Г. А. ФЕДОРОВ-ДАВЫДОВ

СТАТИСТИЧЕСКИЕ МЕТОДЫ В АРХЕОЛОГИИ

Допущено
Министерством высшего и среднего
специального образования СССР
в качестве учебного пособия
для студентов высших учебных заведений,
обучающихся по специальности «История»

Москва
«Высшая школа»
1987

ББК 63.4
Ф 33

Рецензенты:

кафедра истории стран Азии, Африки и Латинской Америки историко-филологического факультета Горьковского государственного университета им. Н. И. Лобачевского (и. о. зав. кафедрой доцент *Т. В. Гусева*); *В. Б. Ковалевская,* ст. научный сотрудник Института археологии АН СССР

Федоров-Давыдов Г. А.

Ф 33 Статистические методы в археологии: Учеб. пособие для вузов по спец. «История». — М.: Высш. шк., 1987. — 216 с.

В пособии излагаются основные понятия теории вероятностей и математической статистики применительно к археологии. Показаны методы работы с массовыми количественными и качественными признаками на археологических объектах, способы статистического выявления взаимосвязей между признаками, объектами и группами объектов и их археологическое истолкование.

Ф $\frac{0507000000-104}{001(01)-87}$ 78—87

ББК 63.4
902.6

ПРЕДИСЛОВИЕ РЕДАКТОРА

Статистика — слово непростое. Оно относится и к способам регистрации таких массовых явлений, как рождаемость или миграция, и к методам измерения показателей роста народного хозяйства в текущем году, и к массивам информации, собранной в результате экспериментов, наблюдений, раскопок, опросов, и к методам сбора, обработки и анализа такого рода данных, и даже к получаемым результатам. Наконец, под статистикой понимается наука о способах получения, упорядочения и обработки массовых данных, методов формирования и проверки гипотез на их основе, о правилах введения статистических фактов в конкретную теоретическую дисциплину.

До недавнего времени в статистической теории господствовали центробежные тенденции, приведшие, в частности, к формированию статистических дисциплин, получивших в вузовских курсах наименования: «Общая теория статистики» и «Математическая статистика». Первая из них специализируется в основном на методологическом анализе основных статистических понятий и условий их применимости в социально-экономических исследованиях, что подчас рассматривается как основание для отнесения общей теории статистики к разряду общественных наук. Вторая, напротив, объявляется разделом математики, связанным с задачами оценки тех или иных характеристик вероятностных распределений на основе моделей случайных наблюдений.

Различия в этих двух подходах продемонстрируем на примере понятий среднего значения и группировки. Согласно общей теории статистики среднее — это обобщающая характеристика общественных явлений по одному количественному признаку, и основные проблемы связаны с условиями и принципами использования средних в конкретных условиях. Согласно математической статистике среднее — это выборочная оценка одного из важнейших показателей вероятностного распределения — «математического ожидания», и основные проблемы связаны с исследованием точности этой оценки при тех или иных моделях формирования данных. Аналогично, по общей теории статистики, группировка — это разделение совокупности явлений на однородные группы, выполняемое исследователем на основе теоретических представлений об изучаемом процессе; данные используются лишь для анализа рассматриваемых статистиком групп. По

определению математической статистики группировка — это операция оценки параметров вероятностных распределений, соответствующих отдельным группам, проводимая исключительно по данным, рассматриваемым как случайная выборка из «смеси» этих распределений.

В обоих случаях подходы общей теории и математической статистики взаимоисключающие. Математик обвинит теоретика статистики в ненаучности и субъективности, а теоретик, как дважды два, докажет, что рафинированные математические модели не применимы практически ни в одной реальной ситуации.

Последние годы ознаменованы появлением противоположных, центростремительных тенденций. На базе широкого использования электронно-вычислительной техники развивается и крепнет интегрирующая концепция прикладной статистики, называемая также «теорией анализа данных». В центре внимания этой концепции — потребности специалиста в той или иной конкретной области (в данном случае — археолога), который хотел бы использовать имеющиеся у него конкретные данные в своей предметной сфере. Такого специалиста интересуют способы упорядочения и формализации тех видов данных, с которыми он сталкивается в своей работе, методы обработки этих данных и интерпретации получаемых результатов. Все эти средства и должна предоставить ему прикладная статистика, используя существующие и при необходимости создавая новые инструменты и методы.

Подобно общей теории статистики, теория анализа данных во главу угла ставит специалиста-предметника, но дает ему продвинутые математические средства (подобно математической статистике), используя при этом «инженерную» точку зрения на математические методы как инструменты познания, а не слепки реальности.

Книга Г. А. Федорова-Давыдова принадлежит именно этому третьему направлению. В ней использован и творчески преломлен богатый опыт работы с массивами данных, получаемых в процессе археологических раскопок. Пособие рассчитано на читателя гуманитарного профиля. Автору пришлось немало потрудиться для того, чтобы в краткой, популярной, но в то же время достаточно строгой форме преподнести подчас сложные математические конструкции. Это ему удалось. Книга является уникальной по уровню и широте стыковки статистического аппарата с археологической проблематикой.

Б. Г. Миркин

ВВЕДЕНИЕ

Применение методов математической статистики в археологии обусловлено резким количественным ростом археологического материала. Археология относится к числу тех наук, в которых закономерности проявляются как тенденция, прокладывающая себе путь среди множества отклонений и случайных сдвигов. Методы математической статистики позволяют выявить тенденции, которые в многочисленном материале обычными способами не выявляются, т. е. получить «скрытую» информацию.

Там, где традиционные методы археологии дают возможность оценивать какое-либо явление в приблизительных понятиях «больше — меньше», «сходно», «малосходно», «несходно» и т. п., статистические методы позволяют более точно выразить степень интенсивности, силы этого явления, дают количественное выражение какого-либо качества. При этом окончательный вывод, интерпретация полученных результатов всегда остаются делом собственно археолога, анализирующего всю совокупность явлений. Исторический вывод не может быть получен автоматически, с помощью какого-либо алгоритма, правила. «...В статистических исследованиях использование математических методов... способствует объективизации, но не вследствие однозначного выбора, а с помощью лучшего представления материала и его особенностей» [1].

Применение математико-статистических методов в археологии будет успешным, если правильно поставлена задача и для решения ее подобраны соответствующие математические методы. Эти методы должны быть адекватны сути явлений, которую исследователь стремится раскрыть [2].

[1] *Миркин Б. Г.* Группировки в социально-экономических исследованиях. М., 1985. С. 172.

[2] Основные принципы применения математических методов при изучении письменных источников изложены И. Д. Ковальченко (см.: Количественные методы в исторических исследованиях. М., 1984. С. 62 и далее).

Применение математико-статистических методов в археологии целесообразно, а иногда и необходимо при решении следующих задач:

1. Установление средних размеров археологических объектов как стандартов, которым следовали древние мастера. Установление степени «вариабельности» этих размеров, т. е. их колебаний вокруг средних, определение случайных и закономерных отклонений от этих стандартов. Подобные задачи возникают при углубленном исследовании объектов и при сравнении их между собой. Этим сюжетам посвящена глава I.

2. Исследование частоты встречаемости и взаимовстречаемости объектов и признаков, тех или иных явлений, оценка степени взаимосвязи между ними. Эти задачи возможны при выявлении типических видов вещей, в которых признаки закономерно и тесно связаны между собой, при построении хронологических систем, когда нужно установить, какие виды вещей закономерно встречаются вместе, при выявлении локальных вариантов или культур, когда требуется установить, что виды вещей и объекты закономерно встречаются на одной территории. Всему этому посвящены главы II, III и V.

3. Исследование сходства между объектами, культурами или их локальными вариантами, полностью не совпадающими друг с другом, но имеющими набор совпадающих и несовпадающих признаков. Эти задачи встают перед исследователем при выделении культур и их локальных вариантов, при построении эволюционных рядов, при разбиении множества каких-либо объектов на группы, т. е. при классификации. Об этом главы IV и V.

4. Исследование структуры взаимосвязей признаков археологических объектов с целью выявления более или менее существенных признаков необходимо при построении археологической классификации и при более углубленном исследовании сходства объектов между собой. Об этом написано в главах IV, V и VI.

Предлагаемое пособие ставит своей целью показать применение самых простых статистических методов в археологии. Оно рассчитано на читателя, владеющего математическими знаниями в пределах курса средней школы.

Не следует рассматривать математическую статистику как средство построения некоей «математической археологии». Археология остается специфической наукой, а математическая статистика только инструмент в ее руках для выявления скрытых тенденций и оценки степени достоверности тех или иных выводов. Статистические методы в археологии, так же как и в истории вообще, являются лишь основой для содержательного исторического анализа. К этим методам по мере накопления фактов и наблюдений археология неизбежно приходит, так же как пришли к ним биология, медицина, психология и другие дисциплины. Уместно вспомнить слова К. Маркса: «Наука только тогда достигает совершенства, когда ей удается пользоваться математикой» [3].

В заключение хочу поблагодарить за весьма существенную помощь в работе над пособием Л. И. Бородкина и Б. Г. Миркина.

1. Описание археологических объектов. Признаки

То, что сделано древним человеком (вещи, погребальные сооружения, жилища, архитектурные сооружения) и что сохранилось до нашего времени в виде материальных остатков, составляет археологические объекты (в теоретической археологической литературе их называют часто артефактами). Каждая категория объектов должна быть единообразно и правильно описана. В настоящее время выделяется специальный раздел археологии — дескриптивная, т. е. описательная, археология, которая дает рекомендации для наиболее удобного формализованного и единообразного описания археологических объектов с помощью *признаков* [4]. Признак — это какая-либо сторона, деталь, часть объекта. Признаков у каждого объекта бесконечное множество. Исследователь сначала интуитивно, а затем руководствуясь опытом изучения своего материала, ограничивается определенным набором наиболее важных признаков. Само определение важности, существенности того или иного признака в рамках заданного контекста — задача в значительной мере статистическая.

Существует глубоко разработанная теория измерений, определяющая виды признаков и шкал. Археология, как и всякая наука, имеет свою специфику. С учетом ее мы рассматриваем следующие виды признаков.

---

[3] Воспоминания о Марксе и Энгельсе. М., 1956. С. 66.

[4] См. об этом: *Каменецкий И. С., Маршак Б. И., Шер Я. А.* Анализ археологических источников (Возможности формализованного подхода). М., 1975.

Каждый признак имеет набор *значений* (вариантов), так что на каждом объекте рассматриваемой совокупности реализуется то или иное значение. Признак называется *количественным* в том случае, когда его значение может быть измерено или исчислено. Количественные признаки могут быть *мерными, непрерывными* (например, ширина могилы, вес монеты, содержание олова в бронзе и т. д.) или *счетными, дискретными,* содержащими какое-то количество единиц чего-либо, составляющего признак (например, число бус в погребении). Непрерывными признаками являются такие количественные признаки, у которых значения в принципе могут отличаться друг от друга на сколь угодно малую величину.

Отдельные значения дискретных признаков отличаются на какую-то конечную величину.

Всякий непрерывный признак может быть измерен с ограниченной степенью точности и тем самым множество его значений разбивается на *интервалы.* Исследователь выбирает размер минимального интервала, шкалы, исходя из целей своей работы и допущений и соглашений, установившихся в данной отрасли науки.

Признак является *качественным* в тех случаях, когда он не может быть измерен, сосчитан, выражен числом или какой-либо мерой, когда он выражает какое-либо качество, свойство или состояние предметов.

При измерении количественного признака определяется степень интенсивности (мера) какого-то свойства на объекте, при определении качественного признака констатируется только словесное выражение конкретного варианта на объекте.

Пример 1.

Дано определенное количество светлоглиняных амфор, полученных в результате раскопок позднеэллинистических и римского времени слоев археологических памятников Причерноморья. Требуется описать их формализованно каким-то набором количественных и качественных признаков. Количественные признаки были приняты такие:

признак: переход боковой поверхности донной части в нижнюю горизонтальную поверхность;

его значения:

1 — четко выраженный;

2 — заглаженный, плавный;

3 — низ боковой поверхности срезан под углом;

4 — низ боковой поверхности небрежно замят внутрь;

признак: поддон;

его значения:

1 — не выражен, плавный переход ко дну;

2 — выражен;

3 — с вмятиной на дне, образующей подобие поддона;

4 — вместо поддона цилиндрический выступ [5].

Пример 2.

При изучении античных бус в числе других определен такой качественный признак:

признак: материал;

его значения:

1 — стекло: 1.1 — стекло, 1.2 — египетский фаянс;

2 — металл: 2.1 — золото, 2.2 — серебро, 2.3 — бронза, 2.4 — свинец;

3 — смолистые вещества: 3.1 — агар, 3.2 — янтарь;

4 — драгоценные и полудрагоценные камни: 4.1 — горный хрусталь, 4.2 — аметист, 4.3 — халцедон, 4.4 — сердолик, 4.5 — сардер, 4.6 — хризопраз, 4.7 — агат, 4.8 — гранит, 4.9 — яшма;

5 — прочие материалы: 5.1 — кремнистая порода, 5.2 — мрамор, 5.3 — меловая порода, 5.4 — коралл, 5.5 — раковина, 5.6 — кость, 5.7 — глина;

6 — комбинированные материалы: 6.1 — стекло с внутренней позолотой, 6.2 — стекло с внутренним посеребрением, 6.3 — глина с позолотой, 6.4 — золото с серебром [6].

Таким образом, признаку «материал» можно придать шесть «укрупненных» градаций (значений). Но если требуется более детальная и углубленная характеристика объектов по этому признаку, каждой из этих градаций можно придать более мелкие значения.

Качественным признаком может быть и категория археологических вещей или объектов, тогда значениями его будут разновидности (виды) этих вещей или объектов. Например, признак: стремена; его значения: стремена типа 1 — AI, 2 — AII, 3 — БI и т. д. [7]

Качественные признаки могут быть *простыми* и *со-*

---

[5] См.: *Деопик Д. В., Карапетянц А. М.* Некоторые принципы описания применительно к возможностям статистического анализа // СКМА. М., 1970.

[6] См.: *Алексеева Е. М.* Классификация античных бус // СКМА. М., 1970.

[7] См.: *Федоров-Давыдов Г. А.* Кочевники Восточной Европы под властью золотоордынских ханов. М., 1966.

ставными. Признак «материал» может рассматриваться как простой; признак «стремена» как составной, в котором интегрировано несколько признаков. Когда мы выделяем, как нам кажется, простые признаки, они состоят из сочетаний еще более простых признаков, но объединение этих последних мы производим не открыто, формализованно, а скрыто, интуитивно. Например, признак «материал» даже при дробных, мелких своих градациях (значениях) может быть рассмотрен как составной, как химическое соединение элементов, как определенная структура молекул и т. п. Но для того уровня исследования, которое предпринимается, исследователь этим пренебрегает и принимает признак «материал» как простой. Таким образом, говоря, что признак «простой», мы определяем тем самым границу между формализованным и интуитивным исследованием объекта.

Признаками объекта могут быть и его географическое положение, и такие сложные признаки, устанавливаемые после исследования простых, как принадлежность к археологической культуре, дата объекта и т. п.

Установление признака предполагает однородность каждого его значения. Например, значение материала «золото» предполагает, что все бусы «одинаково золотые». Если исследователь считает, что такая характеристика не выявляет какой-то существенной стороны объекта, то он вводит иные значения признака: «высокопробные золотые», «низкопробные золотые», или иначе определяет эти значения.

Качественные признаки могут быть *ранжированными (ранговыми)* и *неранжированными (номинативными)*.

Большинство простых признаков в археологии неранжировано, т. е. их нельзя расположить в каком-либо порядке, который будет отражать существенную черту объекта. Конечно, признак «цвет» можно расположить в порядке солнечного спектра, но этот порядок будет чужд археологическому объекту, так как древние, придавая предмету тот или иной цвет, действовали из других побуждений и не думали при этом о порядке цветов в спектре. Это будет «некорректное» применение количественного метода.

Расположение типов жилища как значений признака «конструкция жилища» по степени их усложнения и комфортабельности или погребений по богатству их инвентаря представляется обоснованным, так как это соответствует археологической сути этого признака. Такой признак можно отнести к разряду качественных ранжированных.

Такие признаки, как «географическое положение», могут быть ранжированы по какому-либо направлению (с севера на юг или по мере удаленности от какого-либо центра). Ранжируемыми признаками обычно оказываются сложные вторичные составные признаки. Например, значения признака «принадлежность к той или иной археологической культуре» можно расположить в хронологическом порядке.

Если признак «принадлежность к археологической культуре» позволяет достаточно дробно и точно датировать объект, то этот признак превращается в количественный признак с интервалами, соответствующими длительности каждой культуры. При этом интервалы могут быть неравными.

Качественные признаки могут быть представлены двояким образом. 1) Мы называем признаком какую-либо черту, сторону объекта, которая может принимать различные значения (или варианты признака). Мы их будем называть просто качественными.

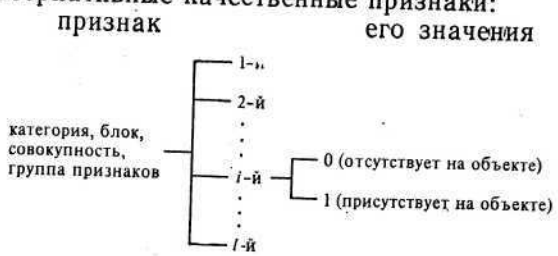
2) Каждое значение (вариант) признака можно представить как отдельный признак с двумя значениями: присутствует (1) или отсутствует (0) на объекте. Например, признак «материал бусины — янтарь». Его значения «да, присутствует» (1) — т. е. бусина из янтаря, или «нет, отсутствует» (0), т. е. бусина не из янтаря. Определенные таким образом признаки будем называть *альтернативными (дихотомическими)*.

В археологической литературе встречаются оба способа представления и записи признаков. Они равноправны. Терминология этих способов соотносится так:

Качественный признак:

его значения

Альтернативные качественные признаки:



При применении статистических методов исследования эти два способа определения признаков требуют, как мы увидим ниже, некоторого изменения приемов, хотя суть методики остается.

При определении качественных признаков и их значений следует соблюдать следующие требования.

1) Обязательность присутствия какого-либо значения на всех взятых для исследования объектах. Этому требованию удовлетворяет, например, признак «форма» для сосудов — все сосуды имеют форму. Однако признак «ручка» не всегда удовлетворяет этому требованию. Для таких признаков следует вводить 0-е значение признака, не смешивая с отсутствием сведений или неполной сохранностью. Альтернативные качественные признаки всегда удовлетворяют этому требованию, так как одно из двух возможных значений (0, 1) всегда может быть зарегистрировано на объекте.

2) Несовместность значений признаков. На одном объекте не должно быть двух или более значений одного признака. Признак «форма» для бус удовлетворяет этому требованию: у всех бус имеется одна какая-либо форма. Однако признак «цвет» не удовлетворяет этому требованию, если исследуются не только одноцветные, но и многоцветные бусы. Тогда следует заменить простой признак на более сложный, значения которого будут сочетаниями значений простого.

Например, признак «цвет» нужно заменить признаком «сочетания цветов». Его значения: 1 — красный монохромный, 2 — черный монохромный, ..., i-е — черно-красный, $i+1$-е — черно-белый, ..., j-е — черно-красно-белый и т. д. Или признак: набор стремян (максимальное число стремян, встречающихся в исследуемом массиве погребений,— 3);

его значения: 1 — 000, 2 — 00AI, 3 — 00AII, ..., i-е — AI, AI, AI, $(i+1)$-е — AI, AI, AII, ... и т. д., где AI, AII, ... — виды стремян. В некоторых случаях целесообразно заменить значения качественного признака на качественный альтернативный признак: стремена, AI; его значения: присутствует (1), и отсутствует (0) и т. д. Качественный альтернативный признак может быть представлен в объекте несколько раз: несколько стремян одного типа в погребении. Тогда он превращается в количественный счетный дискретный признак, значениями которого являются число единиц на объекте. Например, признак: «стремена типа AI»», его значения: 0, 1, 2, 3 стремени.

Или признак: «бусы типа I в погребении». Его значения: 0, 1, 2, 3, ... количество бус.

Пример 3.

Дано 7283 бусины из раскопок 59 погребений I Поломского могильника в Удмуртии (IX—X вв.). В разных погребениях разные виды бус встречались в разных количествах. Требуется определить признак или набор признаков, которыми можно было бы описать набор бус в каждом погребении этого могильника. Как это сделать правильно? Если взять признак «бусы», а значениями его — вид бус, то погребение будет описано так: бусы вида А, бусы вида В и т. д. Мы получим лишь представление о том, какие виды бус присутствуют в погребении.

Если мы каждый вид бус примем за альтернативный признак и определим его значения так: 1 — присутствует в погребении хотя бы одна бусина данного вида и 0 — отсутствует в погребении бусины этого вида, то совершим ту же ошибку. Встречаемости в погребении 1 бусины или 100 бус окажутся равнозначны[8]. Нарушится принцип однородности значения признака. Объект, имеющий 100 бус вида «А», попадает в то же значение качественного признака, что и объект, имеющий 1 бусину вида «А». Видимо, здесь следует ввести количественный признак, значения которого будут равны числу бус вида «А» в погребении или доли бус вида «А» среди всех бус в погребении. Наборы бус тогда будут описываться некоторым числом (равным числу видов бус) количественных счетных признаков, чьи значения равны числу бус каждого вида или доли бус каждого вида среди всех бус в погребении.

[8] См.: *Львова З. А.* Бусы I Поломского могильника // Вопросы археологии Удмуртии. Ижевск, 1976. Табл. I. (Подсчет так называемой «частоты встречаемости бус данного вида»).

Итак, исходные археологические данные могут быть представлены в виде признаков разных видов. Каждому виду соответствует своя шкала измерений. Шкала измерений — это способ придания каждому значению признака определенного числового значения.

В археологии типична ситуация, когда материал доходит до исследователя в виде обломков. Например, среди огромного числа фрагментов керамики обычно мало таких, которые дают полный профиль сосуда, так называемые «археологические целые» сосуды. Форма и орнамент сосуда обычно описываются набором каких-то признаков. Полный набор этих признаков можно получить только у сосудов, сохранивших полный профиль. Для остальных сосудов, сильно фрагментированных, этот набор признаков будет неполным. Можно ограничиться только «археологически целыми» сосудами. Но тогда останется за бортом большое количество информации, заключенной в фрагментированных сосудах. Можно недостающие данные дополнить предположительными, наиболее вероятными значениями утраченных признаков, рассчитанных на основе имеющейся информации о целых сосудах. Но при этом мы можем ошибиться, так как будем иметь дело с предположениями, допускающими ошибки с той или иной степенью вероятности. Возможна оптимизация этого подсчета, т. е. отбор таких фрагментированных сосудов, при котором потери информации и риск ошибок от гипотетического восстановления утраченных значений некоторых признаков будут минимальными.

Эта сложная статистическая процедура нами не рассматривается. Но мы предлагаем некоторые статистические приемы, учитывающие информацию, заключенную в фрагментированном археологическом материале (см., например, гл. III, § 4).

2. Выборочный метод и нулевая гипотеза

Одним из основных понятий математической статистики является понятие *«генеральная совокупность»*, представляющее собой модель источника данных. Применительно к археологии это понятие можно интерпретировать как совокупность объектов, объединенных какими-либо признаками или качествами, бывших в употреблении, обиходе, «бытовавших» в определенный период на определенной территории. Эта «генеральная совокупность» предметов или явлений недоступна для полного

изучения. Некоторое количество элементов этой совокупности превращается в археологический материал и образует *выборку* из генеральной совокупности. Выборку из этой выборки составляет тот археологический материал, который попал в поле зрения исследователя. Таким образом, археологи по выборке пытаются, используя методы и приемы математической статистики, судить о тех или иных свойствах генеральной совокупности.

Упрощенно говоря, генеральная совокупность — это то, из чего произведена выборка. Если выборка — могильник, то генеральная совокупность — все погребения того населения, которое хоронило в могильнике. Если выборка — сосуды из раскопанной части могильника, то генеральная совокупность — все сосуды могильника, но также и все сосуды населения, хоронившего в могильнике. Если выборка — клад монет, то генеральная совокупность — денежное обращение в момент зарытия клада, если выборка — ножи из определенного слоя данного раскопа в городе, то генеральная совокупность — все ножи в культурном слое города или все ножи, бывшие в употреблении в этом городе или в данном районе города в данный момент.

Для статистического исследования особенно важна и ценна *случайная выборка*. Если все элементы генеральной совокупности имеют одинаковые шансы попасть в выборку, то такую выборку мы называем случайной. Случайная выборка статистически отражает генеральную совокупность. Но при этом должны быть устранены все субъективные, неслучайные факторы образования выборки. Случайной выборкой из клада монет можно признать его часть, оставшуюся после хищения клада находчиками, в монетах не разбирающимися. Эта часть отражает состав клада. Не является случайной выборкой часть клада, отобранная коллекционерами. Она отражает не столько состав клада, сколько интерес коллекционера.

Главное, что отличает археологическую выборку от статистических выборок в других областях науки,— это невозможность самому исследователю организовать выборочный отбор. К нему попадают вещи из современных раскопок, из старых раскопок, от которых случайно сохранилась лишь часть коллекций, из случайных находок. Чтобы исследовать, скажем, сосуды какой-то культуры или ее локального варианта или одного памятника, археолог берет обычно не только сосуды из хорошо до-

кументированных находок, но и из старых, не полностью сохранившихся коллекций и сборов, или случайно добытые сосуды, по ряду признаков относящиеся к исследуемому комплексу. Образуется то, что принято называть *«естественной» выборкой.*

Даже если исследователь ограничивается только объектами из строго документированных раскопок и берет оттуда все объекты, для него остается большей частью скрытым сам механизм попадания вещей в культурный слой или могилу, механизм далеко не всегда случайный. Близким к случайной выборке можно считать археологический материал, извлеченный из раскопа в той части культурного слоя, которая не является жилищем, куда обломки вещей данной категории попадают случайно, например были потеряны, выкинуты. Раскоп, пришедшийся на жилище, в меньшей степени случаен, так как извлеченный из него материал отражает подбор вещей, произведенный хозяином этого жилища в соответствии с его профессией, уровнем жизни, вкусами, этнической принадлежностью.

На формирование «естественной» археологической выборки влияет столь большое количество различных и в большинстве случаев взаимно независимых причин, что с известной мерой осторожности можно считать ее простой случайной выборкой, в которой находят отражение некие усредненные, обобщенные характеристики генеральной совокупности. Осторожность же эта заключается в том, что выборка не должна рассматриваться как слишком большая. Методы статистики позволяют исследователю определять минимальную численность выборки для того, чтобы получить выводы с достаточно малой вероятностью ошибки. Но риск ошибки вычисляется для модели случайной выборки. Для «естественной» выборки он несколько выше. Поэтому исследователь-археолог не должен ограничиваться частью имеющейся у него «естественной» выборки, а должен исследовать ее всю, более того, он должен стремиться отыскать дополнительные материалы, чтобы увеличить ее объем.

Методы математической статистики предназначаются для случайной выборки. Потому, прилагая их к «естественной» выборке, нужно стремиться приблизить ее к случайной и по возможности уменьшить влияние неслучайных факторов ее формирования. Так, если известно, что при отборе вещей из каких-либо раскопок брались выборочно только некоторые категории сосудов (например, только кувшины с носиком), следует материалы этих раскопок исключить из подсчета видов керамики для всей культуры. Но допустим, от этих раскопок остался не полный керамический материал, а только случайно отобранная его часть или при раскопках брались лишь «археологически целые» сосуды, то такой материал может быть включен в статистическое исследование («археологические целые» сосуды можно считать выборкой более или менее случайной из всей керамики, извлеченной при раскопках из культурного слоя).

Раскопанная часть памятника может рассматриваться как выборка из генеральной совокупности, которая нам неизвестна, поскольку она остается под землей. Если эта совокупность однородна, ее элементы имеют одинаковые шансы попасть в ту часть могильника или поселения, которая подверглась раскопкам. Допустим, ножи разных видов за время жизни людей перемешались между собой и распространились равномерно на какой-то территории. Тогда «естественная» выборка — все ножи из раскопа — будет близка к случайной выборке из генеральной совокупности всех ножей на данной территории. Допустим теперь, что ножи не достаточно хорошо перемешались. Тогда раскопы в разных частях поселения будут случайными выборками из каких-то разных генеральных совокупностей, распространенных только на окрестности данного раскопа. Несколько раскопов на могильнике — это также несколько выборок из генеральной совокупности. Математико-статистические методы позволяют определить, однородна ли эта генеральная совокупность или же неоднородна, так что одни раскопы являются выборкой из одной генеральной совокупности объектов, другие — из другой.

По отношению к одним предметам раскопы могут быть выборками из однородной генеральной совокупности, по отношению к другим — выборками из разных генеральных совокупностей. Мы будем в дальнейшем считать всякую генеральную совокупность однородной. В тех случаях, когда устанавливается неоднородность генеральной совокупности, мы будем говорить о нескольких однородных генеральных совокупностях.

Совокупность ножей из одного раскопа на могильнике — случайная выборка из однородной генеральной совокупности ножей в окрестностях данного раскопа, но она не является случайной выборкой из генеральной совокупности всех ножей, попавших в землю в данном мо-

гильнике, если эта генеральная совокупность «плохо перемешана», т. е. неоднородна. Но степень однородности совокупности и случайности выборки мы не знаем заранее. Если устранены все субъективные, зависящие от исследователя факторы, нарушающие случайный характер выборки (преднамеренный отбор вещей, разные программы и способы описания и т. п.), то об однородности генеральной совокупности можно статистически судить по самим выборкам. Если несколько выборок из разных частей могильника дадут по отношению к ножам сходные статистические данные, то тогда с определенной долей уверенности можно говорить об однородности генеральной совокупности. Если установлено, что в погребения клали специально изготовленные для этого сосуды, то сосуды из могильников будут случайной выборкой из генеральной совокупности всех погребальных сосудов у данного населения в данный момент времени, но не будут случайной выборкой из всех сосудов вообще у данного населения в данный момент, так как эта последняя совокупность будет неоднородна.

Различают возвратную и безвозвратную выборки. В первом случае каждый взятый объект возвращается в генеральную совокупность и только потом производится случайный выбор другого объекта, так что генеральная совокупность не изменяется от выбора к выбору. При безвозвратной выборке отобранные единицы не возвращаются в генеральную совокупность и она численно изменяется. Величина генеральной совокупности в археологии по сравнению с выборками так велика, что мы можем считать все выборки возвратными, т. е. не меняющими состав генеральной совокупности.

3. Вероятность. Теоремы о сложении и умножении вероятностей

Вероятность — понятие сложное. Его трактуют и как меру возможности, и как меру нашей уверенности, и как меру сложности, и как частоту наступления события в повторяющихся экспериментах и т. п. Математическая трактовка этого понятия в настоящий момент использует сложные математические конструкции. Мы же рассмотрим это понятие в следующих терминах.

Если из N объектов случайно выбирается один объект так, что при этом не отдается предпочтения никакому из этих объектов, мы будем считать, что каждому

объекту обеспечена равная возможность быть выбранным. В этом случае вероятность отбора любого объекта равна $1/N$.

Какое-либо событие *«A»* может осуществляться или неосуществляться в тех условиях, в которых производится опыт (испытание). Такое событие мы будем называть *случайным событием.* Событие может быть *достоверным.* Достоверное событие обязательно должно произойти в данных условиях опыта (испытания). *Невозможное* событие — это событие, которое в данных условиях произойти не может.

Всякое событие *«A»,* связанное с данными объектами, соответствует подмножеству тех объектов, на которых оно реализуется. Например, событие «быть стеклянной» (значение признака «материал» на множестве бус) соответствует подмножеству стеклянных бус.

Выше для признаков мы ввели понятие несовместимости. По отношению к событиям содержание этого понятия определяется следующим образом.

События *«A», «B», «C»* будут тогда несовместными, когда в условиях данного опыта, испытания оказывается возможным осуществление только одного из этих событий. Другими словами, подмножества объектов, отвечающих событиям *«A, «B», «C»,* не пересекаются, поэтому в результате осуществления испытания эти события не могут появиться одновременно. Если же появление одного события не исключает возможности появления другого (существуют объекты, относящиеся к обоим событиям), то такие события мы будем называть совместными.

Если равновозможность и несовместность исходов опыта (испытания) соблюдены, то случайный выбор каждого объекта, т. е. каждый исход опыта, каждое событие — это один шанс из N. Если из N объектов A объектов обладают каким-либо признаком, то событие *«A»,* состоящее в случайном взятии одного объекта именно с этим признаком, имеет A шансов появиться как исход опыта.

Теперь мы можем определить математическую вероятность случайного события таким образом: математическая вероятность случайного события *«A»,* т. е. появления объекта с данным признаком в результате данного опыта (испытания), представляет собой отношение числа шансов, благоприятствующих событию *«A»,* к общему числу шансов, благоприятствующих или неблагоприятствующих событию.

В ящике имеется 10 шаров, из которых 8 — черных. Опыт состоит в вынимании, не глядя, одного шара. Равновозможность исхода здесь в том, что вынимается шар случайно. Несовместность в том, что шар может быть только или черным, или нечерным. Вероятность того, что вынутый из ящика шар будет черным, равна

$$p=\frac{A}{N}=\frac{8}{10}=0{,}8.$$

Вероятность противоположного события равна

$$q=1-p,$$

т. е. вероятность вынуть нечерный шар равна

$$q=1-0{,}8=0{,}2.$$

Вероятность достоверного события равна 1 (вероятность вынуть один черный шар из 10 черных), вероятность невозможного события равна 0 (вероятность вынуть белый шар из 10 черных). Эксперименты показывают, что в такой «урновой» схеме частота наступления события при большом числе испытаний соответствует его вероятности.

Попадание в культурный слой предметов из быта людей аналогично, до некоторой степени, случайному вниманию шаров из ящика. Аналогия эта не полная, так как в некоторых частях культурного слоя (например, обстановка внутри дома) специально подбирали предметы, в других частях они оказались случайно выкинутые или потерянные. Последние случаи преобладают.

Аналогия приближается к полной, когда мы статистически устанавливаем, что выборка является случайной, а исследуемая генеральная совокупность — однородной.

Если события «A» и «B» несовместны, то вероятность того, что произойдет или одно, или другое событие, равна сумме вероятностей каждого из них, взятого отдельно (*теорема сложения вероятностей*).

Совокупность событий, в которой каждое событие исключает другое, причем какое-то из них обязательно произойдет в результате испытания, называется *полной системой событий.*

Сумма вероятностей событий, составляющих полную систему событий, равна единице.

В генеральной совокупности из N объектов имеется A объектов с значением признака «A» и B объектов со значением признака «B». Тогда вероятность того, что на случайно взятом объекте из этой генеральной совокупности будет значение признака «A», равна $p(A)=A/N$, а вероятность того, что на нем будет значение признака «B», равна $p(B)=B/N$. Допустим, что значения «A» и «B» принадлежат одному признаку и значит несовместны. Тогда если мы вынем случайно взятый из генеральной совокупности объект, то вероятность того, что на нем окажется или «A» или «B», равна

$$p(A+B)=p(A)+p(B).$$

События «A» и «B» будут *независимыми* друг от друга, если вероятность появления одного из них не изменяется в результате появления или непоявления другого события. Если вероятность изменяется, то эти события будут зависимыми друг от друга.

Вероятность появления сложного события, состоящего из последовательного осуществления двух независимых друг от друга событий, равна произведению вероятностей каждого из них в отдельности (*первая теорема произведения вероятностей*).

Если мы вынем случайно два объекта из генеральной совокупности, каждый раз возвращая предмет в совокупность, то вероятность того, что первый из них будет со значением признака «A», а второй — «B» (т. е. сложное событие «AB»), равна

$$p(AB)=p(A)\,p(B).$$

Пусть значения «A» и «B» принадлежат разным признакам, тогда они совместимы. Вероятность того, что вынутый из генеральной совокупности объект будет обладать одновременно значением одного признака «A» и значением другого признака «B», тоже равна

$$p(AB)=p(A)\,p(B).$$

Но это будет так, если значения «A» и «B» в генеральной совокупности независимы. Это значит, что вынимание объекта с «A» не меняет вероятность вынуть во второй раз объект с «B» на большую или меньшую, чем если бы это делалось в первый раз.

Связь событий «A» и «B» улавливается понятием «условной вероятности» $p(B/A)$, т. е. вероятности появ-

ления события «B» при условии, что уже осуществилось событие «A».

В рамках данного выше определения условия вероятность $p(B/A)$ определяется как доля объектов со значениями «A» и «B» среди всех объектов со значением «A». Вероятность того, что случайно вынутый объект будет обладать значениями «A» и «B», если между этими значениями есть какая-то зависимость, равна

$p(AB)=p(A)p(B/A)$ *(вторая теорема произведения вероятностей).*

Если события «A» и «B» независимы друг от друга, то

$$p(B)=p(B/A) \text{ и } p(A)=p(A/B).$$

4. Статистическая проверка гипотез

Выборочное исследование строится в виде так называемой *проверки статистических гипотез.*

Статистическая гипотеза — это какое-то предположение о свойствах генеральной совокупности, которое можно проверить (принять или отвергнуть), применяя методы математической статистики к данным выборки.

Задача статистической проверки гипотезы состоит в том, чтобы установить, насколько выборочные данные согласуются с предположениями, сделанными относительно генеральной совокупности, и если они не согласуются, то установить, может или не может это быть вызвано только случайностями образования выборки.

Формулируется какая-то гипотеза относительно генеральной совокупности. Выбирается заранее (обычно единый для всего исследования) *доверительный уровень* (иначе «уровень значимости»).

Доверительный уровень — это такое значение вероятности, при котором событие, имеющее вероятность выше этого значения (уровня), считается практически достоверным. Доверительный уровень определяет, какова вероятность ошибки отвергнуть гипотезу как неправильную, в то время как она на самом деле правильна (ошибка I рода). Обычно в археологии считают достаточным доверительный уровень 0,95. Это значит, что событие, осуществляющееся с вероятностью выше 0,95, считается практически достоверным, а противоположное событие (вероятность ниже $1-0,95=0,05$) считается практически невозможным[9]. Если гипотеза отвергается при доверительном уровне 0,95, это значит, что вероятность отвергнуть ее в то время, как она правильна (это и есть ошибка I рода), составляет всего 0,05.

Затем применяют тот или иной *статистический критерий.* Статистический критерий — это комплекс правил, выполнение которых определяет, в каких случаях, при каких результатах исследования выборок статистическая гипотеза может быть принята как правильная, допустимая, согласующаяся с данными выборки, а при каких должна быть отклонена как неправильная, недопустимая, противоречащая данным выборки. Статистический критерий обычно основан на той или иной функции, которая вычисляется по выборочным данным. Ее для краткости и называют критерием. Полученное значение критерия сравнивают с табличными, теоретически вычисленными его значениями. Из этих таблиц мы узнаем для того или иного доверительного уровня (кроме 0,95 применяют 0,90; 0,99) так называемую *«критическую область»* значений — значения большие (в некоторых случаях меньшие) некоторой определенной граничной величины. Если по выборке получено значение критерия, которое попадает в «критическую область» для избранного доверительного уровня, то гипотеза отклоняется, как маловероятная, практически невозможная и принимается альтернативная гипотеза, заключающаяся в противоположном утверждении.

Если же полученное значение критерия не достигает табличного значения (т. е. попадает в область «принятия гипотезы»), то считается, что нулевая гипотеза не опровергнута. При этом, правда, возможна ошибка принять неопровергнутую нулевую гипотезу за истинную, в то время как она ложна (ошибка II рода). Вероятность того, что не будет совершена ошибка II рода определяет так называемая «мощность» критерия.

Табличные значения статистических критериев, как правило, строятся на основе предположения о том, что выборочные данные случайно и независимо извлечены из генеральной совокупности с заданным распределением вероятностей. Это предположение позволяет рассчитать распределение вероятностей тех или иных значений критерия, его «плотность» и «функцию распреде-

[9] Обратим внимание на произвольность выбора конкретного значения доверительного уровня.

ления» (об этих понятиях см. гл. I, § 6), которые задаются табличным образом, так что критическая область значений критерия определяется вероятностью того, что значение критерия попадает в нее.

Соотношение правильности нулевой гипотезы и ее проверки

Гипотеза	Гипотеза зерна	Гипотеза неверна
Отвергнута	Ошибка I рода	Правильное решение
Не отвергнута	Правильное решение	Ошибка II рода

Статистические гипотезы обычно называют *нулевыми гипотезами,* потому что они, как правило, состоят в предположении об отсутствии различия между какими-то значениями или об отсутствии связи между признаками, объектами или явлениями. Так же и в археологии. Совершить ошибку II рода, т. е. не отвергнуть ложную нулевую гипотезу (констатировать отсутствие этих различий и связей, когда они на самом деле имеются),— это значит быть излишне осторожным и воздержаться от выводов. Совершить ошибку I рода и отвергнуть правильную гипотезу (констатировать наличие различий и связей, когда их нет на самом деле) значительно опаснее. Это означает сделать неверные выводы. В археологии обычно не подсчитывают и не указывают вероятность ошибки II рода, ограничиваясь указанием на то, что при неопровержении нулевой гипотезы возможно в дальнейшем, при накоплении дополнительных материалов, она будет все же отвергнута. Наиболее верный путь уменьшить риск ошибки обоих родов — это увеличить объем выборки и связать выводы, основанные на данной выборке, с выводами других исследований.

Археологи всегда понимали, что изучают материальную культуру того или иного общества по малой ее части, извлеченной из земли, т. е. по выборке. Им всегда было ясно, что в этой «выборке» могут быть случайности, искажающие представление об этой материальной культуре. Статистический метод позволяет учитывать эти случайности выборки и оценивать риск (вероятность) ошибки при суждении о генеральной совокупности.

Наиболее часто нулевая гипотеза состоит в утверждении, что то или иное различие между величинами, та или иная связь между объектами или признаками несущественна или незначима. Под этим понимается то, что эти статистические явления имеют причину не в генеральной совокупности, а порождены случайностями выборки. «*Существенным*», или «*значимым*», мы будем называть всякое статистическое явление, которое не является порождением только случайностей выборки, а отражает особенности генеральной совокупности.

ГЛАВА I

КОЛИЧЕСТВЕННЫЕ ПРИЗНАКИ

1. Вариационный ряд. Полигон. Гистограмма

Перед нами выборка в виде n элементов (n — *объем выборки*) из генеральной совокупности, представляющая серию археологических объектов. На каждом из этих объектов измерено значение какого-то количественного признака. Для каждого значения признака (x) мы можем подсчитать, сколько раз оно встречается, т. е. сколько объектов обладает именно этим значением признака. Число, означающее, сколько раз встречено то или иное значение, будем называть *частотой* (m). Если расположить значения признака в каком-либо порядке (возрастающем или убывающем) и с каждым значением сопоставить его частоты, то получим *вариационный ряд*, т. е. сопоставление значений признака и числа объектов, на которых оно зарегистрировано. Вариационный ряд тем самым показывает распределение объектов по значениям признака, т. е. показывает, как часто встречается то или иное значение признака. *Распределением признака* в данной совокупности называется соотношение численностей отдельных частей данной совокупности, каждая из которых характеризуется тем или иным значением этого признака.

Долю того или иного значения признака среди всей совокупности произведенных измерений (m/n) будем называть относительной частотой или *частостью* (w). Сумма всех частостей равна 1 (или 100%). Сумма всех частот равна объему выборки n. Таким образом, имеем:

Значения признака	$x_1, x_2, \ldots, x_i, \ldots, x_l$
Частота	$m_1, m_2, \ldots, m_i, \ldots, m_l$
Частость	$w_1, w_2, \ldots, w_i, \ldots, w_l$

Кроме того, мы можем записать, что

$$\sum_{i=1}^{l} m_i = n; \quad \sum_{i=1}^{l} w_i = 1,$$

где l — число значений признака; знак Σ здесь и далее означает сумму пронумерованных членов (i — номер члена по порядку), начиная с номера, указанного внизу, и кончая номером, указанным на верху этого знака.

Как мы знаем, количественные признаки могут быть мерными (непрерывными) и счетными (дискретными). Для построения вариационного ряда при мерном (непрерывном) признаке его наблюденные значения следует представить как интервальные. Берется какой-то интервал α, область значений разбивается на отрезки длиной α и подсчитывается число тех случаев, при которых значение признака входит в каждый из этих отрезков.

Рекомендуют формулу для выбора оптимального интервала, т. е. такого, чтобы он не был слишком узким, а потому слишком подверженным случайным отклонениям и вместе с тем не был бы слишком широким, чтобы не потерялись характерные черты распределения признака. Можно использовать *формулу Стерджесса*:

$$\alpha = \frac{x_{\max} - x_{\min}}{1 + \log_2 n} = \frac{x_{\max} - x_{\min}}{1 + 3{,}32 \lg n},$$

где $x_{\max}$ — самое большое значение признака; $x_{\min}$ — самое малое его значение.

Величину $P = x_{\max} - x_{\min}$ будем называть *вариационным размахом*. Значения дискретного признака могут также быть сгруппированы в интервалы.

Плотностью распределения называется отношение частоты (абсолютная плотность) или частости (относительная плотность) к величине интервала:

$$f_a = \frac{m}{\alpha} \text{ или } f_{отн} = \frac{w}{\alpha}.$$

Для графического выражения распределения признака или вариационного ряда строят *полигон распределения*. На оси абсцисс откладывают значения дискретного признака или интервалы, а на оси ординат — значения частот (или частостей). Точки пересечения перпендикуляров, построенных к осям координат в этих точках (в случае непрерывного признака — к середине

интервалов), будут соответствовать значению варианта и его частоте (или частости). Соединив отрезками точки пересечения, получают полигон распределения.

Существует другой способ графического выражения распределения признака — *гистограмма.*

При построении гистограммы на оси абсцисс откладывают значение признака, выраженное в его интервалах, а на оси ординат — плотность распределения. Полученный над каждым интервалом, отложенным на оси абсцисс, прямоугольник с высотой, равной плотности распределения, имеет площадь, равную частоте (или частости). Гистограмма дает ту же картину распределения, что и полигон в случае равномерных интервалов: интервал принимается равным 1 и плотность частот (или частостей) не отличается от самих частот (или частостей). Но картину, отличную от полигона и более правильно выражающую характер распределения признака, гистограмма дает в случае неравных интервалов.

2. Метод «скользящей средней»

Часто вариационный ряд, или распределение признака, оказывается сильно колеблющимся и тогда полигон получается как бы угловатым, имеющим характер ломаной во многих точках линии. Правильный выбор интервала «сглаживает» до некоторой степени эти случайные колебания полигона. Но, кроме того, для более ясного выражения общей тенденции вариационного ряда можно прибегнуть к выравниванию, «сглаживанию», например методом «*скользящей средней*».

Сначала суммируются попарно частоты:

$$\Sigma'_{1,2}=m_1+m_2;\ \Sigma'_{2,3}=m_2+m_3 \text{ и т. д.}$$

Затем суммируются попарно эти суммы:

$$\Sigma''_{1,2,2,3}=\Sigma'_{1,2}+\Sigma'_{2,3};\ \Sigma''_{2,3,3\cdot4}=\Sigma'_{2,3}+\Sigma'_{3,4} \text{ и т. д.}$$

Такая операция проводится a раз. Затем каждая из найденных сумм делится на 2^a. Полученные значения откладываются на оси ординат.

Можно складывать по три члена вариационного ряда, но тогда нужно делить на 3^a и т. д. Существенно важным при этом представляется равенство интервалов. При выравнивании нельзя брать слишком большое значение «a» и число складываемых членов, так как в этом случае можно скрыть характерные черты вариационного ряда (как при выборе слишком большого интервала). Выбор этих значений (a) зависит каждый раз от самого вариационного ряда и целей исследования.

Пример 4.

Даны результаты взвешивания 203 серебряных монет с изображением барса на лицевой стороне и именем хана Тактамыша на оборотной, чеканенных при Василии Дмитриевиче в Москве в конце XIV в.[1] Вариационный ряд и полигон распределения оказался с сильными колебаниями и отклонениями. Требуется его выравнить, «сгладить». Был применен способ «скользящей средней» (см. табл. 1, где x — значение признака, m — частота; Σ — сумма частот).

Таблица 1

x, г	m	Σ'	Σ''	Σ'''	$\Sigma/2^3$	x, г	m	Σ'	Σ''	Σ'''	$\Sigma/2^3$
0,76	1	—	—	—	—	0,88	11	34	62	126	15,75
0,77	2	3	5	—	—	0,89	17	28	90	152	19,00
0,78	0	2	2	7	0,88	0,90	45	62	131	221	27,62
0,79	0	0	0	2	0,25	0,91	24	69	126	257	32,12
0,80	0	0	1	1	0,12	0,92	33	57	101	227	28,38
0,81	1	1	3	4	0,50	0,93	11	44	61	162	20,25
0,82	1	2	4	7	0,88	0,94	6	17	32	93	11,62
0,83	1	2	5	9	1,12	0,95	9	15	26	58	7,25
0,84	2	3	10	15	1,88	0,96	2	11	13	39	4,88
0,85	5	7	19	29	3,62	0,97	0	2	3	16	2,00
0,86	7	12	42	61	7,62	0,98	1	1	3	6	0,75
0,87	23	30	64	106	13,25	0,99	1	2			

Таблица 2

x	m	x	m	x	m
0,75—0,77	3	0,84—0,86	14	0,93—0,95	26
0,78—0,80	0	0,87—0,89	51	0,96—0,98	3
0,81—0,83	3	0,90—0,92	102	0,99—1,01	1

[1] См.: *Федоров-Давыдов Г. А.* Монеты Московской Руси. М., 1981. С. 51. Табл. 7.

Выравнивания в данном примере не потребуется, если признак разбить на более удобные интервалы. Максимальная масса — 0,99 г, минимальная — 0,76. Находим

$$a=\frac{x_{max}-x_{min}}{1+3{,}32\lg n}=\frac{0{,}99-0{,}76}{1+3{,}32\lg 203}\approx 0{,}03.$$

Этот интервал дал следующий вариационный ряд, который не требует выравнивания (табл. 2).

3. Характеристики вариационного ряда

Средняя арифметическая $\bar{x}$ вычисляется как сумма произведений каждого значения признака x на соответствующее ему значение частости w:

$$\bar{x}=\sum_{i=1}^{l} x_i w_i,$$

где l — количество значений.

Если все $w_i=1/n$, т. е. каждый вариант представлен один раз, то

$$\bar{x}=\frac{x_1+x_2+x_3+\dots}{n}=\frac{1}{n}\sum_{i=1}^{l} x_i.$$

Если значение признака выражено интервалом, то при вычислении средней арифметической в качестве x_i берется середина i-го интервала.

Средняя арифметическая представляет собой важнейшую характеристику вариационного ряда, определяет тот стандарт, норму, вокруг которых колеблются значения признака, т. е. дает типичный для совокупности размер признака. Она имеет смысл только для однородного признака, например, при измерении сосудов одного вида для однородной совокупности. Средняя при измерении, скажем, высот амфор и пифосов не имеет смысла. Для того чтобы определить, однороден признак или нет, нужно построить его полигон или гистограмму. Они должны быть одновершинными и примерно симметричными. Другими словами, значения признака (варианты) должны концентрироваться вблизи одного какого-то значения. Мы будем называть *однородным* такой количественный признак, у которого имеются только случайные отклонения от одного нормативного значения. Например, длина ножей одного вида, диаметр горла сосудов одного вида и т. п. Такое распределение хорошо моделируется так называемым нормальным распределением, о котором пойдет речь ниже (см. гл. I, § 6). Смешанным количественным признаком будет признак, имеющий несколько нормативных стандартов, вокруг которых колеблются в силу случайных отклонений его значения. Таким будет длина ножей разных видов, высота сосудов разных видов и т. п.

Среднее квадратическое отклонение σ вычисляется как квадратный корень из суммы отклонений значений признака x от средней $\bar{x}$, возведенных в квадрат и умноженных на частости ω этого значения, т. е.

$$\sigma=\sqrt{\sum_{i=1}^{l}(x_i-\bar{x})^2 w_i}.$$

Величина σ^2 называется *дисперсией*.

Среднее квадратическое отклонение может быть также вычислено по следующей формуле, равносильной вышеприведенной формуле:

$$\sigma=\sqrt{\overline{x^2}-\bar{x}^2},$$

т. е. как квадратный корень из разности средней арифметической квадратов всех значений признака и квадрата средней арифметической.

Среднее квадратическое отклонение показывает, насколько «вариабелен» признак, т. е. насколько сильно он колеблется (рассеян) вокруг своей арифметической средней.

Среднее квадратическое отклонение, таким образом, есть мера рассеивания признака. Есть и другие меры рассеивания, например: вариационный размах $P=x_{max}-x_{min}$, среднее линейное абсолютное отклонение $Л=\sum_{i=1}^{l}|x_i-\bar{x}|w_i$, энтропия $\Sigma w_i \log_a w_i$ (см. гл. III, § 6), но чаще всего используются σ и σ^2, поскольку именно они характерны для нормального распределения.

Средний квадрат отклонений может быть вычислен от любого значения признака. Но среднее квадратическое отклонение, вычисленное от средней арифметической, будет минимальным.

Одинаковые σ при колебании сравнительно малого по размерам признака (например, ширины ручки сосуда) и при колебании большеразмерного признака (на-

пример, высота сосуда) говорят о разной вариабельности признака. Для маломерного признака даже малая σ может говорить о сильной его колеблемости вокруг его средней, а при большеразмерном признаке даже значительная σ может свидетельствовать лишь о незначительной колеблемости значений признака вокруг своей средней. Эти соображения учитываются в относительных показателях разброса, в частности в *коэффициенте вариации*: $V=\sigma/\bar{x}$.

Важной характеристикой вариационного ряда является *мода* (Mod) — то значение признака, которое встречается чаще всех других, т. е. которому соответствует наибольшая частота. В случае дискретности признака определение моды не вызывает трудностей. В случае интервального вариационного ряда берется модальный интервал (по наибольшей частоте, если интервалы равны, и по наибольшей плотности, если интервалы не равны). Далее берут середину модального интервала или, более точно, среднюю арифметическую модального интервала.

Если средняя арифметическая совпадает с модой, то вариационный ряд может быть симметричным, если не совпадает, то он асимметричен. Чтобы выразить степень асимметрии, применяют *коэффициент асимметрии*

$$K_a=\frac{\sum_{i=1}^{l}(x_i-\bar{x})^3 w_i}{\sigma^3}$$

или более простой коэффициент

$$K'_a=\frac{\bar{x}-\mathrm{Mod}}{\sigma}.$$

Если $K_a<0$, то полигон имеет вытянутость влево (левосторонняя асимметрия), если $K_a>0$, то — вправо (правосторонняя асимметрия).

Полигон распределения (вариационного ряда) может быть плоским или крутым, островершинным. Чтобы выразить степень уплощенности или крутизны, применяют *коэффициент эксцесса*:

$$K_э=\frac{\sum_{i=1}^{l}(x_i-\bar{x})^4 w_i}{\sigma^4}-3.$$

Чем $K_э$ больше, тем более крутой полигон имеет вариационный ряд.

Пример 5.

Даны результаты замеров горловин 222 сосудов одного вида. Требуется вычислить среднюю арифметическую величину, среднеквадратическое отклонение и коэффициент вариации. Все данные и промежуточные результаты вычислений представлены в табл. 3.

Таблица 3

x	m	w	xw	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^2 w$	x^2	$x^2 w$
4,0	5	0,02	0,08	—0,55	0,3025	0,006050	16,00	0,3200
4,1	7	0,03	0,12	—0,45	0,2025	0,006075	16,81	0,5043
4,2	13	0,06	0,25	—0,35	0,1225	0,007350	17,64	1,0584
4,3	20	0,09	0,39	—0,25	0,0625	0,005625	18,49	1,6641
4,4	24	0,11	0,48	—0,15	0,0225	0,002475	19,36	2,1296
4,5	35	0,16	0,72	—0,05	0,0025	0,000400	20,25	3,2400
4,6	40	0,18	0,83	0,05	0,0025	0,000450	21,16	3,8088
4,7	36	0,16	0,75	0,15	0,0225	0,003600	22,09	3,5344
4,8	25	0,11	0,53	0,25	0,0625	0,006875	23,04	2,5344
4,9	11	0,05	0,25	0,35	0,1225	0,006125	24,01	1,2005
5,0	6	0,03	0,15	0,45	0,2025	0,006075	25,00	0,7500
Σ	222	1,00	4,55			0,051100		20,7445

Таким образом вычисляются среднеарифметическая величина

$$\bar{x}=\sum_{i=1}^{l} x_i w_i=4{,}55,$$

дисперсия

$$\sigma^2=\sum_{i=1}^{l}(x_i-\bar{x})^2 w_i=0{,}0511$$

и среднеквадратическое отклонение

$$\sigma=\sqrt{0{,}0511}=0{,}23.$$

Дисперсия и среднеквадратическое отклонение могут быть получены другим путем:

$$\overline{x^2}=\sum x_i^2 w_i=20{,}7445;$$

$$\bar{x}^2=4{,}55^2=20{,}7025;$$

$$\sigma^2 = 20{,}7445 - 20{,}7025 = 0{,}0420;$$

$$\sigma = \sqrt{0{,}0420} = 0{,}20.$$

Некоторое небольшое расхождение между σ, полученной вторым способом, и σ, полученной первым способом, объясняется округлениями при подсчетах ω_i и $x_i w_i$. Коэффициент вариации

$$V = \frac{\sigma}{\bar{x}} = \frac{0{,}23}{4{,}55} = 0{,}05.$$

Подсчитаем теперь коэффициенты асимметрии и эксцесса. Все данные и промежуточные результаты вычислений представлены в табл. 4.

Таблица 4

x	w	$x-\bar{x}$	$(x-\bar{x})^3$	$(x-\bar{x})^3 w$	$(x-\bar{x})^4$	$(x-\bar{x})^4 w$
4,0	0,02	—0,55	—0,16637	—0,003328	0,0915	0,001830
4,1	0,03	—0,45	—0,09112	—0,002734	0,0410	0,001230
4,2	0,06	—0,35	—0,04288	—0,002572	0,0150	0,000900
4,3	0,09	—0,25	—0,01562	—0,001406	0,0039	0,000351
4,4	0,11	—0,15	—0,00338	—0,000371	0,0005	0,000056
4,5	0,16	—0,05	—0,00012	—0,000020	0,0000	0,000001
4,6	0,18	0,05	0,00012	0,000022	0,0000	0,000001
4,7	0,16	0,15	0,00338	0,000540	0,0005	0,000081
4,8	0,11	0,25	0,01562	0,001719	0,0039	0,000430
4,9	0,05	0,35	0,04288	0,002144	0,0150	0,000750
5,0	0,03	0,45	0,09112	0,002734	0,0410	0,001230
Σ	1,00			—0,003272		0,006860

Таким образом, коэффициент асимметрии

$$K_a = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^3 w_i}{\sigma^3} = \frac{-0{,}003272}{0{,}23^3} = -0{,}2689 \approx -0{,}27.$$

Имеется некоторая незначительная левосторонняя асимметрия. Подсчитаем теперь коэффициент эксцесса:

$$K_э = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^4 w_i}{\sigma^4} - 3 = \frac{0{,}006860}{0{,}23^4} - 3 = 2{,}4514 - 3 =$$

$$= -0{,}5486 \approx -0{,}55.$$

Наблюдается некоторый отрицательный эксцесс, т. е. небольшая уплощенность полигона. Строим полигон распределения (рис. 1).

4. Кумуляты

Вычислим для каждого значения вариационного ряда сумму частот (или частостей) для всех значений, не превышающих данное значение. Эти суммы называются

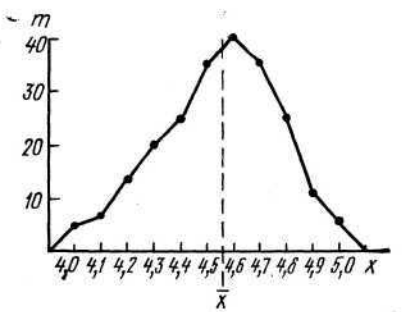


Рис. 1. Полигон распределения признака: ширина горла сосудов (по частотам)

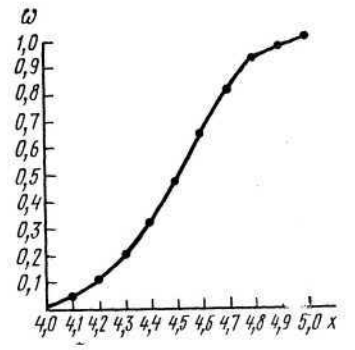


Рис. 2. Кумулята того же распределения

накопленными частотами (или *накопленными частостями*). Если на оси абсцисс отложить значения признака, а на оси ординат — накопленные частоты, то получим кривую, которую называют *кумулятой* (рис. 2). Она исходит из начала координат и заканчивается в правом верхнем углу, в точке, которая соответствует самому большому значению признака и общей численности выборки (для частот) или 1 (для частостей).

Накопленные частоты находят применение в так называемых *кривых Лоренца*, наглядно отражающих степень концентрации каких-либо объектов по группам.

Кривая Лоренца строится следующим образом. Имеется ряд интервальных значений дискретного признака с частотами и частостями (т. е. вариационный ряд). Подсчитывают накопленные частости и откладывают их на оси абсцисс. Подсчитывают, сколько элементов, которые составляют единицу исчисления признака, содержит группа объектов, относящаяся к каждому интервалу. Далее полученные числа переводят в

частости с помощью деления каждого из них на их сумму. По этим частостям высчитывают второй ряд накопленных частостей для каждого интервала. Откладывают их на оси ординат. Для каждого интервала получают таким образом два значения накопленных частостей, каждой такой паре соответствует точка. Точки соединяют и получают кривую, начинающуюся так же, как кумулята, в начале координат, а заканчивающуюся в правом верхнем углу графика.

Таблица 5

x	m	x	m	x	m	x	m
1	4	32	1	64	1	206	1
2	2	34	3	69	1	232	1
3	1	35	1	79	1	267	1
11	1	38	1	88	1	360	1
12	1	41	1	97	1	391	1
13	3	42	1	101	1	450	2
16	3	43	2	103	1	632	1
17	1	44	2	122	1	1928	1
23	2	45	2	141	1		
24	1	48	1	163	1		
25	1	58	1	170	1		
31	1	61	1	185	1		

Таблица 6

x	m	w	$w_н$	m'	w'	$w'_н$
1—50	36	0,61	0,61	866	0,12	0,12
51—100	7	0,12	0,73	516	0,07	0,19
101—150	4	0,07	0,80	467	0,06	0,25
151—200	3	0,05	0,85	518	0,07	0,32
201—250	2	0,03	0,88	438	0,06	0,38
251—300	1	0,02	0,90	267	0,04	0,42
. . .	...	...	...	...	...	...
351—400	2	0,03	0,93	751	0,10	0,52
401—450	2	0,03	0,96	900	0,12	0,64
. . .	...	...	...	...	...	...
601—650	1	0,02	0,98	632	0,09	0,73
. . .	...	...	...	...	...	...
1901—1950	1	0,02	1,00	1928	0,27	1,00
Σ	59	1,00		7283	1,00	

Если дискретный признак распределен равномерно, кривая Лоренца превращается в диагональ. Чем больше концентрация, т. е. неравномерность распределения признака, тем сильнее кривая Лоренца выгибается вправо вниз от этой диагонали.

Пример 6.

Дано распределение бус по погребениям в I Поломском могильнике IX—X вв.[2] Требуется построить кри-

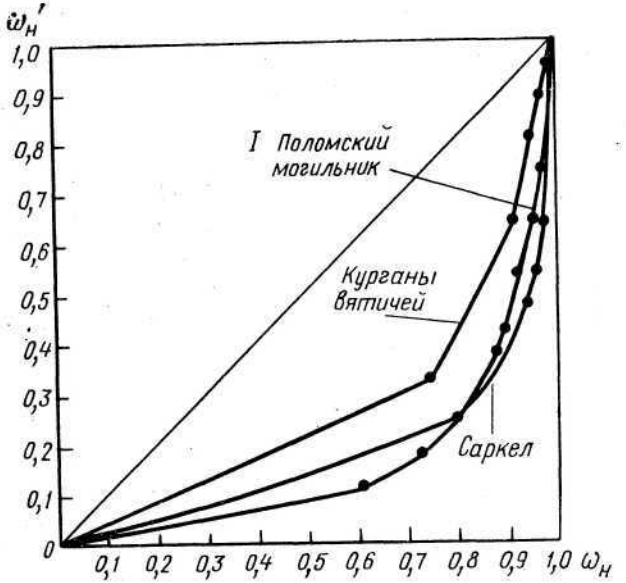


Рис. 3. Кривые Лоренца для распределения бус по погребениям

вую Лоренца и тем самым показать степень концентрации бус в погребениях.

В табл. 5 дано распределение бус по погребениям (x — количество бус, m — число погребений с этим количеством бус).

Этот вариационный ряд может быть разбит на интервалы по 50 бус. Подсчитываем для каждого интервала x число погребений с данным числом бус, т. е. частоту m; долю погребений с данным числом бус, т. е. частость w; накопленную частость $w_н$; количество бус

[2] См.: *Львова З. А.* Указ. соч.

m' во всех погребениях, относящихся к данному интервалу; долю того количества бус, которое содержится во всех погребениях, относящихся к данному интервалу, т. е. частость w'; накопленную частость $w_н'$. Все эти результаты сведены в табл. 6.

Данные этой таблицы переносятся на график. Получится довольно вогнутая кривая, показывающая значительную концентрацию. Действительно, около 40% всех бус сосредоточено в двух погребениях, около 60% всех бус сосредоточено в шести погребениях, что составляет только 10% всех погребений.

Площадь между кривой Лоренца и диагональю может быть мерой этой концентрации.

Кривая Лоренца, построенная для бус из могильника около Саркела-Белой Вежи[3] дает почти такую же степень концентрации, для вятических курганов[4] — существенно более низкую (рис. 3).

Подобные исследования могут выявить различную степень концентрации находок в погребении, определенную степень имущественного неравенства, при исследовании численности монетных кладов — особенности денежного обращения и кладообразования.

5. Временной ряд

Одним из видов вариационного ряда может быть *временной* ряд, часто встречающийся в археологии. Его признак — это мерный количественный признак — отрезок времени, разбитый на какие-то интервалы, а частота — количество единиц какого-то объекта, относящихся к тому или иному интервалу времени. Временные ряды чаще всего возникают в археологии при исследовании распределения какого-либо объекта или признака по хронологическим периодам, соответствующим прослойкам культурного слоя. Дата объекта рассматривается как его признак.

На полигоне (или гистограмме) на оси абсцисс откладываются отрезки времени, на оси ординат — частота (или плотность распределения) данного объекта. Временной ряд показывает, как тот или иной объект распределяется во времени, и может быть построен, когда материал хорошо делится хронологически. Мода этого ряда — время наибольшего распространения этого объекта.

Заметим, что временной ряд перестает иметь вид вариационного ряда, когда с каким-либо отрезком времени сопоставлены не частоты, а другие показатели.

Для выравнивания временных рядов удобен способ скользящей средней[5].

Пример 7.

Керамические формы из поселения Змейского в Северной Осетии (эпоха бронзы, X—VIII вв. до н. э.) были классифицированы на шесть видов, из них четыре вида с орнаментом (II, III, V, VI)[6]. Исследователей интересовало соотношение этих типов керамики и его изменение во времени. Культурный слой был разбит на 10 условных прослоек одинаковой толщины и было сделано весьма вероятное допущение, что при равномерности накапливания культурного слоя и одинаковой интенсивности жизни эти прослойки соответствуют одинаковым отрезкам времени. Сосчитано было количество черепков каждого вида керамики для каждой прослойки. Можно построить для каждого вида временной ряд, взяв абсолютные значения числа керамики в каждой прослойке. Такой временной ряд аналогичен вариационному: признак — датировка черепка по слою, часто-

Таблица 7

Виды керамики	Слой				
	I—II	III—IV	V—VI	VII—VIII	IX—X
II	0,45	0,44	0,49	0,51	0,57
III	0,10	0,08	0,10	0,10	0,12
V	0,14	0,06	0,06	0,07	0,04
VI	0,22	0,30	0,21	0,19	0,16
Прочие	0,09	0,12	0,14	0,13	0,13

та — число черепков данного вида в слое. Но при этом мы рискуем ошибиться в оценке степени распространенности данного вида в тот или иной отрезок времени. Ошибка может возникнуть от несопоставимости значе-

[3] См.: *Артамонова О. А.* Могильник Саркела-Белой Вежи // МИА. М., Л., 1963. № 109.

[4] См.: *Арциховский А. В.* Курганы вятичей. Л., 1930.

[5] Подробнее о временных рядах см.: *Кендэл М.* Временные ряды. М., 1981.

[6] См.: *Деопик Д. В., Узянов А. А.* Статистический анализ керамического комплекса (Орнаменты сосудов с поселения Змейского) // Математические методы в исторических исследованиях. М., 1972.

ний для разных прослоек (периодов времени), так как в каждый период могло попасть в слой разное количество керамики. Заметим, что другая возможность несопоставимости абсолютных данных по слоям заключается в том, что в каждом слое могли быть раскопаны разные площади (в нижних меньшие, в верхних большие). Чтобы устранить эту несопоставимость, следует взять относительные количества, т. е. долю данного вида керамики среди всей керамики из прослойки.

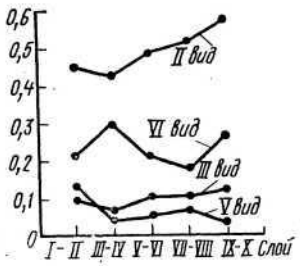


Рис. 4. Полигон распределения долей керамики различного вида в поселке Змейском

Были получены следующие значения доли керамики каждого вида среди всей керамики по прослойкам (табл. 7):

По этим значениям построены временные ряды, значительно более точно отражающие динамику соотношения четырех видов керамики для каждого отрезка времени. Графически они могут быть выражены следующими полигонами на одном графике (рис. 4). На полигонах видно снижение с течением времени доли керамики V вида и некоторое повышение, а затем постепенное понижение доли керамики VI вида.

Пример 8.

При раскопках Саркела-Белой Вежи было выделено 6 слоев разной мощности и продолжительности и предшествующий им комплекс ранних ям, подсчитано количество фрагментов керамики разных видов. Для амфор были получены следующие результаты (табл. 8).

Таблица 8

Слой	Толщина слоя, см	Количество фрагментов амфор, тыс.	Количество всех фрагментов керамики, тыс.	Слой	Толщина слоя, см	Количество фрагментов амфор, тыс.	Количество всех фрагментов керамики, тыс.
Ямы	50	0,8	3,7	III	50	5,3	12,9
I	50	2,9	24,8	IV	50	3,6	7,3
IIа	25	0,6	6,1	V	50	2,9	5,8
IIб	25	0,9	5,2				

Были построены гистограммы временных рядов с учетом неодинаковости временных интервалов, которые были вполне обоснованно приравнены к толщине культурного слоя[7]. Абсолютные количества черепков какого-либо вида керамики были разделены на толщину слоя. Полученные частные (плотности распределения) оказались следующие (в тыс.; 25 см взяты за 1):

Ямы	0,40		
Слой		Слой	
I	—1,45	III	—2,65
IIа	—0,60	IV	—1,80
IIб	—0,90	V	—1,45

Далее строят прямоугольники с основанием, равным толщине слоя, и высотой, равной этому частному. Площадь этих прямоугольников соответствует числу черепков из слоя (рис. 5).

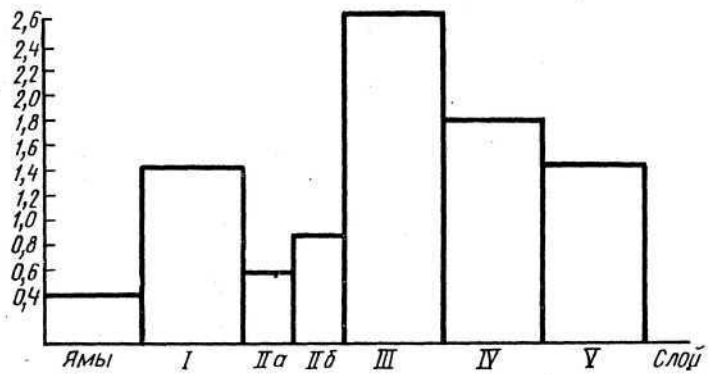


Рис. 5. Гистограмма распределения амфор в Саркеле-Белой Веже

Следует заметить, что толщина слоя не обязательно соответствует длительности его накопления. Приведенные в том же исследовании датировки слоев дают такую длительность их накопления:

ямы — первая половина IX в., т. е. примерно 50 лет;

I слой — вторая половина IX — первая половина X в. — ≈50 лет;

IIа слой — 40—60-е годы X в. (до 965 г.) — ≈30 лет;

IIб слой — вторая половина X в. — ≈30 лет;

III слой — XI в. — ≈80 лет;

[7] См.: *Плетнева С. А.* Керамика Саркела-Белой Вежи // МИА, М., Л., 1959, № 75.

IV слой — конец XI — начало XII в. — ≈30 лет;
V слой — первая половина XII в. — ≈30 лет.

С учетом этих дат может быть построена другая гистограмма для долей амфор, несколько отличающаяся от вышеприведенной. В ней частное от деления абсолютного числа черепков данного вида в слое на толщину слоя заменено частным от деления абсолютного числа черепков данного вида в слое на хронологическую длительность слоя. В результате этого деления получены следующие числа (в тыс.; 5 лет взяты на 1):

Ямы	—0,08		
Слой		Слой	
I	—0,29	III	—0,33
IIа	—0,10	IV	—0,60
IIIб	—0,15	V	—0,48

Далее, на оси абсцисс откладываются отрезки времени, соответствующие длительности отложения слоя. На этих отрезках, как на основаниях, были построены прямоугольники с высотами, равными полученным выше цифрам. Площадь прямоугольника равна количеству фрагментов амфор в соответствующем слое (рис. 6).

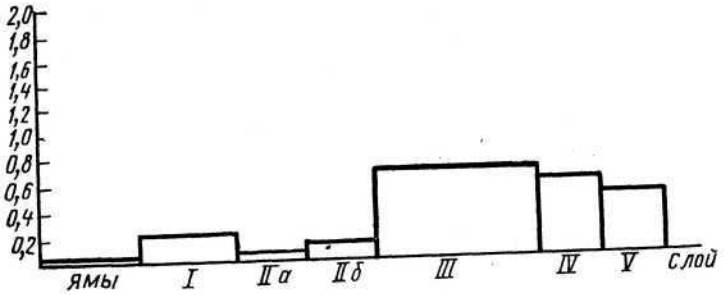


Рис. 6. Гистограмма распределения амфор в Саркеле-Белой Веже

Частное от деления абсолютного числа черепков данного вида в слое на толщину слоя, т. е. плотность распределения, можно заменить частным от деления абсолютного числа черепков данного вида в слое на общее число керамики в этом слое (т. е. долей черепков данного вида среди всей керамики). Получатся следующие цифры:

Ямы	0,22		
Слой		Слой	
I	—0,12	III	—0,41
IIа	—0,10	IV	—0,49
IIб	—0,17	V	—0,50

Доля амфор в какой-то степени аналогична плотности их распределения во времени: предполагается, что толщина слоя пропорциональна числу попавших в него черепков. Значит, безразлично, на что делить частоту — на толщину слоя или на общее число черепков в нем, чтобы получить плотность распределения.

Теперь гистограмма может быть построена так, чтобы высоты прямоугольников были равны этим долям, а основания — длительности накопления слоя. Получаем несколько другую по своему виду гистограмму (рис. 7).

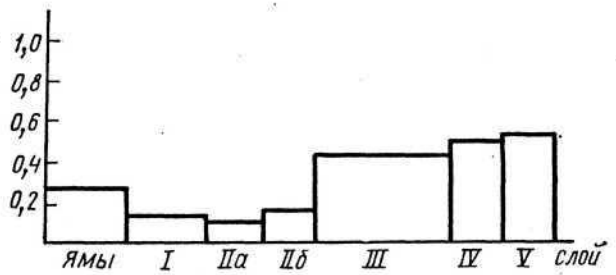


Рис. 7. Гистограмма распределения амфор в Саркеле-Белой Веже

Это объясняется тем, что число оседавших в слое черепков всех видов на самом деле не вполне пропорционально мощности культурного слоя. В I слое отложилось черепков почти в 3 раза больше, чем в IV слое, и почти в 2 раза больше, чем в III, а мощности этих слоев одинаковы. Предпочтение должно быть отдано, видимо, третьей гистограмме, так как нас прежде всего интересует распространение амфор среди других видов керамики.

6. Случайная величина. Закон распределения случайной величины. Нормальный закон распределения

Во введении уже говорилось о случайном событии и связанном с ним понятии вероятности случайного события. Введем теперь понятие *случайной величины*.

В качестве примера случайной величины рассмотрим образование одновершинного вариационного ряда, кото-

рое может быть представлено следующим образом. Существует какой-то стандарт, типическое значение признака (среднее арифметическое значение признака). Каждый раз при его реализации действительное значение признака в силу множества случайных факторов отклоняется от этого среднего. Значения признака x можно рассматривать как реализацию случайной величины, т. е. такой величины, которая принимает те или иные значения случайно, с той или иной вероятностью.

Чтобы описать дискретную случайную величину, нужно указать не только те значения, которые она может принять, но и те вероятности, с которыми она принимает эти значения. Эти данные образуют *распределение* случайной величины. Сумма вероятностей всех возможных значений случайной величины равна 1.

Распределение случайной величины задается *законом распределения.* Закон распределения может иметь табличную форму. В такой таблице будут сопоставлены возможные значения случайной величины и их вероятности.

Если случайная величина непрерывна, такая таблица не может быть построена, так как число значений непрерывной случайной величины бесконечно. Тогда для описания закона распределения прибегают к *функции распределения.*

Вероятность того, что случайная величина примет значение, меньше некоего X, называется функцией распределения.

Вероятность, с которой реализуется какой-то малый интервал значений случайной величины, называют *плотностью распределения вероятностей.* Введение этого понятия связано с тем, что случайная величина может быть как дискретной, так и непрерывной. Закон распределения случайной величины часто выражается формулой, связывающей плотность распределения вероятности со значением случайной величины. Графически закон распределения случайной величины выражается полигоном (если значения случайной величины дискретны) или кривой (если значения случайной величины непрерывны), которая строится следующим образом: на оси абсцисс откладываются значения случайной величины, а на оси ординат — плотности распределения вероятностей.

Следует различать *эмпирическое* и *теоретическое распределения.*

Эмпирическое распределение — это результат выборочного наблюдения и выражается в виде вариационного ряда. Полигон и гистограмма этого вариационного ряда — не что иное, как графическое выражение эмпирического распределения.

Теоретическое распределение — это математическая абстракция, к которой приближается распределение случайной величины в генеральной совокупности большого и бесконечного объема. Анализируя сам механизм образования случайной величины и теоретически подсчитывая вероятности реализаций его значений, математики определяют закон теоретического распределения той или иной случайной величины и дают аналитическое выражение плотности распределения как функции от значения случайной величины, а затем строят полигон, гистограмму или кривую этого распределения.

Плотность распределения в теоретическом распределении аналогична частости в эмпирическом распределении (вариационном ряду).

Так же как эмпирическое распределение (вариационный ряд), теоретическое распределение имеет свои характеристики. Сумма произведений всех значений случайной величины на соответствующие им плотности распределения носит название *математического ожидания* M. В эмпирическом распределении аналогичная математическому ожиданию величина — это средняя арифметическая. Дисперсией теоретического распределения случайной величины является сумма произведений квадратов всех отклонений значений случайной величины от математического ожидания на соответствующие им плотности распределения. Дисперсия случайной величины в теоретическом распределении аналогична дисперсии признака в эмпирическом распределении. Соответственно среднее квадратическое отклонение теоретического распределения (квадратный корень из дисперсии) аналогично среднему квадратическому отклонению в эмпирическом распределении.

Большое количество встречающихся на практике эмпирических распределений приближается к наиболее распространенному виду теоретического распределения, которое носит название *нормального* (или Гауссового) *распределения.* Если на колебания признака вокруг его средней оказывает влияние множество мелких, неучитываемых, неуловимых случайных причин, то распределение признака может быть описано нормальным за-

коном распределения. Другими словами, если колебания случайной величины складываются как суммы колебаний многих случайных величин с произвольным распределением, то такая случайная величина распределена по нормальному закону или близко к нему. Нормальное распределение, как и все теоретические распределения — математическая абстракция. Действительные распределения признаков лишь в большей или меньшей степени приближаются к нему.

Доказано, что случайная величина, состоящая из достаточно большого числа взаимно независимых слагаемых случайных величин, из которых ни одна не выделяется резко величиной своей дисперсии, имеет распределение, приближающееся к нормальному.

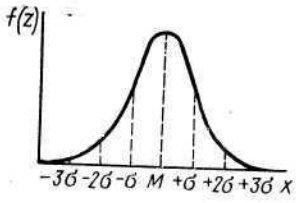


Рис. 8. Кривая нормального распределения

Если на оси абсцисс мы поместим значения нормально распределенной случайной величины, а на оси ординат — вероятность, с которой то или иное значение случайной величины (признака) может осуществиться, то получим кривую плотности нормального распределения $f(x)$ (рис. 8)[8]. Она будет иметь одну вершину (ей соответствует на оси абсцисс то значение случайной величины, которое называется математическим ожиданием) и симметричные ветви справа и слева.

График плотности распределения вероятности нормального закона имеет следующее аналитическое выражение:

$$f(z)=\frac{1}{\sigma\sqrt{2\pi}}e^{-z^2/2},$$

где $z=(x-M)/\sigma$.

Величина $f(z)$ табулирована и ее можно найти по таблице в любом пособии по математической статистике для каждого z при $\sigma=1$ (см. табл. 1).

Функция $f(z)$ достигает максимума при $z=0$ $(x=M)$, т. е. в той точке оси абсцисс, которая соответствует математическому ожиданию.

[8] Из-за парадоксальных свойств бесконечного числового континуума эту фразу следует понимать в том смысле, что значения $f(x)h$ выражают вероятности интервалов x, $x+h$ (h — малое положительное число) значений случайной величины.

Нормальное распределение случайной величины (признака) реализуется в том или ином опыте или выборке со случайными отклонениями. В этом случае чем больше замерено объектов, чем больший объем выборки, тем ближе будет эмпирическое распределение значений признака к нормальному распределению, тем ближе будет средняя арифметическая вариационного ряда (выборки) к математическому ожиданию теоретического распределения, а квадратическое отклонение (с некоторой поправкой) вариационного ряда (выборки) — к среднему квадратическому отклонению теоретического распределения.

7. Оценки характеристик генеральной совокупности по выборочным данным

Теоретическое распределение — это распределение вероятностей случайной величины. Его нельзя наблюдать непосредственно.

В большинстве статистических исследований генеральная совокупность также не доступна исследователю и обычно весьма велика. О характеристиках этой генеральной совокупности и о том теоретическом распределении, которому подчиняется признак в генеральной совокупности как случайная величина, мы можем судить только по выборке.

Если из генеральной совокупности произведена выборка достаточно большого объема, можно утверждать с вероятностью, сколь угодно близкой к 1, что при ограниченной дисперсии разность между выборочной средней арифметической $\bar{x}$ и генеральной средней арифметической $\bar{X}$ будет сколь угодно мала. Это дает основание утверждать, что среднеарифметическая выборки может быть использована как оценка генеральной средней. Одновременно среднеарифметическая выборки будет оценкой и математического ожидания M того теоретического распределения, которому подчиняется в генеральной совокупности случайная величина.

Аналогично можно утверждать, что среднеквадратическое отклонение выборки есть оценка среднеквадратического отклонения генеральной совокупности и среднеквадратического отклонения того распределения, которому подчиняется в генеральной совокупности случайная величина, но с некоторой поправкой: $\sigma\sqrt{n/(n-1)}$ выборки есть оценка σ генеральной совокупности.

При больших n отношение $n/(n-1)$ близко к 1 и может не учитываться.

Приводимые в § 6 и 7 математические положения наряду с некоторыми другими выражают так называемый закон больших чисел. Этот закон состоит в том, что совместное действие большого числа независимых случайных факторов приводит к тому, что результат становится мало зависящим от случайностей.

Закон больших чисел утверждает, что если в отдельном испытании результат сильно зависит от случая, то при достаточно большом количестве испытаний средний результат становится практически не зависимым от него. Именно этот закон дает возможность по выборке оценивать характеристики генеральной совокупности.

8. Правило «трех сигм»

Каждая выборка отражает генеральную совокупность с некоторой ошибкой. Если выборка не случайна и на ее формирование оказал воздействие какой-то фактор, то имеет место систематическая ошибка. Если выборка близка к случайной, то можно считать, что ошибки в ней случайные.

Для оценки характеристик распределения признака в генеральной совокупности и того теоретического распределения, к которому оно приближается, как мы видели, используются данные выборочного распределения. Но случайности выборки могут повлиять на точность этих оценок. Как определить степень этих ошибок, степень возможных случайных отклонений выборочных данных от характеристик теоретического генерального распределения? Существует теорема (неравенство Чебышева), также входящая в систему теорем закона больших чисел, следствием которой является так называемое *правило «трех сигм»*. Оно заключается в том, что с весьма большой вероятностью ($p=0{,}89$) можно утверждать, что отклонение случайной величины от ее математического ожидания будет по абсолютной величине меньше трехкратного квадратического отклонения, т. е. меньше 3 σ. Распределение случайной величины может быть каким угодно, но должно иметь математическое ожидание и ограниченную дисперсию. В случае, если случайная величина распределена по нормальному закону, вероятность повышается до 0,997.

Это правило позволяет утверждать, что при реализации случайной величины в выборке те значения x, которые отличаются более чем на 3 σ от математического ожидания M, будут появляться очень редко. Вероятность появления такого значения менее 0,89 (если случайная величина распределена как нормальная — менее 0,997).

В некоторых случаях вариационный ряд может содержать единичные, резко отличающиеся от остальных значения (или очень маленькие или слишком большие). Они могут оказать сильное влияние при подсчете средней арифметической и среднего квадратического отклонения. Вместе с тем интуитивно ясно, что эти значения варианта попали в вариационный ряд из какой-то другой генеральной совокупности, являются чужеродными ему. Они нарушают однородность вариационного ряда, его нормальное распределение. Их следует исключить, они являются «выбросом». Для определения «выброса» чаще всего прибегают к правилу «трех сигм». Отбрасывают те крайние значения, которые отстоят от среднеарифметического более чем на 3 σ (σ оценивается по данным выборки). При этом вероятность отбросить значение, принадлежащее однородному вариационному ряду, не более 0,01, если признак распределен приблизительно по закону нормального распределения. Если распределение не известно, эта вероятность оказывается не более 0,11. Но правило это применимо при достаточно большой выборке, не менее 50 единиц.

После исключения крайних значений среднее арифметическое значение и среднеквадратическое отклонение следует пересчитать.

9. Средняя ошибка выборки

В предыдущем разделе была определена вероятность того или иного отклонения случайной величины в эмпирическом распределении (выборке) от ее теоретического математического ожидания. Зададимся теперь вопросом, какова вероятность того или иного отклонения выборочной средней арифметической $\bar{x}$ от генеральной средней арифметической $\bar{X}$. Для этого введем понятие *средней ошибки выборки.*

Средняя ошибка выборки μ определяется как отношение среднего квадратического отклонения σ к квадратному корню из объема выборки n:

$$\mu = \sigma/\sqrt{n}.$$

При малых ($n<25$) выборках вводится, как мы видели, поправка для вычисления средней квадратической.

Тогда средняя ошибка малой выборки будет равна

$$\mu_{м.в} = \sigma/\sqrt{n-1}.$$

Выборочная средняя является оценкой средней генеральной совокупности, но эта генеральная средняя $\bar{X}$ может быть оценена выборочным методом только с какой-то ошибкой.

Таблица 9

n	t	n	t
4	3,2	9—10	2,3
5	2,8	11—14	2,2
6	2,6	15—26	2,1
7—8	2,4	27—200	2,0
		Свыше 200	1,96

Теория устанавливает вероятность $S(t)$ отклонения выборочной средней $\bar{x}$ от генеральной средней нормального распределения $\bar{X}$ на величину $t\mu$. Она определяет зависимость $S(t)$ от t при различных n. Это так называемое распределение Стьюдента для величины $S(t)$. Для каждого t высчитана величина $S(t)$ при соответствующих n.

При $S(t)=0{,}975$ n и t зависят друг от друга следующим образом (см. табл. 9; каждому n соответствует т. н. число степеней свободы, равное $n-1$).

10. Доверительный интервал

Зададимся определенным доверительным уровнем (0,95; вероятность совершить ошибку при нем равна $1-0{,}95=0{,}05$). Каково же должно быть t, чтобы, отложив на числовой оси по обе стороны от арифметической средней $\bar{x}$ выборки t раз по μ (т. е. $t\mu$), получить такой отрезок, в который генеральная средняя $\bar{X}$ попадет с вероятностью в 0,95?

[9] Таблица составлена по: *ван дер Варден Б. Л.* Математическая статистика. М., 1960. С. 410. Табл. 7.

Величина $\bar{x}$ может оказаться больше $\bar{X}$ на величину, большую $t\mu$, с вероятностью $1-S(t)$.

Аналогично $\bar{x}$ может оказаться меньшим $\bar{X}$ на величину, большую, чем $t\mu$, с вероятностью $1-S(t)$. Таким образом, x может оказаться одновременно или большим или меньшим X на величину, большую, чем $t\mu$, с вероятностью $2\,[1-S(t)]$ (по теореме сложения вероятностей). Задавшись доверительным уровнем 0,95, мы получаем

$$2[1-S(t)]=1-0{,}95;$$

откуда

$$S(t)=1-\frac{1-0{,}95}{2}=0{,}975.$$

По табл. 9 находим, что при достаточно большой выборке ($n>27$) при $S(t)=0{,}975$ $t\approx 2$.

Следовательно, если на числовой оси мы отложим от $\bar{x}$ в обе стороны отрезки, равные 2μ, то получим так называемый *доверительный интервал* (x', x''), в котором с вероятностью 0,95 будет заключена генеральная средняя $\bar{X}$.

Если выборка мала ($n<27$), то нужно по обе стороны от $\bar{x}$ на числовой оси откладывать $t\mu$. При этом t следует брать соответствующим объему выборки n.

Доверительный интервал является мощным статистическим средством при анализе выборочных данных. Он показывает, в каких пределах вокруг выборочной средней может быть заключена генеральная средняя признака. Чем больше выборка, тем интервал уже, и чем большим мы зададимся доверительным уровнем, тем интервал шире.

11. Приблизительное определение среднего квадратического отклонения

Знание характера распределения признака облегчает приблизительное вычисление характеристик его эмпирического распределения. Если есть основания считать какой-либо признак распределенным хотя бы примерно по нормальному закону и если выборка не очень мала, тогда $\sigma\approx 1{,}25Л$ ($Л$ — среднее линейное абсолютное отклонение).

Пример 9.

По данным примера 5 определим приближенное среднее квадратическое отклонение (табл. 10).

Таблица 10

x	w	$\lvert x-\bar{x}\rvert$	$\lvert x-\bar{x}\rvert w$
4,0	0,02	0,55	0,0110
4,1	0,03	0,45	0,0135
4,2	0,06	0,35	0,0210
4,3	0,09	0,25	0,0225
4,4	0,11	0,15	0,0165
4,5	0,16	0,05	0,0080
4,6	0,18	0,05	0,0090
4,7	0,16	0,15	0,0240
4,8	0,11	0,25	0,0275
4,9	0,05	0,35	0,0175
5,0	0,03	0,45	0,0135
Σ	1,00		0,1840

Таблица 11

n	γ	n	γ
2	0,89	12	0,31
3	0,59	14	0,29
4	0,49	16	0,28
5	0,43	18	0,27
6	0,40	20	0,27
7	0,37	30	0,25
8	0,35	40	0,23
9	0,34	50	0,22
10	0,33	100 и выше	0,20

Отсюда $\sigma \approx 1{,}25 \cdot 0{,}184 \approx 0{,}23$.

Найденное значение σ мало отличается от значения σ, определенного в примере 5 даже при некоторой асимметрии распределения, отличающего его от нормального.

Среднее квадратическое отклонение может быть грубо определено в выборке с примерно нормальным распределением по вариационному размаху:

$$\sigma \approx (x_{\max} - x_{\min})\gamma,$$

где γ — число, которое можно найти по табл. 11[10]; n — объем выборки.

Пример 10.

По данным примера 5 определим приближенно среднее квадратическое отклонение

$$\sigma = (5{,}0 - 4{,}0)\,0{,}2 = 0{,}2.$$

Найденное значение также мало отличается от σ, определенного в примере 5.

12. Определение достаточного объема выборки

Может возникнуть задача определения необходимого объема выборки, при котором с принятым доверительным уровнем можно утверждать, что выборочная средняя какой-то величины не отклонится от генеральной совокупности более чем на величину Δ. Доверительный уровень берется, как обычно, 0,95.

Мы знаем из распределения Стьюдента, что вероятность отклонения средней выборочной от средней генеральной совокупности не более чем на $1{,}96\,\mu \approx 2\mu$ равна 0,95. Следовательно,

$$\Delta = 2\mu = 2\sigma/\sqrt{n};\quad n = 4\sigma^2/\Delta^2.$$

Но величина σ нам неизвестна. Она может быть примерно определена по небольшой предварительной выборке. Допустимо грубое определение по вариационному размаху.

Пример 11.

Сколько нужно иметь однотипных монет, чтобы получить среднюю их массу с точностью 0,03 г? Предварительно взятые 20 монет показали разность масс самой легкой и самой тяжелой монет, равную 0,8 г. Доверительный уровень принимается равным 0,95.

Имеем

$$\sigma = 0{,}8 \cdot 0{,}27 = 0{,}216;\quad n = \frac{4\sigma^2}{\Delta^2} = \frac{4 \cdot 0{,}216^2}{0{,}03^2} \approx 207.$$

Это значит, что надо иметь не менее 207 монет, чтобы с требуемой точностью установить средний вес монет.

13. Сравнение двух выборок

Перед нами две независимые друг от друга выборки объемом n_x и n_y. В них даны значения одного и того же признака, собранные в вариационные ряды (x и y). Получены два разных средних арифметических. Как определить, существенно ли расхождение этих средних? Формулируем нулевую гипотезу: расхождение между двумя средними несущественно и обе выборки отражают одну и ту же генеральную среднюю, т. е. $\bar{X} = \bar{Y}$, где $\bar{X}$ и $\bar{Y}$ — генеральные средние.

Проверить эту нулевую гипотезу можно с помощью доверительных интервалов. По $\bar{x}$ и σ_x определим доверительный интервал для $\bar{X}$ и аналогично по $\bar{y}$ и σ_y определим доверительный интервал для $\bar{Y}$. Если на числовой оси они перекрывают друг друга, то нулевая гипотеза не опровергается, а если не перекрывают, то нулевая гипотеза опровергается.

[10] См.: *Снедекор Д. У.* Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., 1961. С. 54. Табл. 10.

Более предпочтительным является другой путь проверки нулевой гипотезы. Подсчитывается средняя ошибка разности двух выборочных средних:

$$\mu_\Delta = \sqrt{\mu_x^2 + \mu_y^2} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}.$$

Вычислим

$$t = |\bar{x} - \bar{y}| / \mu_\Delta.$$

При доверительном уровне 0,95 нулевая гипотеза отвергается, если $t > 1{,}96 \approx 2$.

Если нулевая гипотеза опровергнута, значит расхождение между двумя выборочными средними существенно, т. е. эти две выборки отражают разные генеральные совокупности с разными генеральными средними: $\bar{X} \neq \bar{Y}$.

Пример 12.

Дана выборка кирпичей из одного комплекса на золотоордынском городище, у которых была замерена толщина x (см. табл. 12, где x — значение признака, в данном случае толщина кирпича в см, m_x — частота значения признака, w_x — частость значения признака).

Таблица 12

x	m_x	w_x	xw_x	$\lvert x-\bar{x}\rvert$	$(x-\bar{x})^2 w_x$
3,0	10	0,06	0,18	1,28	0,0983
3,5	28	0,16	0,56	0,78	0,0973
4,0	40	0,23	0,92	0,28	0,0180
4,5	53	0,30	1,35	0,22	0,0145
5,0	39	0,22	1,10	0,72	0,1140
5,5	5	0,03	0,17	1,22	0,0446
Σ	175		4,28		0,3867

По этим данным определяем: $\sigma_x = 0{,}62$; $\mu_x = 0{,}047$.

Дана также выборка кирпичей из другого комплекса, у которых тоже была замерена толщина (см. табл. 13, где y — значение признака, в данном случае толщина кирпича в см, m_y — частота значения этого признака, w_y — частость значения признака).

По этим данным определяем: $\sigma_y = 0{,}79$; $\mu_y = 0{,}059$.

Требуется определить, существенно или нет различие между двумя средними этих выборок, т. е. между $\bar{x} = 4{,}28$ и $\bar{y} = 4{,}48$. Для этого подсчитаем среднюю ошибку разности двух выборочных средних:

$$\mu_\Delta = \sqrt{\mu_x^2 + \mu_y^2} = \sqrt{0{,}047^2 + 0{,}059^2} \approx 0{,}076;$$

$$t = \frac{\bar{y} - \bar{x}}{\mu_\Delta} = \frac{4{,}48 - 4{,}28}{0{,}076} = 2{,}63 > 2.$$

С уверенностью в 0,95 можем считать, что различие между двумя выборочными средними существенно: второй комплекс имел кирпичи в среднем более толстые, чем первый.

Таблица 13

y	m_y	w_y	yw_y	$\lvert y-\bar{y}\rvert$	$(y-\bar{y})^2 w_y$
3,0	15	0,08	0,24	1,48	0,1752
3,5	18	0,10	0,35	0,98	0,0960
4,0	28	0,16	0,64	0,48	0,0369
4,5	40	0,23	1,04	0,02	0,0001
5,0	48	0,27	1,35	0,52	0,0730
5,5	15	0,08	0,44	1,02	0,0832
6,0	13	0,07	0,42	1,52	0,1617
Σ	177		4,48		0,6261

Если объемы выборок n_x и n_y сильно отличаются друг от друга, то предпочтительнее среднюю ошибку разности вычислять по формуле

$$\mu_\Delta = \sqrt{\frac{n_x}{n_y}\mu_x^2 + \frac{n_y}{n_x}\mu_y^2}.$$

Этот критерий применяется тогда, когда выборка содержит не меньше 25 элементов каждая. Если $n < 25$, то применяется критерий Стьюдента для малых выборок.

Определяется средняя ошибка разности двух малых выборок с объемами n_x и n_y:

$$\mu_\Delta = \sqrt{\frac{\left[\sum_{i=1}^{l_x}(x_i - \bar{x})^2 + \sum_{i=1}^{l_y}(y_i - \bar{y})^2\right](n_x + n_y)}{(n_x + n_y - 2)\, n_x n_y}}$$

или, что то же самое,

$$\mu_\Delta = \sqrt{\left[\left(\sum_{i=1}^{l_x} x_i^2 - \frac{\left(\sum_{i=1}^{l_x} x_i\right)^2}{n_x}\right) + \left(\sum_{i=1}^{l_y} y_i^2 - \frac{\left(\sum_{i=1}^{l_y} y_i\right)^2}{n_y}\right)\right] \times} \times \sqrt{\frac{n_x + n_y}{(n_x + n_y - 2)\, n_x n_y}},$$

где $\bar{x}$ — среднее арифметическое первой выборки; $\bar{y}$ — среднее арифметическое второй выборки; n_x — объем первой выборки; n_y — объем второй выборки; l_x — число значений признака в первой выборке; l_y — число значений признака во второй выборке.

Вычисляется t:

$$t = \frac{|\bar{x} - \bar{y}|}{\mu_\Delta}.$$

Мы выше приводили табличные значения t для разного числа степеней свобод (об этом подробнее см. гл. I, § 16, в данном случае оно определяется $n_x + n_y - 2$). Если получено t больше этого значения для соответствующей степени свободы, то гипотеза опровергается с доверительным уровнем 0,95, т. е. вероятностью совершить ошибку 0,05.

Строго говоря, этот критерий применяется при предположении, что обе выборки отражают одну σ генеральной совокупности. При разных выборочных σ следует брать предельное значение для t несколько большим, так как расчет числа степеней свободы иной и приводит к числу меньшему, чем $n_x + n_y - 2$. Но если выборки не очень малы и выборочные средние квадратические отклонения не очень сильно отличаются друг от друга, то этим можно пренебречь.

Если объемы выборок равны, может быть применен следующий критерий, не требующий больших объемов вычислений, связанных с определением μ. Определяется для обеих выборок вариационный размах P, т. е. разность максимального и минимального значений признака. Далее определяется статистика:

$$u = \frac{2\,|\,\bar{x} - \bar{y}\,|}{P_x + P_y}$$

и сравнивается с табличным значением для доверительного уровня 0,95 (табл. 14) [11].

[11] Таблица взята из: *Снедекор Д. У.* Указ. соч. С. 117. Табл. 35.

Если u превышает табличное значение для соответствующего объема выборки, то нулевая гипотеза о том, что обе выборки взяты из генеральных совокупностей с одним средним арифметическим (т. е. о несущественности различия двух средних), отвергается с доверительным уровнем 0,95.

Таблица 14

n	u	n	u	n	u
3	1,272	9	0,334	15	0,216
4	0,813	10	0,304	16	0,205
5	0,613	11	0,280	17	0,195
6	0,499	12	0,260	18	0,187
7	0,426	13	0,243	19	0,179
8	0,373	14	0,228	20	0,172

Описанные выше процедуры построения доверительного интервала и сравнения двух выборок с использованием распределения Стьюдента рассчитаны на те случаи, когда признак по своим значениям распределен хотя бы приближенно по нормальному закону.

При достаточно больших выборках характер распределения признака вообще становится несущественным при применении этих процедур, так как в типичной ситуации при достаточно больших объемах выборки выборочная среднеарифметическая приобретает свойства распределенной по нормальному закону случайной величины. Таким образом, требование о приближенно нормальном распределении признака возникает при малых выборках.

14. Критерий Вилкоксона

Существуют непараметрические критерии, например критерий Вилкоксона для оценки существенности расхождения средних в двух выборках. Они удобны тем, что не нужно много считать. Но главное их достоинство в том, что они могут быть применены тогда, когда характер распределения нам неизвестен, а выборка мала. Основной недостаток непараметрических критериев — их малая мощность, т. е. существует сравнительно большой риск ошибки II рода.

Все значения $\dot{x}$ и $\dot{y}$ выстраиваются в порядке их возрастания в один ряд независимо от того, из какой выборки они взяты. Около каждого значения пишется, из какой выборки оно происходит. Затем подсчитывается число инверсий U для той выборки, у которой среднее арифметическое меньше, т. е. количество значений из другой выборки, которое предшествует каждому значению признака из первой выборки. Это число сопоставляется со специальной таблицей, и если оно меньше табличного, то нулевая гипотеза отвергается на том доверительном уровне, для которого составлена эта таблица. Критерий Вилкоксона удобен и дает хорошие результаты для малых выборок. Критические значения числа инверсий U для различных объемов выборок n_1 и n_2 при доверительном уровне 0,95 приведены в табл. 15.[12]

Таблица 15

n_1	n_2	U	n_1	n_2	U
2	5	0	5	5	4
2	6	0	5	6	5
2	8	1	5	7	6
2	9	1	5	8	8
2	10	2	5	9	9
3	3	0	5	10	11
3	4	0	6	6	7
3	5	1	6	7	8
3	6	2	6	8	10
3	7	2	6	9	12
3	8	3	6	10	14
3	9	4	7	7	11
3	10	4	7	8	13
4	4	1	7	9	15
4	5	2	7	10	17
4	6	3	8	8	15
4	7	4	8	9	18
4	8	5	8	10	20
4	9	6	9	9	21
4	10	7	9	10	24
			10	10	27

Пример 13.

Результаты замеров высоты кувшинов в двух выборках приведены в табл. 16, где x — значение признака в первой выборке, m_x — его частота, y — значение признака во второй выборке, m_y — его частота. Требуется установить, существенно ли различие между средними этих двух выборок.

[12] Таблица составлена по: *ван дер Варден Б. Л.* Указ. соч. С. 418. Табл. 10.

Таблица 16

x	m_x	x^2	y	m_y	y^2
10,1	1	102,01	7,4	1	54,76
12,8	1	163,84	9,0	1	81,00
15,0	1	225,00	9,4	1	88,36
16,2	1	262,44	9,9	1	98,01
16,5	1	272,25	10,5	1	110,25
18,9	1	357,21	10,7	1	114,49
19,3	1	372,49	11,3	1	127,69
20,1	1	404,01	13,9	1	193,21
			16,1	1	259,21
			20,0	1	400,00
Σ 128,9		2159,25	118,2		1526,98

Используя данные табл. 16, находим:

$$\bar{x}=16{,}11; \qquad \bar{y}=11{,}82;$$

$$\left(\sum_{i=1}^{l_x} x_i\right)^2=16615{,}21; \qquad \left(\sum_{i=1}^{l_y} y_i\right)^2=13971{,}24;$$

$$\mu_\Delta=\sqrt{2159{,}25-\frac{16615{,}21}{8}+1526{,}98-\frac{13971{,}24}{10}}\times$$

$$\times\sqrt{\frac{18}{16\cdot 8\cdot 10}}\approx 1{,}73\cdot$$

$$t=\frac{\bar{x}-\bar{y}}{\mu_\Delta}=\frac{16{,}11-11{,}82}{1{,}73}=2{,}48>2{,}1.$$

Число степеней свободы $n_x+n_y-2=16$. При этом числе степеней свободы для доверительного уровня $S(t)=0{,}95$ t не должно превышать 2,1. Иначе нулевую гипотезу, состоящую в том, что между двумя выборочными средними нет существенной разницы, следует опровергнуть. В данном случае нулевая гипотеза опровергается и заключается с вероятностью ошибиться 0,05, что между двумя выборочными средними есть существенное различие.

Проверим ту же нулевую гипотезу на тех же выборках с помощью критерия Вилкоксона.

Располагаем все значения x и y в порядке возрастания:

7,4	9,0	9,4	9,9	10,1	10,5	10,7	11,3	12,8
(y)	(y)	(y)	(y)	(x)	(y)	(y)	(y)	(x)
13,9	15,0	16,1	16,2	16,5	18,9	19,3	20,0	20,1
(y)	(x)	(y)	(x)	(x)	(x)	(x)	(y)	(x)

Количество инверсий U для y равно $1+1+1+2+3+7=15<20$, критического значения U для выборок объемом 8 и 10.

Следовательно, вероятность того, что нулевая гипотеза правильна, меньше 0,05, т. е. с доверительным уровнем 0,95 нулевая гипотеза может считаться отвергнутой.

15. Сравнение коэффициентов вариации двух выборок

Задача, аналогичная сравнению средних арифметических двух разных выборок, возникает тогда, когда мы хотим сравнить коэффициент вариации двух разных выборок V_x и V_y.

Если принять

$$t=\frac{V_x-V_y}{\sqrt{V_x^2/2n_x+V_y^2/2n_y}},$$

то при $t>2$ различие коэффициентов вариации считается существенным (с доверительным уровнем 0,95).

В двух выборках, даже если их средние арифметические не различаются существенно, могут быть существенно разные коэффициенты вариации. Изменение степени вариабельности, разброса какого-либо признака свидетельствует о важных изменениях в производстве и потреблении вещи или в сооружении того или иного объекта. Обычно увеличение степени вариабельности, увеличение даже при сохранении $\bar{x}$, т. е. стандарта, нормативного, типического значения признака, свидетельствует о том, что мастера и потребители стали меньше внимания уделять этому признаку на предмете. Это наблюдается при смене типов, при сменах культур и тому подобных ситуациях: увеличение V иногда предшествует существенному изменению признака, т. е. изменению его $\bar{x}$.

16. Сравнение эмпирического (выборочного) и теоретического распределений

Большинство археологических мерных признаков для каких-либо однородных объектов, представляющих их размер, объем, вес и т. п., колеблется вокруг своей средней, подчиняясь приближенно закону нормального распределения. В этом случае вариационный ряд, отражающий это распределение, — одновершинный и симметричный. Многие археологические временные ряды тоже имеют характер нормального распределения.

Можно вычислить среднюю арифметическую и среднее квадратическое отклонение этих вариационных рядов. По этим данным можно теоретически построить распределение случайной величины, подчиненное нормальному распределению, т. е. вычислить теоретически, исходя из формулы закона нормального распределения, вероятности тех значений признака, которые представлены в вариационном ряду. Действительные, эмпирические, полученные путем наблюдений над признаком частости значений признака могут отклоняться от этих теоретически вычисленных вероятностей. Как проверить, существенны ли эти отклонения или они — лишь игра случая, так как наблюдения над признаком производились по выборке, т. е. по части генеральной совокупности.

Чтобы найти значения вероятностей p (теоретических частостей) для каждого значения признака x (случайной величины), следует вычислить среднюю арифметическую вариационного ряда $\bar{x}$. Это будет оценкой математического ожидания M теоретического распределения. Затем вычислить его среднее квадратическое отклонение σ. Это будет оценкой среднего квадратического отклонения теоретического распределения.

Затем вычислить $z=(x-\bar{x})/\sigma$ и по нему в таблице 1 найти $f(z)$. Далее, чтобы вычислить теоретические частоты $\tilde{m}$ для каждого значения варианта, нужно умножить $f(z)$ на общее число объектов n с коэффициентом α/σ, где α — размер интервала. В итоге получаем два ряда частот: действительные, реальные частоты m значений признака в выборке и теоретические $\tilde{m}$, вычисленные при гипотезе, что значения признака распределены по нормальному закону. Формулируем нулевую гипотезу, заключающуюся в утверждении, что между распределением значений признака в вариационном ряду (выборке) и теоретическим распределением по нормальному закону

с теми математическим ожиданием и средним квадратическим отклонением, которые были оценены по вариационному ряду (выборке), нет существенного (значимого) различия.

Для проверки этой нулевой гипотезы используется так называемый *критерий согласия* Пирсона:

$$\chi^2=\sum_{i=1}^{l}(m_i-\tilde{m}_i)^2/\tilde{m}_i,$$

где m_i — эмпирические, а $\tilde{m}_i$ — теоретически вычисленные частоты, а l — число значений признаков. Если значение χ^2 превосходит табличное (см. табл. II) для соответствующего доверительного уровня и соответствующего числа степеней свободы, то нулевая гипотеза о том, что значения признака распределены по нормальному закону, отвергается. Число степеней свободы вычисляется как общее число значений признака минус 1 (l—1). Это число указывает, по какому количеству значений признака происходит независимое колебание его частостей. Так как сумма частостей всегда равна 1, то независимо могут изменяться частости всех значений признаков, кроме одного. Из l—1 следует вычесть 2 — число параметров теоретического распределения, если они были определены по выборке.

Существенным ограничением применения критерия χ^2 является то, что он может быть применен только тогда, когда объем выборки не слишком мал, и число степеней свободы не слишком мало. При этом частости не должны быть очень малыми. Для выполнения этих условий следует варианты с малыми частотами объединять с соседними или между собой.

Некоторые ученые считают, что чрезмерная осторожность в применении χ^2 является излишней: этот критерий может быть применен при $n\geqslant 10$, числе степеней свободы не меньше 9 и при ожидаемых частотах не менее 1. Подобные соображения имеют место и в других случаях использования этого критерия с числом степеней свободы большим 1 (см. гл. III, § 5, гл. IV, § 3) [13].

Возможен более простой, хоть и менее точный прием определения нормальности распределения. Сначала определяется среднее арифметическое $\bar{x}$ и среднее квадратическое отклонение σ признака. Затем проверяются следующие условия: 1) сумма частот всех значений от $\bar{x}$—0,3σ до $\bar{x}$+0,3σ должна составлять примерно 0,25 объема всей выборки;

2) сумма частот всех значений от $\bar{x}$—0,7σ до $\bar{x}$+0,7σ должна составлять примерно 0,5 всей выборки;

3) сумма частот всех значений от $\bar{x}$—1,1σ до $\bar{x}$+1,1σ должна составлять 0,75 объема всей выборки;

4) сумма частот всех значений от $\bar{x}$—3σ до $\bar{x}$+3σ должна составлять 0,998 объема всей выборки.

Если эти условия соблюдаются, то данное эмпирическое, выборочное распределение близко к нормальному.

[13] См.: *ван дер Варден Б. Л.* Указ. соч. С. 275.

17. Отклонение вариационного ряда эмпирического распределения от нормального распределения

Если вариационный ряд для какого-либо количественного признака не соответствует нормальному распределению, то это означает, на его формирование оказали влияние кроме множества мелких случайных факторов один или несколько факторов настолько сильно действующих, что они нарушили совокупное действие мелких. Следует помнить, что отклонения от нормального распределения могут быть вызваны и неслучайностью самой выборки.

Отклонения вариационного ряда от нормального распределения могут быть связаны с его асимметрией. Асимметрия выражается коэффициентом асимметрии, а неслучайность ее может быть проверена с помощью критерия согласия χ^2. Но и без этой проверки можно считать, что асимметрия тогда значима и не объяснима случайностью выборки (с доверительным уровнем 0,95), когда

$$|K_a'|>\frac{2}{\sqrt{n}}.$$

В примере 5

$$|K_a'|=0{,}22>\frac{2}{\sqrt{222}},$$

следовательно, асимметрия не может быть признана случайной. Уже этого достаточно, чтобы отвергнуть предположение о соответствии распределения случайной величины — ширины горла сосудов — нормальному распределению. Очевидно, перед нами выборка из неоднородной генеральной совокупности. Можно предположить, что в

ней смешаны сосуды с большим и меньшим в среднем размером горла. Но эта разница незначительна и так просто не видна. Требуется проверить гипотезу о нормальном распределении этой величины и опровергнуть ее, чтобы прийти к такому заключению. Дальше уже дело собственно археологического исследования — признать и объяснить выявленную неоднородность казавшейся ранее однородной совокупности сосудов, дать ей какое-либо историческое толкование или игнорировать ее. Нужно выяснить, соответствуют ли эти различия хронологической группировке сосудов, если сосуды происходят из разных слоев или разных памятников. Может оказаться, что взятые из одного слоя или памятника сосуды дадут более однородные выборки.

Могут быть приведены и другие примеры. Так, наблюдаемая левосторонняя асимметрия распределения веса монет часто объясняется тем, что тяжеловесные монеты изымались из обращения. Асимметрия в динамических рядах, показывающих распространение того или иного вида предмета во времени, объясняется тем, что этот вид медленно входил в моду и быстро из нее выходил или наоборот. Иногда левосторонняя асимметрия объясняется тем, что бытование предмета было прервано какой-то катастрофой, завоеванием и т. п. (рис. 9, *а*).

Отклонения вариационного ряда от нормального распределения могут быть связаны с его эксцессом. Эксцесс выражается коэффициентом эксцесса, а неслучайность его проверяется критерием согласия χ^2. В динамическом ряду положительный коэффициент эксцесса может быть вызван быстрым вхождением вида предмета в моду и быстрым его исчезновением, или наоборот, отрицательный коэффициент эксцесса может быть вызван долговременным бытованием какого-либо вида во времени (рис. 9, *б*).

Значительный положительный коэффициент эксцесса может быть вызван массовым завозом какого-либо изделия; до и после этого изделие поступало в небольших количествах, случайно.

Отклонение вариационного ряда от нормального распределения иногда выражается в появлении зубца в какой-то части полигона распределения. В динамических рядах такие зубцы или понижения в полигонах распределения могут быть вызваны какими-либо политическими событиями, нарушившими обычный ритм производства (или привоза) исследуемых предметов в данном месте или, наоборот, какими-либо обстоятельствами, стимулировавшими это производство или привоз (рис. 9, *в*).

Как определить, появился этот зубец случайно и это лишь случайность выборки или же он отражает действительное нарушение закона нормального распределения?

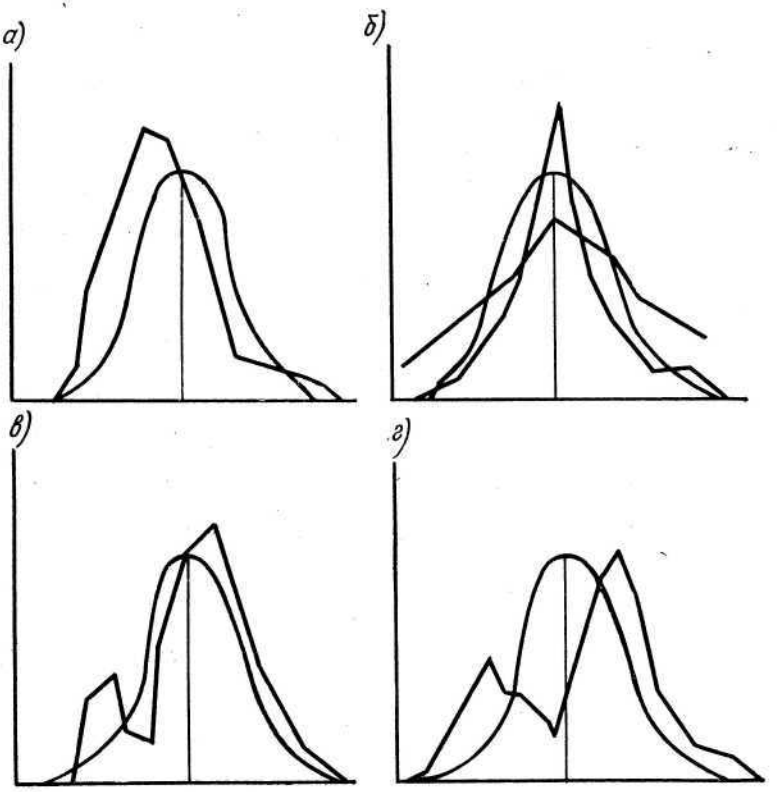


Рис. 9. Отклонения эмпирического распределения от нормального:

а — правосторонняя асимметрия; *б* — положительный и отрицательный эксцессы; *в* — «зубец» в полигоне; *г* — многовершинность полигона

Следует подсчитать по всему вариационному ряду среднее арифметическое и среднее квадратическое отклонение и применить затем критерий согласия χ^2 к тому отрезку вариационного ряда, который содержит эту аномалию, соответственно определив число степеней свободы.

Если зубцы большие, то отклонения вариационного ряда от нормального распределения могут выразиться в

многовершинности полигона (рис. 9, *г*). Это свидетельствует о неоднородности признака, т. е. о том, что в выборке смешались несколько генеральных совокупностей, каждая со своим математическим ожиданием, что и находит выражение в вершинах полигона. В динамических рядах такая двухвершинность может объясняться временным прекращением (производства или привоза) исследуемых предметов в данном месте.

Если эта многовершинность выражена неясно, то остается сомнение, не вызвана ли она случайностями выборки. Тогда следует проверить нулевую гипотезу о соответствии выборки нормальному распределению значений признака. В случае ее опровержения следует прийти к заключению, что в выборке смешаны объекты с разными стандартами.

Анализ временного ряда, который показывает распределение того или иного вида или варианта предмета в культурном слое, может дать представление о нарушениях этого слоя, его перемещениях. Симметричное, близкое к нормальному, распределение видов предметов по уровням культурного слоя может свидетельствовать о равномерности отложения слоя и об отсутствии в нем значительных нарушений и смещений. Вместе с тем нормальное распределение какого-либо вида предмета во времени свидетельствует о его равномерном и обычно местном производстве. Деформация этого распределения говорит о различного рода нарушениях и неравномерностях роста культурного слоя [14] или о сложном характере производства, привоза и бытования исследуемых вещей.

Пример 14.

Дано распределение 66 овальных с прямой прорезью кресал по ярусам Неревского раскопа в Новгороде (см. табл. 17, где x — номер яруса, m — частота кресал в ярусе, w — их частость) [15]. Мы допускаем, что ярусы имеют одинаковую толщину, соответствуют примерно одинаковым отрезкам времени и номер яруса равен высоте верхней границы яруса над материком в условных единицах, равных толщине яруса (α). Требуется установить, значимо или нет отклонение этого распределения от нормаль-

[14] См.: *Каменецкий И. С.* К теории слоя // СКМА. М., 1970.

[15] См.: *Колчин В. А.* Железообрабатывающее ремесло Новгорода Великого // МИА. М., Л., 1959. № 65. С. 102.

Таблица 17

x	m	w	xw	$\lvert x-\bar{x}\rvert$	$(x-\bar{x})^2\,w$	z	$f(z)$	$\tilde{m}$	$\approx\tilde{m}$	$(m-\tilde{m})^2$	$\frac{(m-\tilde{m})^2}{\tilde{m}}$
3	1	0,01	0,03	5,68	0,3226	2,27	0,03	0,79	1	0	0,00
4	2	0,03	0,12	4,68	0,6571	1,87	0,07	1,85	2	0	0,00
5	5	0,08	0,40	3,68	1,0834	1,47	0,13	3,43	3	4	1,33
6	6	0,09	0,54	2,68	0,6464	1,07	0,22	5,81	6	0	0,00
7	3	0,04	0,28	1,68	0,1129	0,67	0,32	8,45	8	25	3,12
8	14	0,21	1,68	0,68	0,0971	0,27	0,38	10,03	10	16	1,60
9	14	0,21	1,89	0,32	0,0215	0,13	0,40	10,56	11	9	0,82
10	4	0,06	0,60	1,32	0,1045	0,53	0,35	9,24	9	25	2,78
11	5	0,08	0,88	2,32	0,4306	0,93	0,26	6,86	7	4	0,57
12	6	0,09	1,08	3,32	0,9920	1,33	0,16	4,22	4	4	1,00
13	5	0,08	1,04	4,32	1,4930	1,73	0,09	2,38	2	9	4,50
14	1	0,01	0,14	5,32	0,2830	2,13	0,04	1,06	1	0	0,00
Σ	66		8,68		6,2441						15,72

18. Превращение количественных признаков в качественные

В некоторых задачах, например при совместном исследовании нескольких признаков, из которых одни являются количественными, а другие — качественными, или при выявлении связи между какими-то значениями количественного признака и значениями другого признака, требуется так разбить значения количественного признака на интервалы, чтобы они наилучшим образом отражали сущность явления. Эти полученные интервалы рассматриваются далее как ранжированные значения качественного признака.

Чтобы правильно разбить количественный признак на интервалы, следует построить полигон частот его значений и установить, однородный он или нет.

а) Признак однороден и его полигон близок к кривой нормального распределения. Его среднее квадратическое отклонение σ не слишком велико. Этот признак можно рассматривать как признак с одним значением.

б) Признак имеет большое среднее квадратическое отклонение. Его полигон оказывается пологим. Признак может иметь полигон, приближающийся к горизонтальной линии, т. е. все значения признака примерно одинаково часто встречаются. В таком случае следует брать равные интервалы.

в) Признак может иметь многовершинный полигон. Это значит, что признак смешанный. Нужно выделить интервалы, примерно соответствующие областям, расположенным на оси абсцисс около средних значений тех однородных признаков, которые составляют этот смешанный признак. Это можно сделать, проведя границы между интервалами по резким «понижениям» полигона. Интервалы могут быть неравными. Если «понижения» и «вершины» выражены не четко, целесообразнее разбивать значения количественного признака на равные интервалы независимо от полигона его распределения.

Следует помнить, что при замене количественных признаков качественными всегда происходит некоторая потеря информации. Но необходимость иметь при описании объектов однородные признаки заставляет мириться с этой потерей.

КАЧЕСТВЕННЫЕ ПРИЗНАКИ

Значения качественных признаков не могут быть представлены в виде вариационного ряда, они не сравнимы друг с другом по количественному своему выражению или составу элементов. Поэтому порядок значений неранжированных качественных признаков случаен, произволен, эти значения могут быть расположены в любой последовательности. Такие понятия, как средняя арифметическая или среднеквадратическое отклонение, не могут быть применимы к ряду значений качественных признаков.

При изучении какой-либо совокупности объектов подсчитывается количество объектов, имеющих то или иное значение качественного признака, т. е. частота m. При этом получают смысл требования обязательности и несовместимости значений этого признака. Каждый объект должен иметь одно и только одно значение признака. Если эти требования выполняются, то сумма частот признака оказывается равной количеству всех взятых для исследования объектов. Если количество объектов, обладающих данным значением признака, разделить на количество всех объектов, то получим частость w значения признака. Сумма частостей для данного количества объектов равняется 1 (100%).

Построение полигона или гистограммы частот, частостей или плотности распределения неранжированных качественных признаков, когда на оси абсцисс откладываются качественные признаки, не имеет смысла и не показывает особенностей ряда качественных признаков, так как их порядок произволен.

Для графической интерпретации распределения качественных признаков более удобны круговые диаграммы, у которых угол секторов соответствует частости значения признака.

Таким образом, в ряду значений качественного признака имеется только одна количественная характеристика — частота появления каждого значения признака, или частость как нормированная частота, т. е. частота, отнесенная к общему количеству наблюдений.

Номер значения признака	$1, 2, \ldots, i, \ldots, l$
Частота	$m_1, m_2, \ldots, m_i, \ldots, m_l$
Частость	$w_1, w_2, \ldots, w_i, \ldots, w_l$

$$\sum_{i=1}^{l} m_i = n, \quad \sum_{i=1}^{l} w_i = 1,$$

где n — объем выборки; l — количество значений признака.

1. Распределение частот каждого признака. Биномиальное распределение

В генеральной совокупности имеется какое-то количество объектов, описываемых различными значениями данного признака. Допустим, производится n независимых испытаний, состоящих в том, что из генеральной совокупности случайно извлекается какой-то объект. Известно, что вероятность появления каждый раз на этом объекте какого-то определенного значения признака (случайное событие) равна p, так как в генеральной совокупности доля объектов, обладающих именно этим значением, составляет p. Тогда вероятность того, что на извлеченном объекте не будет этого значения признака, равна $q=1-p$.

Допустим, что объект возвращается в совокупность или же совокупность настолько велика, что извлечение из нее объекта не отражается на оставшейся части совокупности. При n независимых испытаниях (они составляют выборку из n элементов) вероятность того, что интересующее нас значение появится ровно m раз, равна:

$$p_m = C_n^m p^m q^{n-m} = \frac{n!}{m!\,(n-m)!}\, p^m q^{n-m},$$

где C_n^m — число сочетаний из n элементов по m. Легко видеть, что эта вероятность равна m-му члену разложения бинома Ньютона $(p+q)^n$. А так как $p+q=1$, то сумма всех вероятностей

$$p_0 + p_1 + \ldots + p_n = 1.$$

Это согласуется со смыслом опыта и с теоремой сложения вероятностей: в этом опыте, т. е. в выборке объемом n, обязательно должен появиться признак или 0 раз, или 1 раз, или 2 раза, ..., или n раз. Эта полная система событий.

Будем считать частоту появления какого-либо значения признака случайной величиной, которая реализуется с той или иной вероятностью. Тогда эти частоты могут быть оценены исходя из *биномиального распределения*. Это значит, что в выборке из n элементов данное значение признака будет иметь вероятность:

появиться 0 раз, равную $C_n^0 p^0 q^{n-0} = q^n$;

появиться 1 раз, равную $C_n^1 p^1 q^{n-1} = npq^{n-1}$;

. . .

появиться i раз, равную $C_n^i p^i q^{n-i}$;

. . .

появиться n раз, равную $C_n^n p^n q^{n-n} = p^n$.

Это распределение симметрично только при $p=1/2$. Чем ближе p к 0 или к 1, тем больше график распределения будет сдвинут влево или вправо.

Также распределены и частости w.

Следует отметить, что так распределены не частоты ряда значений качественного признака $m_1, m_2, \ldots, m_i, \ldots, m_l$ и не его частости $w_1, w_2, \ldots, w_i, \ldots, w_l$. Рассматривается только одно значение признака (его рассматривают как отдельный альтернативный признак). Каждому значению признака (например, цвет бусин — красный, синий, зеленый) соответствует случайная величина — частота (частость) этого признака в выборке. В разных выборках из одной генеральной совокупности она реализуется по-разному. В одной выборке из 100 бус содержится 10 зеленых (10%), в другой — 9, в третьей — 11%. Произведя выборку из n археологических объектов, мы получим определенную частоту m для i-го значения признака. Но при другой выборке из той же генеральной совокупности в силу случайных обстоятельств, неизбежных при всякой выборке, мы могли бы получить другую частоту для того же значения признака. Случайная величина — частота данного значения признака в выборке — реализуется с той или иной вероятностью каждый раз, когда мы производим выборку. Эта случайная величина распределена по биномиальному закону. Аналогично дело обстоит и с частностью w.

Математическое ожидание этой случайной величины — частоты — равно

$$M = np,$$

а ее среднеквадратическое отклонение

$$\sigma = \sqrt{npq}.$$

Математическое ожидание частости равно p. Ее среднеквадратическое отклонение

$$\sigma = \sqrt{pq}.$$

При достаточно большом числе независимых испытаний можно с вероятностью, сколь угодно близкой к 1, утверждать, что частости появления событий сколь угодно мало будут отличаться от вероятности его появления в отдельном испытании. Таким образом, чем больше объем выборки n, тем ближе будет величина ω к вероятности p, т. е. доли признака в генеральной совокупности, и значительные отклонения w от p будут при этом маловероятны. Следовательно, w выборки является оценкой p генеральной совокупности.

2. Доверительный интервал для частостей

Пусть в генеральной совокупности из N объектов, описываемых данным признаком, A объектов имеет определенное значение «A» этого признака. Вероятность того, что при случайном извлечении из генеральной совокупности одного объекта на нем будет именно это значение признака, равна $p = A/N$.

Но генеральная совокупность для исследователя недоступна и, следовательно, величину p он знать не может. Он ее может только оценить по частости w. Эта оценка будет тем более точной, чем больше объектов содержит выборка.

Но так как выборка подвержена разного рода случайностям, то ω колеблется вокруг p в определенных пределах.

Мы должны как-то оценить степень этих возможных случайных колебаний. Для этого выбрав какой-то доверительный уровень, построим такой доверительный интервал, чтобы с определенной долей уверенности можно было бы утверждать, что вероятность p (т. е. доля признака в генеральной совокупности) находится внутри этого интервала.

Этот доверительный интервал подсчитывается следующим образом:

$$p',\ p'' = \frac{wn + \frac{1}{2}t^2 \pm t\sqrt{w(1-w)n + \frac{1}{4}t^2}}{n + t^2},$$

где t может быть определено при больших выборках по n и по избранному доверительному уровню из таблицы t-распределения Стьюдента (см. табл. 9). Значения p' и p'' будут границами доверительного интервала. Они располагаются тем более несимметрично относительно w, чем более p отличается от $1/2$. Вероятность того, что доля признака в генеральной совокупности является величиной, лежащей вне этих границ, мала и игнорируется исследователем.

Так определяется доверительный интервал для p. При доверительном уровне 0,95 и не слишком малом w или $1-w$ и при достаточно большом n (желательно, чтобы $nw(1-w) \geqslant 9$) $t \approx 2$. Доверительные интервалы для p по частотам удобно (а если указанные условия не выполняются, то необходимо) определять по специальным таблицам, составленным для разных доверительных уровней, в которых по n и m сразу определяются p' и p'' (см. табл. III [1]).

Если выборка велика, можно обойтись и без доверительных интервалов, так как они будут весьма узкими. Если выборка небольшая, доли следует давать в доверительных интервалах. В тех случаях, когда доверительные интервалы для разных объектов не перекрывают друг друга, можно говорить, что доли этих объектов в генеральной совокупности значимо различаются.

Пример 16.

Дано распределение красноглиняных ножек амфор в Танаисе в одном из раскопов по слоям [2]. Были подсчитаны доли этих ножек среди всех амфорных ножек по слоям и определены доверительные интервалы для этих долей. Все данные и результаты подсчетов представлены в табл. 19, где арабскими цифрами обозначены эллинистические слои, римскими цифрами — слои римской эпохи, m — число красноглиняных амфорных ножек в дан-

[1] Более обширную табл. см.: *Большов Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., 1965. Табл. 5.2.

[2] См.: *Деопик Д. В.* Керамический комплекс и культурный слой. С. 256. Табл. 5.1.

ном слое, n — число всех амфорных ножек в данном слое, w — доля красноглиняных ножек среди всех амфорных ножек в данном слое и p', p'' — доверительный интервал для этой доли [3].

Таблица 19

Слой	m	n	w	p', p''
VIII—VII	6	19	0,316	0,126—0,565
VI—V	16	45	0,355	0,214—0,502
IV—III	13	40	0,325	0,181—0,481
II—I	7	60	0,117	0,051—0,237
1—2	5	87	0,057	0,019—0,132
3—4	1	12	0,083	0,002—0,385

Требуется определить, менялась ли существенно с течением времени доля красноглиняных амфор в культурном слое Танаиса. Мы видим, что в эллинистических и первом римском слоях доля красноглиняных амфор дает перекрывающиеся доверительные интервалы и, следовательно, нельзя говорить об изменении этой доли на протяжении всего периода, но с III—IV слоя доля этих амфор резко возрастает по сравнению с первым римским слоем: нижняя граница доверительного интервала доли красноглиняных амфор в VI—V слоях чуть ниже верхней границы доверительного интервала для доли той же керамики в II—I слое. Но на протяжении всех остальных слоев римского периода доля красноглиняной керамики дает пересекающиеся доверительные интервалы и поэтому также нет оснований говорить о какой-то динамике доли этой керамики на протяжении этого времени.

3. Сравнение частостей двух выборок

Мы имеем две выборки объемом n_1 и n_2 и в них данный признак встретился с частотами m_1 и m_2 и частостями w_1 и w_2. Необходимо выяснить, существенно или несущественно расхождение этих частотей, т. е. узнать, отражают ли эти две выборки одну генеральную совокупность, в которой признак имеет одну вероятность p, или они отражают две совокупности, в которых признак имеет разные вероятности p_1 и p_2.

[3] Доверительные интервалы определены по: *Большов Л. Н., Смирнов Н. В.* Указ. соч. Табл. 5.2.

Это сравнение можно произвести, построив два доверительных интервала. Если они не перекрывают друг друга, то нулевая гипотеза о том, что обе выборки произведены из одной генеральной совокупности, отвергается. Таким способом мы сравнивали частоты красноглиняных амфор в разных слоях в примере 16.

Однако для оценки степени расхождения двух частостей предпочтительнее другой способ, использующий критерий χ^2. Он изложен в гл. III, § 4.

4. Определение достаточности объема выборки

Мы хотим определить, какой величины должна быть выборка для того, чтобы полученная опытная частость какого-либо признака правильно отражала (с данным доверительным уровнем) генеральную частость этого признака. Другими словами, как велико должно быть n, чтобы с вероятностью (доверительный уровень) 0,95 можно было сказать, что доля данного признака в совокупности (вероятность p) отличается от эмпирической частости не более, чем на величину Δ, которую мы условились считать несущественным отклонением. Теория устанавливает, что отклонение выборочного w от p при больших выборках не должно (с вероятностью 0,95) превышать $2\sqrt{pq/n}$. Каково же должно быть n, чтобы выполнялось неравенство

$$\Delta \leqslant 2\sqrt{\frac{pq}{n}}.$$

Ясно, что оно выполняется при $n \geqslant 4\,pq/\Delta^2$. Заменяя p и q их эмпирическими оценками w и $1-w$, получаем

$$n = \frac{4w(1-w)}{\Delta^2}.$$

Пример 17.

Рассматривается керамика из одного слоя раскопа с одинаковым характером его на всей вскрытой площади. Нужно определить для каждого черепка, к лепной или гончарной керамике он принадлежит. Сколько нужно просмотреть черепков, чтобы утверждать, что эмпирическая частость фрагментов лепной керамики в этом слое не отклонится от доли фрагментов лепной керамики в генеральной совокупности более чем на 0,02? Эмпирическая частость по просмотренным 100 фрагментам оказа-

лась равной 0,27. Доверительный уровень принят 0,95. Подставляя эти значения в формулу, получим

$$n = \frac{4 \cdot 0{,}27 \cdot 0{,}73}{0{,}02^2} = 1971.$$

Следовательно, для выполнения поставленных условий нужно просмотреть около 2000 черепков.

Если предварительно p нельзя оценить даже приближенно, то берут максимальное из возможных значений произведения

$$w(1-w) = 0{,}25.$$

Таким образом, заранее можно сказать, что если взять из генеральной совокупности черепков не менее

$$n = \frac{4 \cdot 0{,}25}{0{,}02^2} = 2500,$$

то это позволит установить с достаточной точностью долю лепной керамики.

ГЛАВА III

СВЯЗИ МЕЖДУ ПРИЗНАКАМИ И ИХ ЗНАЧЕНИЯМИ

Изучение связей между явлениями — одна из основных задач каждой науки. В археологии связи между явлениями, как правило, статистические. Статистика только устанавливает наличие связей, определяет их силу. Но истолкование смысла и причины возникновения связи — дело археологии как таковой.

В этой главе рассматриваются парные связи. Статистическая взаимозависимость, связь между двумя явлениями «А» и «В» (событиями, признаками, объектами) понимается как более вероятное осуществление «А» при условии, что осуществилось «В», чем без этого условия. При исследовании связей необходимо различать четыре случая: 1) связи между количественными признаками; 2) связи между качественными признаками, поддающимися ранжировке; 3) связи между качественными признаками, не поддающимися ранжировке; 4) связи между количественными и качественными признаками. В этом последнем случае необходимо: или качественным признакам придать вид количественных, т. е. как-то ранжировать и «оцифровать» их; или количественным признакам путем разбиения их на интервалы придать вид качественных ранжированных признаков. Обычно, когда одно исследование охватывает и качественные и количественные признаки, идут именно по этому пути.

Кроме того, следует различать:

а) исследование связей между одним каким-либо значением одного признака и одним каким-либо значением другого признака;

б) исследование связи между всеми значениями одного признака и всеми значениями другого признака, т. е. связи между признаками со множеством значений;

в) исследование связи одного признака с несколькими другими признаками.

Связь между всеми значениями одного признака и всеми значениями другого — это как бы «интегральная», обобщенная, связь «в целом». Если такая связь выявлена, это не значит еще, что между любым значением одного признака и любым значением другого есть взаимозависимость. Эта последняя — как бы «локальная» связь. «Локальные» связи между значениями признаков могут быть сильными, слабыми или вообще отсутствовать для каких-либо пар значений. В сумме же все эти связи (и сильные и слабые) определяют, есть ли между признаками связь и какой она силы.

В ряде случаев сама природа признаков должна указывать на то, какой признак является в выявленной взаимосвязи независимым, т. е. оказывающим воздействие, *факторным,* а какой зависящим от этого факторного признака, т. е. *результативным.* Это определяет «направление связи».

1. Связь между количественными признаками (случай 1б). Корреляционный анализ. Коэффициент корреляции. Дисперсионный анализ

Различают функциональные и корреляционные (случайные, стохастические) связи (зависимости). Функциональная связь между двумя величинами (аргументом и его функцией) является «твердой» связью: всякий раз как берется одно какое-либо значение аргумента, ему соответствует строго определенное значение функции. В археологии такие связи почти не встречаются. При корреляционной связи одному значению одной величины могут

соответствовать каждый раз разные значения другой. Они появляются с той или иной вероятностью. На практике получается так, что одному значению одного признака на разных объектах соответствуют различные значения другого признака. Чем меньший разброс этих значений, тем более тесная связь между значениями двух этих признаков.

Если нанести на координатную плоскость значения двух признаков, полученных в результате исследований серии объектов (корреляционное поле), то мы получим графики, выражающие характер зависимости этих признаков друг от друга. Может быть полное отсутствие зависимости. Тогда график будет покрыт беспорядочно расположенными точками или точки образуют круг или эллипс вдоль одной из осей координат (рис. 10, *а*). Точки могут группироваться «вдоль» осей координат так, что увеличению значений одного признака будет соответствовать уменьшение значений другого признака. Точки в таком случае строятся вдоль гиперболы, т. е. отражают обратно пропорциональный характер связи между признаками (рис. 10, *б*). Точки могут выстраиваться и вдоль любой другой линии. В этом случае требуется специальное осмысление характера связи между признаками. Точки, наконец, могут группироваться в виде более или менее вытянутого эллипса, т. е. вдоль прямой линии, наклонно размещенной по отношению к осям координат (рис. 10, *в*, 10, *г*).

Прямая линия — наиболее употребительный тип зависимости, который избирается при первоначальном исследовании объектов. Прямолинейная зависимость между признаками x и y выражается алгебраически линейной функцией вида $y=\alpha x+\beta$, где α и β — некоторые постоянные величины. В качестве теоретической зависимости $y=\alpha x+\beta$ выбирают такую прямую, которая минимально удалена (по y) от всех точек корреляционного поля.

При этом величины α и β определяются формулами:

$$\alpha=r\frac{\sigma_y}{\sigma_x},\quad \beta=\bar{y}-\alpha\bar{x},$$

где r — *коэффициент корреляции* (*Пирсона*) между x и y. Этот коэффициент подсчитывается следующим образом. В результате каких-то измерений n объектов получают

два параллельных ряда значений количественных признаков:

$$x_1,\ x_2,\ldots,\ x_i,\ldots,\ x_n;$$

$$y_1,\ y_2,\ldots,\ y_i,\ldots,\ y_n.$$

Каждые x_i и y_i принадлежат одному i-му объекту.

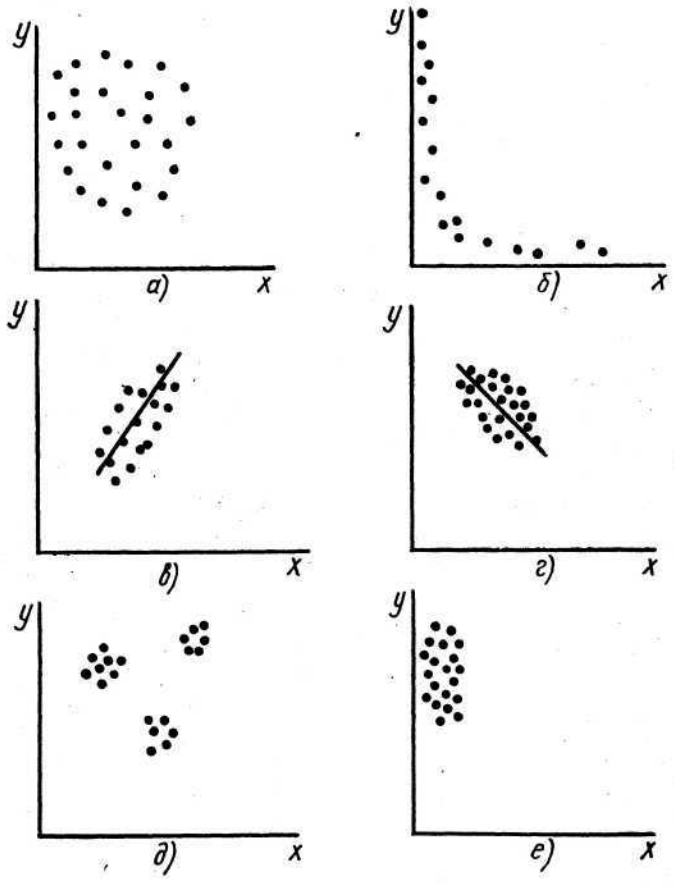


Рис. 10. Различные виды зависимости между количественными признаками:

а — нет зависимости; *б* — зависимость, близкая к обратно пропорциональной; *в* — зависимость, приближающаяся к линейной с положительным коэффициентом корреляции; *г* — зависимость, приближающаяся к линейной с отрицательным коэффициентом корреляции; *д* — кучное расположение точек без приближения к линейной зависимости; *е* — нулевой коэффициент корреляции

Тогда коэффициент корреляции равен:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=$$

$$=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right]\left[n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right]}}=$$

$$=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y},$$

где $\overline{xy}$ — средняя арифметическая произведений двух совместно встречающихся значений признаков.

Коэффициент корреляции применим только к такого рода зависимостям, которые приближаются к линейной, выражаемой уравнением $y=\alpha x+\beta$. Если коэффициент имеет положительное значение, то связь близка к прямой (чем выше значение одного, тем выше значения другого признака) (см. рис. 10, *в*). Если он отрицательный, то зависимость тоже линейная, но со знаком минус — чем выше значение одного признака, тем ниже значение другого (см. рис. 10, *г*). Однако это не обратно пропорциональная зависимость, выражаемая гиперболой и формулой $y=1/x$, а линейная зависимость типа $y=\alpha x+\beta$, где $\alpha<0$.

Если коэффициент корреляции оказывается равным 1 или —1, то корреляционная связь между значениями двух признаков превращается в функциональную, а корреляционный эллипс превращается в прямую линию. Коэффициент корреляции приближается к 0, если точки образуют фигуру, близкую к кругу, или выстраиваются вдоль прямой, параллельной какой-либо оси координат (рис. 10, *е*).

Коэффициент корреляции может быть очень низок, а связь между признаками может быть довольно тесная в том случае, если эта связь аналитически выражается нелинейной зависимостью. На корреляционном поле точки выстраиваются в этом случае вдоль какой-либо кривой, более или менее сильно отличающейся от прямой линии (например, на рис. 10, *б*).

При проведении корреляционного анализа важно установить надежность коэффициента корреляции. Чем больше выборка, тем больше надежность коэффициента корреляции. Эту надежность можно оценить.

По следующей таблице (табл. 20) можно определить значения, которые должен превысить коэффициент корреляции r, чтобы при соответствующем числе пар n и на доверительном уровне 0,95 можно было отвергнуть нулевую гипотезу о независимости признаков [1].

Таблица 20

n	r	n	r	n	r
3	1,00	15	0,51	27	0,38
4	0,95	16	0,50	28	0,37
5	0,88	17	0,48	29	0,37
6	0,81	18	0,47	30	0,36
7	0,75	19	0,46	31	0,36
8	0,71	20	0,44	32	0,35
9	0,67	21	0,43	35	0,33
10	0,63	22	0,42	40	0,30
11	0,60	23	0,41	50	0,27
12	0,58	24	0,40	60	0,25
13	0,55	25	0,40		
14	0,53	26	0,39		

Но для проведения корреляционного анализа и, в частности, для установления степени надежности коэффициента корреляции необходимо выполнение некоторых требований.

Каждая пара наблюдений должна быть статистически независимой от другой пары наблюдений. Совместное распределение признаков должно быть приблизительно нормально (в этом случае на корреляционном поле наблюдения образуют эллипс). При этом колебания второй случайной величины вокруг своей средней при каждом фиксированном значении первой должны подчиняться нормальному закону распределения.

Эти условия могут и нарушаться, но не очень сильно. На практике коэффициент Пирсона дает неплохие результаты и при некоторых отклонениях от этих условий.

Для быстрой и приблизительной оценки тесноты связи между двумя количественными признаками удобен *коэффициент Фехнера*. Подсчитывается, сколько раз (*а*) оба значения одновременно боль-

[1] См.: *Снедекор Д. У.* Указ. соч. С. 173. Табл. 57.

ше своих средних или меньших их, затем сколько раз (b) одно значение больше средней, а другое — меньше средней или наоборот, а затем определяется коэффициент по формуле

$$\Phi=\frac{a-b}{a+b}.$$

Этот коэффициент основан на том, что чем сильнее корреляционная связь между признаками, тем чаще их значения одновременно будут или подниматься над их средними значениями, или опускаться ниже их. И чем слабее эта связь, тем чаще будет возникать такое явление, что один признак будет иметь значение выше своего среднего, а другой — ниже. Пример применения коэффициента Фехнера см. в этой главе (пример 29).

Одним из видов статистической зависимости является такая зависимость, при которой точки на корреляционном поле будут группироваться кучно. Эти кучки точек не обязательно будут располагаться вдоль одной линии.

Если кучки на графике располагаются примерно на одной прямой, то возможно применение коэффициента корреляции r для всех значений признаков. Если они не располагаются так, то применение коэффициента корреляции для всех значений не допустимо (рис. 10, *д*).

Пример 18.

Измерены высота тулова и высота горла для сероглиняных кувшинчиков V в. Пашковского могильника[2]. Получено два ряда значений признаков: высота тулова x и высота горла y в см.

Требуется установить, зависимы ли эти два признака друг от друга и какова степень этой зависимости. Все данные и промежуточные расчеты сведены в табл. 21.

Средние арифметические величины равны:

$$\bar{x}=8{,}15,\quad \bar{y}=6{,}43.$$

Подсчитываем коэффициент корреляции:

$$r=\frac{20{,}39}{\sqrt{26{,}48\cdot 25{,}42}}\approx 0{,}79,\quad r>0{,}60.$$

Следовательно, гипотезу о независимости этих двух признаков следует отвергнуть.

Могут быть применены другие способы выявления и оценки тесноты связи между признаками, например *дисперсионный анализ*. В ходе его выявляется более универсальная мера тесноты связи между двумя признаками, которая может быть применена не только к линейным видам зависимости.

[2] См.: *Ковалевская (Деопик) В. Б.* Применение статистических методов к изучению массового археологического материала // Археология и естественные науки. М., 1965.

Допустим, нужно определить, насколько влияет признак x — факторный — на значения признака y — результативного. Для этого применяют дисперсионный анализ.

Таблица 21

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
5,5	5,0	—2,65	7,02	—1,43	2,04	3,79
6,0	4,0	—2,15	4,62	—2,43	5,90	5,22
7,3	6,3	—0,85	0,72	—0,13	0,02	0,11
7,8	6,5	—0,35	0,12	0,07	0,00	—0,02
8,0	4,9	—0,15	0,02	—1,53	2,34	0,23
8,0	8,2	—0,15	0,02	1,77	3,13	—0,26
8,5	5,6	0,35	0,12	—0,83	0,69	—0,29
8,6	5,7	0,45	0,20	—0,73	0,53	—0,33
9,2	7,9	1,05	1,10	1,47	2,16	1,54
9,3	7,4	1,15	1,32	0,97	0,94	1,12
11,5	9,2	3,35	11,22	2,77	7,67	9,28
Σ 89,7	70,7		26,48		25,42	20,39

Его сущность заключается в том, что все значения результативного признака разбивают в зависимости от факторного признака на несколько групп и анализируют, какую часть разброса значений этого признака можно объяснить с помощью групп. Если бы не было никакого влияния факторного признака на результативный, то средние арифметические значения результативного признака по группам $\bar{y}_i$ очень мало бы отличались от общей средней арифметической этого признака y. Но так как они все же отличаются, то это отличие может быть показателем влияния факторного признака на результативный.

Вычисляют три дисперсии:

1) общую дисперсию

$$D_0=\sum_{j=1}^{n}(y_j-\bar{y})^2,$$

т. е. суммируют квадраты отклонений всех значений y от общей средней $\bar{y}$, без учета разбиения по группам, n — число всех объектов;

2) факторную дисперсию

$$D_{\text{ф}}=\sum_{l=1}^{k} n_l(\bar{y}_l-\bar{\bar{y}})^2,$$

т. е. подсчитывают квадрат отклонений каждой i-й групповой средней $\bar{y}_i$ от общей средней $\bar{\bar{y}}$ и умножают на численность i-й группы. После этого все полученные произведения суммируют; k — число групп;

3) случайную дисперсию

$$D_{\text{с}}=\sum_{i=1}^{k}\sum_{j=1}^{n_l}(y_{ij}-\bar{y}_l)^2,$$

при этом суммируют квадраты отклонений всех значений результативного признака, включенных в i-ю группу, от соответствующей групповой средней $\bar{y}_i$. Полученные по всем группам результаты суммируют.

Влияние факторного признака на результативный определяется величиной так называемого корреляционного отношения

$$\eta_{\text{ф}}=\frac{D_{\text{ф}}}{D_{\text{о}}}.$$

Влияние остальных, неучтенных факторов на результативный признак определяется отношением

$$\eta_{\text{с}}=\frac{D_{\text{с}}}{D_{\text{о}}},$$

при этом

$$\eta_{\text{ф}}+\eta_{\text{с}}=1.$$

Чтобы проверить, не определяются ли полученные показатели случайностью выборки, применяется следующий критерий.

Подсчитывают дисперсии, приходящиеся на одну степень свободы в каждом случае. Для случайной дисперсии $D_{\text{с}}$ число степеней свободы $K_{\text{с}}$ определяется как число всех значений результативного признака без числа групп, для факторной дисперсии $D_{\text{ф}}$ число степеней свободы $K_{\text{ф}}$ равно числу групп, т. е. числу значений факторного признака, без 1.

Получают

$$F_{\text{ф}}=\frac{\sum n_l(\bar{y}_i-\bar{\bar{y}})^2}{K_{\text{ф}}} \quad \text{и} \quad F_{\text{с}}=\frac{\sum\sum(\bar{y}_{ij}-\bar{y}_i)^2}{K_{\text{с}}}.$$

Далее вычисляют отношение

$$F=\frac{F_{\text{ф}}}{F_{\text{о}}}.$$

Если F больше табличного значения распределения Фишера для чисел степеней свободы $K_{\text{о}}$ и $K_{\text{ф}}$ и соответствующего доверительного уровня, то нулевая гипотеза о случайности установленного корреляционным отношением влияния факторного признака на результативный отвергается с соответствующим доверительным уровнем [3].

Дисперсионный анализ позволяет установить также степень влияния нескольких факторных признаков на результативный. Факторный признак может быть как количественным (случай 1б), так и качественным (случай 4б), результативный — обязательно количественным.

Пример применения дисперсионного анализа в случае качественного факторного признака (случай 4б) см. ниже (пример 31).

3. Исследование связи между одним количественным признаком и несколькими другими количественными признаками (случай 1в)

Иногда нужно выявить связь и оценить тесноту связи между двумя признаками (x и y), с одной стороны, и каким-либо третьим (z) — с другой. Вычисляют коэффициенты корреляции по парам и получают совокупный коэффициент корреляции

$$r_{xy/z}=\sqrt{\frac{r_{xz}^2+r_{yz}^2-2r_{xy}r_{xz}r_{yz}}{1-r_{xy}^2}}.$$

Возможен также подсчет *множественного коэффициента корреляции* одного количественного признака и большего чем два числа других количественных признаков.

3. Связь между качественными признаками, поддающимися ранжированию (случай 2б)

В этом исследовании применимы непараметрические коэффициенты, в частности коэффициент Спирмена. Зна-

[3] О дисперсионном анализе см., например: *Венецкий И. Г., Кильдишев Г. С.* Основы математической статистики. М., 1963. С. 189 и далее; см. также: *Хьюстон А.* Дисперсионный анализ. М., 1971.

чения первого и второго признака независимо друг от друга ранжируются и нумеруются. Каждой паре признаков, совмещенных на объекте, таким образом приписываются два числа, обозначающих номера (ранги) значений присутствующих на нем признаков. Вычисляются разности этих номеров (d) для каждого объекта. *Коэффициент Спирмена* равен

$$R_C = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

где n — число объектов.

Чем больше n, тем меньшее значение R_c требуется, чтобы отвергнуть нулевую гипотезу о независимости двух признаков. Используют следующую таблицу (табл. 22)

Таблица 22

n	R_C	n	R_C	n	R_C	n	R_C
5	1,0	11	0,60	17	0,48	24	0,40
6	0,89	12	0,58	18	0,47	25	0,40
7	0,75	13	0,55	19	0,46	26	0,39
8	0,71	14	0,53	20	0,44	27	0,38
9	0,67	15	0,51	21	0,43	28	0,37
10	0,63	16	0,49	22	0,42	29	0,37
				23	0,41	30	0,36

значимости рангового коэффициента корреляции Спирмена для доверительного уровня 0,95 [4]. Если R_c больше табличного, а при $n>30$, если

$$R_C > \frac{1,96}{\sqrt{n-1}},$$

то гипотезу о независимости двух признаков следует отвергнуть с доверительным уровнем 0,95.

Аналогом коэффициента Спирмена является *коэффициент Кенделла* R_K [5]. Непараметрические коэффициенты корреляции Спирмена и Кенделла могут применяться также для быстрого ориентировочного определения степени тесноты связи между двумя количественными признаками (случай 1б). Каждый интервал при этом рассматривается как значение качественного ранжированного признака. Он удобен при исследовании тесноты взаимосвязи между ранжированным качественным и количественным признаками (интервальные значения последнего также рассматриваются как значения качественного ранжированного признака). Но как у всяких непараметрических методов, у них малая мощность: они допускают большую вероятность ошибки II рода, т. е. могут не отвергнуть неправильную гипотезу.

[4] См.: *Снедекор Д. У.* Указ. соч. С. 173. Табл. 57; С. 187. Табл. 65.

[5] См.: Количественные методы в исторических исследованиях. С. 218 и далее.

Пример 19.

Подсчитаем коэффициент Спирмена для двух количественных признаков примера 18.

Первый ряд x упорядочен и значения признака получили номера от 1 до 11. Упорядочиваем второй ряд y:

	4,0;	4,9	5,0	5,6;	5,7;	6,3;	6,5;	7,4;	7,9;	8,2;	9,2
Ранг	1	2	3	4	5	6	7	8	9	10	11

Располагаем номера значений упорядоченного второго ряда в том порядке, в каком эти значения располагались при упорядочении первого ряда:

Ранг	1	2	3	4	5	6	7	8	9	10	11
Ранг	3	1	6	7	2	10	4	5	9	8	11

Подсчитываем квадраты разности номеров для каждой пары признаков:

$$\sum_{i=1}^{n} d_i^2 = 2^2 + 1^2 + 3^2 + 3^2 + 3^2 + 4^2 + 3^2 + 3^2 + 0^2 + 2^2 + 0^2 = 70;$$

$$R_C = 1 - \frac{6 \cdot 70}{11 \cdot 120} \approx 0,68.$$

Замечаем, что $R_c \approx 0,68$ достаточен для отклонения гипотезы о независимости признаков на доверительном уровне 0,95, так как

$$R_C \approx 0,68 > 0,60.$$

Таким образом, применение непараметрических коэффициентов и критериев корреляции для количественных признаков возможно, оно облегчает счетную работу. Но в случае неопровержения гипотезы об отсутствии статистической значимости связи между признаками, непараметрический критерий следует проверить параметрическим.

Пример 20.

На монетах Москвы и ее уделов времени Василия I, чеканенных до 1400 г., было зарегистрировано в легендах 8 значений признака, выражающего вассалитет по отношению к Орде (I признак), которые можно было ранжировать по степени увеличения выраженности этого вассалитета, и 5 значений признака, выражающего русский суверенитет (II признак), которые также оказалось возможным ранжировать по степени выраженности этого суверенитета [6]. Значения этих признаков на монетах встречаются в 15 сочетаниях.

Таким образом, дано 15 следующих сочетаний:

Значение I признака	1	2	2	3	3	3	4	5	6	7	8	8	8	8	8
Значение II признака	2	4	2	1	4	3	4	3	3	3	I	5	4	3	2

Требуется определить коэффициент корреляции Спирмена и узнать, имеется ли существенная связь между этими двумя признаками.

Сочетания в данном случае играют роль объектов. Расположим эти 15 сочетаний по порядку значений I признака. Тем сочетаниям, у которых значения этого признака оказываются одинаковыми, дадим средний номер. Получим следующий ряд сочетаний, упорядоченных по I признаку:

Значение I признака	1	2	2	3	3	3	4	5	6	7	8	8	8	8	8
Номер	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ранг	1	2,5	2,5	5	5	5	7	8	9	10	13	13	13	13	13

Аналогично получаем ряд сочетаний, упорядоченный по II признаку:

Значение II признака	1	1	2	2	2	3	3	3	3	3	4	4	4	4	5
Номер значения	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ранг	1,5	1,5	4	4	4	8	8	8	8	8	12,5	12,5	12,5	12,5	15

Расположим номера рангов сочетаний, упорядоченных по II признаку, в том порядке, в котором эти сочетания располагались при упорядочении сочетаний по I признаку:

Ранг по I признаку	1	2,5	2,5	5	5	5	7	8	9	10	13	13	13	13	13
Ранг по II признаку	4	12,5	4	1,5	12,5	8	12,5	8	8	8	1,5	15	12,5	8	4

Применим коэффициент Спирмена. Подсчитаем

$$\sum_{i=1}^{n} d_i^2 = 3^2 + 10^2 + 1{,}5^2 + 3{,}5^2 + 7{,}5^2 + 3^2 + 5{,}5^2 + 0^2 + 1^2 + 2^2 + 1{,}5^2 + 2^2 + 0{,}5^2 + 5^2 + 9^2 = 466{,}5.$$

$$R_C = 1 - \frac{6 \cdot 466{,}5}{15 \cdot 224} \approx 0{,}17 < 0{,}51.$$

Коэффициент R_C получился низкий (еще ниже точное значение коэффициента — 0,12)[7]. Гипотеза о независимости двух признаков не опровергается. Очевидно, прямой зависимости между степенью выраженности ордынского вассалитета и степенью выраженности русского суверенитета при составлении монетного типа не было. Детальное исследование показывает, что связь эта все же была и весьма сложная.

4. Связь между значениями качественных признаков (случай 2а, 3а)

Допустим, на каком-то объекте i-е значение признака I должно появиться с вероятностью $p(\mathrm{I})_i$, а j-е значение признака II должно появиться с вероятностью $p(\mathrm{II})_j$.

Примем в качестве нулевой гипотезы предположение о том, что между i-м значением I признака и j-м значением признака II нет никакой связи. Тогда по теореме о произведении вероятностей взаимонезависимых событий вероятность одновременного появления i-го значения признака I и j-го значения признака II на каком-то объекте равна

$$p[(\mathrm{I})_i, (\mathrm{II})_j] = p(\mathrm{I})_i \, p(\mathrm{II})_j.$$

[6] См.: *Федоров-Давыдов Г. А.* Монеты Московской Руси. С. 105 и далее.

[7] В случае совпадающих значений вышеприведенная формула коэффициента Спирмена становится неточной и дает завышенные значения. Погрешность тем выше, чем большую долю составляют совпадения значений.

Если нулевая гипотеза не верна и эти значения признаков зависимы друг от друга, то
или $p[(\mathrm{I})_i, (\mathrm{III})_j] > p(\mathrm{I})_i p(\mathrm{II})_j$ в случае,

если зависимость такова, что заставляет эти значения признака встречаться на одном объекте чаще, чем это было бы, если значения признаков были независимы, т. е. сближает эти значения признаков (положительная связь),

или $p[(\mathrm{I})_i, (\mathrm{II})_j] < p(\mathrm{I})_i p(\mathrm{II})_j$ в случае,

если зависимость такова, что, наоборот, препятствует этим значениям признаков встречаться на одном объекте, т. е. отделяет их друг от друга (отрицательная связь).

Таким образом, нам для проверки нулевой гипотезы нужно сравнить теоретически вычислительную вероятность встречи на одном объекте i-го значения признака I и j-го значения признака II с реально наблюденной по выборке частостью этого события. Но $p(I)_i$ и $p(\mathrm{II})_j$ нам не известны в точности, а оценены по выборкам, и реальная, наблюденная по выборке, частость встречи этих двух признаков на одном объекте также подвержена случайностям выборки и является лишь приближенной оценкой частости соединения указанных значений двух признаков на одном объекте в генеральной совокупности. Возникает вопрос: насколько существенно расхождение теоретической частости, вычисленной при допущении нулевой гипотезы, и реально наблюденной эмпирической частости, опровергает ли оно нулевую гипотезу и тем самым показывает действительную, существующую в генеральной совокупности взаимосвязь значений признаков, или же оно является следствием случайных обстоятельств, влияющих на выборку.

При исследовании какого-либо многочисленного, но сильно фрагментированного материала, например керамики, мы хотим установить взаимосвязь значений двух признаков: например, венчика вида «A» и орнамента на плечиках сосуда вида «B».

Мы имеем весьма много (n_1) венчиков (фрагментов) и среди них m_A венчиков вида «A» (аналогично n_2 и m_B).

Таким образом,

$$w_A = \frac{m_A}{n_1}; \quad w_B = \frac{m_B}{n_2}.$$

Так как n_1 и n_2 велики, мы w_A и w_B можем принять как оценки $p(A)$ и $p(B)$, без учета отклонений из-за случайностей выборок.

Естественно, значительно меньше таких фрагментов сосудов, где есть и венчик и плечики. Их всего $n_{1,2}$, из них на m_{AB} имеется венчик вида «A» и орнамент вида «B». Тогда эмпирическая частость этих фрагментов будет равна

$$w_{AB} = \frac{m_{AB}}{n_{1,2}}$$

По эмпирической частности события «AB» (совмещения на одном сосуде венчиков типа «A» и орнамента типа «B») оценим его вероятность в генеральной совокупности, которая должна находиться в пределах доверительного интервала (p', p''). Последний мы можем найти в таблице или построить для избраниого доверительного уровня (0,95). Теперь остается сопоставить теоретически вычисленную вероятность события «AB» при допущении нулевой гипотезы о независимости события «A» и события «B» с доверительным интервалом для вероятности этого события, оцененной по его эмпирической частости. Если

$$p' \leqslant \frac{m_A}{n_1} \cdot \frac{m_B}{n_2} = w_A w_B \leqslant p'',$$

то нулевая гипотеза остается не опровергнутой. В противном случае она считается опровергнутой с риском ошибки в 0,05. При этом, если

$$w_A w_B < p',$$

то между видом венчика («A») и видом орнамента («B») связь положительная, а если

$$w_A w_B > p'',$$

то связь между этими значениями признака отрицательная.

Достоинством рассмотренного метода является то, что мы можем в той или иной мере учесть информацию, которая содержится в сильно фрагментированном материале, а не только в «археологически целых» изделиях.

Пример 21.

На поселении обнаружен 31 фрагмент венчиков вида «A» из 108 фрагментов всех венчиков и 370 фрагментов плечиков сосудов с орнаментом вида «B» из 805 фрагментов всех плечиков. Кроме того, имеется 42 крупных фрагмента, на которых сохранился и венчик и плечико

сосуда, и из этих фрагментов на 12 присутствует венчик вида «A» и орнамент вида «B». Требуется установить, имеется ли связь между этими признаками, т. е. венчиком вида «A» и орнаментом вида «B».

Так как выборки эти велики, то

$$p(A)\approx\frac{31}{108}\approx 0{,}29;\quad p(B)\approx\frac{370}{805}\approx 0{,}46.$$

Тогда теоретическая вероятность появления на одном сосуде венчика «A» и орнамента «B» при допущении (нулевая гипотеза), что эти признаки между собой никак не связаны, равна

$$p(AB)=p(A)\,p(B)\approx 0{,}29\cdot 0{,}46=0{,}1334.$$

В действительности же мы имеем 12 таких фрагментов сосудов, на которых есть венчик вида «A» и орнамент вида «B» из 42 всех фрагментов, у которых сохранился венчик и плечики, т. е. эмпирическая частость соединения венчика вида «A» и орнамента вида «B» равна

$$w_{AB}=\frac{12}{42}\approx 0{,}28.$$

Замечаем, что $p(A)p(B)<w_{AB}$. Как убедиться, что это неравенство не результат случайности выборки? Находим доверительный интервал для генеральной вероятности по выборочной частости 12/42. Он составляет интервал 0,157 — 0,446 [8]. Следовательно, с вероятностью 0,95 можно утверждать, что генеральная вероятность, равная при допущении нулевой гипотезы теоретической, не выйдет за пределы от 0,157 до 0,446, а полученная теоретическая вероятность ниже нижней границы доверительного интервала. Следовательно, нулевая гипотеза отвергнута с доверительным уровнем 0,95 в пользу гипотезы о наличии положительной связи венчика типа «A» и орнамента типа «B».

* * *

Как произвести исследование связей i-го значений признака I и j-го значений признака II по одной выборке состоящей из n объектов?

[8] См.: *Большов Л. Н., Смирнов Н. В.* Указ. соч. Табл. 5,2.

Для этого удобно рассматривать i-е значение I признака как альтернативный признак «A» со значениями: A — присутствует, $\bar{A}$ — отсутствует; аналогично j-е значение II признака как альтернативный признак «B» со значениями: B — присутствует и $\bar{B}$ — отсутствует. Составляем 4-польную таблицу:

	B	$\bar{B}$	
A	a	b	$a+b$
$\bar{A}$	c	d	$c+d$
	$a+c$	$b+d$	n

где a — число объектов, на которых есть и признак «A» и признак «B» (событие AB);

b — число объектов, на которых есть признак «A», но нет признака «B» (событие $A\bar{B}$);

c — число объектов, на которых нет признака «A», но есть признак «B» (событие $\bar{A}B$);

d — число объектов, на которых нет ни признака «A», ни признака «B» (событие $\bar{A}\bar{B}$);

$a+b$ — число объектов, на которых есть признак «A» (событие A);

$a+c$ — число объектов, на которых есть признак «B» (событие B);

$c+d$ — число объектов, на которых нет признака «A» (событие $\bar{A}$);

$b+d$ — число объектов, на которых нет признака «B» (событие $\bar{B}$).

Нулевая гипотеза с доверительным уровнем 0,95 будет состоять в том, что между признаком «A» и признаком «B» нет никакой связи, взаимозависимости.

Тогда должно иметь место равенство

$$\frac{a}{c}=\frac{b}{d}.$$

Если это равенство нарушается, то между «A» и «B» имеется взаимозависимость.

Но это нарушение мы наблюдаем только в выборке, а не в генеральной совокупности. Оно может быть случайным, связанным с выборочностью материала. Задача сводится к сравнению двух частостей и установлению, значимо или незначимо различие между ними. Эта задача решалась уже во II главе путем сравнения доверительных интервалов. Здесь применим иной способ, предпочтительный в данном случае. Для того чтобы отвергнуть нулевую гипотезу о независимости этих значений

признака, применяется критери χ^2, уже знакомый нам по I главе. В данном случае он имеет вид

$$\chi^2=\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Число степеней свободы высчитывается как произведение числа возможных независимых частостей по значениям признака «A» (2—1=1) и по значениям признака «B» (2—1=1), т. е. как $1\cdot 1=1$.

Если χ^2 превосходит табличное (см. табл. II) значение для I степени свободы и доверительного уровня 0,95 (в этих условиях 3,84), то нулевая гипотеза может считаться опровергнутой (с риском ошибки в 0,05). Мы можем утверждать, что между признаками «A» и «B» есть какая-то связь.

Для применения критерия χ^2 для 4-польной таблицы взаимовстречаемости признаков желательно, чтобы величины

$$\frac{a+c}{n}(a+b);\ \frac{b+d}{n}(a+b);\ \frac{(a+c)}{n}(c+d);\ \frac{b+d}{n}(c+d)$$

были бы не менее четырех [9].

При малых выборках ($n\leqslant 25$) рекомендуется употреблять для проверки нулевых гипотез на 4-польной таблице взаимовстречаемости двух признаков, так называемый *точный критерий Фишера.*

При этом рядом с исходной таблицей образуются дополнительные 4-польные таблицы, в которых число a уменьшается в каждой таблице на 1 (до нулевого значения), а суммы по всем строкам и столбцам остаются прежними. Для 4-польной таблицы число степеней свободы всегда равно 1, и изменение в одной клетке необходимо влечет соответствующие изменения во всех остальных трех клетках, так как суммы столбцов и строк остаются неизменными.

Для каждой таблицы

$$p=\frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{n!\,a!\,b!\,c!\,d!},$$

и суммируем все такие p. Этим мы подсчитываем вероятность того, что при данных $a+b$, $a+c$ и n число a будет не меньше того, которое получено в выборке при условии, что между признаками «A» и «B» нет никакой связи.

[9] См.: *ван дер Варден Б. Л.* Указ. соч. С. 286.

Если сумма p для всех таблиц не превышает половину допустимой возможности ошибки (0,05 : 2=0,025), то гипотеза об отсутствии связи между признаками отвергается в пользу гипотезы о наличии отрицательной связи между этими признаками. Если проделать ту же процедуру с дополнительными таблицами, но последовательно уменьшать число (b или c) и прийти к тому, что суммы полученных вероятностей не будут превосходить 0,025, то нулевая гипотеза будет отвергнута в пользу гипотезы о наличии положительной односторонней связи между признаками.

* * *

Рассмотренные критерии для 4-польной таблицы взаимовстречаемости двух признаков показывают только с той или иной вероятностью ошибки (доверительный уровень), что между признаками существует связь. Но они ничего не говорят о том, насколько эта связь сильна. Для количественного выражения тесноты этой связи можно использовать так называемые эмпирические коэффициенты, например *коэффициент сопряженности:*

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

Замечаем, что этот коэффициент есть квадратный корень значения χ^2/n. С другой стороны, его величина равна коэффициенту корреляции между качественными альтернативами признаками «A» и «B», рассматриваемыми как количественные после приписывания объекту значения I при наличии признака и 0 при отсутствии его на объекте.

Даже при относительно больших значениях Q, но малых выборках n мы можем получить слишком низкое значение критерия χ^2, чтобы утверждать о наличии связи. В таком случае высокий уровень коэффициента связи может быть результатом случайности выборки.

Так как для опровержения нулевой гипотезы об отсутствии связи между «A» и «B» необходимо, чтобы $\chi^2>3{,}84\approx 4$ (при вероятности ошибки 0,05), то должно быть

$$Q^2n>4,\ Q>\frac{2}{\sqrt{n}}.$$

Например, при $n=49$, Q должно быть больше $2/7\approx 0{,}29$, чтобы считать связь между «A» и «B» существенной.

Коэффициент Q колеблется от —1 до 1, т. е. от максимальной отрицательной связи до максимальной положительной связи. Эти значения он принимает, когда a и d оказываются равными 0 ($Q=-1$) или b и c оказываются равными 0 ($Q=1$); $Q=0$, если $ad=bc$. Таким образом, увеличение значений a и d ведет к увеличению показателя силы связи, а увеличение b и c ведет к уменьшению показателя силы связи. Действительно, чем чаще «A» встречается с «B» на одном объекте, тем больше «сила», их соединяющая. Чем чаще на объекте нет ни «A», ни «B», тем теснее связь между «A» и «B», но осуществляется она на других объектах. О слабости связи между «A» и «B» говорят те случаи, когда на объекте есть «A», но нет «B» или наоборот.

Часто используют другой показатель силы связи между двумя признаками — *коэффициент ассоциации* (Юла):

$$K=\frac{ad-bc}{ad+bc}.$$

Увеличение значений a и d ведет к увеличению показателя ассоциации, увеличение значений b и c ведет к уменьшению этого показателя. $K=-1$ (полная отрицательная связь) возникает, когда a или d обращаются в 0, $K=1$, когда или b, или c обращаются в 0. $K=0$, если ad равно cd, т. е. тогда же, когда $Q=0$. Оба коэффициента дают высокие значения, когда имеется двусторонняя связь, т. е. появление признака «A» влечет более частое появление признака «B», а появление признака «B» влечет также более частое появление признака «A». Если появление признака «A» влечет более частое появление признака «B», а появление признака «B» не влечет существенного увеличения частоты признака «A», то коэффициент Q будет ниже, чем коэффициент K. Другими словами, коэффициент Q чувствителен к обоим направлениям связи между «A» и «B» и отсутствие одного из этих направлений существенно его снижает. Коэффициент K отражает наличие хотя бы односторонней связи, даже при слабой связи в противоположном направлении.

В археологических исследованиях предпочтительнее коэффициент Q, так как чаще ставится задача выявления двусторонних связей между признаками.

Пример 22.

В Пенджикенте были подсчитаны случаи находок в зданиях лепной и гончарной керамики [10]: A — гончарная керамика; $\bar{A}$ — лепная керамика; B — находки в раннем здании; $\bar{B}$ — находки в позднем здании.

Таким образом, дана следующая 4-польная таблица взаимовстречаемости двух признаков: техники изготовления керамики (гончарная или лепная) и местонахождение ее (в раннем или позднем здании):

	A	$\bar{A}$	
B	16	15	31
$\bar{B}$	8	28	36
	24	43	67

Требуется установить, имеется ли между этими признаками существенная связь. Подсчитываем:

$$\chi^2=\frac{(16\cdot 28-8\cdot 15)^2\,67}{31\cdot 36\cdot 24\cdot 43}=6{,}26>3{,}84.$$

Следовательно, связь между гончарной керамикой и находками ее в раннем здании не случайна. Ошибиться, утверждая это, можно с вероятностью менее 0,05.

$$Q=\frac{16\cdot 28-8\cdot 15}{\sqrt{31\cdot 36\cdot 24\cdot 43}}\approx 0{,}31.$$

Хотя коэффициент связи невелик, но связь все же есть.

Пример 23.

Из погребений Саркельского грунтового могильника XI в. было извлечено значительное количество бус [11]. Исследовались они с точки зрения техники исполнения и формы.

A — бусы в технике навивки; $\bar{A}$ — бусы в другой технике; B — бусы шаровидно-зонной формы; $\bar{B}$ — бусы другой формы.

Таким образом, дана следующая 4-польная таблица взаимовстречаемости двух признаков: техника навивки и шаровидно-зонная форма:

	A	$\bar{A}$	
B	158	236	394
$\bar{B}$	469	1626	2095
	627	1862	2489

[10] См.: *Распопова В. И.* Керамика и слой поселения (на материалах VII—VIII вв. Пенджикента) //СКМА. М., 1970.

[11] См.: *Федоров-Давыдов Г. А.* О статистическом исследовании взаимовстречаемости признаков и типов в археологических комплексах // СКМА. М., 1970.

Требуется установить, имеется ли между этими двумя признаками существенная связь. Подсчитываем:

$$\chi^2 \approx 55 > 3{,}84, \quad Q \approx 0{,}15, \quad K \approx 0{,}4.$$

Хотя показатели силы связи не очень велики, связь между двумя значениями признаков несомненна, и ошибка при утверждении этого возможна с очень малой вероятностью, много меньшей не только 0,05, но и 0,01 (при этом уровне χ^2 должен быть больше 11, а он равен 55,5).

Пример 24.

Еще Б. Н. Граков заметил, что скифские катакомбы II вида не впускаются в насыпи курганов с основными погребениями, имеющими катакомбы III вида [12]. Проверим этот тезис статистически.

A — впускные катакомбы II вида; $\bar{A}$ — впускные катакомбы других видов; B — основные катакомбы *III* вида; $\bar{B}$ — основные катакомбы других видов.

В результате сбора материалов о раскопках скифских курганов получена следующая 4-польная таблица взаимовстречаемости двух признаков:

	A	$\bar{A}$	
B	0	8	8
$\bar{B}$	12	77	89
	12	85	97

Требуется установить, имеется ли между этими двумя признаками существенная связь. Подсчитаем:

$$Q = \frac{-12 \cdot 8}{\sqrt{12 \cdot 85 \cdot 8 \cdot 89}} \approx -0{,}11.$$

Коэффициент сопряженности получился очень низким и не свидетельствует о наличии значительной отрицательной связи между «A» и «B».

Проверим нулевую гипотезу об отсутствии отрицательной значимой связи между «A» и «B». Применим критерий Фишера. Так как $a=0$, то строить дополнительные таблицы не нужно. Это облегчает вычисления, сложные при подсчете этого критерия для больших a.

$$p = \frac{12!\,85!\,89!\,8!}{97!\,0!\,8!\,12!\,77!} \approx 0{,}33.$$

Вероятность того, что полученное распределение частот возможно при нулевой гипотезе, высока (0,33). Это не дает возможности отвергнуть нулевую гипотезу с доверительным уровнем, равным даже 0,8, и тезис о несовместимости впускных катакомб II типа и основных III типа остается не подтвержденным статистически. Возможно, этот тезис подтвердится в будущем с появлением нового материала.

* *
*

Некоторые исследователи в археологии применяют следующий коэффициент для оценки связи между двумя значениями признаков. В 4-польной таблице берутся только величины a, $a+c$, $a+b$.

$$Q' = \frac{a^2}{(a+c)(a+b)}.$$

Этот коэффициент дает приблизительную оценку связи между двумя значениями признаков. Однако он имеет крупный недостаток: коэффициент не учитывает d, т. е. безразличен к тому, сколько объектов не имеют ни «A», ни «B». Однако это значение d весьма важно, и в некоторых случаях даже при малом a регистрируется существенная связь. На примере 25 будет показано, как от учета d зависит решение вопроса о наличии или отсутствии связи и истолкование ее. Эти соображения заставляют рассматривать коэффициент Q' как неоправданно упрощенный. Q' всегда больше 0. Отрицательная связь выражается в значениях Q', близких к 0, и выявляется с трудом.

* *
*

После того как связь между значениями признаков установлена, дело собственно археологии истолковать ее. Она может быть функциональной, хронологической или территориальной; связь между значениями признаков может быть заложена в самой природе признаков, например цвет (красный) и материал (сердолик) для бус (часто в таком случае эта односторонняя связь абсолютна и выражается коэффициентом ассоциации, равным 1) и т. п. В значительной мере на истолкование характера

[12] См.: *Ольховский В. С.* Скифские катакомбы в Северном Причерноморье // СА. 1977. № 4.

связи оказывает влияние то, из какой совокупности взята выборка. Так, если установлена связь между видом стремян «*A*» и видом удил «*B*», то в этой связи отражены два фактора: 1) стремена вообще связаны с удилами по своим функциям и 2) стремена именно вида «*A*» связаны с удилами именно вида «*B*», так как они одновременны или относятся к одному локальному варианту. Если взять выборку из всех погребений данной культуры (безразлично, есть там стремена или удила), то главное воздействие на возможное цифровое выражение связи стремян вида «*A*» с удилами вида «*B*» окажет не связь между видами, а связь между категориями, которая «забьет» связь между видами. Если мы возьмем в качестве выборки только те погребения, в которых есть стремена любого вида и удила тоже любого вида, то на возможное выражение связи между стременами вида «*A*» и удилами вида «*B*» окажет влияние только связь между этими видами, а функциональная связь между стременами вообще и удилами вообще будет элиминирована, устранена.

Исследовалась взаимовстречаемость различного вида стремян и удил в кочевнических погребениях XI—XIV вв. в степях Восточной Европы. Задача состоит в том, чтобы установить взаимосвязь стремян вида «*A*» и удил вида «*B*» и оценить ее силу. При этом возможны два варианта выборки.

Пример 25.

Выборка состоит из всех известных нам погребений данной культуры. *A* — стремена вида «*A*»; $\bar{A}$ — нет стремян вида «*A*»; *B* — удила вида «*B*»; $\bar{B}$ — нет удил вида «*B*».

Дана следующая 4-польная таблица взаимовстречаемости двух признаков: стремена вида «*A*» и удила вида «*B*»:

	A	$\bar{A}$	
B	18	50	68
$\bar{B}$	46	900	946
	64	950	1014

Подсчитываем: $\chi^2=50>3{,}84$; $Q=0{,}22$; $Q'=0{,}07$.

Выборка состоит из известных нам погребений, имеющих стремена и удила. *A* — стремена вида «*A*»; $\bar{A}$ — стремена других видов; *B* — удила вида «*B*»; $\bar{B}$ — удила других видов.

Дана следующая 4-польная таблица взаимовстречаемости тех же признаков:

	A	$\bar{A}$	
B	18	50	68
$\bar{B}$	46	210	324
	64	260	

Подсчитываем: $\chi^2=2{,}4<3{,}84$; $Q=0{,}09$; $Q'=0{,}07$.

В этих двух случаях *a*, *b*, *c* — одинаковы, а *d* — различны. В первом случае связь устанавливается, во втором случае — не устанавливается. Связь в первом случае была функциональной связью между стременами и удилами вообще, как взаимосвязанными принадлежностями конской сбруи.

Заметим, что коэффициент Q' одинаков для обеих выборок, так как он не учитывает *d*.

Как мы должны понимать влияние хронологического или географического фактора на образование того или иного показателя связи двух явлений (признаков)?

Представим себе совокупность всех объектов, на которых есть хотя бы один из признаков «А» или «В». Эти объекты бытовали в одном каком-то месте в течение определенного периода. По вертикали откладывается количество всех объектов с «А» и всех объектов с «В», бытовавших в какой-то момент времени, по горизонтали — время. Получаются две кривые — графики двух временных рядов. Площадь под каждой кривой — общее количество объектов с «А» или объектов с «В». Отрезок хронологической оси, находящийся и под первой кривой и под второй, — это периоды времени, когда имелись и объекты с «А» и объекты с «В» и существовала возможность соединения на одном объекте «*A*» с «*B*» или соединения объектов с «*A*» и объектов с «*B*» в одном комплексе. Остальные участки хронологической оси под кривыми — это периоды раздельного существования «*A*» и «*B*».

Участки хронологической оси, не находящиеся под кривыми, — это время, когда не было ни объектов с «*A*», ни объектов с «*B*».

Но эти временные ряды и их графики нам не известны; мы их строим гипотетически исходя из данных выборки. В зависимости от того, какими окажутся *a*, *b*, *c*, *d*, мы можем представить время, охватываемое выборкой (обозначено t_1, t_2), и взаимное расположение кривых.

Если мы представим себе совокупность всех объектов с «*A*» и с «*B*», бытовавших на обширной территории в один какой-то краткий момент, то должны будем изобра-

зить это набором ареалов. В каком-то ареале бытуют объекты с признаком «A», в другом — с признаком «B», На пересечении этих зон бытуют объекты и с «A», и с «B»: или имеется возможность соединения этих признаков на одном объекте или соединения этих объектов с «A» и объектов с «B» в одном комплексе.

Но эти ареалы — также только предположения, основанные на данных выборки. Исходя из них, мы может

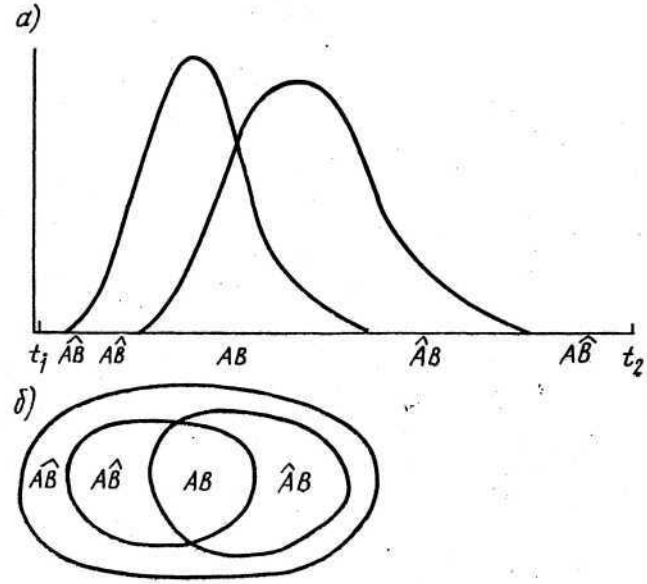


Рис. 11. Интерпретация случая 1 зависимости между качественными признаками:
а — хронологическая; *б* — географическая

представить территориальные границы выборки (внешняя линия) и взаиморасположение ареалов.

Пример 26.

1) Низкий коэффициент показывает слабую или нулевую связь между «A» и «B». Например:

	A	$\bar{A}$	
B	140	150	290
$\bar{B}$	100	110	210
	240	260	500

$Q=0{,}01$; $K=0{,}01$; $\chi^2=0{,}02<3{,}84$; $Q'=0{,}28$.

Такая ситуация возникает, когда a, b, c и d примерно одинаковы, не слишком сильно отличаются друг от друга. Периоды бытования «A» и «B» (или их географические ареалы) близки друг к другу, но недостаточно близки для того, чтобы утверждать, что «A» и «B» синхронны (или имели один ареал). Хронологически эту ситуацию можно представить так, как изображено на рис. 11, *а*, а территориально — на рис. 11, *б*.

Заметим, что Q', не учитывающий d, показывает сравнительно высокую связь.

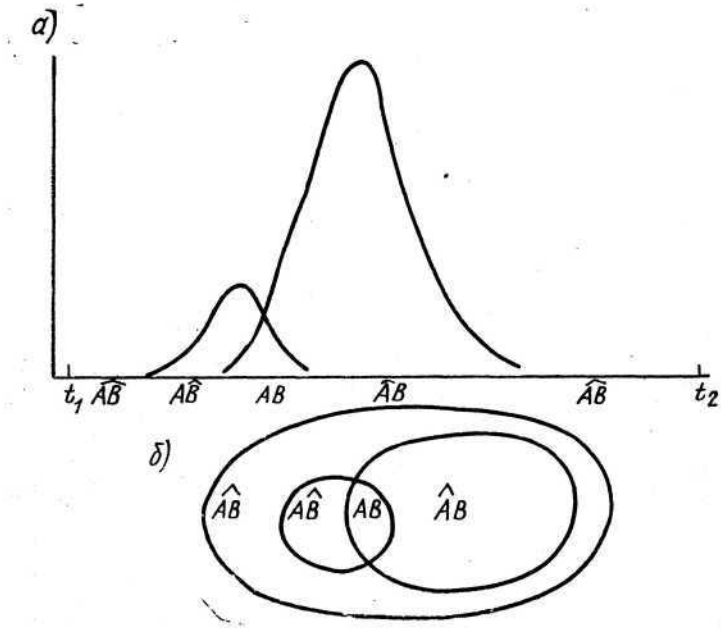


Рис. 12. Интерпретация случая 2 зависимости между качественными признаками:
а — хронологическая; *б* — географическая

2) Низкий уровень у коэффициента связи может возникнуть также за счет низких a и c при высоких b и d. Например,

	A	$\bar{A}$	
B	20	180	200
$\bar{B}$	30	270	300
	50	450	500

$Q=0$; $K=0$; $\chi^2=0$; $Q'=0{,}04$.

Такая ситуация возникает, когда $a+e$ вообще мало («A» — редкий признак). Большая часть «B» бытует без

«*A*». Эту ситуацию на хронологической оси можно представить так, как показано на рис. 12, *а*.

Как это можно представить географически, см. на рис. 12, *б*.

Таким образом, низкий коэффициент связи не дает возможность утверждать ни синхронность, ни разновременность «*A*» и «*B*» (аналогично ни территориальную близость, ни разобщенность «*A*» и «*B*»).

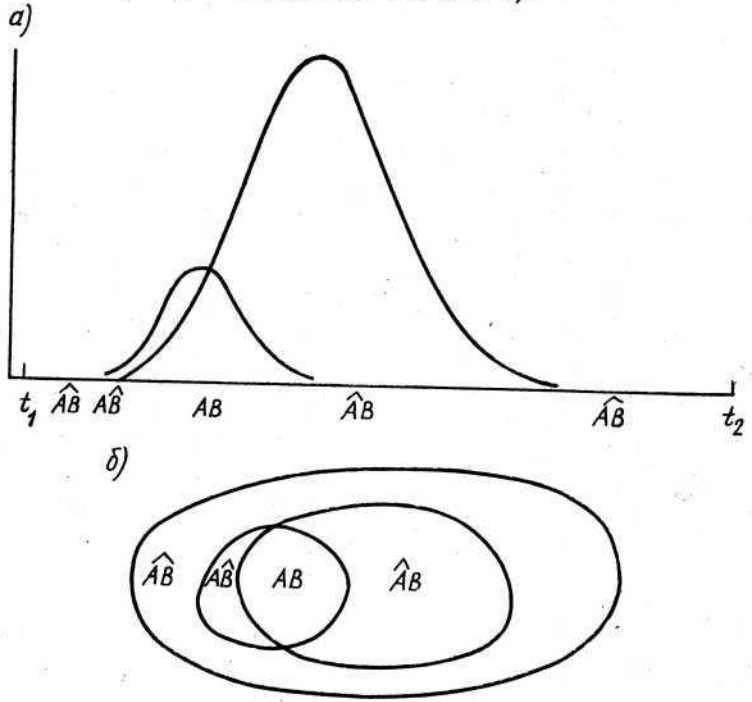


Рис. 13. Интерпретация случая 3 зависимости между качественными признаками:
а — хронологическая; *б* — географическая

3) При малом $a+c$ может возникнуть высокий коэффициент связи Q, если c относительно a тоже мало (a составляет большую часть величины $a+c$).

	A	$\bar{A}$	
B	40	160	200
$\bar{B}$	10	290	300
	50	450	500

$Q=0{,}27$; $K=0{,}76$; $\chi^2=37>3{,}84$; $Q'=0{,}16$.

Высокий уровень коэффициента K возникает за счет односторонней связи $A\rightarrow B$ (почти — 80% «*A*» соединяются с «*B*»). Обратная связь мала, так как только 20% «*B*» соединяются с «*A*». Показатель двусторонней связи Q поэтому сравнительно низок.

Все это представлено на рис. 13, *а* — хронологически и на рис. 13, *б* — географически.

Утверждать синхронность «*A*» и «*B*» можно только с оговорками. Тогда, когда бытовал «*B*», не обязательно бытовал «*A*», хотя когда бытовал «*A*», бытовал и «*B*».

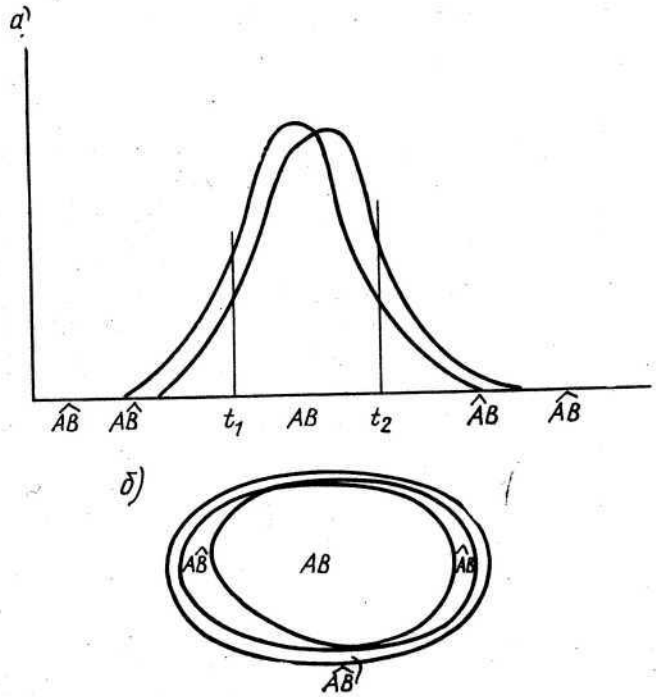


Рис. 14. Интерпретация случая 4 зависимости между качественными признаками:
а — хронологическая; *б* — географическая

Аналогично: на территории, где распространен был «*A*», встречается и «*B*», но не на всей территории распространения «*B*» встречается «*A*». Какой-то «*C*» может сосуществовать с «*B*», но из этого не следует, что «*C*» синхронен «*A*».

4) Низкий уровень коэффициента связи Q может возникнуть также за счет очень низкого d, даже при высоких a и низких b и c (так называемый d-эффект).

	A	$\bar{A}$	
B	260	10	270
$\bar{B}$	20	1	21
	280	11	291

$Q=0{,}01$; $K=0{,}13$; $Q'=0{,}89$.

Периоды бытования «A» и «B» близки, но распространены почти на все время, охваченное выборкой. Не имеется какого-то ограниченного периода синхронности «A» и «B» внутри этого времени, т. е. в пределах взятого периода (или на определенной территории) «A» и «B» равномерно сосуществуют и не обнаруживается ничего, что заставляло бы их в одно время (или в одном месте) встречаться вместе чаще, а в другое — реже. Этот случай важен и интересен тем, что показывает, что взаимовстречаемость выявляет не просто сосуществование признаков, а их сосуществование в какой-то ограниченный период на фоне раздельного их существования и полного отсутствия обоих признаков.

На хронологической оси это можно представить так, как показано на рис. 14, *а*, географически — на рис. 14, *б*.

Заметим, что в этих случаях коэффициенты Q' и K будут ненадежны, так как могут показывать довольно высокую связь, в то время как ее на самом деле в пределах изучаемой выборки нет (из-за малого d).

5) Высокий уровень коэффициента Q показывает существенную связь между «A» и «B». Например:

	A	$\bar{A}$	
B	260	40	300
$\bar{B}$	30	170	200
	290	210	500

$Q=0{,}71$; $K=0{,}95$; $\chi^2=253>3{,}84$; $Q'=0{,}78$.

Высокий уровень коэффициента Q образуется за счет больших a, d и малых b, c. Это показывает значительное сближение периодов (или ареалов) бытования «A» и «B». На хронологической оси это представлено рис. 15, *а*, географически — рис. 15, *б*.

Заметим, что коэффициент Q', который не учитывает величину d, не выявляет различия между такими ситуациями: в случае 4 «A» и «B» совместно встречаются на всем протяжении периода времени, охваченного выборкой, т. е. имеют по отношению к нему длительный пе-

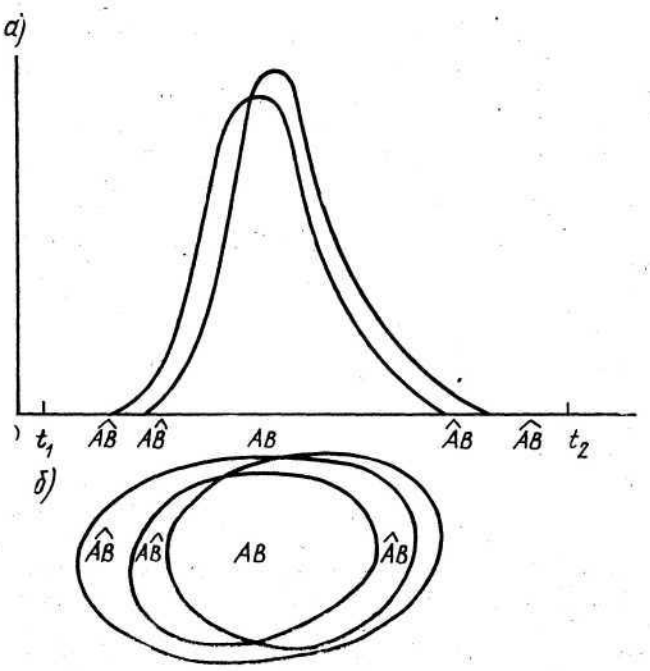


Рис. 15. Интерпретация случая 5 зависимости между качественными признаками:

а — хронологическая; *б* — географическая

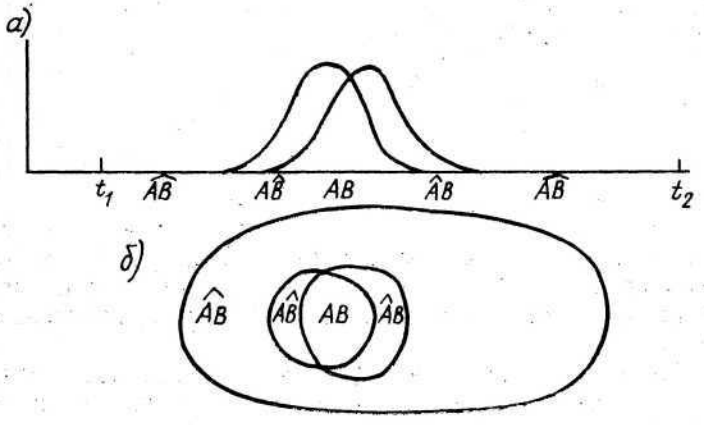


Рис. 16. Интерпретация случая 6 зависимости между качественными признаками:

а — хронологическая; *б* — географическая

риод бытования и не являются характерными для какого-то частного периода, т. е. не являются датирующими; в случае 5 имеется синхронность «*A*» и «*B*» для какого-то частного периода из всего времени, охваченного выборкой, т. е. можно говорить о характерности «*A*» и «*B*» для какого-то периода.

6) Высокий уровень коэффициент Q, показывающий существенную связь между «*A*» и «*B*», может возникнуть и при небольшом a, но при большом d (d-эффект). Например:

	A	$\bar{A}$	
B	70	30	100
$\bar{B}$	30	370	400
	100	400	500

$Q=0{,}63$; $K=0{,}93$; $\chi^2=195>3{,}84$; $Q'=0{,}49$.

Такая ситуация возникает тогда, когда период (или ареал), охватываемый учтенным материалом, велик по сравнению с периодом (ареалом) бытования «*A*» и «*B*». На его фоне соседство во времени (или территориальное) «*A*» и «*B*» воспринимается как свидетельство о синхронности (или территориальной близости) «*A*» и «*B*». Этот случай обратен случаю 4.

На хронологической оси можно представить так, как показано на рис. 16, *а*; географически — на рис. 16, *б*.

Коэффициент Q' в этом случае оказывается заниженным, так как не учитывает d.

7) Крайний случай представляет собой высокий уровень коэффициента Q, образующийся за счет того, что $b=0$ или $c=0$. Например:

	A	$\bar{A}$	
B	100	10	110
$\bar{B}$	0	390	390
	100	400	500

$Q=0{,}94$; $K=1$; $\chi^2=443>3{,}84$; $Q'=0{,}91$.

Это значит, что «*A*» почти всегда бытовало вместе с «*B*», а «*B*» почти всегда с «*A*». Периоды бытования (или ареалы) «*A*» и «*B*» почти совпадают. Хронологически это представлено на рис. 17, *а*, географически — на рис. 17, *б*.

Таким образом, при высоком коэффициенте связи можно утверждать примерную синхронность «*A*» и «*B*» (аналогично их территориальную близость).

8) Если значение коэффициента отрицательно и велико по абсолютной величине, то это показывает отри-

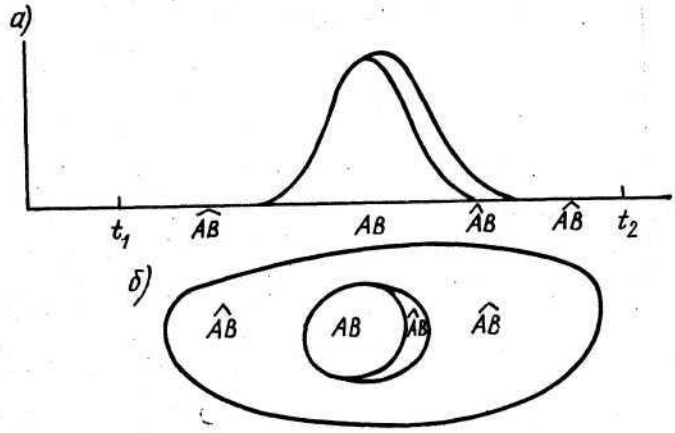


Рис. 17. Интерпретация случая 7 зависимости между качественными признаками:

а — хронологическая; *б* — географическая

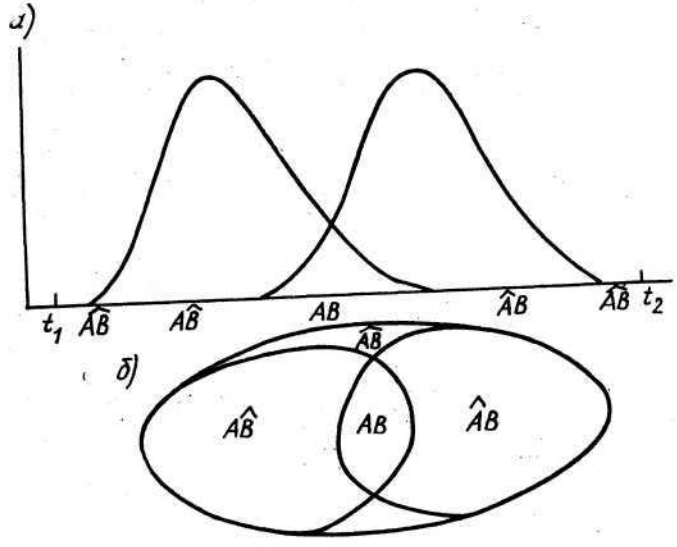


Рис. 18. Интерпретация случая 8 зависимости между качественными признаками:

а — хронологическая; *б* — географическая

цательную связь между «A» и «B», т. е. их систематическую невстречаемость. Например:

	A	$\bar{A}$	
B	50	250	300
$\bar{B}$	190	10	200
	240	260	500

$Q=-0{,}77$; $K=-0{,}98$; $\chi^2=295>3{,}84$; $Q'=0{,}03$.

Такой коэффициент может возникнуть за счет низких a и d и высоких b и d. Это показывает существенное расхождение периодов бытования (или ареалов) «A» и «B». Географически это можно представить рис. 18, *а*, хронологически — рис. 18, *б*.

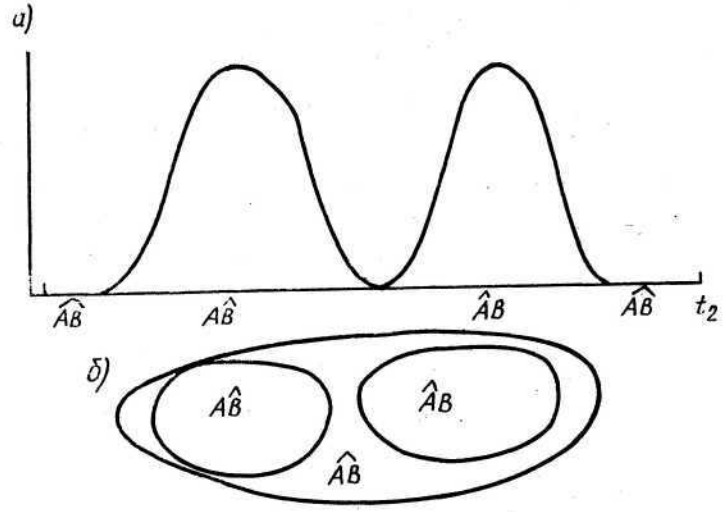


Рис. 19. Интерпретация случая 9 зависимости между качественными признаками:

а — хронологическая; *б* — географическая

9) Крайний случай представляет собой высокий отрицательный коэффициент, образующийся за счет того, что $a=0$. Например:

	A	$\bar{A}$	
B	0	200	200
$\bar{B}$	200	100	300
	200	300	500

$Q=-0{,}67$; $K=-1$; $\chi^2=222>3{,}84$; $Q'=0$.

Периоды бытования (или ареалы) «A» и «B» не перекрывают друг друга, разобщены, так что почти нет случаев встречи «A» и «B». На хронологической оси это может быть представлено так, как показано на рис. 19, *а*, географически — на рис. 19, *б*.

Таким образом, высокий отрицательный коэффициент связи свидетельствует о разновременности «A» и «B» (и о территориальной их разобщенности).

Коэффициент Q' становится близким к 0, но в случае 2 он тоже близок к 0. Однако случаи 2 и 8, 9 принципиально различны. В случае 2 имеет место случайная встречаемость «A» и «B», в случаях 8, 9 — закономерная их невстречаемость.

В приведенных случаях было рассмотрено отдельно влияние хронологического и территориального факторов. В действительности может иметь место сложное сочетание хронологического, географического и каких-то других факторов. На примере 25 мы видели, как, правильно произведя выборку (взяв правильное d), мы устранили влияние функционального фактора. Если выборка взята из одного места (из одного или группы близко расположенных друг к другу памятников), мы снимаем или сильно ослабляем воздействие географического фактора. Если выборка взята из памятников синхронных и с кратким периодом функционирования, то ослабляется действие хронологического фактора.

5. Связи между двумя качественными признаками с несколькими значениями (случай 3б)

Выше мы исследовали связи между одним значением одного качественного признака и другим значением другого качественного признака. Значение каждого признака рассматривалось как самостоятельный признак с двумя альтернативными значениями: присутствует или отсутствует. Усложним задачу: мы хотим узнать, связан ли один признак с несколькими значениями с другим признаком тоже с несколькими значениями.

Допустим, мы имеем k значений признака I и l значений признака II. Все случаи взаимовстречаемости каждого значения признака I с каждым значением признака II записаны в lk клетках табл. 23, где m_{ij} — число объектов с i-м значением I признака и j-м значением II признака, m_{i0} — число всех объектов с i-м значением

I признака, m_{0j} — число всех объектов с j-м значением II признака, n — число всех объектов.

Если мы будем два какие-либо значения признаков рассматривать как альтернативные признаки (например, i-е значение I признака и j-е значение II признака), то можем получить 4-польную таблицу для их взаимовстречаемости путем сжатия таблицы:

$$\begin{matrix} m_{ij} & m_{i0}-m_{ij} \\ m_{0j}-m_{ij} & n-m_{i0}-m_{0j}+m_{ij} \end{matrix}$$

Табл. 23 отражает «интегральную» связь между признаками I и II, полученная из нее сжатием 4-польная таблица отражает локальную связь между одним значением I признака и одним значением II признака.

Таблица 23

	1	2 ... j ... l	Σ
1	m_{11}	$m_{12}\dots m_{1j}\dots m_{1l}$	m_{10}
2	m_{21}	$m_{22}\dots m_{2j}\dots m_{2l}$	m_{20}
⋮			
i	m_{i1}	$m_{i2}\dots m_{ij}\dots m_{il}$	m_{i0}
⋮			
k	m_{k1}	$m_{k2}\dots m_{kj}\dots m_{kl}$	m_{k0}
Σ	m_{01}	$m_{02}\dots m_{0j}\dots m_{0l}$	n

Для проверки гипотезы о независимости двух признаков подсчитывается значение критерия

$$\chi^2=n\left[\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{m_{ij}^2}{m_{i0}m_{0j}}-1\right]$$

и сравнивается с критическим при $(k-1)(l-1)$ степенях свободы. При 4-польной таблице взаимовстречаемости эта формула совпадает с ранее выписанной (см. с. 96).

Техника вычислений следующая.

В каждой клетке записываются частота взаимовстречаемости (m_{ij}), ее квадрат и квадрат частоты, деленный на сумму частот по столбцу (m_{oj}). В итоговом столбце записываются суммы частот по строке (m_{io}), сумма по строке результатов деления квадратов частот на сумму частот по столбцу, а также результат деления второго числа на первое. Эти последние результаты в итоговой таблице суммируются (Σ). Далее получают $\chi^2=n(\Sigma-1)$. Можно подсчитать *коэффициент сопряженности Чупрова* для всех признаков:

$$R_{\text{ч}}^2=\frac{1}{n}\frac{\chi^2}{\sqrt{(k-1)(l-1)}}.$$

Этот коэффициент показывает усредненную, «интегральную» связь всех значений признака I со всеми значениями признака II; при этом некоторые пары значений этих признаков могут быть более тесно связаны, а некоторые связаны весьма слабо.

Если k сильно отличается от l, то предпочтительнее *коэффициент Крамера:*

$$R_{\text{Кр}}^2=\frac{\chi^2}{n\min[(l-1),\ (k-1)]}.$$

$R_{\text{Кр}}$ немного больше $R_{\text{ч}}$ при $l\neq k$.

При полной связи между признаками оба коэффициента равны 1, при отсутствии всякой связи между признаками, т. е. когда признаки полностью не зависимы, оба коэффициента превращаются в 0.

Коэффициенты Чупрова и Крамера дают хорошие результаты, если размер таблицы не меньше 5×5 и $n\geqslant 100$. Желательно, чтобы ожидаемые частоты в клетках таблицы, т. е. $\frac{m_{i0}}{n}m_{0j}$ не были очень малы (см. гл. I, § 16). В противном случае слишком редкие значения признака следует объединять друг с другом.

Пример 27.

Подсчитаны случаи взаимовстречаемости 15 значений признака «орнамент» и 13 значений признака «форма» для курильниц катакомбной культуры[13]. Объединив некоторые значения признаков и получив таблицу 6×6 для выполнения указанных выше требований, имеем следующие данные (табл. 24).

Подсчитываем:

$$\chi^2=419(1{,}60-1)=251{,}4,$$

что намного превышает критическое значение для 25 степеней свободы (37,6) при 95%-ном доверительном

[13] См.: *Егоров В. Г.* Классификация курильниц катакомбной культуры // СКМА. М., 1970. С. 157.

Таблица 24

Форма	Виды орнамента	1, 2, 4		3, 5		6, 7		8, 9		10, 13		11, 12, 14, 15		Σ	
1		18	324	20	400									38	11,41
			5,06		6,35										0,30
2		19	361	20	400									39	11,99
			5,64		6,35										0,31
4		11	121	11	121	32	1024	51	2601	38	1444	74	5476	217	136,95
			1,89		1,92		15,51		35,63		27,24		54,76		0,63
3, 5		7	49	3	9	13	169	8	64	5	25	15	225	51	7,06
			0,76		0,14		2,56		0,88		0,47		2,25		0,14
6, 7		1	1	4	16	12	144	7	49	6	36	8	64	38	4,43
			0,01		0,25		2,18		0,67		0,68		0,64		0,12
8, 9, 10,		8	64	5	25	9	81	7	49	4	16	3	9	36	3,69
			1,00		0,40		1,23		0,67		0,30		0,09		0,10
11, 12, 13															
Σ		64		63		66		73		53		100		419	1,6C

уровне. Следовательно, гипотеза о независимости признаков опровергнута с доверительным уровнем 0,95. Вместе с тем

$$R_{ч}^2 = R_{Кр}^2 = \frac{60}{\sqrt{5 \cdot 5}} = 0{,}12; \quad R_{ч} = \sqrt{0{,}12} \approx 0{,}35.$$

Мы получили малый коэффициент связи. Это говорит о том, что между орнаментом и формой курильниц была зависимость, но в среднем слабая.

Это не значит, что между отдельными значениями этих признаков связь тоже очень мала. Например, между разновидностью 1 формы («*A*») и разновидностями 3 и 5 орнамента («*B*») есть более значительная связь:

	A	$\bar{A}$	
B	20	43	63
$\bar{B}$	18	338	356
	38	381	419

Подсчитываем: $Q = 0{,}33$; $\chi^2 \approx 46 > 3{,}84$.

Таким образом, при малой в целом (в среднем) связи между признаками «форма» и «орнамент» между некоторыми значениями этих признаков связь устанавливается значимая. Это следует понимать в том смысле, что между многими другими парами значений признаков связь отсутствует или слаба и в среднем это столь сильно влияет на общий показатель связи, что не фиксирует сильной связи между признаками.

6. Теория информации. Информативность признака и связь между двумя качественными признаками с несколькими значениями (случаи 2б, 3б) и между одним и несколькими качественными признаками (случаи 2в, 3в)

Если мы знаем, каково значение одного признака на данном объекте, то как оценить, насколько много мы знаем о значении второго признака на том же объекте, т. е. с какой уверенностью мы может сказать, что значение второго признака будет таким-то, а не другим? Рассмотренные выше коэффициенты в определенной степени дают ответ на этот вопрос. Несколько в иной плоскости ставит этот вопрос теория информации.

Допустим, мы имеем n исходов какого-либо опыта который может наступить с той или иной вероятностью
1-й исход — с вероятностью p_1;
2-й исход — с вероятностью p_2;
...;
i-й исход — с вероятностью p_i;
...;
l-й исход — с вероятностью p_l.

При этом

$$\sum_{i=1}^{l} p_i = 1.$$

Согласно теории информации неопределенность (*энтропия*) этого опыта составляет величину

$$Э = -\sum_{i=1}^{l} p_i \log_a p_i.$$

Заметим, что если $p_1 = p_2 = p_3 \ldots = p_l = 1/l$, то энтропия оказывается максимальной и равна

$$Э_{max} = -\sum_{i=1}^{l} p_i \log_a p_i = -\log_a \frac{1}{l} = \log_a l.$$

После того как мы узнали, какой исход наступил в результате проведения этого опыта, неопределенность ликвидирована и мы получили *информацию* о нем, равную снятой неопределенности, энтропии, т. е.

$$И = Э.$$

Пусть археологический объект описывается L признаками. Представим, что из генеральной совокупности мы вынимаем случайно один археологический объект. Это тоже опыт. Исходами этого опыта будут те значения признака, которые могут оказаться на вынутом объекте.

Вероятность, что на объекте окажется 1-е значение признака I, равна $p(I)_1$.

Вероятность, что на объекте окажется 2-е значение признака I, равна $p(I)_2$, и т. д.

Следовательно, мы можем оценить неопределенность этого опыта, как

$$Э(I) = -\sum_{i=1}^{l_I} p(I)_i \log_a p(I)_i,$$

где $p(I)_i$ — вероятность появления на объекте i-го значения признака I; эту вероятность мы можем оценить как выборочную частность i-го значения признака I $(w(I)_i)$; l_I — число значений признака I.

Аналогично определяем Э(II) для признака II.

Аналогично определяем Э(I, II) для сочетаний значений признаков I и II:

$$Э(I, II) = -\sum_{i=1}^{l_{I,II}} p(I, II)_i \log_a p(I, II)_i,$$

где $p(I, II)_i$ — вероятность каждого i-го сочетания значений I и II признаков в выборке. Эту вероятность мы можем также оценить как выборочную частость i-го сочетания значений I и II признаков $(w(I, II)_i)$; $l_{I,II}$ — число этих сочетаний.

Взаимная информация признаков I и II определяется так:

$$И(II/I) = И(I/II = Э(I) + Э(II) - Э(I, II),$$

И (II/I) или И (I/II) — это средний объем информации о признаке I, содержащийся в значениях признака II (и наоборот), или, другими словами, мера, показывающая, насколько хорошо значения I признака определяются значениями II признака и наоборот.

Эта информация тем больше, чем больше взаимосвязь признаков I и II.

И (I/II) обращается в 0, когда между признаками нет статистической связи. Если связь между этими признаками имеется и И (I/II) $\neq 0$, то можно оценить, существенна она или несущественна, с помощью критерия χ^2. Если И (I/II) $> \chi^2/2n$ для критического значения χ^2 при доверительном уровне 0,95 (таблица III) при $(l_I - 1) \times (l_{II} - 1)$ степенях свободы, то гипотеза о несущественности связи между признаками отвергается (n — объем выборки).

Мера зависимости между двумя признаками I и II определяется как нормированная взаимная информативность (*коэффициент Райского*):

$$R_{инф}(I/II) = \frac{И(I/II)}{Э(I, II)}.$$

Так же можно оценить информативность признака I по отношению к любому числу других признаков, в частности ко всем остальным (случай 2в, 3в). В формуле следует заменить сочетание двух признаков I и II соче-

таниями (L—I) признаков, т. е. всех, кроме признака I. Тогда:

$$И(I/II,\ III \dots L) = Э(I) + Э(II,\ III \dots L) - Э(I,\ II \dots L);$$

$$R_{инф} = \frac{И(I/II,\ III \dots L)}{Э(I,\ II \dots L)},$$

где L — номер последнего признака.

Полученная информационная мера зависимости двух признаков двусторонняя, т. е.

$$R_{инф}(I/II) = R_{инф}(II/I).$$

Но если $Э(I) > Э(II)$, то значение признака I в большей степени определяет значения признака II, чем наоборот. Чтобы узнать относительную информативность признака I по отношению к признаку II, вводится односторонняя направленная мера зависимости путем деления взаимной информации И (I/II) на энтропию того признака, степень влияния на который другого признака хотят оценить:

$$R_{инф}(I \to II) = \frac{И(I/II)}{Э(II)};$$

$$R_{инф}(II \to I) = \frac{И(I/II)}{Э(I)}.$$

Теория информации может быть применена и для оценки степени неравномерности какого-либо распределения. Если объекты распределены по значениям какого-то признака равномерно, т. е. на каждое значение приходится одно и то же количество объектов, то неравномерность такого распределения будет минимальной ($p_i = 1/l$). Если на одно или несколько значений признака приходится большое количество объектов, а на другие — малое, то неравномерность будет большей. Если все объекты приходятся на одно значение признака — неравномерность будет максимальной. Энтропия равномерного распределения максимальна.

Но в действительности обычно $p_i \neq 1/l$, так как значения признака имеют разные частоты. *Коэффициент неравномерности* определяется так:

$$R_н = \frac{Э_{max} - Э}{Э_{max}} = \frac{\log_a l - Э}{\log_a l}.$$

Пример 28.

При раскопках Саркела-Белой Вежи (IX—XI вв.) и Саркельского грунтового могильника (XI в.) было найдено значительное количество бус. Дано распределение этих бус по значениям признака «техника» и «цвет»[14]. Требуется определить нормированную взаимную информативность признаков «техника» и «цвет», а также нормированную информативность каждого из этих признаков и их коэффициенты неравномерности.

1. Вычисляем энтропию распределения бус по значениям признака «техника» (Э(Т)). Все данные и промежуточные результаты приведены в табл. 25, где x —

Таблица 25

№	x	m	w	$-\lg w$	$-w \lg w$
1	0101	722	0,2538	0,6	0,151
2	0102	166	0,0583	1,23	0,072
3	0103	23	0,0081	2,09	0,017
4	0104	13	0,0046	2,34	0,011
5	0106	1	0,0004	3,45	0,001
6	0107	5	0,0017	2,76	0,005
7	0108	56	0,0197	1,70	0,034
8	0109	1	0,0004	3,45	0,001
9	0110	38	0,0133	1,87	0,025
10	0116	3	0,0011	2,98	0,003
11	0118	401	0,1409	0,85	0,120
12	0119	1	0,0004	3,45	0,001
13	0201	1163	0,4087	0,39	0,159
14	0202	1	0,0004	3,45	0,001
15	0204	1	0,0004	3,45	0,001
16	0206	26	0,0091	2,04	0,018
17	0208	1	0,0004	3,45	0,001
18	0402	1	0,0004	3,45	0,001
19	0501	2	0,0007	3,15	0,002
20	0502	4	0,0014	2,85	0,004
21	0503	11	0,0038	2,41	0,009
22	0601	3	0,0011	2,98	0,003
23	0602	2	0,0007	3,15	0,002
24	0801	177	0,0622	1,21	0,075
25	0802	6	0,0021	2,68	0,006
26	0803	3	0,0010	2,98	0,003
27	0901	13	0,0046	2,34	0,011
28	1902	1	0,0004	3,45	0,001
Σ		2845	1,0000		0,738

[14] См.: *Федоров-Давыдов Г. А.* Археологическая типология и процесс типообразования. Математические методы в социально-экономических и археологических исследованиях. М., 1981.

значение признака (в данном случае код, обозначающий технику), m — число бус с данным значением признака, w — частость этого значения признака.

Таким образом, Э(Т) = 0,738.

2) Аналогично вычисляем энтропию распределения бус по значениям признака «цвет» — Э(Ц).

Таблица 26

№	x	m	w	$-\lg w$	$-w \lg w$
1	0101	112	0,0394	1,41	0,055
2	0103	388	0,1364	0,86	0,118
...	...	...	...		
47	1105	1	0,0004	3,45	0,001
Σ		2845	1,0000		0,965

Таким образом, Э(Ц) = 0,965.

3) Так же вычисляем энтропию распределения бус по сочетаниям всех значений признаков «техника» (x_1) и «цвет» (x_2), встречающихся на бусах, — Э(Т, Ц).

Таблица 27

x_1	x_2	m	w	$-\lg w$	$-w \lg w$
0101	0101	5	0,0017	2,76	0,005
0101	0103	293	0,1030	0,99	0,102
...	...	...	...		
1002	0107	1	0,0004	3,45	0,001
Σ		2845	1,0000		1,331

Таким образом, Э(Т, Ц) = 1,331.

4) Вычисляем нормированную взаимную информативность признаков «техника» и «цвет»:

$$R_{\text{инф}}(\text{Т/Ц}) = \frac{0{,}738 + 0{,}965 - 1{,}331}{1{,}331} \approx 0{,}28.$$

5) Вычисляем направленную информативность признака «техника» на признак «цвет»:

$$R_{\text{инф}}(\text{Т} \to \text{Ц}) = \frac{0{,}738 + 0{,}965 - 1{,}331}{0{,}965} \approx 0{,}38.$$

6) Вычисляем нормированную информативность признака «цвет». Для этого проводим следующие вычисления.

Энтропия признака «цвет» уже нам известна: Э(Ц) = = 0,965. Вычисляем энтропию распределения бус по сочетаниям значений всех признаков (см. табл. 28, где x_1 — код значений признака «количество отверстий», x_2 — код значений признака «материал», x_3 — код значений признака «техника», x_4 — код значений признака «форма и пропорции», x_5 — код значений признака «цвет», x_6 — код значений признака «прозрачность», x_7 — код значений признака «размер»).

Таблица 28

x_1	x_2	x_3	x_4	x_5	x_6	x_7	m	w	$-\lg w$	$-w \lg w$
1	101	0101	010101	0105	1	2	2	0,0007	3,15	0,002
1	101	0101	010101	0106	2	2	1	0,0004	3,45	0,001
...	...	...	...	...	...	...	...	...	...	...
1	702	0802	030207	0112	2	3	1	0,0004	3,45	0,001
Σ							2845	1,0000		1,699

Таким образом, Э (по всем признакам) = 1,699.

Аналогично вычисляем энтропию распределения бус по сочетанию всех значений всех признаков, за исключением признака «цвет».

Таблица 29

x_1	x_2	x_3	x_4	x_6	x_7	m	w	$-\lg w$	$-w \lg w$
1	101	0101	010101	1	2	2	0,0007	3,15	0,002
1	101	0101	010101	2	2	18	0,0063	2,20	0,014
...	...	...	...	...	...	...	...	...	...
1	702	0802	36207	2	3	1	0,0004	3,45	0,001
Σ						2845	1,0000		1,461

Таким образом, Э (по всем признакам без Ц) = 1,461.

Теперь можем определить нормированную информативность признака «цвет» по отношению ко всем другим признакам:

$$R_{\text{инф}}(\text{Ц/все остальные признаки}) = \frac{0{,}965 + 1{,}461 - 1{,}699}{1{,}699} \approx 0{,}43.$$

7) Вычисляем коэффициент неравномерности признаков «техника» и «цвет»:

$$R_{\text{н}}(T) = \frac{\text{Э}(\text{T})_{max} - \text{Э}(\text{T})}{\text{Э}(\text{T})_{max}} = \frac{\lg 28 - 0{,}738}{\lg 28} =$$

$$= \frac{1{,}447 - 0{,}738}{1{,}447} \approx 0{,}49.$$

$$R_{\text{н}}(\text{Ц}) = \frac{\text{Э}(\text{Ц})_{max} - \text{Э}(\text{Ц})}{\text{Э}(\text{Ц})_{max}} = \frac{\lg 47 - 0{,}965}{\lg 47} =$$

$$= \frac{1{,}672 - 0{,}965}{1{,}672} \approx 0{,}42.$$

Коэффициенты $R_{\text{инф}}$ и $R_{\text{н}}$ могут быть использованы, например, при археологической классификации (см. гл. V, § 6; гл. VI).

7. Связь между отдельными значениями количественных признаков (случай 1а)

Выше рассмотренные способы установления взаимосвязи между отдельными значениями признаков относились к качественным признакам. Но эти способы могут быть использованы и для работы с количественными признаками.

Характер и теснота взаимосвязи одного количественного признака с другим может быть установлена и оценена корреляционным или дисперсионным анализом (случай 1б). Часто возникает необходимость исследовать взаимосвязь каждого интервального значения одного количественного признака с каждым значением другого тоже количественного признака (случай 1а). Прибегают к разбиению количественных признаков на интервалы (об этом см. гл. I, § 18). Полученные интервалы рассматриваются как значения качественных ранжированных признаков.

Теперь можно применить те коэффициенты сопряженности Q и ассоциации K и критерии χ^2 и Фишера, которые мы применяли для качественных признаков (случай 3а). См. примеры 29, 30.

8. Связи между качественными и количественными признаками и их значениями (случаи 4а, 4б, 4в)

При археологических исследованиях часто возникает задача исследования или «локальной» связи между значениями признаков, или «обобщенной» связи между признаками, или связи одного признака с несколькими другими при наличии как количественных, так и качественных признаков. Эта ситуация возникает очень часто, так как большинство археологических объектов описываются признаками разного рода, исчисляемыми в разных шкалах.

Во всех этих случаях количественный признак можно превратить в качественный ранжируемый признак путем разбиения его на интервальные значения. В случае, когда задача стоит в выявлении и оценке «локальной» связи между одним значением качественного и одним значением количественного признака (случай 4а), могут быть применимы те методы исследования, которые применялись в случае 3а (коэффициенты ассоциации (K) и сопряженности (Q), критерии χ^2 и Фишера.

В случае, когда задача стоит в выявлении и оценке «обобщенной» связи между качественным и количественным признаками (случай 4б), применимы те методы исследования, которые применялись в случае 3б (коэффициент сопряженности $R_{\text{ч}}$, $R_{\text{Кр}}$ и коэффициент нормированной информативности $R_{\text{инф}}$). Применим также дисперсионный анализ (пример 31).

В случае, когда задача стоит в выявлении и оценке связи одного качественного или количественного признака с несколькими количественными и качественными признаками (случай 4в), применимы те методы исследования, которые использовались в случае 3в (коэффициент нормированной информативности $R_{\text{инф}}$).

Пример 29.

При исследовании курганов некрополя Бирка IX—X вв. в Скандинавии были замерены их диаметры и высоты. Выяснилось, что мерный признак — высота курганов x может быть разбит на 4 группы по равным интер-

валам x_1 — 0,3—0,6 м; x_2 — 0,6—0,9 м; x_3 — 0,9—1,2 м; x_4 — 1,2—1,5 м. Другой мерный признак — диаметр курганов y — был разбит в соответствии с полигоном его распределения на 4 группы с неравными интервалами: y_1 — 3—4,5 м; y_2 — 4,5—6,5 м; y_3 — 6,5—7,5 м; y_4 — 7,5—9,5 м.

Требуется установить, имеется ли какая-либо существенная связь между значениями одного признака и значениями другого. Были подсчитаны взаимовстречаемости этих значений (табл. 30).

Таблица 30

	y_1	y_2	y_3	y_4	Σ
x_1	22	186	8	7	223
x_2	3	36	10	18	67
x_3	0	9	17	50	76
x_4	0	8	3	17	28
Σ	25	239	38	92	394

Таблица 31

	y_1	y_2	y_3	y_4
x_1	0,16	0,53	—0,23	—0,54
x_2	—0,03	—0,06	0,08	0,04
x_3	—0,13	—0,49	0,21	0,49
x_4	—0,07	—0,18	0,01	0,24

Далее был подсчитан коэффициент сопряженности Q. A — значение признака «диаметр»; $\bar{A}$ — все другие значения этого признака; B — значение признака «высота»; $\bar{B}$ — все другие значения этого признака.

Получились следующие значения этого коэффициента (табл. 31).

Выявилась в целом связь, близкая к прямой зависимости: малые высоты связаны с малыми диаметрами, большие высоты — с большими диаметрами. Вывод оказался почти тривиальным, его следовало ожидать до исследования. Но все же наметились некоторые отклонения — средние диаметры 4,5—6,5 оказались связанными с малыми высотами [15].

Теснота связи между диаметром x и высотой курганов y может быть выражена коэффициентом корреляции. Для быстрой его оценки применим коэффициент Фехнера. Для этого берем в качестве значений признаков середины интервалов и подсчитываем среднюю высоту $\bar{x}$ и средний диаметр $\bar{y}$ (табл. 32, 33).

[15] См.: *Лебедев Г. С.* Разновидности обряда трупосожжения в могильнике Бирка // СКМА. М., 1970.

$$a=22+186+10+18+17+50+3+17=323;$$

$$b=8+7+3+36+8+9=71;$$

$$\Phi=\frac{323-71}{323+71}=\frac{252}{394}\approx 0{,}64.$$

Таблица 32

x	m_x	w_x	xw_x
0,45	223	0,57	0,256
0,75	67	0,17	0,128
1,05	76	0,19	0,200
1,35	28	0,07	0,094
Σ	394	1,00	0,678

Таблица 33

y	m_y	w_y	yw_y
3,75	25	0,06	0,225
5,5	239	0,61	3,355
7,0	38	0,10	0,700
8,5	92	0,23	1,955
Σ	394	1,00	6,235

Теснота связи оказалась средней, не очень большой из-за указанных выше отклонений.

Пример 30.

При изучении лепной керамики из городища Камно было получено несколько керамических показателей. Два из них — высотно-горловинный (ФБ) и широтно-горловинный (ФБ) дали полигон, близкий к нормальному распределению, что было интерпретировано как свидетельство случайности отклонений от стандарта. Полигон распределения указателя профилировки шейки сосуда (ФГ) дал три вершины и было установлено три

Таблица 34

Признак	1	2	3	4	5	6	7	8
1. ФГ<0,03	x							
2. 0,03≤ФГ≤0,15	x	x						
3. ФГ>0,15	x	x	x					
4. 0,05≤ФЖ≤0,17	19	17	2	x				
5. ФЖ>0,17	47	50	40	x	x			
6. ФК<10	12	3	1	7	9	x		
7. 10≤ФК≤50	47	39	43	34	106	x	x	
8. ФК<50	9	9	1	5	14	x	x	x

интервала: меньше 0,03; 0,04—0,15; больше 0,15. Распределение указателя выпуклости плечика (ФЖ) позволило определить два интервала (0,05—0,17; больше 0,17), распределение указателя конфигурации придонной части (ФК) дало три интервала (меньше 10; 10—50; больше 50). Таким образом, получены были следующие исходные данные (табл. 34).

Требуется установить, имеются ли между значениями этих признаков существенные связи. Были подсчитаны взаимовстречаемости этих значений признаков, например, для двух — ФГ и ФК (табл. 35).

Таблица 35

	ФК<10	10<ФК<50	ФК>50	Σ
ФГ<0,03	12	47	9	68
0,04≤ФГ≤0,15	3	39	9	51
ФГ>0,15	1	43	1	45
Σ	16	129	19	164

Далее были подсчитаны для каждой пары значений признаков коэффициенты ассоциации K и более предпочтительный в данном случае коэффициент сопряженности Q в скобках (табл. 36).

Таблица 36

	ФК<10	10<ФК<50	ФК>50
ФГ<0,03	0,66 (0,22)	—0,45 (—0,19)	0,13 (0,04)
0,04≤ФГ≤0,15	—0,35 (—0,09)	—0,09 (—0,04)	0,38 (0,04)
ФГ>0,15	—0,73 (—0,16)	0,78 (0,25)	—0,77 (—0,18)

В данном случае связь между количественными признаками не приближается к прямой зависимости. Малое ФГ связано с малым ФК, но большие ФГ не только не связаны с большими ФК, но дают отрицательную связь. Зато среднее ФК связано с большим ФГ.

Здесь разбиение значений признака на интервал и исследование взаимовстречаемости значений по этим интервалам дает дополнительную существенную информацию о взаимосвязи значений признаков. Эта информация была использована исследователем для группировки сосудов [16].

Пример 31.

При исследовании Безводнинского могильника V—VIII вв. в Горьковской обл. было сделано заключение, что размеры могил, в частности ширина, зависят от вида погребения, т. е. от того, имело ли место: 1 — трупоположение, 2 — трупосожжение или 3 — вторичное захоронение [17]. Проверим это с помощью дисперсионного анализа (табл. 37).

Таблица 37

Ширина ямы, см	Середина интервала, см	Факторный признак (вид погребения) 1	2	3	Σ
26—50	38	1	0	0	1
51—75	63	29	7	5	41
76—100	88	21	4	9	34
101—125	113	5	1	3	9
126—150	138	1	0	0	1
151—175	163	0	0	2	2
		57	12	19	88

Подсчитываем D_0:

$$\bar{y}=\frac{38\cdot1+63\cdot41+88\cdot34+113\cdot9+138\cdot1+163\cdot2}{88}=80{,}6;$$

$$D_0=(80{,}6-38)^2\,1+(80{,}6-63)^2\,41+(80{,}6-88)^2\,34+$$
$$+(80{,}6-113)^2\,9+(80{,}6-138)^2\,1+(80{,}6-163)\,2=$$
$$=42698{,}88.$$

Подсчитываем D_Φ:

$$\bar{y}_1=\frac{38\cdot1+63\cdot29+88\cdot21+113\cdot5\cdot138\cdot1}{57}=77{,}5;$$

[16] См.: *Белецкий С. В.* К вопросу о возможностях статистического изучения лепной керамики // Новое в применении физико-математических методов в археологии. М., 1979.

[17] См.: *Краснов Ю. А.* Безводненский могильник. М., 1980. С. 16—18.

$$\bar{y}_2 = \frac{63 \cdot 7 + 88 \cdot 4 + 113 \cdot 1}{12} = 75{,}5;$$

$$\bar{y}_3 = \frac{63 \cdot 5 + 88 \cdot 9 + 113 \cdot 3 + 163 \cdot 2}{19} = 93{,}3;$$

$$D_\phi = (80{,}6 - 77{,}5)^2\, 57 + (80{,}6 - 75{,}5)^2\, 12 + (80{,}6 - 93{,}3)^2 \times \times 19 = 3924{,}40.$$

Подсчитываем D_c:

$$D_1 = (77{,}5 - 38)^2\, 1 + (77{,}5 - 63)^2\, 29 + (77{,}5 - 88)^2\, 21 + + (77{,}5 - 113)^2\, 5 + (77{,}5 - 138)^2\, 1 = 19934{,}25;$$

$$D_2 = (75{,}5 - 63)^2\, 7 + (75{,}5 - 88)^2\, 4 + (75{,}5 - 113)^2\, 1 = 3125;$$

$$D_3 = (93{,}3 - 63)^2\, 5 + (93{,}3 - 88)^2\, 9 + (93{,}3 - 113)^2\, 3 + + (93{,}3 - 163)^2\, 2 = 15723{,}71;$$

$$D_c = 19934{,}25 + 3125 + 15723{,}71 = 38782{,}96.$$

Определяем теперь отношения:

$$\eta_\phi = \frac{D_\phi}{D_0} = \frac{3924{,}4}{42698{,}88} = 0{,}09, \quad \eta_c = \frac{D_c}{D_0} = \frac{38782{,}96}{42698{,}88} = 0{,}91;$$

$$\eta_\phi + \eta_c = 0{,}09 + 0{,}91 = 1;$$

$$K_c = 88 - 3 = 85, \quad K_\phi = 3 - 1 = 2;$$

$$F = \frac{D_\phi}{K_\phi} : \frac{D_c}{K_c} = \frac{3924{,}4}{2} : \frac{38782{,}96}{85} = 4{,}3.$$

Мы видим, что η_ϕ очень мало, F больше табличного значения при соответствующих степенях свободы, хотя η_c — велико. Следовательно, факторный признак (вид погребения) как то влияет на результативный, но в малой степени. Сделанное предварительное без применения статистических методов заключение подтвердилось.

В заключение сведем все рассмотренные выше 12 случаев парных (двухмерных) связей между значениями признаков и самими признаками в табл. 38.

Мы видели, что для исследования взаимосвязей между явлениями предлагается несколько способов с различными коэффициентами связи. Ни один из них не является идеальным, и дело исследователя выбрать наиболее подходящий для его материала и целей исследования.

	1	2	3	4
	Связи между количественными признаками	Связи между ранжированными и качественными признаками	Связи между неранжированными качественными признаками	Связи между количественными и качественными признаками
а. Связи между одним значением одного признака и одним значением другого	Коэффициенты сопряженности и ассоциации (Q и K) и критерии χ^2 и Фишера			
б. Связи между всеми значениями одного признака и всеми значениями другого признака	Коэффициенты корреляции r, непараметрические коэффициенты Φ, R_C, R_K	Непараметрические коэффициенты R_C, R_K	Коэффициенты сопряженности $R_Ч$ и $R_{Кр}$, нормированная информативность $R_{инф}$, критерий χ^2	Коэффициенты сопряженности $R_Ч$ и $R_{Кр}$, нормированная информативность $R_{инф}$, критерий χ^2 и дисперсионный анализ
в. Связи между одним признаком и несколькими другими признаками	Коэффициент множественной корреляции	Нормированная информативность $R_{инф}$		

В заключение этой главы заметим, что выявленные статистическим путем связи между объектами, явлениями, признаками не всегда отражают прямую взаимозависимость между ними. Может иметь место так называемая «связь сопутствия», когда оба явления или объекта не прямо связаны друг с другом, а опосредованно. Может иметь место явление «ложной корреляции», возникающей от неправильного измерения или выбора признаков.

Существенно искажает реальные связи между объектами неправильное объединение объектов в группы, неправильное разбиение признака на интервалы. Возможна также ситуация, при которой, получив какие-то значения показателя связи между явлениями или объектами, исследователь не удовлетворен ими или они оказываются негодными для интерпретации. Это заставляет исследователя корректировать список признаков и их значений, а иногда и выбор показателя связи.

ГЛАВА IV

ПРОСТРАНСТВО ПРИЗНАКОВ И СХОДСТВО ОБЪЕКТОВ

1. Пространство количественных признаков

Возьмем систему двух прямоугольных осей координат. На одной оси будем откладывать значения признака (x_1), на другой — второго признака (x_2). В местах пересечения перпендикуляров, восставленных в этих точках, получим точки. Каждая такая точка плоскости будет соответствовать объекту, на котором совмещается пара отложенных на осях значений этих двух признаков.

Если мы возьмем систему трех прямоугольных осей координат и будем на них откладывать значения трех признаков (x_1, x_2, x_3), то полученные точки будут соответствовать объектам, на которых совмещаются тройки отложенных на осях значений этих трех признаков. Эти точки будут находиться в трехмерном пространстве Если признаков много (L), то мы должны построить

L-мерное пространство с L осями координат, представляющее собой математическую абстракцию. На каждой оси будут откладываться значения одного какого-либо признака. Тогда в этом L-мерном гиперпространстве точка ($x_1, x_2, \ldots, x_L$) будет соответствовать объекту, на котором совместилось именно данное сочетание значений $x_1, x_2, \ldots, x_L$ всех L признаков, описывающих объект.

2. Коэффициент сходства объектов, описанных мерными непрерывными признаками

Пусть мы имеем два объекта, которые описываются двумя признаками. У одного значение первого признака x_{11}, второго — x_{12}, у другого объекта значение первого признака x_{21}, второго — x_{22}. Тогда, по теореме Пифагора, расстояние между двумя точками 1 (x_{11}, x_{12}) и 2 (x_{21}, x_{22}) (рис. 20), соответствующими этим объектам, равно

$$g=\sqrt{(x_{11}-x_{21})^2+(x_{12}-x_{22})^2}.$$

Распространим это на случай, когда объект описывается не двумя, а L признаками. Объект 1 описывается следующими значениями количественных мерных признаков: x_{11}, x_{12}, ..., x_{1i}, ..., x_{1L}. Объекту 2 соответствуют следующие значения тех же признаков x_{21}, x_{22}, ..., x_{2i}, ..., x_{2L}. Тогда расстояние между двумя точками, соответствующими этим объектам, равно

$$g=\sqrt{\sum_{l=1}^{L}(x_{1l}-x_{2l})^2}.$$

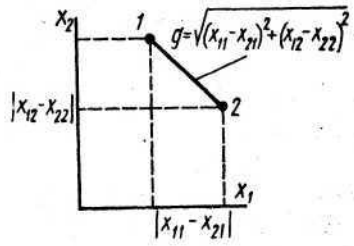


Рис. 20. Вычисление эвклидова расстояния между двумя объектами

Это так называемое *эвклидово расстояние*. Его можно использовать как меру сходства между парой объектов.

Однако при таком подсчете эвклидова расстояния возможны существенные ошибки в определении меры сходства двух объектов. Признаки могут иметь большие среднеквадратические отклонения (σ) и малые среднеквадратические отклонения. Может случиться так, что расхождения по признакам с большим σ «забьют», «за-

темнят» расхождения по признакам с малым σ, а последние могут быть так же важны в описании объектов, как и первые. Расстояние между объектами будет определяться главным образом по признакам с большим среднеквадратическим отклонением.

Для того чтобы избежать этой особенности, следует каждое значение признака нормировать, т. е. разделить на среднеквадратическое отклонение этого признака. Теперь эвклидово расстояние будет подсчитываться следующим образом:

$$g_н = \sqrt{\sum_{l=1}^{L} \left(\frac{x_{1l} - x_{2l}}{\sigma_l}\right)^2}$$

Эвклидово расстояние само по себе может служить мерой сходства двух объектов. Чем оно больше, тем сходство двух объектов меньше. Но можно также рассматривать коэффициент сходства, заключенный между 0 и 1, который имеет вид

$$G = 1 - \frac{g_н}{g_{н(\max)}},$$

так что $G=1$, когда $g_н=0$, т. е. когда объекты совпадают по всем L признакам. $G=0$, когда $g_н=g_{max}$, т. е. когда точки удалены друг от друга на максимально возможное в данных условиях расстояние. В первом случае полное сходство, во втором — полное несходство по всем признакам.

Этот коэффициент сходства удобен для сравнения двух объектов, описанных только мерными (непрерывными) признаками.

Пример 32.

Кувшины нижневолжских золотоордынских городищ описаны следующими признаками: x_1 — диаметр горла свеху, x_2 — диаметр горла под венчиком, x_3 — диаметр горла у основания, x_4 — наибольший диаметр тулова, x_5 — диаметр дна, x_7 — высота горла, x_9 — общая высота сосуда, x_{10} — высота наибольшего расширения тулова, x_{12} — высота прикрепления ручки внизу. Три сосуда имеют следующие значения этих признаков (в мм):

Номер сосуда	x_1	x_2	x_3	x_4	x_5	x_7	x_9	x_{10}	x_{12}
60	90	85	90	130	90	50	200	120	140
56	80	80	60	120	80	60	170	80	100
68	90	70	100	170	90	50	270	100	130

Подсчитываем по всем объектам (75 сосудов) дисперсии для каждого признака: $\sigma_1^2=373$, $\sigma_2^2=353$, $\sigma_3^2=599$, $\sigma_4^2=2457$, $\sigma_5^2=348$, $\sigma_7^2=520$, $\sigma_9^2=13\,421$, $\sigma^2_{10}=2771$, $\sigma^2_{12}=6304$.

Подсчитываем теперь эвклидовы расстояния между объектами:

$$g_н^2(60,56) = \frac{(90-80)^2}{373} + \frac{(85-80)^2}{353} + \frac{(90-60)^2}{599} + \frac{(130-120)^2}{2457} + \frac{(90-80)^2}{348} + \frac{(50-60)^2}{520} + \frac{(200-170)^2}{13421} + \frac{(120-80)^2}{2771} + \frac{(140-100)^2}{6304} = 3{,}26;$$

$$g_н(60,56) = 1{,}80.$$

$$g_н^2(60,68) = \frac{(90-90)^2}{373} + \frac{(85-70)^2}{353} + \frac{(90-100)^2}{599} + \frac{(130-170)^2}{2457} + \frac{(90-90)^2}{348} + \frac{(50-50)^2}{520} + \frac{(200-270)^2}{13421} + \frac{(120-100)^2}{2771} + \frac{(140-130)^2}{6304} = 1{,}98;$$

$$g_н(60,68) = 1{,}41.$$

$$g_н^2(56,68) = \frac{(80-90)^2}{373} + \frac{(80-70)^2}{353} + \frac{(60-100)^2}{599} + \frac{(120-170)^2}{2457} + \frac{(80-90)^2}{348} + \frac{(60-50)^2}{520} + \frac{(170-270)^2}{13421} + \frac{(80-100)^2}{2771} + \frac{(100-130)^2}{6304} = 5{,}75;$$

$$g_н(56,68) = 2{,}40.$$

Определяем вариационный размах по всем признакам:

$P_1 - 120 - 33$; $P_2 - 110 - 30$; $P_3 - 130 - 30$; $P_4 - 300 - 75$; $P_5 - 110 - 30$; $P_7 - 120 - 15$; $P_9 - 740 - 65$; $P_{10} - 450 - 30$; $P_{12} - 230 - 25$.

$$g^2_{н(max)} = \frac{(120-33)^2}{373} + \frac{(110-30)^2}{353} + \frac{(130-30)^2}{599} +$$

$$+ \frac{(300-75)^2}{2457} + \frac{(110-30)^2}{348} + \frac{(120-15)^2}{520} + \frac{(740-65)^2}{13421} +$$

$$+ \frac{(450-30)^2}{2771} + \frac{(230-25)^2}{6304} = 219{,}59;$$

$$g_н(\max) = 14{,}82.$$

Коэффициенты сходства получаются следующие:

$$G(60,56) = 1 - \frac{1{,}80}{14{,}82} \approx 0{,}88;$$

$$G(60,68) = 1 - \frac{1{,}41}{14{,}82} \approx 0{,}91;$$

$$G(56,68) = 1 - \frac{2{,}40}{14{,}82} \approx 0{,}84.$$

Теперь можно сказать, что все три сосуда мало отличаются друг от друга, так как имеют попарно довольно высокий коэффициент сходства. Но все же 56-й сосуд от 60-го и 68-го более отличается, чем 60-й от 68-го.

3. Коэффициенты сходства объектов, описанных счетными (дискретными) признаками

Если объект описывается количественными счетными признаками, может быть применен тот же коэффициент сходства. Он может быть использован также и в случае, когда объекты описываются одновременно как мерными, так и счетными признаками.

Пример 33. Допустим, мы сравниваем состав инвентаря нескольких погребений. Нас не интересует, какие виды тех или иных предметов попали в могилу. Нас интересует, какие вещи и в каком количестве были положены в погребение. Именно функциональные категории этих вещей характеризуют и обряд погребения, и социальное, профессиональное или имущественное положение погребенного, его пол, возраст и т. п. Для этого сравнения каждую функциональную категорию предметов будем рассматривать как признак, а число предметов этой категории будем считать значением соответствующего признака.

В Безводнинском могильнике среди 120 ненарушенных человеческих погребений выберем три могилы

Таблица 39—40

Категории вещей	Погребения 15	81	153
1. Сюльгамы	1	0	1
2. Кельты	1	0	1
3. Ножи	1	1	1
4. Поясные пряжки	2	0	1
5. Копья	1	0	1
6. Кресала или их принадлежности	1	2	0
7. Браслеты	0	0	1
8. Поясные накладки	0	0	10

Таблица 41

x	m	w	xw	$\lvert x-\bar{x}\rvert$	$(x-\bar{x})^2 w$
0	57	0,47	0	0,89	0,37
1	33	0,28	0,28	0,11	0,00
2	19	0,16	0,32	1,11	0,20
3	8	0,07	0,21	2,11	0,31
4	3	0,02	0,08	3,11	0,19
Σ	120	1,00	0,89		1,07

(№ 15, 81 и 153). Состав их инвентаря см. в табл. 39—40. Заметим, что некоторые признаки похожи на альтернативные качественные признаки с двумя значениями — присутствует (1) или отсутствует (0). Но на самом деле это количественные признаки — ведь они могут принимать значения и больше чем 1.

Требуется установить степень сходства между этими погребениями попарно. Подсчитаем дисперсии каждого признака по всем 120 погребениям (см. табл. 41, где x — число сюльгам, m — количество погребений с этим числом сюльгам, w — доля этих погребений среди всех погребений, т. е. частость).

Таким образом, для сюльгам $\sigma^2 = 1{,}07$. Аналогично подсчитаем: для кельтов $\sigma^2 = 0{,}13$; для ножей $\sigma^2 = 0{,}26$; для пряжек $\sigma^2 = 0{,}66$; для копий $\sigma^2 = 0{,}15$; для кресал $\sigma^2 = 0{,}17$; браслетов $\sigma^2 = 0{,}43$; для поясных накладок $\sigma^2 = 143{,}18$. Далее определим максимально возможное эвклидово расстояние. Вариационный размах для сюльгам $P = 4$, для кельтов $P = 1$, для ножей $P = 2$, для пряжек $P = 4$, для копий $P = 1$, для кресал $P = 2$, для браслетов $P = 3$, для поясных накладок $P = 68$.

$$g^{2*}_н(\max) = \frac{4^2}{1{,}07} + \frac{1^2}{0{,}13} + \frac{2^2}{0{,}26} + \frac{4^2}{0{,}66} + \frac{1^2}{0{,}15} + \frac{2^2}{0{,}17} + \frac{3^2}{0{,}43} +$$

$$+ \frac{68^2}{143{,}18} = 145{,}694;$$

$$g_н(\max) = 12{,}07.$$

Далее определим эвклидовы расстояния между погребениями:

$$g_н^2(15,81) = \frac{1^2}{1,07} + \frac{1^2}{0,13} + \frac{0^2}{0,26} + \frac{2^2}{0,66} + \frac{1^2}{0,15} + \frac{1^2}{0,17} + \frac{0^2}{0,43} + \frac{0^2}{143,18} = 27,24;$$

$$g_н(15,81) = 5,22.$$

$$g_н^2(15,153) = \frac{0^2}{1,07} + \frac{0^2}{0,13} + \frac{0^2}{0,26} + \frac{1^2}{0,66} + \frac{0^2}{0,15} + \frac{1^2}{0,17} + \frac{1^2}{0,43} + \frac{10^2}{143,18} = 10,42;$$

$$g_н(15,153) = 3,23.$$

$$g_н^2(81,153) = \frac{1^2}{1,07} + \frac{1^2}{0,13} + \frac{0^2}{0,26} + \frac{1^2}{0,66} + \frac{1^2}{0,15} + \frac{2^2}{0,17} + \frac{1^2}{0,43} + \frac{10^2}{143,13} = 43,36;$$

$$g_н(81,153) = 6,58.$$

Теперь можно подсчитать коэффициенты сходства:

$$G(15,81) = 1 - \frac{5,22}{12,07} = 1 - 0,43 = 0,57;$$

$$G(15,153) = 1 - \frac{3,23}{12,07} = 1 - 0,27 = 0,73;$$

$$G(81,153) = 1 - \frac{6,58}{12,07} = 1 - 0,54 = 0,46.$$

Таким образом, мы видим, что погребения мало сходны друг с другом. При этом погребения № 15, 81 и 81, 153 мало похожи по составу инвентаря, а погребения № 15 и 153 обнаруживают бóльшее сходство. Подсчет коэффициентов сходства по составу инвентаря подтверждает интуитивное заключение, что погребения относятся к разным имущественным группам (81 — «бедное», 15 — «среднее», 153 — «богатое»), и при этом дополнительно показывает, что «среднее» и «богатое» погребения более сходны между собой, чем «бедное» и «среднее». Естественно, что наиболее несходными оказались «бедное» и «богатое» погребения[1].

[1] См.: *Краснов Ю. А.* Безводнинский могильник. М., 1980.

Часто в археологии возникает такая ситуация, когда важно сравнить два объекта не по абсолютному количеству каких-либо признаков в них, а по относительному количеству однородных признаков. Например, в одном погребении 5 бус какого-либо вида, а в другом — 6 бус этого же вида. Казалось бы, наблюдается большое сходство по бусам этого вида. Но в первом погребении всего бус 100, а в другом только 40. Приходится сравнивать доли бус определенного типа среди всех бус в погребении, т. е. делать значения признаков сопоставимыми. В таком случае следует сравнивать не цифры 5 и 6, а цифры 5% и 15%. Аналогично, при сравнении двух стоянок или двух слоев одного памятника по содержанию в них находок приходится сравнивать не абсолютные численности видов находок, а их доли среди находок определенной категории. Существенным представляется, чтобы за 1 принималось общее число однородных объектов. Каждый вид данной категории предметов становится признаком. Доля этого вида среди всех предметов данной категории является значением признака.

Пусть в объекте 1 содержится первого признака x_{11} единиц, второго признака x_{12} единиц и т. д., а в объекте 2 содержится первого признака x_{21} единиц, второго x_{22} единиц и т. д. Следовательно, объект 1 описывается следующими значениями признаков:

$$x_{11}, x_{12}, \ldots, x_{1l}, \ldots, x_1L; \quad \sum_{l=1}^{L} x_{1l} = n_1,$$

а объект 2 описывается следующими значениями L признаков:

$$x_{21}, x_{22}, \ldots, x_{2l}, \ldots, x_{2L}; \quad \sum_{l=1}^{L} x_{2l} = n_2.$$

Выразим значения каждого признака у обоих объектов через их доли:

$$v_{1l} = \frac{x_{1l}}{n_1}; \quad v_{2l} = \frac{x_{2l}}{n_2}.$$

Это делает признаки сопоставимыми. Тогда объект 1 будет описываться следующими значениями признаков:

$$v_{11}, v_{12}, \ldots, v_{1l}, \ldots, v_{1L};$$

объект 2 будет описываться следующими значениями тех же признаков:

$$v_{21}, v_{22}, \ldots, v_{2l}, \ldots, v_{2L}.$$

Другими словами, мы принимаем за 1 общее количество однородных предметов (одной категории — бус) в объекте и подсчитываем, какую долю от этого количества составляет каждый вид этого предмета. Мерой сходства объектов в этом случае также может быть нормированное эвклидово расстояние. Но обычно в археологии применяют более простую меру сходства: сумму абсолютных значений разностей долей для каждого признака:

$$s = \sum_{l=1}^{L} |v_{1l} - v_{2l}|.$$

Объекты 1 и 2 удалены на максимальное расстояние тогда, когда на объекте 1 имеются только такие признаки, которые полностью отсутствуют на объекте 2 и обратно. Тогда $s_{max}=2$. Коэффициент сходства будет иметь вид

$$S = 1 - \frac{s}{2}.$$

$S=1$, когда $s=0$, т. е. когда у объектов доли каждого признака совпадают, и объекты полностью сходны по этим признакам друг с другом.

$S=0$, когда $s=s_{max}=2$, т. е. объекты максимально несходны друг с другом.

Подсчет S может быть облегчен. Суммируются величины $\min(v_{1i}, v_{2i}) = v_i(\min)$, каждая из которых представляет минимальное значение из пары значений v_{1i} и v_{2i}.

Для каждого признака берут его долю в том объекте, у которого она оказывается наименьшей, и затем суммируют их. Именно в таком виде этот коэффициент применяется наиболее часто. Иногда его называют *«индексом родственности»* (ИР). Этот коэффициент был предложен В. Робинсоном.

Значения показателя сходства при малых объемах выборки могут показывать случайные расхождения или сближения. Как проверить существенность расхождения, т. е. отвергнуть нулевую гипотезу о случайности расхождения значений признаков в двух объектах?

Проверка существенности расхождения двух объектов по значениям счетных (дискретных) признаков, представленных в них, в данном случае сводится к проверке существенности расхождения двух распределений, известных нам по двум выборкам. Для этого вычисляется критерий

$$\chi^2 = \left[\sum_{l=1}^{L} \frac{(x_{1l}n_2 - x_{2l}n_1)^2}{x_{1l} + x_{2l}} \right] \frac{1}{n_1, n_2}.$$

Если χ^2 превышает табличное значение для соответствующего числа степеней свободы, которое равно $L-1$, то нулевая гипотеза об отсутствии существенного различия между объектами отвергается. При использовании этого критерия желательно, чтобы все математические ожидания (в предположении об однородности распределений) для значений каждого признака x были бы не очень малы (см. об этом гл. I, § 16). Эти математические ожидания подсчитываются как $n_1 \frac{x_{1l} + x_{2l}}{n_1 + n_2}$ для первой выборки и $n_2 \frac{x_{1l} + x_{2l}}{n_1 + n_2}$ для второй.

Если ожидаемые величины частот малы, то следует несколько признаков, по которым производится сравнение объектов, объединить.

Пример 34.

На двух участках Самаркандской палеолитической стоянки были подсчитаны кремневые орудия по классам [2].

Эти данные сведены в табл. 42, где x_1 — число орудий данного типа на I участке, v_1 — доля орудий этого типа среди всех орудий I участка, x_2 — число орудий данного типа на II участке, v_2 — доля орудий этого типа среди всех орудий II участка.

Требуется по этим данным установить степень сходства I и II участков Самаркандской палеолитической стоянки.

$$S = 1 - \frac{1}{2} \sum_{l=1}^{L} |v_{1l} - v_{2l}| = 1 - \frac{1}{2} (0{,}0042 + 0{,}0037 +$$

$$+ 0{,}1275 + 0{,}0071 + 0{,}0038 + 0{,}1057 + 0{,}0446 + 0{,}1046 +$$

$$+ 0{,}0218 + 0{,}0697 + 0{,}0713 + 0{,}0075 + 0{,}0107 + 0{,}0037 +$$

$$+ 0{,}0074 + 0{,}0074 + 0{,}1028 + 0{,}0037) =$$

[2] См.: *Холюшкин Ю. П.* Проблемы корреляции позднепалеолитических индустрий Сибири и Средней Азии. Новосибирск, 1981.

$$= \sum_{l=1}^{L} v_l(\min) = 0{,}0367 + 0{,}0037 + 0{,}1078 + 0{,}0149 + 0{,}0074 +$$

$$+ 0{,}0855 + 0{,}0074 + 0{,}0367 + 0{,}0149 + 0{,}0186 + 0{,}0588 +$$

$$+ 0{,}0074 + 0{,}0260 + 0{,}0000 + 0{,}0000 + 0{,}0000 + 0{,}2206 +$$

$$+ 0{,}0000 = 0{,}6464.$$

Небольшое сходство есть. Возникает вопрос, существенно ли различие? Применим критерий χ^2. Для его корректного применения объединим признаки 2, 14, 15, 16, 18

Таблица 42

Классы орудий	x_1	v_1	x_2	v_2
1. Резцы	11	0,0409	5	0,0367
2. Скребки-резцы	1	0,0037	1	0,0074
3. Скребки	29	0,1078	32	0,2353
4. Скребла	4	0,0149	3	0,0220
5. Скребки-скобели	3	0,0112	1	0,0074
6. Выемчатые	23	0,0855	26	0,1912
7. Остроконечники	14	0,0520	1	0,0074
8. Пластины с ретушью	38	0,1413	5	0,0367
9. Ножи на отщепах	4	0,0149	5	0,0367
10. Нуклевидные скребки	5	0,0186	12	0,0883
11. Галечные прочие	35	0,1301	8	0,0588
12. Чопперы	4	0,0149	1	0,0074
13. Тесла и топоры	7	0,0260	5	0,0367
14. Долота из галек	1	0,0037	0	0,0000
15. Рубильца	2	0,0074	0	0,0000
16. Мотыжки	0	0,0000	1	0,0074
17. Пластины	87	0,3234	30	0,2206
18. Плитки	1	0,0037	0	0,0000
Σ	269	1,0000	136	1,0000

(для каждого из них математические ожидания оказываются меньше 1): χ^2 оказывается примерно равен 53,5, что больше табличного (22,40) для доверительного уровня 0,95 и 13 степеней свободы. Следовательно, критерий χ^2 нулевую гипотезу об отсутствии существенного различия отвергает.

4. Пространство качественных признаков. Коэффициенты сходства

Если объект описывается качественными признаками, мы также можем представить эти объекты как точки в воображаемом L-мерном гиперпространстве признаков, где L равно количеству признаков. На осях координат в этом случае будет откладываться качественные признаки и соответствующие им точки будут соответствовать объектам. Так как расположение качественных неранжируемых признаков на осях координат произвольно, то и точки окажутся в произвольных положениях одна относительно другой, и геометрически расстояния между ними также будут произвольными, не будут свидетельствовать о сходстве объектов или об их различии. Поэтому для вычисления расстояния между объектами удобно значения качественных признаков рассматривать как самостоятельные альтернативные признаки с двумя значениями 1 и 0 (присутствует или отсутствует на объекте).

Пусть объект 1 описывается следующими значениями качественных альтернативных признаков: $x_{11}, x_{12}, \ldots, x_{1i}, \ldots, x_{1L}$. Объект 2 описывается следующими значениями тех же качественных альтернативных признаков: $x_{21}, x_{22}, \ldots, x_{2i}, \ldots, x_{2L}$. При этом x_1 и x_2 принимают значения только 0 или 1.

Представим пространство с L осями координат (L — общее количество альтернативных признаков). На каждой оси откладывается одно из значений альтернативного признака — 0 или 1. Тогда расстояние между точками по аналогии с пространством количественных признаков будет равно:

$$c_1 = \sqrt{\sum_{l=1}^{L} (x_{1l} - x_{2l})^2},$$

$x_{1i} - x_{2i} = 1$, если признак присутствует на одном и отсутствует на другом объекте. Тогда коэффициент сходства

$$C_1 = 1 - \frac{c_1}{c_1(\max)} = 1 - \frac{c_1}{\sqrt{L}},$$

так как

$$c_1(\max) = \sqrt{L},$$

Мы видим, что $C_1 = 1$, когда $c_1 = 0$, т. е. для каждого признака наблюдается, что он или одновременно присутству-

ет, или отсутствует на обоих объектах — имеется полная схожесть.

$C_1=0$, когда $c_1=\sqrt{L}$, т. е. когда для каждого признака наблюдается присутствие его на одном объекте и отсутствие на другом, т. е. имеется полная несхожесть.

Для выявления сходства и выражения его численно часто прибегают к другому пониманию расстояния между точками в пространстве качественных признаков. Это понимание расстояния между точками-объектами заключается в том, что расстояние между ними признается тем меньшим (т. е. сходство тем более сильным), чем большее количество признаков совпадает по своим значениям. При этом может быть применен коэффициент, в котором определяется так называемое расстояние «по Хеммингу»:

$$c_2=\sum_{i=1}^{L}|x_{1i}-x_{2i}|=c_1^2;$$

$$C_2=1-\frac{c_2}{L}=\frac{a+d}{L},\quad c_1=\sqrt{\frac{a+d}{L}},$$

где a — количество признаков, присутствующих на обоих объектах; d — количество признаков, одновременно отсутствующих на обоих объектах. Мы применяем здесь обозначения 4-польной таблицы взаимовстречаемости (см. гл. III, § 4), только вместо объектов подсчитываем численности признаков.

Заметим, что коэффициенты C_1 и C_2 учитывают сходство между объектами и по одновременному отсутствию какого-либо признака на них. Но таких признаков может быть, вообще говоря, для объектов названо бесконечное количество. Поэтому в расчетах ограничиваются только теми признаками, которые имеются хоть на одном объекте из всех объектов, подвергаемых попарному сравнению.

Отметим еще одну особенность коэффициентов C_1 и C_2. Чем больше блоки альтернативных признаков, получившихся из представления одного качественного признака множеством качественных альтернативных, и, следовательно, чем больше L, тем больше будут коэффициенты C_1 и C_2. Это вполне соответствует интуитивному пониманию сходства. Действительно, совпадение значений какого-либо качественного признака, тем более важно и

значительно для оценки сходства, чем более «вариабелен» этот признак, т. е. чем больше может он принимать значений.

Часто употребляют коэффициент сходства вида

$$C_3=\frac{a^2}{L_1L_2},$$

где a — количество совпадающих признаков на обоих объектах; L_1 — количество признаков, присутствующих на одном объекте; L_2 — количество признаков, присутствующих на другом объекте. Этот коэффициент также дает представление о сходстве объектов. Однако он имеет недостаток: он не учитывает общее количество признаков (L). Если какие-то признаки отсутствуют и на I, и на II объектах, это также может сближать эти объекты. Коэффициент C_3 полностью аналогичен коэффициенту Q', рассмотренному выше для 4-польной таблицы.

Нет какого-то правила, предписывающего применять тот или иной показатель сходства для качественных признаков в тех или иных ситуациях. Чаще всего применяется коэффициент C_3. Но его применение допустимо только в тех случаях, когда сходство по отсутствию признаков не принимается во внимание, признано несущественным. Если же признается, что сходство по одновременному отсутствию признаков на объектах существенно, важно, то следует использовать один из коэффициентов C_1, C_2. В этом случае может быть рекомендован коэффициент C_2 как интуитивно более простой.

При сравнении объектов важно правильно определить, по каким признакам проводится сравнение и подсчет степени сходства. Например, при исследовании сходства памятников по орнаментации керамики иногда берут качественный альтернативный признак «керамики данного вида на памятнике» — присутствует или отсутствует. Это неправильно. Следует брать количественный признак со значениями — доля керамики с данной орнаментацией среди всей керамики данного памятника. Ведь значение «присутствует такой-то орнамент» на памятнике может означать и то, что вся керамика данного памятника имеет этот вид орнамента, и то, что только на единичных черепках этот вид орнамента встречается. Ошибка в неправильном определении шкалы признака аналогична ошибке, рассмотренной во введении (пример 3).

Пример 35.

По группам памятников неолитической эпохи[3] фрагменты керамики с разными видами орнаментации распределялась таким образом (табл. 43, в %)

Таблица 43

Группа памятников	гребенчатый	ямочный	нарезной	без орнамента	накольчатый	другие виды
1. Смоленская	57,7	17,6	14,7	7,9	0,0	2,1
2. Могилево-Гомельская	57,6	22,9	5,7	13,8	0,0	0,0
3. Гомельская	53,0	35,0	1,8	2,9	5,3	2,0
4. Киево-Черниговская	36,5	15,0	10,4	31,5	4,7	1,9
5. Днепро-Донецкая	20,9	0,0	30,0	9,6	38,0	1,5

Таблица 44

	1	2	3	4	5
1	1	0,89	0,77	0,72	0,45
2		1	0,81	0,71	0,36
3			1	0,63	0,32
4				1	0,47
5					1

Получены следующие парные коэффициенты сходства S для каждой пары групп (табл. 44).

Таблица 45

Группы памятников	гребенчатый	ямочный	нарезной	без орнамента	накольчатый	другие виды
1. Смоленская	1	1	1	1	0	1
2. Могилевско-Гомельская	1	1	1	1	0	0
3. Гомельская	1	1	1	1	1	1
4. Киево-Черниговская	1	1	1	1	1	1
5. Днепро-Донецкая	1	0	1	1	1	1

[3] См.: *Третьяков В. П.* Неолитическая керамика и археологические культуры (по материалам лесной и лесостепной зоны европейской части СССР) // СА. 1984. № 1.

А если встречаемость каждого мотива орнамента будет определена только двумя значениями «присутствует» или «отсутствует» среди керамики, то коэффициенты сходства будут совсем иные. Таблица исходных данных будет иметь следующий вид (табл. 45).

По данным табл. 45 получены следующие парные коэффициенты сходства между группами $C_3 = \frac{a^2}{L_1 L_2}$ (табл. 46).

Таблица 46

	1	2	3	4	5
1	1	0,8	0,83	0,83	0,64
2		1	0,67	0,67	0,45
3			1	1,00	0,83
4				1	0,83
5					1

Если в первом случае наиболее близки друг к другу первые четыре группы, а пятая несколько отдалена от них, то во втором случае мы получили, что наиболее тесно связаны последние три группы. Очевидно, что упрощенный способ подсчета степени сходства с заменой количественного признака на качественный альтернативный неприемлем. Сам автор в начале своего исследования проводит подсчет степени сходства керамических групп по орнаменту по количественному признаку — по доле керамики с той или иной орнаментацией среди всей

Таблица 47

Памятник	Виды орнаментации															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Хуторская	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1
Кряжская	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0

керамики. Но в конце своего исследования он переходит к подсчету сходства групп керамики по наличию или отсутствию орнамента. Например, керамика двух стоянок сравнивается так (табл. 47):

Подсчитаем коэффициент сходства:

$$C_3 = \frac{10^2}{10 \cdot 12} = 0{,}83.$$

Этот коэффициент правильно отражает степень сходства, если доли керамики той или иной орнаментации на стоянках одинаковы хотя бы примерно. Но если, скажем, один орнаментальный узор присутствует на 50% керамики на Хуторской и только на 5% керамики из Кряжской, то полученный показатель сходства существенно исказит действительную ситуацию.

ГЛАВА V

ГРУППИРОВКИ ОБЪЕКТОВ И ПРИЗНАКОВ

Группировкой применительно к археологии мы будем называть разбиение совокупности объектов или признаков на группы, классы, кластеры таким образом, чтобы ни один объект из одной группы не оказался бы одновременно в другой группе, т. е. полученные группы не должны пересекаться. Получаемые в результате группировки группы должны включать в себя объекты до определенной степени однородные, т. е. сходные между собой до какого-то принятого исследователем уровня. Сходство между объектами, входящими в разные группы, должно быть достаточно малым. Другими словами, при группировке стремятся получить максимальную степень разнородности разных классов. Таким образом, группировка осуществляется путем объединения объектов, сильно сходных или сильно связанных между собой. Поэтому необходимым этапом группировки является построение так называемой матрицы каких-либо коэффициентов сходства или связи между каждой парой объектов.

Методы построения группировки могут быть организованы таким образом, чтобы вычислять не все коэффициенты связи сразу, а только те, которые необходимы на каждом этапе, что важно с точки зрения экономики памяти ЭВМ, но в нашем изложении нет необходимости специально рассматривать эту возможность.

1. Матрицы

Выше (в гл. III и IV) были изложены различные методы, устанавливающие связи (взаимозависимости) и

сходства между объектами, значениями признаков, самими признаками. Эти взаимозависимости и сходства устанавливаются главным образом попарно, хотя есть способы оценки и тройственных связей, и связей более высокого порядка. Наиболее удобным способом записи и графического выражения попарных связей являются таблицы (*матрицы*) и *графы*.

Исходные матрицы представляют собой таблицы, где первичные данные располагаются следующим образом:

$$\begin{array}{l} a_{11}a_{12}\ldots a_{1j}\ldots a_{1l}, \\ a_{21}a_{22}\ldots a_{2j}\ldots a_{2l}, \\ \ldots\ldots\ldots\ldots \\ a_{i1}a_{i2}\ldots a_{ij}\ldots a_{il}, \\ \ldots\ldots\ldots\ldots \\ a_{k1}a_{k2}\ldots a_{kj}\ldots a_{kl}, \end{array}$$

где a_{ij} — j-е значение признака на i-м объекте; l — число признаков, k — число объектов. В случае, если признак количественный, то a_{ij} — число; если ранговый качественный, то a_{ij} — номер значения признака; если качественный, то a_{ij} — условный код значения признака. В последнем случае каждое значение признака может быть представлено в виде альтернативного признака с двумя значениями: 1 — присутствует, 0 — отсутствует.

Если исследуются связи между значениями двух признаков, то составляется матрица взаимовстречаемости, имеющая тот же вид, но в ней a_{ij} — частота соединения i-го значения одного признака с j-м значением другого признака, k — число значений одного признака, l — число значений другого признака. Таблица взаимовстречаемости значений одного качественного признака со всеми значениями другого качественного признака уже составлялась для подсчета коэффициента Чупрова $R_ч$ или Крамера $R_{Кр}$. Сжатием ее мы получали 4-польную таблицу взаимовстречаемости для двух альтернативных качественных признаков (см. гл. III, § 5). Все это матрицы, т. е. двухвходовые таблицы без пробелов.

После того как подсчитаны попарные коэффициенты связи, матрица взаимовстречаемости превращается в матрицу парных связей между значениями признаков. Она имеет тот же вид, но в ней a_{ij} — коэффициент (показатель) связи между i-м значением одного признака и j-м значением другого признака, k — число значений одного признака, l — число значений другого признака.

Если исследуются двусторонние связи или сходство между объектами, или связи между значениями признака, их удобно представлять треугольной матрицей. Треугольная матрица — это половина квадратной матрицы, симметричная относительно так называемой главной диагонали, идущей из левого верхнего угла в правый нижний.

$$a_{11}a_{12}\ldots a_{1j}\ldots a_{1k},$$
$$a_{22}\ldots a_{2j}\ldots a_{2k},$$
$$\ldots\ldots$$
$$a_{ij}\ldots a_{ik},$$
$$\ldots$$
$$a_{kk},$$

где a_{ij} — показатель сходства или связи i-го и j-го объектов или значений признака. Все a_{11}, a_{22}, ..., a_{kk}, т. е. члены главной диагонали, равны 1, так как каждый объект (или значение признака) полностью похож сам на себя или связан сам с собой; k — число объектов или значений признака.

2. Графы и методы группировки

Записанные в матрицу попарные показатели связи наглядно могут быть выражены *графом*. В зависимости от смысла строк матрицы связей каждая *вершина* графа — признак, значение признака или объект; линия, соединяющая вершины (*ребро*), — связь или сходство. Граф наглядно отражает структуру.

Всякая структура — это набор элементов и связей между ними. Элементы — это вершины графа, связи — это его ребра. Но степень силы связи различна. Очевидно, следует как-то учесть эти различия. Прибегают к способу построения нескольких последовательных графов или составления одного графа с показом различных по степени силы связей разными условными линиями или цифрами вдоль линий. Сначала показывают более сильные связи, потом дорисовывают связи послабее, далее — все более и более слабые связи, пока не достигают минимальных показателей. На каждом уровне ослабления связей граф показывает различные степени объединения вершин (признаков, объектов). Мы как бы «по слоям» исследуем структуру связей между признаками или объектами. Для конкретных целей выбираются одно или несколько пороговых значений соответствующего показателя связи (сходства). Он должен быть выбран для данного материала исходя именно из его организации, степени упорядоченности и не может быть определен постоянно, раз навсегда для всякого исследования.

Такое выявление структуры парных связей есть по сути дела группировка объектов или признаков. Сначала объединяются признаки и объекты, более сильно связанные или более похожие друг на друга, затем — менее сильно связанные, менее похожие друг на друга и т. д.

После того как произведено разделение на группы взаимосвязанных объектов (*плеяд*), можно разными способами оценить качество такого разбиения. Один из них — подсчет так называемого *индекса связанности плеяд* для каждой группы:

$$J_1 = \frac{2\,[k\,(\beta) - n + 1]}{(n-1)\,(n-2)},$$

где n — количество элементов в плеяде; $k(\beta)$ — число связей, которые по своему значению выше некоторого порогового значения β между объектами внутри плеяды. Если между каждой парой значений элементов внутри группы имеется связь, то $J_1 = 1$, т. е. связанность плеяды максимальная.

Этот показатель дает возможность оценить, как много связей между объектами находится внутри группы. Разбиение будет тем лучше, чем больше связей внутри группы (*внутренних*) и чем меньше связей будет между объектами из разных групп (*внешних*). Качество разбиения, таким образом, может быть также оценено простым отношением количества внешних связей к количеству внутренних:

$$J_2 = 1 - \frac{k_1}{k_2},$$

где k_1 — число внешних связей; k_2 — число внутренних связей. $J_2 = 1$, когда внешние связи отсутствуют вообще; $J_2 = 0$, когда число внешних связей равно числу внутренних. При этом желательно, чтобы связи внутри группы были бы как можно более сильными, а связи между объектами из разных групп — как можно более слабыми. Но при этом нельзя ограничиваться одними только очень сильными связями, так как они не дают еще группировки. Представим, например, граф при очень высоком пороге β (объединенными оказываются только две

пары объектов). Получатся две группы по два объекта в каждой, связанные максимальным числом связей (для группы из двух объектов это число связей равно 1), а остальные объекты окажутся необъединенными. Такая группировка будет трюизмом и не вскроет структуры материала. Исследователь должен найти такой порог β, чтобы большая часть объектов была охвачена группировкой, чтобы связи были не слишком слабыми, чтобы группы, получаемые при таком пороге, были бы достаточно компактными и, наконец, чтобы межгрупповые связи были бы достаточно малочисленными.

Существуют математические средства оптимизации группировки по всем этим направлениям для оценки качества разбиения на группы, которые здесь не рассматриваются. Может быть рекомендован такой способ нахождения оптимального порога β. Постепенно понижая порог для показателя силы связи, получают несколько графов. На графе, соответствующем очень сильным связям, объединяются в группы только несколько объектов. На графах с некоторым понижением порога объединение охватывает большее число объектов, и, наконец, после какого-то очередного понижения порога граф покрывается беспорядочной сетью ребер. При этом качество разбиения резко падает. Очевидно, перед этим шагом следует определить наилучший для данного материала порог для показателя связи между объектами. Все показатели ниже этого порога расцениваются как несущественные и отбрасываются. Назовем этот способ *простой последовательной группировкой.*

В некоторых случаях при группировке сохраняются ребра, соответствующие слабым связям, и отбрасываются ребра, соответствующие сильным связям. Это делается с учетом ошибок выборки, искажающих показатели связей. Кроме того, учитывают при этом, что несколько слабых связей в тех случаях, когда они создают значительную компактность разбиения, полезней, чем одна сильная, противоречащая выявляемой картине.

Мы будем рассматривать такие группировки, в которых принимаются во внимание, учитываются и переносятся на граф в виде ребер только сильные (выше избранного порога β) связи между каждой парой объектов. Слабые связи при этом оказываются полезными для выявления взаимоотношений групп между собой.

Другим способом группировки является метод *корреляционных плеяд.* Он заключается в следующем. В матрице показателей связи находят максимальный недиагональный элемент. Выписывают номер строки, в которой он расположен. Под каждым значением ставят номер строки, над каждым значением — номер столбца. Находят столбец, в котором располагается это максимальное значение. Из строки, номер которой совпадает с номером этого столбца, выписывают те значения, которые больше значений предыдущей строки в соответствующем столбце. Этим большим значениям придается внизу индекс новой строки. Там, где значения меньше, оставляют значения предыдущей строки, сохраняя их старый индекс. В новой строке находят наибольшее значение и повторяют процедуру со строкой, номер которой соответствует номеру того столбца, в котором располагается это новое максимальное значение, и т. д. Максимальные значения в каждой строке помечают, и под помеченными значениями уже не пишут новые. Так перебирают все строки и столбцы. Получают серию помеченных значений, и они связывают пары объектов. Таких связей на n объектов оказывается $n-1$.

Переносят эти связи на граф. Получается граф, где все n вершин связаны $n-1$ ребрами. Такой граф называют *«деревом максимальной длины»*, так как сумма показателей связи, соответствующих этим ребрам, максимальна из всех возможных сумм $n-1$ показателей связи, соединяющих все n вершин. Затем находят те ребра, которые из этих $n-1$ ребер означают наименьшие коэффициенты связи, и опускают их. Тем самым «разрезают» дерево на несколько «плеяд».

3. Построение эволюционных рядов

Покажем, как группировка и построение графа могут быть использованы для формализованного построения эволюционного ряда.

Пример 36.

Орнаментация крышек котлов из Пенджикента была описана 20 признаками, имеющими только два значения — присутствует или отсутствует, т. е. альтернативными качественными признаками. Получены следующие данные[1] (см. исходную матрицу — табл. 48). По этим

[1] См.: *Каменецкий И. С., Маршак Б. И., Шер Я. А.* Указ. соч. С. 56 и далее.

Таблица 48

Номер орнаментальной композиции	Элементы орнамента																			
	86	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
II	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
III	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	1	0	1	0	0
IV	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
V	0	1	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
VI	0	1	1	0	0	0	0	1	1	0	1	0	0	0	0	1	1	0	0	0
VII	0	1	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
VIII	1	1	0	1	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0
IX	1	1	0	1	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0
X	0	1	0	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0
XI	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0
XII	0	1	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0
XIII	0	1	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	1
XIV	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Таблица 49

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
I	1	0,67	0,10	0,00	0,00	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
II	0,67	1	0,07	0,00	0,00	0,05	0,00	0,05	0,04	0,00	0,00	0,00	0,00	0,00
III	0,10	0,07	1	0,60	0,45	0,26	0,04	0,03	0,02	0,03	0,11	0,05	0,04	0,00
IV	0,00	0,00	0,60	1	0,75	0,19	0,07	0,05	0,04	0,05	0,19	0,08	0,07	0,00
V	0,00	0,00	0,45	0,75	1	0,32	0,20	0,14	0,12	0,14	0,14	0,25	0,20	0,00
VI	0,07	0,05	0,26	0,19	0,32	1	0,26	0,18	0,16	0,18	0,08	0,32	0,26	0,14
VII	0,00	0,00	0,04	0,07	0,20	0,26	1	0,46	0,40	0,46	0,11	0,45	0,36	0,00
VIII	0,00	0,05	0,03	0,05	0,14	0,18	0,46	1	0,87	0,51	0,18	0,32	0,26	0,00
IX	0,00	0,04	0,02	0,04	0,12	0,16	0,40	0,87	1	0,45	0,16	0,28	0,22	0,00
X	0,00	0,00	0,03	0,05	0,14	0,18	0,46	0,51	0,45	1	0,33	0,32	0,26	0,00
XI	0,00	0,00	0,11	0,19	0,14	0,08	0,11	0,18	0,16	0,33	1	0,14	0,11	0,14
XII	0,00	0,00	0,05	0,08	0,25	0,32	0,45	0,32	0,28	0,32	0,14	1	0,80	0,00
XIII	0,00	0,00	0,04	0,07	0,20	0,26	0,36	0,26	0,22	0,26	0,11	0,80	1	0,00
XIV	0,00	0,00	0,00	0,00	0,00	0,14	0,00	0,00	0,00	0,00	0,14	0,00	0,00	1

данным требуется формализованным образом построить эволюционный ряд. Для этого был посчитан по всем парам коэффициент сходства C_3 (см. гл. IV) и составлена матрица парных коэффициентов сходства, симметричная относительно главной диагонали (табл. 49).

Далее составляется граф (рис. 21) по следующему принципу: вначале соединяются объекты, у которых коэффициент сходства равен 1,0—0,71, затем те, у которых коэффициент сходства равен 0,70—0,41, и т. д. Легко заметить, что коэффициенты сходства выше 0,40 группируют объекты на три четкие группы I—II, III—V и VII—X, XII, XIII. Коэффициенты сходства 0,40—0,21 соединяют две последние группы через объект VI, а коэффициенты ниже 0,20 покрывают

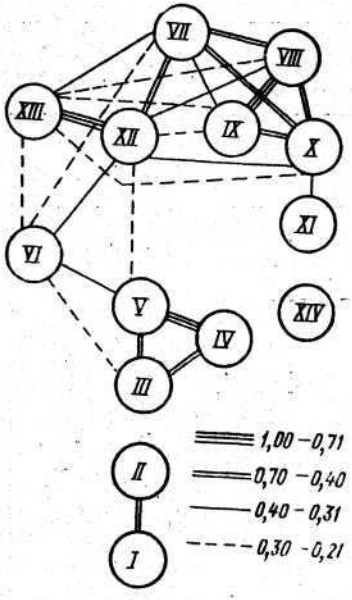


Рис. 21. Граф сходства объектов-крышек из Пенджикента

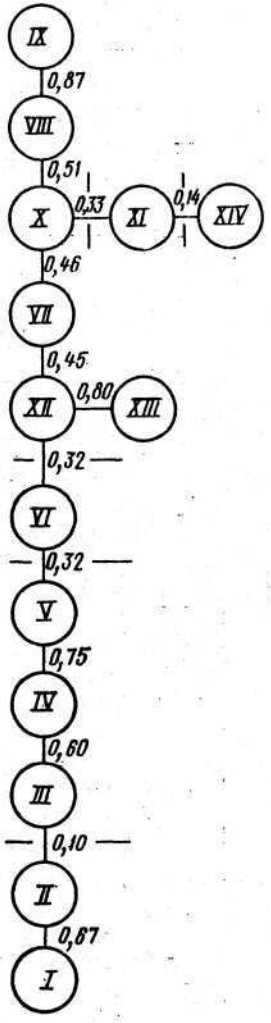


Рис. 22. Граф сходства объектов-крышек из Пенджикента (по методу корреляционных плеяд)

весь граф сетью линий, затушевывающих эту картину. На основании этого наблюдения можно считать значения коэффициента выше 0,40 существенными для группировки объектов. Слабые (ниже 0,40) показатели сходства также имеют значение. Между первой и второй группами нет слабых связей (0,40—0,21), а между второй и третьей — есть. Значит, первая группа дальше отстоит от второй, чем вторая от третьей.

Применим к матрице попарных коэффициентов сходства метод корреляционных плеяд (для этого выписана

Таблица 50

I	II	III	IV	V	VI	VII	IX	X	XI	XII	XIII	XIV
0,00	0,05	0,03	0,05	0,14	0,18	0,46	0,87	0,51	0,18	0,32	0,26	0,00
VIII	VIII	VIII	VIII	VIII	VIII	VIII	VIII	VIII	VIII	VIII	VIII	VIII
0,00	0,05	0,03	0,05	0,14	0,18	0,45		0,51	0,18	0,32	0,26	0,00
VIII	VIII	VIII	VIII	VIII	VIII	VIII		VIII	VIII	VIII	VIII	VIII
0,00	0,05	0,03	0,05	0,14	0,18	0,46			0,33	0,32	0,26	0,00
VIII	VIII	VIII	VIII	VIII	VIII	X			X	VIII	VIII	VIII
0,00	0,05	0,04	0,07	0,20	0,26				0,33	0,45	0,36	0,00
VIII	VIII	VII	VII	VII	VII				X	VII	VII	VIII
0,00	0,05	0,05	0,08	0,25	0,32				0,33		0,80	0,00
VIII	VIII	XII	XII	XII	XII				X		XII	VIII
0,00	0,05	0,05	0,08	0,25	0,32				0,33			0,00
VIII	VIII	XII	XII	XII	XII				X			VIII
0,00	0,05	0,11	0,19	0,25	0,32							0,14
VIII	VIII	XI	XI	XII	XII							XI
0,07	0,05	0,26	0,19	0,32								0,14
VI	VIII	VI	XI	VI								XI
0,07	0,05	0,45	0,75									0,14
VI	VIII	V	V									XI
0,07	0,05	0,60										0,14
VI	VIII	IV										XI
0,10	0,07											0,14
III	III											XI
0,10	0,07											
III	III											
	0,67											
	I											

вторая, симметричная часть матрицы). Находим максимальное значение коэффициента сходства (0,87). Оно расположено в VIII строке. Выписываем эту строку и над каждым значением коэффициента помещаем номер столбца, а под каждым значением — номер строки. В

данном случае под всеми значениями будет стоять VIII. Помечаем это максимальное значение. Оно стояло в IX столбце. Берем IX строку и выписываем те ее значения, которые больше значений VIII строки в соответствующих столбцах, помечая их теперь цифрой IX. В тех случаях, когда значения IX строки в соответствующем столбце ниже значений предыдущей строки (т. е. VIII строки), то оставляем значения этой предыдущей строки, помечая их прежним индексом (VIII). В данном случае все элементы IX строки оказались меньшими, чем элементы VIII строки, поэтому в новой строке все осталось из VIII строки. Снова находим максимальное значение строки — 0,51, которое оказалось в X столбце. Помечаем его. Повторяем операцию. Теперь из X строки в новой строке оказалось значение 0,46 в VII столбце. Продолжая эту процедуру, получаем следующий граф (рис. 22).

Теперь, если разрезать его в местах наиболее низких значений коэффициента сходства (0,32; 0,10; 0,14; 0,32; 0,33), то получим три плеяды I—II, III—V и VII, VIII, IX, X, XII, XIII. Три объекта остались вне группировки (VI, XI, XIV), что соответствует их промежуточному значению. Этот результат хорошо согласуется с простой группировкой, проведенной выше.

Приведем запись метода корреляционных плеяд (табл. 50).

4. Синхронизация эволюционных рядов. Построение системы хронологии

Типологический метод в археологии давно выработал способ установления относительной хронологии путем визуального построения параллельных эволюционных рядов и синхронизации его звеньев. Синхронизация эта может быть уточнена и проверена с помощью показателей связи, вычисленных на основании подсчета взаимовстречаемости видов вещей в погребальных комплексах.

Эволюционные ряды вещей, построенные на основании изменения одного или нескольких признаков, образуют цепочку, вытянутую во времени. Несколько параллельных рядов сопоставляются для исследования взаимовстречаемости их звеньев. Если начальные звенья одного ряда связаны с начальными звеньями другого,

срединные — со срединными, конечные — с конечными, то это будет мощным подтверждением правильности построения эволюционных рядов в целом и правильности их взаимной ориентировки.

Пример 37.

Из могильников древней мордвы Присурья и Примокшанья были извлечены вещи разных категорий и видов. Требуется установить группы синхронных видов этих вещей.

Для выявления связи между видами вещей по их совместной встречаемости в одних погребениях был применен коэффициент сопряженности (см. гл. III, § 4) и было выделено три уровня связей:

$0{,}1 < Q \leqslant 0{,}2$ — слабые связи;

$0{,}2 < Q \leqslant 0{,}3$ — сильные связи;

$0{,}3 < Q \leqslant 1$ — очень сильные связи.

Кроме того, принято было, что

$-0{,}1 < Q \leqslant 0{,}1$ — связь отсутствует;

$-1 < Q \leqslant -0{,}1$ — связь отрицательная.

Были визуально построены эволюционные ряды височных подвесок и блях. Пять звеньев последнего ряда оказались связанными сильными связями по взаимовстречаемости, и граф этих связей составил вытянутую цепочку. Сюльгамы и гривны не дали цепочки эволюционных рядов, а разбились по сильным связям между ними на несколько последовательно сменяющих друг друга групп.

Матрица данных о взаимостречаемости и силе связи каждой пары разновидностей была преобразована в граф (рис. 23).

Взаимосвстречаемость звеньев эволюционных рядов подтвердила правильность их построения и разбила на уровне сильных и очень сильных связей все виды этих вещей на две группы, которым был придан хронологический смысл: могильники армиевского типа и могильники селиксенского типа. Далее полученный граф был упорядочен на основании следующих принципов-правил.

1) Все виды одной категории, образующие связь в виде цепочек, располагаются вертикально в порядке их эволюции или хронологической последовательности.

2) Связи между разновидностями одной категории, принимающие форму графа, который приближается к полному (т. е. такому графу, где вершины соединены

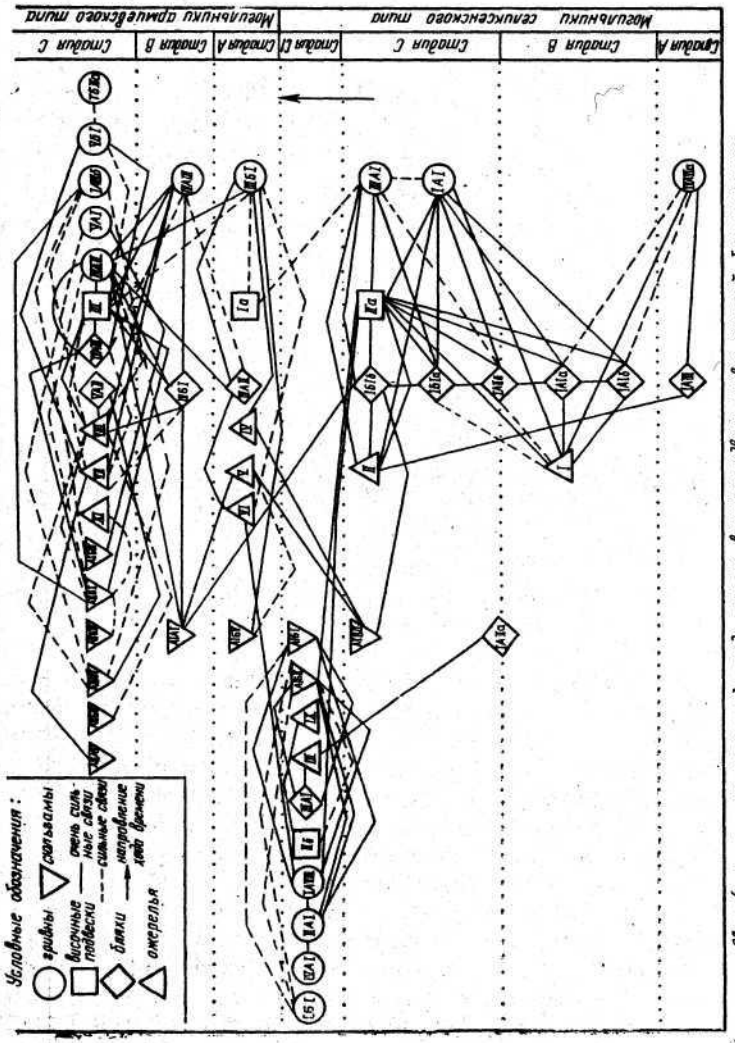


Рис. 23. Граф сильных связей между видами вещей из могильников селикенского и армиевского типа (по В. И. Вихляеву)

максимально возможным числом линий), означают примерно синхронную группу вещей. Располагаются эти разновидности по возможности горизонтально.

3) Связи между долгобытующими видами и другими видами изображаются пучком наклонных линий, выходящих из той вершины графа, которая обозначает долгобытующий тип. Более сильные связи должны иметь меньший наклон по отношению к горизонтальной линии.

За основу был положен наиболее четкий эволюционный ряд блях. Все разновидности других категорий располагались по связям со звеньями этого ряда. Чем сильнее была связь, тем более близкой к горизонтали линией она обозначалась.

Руководствуясь этими правилами, оказалось возможным построить десять горизонтальных линий разновидностей, которые были затем объединены в 7 периодов. Одна из линий, не имеющая значительных связей с другими и отличающаяся своей обособленностью, была выделена в локальный вариант. Каждый хронологический период по аналогиям и по находкам, встречаемости в его погребениях других, хорошо датирующих вещей, может быть абсолютно датирован [2].

Пример 38.

Вещи из Безводнинского могильника (Горьковская область) были расклассифицированы, каждая категория на несколько видов. Исследовалась попарная взаимовстречаемость вещей в погребениях. Как в примере 37, здесь также требовалось выявить группы синхронных вещей. Применялся способ, изложенный выше: теоретическая вероятность встречаемости вещей вместе в одной могиле подсчитывалась как произведение их частостей и сравнивалась с доверительным интервалом, определенным исходя из действительной частоты случаев совместных находок данной пары видов (см. гл. III,§ 4). В результате были выявлены сильные связи. Они были перенесены на графы. Получилось 4 группы связанных видов вещей.

В I группу вошли сюльгамы вида ЖIА1, IА3, головные жгуты вида I, пряжки видов IА1, IБ1, IБ2, IБ3, нагрудные бляхи вида А1 и В4, браслеты вида IVБ1, железные наконечники копий видов I и II, удила вида I.

---

[2] См.: *Вихляев В. И.* Древняя мордва Посурья и Примокшанья. Саранск, 1977.

II группа составилась из сюльгам видов IA1, IA2, IA4, IБ3, IB3, головных жгутов вида II, крестовидных и двухпластинчатых фибул, пряжек вида IA2, IБ4, поясных накладок вида I, II, IV, нагрудных блях вида A2, Б4, браслетов вида IVБ2.

III группа состояла из сюльгам группы II и видов CIA1, IB1, пряжек видов IA3, IIIБ2, IIIБ3, IIIБ4, IIIБ5, поясных накладок видов III, V, VI, VII, VIII, IX, X, браслетов видов IA2, IIБI, наконечников копий вида I, наконечников дротиков вида I и II.

IV группа включила сюльгамы групп вида III и IV, головные жгуты вида III, пряжки вида IA4, IIБ3, IIIБ6, IVA1, поясные накладки видов XIII—XVII, железные наконечники стрел вида II и III, бляхи с крылатой иглой, наконечники копий видов IV и VI, удила вида IV, стремена.

Из этого перечня были исключены виды, которые оказались связанными слабыми связями с вещами из другой группы: из I группы были исключены сюльгамы вида IA3 и нагрудные бляхи вида Б4, которые имели слабые связи с вещами II группы, а также удила вида I и наконечники копий вида I, имеющие слабые связи с вещами всех остальных групп; из II группы исключены были сюльгамы видов IA1 и IБ3, нагрудные бляхи и пряжки вида IБ4, которые имели слабые связи с вещами I и III групп; из III группы исключены были наконечники дротиков вида I, имеющие слабые связи с вещами IV группы, и наконечники копий вида I, имеющие слабые связи с вещами I и II групп.

Исключенные виды рассматривались как долгобытующие виды предметов. В итоге были получены компактные группы видов, которые рассматривались как синхронные. Для выяснения последовательности расположения этих групп были использованы указанные выше долгобытующие разновидности, как бы «объединяющие» близкие по времени группы. I и II группы оказались, таким образом, «объединенными» сюльгамами вида IA3 и нагрудными бляхами типа Б4; II и III группы объединялись удилами вида III; III и IV группы объединялись кельтами группы II, наконечниками дротиков типа I, браслетами вида IVA1. Есть видны, которые слабыми связями связывали три также, видимо, соседние по времени группы: сюльгамы видов IA1, IБ1, IБ3, пряжки видов IБ4, наконечники копий вида I объединяют I, II и III группы. Нет вещей, которые бы объединяли I и IV

группы, не будучи связанными с II и III группами (рис. 24).

Все это позволяет считать, что группы должны быть расположены в такой последовательности: I—II—III—IV или наоборот. Направление этого хронологического ряда подтверждается эволюционными рядами некоторых вещей. Так, хорошо известно, что развитие сюльгам шло от отдела I к отделу II и отделу III, т. е. от коротких

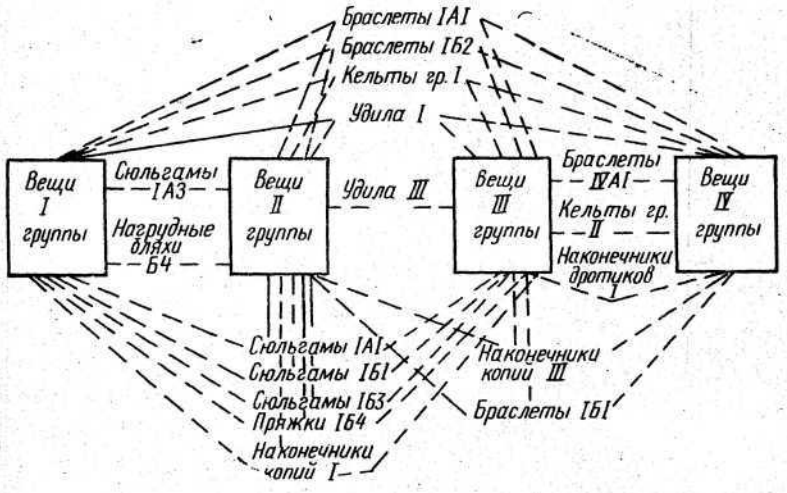


Рис. 24. Граф сильных связей между видами вещей из Безводнинского могильника (по Ю. А. Краснову)

усов к более длинным. Следовательно, II группа с сюльгамами отдела I раньше III группы с сюльгамами отдела II, а та в свою очередь раньше IV группы с сюльгамами отдела III. То же направление хронологического ряда было подтверждено эволюционными рядами кельтов и стратиграфией могильника. Так, были получены 4 группы видов с узкими периодами бытования, ставшими средством датировки погребений, которые оказалось возможным разбить на 4 стадии[3].

Затем все погребения могильника, содержащие датированные виды предметов, располагаются по столбцам, а датирующие виды предметов — по строкам. Получается таблица распределения погребений по стадиям.

[3] См.: *Краснов Ю. А.* Указ. соч.

В месте пересечения соответствующего столбца и строки точкой обозначается наличие в данном погребении данного датирующего вида. Так как датирующие предметы и датированные ими виды погребения расположены в хронологической последовательности по стадиям, то получается несколько блоков, в которых клетки заполнены с большой плотностью. Эти блоки располагаются по линии, соединяющей левый верхний и правый нижний углы таблицы. Вне этих блоков точки оказываются редкими. Некоторым разновидностям, имевшим длительный период бытования, соответствуют точки в нескольких блоках. Такие разновидности связывают соседние по времени блоки.

5. Выделение локальных групп памятников или культур

Пример 39.

Металлический инвентарь ряда культур бронзового века был описан признаками, каждый из которых представлял морфологический (по форме) вид металлического предмета, а значения, которые признак принимал в каждой культуре, — были частости, с которыми вид в ней встречался.

Для каждой пары культур подсчитан был коэффициент сходства $S(ф)$ (см. гл. IV, § 3), где признаком был морфологический вид вещи, а значением его была частота встречаемости этого вида в данной культуре.

Аналогично были подсчитаны коэффициенты сходства $S(м)$ для каждой пары культур, где признаком был металлургический тип вещи, а значениями его были также частости встречаемости этого типа в данной культуре. В результате были получены такие цифры табл. 51, вверху — $S(ф)$, внизу — $S(м)$.

Эта матрица дала основания для построения графа связей металлургических и металлообрабатывающих очагов Волго — Урала по степени морфологической и химико-металлургической близости.

Коэффициенты сходства по химико-металлургической близости дали возможность выделить две группы культур: северную (абашевская, турбино, баланбашская) и южную (сейма, срубная, приказанская, поздняковская, андроновская) и т. п. (рис. 25).

Затем коэффициенты сходства $S(ф)$ и $S(м)$ были рассмотрены как два признака для каждой культуры.

Принадлежностью к одной культуре их значения соединены попарно. Был получен на корреляционном поле корреляционный эллипс (эллипс разброса), показывающий, что связь между $S(ф)$ и $S(м)$ близка к линейной

Таблица 51

Культуры	1	2	3	4	5	6	7	8
1. Срубная	x							
2. Андроновская	0,53 0,84	x						
3. Приказанская	0,26 0,87	0,26 0,56	x					
4. Поздняковская	0,36 0,81	0,54 0,80	0,42 0,71	x				
5. Сейма	0,35 0,70	0,18 0,33	0,17 0,91	0,23 0,49	x			
6. Турбино	0,24 0,19	0,13 0,12	0,12 0,29	0,19 0,00	0,44 0,40	x		
7. Баланбашская	0,28 0,27	0,24 0,16	0,16 0,38	0,07 0,00	0,14 0,37	0,35 0,86	x	
8. Абашевская	0,09 0,36	0,19 0,20	0,09 0,41	0,08 0,00	0,02 0,38	0,14 0,65	0,48 0,83	x

(рис. 26). Подсчитан коэффициент корреляции между $S(ф)$ и $S(м)$. Он оказался равен 0,62.

Из этого был сделан вывод о том, что металлический импорт, связанный с металлургической характеристикой вещей, определял в значительной степени типологию инвентаря, но все же значительный разброс значений (сравнительно низкий коэффициент корреляции) говорит о том, что взаимосвязь между этими явлениями была сложной и определялась многими причинами.

Были подсчитаны «частные» коэффициенты корреляции между $S(ф)$ и $S(м)$ для каждой культуры. Между $S(ф)$ и $S(м)$ получены следующие коэффициенты:

Срубная	—0,60	Сейма	—0,14
Андроновская	—0,92	Турбино	—0,39
Приказанская	—0,51	Баланбашская	—0,81
Поздняковская	—0,91	Абашевская	—0,65

Только для андроновской, поздняковской и баланбашской культур коэффициент корреляции оказался вы-

соким, говорящим о тесной зависимости $S(ф)$ и $S(м)$, в остальных — низким [4].

Выделение ареалов в приведенном примере было произведено только по сходству, учитывавшему металлургический тип изделий. В археологии возникает необходимость выделения культурных ареалов по многим показателям.

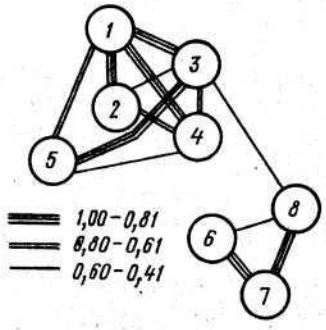


Рис. 25. Граф сходства между культурами по химико-технологическим признакам (по Е. Н. Черныху)

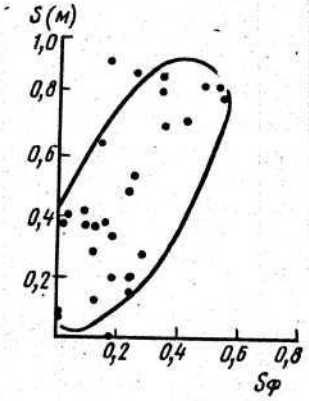


Рис. 26. Корреляция между S (ф) и S (м) для культур бронзового века (по Е. Н. Черныху)

Сходство между памятниками по распределению одной группы вещей может быть большим, а по другой группе вещей — меньшим. Так, сходство (коэффициент S) позднепалеолитических памятников одного культурно-хозяйственного типа в Сибири по кремневому инвентарю колебалось в пределах 0,59—0,74, по костяным орудиям — 0,37—0,7, по украшениям и предметам искусства — 0,20—0,50. Можно получить усредненный показатель сходства между памятниками по нескольким проявлениям материальной культуры и на основании этого показателя произвести группировку памятников. Эта группировка будет означать выделение культурно-исторических общностей, археологических культур и их локальных вариантов. Д. Кларк считал, что сходство

[4] См.: *Черных Е. Н.* Связи типологических и химико-металлургических признаков // СКМА. М., 1970.

памятников внутри локального варианта должно быть 1,00—0,65, внутри археологической культуры 0,65—0,30, внутри общности 0,30—0,05 [5].

6. Исследование структуры множества признаков

Исследовать структуру множества признаков, которыми описывается какая-то серия предметов или объектов, — значит установить связи между этими признаками и сгруппировать на основании этих связей наиболее сильно связанные признаки. При этом открывается возможность расположить признаки по степени их «важности», «значимости», и т. п.

Для исследования структуры в множестве качественных признаков установление связи между признаками осуществляется разными путями. Может быть применен информационный коэффициент связи ($R_{инф}$, см. гл. III, § 6) — так называемый коэффициент нормированной взаимной информативности. Может быть применен и другой показатель, например коэффициент сопряженности Чупрова или Крамера ($R_{Ч}$ или $R_{Кр}$, см. гл. III, § 5). Подсчитав попарно какой-либо показатель связи и составив матрицу его значений, можем построить граф, в котором вершинами будут признаки, а ребрами — связи между ними, показанные условно, в зависимости от силы этой связи.

Пример 40.

Бусы Северного Кавказа VI—VII и VIII—IX вв. были исследованы по 8 признакам: I — количество отверстий, II — материал, III — техника, IV — форма, V — пропорции, VI — цвет, VII — прозрачность, VIII — наибольший размер. Все признаки, кроме V и VIII, — качественные. V и VIII были разбиты на равные интервалы и также рассматривались как качественные признаки. Требовалось выявить структуру связей между этими признаками.

Была подсчитана нормированная взаимная информативность и получена для бус Северного Кавказа VI—VII вв. матрица значений этого показателя для каждой пары признаков (табл. 52).

Аналогично была подсчитана нормированная взаимная информативность признаков для каждой пары и для

[5] См.: *Clark D. L.* Analytical Archacology. L., 1968. P. 398.

Таблица 52

	I	II	III	IV	V	VI	VII	VIII	Σ
I	1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,00
II	0,0	1	0,51	0,42	0,31	0,49	0,10	0,10	2,93
III	0,0	0,51	1	0,55	0,32	0,55	0,20	0,19	3,32
IV	0,0	0,42	0,55	1	0,68	0,53	0,11	0,10	3,39
V	0,0	0,31	0,32	0,68	1	0,30	0,20	0,10	2,91
VI	0,0	0,49	0,55	0,53	0,30	1	0,20	0,10	3,17
VII	0,0	0,10	0,20	0,11	0,20	0,20	1	0,10	1,91
VIII	0,0	0,10	0,19	0,10	0,10	0,10	0,10	1	1,69

бус Северного Кавказа VIII—IX вв. и составлена матрица попарных значений. Далее были построены два графа — для бус Северного Кавказа VI—VII и VIII—IX вв. (рис. 27).

Видно, что признаки II, III, IV, V и VI в первом комплексе и признаки III, IV, V, VI во втором более

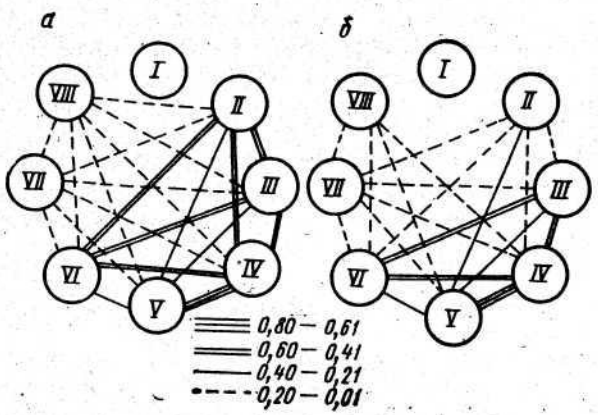


Рис. 27. Граф связей между признаками для бус Северного Кавказа:

а — VI—VII вв.; *б* — VIII—IX вв.

тесно связаны между собой, чем с другими признаками. Особенно тесно связаны признаки V и VI, т. е. форма и пропорции, что и следовало ожидать. Видимо, эти два признака целесообразно объединить и рассматривать как сложный признак, имеющий значениями все сочетания значений IV и V признаков. Остальные признаки

объединять не следует, так как связь между ними или малая, или средняя.

Полученная группировка признаков дает некоторое представление о том, какие признаки более информативны, так как выявляет признаки, более связанные между собой, более зависящие друг от друга. Такими признаками являются в нашем примере III, IV—V, VI. Зная значение одного из них, можно с определенной долей вероятности предполагать то или иное значение других признаков из этой группы. Следовательно, признаки эти информативнее других признаков. Именно они имеют наибольшие суммы коэффициентов по строкам (или столбцам). Они же имеют наибольшие коэффициенты нормированной информативности по отношению ко всем другим признакам. Это позволяет сгруппировать признаки и расположить их в определенном порядке по степени их информативности [6].

Пример 41.

Было взято 7 достаточно многочисленных собраний бус и подсчитаны коэффициенты нормированной информативности для каждого признака: I — количество от-

Таблица 53

Признаки	Кердер (VII—VIII вв.)	Сев. Кавказ (VI—VII вв.)	Сев. Кавказ (VIII—IX вв.)	I Поломский могильник (VIII—X вв.)	Танкеевский могильник (IX—X вв.)	Саркел (XI в.)	Селитренное городище (XIV в.)
I	0,00	0,00	0,00	0,00	0,00	0,00	0,00
II	0,21	0,22	0,13	0,04	0,10	0,10	0,01
III	0,31	0,30	0,30	0,33	0,32	0,45	0,40
IV—V	0,64	0,39	0,56	0,30	0,30	0,34	0,39
VI	0,60	0,60	0,60	0,51	0,55	0,43	0,27
VII	0,07	0,13	0,07	0,05	0,05	0,14	0,01
VIII	0,18	0,15	0,12	0,09	0,09	0,18	0,20

верстий, II — материал, III — техника, IV—V — форма и пропорция, VI — цвет, VII — прозрачность, VIII — размер. Были получены следующие результаты (табл. 53).

Наибольшей информативностью обладают признаки III, IV—V, VI. Признак I не дает вообще никакой информации.

[6] См.: *Федоров-Давыдов Г. А.* Археологическая типология и процесс типообразования //Математические методы в социально-экономических и археологических исследованиях. М., 1981.

Но малоинформативные признаки неодинаковы. Они могут быть разделены на две группы. У одних малая информативность есть следствие того, что подавляющая масса бус относится к одному значению этого признака, и потому значение его никак не определяет другие признаки. Такое явление наблюдается в признаке I (количество отверстий): почти все бусы принадлежат к одному значению этого признака — одно отверстие. Такое же явление наблюдается и для признака II (материал): подавляющая часть бус приходится на один признак — стекло.

Иначе обстоит дело с признаками VII (прозрачность) и VIII (размер). Бусы примерно одинаково, равномерно распределены по значениям этих признаков. Поэтому, если мы знаем, прозрачна бусина или нет, или знаем ее размер, то это почти ничего не дает для нашего знания об остальных признаках этой бусины.

Нужна дополнительная характеристика признака, дающая представление о его неравномерности.

Такой характеристикой может быть коэффициент неравномерности (см. гл. III, § 6).

Для тех же комплексов были подсчитаны для всех признаков коэффициенты неравномерности (табл. 54).

Таблица 54

Признаки	Кердер (VII—VIII вв.)	Сев. Кавказ (VI—VII вв.)	Сев. Кавказ (VIII—IX вв.)	I Поломский могильник (VIII—IX вв.)	Танкеевский могильник (IX—X вв.)	Саркел (XI в.)	Селитренное городище (XIV в.)
I	1,00	1,00	1,00	1,00	1,00	1,00	1,00
II	0,43	0,55	0,70	0,95	0,82	0,90	1,00
III	0,38	0,51	0,43	0,50	0,59	0,47	0,40
IV—V	0,23	0,69	0,66	0,74	0,62	0,21	0,35
VI	0,29	0,41	0,15	0,44	0,50	0,42	0,40
VII	0,27	0,03	0,03	0,85	0,03	0,03	0,90
VIII	0,31	0,27	0,27	0,50	0,29	0,30	0,27

В результате исследования каждого признака на информативность и неравномерность оказалось возможным произвести следующую группировку признаков:

1) с малой информативностью (0,00—0,13) и высоким коэффициентом неравномерности (0,70—1,00). К этой группе относятся признаки I и II;

2) с большой информативностью (0,21—0,60) и средним коэффициентом неравномерности (0,15—0,74). К ним относятся признаки III, IV—V и VI;

3) с малой информативностью (0,00—0,20) и малым коэффициентом неравномерности (0,00—0,31). К ним относятся признаки VII и VIII.

Исключением оказались признак II (материал) в двух первых комплексах (относятся ко второй группе признаков), признак VII в I Поломском и Селитренном комплексах (относится к 1-й группе признаков).

Такое исследование дает возможность построить иерархию признаков.

Первую группу признаков следует отнести к признакам, характеризующим категорию. Они имеют одно преобладающее значение, которое в данном комплексе все время сопутствует функциональному назначению вещи — бусы обычно с одной дыркой и почти всегда стеклянные. В Кердере и Северо-Кавказском комплексе VI—VII вв., где было много каменных бус, признак II (материал) выпал и переместился во вторую группу. В I Поломском могильнике и на Селитренном городище признак категории дополнен признаком VII (прозрачность) — там бусы однодырчатые, по преимуществу стеклянные и непрозрачные.

Признаки второй группы характеризуют тип — это типообразующие признаки, которые определяют наиболее важные изменения и различия объектов одной категории.

Признаки третьей группы характеризуют малосущественные стороны объекта. Это признаки, определяющие варианты [7].

Пример 42.

При исследовании выборки в 365 погребений многоваликовой керамики из Донецкой степи было выделено четыре признака со следующими значениями:

I — наличие сосуда в погребении: 1 — есть сосуд, 2 — нет сосуда;

II — ориентировка: 1 — Восток, 2 — Запад;

III — положение рук: 1 — обе согнуты, кисти у лица, 2 — одна рука вытянута, вторая согнута под тупым или прямым углом;

---

[7] См.: *Федоров-Давыдов Г.-А.* Археологическая типология и процесс типообразования.

IV — могильное сооружение: 1 — сруб, 2 — каменный ящик, 3 — яма, 4 — в насыпи.

Для выявления случаев закономерной встречаемости этих значений признаков И. А. Писларий[8] подсчитывает доли (в %) каждого значения признаков, доли погребений с двумя значениями признаков и, наконец, доли погребений с сочетанием значений трех признаков. Дается также встречаемость значений четвертого признака с группами погребений, определенными по сочетаниям значений трех первых признаков. В итоге исследователь приходит к выводу, что те два сочетания признаков, которые часто встречаются и составляют большую долю среди остальных сочетаний, — типичны, а так как они стратиграфически имеют тенденцию быть одна раньше, а другая — позже, то он относит их к разным культурным эпохам, рассматривая другие редкие сочетания как переходные. Автор правильно заключает, что IV признак мало связан с тремя первыми. Все эти выводы имеют, однако, полуинтуитивный характер. Более предпочтителен в данном случае другой путь: исследование связей каждого значения признака со всеми другими значениями попарно.

Будем рассматривать четыре значения IV признака как четыре признака с альтернативными значениями. Тогда можно применять коэффициент Q.

Составляется матрица попарных коэффициентов сопряженности Q (табл. 55, в правой верхней половине — значение коэффициента Q, в левой нижней — число случаев встречаемости).

Матрица дает следующую картину взаимосвязи между признаками. Признаки I, II, III тесно между собой связаны. Все коэффициенты взаимосвязи между значениями трех признаков, взятых попарно, колеблются в пределах 0,45—0,59 с положительным или отрицательным знаком. Признак IV мало связан с первыми тремя признаками, так как коэффициент всех значений признака IV со всеми значениями признаков I, II и III только в двух случаях достигает 0,16—0,20, а в остальных — не сильно отличается от 0,1. Этого достаточно для того, чтобы судить о структуре признаков.

Таблица 55

		I		II		III		IV			
		1	2	1	2	1	2	1	2	3	4
I	1	×	×	+0,59	−0,59	+0,51	−0,51	−0,07	+0,12	−0,08	+0,07
	2	×	×	−0,59	+0,59	−0,51	+0,51	+0,07	−0,12	+0,08	−0,07
II	1	236	51	×	×	+0,45	−0,45	−0,16	+0,12	−0,10	+0,15
	2	12	66	×	×	−0,45	+0,45	+0,16	−0,12	+0,10	−0,15
III	1	220	47	240	27	×	×	−0,20	+0,10	−0,06	−0,14
	2	28	70	47	51	×	×	+0,20	−0,10	+0,06	+0,14
IV	1	25	18	26	17	21	22	×	×	×	×
	2	16	1	17	0	16	1	×	×	×	×
	3	54	34	63	25	60	28	×	×	×	×
	4	153	64	181	36	170	47	×	×	×	×

[8] См.: *Писларий И. А.* О методе проверки однородности массива археологических памятников // Новые методы археологических исследований. Киев, 1982.

Признаки I, II, III — важные, информативные, составляют ядро типообразующих признаков, признак IV — менее существенный, малоинформативный, т. е. вариантный. Таким образом мы выявили структуру признаков.

Структура признаков показывает отношение человека к данному объекту, что он усматривал главное в этом объекте, что в объекте играло функциональную роль, т. е. определяло непосредственную полезность вещи, что связано с другим аспектом его употребления и что было случайным и неважным в предмете или явлении.

Структура признаков, распределение их на указанные выше группы оказываются устойчивыми для данной категории объектов в определенный длительный исторический период и в большом регионе. В разные времена и на разных территориях бытовали данные объекты с разными значениями признаков, но иерархия признаков остается примерно одинаковой. Это происходит потому, что она отражает отношение людей к этому объекту, которое оказывается длительным и устойчивым, характеризует целые эпохи в истории. Распределение признаков на группы в зависимости от их информативности и неравномерности оказывается как бы внутренней формулой объекта, определяющей структуру его признаков.

7. Факторный анализ

Что связывает признаки, что объединяет их в группы взаимосвязанных, взаимозависимых признаков? Видимо, на них действует какой-то фактор, который и заставляет признаки на объекте принимать не полностью случайные значения, а такие, которые так или иначе зависят друг от друга.

Предположим, мастер избирает определенный размер для какой-то детали изделия (для нас — объекта), а для другой детали он берет не случайный размер, а соответствующий каким-то образом первому, как-то с ним связанный, зависящий от него. Для третьей детали того же изделия он берет размер без всякой зависимости от двух первых. В уме мастера имеется образ готового изделия, его прототип. Этот образ определяет тот или иной набор признаков в предмете. Так, для первых двух взаимосвязанных размеров существует какой-то прототип, сложившийся в голове мастера, который и заставляет

его подбирать размеры так, чтобы они соответствовали друг другу. Эти прототипы и есть *факторы*.

По-видимому, каждое явление, каждый объект может быть описан сравнительно небольшим числом таких характеристик-факторов, скрытых от непосредственного наблюдения. Эти скрытые факторы проявляются в тех признаках, которые доступны наблюдению, измерению, подсчету. Каждый фактор входит в эти признаки по-разному. На одну группу признаков он может влиять сильно, на другую — слабо, так как на нее влияет другой фактор. Поэтому признаки могут быть сгруппированы по факторам.

Фактор проявляется в группе взаимосвязанных признаков, присутствует скрыто в них в качестве так называемой *факторной нагрузки*. Факторная нагрузка показывает степень тесноты связи между фактором и признаком, степень влияния фактора на признак.

Существуют методы подсчета факторных нагрузок и выделения таким образом факторов. Это так называемые методы *факторного анализа*. Один из них — «центроидный» метод определения факторов.

В основе факторного анализа лежит квадратная матрица размером $L \times L$ коэффициентов корреляции между каждой парой признаков, где r_{ij} — коэффициент корреляции между i-м и j-м признаком, а L — количество признаков.

$$
\begin{array}{l}
r_{11}\ r_{12}, \ldots\ r_{1j} \ldots r_{1L}, \\
r_{21}\ r_{22}, \ldots\ r_{2j} \ldots r_{2L}, \\
\cdots\cdots\cdots\cdots \\
r_{i1}\ r_{i2} \ldots\ r_{ij} \ldots r_{iL}, \\
\cdots\cdots\cdots\cdots \\
r_{L1}\, r_{L2} \ldots\ r_{Lj} \ldots r_{LL}.
\end{array}
$$

Все коэффициенты на главной диагонали, т. е. коэффициенты, у которых индексы $i=j$, равны 1. В результате определенной процедуры над этими коэффициентами[9] получают новую матрицу размером $L \times K$, где L — число признаков, K — число факторов (табл. 56).

---

[9] См.: *Окунь Я.* Факторный анализ. М., 1974.

Таблица 56

	1	2	...	j	...	K	h_i^2
1	a_{11}	a_{12}	...	a_{1j}	...	a_{1K}	$\sum_{j=1}^{K} a_{1j}^2$
2	a_{21}	a_{22}	...	a_{2j}	...	a_{2K}	$\sum_{j=1}^{K} a_{2j}^2$
⋮	⋮						
l	a_{l1}	a_{l2}	...	a_{lj}	...	a_{lK}	$\sum_{j=1}^{K} a_{lj}^2$
⋮	⋮						
L	a_{L1}	a_{L2}	...	a_{Lj}	...	a_{LK}	$\sum_{j=1}^{K} a_{Lj}^2$
w_j^2	$\sum_{i=1}^{L} a_{i1}^2$	$\sum_{i=1}^{L} a_{i2}^2$	...	$\sum_{i=1}^{L} a_{ij}^2$	...	$\sum_{i=1}^{L} a_{iK}^2$	$D=\sum_{j=1}^{K} w_j^2$

где a_{ij} — так называемые факторные нагрузки, могут рассматриваться как коэффициент корреляции между i-м признаком и j-м фактором. Величина h_i^2 оценивает «общность признаков», ту степень, с которой все факторы «воздействуют» на данный i-й признак. Если эта величина высока, значит данный признак в большей мере определяется выделенными общими факторами, если низка — значит, в системе выделенных при анализе факторов нет такого фактора, который определял бы данный признак.

Величина w_j^2, так называемый *вклад фактора*, показывает воздействие j-го фактора на все признаки, это как бы весомость, действенность фактора. Она позволяет расположить факторы по степени их важности, значимости в формировании всех признаков объекта.

Величина D — *общий вклад* всех факторов. Она оценивает суммарное воздействие всех общих факторов на все признаки. Чем больше D, тем больше связей имеется между признаками, тем более внутреннее организован материал, подвергнутый факторному анализу, тем больше он представляет собой какую-то внутреннюю структуру.

Несколько признаков с высокими нагрузками какого-либо фактора можно выделить в группу признаков, связанных с этим фактором, другие признаки, имеющие высокие нагрузки другого фактора, — в другую группу и т. д. Дело собственно археологического исследования истолковать эти факторы, определить их природу.

Таким образом, факторный анализ может использоваться для группировки сильнокоррелированных признаков. Тем самым выявляется и структура множества признаков, так как факторы и соответствующие группы признаков можно расположить в каком-то определенном порядке.

Следует иметь в виду, что факторы $z_1, z_2, \ldots, z_k$ и их нагрузки a_{ij} вначале не заданы; мы их ищем по заданным признакам $x_1, x_2, \ldots, x_L$ таким образом, чтобы каждый признак x_i был приближен, но представлен как взвешенная сумма $a_{i1}z_1+a_{i2}z_2+\ldots+a_{ik}z_k (i=1, 2, \ldots, L)$.

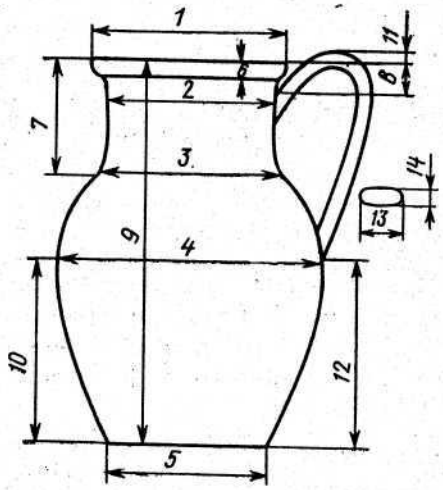


Рис. 28. Количественные (мерные) признаки у кувшинов из нижневолжских городищ XIV в.

Пример 43.

75 сосудов из городов Нижнего Поволжья XIV в. были обмерены по 14 признакам (рис. 28). Среди этих 14 мерных признаков требуется выявить группы тесно связанных признаков.

Подсчитаны коэффициенты корреляций между каждой парой полученных количественных мерных признаков (табл. 57).

Таблица 57

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	0,98	0,85	0,82	0,98	0,27	0,74	0,42	0,73	0,60	0,26	0,76	0,27	0,21
2		1	0,77	0,76	0,94	0,25	0,58	0,28	0,76	0,64	0,13	0,66	0,09	0,14
3			1	0,91	0,82	0,20	0,79	0,43	0,87	0,73	0,17	0,84	0,35	0,29
4				1	0,79	0,19	0,78	0,52	0,87	0,74	0,26	0,90	0,36	0,42
5					1	0,22	0,60	0,34	0,71	0,62	0,23	0,73	0,19	0,21
6						1	0,19	0,23	0,16	0,08	0,04	0,13	0,04	0,16
7							1	0,47	0,89	0,72	0,39	0,84	0,38	0,50
8								1	0,60	0,24	0,71	0,57	0,41	0,48
9									1	0,72	0,41	0,96	0,48	0,52
10										1	0,07	0,66	0,12	0,99
11											1	0,37	0,23	0,49
12												1	0,51	0,49
13													1	0,78
14														1

Эта матрица была подвергнута процедуре центроидного метода факторного анализа с вращением для I фактора. Получены следующие значения факторных нагрузок (табл. 58).

Таблица 58

Признак	Факторы				
	I	II	III	IV	h_i^2
1	0,900	0,196	—0,336	—0,200	1,000
2	0,904	0,045	—0,323	—0,171	0,952
3	0,914	0,217	0,108	—0,032	0,895
4	0,887	0,306	0,095	0,049	0,892
5	0,892	0,139	—0,307	—0,193	0,946
6	0,180	0,173	—0,191	—0,139	0,118
7	0,765	0,414	0,207	0,254	0,864
8	0,277	0,759	—0,212	0,316	0,798
9	0,842	0,458	0,189	0,147	0,976
10	0,793	0,024	0,232	0,147	0,705
11	0,116	0,651	—0,225	0,355	0,614
12	0,824	0,440	0,190	0,122	0,924
13	0,168	0,674	0,356	—0,347	0,729
14	0,146	0,834	0,302	—0,214	0,854
w_j^2	6,819	2,955	0,850	0,644	11,268

Большая часть h^2 выше 0,7. Это показывает, что признаки хорошо описаны выделенными факторами.

Нагрузки I фактора оказались значительно большими у признаков 1—5, 7, 9, 10, 12, чем у других. Очевидно, это тот фактор, который определяет общие пропорции сосуда. Нагрузки II фактора оказались значительными у признаков 8, 11, 13, 14. Этот фактор оказывает воздействие на формирование ручки, на выбор места ее прикрепления к горлу, на ее изгиб. III — фактор сечения ручки (признаки 13, 14). IV — фактор, определяющий места прикрепления ручки к горлу и ее изгиб (признаки 8, 11). III и IV факторы как бы расчленяют смешанный II фактор. Остается 6-й признак. Он не имеет значительной нагрузки ни одного из факторов.

Таким образом, многофакторным анализом признаки разбиваются на три группы сильно коррелированных между собой признаков.

I группа: признаки 1—5, 7, 9, 10, 12 соответствуют I фактору.

II группа: признаки 13, 14 соответствуют II и III факторам.

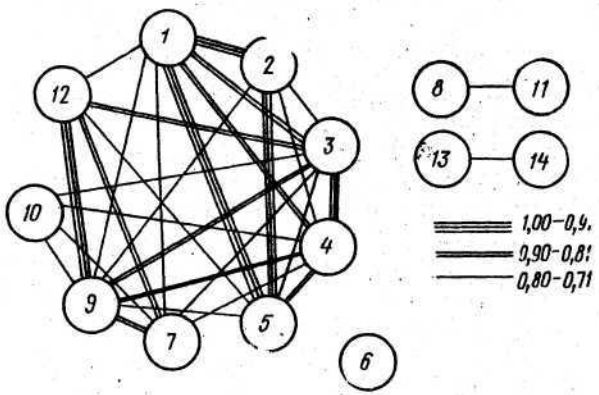


Рис. 29. Граф связей признаков у кувшинов из нижневолжских городищ XIV в.

III группа: признаки 8, 11 соответствуют II и IV факторам.

Остается один признак — 6, не связанный с другими признаками.

Первая группа признаков (1—5, 7, 9, 10, 12) определяет общий облик сосуда, вторая (8, 11) — место при-

крепления ручки к горлу и изгиб ручки, третья (13, 14) — сечение ручки.

Результаты многофакторного анализа по группировке признаков, как правило, вполне соответствуют простой последовательной группировке этих признаков по коэффициентам корреляции. Действительно, если из матрицы коэффициентов корреляции выбрать те пары, у которых значения коэффициента корреляции равны или больше 0,70, и перенести их на граф, то получим те же три группы признаков (рис. 29).

Факторный анализ в описанном виде применим только для количественных признаков, между которыми может быть подсчитан коэффициент корреляции. Однако разработаны приемы факторного анализа и для качественных признаков. Упрощенный факторный анализ качественных признаков дает пример 40, в котором на основании матрицы попарных коэффициентов взаимной информативности выявлялась структура признаков путем объединения наиболее тесно связанных признаков в один.

8. Кластер-анализ

Методы автоматической группировки часто объединяют под общим названием *кластер-анализ*. С помощью кластер-анализа можно провести автоматическое разбиение объектов на классы (кластеры) по какому-либо показателю связи между ними. Возможно применение рассмотренных методов группировки или более сложных, например так называемого агломеративно-иерархического метода для группировки признаков по коэффициенту корреляции между каждой парой признаков или для группировки объектов по коэффициенту сходства, или по расстоянию между ними.

Кластером называют такую группу объектов, между которыми среднее расстояние (связь) превышает среднее расстояние (связь) этих элементов с остальными элементами.

На первом шаге алгоритма агломеративно-иерархического метода кластерного анализа находят минимальное значение расстояния (минимальный коэффициент корреляции, сходства) между объектами и объединяют их в один кластер. Затем определяют расстояние от этого кластера до каждого из всех остальных объектов. Это расстояние вычисляется как среднее расстояние от объединенных в первый кластер объектов до какого-либо из оставшихся необъединенными. Выбирают наименьшее из этих расстояний и присоединяют к первому кластеру тот объект, расстояние от которого до этого кластера оказалось наименьшим. Если имеется расстояние между какими-нибудь двумя необъединенными объектами меньшее, чем расстояние между необъединенными объектами и первым кластером, то образуют новый кластер. Далее подсчитывают новые средние расстояния от полученных кластеров (одного или двух) до оставшихся необъединенными объектов или если получен новый кластер, то подсчитывают и расстояние между ними (расстояние между кластерами получают как среднее расстояние между всеми парами объектов, у которых один объект принадлежит одному кластеру, другой — другому). Из всех полученных расстояний между объектами, между объектами и кластерами или между кластерами выбирают наименьшее и формируют новый кластер, объединяя объекты или присоединяя к ранее полученным кластерам новый объект или, наконец, объединяя два кластера. Так перебирают все объекты. Кластер, полученный на определенном шаге алгоритма, рассматривается как новый объект. На каждом шаге происходит уменьшение числа объектов на 1. Всего шагов в алгоритме $n-1$, если n — число объектов. В результате получают граф в виде «дерева». Его вершины — это объекты, а ветви, объединяющие на разных уровнях эти вершины, — связи между объектами и группами объектов с разными показателями связи.

Обычно заранее задается какое-то число групп, и на этом графе проводят линию так, чтобы она пересекала соответствующее количество ветвей. Все ответвления ниже этой линии не учитываются и рассматриваются как один класс. Ветви выше этой линии показывают разбиение объектов на кластеры.

В археологии число классов не может быть названо заранее, и потому линия проводится интуитивно, т. е. интуитивно устанавливается порог для значений показателя связей. Существуют и более тонкие методы кластер-анализа, позволяющие получать группировки с не заданным заранее количеством классов.

Пример 44.

В примере 43 мы показали, как 14 мерных признаков, которыми описываются 75 кувшинов из городищ Нижнего Поволжья XIV в., группируются в три группы тесно-

коррелированных признаков, соответствующие трем факторам.

Требуется провести группировку сосудов по их сходству по каждой группе признаков. Для этого проведем кластер-анализ 75 кувшинов по I группе из 9 признаков, тесно коррелированных между собой.

Сначала подсчитываются эвклидовы расстояния между каждой парой сосудов. (В примере 32 мы определили эти расстояния между тремя сосудами.) Наименьшее расстояние оказывается между 13-м и 16-м сосудами (0,000). Они объединяются. Это первый шаг алгоритма. Далее, на втором шаге, к этой группе присоединяется 10-й сосуд, среднее расстояние которого от сосудов 13 и 16 оказывается наименьшим (0,042) из всех вычисленных расстояний. На третьем шаге к этой группе присоединяется сосуд 11 (его среднее расстояние от сосудов 13, 16, 10 равно 0,087). На четвертом шаге объединяются в новый кластер сосуды 25 и 75 (расстояние между ними 0, 207). На пятом шаге к группе сосудов 13, 16, 10, 11 присоединяется сосуд 4 (среднее расстояние его от указанных сосудов — 0,231) и т. д.

Таким образом, получаем группировку объектов, имеющую вид дерева. Оно ветвится на разных уровнях схожести объектов и конечными ветвями его являются 75 сосудов, взятые для обработки. Необходимо теперь выбрать такое расстояние, при котором получилась бы удовлетворяющая нас группировка объектов. Если взять очень высокий порог, то можно получить вообще одну группу, если его снизить, то получим две группы, если еще снизить, — то получим три группы и т. д. Очевидно, нужно стремиться к тому, чтобы было не очень много мелких групп.

При пороговом расстоянии 3,741 совокупность сосудов делится на 7 кластеров. Если взять большее расстояние в качестве порогового, то будет получаться слишком мало кластеров, и классификация будет слишком общей, неприемлемой. Полученные 7 кластеров очень неравномерны по составу: 1-й кластер включает 31 сосуд; 2-й — 1 сосуд; 3-й — 1 сосуд; 4-й — 2 сосуда; 5-й — 34 сосуда; 6-й — 2 сосуда и 7-й кластер — 4 сосуда.

Вся совокупность разделилась на две большие группы и на 5 маленьких (1—4 сосуда) группок. Если снизить пороговое расстояние до 2,4, из 1-го кластера выделится небольшая группка в 4 сосуда, а 5-й кластер разобьется на 2 группы в 8 и 26 сосудов. Получится 9

кластеров. Остаток 1-го кластера разбивается на 3 группы только при понижении порогового расстояния до 1,269, но при этом пороговом значении вся совокупность делится на 26 слишком мелких кластеров, что также неприемлемо. Видимо, следует остановиться на пороговом значении расстояния 2,4, разбивающем совокупность на 9 кластеров.

Подсчитаем средние значения каждого признака для каждого кластера в мм (табл. 59).

Таблица 59

Кластер	Признаки									Численность кластера
	1	2	3	4	5	7	9	10	12	
1	48,3	46,2	39,4	86,3	46,9	27,6	98,2	48,8	49,3	27
2	38,2	30,0	51,2	106,2	42,7	45,0	185,0	102,5	110,0	4
3	120,0	80,0	130,0	300,0	110,0	120,0	740,0	450,0	530,0	1
4	120,0	120,0	150,0	240,0	120,0	120,0	510,0	300,0	300,0	1
5	110,0	105,0	120,0	225,0	105,0	95,0	465,0	210,0	245,0	2
6	87,5	83,1	72,1	133,4	86,3	49,3	193,5	99,3	111,2	8
7	66,1	61,5	54,8	103,7	64,4	35,5	137,8	69,6	70,6	26
8	75,0	77,5	110,0	235,0	72,5	75,0	342,5	167,5	190,0	2
9	77,5	70,3	71,2	151,0	73,2	70,0	316,2	201,2	197,2	4

Таблица 60

Кластер	Признаки		Численность кластера
	8	11	
1	0,00	—70,0	1
2	0,6	—4,3	8
3	2,0	3,5	56
4	25,0	14,0	5
5	35,0	35,0	4
6	60,0	60,0	1

Таблица 61

Кластер	Признаки		Численность кластера
	13	14	
1	12,0	10,0	1
2	21,2	6,5	5
3	25,0	10,2	55
4	30,1	13,6	13
5	50,0	20,0	1

Средние значения определяют как бы «центр тяжести» кластера в соответствующем пространстве признаков. Теперь мы имеем один составной признак, объединяющий 9 простых мерных количественных признаков (I группа признаков). Он имеет характер качественного признака с 9 значениями.

Аналогично могут быть определены значения составных признаков для других групп теснокоррелированных признаков.

Средние значения признаков II группы — 8 и 11 приведены в табл. 60 в мм.

Средние значения признаков III группы — 13 и 14 приведены в табл. 61 в мм.

Одиночный, нескоррелированный с другими, 6-й признак легко делим по полигону распределения его значений на три интервала (мм) (табл. 62).

Таблица 62

Интервал	Среднее значение	Численность интервала
1	0,1	47
2	6,8	26
3	14,0	2

После того как произведена группировка объектов, возможно получение, поступление новых объектов. Нужно определить, к какой группе относится этот новый объект. Можно сравнить значения признаков нового объекта со средними значениями для каждой группы полученных в результате группировки и определить, к средним какого класса ближе всего новый объект.

Пример 45.

После того как был произведен кластер-анализ 75 золотоордынских кувшинов (см. пример 44), при дальнейших раскопках были найдены еще 8 археологически целых кувшинов. Размеры этих новых сосудов (мм) и отнесение их к кластерам приведены в табл. 63.

Сосуды 77, 78, 83 весьма близки по значениям признаков I группы к средним значениям этих признаков 1-го кластера. Несколько дальше отстоит по значениям этих признаков от средних 1-го кластера сосуд 76, а сосуд 81 — от средних 8-го кластера. Но все же мы можем пренебречь этими отличиями.

Сосуды 79 и 80 по I группе признаков стоят между средними значениями 6-го и 7-го кластеров. Чтобы решить, к какому кластеру все же их отнести, определим эвклидовы расстояния их от «центров» этих классов: $g_н$ (79,6 кл.) = 1,2; $g_н$ (79,7 кл.) = 2,2; $g_н$ (80,6 кл.) = 1,3; $g_н$ (80,7 кл.) = 2,6.

Таким образом, сосуды 79 и 80 по признакам I группы «ближе» к «центру» 6-го кластера. Остается 82-й сосуд. Он не подходит ни к одному из кластеров. Действительно, расстояние его до ближайшего «центра» кластера равно 4,42, что значительно превышает порог близости, принятый нами при кластерном анализе (2,4). Сосуд 82, таким образом, составит новый кластер.

Аналогично определяется принадлежность к выделенным уже кластерам по II и III группам признаков.

Таблица 63

Номера сосудов	Признаки: I группа									II группа		III группа			Кластеры по группам признаков			
	1	2	3	4	5	7	9	10	12	8	11	13	14	6	I	II	III	6
76	60	58	55	90	45	30	100	50	50	5	0	25	10	5	1	3	3	2
77	50	50	50	80	45	35	100	45	45	5	0	25	10	5	1	3	3	2
78	50	50	45	80	40	32	100	53	53	4	1	15	7	4	1	3	1	2
79	75	80	70	115	75	45	160	85	85	5	—1	25	10	5	6	3	3	2
80	79	80	75	120	70	45	163	80	90	5	0	22	7	5	6	3	3	2
81	100	80	80	200	80	100	340	180	180	30	40	35	15	11	8	5	4	3
82	140	140	140	170	110	40	210	85	85	0	—2	32	7	0	10	2	3	1
83	50	45	45	80	40	25	100	50	50	7	0	23	7	5	1	3	3	2

В некоторых случаях просто так, «на глазок», это сделать трудно, так как трудно оценить по каждому признаку отклонения от этих групповых средних. Например, новый кувшин будет по одному признаку ближе к одной группе, а по другому будет приближаться ко второй группе, а по нескольким признакам — к третьей группе. Для сведения к минимуму ошибки при определении, к какой группе принадлежит новый объект, существуют модели теории распознавания образов, где каждая группа (класс) рассматривается как образ и решается задача, как с наименьшей ошибкой отнести новый объект к тому или иному образу [10].

Группировкой, в частности кластер-анализом, можно, как мы видели на примере 44, получить из группы количественных мерных признаков один составной качественных признак. Выше мы рассматривали способы превращения количественных признаков в качественные и методы разбиения отдельного количественного признака на такие интервалы, которые могли бы рассматриваться

[10] См.: *Загоруйко Н. Г.* Методы распознавания и их применение. М., 1972; *Ту Дж., Гонсалес Р.* Принципы распознавания образов. М., 1978.

как значения качественного признака. Но если объект описывается несколькими количественными признаками, то трансформация количественных признаков в качественные по отдельности несет значительную потерю информации и существенное искажение действительности.

Если качественные признаки имеют четкие границы между значениями, то разбитые на интервалы количественные признаки таких четких и как бы абсолютно отдельных одно от другого значений не получают. Полигон распределения количественного признака есть проекция на одну из осей координат действительной группировки объектов в данном пространстве признаков. Возможна такая группировка объектов, при которой проекция их на одну из осей не дает четких «вершин» или вообще не

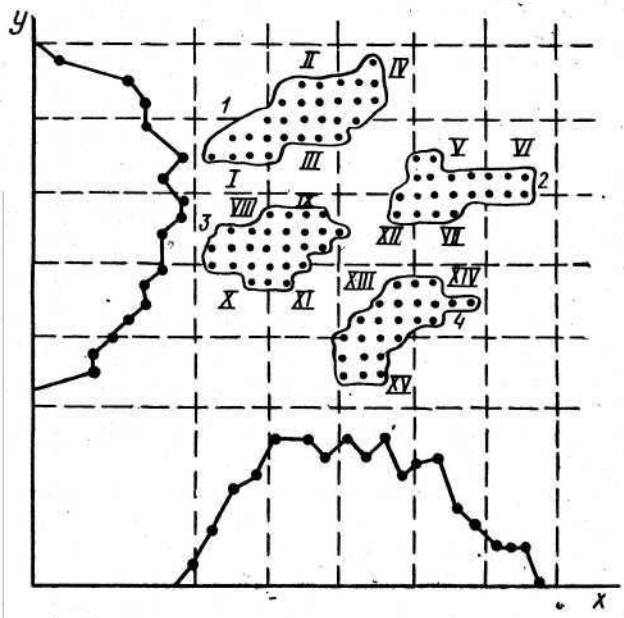


Рис. 30. Соотношение групп, определенных по интервалам признаков и действительной группировкой объектов

дает «вершин». В некоторых случаях даже четкие «вершины» и «понижения» в полигоне признака не соответствуют реальным группам объектов. Проиллюстрируем это на примерах, в которых для наглядности взят случай с двумя количественными признаками.

На рис. 30 показано, как группируются объекты в четыре класса (обозначены арабскими цифрами). Полигоны, построенные на осях x и y, не дают возможности определить какие-либо интервалы по «понижениям». Приходится интервалы брать равными. Эти интервалы показаны пунктирами. Выделим группы объектов по по-

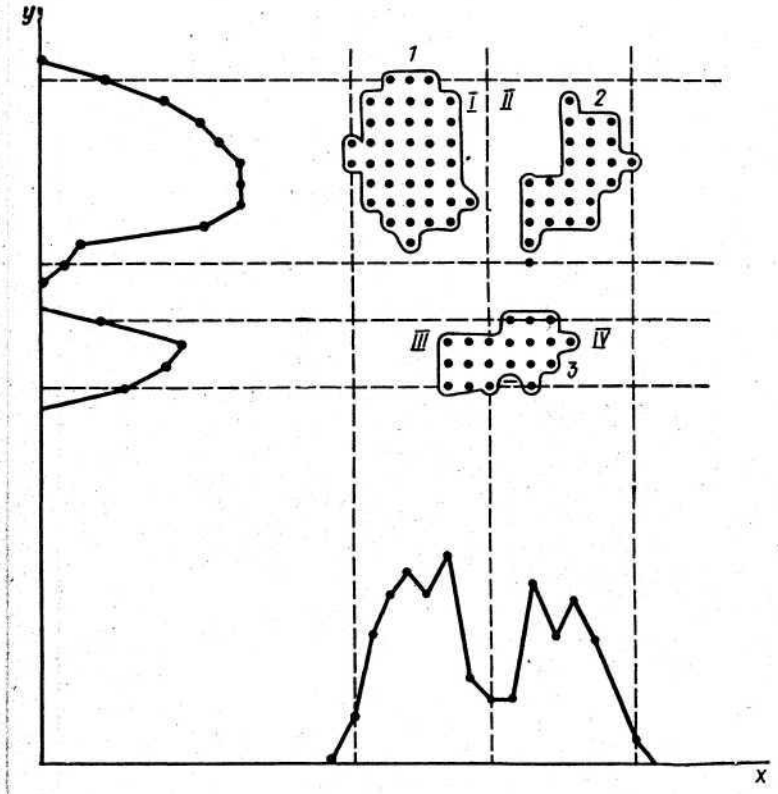


Рис. 31. Соотношение групп, определенных по интервалам признаков и действительной группировкой объектов

паданию объектов в пересечения этих интервалов. Получим 15 групп, обозначенных римскими цифрами. Такая группировка по интервалам может дать сильное искажение действительной группировки объектов в пространстве признаков.

На рис. 31 показано, как группируются объекты в 3 класса (обозначены арабскими цифрами). Полигоны,

построенные на осях x и y, позволяют определить по два интервала для каждого признака. Эти интервалы показаны пунктирами. Будем выделять группы объектов по попаданию объектов в пересечение этих интервалов. Получим 4 группы, обозначенные римскими цифрами. Подобная группировка по интервалам также может привести к сильному искажению действительной группировки объектов в признаковом пространстве.

Однако в археологической практике мы встречаемся именно с такой группировкой. Разбивают количественные признаки на интервалы (по полигону или на равные интервалы) и производят группировку, выделяя группу объектов, у которых первый признак имеет значение, попадающее в определенный интервал, второй признак имеет значение, попадающее в другой определенный интервал, и т. д. И хотя с оговорками, такая практика может быть принята. Все же предпочтительнее группировка, исходящая из близости самих объектов друг к другу, т. е. исходящая из действительного расположения объектов в многомерном пространстве, а не из проекций на оси пространства.

Таким образом, группировка — это средство образования значений составного признака. В этой связи следует заметить, что одной из задач группировок является выявление скрытой (латентной) переменной, которая не может быть выявлена непосредственным наблюдением или измерением. Например, если в одном могильнике удалось сгруппировать погребения так, что обеспечивается достаточная однородность их, то мы можем считать, что каждая группа отвечает значению какой-то скрытой переменной. Ее только нужно правильно интерпретировать. Допустим, при этом нам как-то удалось элиминировать, устранить влияние этноса: предварительно все погребения разбиты на группы по этническим признакам. Тогда если мы по другим признакам сгруппируем погребения внутри этих этнических групп, то полученные вторичные группы можно рассматривать как значения латентной переменной — времени и придавать им хронологический смысл. Наоборот, если предварительно сформировать группы погребений по хронологическому признаку, а затем внутри каждой временной группы сгруппировать погребения по другим признакам, то можно считать, что полученные вторичные группы отражают значения другой латентной переменной — этноса.

Группировки могут быть и многостепенными, например по хронологическим, социальным, этническим признакам и т. п. В таких случаях группировка превращается в иерархическую классификацию. Подобные группировки позволяют правильно интерпретировать выявляемые связи между объектами. Например, предварительная группировка погребений по признаку «есть сбруя» или «нет сбруи коня» (см. пример 25 в гл. III), связанному с социальной принадлежностью погребенного и ритуалом, позволила устранить функциональный фактор и выявить наличие или отсутствие связи между отдельными типами стремян и удил.

Группировки, таким образом, являются важнейшим инструментом исследования в археологии, как и во всех тех науках, где нет строгих функциональных зависимостей, количественно выраженных закономерностей (в математике, физике, химии).

ГЛАВА VI

АРХЕОЛОГИЧЕСКАЯ КЛАССИФИКАЦИЯ И ТИПОЛОГИЯ

Мы не предполагаем рассмотреть здесь проблемы археологической *классификации* во всей их сложности[1]. Ограничимся лишь применением математико-статистических методов в некоторых приемах классификации и в определении и выявлении типов.

В основе классификации лежит группировка. Однако группировкой объектов классификация не исчерпывается. Иногда классификация ограничена какой-либо частной задачей — например, создание классификации объектов по одному или небольшому количеству признаков — так называемая частная классификация. Например, группировка металлических изделий по составу металла, классификация керамики по составу теста и т. п. Эти частные классификации предназначены для какого-то специального изучения объектов. В этих случаях классификация совпадает с группировкой.

[1] О классификации, основанной на математических методах, подробнее см., например: *Айвазян С. А., Бежаева З. И., Староверов О. В.* Классификация многомерных наблюдений. М., 1974; *Миркин Б. Г.* Группировки в социально-экономических исследованиях. М., 1985.

Но в археологии часто ставится задача создания общей классификации археологических объектов, в которой бы использовались все или почти все интуитивно выявляемые признаки. Например, классификация керамики и по составу теста, и по технике формовки, и по профилю, и по орнаменту и т. п. или классификация погребений и по полу, и по возрасту, и по размеру ямы и по устройству ее, и по надмогильным конструкциям, и по инвентарю, и по антропологическому типу погребенного.

В этих случаях классификация не ограничивается одной только группировкой, а использует целый набор приемов группировки. Группировке подвергаются и признаки объектов и сами объекты. Различное использование методов группировки создает разные классификации. Проверкой классификации является то, в какой мере классификация выявляет закономерности, заложенные в исследуемой массе объектов. На этой стадии классификация переходит в *типологию*. Можно сказать, что классификация тогда является типологией, когда выявленные ею группы объектов показывают какие-то закономерности [2].

В реальности существует скрытое от нас действительное объединение объектов в группы, отражающие те закономерности и те процессы, которые в природе этих объектов имеются. Методы группировки лишь приближенно, а часто и весьма приближенно отражают эту реальную группировку. Классификация, используя комплекс методов группировок, вырабатывает такое разбиение на классы, которое наилучшим образом приближается к реальной, действительной группировке объектов. Типология — это проверка соответствия классификации реальности.

1. Формирование пространства признаков

Почему мы не можем при классификации объектов по нескольким признакам ограничиться группировкой, т. е. подсчетом степени близости объектов и разбиением их на группы одним из методов группировок?

Простой подсчет степени близости между объектами по всем выделенным признакам и последующее разбиение по этим показателям сходства может привести к грубым ошибкам: вместо выявления действительно существующих групп объектов упрощенная группировка исказит их. Опасности искажения действительности при простой группировке таятся в следующих двух моментах. Первый: выбранные признаки сами зависят друг от друга, влияют друг на друга. Если два сравниваемых объекта совпадают или близки по нескольким зависимым друг от друга признакам и мы будем все признаки, в том числе и эти, одинаково учитывать при определении степени сходства, то показатель сходства окажется завышенным по сравнению с теми объектами, которые близки между собой по независимым признакам. Действительно, если объекты совпадают или близки по одному какому-либо признаку, то они будут совпадать или будут близкими друг к другу и по другим признакам, сильно связанным с этим первым или его дублирующим. И это обстоятельство сделает высоким уровень показателя сходства, даже если по ряду других признаков объекты совершенно не похожи. Возникает риск отнесения этих объектов к одному классу, хотя по сути они сильно разнятся друг от друга.

Поэтому существует общеметодическое требование: при определении степени сходства, близости между объектами следует отбирать независимые или одинаково зависимые между собой признаки. Можно из сильно связанных, сильно коррелированных признаков отобрать какой-то один и учитывать только его при вычислении показателя сходства. Но при этом возможны существенные пропуски в классификации, особенно если признаков много и связь между ними не очень сильна. Ведь даже сильнокоррелированные, тесно взаимосвязанные признаки связаны между собой все же не функциональной зависимостью (при которой одному значению одного признака соответствует строго определенное значение другого). Одному значению одного признака могут соответствовать несколько значений другого, которые каждый раз реализуются с разной степенью вероятности. Поэтому могут быть упущены какие-то разновидности, если при классификации мы ограничимся лишь одними независимыми признаками, т. е. будем сравнивать объекты между собой лишь по независимым признакам — «представителям» и не принимать во внимание другие признаки, даже с ними сильно коррелированные.

[2] См.: Типология и классификация в социологических исследованиях. М., 1982. Гл. 1.

Другим путем преодоления указанной опасности искажения действительной группировки объектов при простом подсчете сходства по всем признакам является придание каждому признаку определенного весового коэффициента, снижающего его роль в образовании общего показателя сходства, если он сильно коррелирован с каким-либо другим.

Третий путь, который нам кажется более предпочтительным в археологии, заключается в формировании из группы теснокоррелированных признаков одного составного признака. Объекты разбиваются на группы по теснокоррелированным признакам и каждый полученный класс рассматривается как значение этого нового составного признака.

Такая группировка проводится независимо столько раз, сколько выявлено групп сильнокоррелированных признаков и соответственно столько же формируется новых составных признаков.

Вторая опасность искажения действительности при классификации по степени сходства состоит в следующем. Признаки и группы признаков могут быть важными и не важными. Сходство и совпадение по более важным признакам должно быть рассмотрено как более существенное, а сходство по маловажным признакам — как менее существенное или вообще несущественное. Для преодоления этой трудности в определении степени сходства между объектами необходимо построение иерархии признаков, т. е. выявление более существенных, важных и менее существенных, менее важных признаков. Это возможно на так называемом *содержательном* уровне, при котором исследователь, исходя из своего опыта, ценностных ориентаций своей эпохи, интуиции, этнографических параллелей и т. п., неформализованно определяет степень важности признаков. Но при этом возможен субъективизм, перенесение собственных оценок, определяемых тем временем, в котором живет исследователь, на явления древности. Поэтому предпочтителен формализованный путь.

Прибегнем к методам исследования структуры признаков. Ими мы определяли не только группы сильносвязанных признаков, но и иерархию признаков по степени их информативности. Эта иерархия представляется нам как некая «формула» вещи, устойчивая для длительного отрезка времени, но исторически преходящая. Нельзя ее переносить из одной большой эпохи в другую. Одни

признаки для людей определенной эпохи были более существенными, важными, другие — менее существенными, менее важными. Одни определяли функциональное назначение вещи, другие — ее хронологические модификации, соответствие вкусам эпохи, этносу и т. п. Нам представляется, что эта «формула» вещи, заключающаяся в информативной иерархии признаков, является основой для оценки важности признаков.

Представляется, что наиболее важными должны быть так называемые «типообразующие» признаки (пример выявления таких типообразующих признаков см. в гл. V, § 6).

Таким образом, для более правильного отражения действительной группировки объектов следует предварительно сформировать пространство признаков, наиболее адекватно отражающее реальность. На этом этапе должны подвергнуться исследованию связи между самими признаками, должны быть выявлены группы сильносвязанных, сильнокоррелированных признаков и установлена их ирерахия. Если в описании объектов участвуют главным образом качественные признаки, а количественных мало, то они легко переводятся по своим интервалам в качественные.

В тех случаях, когда объекты описываются по преимуществу количественными, но и некоторым числом качественных признаков, возникают существенные трудности. Структура множества признаков может быть изучена, а следовательно, и порядок признаков может быть определен, только если все признаки будут однотипными и показатели связи между ними будут однородными[3]. Один из способов преодоления этой трудности заключается в переводе всех количественных признаков в качественные путем разбиения их на интервалы по «вершинам» и «понижениям» полигона.

Каждый интервал может рассматриваться как значение качественного признака. Мы рассмотрели этот способ в гл. I и III. Однако прямой переход всех количественных признаков в такое же количество качественных ранжированных признаков грозит существенными ошиб-

---

[3] В последнее время разработаны приемы классификации объектов, описанных разнотипными признаками. См., например: *Миркин Б. Г.* Указ. соч.; *Лбов Г. С.* Методы обработки разнотипных экспериментальных данных. Новосибирск, 1981.

ками. Группировка по таким признакам может искажать действительную группировку объектов.

Более предпочтительным (см. гл. V, § 7, 8) является другой способ преобразования количественных признаков в качественные: выявление сильно коррелированных между собой количественных признаков и образование путем группировки составного качественного признака для каждой такой группы.

Полученные признаки, которые могут рассматриваться теперь как однородные качественные, должны быть исследованы для установления иерархии признаков. Заметим, что на этом этапе происходит снижение размерности пространства признаков, так как несколько тесносвязанных признаков заменяются одним.

2. Собственно классификация

После того как выявлены важные «типообразующие» признаки, можно их выделить придавая им повышенный коэффициент, а остальным признакам — пониженный. Если признак важен, коэффициент будет высоким и повысит удельный вес этого признака при образовании общего суммарного показателя сходства между объектами. Если признак несущественный, неважный, коэффициент будет мал и понизит этот вес. Далее вся совокупность разбивается одним из указанных выше способов на классы по матрице показателей сходства, вычисленных с учетом этих коэффициентов.

Если мы существенным, «важным» признакам придадим вес, равный 1, а несущественным, «не важным» признакам — 0, то получим группировку только по существенным признакам. При этом удобно построить дендрограмму — «дерево» классификации, при котором вся совокупность объектов разбивается на группы — классы первой ступени классификации по одному признаку, затем каждый полученный класс разбивается на классы второй ступени классификации, затем на классы третьей ступени классификации и т. д. Какая-то ступень классификации, которая использует последний «важный» признак, дает разбиение на классы, соответствующее классам, полученным группировкой по степени сходства с приданием важным признакам веса 1, а «не важным» веса 0. Это основные классы классификации. Назовем их *видами*. Если продолжить это дерево классификации, то получим классы, соответствующие «не

важным» признакам, — как бы *варианты* основных классов. Этот способ классификации предпочтительнее в археологии, так как создает удобство для описания и кодификации объектов по признакам.

Каждая разновидность может быть записана в виде числового кода. Количество позиций равно количеству окончательно выделенных признаков, числа означают номер значения признака.

На каждой ступени классификации используется один признак или несколько тесно связанных признаков, которые рассматриваются как единый составной признак. На каждом этапе такой классификации разбиение происходит по значениям этого признака, простого или составного. Тем самым на каждой ступени классификации разбиение происходит по признакам, примерно одинаково зависимым друг от друга. Потому указанное выше методическое требование соблюдается.

Почему следует тесносвязанные признаки рассматривать как один сложный, составной признак с значениями, образующимися из всех сочетаний значений составляющих его простых признаков?

Как мы видели, это лучше отражает действительную группировку объектов, чем классификация по интервалам каждого признака. Кроме того, если несколько теснокоррелированных признаков мы будем использовать для нескольких ступеней классификации, то классы не будут делиться на более мелкие при последующих ступенях классификации, а в большинстве случаев повторят предшествующее разбиение. Классификация получится излишне усложненной и удлиненной. На «дереве» классификации образуются при этом много длинных неветвящихся стволов.

Пример 46.

Подвергнем классификации 83 кувшина из нижневолжских золотоордынских городищ. Выше мы видели, что они описываются 14 мерными количественными признаками. Факторный анализ показал, что эти признаки определяются тремя факторами и их можно сгруппировать в три группы сильно коррелированных между собой признаков. При этом остается один необъединяемый признак, слабо коррелированный со всеми остальными. С помощью кластер-анализа каждая группа простых количественных признаков была преобразована в составные качественные (примеры 44, 45).

Таким образом, мы имеем три новых сложных и один простой признаки. Каждый признак имеет несколько значений и может рассматриваться как качественный.

Кроме того, сосуды описываются еще тремя качественными признаками: 15 — наличие или отсутствие орнаментации, 16 — наличие или отсутствие широкой углубленной полосы на середине тулова, 17 — наличие или отсутствие слива. Теперь для классификации мы можем использовать 7 однородных (качественных) признака. Каждый этап классификации будет состоять в разбиении на классы по одному из этих 7 признаков. Остается установить порядок, по которому должны вступать эти признаки в классификацию, т. е. иерархию признаков.

Применим употреблявшийся выше метод подсчета коэффициента нормированной информативности $R_{инф}$ и коэффициента неравномерности $R_н$ для каждого признака. Были получены следующие значения (табл. 64).

Таблица 64

Признак	$R_{инф}$	$R_н$
Признаки I группы (1—5, 7, 9, 10)	0,28	0,29
Признаки II группы (8, 11)	0,22	0,41
Признаки III группы (13, 14)	0,17	0,48
Признак 6	0,13	0,27
» 15	0,11	0,01
» 16	0,09	0,06
» 17	0,06	0,63

Признаки могут быть разделены на три группы:

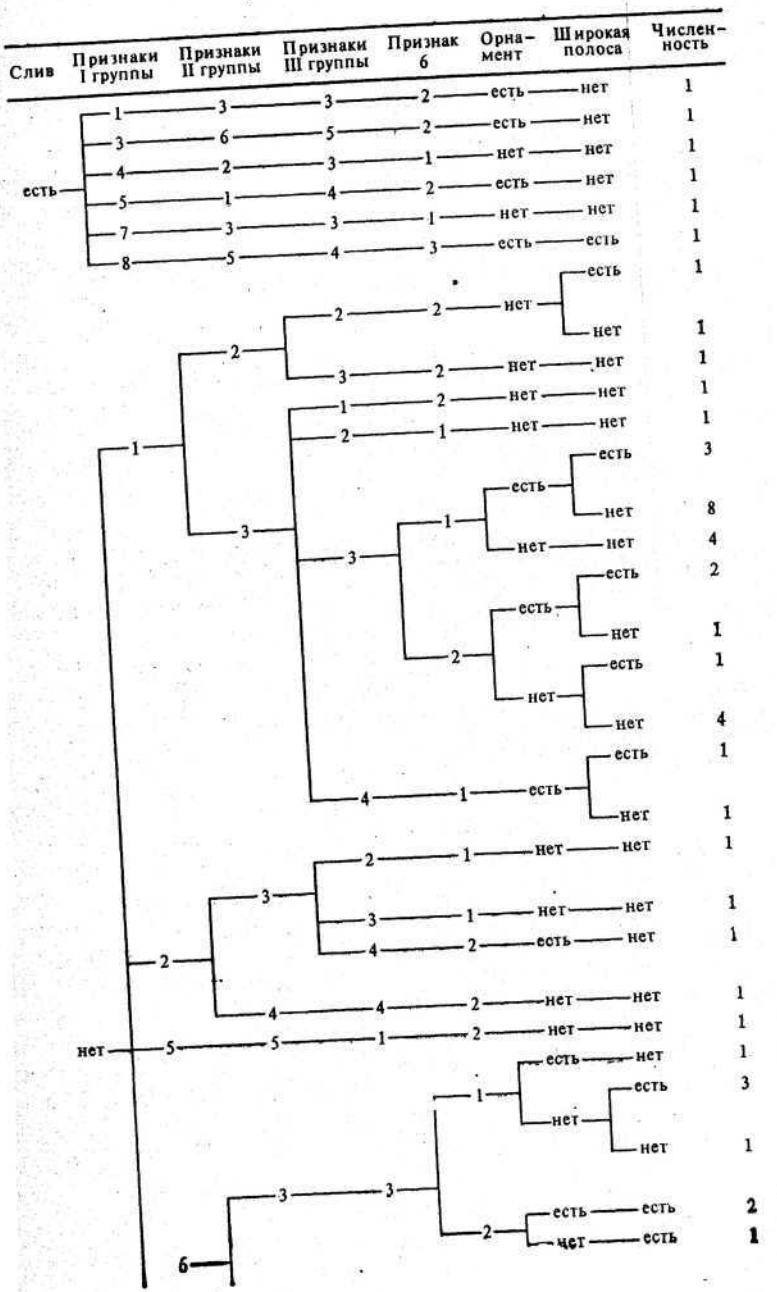
1. С низким $R_{инф}$ и высоким $R_н$ (17-й признак). Это признаки, связанные с категорией предметов.

2. С средним и высоким $R_{инф}$ и средним $R_н$ (признаки I, II группы). Это типообразующие признаки.

3. С низким $R_{инф}$ и средним или низким $R_н$ (признаки III группы, 6, 15, 16). Это вариантные признаки.

К этому следует добавить еще одно указание на иерархию признаков. Из результатов факторного анализа следовало, что признаки I группы более общи, чем признаки II и III групп. Это позволяет поместить признаки I группы перед признаками II и III групп.

Таблица 65



Продолжение табл. 65

Слив	Признаки I группы	Признаки II группы	Признаки III группы	Признак 6	Орнамент	Широкая полоса	Численность
7	4	3	2		есть	нет	1
		4	1		есть	есть	1
	2	3	1		есть	нет	1
			2		есть	есть	1
						нет	1
	2	1			есть	есть	1
	3	3	1		есть	есть	5
						нет	2
					нет	нет	4
			2		есть	есть	2
						нет	1
					нет	есть	1
						нет	1
			3		нет	есть	1
		4	1		нет	нет	1
			3		есть	нет	1
	4	3	2		есть	есть	1
		4	1		есть	есть	1
8	2	3	2		есть	нет	1
	3	4	1		есть	нет	1
9	3	4	1		есть	есть	1
	5	4	1		есть	нет	1
			2		нет	нет	1
			3		нет	нет	1
10	2	3	1		нет	нет	1

В итоге получаем следующую иерархию признаков:

Признаки, связанные с категорией		наличие или отсутствие слива (17)
Признаки	типообразующие	признаки I группы признаки II группы
Признаки	вариантные	признаки III группы признак 6 наличие или отсутствие орнамента (15) наличие или отсутствие полос (16)

Вводя признак в алгоритм классификации именно в этом порядке, строим «дерево» классификации.

Первое ветвление «дерева» будет соответствовать признаку, связанному с категорией (есть или нет слив), второе ветвление соответствует признакам I группы (значения его — это номера кластеров), третье ветвление соответствует признакам II группы. На этом уровне получаются основные классы-виды. Далее идут варианты: четвертое ветвление — признакам III группы, пятое соответствует признаку 6, шестое — наличию или отсутствию орнамента, седьмое — наличию или отсутствию широкой полосы на тулове.

Получаем следующее «дерево» классификации сосудов (табл. 65).

3. Типология

Тип — одно из основных понятий в археологии. В последнее время все более побеждает статистическое понимание археологического типа. Понятие типа связано с взаимозависимостью признаков. Тип есть устойчивое сочетание признаков.

Эта устойчивость понимается, однако, по-разному. Иногда считают, что устойчивость какого-либо сочетания признаков — это ее частная повторяемость. Таким образом, ставится знак равенства между часто встречаемым и типически устойчивым. Типическими разновидностями объявляются те разновидности, которые часто встречаются (так называемая *устойчивая разновидность формы* — УРФ).

Могут ли быть типами редкие разновидности? Мы полагаем, что могут, и предлагаем другое понимание устойчивости сочетания признаков.

Устойчивым сочетанием признаков является такое сочетание, в котором значения признаков связаны закономерно значимыми связями, т. е. сильными связями.

Мы можем определить взаимозависимость значений признаков по их взаимовстречаемости. Если она выше принятого порога, то такая связь может считаться сильной.

Для того чтобы разновидность считалась типической, нужно, чтобы у нее было ядро типообразующих признаков, связанных значимыми связями. Могут быть типы сильные и слабые. Сильные — это те типы, у которых имеются многосторонние связи или полный набор двусторонних связей в ядре, что говорит о высокой степени устойчивости этого сочетания признаков. Слабые — это те типы, у которых в ядре типообразующих признаков отсутствуют сильные многосторонние связи или же нет полного набора парных связей. Если таких связей мало, если они единичны, разновидность вообще не является типической.

При таком определении типа могут быть и редкие типы и часто встречающиеся нетипические разновидности.

Если для отдельных признаков понятие «массовость» и «типичность» совпадают, то для более сложного образования — совокупности признаков (т. е. для разновидности или совокупности разновидностей) этого совпадения нет. В массовых нетипических разновидностях значения признаков соединяются часто и все же случайно, только в силу того, что эти значения признаков вообще частые. Естественно, если признак «*А*» частый и признак «*В*» частый, то и соединяться на объектах эти признаки будут часто. Но частота их соединений не выходит за те пределы, когда мы смогли бы сказать, что эти признаки соединяются чаще, чем если бы это диктовалось только волей случая. Мастер не подбирал специально к признаку «*А*» признак «*В*». Если признаки «*А*» и «*В*» редкие, то и соединяться они должны на редких объектах или вообще маловероятно их соединение даже на одном объекте. А если все же такие объекты имеются, можно говорить, что это не случайное соединение, а закономерное, т. е. мастер специально к признаку «*А*» подобрал именно признак «*В*».

Нетипическая разновидность возникает тогда, когда мастер, изготовлявший вещь (или строящий дом, или устраивающий погребение и т. п.) в значительной мере случайно соединяет признаки на одном изделии. Если он их соединяет не случайно, а сознательно подбирает к одному значению одного признака определенное значение другого и делает это с теми признаками, которые в представлении людей того времени были существенно важными для данного предмета, то возникает типическая разновидность. Специальный подбор одного значения признака к другому означает, что в голове мастера уже сложился стереотип, образец, какой-то устойчивый праобраз изделия, которому он следует.

Определив таким образом тип, нужно в «дереве» классификации провести такую черту, где следует искать типы. Очевидно, она пройдет между типообразующими, информативными, признаками и малоинформативными, вариантными, т. е. между «важными» и «не важными» признаками. До этой черты группировка производится только по важным типообразующим признакам. За ней разбиение на варианты происходит по маловажным признакам.

Эти варианты могут быть у группы, являющейся типом или не являющейся типом. В этом смысле варианты — это «окрестности» типа, случайные отклонения от типа. Их образуют те признаки, которые мастер подбирает почти всегда более или менее случайно, в отличие от типообразующих, которые подбираются для типических разновидностей закономерно. Именно этим объясняется высокая информативность типообразующих признаков и низкая информативность вариантных признаков.

Уровень группировки, который определяет основные классы классификации, выявляет типы и нетипы, — мы назвали выше группировкой по видам. После него будут располагаться варианты. До него группировка происходит на отделы, подотделы, семейства, подсемейства и т. п. Чтобы не ошибиться в том, как разбивается совокупность объектов на виды, т. е. на типы и нетипы — основной уровень классификации, и проделывается сложная работа по исследованию структуры признаков. Именно типические виды — главные показатели всей материальной и духовной культуры общества в той степени, в какой она выражается в остатках материальной культуры. Для характеристики той или иной эпохи археологи должны прежде всего брать типы вещей, жилищ, могильных сооружений, ритуала и т. п.

Одна категория вещей какого-либо хронологически узкого памятника содержит типические и нетипические виды. Соотношение их показывает, какая часть предметов охвачена процессом типообразования, как интенсив-

но этот процесс идет в данном обществе. Это позволяет более точно и полно охарактеризовать это общество.

Каждая разновидность проходила, очевидно, несколько стадий в своем развитии, доходила до определенной степени массовости и типичности или не достигала ни того, ни другого или достигала одного из этих качеств. Достигнув массовости и типичности, разновидность затем выходила из моды, сохраняя или постепенно теряя черты типичности. Производство могло по каким-либо причинам начаться сразу с типических и массовых изделий, т. е. период вхождения в моду и становления типического ядра мог быть коротким. Могли быть завезены большие партии одного вида изделий. Тогда они становились на месте завоза аналогом такого производства с коротким или отсутствующим периодом становления массовости и типичности. Завоз отдельных редких изделий, отличающихся от местных набором необычных черт, аналогичен выпуску на месте редких типичных вещей. Хронологическим срезом этого процесса типообразования и типоразвития является подсчет доли массовых и редких типов и массовых и редких нетипов в каком-либо производстве.

Итак, типология — это проверка классификации, как бы ее последний этап. Если классификация формирует в группы объекты таким образом, что какая-то часть групп является типическими, значит классификация будет отражать определенные закономерности, заложенные в материале. В зависимости от того, как мы сформируем признаковое пространство и какой алгоритм классификации применим, будут получаться разные классификации. Чтобы оценить их, мы должны выделить в каждой классификации типические группы объектов. Классификация, в которой больше групп окажутся типами, будет лучшей. Такая проверка качества классификации требует большой работы; ее выполнение возможно только в далекой перспективе. Мы считаем, что соблюдение указанных выше условий при формировании пространства признаков и рекомендуемые алгоритмы собственно классификации обеспечивают приемлемый уровень классификации, который может служить исходным пунктом дальнейшей работы.

Другая возможная проверка классификации, т. е. выявление закономерностей, это проверка соответствия типов, полученных в результате классификации, каким-то другим признакам, не участвующим в классификации,

например временной или территориальной принадлежности вещей. Можно провести параллельную классификацию материала, скажем, по двум признакам — по времени и месту, или отдельно — по времени или по месту. Получается группировки материала по этим признакам. Наложение классификаций друг на друга выявит степень пригодности классификации, т. е. покажет, насколько выявленные типы соответствуют группировке по времени и месту или отдельно по времени или по месту, покажет, что тип не есть искусственно выявленная группа объектов, а отражает какие-то закономерности их исторического бытования. Но такой путь проверки классификации возможен, если рассматривается материал, расчленимый по времени или из разных мест. Комплекс материала из одного могильника с узкой датой или однослойного памятника может быть проверен только типологией.

Пример 47.

В примере 46 были установлены типообразующие признаки для 83 кувшинов из золотоордынских городов Нижнего Поволжья. Получено было «дерево» классификации, которое определило набор видов кувшинов. Требуется на этом уровне — основном уровне классификации — выявить типические виды.

Для каждой пары значений типообразующих признаков составим 4-польные таблицы взаимовстречаемости и подсчитаем коэффициент сопряженности Q (А — признак I группы, В — признак II группы).

Достаточно сильными связями мы будем считать такие, которые имеют коэффициент больше 0,22 ($\chi^2>4$). В результате выявляются следующие типы, у которых два типообразующих признака связаны достаточно сильными связями. Подсчет коэффициента сопряженности Q значений признаков I и II групп показан в табл. 66.

Некоторые связи, образующие типы ввиду того, что ожидаемые частности в 4-польной таблице взаимовстречаемости слишком малы, следует проверить точным критерием Фишера. Например, проверим связь между значением 3 первой группы признаков и значением 6 второй группы признаков:

	6	$\hat{6}$	
3	1	0	1
$\hat{3}$	0	82	82
	1	82	83

$$p=\frac{1182182111}{831110101821}=0{,}01<0{,}025.$$

Таблица 66

A	B	a	в	c	d	a+в	c+d	a+c	в+d	ad−вc	$[(a+в)\times(c+d)\times(a+c)\times(в+d)]^{\frac{1}{2}}$	Q
1	1	0	31	1	51	31	52	1	82	−31	363	−0,09
	2	3	28	6	46			9	74	−30	1036	−0,03
	3	28	3	34	18			62	21	+402	1449	+0,28
	4	0	31	5	47			5	78	−155	793	−0,19
	5	0	31	5	47			5	78	−155	793	−0,19
	6	0	31	1	51			1	82	−31	363	−0,19
2	1	0	4	1	72	4	79	1	82	−4	161	−0,02
	2	0	4	9	70			9	74	−36	459	−0,08
	3	3	1	59	20			62	21	+1	641	0,00
	4	1	3	4	75			5	78	+63	351	+0,18
	5	0	4	5	74			5	78	−20	351	−0,06
	6	0	4	1	78			1	82	−4	161	−0,02
3	1	0	1	1	81	1	82	1	82	−1	82	−0,01
	2	0	1	9	73			9	74	−9	234	−0,04
	3	0	1	62	20			62	21	−62	327	−0,19
	4	0	1	5	77			5	78	−5	179	−0,03
	5	0	1	5	77			5	78	−5	179	−0,03
	6	1	0	0	82			1	82	+82	82	+1,00
4	1	0	1	1	81	1	82	1	82	−1	82	−0,01
	2	1	0	8	74			9	74	+74	234	+0,32
	3	0	1	62	20			62	21	−62	327	−0,19
	4	0	1	5	77			5	78	−5	179	−0,03
	5	0	1	5	77			5	78	−5	179	−0,03
	6	0	1	1	81			1	82	−1	82	−0,01
5	1	1	2	0	81	2	81	1	82	+81	115	+0,70
	2	0	2	9	72			9	74	−18	328	−0,05
	3	0	2	62	19			62	21	−124	459	−0,27
	4	0	2	5	76			5	78	−10	251	−0,04
	5	1	1	4	77			5	78	73	251	+0,29
	6	0	2	1	80			1	82	−2	115	−0,02

Продолжение табл. 66

A	B	a	в	c	d	a+в	c+d	a+c	в+d	ad−вc	$[(a+в)\times(c+d)\times(a+c)\times(в+d)]^{\frac{1}{2}}$	Q
6	1	0	10	1	72	10	73	1	82	−10	245	−0,04
	2	0	10	9	64			9	74	−90	697	−0,13
	3	8	2	54	19			62	21	+44	975	+0,04
	4	2	8	3	70			5	78	+116	534	+0,22
	5	0	10	5	68			5	78	−50	574	−0,09
	6	0	10	1	72			1	82	−10	245	−0,04
7	1	0	26	1	56	26	57	1	82	−26	349	−0,07
	2	3	23	6	51			9	74	+15	993	+0,02
	3	21	5	41	16			62	21	+131	1389	+0,09
	4	2	24	3	54			5	78	+36	760	+0,05
	5	0	26	5	52			5	78	−130	760	−0,17
	6	0	26	1	56			1	82	−26	349	−0,07
8	1	0	3	1	79	3	80	1	82	−3	140	−0,02
	2	1	2	8	72			9	74	+56	400	+0,14
	3	1	2	61	19			62	21	−103	559	−0,18
	4	0	3	5	75			5	78	−15	306	−0,05
	5	1	2	4	76			5	78	+68	306	+0,22
	6	0	3	1	79			1	82	−3	140	−0,02
9	1	0	4	1	78	4	79	1	82	−4	161	−0,02
	2	0	4	9	70			9	74	−36	459	−0,08
	3	1	3	61	18			62	21	−165	641	−0,26
	4	0	4	5	74			5	78	−20	351	−0,06
	5	3	1	2	77			5	78	229	351	+0,65
	6	0	4	1	78			1	82	−4	161	−0,02
10	1	0	1	1	81	1	82	1	82	−1	82	−0,01
	2	1	0	8	74			9	74	+74	234	+0,32
	3	0	1	62	20			62	21	−62	327	−0,19
	4	0	1	5	77			5	78	−5	179	−0,03
	5	0	1	5	77			5	78	−5	179	−0,03
	6	0	1	1	81			1	82	−1	82	−0,01

Следовательно, вероятности иметь 0 в клетках «*b*» и «*c*» очень малы и гипотеза о независимости значений этих признаков отклоняется в пользу гипотезы о наличии значимой связи между этими признаками.

Таблица 67

Наличие слива	Группы признаков		Численность
	I	II	
Есть	1	3	1
»	3	6	1
Нет	1	3	27
»	5	1	1
»	9	5	3

Аналогично проверяем связи I4—II2, I5—III, I5—II5, I6—II4, I9—II5, I10—II2. Проверку выдерживают только связи I5—III, I9—II5. Таким образом, выявляются следующие типы (табл. 67).

Пример 48.

В примере 42 мы исследовали структуру признаков погребений культуры многоваликовой керамики. Было показано, что признаки I, II и III — типообразующие, а признак IV — вариантный. Требуется выявить типические разновидности погребений.

Подсчет взаимовстречаемости каждой пары признаков был произведен уже в примере 42. Полный набор сильных положительных связей обнаруживают следующие сочетания значений этих трех типообразующих признаков, т. е. выявляется два типа: I1—III—IIII (211 экз.) и I2—II2—III2 (48 экз.).

Остальные виды нетипические.

Соотношение типических и нетипических разновидностей погребений составляет 259 типических и 106 нетипических, т. е. 71% типических.

Если принять относительные датировки по стратиграфическому залеганию погребений, то в первый период типические составляют 41% всех разновидностей погребений, а во второй период типических разновидностей было уже 85%.

Правильность выделения этих двух типических видов подтверждается тем, что в двух хронологических стадиях культуры многоваликовой керамики они распределяются так, что в I стадии преобладает тип I1—III—IIII, а во второй стадии — тип I2—II2—III2. Нетипические виды примерно одинаково часто встречаются в обеих хронологических группах погребений.

Пример 49.

В примере 41 были установлены типообразующие признаки для бус семи археологических комплексов. Затем для каждого комплекса были выявлены типические и нетипические сочетания этих признаков разной массовости. Массовыми считались разновидности, составлявшие не менее 1,5% всего комплекса. Получены следующие данные (табл. 68).

Сравнивая Танкеевский, Саркельский и I Поломский комплексы, мы замечаем, что процесс типообразования в первом из них шел слабее, чем во втором, и еще слабее — в третьем. В Саркеле процесс охватывает большое количество бус, что говорит о силе типообразующего начала, значительно более высокой организованности и упорядоченности признаков. Сильные типы составляют здесь около 38% всех бус, а в Танкеевском из 23%, в I Поломском — 22%.

Таблица 68

Характер разновидности	Кердер (VII—VIII вв.)	Северный Кавказ (VI—VII вв.)	Северный Кавказ (VIII—IX вв.)	I Поломский могильник (VIII—X вв.)	Танкеевский могильник (IX—X вв.)	Саркел (XI в.)	Селитренное городище (XIV в.)
Массовые сильные типы	0,00	0,11	0,30	0,14	0,22	0,36	0,25
Редкие сильные типы	0,00	0,02	0,08	0,08	0,01	0,02	0,06
Массовые слабые типы и нетипы	0,33	0,57	0,29	0,69	0,43	0,29	0,63
Редкие слабые типы и нетипы	0,67	0,30	0,33	0,09	0,34	0,33	0,06

Параллельно увеличению доли сильных типов уменьшается доля массовых слабых типов и нетипов. Характерно, что увеличивается доля редких нетипов и слабых типов. Это как бы «осколки» процесса типообразования, неудавшиеся типы, и чем интенсивнее шел процесс типообразования, тем их должно было быть больше. Таким образом, I Поломский и Саркельский могильники дают нам как бы два состояния процесса типообразования. В первом случае мало массовых сильных типов и редких слабых типов и нетипов, много массовых слабых типов

и нетипов. Во втором случае много массовых сильных типов и редких слабых типов и нетипов, мало массовых слабых типов и нетипов. Танкеевский могильник находится на промежуточном месте между ними. В Саркеле, расположенном в более интенсивно обживаемой зоне, в государстве с развитой торговлей и ремеслом, процесс типообразования шел быстрее. В I Поломском могильнике отражена культура северных лесных племен, только еще преодолевших последние этапы первобытно-общинного строя.

В комплексе бус из Селитренного городища много массовых сильных типов (в этом сходство с Саркелом), но много массовых слабых типов и нетипов (в этом сходство с I Поломским могильником). Можно думать, что здесь процесс типообразования протекал по-другому.

ПРИЛОЖЕНИЕ

Таблица I

Значения функции плотности нормального распределения

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

z	0	1	2	3	4	5	6	7	8	9
0,0	3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3725	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
4,0	0001	0001	0001	0001	0001	0001	0001	0001	0001	0001

Примечание. Все значения умножены на 10 000.

Таблица II

Критические значения для критерия χ^2 при доверительном уровне 0,95 [1]

Число степеней свободы	χ^2	Число степеней свободы	χ^2	Число степеней свободы	χ^2
1	3,8	13	22,4	25	37,6
2	6,0	14	23,7	26	38,9
3	7,8	15	25,0	27	40,1
4	9,5	16	26,3	28	41,3
5	11,1	17	27,6	29	42,6
6	12,6	18	28,9	30	43,8
7	14,1	19	30,1	31	45,0
8	15,5	20	31,4	32	46,2
9	16,9	21	32,7	33	47,4
10	18,3	22	33,9	34	48,6
11	19,7	23	35,2	35	49,8
12	21,0	24	36,4		

[1] Таблица взята из: *Снедекор Д. У.* Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. М., 1961. С. 46. Табл. 6.

Таблица III

0,95 доверительный интервал в % для биномиального распределения [1]

Число наблюдений m	Объем выборки n					
	10	15	20	30	50	100
0	0—31	0—22	0—17	0—12	0—07	0—4
1	0—45	0—32	0—25	0—17	0—11	0—5
2	3—56	2—40	1—31	1—22	0—14	0—7
3	7—65	4—48	3—38	2—27	1—17	1—8
4	12—74	8—55	6—44	4—31	2—19	1—10
5	19—81	12—62	9—49	6—35	3—22	2—11
6	26—88	16—68	12—54	8—39	5—24	2—12
7	35—93	21—73	15—59	10—43	6—27	3—14
8	44—97	27—79	19—64	12—46	7—29	4—15
9	55—100	32—84	23—68	15—50	9—31	4—16
10	69—100	38—88	27—73	17—53	10—34	5—18
11		45—92	32—77	20—56	12—36	5—19
12		52—96	36—81	23—60	13—38	6—20
13		60—98	41—85	25—63	15—41	7—21
14		68—100	46—88	28—66	16—43	8—22
15		78—100	51—91	31—69	18—44	9—24
16			56—94	34—72	20—46	9—25
17			62—97	37—75	21—48	10—26
18			69—99	40—77	23—50	11—27
19			75—100	44—80	25—53	12—28
20			83—100	47—83	27—55	13—29
21				50—85	28—57	14—30
22				54—88	30—59	14—31

Продолжение табл. III

Число наблюдений m	Объем выборки n					
	10	15	20	30	50	100
23				57—90	32—61	15—32
24				61—92	34—63	16—33
25				65—94	36—64	17—35
26				69—96	37—66	18—36
27				73—98	39—68	19—37
28				78—99	41—70	19—38
29				83—100	43—72	20—39
30				88—100	45—73	21—40
31					47—75	22—41
32					50—77	23—42
33					52—79	24—43
34					54—80	25—44
35					56—82	26—45
36					57—84	27—46
37					59—85	28—47
38					62—87	28—48
39					64—88	29—49
40					66—90	30—50
41					69—91	31—51
42					71—93	32—52
43					73—94	33—53
44					76—95	34—54
45					78—97	35—55
46					81—98	36—56
47					83—99	37—57
48					86—100	38—58
49					89—100	39—59
50					93—100	40—60

[1] Таблица взята из: *Снедекор Д. У.* Статистические методы в применении к исследованиям в сельском хозяйстве и биологии. С. 23. Табл. 1.

Примечание. Если m превосходит 50, то следует брать за наблюденное число $100 - m$ и вычитать каждый доверительный предел из 100. Если m/n превосходит 0,5, то за наблюденную долю следует брать $1{,}00 - m/n$ и вычитать каждый доверительный предел из 100.

СПИСОК СОКРАЩЕНИЙ

ГИМ — Государственный Исторический музей
КСИА — Краткие сообщения Института археологии АН СССР
МИА — Материалы и исследования по археологии СССР
СА — Советская археология
СКМА — Статистико-комбинаторные методы в археологии

ЛИТЕРАТУРА

Буйнов Ю. В., Кузьменко А. О. Опыт применения многофакторного анализа данных о погребальном обряде для реконструкции социальной структуры населения Среднего Приднепровья в VIII—XI вв. до н. э. //Вестник Харьковского ун-та. 1985. Сер. История. № 268.

Венецкий И. Г., Кильдишев Г. С. Основы математической статистики. М., 1963.

Деопик Д. В. Соотношение статистических методов, классификаций и культурно-стратиграфических характеристик в археологическом исследовании //Краткие сообщения института археологии. М., 1977. № 148.

Джуракулов М. Д., Холюшкин Ю. П. Некоторые вопросы применения методов математической статистики в археологии каменного века //Материалы по археологии Узбекистана. Тр. Самаркандского ун-та. Самарканд, 1975. Вып. 270.

Каменецкий И. С. Датировка слоев по процентному соотношению типов керамики //Археология и естественные науки. М., 1965.

Каменецкий И. С. К теории слоя //Статистико-комбинаторные методы в археологии. М., 1970.

Каменецкий И. С., Маршак Б. И., Шер Я. А. Анализ археологических источников (Возможности формализованного подхода). М., 1975.

Ковалевская В. Б. Применение статистических методов к изучению массового археологического материала //Археология и естественные науки. М., 1970.

Ковалевская В. Б., Погожев И. Б., Погожева А. П. Количественные методы оценки степени близости памятников по процентному содержанию массового материала //Советская археология. 1970, № 3.

Количественные методы в исторических исследованиях. М., 1984.

Крамер Г. Математические методы статистики. М., 1958.

Лесман Ю. М. К применению методики распознавания образов для анализа керамического комплекса //Новое в применении физико-математических методов в археологии. М., 1979.

Математические методы в исторических исследованиях. М., 1972.

Миркин Б. Г. Группировки в социально-экономических исследованиях. Методы построения и анализа. М., 1985.

Федоров-Давыдов Г. А. Археологическая типология и процесс типообразования (на примере средневековых бус) //Математические методы в социально-экономических и археологических исследованиях. М., 1981.

Холюшкин Ю. П. Проблемы корреляции позднепалеолитических индустрий Сибири и Средней Азии. Новосибирск, 1981.

Шер Я. А. Интуиция и логика в археологическом исследовании //СКМА. М., 1970.

УКАЗАТЕЛЬ ТЕРМИНОВ

«Звездочкой» помечены термины, употребляемые в книге повсеместно. Дается только номер страницы с определением этого термина.

ОГЛАВЛЕНИЕ

УЧЕБНОЕ ИЗДАНИЕ

ГЕРМАН АЛЕКСЕЕВИЧ ФЕДОРОВ-ДАВЫДОВ

СТАТИСТИЧЕСКИЕ МЕТОДЫ В АРХЕОЛОГИИ

Заведующая редакцией Т. Г. Липкина. Научный редактор Б. Г. Миркин. Редактор издательства И. В. Павлова. Младшие редакторы С. М. Ерохина и Л. С. Макаркина. Художественный редактор Т. А. Коленкова. Художник А. В. Алексеев. Технический редактор З. А. Муслимова. Корректор Р. К. Косинова

ИБ № 6484

Изд. № ИСТ-422. Сдано в набор 09.10.86 Подп. в печать 07.01.87. А—03304. Формат 84×108¹/₃₂. Бум. кн.-журн. Гарнитура литературная. Печать высокая. Объем 11,34 усл.-печ. л. 11,45 усл. кр.-отт. 11,23 уч.-изд. л. Тираж 4000 экз. Зак. № 634. Цена 35 коп.

Издательство «Высшая школа». 101430, Москва, ГСП-4, Неглинная ул., д. 29/14.

Московская типография № 8 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 101898, Москва, Центр, Хохловский пер., 7.