1/2} p_{2/1},$$

$$En_{1/1}{}^2 n_{2/2}{}^2 = N^{4\cdot} p_{1/1}{}^2 p_{2/2}{}^2 + N^{3\cdot} p_{1/1} p_{2/2}[p_{1\cdot 1} + p_{2/2}] + N^{2\cdot} p_{1/1} p_{2/2},$$

$$En_{1/2}{}^2 n_{2/1}{}^2 = N^{4\cdot} p_{1/2}{}^2 p_{2/1}{}^2 + N^{3\cdot} p_{1/2} p_{2/1}[p_{1/2} + p_{2/1}] + N^{2\cdot} p_{1/2} p_{2/1},$$

$$N^4 E[\delta']^2 = En_{1/1}{}^2 n_{2/2}{}^2 + En_{1/2}{}^2 n_{2/1}{}^2 - 2En_{1/1} n_{2/2} n_{1/2}\, n_{2/1} =$$

$$= N^{4\cdot}[p_{1/1}p_{2/2} - p_{1/2}p_{2/1}]^2 + N^{3\cdot}\{p_{1/1}p_{2/2}[p_{1/1}+p_{2/2}] + p_{1/2}p_{2/1}[p_{1/2}+p_{2/1}]\} +$$

$$+ N^{2\cdot}[p_{1/1}p_{2/2} + p_{1/2}p_{2/1}] = N(N-1)\{N^2\delta^2 + N[p_{1/1}p_{2/2}(p_{1/1}+p_{2/2}) +$$

$$+ p_{1/2}p_{2/1}(p_{1/2}+p_{1/2}) - 5\delta^2] + [p_{1/1}p_{2/2} + p_{1/2}p_{2/1} - 2p_{1/1}p_{2/2}(p_{1/1}+p_{2/2}) -$$

$$- 2p_{1/2}p_{2/1}(p_{1/2}+p_{2/1}) + 6\delta^2]\}.$$

$$\sigma^2{}_{\frac{N}{N-1}\,\delta'} = E\left[\frac{N}{N-1}\,\delta'\right]^2 - \delta^2 =$$

$$= \frac{1}{N-1}\{[p_{1/1}p_{2/2}(p_{1/1}+p_{2/2}) + p_{1/2}p_{2/1}(p_{1/2}+p_{2/1}) - 4\delta^2] +$$

$$+ \frac{1}{N}[(p_{1/2}+p_{2/1})\delta - (p_{1/1}+p_{2/2})\delta + 6\delta^2]\}.$$

§ IV, 1), B). При неизменности закона распределения и взаимной независимости испытаний

$$E\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}=E\frac{[x^{[d]\prime}-x_0']^2}{\Sigma_1^2},$$

$$E\frac{[x^{[f]\prime}-x_0'][x^{[d]\prime}-x_0']}{\Sigma_1^2}=E\frac{[x^{[h]\prime}-x_0'][x^{[c]\prime}-x_0']}{\Sigma_1^2}$$

и, следовательно,

$$E\frac{[x^{[f]\prime}-x_0'][x^{[d]\prime}-x_0']}{\Sigma_1^2}=\frac{1}{N-1}E\frac{[x^{[f]\prime}-x_0']\sum\limits_{d\neq f}[x^{[d]\prime}-x_0']}{\Sigma_1^2}=$$

$$=\frac{1}{N-1}E\frac{[x^{[f]\prime}-x_0']\{\sum\limits_{d\neq f}[x^{[d]\prime}-x_0']+[x^{[f]\prime}-x_0']-[x^{[f]\prime}-x_0']\}}{\Sigma_1^2}=$$

$$=-\frac{1}{N-1}E\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}.$$

Сходным путем убеждаемся, что

$$E\frac{[y^{[f]\prime}-y_0'][y^{[d]\prime}-y_0']}{\Sigma_2^2}=-\frac{1}{N-1}E\frac{[y^{[f]\prime}-y_0']^2}{\Sigma_2^2}.$$

Принимая во внимание, что переменные X и Y взаимно независимы, находим, таким образом,

$$E[r_{1/1}']^2=E\left\{\sum_{f=1}^{N}\frac{[x^{[f]\prime}-x_0'][y^{[f]\prime}-y_0']}{\Sigma_1\,\Sigma_2}\right\}^2=$$

$$=E\sum_f\frac{[x^{[f]\prime}-x_0']^2[y^{[f]\prime}-y_0']^2}{\Sigma_1^2\,\Sigma_2^2}+$$

$$+E\sum_f\sum_{d\neq f}\frac{[x^{[f]\prime}-x_0']^2[x^{[d]\prime}-x_0'][y^{[f]\prime}-y_0'][y^{[d]\prime}-y_0']}{\Sigma_1^2\,\Sigma_2^2}=$$

$$=N\left\{E\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}\right\}\left\{E\frac{[y^{[f]\prime}-y_0']^2}{\Sigma_2^2}\right\}+$$

$$+N(N-1)\left\{E\frac{[x^{[f]\prime}-x_0'][x^{[d]\prime}-x_0']}{\Sigma_1^2}\right\}\left\{E\frac{[y^{[f]\prime}-y_0'][y^{[d]\prime}-y_0']}{\Sigma_2^2}\right\}=$$

$$=\frac{N^2}{N-1}\left\{E\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}\right\}\left\{E\frac{[y^{[f]\prime}-y_0']^2}{\Sigma_2^2}\right\}.$$

Но

$$E\frac{[x^{[f]\prime}-x_0']^2}{\Sigma_1^2}=\frac{1}{N},\quad E\frac{[y^{[f]\prime}-y_0']^2}{\Sigma_2^2}=\frac{1}{N}.$$

Следовательно

$$E[r_{1/1}']^2=\frac{1}{N-1}.$$

§ IV, 2). Приведенные в тексте формулы даны в некоторых случаях в составе до членов порядка $\left(\frac{1}{N}\right)^2$ включительно, но список вспомогательных формул я ограничиваю теми, которые необходимы для вычисления членов порядка $\frac{1}{N}$, дабы не загромождать чрезмерно Приложения.

$$E[dp_{ij}']^2 = \frac{1}{N} p_{ij}(1-p_{ij}), \qquad E[dp_{ii}']^2 = \frac{1}{N} p_{ii}(1-p_{ii}),$$

$$E[dp_{ij}'\, dp_{fg}'] = E[p_{ij}'\, p_{fg}'] - p_{ij}\, p_{fg} = \frac{N-1}{N} p_{ij}\, p_{fg} - p_{ij}\, p_{fg} = -\frac{1}{N} p_{ij} p_{fg},$$

$$E[dp_{ij}'\, dp_{ig}'] = -\frac{1}{N} p_{ij}\, p_{ig}, \qquad E[dp_{ii}'\, dp_{fi}'] = -\frac{1}{N} p_{ii}\, p_{fi},$$

$$E[dp_{ii}'\, dp_{ij}'] = E\{[\sum_{g=1}^{l} dp_{ig}']\, dp_{ij}'\} = E\{[dp_{ij}']^2 + \sum_{g \neq j} dp_{ig}'\, dp_{ij}'\} =$$

$$= \frac{1}{N} p_{ij}(1-p_{ij}) - \frac{1}{N} \sum_{g \neq j} p_{ig}\, p_{ij} = \frac{1}{N} p_{ij}(1-p_{ii}),$$

$$E[dp_{ij}'\, dp_{ii}'] = \frac{1}{N} p_{ij}(1-p_{ij}),$$

$$E[dp_{ii}'\, dp_{fi}'] = E\{[dp_{ij}' + \sum_{g \neq j} dp_{ig}'][dp_{ii}' + \sum_{f \neq i} dp_{fj}']\} =$$

$$= \frac{1}{N} p_{ij}(1-p_{ij}) - \frac{1}{N} p_{ij} \sum_{f \neq i} p_{fj} - \frac{1}{N} p_{ij} \sum_{g \neq j} p_{ig} - \frac{1}{N} [\sum_{f \neq i} p_{fj}][\sum_{g \neq j} p_{ig}] =$$

$$= \frac{1}{N} [p_{ij} - p_{ii}\, p_{ij}].$$

Доказательство, что

$$E\{[dp_{ij}']^h [dp_{ig}']^{2s-1-h}\} \text{ и } E\{[dp_{ij}']^h [dp_{fg}']^{2s-h}\}$$

не содержат членов порядка ниже $\left(\frac{1}{N}\right)^s$, — см. A. Tchouproff, On the mathematical expectation of the moments of frequency distributions, p. 194-200 (Biometrika, vol. XII). Отсюда вытекает, что и

$$E\{[dp_{ij}']^h [dp_{ii}']^{2s-1-h}\}$$

и т. д. не содержат членов порядка ниже $\left(\frac{1}{N}\right)^s$.

§ IV, 2), C). Замечая, что в случае стохастической независимости переменных $\varphi^2 = 0$, а $p_{ij} = p_i p_j$ для всех значений i и j, находим:

$$E[\varphi']^2 = \frac{1}{N}\left\{\sum_i \sum_j [1-p_i][1-p_j]\right\} + \frac{1}{N^2}\left\{\sum_i \sum_j [1-p_i][1-p_j]\right\} + \cdots =$$

$$= \frac{1}{N}[k-1][l-1] + \frac{1}{N^2}[k-1][l-1] + \cdots$$

$$\sigma_{[\varphi']^2}^2 = \frac{1}{N}\left\{4\sum_i \sum_j p_{ij} - 3\sum_i \left[\frac{1}{p_i^3}\left(p_i^2 \sum_j p_j\right)^2\right] - 3\sum_j \left[\frac{1}{p_j^3}\left(p_j^2 \sum_i p_i\right)^2\right] + \right.$$

$$\left. + 2\sum_i \sum_j \left[\frac{1}{p_i p_j}\left(p_j^2 \sum_f p_f\right)\left(p_i^2 \sum_g p_g\right)\right]\right\} + \cdots =$$

$$= \frac{1}{N}[4-3-3+2] + \cdots = \frac{0}{N} + \cdots$$

При выводе общей формулы для рассеяния $[\varphi']^2$ всего удобнее исходить из соотношений

$$\sigma_{[\varphi']^2}^2 = E[\varphi']^4 - \{E[\varphi']^2\}^2,$$

$$[\varphi']^4 = \left\{\sum_i \sum_j \frac{n_{ij}^2}{n_i n_j} - 1\right\}^2 = \left[\sum_i \sum_j \frac{n_{ij}^2}{n_i n_j}\right]^2 - 2\sum_i \sum_j \frac{n_{ij}^2}{n_i n_j} + 1 =$$

$$= \left[\sum_i \sum_j \frac{n_{ij}^2}{n_i n_j}\right]^2 - 2[\varphi']^2 - 1,$$

$$\left[\sum_i \sum_j \frac{n_{ij}^2}{n_i n_j}\right]^2 = \sum_i \sum_j \frac{n_{ij}^4}{n_i^2 n_j^2} + \sum_i \sum_j \sum_{h \neq j} \frac{n_{ij}^2 n_{ih}^2}{n_i^2 n_j n_h} +$$

$$+ \sum_i \sum_j \sum_{f \neq i} \frac{n_{ij}^2 n_{fj}^2}{n_i n_f n_j^2} + \sum_i \sum_j \sum_{f \neq i} \sum_{h \neq j} \frac{n_{ij}^2 n_{fh}^2}{n_i n_j n_f n_h}.$$

Выкладки весьма громоздки, но не представляют особых трудностей, требуя лишь внимания.

§ IV, 3). Исходя из приведенных выше (§ IV, 2)) вспомогательных формул, находим без особых затруднений следующие формулы, имеющие отчасти значение вспомогательных формул, отчасти же представляющие и самостоятельный интерес:

$$Em'_{fg} = E\left\{\sum_i \sum_j p_{ij}' x_i^f y_j^g\right\} = \sum_i \sum_j p_{ij} x_i^f y_j^g = m_{fg},$$

$$dm'_{fg} = \sum_i \sum_j [p_{ij}' - p_{ij}] x_i^f y_j^g = \sum_i \sum_j dp_{ij}' x_i^f y_j^g,$$

$$E[dm_{fg}' dm_{cd}'] = \frac{1}{N}[m_{f+c,g+d} - m_{fg} m_{cd}],$$

$$E[dm'_{f/g}]^2 = \frac{1}{N}[m_{2f/2g} - m_{f/g}{}^2],$$

$$E\mu''_{f/g} = \mu_{f/g}, \quad E[d\mu''_{f/g}\, d\mu''_{c/d}] = \frac{1}{N}[\mu_{f+c/g+d} - \mu_{f/g}\,\mu_{c/d}],$$

$$E[d\mu''_{f/g}]^2 = \frac{1}{N}[\mu_{2f/2g} - \mu_{f/g}{}^2],$$

$$E[dp_{i/j}'\, dm_{0/1}'] = \frac{1}{N} p_{i/j}[y_j - m_{0/1}],$$

$$E[dp_{i/}'\, dm_{0/1}'] = \frac{1}{N} p_{i/}[m_{/1}^{(i)} - m_{0/1}], \quad E[dp_{/j}'\, dm_{0/1}'] = \frac{1}{N} p_{/j}[y_j - m_{0/1}],$$

$$E[dp_{i/}'\, d\mu_{1/1}''] = \frac{1}{N}\{p_{i/}[x_i - m_{1/0}]\,[m_{/1}^{(i)} - m_{0/1}] - p_{i/}\,\mu_{1/1}\},$$

$$E[dp_{i/}'\, d\mu_{0/2}''] = \frac{1}{N}\{p_{i/}[m_{/1}^{(i)} - m_{0/1}]^2 + p_{i/}[\mu_{/2}^{(i)} - \mu_{0/2}]\},$$

$$E[dp_{i/}'\, d\mu_{2/0}''] = \frac{1}{N}\{p_{i/}[x_i - m_{1/0}]^2 - p_{i/}\,\mu_{2/0}\},$$

$$Em_{/1}^{(i)'} = E\left\{\sum_j \frac{n_{i/j}}{n_{i/}}\, y_j\right\} = \sum_j \frac{p_{i/j}}{p_{i/}}\, y_j = m_{/1}^{(i)} \text{ (ср. выше § IV, 1))},$$

$$dm_{/1}^{(i)'} = \sum_j \left[\frac{p_{i/j}'}{p_{i/}'} - \frac{p_{i/j}}{p_{i/}}\right] y_j = \sum_j \frac{p_{i/j}}{p_{i/}}\left[\frac{dp_{i/j}'}{p_{i/j}} - \frac{dp_{i/}'}{p_{i/}} + \cdots\right] y_j,$$

$$E[dm_{/1}^{(i)'}]^2 = \frac{1}{Np_{i/}}\,\mu_{/2}^{(i)} + \cdots,$$

$$E[dm_{/1}^{(i)'}\, dm_{/1}^{(f)'}] = \frac{0}{N} + \cdots,$$

$$E[dp_{i/j}'\, dm_{/1}^{(i)'}] = \frac{1}{Np_{i/}}\, p_{i/j}[y_j - m_{/1}^{(i)}] + \cdots,$$

$$E[dp_{f/j}'\, dm_{/1}^{(i)'}] = \frac{0}{N} + \cdots,$$

$$E[dp_{i/}'\, dm_{/1}^{(i)'}] = \frac{0}{N} + \cdots,$$

$$E[dp_{/j}'\, dm_{/1}^{(i)'}] = \frac{1}{Np_{i/}}\, p_{i/j}[y_j - m_{/1}^{(i)}] + \cdots,$$

$$E[dm_{/1}^{(i)'}\, dm_{0/1}'] = \frac{1}{N}\,\mu_{/2}^{(i)} + \cdots,$$

$$E[dm_{/1}^{(i)'}\, d\mu_{2/0}''] = \frac{0}{N} + \cdots,$$

$$E[dm_{/1}^{(i)\prime}\, d\mu_{0/2}''] = \frac{1}{N}\left\{\frac{1}{p_{i/}}\sum_j p_{i/j}[y_j - m_{0/1}]^3 - \right.$$

$$\left. - [m_{/1}^{(i)} - m_{0/1}]\,\mu_{/2}^{(i)} - [m_{/1}^{(i)} - m_{0/1}]^3\right\} + \cdots =$$

$$= \frac{1}{N}\left\{\mu_{/3}^{(i)} + 2\,[m_{/1}^{(i)} - m_{0/1}]\,\mu_{/2}^{(i)}\right\} + \cdots,$$

$$E[dm_{/1}^{(i)\prime}\, d\mu_{1/1}''] = \frac{1}{Np_{i/}}\left\{\sum_j p_{i/j}[x_i - m_{1/0}][y_j - m_{0/1}]^2 - \right.$$

$$\left. - [m_{/1}^{(i)} - m_{0/1}]\sum_j p_{i/j}[x_i - {}_{1/0}][y_j - m_{0/1}]\right\} + \cdots =$$

$$= \frac{1}{N}\left\{[x_i - m_{1/0}]\sum_j p_{j/}^{(i)}[y_j - m_{0/1}]^2 - [x_i - m_{1/0}][m_{/1}^{(i)} - m_{0/1}]^2\right\} + \cdots =$$

$$= \frac{1}{N}[x_i - m_{1/0}]\,\mu_{/2}^{(i)} + \cdots,$$

$$E\,\mu_{f/g}' = \mu_{f/g} - \frac{1}{N}\left\{\left(\quad + g\right)\mu_{f/g} - \frac{1}{2}f(f-1)\,\mu_{f-2/g}\,\mu_{2/0} - \right.$$

$$\left. - \frac{1}{2}g(g-1)\,\mu_{f/g-2}\,\mu_{0/2} - fg\,\mu_{f-1/g-1}\,\mu_{1/1}\right\} + \cdots,$$

$$E[d\,\mu_{f/g}']^2 = \frac{1}{N}\left\{\mu_{2f/2g} - \mu_{f/2}^{\,2} + f^2\,\mu_{f-1/g}^{\,2}\,\mu_{2/0} + g^2\,\mu_{f/g-1}^{\,2}\,\mu_{0/2} - \right.$$

$$\left. - 2f\,\mu_{f+1/g}\,\mu_{f-1/g} - 2g\,\mu_{f/g+1}\,\mu_{f/g-1} + 2fg\,\mu_{f-1/g}\,\mu_{f/g-1}\,\mu_{1/1}\right\} + \cdots.$$

§ IV, 3) B) и C). Замечая, что в случае прямолинейной регрессии Y по X,

$$m_{/1}^{(i)} - m_{0/1} = r_{1/1}\sqrt{\frac{\mu_{0/2}}{\mu_{2/0}}}\,[x_i - m_{1/0}],$$

находим:

$$\sum_i p_{i/}[x_i - m_{1/0}]\,[m_{/1}^{(i)} - m_{0/1}]^2 = \frac{\mu_{0/2}}{\mu_{2/0}}\,r_{1/1}^{\,2}\,\mu_{3/0} = \mu_{0/2}\sqrt{\mu_{2/0}}\,r_{1/1}^{\,2}\,r_{3/0},$$

$$\sum_i p_{i/}[m_{/1}^{(i)} - m_{0/1}]^4 = \mu_{0/2}^{\,2}\,r_{1/1}^{\,4}\,r_{4/0},$$

$$\sum_i p_{i/}[x_i - m_{1/0}]^2\,[m_{/1}^{(i)} - m_{0/1}]^2 = \mu_{0/2}\,\mu_{2/0}\,r_{1/1}^{\,2}\,r_{4/0},$$

$$\sum_i p_{i/}[x_i - m_{1/0}]\,[m_{/1}^{(i)} - m_{0/1}]^3 = \sqrt{\mu_{2/0}\,\mu_{0/2}^{\,3}}\,r_{1/1}^{\,3}\,r_{4/0}.$$

Исходя из тожества

$$\mu_{2/2} = \sum_i \sum_j p_{i/j} [x_i - m_{1/0}]^2 [y_j - m_{0/1}]^2 =$$

$$= \sum_i \left\{ p_{i/} [x_i - m_{1/0}]^2 \sum_j p_{j/}^{(i)} [y_j - m_{0/1}]^2 \right\} =$$

$$= \sum_i \left\{ p_{i/} [x_i - m_{1/0}]^2 \left(\sum_j p_{j/}^{(i)} [y_j - m_{/1}^{(i)}]^2 + [m_{/1}^{(i)} - m_{0/1}]^2 \right) \right\} =$$

$$= \sum_i p_{i/} [x_i - m_{1/0}]^2 \mu_{/2}^{(i)} + \sum_i p_{i/} [x_i - m_{1/0}]^2 [m_{/1}^{(i)} - m_{0/1}]^2,$$

получаем, таким образом, для случая прямолинейной регрессии Y по X

$$\sum_i p_{i/} [x_i - m_{1/0}]^2 \mu_{/2}^{(i)} = \mu_{2/0} \mu_{0/2} [r_{2/2} - r_{1/1}^2 r_{4/0}],$$

$$\sum_i p_{i/} [m_{/1}^{(i)} - m_{0/1}]^2 \mu_{/2}^{(i)} = \frac{\mu_{0/2}}{\mu_{2/0}} r_{1/1}^2 \sum_i p_{i/} [x_i - m_{1/0}]^2 \mu_{/2}^{(i)} =$$

$$= \mu_{0/2}^2 r_{1/1}^2 [r_{2/2} - r_{1/1}^2 r_{4/0}],$$

$$\sum_i p_{i/} [x_i - m_{1/0}] [m_{/1}^{(i)} - m_{0/1}] \mu_{/2}^{(i)} \sqrt{\mu_{2/0} \mu_{0/2}^3} r_{1/1} [r_{2/2} - r_{1/1}^2 r_{4/0}].$$

Сходным образом, исходя из тожества

$$\mu_{1/3} = \sum_i p_{i/} [x_i - m_{1/0}] \mu_{/3}^{(i)} + 3 \sum_i p_{i/} [x_i - m_{1/0}] [m_{/1}^{(i)} - m_{0/1}] \mu_{/2}^{(i)} +$$

$$+ \sum_i p_{i/} [x_i - m_{1/0}] [m_{/1}^{(i)} - m_{0/1}]^3,$$

находим для случая прямолинейной регрессии Y по X

$$\sum_i p_{i/} [x_i - m_{1/0}] \mu_{/3}^{(i)} = \sqrt{\mu_{2/0} \mu_{0/2}^3} [r_{1/3} - 3 r_{1/1} r_{2/2} + 2 r_{1/1}^3 r_{4/0}],$$

$$\sum_i p_{i/} [m_{/1}^{(i)} - m_{0/1}] \mu_{/3}^{(i)} = \mu_{0/2}^2 r_{1/1} [r_{1/3} - 3 r_{1/1} r_{2/2} + 2 r_{1/1}^3 r_{4/0}],$$

а исходя из тожества

$$\mu_{1/2} = \sum_i p_{i/} [x_i - m_{1/0}] \mu_{/2}^{(i)} + \sum_i p_{i/} [x_i - m_{1/0}] [m_{/1}^{(i)} - m_{0/1}]^2,$$

получаем

$$\sum p_{i/}\,[x_i - m_{1/0}]\,\mu_{/2}^{(i)} = \mu_{0/2}\,\sqrt{\mu_{2/0}}\,[r_{1/2} - r_{1/1}{}^2\,r_{3/0}].$$

Замечая, что

$$\eta_{Y/X}{}^2 = 1 - \frac{1}{\mu_{0/2}} \sum_i p_{i/}\,\mu_{/2}^{(i)},$$

находим:

$$\sum_i p_{i/}\,\mu_{/2}^{(i)} = \mu_{0/2}\,[1 - \eta_{Y/X}{}^2].$$

В случае гомоскедастичной связи Y с X получаем отсюда:

$$\mu_{/2}^{(i)} = \mu_{0/2}\,[1 - \eta_{Y/X}{}^2].$$

Если регрессия Y по X прямолинейна, то $\eta_{Y/X}{}^2 = r_{1/1}{}^2$ и

$$\sum_i p_{i/}\,\mu_{/2}{}^{(i)} = \mu_{0/2}\,[1 - r_{1/1}{}^2];$$

если при этом связь Y с X гомоскедастична, то

$$\mu_{/2}^{(i)} = \mu_{0/2}\,[1 - r_{1/1}{}^2].$$

Подставляя

$$\mu_{/2}^{(i)} = \mu_{0/2}\,[1 - r_{1/1}{}^2]$$

в

$$\sum_i p_{i/}\,[x_i - m_{1/0}]^2\,\mu_{/2}^{(i)} = \mu_{2/0}\,\mu_{0/2}\,[r_{2/2} - r_{1/1}{}^2\,r_{4/0}]$$

и в

$$\sum_i p_{i/}\,[x_i - m_{1/0}]\,\mu_{/2}^{(i)} = \mu_{0/2}\,\sqrt{\mu_{2/0}}\,[r_{1/2} - r_{1/1}{}^2\,r_{3/0}],$$

получаем для случая прямолинейной регрессии Y по X и гомоскедастичной связи Y с X:

$$1 - r_{1/1}{}^2 = r_{2/2} - r_{1/1}{}^2\,r_{4/0},$$

$$0 = r_{1/2} - r_{1/1}{}^2\,r_{3/0}.$$

§ IV, 3), D). Если требуется найти значения

$$EU',\ \sigma_{U'},\ E[U']^2,\ \sigma_{[U']^2},$$

нет надобности выполнять все выкладки четыре раза. Задачу можно осязательно облегчить, исходя из следующих соотношений. Представим U' в виде

$$U' = C + d^{\mathrm{I}} + d^{\mathrm{II}} + \cdots,$$

причем C обозначает сумму членов, не содержащих разностей $dp_{i'}$ и т. д., d^{I} — сумму членов, содержащих эти разности в первой степени, d^{II} — сумму членов, содержащих их во второй степени и т. д. Замечая, что

$$[U']^2 = C^2 + 2Cd^{\mathrm{I}} + \{[d^{\mathrm{I}}]^2 + 2Cd^{\mathrm{II}}\} + \cdots,$$
$$[U']^4 = C^4 + 2C^3 d^{\mathrm{I}} + \{6C^2[d^{\mathrm{I}}]^2 + 4C^3 d^{\mathrm{II}}\} + \cdots,$$

находим:

$$EU' = C + Ed^{\mathrm{II}} + \cdots, \quad U' - EU' = d^{\mathrm{I}} + \{d^{\mathrm{II}} - Ed^{\mathrm{II}}\},$$
$$E[U']^2 = C^2 + \{E[d^{\mathrm{I}}]^2 + 2CEd^{\mathrm{II}}\} + \cdots,$$
$$E[U']^4 = C^4 + \{6C^2 E[d^{\mathrm{I}}]^2 + 4C^3 Ed^{\mathrm{II}}\} + \cdots,$$
$$\sigma_{U'}{}^2 = E[U' - EU']^2 = E[d^{\mathrm{I}}]^2 + \cdots,$$
$$\sigma_{(U')^2}{}^2 = E\{[U']^2 - E[U']^2\}^2 = E[U']^4 - \{E[U']^2\}^2 = 4C^2 E[d^{\mathrm{I}}]^2 + \cdots.$$

Если выкладки не идут дальше членов порядка $\frac{1}{N}$, то для нахождения математического ожидания и рассеяния $[U']^2$ достаточно располагать значениями $E[d^{\mathrm{I}}]^2$ и Ed^{II}, которые необходимы для нахождения математического ожидания и рассеяния U'.

Сходным образом может быть облегчено вычисление рассеяния разности

$$[\zeta_{r/x}']^2 = [\eta_{r/x}']^2 - [r_{1/1}']^2.$$

Опуская для упрощения начертаний пометки при η', r' и ζ' и исходя из аналогичных выше использованному разложений

$$r' = C + d^{\mathrm{I}} + d^{\mathrm{II}} + \cdots,$$
$$[\eta']^2 = K + \Delta^{\mathrm{I}} + \Delta^{\mathrm{II}} + \cdots,$$

находим:

$$[r']^2 = C^2 + 2Cd^{\mathrm{I}} + \{[d^{\mathrm{I}}]^2 + 2Cd^{\mathrm{II}}\} + \cdots,$$
$$[\eta']^2[r']^2 = KC^2 + \{C^2\Delta^{\mathrm{I}} + 2KCd^{\mathrm{I}}\} + \{C^2\Delta^{\mathrm{II}} + 2C\Delta^{\mathrm{I}}d^{\mathrm{I}} + K[d^{\mathrm{I}}]^2 + 2KCd^{\mathrm{II}}\} + \cdots.$$
$$E\{[\eta']^2 - E[\eta']^2\}\{[r']^2 - E[r']^2\} = E\{[\eta']^2[r']^2\} - \{E[\eta']^2\}\{E[r']^2\} =$$
$$= 2CE[\Delta^{\mathrm{I}}d^{\mathrm{I}}] + \cdots,$$
$$\sigma_{[\zeta']^2}{}^2 = E\{([\eta']^2) - E([\eta']^2) - ([r']^2 - E[r']^2)\}^2 =$$
$$= \sigma_{[\eta']^2}{}^2 + \sigma_{[r']^2}{}^2 - 2E\{[\eta']^2 - E[\eta']^2\}\{[r']^2 - E[r']^2\} =$$
$$= \sigma_{[\eta']^2}{}^2 + \sigma_{[r']^2}{}^2 - 4CE[\Delta^{\mathrm{I}}d^{\mathrm{I}}] \cdots.$$

§ IV, 4). Полагая

$$EZ' = m_{1/0}, \quad EZ'W' = m_{1,1}, \quad EW' = m_{0/1}, \quad E[W']^2 = m_{0/2},$$

и замечая, что

$$\frac{1}{C}EY' - \frac{1}{C^2}EZ'(W' - C) = \frac{2m_{1,0}}{C} - \frac{m_{1,1}}{C^2},$$

$$\frac{1}{C}EZ' - \frac{1}{C^2}EZ'[W' - C] + \frac{1}{C^2}E[W' - C]^2 = \frac{2m_{1,0}}{C} - \frac{2m_{0,1}}{C} - \frac{m_{1,1}}{C^2} + \frac{m_{0,2}}{C^2} + 1,$$

находим из условия

$$\frac{\partial\left[\frac{2m_{1,0}}{C} - \frac{m_{1,1}}{C^2}\right]}{\partial C} = 2\left[\frac{m_{1,1}}{C^3} - \frac{m_{1,0}}{C^2}\right] = 0$$

значение

$$C = \frac{m_{1,1}}{m_{1,0}}.$$

а из условия

$$\frac{\partial\left[\frac{2m_{1,0}}{C} - \frac{2m_{0,1}}{C} - \frac{m_{1,1}}{C^2} + \frac{m_{0,2}}{C^2} + 1\right]}{\partial C} = 2\left[\frac{m_{0,1}}{C^2} - \frac{m_{1,0}}{C^2} + \frac{m_{1,1}}{C^3} - \frac{m_{0,2}}{C^3}\right] = 0$$

значение

$$C = \frac{m_{0,1} - m_{1,0}}{m_{0,2} - m_{1,1}}.$$

§ IV, 5). Вывод формул для $\mu_{2,0}$, $\mu_{0,2}$, $r_{1,1}$ для случая, когда обе переменные способны принимать лишь по два разных значения, указан выше (глава четвертая, § VI). Замечая, что

$$p^3 + (1-p)^3 = 1 - 3p(1-p),$$

находим:

$$\mu_{4,0} = [p_{1'}p_{2'}^4 + p_{2'}p_{1'}^4][x_1 - x_2]^4 = p_{1'}p_{2'}[1 - 3p_{1'}p_{2'}][x_1 - x_2]^4,$$

$$r_{4,0} = \frac{\mu_{4,0}}{\mu_{2,0}^2} = \frac{1}{p_{1'}p_{2'}} - 3.$$

Замечая, что

$$p_{1,1} = \delta + p_{1'}p_{1'}, \quad p_{1,2} = -\delta + p_{1'}p_{2'}$$

и т. д., получаем:

$$\mu_{3,1} = [x_1 - x_2]^3[y_1 - y_2][p_{1,1}p_{2'}^3p_{2'} - p_{1,2}p_{2'}^3p_{1'} - p_{2,1}p_{1'}^3p_{2'} + p_{2,2}p_{1'}^3p_{1'}] =$$

$$= [x_1 - x_2]^3[y_1 - y_2]\{\delta[p_{2'}^3p_{2'} + p_{2'}^3p_{1'} + p_{1'}^3p_{2'} + p_{1'}^3p_{1'}] +$$

$$+ p_{1'}p_{2'}p_{1'}p_{2'}[p_{2'} - p_{2'} - p_{1'} + p_{1'}]\} = [x_1 - x_2]^3[y_1 - y_2]\delta[p_{1'}^3 + p_{2'}^3],$$

$$r_{3,1} = r_{1,1}r_{4,0}$$

$$\mu_{2,2} = [x_1 - x_2]^2[y_1 - y_2]^2\{\delta[p_{1'} - p_{2'}][p_{1'} - p_{2'}] + p_{1'}p_{2'}p_{1'}p_{2'}\},$$

$$r_{2,2} = 1 + \frac{[p_{1'} - p_{2'}][p_{1'} - p_{2'}]\delta}{p_{1'}p_{2'}p_{1'}p_{2'}}.$$

Так как

$$r_{3/1} = r_{1.1}\, r_{4\,0} \quad \text{и} \quad r_{1/3} = r_{1/1}\, r_{0/4},$$

то полученные значения параметров r при вычислении математического ожидания и рассеяния $r_{1/1}'$ могут быть подставлены не в общие формулы, а в формулы, полученные для случая, когда обе линии регрессии прямые.

Глава седьмая.

§ I. Применительно к схеме невозвращаемого билета находим (ср. глава шестая, § II, 1)):

$$E[dp_{i/}']^2 = \frac{A-N}{A-1}\frac{1}{N} p_{i/}(1-p_{i/}),$$

$$E[dp_{i/}']^2 = \frac{A-N}{A-1}\frac{1}{N} p_{i/}(1-p_{i/}),$$

$$E[dp_{/j}']^2 = \frac{A-N}{A-1}\frac{1}{N} p_{/j}(1-p_{/j}).$$

Замечая, что

$$E\, n_{i/} n_{f/g} = \sum_h P_h h(N-h)\frac{p_{f/g}}{1-p_{i/}} = \frac{p_{f/g}}{1-p_{i/}}\left\{N\sum_h P_h h - \sum_h P_h h^2\right\} =$$

$$= \frac{p_{f/g}}{1-p_{i/}}\left\{N^2 p_{i/} - N^2 p_{i/}\frac{B-1}{A-1} - N p_{i/}\left[1-\frac{B-1}{A-1}\right]\right\} =$$

$$= \frac{A}{A-1} N(N-1) p_{i/} p_{f/g},$$

находим

$$E[dp_{i/}'\, dp_{f/g}'] = \frac{1}{N^2}\{E[n_{i/} n_{f/g}] - [E n_{i/}][E n_{f/g}]\} = -\frac{A-N}{A-1}\frac{1}{N} p_{i/} p_{f/g};$$

отсюда получаем

$$E[dp_{i/}'\, dp_{f/}'] = -\frac{A-N}{A-1}\frac{1}{N} p_{i/} p_{f/},$$

$$E[dp_{i/}'\, dp_{i/}'] = \frac{A-N}{A-1}\frac{1}{N} p_{i/}[1-p_{i/}],$$

$$E[dp_{i/}'\, dp_{/j}'] = \frac{A-N}{A-1}\frac{1}{N}[p_{i/} - p_{i/} p_{/j}].$$

Математические ожидания вторых степеней разностей $dp_{i/}'$ и т. д. оказываются, таким образом, в случае извлечений без возвращения равными произведению соответственных математических ожиданий для случая независимых испытаний на $\frac{A-N}{A-1}$.

Исходя из разложения функции эмпирических значений в ряд (см. глава шестая, § IV, 3), D)).

$$U' = C + d^{I} + d^{II} + \cdots$$

и обозначая через $\frac{1}{N} D^{II}$ математическое ожидание d^{II} в случае независимых испытаний, мы находим для случая извлечений без возвращения

$$EU' = C + \frac{A-N}{A-1}\frac{1}{N} D^{II} + \cdots$$

Сходным путем убеждаемся, что в случае извлечений с прикладыванием

$$E[dp_{ij}']^2 = \frac{A+N}{A+1}\frac{1}{N} p_{ij}(1-p_{ij})$$

и т. д. и что, следовательно, в этом случае

$$EU' = C + \frac{A+N}{A+1}\frac{1}{N} D^{II} + \cdots$$

При этом необходимо, однако, иметь в виду, что $\frac{A+N}{A+1} > 1$, а при сравнительно небольшом A $\frac{A+N}{A+1}$ может быть и порядка величины N. Таким образом, второй член разложения может даже при большом N быть того же порядка величины, как предшествующий, и приближение, получаемое в результате использования этого приема при наличии связи между испытаниями, может быть иллюзорным.

§ II. Замечая, что для всех трех схем

$$\mu_{0/2}' = \sum_j p_j' [y_j - m_{0/1}']^2 = \sum_j p_j' [y_j - m_{0/1}]^2 -$$

$$- 2\sum_j p_j' [m_{0/1}' - m_{0/1}][y_j - m_{0/1}] + [m_{0/1}' - m_{0/1}]^2 = \mu_{0/2}'' - [dm_{0/1}']^2$$

$$E\mu_{0/2}'' = \mu_{0/2}$$

и что

$$E[dm_{0/1}']^2 = \frac{1}{N}\mu_{0/2}$$

для схемы возвращаемого билета,

$$E[dm_{0/1}']^2 = \frac{A-N}{A-1}\frac{1}{N}\mu_{0/2}$$

для схемы невозвращаемого билета,

$$E[dm_{0/1}']^2 = \frac{A+N}{A+1}\frac{1}{N}\mu_{0,2}$$

для схемы прикладываемого билета,
находим:

$$E\mu_{0/2}' = \mu_{0/2} - \frac{1}{N}\mu_{0/2} = \frac{N-1}{N}\mu_{0/2}$$

для схемы возвращаемого билета,

$$E\mu_{0/2}' = \mu_{0/2} - \frac{A-N}{A-1}\frac{1}{N}\mu_{0/2} = \frac{A}{A-1}\frac{N-1}{N}\mu_{0/2}$$

для схемы невозвращаемого билета,

$$E\mu_{0/2}' = \mu_{0/2} - \frac{A+N}{A+1}\frac{1}{N}\mu_{0/2} = \frac{A}{A+1}\frac{N-1}{N}\mu_{0/2}$$

для схемы прикладываемого билета.

§ III, 1). Замечая, что

$$\sum_{f=1}^{N}\{x^{[f]'} - x_0'\}\{y^{[f]'} - y_0'\} =$$

$$= \sum_{f=1}^{N}\{x^{[f]'} - m_{1/0}\}\{y^{[f]'} - m_{0/1}\} - N[x_0' - m_{1/0}][y_0' - m_{0/1}]$$

и что, в предположении взаимной независимости испытаний,

$$E[x_0' - m_{1/0}][y_0' - m_{0/1}] = \frac{1}{N^2}E\{\sum_{f=1}^{N}[x^{[f]'} - m_{1/0}]\}\{\sum_{f=1}^{N}[y^{[f]'} - m_{0/1}]\} =$$

$$= E\{\sum_{f=1}^{N}[x^{[f]'} - m_{1/0}][y^{[f]'} - m_{0/1}] + \sum_{f=1}^{N}\sum_{g\neq 1}[x^{[f]'} - m_{1/0}][y^{[g]'} - m_{0/1}]\} =$$

$$= \frac{1}{N}\mu_{1/1}.$$

находим:

$$E\{\sum_{f=1}^{N}[x^{[f]'} - x_0'][y^{[f]'} - y_0']\} = N\mu_{1/1} - \mu_{1/1} = (N-1)\mu_{1/1}.$$

§ III, 2). Замечая, что $x_0^{[h]'}$ и $y_0^{[h]'}$ — средние арифметические значений X и Y для n взаимно независимых испытаний h-ой серии, находим для каждой из r серий

$$E[x_0^{[h]'} - m_{1,0}][y_0^{[h]'} - m_{0,1}] = \frac{1}{n}\mu_{1,1}$$

и отсюда

$$E\{\sum_{h=1}^{r}[x_0^{[h]'} - x_0'][y_0^{[h]'} - y_0']\} = (r-1)\frac{1}{n}\mu_{1,1}.$$

§ III, 2). Обозначая числителя Q через Z, а знаменателя через T, легко доказать, что, при любом k, $EZT^k = ET^{k+1}$ и, следовательно, при $k = -1$,

$$E\frac{Z}{T} = EQ = 1.$$

Замечая, что

$$\sum_{f=1}^{rn}[x^{[f]'} - x_0'][y^{[f]'} - y_0'] = \sum_{f=1}^{rn} x^{[f]'}y^{[f]'} - rn\,x_0'\,y_0'$$

и что закон связи остается неизменным, а все испытания взаимно независимы, находим:

$$ET^{k+1} = ET\cdot T^k =$$

$$= \frac{1}{rn-1}\{E[\sum_{f=1}^{rn} x^{[f]'}y^{[f]'}]T^k - \frac{1}{rn}E[\sum_{f=1}^{rn}x^{[f]'}][\sum_{f=1}^{rn}y^{[f]'}]T^k\} =$$

$$= \frac{1}{rn-1}\{rn\,Ex^{[f]'}y^{[f]'}T^k - \frac{1}{rn}E[\sum_{f=1}^{rn}x^{[f]'}y^{[f]'} +$$

$$+ \sum_{f=1}^{rn}\sum_{g\neq f} x^{[f]'}y^{[g]'}T^k\} = Ex^{[f]'}y^{[f]'}T^k - Ex^{[f]'}y^{[g]'}T^k.$$

Сходным образом доказываем, что

$$EZT^k = Ex^{[f]'}y^{[f]'}T^k - Ex^{[f]'}y^{[g]'}T^k$$

и убеждаемся, что

$$EZT^k = ET^{k+1}.$$

Обзор литературы. [1])

В предисловии я указал, что настоящий труд не ставит себе целью ввести начинающего в практику статистического исследования. Для этого лучшим пособием служит превосходная книга G. Udny Yule'a „An introduction to the theory of statistics“ (London 1911; впоследствии много раз переиздававшаяся).

В качестве полезного руководства можно затем отметить учебник Truman L. Kelley „Statistical method“ (New York 1923), отличающийся как наглядно составленной сводкой методов и показателей, предложенных различными исследователями, так и богатыми литературными указаниями.

Тем из читателей, которые охотнее пользуются немецким языком, можно указать книгу E. Czuber'a „Die statistischen Forschungsmethoden“ (Wien 1921). Руководство Чубера примыкает к Юл'овскому, хотя все же отстает от своего образца в отношении законченности и дидактической силы и вместе с тем предполагает более высокий уровень математических познаний; однако и мало подготовленный читатель вынесет из изучения первых двух отделов много для себя полезного; что же касается последнего отдела, опирающегося на теорию вероятностей, то без достаточной подготовки за него лучше не браться.

Руководства, составленные немецкими статистиками, обычно обходят интересующие нас вопросы, так как „математические“ приемы работы в среде германских профессиональных статистиков до сих пор еще почитаются скорее причудой любителей, чем общим достоянием всех статистически работающих исследователей. Между тем, немецкая литература по тем отраслям естествознания, которые применяют статистические методы, содержит целый ряд подчас весьма удачных руководств по применению „математических“ приемов исследования. На первом месте в качестве введений, дающих систематические обзоры методов, следует назвать следующие труды: W. Johannsen „Elemente der exakten Erblichkeitslehre. Mit Grundzügen der biologischen Variationsstatistik“ (2 Aufl., Jena 1913); A. Lang „Die experimentelle Vererbungslehre in der Zoologie seit 1900“. С отделом „Anfangsgründe der Biometrik der Variation und Korrelation“ (Erste Hälfte, Jena 1914) и W. Wirth „Spezielle psychophysische Massmethoden“ (Handbuch der biologischen Arbeitsmethoden, herausgegeben von E. Abderhalden, Abt. 6, A, Heft 1. Berlin — Wien 1920).

---

1) Вследствие болезни автор не мог переработать текст этого отдела для русского издания. Согласно указаниям автора, сделанным в письме к Н. С. Четверикову, последним был выполнен перевод настоящей главы с немецкого издания, дополненный некоторыми примечаниями, помеченными инициалами Н. С. Ч. [Изд.].

Особенно много интересного и поучительного для статистика, не интересующегося специально биологией, содержит в себе книга Иогансена. Книгу Ланга, которая в соответственной своей части (стр. 201-464) опирается, главным образом, на изложение Иогансена, можно в особенности рекомендовать начинающему в виду подробно разработанных числовых примеров. К сожалению, как у того, так и у другого автора теория корреляции излагается на основе способа наименьших квадратов. Книга Вирта предполагает значительно больший объем математической подготовки у читателя, но вместе с тем гораздо шире охватывает новые построения в области теории корреляции; самое изложение, принимающее подчас характер монографического исследования, рассчитано на более тесный круг читателей, примыкающих по своей специальности к автору. Для тех, кто интересуется статистикой, работая в иных областях, этот труд окажется менее доступным, чем „Основы“ Иогансена.

[На русском языке наилучшим руководством для ознакомления с теорией корреляции является книга Е. Е. Слуцкого „Теория корреляции и элементы учения о кривых распределения“ (Известия Киевского Коммерческого Института, кн. XVI, Киев 1912, а также отдельным оттиском), хотя и в ней автор подводит читателя к коэффициенту корреляции со стороны принципа наименьших квадратов. Книга Е. Е. Слуцкого весьма полно отображает современное ей состояние теории корреляции и предполагает повышенный уровень математической подготовки][1]).

Тем статистикам, которым приходится производить массовые вычисления коэффициентов корреляции, корреляционных отношений и т. п. я настоятельно советую просмотреть новейшую американскую литературу, уделяющую много внимания упрощению и усовершенствованию техники вычисления. Об относящихся сюда работах легче всего осведомиться в библиографических обзорах в Journal of the American Stasistical Association: в этом журнале встречаются и самостоятельные статьи на эти темы; как на пример укажем на A. I. Holzinger „A combination form for calculating the correlation coefficient and ratios“ (См. XVIII, № 141, стр. 623—627, 1923).

К главе первой.

Изобретением статистических приемов уловления связей, которые ныне причисляются к „элементарным“-усердно занимались во второй четверти и ближе к середине XIX столетия; в особенности следует упомянуть о французах Guerry и Dufau. С большой любовью останавливаются на этих приемах авторы русских курсов по статистике, причем до недавнего времени их считали, примыкая к „Логике“ Д. С. Милля, за разновидности метода сопутствующих изменений. В качестве примера, взятого из новой немецкой литературы, укажем на статью F. Tönnies'a „Eine neue Methode der Vergleichung statistischer Reihen“ (Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft, Jahrgang 33, 1909).

---

1) [Прим. Н. С. Ч.].

§ II, B).—Сходно со мною к коэффициенту корреляции подходит от Фехнеровских показателей также и L. March в его интересной статье „Comparaison numérique de courbes statistiques“ (Journal de la Société de Statistique de Paris, 1905)[1]).

C). — Показатель ρ был введен американским психологом C. Spearman'ом. Пирсон ввел проблему корреляции рангов в общую систему учения о корреляции и дал доказательство, что в случае нормальной корреляции коэффициент корреляции r связан с показателем ρ Спирмэн'а соотношением

$$r = 2 \sin \left(\frac{\pi}{6} \rho \right).$$

Ср. K. Pearson „On further methods of determining correlation“ (Drapers' Company Research Memoirs, Biometric Series, IV, 1907).

К главе второй.

Изложенное здесь понимание существа дела и задач статистического корреляционного исследования развито подробнее в моих „Очерках по теории статистики“ (2-е изд., С.-Петербург 1910); ср. A. Tschuprow „Die Aufgaben der Theorie des Statistik“ (Jahrbuch für Gesetzgebung, Verwaltung u. s. w., 1903), а также А. Кауфман „Теория и методы статистики“; часть I, глава V. (4-е изд., Москва 1922).

К главе третьей.

§ IV. Изложение, данное мною, соответствует, в априорном аспекте, „обобщенной идеи корреляции“ („generalised idea of correlation“), развитой K. Pearson'ом в его мемуаре „On the general theory of skew correlation and non-linear regression“ (Drapers' Company Research Memoirs, Biometric Series II, 1905). Эту основоположную работу Пирсона G. Dunker реферировал в Biologisches Zentralblatt (Bd. 42, 1922). Первоначально Пирсон, примыкая к Ф. Гальтону исходил из вышеприведенного (§ IV, 3) определения коррелированности; ср. K. Pearson „Regression, heredity and panmixia“, стр. 256-257 (Phil. Trans., A, vol. 187, 1897). Более подробные указания литературы см. А. А. Чупров „Стохастические основы теории статистики“, § II, 3 (Сборник „Памяти Николая Алексеевича Каблукова“, Москва 1925. Также: Вестник Статистики, кн. XXII, Москва 1925).

§ V. — То обстоятельство, что при обработке искаженных случайными ошибками наблюдений над функционально связанными величинами получаются несовместимые уравнения (если в одном случае за независимую переменную принять X, а в другом Y) — впервые, повидимому, подметил J. Kleiber „Über

---

1) См. русский перевод в Статистическом Бюллетени Киевского Губ. Стат. Бюро, № 3, № 4-5, (Киев 1922). [Прим. Н. С Ч.].

den Abrundungsfehler meteorologischer Zahlen“ (Meteor. Zeitschr., Bd. 5, 1888). Клейбер, однако, не дал себе отчета в причинах такого расхождения результатов и пытался свести ее на неточность округленных чисел. Что дело здесь не в округлении чисел показал Срезневский в следующем номере того же журнала; Sresnewski „Über Abrundungsfehler“ (Meteor. Zeitschr., Bd. 6, 1889). Строгую постановку вопроса и его решение при известных ограничительных допущениях дал K. Pearson „On lines and planes of closest fit to systems of points in space“ (Phil. Mag., Ser. 6, vol. 2, 1901). Отчасти примыкая к Пирсону, отчасти независимо от него, эту задачу решали в том же направлении различные исследователи; ср., например, работы C. Gini „Sull' interpolazione di una retta quando i valori della variabile independente sono affetti da errori accidentali“ и L. J. Reed „Fitting straight lines“ во втором томе Metron'a (1921). Обособленное место занимает Вирт, который видит здесь не задачу отыскания закона функциональной зависимости между неслучайными переменными, а прием уловления существенного — по его мнению — свойства стохастической связи между двумя случайными переменными (если выразить его мысль в принятых мною терминах); ср. W. Wirth „Spezielle psycho-physische Massmethoden“ (Handbuch der biologischen Arbeitsmethoden, hgg. von E. Abderhalden, Berlin-Wien 1920) и примыкающие работы E. Czuber'a и W. Wirth'a в Archiv für die gesamte Psychologie (Bde 41, 1921 и 44, 1923), в особенности статью W. Wirth'a „K. Pearsons angepasste Gerade (Best fitting straight line) und die mittlere Regression“ (Archiv für die ges. Psychologie, Bd. 44).

Различие между статистическим пониманием корреляции и точкой зрения тех математиков-нестатистиков, которые пришли к тем же формулам, очень поучительно выясняется в статье K. Pearson'a „Notes on the history of correlation“ (Biometrika, vol. XIII) путем сравнения работ Гаусса, Браве и Гальтона в области теории корреляции.

К главе четвертой.

Текст опирается на мою статью во втором томе Трудов русских ученых за границей (Берлин, 1923). Проблемы, рассмотренные в четвертой и пятой главах обычно разрабатываются монографически в связи с теми вопросами, которые в системе моего изложения отнесены к главе шестой. Вследствие этого представляется более целесообразным соответствующие литературные указания также приурочить к главе шестой.

§ II, 1) (ср. главу пятую, § VI) — W. F. Sheppard в его мемуаре „On the application of the theory of error to cases of normal distribution and normal correlation“, § 21 (Phil. Trans.. A, vol. 192, 1899) предложил сравнивать разность $p_{ij} - p_{i}' p_{j}'$ с ее квадратической ошибкой для того, чтобы таким путем устанавливать отсутствие независимости между переменными; причем этот прием сам Шеппард рассматривает как „an extension of the ordinary method (used largely by Prof. Lexis and Prof. Edgeworth) for testing the stability of stati-

stical ratios“ [1]). Шеппард ограничивается рассмотрением ряда отдельных значений разностей, не объединяя их в общий показатель. Показатель связи φ (Mean Square Contingency) ведет начало от K. Pearson'a „On the theory of contingency and its relation to association and normal correlation“ (Drapers' Company Research Memoirs, Biometric Series, I, 1904).

§ II, 1) и § VII. (Ср. главу пятую, § VII).—Значение коэффициента корреляции для четырехраздельной (тетрахорической) таблицы определяется способом, сходным с вышеуказанным, но в предположении, что переменные X и Y могут принимать лишь значения 0 и 1; ср. G. U. Yule „On the methods of measuring association between two attributes“, стр. 596, 606 и сл. (Journal of the Royal Statistical Society, vol. 75, 1912); L. v. Bortkiewicz библиографическая заметка о книге Charlier, стр. 346—347 (Nordisk Statistisk Tidskrift, Bd. 1, 1922); ср. далее C. Gini „Indici di omofilia e loro relazioni col coefficienti di correlazione e con gli indici di attrazione“, стр. 600 и прим. 1-е к стр. 602 (Atti del Reale Instituto Veneto, t. 74, parte seconda, Venezia 1915). C. V. L. Charlier „Vorlesungen uber die Grundzüge der mathematischen Statistik“ стр. 105-114 (1920) получает то же выражение для коэффициента корреляции обходным путем через уравнение корреляционной поверхности. S. D. Wicksell „Some theorems in the theory of probability, with special reference to their importance in the theory of homograde correlation“ (Svenska Actuarieföreningens Tidskrift, 1916) приходит к нему же через уравнения регрессий.

Подход к проблеме, изложенный в настоящей книге и исходящий из допущения, что обе переменные могут принимать лишь по два разных значения, следует отличать от такой постановки вопроса, которая предполагает произвольно большое число возможных значений, стянутых в две группы для каждой из переменных, причем исследователю ставится задача показать, каким путем можно найти значения показателей, характеризующих закон связи, если материал сгруппирован в подобную тетрахорическую таблицу. Такова именно позиция Пирсона, когда он исследует тетрахорическую схему. Для того, чтобы можно было эту задачу решить, необходимо присовокупить какое-либо допущение относительно закона связи; понятно, что на первую очередь здесь выдвинулось предположение нормальной корреляции; ср. K. Pearson „On the correlation of characters not quantitatively measurable“ (Phil. Trans., A, vol. 195, 1901); K. Pearson „On a novel method of regarding the association of two variates classed solely in alternate categories“ (Drapers' Company Research Memoirs, Biometric Series, VII, 1912); K. Pearson and D. Heron „On theories of association“ (Biometrika, vol. IX); ср. W. F. Sheppard „On the application of the theory of error to cases of normal distribution and normal correlation“ (Phil. Trans., A, vol. 192, 1899).

То, что вычисление Mean Square Contingency для четырехраздельной таблицы приводит к той же формуле, было обнаружено Пирсоном одновременно с построением самого понятия Mean Square Contingency; ср. K. Pearson

---

1) «Обобщение обычного метода (широко использованного проф. Лексисом и проф. Эджвортом) для определения устойчивости статистических относительных чисел». [Прим. Н. С. Ч.].

„On the theory of contingency and its relation to association and normal correlation", стр. 21 (Drapers' Company Research Memoirs, Biometric Series, I, 1904).

§ II, 2). — Разнообразные искусственные приемы, которыми пользуются статистики, имея дело с качественными признаками, интересно изложены в заключительных главах (XVI — XIX) руководства A. Niceforo „Il metodo statistico. Teoria e applicazione alle scienze naturali, alle scienze sociali, e all'arte" (Messina 1923).

§ III, 1). — Название „problème des moments" исходит от Stieltjes'a „Recherches sur les fractions continues", стр. 48 (Annales de la faculté de Toulouse, 1894).

Новейшую литературу вопроса см. у G. Pólya „Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem" (Math. Zeitschrift, Bd. 8, Berlin 1920) и у M. Riesz „Sur le problème des moments et le théorème de Parseval correspondant" (Skandinavisk Aktuarietidskrift, 1924). Ср. также J. F. Steffensen „Matematisk iagttagelseslaere", стр. 41—42 (Kobenhavn 1923).

§ III, 1). — Неравенство $r_{1/1}^2 \leqq 1$ может быть выведено различными способами. Приведенное в тексте доказательство воспроизводит вывод Юла; ср. G. U. Yule „On the significance of Bravais' formulae for regression ... in the case of skew correlation" стр. 482 (Proc. Roy. Soc., vol. 60, 1897); G. U. Yule „On the theory of correlation", стр. 820 (Jour. Roy. Stat. Soc., vol. 60, 1897). Можно исходить также из соотношения:

$$E\left[\frac{X-m_{1/0}}{\sigma_X} \mp \frac{Y-m_{0/1}}{\sigma_Y}\right]^2 \geqq 0,$$

(ср. L. Tr. Kelley „Statistical method" стр. 191, New-York 1923) или же из

$$\mu_{2/0}\,\mu_{0/2} - \mu_{1/1}^2 = \sum_i \sum_j [(x_i - m_{1/0})(y_j - m_{0/1}) - (x_j - m_{1/0})(y_i - m_{0/1})]^2$$

(ср. A. Bowley „Elements of statistics", 4 ed., стр. 354). Можно взять за исходное положение также и тождество (см. мою работу „Исходная задача математической теории приемов статистического исследования связи между двумя случайными переменными", стр. 298. Труды русских ученых за границей, т. II. Берлин 1923)

$$\mu_{2/0}\,\mu_{0/2} - \mu_{1/1}^2 = \mu_{2/0} \sum_i p_{i/}\,\mu_{/2}^{(i)} +$$

$$+ \sum_i \sum_f \sum_h p_{i/}^2\, p_{f/}\, p_{h/} \{[x_i - x_f]\,[m_{/1}^{(i)} - m_{/1}^{(h)}] - [x_i - x_h]\,[m_{/1}^{(i)} - m_{/1}^{(f)}]\}^2 +$$

$$+ \sum_i \sum_g \sum_f \sum_h p_{i/}\, p_{g/}\, p_{f/}\, p_{h/} \{[x_i - x_g]\,[m_{/1}^{(f)} - m_{/1}^{(h)}] -$$

$$- [x_f - x_h]\,[m_{/1}^{(i)} - m_{/1}^{(g)}]\}^2,$$

которое вскрывает и некоторые соотношения между коэффициентом корреляции и корреляционным отношением. К неравенству $r_{1/1}^2 \leqq 1$ можно, наконец, притти и окольным путем, доказав, что

$$r_{1/1}^2 \leqq \eta_{Y/X}^2, \text{ а } \eta_{Y/X}^2 \leqq 1.$$

(Ср. главу четвертую, § IV, 2) и § IV, 3). Последний путь доказательства особенно удобен в случае более чем двух переменных, когда мы хотим подобное же предложение установить для так назыв. коэффициента совокупной корреляции.

Такими же способами можно доказывать и относительно эмпирического коэффициента корреляции (ср. главу пятую, § II, 2), § IV, § V), что он по своей абсолютной величине не может превысить единицы.

§ III, 2). (Ср. главу пятую, § III). — Ср. K. Pearson „On the general theory of skew correlation and non-linear regression“ (Drapers' Company Research Memoirs, Biometric Series, II, 1905); ср. K. Pearson „On a general method of determining the successive terms in a skew regression line“ (Biometrika, vol. XIII).

§ III, 4). (Ср. главу пятую, § III). — Изложение, приведенное в тексте, соответствует — в плоскости априорных построений — воззрениям Юл'а, как они выявились в его работах: G. U. Yule „On the significance of Bravais' formulae for regression etc. in the case of skew correlation“ (Proc. Roy. Soc., vol. 60, 1897); G. U. Yule „On the theory of correlation“ (Jour. Roy. Stat. Soc., vol. 60, 1897).

§ IV, 3). — Эксперименты с игральными костями, часть которых не участвует при повторном подбрасывании, могут быть превосходно использованы преподавателем для наглядного выяснения смысла коэффициента корреляции также, как и взаимного соотношения между эмпирическим коэффициентом корреляции и лежащим в его основе априорным. Ценнейший материал для подобных иллюстраций дает статья A. D. Darbishire'a „Some tables for illustrating statistical correlation“ (Mem. and Proc. Manchester Lit. and Phil. Soc., vol. 51, 1907).

К главе пятой.

Ограничившись рассмотрением прерывно изменяющихся переменных, мы получили возможность при нашем теоретическом анализе исходить из допущения, что, обрабатывая значения переменных, мы не сводим их в классы. Вопросы, возникающие при группировке по классам, которые входят существенной частью в теорию правильной обработки непрерывно изменяющихся переменных, здесь — в случае прерывной изменчивости — представляют лишь узко-технический интерес. Поэтому я и не останавливаюсь на этой проблеме, т. е. на вопросе о так наз. „Шеппардовских поправках“. О них см. работу E. Pairman and K. Pearson „On corrections for the moment-coefficients of limited range frequency distributions when there are finite or infinite ordinates and any slopes at the terminals of the range“ (Biometrika, vol. XII).

К главе шестой.

§ I и § II. — Ср. А. А. Чупров „Основные задачи стохастической теории статистики“ (Сборник „Памяти Николая Алексеевича Каблукова“, Москва 1925. а также Вестник Статистики, кн. XIX, Москва 1925).

§ III, 2). (Ср. главу четвертую, § II; главу пятую, § VII). — Величина δ' названа Пирсоном „the transfer per unit“ [1]); см. K. Pearson „On the correlation of characters not quantitatively measurable“ стр. 14 (Phil. Trans., A, vol 195, 1901). Для ознакомления с теорией показателей, опирающихся на величину δ, кроме трудов Шеппард'а и Пирсон'а, отмеченных в указателе к главе четвертой § II, 1), следует прежде всего обратиться к работам Юл'а, из которых упомянем лишь две: G. U. Yule „On the association of attributes in statistics“ (Phil. Trans., A, vol. 194, 1900) и G. U. Yule „On the methods of measuring association between two attributes“ (Jour. Roy. Stat. Soc., vol. 75, 1912). Ср. А. А. Чупров „О математическом ожидании частного двух взаимно зависимых случайных переменных“. (Труды русских ученых за границей, т. I, стр. 240-271. Берлин, 1922); В. И. Романовский „О вероятностях связанных признаков и о их применении в статистике“. (Вестник Статистики, кн. XII, Москва 1923).

§ IV. — Задачей вычисления $E\frac{X}{Y}$ занимался K. Pearson в статье „On a form of spurious correlation which may arise when indices are used in the measurment of organs“ (Proc. Roy. Soc., vol. 60, 1897); ср. K. Pearson „On the constants of index distributions as deduced from the like constants for the components of ratio, with special reference to the opsonic index“ (Biometrika, vol. VII). Впоследствии ею занимались и другие исследователи, которые сходным же образом, не зная статей Пирсона, выводили некоторые из полученных им формул; ср. напр., E. Czuber „Über Funktionen von Variablen, zwischen welchen Korrelationen bestehen“ (Metron, vol. I, 1920). Систематический обзор известных к тому времени приемов дан в моей работе „О математическом ожидании частного двух взаимно зависимых случайных переменных“ (Труды русских ученых за границей, т. I, Берлин 1922).

§ IV, 1), A). — Ср. А. А. Чупров „О математическом ожидании частного“... стр. 244-245. Вытекающее из равенства

$$E\frac{n_{ij}}{n_{i'}} = \frac{p_{ij}}{p_{i'}}$$

соотношение $Em_1^{(i)'} = m_1^{(i)}$ было получено как приближенное K. Pearson'ом в его работе „On the application of „goodness of fit“ tables to test regression curves and theoretical curves used to describe observational and experimental data“ стр. 240 (Biometrika, vol. XI); ср. редакторскую статью „Peccavimus!“ стр. 267 (Biometrika vol. XII), где приведено отличное от моего дока-

1) «Относительная переброска, приведенная к единице наблюдения» [Прим. Н. С. Ч.].

зательство, что это равенство справедливо „not merely to a high order of approximation, but absolutely“ [1]).

§ IV, 1), B). — Ср. А. А. Чупров „О математическом ожидании частного...“ стр. 246—251.

§ IV, 2). (Ср. главу четвертую § II, 1) и главу пятую § VI). — Ср. А. А. Чупров „О математическом ожидании частного...“ стр. 256—261. Показатель взаимной сопряженности φ (Mean Square Contingency) введен K. Pearson'ом в работе „On the theory of contingency and its relation to association and normal correlation“ (Drapers' Company Research Memoirs, Biometric Series, I, 1904). Возможные формы показателя первоначально различались недостаточно отчетливо. В настоящее время Пирсон проводит различие между „the mean square contingency for the whole population“ [2]) (мое φ), между „the approximate value of the mean square contingency“ [3]) (мое φ') и между „its true value“ [4]), определяемое формулой

$$\sum_i \sum_j \frac{(p_{ij}' - p_i p_j)^2}{p_i p_j}.$$

Ср. K. Pearson „On the probable error of a coefficient of mean square contingency“ (Biometrika, vol. X) и K. Pearson and A. W. Young „On the probable error of a coefficient of contingency without approximation“ (Biometrika, vol. XI, некоторые из этих формул не вполне точны; ср. „Peccavimus“! стр. 259-260. Biometrika, vol. XII).

Квадратическая ошибка для φ' в первом приближении выведена (иным способом чем у меня) в работе J. Blakeman and K. Pearson „On the probable error of mean square contingency“ (Biometrika, vol. V).

§ IV, 3), A). — Ср. А. А. Чупров „О математическом ожидании частного...“ стр. 267-269.

Систематическую ошибку и квадратическое отклонение для случая нормальной корреляции сходным путем нашел H. E. Soper „On the probable error of a correlation coefficient to a second approximation“ (Biometrika, vol. IX). P. A Фишер под влиянием работы „Student“'a „Probable error of a correlation coefficient“ (Biometrika, vol. VI) исследовал закон распределения эмпирического коэффициента корреляции для случая нормальной корреляции; R. A. Fisher „Frequency distribution of the values of the correlation coefficient in samples from indefinitely large population“ (Biometrika, vol. X); ср. Soper, Young, Cave, Lee and Pearson „On the distribution of the correlation coefficient in small samples“ (Biometrika, vol. XI); R. A. Fisher „On the probable error of a coefficient of correlation deduced from a small sample“ (Metron, vol. I, 1921). Общая формула квадратической ошибки эмпирического коэффициента корреляции впервые дана была в работе W. F. Sheppard'a „On the application

---

1) «Не только с большой но и с полной точностью». [Прим. Н. С. Ч.].

2) «Показателем взаимной сопряженности признаков в безграничной совокупности» [Прим Н. С. Ч.].

3) «Приближенное значение показателя взаимной сопряженности» [Прим. Н С. Ч.].

4) «Его истинное значение» [Прим. Н. С. Ч.].

of the theory of error to cases of normal distribution and normal correlation", стр. 128 (Phil. Trans., A, vol. 192, 1899); наиболее употребительная форма квадратической ошибки для $r_{1/1}'$ в случае нормальной корреляции ведет свое начало от работы K. Pearson and L. N. G. Filon „On the probable errors of frequency constants and on the influence of random selection on variation and correlation", p. 245 (Phil. Trans., A, vol. 191, 1898). В более ранних статьях можно еще встретить указания на другое значение квадратической ошибки эмпирического коэффициента корреляции, а именно

$$\frac{1 - r_{1/1}^2}{\sqrt{N(1 + r_{1\,1}^2)}}.$$

Эта формула была дана K. Pearson'ом в мемуаре „Regression, heredity and panmixia", стр. 266 (Phil. Trans., A, vol. 187, 1897); в вышеупомянутой работе Пирсона и Филона она была заменена правильной приближенной формулой.

§ IV, 3), B). — Квадратическая ошибка для $A_{/1}'$ в предположении нормальной корреляции была выведена в работе K. Pearson'a и L. N. G. Filon'a „On the probable errors of frequency constants and on the influence of random selection on variation and correlation", стр. 245 (Phil. Trans., A, vol. 191, 1898); в редакторской статье „On the probable errors of the frequency constants" стр. 9 (Biometrika, vol. IX) эта формула воспроизводится, причем отбрасывается требование нормальной корреляции и остается лишь допущение, что „the frequencies are simmetrical and the regression linear" [1]); там же приводится и квадратическая ошибка для $A_{/0}'$ в форме

$$\sigma_{A/1}' \sqrt{m_{1/0}^2 + \mu_{2/0}}.$$

§ IV, 3), C). (Ср. главу четвертую § IV и главу пятую § IV). — Ср. В. Романовский „О корреляционном отношении" (Вестник Статистики, кн. XII, Москва 1923). Величина η' под названием „Correlation Ratio" введена K. Pearson'ом „On the general theory of skew correlation and non linear regression" (Drapers' Company Research Memoirs, Biometric Series, II, 1905). В этой же статье на стр. 19 приведена и квадратическая ошибка для η'. Систематическая ошибка оценки корреляционного отношения при помощи η' была вычислена лишь позднее; ср. K. Pearson „On the correction necessary for the correlation ratio η" (Biometrika, vol. XIV).

§ IV, 3), D). — Ср. J. Blakeman „On tests for linearity of regression in frequency distributions" (Biometrika, vol. IV).

§ IV, 5). — Квадратическая ошибка для

$$\frac{\delta'}{\sqrt{p_{1/}'\, p_{2/}'\, p_{/1}'\, p_{/2}'}}$$

1) «Распределения симметричны, а регрессия линейна» [Прим. Н. С. Ч.].

была найдена Юл'ом способом, отличным от моего; см. G. U. Yule „On the methods of measuring association between two attributes", стр. 603 (Jour. Roy. Stat. Soc., vol. 75, 1912).

§ V. — Ср. Е. Слуцкий „О некоторых схемах корреляционной связи и о систематической ошибке эмпирического значения коэффициента корреляции" (Вестник Статистики, кн. XIII, Москва 1923). Ср. постановку вопроса у L. Isserlis'a „The variation of the multiple correlation coefficient in samples drawn from an infinite population with normal distribution" (Phil. Mag., 6-th series, vol. 34, 1917).

К главе седьмой.

Изложенные здесь взгляды подробнее развиты в моих работах: „Основные задачи стохастической теории статистики" §§ VII и VIII, (Сборник „Памяти Николая Алексеевича Каблукова", Москва 1925, а также Вестник Статистики, кн. XIX, Москва 1925); „Über normal stabile Korrelation" (Skandinavisk Aktuarietidskrift, 1924), „Ist die normale Stabilität empirisch nachweisbar?" (Nordisk Statistisk Tidskrift, Bd. I, 1922). Ср. Я. Мордух „О связанных испытаниях, отвечающих условию стохастической коммутативности" (Труды русских ученых за границей, т. II, Берлин 1923).

К главе восьмой.

§ I, 2). — По вопросу о требованиях, которые следует предъявлять показателям связи, очень поучительно сопоставить статьи G. U. Yule'a „On the association of attributes in statistics" (Phil. Trans., A, vol. 194, 1900) и „On the methods of measuring association between two attributes" (Jour. Roy. Stat. Soc., vol. 75, 1912) со статьей K. Pearson'a и D. Heron'a „On theories of association" (Biometrika, vol. IX).

§ II, 1). — В качестве примера картографического использования корреляционных исчислений в метеорологии можно указать на работы F. Exner'a „Über monatliche Witterungsanomalien auf der nordlichen Erdhalfte im Winter" (Sitz.-Ber. der Akad. d. Wissensch., Bd, 122, Abt. IIa; Wien 1913; реферировано с воспроизведением чертежей A. Defaut „Wetter und Wettervorhersage", Leipzig und Wien 1918) и A. Schmaus'a „Korrelationen von Marz bis September" (Meteor. Zeitschrift, Bd. 41, 1924); ср. также Brit. Antarctic Exped. 1910—1913, Meteorology, vol. I (Calcutta 1919).

Указатель обозначений.

a_{jr} (44) A_{j0}, A_{j1} (49) A_{j0}', A_{j1}' (69)

b_{jr} (45)

E (28)

i (7) I (8)

m_{fg} (42) m_{fg} (66)

$m_{j1}^{(i)}$ (43) $M_{j1}^{(i)}$ (49) $\mathfrak{M}_{j1}^{(i)}$ (46) $m_{jg}^{(i)'}$ (68) $M_{j1}^{(i)'}$ (69)

n_{ir}, n_{js}, n_{sr} (64—65) N (64)

p_{ij}, p_{js}, p_{ir} (37) p_{ij}', p_{js}', p_{ir}' (65)

$p_{ij}^{(s)}$ (37) $p_{ij}^{(s)}$ (65)

r_{fg} (42) $r_{1/1}$ (42) r_{fg}' (67) $r_{1/1}'$ (67)

$x^{[r]'}$ (65) x_0' (65) $x_0^{(i)'}$ (65) $x_0^{[k]'}$ (114)

$\mathfrak{X}$ (46)

δ (39) δ' (74)

$\eta_{Y/X}$ (51) $\eta_{Y/X}'$ (70)

$\zeta_{Y/X}$ (97) $\zeta_{Y/X}'$ (97)

μ_{fg} (42) μ_{fg}' (67) μ_{fg}'' (90)

$\mu_{j2}^{(i)}$ (43) $\mu_{j2}^{(i)'}$ (68)

ρ (11)

σ (28)

φ (39) φ' (73)

Предметный указатель.

Гетероскедастичность 30.
Гомоскедастичность 30.
Зависимость стохастическая 24—27, 41.
Закон больших чисел 75—76.
Закон малых чисел 101.
Закон распределения значений переменной 28;
условный 29.
Закон связи 29;
условный 109.
Извлечения без возвращения 77, 110—111;
с возвращением 76—77, 78—79, 110;
с прикладыванием 111—112.
Индексы Фехнера 6—8.
Индукция, методы 19.
Истолкование связи 120—125.
Координаты нормальные 46—47.
Корреляционная таблица 2, 64—65.
Корреляционное исследование и «сглаживание» рядов 27, 34—36.
Корреляционное отношение:
априорное 51—54;
математическое ожидание 95—97;
рассеяние 96—97;
условное 109;
эмпирическое 70.
Корреляция:
кратная 109, 119;
нормальная 56—59;
частичная 109.
Корреляция рангов 11.
Коэффициент дисперсии 83, 113—114.
Коэффициент корреляции:
априорный 42, 53—55, 56—57, 62—63;
математическое ожидание 85, 90—93;
рассеяние 88, 93—94;
условный 109,
эмпирический 8, 67, 69, 71—73, 74.
Коэффициент регрессии:
априорный 46—47;
математическое ожидание 94—95;
рассеяние 95;
эмпирический 69.
Линия регрессии:
априорная 30, 44—48;
эмпирическая 3—5, 69.
Независимость стохастическая:
определение понятия 31—32;
условия 38, 43—44.

Некоррелированность:
определение понятия 32—34;
условия 48.
Ожидание математическое 28;
условное 29;
частного 80, 83, 87, 101—103.
Ошибка средняя квадратическая 28;
условная 30.
Ошибка оценки, систематическая 78—80, 105—107, 111.
Переменная случайная 20—24.
Поверхность регрессии 109.
Признак случайно изменчивый 40—41.
Предположительное значение 77, 77—79.
Проблема моментов (problème des moments) 42.
Пропорциональность между причиной и следствием, прямая 13—14.
Прямая регрессии, наилучшая:
априорное уравнение 49;
эмпирическое уравнение 68—69.
Рассеяние:
априорное 28;
условное 29, 43, 50, 68;
эмпирическое 67.
Регрессия параболическая, второго порядка:
корреляционное отношение 53;
уравнение 47;
условия 48.
Регрессия прямолинейная:
априорное уравнение 46;
критерий априорный 53—54, 97, 117;
математическое ожидание 97.
рассеяние 98—99.
эмпирический 97.
условия 46—47.
Скедастичность 30.
Сопряженность признаков, показатель взаимной —, (Mean Square Contingency):
априорный 39—40;
математическое ожидание 88—89;
рассеяние 89—90;
эмпирический 73—74.
Таблица тетрахорическая (четырехраздельная) 38—39, 41, 62—63, 74, 81, 103—105, 132—133.
Теснота связи 12, 14—17, 19;
измерение и оценка тесноты связи 116—117, 118;
истолкование числовых значений показателей 18—19, 55—56, 117—118;
и формы линий регрессии 59—61, 107—109.
Transfer per unit 157;
априорный 39;
математическое ожидание 81;
рассеяние 81—82;
эмпирический 74.
Уравнение регрессии и функциональная зависимость 34—36, 50, 59—61.
Устойчивость нормальная 113—115.

Оглавление.

стр

Предисловие . . . . . III—VII

Глава первая. «Математическая» теория корреляции и «элементарные» приемы уловления связи между явлениями . . . . . 1—12
§§ I (1), II (2), III (11).

Глава вторая. Объект и задачи стохастического изучения связей между явлениями. Неразрывная причинная связь и корреляция . 13—19

Глава третья. Стохастическая связь и функциональная зависимость . . . . . 20—36
§§ I (20), II (24), III (27), IV (29), V (34).

Глава четвертая. Априорный закон связи и система его сводных характеристик . . . . . 37—63
§§ I (37), II (38), III (41), IV (51), V (56), VI (59), VII (61).

Глава пятая. Сводные характеристики совокупности эмпирических данных . . . . . 64—74
§§ I (64), II (64), III (68), IV (70), V (71), VI (73), VII (74).

Глава шестая. Оценка априорных искомых по эмпирическим данным . . . . . 75—109
§§ I (75), II (76), III (80), IV (83), V (105), VI (107), VII (109).

Глава седьмая. Стохастические предпосылки числовой характеристики связи . . . . . 110—115
§§ I (110), II (112), III (113), IV (114).

Глава восьмая. Задачи и выгоды числовой характеристики связи . . . . . 116—125
§§ I (116), II (119).

Приложение . . . . . 126—148

Обзор литературы . . . . . 149—159

Указатель обозначений 161

Предметный указатель . . . . . 163—164

Существенные опечатки.

Страница.	Строка.		Напечатано.	Следует
42	5	сверху	μ^{fg}	$\mu_{f/g}$
45	1	снизу	$m_{/1}' - m_{1/0} =$	$m_{/1}' - m_{0/1} =$
46	18	сверху	$m_{1/}^{(i)}$	$m_{1/}^{(i)}$
46	3	снизу	$m_{/01}$	$m_{0/1}$
53	13	снизу	$\eta_{1/X}^2 = \frac{1}{\mu_{0/2}} \sum_i p_{i/} [\mathfrak{M}_{/1}^{(i)}]^2$	$\eta_{1/X}^2 = \sum_i p_{i/} [\mathfrak{M}_{/1}^{(i)}]^2$
62	12	сверху	$\delta = p_{1/1} p_{2/2} p_{1/2} p_{2/1}$	$\delta = p_{1/1} p_{2/2} - p_{1/2} p_{2/1}$
67	2	сверху	$r_{f/g}$	$r_{f/g}'$
67	4	сверху	$r_{f/g}$	$r_{f/g}'$
67	8	сверху	$\mu_{0/2}' = \sum_i p_{i/}' [x_i - m_{1/0}']^2$	$\mu_{2/0}' = \sum_i p_{i/}' [x_i - m_{1/0}']^2$
70	1	снизу	$\sum_i p_{1/}' [m_{/1}^{(i)'} - \cdots$	$\sum_i p_{i/}' [m_{/1}^{(i)'} - \cdots$
71	17	сверху	$M_{/1}^{(i)} - m_{0/1}'$	$M_{/1}^{(i)'} - m_{0/1}'$
95	4	сверху	$[r_{2/2} + r_{4/0} r_{1/1} - 2 r_{3/1} r_{1/1}]$	$[r_{2/2} + r_{4/0} r_{1/1}^2 - 2 r_{3/1} r_{1/1}]$
95	6	сверху	$m_{1/0} [r_{2/2} + \cdots$	$m_{1/0}^2 [r_{2/2} + \cdots$
99	10	сверху	$- 4 [\eta_{1/X} - r_{1/1}^2]$	$- 4 [\eta_{1/X}^2 - r_{1/1}^2]$
100	5	сверху	$Z' - EZ$	$Z' - EZ'$
102	3	сверху	$\frac{K^3}{C^2}$	$\frac{K^2}{C^2}$
127	5	сверху	$=$	$+$
133	8	сверху	$= \frac{1}{p_{2/}} \delta [y_1 - y_2]$	$= - \frac{1}{p_{2/}} \delta [y_1 - y_2]$

Страница.	Строка		Напечатано	Следует
136	8	снизу	$[x^{[f]'} - x_0']^2$	$[x^{[f]'} - x_0']$
138	5	сверху	$p_{/i}$	$p_{/i}$
140	5	сверху	$[x_i - {}_{1/0}]\, y_i - m_{0\,1}]$	$[x_i - m_{1/0}]\, [y_i - m_{0/1}]$
140	8	сверху	$\{(+ g)$	$\{(f + g)$
140	10	сверху	$\mu_{f,2}$	$\mu_{f/g}$
141	10	сверху	$\mu_{/2}^{(i)} \sqrt{\mu_{2/0}} \cdot \cdot$	$\mu_2^{(i)} = \sqrt{\mu_{2/0}} \cdots$
144	2	сверху	EY'	EZ'
147	4	снизу	$= E\{$	$= \frac{1}{N^2} E\{$
148	5	снизу	$+ \sum\sum x^{[f]'} y^{[g]'} T^k\}$	$+ \sum\sum x^{[f]'} y^{[g]'}]\, T^k\}$
154	3	снизу	$[x_i - x_j]$	$[x_i - x_f]$

БИБЛИОТЕКА
... СТИ...
...АДЕМИИ.