

Методы статистического анализа в прикладных социальных исследованиях.  
1955 – 2005 гг.

Кутлалиев А.Х.

Международный институт маркетинговых и социальных исследований «ГфК Русь»

История социологии насчитывает около полутора веков. И с самого её зарождения методы количественного статистического анализа активно использовались как в работах самого отца–основателя науки, Огюста Конта, так и других выдающихся социологов того времени – Карла Маркса, Макса Вебера и Эмиля Дюркгейма. Например, знаменитая работа Дюркгейма «Суицид», опубликованная в 1897 году, является блестящим примером анализа, объединившего эмпирические статистические данные и теоретические конструкты. Не менее знаменитая работа «Протестантская этика и дух капитализма» 1904 года представляла собой попытку Вебера использовать для доказательства своей теории как религиозную доктрину протестантизма, так и не в меньшей степени доступные ему статистические данные о развитии капитализма в Западной Европе с периода Реформации.

Тем не менее до Второй мировой войны методы статистического анализа, применяемые в социологии, носили в большей степени описательный характер и были достаточно просты. Чтобы понять причины, необходимо вспомнить некоторые особенности развития теории вероятности и математической статистики и их применение в различных отраслях естествознания в XVII-XIX веках [1]. На начальном этапе развития теории вероятностей во второй половине XVII века Паскаль, Ферма, Гюйгенс, Я. Бернулли уделяли большое внимание соблюдению критериев математической строгости при доказательстве теорем и практическом применении в различного рода задачах из самых разных областей науки. Однако уже в XVIII и XIX веках большинство ученых не смогли избежать искушения применить теорию вероятности к исследованию очень широкого круга проблем, некритично смешивая строгие математические предпосылки с содержательными аспектами различных областей общественной жизни. Этот период развития теории вероятностей можно назвать романтическим, так как в нём причудливо сочеталось получение интереснейших математических результатов и обоснованные их применения с наивными, а то и нелепыми с современной точки зрения попытками приложения вероятностных моделей к моральной проблематике. Такой подход к теории

вероятностей во многом сложился под влиянием таких выдающихся ученых, как Лаплас [7] и Пуассон [8]. Например, знаменитая теорема Пуассона, вошедшая во все учебники по математической теории вероятностей, была опубликована в его работе под говорящим названием «Исследование о вероятности судебных приговоров по уголовным и гражданским делам». С одной стороны, достижения Лапласа и Пуассона в области теории вероятностей привлекли внимание научного мира к развитию этой дисциплины. С другой стороны, их необоснованные заявления о перспективах решения с помощью данного подхода политических, правовых и нравственных проблем привели к огромному числу работ, посвященных приложениям к различным областям естествознания и общественной жизни. Многие из них были настолько мало обоснованы, что впоследствии воспринимались в качестве математического скандала. В результате упомянутые увлечения сменились глубоким разочарованием и полным скептицизмом в отношении возможности использования теории вероятностей в качестве метода научного познания. Различие между теорией вероятностей и другими математическими теориями в аспекте обоснованности результатов и приложений стало особенно остро осознаваться после того, как Коши (1821) в начале XIX века привел строгое обоснование анализа бесконечно малых, и аксиоматический подход при построении математических теорий стал общепринятым. Всё это в конце концов привело большинство математиков (и прежде всего в Западной Европе) к отказу от признания за теорией вероятности статуса математической дисциплины вплоть до 1920-х годов. Этот неудачный исторический опыт применения математико-статистических подходов в области социальных наук сказывался и в дальнейшем. Тем не менее в конце девятнадцатого – первой половине двадцатого столетия в развитии статистической науки произошел резкий рывок, во многом благодаря российской школе статистиков. Упомяну здесь только самые знаковые фигуры из этой когорты: Чебышев, Ляпунов, Марков, Слуцкий, Колмогоров. Крамер в своей книге "Полвека с теорией вероятностей" [5] отметил, что «если в 1920 году она едва ли заслуживала названия математической теории, то в 1945 году вступила в послевоенный мир в качестве хорошо организованного раздела математики, ... с постоянно расширяющимися сферами приложения в других науках, так же как и в различных видах практической деятельности». В том числе, несмотря на трудности ручного счета, статистические методы стали массово применяться в социологии. Тем не менее надо заметить, что число математиков, работающих в области решения социологических проблем, было относительно мало по сравнению с теми, кто работал в области технических наук, медицины и даже биологии. Неудивительно, что длительное время

наибольший вклад в развитие статистических методов в социологии вносили сами социологи, но мы можем видеть, что данная тенденция начинает меняться. Вряд ли я смогу перечислить всех математиков, которые внесли тот или иной вклад в социологию, хочу просто вспомнить того, кому посвящена данная конференция – Александра Олеговича Крыштановского.

Он был одним из тех, кто пришел в социологию из профессиональной математической среды – с дипломом выпускника факультета прикладной математики Московского института электронного машиностроения. Не замыкаясь в рамках проблем обработки данных, он глубоко вникал в суть решаемых им задач, прекрасно знал предметную область, и воспринимался «классическими» социологами абсолютно органично, как профессионал, имеющий полное право возглавить факультет социологии и строить процесс социологического образования в соответствии со своими представлениями о будущем социологии.

Данный обзор – попытка описать картину развития математико-статистических методов за последние полвека, которая сложилась в голове автора под влиянием неспешных бесед, отдельных разговоров и кратких реплик Крыштановского, касающихся методов количественного анализа в прикладных социальных исследованиях, а также представить некоторые новые подходы к анализу данных в эмпирических социальных науках.

### **Три поколения количественных методов анализа**

Условно можно разделить развитие количественных методов на три поколения, во многом связанных с доступной исследователям на той или иной стадии вычислительной мощностью имеющихся инструментов (от счетных линеек до персональных компьютеров), уровнем агрегирования собираемых данных и представлениями о том, что можно анализировать количественно. При этом нельзя провести четких границ, когда развитие одних методов сменяется новым поколением – как и в человеческом обществе, эти поколения сосуществуют и дополняют друг друга, время от времени обогащая друг друга новыми идеями и подходами. Первое поколение можно назвать поколением таблиц сопряженности или кросс-таблиц. Большинство данных, с которыми имели дело социологи, имело вид таблиц частот сопряженных уровней двух (и более) наблюдаемых признаков. Второе поколение во многом связано с развитием электронной вычислительной техники, когда стал возможен анализ баз данных, включавших в себя большое количество переменных по каждому

независимому наблюдению. Третье поколение, отходя от вероятностно-статистической доминанты, обращается к анализу информации, представленной в самой разнообразной форме, например текстовой, пространственной (топологические данные) и т.п. Все перечисленные направления продолжают развиваться и использоваться и в настоящее время, и более того, идет процесс постоянного взаимопроникновения и перекрытия методов.

### **Первое поколение: Анализ на агрегированном уровне.**

Социология – это изучение  
социального явления на  
агрегированном уровне.

Франклин Гиддингс

Анализ на агрегированном уровне – анализ распределений частот и кросстабуляций (таблиц сопряженности). Фактически кросстабуляция - это процесс объединения двух (или нескольких) таблиц частот так, что каждая ячейка в построенной таблице представляется единственной комбинацией значений или уровней табулированных переменных. Таким образом, кросстабуляция позволяет совместить частоты появления наблюдений на разных уровнях рассматриваемых факторов. Исследуя эти частоты, можно определить связи между переменными. Традиционным считается анализ независимости факторов, основанный на статистике хи-квадрат. При всей простоте и мощности данного критерия он не отвечает на вопрос, каков вид связи между факторами, если гипотеза об их независимости отвергается. Исследователями было предложено множество мер связи самого разного характера. Отголосок этого научно-практического поиска можно наблюдать в статистическом пакете SPSS, где в анализе кросстаблиц предлагается более десятка мер связи, многие из которых в настоящее время представляют больше академический, чем практический интерес. Более плодотворными оказались подходы Гудмана и Лазерсфельда, предложивших для анализа таблиц логлинейные модели и латентно-структурный подход соответственно.

Модель логлинейного анализа основана на мультипликативном определении понятия взаимосвязи, которое записывается обычно в виде разложения логарифма частоты в каждой клетке таблицы сопряженности на сумму эффектов от учтенных в гипотезе взаимосвязей (по значениям признаков, соответствующим данной клетке). Это разложение по форме аналогично модели дисперсионного анализа. Опыт применения логлинейного анализа показал его эффективность как способа анализа многомерных

таблиц сопряженности. Метод позволяет сжато описать эти таблицы (в виде гипотезы о связях) и в то же время детально проанализировать конкретную взаимосвязь. Тем не менее в прикладных исследованиях логлинейные модели встречаются нечасто в связи с тем, что требуют достаточно высокой методологической подготовки исследователя и значительного объема работ в интерактивном режиме человек-компьютер.

В этом плане чуть больше «повезло» латентно-структурному анализу. Основной постулат этого метода заключается в следующем - в основе латентно-структурного анализа лежит вероятностно-статистическая модель, которая подразумевает, что за корреляцией (статистической связью) наблюдаемых величин стоит латентная переменная, которая эту связь определяет.

Входными данными для анализа являются частоты – вероятности ответов и их комбинаций, а выходными – вероятности того, что по тому или иному набору ответов респонденту можно приписать некоторое значение латентной переменной.

Надо отметить, что на начальных этапах развития ЛСА идея статистической оценки качества полученной модели не рассматривается. В работах 50-х годов по ЛСА такие понятия статистики, как стандартные ошибки, несмещенная оценка, максимальное правдоподобие отсутствуют. И только в работе Лазерсфельда и Генри 1968 года «Латентно-структурный анализ» [17] мы видим привычный статистический подход к формулировке, оценке и проверке стохастических моделей.

Что касается практического применения ЛСА, то, с одной стороны, долгое время упоминания о нем в основном встречались в работах статистиков и методологов, получивших с его помощью целый ряд интересных моделей в самых разных областях: от изучения общественного мнения до криминалистики. Ситуация начала меняться в 90-е годы, в связи с появлением специализированных программных пакетов и массовым распространением персональных компьютеров. С другой стороны, анализ латентных переменных имеет уже более чем вековую историю [19] и широко используется в анализе данных. И многие специалисты по обработке данных, привычно пользующиеся факторным анализом, с великим для себя удивлением открывают, что это всего лишь частный случай латентно-структурного анализа с несколькими латентными переменными (Андерсон [11], Гибсон [13]). Так мольеровский персонаж Журден с восторгом узнаёт, что вот уже сорок лет говорит прозой.

Лазерсфельд сетовал, что столь необходимый для нужд социологии статистический анализ дискретных категориальных данных игнорируется статистиками его времени. Однако в середине 70-х положение изменилось. Гудман [14] в своей

статье «Analyzing Qualitative / Categorical Data: Log-linear Models and Latent Structure Analysis» показал, как латентно–структурная модель может быть проинтерпретирована в терминах логлинейной модели. Эта статья и публикация работы Бишопа, Фейнберга и Холланда [12] «Дискретный множественный анализ» привлекла внимание авторитетных статистиков к латентно-структурному анализу и было предложено множество интерпретаций и применений метода. В свою очередь это привело к созданию ряда компьютерных программ по ЛСА, которые облегчили доступ к его использованию рядовых аналитиков, не владеющих тонкостями вычислительного алгоритма метода.

Отметим также своеобразный подход к анализу данных одного из создателей анализа соответствий французского исследователя Жана Поля Бензекри, выраженный в лозунге: «Модель должна соответствовать данным, а не наоборот!» Действительно, нередко при анализе данных мы имеем минимум априорных идей относительно их структуры и вынуждены исходить из имеющихся данных как таковых. Далее различного рода поисковыми методами мы стремимся понять, как организованы эти данные, какие переменные или группы переменных связаны (коррелируют) между собой, иными словами, стремимся понять структуру данных, исходя из них самих.

Анализ соответствий - метод поискового анализа данных, созданный для исследования как простых двумерных, так и более сложных таблиц на предмет наличия связи между строками и столбцами. В двух словах СА можно определить как особый случай метода анализа главных компонент (РСА) строк и столбцов матрицы данных (таблиц сопряженности). Тем не менее СА и РСА должны применяться для разных типов данных. РСА используется для анализа непрерывных величин, в то время как СА - в основном для анализа категориальных переменных.

Как уже упоминалось ранее, в отличие от традиционных методов тестирования гипотез, в которых исследуется а priori сформулированные гипотезы о связях между переменными, поисковый анализ данных используется для выявления связей при отсутствии (или недостатке) априорной информации о связях между переменными. В то же время анализ соответствий – это метод дескриптивного анализа данных. Исходное многомерное пространство таблицы сопряженности он переводит в пространство меньшей размерности (чаще всего двумерное), а исходное распределение частот – в расстояния между строками (столбцами) и их взаимное пространственное расположение. В высшей степени упрощая представление данных, метод позволяет компактно представить структуру большей части исходной информации, обеспечивая

простой по форме, но всесторонний и исчерпывающий анализ. В настоящее время СА является наиболее востребованным методом из большого семейства методов многомерного шкалирования вследствие самой простой процедуры сбора данных, необходимых для анализа.

### **Второе поколение: Анализ на индивидуальном уровне.**

"Душа - это латентный фактор!"

А.О. Крыштановский

Дальнейшее развитие статистические методы анализа данных получили в 60-е-70-е годы. Данные для анализа использовались на индивидуальном уровне, при этом база данных могла содержать гораздо большее количество переменных по каждому независимому наблюдению, что привело к усложнению и увеличению форм сбора данных. Всё это привело к расширению круга статистических методов, применяемых для анализа. Перечислим лишь некоторые из них: регрессионный анализ и обобщенные линейные модели, моделирование структурных уравнений, путевой анализ, совместный анализ, непараметрические методы и т.д. Вкратце рассмотрим характерные черты данного поколения методов, которое является mainstream-ом в количественном анализе данных и оказало большое влияние на развитие прикладных социальных исследований.

Построение регрессионных моделей на сегодняшний день несомненно являются наиболее широко применяемым методом статистического анализа данных социальных исследований. В публикациях по анализу данных регрессионный анализ используется более чем в половине случаев. Если учесть, что во многие методы оценка регрессионных параметров модели входит неявным образом, то данный метод можно назвать господствующим инструментом в анализе данных социальных исследований. Аппарат регрессионного анализа подробно проработан, сама идея регрессии хорошо понимается как исследователями, так и конечными пользователями результатов исследований, тем не менее на практике приходится сталкиваться с существенными ограничениями при использовании данного метода. Это связано с целым рядом причин. Обозначим некоторые из них [6]:

- В регрессионном анализе зависимость одной переменной от других считается единой для всей совокупности исследуемых объектов. По этому поводу Крыштановский как-то заметил: «Все же нас не покидает надежда, что есть единая модель, которая описывает всех

людей, обезьян, крыс, мышей». Это справедливо при однородности исследуемой совокупности, но чаще всего мы имеем дело с сильной неоднородностью. Единая модель в таком случае сильно огрубляет реальную зависимость, качество модели неизбежно становится низким (типичные коэффициенты детерминации 0.2-0.3 сильно изумляют коллег–эконометристов, привыкших работать с моделями с коэффициентами не менее 0.9-0.95). Попытки разбить исходную совокупность на более однородные группы (сегменты) выглядят разумными, но сами критерии разбиения целиком зависят от субъективизма конкретного исследователя и фактически количество подобных разбиений ограничивается только размерами исследуемой совокупности.

- Регрессионные модели строятся в предположении, что наши данные непрерывны, дифференцируемы и неограниченны по величине. Практически никогда эти требования к данным не соблюдаются. Чаще всего это данные нечисловой природы, которые только для удобства их представления имеют числовой вид. Попытки создания теории измерения в социальных науках, начиная с послевоенной работы американского психолога Стевенса (Stevens S.S.), всё еще не привели исследователей к единой точке зрения. Во многом это связано с тем, что несмотря на объявленный тип, статистическая интерпретация данных зависит от содержания вопроса, задаваемого исследователем до начала обработки этих данных. Простой пример: в панельном исследовании домохозяйствам присваивается цифровой код – уникальный идентификатор домохозяйства. С позиции теории измерений идентификатору присваивается номинальный тип данных, так как его задача – дать возможность однозначно определить конкретное домохозяйство. Поэтому практически единственный вид обработки идентификатора – проверка базы на дублирующие коды. С другой стороны, учитывая, что чаще всего в такого рода базах коды присваиваются последовательно, мы можем трактовать их как порядковые величины и рассмотреть, например вопрос, зависит ли характер потребления домохозяйств от порядка их попадания в панельную базу. Кроме того, нельзя забывать о том, что полученные в результате исследования данные чаще всего дискретны и конечны. Это

накладывает дополнительные ограничения на применение методов континуальной математики и вызывает острую потребность в переходе на методы анализа данных, базирующихся на дискретной математике.

- Проблема мультиколлинеарности или взаимозависимости предикторов регрессионной модели. В большинстве практических случаев наблюдается сильная корреляция между наблюдаемыми факторами, что неудивительно, так как запись в базе данных состоит из ответов на свои самые разнообразные вопросы из одного и того же источника – респондента. Являясь всего лишь внешней реакцией, проецирующей внутренние представления и установки респондента, эти ответы не могут не быть связанными между собой. То есть в своей основе взаимосвязь наблюдаемых признаков не является проявлением недостатков той или иной методики, а присуща самому объекту исследования.

Господство регрессионного подхода в анализе данных мешало и продолжает мешать более широкому распространению методов, основанных на других принципах. Например, модели марковских цепей случайных событий имеют большое значение при анализе многомерных взаимосвязей, позволяя, например, восстанавливать недостающую или пропущенную информацию и находят практическое применение при создании экспертных систем для медицинской диагностики. Но использующийся в этих моделях подход, базирующийся на понятии условной вероятности, считается социологами, привыкшим формулировать свои модели в терминах регрессии или причинно-следственных связей, слишком трудным для интерпретации.

Попытки уйти от перечисленных недостатков породили ряд новых подходов к анализу данных, среди которых хотелось бы буквально парой фраз сказать о двух из них: совместном анализе [15], модели структурных уравнений.

Совместный анализ - это любой декомпозиционный метод, классифицирующий предпочтения различных объектов (например, полезность, важность) на основе общих оценок альтернативных вариантов, описанных в терминах уровней различных атрибутов продукта. На основе ответов респондента выводятся частные полезности и относительные важности для каждого атрибута, что позволяет наилучшим из возможных способов реконструировать порядок предпочтений респондента. Не вдаваясь в подробности, хотелось бы отметить, что характерной чертой метода является то, что он позволяет от дискретных актов выбора респондентов

того или иного варианта ответа перейти к непрерывным значениям, выраженным в форме полезности того или иного атрибута объекта. В настоящее время разработано несколько вариантов совместного анализа и метод широко применяется в исследованиях потребителей.

Моделирование структурными уравнениями [16] (МСУ) является мощным методом статистического анализа, возможности которого российскими социологами пока недооценены. Этот подход можно рассматривать, как комбинацию факторного, регрессионного и путевого анализа. Особый интерес к моделям структурных уравнений основан на возможности рассмотрения теоретических конструкторов, представляющих латентную факторную структуру. Взаимосвязи внутри теоретических моделей представляются регрессионными или путевыми коэффициентами между факторами. Модель структурных уравнений подразумевает анализ ковариационной структуры между наблюдаемыми переменными, из чего происходит альтернативное название техники – моделирование ковариационной структуры. МСУ предоставляет широкое и удобное поле для статистического анализа поскольку включает в себя нескольких традиционных многомерных процедур, таких как факторный, регрессионный, дискриминантный анализ и каноническую корреляцию как частные случаи. К тому же модели, как правило, визуализируются посредством построения путевых диаграмм.

Общей чертой второго поколения можно назвать пристальное внимание развитию методов, направленных на выявление ненаблюдаемых показателей (латентных факторов), причем как на уровне моделей (МСУ), так и на индивидуальном уровне (совместный анализ).

### **Третье поколение: Анализ информации. Новые задачи, новые типы данных, новые методы.**

"Человек - целый мир"

Ф.М. Достоевский

Сверхзадача любого анализа – извлечь максимум информации из имеющихся данных. На самом раннем этапе сбора данных в социальных исследованиях они чаще всего имеют форму естественного языка. Это ответы респондентов, записи наблюдений, групповые дискуссии и т.д. При существующей технологии исследовательского процесса на промежуточных этапах мы теряем значительную часть информации, предоставленной нам респондентом. При фиксации ответов исчезает

невербальная составляющая (самый простой ответ «да» можно произнести с самой разной интонацией). Мы кодируем открытые вопросы, вгоняя высказанные мнения респондентов в жесткую схему пре-кодированных вариантов ответов. Мы пишем отчеты по фокус-группам и глубинным интервью, вставляя туда выжимки из протоколов расшифровки, при этом существующая практика предоставления заказчику аудио- видео-записей глубинных интервью и групповых дискуссий проблему глубокого анализа полученного материала не решает, она всего лишь перекладывает её на плечи самого заказчика. Довольно часто упоминают субъективизм качественных методов, решающее влияние исследователя-«качественника» на полученные выводы, но, как мне представляется, это общая проблема всех исследований, без деления на количественные и качественные. Допустим, мы проводим 1000 фокус-групп представителей некой целевой группы, при этом варьируем состав этих групп с учетом представленности этих представителей в генеральной совокупности. Означает ли это, что мы имеем право обрабатывать полученные данные статистически? Вступит ли полученная таким образом информация в противоречие с информацией, полученной традиционным для анализа фокус-групп способом? Если нет, то будут ли результаты дополнять или же просто дублировать друг друга?

Я акцентировал ваше внимание на понятии «информация» потому, что считаю, что имеющиеся в распоряжении исследователя данные не могут искусственно делиться на те, которые возможно подвергнуть статистической обработке и все остальные. Природа данных одна и такого рода деление свидетельствует об ограниченности нашего подхода к их анализу. Чтобы пояснить свою мысль, вспомним историю создания современной теории информации.

Становление математической теории информации базировалось на вероятностном подходе. Это проявилось уже в том, что основное понятие теории информации - «количество информации» - определялось через понятие вероятности (Шеннон, [2]), что сразу же позволило получить целый ряд важных результатов на базе аппарата теории вероятностей. Прежде всего они касались проблем «передачи по каналам связи массовой информации, состоящей из большого числа сообщений, подчиненных определенным вероятностным закономерностям» [4]. Но возможность введения понятия количества информации на основе понятия вероятности не гарантирует факта первичности понятия вероятности по отношению к понятию информации (или количества информации). Скорее можно говорить о том, что информация и вероятность являются равноправными и независимыми друг от друга

понятиями. Колмогоров ярко проиллюстрировал эту мысль, обсуждая шенноновский подход в теории вероятности: «Какой реальный смысл имеет, например, говорить о «количестве информации», содержащейся в тексте «Войны и мира»? Можно ли включить разумным образом это роман в совокупность «возможных романов», да еще постулировать наличие в этой совокупности некоторого распределения вероятностей ...» [4]. В другой статье он определенно настаивал на том, что «информация по своей природе не вероятностное явление» [3].

Рассмотрим пример. Если кто-то вам скажет, что в результате 100 бросков симметричной монетки у него 100 раз выпал «орёл», то мы откажемся считать такой результат случайным, подозревая говорящего в нечестности. При этом мы с вами согласимся считать случайной последовательность единичек и нулей, выданных нам генератором случайных чисел. Парадокс заключается в том, что с точки зрения вероятностного подхода оба события равноправны и имеют вероятность  $1/2^{100}$ . Если мы рассмотрим пример из более близкой предметной области, то в ситуации, когда в ответах респондента по шкале 1-5 мы имеем только одну оценку (например, 5) для всей батареи высказываний, мы справедливо усомнимся в том, что интервьюер соблюдал процедуру опроса.

Не ограничившись критикой, Колмогоров (1965) [3] и чуть ранее Соломонофф (США, 1964) [18], заложили основы независимой от теории вероятностей алгоритмической теории информации. Два важных вывода из этой теории имеют значения для дальнейшего обсуждения.

- Понятия «энтропии» и «количества информации» оказываются применимы к индивидуальным объектам.
- Эти понятия могут лечь в основу новой концепции случайного, соответствующей естественной мысли о том, что случайность есть отсутствие закономерности.

Если какой-либо объект устроен «просто», то его можно описать с помощью достаточного небольшого количества информации. И напротив, если объект сложен, то его описание связано с использованием большого объема информации. Если у нас есть конечный класс конечных объектов и  $N$  – число элементов этого класса, то для задания любого элемента достаточно  $\log_2 N$  двоичных знаков. Идентификацией элемента может служить двоичная запись его номера. Если элементы обладают какими-то особыми свойствами, другими словами, закономерностями, то возможно его более краткое описание. Проще говоря, если последовательность бросков случайна, то она не может

быть задана иначе, как выписыванием целиком всех результатов бросков, её сложность максимальна среди всех последовательностей равной длины. Если вся последовательность состоит из одних единиц (выпадение «орла»), то её сложность минимальна. Ранее мы говорили, что с точки зрения вероятностного подхода оба события одинаковы и имеют вероятность  $1/2^{100}$ . Но с точки зрения алгоритмического подхода последовательность из одних и тех же цифр (событий) максимально неслучайна, ибо имеет минимальную сложность (описывается с помощью одного элемента).

В рамках информационного подхода возможно уйти от вероятностной доминанты в анализе данных и существенно расширить область применения математических методов.

Во-первых, это позволяет анализировать данные на индивидуальном уровне, не прибегая к массовым опросам и выявлять закономерности, не используя понятие вероятности, а используя понятие информации. Информационный подход позволяет снять проблему искусственного разделения данных на количественные и качественные. Скажем, в Интернете существует такое социальное явление, как Livejournal (Живой журнал), созданный в марте 1999 года Брэдом Фитцпатриком (Brad Fitzpatrick). Люди ведут дневниковые записи либо в частном порядке, либо, что более интересно - открывают к ним публичный доступ (блог – открытый сетевой дневник) и создают сетевые сообщества, посвященные практически любой мыслимой теме. В настоящее время только на сайте [www.livejournal.com](http://www.livejournal.com) существует более 10 миллионов дневников и сообществ, пополняющиеся более чем двумястами тысячами записей в день. Анализ, основанный на вероятностном подходе, такого объема представленной на естественном языке информации, с грамматическими ошибками, своим жаргоном, имеющей вид пространственной социальной сети (гипертекст в гиперсети) практически невозможен не только по причине трудностей вычислительного характера, а также потому, что каждое из сообществ по своему уникально и имеет свои законы развития и функционирования.

Во-вторых, в социологии мы имеем дело с феноменом, с которым физики столкнулись лишь с началом исследования микромира – «принципом неопределенности». В интерпретации Нильса Бора он гласит, что "измерение вносит возмущения и изменения в то, что наблюдается", т.е. принцип неопределенности ограничивает точность измерения величин. Это становится особенно заметно, когда влияние инструмента и потенциал измеряемого объекта становятся сопоставимыми. В

прикладных социальных исследованиях влияние инструмента особенно заметно. Вопросы, задаваемые респонденту, влияют на то, как он будет на них отвечать – хорошо известный факт, и о минимизации этого влияния написаны работы, посвященные как теоретическому анализу опросного инструмента [9], так и обобщению накопленного эмпирического опыта [10]. Но даже при минимизации влияния исследователя изменения в состоянии респондента имеют место быть. Исследователям потребительского поведения давно известен "панельный эффект", заключающийся в том, что домохозяйки, регистрирующие свои покупки в мониторинговых исследованиях потребления в домашних хозяйствах, со временем меняют свое потребительское поведение. Не будем забывать и о том, что и поступает респондент не так, как он высказывался о своих намерениях во время опроса. Поэтому надо ясно представлять, что в ходе наших исследований мы измеряем не состояние респондента, а состояние системы «исследователь–респондент», где респондент либо подтверждает, либо опровергает точку зрения исследователя, который в то же время воздействует своим инструментом измерения на респондента, изменяя его состояние. И разделить собственное состояние респондента и внесенные исследователем возмущения чисто статистическими приемами не удастся.

Всё это должно стимулировать социологов на развитие новых методов, основанных на информационном подходе и направленных на извлечение максимума информации из самых разных, а не только структурированных источников данных. В качестве примера можно привести анализ текстов, ответы на открытые вопросы, социальные сети, пространственные (топологические) данные, цепочки событий и т.п.

### Заключение

Подводя итог данного обзора, можно сказать, что количественные методы в социальных науках, переходя от поколения к поколению ко всё более глубокому анализу собираемой в процессе исследований информации, еще только приблизились к периоду зрелости, когда в среде исследователей возникает понимание, что деление информации на качественную и количественную во многом искусственно, что деление данных по типам есть следствие не их внутренней природы, а грубости и несовершенства инструментов измерения. Также хотелось бы отметить, что классический прием естественнонаучных дисциплин, заключающийся в том, что исследователь выдвигает гипотезу о природе явления, а затем проверяет её, таит в себе большую опасность в социальных науках. Молекуле в общем-то всё равно, что о ней

думает химик–аналитик, бактерия не меняет своего поведения под микроскопом биолога, но респондент, которому исследователь задал вопрос – это уже не тот респондент, каким он был до вопроса и мы получаем ответ о его измененном состоянии – хорошо известный в физике «принцип неопределенности» Гейзенберга.

#### Список литературы

1. Б.В. Гнеденко. Очерк по истории теории вероятностей, М.: УРСС, 2001
2. Клод Шеннон Математическая теория связи, 1948
3. Колмогоров А.Н. Проблемы теории вероятностей и математической статистики. //Вестник АН СССР, 1965, №5, с.95
4. Колмогоров А.Н. Три подхода к определению «количества информации» // Проблемы передачи информации, Том 1, Вып.1, 1965, с.6
5. Крамер Х. Полвека с теорией вероятностей: наброски воспоминаний. Современные проблемы математики. Перев. с англ.-М.: Знание, 1979.-
6. Крыштановский А.О. Ограничения метода регрессионного анализа. // Социология:4М, 2000, №12
7. Лаплас Пьер Симон "Аналитическая теория вероятностей" (1812, 1814, 1820гг.)
8. Пуассон Анри «Исследование о вероятности судебных приговоров по уголовным и гражданским делам»
9. Рогозин Д.М. Когнитивный анализ опросного инструмента. М.: Институт Фонда "Общественное мнение", 2002.
10. Сеймур Садмен, Норман Бредбери Как правильно задавать вопросы: введение в проектирование массовых обследований. – М.: Институт Фонда «Общественное мнение», 2005
11. Anderson, T. W. (1959) "Some Scaling Methods and Estimation Procedures in the Latent Class Model", in Probability and Statistics, U. Grenander (ed.). New York: John Wiley & Sons.
12. Bishop, Y.M.M., S. E. Fienberg, and P.W. Holland (1975) Discrete Multivariate Analysis, Cambridge MA: MIT Press.
13. Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. //Psychometrika, 24, 229-252.
14. Goodman, Leo A. (1978) Analyzing Qualitative / Categorical Data: Log-linear Models and Latent Structure Analysis. Cambridge, MA: Abt Books.
15. Green, P. E., A. M. Krieger, and Y. Wind (2001), "Thirty Years of Conjoint Analysis: Reflections and Prospects," //Interfaces, 31 (May-June)
16. Jöreskog, K., 1969. A General Approach to Confirmatory Maximum Likelihood Factor Analysis. //Psychometrika. 34, 183-202. а также Structural equation modeling: Present and future. A Festschrift in honor of Karl Jöreskog. Chicago: Scientific Software International, pp. 139–168.
17. Lazarsfeld, Paul F. and Neil W. Henry (1968) Latent Structure Analysis, Boston: Houghton Mifflin.
18. Solomonoff R. J. «A formal theory of inductive inference», //Inform. Contr., vol. 7, pp. 1-22, Mar. 1964; also, pp. 224-254, June 1964
19. Spearman, Charles (1904) " 'General Intelligence', Objectively Determined and Measured", //American Journal of Psychology, 15: 201-293.