

НОВОЕ
В ЖИЗНИ,
НАУКЕ,
ТЕХНИКЕ

А. Н. Ефимов,
профессор,
доктор технических наук

Серия
«Математика,
кибернетика»
№ 2, 1980 г.

Издается ежемесячно
с 1967 г.

ПОРЯДКОВЫЕ
СТАТИСТИКИ –
ИХ СВОЙСТВА
И ПРИЛОЖЕНИЯ

Издательство
«Знание»
Москва
1980

Ефимов А. Н.

E91 Порядковые статистики — их свойства и приложения. М., «Знание», 1980.

64 с. (Новое в жизни, науке, технике. Серия «Математика, кибернетика», 2. Издается ежемесячно с 1967 г.)

В брошюре популярно рассказывается о порядковых статистиках — интенсивно развивающемся в настоящее время разделе непараметрической статистики. Демонстрируются специфические свойства порядковых распределений. Особое внимание уделено прикладным задачам — измерению, классификации и идентификации, решаемым с помощью аппарата порядковых статистик.

Рассчитана на тех, кто интересуется математической статистикой и ее приложениями.

20204

22.172

© Издательство «Знание», 1980 г.

ВВЕДЕНИЕ

Для многих отраслей прикладной науки появление ЭВМ явилось важным стимулом к дальнейшему развитию, открыло новые возможности для решения давно назревших задач, позволило поставить совершенно новые проблемы. При этом важно отметить, что задачи, решаемые на ЭВМ, образуют, как и сами ЭВМ, сменяющие друг друга поколения. Задачи первого поколения составляют период, когда ЭВМ выступала в роли мощного арифмометра и предназначалась, главным образом, для расчета по формулам. Второе поколение задач, безусловно, формировалось под девизом «оптимизация». Оно основывалось на возросшей гибкости программного обеспечения ЭВМ второго поколения и ознаменовалось введением в практику широкого круга задач назначения, распределения ресурсов, оптимального планирования, раскroя и размещения. Среди задач нынешнего, третьего, поколения, очевидно, доминируют задачи, связанные с управлением в реальном масштабе времени, с применением ЭВМ в замкнутом контуре управления, в рамках АСУТП или АСУ.

В любопытном положении оказалась математическая статистика — научная дисциплина, для которой вычислительная техника с первых своих шагов была особенно полезной и перспективной.

Так, в первый период «большой арифмометр» оказался просто необходим — математическая статистика впервые смогла продемонстрировать свои возможности при анализе случайных процессов и больших массивов данных.

Далее статистические алгоритмы оценивания, проверки гипотез, фильтрации стали входить в математическое обеспечение систем управления, контроля качества продукции, обработки радиосигналов, начали работать, таким образом, в системах реального времени и стали характери-

зоваться новыми для математической статистики параметрами — быстродействием, объемом необходимой памяти, «иепривередливостью» по отношению к исходным данным.

Вот здесь и начались неприятности. Нельзя сказать, что они оказались неожиданными.

Раньше, в «домашнюю эру», статистическая обработка наблюдений была «расскошью», приемлемой лишь в условиях научного эксперимента и некоторых специальных приложений, таких, как геодезия или демография. При этом наблюдения специально организовывались так, чтобы удовлетворить требованиям, предъявляемым статистикой. Как и всякая научная дисциплина, математическая статистика оперировала с набором несколько идеализированных моделей реальных объектов и явлений. Так, статистический материал считался однородным — выборки формировались из объектов, принадлежащих одной и той же исходной совокупности; статистический материал был обильным — можно было говорить об асимптотических свойствах оценок; наблюдатель был хозяином положения, он мог вести эксперимент активно — брать отсчеты с нужной (равной) точностью в нужный момент, обеспечивать независимость отсчетов и т. д.

Совершенно понятно, что в «естественных» условиях выяснилось, насколько эти модели далеки от реальности. Во-первых, статистический материал, которым может располагать наблюдатель, зачастую весьма ограничен по объему. Так, при контроле или испытаниях речь может идти о единицах испытанных изделий, по свойствам которых необходимо судить о целой их партии. В системах управления воздействия вырабатываются по результатам наблюдения координат объекта, причем, чем дальше их наблюдать, тем лучше можно оценить значения и вероятностные свойства процесса изменения координат. Однако долго наблюдать нельзя — управляющее воздействие должно быть выдано вовремя, иначе оно уже может и не понадобиться! Ясно что в такой ситуации длительность реализации, по которой должны быть сделаны статистические выводы о процессе, не может быть большой. Далее, вероятностные характеристики генеральной совокупности, из которой поступает статистический материал, нестабильны во времени. В задачах радиоприема и обработки сигналов при локации характер помехи меняется очень существенно как по естественным причинам: время суток, характер местности и т. п., так и в результате специально организованного

противодействия. При этом требуется, чтобы алгоритмы, по которым производится фильтрация, оценивание, принятие статистических решений, имели стабильные точностные характеристики.

Не лучше обстоит дело и с традиционными требованиями к независимости статистического материала.

Таким образом, вычислительная техника, предоставив математической статистике поистине неограниченные возможности для проникновения в практику, потребовала, со своей стороны, существенного обновления теоретического аппарата, выдвинув ряд неотложных и срочных задач.

Ограничность статистического материала породила нарастающий поток работ, посвященных «проблеме малых выборок».

Бурно развивается теория «устойчивого оценивания» (*robust estimation*), имеющая целью выработать оценки пусть менее точные, чем оптимальные, но приемлемые в условиях, когда статистические свойства объекта неизвестны или переменны.

Специфической проблеме нарушения независимости между исследуемыми объектами и особенностями их статистических свойств и будет посвящена эта книжка. Речь пойдет о случаях, когда материал, которым располагает наблюдатель, — результаты измерений, подлежащие контролю объекты, наблюдаемые явления — упорядочен по какому-то определенному признаку. Так, специалист по надежности, изучая причины разрушения каких-то деталей, будет иметь дело с группой образцов, объединенных уже тем, что все они разрушились при определенной нагрузке. Напротив, тренер сборной команды имеет дело с группой спортсменов, каждый из которых оказался лучшим в своем клубе или превысил установленный норматив.

Раздел математической статистики, изучающий свойства объектов, занимающих определенные места (ранги) в упорядоченной выборке, называется теорией порядковых статистик.

Любопытно, что, начав с исследования естественно упорядоченных совокупностей, теория продемонстрировала, в конце концов, что в ряде случаев следует упорядочить выборку специально и что такое упорядочение может существенно улучшить статистические оценки.

Мы увидим далее, что между значением элемента выборки и местом, которое он занимает после упорядочения, существует связь настолько значительная, что в ряде слу-

чаев можно, ранжировав выборку, делать статистические оценки и выводы лишь по рангам элементов.

Более того, поскольку упорядочить возможно (по принципу «больше-меньше») и те объекты, параметр которых неизвестен (даже неизмерим), теория порядковых статистик позволяет поставить и решить совершенно новые прикладные задачи, такие, как, например, идентификация объекта с ненаблюдаемым входом.

1. ОБРАЗОВАНИЕ И УПОРЯДОЧЕНИЕ ВЫБОРКИ

Оперировать выборкой вместо совокупности — обычный исследовательский прием. Прибегают к нему по нескольким причинам. Во-первых, зачастую оказывается физически невозможным подвергнуть контролю всю генеральную совокупность. Кроме того, нередки случаи, когда испытание образца — элемента совокупности — связано с его порчей или утратой. Наконец, «выборочный метод» [1], если он правильно применен, может дать вполне приемлемую точность.

Поскольку нашей темой являются свойства упорядоченной выборки, начнем с того, как получить и упорядочить выборку.

«Будем оценивать свойства закона распределения генеральной совокупности X по выборке из n независимых значений x_1, x_2, \dots, x_n ».

«Образовав выборку, ранжируем ее значения»...

Для тех, кто часто общается с литературой по теории вероятностей и математической статистике, приведенные формулировки настолько привычны, что не задерживаются в сознании, а для тех, кто встречается с ними нечасто, они представляют определенную опасность, поскольку могут создать иллюзию, будто образовать выборку для последующих (законных!) выводов по ней легко и просто.

Рассмотрим повнимательнее понятия «образование выборки» и «ранжирование выборки».

Образование выборки

Вам укололи палец, выдавили из него капельку крови и через некоторое время объявили, что содержание гемоглобина в норме. Где же «все в порядке» с гемоглобином — в паль-

це, откуда бралась проба крови? Да нет, результат анализа мы смело переносим на весь объем крови в организме. Эта уверенность базируется, во-первых, на допущении, что в процессе кровообращения кровь перемешивается достаточно хорошо, а во-вторых, на том, что предыдущее предположение не противоречит практике.

Таким образом, перед нами пример того, как малая проба хорошо отражает свойства некоторого объекта в целом. К сожалению, так бывает далеко не всегда. Когда, например, на элеватор приходит машина с хлебом и требуется определить его влажность, пробы, взятые из разных точек объема зерна, могут дать существенно различные результаты хотя бы в силу того, что верхний слой мог просохнуть в пути либо, напротив, намокнуть.

Если поручить неподготовленному человеку набрать пробу кускового материала — щебня, угля, руды и т. д. — для определения состава по крупности, можно наблюдать интересную картину того, как образцы, заполнившие контейнер, будут бессознательно подобраны им «по руке» — ни мелочь, ни крупные глыбы не войдут в пробу.

Таким образом, для организации «репрезентативной» (представительной) выборки, выборки, хорошо отражающей свойства всей «генеральной совокупности», необходимо затратить специальные усилия. Существуют инструкции по отбору проб воздуха, воды, сыпучих веществ и т. д., обеспечивающие представительность выборок.

Кроме представительности, «хорошая» выборка должна состоять из независимых элементов. Это свойство также нелегко обеспечить в целом ряде испытаний. При отборе пробы кускового материала человек, положив в контейнер несколько крупных камней, как правило, «компенсирует» их целой пригоршней мелочи, причем делает это бессознательно.

В книге У. Кокрена [1] приводится несколько распространенных способов отбора, различающихся принципом, на котором построена процедура.

1. Отбор, ограничивающийся легко доступной частью совокупности. Например, выборка угля из открытого вагона берется с небольшой глубины.

2. Отбор производится беспорядочно. Исследователь, выбирая десять кроликов из большой клетки в лаборатории, может делать это без продуманного плана.

3. Имеется небольшая, но неоднородная совокупность. Исследователь просматривает всю совокупность и отби-

рает небольшое число «типичных» единиц, т. е. единиц, отвечающих его представлению о среднем для совокупности. Такой метод называют иногда предвзятым или направленным отбором.

4. Выборка состоит преимущественно из добровольцев в исследованиях, где процесс измерения опасен или неприятен для обследуемого.

Очевидно, наилучшие результаты давала бы процедура отбора, исключающая участие человека в процессе принятия решения — включать или не включать данный образец в выборку. Существуют устройства вроде тех, что выбрасывают шары с номерами «Спортлото», либо программы, генерирующие случайные последовательности чисел, все они призваны выдавать статистически независимые и представительные выборки. В нашем примере с отбором пробы кускового материала следовало бы поступить следующим образом. Во-первых, все объекты исходной совокупности, куски, пронумеровать. Далее, обратиться к генератору случайных чисел и извлекать из кучи щебня те камни, чьи номера названы генератором. Эта процедура должна повторяться до тех пор, пока не будет сформирована выборка нужного объема.

Понятно, что описанный эксперимент нереален и может рассматриваться как некоторая «идеальная» процедура, к которой мы, впрочем, будем обращаться в дальнейшем.

Вот как образуют случайную выборку на практике при переписи населения [1], когда нужно задать дополнительные вопросы небольшой группе граждан: «В США 5 %-ная выборка была впервые применена в переписи 1940 г., когда дополнительные вопросы о роде занятий, происхождении, числе детей и т. д. задавали лицам, чьи фамилии падали на две из каждого 40 строк на лицевой и оборотной сторонах переписного листа. При переписи 1950 г. по 20 %-ной выборке (каждая пятая строка переписного листа) были получены сведения по таким данным, как доход, число лет обучения, миграции, служба в вооруженных силах. Путем отбора из этой 20 %-ной выборки каждого шестого человека дополнительно была получена выборка, дающая сведения о браках и числе рожденных детей. Кроме того, группа вопросов, касающихся сроков службы и состояния жилища, была разбита на пять подгрупп и ответы на вопросы были получены в каждом пятом доме».

Обратим внимание на то, что, генерируя выборку при помощи того или иного источника случайных чисел, мы

получаем набор значений (x_1, x_2, \dots, x_n) из некоторой генеральной совокупности значений x , а извлекая камешки из кучи щебня (множества K), мы образуем набор объектов x_1, x_2, \dots, x_n .

Измеряя, взвешивая, сжигая затем объекты x_i , мы получаем соответствующие им значения линейных размеров, масс, количества выделяющегося тепла — значения случайных параметров, описывающих множество K .

В руководствах по выборочному методу, как, например, у У. Кокрена, отмечается, что большое значение имеет и правильный выбор элементов x — единиц отбора: «Эти единицы должны вместе исчерпывать всю совокупность и не должны перекрывать одна другую, т. е. каждый элемент совокупности должен принадлежать одной и только одной единице. Иногда единицы выделяются очевидным образом, как, например, в совокупности электрических лампочек, где единицей отбора служит отдельная лампочка. Иногда приходится выбирать из нескольких возможных единиц отбора. Например, при обследовании людей в городе единицей отбора может быть отдельный человек, члены одной семьи или же все жители городского квартала. При выборочном изучении урожая сельскохозяйственных культур единицами отбора могут служить поля, фермы или же участки земли, форма и размеры которых заранее известны».

Здесь же существенно то, что, образовав выборку объектов $\{x\}$, нам понадобится воспользоваться еще и измерительным прибором для того, чтобы образовать выборку $\{x_i\}$ присущих этим объектам значений.

Кое-что о порядке

Прежде чем «наводить порядок» в выборке, вспомним кое-что о понятии «порядок», для чего будем рассматривать нашу выборку как некоторое множество, состоящее из n элементов.

Между элементами всякого множества можно устанавливать некоторые отношения, задаваемые набором определенных свойств. Так, отношение между элементами множества именуется отношением эквивалентности, если оно обладает свойствами рефлексивности, симметричности и транзитивности. Например, обычное равенство на множестве чисел есть отношение эквивалентности. Мы будем говорить в дальнейшем об отношениях порядка.

Отношений порядка между элементами множества мож-

но установить несколько, в зависимости от требований к характеру упорядоченности [2].

Говорят о порядках строгих и нестрогих, совершенных и несовершенных, линейных и древовидных [3].

Разный смысл можно вложить и в само слово «порядок». Расположение вещественных чисел по возрастанию есть порядок по отношению «больше».

Пусть M — некоторое множество, а 2^M — множество его подмножеств. Включение $M_1 \subseteq M_2$ является отношением, устанавливающим порядок на 2^M . Это — порядок по отношению «быть подмножеством».

Тип упорядочения, очевидно, будет зависеть от того, считаем ли мы возможным, чтобы каждый объект мог быть подчинен самому себе (как в случае нестрогого неравенства \leqslant или нестрогого включения \subseteq), или, наоборот, считаем, что объект не может быть старше самого себя (строгие неравенство $<$ и включение \subset).

В зависимости от этих предпосылок упорядоченность может быть «частичной» и «строгой».

Отношение на множестве называется *частичным порядком*, если оно рефлексивно, транзитивно и антисимметрично.

Отношения (\leqslant) и (\subseteq) являются отношениями частичного порядка.

Определим отношение $<$ на множестве M : для $a, b \in M$ отношение $<$ имеет место тогда и только тогда, когда $a < b$ и $a \neq b$. Если $a < b$, мы говорим, что a предшествует b (меньше b) или что b следует за a (больше a).

Отношение $<$ задает на множестве M строгий порядок. Оно антирефлексивно, транзитивно и антисимметрично. Это означает:

ни для какого $x \in M$ не выполнимо xAx ;
если xAy и yAz , то выполнено xAz ;
если выполнено xAy , то невозможно yAx .

Первые два свойства образуют определение строгого порядка, а третье из них следует.

Дальнейшее развитие понятия «порядок» связано с вопросом, все ли элементы множества попарно сравнимы между собой. Рассмотрим его на примере упорядочивания по включению [2].

Пусть M — множество русских слов. Слово x «старше» слова y , если y можно получить из x вычеркиванием нескольких букв слева и (или) справа. Это отношение (обозначим его $y < x$) задает на множестве русских слов неко-

торую упорядоченность. Например, «стол» < «столовая», «беда» < «победа». Но слова «облако» и «облатка» несравнимы. На рис. 1 показан фрагмент графа, изображающего старшинство (в указанном смысле) слов. Мы видим, что все слова, входящие в граф, младше слова «коловорот», но не все сравнимы между собой. Отношение старшинства не может быть установлено для слов «кол», «олово», «ворот» и т. д.

Существуют множества (например, множество вещественных чисел), для которых отношение порядка (строгого или нестрогого) может быть установлено между любыми элементами.

В зависимости от этого порядок подразделяют на «совершенный» и «несовершенный».

Отношение частичного (строгого) порядка $\leqslant (<)$ называется совершенным (иначе — линейным) порядком тогда и только тогда, когда $a \leqslant b$ ($a < b$) или $b \geqslant a$ ($b > a$) для всех $a, b \in M$. Если $\leqslant (<)$ совершенный порядок в M , упорядоченная пара $\langle M, \leqslant \rangle$ называется линейно упорядоченным множеством или цепью.

В соответствии с этим определением частичный порядок на множестве натуральных чисел — совершенный, а строгий порядок на рис. 1 не является совершенным.

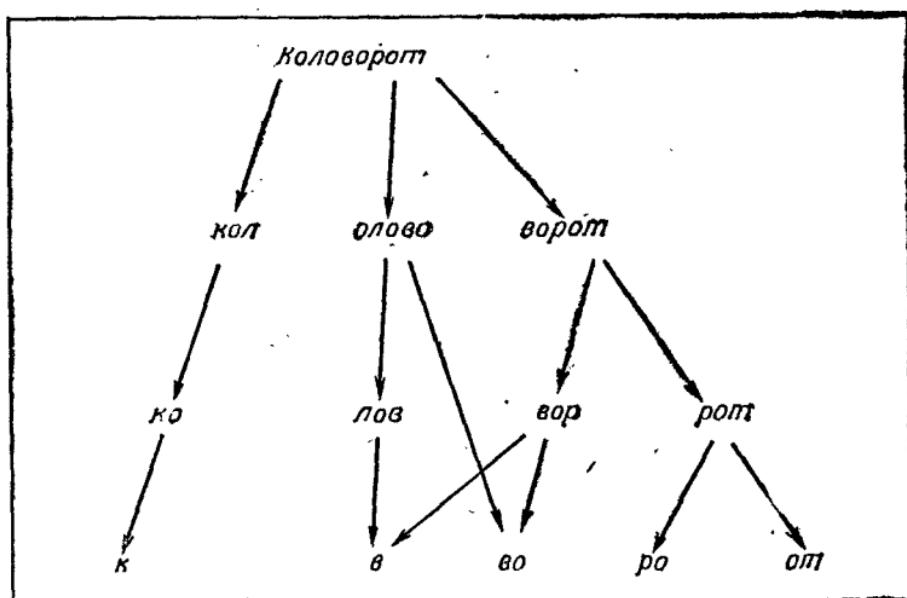


Рис. 1

Наличие несравнимых элементов заставляет определить и понятия непосредственного следования элементов друг за другом, а также понятия минимальный (максимальный) и наименьший (наибольший) элементы.

Если \leqslant ($<$) частичный (строгий) порядок на множестве M , элемент $b \in M$ называют непосредственно следующим за элементом $a \in M$ тогда и только тогда, когда $a < b$, и не существует такого $i \in M$, что $a < i < b$. Элемент a называют непосредственно предшествующим b .

Пусть множество $\{(1), (2), (1,2), (1, 2, 3)\}$ частично упорядочено по \subseteq . Здесь $(1, 2, 3)$ непосредственно следует за $(1, 2)$, а $(1, 2)$ непосредственно следует как за (1) , так и за (2) . Но $(1, 2, 3)$ не следует непосредственно ни за (1) ни за (2) , поскольку $(1) \subset (1, 2) \subset (1, 2, 3)$ и $(2) \subset (1, 2) \subset (1, 2, 3)$. Подмножества же (1) и (2) несравнимы и ни одно из них не может непосредственно следовать за другим.

Очевидно, лишь при строгом совершенном порядке все элементы упорядоченного множества являются непосредственно следующим (либо предшествующими).

Далее, на множестве с заданным отношением порядка элемент b называют минимальным, если не существует другого элемента $a \in M$, такого, что $a < b$.

Ясно, что для цепей понятие минимального элемента совпадает с понятием наименьшего элемента. В противном случае может оказаться, что элемент минимален, но не находится в соотношении $b < a$ с какими-либо иными элементами. Так, на рис. 1 слова «к», «в», «ро», и «от» — минимальные элементы, но не находятся друг с другом в отношении порядка (несравнимы). Слово же «коловорот» является одновременно и максимальным, и наибольшим элементом.

Если на множестве M задан совершенный строгий порядок и множество M конечно, то существует единственный максимальный (минимальный) элемент.

Пусть на множестве M определена функция $f : M \rightarrow R$, принимающая вещественные числовые значения.

Функция называется сохраняющей порядок относительно порядка \leqslant на M и порядка \leqslant на R в том случае, если $a \leqslant b$ в M влечет за собой $f(a) \leqslant f(b)$ в R .

Наконец, приведем одну важную теорему, позволяющую связать значение элемента выборки с его местом в упорядоченном ряду.

Пусть на конечном множестве M существует отношение совершенного строгого порядка. Тогда для подмножества $\{x_1, x_2, \dots, x_n\}$ можно выбрать такую нумерацию элементов

1, 2, ..., i, ..., j, n, что соотношение $x_i < x_j$ будет выполняться в том и только в том случае, если $i < j$.

Эта теорема утверждает, что любой совершенный строгий порядок на конечном множестве равносителен обычному порядку на некотором отрезке натурального ряда. Если же порядок на M не является совершенным, то, очевидно, элементы этого множества нельзя перенумеровать так, чтобы большим номерам соответствовали старшне элементы.

Как упорядочить выборку?

Нас в дальнейшем будут интересовать выборки, упорядоченные по значениям их элементов.

Условимся для определенности считать все элементы выборки сравнимыми между собой попарно, но несовпадающими по значениям. Это дает возможность говорить о совершенном строгом порядке. Для нумерации упорядоченных элементов будем использовать числа натурального ряда 1, 2, ..., n, хотя могли бы в принципе использовать любой другой его отрезок соответствующей длины.

Как мы действуем, упорядочивая последовательность чисел? Такой вопрос может поставить человека в тупик, во всяком случае, не подумав, не перечислишь операции, которые приходится проделать, превращая ряд чисел 2, 5, 1, 4, 3, 8 в ранжированную последовательность 1, 2, 3, 4, 5, 8. Да и нужно ли их перечислять? Ведь и так, неосознанно, мы справляемся с этой задачей!

Формализовать процедуру упорядочения, однако, необходимо, если мы хотим, чтобы эту работу выполняла ЭВМ. Это как раз работа для машины, особенно, если упорядочивать приходится большие массивы чисел. Познакомимся, следя [3], с некоторыми алгоритмами упорядочивания, именуемыми еще «алгоритмами сортировки»

Сортировка выбором. Пусть нужно ранжировать выборку x_1, x_2, \dots, x_n . Выбираем наибольший элемент из x_1, \dots, x_n и меняем его местами с x_n , затем выбираем наибольший элемент из последовательности x_1, \dots, x_{n-1} и меняем его местами с x_{n-1} и т. д. Этот метод требует $(n-1) + (n-2) + \dots + 1 = (n^2 - n)/2$ сравнений элементов между собой.

Сортировка попарной перестановкой. При первом просмотре каждое значение x_i сравнивается с x_{i+1} ; пары элементов, для которых справедливо неравенство $x_i > x_{i+1}$, меняются местами. Таким образом,

наибольший элемент попадает в положение x_n и в последующих просмотрах уже не участвует. Количество сравнений здесь примерно такое же, как и в первом способе, но зависит от начальных свойств выборки. Так, для выборки 3, 2, 1, 6, 4, 5 сортировка выполняется за три просмотра (при третьем просмотре перестановок нет), а сортировка выборки 2, 3, 4, 5, 6, 1 требует 5 просмотров только ради того, чтобы один последний элемент передвинуть на первое место. Видно, что «более ранжированная» исходная выборка требует меньшего количества просмотров.

Сортировка объединением именно это и использует. Элементы выборки соединяются попарно, причем меньший элемент ставится на первое место. Затем эти упорядоченные пары объединяются в четверки, четверки ранжируются и объединяются в восьмерки и т. д.

Метод наиболее эффективен, когда объем выборки равен степени двойки: $n=2^m$. При этом необходимое количество просмотров равно $\log_2 n$, а число сравнений не превосходит $n \log_2 n$.

При $n=1000$ попарное объединение потребует около 10 000 сравнений, а сортировка попарной перестановкой потребует их в 50 раз больше.

Чем оплачивается такая экономия? Метод сортировки объединением требует запоминания промежуточных групп элементов — нужно $2 n$ ячеек, — а метод попарной перестановки не требует промежуточного запоминания вовсе.

Эти же алгоритмы или их модификации используются, когда над выборкой требуется совершить какую-либо иную, связанную с упорядочиванием операцию — выбор наименьшего или наибольшего ее члена, элемента, занимающего среднее положение или имеющего определенный номер в упорядоченном ряду.

Нужно сказать, что в попытках формализовать процедуру упорядочения были перепробованы различные методы. Так, например, если смотреть на логическую операцию дизъюнкцию как на выбор максимального значения, а конъюнкцию придать смысл выбора минимального значения, то при помощи логических дизъюнктивно-конъюнктивных форм можно описать и процесс ранжирования, и другие процедуры сортировки [4].

Покажем это на примере выбора срединного значения — медианы выборки, предположив для простоты, что число членов выборки нечетно и что равных значений в выборке нет.

Итак, если объем выборки нечетен, $n=2h+1$, медианное значение в ранжированной выборке займет место с номером $h+1$. При этом оно будет больше всех предыдущих и меньше последующих, количество которых одинаково — h , и может быть найдено, как

$$\begin{aligned} z &= \text{med } (x_1, x_2, \dots, x_{2h+1}) = \\ &= \min (v_1, v_2, \dots, v_l) = \\ &= \max (u_1, u_2, \dots, u_l). \end{aligned}$$

Здесь v_i — наибольшие из всевозможных сочетаний элементов по $h+1$: $v_i = \max (x_1, x_2, \dots, x_h, \dots, x_{h+k})$.

Величины u_i определяются как наименьшие из сочетаний по $h+1$: $u_i = \min (x_1, x_2, \dots, x_h, \dots, x_{h+k})$.

Количество u_i и x_i одинаково: C_{2h+1}^{h+1} .

Пусть $n=3$, $x_1=1$, $x_2=2$, $x_3=3$. Процедура выделения медианы будет выглядеть так:

$$\begin{aligned} z &= \text{med } (x_1, x_2, x_3) = \max \{ \min (x_1, x_2), \min (x_2, x_3), \\ &\quad \min (x_1, x_3) \} = \max \{ \min (1, 2), \min (2, 3), \min (1, 3) \} = \\ &= \max \{ 1, 2, 1 \} = 2. \end{aligned}$$

Или так:

$$\begin{aligned} z &= \text{med } (x_1, x_2, x_3) = \min \{ \max (x_1, x_2), \max (x_2, x_3), \\ &\quad \max (x_1, x_3) \} = \min \{ 2, 3, 3 \} = 2. \end{aligned}$$

В рассмотренных алгоритмах сортировки мы оперировали с выборками значений x_1, \dots, x_n , заданных в виде чисел, и имели в качестве элементарной операции операцию сравнения двух элементов выборки x_i и x_j (двух чисел). Сравнение может быть осуществлено путем вычитания $x_i - x_j = y$. Положительный знак разности влечет вывод « x_i больше, чем x_j », отрицательный приводит к противоположному заключению.

А как производить ранжирование выборки, состоящей не из чисел, а из объектов $\{x_i\}$, принадлежащих множеству K ?

Очевидный путь — измерить $\{x_i\}$ и, определив значения $\{x_i\}$, оперировать ими, как было показано выше. К сожалению, очевидными путями не всегда удается воспользоваться. Как быть, если измерять объекты x_i нельзя? Ведь может оказаться, что интересующий нас параметр этих объектов неизмерим в принципе, как неизмерима, например, «сила игры» шахматиста (шахматной программы) или доблесть рыцаря.

Возможен и другой случай, когда параметр объекта нечем измерить из-за отсутствия измерительного прибора.

Так обстоит дело, например, с определением интенсивности запаха.

Однако мы видели, что алгоритмы сортировки используют значения x_i для вынесения суждений «больше — меньше» в процессе парных сравнений, а парные сравнения зачастую могут быть произведены и между элементами выборки $\{x_i\}$ непосредственно, без определения их значений.

Действительно, если речь идет о «силе игры» — параметре неизмеримом, то процедура типа турнира, основанная на парных взаимодействиях участников, дает материал именно для ранжирования их по значениям неизмеримого параметра.

Другой способ упорядочения — экспертные оценки — особенно эффективен при вынесении решений о парных соотношениях.

Наконец, для непосредственного сравнения объектов могут быть использованы специальные устройства сравнения — компараторы. Так, при ранжировании по весу компараторами могут быть рычажные весы без гирь. Электрические сигналы могут сравниваться между собой, например, на мостовых схемах и т. д.

Ранжированная выборка — объект с новыми свойствами

Почему выборка из X — вероятностный объект? Да потому, возникает естественный ответ, что ее элементы имеют случайные значения, подчиненные определенному закону распределения. Но не только по этой причине. Заинтересовавшись вопросами сортировки, выделения крайних, средних, да и вообще любых номеров, которые получают элементы при любом упорядочении выборки, мы немедленно столкнемся с новыми вероятностными свойствами ее элементов.

Действительно, выборка, содержащая одни и те же элементы, может быть реализована в опыте $n!$ разными способами в зависимости от порядка следования элементов. Если опыт поставлен правильно, все возможные реализации равновероятны и среди них с вероятностью $1/n!$ может оказаться выборка, ранжированная уже в процессе формирования. Напомним, что вероятность эта очень мала даже для небольших объемов выборки. При $n=12$ она равна $0,00000000208$ и, если на одну выборку уходит 1 с, требуемую реализацию можно ждать в течение нескольких ты-

сяч лет. Мы приводим этот пример для того, чтобы подчеркнуть, что операция упорядочения превращает ранжированную выборку в уникальный объект, настолько же редкий среди «естественных» реализаций, как может быть редок осмысленный текст среди случайных последовательностей букв.

Отличие ранжированной выборки от исходной, «естественней», количественно можно оценить общепринятой мерой беспорядка — энтропией.

Нам известно, что число реализаций выборки $n!$, а все реализации равновероятны. Энтропия в этом случае

$$H = -\log_2(1/n!) = \log_2 n!$$

Ранжированная выборка обладает энтропией, равной нулю. Уменьшение энтропии происходит, очевидно, в процессе упорядочения. Известно, что энтропия системы убывает в результате поступления информации, причем изменение энтропии равно количеству поступившей информации

$$= H = \log_2 n!$$

Откуда берется эта информация, где ее источник? Вспомним, что процесс упорядочения базируется на элементарной операции — парном сравнении, в случае сортировки естественной выборки — на сравнении двух равновероятных объектов. Такое сравнение сопровождается порождением одной двоичной единицы информации.

Мы рассмотрели несколько ранжирующих процедур и видели, что верхняя граница числа инверсий равна $n \log_2 n$, что и определяет то количество информации, которое (в принципе) может быть «закачано» в выборку в процессе упорядочения.

Весьма существенным в дальнейшем окажется то, что процедура упорядочения эквивалентна сложным нелинейным операциям над выборочными значениями. Покажем это для простейшей операции — выбора медианного значения [4]. Воспользуемся единичной ступенчатой функцией от разности выборочных значений $y = x_i - x_j$. Пусть функция $1(y)$ может принимать три значения в зависимости от соотношения x_i и x_j : $1(y) = 1$ при $y > 0$, $1(y) = 0$ при $y < 0$ и $1(y) = 1/2$ при $y = 0$. Тогда процедура выбора медианного значения в выборке из трех элементов может быть выражена в виде нелинейной функции от выборочных значений

$$z = \text{med}(x_1, x_2, x_3) = \sum_{l=1}^3 x_l [1(x_l - x_j) 1(x_k - x_l) + \\ + 1(x_j - x_i) 1(x_i - x_k)], x_i \neq x_j \neq x_k.$$

Итак, номер места, которое в результате ранжирования займет выборочное значение, может быть определен путем нелинейного преобразования над выборкой. Возможно, следует подчеркнуть, что упорядочение объектов может совершаться человеком неосознанно, на основе «скрытого знания». Эту операцию способны выполнять и животные, причем нам неизвестно, по каким алгоритмам действует их мозг. Описанные выше формальные процедуры также достаточно эффективны, если сопоставлять, например, необходимое число парных сравнений с числом выборок, которое пришлось бы на одну «естественно» ранжированную.

Мы познакомились, таким образом, с процедурами формирования и упорядочения выборок, выяснили, что стоит за беглой фразой «образуем и упорядочим выборку», и готовы воспользоваться, наконец, этой выборкой для того, чтобы судить по ней о свойствах того множества, которое она представляет.

Но мы уже так основательно потрудились над естественной выборкой, искусственно придали ей столько новых черт, что возникает мысль, а осталось ли в ней что-нибудь естественное, несет ли она еще черты генеральной совокупности породившего ее объекта, можно ли по ней делать адекватные этому объекту выводы?

Подобные сомнения обычны в науке. Известно, например, что на них базировались возражения против применения оптических средств в астрономии. Как можно верить в то, что линзы дают объективную картину небесных явлений, говорили скептики, если каждому известно, что они неспособны даже просто показать объект в его естественном положении, а переворачивают его вверх ногами! Одной из заслуг Галилея как раз и является то, что он развеял эти сомнения относительно телескопа.

Действительно, «естественная» выборка в результате всех произошедших с нею трансформаций не могла, став ранжированной, не приобрести новых свойств. Она подвергнута нелинейным преобразованиям, она «накачана» информацией, она — очень интересный объект. Исследуем его.

2. ВЕРОЯТНОСТНЫЕ СВОЙСТВА ПОРЯДКОВОЙ СТАТИСТИКИ

Итак, выяснив, что значит «образовать и упорядочить выборку», проделаем это применительно к генеральной совокупности с законом распределения вероятностей $F(x)$.

Пусть генеральная совокупность представляется в виде кучи щебня, отдельные объекты которой — камни — характеризуются случайным параметром X .

Выбрав n камней, мы образуем случайную выборку и можем ранжировать их по весу при помощи компаратора (весов). Как было выяснено выше, упорядочение типа «ранжирование» может быть выполнено путем парных сравнений, так что гири нам не понадобятся.

Случайная выборка имеет вид x_1, x_2, \dots, x_n , а после ранжирования — $x_{(1)}, x_{(2)}, \dots, x_{(m)}, \dots, x_{(n)}$. Индекс (m) означает номер значения в ранжированном ряду и называется «раингом». Ясно, что $1 \leq m \leq n$.

Проделаем следующий мысленный эксперимент. Образовав и упорядочив выборку, разложим камни в n ящиков и на каждом ящике обозначим ранг положенного туда камня. Повторим всю процедуру: опять образуем и ранжируем выборку и также отправим камни из второй выборки в ящики соответствующих рангов.

Проделав все это в третий и в четвертый и т. д. разы, мы обнаружим, что в ящике № 1 содержатся самые легкие камни из всех выборок, а в ящике № n — все самые тяжелые.

Такой процедурой, продолжая ее неограниченно долго, мы смогли бы в принципе разложить по ящикам всю генеральную совокупность, а затем, сложив всю ее вместе и смешав, восстановить снова.

Рассмотрим содержимое каждого ящика более внимательно. Разумеется, веса камней одного и того же ранга в различных выборках вовсе не обязательно одинаковы. Более того, они случайны и подчинены определенному закону распределения. Иначе говоря, они являются значениями некоторой случайной величины, называемой порядковой статистикой, которую мы будем обозначать $X_{(m)}$, где m — ранг.

Ранжированная выборка, таким образом, принимает вид $X_{(1)}, X_{(2)}, \dots, X_{(m)}, \dots, X_{(n)}$.

Ее элемент $X_{(m)}$ (совокупность значений x с рангом m) называют m -й порядковой статистикой. Элементы $X_{(1)}$

и $X_{(n)}$ называются «крайними», или «экстремальными» порядковыми статистиками. Если n нечетно, значения с номером $m = (n+1)/2$ являются центральными. Если m порядка $n/2$, соответствующие порядковые статистики называются « m -ми центральными». Как определить понятие «крайний» для случая, когда выборка неограниченно увеличивается и $n \rightarrow \infty$? Если n растет, а вместе с ним растет и m , так что $m/n \rightarrow \lambda$ при $n \rightarrow \infty$, а $0 < \lambda < 1$, соответствующая порядковая статистика считается центральной. Если же при $n \rightarrow \infty$ $m/n \rightarrow 0$ либо $m/n \rightarrow 1$, порядковые статистики относятся к крайним. В этом определении можно усмотреть элемент произвола, заключающийся в том, что предполагаются сходными свойства элементов, оказавшихся, например, десятыми в сотне, сотыми в тысяче и т. д.

Закон распределения порядковой статистики

Выведем плотность распределения m -й порядковой статистики при объеме выборки n , предполагая закон распределения исходной совокупности таким, что его интегральная функция и плотность $F(x)$ и $f(x)$ непрерывны почти всюду. Мы будем иметь дело со случайной величиной $X_{(m)}$, область определения которой совпадает с областью определения исходной случайной величины X . Действительно, если X

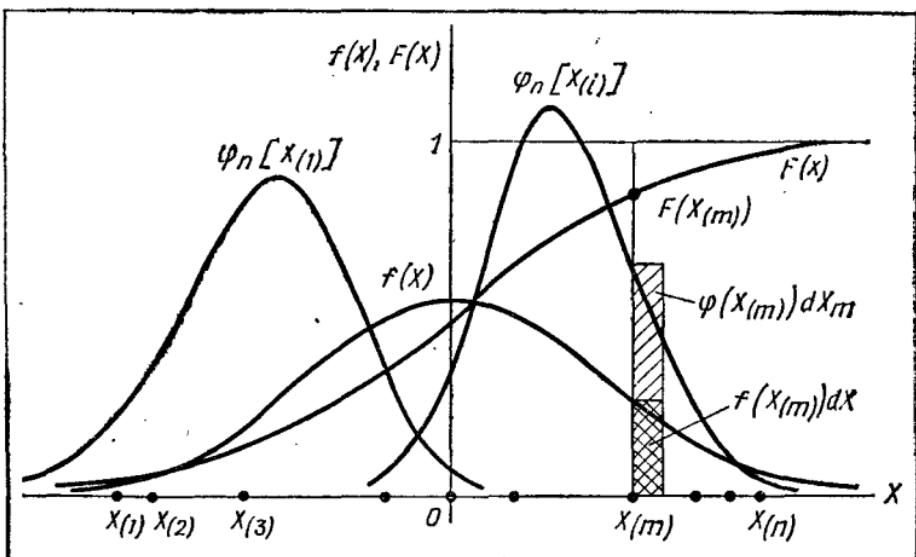


Рис. 2

ограничена, то не могут появиться значения $X_{(m)}$ (принадлежащие выборкам из X), выходящие за ее границы, и, наоборот, если X не ограничена, ничто не удержит $X_{(m)}$ в каких-либо точных пределах.

Обратимся к рис. 2, на котором изображены функции $F(x)$, $f(x)$ и искомая плотность распределения порядковой статистики $\Phi_n(x_{(m)})$. Индекс n указывает объем выборки. На ось x нанесены значения $x_{(1)}, \dots, x_{(m)}, \dots, x_{(n)}$, принадлежащие некоторой конкретной выборке. Запишем элемент вероятности $dF_n(x_{(m)})$, равный вероятности для порядковой статистики $X_{(m)}$ оказаться вблизи точки $x_{(m)}$:

$$p[x_{(m)} < X_{(m)} < x_{(m)} + dx_{(m)}] = \varphi_n(x_{(m)})dx_m. \quad (1)$$

Выразим эту же вероятность через исходный закон распределения, связав таким образом $\varphi_n(x_{(m)})$ с $F(x)$. Будем считать, что процесс образования выборки $x_1, \dots, x_t, \dots, x_n$ — процесс независимых испытаний, в которых «успехом» считается появление значения $X < x_{(m)}$, а «неуспехом» — $X > x_{(m)}$. Очевидно, что вероятность успеха $F(x_{(m)})$, а неуспеха $1 - F(x_{(m)})$ (рис. 2). Количество «успехов» равно $m-1$, а «неуспехов» $n-m$, поскольку m -е значение x_m в выборке объема n таково, что $m-1$ значений меньше и $n-m$ значений больше его.

Нетрудно видеть, что речь идет о процедуре подсчета вероятностей, сходной с той, которая приводит к биномиальному закону распределения.

Вероятность для исходной случайной величины принять значение, близкое к $x_{(m)}$, есть элемент вероятности $dF(x_{(m)}) = f(x_{(m)})dx$.

Вероятность расположения выборки вокруг значения $x_{(m)}$ так, что $m-1$ элементов ее окажутся слева, $n-m$ справа, а сама случайная величина X вблизи него и будет равной

$$C_{n-1}^{m-1} [F(x_{(m)})]^{m-1} [1 - F(x_{(m)})]^{n-m} f(x_m) dx. \quad (2)$$

Но именно эта вероятность и определяется выражением (1)! Поэтому, приравняв (1) и (2), получим

$$\varphi_n(x_{(m)}) dx_{(m)} = C_{n-1}^{m-1} [F(x_{(m)})]^{m-1} [1 - F(x_{(m)})]^{n-m} f(x_m) dx.$$

Если при переходе от плотности $f(x)$ к $\varphi_n(x_{(m)})$ сохранить масштаб по оси x , то

$$\varphi_n(x_{(m)}) = C_{n-1}^{m-1} [F(x)]^{m-1} [1 - F(x)]^{n-m} f(x). \quad (3)$$

Последнее выражение показывает, что плотность распределения порядковой статистики зависит от исходного распределения и ранга m и изменяется с изменением объема выборки n . Формула (3) позволяет определить, в частности, как распределены значения крайних членов выборки, имеющих ранги $m=1$ и $m=n$.

Крайний справа максимальный член имеет функцию распределения $F^n(x)$, а минимальный — $1 - [1 - F(x)]^n$. Для примера продемонстрируем плотности порядковых статистик с рангами $m=1, 2, 3$ при объеме выборки $n=3$ из равномерно распределенной на отрезке $[0, 1]$ совокупности (рис. 3). В соответствии с (3) при исходной плотности $f(x)=1$ (и значит $F(x)=x$) получаем распределение наименьшего члена

$$\varphi_3(x_{(1)}) = 3(1 - 2x + x^2);$$

среднего члена

$$\varphi_3(x_{(2)}) = 6(x - x^2)$$

и максимального

$$\varphi_3(x_{(3)}) = 3x^2.$$

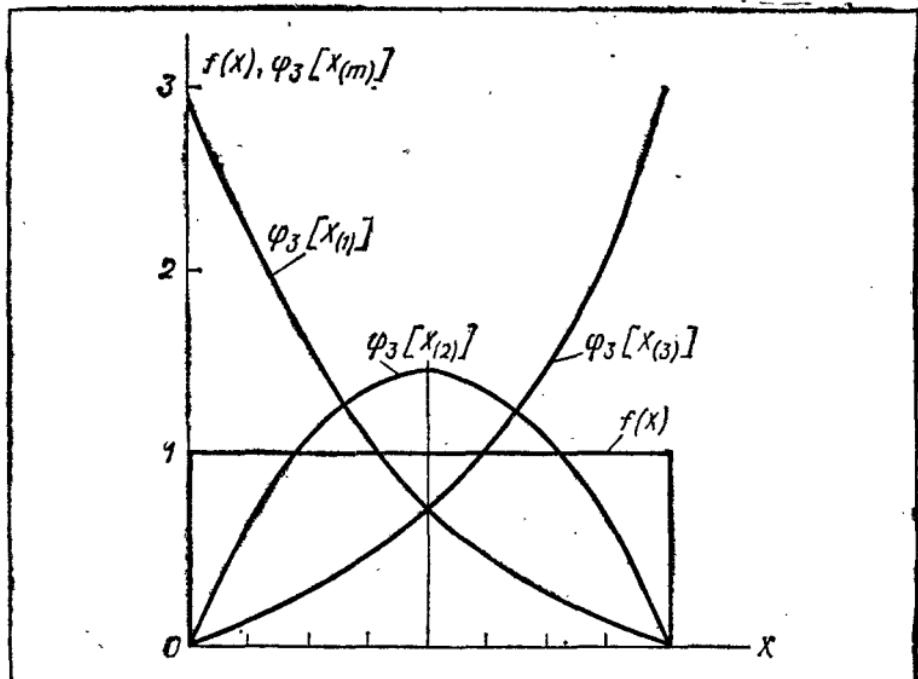


Рис. 3

Эти плотности изображены на рис. 3. В полном согласии с интуитивными представлениями плотность центрального члена выборки симметрична относительно медианы исходного распределения, а плотности крайних ограничены тем же интервалом, что и $f(x)$, и возрастают к соответствующей границе.

Продемонстрируем на этом же примере еще одно любопытное свойство распределений порядковых статистик. Сложим плотности $\varphi_3(x_{(1)})$, $\varphi_3(x_{(2)})$, $\varphi_3(x_{(3)})$ и разделим результат на их число:

$$\frac{1}{3} \sum_{m=1}^3 \varphi_3(x_{(m)}) = \frac{1}{3} (3 - 6x + 3x^2 + \\ + 6x - 6x^2 + 3x^2) = 1 = f(x)$$

на отрезке $[0, 1]$.

Сумма (нормированная) плотностей порядковых статистик оказалась равной исходной плотности $f(x)$! Это значит, что генеральная совокупность X является смесью порядковых статистик $X_{(m)}$. Этого результата следовало ожидать. Выше мы уже упоминали, что, рассортировав исходную совокупность по рангам, мы могли бы, вновь смешав объекты с различными рангами, восстановить ее. Теперь мы убедились в этом на примере, но можно было бы привести и строгое доказательство этого свойства.

Числовые характеристики порядковой статистики

Порядковая статистика, как всякая случайная величина, может описываться рядом числовых характеристик, в том числе и моментами распределения.

Значения моментов определяются видом исходного распределения $F(x)$, объемом выборки n и рангом m .

Обозначив через $\mu_{n,m}^k$ начальный момент k -го порядка m -й порядковой статистики в выборке объема n , получаем, что

$$\mu_{n,m}^k = \int_{-\infty}^{\infty} x_{(m)}^k \varphi_n(x_{(m)}) dx_{(m)}. \quad (4)$$

Аналогично $\mu_{n,i,j}$ — смешанный начальный момент второго порядка i -й и j -й порядковых статистик:

$$\mu_{n, i, j} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_{(i)} x_{(j)} \varphi_n(x_{(i)}, x_{(j)}) dx_{(i)} dx_{(j)}.$$

Отсюда ковариация (момент, характеризующий степень связности двух величин) $x_{(i)}$ и $x_{(j)}$

$$V_n[X_{(i)} X_{(j)}] = \mu_{n, i, j} - \mu_{n, i} \mu_{n, j}. \quad (5)$$

Эту характеристику мы обсудим несколько позже.

Из (4) при $k=1$ получается математическое ожидание m -й порядковой статистики $\mu_{n, m} = E_n[X_{(m)}]$, а из (5) при $i=j=m$ — ее дисперсия

$$D_n[X_{(m)}] = V_n[X_{(m)} X_{(m)}].$$

Выражения для средних $E_n[X_{(m)}]$ и дисперсий $D_n[X_{(m)}]$ очень просты при равномерном распределении исходной совокупности

$$E_n[X_{(m)}] = \frac{m}{n+1};$$

$$D_n[X_{(m)}] = \frac{m(n-m+1)}{(n+1)^2(n+2)}.$$

Вычислим эти характеристики при $n=3$ и $m=1, 2, 3$. Получаем $E_3[X_{(1)}]=1/4$ — математическое ожидание крайнего слева, $E_3[X_{(2)}]=1/2$ — центрального, $E_3[X_{(3)}]=-3/4$ — крайнего справа членов выборки.

При $n=7$ последовательность средних $E_7[X_{(m)}]$ выглядит так: $1/8, 2/8, \dots, 7/8$. Точки, соответствующие математическим ожиданиям при $n=3$ и $n=7$, нанесены на рис. 3. Разумеется, при любом n средние не выйдут за границы распределения.

Значения математических ожиданий образуют на оси x систему точек, такую, что их взаимное положение не зависит от параметра сдвига (средней величины) исходного распределения. Более того, при изменении масштабного параметра (дисперсии) исходного закона не меняются относительные расстояния между $E_n[X_{(m)}]$. Системы средних $E_n[X_{(m)}]$ для нормированного и центрированного нормального закона $F(x)$ называют «нормальными метками» [5]. Ниже мы покажем, как можно использовать шкалы, подобные «нормальным меткам».

Дисперсии крайних равны и при $n=3$ $D_3[X_{(1)}] = D_3[X_{(3)}] = 3/80$, дисперсия центрального значения $D_3[X_2] = 4/80$.

При увеличении выборки до $n=7$ дисперсии как крайних, так и центрального значения уменьшается: $D_7[X_{(1)}] = D_7[X_{(7)}] = 7/596$, $D_7[X_{(4)}] = 16/596$.

Отметим, что дисперсия крайних членов больше, чем центрального значения. Вообще, существует закономерность относительно порядковой статистики с наименьшей дисперсией: если симметричная функция $f(x)$ имеет максимум (или минимум) в медиане, то дисперсия m -го значения обладает минимумом (максимумом) в медиане. Дисперсия увеличивается (уменьшается), если мы движемся от центра симметрии к периферии распределения.

Распределения при неограниченном увеличении выборки. Распределения центральных значений

В рассмотренном выше примере исходное распределение было ограничено отрезком $[0, 1]$, а выборка мала — $n=3$. Центральное значение было единственным, распределение его — симметричным. При увеличении n и неограниченном x центральные значения (такие, что $m/n \rightarrow \lambda$ при $n \rightarrow \infty$) постепенно меняют свой вид, приближаясь к нормальным.

Параметры этих распределений зависят от m и n , а плотности имеют вид:

$$\varphi_n(x_{(m)}) = C \exp \left[-\frac{n(x - \hat{x})^2 f^2(\hat{x})}{2F(\hat{x})[1 - F(\hat{x})]} \right].$$

Здесь $C=\text{const}$, а \hat{x} определяется из уравнения

$$F(\hat{x}) = m/(n+1).$$

Эта формула справедлива асимптотически для симметричных и слабо асимметричных распределений, у которых медиана находится вблизи моды и для рангов в окрестностях медианы. Обычно предполагают, что приближение, даваемое формулой, достаточно в пределах интервала $0,15 \leq F(x) \leq 0,85$, хотя оно и зависит от вида распределения и объема выборки. Как мы видели на примере, распределение медианного значения выборки всего из трех элементов уже несет черты внешнего сходства с нормальным.

Вне указанного интервала распределения порядковых статистик уже не похожи на нормальные.

Распределения крайних значений

Крайние члены выборки представляют большой интерес, поскольку они зачастую играют в группе особую роль. Так, сколько бы раз мы ни растягивали цепь, разорвавшее звено — крайнее слева в прочностном ряду звеньев. Забеги и заплыwy на соревнованиях выделяют сильнейших, которые соревнуются между собой на следующих этапах соревнований. Есть такие экзотические виды спорта, где зачет идет по слабейшему члену команды и побеждает та, которая обзавелась «лучшим худшим».

Рассмотрим вероятностные свойства крайних. Они оказываются весьма интересными.

Если образована выборка объема n , то вероятность, что все ее элементы окажутся меньше определенного значения x , очевидно, равна $F^n(x)$. Вероятность $\Phi_n(x)$ того, что наибольший элемент этой выборки окажется меньше того же значения x , такая же

$$\Phi_n(x) = F^n(x). \quad (6)$$

Напомним, что наибольшая наблюдаемая величина в выборке не фиксированное значение, а самостоятельная случайная величина. Свойства функции $\Phi_n(x)$ зависят от характера закона распределения $F(x)$, в основном в области больших значений x , а так как $x_{(n)}$ велико именно в больших выборках, то, значит, при $n \rightarrow \infty$. Если n и $\Phi_n(x)$ заданы тем или иным способом, то $\Phi_n(x)$ тем самым уже установлена. При увеличении n $\Phi_n(x)$ сдвигается вправо и, как оказывается, стремится к асимптотической форме.

Рассуждения, аналогичные тем, что привели к (2), позволяют получить вероятность того, что наименьшее значение превзойдет x :

$$\Pi_n(x) = [1 - F(x)]^n. \quad (7)$$

Выражения (6) и (7) определяют и функции плотности:

$$\varphi_n(x) = n [F(x)]^{n-1} f(x)$$

$$\text{и} \quad \pi_n(x) = n [1 - F(x)]^{n-1} f(x).$$

Если плотность исходной совокупности симметрична, наибольшее и наименьшее значения распределены взаимно симметрично: $\varphi_n(x) = \varphi_n(-x)$, так что, располагая распределением наибольшего, можно найти распределение наименьшего значения. Наш пример с равномерным распределением (рис. 3) иллюстрирует это кривыми $\varphi_3(x_{(1)})$ и

$\varphi_3(x_{(3)})$. Это позволяет исследование экстремальных значений свести к исследованию только наибольшего. Если исходное распределение симметрично относительно нуля, вероятность, что наибольшее значение будет положительным, быстро приближается к единице при увеличении объема выборки.

Закон распределения наибольшего значения (6) можно представить в виде $\Phi_n(x) = \exp [n \ln F(x)]$ и проследить за изменением n . При больших n значение $F(x)$ приближается к единице, а показатель экспоненты превращается в неопределенность типа $\infty \cdot 0$ и, значит, $\Phi_n(x)$ определяется характером приближения $F(x)$ к единице. Таким образом, существование асимптотических выражений связано с некоторыми условиями, накладываемыми на исходное распределение при больших значениях.

При анализе экстремальных значений вводят новую характеристику крайних: период повторяемости. Если считать, что наблюдения проводятся через равные промежутки времени, число наблюдений приобретает размерность времени, а выражение

$$T(x) = \frac{1}{1 - F(x)}$$

называется периодом повторяемости значения, большего или равного x . Период повторяемости представляет собой ожидаемое число наблюдений, после которых появится одно значение, превосходящее x . Функция $T(x)$ удовлетворяет условию

$$\lim_{x \rightarrow +\infty} T(x) = 1.$$

Период повторяемости неограниченно растет с ростом x .

С периодом повторяемости связаны «характеристические» наибольшее и наименьшее значения

$$u_n = u_n(n), \quad u_1 = u_1(n).$$

Они представляют собой квантили исходного распределения

$$F(u_n) = 1 - 1/n, \quad F(u_1) = 1/n.$$

По определению ожидаемое число больших u_n или меньших u_1 значений равно единице. С ростом n u_n растет, а u_1 убывает. Для симметричных распределений

$$u_1 = -u_n.$$

Еще одна статистическая функция, важная для анализа экстремальных значений, называемая интенсивностью, определяется выражением

$$\eta(x) = f(x)/(1 - F(x))$$

и для характеристических экстремальных значений равна

$$\alpha_n = \eta(u_n) = nf(u_n);$$

$$\alpha_1 = \eta(u_1) = nf(u_1).$$

Эти величины называют экстремальными интенсивностями. С ростом n функция интенсивности, в зависимости от вида закона распределения, ведет себя по-разному: может возрастать, оставаться постоянной или убывать.

Установлено, что существуют всего три типа асимптотических распределений крайних. Все возможные исходные распределения классифицируются по тому, к распределению какого типа принадлежат их крайние.

К распределению первого типа приводят такие неограниченные исходные, у которых $P(x) = 1 - F(x)$ — вероятность, что значение превзойдет x , сходится к нулю не медленнее, чем e^{-x} . К этой категории принадлежит большинство применяемых на практике распределений — нормальное, экспоненциальное, логарифмически нормальное, логистическое и гамма-распределение.

На рис. 4 изображены плотности распределения крайних членов из выборки с нормальным исходным распределением при $n=5; 10$. Таков же характер изменения плотности крайних и для других законов, приводящих к распределению первого типа. Мы видим, что с увеличением n распределение все более тесно концентрируется относительно среднего, в свою очередь неограниченно возрастающего вместе с объемом выборки. Рост крайних происходит медленно, но известно, что при $n \rightarrow \infty$ максимальный член с вероятностью единица превзойдет любое наперед заданное число. Математическое ожидание крайних членов будет приближенно выражаться следующим равенством:

$$M_n [X_{(1), (n)}] = a - \sigma \sqrt{\ln n} + \sigma \frac{\ln \ln n + \ln 4\pi}{2 \sqrt{2 \ln n}}.$$

Таким образом, «раздвигание границ» выборочного распределения с возрастанием объема выборки происходит крайне медленно, пропорционально $\sqrt{\ln n}$, при этом дисперсия крайних членов с ростом n может стремиться к нулю

(это видно на рис. 4), хотя также очень медленно. Это дает возможность гарантировать с вероятностью, сколь угодно близкой к единице, малость уклонений $|x_{(1)} - M_n [X_{(1)}]|$ и $|x_{(n)} - M_n [X_{(n)}]|$ при возрастании n .

Как показал Б. В. Гнеденко, предельные распределения первого типа выражаются так называемым «двойным показательным законом»

$$\Phi_n(x_{(n)}) \rightarrow e^{-e^{-x}}$$

и

$$\Phi_n(x_{(1)}) \rightarrow 1 - e^{-e^{-x}},$$

если x нормирован и центрирован.

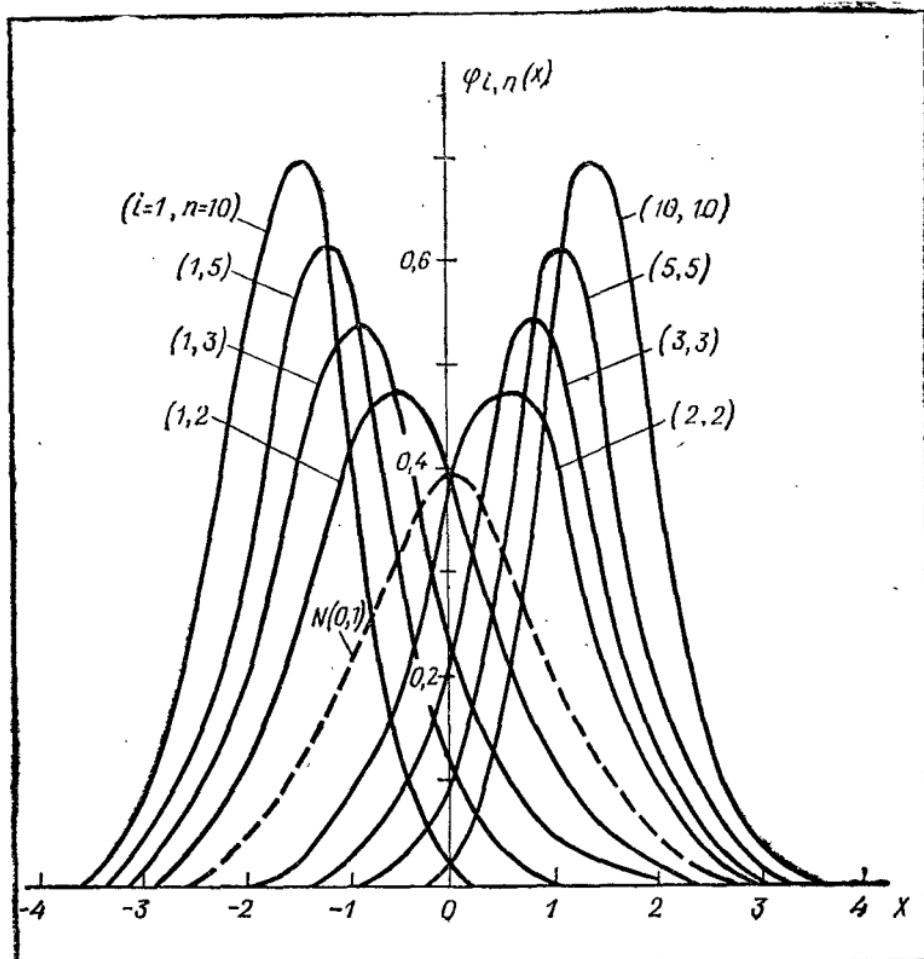


Рис. 4

Дисперсия наибольшего значения при этом имеет простой вид:

$$D_n[X_{(n)}] = \frac{\pi^2}{6\alpha_n},$$

где α_n — экстремальная интенсивность.

Экстремальная интенсивность с ростом n может возрастать, а дисперсия стремиться к нулю (нормальное распределение), оставаться постоянной, как и дисперсия (экспоненциальное и логистическое распределения), или убывать, что влечет за собой рост дисперсии (логарифмически нормальное распределение).

К распределению крайних третьего типа приводят ограниченные исходные распределения. Если распределение ограничено справа, т. е. $F(x)=1$ при $x \geq b$ (тогда как при $x < b$ $F(x) < 1$), то при достаточно больших n возможные значения максимального члена будут в основном сконцентрированы в соседстве с точкой b в интервале $(b-\epsilon, b)$ при любом $\epsilon > 0$.

На рис. 3 изображены распределения крайних выборок ($n=3, 7$) из исходной совокупности, подчиненной равномерному распределению. Видно, как кривые плотностей с ростом n приближаются к границам распределения. Дисперсия с ростом n неограниченно уменьшается.

Асимптотическое распределение имеет вид:

$$\Phi_n(t_n) \rightarrow e^{-|t|^{\alpha}}, t < 0,$$

где t — нормированное расстояние до границы b , а $\alpha = \text{const}$.

Существуют распределения, ограниченные с одной стороны и не ограниченные с другой. Так, например, показательное распределение, распределение Рэлея ограничены слева, но не ограничены справа. В выборках из таких исходных совокупностей распределение максимальных значений будет относиться к первому асимптотическому типу, а минимальных — к третьему.

К распределению крайних второго типа приводят неограниченные исходные распределения, плотность которых убывает с ростом n так медленно, что моментов не существует. К таким относятся, например, распределения Коши и Парето. Отметим лишь, что логарифмы наибольших значений, взятых из распределения типа Коши, распределены как наибольшие значения, взятые из распределения экспоненциального типа. Важно также, что с ростом n

дисперсии крайних не убывают, как в предыдущих случаях, а неограниченно растут.

Вернувшись теперь к нашей модели — куче щебня и ящикам с номерами, по которым раскладывали упорядоченные по весу камни, мы увидим, что если распределение веса камней приводит к распределению крайних I и III типа, то с увеличением n (количество ящиков) весовой состав щебня в них выравнивается — дисперсия убывает. Если бы мы имели дело с распределением II типа, то в ящиках с большими номерами оказались бы очень сильно отличающиеся по весу камни.

Статистике крайних посвящена подробная и известная монография Э. Гумбеля. [6].

Совместные распределения порядковых статистик

Напомним, что из исходной совокупности выбиралась случайная выборка. Если бы мы захотели написать выражение для закона совместного распределения двух каких-либо членов этой выборки, например X_i и X_j , нам следовало бы просто перемножить их плотности, так как эти случайные величины (если с выборкой «все в порядке») независимы.

По-другому обстоит дело в упорядоченной выборке. Порядковые статистики оказываются статистически связанными случайными величинами и совместная плотность распределения любой пары из них не сводится к произведению плотностей. Совместная плотность $\Psi_n(x_{(i)}, x_{(j)})$ i -й и j -й порядковых статистик зависит от объема выборки, значений i и j и исходного закона распределения [7]

$$\begin{aligned}\Psi_n(x_{(i)}, x_{(j)}) &= \frac{n!}{(i-1)! (j-i-1)! (n-j)!} \times \\ &\times [F(x_{(i)})]^{i-1} [F(x_{(j)}) - F(x_{(i)})]^{j-i-1} \times \\ &\times [F(x_{(j)})]^{n-j} f(x_{(i)}) f(x_{(j)}).\end{aligned}$$

Отсюда можно найти и второй начальный смешанный момент, и ковариацию.

Все ли порядковые статистики связаны? Запишем плотность совместного распределения крайних в выборке объема n

$$\Psi_n[X_{(1)}, X_{(n)}] = n(n-1) [F(x_{(n)}) - F(x_{(1)})]^{n-1} f(x_{(1)}) f(x_{(n)}).$$

Нетрудно видеть, что если неограниченно увеличивать n , то разность $F(x_n) - F(x_{(1)}) \rightarrow 1$, а $\psi_n(x_{(1)}, x_{(n)})$ будет приближаться к произведению $f(x_{(1)}) f(x_{(n)})$. Но если плотность совместного распределения двух случайных величин представляет собой произведение плотностей этих величин, то они статистически независимы. Таким образом, крайние члены вариационного ряда асимптотически независимы. Итак, при конечном n все порядковые статистики связаны и, в отличие от элементов случайной выборки, несут информацию друг о друге. Каким образом и для каких целей можно использовать эту информацию, мы увидим ниже.

Сейчас посмотрим, как связаны элементы выборки объемом $n=3$ из равномерно распределенной совокупности (предыдущий пример, рис. 3).

Подсчитаем величину ковариации V между центральным и крайними элементами выборки, а затем между крайними.

Для равномерного закона ковариация определяется выражением

$$V_n [X_{(i)} X_{(j)}] = \frac{i(n-j+1)}{(n+1)^2(n+2)}, \quad i < j.$$

В нашем случае, когда $n=3$, имеем для центрального и крайнего элементов $V_3 [X_{(1)} X_{(3)}] = V_3 [X_{(2)} X_{(3)}] = -2/21$, для двух крайних $V_3 [X_{(1)} X_{(3)}] = 1/21$.

С увеличением n связь между крайними быстро затухает, так как и минимальный и максимальный члены приближаются к границам распределения.

В заключение напомним, что выделить из генеральной совокупности X порядковые статистики — новые случайные величины X_m — можно при помощи процедур случайного выбора (так мы образуем выборки объема n) и парных сравнений для ранжирования элементов выборки. Сами значения при этом могут оставаться неизвестными.

3. ВЫБОРОЧНЫЕ ЗНАЧЕНИЯ И РАНГИ

Обратим теперь внимание на «ранги» — числа, входящие в качестве индексов в обозначения элементов упорядоченной выборки. Совокупность этих индексов обладает интерес-

ными вероятностными свойствами и, кроме того, может быть с пользой применена в практических целях.

Ранг, как уже было сказано, означает место элемента (или соответствующего ему значения) в ранжированной выборке. Наименьший элемент получает ранг $r = 1$, наибольший $r = n$. Очевидно, что ранги — целые положительные числа. Покажем сейчас, следуя [8], менее очевидную вещь, то, что ранг представляет собой случайную величину, и продемонстрируем ее вероятностные свойства.

Пусть образуется выборка объема n и из исходной совокупности извлекается очередной элемент, которому предстоит занять свое место в ранжированном ряду. Каков же будет его ранг? Очевидно, если значение x_i , присущее элементу, еще неизвестно или с другими элементами он не сравнивался, объективная возможность для него занять любое из мест в выборке одинакова. Это значит, что совокупность рангов — случайная n -мерная величина R — дискретна и распределена равновероятно так, что для всех i $p(r_i) = 1/n$.

Часто совокупность рангов выборки $R = \{r_1, \dots, r_n\}$ называют ранговым вектором. Он также случаен, его реализации являются перестановками чисел $1, 2, \dots, n$, число возможных реализаций равно $n!$

Ранжируя выборку, мы вместо исходной $\{x_i\}$ получаем пару вероятностных объектов — ранжированную выборку $\{x_{(i)}\}$ и вектор рангов R , состоящих с исходной выборкой в однозначном соответствии. Естественно, что по паре $(\{x_{(i)}\}, R)$ можно восстановить выборку $\{x_i\}$. Это значит, что упорядоченная выборка и вектор рангов содержат ту же информацию, что и исходная выборка.

Как упорядоченная выборка, так и вектор рангов описывают один и тот же объект — исходную выборку $\{x_i\}$, причем $\{x_i\}$ и R статистически независимы.

Совместное n -мерное распределение непрерывной величины $\{x_{(i)}\}$ и дискретной R представляет собой произведение распределений этих величин.

Это утверждение составляет содержание известной теоремы Гаека. Оно, на первый взгляд, противоречит интуитивному ощущению, что большее наблюдение x_i должно получить в выборке и больший ранг. Дело в том, что выше мы говорили о рангах в предположении, что значения элементов выборки x_i нам неизвестны. Как только появляется условие в виде значения x_i или ранга r_i , наблюдения и их ранги становятся статистически связанными.

Связь между значениями и их рангами

Если становится известным значение ранга r_i , то элемент выборки X_i при этом условии становится i -й порядковой статистикой $X_{(i)}$ с распределением

$$f(x_i/r_i) = \Phi_n(x_{(i)}) = C_{n-1}^{i-1} [F(x_{(i)})]^{i-1} [1 - F(x_{(i)})]^{n-i},$$

что совпадает с ранее полученным (5).

Рассмотрим некоторые количественные характеристики этой связи.

Выясним сначала, как ведут себя линии регрессии — условные математические ожидания значений и рангов. Они определяются уравнениями

$$E(R/x_i) = 1 + (n - 1) F(x_i)$$

$$\text{и} \quad E(X/r) = \int_{-\infty}^{\infty} x_i f(x_i/r) dx_i.$$

Последнее выражение при равномерном распределении исходной совокупности принимает простой и уже знакомый

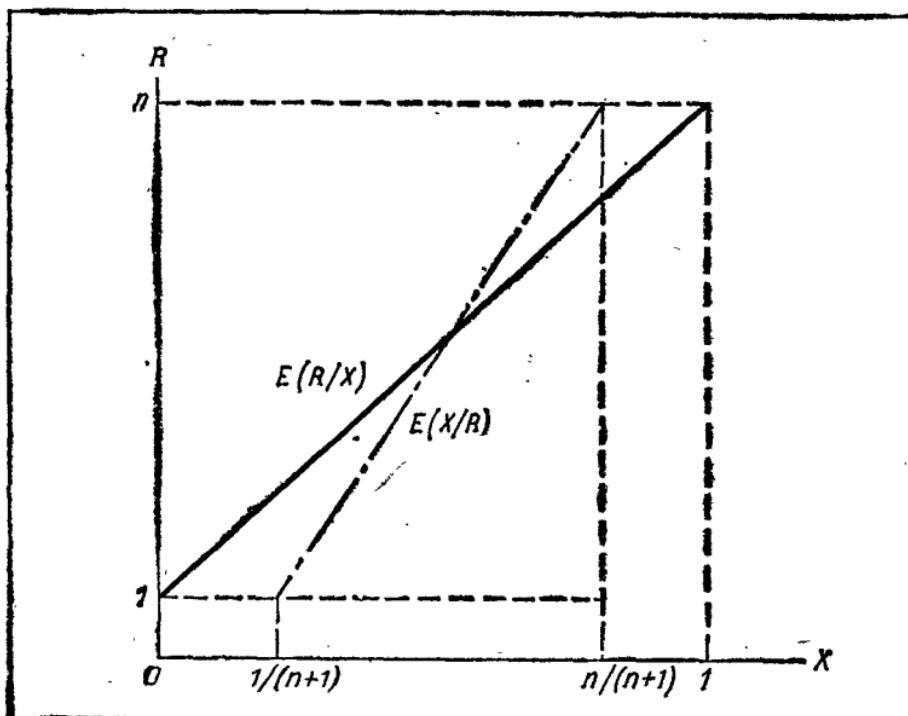


Рис. 5

нам вид: $E(X_i/r) = r/(n+1)$. Это не что иное, как среднее r -й порядковой статистики. Нанесем обе линии регрессии на один график (рис. 5). Они не совпадают, а ведь именно совпадение линий регрессии говорит, как известно, о функциональной зависимости между случайными величинами, образующими систему. Однако они и не перпендикулярны, что свидетельствовало бы о некоррелированности величин X и R .

Обратим внимание на то обстоятельство, что с увеличением объема выборки n угол между линиями неограниченно убывает, что говорит об усилении связи. Статистическая зависимость приближается к зависимости функциональной. Другая мера статистической связи — коэффициент корреляции, естественно, имеет ту же тенденцию — возрастать с увеличением n , но кроме того, он демонстрирует еще одно важное свойство зависимости между X_i и R .

Коэффициент корреляции между X и R по определению равен

$$\rho_{XR}(n) = \frac{E[(x - E_x)(r - E_R)]}{\sqrt{D_x} \sqrt{D_R}}.$$

Преобразование этого выражения с учетом того, что распределение R равномерно, приводит к

$$\rho_{XR}(n) = \sqrt{\frac{n-1}{n+1}} \rho[F(x)].$$

Произошла факторизация коэффициента корреляции. Один сомножитель зависит только от объема выборки и стремится к единице при $n \rightarrow \infty$, второй же определяется лишь видом исходного распределения $F(x)$. Замечательно то, что конструкция зависимости $\rho[F(x)]$ такова, что выполняется следующее равенство при любых a и σ :

$$\rho[F(x)] = \rho\left[F\left(\frac{x-a}{\sigma}\right)\right].$$

Это значит, что коэффициент корреляции между значением и рангом не зависит от параметров расположения и рассеяния этих случайных величин. Интуитивно это понятно. Действительно, «сдвинем» ранги, т. е. проинумеруем элементы выборки при $n=10$ не от 1 до 10, а от 11 до 20. Очевидно связь между значениями и их местами в ряду останется прежней. Выясним еще одно свойство связи между величинами X и R , вычислив количество информации,

которое они несут друг о друге. Преобразование общего выражения

$$J(X, R) \doteq \sum_{r=1}^n \int_X f(x, r) \ln \frac{f(x, r)}{\bar{f}(x) p(r)} dx, r = 1, 2, \dots, n,$$

где $p(r)$ — вероятность ранга r , приводит при $n \rightarrow \infty$ к асимптотической формуле

$$J(X, R) = \frac{1}{2} \ln n + \frac{1}{2} \ln \frac{e}{2\pi}. \quad (8)$$

Нетрудно видеть, что при неограниченном возрастании выборки количество информации в R о X (и наоборот) также неограниченно растет, причем принимает одинаковые значения для всех непрерывных распределений, поскольку $F(x)$ в (8) не входит.

Суммируем результаты. Итак, ранги и выборочные значения статистически связаны. С ростом объема выборки эта связь становится жестче, асимптотически переходя в однозначную функциональную зависимость, при которой одна величина полностью определяет другую. Зависимость между выборочными значениями и рангами не только не меняется с параметрами сдвига и масштаба, но и свободна (при больших n) от вида распределения исходной совокупности.

С выяснением этих свойств немедленно возникает вопрос: как можно и где нужно их использовать? Сейчас мы посмотрим, как можно пользоваться обнаруженной связью, а область применения ранговых процедур выясним позже.

Ранговая корреляция

Наберем камешков из кучи щебня (камешки, как мы помним, пронумерованы), взвесим их, получив выборку значений $\{x_i\}$, и измерим наибольший линейный размер каждого, образовав выборку значений $\{y_i\}$ другой случайной величины Y — максимального линейного размера. По физической сущности объекта ясно, что X и Y статистически связаны — более тяжелый образец в среднем и больший. Мерой связи выступает коэффициент корреляции

$$\rho_{XY} = \frac{E[(x - E_x)(y - E_y)]}{\sqrt{D_x} \sqrt{D_y}}, \quad (9)$$

оценка которого может быть определена по выборкам $\{x_i\}$ и $\{y_i\}$.

Мы выяснили, однако, что, упорядочив выборки $\{x_i\}$ и $\{y_i\}$ и оперируя их рангами R и K , мы можем в принципе располагать той же информацией, что содержалась и в исходных выборках. Значит ли это, что коэффициент корреляции между случайными величинами R и K отразит степень зависимости X и Y ?

Да, дело обстоит именно так. Выясним только, что значит «упорядочим выборки $\{x_i\}$ и $\{y_i\}$ ». Ранжировав $\{x_i\}$ и записав последовательность r_1, r_2, \dots, r_n , мы получим случайную перестановку значений вектора K . Попытавшись проделать эту процедуру с конца, мы придем к тому же: ранжированной станет выборка $\{y_i\}$, а последовательность значений R превратится в случайную перестановку.

Поступим следующим образом. Поставим в обеих выборках на места x_i и y_i соответствующие ранги r_i и k_i , получим две случайные перестановки (r_1, r_2, \dots, r_n) и (k_1, k_2, \dots, k_n) и вычислим коэффициент корреляции между ними — ранговый коэффициент корреляции Спирмена.

$$\rho(RK) = -\frac{\sum_{i=1}^n \left(r_i - \frac{n+1}{2}\right) \left(k_i - \frac{n+1}{2}\right)}{\frac{n}{12} (n^2 - 1)}. \quad (10)$$

Коэффициент этот «устроен» обычно, сравним его с (9). Здесь r_i и k_i — выборочные значения случайных величин, $(n+1)/2$ — среднее арифметическое рангов (первых n чисел натурального ряда)

$$(r_1 + \dots + r_n)/n = (1 + \dots + n)/n = (n+1)/2,$$

а $(n^2 - 1)/12$ — их дисперсия.

Есть смысл переписать (10) более компактно

$$\rho(RK) = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (r_i - k_i)^2. \quad (11)$$

Коэффициент ρ носит имя психолога Спирмена, который ввел его для обнаружения связи между признаками, не имеющими количественного выражения. Если упорядочения по этим признакам статистически связаны, коэффициент ρ принимает положительные значения, когда большая выраженность одного признака соответствует большей

му проявлению другого. В противном случае ρ отрицателен. Так [9], если речь идет об успехах детей в учебе и спорте, можно составить два списка учащихся класса, в которых фамилии пойдут в порядке возрастания успеваемости в учебе и успехов в спорте. Возьмем за основу нумерацию, данную первым списком. Номера во втором списке окажутся при этом рангами, которые ученики получают при упорядочении по второму признаку. Если между двумя признаками нет никакой зависимости, не будет никакой связи и между номерами в двух списках — в качестве второй нумерации с равными шансами могла бы появиться любая.

Коэффициент Спирмена наиболее употребителен среди возможных ранговых мер статистической связи вследствие своей простоты, но он далеко не единственный. Применяются еще и «коэффициенты беспорядка» [5]. Пусть значения, рангов R располагаются в натуральном порядке $1, 2, \dots, n$ а соответствующие ранги K , образующие перестановку $1, 2, \dots, n$, равны k_1, k_2, \dots, k_n . Естественный метод измерения беспорядка K -рангов, т. е. отклонений их от порядка $1, 2, \dots, n$, состоит в подсчете числа инверсий между ними. Например, при $n = 4$ в K -ранжировке 3214 имеются три инверсии, а именно 3—2, 3—1, 2—1. Число таких инверсий, обозначаемое Q , может изменяться от 0 до $n(n - 1)/2$. Коэффициент беспорядка

$$t = 1 - \frac{4Q}{n(n - 1)}$$

заключен в пределах $(-1, +1)$.

Коэффициент t можно «улучшить», если придавать инверсиям различный вес, например в ранжировке 24351 чувствуется, что инверсия 5—1 должна иметь больший вес, чем инверсия 4—3, поскольку она представляет собой более серьезное отклонение от натурального порядка $1, 2, \dots, n$. Наиболее простой способ взвешивания — «измерение» расстояния между рангами, образующими инверсию; в приведенном только что примере это дало бы соответственное веса 4 и 1 двум инверсиям. Однако использование весов $(r_i - k_i)$ возвращает нас к коэффициенту Спирмена, основанному, как мы видим, на сумме квадратов этих весов.

4. УЛУЧШЕНИЕ ОЦЕНОК ПУТЕМ ЦЕНЗУРИРОВАНИЯ ВЫБОРОК

Говоря о представительности выборки, мы обсуждали способы страховки от ошибок путем организации такой процедуры отбора, которая достаточно полно отразила бы в небольшой по объему выборке всю (в принципе бесконечную) генеральную совокупность.

Но вот выборка сформирована, и на ее основе будет сделан статистический вывод, например оценен какой-то параметр распределения генеральной совокупности. Точность при этом будет зависеть, при заданном объеме выборки, от свойств оценки и вида закона распределения. Но быть уверенным в том, что подсчитанная в такой ситуации точность оценки отвечает действительности, можно лишь в том случае, если высока надежность оценки и если мы полностью уверены в представительности и независимости выборки. Действительно, применение мощного математического аппарата, сулящего высокую точность оценок, к ненадежным данным не только бессмысленно, но и опасно, так как порождает необоснованные надежды «замолить грехи» эксперимента. С другой стороны, нерационально результаты надежного эксперимента обрабатывать «на глазок», с низкой точностью. Очевидно, между надежностью данных и точностью обработки существует некоторое оптимальное соотношение. К сожалению, даже относительно выборок, сформированных «по всем правилам», обычно остаются сомнения, причем касаются они, как правило, тех ее членов, которые занимают в вариационном ряду значений крайние места.

Сомнения эти порождаются двумя основными причинами.

Во-первых, нередки случаи, когда генеральная совокупность, из которой формируется выборка, в принципе неоднородна — представляет собой смесь двух или нескольких совокупностей, отличающихся, например, средними значениями. Ясно, что попадание в выборку нескольких «сорных» значений способно лишь исказить оценку среднего. Поэтому, если бы удалось «узнать» эти «сорные» значения и отбросить их, качество оценки могло бы возрасти. Разумеется, если вместо «сорных» отбросить «правильные» элементы выборки, перепутав их, точность понизится еще и за счет укорочения выборки.

Во-вторых, измерительные приборы имеют, как правило, наилучшие точностные характеристики в середине

шкалы. Наибольшие и наименьшие результаты измерений могут быть искажены нелинейностью градуировочной характеристики и действием помех, почему и вызывают подозрения относительно их надежности.

В упоминавшейся книге М. Кендалла и А. Стьюарта приводится очень удачный пример эксперимента, в котором происходит отбрасывание крайних (цензурирование выборки).

Пусть производится стрельба по круглой мишени радиуса R и регистрируется расстояние между точками попадания и центром мишени. При n выстрелах и m попаданиях разность $r = n - m$ составляет количество отброшенных (справа!) наблюдений, а уровень $x = R$ является уровнем урезания распределения X . В экспериментальной практике исследователи, вначале руководствуясь интуицией, издавна стали просто отбрасывать крайние наблюдения. Этот способ позднее подвергся строгому анализу и получил название «цензурирование выборок». Анализ, как это обычно и бывает при анализе интуитивных предположений, дал количественные оценки эффективности метода и указал границы его применимости. Мы покажем, что результаты анализа, а это бывает редко, прошли дальше и смелее интуитивных выводов: оказалось, например, что в некоторых случаях нужно отбрасывать не крайние, а средние наблюдения!

Цензурирование выборок — общая идея

Среднее арифметическое выборки (x_1, \dots, x_n) , являясь оценкой среднего, имеет вид:

$$\hat{E} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Понятно, что вычисление среднего арифметического ранжированной выборки $(x_{(1)}, \dots, x_{(n)})$ дает тот же результат

$$\hat{E} = \frac{1}{n} \sum_{i=1}^n x_{(i)}. \quad (12)$$

Отбрасывание крайних членов выборки (наименьших r_1 и наибольших r_2), выразится в том, что в (12) суммирование будет производиться в пределах $(r_1 + 1, n - r_2)$:

$$\hat{E} = \frac{1}{n - r_1 - r_2} \sum_{r_1+1}^{n-r_2} x_{(i)}, \quad (13)$$

что естественно, отразится на свойствах \hat{E} , как оценки среднего.

Выражение (13) можно трактовать, как суммирование порядковых статистик с весами. Наименьшие и наибольшие статистики получили веса, равные нулю, а все остальные — единице. К такому цензурированию приводило основанное на интуиции отбрасывание крайних членов. Гораздо более гибким оказалось взвешивание всех слагаемых (12), приводящее к линейной операции над порядковыми статистиками, когда вычисляется оценка

$$\hat{E} = \sum_{-1}^n c_i x_{(i)}, \quad \sum_{-1}^n c_i = 1. \quad (14)$$

Наборы коэффициентов c_i характеризуют степень доверия к значениям $x_{(i)}$, но выбираются не произвольно, а в зависимости от критерия точности оценки, закона распределения совокупности X , объема выборки n . Укажем следуя [4, 7], некоторые свойства коэффициентов c_i в зависимости от типов оценок \hat{E} .

Во-первых, часть коэффициентов принимает постоянные и равные значения, а остальные равны нулю. При этом часть порядковых статистик равномерно усредняется, а часть отбрасывается. Если отбрасываются крайние порядковые статистики, говорят что оценки получены по усеченным выборкам. Так могут быть получены выборочная медиана, среднее арифметическое трех или пяти порядковых статистик.

Известна простая и довольно точная оценка математического ожидания — среднее арифметическое двух «наилучших» порядковых статистик, симметрично расположенных относительно медианы:

$$\hat{E} = [x_{(i)} + x_{(j)}]/2,$$

где $i = [0,27 n] + 1$, $j = n - i + 1$, а $[]$ — целая часть числа.

Оценка, усредняющая все наблюдения, кроме двух крайних ($r_1 = 1$, $r_2 = 1$ в (13)), имеет вид:

$$\hat{E} = \frac{1}{n-2} \sum_2^{n-1} x_{(i)}.$$

Далее, коэффициенты c_i могут быть постоянными и зависеть от номера i — положения наблюдения в вариационном ряду. Так, существует формула

$$c_i = \frac{1}{n} h \left(\frac{i}{n+1} \right).$$

Функция h непрерывна и определена на отрезке $[0, 1]$. Она определяется законом распределения F и критерием оценивания. Так, для нормального закона и среднеквадратичного критерия точности при оценивании по усеченной выборке для $n = 7$, $r_1 = r_2 = 1$ оценка имеет вид:

$$\hat{E} = 0,2718 x_{(2)} + 0,1520 x_{(3)} + 0,1524 x_{(4)} + \\ + 0,1520 x_{(5)} + 0,2718 x_{(6)}.$$

Как отразилось усечение выборки на точности оценки? Дисперсия этой оценки равна $0,153 \sigma^2$, а оптимальная оценка по полной выборке $n = 7$ имеет дисперсию, равную $0,143 \sigma^2$. Таков проигрыш в точности. Выигрыш может быть получен в надежности результата и определяется тем, насколько справедливы сомнения относительно достоверности отброшенных крайних. Если эти сомнения несправедливы, проигрыш в точности ничем не компенсирован.

Отметим любопытный факт. Линейная по отношению к ранжированной выборке операция (14) над порядковыми статистиками соответствует нелинейному преобразованию над исходной (неупорядоченной) выборкой

$$\sum_1^n c_i x_{(i)} = \Phi(x_1, x_2, \dots, x_n),$$

поскольку уже сама операция упорядочивания выборки нелинейна.

Оценки параметров нормального распределения по усеченным выборкам

Оценивание параметров нормального распределения по малым усеченным выборкам сочетает высокую эффективность и простоту вычислений. Будем рассматривать выборки, усеченные с двух сторон — на r_1 наименьших и r_2 наибольших членов. Для них оценки математического ожидания и среднеквадратического отклонения вычисляются по формулам:

$$\hat{E} = \sum_{r_1+1}^{n-r_2} a_i x_{(i)} \quad (15)$$

и

$$\hat{\sigma} = \sum_{r_1+1}^{n-r_2} b_i x_{(i)}. \quad (16)$$

Весовые коэффициенты a_i и b_i определяются из выражений:

$$a_i = \frac{1}{n - r_1 - r_2} + \frac{\bar{u} (u_i - \bar{u})}{\sum_{r_1+1}^{n-r_2} (u_i - \bar{u})^2}$$

и

$$b_i = \frac{u_i - \bar{u}}{\sum_{r_1+1}^{n-r_2} (u_i - \bar{u})^2},$$

где u_i — математическое ожидание нормированной i -й порядковой статистики, а

$$\bar{u} = \frac{1}{n - r_1 - r_2} \cdot \sum_{r_1+1}^{n-r_2} u_i$$

— среднее неусеченные.

Таким путем получаются несмешанные оценки \hat{E} и $\hat{\sigma}$ не только при несимметричном, но даже и при одностороннем усечении выборки, когда $r_1 = 0$ или $r_2 = 0$.

Для получения оценок (15) и (16) достаточно, как мы видим, знать только математическое ожидание порядковых статистик.

Коэффициенты a_i растут от края ранжированной выборки к медианному (центральному) значению, которое и получает наибольший вес.

Эффективность оценки \hat{E} сравнивается с эффективностью оптимальной оценки и оказывается, как уже отмечалось, хуже. Проследим, как изменяется эффективность по мере отбрасывания крайних членов. Оказывается, что оценка \hat{E} мало чувствительна к присутствию крайних (их вес мал).

Так, если отброшены крайние и оценкой служит просто среднее из оставшихся (даже без вычисления a_i), ее эффе-

тивность может достигнуть 99 %. Если количество элементов в выборке n и усечены все (!) ее элементы, кроме центрального и соседнего с ним, оценкой \hat{E} служит центральное значение, соседнее можно не учитывать вовсе, а эффективность при этом составляет более 65 %. Например, если при $n = 10$ оставлены всего два наблюдения $x_{(5)}$ и $x_{(6)}$, а остальные усечены, эффективность оценки остается равной 72 %. Но стоит потерять хотя бы одно из этих центральных наблюдений ($x_{(6)}$, например), как эффективность падает за 60 %, даже если, с другой стороны, добавить все недостающие до полной выборки наблюдения.

Для оценки математического ожидания, таким образом, одно центральное наблюдение значит больше, чем половина выборки.

Оценка среднеквадратичного отклонения $\hat{\sigma}$ ведет себя в этом смысле по-другому: при любом фиксированном r_2 эффективность убывает на одну и ту же величину при каждом шаге уменьшения r_1 . Здесь усечения следует производить осторожно. Так, если при $n = 20$ отбрасывать два наблюдения с одного края или по одному с противоположных, эффективность снижается до 85 %. Отбрасывание третьего наблюдения (практически любого) снижает ее до 78 %, а каждого последующего — еще на 3 %.

Оценки параметров равномерного распределения

Равномерное распределение имеет функцию плотности $f(x) = \frac{1}{\vartheta}$, $E - \vartheta/2 \leq x \leq E + \vartheta/2$, где E — математическое ожидание, ϑ — размах.

Оценка среднего типа (14) с весовыми коэффициентами для неусеченной выборки имеет вид:

$$\hat{E} = [x_{(n)} + x_{(1)}]/2 \quad (17)$$

и представляет собой функцию лишь крайних членов выборки. Здесь «усечение» произвел алгоритм выбора весовых коэффициентов, обратив все, кроме крайних, в нули.

Оценка размаха

$$\hat{\vartheta} = \frac{n+1}{n-1} [x_{(n)} - x_{(1)}]$$

также зависит лишь от крайних значений. Дисперсия ошибки оценки среднего

$$D[\hat{E}] = \frac{\theta^2}{2(n+1)(n+2)}.$$

Мы видим, что точность оценки \hat{E} выше точности выборочного среднего. Напомним, что выборочное среднее — эффективная оценка для математического ожидания нормального распределения. Выражение (17) — нелинейная функция выборки (x_1, x_2, \dots, x_n) , дающая так называемую «суперэффективную» оценку, оценку, наилучшую среди всех возможных несмешанных оценок.

Если крайние наблюдения приходится отбрасывать ($r_1 \neq 0, r_2 \neq 0$), оценки имеют вид:

$$\begin{aligned}\hat{E} = & \frac{1}{2(n-r_1-r_2-1)} [(n-2r_2-1)x_{r_1+1} + \\ & + (n-2r_1-1)x_{(n-r_2)}]\end{aligned}$$

и

$$\hat{\theta} = \frac{n+1}{n-r_1-r_2-1} [x_{(n-r_2)} - x_{(r_1+1)}],$$

вновь являясь функцией лишь крайних из оставшихся членов.

Мы видим новый по сравнению с нормальным законом распределения эффект — ценность приобрели именно крайние члены, именно они определяют эффективность оценок.

Эффективность оценки \hat{E} не зависит от того, с какого края и какое количество наблюдений усечено. Эффективность наибольшая, когда при одинаковом общем числе отсутствующих наблюдений с обоих краев усечено их поровну. Для оценки размаха вообще важно знать лишь общий объем отброшенных наблюдений.

Очень чувствительны к потере крайних значений оценки начальной a и конечной b точек интервала распределения:

$$\hat{a} = x_{(r_1+1)} - \frac{r_1+1}{n-r_1-r_2-1} [x_{(n-r_2)} - x_{(r_1+1)}]$$

и

$$\hat{b} = x_{(n-r_2)} + \frac{r_1+1}{n-r_1-r_2-1} [x_{(n-r_2)} - x_{(r_1+1)}].$$

Стоит здесь утратить одно или два крайних значения, как точность оценки падает в несколько раз. Наблюдения же с противоположного края выборки не оказывают сколько-нибудь заметного влияния на точность оценки и не могут компенсировать потерю крайних, «с другого берега».

Эти свойства равномерного распределения (и целого класса других) должны предостеречь экспериментаторов от необдуманного отбрасывания крайних наблюдений. Всегда нужно иметь в виду, что правомочность этой процедуры определяется видом закона распределения наблюданной в эксперименте случайной величины.

Выборка из трех наблюдений

Изложим, следуя [7], практически важный случай трех наблюдений.

Пусть сделаны три измерения случайной величины с функцией плотности $f(x)$. Их порядковые статистики

$$x_{(1)} < x_{(2)} < x_{(3)}$$

могут стать основой для вычисления других интересных статистик

$$x' > x'' > x''',$$

которые определяются следующим образом:

если $x_{(2)} - x_{(1)} < x_{(3)} - x_{(2)}$, то $x' = x_{(2)}$, $x'' = x_{(1)}$, $x''' = x_{(3)}$,
а если $x_{(2)} - x_{(1)} > x_{(3)} - x_{(1)}$, то $x' = x_{(3)}$, $x'' = x_{(2)}$, $x''' = x_{(1)}$.

Другими словами, наблюдения x' и x'' лежат ближе друг к другу, а третье x''' — в стороне от них и может быть наименьшим или наибольшим из всех трех.

Эти статистики определяются не только своим номером после упорядочения, как порядковые статистики, а представляют собой в некотором смысле их обобщения.

Тройные наблюдения — весьма распространенный в практике экспериментирования прием. Третье измерение иногда делают, чтобы выяснить, какое из первых двух ошибочно. Если при этом два из трех наблюдений хорошо сходятся между собой, у исследователя возникает искушение отвергнуть «отскочившее» третье, как ошибочное. При этом он полагает, что улучшит результат, если сделает дополн-

нительное измерение, а худшее отбросит, как если бы его и не было.

Посмотрим, как влияет такая отбраковка наблюдения на оценки математического ожидания и дисперсии.

Оказывается, что при нормальном и равномерном распределениях пара ближайших друг к другу наблюдений обладает следующими свойствами.

Математическое ожидание разности $x' - x''$ составляет менее половины математического ожидания размаха выборки объемом $n = 2$. В таком же соотношении находятся среднеквадратичные отклонения этих разностей. Итак, отбрасывание «третьего лишнего» исказило свойство такой важной статистики, как размах. Лучше было бы пользоваться естественной выборкой объема $n = 2$.

Среднее из двух ближайших $(x' + x'')/2$ — состоятельная оценка математического ожидания, как и среднее из выборки объема $n = 2$, но с несколько большей дисперсией. Особенно интересно, что среднее из двух наиболее расходящихся наблюдений $(x_{(1)} + x_{(3)})/2$ служит лучшей оценкой математического ожидания, чем среднее из двух ближайших друг к другу наблюдений.

Итак, отбраковка «отскочившего» наблюдения ухудшает свойства и этой статистики.

Это, впрочем, не значит, что отбраковку вообще не следует делать. Отбраковка необходима, если есть основания считать «отскочившее» наблюдение «сорным» — принадлежащим чужому исходному распределению.

Само по себе «отскочившее» наблюдение x''' объединяет свойства крайних порядковых статистик $x_{(1)}$ и $x_{(3)}$. В случае равномерного на $[0, 1]$ исходного распределения случайная величина $\omega = x'''$ имеет плотность

$$\varphi(\omega) = 3(\omega^3 - \omega + 1/2),$$

равную нормированной сумме плотностей крайних порядковых статистик при $n = 3$, изображенных на рис. 2. Среднее $E(\omega) = 1/2$, т. е. совпадает с $E(X)$.

5. БЕЗЭТАЛОННЫЕ ПРОЦЕДУРЫ ИЗМЕРЕНИЯ, ИДЕНТИФИКАЦИИ И КЛАССИФИКАЦИИ

В предыдущих разделах было показано, как методы порядковых статистик работают «внутри» статистики, помогая

решать задачи оценивания, проверки гипотез, редактирования наблюдений.

Здесь мы продемонстрируем, как с их помощью могут быть решены некоторые прикладные задачи.

Измерение без эталона

Если мы хотим произвести измерение некоторого параметра, то должны располагать, во-первых, материальным носителем этого параметра — объектом измерения, во-вторых, эталоном — носителем единицы, в которой должен выражаться результат измерения, и, в-третьих, компаратором — устройством, при помощи которого объект измерения будет сравниваться с эталоном.

Так, если измеряемый параметр X — это вес (точнее, масса) тела, то объект измерения должен быть тяготеющей массой. Говорить о весе меридиана, улыбки или квадратного трехчлена бессмысленно. При взвешивании эталоном выступает образцовая мера веса, воплощенная в наборе гирь — разновесок. Значения разновесок образуют шкалу — систему эталонных мер. Компараторм являются привычные нам рычажные весы.

Объект X может быть случайной величиной либо процессом, принимать дискретные либо непрерывные значения.

Систему эталонов представим в виде вектора эталонных мер

$$\Delta = \langle \Delta x_1, \dots, \Delta x_i, \dots, \Delta x_k \rangle,$$

где величина i -го эталона в простейшем случае определяется, как

$$\Delta x_i = n^{-(i-1)}.$$

Здесь Δx_i — то, что принято именовать «ценой разряда»;

n — основание принятой системы счисления, в которой формируется результат, обычно $n = 10$ либо 2 ;

k — количество разрядов результата измерения;

Δx_k — младший разряд результата, имеющий смысл разрешающей способности.

Измерительная процедура заключается в том, что значение измеряемой величины последовательно локализуют в сужающихся областях $\Delta x_1, \dots, \Delta x_k$, образуя результат измерения — число $A = \langle \alpha_1, \dots, \alpha_i, \dots, \alpha_k \rangle$.

Измеренное значение с точностью до Δx_k представляется в виде скалярного произведения Δ и A

$$x = \Delta \cdot A = \sum_1^k \alpha_i x_i.$$

При взвешивании компаратор после добавления гири позволяет вынести суждение «больше» или «меньше» и решить, убрать или добавить следующую гирю. Δ_k здесь — наименьшая гиря разновеска, а $\alpha_1, \alpha_2, \dots$ число килограммов, сотен, десятков граммов и т. д.

Существенно, что процесс измерения заканчивается лишь указанием наблюдателю, где, в какой части диапазона измерения находится значение измеряемого объекта.

После этого наблюдатель волен сам выбрать точечную оценку измеряемого значения x . Ею может быть середина, правый или левый конец отрезка Δx_k или любая другая фиксированная его точка. Очевидно, что область возможных оценок совпадает с Δx_k , а расстояние между оценкой и точным значением x представляет собой ошибку — погрешность измерения. На рис. 6 изображен закон распределения измеряемой величины (зачем он нужен при измерении, мы поговорим позже), шкала и заключенное в разряде Δx_k измеряемое значение. В качестве оценки выбран левый конец отрезка Δx_k . Расстояние между оценкой и значением измеряемой величины является ошибкой ξ — случайной величиной, распределенной внутри Δx_k .

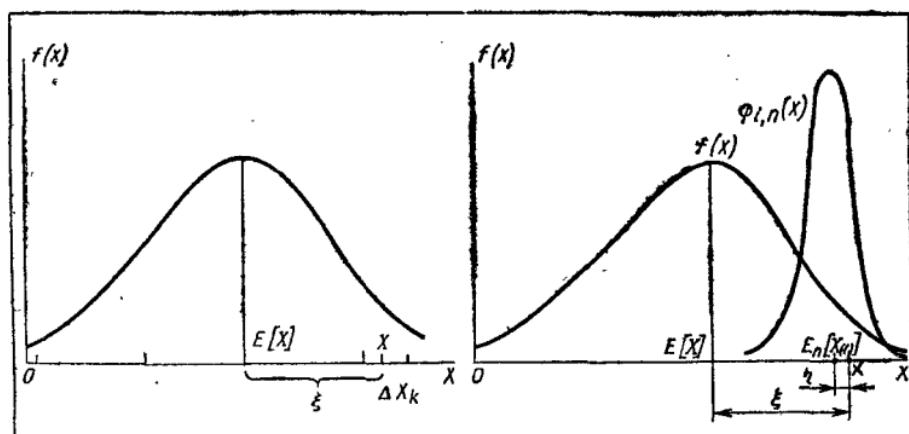


Рис. 6

Как мы видели, закон распределения не учитывался в процессе измерения. Если он известен, то в принципе можно построить оптимальную шкалу и получить ту же разрешающую способность за меньшее число сравнений.

Описанная измерительная процедура является основой для экспериментальной базы точных и естественных наук.

К сожалению, осуществить измерения возможно далеко не всегда. Можно указать две принципиальные причины, по которым измерение не может быть выполнено: отсутствие эталона и невозможность сравнения измеряемой величины с эталоном, пусть даже существующим.

Попытки преодолеть эти трудности привели к различным методам оценивания, называемым «неметрическим шкалированием». Термин «шкалирование» означает получение количественной информации об объекте в тех случаях, когда объективное измерение невозможно.

Ряд научных дисциплин обрел экспериментальную базу, разработав методы интервального неметрического шкалирования, создав, таким образом, суррогат измерения.

Шкалирование или ранжирование при помощи экспериментальных оценок представляет собой попытку справиться с трудностью второго рода. Подразумевается, что эксперт при этом выступает в роли носителя эталона (шкалы), с которым эвристически соотносит оцениваемую величину. Продводя аналогию между экспертом и измерительным прибором, можно, очевидно, говорить о «разрешающей способности» эксперта. Этот термин характеризует «систему эталонов» — «шкалу», которой вооружены органы чувств или интуиция эксперта, позволяющая ему проводить более или менее тонкие градации предъявляемых значений величины — от суждений «больше — меньше» до присвоения числовых значений.

Сложнее оценить количественно явления, для которых не существует эталона, даже неформального. Как, например, количественно оценить уже упоминавшиеся «добрость рыцаря», «силу игры» футбольной команды или «эвристическую силу» интеллектуальной (шахматной, например) программы?

Если эталона для измерения подобного свойства в принципе не существует, ранжирование объектов по значениям этой величины может в ряде случаев быть осуществлено путем организации специальной процедуры взаимодействия между объектами — процедуры типа игры, поединка, конфликта. При этом подразумевается, что каж-

дому из соперников присуще значение некоторого параметра — «силы игры» («доброты»), принципиально неизмеримого (иначе не нужен был бы турнир!); но объективно существующего и подчиненного некоторому закону распределения.

Теория измерений довольно скучно вознаграждает наблюдателя, снабженного прибором, за знание им закона распределения: с его помощью можно лишь несколько улучшить измерительную процедуру, повысив (на практике не намного) разрешающую способность.

Если же прибора нет, а закон $F(x)$ известен, лучшее, что может предпринять наблюдатель, выбрать в качестве оценки любого значения среднее — $E[X]$, что, конечно, представляет собой очень грубую оценку: дисперсия ошибки $D[\xi]$ равна при этом дисперсии величины X и улучшена быть не может.

Здесь мы продемонстрируем, как, пользуясь свойствами порядковых статистик, можно получить количественные оценки (иногда в принципе сколь угодно точные) величин, для измерения которых не существует эталона, но известен закон распределения и допустимо ранжирование. Применительно к нашему примеру со щебнем это означает, что мы, зная $F(x)$, будем пытаться взвесить отдельные камешки на весах, не имея... гирь!

Обратимся к идеи безэталонного измерения [10]. Пусть К-генеральная совокупность образцов κ ; X — генеральная совокупность присущих им значений x , подчиненная закону $F(x)$.

Образуем выборку из n -образцов $\kappa_1, \kappa_2, \dots, \kappa_n$ и ранжируем ее по параметру X , используя компаратор $\kappa_{(1)} < \kappa_{(2)} < \dots < \kappa_{(n)}$. Теперь и значения x_i , соответствующие образцам κ_i , образуют вариационный ряд $x_{(1)} < x_{(2)} < \dots < x_{(n)}$. Обратим внимание на то, что значения членов этого ряда остаются неизвестными для наблюдателя, причем он хочет оценить их. Элементы ряда $x_{(1)}, x_{(2)}, \dots$ представляют собой значения порядковых статистик $X_{(1)}, X_{(2)}, \dots$

В качестве точечной оценки $x_{(i)}$ неизвестного значения, занимающего i -е место в вариационном ряду, естественно принять математическое ожидание соответствующей порядковой статистики

$$\hat{x}_{(i)} = E_n[X_{(i)}]. \quad (18)$$

При этом допускается ошибка, значение которой $\eta =$

$= x_{(i)} - E_n [X_{(i)}]$, а дисперсия равна дисперсии i -й порядковой статистики.

Обратимся к рис. 6. На нем изображены кривые $f(x)$, плотность распределения i -й порядковой статистики $\varphi_n [X_{(i)}]$, неизвестное значение x_i и совокупность средних $E_n [X_{(i)}]$, образующая систему точек — шкалу. Оценка — среднее — $E_n [X_{(i)}]$ отстоит от оцениваемого значения x_i на величину ошибки η . На том же рисунке видна ошибка $\xi = x_i - E [X]$.

Разумеется, вводить $E_n [X_{(i)}]$ в качестве оценки имеет смысл лишь в том случае, когда дисперсия $D_n [X_{(i)}]$ оказывается существенно меньше, чем $D [X]$. Как было отмечено выше, подобный эффект наблюдается у широкого круга практически важных распределений, среди которых нормальное и все усеченные.

Рис. 7 и 8 демонстрируют, что при сравнительно небольших объемах выборок дисперсии порядковых статистик могут оказаться в десятки раз меньше, чем дисперсия исходной совокупности. Это значит, что, не сравнивая образец с эталонами, а лишь упорядочивая выборку образцов $\{x_i\}$ достаточного объема, можно, пользуясь оценкой (18), оценить значения $\{x_i\}$ элементов выборки сколь угодно

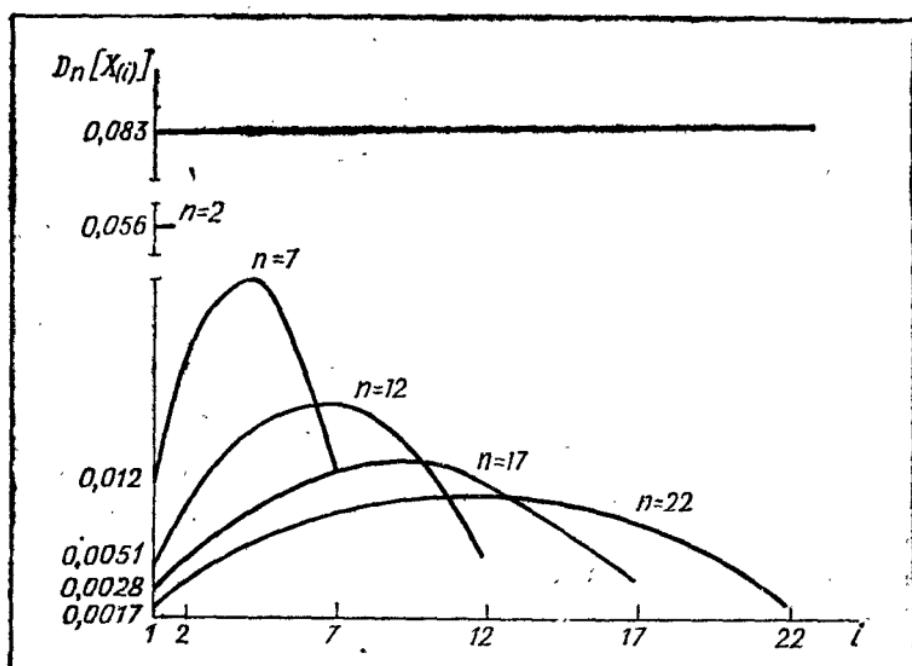


Рис. 7

точно в смысле дисперсии ошибки. Роль системы эталонов, роль «шкал» при этом играет множество средних $\{E_n[X_{(i)}]\}$. При ограниченном диапазоне распределения измеряемой величины и неограниченном росте объема выборки n расстояние между точками этой «шкалы» (ее «разрешающая способность») неограниченно уменьшается. Точность при таком способе измерений будет ограничиваться чувствительностью компаратора.

Рассмотрим пример. Пусть величина X распределена равномерно в промежутке $[0, 1]$. Требуется оценить значения элементов выборки объемом $n = 7$. Среднее $E[X] = 1/2$, а оценка любого выборочного значения средним $E[X]$ приводит к ошибке с дисперсией $D[X] = 1/12$.

Ранжируем 7 элементов x_i , и, зная закон распределения X , вычислим средние порядковых статистик и, нанося их значения на ось x , образуем шкалу. В нашем случае

$$E_n[X_{(i)}] = i/(n+1), \quad (19)$$

$n = 7, i = 1, 2, \dots, 7$ и

$$E_{7,1} = 1/8, E_{7,2} = 2/8, E_{7,3} = 3/8, \dots, E_{7,7} = 7/8.$$

Применяя оценку $\hat{x}_{(i)} = E_n[X_{(i)}]$, приписываем наименьшему из выборки $\{x_i\}$ значение $\hat{x}_{(1)} = 1/8$, следующему — $\hat{x}_{(2)} = 2/8$ и т. д. до $\hat{x}_{(7)} = 7/8$. Дисперсия ошибки оценивания будет различной для разных членов ранжированной выборки. Для крайних $D_7[X_{(1)}] = D_7[X_{(7)}] = 7/596$, для центрального — $16/596$. Желая увеличить точность оценивания, следует увеличить выборку n .

Как мы видели, ранжирование образцов x_i позволило существенно улучшить точность оценивания присущих им значений, не пользуясь эталоном. Откуда взялась эта дополнительная информация? Не «вечный ли двигатель» перед нами? Вспомним, что ранжируя выборку, упорядочивая ее, мы понижали ее энтропию, «закачивали» в нее информацию. Каждый акт парного сравнения порождал 1 бит информации.

Понижение энтропии упорядоченной выборки по сравнению с неупорядоченной проявилось в том, что ее элементы оказались коррелированными — это было установлено в разделе 2. Далее выяснилось, что ранги — номера элементов в упорядоченной выборке — несут информацию о значениях, причем степень коррелированности элементов выборки, и количество информации о значениях, заключенной в рангах, увеличивалось с ростом n . Таким образом, описанная здесь процедура «безэталонного» из-

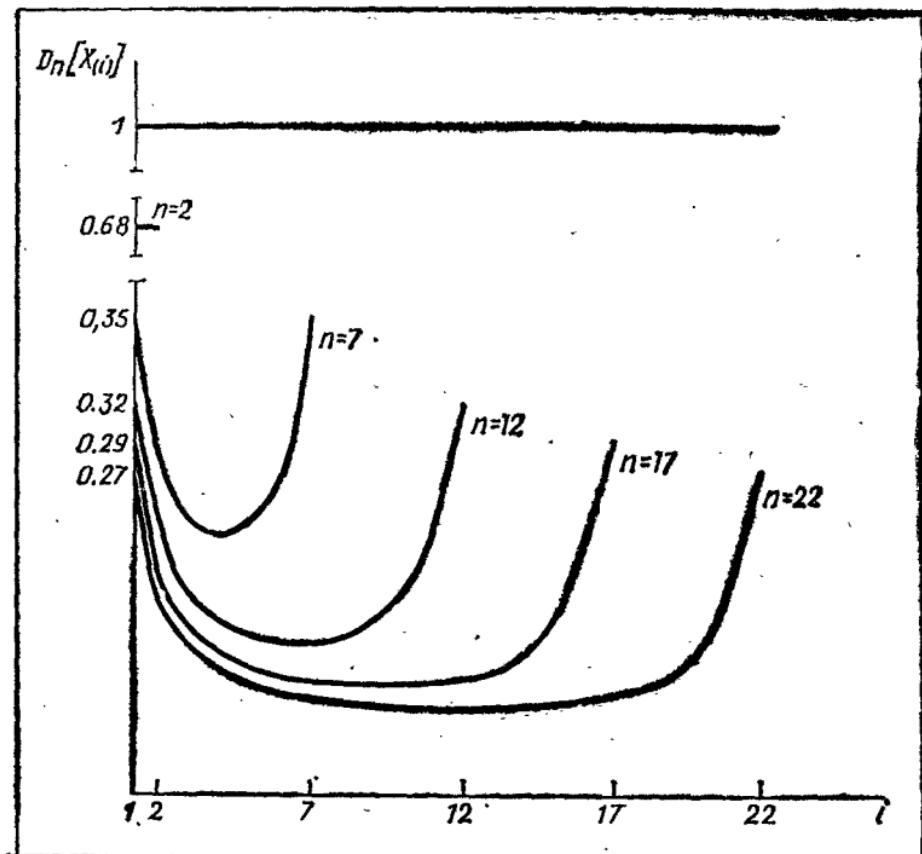


Рис. 8

мерения просто использует ту информацию, которая содержится в ранжированной выборке.

Да и название «безэталонное» не совсем точно. Эталон, оказывается, содержится в законе распределения случайной величины X в виде шкалы — последовательности средних порядковых статистик. Для того чтобы воспользоваться этим эталоном, нужно располагать выборкой $\{x_i\}$ и компаратором для ее упорядочения.

Идентификация объекта с ненаблюдаемым входом

Любимое детище кибернетики — «черный ящик» породил вокруг себя целую литературу и вызвал к жизни научную дисциплину, именуемую «идентификацией». Проблема «черного ящика» в первозданном виде заключалась в том,

чтобы, подавая на вход некоторого объекта пробные сигналы и наблюдая его реакцию — сигналы на выходе, определить структуру и параметры объекта, узнать, что в «ящике».

Десятилетия изучения проблемы [11] позволили охватить широкий класс идентифицируемых объектов, свести к минимуму необходимую информацию, но оставили в неприкосновенности главный принцип — принцип наблюдаемости входа и выхода.

Между тем рассмотренная выше оценка выборочных значений (18) позволяет поставить новую задачу идентификации: определение неизвестной статической (безынерционной) характеристики звена с ненаблюдаемым входом [10].

Пусть значения входной величины X с известным законом распределения $F(x)$ не могут быть измерены непосредственно (рис. 9). Выходная величина $Y = g(X)$ наблюдаема, но зависимость $g(x)$ неизвестна. Известно лишь, что она монотонна.

Измерив n значений y_i и ранжировав выборку $\{y_i\}$, мы получаем возможность ранжировать выборку $\{x_i\}$ и оценить выборочные значения, как это было проделано выше,

$$x_{(i)} = E_n [X_{(i)}].$$

Нанеся (рис. 9) значения $y_{(i)}$ и оценки $\hat{x}_{(i)}$ на оси y и x ,

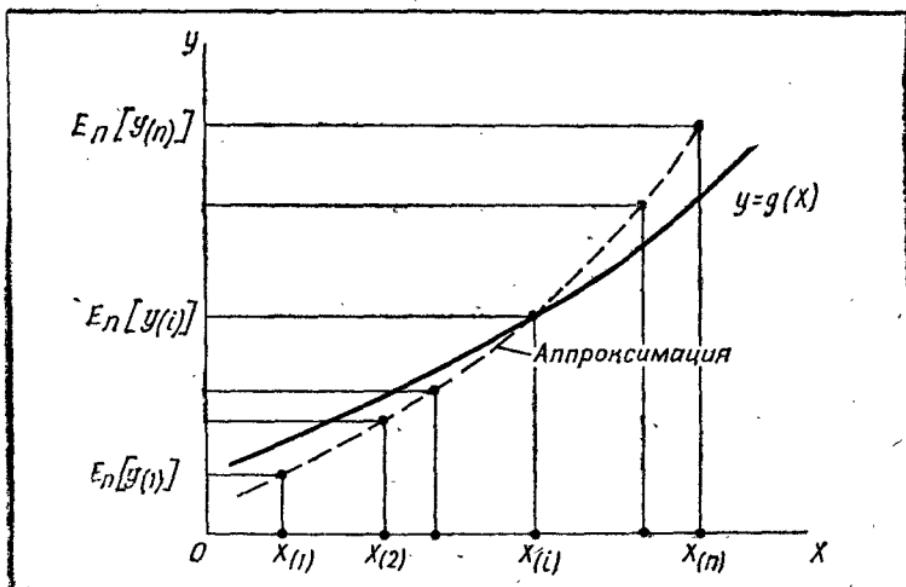


Рис. 9

мы сможем тем, или иным способом установить (идентифицировать) зависимость $y = g(x)$. Проделаем это с помощью метода наименьших квадратов, несколько конкретизировав задачу.

Предположим, что зависимость $y = g(x)$ параметризована, линейна по параметрам и имеет вид:

$$y = \theta_1 g(x) + \theta_2,$$

где $g(x)$ — известная функция, а θ_1 и θ_2 — параметры, которые следует найти. Тогда, имея набор упорядоченных наблюдений $y_{(1)}, y_{(2)}, \dots, y_{(n)}$, получаем $y_{(i)} = \theta_1 g(x_{(i)}) + \theta_2$.

В этом случае математическое ожидание фактических наблюдений $y_{(i)}$ представляет собой линейные функции искомых параметров θ_1 и θ_2 .

$$E_n[Y_{(i)}] = \theta_1 E[g(X_{(i)})] + \theta_2 = \theta_1 \alpha + \theta_2.$$

Ковариации наблюдений известны с точностью до постоянного множителя θ_1^2

$$V_n[Y_{(i)} Y_{(j)}] = \theta_1^2 V_n[g(X_{(i)}) g(X_{(j)})] = \theta_1^2 V_{ij}.$$

Нам предстоит найти оценки $\hat{\theta}_1$ и $\hat{\theta}_2$ неизвестных параметров θ_1 и θ_2 .

Напомним, что метод наименьших квадратов требует иекоррелированности наблюдений. Это требование, как мы видим, не выполняется — упорядочивание превращает элементы выборки в коррелированные случайные величины. Для преодоления трудности такого рода разработан «обобщенный метод», позволивший обрабатывать коррелированные последовательности.

Выражения для оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ определяются из системы нормальных уравнений обобщенным методом наименьших квадратов [7].

Не приводя вывода, дадим окончательные выражения для оценок $\hat{\theta}_1$ и $\hat{\theta}_2$:

$$\hat{\theta}_1 = \frac{n+1}{n-1} [y_{(n)} - y_{(1)}],$$

$$\hat{\theta}_2 = \frac{ny_{(n)} - y_{(1)}}{n-1}.$$

Оценки определяются лишь крайними элементами выборки. Операция ранжирования в этом случае не нужна, достаточно выбрать наибольший и наименьший элементы.

Любопытно, что в соответствии с (17) упорядоченная выборка оказалась «отредактирована»: ее крайние элементы имеют вес 1, все остальные 0.

Дисперсии и ковариация оценок:

$$D [\hat{\theta}_1] = \frac{2\theta_1^2}{(n+1)(n+2)},$$

$$D [\hat{\theta}_2] = \frac{n\theta_1^2}{(n^2-1)(n+2)},$$

$$V [\hat{\theta}_1 \hat{\theta}_2] = -\frac{\theta_1^2}{(n-1)(n+2)}.$$

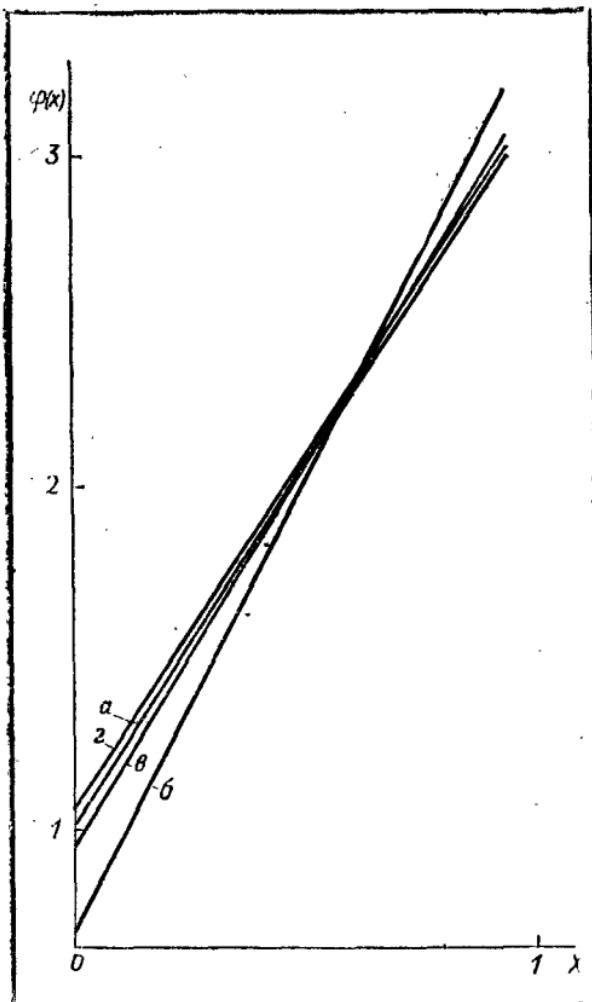


Рис. 10

В таблице приведены значения оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ параметров модели $y = \theta_1 x + \theta_2$, полученных при помощи статистического моделирования для выборок объемов $n = 20; 50; 100$ из равномерной совокупности. Истинные значения коэффициентов $\theta_1 = 2$ и $\theta_2 = 1$. Полученные результаты иллюстрирует рис. 10.

Объем выборок	Истинные значения коэффициентов		Наибольший и наименьший элементы выборки		Оценки коэффициентов		Дисперсии коэффициентов	
	n	θ_1	θ_2	$y_{(1)}$	$y_{(n)}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$D[\hat{\theta}_1]$
20	2	1	1.0224	2.9547	2.136	0.921	0.0173	0.0182
50	2	1	1.0664	2.9922	2.004	1.027	0.003	0.0036
100	2	1	1.005	2.9468	1.998	0.981	0.007	0.0003

Безэталонная классификация

Все знают, как сортируют по крупности помидоры или лимоны: их высыпают на плоскость с отверстиями определенного диаметра, в которые мелкие проваливаются, а более крупные остаются. Провалившиеся мелкие вновь попадают на плоскость с отверстиями еще меньшего диаметра, где процедура повторяется. Так происходит классификация, причем присутствует система эталонов — отверстий с заданными диаметрами. Желая классифицировать любые другие объекты, например камешки из кучи щебня, на мелкие, средние и крупные, мы вновь должны будем обратиться к эталонам. Нам придется конкретизировать понятие «мелкие», сопоставив ему определенное значение наибольшего веса, скажем 0,5 кг, «средние» — до 0,8 кг и «крупные» — свыше 0,8 кг. Далее, нам нужно будет запастись материальными воплощениями этих значений в эталонах — гирях соответствующего веса и, естественно, компаратором — рычажными весами. Процедура классификации будет заключаться в последовательном сравнении образцов со старшим эталоном и отбором «крупных», затем оставшихся с младшим и отбором «средних». В результате останутся «мелкие».

Источником ошибки здесь будет неверное чтение показаний компаратора либо сбои в работе самого компаратора.

А можно ли разделить кучу щебня на мелкие, средние и крупные камешки не имея гирь-эталонов, а лишь сравнивая камешки между собой? Мы уже знаем, что можно, но для этого понадобится знать закон распределения параметра, по которому производится классификация — в нашем случае веса.

Воспользуемся тем, что, хотя отсутствие эталонов не дает возможности получить реализацию вектора X — выборку значений, наличие компаратора позволяет получить реализацию их вектора рангов R . Потребуется, таким образом, разбить на «мелкие», «средние» и «крупные» ранги объектов — номера объектов в ранжированной выборке [12].

Обратимся к рис. 11, где изображена функция распределения $F(x)$, а на оси x нанесены, отмеченные римскими цифрами, значения эталонов. Количество эталонов $k = 4$. На этой же оси расположены значения параметра X в выборке объема $n = 10$, однако нанести их на ось невозможно, так как они неизвестны. Точки, соответствующие значениям x_i , поэтому нанесены условно, а арабские цифры 1, 2, ..., 10 означают их ранги.

Если соединить k эталонов с выборкой из n объектов, т. е. в нашем случае смешать 10 камешков и 4 гири в одну выборку объемом $N = n + k = 14$ и разложить ее, то гири получили бы ранги тем большие, чем больше их

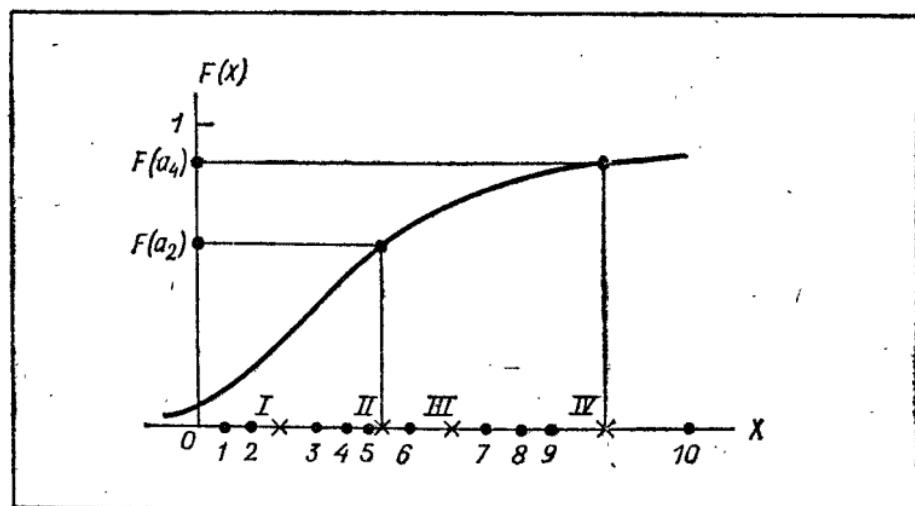


Рис. 11

вес. Если с этими же гилями смешивать все новые и новые десятки камешков, можно заметить, что ранги гилья варьируются от выборки к выборке тем меньше, чем больше n .

Теперь принцип классификации без эталона ясен: нужно, зная n , k и $F(x)$, определить ранги эталонов r в выборке объемом N , а затем разбить пространство рангов на $k + 1$ область, сравнивая ранги объектов с рангами эталонов. При этом произойдет и классификация объектов по значениям параметра X (по-прежнему неизвестным нам), поскольку, как было показано выше, ранги и соответствующие им выборочные значения связаны тем более тесно, чем больше объем выборки.

Источником ошибок классификации является конечность выборки, не позволяющая точно определять ранги значений эталонов. Задавшись, однако, значениями ошибок первого и второго рода, можно определить необходимый для получения требуемой точности объем выборки.

Итак, закон распределения случайной величины заключает в себе шкалу (эталон), а компаратор, воплощенный в той или иной форме, позволяет эту шкалу построить и использовать для оценивания, идентификации и классификации.

Все это очень интересно, скажет искушенный читатель, но где же взять закон распределения параметра, который неизмерим? Ведь закон распределения строится именно путем обработки большого числа измерений случайной величины. В силу такой точки зрения закон распределения неизмеримого параметра считается неизвестным, а вернее, просто не вызывает интереса у исследователей. Принято считать, что если измерение возможно, он не нужен, а если невозможно — бесполезен. Однако физики знают множество примеров, когда распределение величины известно из теории, а измерение ее выборочного значения невозможно или затруднительно. Так, скорость молекул газа в сосуде подчинена закону распределения Максвелла с известными параметрами, если известны давление и температура.

Точно также без наблюдения за отдельными молекулами можно указать закон распределения их удаления от какой-либо исходной точки при диффузии. Кроме того, существуют величины, неизмеримость которых относительна. Сейчас мы фиксируем скорость бегуна секундомером, а у древних греков не было эталона подходящего масштаба для измерения времени бега на спортивных соревнованиях,

поэтому на олимпиадах древности могла быть зафиксирована лишь последовательность бегунов на финише — их ранги. Тем не менее каждый бегун обладал определенным значением неизмеримого тогда параметра — скорости. Сейчас эта «величина в себе» объективизировалась так, что известны и ее закон распределения, и мгновенные значения.

Таким образом, безэталонные процедуры могут, с одной стороны, найти область приложения, где законы распределения уже известны, а с другой — стимулировать процесс выяснения этих законов.

ЛИТЕРАТУРА

1. Ко крен У. Методы выборочного исследования. М., «Статистика», 1976.
2. Шрейдер Ю. А. Равенство, сходство, порядок. М., «Наука», 1971.
3. Берзтисс А. Т. Структуры данных. М., «Статистика», 1974.
4. Гильбо Е. П., Челпанов И. Б. Обработка сигналов на основе упорядоченного выбора. М., «Советское радио», 1975.
5. Кендалл М. Дж., Стьюарт А. Статистические выводы и связи. М., «Наука», 1973.
6. Гумбель Э. Статистика экстремальных значений. М., «Мир», 1965.
7. Введение в теорию порядковых статистик. Сборник. М., «Статистика», 1970.
8. Тарасенко Ф. П. Непараметрическая статистика. Томск, изд-во Томского университета, 1976.
9. Тюрии Ю. Н. Непараметрические методы статистики. М., «Знание», 1978.
10. Ефимов А. Н., Кутеев В. М. Безэталонные измерения и идентификация методами теории порядковых статистик. «Автоматика и телемеханика», 1978 № 12.
11. Эйкхофф П. Основы идентификации систем управления. М., «Мир», 1975.
12. Ефимов А. Н., Кутеев В. М. Ранговые процедуры измерения и классификации без эталона. Transactions of the eight Prague conference of information theory. Academia, Prague, 1978.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1. ОБРАЗОВАНИЕ И УПОРЯДОЧЕНИЕ ВЫБОРКИ	6
Образование выборки	6
Что о порядке	9
Как упорядочить выборку?	13
Ранжированная выборка — объект с новыми свойствами	16
2. ВЕРОЯТНОСТНЫЕ СВОЙСТВА ПОРЯДКОВОЙ СТАТИСТИКИ	19
Закон распределения порядковой статистики	20
Числовые характеристики порядковой статистики	23
Распределения при неограниченном увеличении выборки. Распределения центральных значений	25
Распределения крайних значений	26
Совместные распределения порядковых статистик	31
3. ВЫБОРОЧНЫЕ ЗНАЧЕНИЯ И РАНГИ	34
Связь между значениями и их рангами	34
Ранговая корреляция	36
4. УЛУЧШЕНИЕ ОЦЕНOK ПУТЕМ ЦЕНЗУРИРОВАНИЯ ВЫБОРОК	39
Цензурирование выборок — общая идея	40
Оценки параметров нормального распределения по усеченным выборкам	42
Оценки параметров равномерного распределения	44
Выборка из трех наблюдений	46
5. БЕЗЭТАЛОННЫЕ ПРОЦЕДУРЫ ИЗМЕРЕНИЯ, ИДЕНТИФИКАЦИИ И КЛАССИФИКАЦИИ	47
Измерение без эталона	48
Идентификация объекта с ненаблюдаемым входом	54
Безэталонная классификация	58
Литература	62

Алексей Николаевич ЕФИМОВ ПОРЯДКОВЫЕ СТАТИСТИКИ — ИХ СВОЙСТВА И ПРИЛОЖЕНИЯ

Главный отраслевой редактор В. П. Демьянин
Редактор Г. Г. Карповский. Мл. редактор Т. Г. Иншакова. Обложка художника Л. П. Ромасенко.
Худож. редактор М. А. Бабичева. Техн. редактор
А. М. Красавина. Корректор В. И. Ширяева

ИБ № 2802

T-01238. Индекс заказа 804302.
Сдано в набор 12.12.79 г. Подписано к печати 21.01.80 г.
Формат бумаги 84×108^{1/32}. Бумага типографская № 3. Бум. л. 1.
Печ. л. 2. Усл. печ. л. 3,36 Уч.-изд. л. 3,19. Тираж 35 350 экз.
Зак. 2911

Издательство «Знания», 101835, ГСП,
Москва, Центр, проезд Серова, д. 4.

Цена 11 коп. Заказ 2911.

Чеховский полиграфический комбинат Союзполиграфпрома
Государственного комитета СССР по делам издательства,
полиграфии и книжной торговли,
г. Чехов Московской области

УВАЖАЕМЫЕ ТОВАРИЩИ!

Издательство «Знание» предлагает вам серию
подписных научно-популярных брошюр «Химия»

Брошюры этой серии пропагандируют химические знания среди широкого круга специалистов и всех, кто интересуется химией. Они рассказывают о самых актуальных проблемах химической науки и химического производства, дают читателю самую новую научную информацию в популярном изложении.

Авторы брошюр — виднейшие ученые нашей страны. У нас выступали академики: С. И. Вольфович, В. В. Кафаров, И. Л. Кнусянц, Б. П. Никольский, Ф. Д. Овчаренко, И. В. Петрянов, А. С. Садыков, Н. Н. Семенов, Г. Н. Флеров, А. В. Фокин, К. Б. Яцмирский и другие, члены-корреспонденты АН СССР: И. Н. Азербаев, И. П. Белецкая, И. В. Березин, П. П. Будников, В. И. Гольданский, Б. В. Дерягин, и другие ученые. Министр химической промышленности Л. А. Констандов — тоже наш автор.

Вот те работы, которые получат подписчики серии «Химия» во второй половине 1980 г.:

ОЗОННЫЙ ЩИТ ЗЕМЛИ

СОВРЕМЕННЫЕ ПРОБЛЕМЫ АНАЛИТИЧЕСКОЙ ХИМИИ

ИНГИБИТОРЫ КОРРОЗИИ

МЕТОДЫ ИНТЕНСИФИКАЦИИ ТЕХНОЛОГИЧЕСКИХ
ПРОЦЕССОВ

В 1981 г. подписчики получат 12 брошюр, в том числе:

Г. К. Боресков, академик

ОСНОВНЫЕ НАПРАВЛЕНИЯ РАЗВИТИЯ НАУКИ

О КАТАЛИЗЕ

В. В. Кафаров, академик

ИНФОРМАЦИОННЫЙ АНАЛИЗ В ХИМИИ И ХИМИЧЕСКОЙ
ТЕХНОЛОГИИ

М. Г. Дмитриев, доктор химических наук

АТМОСФЕРА — ОСНОВНОЙ КОМПОНЕНТ ОКРУЖАЮЩЕЙ

СРЕДЫ

Б. К. Соколов, кандидат технических наук

ГАЗЫ ОСОБОЙ ЧИСТОТЫ

А. Я. Кипнес, кандидат химических наук

КЛАСТЕРЫ В ХИМИИ

Серия «Химия» в каталоге «Союзпечати» расположена в разделе «Научно-популярные журналы» под рубрикой «Брошюры издательства «Знание». Индекс серии 70074.

ВЫПИСЫВАЙТЕ, ЧИТАЙТЕ СЕРИЮ НАУЧНО-ПОПУЛЯРНЫХ
БРОШЮР «ХИМИЯ»! В розничную продажу брошюры нашей серии
не поступают.

Подписная цена на год 1 руб. 32 коп.

Подписка принимается во всех отделениях связи, агентствах
«Союзпечати» и общественными распространителями печати по
месту работы и учебы. Без ограничения, с любого месяца.

ИЗДАТЕЛЬСТВО «ЗНАНИЕ»