Министерство образования и науки Российской Федерации
Ярославский государственный университет им. П. Г. Демидова
Кафедра социологии

Н. С. Гаджигасанова

МЕТОДЫ ПРИКЛАДНОЙ СТАТИСТИКИ ДЛЯ СОЦИОЛОГОВ

Методические указания

Рекомендовано
Научно-методическим советом университета
для студентов, обучающихся по направлению Социология

Ярославль
ЯрГУ
2013

УДК 31:316(072)
ББК С60я73
Г13

Рекомендовано
Редакционно-издательским советом университета
в качестве учебного издания. План 2013 года.

Рецензент
кафедра социологии
Ярославского государственного университета
им. П. Г. Демидова

Гаджигасанова, Н. С. Методы прикладной статистики для социологов: метод. указания / Н. С. Гаджигасанова; Яросл. гос. ун-т им. П. Г. Демидова. — Ярославль : ЯрГУ, 2013. — 72 с.

Г 13

В методических указаниях, выполненных в соответствии с федеральным государственным образовательным стандартом, изложены теоретико-методологические основы дисциплины «Методы прикладной статистики для социологов». В работе рассмотрены теоретические и практические аспекты использования статистики в социологии, описаны конкретные методы анализа со-циологических данных. Каждый раздел сопровождается вопросами и задачами для самопроверки, списком литературы, работа с которыми позволит расширить знания материала.

Предназначены для студентов, обучающихся по направлению 040100.62 Социология (дисциплина «Методы прикладной статистики для социологов», цикл Б2), очной формы обучения.

УДК 31:316(072)
ББК С60я73

Предисловие

Целью учебной дисциплины «Методы прикладной статистики для социологов» в соответствии с федеральным государственным образовательным стандартом высшего профессионального образования является овладение основными методами прикладной статистики, наиболее востребованными и интенсивно применяемыми в социологии. Дело в том, что грамотно используемые статистические методы анализа существенно расширяют возможности научного исследования.

Основной задачей данной работы является ознакомление с современными методами статистической обработки и анализа данных, встречающихся в социологических исследованиях и представленных в различных типах шкал измерений. Особое внимание уделено формированию умений содержательно интерпретировать полученные результаты.

Дисциплина «Методы прикладной статистики для социологов» изучается в следующих формах:

- обязательные учебные занятия, являющиеся основной формой обучения по дисциплине и включенные в учебное расписание;
- индивидуальная работа с преподавателем;
- самостоятельная работа студентов.

Студенты обязаны:

- систематически посещать учебные (теоретические и практические) занятия в дни и часы, предусмотренные учебным расписанием;
- выполнять необходимые контрольные задания для определения уровня освоения теоретического материала;
- активно овладевать знаниями по основам теории и методики дисциплины, используя специальную литературу;
- выполнять соответствующие задания по совершенствованию практических навыков.

Студенты, выполнившие рабочую учебную программу, сдают экзамен. Условием допуска к экзамену является регулярность посещения учебных занятий в объеме, предусмотренном расписанием, а также успешное выполнение контрольных заданий.

Тема 1. Методы прикладной статистики и их возможности в социологии

Общая характеристика методов прикладной статистики. Значение этих методов для социологии. Статистическое подтверждение результатов исследования. Проблемы генерализации данных. Эмпирическая и математическая система, их взаимосвязь.

Методические рекомендации для студентов

Данная тема представляет собой введение в проблематику дисциплины и включает вопросы, которые дают общую характеристику методов прикладной статистики. Как объект изучения прикладная статистика — методическая дисциплина, являющаяся центром статистики. Методы прикладной статистики активно применяются во многих областях социогуманитарного знания, в том числе и в социологии.

Методы, применяемые социологами для анализа данных[1], многообразны. В первую очередь выбор конкретного метода обусловлен характером исследовательских гипотез, т. е. от того, на какие вопросы мы хотим получить ответ. Если целью является описание одной характеристики выборки в определенный момент времени, целесообразно ограничиться одномерным анализом. Разнообразные техники многомерного анализа позволяют одновременно исследовать взаимоотношения двух и более переменных и в той или иной форме проверять гипотезы о причинных связях между ними.

В реальном исследовании каждое уточнение исходных гипотез или выдвижение новой гипотезы в ходе анализа результатов приводит к необходимости выбора новой техники анализа данных. Так, если изначальная модель взаимоотношения двух переменных[2] (скажем, профессии и дохода) не позволяет выявить

[1] Данные — под ними мы будем понимать информацию, полученную в результате социологического исследования (ответы респондентов, оценки экспертов, результаты наблюдения и т. п.; совокупность значений переменных, приписанных единицам исследования — объектам).

[2] Понятия признак / переменная могут использоваться как взаимо-

определенную закономерность в собранных данных, исследователь выбирает одну из статистических техник, позволяющих контролировать влияние какой-то третьей переменной, например пола, на интересующее его отношение.

Помимо характера исследовательских гипотез, на выбор методов статистического анализа влияет и природа полученных социологом данных. Дело в том, что разные уровни измерения социологических переменных определяют возможности и ограничения анализа. Для того чтобы охарактеризовать распределение в выборке такого номинального признака, как «пол», мы не можем воспользоваться его среднеарифметическим значением и, следовательно, нам потребуются какие-то другие приемы компактного и точного представления полученной информации.

Методы, используемые для анализа связи между двумя номинальными переменными, также будут отличаться от методов анализа связи между номинальной переменной и переменной, измеренной на интервальном уровне. Таким образом, выбор той или иной статистики будет зависеть и от целей анализа, и от уровня измерения исследуемых переменных.

Существуют два основных класса задач, решаемых с помощью статистических методов анализа. ***Задачей дескриптивной (описательной) статистики*** является описание распределения переменной-признака в конкретной выборке. Совокупность наиболее употребительных приемов получения закономерностей, описывающих изучаемое множество объектов, называется ***описательной, или дескриптивной, статистикой***. Методы дескриптивной статистики позволяют также анализировать взаимосвязь между различными переменными.

Другой класс задач, связанный с необходимостью вывести свойства большой совокупности, основываясь на имеющейся информации о свойствах выборки из этой совокупности, решается

заменяемые. Признак — некоторое общее для всех объектов качество, конкретные проявления которого (значения признака; их называют также альтернативами, градациями) могут меняться от объекта к объекту (например, пол, возраст респондентов, их удовлетворенность свои трудом). В качестве значений признака «возраст» могут выступать 25 лет, 48 лет, 21 год.

с помощью ***методов индуктивной статистики***, или теории статистического вывода, основанной на вероятностном подходе к принятию решений. Воспользовавшись какой-то моделью для анализа полученных выборочных данных, социолог обычно также применяет некоторые методы статистического вывода, позволяющие определить, выполняются ли обнаруженные им при анализе данных отношения на уровне большой совокупности, из которой была извлечена выборка.

Для прикладной статистики роль математического фундамента выполняет математическая статистика. Математическая статистика позволяет выявить широкий круг статистических закономерностей (наборов параметров вероятностных распределений одномерных и многомерных случайных величин): меры средней тенденции, разброса значений случайных величин, связи между признаками и т. д. Статистическая закономерность возникает как результат взаимодействия большого числа элементов, составляющих совокупность, и характеризует не столько поведение отдельного элемента совокупности, сколько всю совокупность в целом. Она адекватно описывает массовые явления случайного характера, а именно такого рода явления и изучает обычно социолог. Анализ данных с помощью математических методов позволяет выявлять статистические закономерности.

Статистическая закономерность — закономерность, проявляющаяся в массе однородных явлений при обобщении данных статистической совокупности. Например, в социологической практике статистическими являются следующие утверждения:

- средний возраст рабочих на предприятии равен 30 годам;
- выбор профессии выпускниками школ не связан с их гендерной / половой принадлежностью;
- такой-то радиоканал (например, «Европа-плюс») имеет самый высокий рейтинг среди слушателей.

Применение математико-статистических методов в социологии опирается на то, что мы считаем возможным:

- выделить некоторый фрагмент реальности;
- построить (посредством измерения) его математическую модель (т. е. получить исходные данные);

- изучить эту модель традиционными для статистики способами (применить тот или иной алгоритм анализа данных) и прийти к некоторым выводам (в результате анализа данных получить математический результат: точное значение коэффициента корреляции, параметры уравнения регрессии и т. д.);

- проинтерпретировать эти выводы и получить, таким образом, новое знание.

Первые два этапа обычно относят к области измерения (шкалирования), последние два — к области анализа данных. Но все этапы тесно связаны друг с другом.

Выделенный фрагмент реальности называется **эмпирической системой** (ЭС). Назовем эмпирической системой (ЭС) интересующую исследователя совокупность реальных (эмпирических) объектов с выделенными соотношениями между ними. Последние часто можно выразить в виде некоторых отношений между объектами (любое отношение есть соотношение, но не наоборот), и тогда говорят об эмпирической системе с отношениями (ЭСО).

Пример ЭСО — совокупность сотрудников какого-либо предприятия, рассматриваемых как «носители» удовлетворенности своим трудом с заданным бинарным (т. е. определенным на парах объектов) отношением: «респондент А больше удовлетворен работой, чем респондент Б». Для одних пар это отношение может выполняться, для других нет. Но мы полагаем, что, каких бы респондентов мы ни взяли, разговор о выполнении этого отношения будет осмысленным. ЭС отражает представление исследователя об изучаемой реальности, процесс ее формирования по существу является моделированием. С учетом этого ЭС можно считать фрагментом реальности.

Процесс перевода всех компонент фрагмента реальности на формальный, математический язык, т. е. процесс измерения, позволяет нам перейти от ЭС к МС — математической системе (в социологии она может быть числовой или нечисловой).

Назовем математической системой (МС) совокупность математических объектов (чаще всего в качестве таковых выступают числа, и тогда МС называется числовой) с выделенными соотношениями между ними. Когда последние задаются в виде некоторых от-

ношений между объектами, говорят о математической системе с отношениями или о числовой системе с отношениями (МСО и ЧСО).

Таким образом, использование математических методов в процессе проведения социологического исследования позволяет достичь следующих целей:

1. Побуждает исследователя четко формулировать свои представления об изучаемом объекте. Необходимым условием успешности здесь является комплексность анализа (использование группы методов). Так, желая сравнить величину связи между какими-либо признаками для разных совокупностей респондентов, мы, пытаясь построить математический критерий такой связи, вынуждены конкретизировать свои представления о ней. Это можно сделать многими способами (только коэффициентов парной связи между номинальными признаками известно более сотни; имея перед собой множество таких коэффициентов, мы можем понять, что есть наша связь в реальности).

2. Позволяет абстрагироваться от большого количества реальных свойств изучаемых объектов.

3. Дает возможность получить содержательные выводы за счет расширения круга логических умозаключений.

4. Дает возможность выявить скрытые механизмы взаимодействий при анализе огромных массивов информации (с которыми обычно и имеет дело социолог) и учете огромного количества факторов (определяющих любое общественное явление).

Характерной задачей, которая решается исследователем в процессе анализа анкетных массивов, является нахождение сочетаний значений признаков, которые детерминируют некоторое поведение респондента (скажем, голосование или неголосование на выборах). Результатом решения подобной задачи может служить, например, вывод, что среди мужчин старше 40 лет с высшим экономическим образованием, живущих в сельской местности, 95 % проголосовало за лидера, т. е. что для респондентов, обладающих названными свойствами, характерна данная модель поведения. Но подобный вывод некорректен, т. к. мы не обнаруживаем всех требующихся групп респон-

дентов. В таком случае могут помочь специфические алгоритмы (например, алгоритмы типа AID, рассматриваемые ниже).

Кроме того, необходимо отметить, что практически перед любым исследователем-социологом очень остро стоит ещё одна проблема — проблема соотнесения выборки и генеральной совокупности. Генерализация (лат. generalis — общий, главный) — метод познания, позволяющий на основании выделения множества элементов, имеющих однотипную характеристику (генеральной совокупности), и выбора единицы анализа изучать массивы (системы) этих элементов.

Прежде всего отметим, что выборочные оценки параметров, рассчитанные на основе частотных распределений, называются статистиками. Выборочные исследования — способ получения статистических данных и важная часть прикладной статистики. В прикладной статистике есть методы определения необходимого объема выборки, которые основаны на разных подходах: 1) на задании необходимой точности оценивания параметров; 2) на явной формулировке альтернативных гипотез, между которыми необходимо сделать выбор; 3) на учете погрешностей измерений[3].

Так как при изучении статистических закономерностей социолога практически всегда интересует задача перенесения полученных им результатов с той совокупности объектов, которая непосредственно была обследована (с выборки), на более широкую совокупность (генеральную), то это делает использование математической статистики еще более привлекательным для социолога, так как с помощью соответствующих подходов можно осуществлять анализ выборочных данных именно с намерением обобщения получаемых результатов на соответствующую генеральную совокупность. Дело в том, что вид закономерности, найденной для выборки будет отличаться от вида ее для генеральной совокупности. В силу этого важную роль должна играть оценка подобного различия, поскольку нас интересуют закономерности, свойственные генеральной совокупности, хотя на практике мы и имеем дело лишь с выборкой. Именно такую оценку мы сможем сделать, пользуясь положениями математической статистики.

[3] См. подробнее: Орлов А. И. Прикладная статистика.

Перейти от статистики к закономерностям генеральной совокупности можно, используя методы математического характера. **Основные методы математической статистики** обычно делят на две группы:

- методы статистической оценки параметров (способы расчета выборочных значений параметров и перехода от выборочных значений к генеральным; математическая статистика говорит о том, какими качествами эти оценки должны обладать, чтобы как можно более походить на их генеральные прообразы, и каким образом надо строить «хорошие» статистики, отражающие параметры вероятностных распределений);

- методы проверки статистических гипотез (оценка степени правдоподобности гипотезы о наличии некоторых соотношений между случайными величинами в генеральной совокупности на основании расчета определенных характеристик соответствующих выборочных распределений).

Важно учитывать, что правила переноса результатов с выборки на генеральную совокупность базируются на рассмотрении некоторых выборочных статистик как случайных величин и изучении определенных параметров их вероятностных распределений (скажем, если статистика — среднее арифметическое значение какого-либо признака, то упомянутое распределение для нее получится, если представить себе бесконечное количество выборок одного и того же размера и расчет для каждой выборки этого среднего; заметим, что, как известно, дисперсия такого распределения средних обычно называется средней ошибкой выборки и очень часто используется в эмпирических исследованиях).

Математический аппарат, используемый в эмпирической и прикладной социологии, предлагает для выявления связи между явлениями, определения ее направления и силы большое число специализированных процедур. Выбор их для конкретного исследования зависит от задач исследования, от уровня его подготовки, от корректности целей.

Таким образом, без применения математического аппарата трудно обойтись при решении практически любой социологической задачи.

Контрольные вопросы к теме

1. Что такое генерализация выводов исследования и при каких условиях она осуществляется?

2. Укажите связь математической и эмпирической системы.

3. Какие методы прикладной статистики вам уже известны из предыдущих курсов? Как они могут использоваться в социологии?

Список литературы

1. Гнеденко, Б. В. Курс теории вероятностей / Б. В. Гнеденко. — М. : Наука, 1965.

2. Епархина, О. В. Математические методы обработки и анализа данных в социологии / О. В. Епархина. — Ярославль : ЯрГУ, 2007.

3. Крамер, Д. Математическая обработка данных в социальных науках / Д. Крамер. —— М. : Академия, 2007.

4. Крыштановский, А. О. Анализ социологических данных с помощью пакета SPSS / А. О. Крыштановский. — М. : ГУ ВШЭ, 2006.

5. Орлов, А. И. Прикладная статистика / А. И. Орлов. — М. : Экзамен, 2004.

6. Основы прикладной социологии / под ред. Ф. Э. Шереги и М. К. Горшкова. — М. : Инфра-М, 1996.

7. Рабочая книга социолога / под ред. Г. В. Осипова. — М. : URSS, 2006.

8. Сидоренко, Е. В. Методы математической обработки в психологии / Е. В. Сидоренко. — СПб. : Речь, 2002.

9. Сикевич, З. В. Социологическое исследование / З. В. Сикевич. — СПб. : Питер, 2005.

10. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

11. Хилл, Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хилл. — СПб. : Питер, 2005.

12. Ядов, В. А. Стратегия социологического исследования / В. А. Ядов. — М. : Омега-Л, 2005.

Тема 2. **Проблемы измерения в социологии и виды шкал**

Понятие измерения в социологии. Виды шкал. Низкие и высокие шкалы. Номинальная, порядковая, интервальная шкала и шкалы отношений. Типы данных. Правила ранжирования. Правило связанных рангов.

Методические рекомендации для студентов

Изучение социальных процессов предполагает выявление не только их качественных, но и количественных характеристик, процесс получения которых основан на так называемой процедуре измерения.

Процесс измерения в наиболее общем виде есть квантификация[4] свойств изучаемого явления, т. е. присвоение им числовых значений по заданным правилам[5].

Измерение в социологии — процедура, при помощи которой объекты исследования, рассматриваемые как носители определенных отношений и как таковые составляющие эмпирическую систему, отображаются в некоторую математическую систему с соответствующими отношениями между ее элементами[6]. Любые интересующие социолога объекты могут выступать в качестве объектов измерения — индивиды, производственные коллективы, условия труда, быта и т. д. При измерении каждому объекту приписывается определенный элемент используемой математической системы. В социологической практике чаще всего используются числовые математические системы, т. е. такие, элементами которых являются действительные числа. Однако возможно использование нечисловых систем: частично упорядоченных множеств, графов, матриц и т. д. В отношения, моделируемые при из-

[4] Квантификация (от лат. quantum — сколько и facere — делаю) — количественное выражение качественных признаков.

[5] Горшков М. К., Шереги Ф. Э. Прикладная социология: методология и методы: интерактивное учеб. пособие. М.: Институт социологии РАН, 2011. 1 CD ROM.

[6] Российская социологическая энциклопедия / под общ. ред. Г. В. Осипова. С. 145.

мерении, объекты вступают как носители определенных свойств. Вследствие этого вместо термина «измерение объектов» часто используется термин «измерение свойств объектов». В процессе познания измерение есть связующее звено между социальным объектом и его математическим представлением.

Подход к осмыслению измерения, который в настоящее время находит наиболее широкое практическое применение в социологии, начал формироваться на рубеже XIX–XX вв. Его возникновение было обусловлено потребностями общественных наук, которые к этому времени достигли уровня, когда дальнейшее интенсивное их развитие без использования формальных моделей изучаемых процессов или явлений стало немыслимым. Последствием этого стал особый интерес к проблеме измерения. Кроме того, непригодность классического подхода для измерения в общественных науках обусловила расширение данного понятия. В результате измерение стало интерпретироваться как способ приписывания чисел объектам независимо от того, использовалась ли при этом единица измерения.

Одним из основоположников нового подхода к пониманию измерения стал американский психолог С. С. Стивенс, автор классификации шкал по уровню измерения. Исследователь первым сформулировал положение о том, что система арифметических отношений между числами, как правило, шире, чем те эмпирические отношения между объектами, которые моделируются с помощью этих чисел.

С. Стивенс предложил классификацию из 4 типов шкал измерения:

1) номинативная, или номинальная, или шкала наименований;

2) порядковая, или ординальная, шкала;

3) интервальная, или шкала равных интервалов;

4) шкала равных отношений.

Номинативная шкала — это шкала, классифицирующая по названию: nomen (лат.) — имя, название. Название не измеряется количественно, оно лишь позволяет отличить один объект от другого или одного субъекта от другого. Номинативная шкала — это способ классификации объектов или субъектов, распределения их по ячейкам классификации.

Простейший случай номинативной шкалы — дихотомическая шкала, состоящая лишь из двух ячеек, например: «имеет братьев и сестер — единственный ребенок в семье»; «иностранец — соотечественник»; «проголосовал "за" — проголосовал "против"» и т. п.

Признак, который измеряется по дихотомической шкале наименований, называется альтернативным. Он может принимать всего два значения. При этом исследователь зачастую заинтересован в одном из них, и тогда он говорит, что признак «проявился», если тот принял интересующее его значение, и что признак «не проявился», если он принял противоположное значение. В принципе номинативная шкала может состоять из ячеек «признак проявился — признак не проявился».

Более сложный вариант номинативной шкалы — классификация из трех и более ячеек, например: «выбор кандидатуры А — кандидатуры Б — кандидатуры В — кандидатуры Г» или «старший — средний — младший — единственный ребенок в семье».

Таким образом, номинативная шкала позволяет нам подсчитывать частоты встречаемости разных «наименований», или значений признака, и затем работать с этими частотами с помощью математических методов.

Типичным признаком, значения которого обычно получаются именно по номинальной шкале, является профессия респондента. Если одному респонденту приписано значение «3» («токарь»), а другому значение «4» («пекарь»), то, имея в руках эти числа, мы можем быть уверенными в том, что рассматриваемые объекты в интересующем нас отношении различны (респонденты имеют разные профессии), но больше ничего мы о них сказать не можем.

Единица измерения, которой при этом пользуется исследователь, — количество наблюдений (испытуемых, реакций, выборов и т. п.) либо частота. Точнее, единица измерения — это одно наблюдение.

Порядковая (ранговая) шкала — это шкала, классифицирующая по принципу «больше — меньше». В порядковой шкале классификационные ячейки образуют последовательность от ячейки «самое малое значение» к ячейке «самое большое значение» (либо наоборот). В данном случае ячейки уместнее назы-

вать классами. В порядковой шкале должно быть не менее трех клас-сов, например «положительная реакция — нейтральная реакция — отрицательная реакция» или «подходит для занятия вакантной должности — подходит с оговорками — не подходит».

В порядковой шкале исследователь знает лишь, что классы образуют последовательность, в то время как истинное расстояния между классами ему неизвестно. Например, классы «подходит для занятия вакантной должности» и «подходит с оговорками» могут быть реально ближе друг к другу, чем класс «подходит с оговорками» к классу «не подходит». От классов легко перейти к числам, если мы сделаем допущение, что низший класс получает ранг 1, средний класс — ранг 2, а высший класс — ранг 3, или наоборот. Чем больше классов в шкале, тем больше у нас возможностей для математической обработки полученных данных и проверки статистических гипотез. Единица измерения в шкале порядка — расстояние в 1 класс или в 1 ранг, при этом расстояние между классами и рангами может быть разным (которое исследователю неизвестно).

При использовании порядковой шкалы исследователь ставит целью отобразить не только некоторое отношение равенства-неравенства между реальными объектами, но и содержательное отношение порядка между ними. В качестве примера может служить анкета с вопросом *«Удовлетворены ли Вы Вашей работой (ходом реформ, президентом РФ...)?»* и веером из 5 (3, 7 и т. д.) вариантов ответов от *«Совершенно не удовлетворен»* до *«Вполне удовлетворен»*, которым ставятся в соответствие числа от 1 до 5 (от 1 до 3, от 1 до 7, от –3 до +3 и т. д.). В этом случае исследователь при осуществлении шкалирования[7] ставит целью отобразить в числах не только отношение равенства респондентов по их удовлетворенности объектом, но и отношение порядка между респондентами по степени «накала» их эмоций, направленных в адрес этого объекта. И если выяснится, что одному респонденту приписано число «2», а другому — «4», то будет предпола-

[7] Шкалирование — совокупность методов измерения, посредством которых эмпирическая система отношений трансформируется в соответствующую числовую систему.

гаться, что упомянутый «накал» второго респондента не просто не равен «накалу» первого, но больше такового.

Интервальная шкала — это шкала, классифицирующая по принципу «больше на определенное количество единиц — меньше на определенное количество единиц». В интервальных шкалах полученные данные похожи на действительные числа, но все же таковыми не являются. Они отображают в числовых отношениях не только некоторые эмпирические отношения равенства и порядка, но и структуру эмпирических интервалов — отношения равенства и порядка для расстояний между объектами.

Интервальная шкала применяется в прикладной социологии для измерения весьма небольшого числа свойств, значения которых в основном можно выразить числом: возраст, стаж работы, число членов семьи, доход и др. Позиции в такой шкале расположены, как правило, через равные интервалы, но иногда могут располагаться и через неравные интервалы, хотя это нежелательно, так как уменьшает точность вычислений. Шкала с *равными интервалами* имеет вид: *«Сколько Вам лет?»*

- от 16 до 25 лет включительно;
- от 26 до 35 лет включительно;
- от 36 до 45 лет включительно;
- от 46 до 55 лет включительно;
- от 56 до 65 лет включительно.

Шкала с *неравными интервалами* имеет вид: *«Сколько лет Вы работаете на данном предприятии?»*

- менее года;
- от 1 года до 3 лет включительно;
- от 4 до 5 лет включительно;
- от 6 до 10 лет включительно;
- свыше 10 лет.

Очень важно следить за тем, чтобы варианты ответа на вопрос соотносились между собой по всем правилам построения соответствующей шкалы, что гарантирует применимость при анализе ответов правил математической, прикладной статистики.

Шкала равных отношений — это шкала, классифицирующая объекты или субъектов пропорционально степени выражен-

ности измеряемого свойства. В шкалах отношений классы обозначаются числами, которые пропорциональны друг другу: 2 так относится к 4, как 4 к 8. Это предполагает наличие абсолютной нулевой точки отсчета. Шкала отношений — подмножество интервальных шкал[8].

Шкалами низкого типа считают шкалы номинальные и порядковые, а шкалами высокого типа — интервальные и шкалы отношений. Шкалы низкого типа (и получаемые с их помощью данные) называют также качественными, а шкалы высокого типа (и соответствующие данные) — количественными, или числовыми.

Возможности использования математической статистики для изучения данных, полученных по шкалам низких типов, подробнее изучаются статистикой объектов нечисловой природы. Для чисел, полученных по шкалам низких типов, не имеет смысла большинство традиционных операций с числами. Например, вряд ли найдется человек, усматривающий что-то рациональное в утверждениях: «Среднее арифметическое значение профессий для рассматриваемой совокупности респондентов равно 3,7, и оно меньше аналогичного среднего значения для другой совокупности, равного 3,9». А потребность использования в социологии шкал низких типов (отвечающих шкальным значениям, являющимся неполноценными числами) заставило исследователей с особым вниманием отнестись к тому, что на множестве чисел возможно задание разных структур.

Данные — под ними мы будем понимать информацию, полученную в результате социологического исследования (ответы респондентов, оценки экспертов, результаты наблюдения и т. п.; совокупность значений переменных, приписанных единицам исследования — объектам).

Построение распределения ряда данных — это разделение первичных данных, полученных на выборке, на классы или категории с целью получить обобщенную упорядоченную картину, позволяющую их анализировать.

---

[8] См.: Российская социологическая энциклопедия / под общ. ред. Г. В. Осипова. С. 614–616; Сидоренко Е. В. Методы математической обработки в психологии.

Существуют ***три типа данных***:

1. *Количественные данные*, получаемые при измерениях (например, данные о весе, размерах, температуре, времени, результатах тестирования и т. п.). Их можно распределить по шкале с равными интервалами.

2. *Порядковые данные*, соответствующие местам этих элементов в последовательности, полученной при их расположении в возрастающем порядке.

3. *Качественные данные*, представляющие собой какие-то свойства элементов выборки или популяции. Их нельзя измерить, и единственной их количественной оценкой служит частота встречаемости.

Из всех этих типов данных только количественные данные можно анализировать с помощью методов, в основе которых лежат параметры (такие, например, как средняя арифметическая, мода, дисперсия и т. д.). Но даже к количественным данным такие методы можно применить лишь в том случае, если число этих данных достаточно, чтобы проявилось нормальное распределение.

При обработке и анализе данных широко применяются методы ранжирования, шкалирования и др.

Ранжирование — это процедура установления относительной значимости (предпочтительности) исследуемых объектов на основе их упорядочивания. Ранг — это показатель, характеризующий порядковое место оцениваемого объекта в группе и других объектов, обладающих существенными для оценки свойствами. Для каждого объекта вычисляют сумму рангов, полученную от всех экспертов, затем упорядочивают эту сумму. Ранг 1 присваивают объекту, получившему наименьшую сумму, самый низкий ранг — объекту с наивысшей суммой.

Правила ранжирования. Использование порядковой шкалы позволяет присваивать ранги объектам по какому-либо признаку. Таким образом, количественные значения переводятся в ранговые. При этом фиксируются различия в степени выраженности свойств. В процессе ранжирования следует придерживаться двух правил.

Правило порядка ранжирования. Надо решить, кто получает первый ранг: объект с самой большей степенью выраженности

какого-либо качества или наоборот. Чаще всего это абсолютно безразлично и не отражается на конечном результате. Традиционно принято первый ранг приписывать объектам с большей степенью выраженности качества (большему значению — меньший ранг). Например, чемпиону присуждают первое место, а не наоборот. Хотя и здесь, если бы был принят обратный порядок, результаты от этого не изменились бы. Так что порядок ранжирования каждый исследователь вправе определять сам. Например, Е. В. Сидоренко[9] рекомендует меньшему значению приписывать меньший ранг.

Правило связанных рангов. Поскольку одинаковым значениям признака нельзя присвоить разные ранги, то каждому совпадающему значению присваивается ранг, равный среднему арифметическому из тех рангов, которые имели бы эти элементы, будь они различны. Продемонстрируем технику расчета связанных рангов в табл. 1.

Наименьшему значению (xi = 9) присвоим наименьший ранг, т. е. 1. Далее идут два одинаковых значения 10. В случае если бы значения были разные (например, 10 и 11), им присвоили бы ранги 2 и 3 соответственно. Но поскольку значения одинаковые, то им присваивается средний из двух этих рангов, а именно (2 + 3) / 2 = 2,5. Далее идут три одинаковых значения xi, равные 12. В случае если бы значения были разные (например, 12, 13 и 14), им присвоили бы ранги 4, 5 и 6 соответственно. Но поскольку значения одинаковые, то им присваивается средний из этих трех рангов, а именно (4 + 5 + 6) / 3 = 5. В результате получаем последовательность значений признака xi и соответствующих рангов (см. табл. 1).

Таблица 1

Расчет связанных рангов

xi	9	10	10	12	12	12	15	17
Ранг xi	1	2,5	2,5	5	5	5	7	8

Два подряд идущих ранга 2,5, равно как и три подряд идущих ранга 5, называются *связанными рангами*.

Ранжирование дополняется, как правило, другими методами экспертных оценок.

[9] См. подробнее: Сидоренко Е. В. Методы математической обработки в психологии.

Контрольные вопросы к теме

1. Раскройте понятие измерения в социологии.

2. Какие типы шкал используются в прикладной статистике и чем они различаются?

3. Приведите примеры номинальной, порядковой, интервальной шкалы и шкал отношений.

4. Какие математические и логические операции можно производить с данными, полученными по каждой шкале?

5. Какие взаимные преобразования допустимы со шкалами разного порядка?

6. Что такое ранжирование?

Список литературы

1. Гнеденко, Б. В. Курс теории вероятностей / Б. В. Гнеденко. — М. : Наука, 1965.

2. Горшков, М. К. Прикладная социология: учеб. пособие для вузов / М. К. Горшков, Ф. Э. Шереги. — М., 2003.

3. Епархина, О. В. Математические методы обработки и анализа данных в социологии / О. В. Епархина. — Ярославль : ЯрГУ, 2007.

4. Илышев, А. М. Общая теория статистики: учеб. пособие / А. М. Илышев. — М. : КНОРУС, 2013.

5. Орлов, А. И. Прикладная статистика / А. И. Орлов. — М. : Экзамен, 2004.

6. Паниотто, В. И. Количественные методы в социологических исследованиях / В. И. Паниотто, В. С. Максименко. — Киев : Наукова думка, 1982.

7. Рабочая книга социолога / под ред. Г. В. Осипова. — М. : URSS, 2006.

8. Российская социологическая энциклопедия / под общ. ред. Г. В. Осипова. — М. : НОРМА-ИНФРА, 1998.

9. Сидоренко, Е. В. Методы математической обработки в психологии / Е. В. Сидоренко. — СПб. : Речь, 2002.

10. Сикевич, З. В. Социологическое исследование / З. В. Сикевич. — СПб. : Питер, 2005.

11. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

12. Толстова, Ю. Н. Измерение в социологии / Ю. Н. Толстова. — М. : Инфра-М, 1998.

13. Толстова, Ю. Н. Математика в социологии: элементарное введение в круг основных понятий (измерение, статистические закономерности, принципы анализа данных) / Ю. Н. Толстова. — М. : ИСАН СССР, 1990.

14. Хилл, Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хилл. — СПб. : Питер, 2005.

15. Ядов, В. А. Стратегия социологического исследования / В. А. Ядов. — М. : Омега-Л, 2005.

Тема 3. **Описательные статистики**

Меры центральной тенденции: мода, медиана, средние. Меры изменчивости: размах, дисперсия, стандартное отклонение. Формулы приближенных вычислений.

Методические рекомендации для студентов

Важной характеристикой при описании поведения отдельных признаков и соответствующих им переменных являются меры средней тенденции. Меры центральной (средней) тенденции — это некоторое значение рассматриваемого признака, которое должно характеризовать всю совокупность (как бы подменять ее). Такие значения переменной, которые выступают точечной оценкой выборочной или генеральной совокупности, называются статистикой. Возможности использования различных мер средней тенденции для шкал различного типа приведены в табл. 2.

Среднее арифметическое (или просто среднее) — сумма значений переменной, поделенная на число значений.

Среднее арифметическое широко используется, но применение лишь этой статистики таит в себе опасность. Говоря о среднем значении переменной, мы подменяем рассмотрение всей совокупности ее значений одним показателем. При этом мы

предполагаем, что значение данного показателя достаточно хорошо описывает поведение анализируемой переменной (т. е. выступает в качестве модели).

Таблица 2

Меры центральной (или средней) тенденции

Тип шкалы	*Допустимые меры средней тенденции*
Номинальный	Мода
Порядковый	
Ранговый	Мода, медиана
Интервальный	Мода, медиана,
среднее арифметическое	

Среднее арифметическое значение, вычисленное для какой-либо группы респондентов, *чаще всего интерпретируется как значение наиболее типичного для этой группы человека*. Но если признак в этой группе распределен неравномерно, то подобная интерпретация неуместна. Так, например, зная среднее значение зарплаты опрошенных, нельзя точно определить размер зарплаты у определенного респондента. Средний доход у предпринимателей будет различаться в зависимости от региона: г. Иваново, г. Москва или Дальний Восток. Следовательно, среднее арифметическое переменной неполно представляет совокупность значений этой переменной, что может привести к ошибкам.

Только в случае, когда все значения переменной одинаковы, среднее значение абсолютно точно отражает поведение признака. Во всех других случаях среднее значение как модель является неточным.

Среднее арифметическое чувствительно к средним значениям (если к посетителям библиотеки добавить 80-летнего читателя, то показатель среднего арифметического возраста читателей вырастет). Следовательно, сами по себе значения средних мало что говорят. Они не отражают качество модели среднего.

Мода. Для номинальных переменных мерой центральной тенденции может выступать только *мода* — наиболее часто встречающееся значение переменной. Мода не имеет какого-

либо показателя разброса. Определенной характеристикой может считаться лишь само процентное значение модальной величины.

Распределение может иметь и не одну моду. Когда все значения встречаются одинаково часто, принято считать, что такое распределение не имеет моды. В случае если распределение имеет несколько мод, то говорят, что оно мультимодально или многомодально (имеет два или более «пика»).

Мультимодальность распределения дает важную информацию о природе исследуемой переменной. Так, например, в социологических опросах, если переменная представляет собой предпочтение либо отношение к чему-то, то мультимодальность может означать, что существует несколько определенно различных мнений.

Медиана. Для переменных, измеренных на порядковом уровне, основной мерой центральной тенденции является медиана. *Медиана* — это значение признака, которое делит *вариационный ряд*, отвечающий этому признаку, пополам. Вариационный ряд — последовательность значений признака, расположенных в порядке их возрастания.

Медиана — это такое значение переменной, меньше которого отметили 50 % респондентов. Таким образом, медиана обладает тем свойством, что половина всех выборочных значений признака меньше ее, а половина — больше.

Данная мера центральной тенденции имеет смысл только для порядковых и интервальных шкал (для номинальных она не подходит, поскольку ее интерпретация будет бессмысленна с содержательной стороны). Например, мы имеем 11 измеренных значений: 3, 7, 8, 5, 4, 6, 3, 9, 2, 8, 4. Вариационный ряд будет представлять собой упорядоченную в порядке возрастания совокупность значений — 2, 3, 3, 4, 4, 5, 6, 7, 8, 8, 9. В этом случае медиана равна 5. В случае если вариационный ряд содержит четное число измерений, например: 12 — 2, 3, 3, 4, 4, 5, 6, 7, 8, 8, 8, 9, — то медиана будет равна среднему арифметическому двух центральных значений: Me = (5+6) / 2 = 5,5.

Меры центральной тенденции позволяют нам судить о концентрации исходных данных на числовой оси. Каждая такая мера дает значение, которое представляет в каком-то смысле

все элементы выборки. В этой ситуации фактически пренебрегают различиями, существующими между отдельными элементами выборки. Поэтому для учета таких различий используются другие описательные статистики, которые называются ***мерами изменчивости*** (рассеяния, разброса). Самой простой мерой изменчивости является *размах выборки*, для вычисления которого необходимо из максимального элемента выборки вычесть минимальный ($R = x_{max} — x_{min}$).

Функцию оценки качества модели среднего выполняют меры разброса: дисперсия, среднеквадратичное отклонение, стандартная ошибка среднего.

Дисперсия — это мера вариации значений признака в среднем и вокруг средней арифметической. Фактически это сумма квадратов остатков, деленная на число наблюдений.

Для того чтобы вычислить значение дисперсии, надо вычесть из каждого наблюдаемого значения среднее, возвести в квадрат все полученные отклонения, сложить квадраты отклонений и разделить полученную сумму на n:

$$D = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n},$$

где x — каждое наблюдаемое значение признака;

$\bar{X}$ (*с черточкой сверху*) — среднее арифметическое значение признака (переменной x);

n — количество наблюдений.

Для того чтобы сделать соответствующую точечную оценку дисперсии несмещенной, величина объема выборки в знаменателе уменьшается на 1:

$$S^2(x) = \frac{\sum_{i}(x_i - \bar{x})^2}{n-1},$$

где xi — каждое наблюдаемое значение признака;

$\bar{X}$ (*с черточкой сверху*) — среднее арифметическое значение признака (переменной x);

n — количество наблюдений.

В зависимости от того, насколько велика (мала) дисперсия или среднеквадратическое отклонение, можно судить, насколько

единодушны были в своих оценках респонденты (при меньшем значении дисперсии) или насколько сильно они расходятся в своих мнениях (при большем значении дисперсии).

Недостатком дисперсии является то, что это величина безразмерная. Мы можем понять размер доходов и единицы измерения остатков, но в данном случае дисперсия равна 4 000 000. Вряд ли можно сказать, большая это величина или маленькая. Кроме того, данное значение не позволяет определить качество модели среднего, поскольку в формуле расчета дисперсии остатки берутся в квадрате.

Для того чтобы преодолеть эти трудности, существуют два производных от дисперсии показателя — стандартное (среднеквадратичное) отклонение и стандартная ошибка среднего.

Стандартное отклонение — это корень квадратный из дисперсии:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n}}$$

где xi — каждое наблюдаемое значение признака;

$\overline{X}$ (*с черточкой сверху*) — среднее арифметическое значение признака (переменной x);

n — количество наблюдений.

Очевидной интерпретацией стандартного отклонения является его способность оценивать «типичность» среднего: тем меньше, чем лучше среднее представляет совокупность.

Зная значение среднеквадратического отклонения, можно сравнивать меры рассеяния разных признаков или одного признака для различных совокупностей. Прямое сравнение дисперсий и среднеквадратических отклонений без сопоставления со средними арифметическими является бессмысленным.

Интерпретировать данные показатели в совокупности можно следующим образом.

Допустим, нами были рассчитаны среднее арифметическое и среднеквадратическое отклонения затрат времени на домашнюю уборку для нескольких групп женщин: домохозяек ($x = 6$, $\sigma = 4$), предпринимателей ($x = 4{,}5$, $\sigma = 3{,}5$), служащих ($x = 5{,}4$ $\sigma = 3{,}5$), вре-менно не работающих ($x = 6$, $\sigma = 2$)[10]. Из полученных данных

[10] Данные взяты произвольно.

видно, что женщины-домохозяйки и женщины, временно не работающие, затрачивают на домашнюю уборку в среднем одинаковое время, но совокупность домохозяек менее однородна, потому что среднеквадратическое отклонение больше. Женщины-служащие затрачивают на домашнюю уборку в среднем больше времени, чем женщины-предприниматели (дисперсия одинакова в этих группах). Когда средние и дисперсии в сравниваемых группах различны, то необходимо рассчитать коэффициент вариации.

Коэффициент вариации определяется просто как процент наблюдений, лежащих вне модального интервала, т. е. процент (доля) наблюдений, не совпадающих с модальным значением. Например, если от модального отличаются 60 % значений, то V = 60 % (или V = 0,6).

Наряду со стандартным отклонением для оценки разброса используется и *стандартная ошибка среднего*. Основной причиной ее активного использования является то, что в интервале (среднее значение) ± две стандартных ошибки среднего должно находиться 95 % от числа всех значений анализируемой переменой. Например, по результатам исследования мы выяснили, что средний доход респондентов равен 3 275 руб., значение стандартной ошибки среднего составило 132 руб. Следовательно, можно говорить, что не менее 95 % всех значений дохода, указанных респондентами, должно лежать в интервале 3 275 ± 2 * 132, т. е. от 3 011 до 3 539 руб.

Важную роль в получении обобщающих характеристик распределения играют перцентили[11] , которые можно рассматривать как показатели, разбивающие данные на определенные части.

Квартильное разбиение делит всех респондентов на 4 части. Так, 1-й квартиль — это значение переменной, меньше которой ответили 25 % респондентов, 2-й квартиль — это медиана, 3 й квартиль — точка, меньше которой ответили 75 % респондентов.

Квартильное отклонение — это разница между 1-м и 3-м квартилями.

---

[11] Перцентиль — величина, делящая совокупность на 100 равных групп и показывающая точку, ниже которой находится определенный процент наблюдений.

Квартильное отклонение является наиболее распространенным показателем, характеризующим разброс значений порядковой переменной.

Кроме того, можно производить разбиение совокупности значений на любое количество равных частей. 5 частей — квинтельное разбиение, 10 частей — децильное разбиение.

Применительно к ним можно использовать и такие меры разброса, как квинтельное отношение или децильное отношение.

Децильное отношение — это отношение границы 10-го дециля к границе 1-го дециля. Данный показатель демонстрирует, насколько больше получают 10 % высокооплачиваемых респондентов в сравнении с 10 % наименее оплачиваемых. Данное отношение в нашем примере составляет 4, что показывает степень неоднородности доходов.

Контрольные вопросы к теме

1. Что такое мода и каковы ограничения в ее применении?
2. Что такое медиана и каковы ее свойства?
3. Для каких данных может быть рассчитана средняя арифметическая простая?
4. Какие меры центральной тенденции можно рассчитать для данных высокого и низкого типа? Обоснуйте свой ответ.
5. Какие меры изменчивости можно рассчитать для данных высокого и низкого типа? Обоснуйте свой ответ.
6. В чем заключается основная функция вычисления дисперсии?
7. Для каких случаев рассчитывается простое среднее арифметическое, а для каких взвешенное?
8. Что характеризует среднеквадратичное отклонение?

Практические задания

Выявление центральных тенденций распределения. Оценка разброса данных.

Цель задания: освоение расчета моды, медианы, среднего арифметического, дисперсии и стандартного отклонения системы упорядоченных событий на ПК. Оценка меры отклонения распределения от нормального на ПК.

Аппаратура: персональный компьютер с лицензионным программным обеспечением.

Математическое обеспечение: операционная система WINDOWS и EXCEL 7.0.

Теоретическое обеспечение: система упорядоченных событий; ранжирование; меры оценки центральной тенденции; оценка разброса данных (дисперсия, стандартное отклонение)

Этапы обработки данных:

1. Занести данные в таблицу Excel (две выборки).
2. Упорядочить данные (по убыванию) в каждой выборке.
3. Рассчитать моду, медиану и среднее.
4. Выполнить сравнительный анализ полученных результатов.
5. Посчитать дисперсию, стандартное отклонение.
6. Сделать интерпретацию результатов.

Задачи

Вариант 1

Индивидуальные значения уровня интеллекта в выборках студентов физического и гуманитарного факультетов распределились следующим образом:

Студенты-физики	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	132	134	124	132	135	132	131	132	121	127	136	129	136	136
Студенты-гуманит.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
	126	127	132	120	119	126	120	123	120	116	123	115	122	125

Дать сравнительную характеристику по уровню вербального интеллекта в двух студенческих группах.

Вариант 2

При определении степени выраженности / диагностического коэффициента стереотипа в двух группах, *основной* и *контрольной*, баллы распределились следующим образом:

основная группа — 19, 16, 17, 12, 15,16, 17,17, 21, 23, 18, 13, 12, 13, 19, 20, 21;

контрольная группа — 27, 9, 12, 13, 26, 23, 14, 15, 22, 21, 16, 16, 18, 17, 10, 12, 17.

Дать сравнительную характеристику степени выраженности указанного коэффициента в данных группах.

Вариант 3

Была исследована группа детей с заболеванием крови до лечения препаратами и после лечения. В таблицу занесены показатели ***НВ*** крови по результатам медицинского обследования. Сделать сравнительный анализ результативности лечения данным препаратом, используя методы описательной статистики.

	№ респондента													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
До лечения	112	60	84	60	60	40	76	60	84	40	112	46	64	70
После лечения	82	78	110	130	130	104	108	129	110	88	105	73	85	80

Список литературы

1. Горшков, М. К. Прикладная социология: учеб. пособие для вузов / М. К. Горшков, Ф. Э. Шереги. — М., 2003.

2. Епархина, О. В. Математические методы обработки и анализа данных в социологии / О. В. Епархина. — Ярославль : ЯрГУ, 2007.

3. Крыштановский, А. О. Анализ социологических данных с помощью пакета SPSS / А. О. Крыштановский. — М. : ГУ ВШЭ, 2006.

4. Орлов, А. И. Прикладная статистика / А. И. Орлов. — М. : Экзамен, 2004.

5. Сидоренко, Е. В. Методы математической обработки в психологии / Е. В. Сидоренко. — СПб. : Речь, 2002.

6. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

Тема 4. **Первичное описание исходных данных**

Статистическая совокупность. Упорядочение массива. Кросстабуляция. Вариационный ряд. Графики. Полигон. Сглаженная кривая. Диаграммы и их виды. Алгоритм построения диаграмм.

Методические рекомендации для студентов

Прикладная статистика оперирует определенными категориями — понятиями, отражающими существенные, всеобщие свойства и основные отношения явлений действительности. Объект конкретного статистического исследования называют статистической совокупностью. *Статистическая совокупность* — это множество единиц (объектов, явлений), объединённых единой закономерностью и варьирующих в пределах общего качества.

Специфическим свойством статистической совокупности является массовость единиц, поскольку явление характеризуется массовым процессом и всем многообразием определяющих его причин и форм.

Общие принципы анализа социологической информации можно свести к следующим: упорядочение, уплотнение, компактное описание собранной информации, которые реализуются в ходе аналитических процедур.

Собранная в ходе полевого этапа первичная социологическая информация не структурирована, а потому не поддается непосредственному изучению. Упорядочение информации осуществляется с помощью статистической группировки данных и типологизации информации.

Метод статистической группировки заключается в том, что обследуемая совокупность расчленяется на однородные группы (отдельные единицы которых обладают общим для всех признаком).

При ***группировке по количественным признакам*** (возраст, стаж работы, размер дохода) весь диапазон изменения переменной разбивают на определенные интервалы с последующим подсчетом числа единиц, входящих в каждый из них.

При ***группировке по качественным признакам*** каждая из единиц анализа относится к одной из выделенных градаций с тем, чтобы суммарное число единиц анализа, отнесенных ко всем градациям, было равно общей численности изучаемой совокупности.

Метод ***типологизации информации*** представляет собой обобщение признаков социальных явлений на основе идеальной теоретической модели и по теоретически обоснованным критериям. В качестве примера можно привести исследование политической ориентации жителей Ярославской области, в ходе которого выделяются такие типы политической ориентации, как демократы, либералы, коммунисты, националисты и т. п.

Существуют две основные формы группировки — *статистические таблицы* и *статистические ряды*. Статистические таблицы представляют собой наиболее распространенную форму группировки, в то время как так называемые *статистические ряды* относят к числу особых форм группировки.

Статистическим называют ряд числовых значений признака, расположенных в определенном порядке. В числе статистических рядов видное место занимают вариационные ряды.

Ряд чисел, получаемый в результате группировки, называется рядом распределения. Ряд распределения — простейшая разновидность структурной группировки по одному признаку, отображенная в групповой таблице с двумя графами, в которых содержатся варианты и частоты признака.

Ряды распределений, отражающие результат группировки респондентов по качественным признакам, называют ***атрибутивными*** (например, деление населения по полу, национальности, семейному положению), а по количественным — ***вариационными***. В соответствии с характером количественных признаков вариационные ряды делятся на ***дискретные***[12] и ***непрерывные***

[12] Дискретный (от лат. discretus — прерывистый) — дробный, состоящий из отдельных частей

(которые, как правило, носят интервальный характер, в связи с чем их также называют ***интервальными***).

Примером дискретного вариационного ряда может служить распределение российских семей по числу имеющихся детей.

Интервальные вариационные ряды объединяют варианты либо непрерывных признаков, либо изменяющихся в широких пределах дискретных признаков. Интервальным является вариационный ряд распределения населения России по величине среднедушевых денежных доходов.

Кросс-табуляция (Cross-tabulation) — статистический метод, при котором одновременно характеризуются значения двух или более переменных. Кросс-табуляция заключается в создании таблиц сопряженности признаков, отражающих совместное распределение двух или более переменных с ограниченным количеством категорий или определенными значениями.

Таблица сопряженности — двумерное распределение единиц совокупности по признакам *x* и *y*. Относится к наиболее часто используемым инструментам изучения взаимосвязи двух переменных[13]. Очень важно помнить, что при анализе таблиц сопряженности исследователь ориентирован на поиск присутствия (либо отсутствия) только определенных статистических зависимостей.

Графиками в статистике называют условные изображения числовых величин и их соотношений в виде различных геометрических образов — точек, линий, плоских фигур и т. п. Графический способ отображения социологических данных широко применяется в целях наглядности. Существуют три основных метода графического представления данных — гистограмма (столбиковая диаграмма), полигон частот и сглаженная кривая (огива).

Полигон используется преимущественно для графического отображения непрерывных рядов распределения, а гистограмма — дискретных. Графики строятся в прямоугольной системе координат, где на оси у отмечается общая численность, или доля, респондентов (в %) по группам, на оси х — значения, или порядок, признака.

Гистограмма представляет последовательность столбцов, каждый из которых опирается на один интервал группирования

[13] См. подробнее: Крыштановский А. О. Анализ социологических данных с помощью пакета SPSS. С. 40–47.

данных, а высота столбца соответствует количеству элементов выборки, попавших в этот интервал группирования. Для построения гистограммы по горизонтальной оси откладываются границы интервалов группирования данных, а по вертикальной оси — частоты попадания наблюдений в интервалах (рис. 1).

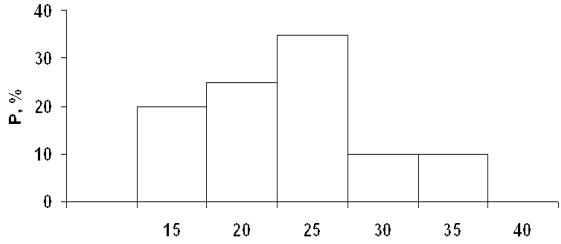

Рис. 1. Гистограмма плотности распределения непрерывного признака

Полигон частот — построение полигона частот во многом напоминает построение гистограммы, только в этом случае по горизонтальной оси откладываются значения середин интервалов группирования данных (по вертикальной то же самое). После этого на координатной плоскости наносятся точки. Первая координата соответствует середине интервала группирования и вторая — частоте. Для окончательного построения полигона частот точки соединяются отрезками прямых (рис. 2)[14].

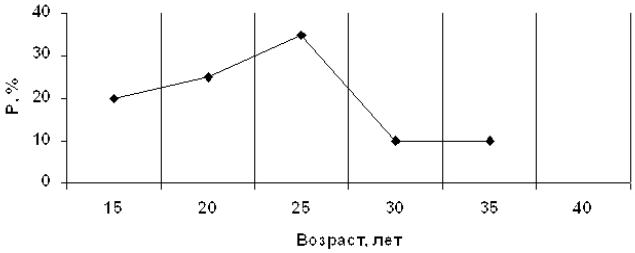

Рис. 2. Полигон плотности распределения непрерывного признака

[14] См. подробнее: Толстова Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками. С. 95–99.

Сглаженная кривая, или огива. Основное отличие сглаженной кривой в том, что она проводится по точкам таким образом, чтобы график не имел острых углов или зубцов. Для ее построения по горизонтальной оси всегда откладываются значения от 0 до 100 (они соответствуют процентам). По вертикальной оси откладываются границы интервалов группирования данных. После этого на координатной плоскости наносятся точки, вторая координата которой соответствует границе интервала, а первая координата накопленной частоте попадания, выраженной в процентах. Для окончательного построения нанесенные точки соединяются гладкой кривой (рис. 3).

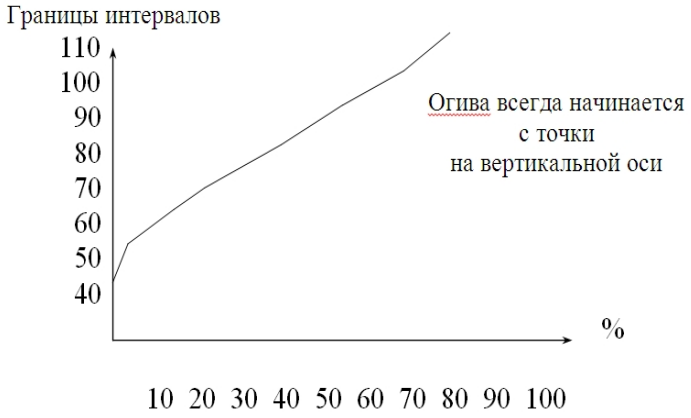

Рис. 3. Сглаженная кривая

Сглаженная кривая представляет собой неубывающую функцию. С помощью сглаженной кривой можно находить приближенно процентили. Кроме того, огива — неубывающая функция. С помощью сглаженной кривой можно судить о наличии малых и больших значений исследуемого показателя.

Современные компьютерные программы позволяют получить более разнообразные графические отображения социологической информации.

Контрольные вопросы к теме

1. Как можно упорядочить массив данных?

2. Что собой представляет статистическая группировка данных? Ее виды.

3. Для чего используется кросс-табуляция?

4. Что такое таблица сопряженности и как она анализируется?

5. Перечислите виды рядов распределения. Понятие вариационного ряда.

6. Чем отличается отображение данных в таблицах от их отображения на графиках и диаграммах?

7. Чем полигон отличается от иных видов графиков? На данных какого типа его можно построить?

8. Что показывает диаграмма рассеивания?

9. В чем смысл ящичковых диаграмм?

Список литературы

1. Аптон, Г. Анализ таблиц сопряженности / Г. Аптон. — М. : Финансы и статистика, 1982. (Upton G. J. G. The analysis of cross-tabulated data. N.-Y.: J. Wiley&Sons, 1978.)

2. Горшков, М. К. Прикладная социология: учеб. пособие для вузов / М. К. Горшков, Ф. Э. Шереги. — М., 2003.

3. Епархина, О. В. Математические методы обработки и анализа данных в социологии / О. В. Епархина. — Ярославль : ЯрГУ, 2007.

4. Илышев, А. М. Общая теория статистики: учеб. пособие / А. М. Илышев. — М. : КНОРУС, 2013.

5. Наследов, А. Компьютерный анализ данных в психологии и социальных науках / А. Наследов. — СПб. : Питер, 2005.

6. Сечко, В. В. Математические методы обработки психологических данных / В. В. Сечко. — Минск, 2002.

7. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

8. Швецова, С. В. Прикладная статистика для психологов: учеб. пособие для вузов / С. В. Швецова. — Ярославль : ЯрГУ, 2003.

Тема 5. **Распределение социологических данных**

Виды распределений: нормальное распределение, распределение Стьюдента, биномиальное распределение, распределение Пуассона. Проверка распределения на нормальность.

Методические рекомендации для студентов

Распределением признака называется закономерность встречаемости разных его значений. В эмпирических исследованиях чаще всего ссылаются на нормальное распределение.

Нормальное распределение характеризуется тем, что крайние значения признака в нем встречаются достаточно редко, а значения, близкие к средней величине, — достаточно часто. Нормальным такое распределение называется потому, что оно очень часто встречалось в естественно-научных исследованиях и казалось «нормой» всякого массового случайного проявления признаков. Это распределение следует закону, открытому тремя учеными в разное время: А. Муавром в 1733 г. в Англии, К. Ф. Гауссом в 1809 г. в Германии и П.-С. Лапласом в 1812 г. во Франции. График нормального распределения представляет собой так называемую колоколообразную кривую (рис. 4).

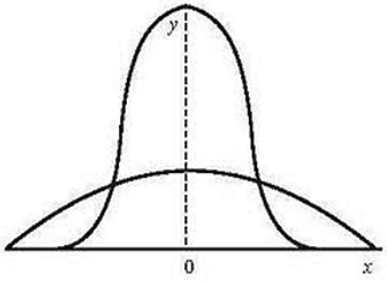

Рис. 4. Кривая нормального распределения

Нормальное распределение асимптотически приближается к оси X (то есть может принимать сколь угодно малые значения по ординате при стремлении икс-значений к плюс или минус бес-

конечности), значения моды, медианы и среднего арифметического равны между собой. Свойством нормальных распределений является наличие определенного количества случайной величины (случаев, испытуемых), приходящегося на интервалы между значениями σ, обычно это количество измеряют в процентах от общего числа случаев, испытуемых. Принято считать, что нормальное распределение характеризует такие случайные величины, на которые воздействует большое количество разнообразных факторов, причем сила воздействия одного отдельно взятого фактора значительно меньше суммы воздействий остальных факторов. В результате получается, что чаще наблюдаются некоторые средние значения измеряемого параметра, реже крайние, и чем сильнее отличается какое-то значение от среднего, тем реже оно встречается. Многие биологические параметры распределены подобным образом (рост, вес и т. п.).

Параметры распределения — это его числовые характеристики, указывающие, где «в среднем» располагаются значения признака, насколько эти значения изменчивы и наблюдается ли преимущественное появление определенных значений признака. Наиболее важными параметрами являются математическое ожидание, дисперсия, стандартное отклонение (σ — сигма), показатели асимметрии и эксцесса.

Асимметрия — это мера отклонения распределения признака от симметричного распределения. Если все значения признака находятся в интервале от — 3σ до + 3σ, то считается, что закон распределения признака нормальный. При *нормальном законе распределения* A = 0. Для симметричных распределений A = 0 (рис. 5.1).

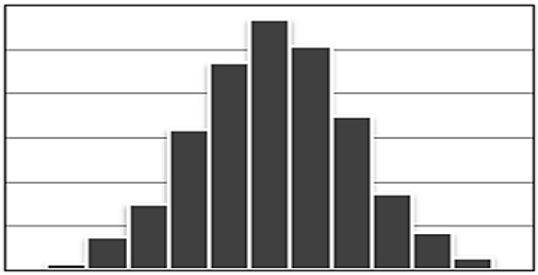

Рис. 5.1. Симметричное распределение

В тех случаях, когда какие-нибудь причины благоприятствуют частому появлению значений, которые выше или, наоборот, ниже среднего, образуются асимметричные распределения. Пока-затель асимметрии (А) вычисляется по формуле:

$$A = \frac{\sum \left(x_i - \bar{x}\right)^3}{n \cdot \sigma^3}$$

Для симметричных распределений A=0.

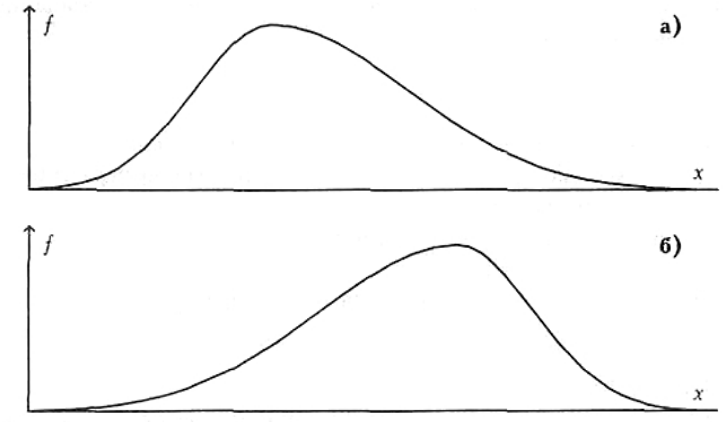

Рис. 5.2. Асимметрия распределений:

а) левая, положительная; б) правая, отрицательная

При *положительной (левосторонней) асимметрии* в распределении чаще встречаются более низкие значения признака (рис. 5.2а).

При *отрицательной (правосторонней) асимметрии* в распределении чаще всего встречаются более высокие значения признака (рис. 5.2б).

Таким образом, если величина А меньше 0, то распределение растянуто влево, если больше 0, то — вправо.

Эксцесс — коэффициент вариации — показывает, является ли распределение пологим (при большом значении) или крутым. Если V значительно отличается от 0, то гипотеза о нормальном распределении отвергается. При нормальном распределении V = 0.

Показатель эксцесса (E) определяется по формуле:

$$E = \frac{\sum_{i=1}^{n}(a_i - a)^4}{n\sigma^4} - 3;$$

E — показатель эксцесса, σ — среднеквадратическое отклонение, a — среднее арифметическое, n — число измерений параметра, ai — каждое наблюдаемое значение признака.

В тех случаях, когда появляются преимущественно средние значения или близкие к средним, образуется распределение с положительным эксцессом (рис. 5.3а).

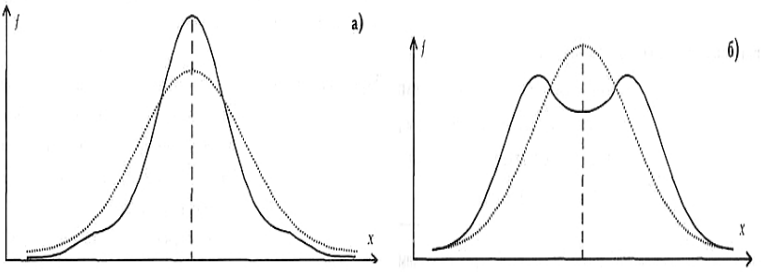

Рис. 5.3. Эксцесс: а) положительный; б) отрицательный

Если же в распределении преобладают крайние значения, причем одновременно и более низкие, и более высокие, то такое распределение характеризуется отрицательным эксцессом, в центре распределения может образоваться впадина, превращающая его в двухвершинное (рис. 5.3б).

Рассмотренные параметры распределения можно определить только по отношению к данным, представленным, по крайней мере, в интервальной шкале.

Критерий нормальности Колмогорова — Смирнова обладает достаточной чувствительностью даже при малом числе значений, поэтому его можно применять также для проверки соответствия любому распределению (например, равномерному). Однако следует иметь в виду, что функция распределения, установленная гипотезой, должна быть непрерывной.

С помощью нормального распределения определяются три распределения, которые в настоящее время часто используются при статистической обработке данных.

Распределение χ^2 (хи-квадрат) — распределение случайной величины: $X = X_1^2 + X_2^2 + ... + X_n^2$ где случайные величины X_1, X_2,…, X_n независимы и имеют одно и то же распределение $N(0,1)$. При этом число слагаемых, т. е. n, называется «числом степеней свободы» распределения хи-квадрат.

Распределение t Стьюдента — это распределение случайной величины

$$T = \frac{U\sqrt{n}}{\sqrt{X}},$$

где случайные величины U и X независимы, U имеет распределение стандартное нормальное распределение $N(0,1)$, а X — распределение *хи*-квадрат с n степенями свободы. При этом n называется «числом степеней свободы» распределения Стьюдента. Это распределение было введено в 1908 г. английским статистиком В. Госсетом. Вероятностно-статистические методы использовались для принятия экономических и технических решений на фабрике, поэтому ее руководство запрещало В. Госсету публиковать научные статьи под своим именем. Таким способом охранялась коммерческая тайна, «ноу-хау» в виде вероятностно-статистических методов, разработанных В. Госсетом. Однако он имел возможность публиковаться под псевдонимом «Стьюдент». Распределение Стьюдента используется в статистике для точечного оценивания, построения доверительных интервалов и тестирования гипотез, касающихся неизвестного среднего статистической выборки из нормального распределения.

Распределение Фишера — это распределение случайной величины

$$F = \frac{\frac{1}{k_1} X_1}{\frac{1}{k_2} X_2},$$

где случайные величины X_1 и X_2 независимы и имеют распределения *хи*-квадрат с числом степеней свободы k_1 и k_2 соответст-венно. При этом пара (k_1, k_2) — пара «чисел степеней свободы» распределения Фишера, а именно, k_1 — число степеней свободы числителя, а k_2

— число степеней свободы знаменателя. Распределение случайной величины F названо в честь великого английского статистика Р. Фишера (1890–1962), активно использовавшего его в своих работах[15].

В вероятностно-статистических методах наиболее часто используют три семейства дискретных распределений — биномиальных, гипергеометрических и Пуассона. *Биномиальное распределение* имеет место при независимых испытаниях, в каждом из которых с вероятностью p появляется событие *A*. Если общее число испытаний n задано, то число испытаний *Y*, в которых появилось событие *A*, имеет биномиальное распределение. Для биномиального распределения вероятность принятия случайной величиной *Y* значения *y* определяется формулой

$$P(Y = y \mid p,n) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0,1,2,\ldots,n, \quad \text{где}$$

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} = C_n^y -$$

— число сочетаний из n элементов по y, известное из комбинаторики. Для всех *y*, кроме 0, 1, 2, …, *n*, имеем *P(Y=y)=0*. Биномиальное распределение при фиксированном объеме выборки *n* задается параметром *p*, т. е. биномиальные распределения образуют однопараметрическое семейство. Они применяются при анализе данных выборочных исследований, в частности, при испытаниях совокупностей индивидуумов в социологии.

Распределение Пуассона. Случайная величина Y имеет распределение Пуассона, если

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0,1,2,\ldots,$$

где λ — параметр распределения Пуассона, и *P(Y=y)=0* для всех прочих y (при y=0 обозначено 0! =1). Для распределения Пуассона

$$M(Y) = \lambda,\ D(Y) = \lambda.$$

Это распределение названо в честь французского математика С. Д. Пуассона (1781–1840), впервые получившего его в 1837 г.

[15] См. подробнее: Орлов А. И. Прикладная статистика.

Распределение Пуассона является предельным случаем биномиального распределения, когда вероятность p осуществления со-бытия мала, но число испытаний *n* велико, причем $np = \lambda$. Точнее, справедливо предельное соотношение

$$\lim_{n\to\infty, np\to\lambda} P(Y = y \mid p, n) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0,1,2,\ldots$$

Поэтому распределение Пуассона (в старой терминологии «закон распределения») часто называют также «законом редких событий».

Контрольные вопросы к теме

1. В чем специфика мер средней тенденции в нормальном распределении?

2. Чем отличаются нормальное распределение и распределение Стьюдента?

3. Что такое правосторонняя и левосторонняя асимметрия распределения?

4. Как определить тип асимметрии по обобщающим статистическим показателям (моде, медиане, средней арифметической)?

5. Что характеризует эксцесс распределения и как с его помощью можно охарактеризовать форму распределения?

6. Для каких событий целесообразно использовать биномиальное распределение? Приведите примеры.

7. В каких случаях используют распределение Пуассона?

Список литературы

1. Бутенко, И. А. Прикладная социология: наука и искусство / И. А. Бутенко. — М. : Наука, 1999.

2. Горшков, М. К. Прикладная социология: учеб. пособие для вузов / М. К. Горшков, Ф. Э. Шереги. — М., 2003.

3. Епархина, О. В. Математические методы обработки и анализа данных в социологии / О. В. Епархина. — Ярославль : ЯрГУ, 2007.

4. Илышев, А. М. Общая теория статистики: учеб. пособие / А. М. Илышев. — М. : КНОРУС, 2013.

5. Орлов, А. И. Прикладная статистика / А. И. Орлов. — М. : Экзамен, 2004.

6. Рабочая книга социолога / под ред. Г. В. Осипова. — М. : URSS, 2006.

7. Сикевич, З. В. Социологическое исследование / З. В. Сикевич. — СПб. : Питер, 2005.

8. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

9. Хилл, Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хилл. — СПб. : Питер, 2005.

10. Ядов, В. А. Стратегия социологического исследования / В. А. Ядов. — М. : Омега-Л, 2005.

Тема 6. **Основные понятия математической статистики**

Статистическая значимость. Статистические гипотезы. Принятие и отвержение гипотез. Зависимые и независимые выборки. Степени свободы. Статистические критерии.

Методические рекомендации для студентов

В математической статистике выделяют два фундаментальных понятия: ***генеральная совокупность*** и ***выборка***.

Совокупностью называется практически счетное множество некоторых объектов или элементов, интересующих исследователя.

Свойством совокупности называется реальное или воображаемое качество, присущее некоторым или всем ее элементам. Свойство может быть случайным или неслучайным.

Параметром совокупности называется свойство, которое можно квантифицировать в виде константы или переменной ве-личины.

Простая совокупность характеризуется:

• отдельным свойством (например, все студенты России);

• отдельным параметром в виде константы или переменной (все студенты женского пола);

• системой непересекающихся (несовместных) свойств (все учителя и ученики школ г. Ярославля).

Сложная совокупность характеризуется:

• системой, хотя бы частично пересекающихся свойств (студенты психологического и математического факультетов МГУ, окончивших школу с золотой медалью);

• системой параметров независимых и зависимых в совокупности; при комплексном исследовании личности.

Гомогенной, или однородной, называется совокупность, все характеристики которой присущи каждому ее элементу.

Гетерогенной, или неоднородной, называется совокупность, характеристики которой сосредоточены в отдельных подмножествах элементов.

Важным параметром является ***объем*** совокупности — количество образующих ее элементов. Величина объема зависит от того, как определена сама совокупность и какие вопросы конкретно интересуют исследователя.

В случаях, когда объект социологического исследования насчитывает 500 человек и более, единственно правильным следует признать применение выборочного метода, теория которого заимствована из математической статистики. Этот метод довольно широко практиковался в России еще во времена Петра Великого для оценки ожидаемой урожайности злаковых; использовался он и земскими статистиками, проводившими перепись населения. Процесс выборки основан 1) на взаимосвязи и взаимообусловленности качественных характеристик и признаков социальных объектов; 2) на правомерности выводов о целом на основании изучения его части при условии, что по своей структуре эта часть представляет собой микромодель целого. ***Выборка*** — это по сути микромодель, которая является одним из наиболее экономных средств для проверки предположений или гипотез о свойствах предметов, явлений.

При рассмотрении основных методов выборки, применяемых в исследовании социальных проблем, используются такие ключевые понятия, как объект исследования, генеральная совокупность, выборочная совокупность, единица отбора, единица наблюдения.

Объектом социологического исследования выступает носитель той или иной социальной проблемы. *Генеральной совокупностью* называется объект исследования, который локализован территориально, во времени, по демографическим или социальным признакам и на который распространяются выводы исследования. Локализация объекта территориально чаще всего происходит по административному делению — регион, область, район, поселение. Данный критерий лежит в основе локализации объекта в электоральных исследованиях. Локализация по демографическим признакам осуществляется в исследованиях отдельных групп, например молодежи, женщин, пенсионеров, этносов; по социальным признакам — в целевых исследованиях профессиональных групп, конфессий, политических движений и др. Локализация объекта во времени осуществляется в длительных социальных экспериментах, контент-анализе средств массовой коммуникации или изучении их аудитории, при проведении повторных исследований.

Выборочную совокупность составляет определенное число элементов генеральной совокупности, отобранных по строго заданному правилу. Необходимо, чтобы структура выборочной совокупности максимально совпадала со структурой генеральной совокупности по основным изучаемым качественным характеристикам и контрольным признакам. Чтобы добиться этого, нужно строго соблюдать правила выборки.

Единицы наблюдения есть элементы выборочной совокупности (респонденты), подлежащие изучению (например, опросу). Такими единицами могут выступать отдельные индивиды и целые группы (семья, посетители кинотеатра и т. д.).

Правила формирования выборочной совокупности таковы, что в процессе отбора основными элементами не всегда выступают единицы наблюдения, т. е. непосредственно опрашиваемые. Так, вначале могут быть отобраны те или иные административные регионы (области, края, республики), потом в них — города, в последних — семьи, в которых опрашиваются либо все взрослые члены, либо один член семьи, отобранный по заданному принципу (глава, «распорядитель» бюджета, старший ребенок и т. д.). Элементы (регионы, поселения, семьи, респонденты),

отбираемые на каждом этапе выборки по особому плану, называются *единицами отбора*.

Выборки классифицируются по репрезентативности, объему, способу отбора и схеме испытаний.

Репрезентативная (представительная) выборка, адекватно отображающая генеральную совокупность в качественном и количественном отношениях. Выборка должна адекватно отображать генеральную совокупность, иначе результаты не совпадут с целями исследования.

По способу отбора

Случайная — если элементы отбираются случайным образом. Так как большинство методов математической статистики основывается на понятии случайной выборки, то, естественно, выборка должна быть случайной.

Неслучайная выборка:

• *механический отбор*, когда вся совокупность делится на столько частей, сколько единиц планируется в выборке, и затем из каждой части отбирается один элемент;

• *типический отбор* — совокупность делится на гомогенные части, и из каждой осуществляется случайная выборка;

• *серийный отбор* — совокупность делят на большое число разновеликих серий, затем делают выборку одной какой-либо серии;

• *комбинированный отбор* — сочетаются рассматриваемые виды отбора, на разных этапах.

По схеме испытаний — выборки могут быть независимые и зависимые.

Обычна ситуация исследования, когда интересующее исследователя свойство изучается на двух или более выборках с целью их дальнейшего сравнения. Эти выборки могут находиться в различных соотношениях — в зависимости от процедуры их организации. Независимые выборки характеризуются тем, что вероятность отбора любого испытуемого одной выборки не зависит от отбора любого из испытуемых другой выборки. Напротив, зависимые выборки характеризуются тем, что каждому испытуемому одной выборки поставлен в соответствие по определенному критерию испытуемый из другой выборки. В общем случае за-

висимые выборки предполагают попарный подбор испытуемых в сравниваемые выборки, а независимые выборки — независимый отбор испытуемых.

Наиболее часто зависимые выборки возникают, когда измерение проводится для нескольких моментов времени. Зависимые выборки образуют значения параметров изучаемого процесса, соответствующие различным моментам времени. Если закономерное и однозначное соответствие между выборками невозможно, эти выборки являются *независимыми*.

Примеры зависимых выборок: пары близнецов, два измерения какого-либо признака до и после экспериментального воздействия, супружеские пары (мужья и жёны) и т. п. Примерами независимых выборок могут быть мужчины и женщины, психологи и математики.

По объему выборки делят на *малые* и *большие*. К малым относят выборки, в которых число элементов $n \leq 30$. Понятие большой выборки не определено, но большой считается выборка, в которой число элементов > 200 и средняя выборка удовлетворяет условию $30 \leq n \leq 200$. Это деление условно.

Малые выборки используются при статистическом контроле известных свойств уже изученных совокупностей. Большие выборки используются для установки неизвестных свойств и параметров совокупности.

Гипотеза есть некое научное предположение, которое необходимо проверить и далее принять или отвергнуть.

Гипотезы в статистике различают простые и сложные:

• *простая* гипотеза полностью задает распределение вероятностей;

• *сложная* гипотеза указывает не одно распределение, а некоторое множество распределений. Обычно это множество распределений, обладающих определенным свойством.

Статистической гипотезой называют предположение о свойстве генеральной совокупности, которое можно проверить, опираясь на данные выборки. Статистические гипотезы подразделяются на нулевые и альтернативные, направленные и ненаправленные.

Нулевая гипотеза — это гипотеза об отсутствии различий. Она обозначается как $\mathbf{H_0}$ и называется нулевой потому, что содержит число 0: $X_1 - X_2 = 0$, где X_1, X_2 — сопоставляемые значения признаков. Нулевая гипотеза — это то, что мы хотим опровергнуть, если перед нами стоит задача доказать значимость различий.

Альтернативная гипотеза — это гипотеза о значимости различий. Она обозначается $\mathbf{H_1}$. Альтернативная гипотеза — это то, что мы хотим доказать, поэтому иногда ее называют *экспериментальной* гипотезой.

Таким образом, принятие нулевой гипотезы $\mathbf{H_0}$ свидетельствует об отсутствии различий, а гипотеза $\mathbf{H_1}$ — о наличии различий.

Проверка гипотез осуществляется с помощью критериев статистической оценки различий. **Статистический критерий** — это решающее правило, обеспечивающее надежное поведение, то есть принятие истинной и отклонение ложной гипотезы с **высокой** вероятностью[16]. Дело в том, что в большинстве случаев для того, чтобы мы признали различия значимыми, необходимо, чтобы эмпирическое значение критерия превышало критическое, в некоторых критериях придерживаются противоположного правила. Эти правила оговариваются в описании каждого критерия.

В некоторых случаях расчетная формула критерия включает в себя количество наблюдений в исследуемой выборке, обозначаемое как *n*. В этом случае эмпирическое значение критерия одновременно является тестом для проверки статистических гипотез. По специальной таблице определяется, какому уровню статистической значимости различий соответствует данная эмпирическая величина.

В большинстве случаев одно и то же эмпирическое значение критерия может оказаться значимым или незначимым в зависимости от количества наблюдений в выборке (***n***) или от так называемого количества *степеней свободы*, которое обозначается как ν.

Число степеней свободы равно числу классов вариационного ряда минус число условий, при которых он был сформирован. К числу таких условий относятся объем выборки, *средние и дисперсии*.

---

[16] См. подробнее: Сидоренко Е. В. Методы математической обработки в психологии.

Если мы расклассифицировали наблюдения по классам какой-либо номинативной шкалы и подсчитали количество наблюдений в каждой ячейке классификации, то получаем так называемый частотный вариационный ряд. Единственное условие, которое соблюдается при его формировании, — объем выборки ***n***.

Предположим у нас три класса: «Умеет работать на ПК — умеет выполнять лишь определенные операции — не умеет работать». Выборка состоит из 50 человек. Если в первом классе 20 человек, во втором классе — 20 человек, то в третьем должны оказаться 10 человек. В данном случае мы ограничены только одним условием — объемом выборки. Мы не свободны в определении количества испытуемых в третьем классе, «свобода» простирается только на первые два класса

$$\nu = c - 1 = 3 - 1 = 2.$$

Аналогичным образом, если бы у нас была классификация из 10 разрядов или классов, мы были бы свободны только в 9 и т. д.

Зная ***n*** и/или *число степеней свободы*, по специальным таблицам можно определить критические значения критерия и сопоставить с ними полученное эмпирическое значение.

Среди возможных статистических критериев выделяют: *односторонние и двусторонние, параметрические и непараметрические, более и менее мощные*.

Односторонние и двусторонние. Понятие *одностороннего* либо *двустороннего* критерия связано с формулировкой гипотез. Если «нулевая» гипотеза формулируется о равенстве ($\mathbf{X_1} = \mathbf{X_2}$), то для проверки используется двусторонний критерий. Если же «нулевая» гипотеза формулируется о неравенстве, то возможны варианты:

1) если $\mathbf{X_1} \neq \mathbf{X_2}$, то используется двусторонний критерий;

2) если $\mathbf{X_1} > \mathbf{X_2}$ или $\mathbf{X_1} < \mathbf{X_2}$, то односторонний критерий.

Параметрические критерии — это некоторые функции от параметров совокупности, они служат для проверки гипотез об этих параметрах или для их оценивания. *Параметрические критерии* включают в формулу расчета параметры распределения, т. е. средние и дисперсии.

Непараметрические критерии — это некоторые функции от функций распределения или непосредственно от вариационного ряда наблюдавшихся значений изучаемого случайного явления. Они служат только для проверки гипотез о функциях распределения или рядах наблюдавшихся значений.

Непараметрические критерии не включают в формулу расчета параметров распределения и основанные на оперировании частотами или рангами.

Параметрические и непараметрические критерии имеют свои преимущества и недостатки.

Параметрические критерии могут оказаться несколько более *мощными*, чем непараметрические, но только в том случае, *если признак измерен по интервальной шкале и нормально распределен*. Лишь с некоторой натяжкой мы можем считать данные, представленные в стандартизованных оценках, как интервальные. Кроме того, проверка распределения «на нормальность» требует достаточно сложных расчетов, результат которых заранее не известен. Может оказаться, что распределение признака отличается от нормального и исследователю так или иначе придется обратиться к непараметрическим критериям.

Непараметрические критерии лишены всех перечисленных ограничений и не требуют таких длительных и сложных расчетов. По сравнению с параметрическими критериями они ограничены лишь в одном — с их помощью невозможно оценить взаимодействие двух или более условий или факторов, влияющих на изменение признака.

Возможность экстраполяции полученных результатов на генеральную совокупность проверяется с помощью тестов на *статистическую значимость*.

Общая логика проведения такой проверки может быть представлена следующим образом.

1. *Выдвигаются две альтернативные гипотезы*:

$\mathbf{H_0}$: коэффициент корреляции статистически незначим (иначе говоря, в генеральной совокупности он может оказаться равным нулю);

$\mathbf{H_1}$: коэффициент корреляции статистически значим (в генеральной совокупности коэффициент корреляции отличается от нуля).

2. *Рассчитывается наблюдаемое значение t-критерия Стьюдента*[17] *(t набл)*.

Если число наблюдений невелико (как правило, меньше 50), то для расчета критерия применяется формула:

$$t_{набл} = \sqrt{\frac{r^2}{1-r^2}} \times (n-2)$$

где *r* — коэффициент корреляции; *n* — объем выборки.

В случае если выборка большого объема, то используется формула:

$$t_{набл} = \sqrt{\frac{r^2}{1-r^2}} \times n$$

3. *Наблюдаемое значение критерия сравнивается с его табличным значением (t набл)*, и выбор в пользу той или иной гипотезы осуществляется по следующему правилу:

- если $t_{набл} < t_{табл}$, принимается гипотеза $\mathbf{H_0}$, т. е. коэффициент корреляции статистически незначим и результаты анализа не могут распространяться на генеральную совокупность;
- если $t_{набл} > t_{табл}$ принимается гипотеза $\mathbf{H_1}$, т. е. коэффициент корреляции статистически значим.

Для нахождения табличного значения критерия следует воспользоваться *таблицами критических значений критерия*[18]. В таблице выбирают число, находящееся на пересечении двух параметров — уровня статистической значимости (α) и числа степеней свободы (*df*). При этом уровень статистической значимости (или вероятность ошибки) задает сам исследователь (обычно не более 5 %, т. е. $\alpha < 0{,}05$). А число степеней свободы для коэффициента парной корреляции определяется как $df = n - 2$.

Уровень значимости (*p — уровень значимости*) — это вероятность того, что мы сочли различия существенными, а они на

[17] t-критерий Стьюдента — общее название для класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента.

[18] См.: Таблица критических значений t-критерия Стьюдента // Ильышев А. М. Общая теория статистики: учеб. пособие. С. 417.

самом деле случайны. Когда мы указываем, что различия достоверны на 5 % уровне значимости, или при $p \leq 0{,}05$, то мы имеем в виду, что вероятность того, что они недостоверны, составляет 0,05. Если же мы указываем, что различия достоверны на 1 % уровне значимости, или при $p \leq 0{,}01$, то имеем в виду, что вероятность того, что они все-таки недостоверны равна 0,01. Иначе, уровень значимости — это вероятность отклонения нулевой гипотезы, в то время как она верна.

Статистическая значимость результата представляет собой меру уверенности в его истинности (в смысле репрезентативности выборки). Более точно, *p*-уровень — это показатель, обратно пропорциональный надежности результата. Более высокий *p*-уровень соответствует более низкому уровню доверия найденным в выборке результатам, например зависимостям между переменными. А именно, *p*-уровень представляет собой вероятность ошибки, связанной с обобщением наблюдаемого результата на всю популяцию. Например, *p*-уровень = 0,05 (т. е. 1/20) показывает, что имеется 5 % вероятности того, что найденная в выборке зависимость между переменными является лишь случайной особенностью данной выборки. Иными словами, если данная зависимость в популяции отсутствует, а вы многократно проводили бы подобные эксперименты, то примерно в одном из двадцати повторений эксперимента можно было бы ожидать такой же или более сильной зависимости между изучаемыми переменными. Во многих исследованиях *p*-уровень 0,05 рассматривается как приемлемая граница уровня ошибки.

Стоит заметить, что уровень значимости при прочих равных условиях выше (значение *p*-уровня меньше), если: 1) величина связи (различия) больше; 2) изменчивость признака (признаков) меньше; 3) объем выборки (выборок) больше.

Контрольные вопросы к теме

1. Какой уровень статистической значимости обычно считается приемлемым в социологии?

2. Приведите пример нуль-гипотезы.

3. Что такое зависимые выборки в социологии? Приведите пример.

4. Зависит ли степень свободы от размера таблицы? Почему?

5. Приведите примеры статистических критериев.

Список литературы

1. Бутенко, И. А. Прикладная социология: наука и искусство / И. А. Бутенко. — М. : Наука, 1999.

2. Гнеденко, Б. В. Курс теории вероятностей / Б. В. Гнеденко. — М. : Наука, 1965.

3. Горшков, М. К. Прикладная социология: учеб. пособие для вузов / М. К. Горшков, Ф. Э. Шереги. — М., 2003.

4. Епархина, О. В. Математические методы обработки и анализа данных в социологии / О. В. Епархина. — Ярославль : ЯрГУ, 2007.

5. Илышев, А. М. Общая теория статистики: учеб. пособие / А. М. Илышев. — М. : КНОРУС, 2013.

6. Кендалл, М. Дж. Статистические выводы и связи / М. Дж. Кендалл, А. Стьюарт. — М. : Наука, 1973.

7. Крамер, Д. Математическая обработка данных в социальных науках / Д. Крамер. — М. : Академия, 2007.

8. Наследов, А. Компьютерный анализ данных в психологии и социальных науках / А. Наследов. — СПб. : Питер, 2005.

9. Паниотто, В. И. Количественные методы в социологических исследованиях / В. И. Паниотто, В. С. Максименко. — Киев : Наукова думка, 1982.

10. Рабочая книга социолога / под ред. Г. В. Осипова. — М. : URSS, 2006.

11. Сидоренко, Е. В. Методы математической обработки в психологии / Е. В. Сидоренко. — СПб. : Речь, 2002.

12. Социология: словарь-справочник. Т. 4. Социологическое исследование: методы, математика и статистика. — М., 1991.

13. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между номинальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

14. Толстова, Ю. Н. Математика в социологии: элементарное введение в круг основных понятий (измерение, статистические закономерности, принципы анализа данных) / Ю. Н. Толстова. — М. : ИСАН СССР, 1990.

15. Хилл, Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хилл. — СПб. : Питер, 2005.

16. Ядов, В. А. Стратегия социологического исследования / В. А. Ядов. — М. : Омега-Л, 2005.

Тема 7. **Исследование взаимосвязи признаков**

Понятие корреляции. Виды корреляции. Классификация коэффициентов корреляции по силе, уровню значимости и их вычисление. Определение значимости корреляции.

Методические рекомендации для студентов

Анализ связей между признаками — главный вид задач, встречающийся практически в любом социологическом исследовании. Изучение связей между переменными интересует исследователя не само по себе, а как отражение соответствующих причинно-следственных отношений.

При изучении *корреляций* стараются установить, существует ли какая-то связь между двумя показателями в одной выборке (например, между ростом и весом детей или между уровнем IQ и школьной успеваемостью) либо между двумя различными выборками (например, при сравнении пар близнецов), и если эта связь существует, то сопровождается ли увеличение одного показателя возрастанием (положительная корреляция) или уменьшением (отрицательная корреляция) другого. Иными словами, корреляционный анализ помогает установить, можно ли предсказывать возможные значения одного показателя, зная величину другого.

Первоначальное значение термина «корреляции» — взаимная связь. Когда говорят о корреляции, используют термины «корреляционная связь» и «корреляционная зависимость».

Корреляционная связь — это согласованные изменения двух признаков или большего количества признаков (множественная корреляционная связь). Корреляционная связь отражает тот факт, что изменчивость одного признака находится в некотором соответствии с изменчивостью другого. «Стохастическая» связь имеется тогда, когда каждому из значений одной случайной величины соответствует специфическое (условное) распределение вероятностей значений другой величины и, наоборот, каждому из значений этой другой величины соответствует специфическое (условное) распределение вероятностей значений первой случайной величины».

Корреляционная зависимость — это изменения, которые вносят значения одного признака в вероятность появления разных значений другого признака. «Стохастическая» означает «вероятностная». Связи между случайными явлениями называют вероятностными, или стохастическими, связями. Этот термин подчеркивает их отличие от детерминированных или функциональных связей в физике или математике (связь площади треугольника с его высотой и основанием, связь длины окружности с ее радиусом и т. п.). В функциональных связях каждому значению первого признака всегда соответствует (в идеальных условиях) совершенно определенное значение другого признака. В корреляционных связях каждому значению одного признака может соответствовать определенное распределение значений другого признака, но не определенное его значение.

Оба термина — корреляционная связь и корреляционная зависимость — часто используются как синонимы. Между тем согласованные изменения признаков и отражающая это корреляционная связь между ними могут свидетельствовать не о взаимозависимости этих признаков, а о зависимости обоих этих признаков от какого-то третьего признака или сочетания признаков, не рассматриваемых в исследовании.

Зависимость подразумевает влияние, связь — любые согласованные изменения, которые могут объясняться сотнями причин.

Говорить в строгом смысле о зависимости мы можем только в тех случаях, когда сами оказываем какое-то контролируемое воздействие на испытуемых или так организуем исследование, что оказывается возможным точно определить интенсивность не зависящих от нас воздействий. Воздействия, которые мы можем качественно определить или даже измерить, могут рассматриваться как независимые переменные. Признаки, которые мы измеряем и которые, по предположению исследователя, могут изменяться под влиянием независимых переменных, считаются зависимыми переменными. Согласованные изменения независимой и зависимой переменной действительно могут рассматриваться как зависимость.

Если в исследование включены независимые переменные, которые мы можем по крайней мере учитывать, например возраст, то можно считать выявляемые между возрастом и психологическими признаками корреляционные связи корреляционными зависимостями. В большинстве же случаев нам трудно определить, что в рассматриваемой паре признаков является независимой, а что — зависимой переменной.

Корреляционные связи различаются по ***форме***, ***направлению*** и ***степени*** (силе).

По форме корреляционная связь может быть прямолинейной или криволинейной. ***Прямолинейной*** может быть, например, связь между количеством тренировок на тренажере и количеством правильно решаемых задач в контрольной сессии. ***Криволинейной*** может быть, например, связь между уровнем мотивации и эффективностью выполнения задачи.

По направлению корреляционная связь может быть положительной («прямой») и отрицательной («обратной»). При ***положительной*** прямолинейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака — низкие значения другого. При ***отрицательной*** корреляции соотношения обратные. При положительной корреляции коэффициент корреляции имеет положительный знак, например r = + 0,207, при отрицательной корреляции — отрицательный знак, напри-

мер r = – 0,207. Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции.

Наглядное представление о характере вероятностной связи дает диаграмма рассеивания — график, оси которого соответствуют значениям двух переменных, а каждый испытуемый представляет собой точку (см. рис. 6).

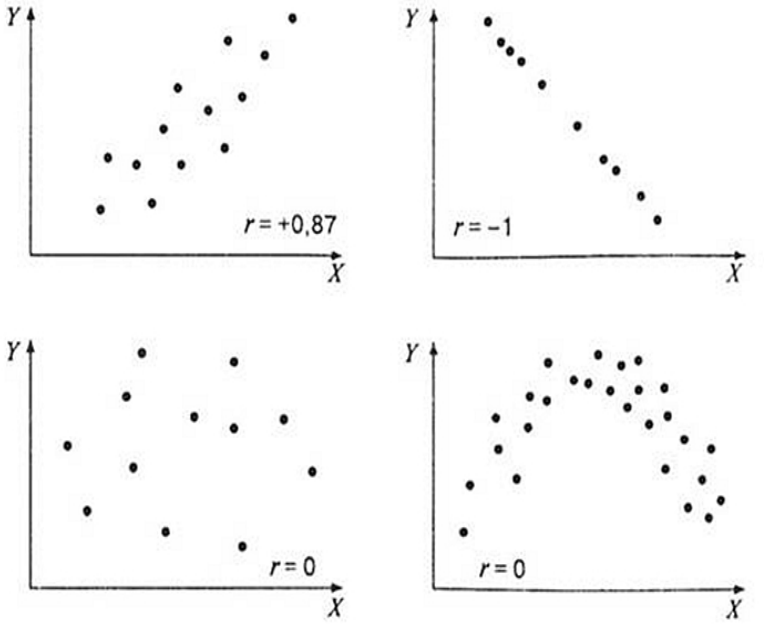

Рис. 5.3. Примеры диаграмм рассеивания и соответствующих коэффициентов корреляции

Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции.

Коэффициент корреляции — это величина, которая может варьировать в пределах от +1 до –1. В случае полной положительной корреляции этот коэффициент равен плюс 1, а при полной отрицательной — минус 1. В случае если коэффициент корреляции равен 0, обе переменные полностью независимы друг от друга.

Используются две системы классификации корреляционных связей по их силе: общая и частная.

Общая классификация корреляционных связей:

1) сильная, или тесная, при коэффициенте корреляции $r > 0,70$;

2) средняя при $0,50 < r < 0,69$;

3) умеренная при $0,30 < r < 0,49$;

4) слабая при $0,20 < r < 0,29$;

5) очень слабая при $r < 0,19$.

Частная классификация корреляционных связей:

1) высокая значимая корреляция при r, соответствующем уровню статистической значимости $\rho \leq 0,01$;

2) значимая корреляция при r, соответствующем уровню статистической значимости $\rho \leq 0,05$;

3) тенденция достоверной связи при r, соответствующем уровню статистической значимости $\rho \leq 0,10$;

4) незначимая корреляция при r, не достигающем уровня статистической значимости.

Необходимо уточнить, что эти классификации не совпадают. Первая ориентирована только на величину коэффициента корреляции, а вторая определяет, какого уровня значимости достигает данная величина коэффициента корреляции при данном объеме выборки. Чем больше объем выборки, тем меньшей величины коэффициента корреляции оказывается достаточно, чтобы корреляция была признана достоверной. В результате при малом объеме выборки может оказаться так, что сильная корреляция окажется недостоверной. В то же время при больших объемах выборки даже слабая корреляция может оказаться достоверной.

Как правило, принято ориентироваться на вторую классификацию, поскольку она учитывает объем выборки. Вместе с тем необходимо помнить, что сильная, или высокая, корреляция — это корреляция с коэффициентом $r > 0,70$, а не просто корреляция высокого уровня значимости.

В качестве мер корреляции используются:

1) эмпирические меры тесноты связи, многие из которых были получены еще до открытия метода корреляции, а именно:

а) коэффициент ассоциации, или тетрахорический показатель связи;

б) коэффициенты взаимной сопряженности Пирсона и Чупрова;

в) коэффициент Фехнера;

г) коэффициент корреляции рангов;

2) линейный коэффициент корреляции r;

3) корреляционное отношение η;

4) множественные коэффициенты корреляции и др.

Критерием для отбора «достаточно сильных» корреляций может быть как абсолютное значение самого коэффициента корреляции (от 0,7 до 1), так и относительная величина этого коэффициента, определяемая по уровню статистической значимости (от 0,01 до 0,1), зависящему от размера выборки. В малых выборках для дальнейшей интерпретации корректнее отбирать сильные корреляции на основании уровня статистической значимости. Для исследований, которые проведены на больших выборках, лучше использовать абсолютные значения коэффициентов корреляции.

Таким образом, задача корреляционного анализа сводится к установлению направления (положительное или отрицательное) и формы (линейная, нелинейная) связи между варьирующими признаками, измерению ее тесноты и, наконец, к проверке уровня значимости полученных коэффициентов корреляции.

В настоящее время разработано множество различных коэффициентов корреляции. Самыми важными и незаменимыми являются три из них: ***r***-Пирсона, ***r***-Спирмена и ***τ***-Кендалла. Современные компьютерные статистические программы (например, SPSS) в меню «Корреляции» предлагают также эти три коэффициента, а для решения других исследовательских задач предлагаются методы сравнения групп.

Выбор метода вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные (см. табл. 3).

Таблица 3

Выбор метода вычисления коэффициента корреляции

Типы шкал		*Мера связи*
Переменная X	***Переменная Y***	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент φ, четырёхполевая корреляция
Дихотомическая	Ранговая	Рангово-бисериальный коэффициент
Дихотомическая	Интервальная или отношений	Бисериальный коэффициент
Интервальная	Ранговая	Не разработан

Для переменных с интервальной и с номинальной шкалой используется ***коэффициент корреляции Пирсона*** (корреляция моментов произведений).

В случае если одна из двух переменных имеет порядковую шкалу либо не является нормально распределённой, то используется ранговая корреляция по Спирмену или τ (тау) Кендала. Если же одна из двух переменных является дихотомической, то можно использовать точечную двухрядную корреляцию (в статистической компьютерной программе SPSS эта возможность отсутствует, вместо нее может быть применён расчёт ранговой корреляции).

В том случае если обе переменные являются дихотомическими, используется четырёхполевая корреляция (данный вид корреляции рассчитываются SPSS на основании определения мер расстояния и мер сходства)[19]. Расчёт коэффициента корреляции

[19] См. подробнее: Крыштановский А. О. Анализ социологических

между двумя недихотомическими переменными возможен только тогда, кода связь между ними линейна (однонаправлена). Если связь, к примеру, U-образная (неоднозначная), то коэффициент корреляции непригоден для использования в качестве меры силы связи: его значение стремится к нулю.

Таким образом, *условия применения коэффициентов* корреляции будут следующими:

1. Переменные, измеренные в количественной (ранговой, метрической) шкале на одной и той же выборке объектов.

2. Связь между переменными является монотонной.

Основная статистическая гипотеза, которая проверяется корреляционным анализом, является ненаправленной и содержит утверждение о равенстве корреляции нулю в генеральной совокупности H_0: $r_{xy} = 0$. При ее отклонении принимается альтернативная гипотеза H_1: $r_{xy} \neq 0$ о наличии положительной или отрицательной корреляции — в зависимости от знака вычисленного коэффициента корреляции.

На основании принятия или отклонения гипотез делаются содержательные выводы. Если по результатам статистической проверки H_0: $r_{xy} = 0$ не отклоняется на уровне α, то содержательный вывод будет: связь между X и Y не обнаружена. Если же H_0: $r_{xy} = 0$ отклоняется на уровне α, то вывод: обнаружена положительная (отрицательная) связь между X и Y. Однако к интерпретации выявленных корреляционных связей следует подходить осторожно. Следует избегать категоричных фраз типа *«переменная X является причиной увеличения показателя Y»*. Подобные утверждения следует формулировать как предположения, которые должны быть строго обоснованы теоретически.

Корреляция метрических переменных. Для изучения взаимосвязи двух метрических переменных, измеренных на одной и той же выборке, применяется **коэффициент корреляции *r*-Пирсона**. Сам коэффициент характеризует наличие только линейной связи между признаками, обозначаемыми, как правило, символами X и Y. Коэффициент линейной корреляции является

данных с помощью пакета SPSS; Наследов А. Компьютерный анализ данных в психологии и социальных науках.

параметрическим методом, и его корректное применение возможно только в том случае, если результаты измерений представлены в шкале интервалов, а само распределение значений в анализируемых переменных отличается от нормального в незначительной степени. Существует множество ситуаций, в которых его применение целесообразно. Например: влияет ли интеллект школьника на его успеваемость; влияет ли настроение на успешность выхода из проблемной ситуации; зависит ли уровень материальной обеспеченности от темперамента и т. п.

При обработке данных «вручную» необходимо вычислить коэффициент корреляции, а затем определить *p*-уровень значимости при помощи критерия ***t***-Стьюдента (в целях упрощения проверки данных пользуются таблицами критических значений $\mathbf{r}_{xy}$, которые составлены с помощью этого критерия). Величина коэффициента линейной корреляции Пирсона не может превышать +1 и быть меньше –1. Эти два числа: +1 и –1 — являются границами для коэффициента корреляции. Когда при расчете получается величина большая +1 или меньшая –1, следовательно произошла ошибка в вычислениях.

При вычислениях на компьютере статистическая программа (SPSS, Statistica) сопровождает вычисленный коэффициент корреляции более точным значением *p*-уровня. Для статистического решения о принятии или отклонении обычно устанавливают $\alpha = 0{,}05$, а для большого объема наблюдений (100 и более) $\alpha = 0{,}01$. Если $p \leq \alpha$ H_0 отклоняется и делается содержательный вывод о том, что обнаружена статистически достоверная (значимая) связь между изучаемыми переменными (положительная или отрицательная — в зависимости от знака корреляции). Когда $p > \alpha$, H_0 не отклоняется, и содержательный вывод ограничен констатацией того, что связь (статистически достоверная) не обнаружена.

Если связь не обнаружена, но есть основания полагать, что связь на самом деле есть, то следует проверить возможные причины недостоверности связи.

1. *Нелинейность связи* — для этого посмотреть график двумерного рассеивания. Если связь нелинейная, но монотонная, перейти к ранговым корреляциям. Если связь не монотонная,

то делить выборку на части, в которых связь монотонная, и вычислить корреляции отдельно для каждой части выборки или делить вы-борку на контрастные группы и далее сравнивать их по уровню выраженности признака.

2. *Наличие выбросов и выраженная асимметрия* распределения одного или обоих признаков. Для этого необходимо посмотреть гистограммы распределения частот обоих признаков. При наличии выбросов или асимметрии исключить выбросы или перейти к ранговым корреляциям.

3. *Неоднородность выборки* (посмотреть график двумерного рассеивания). Попытаться разделить выборку на части, в которых связь может иметь разные направления.

Если же связь статистически достоверна, то, прежде чем делать содержательный вывод, необходимо исключить возможность ложной корреляции.

1. Связь обусловлена выбросами. При наличии выбросов перейти к ранговым корреляциям или исключить выбросы.

2. Связь обусловлена влиянием третьей переменной. Если подобное явление возможно, необходимо вычислить корреляцию не только для всей выборки, но и для каждой группы в отдельности. Если «третья» переменная метрическая — вычислить частную корреляцию.

Коэффициент частной корреляции $\mathbf{r}_{xy\text{-}z}$ вычисляется в том случае, если необходимо проверить предположение, что связь между двумя переменными *X* и *Y* не зависит от влияния третьей переменной — *Z*. Очень часто две переменные коррелируют друг с другом только за счет того, что обе они согласованно меняются под влиянием третьей переменной. Иными словами, на самом деле связь между соответствующими свойствами отсутствует, но проявляется в статистической взаимосвязи под влиянием общей причины.

При интерпретации частной корреляции с позиции причинности следует быть осторожным, так как, если Z коррелирует и с *X* и с *Y*, а частная корреляция $\mathbf{r}_{xy\text{-}z}$ близка к нулю, из этого не обязательно следует, что именно Z является общей причиной для *X* и *Y*.

Корреляция ранговых переменных. Если к количественным данным неприемлем коэффициент корреляции ***r***-Пирсона,

то для проверки гипотезы о связи двух переменных после предварительного ранжирования могут быть применены корреляции ***r*-**Спирмена или **τ**-Кендала.

Для корректного вычисления обоих коэффициентов (Спирмена и Кендалла) результаты измерений должны быть представлены в шкале рангов или интервалов. Принципиальных отличий между этими критериями не существует, но принято считать, что коэффициент Кендала является более «содержательным», так как он более полно и детально анализирует связи между переменными, перебирая все возможные соответствия между парами значений. Коэффициент Спирмена более точно учитывает именно количественную степень связи между переменными.

Коэффициент ранговой корреляции Спирмена является непараметрическим аналогом классического коэффициента корреляции Пирсона, но при его расчете учитываются не связанные с распределением показатели сравниваемых переменных (среднее арифметическое и дисперсия), а ранги. Например, необходимо определить связь между ранговыми оценками качеств личности, входящими в представление человека о своем «Я реальном» и «Я идеальном». Так как данный коэффициент является аналогом ***r*-**Пирсона, то и применение его для проверки гипотез аналогично применению ***r***-Пирсона. То есть проверяемая статистическая гипотеза, порядок принятия статистического решения и формулировка содержательного вывода те же. В компьютерных программах (SPSS, Statistica) уровни значимости для одинаковых коэффициентов ***r***-Пирсона, и ***r***-Спирмена всегда совпадают.

Преимущество ***r***-Спирмена по сравнению с ***r***-Пирсона в большей чувствительности к связи в следующих случаях:

1) существенного отклонения распределения хотя бы одной переменной от нормального вида (асимметрия, выбросы);

2) криволинейной (монотонной) связи.

Ограничением для применения коэффициента r-Спирмена является:

1) по каждой переменной не менее 5 наблюдений;

2) коэффициент при большом количестве одинаковых рангов по одной или обеим переменным дает огрубленное значение.

Коэффициент ранговой корреляции ***τ***-Кендалла является самостоятельным оригинальным методом, опирающимся на вычисление соотношения пар значений двух выборок, имеющих одинаковые или отличающиеся тенденции (возрастание или убывание значений). Этот коэффициент называют еще коэффициентом конкордации. Основной идеей данного метода является то, что о направлении связи можно судить, попарно сравнивая между собой испытуемых: если у пары испытуемых изменение по *X* совпадает по направлению с изменением по *Y*, то это свидетельствует о положительной связи, если не совпадает — то об отрицательной связи. Количество инверсий (нарушений монотонности по сравнению с первым рядом) используется в формуле для корреляционных коэффициентов.

При подсчете ***τ***-Кендалла «вручную» данные сначала упорядочиваются по переменной *X*. Затем для каждого испытуемого подсчитывается, сколько раз его ранг по *Y* оказывается меньше, чем ранг испытуемых, находящихся ниже. Результат записывается в столбец «Совпадения». Сумма всех значений столбца «Совпадение» и есть *P* — общее число совпадений, подставляется в формулу для вычисления коэффициента Кендалла. Коэффициент ***τ***-Кендала более прост в вычислительном отношении, но при возрастании выборки, в отличие от ***r***-Спирмена, объем вычислений возрастает не пропорционально, а в геометрической прогрессии. Так, например, при $N = 12$ необходимо перебрать 66 пар испытуемых, а при $N = 489$ — уже 1 128 пар, т. е. объем вычислений возрастает более чем в 17 раз. При вычислениях на компьютере статистическая программа (SPSS, Statistica), аналогично ***r***-Спирмена и ***r***-Пирсона, сопровождает вычисленный коэффициент корреляции ***τ***-Кендала более точным значением *p*-уровня.

Применение коэффициента Кендалла является предпочтительным, если в исходных данных имеются выбросы.

Особенностью ранговых коэффициентов корреляции является то, что максимальным по модулю ранговым корреляциям (+1, –1) не обязательно соответствуют строгие прямо или обратно пропорциональные связи между исходными переменными *X* и *Y*: достаточна лишь монотонная функциональная связь между

ними. Ранговые корреляции достигают своего максимального по модулю значения, если большему значению одной переменной всегда соответствует большее значение другой переменной (+1) или большему значению одной переменной всегда соответствует меньшее значение другой переменной (–1).

Проверяемая статистическая гипотеза, порядок принятия статистического решения и формулировка содержательного вывода те же, что и для случая ***r***-Спирмена или ***r***-Пирсона.

Если статистически достоверная связь не обнаружена, но есть основания полагать, что связь на самом деле есть, то следует сначала перейти от r-Спирмена к τ-Кендала (или наоборот), а затем проверить возможные причины недостоверности связи.

1. Нелинейность связи: для этого посмотреть график двумерного рассеивания. Если связь не монотонная, то делить выборку на части, в которых связь монотонная или делить выборку на контрастные группы и далее сравнивать их по уровню выраженности признака.

2. Неоднородность выборки (посмотреть график двумерного рассеивания). Попытаться разделить выборку на части, в которых связь может иметь разные направления.

Если же связь статистически достоверна, то, прежде чем делать содержательный вывод, необходимо исключить возможность ложной корреляции (по аналогии с метрическими коэффициентами корреляции).

Подробное описание математической процедуры для каждого коэффициента корреляции дано в учебной литературе по математической статистике[20].

[20] Глас Дж., Стенли Дж. Статистические методы в педагогике и психологии. М.: Прогресс, 1976; Ермолаев О. Ю. Математическая статистика для психологов. М.: Московский психолого-социальный институт, 2003; Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. СПб.: Речь, 2004; Сидоренко Е. В. Методы математической обработки в психологии.

Контрольные вопросы к теме

1. Понятие и виды корреляции.

2. На каком уровне значимости принимаются коэффициенты корреляции?

3. Как определить силу связи по значению коэффициента корреляции?

4. Как определить направление связи по коэффициенту корреляции?

5. Показывает ли коэффициент корреляции 0,5 сильную связь? Почему?

6. Как определить значимость корреляции?

Список литературы

1. Елисеева, И. И. Группировка, корреляция, распознавание образов / И. И. Елисеева, В. О. Рукавишников. — М. : Статистика, 1977.

2. Кендалл, М. Дж. Статистические выводы и связи / М. Дж. Кендалл, А. Стьюарт. — М. : Наука, 1973.

3. Крамер, Д. Математическая обработка данных в социальных науках / Д. Крамер. —— М. : Академия, 2007.

4. Крыштановский, А. О. Анализ социологических данных с помощью пакета SPSS / А. О. Крыштановский. — М. : ГУ ВШЭ, 2006.

5. Наследов, А. Компьютерный анализ данных в психологии и социальных науках / А. Наследов. — СПб. : Питер, 2005.

6. Рабочая книга социолога / под ред. Г. В. Осипова. — М. : URSS, 2006.

7. Толстова, Ю. Н. Анализ социологических данных: Методология, дескриптивная статистика, изучение связей между но-минальными признаками / Ю. Н. Толстова. — М. : Научный мир, 2000.

8. Хилл, Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хилл. — СПб. : Питер, 2005.

9. Шишлянникова, Л. М. Применение корреляционного анализа в психологии / Л. М. Шишлянникова // Психологическая наука и образование. — 2009. — № 1.

10. Яшин, В. П. Корреляционный анализ в социологических и психологических исследованиях / В. П. Яшин. — Н. Новгород : Изд-во НКИ, 1999.

Вопросы к экзамену

1. Методы прикладной статистики и их возможности в социологии.
2. Виды шкал. Низкие и высокие шкалы.
3. Правила ранжирования.
4. Меры центральной тенденции.
5. Мода и ее расчет.
6. Медиана и ее расчет.
7. Расчет средних.
8. Меры изменчивости.
9. Размах и его расчет.
10. Дисперсия и ее расчет.
11. Квадратическое отклонение и его расчет.
12. Статистическая совокупность.
13. Графики и диаграммы.
14. Виды распределений данных в социологии.
15. Основные понятия математической статистики.
16. Понятие и виды корреляции. Коэффициенты корреляции.
17. Понятие таблиц сопряженности и их использование в социологии.
18. Критерий *хи*-квадрат.
19. Понятие ранговой корреляции. Вычисление ранговой корреляции по Спирмену.
20. Сравнение распределений и меры связи для номинальных переменных.
21. Оценка достоверности различий.
22. Оценка достоверности сдвига.
23. Дисперсионный анализ: общие принципы.
24. Однофакторный дисперсионный анализ.

25. Многофакторный дисперсионный анализ.

26. Критерий Фишера и его использование в дисперсионном анализе.

27. Кластерный анализ и его виды.

28. Факторный анализ и его виды.

29. Регрессионные модели в социологии.

30. Логлинейный анализ.

31. Дискриминантный анализ.

32. Понятие остатков в статистическом анализе и их исследование.

33. Параметрическая и непараметрическая статистика.

Список дополнительной литературы

1. Айвазян, С. А. Прикладная статистика и основы эконометрики: учебник / С. А. Айвазян. — М. : ЮНИТИ, 1998.

2. Дэйвисон, М. Многомерное шкалирование / М. Дэйвисон. — М. : Финансы и статистика, 1988.

3. Елисеева, И. И. Статистические методы измерения связей / И. И. Елисеева. — Л. : ЛГУ, 1982.

4. Миркин, Б. Г. Анализ качественных признаков и структур / Б. Г. Миркин. — М. : Статистика, 1980.

5. Миркин, Б. Г. Группировки в социально-экономических исследованиях / Б. Г. Миркин. — М. : Финансы и статистика, 1985.

6. Татарова, Г. Г. Типологический анализ в социологии / Г. Г. Татарова. — М., 1993.

7. Философский энциклопедический словарь / ред. Л. Ф. Ильичев, П. Н. Федосеев, С. М. Ковалев, В. Г. Панов. — М. : Наука. 1983.

8. Чесноков, С. В. Детерминационный анализ социально-экономических данных / С. В. Чесноков. — М. : Наука, 1982.

9. Яглом, А. М. Вероятность и информация / А. М. Яглом, И. М. Яглом. — М. : Гос. Изд-во физ-мат. литературы, 1960.

10. Ядов, В. А. Стратегия социологического исследования: описание, объяснение, понимание социальной реальности / В. А. Ядов. — М. : Добросвет, 1998.

Оглавление

Учебное издание

Гаджигасанова Наиде Сефтеровна

Методы прикладной статистики для социологов

Методические указания

Редактор, корректор М. Э. Левакова
Верстка Е. Б. Половковой

Подписано в печать 25.06.13. Формат 60×84 1/16.
Усл. печ. л. 4,18. Уч.-изд. л. 3,13.
Тираж 100 экз. Заказ

Оригинал-макет подготовлен
в редакционно-издательском отделе ЯрГУ

Ярославский государственный университет им. П. Г. Демидова.
150000, Ярославль, ул. Советская, 14.