ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ СОЦИОЛОГОВ

ЗАДАЧНИК

Ответственный редактор
Ю.Н. Толстова

Издательский дом
Государственного университета — Высшей школы экономики

Москва 2010

УДК 519.2(07)
ББК 22.172я7
М34

Рецензент:
кандидат физико-математических наук, профессор,
заведующий общеуниверситетской кафедрой высшей математики
Государственного университета — Высшей школы экономики
А.А. Макаров

Авторский коллектив:
Ю.Н. Толстова, А.А. Куликова, А.В. Рыжова, Г.Б. Юдин

ISBN 978-5-7598-0626-4

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Со времени становления в стране профессионального социологического образования (как известно, первые социологические факультеты были созданы в современной России в 1989 г.) встал вопрос о том, как преподавать так называемые математические предметы, в том числе теорию вероятностей и математическую статистику. Поясним подробнее суть вопроса.

Необходимость включения в учебный план этих дисциплин с самого начала не вызывала сомнений. Объяснялось это, в первую очередь, тем очевидным фактом, что значительная часть социальных закономерностей носит статистический характер, и их исследование немыслимо без использования теоретико-вероятностного и математико-статистического аппарата. Практика давно подтвердила эффективность такого использования. Более того, изучение истории науки показывает, что само рождение основных положений теории вероятностей и математической статистики не в последнюю очередь было связано именно с потребностями обществознания (вопреки распространенному мнению о том, что толчок к развитию упомянутых ветвей математики исходил только от азартных игр и естественных наук). Примерно до середины XIX в. представители естественных наук практически не интересовались статистическим подходом к получению нового знания. Затем ситуация резко изменилась. В указанный период физики столкнулись с проблемами, чем-то напоминающими те, которые стояли перед обществоведами (например, с теми, которые возникли при изучении газов). Родилась статистическая физика. И именно потребности физики стали служить дальнейшему развитию математической статистики. Это привело к некоторому отрыву этой науки от того, что было нужно ученым-обществоведам. Многие социологи стали отрицать необходимость использования упомянутой науки при изучении общества. Такая ситуация существовала и у нас, и на Западе. Но если в других странах она довольно быстро была преодолена, то в России, к сожалению, серьезного обучения социологов математической статистике не существовало до конца XX в.

К концу 80-х гг. XX в. у нас в стране имелась обширная и основательная литература по теории вероятностей и математической статистике. Парадокс состоял в том, что эта литература оказалась практически непригодной для обеспечения интересующего нас педагогического процесса.

Многочисленные работы по теории вероятностей и математической статистике или были направлены на профессионала-математика (русская теоретико-вероятностная и математико-статистическая школа примерно с середины XIX в. занимала лидирующие позиции в мировой науке), или носили учебно-прикладной характер с расчетом на студента-технаря. Типичный же студент-социолог оказался неспособным воспринимать эту литературу. В сознании такого студента существовала и, к сожалению, до сих пор существует своего рода психологическая «заслонка», мешающая воспринимать любой текст, написанный с использованием математического языка. Настроенные на гуманитарный лад, студенты-социологи, поступив учиться на социологический факультет, с удивлением узнают, что им придется слушать довольно много математических курсов, и возмущаются этим, полагая, что «истинный» социолог должен прежде всего понимать людей, для чего совсем не обязательно знание математики. Любой математический символ встречается «в штыки», и все, что следует за его введением, заведомо отторгается, студент уже как бы не слышит, о чем идет речь. Естественно, подобный настрой студентов обусловливает необходимость разработки специфических учебников.

Предлагаемый задачник является приложением к учебному пособию Ю.Н. Толстовой. «Математико-статистические модели в социологии (математическая статистика для социологов)». Это пособие посвящено изложению основных (ставших классическими) положений математической статистики (статистическое оценивание параметров вероятностных распределений и проверка статистических гипотез) в расчете на читателя-социолога. Основное его отличие от других учебников по математической статистике — нацеленность на преодоление указанной психологической «заслонки» в сознании читателей (имеется и ряд других отличий, но здесь мы об этом не говорим).

Авторы задачника имеют большой опыт обучения студентов-социологов математической статистике. Ими в течение ряда лет

реализовывалась учебная программа, отраженная в названной книге. Соответствующая дисциплина входила в учебный план 2-го курса факультета социологии ГУ ВШЭ. И лектор (Ю.Н. Толстова), и преподаватели, ведущие семинарские занятия (А.В. Рыжова, А.А. Куликова, Г.Б. Юдин), в процессе работы составили довольно много задач, которые предлагались студентам в качестве домашних заданий, включались в контрольные работы, служили основой экзаменационных опросов. Эти наработки и легли в основу предлагаемого задачника.

Насколько нам известно, учебно-методическое обеспечение преподавания основ математической статистики студентам-социологам (в том числе и учебники, и задачники) в России практически отсутствует. Относительно западной литературы можно сказать следующее.

Имеется немало англоязычных учебников, излагающих основы математической статистики для читателя-гуманитария. В них, в частности, включено большое количество задач[1]. Конечно, весьма актуальным является перевод этих задач на русский язык (мы сейчас не говорим о целесообразности перевода каких-то учебников целиком). Но авторы настоящего задачника полагают, что в предлагаемых ими задачах имеется нечто, с одной стороны, отличающее их от задач из англоязычной литературы, а с другой — весьма полезное для социолога. Поясним, о чем идет речь.

Главное методологическое положение, которого придерживались авторы задачника, состояло в том, чтобы формулировка задачи по возможности носила социологический характер. Конечно, трудно избежать того, чтобы просить студента, например, просто рассчитать доверительный интервал при таких-то условиях. Но авторы стремились большинство задач формулировать по-другому: обучающемуся предлагается некоторая содержательная задача, и он сам должен выбрать способ ее решения. При этом могут использоваться любые рассматриваемые в курсе методы. Например, читателю предлагается оценить активность контактов крупнейших российских компаний с Украиной по данным о количестве сделок, заключенных с названной страной несколькими российскими ком-

[1] См., например: *Bluman A.G.* Elementary Statistics: A Step by Step Approach. The McGraw Hill Companies, 1992, 1995, 1998, 2001.

паниями (задача 2 главы 4). Студент сам должен понять, что компании, по которым есть данные, — это некоторая выборка из интересующей генеральной совокупности; что общую интенсивность контактов можно оценить с помощью математического ожидания; что само математическое ожидание нельзя найти, но можно построить соответствующий доверительный интервал. В тех случаях, когда речь идет об изучении причинно-следственных отношений, ре-шения, естественно, могут быть сложнее и, как правило, неоднозначнее.

Другими словами, составители задачника полагают, что уже сама постановка задач должна говорить о том, где и когда социолог может использовать рассматриваемый метод в своей профессиональной деятельности. Более того, предполагается, что та же постановка должна учить студента видеть «подводные камни», встречающиеся при использовании математической статистики в социологии.

Авторы на практике убедились в том, что благодаря такому подходу обучающийся начинает лучше понимать, зачем социологу нужна математическая статистика, в частности, различие между содержательной и математико-статистической гипотезой. А опыт показывает, что осознание такого различия далеко не всегда просто дается студенту.

Второй аспект, которых хотелось бы отметить, касается упомянутой выше неоднозначности выбора способа ответа на поставленный социологией вопрос. Авторы учитывали, что практически для каждой социологической задачи в арсенале социологии существует целый ряд методов ее решения. Решение студентом той или иной предлагаемой задачи несколькими способами, сравнение интерпретаций применения разных методов приветствуется. Во многих задачах прямо предлагается сравнить результаты использования нескольких подходов, иногда предполагается, что студент сам должен догадаться о способе решения задачи. Из указанной посылки логически вытекает, что студент получит более высокую оценку, если он предложит больше вариантов решения одной и той же задачи.

Еще одной особенностью задачника является большое внимание его составителей к проблеме измерения. Книги, рассчитанные на технарей, естественно, не учитывают типов тех шкал, по кото-

рым получаются значения рассматриваемых случайных величин. В западных учебниках, рассчитанных на читателя-гуманитария, коротко говорится о том, что исходные данные могут быть получены по номинальным, порядковым и интервальным шкалам, а выбор метода анализа данных зависит от типа данных. Круг рассматриваемых методов анализа данных при этом ограничивается самыми простейшими операциями типа вычисления показателей средней тенденции. Авторы задачника рассматривают проблему измерения гораздо шире и глубже.

Во-первых, авторы полагали, что от студента требуется знание основных принципов теории измерений, и в первую очередь знакомство с понятием допустимых преобразований шкалы. Это дает возможность четкого понимания студентами того, почему один метод пригоден для той или иной шкалы, а другой — нет (мы имеем в виду определение формальной адекватности метода). Студент, имеющий соответствующие представления, может творчески подходить к решению вопроса о выборе метода в конкретной ситуации.

Во-вторых, при составлении задачника учитывалось, что в социологии положение осложняется тем, что процесс определения типа шкалы в социологическом исследовании зачастую не поддается формализации, определяется самим исследователем. Признаки (случайные величины) очень часто интересуют социолога не сами по себе, а лишь как некие «приборы», позволяющие оценить некие латентные (не поддающиеся явному измерению) переменные. И в таком случае фактически использующийся тип шкалы может не совпадать с тем, который фактически использовался при получении исходных данных. Например, может возникнуть ситуация, когда возраст имеет смысл считать полученным по порядковой или даже по номинальной шкале. В предлагаемых задачах студенту нередко предлагается самому решить, какой тип шкал фактически используется в рассматриваемой задаче.

Еще один аспект измерения — это разбиение диапазонов изменения рассматриваемых признаков на интервалы. Разные разбиения фактически делают исходный признак признаком — «прибором», «настроенным» на измерение разных латентных переменных. Студент привыкает к мысли о том, что при разных разбиениях диапазона изменения одного и того же признака один и тот же

анализ данных может дать совершенно разные результаты. И студенту предлагается подумать об этом.

Таким образом, данный задачник — это оригинальная работа, предназначенная для повышения эффективности процесса преподавания математической статистики студентам-социологам. Авторы надеются, что он поможет сделать соответствующий процесс более эффективным. Конечно, работа наверняка имеет недостатки. И авторы будут благодарны читателям, приславшим свои критические замечания. Кроме того ясно, что круг рассматриваемых задач должен расширяться. И в этом отношении, несомненно, огромную роль могут сыграть контакты авторов задачника с другими отечественными специалистами, занимающимися преподаванием математической статистики социологам.

Задачник состоит из 10 глав. Каждая глава соответствует главам (темам) упомянутого выше учебника.

Глава 1 «Функция распределения вероятностей и функция плотности распределения вероятностей. Основные параметры одномерного и двумерного распределения» — тема 1.

Глава 2 «Социологические шкалы и формальная адекватность методов» — тема 2.

Глава 3 «Стандартизация случайных величин. Основные распределения случайных величин» — тема 3.

Глава 4 «Интервальное оценивание параметров. Определение объема выборки» — тема 6.

Глава 5 «Принципы проверки статистической гипотезы» — тема 7.

Главы 6–7 «Проверка статистических гипотез: о равенстве средних и равенстве дисперсий. Направленные и ненаправленные альтернативные гипотезы», «Проверка статистических гипотез: о равенстве долей, равномерности генерального распределения, о равенстве нулю коэффициента корреляции» — темы 8–10.

Глава 8 «Методология проверки математико-статистических гипотез. Ошибки первого и второго рода» — тема 11.

Глава 9 «Причины и следствия в математической статистике. Корреляционное отношение» — тема 13.

Глава 10 «Дисперсионный анализ» — темы 14–15.

Каждая глава имеет небольшую теоретическую часть, в которой коротко говорится о положениях математической статистики,

лежащих в основе решения входящих в нее задач. Все задачи снабжены ответами, приведенными в конце книги.

Кроме того, в разделе «Дополнительные задачи» собраны задачи, не относящиеся к той или иной теме. Эти задачи предполагают использование нескольких методов для получения ответа на один и тот же или разные вопросы. Этот раздел, как нам кажется, необходим с педагогической точки зрения. Студент должен сам выбрать решение задачи, не пользуясь подсказкой, которая содержится в заголовке раздела.

Главы 1–4, 8, 9 написаны Г.Б. Юдиным, главы 5–7 — А.В. Рыжовой, глава 10 — А.А. Куликовой.

Ю.Н. Толстова

ГЛАВА 1

Функция распределения вероятностей и функция плотности распределения вероятностей. Основные параметры одномерного и двумерного распределения

1. Основные способы представления и плотности распределения вероятностей. Математическая статистика изучает случайные величины и параметры их распределения. Закон распределения случайной величины описывается с помощью функции распределения вероятностей, а также функции плотности распределения вероятностей.

Распределение случайной величины может быть представлено с помощью частотной таблицы, полигона распределения, гистограммы. При построении полигона распределения дискретной случайной величины по оси абсцисс откладываются значения, которые принимает величина, а по оси ординат — частота появления этого значения в рассматриваемой совокупности. При построении полигона распределения непрерывной случайной величины по оси абсцисс располагаются точки, обозначающие интервалы, в которые может попадать случайная величина (это могут быть как серединные значения интервалов, так и другие значения, в зависимости от принимаемой модели), а по оси ординат — частота появления значения признака (рис. 1.1).

Гистограмма используется только для непрерывных признаков: предполагается, что внутри каждого интервала распределение случайной величины равномерно. По оси абсцисс откладываются крайние значения интервалов, по оси ординат — частота появления значения в рассматриваемой совокупности. Таким образом, каждый интервал представляется на графике двумя точками, из

которых опускаются проекции на ось абсцисс, после чего полученные четыре точки соединяются в прямоугольники. Площади полученных прямоугольников представляют собой значения функции плотности распределения вероятностей для данных отрезков[1].

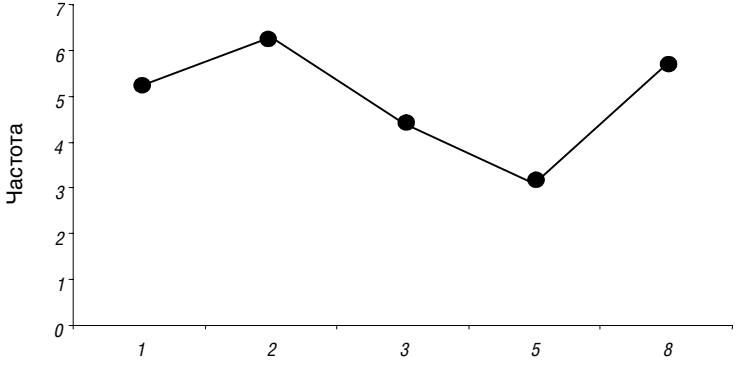

Рис. 1.1. Полигон распределения случайной величины

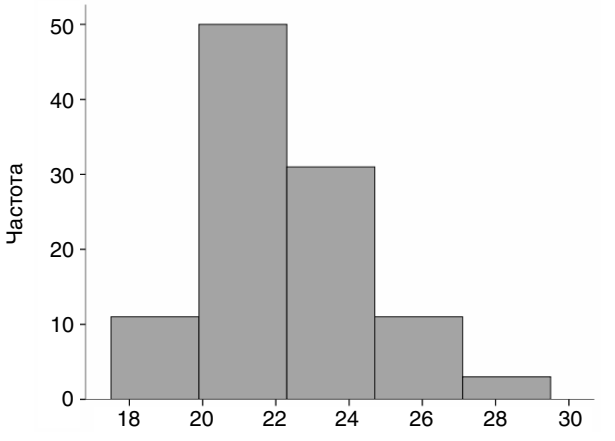

Рис. 1.2. Гистограмма

[1] Гистограмму следует отличать от столбчатой диаграммы — отличие состоит в том, что столбчатая диаграмма отображает частоту через высоту столбца, а гистограмма — через его площадь. Иными словами, столбцы столбчатой диаграммы всегда имеют одинаковую ширину, а столбцы гистограммы могут иметь разную ширину, в зависимости от интервалов.

Также распределение случайной величины может быть представлено в виде кумуляты (рис. 1.3). При построении кумуляты по оси ординат откладывается накопленная частота — частота появления в рассматриваемой совокупности всех значений случайной величины, меньших данного значения. В случае с кумулятой дискретной случайной величины по оси абсцисс располагают все возможные значения признака, а в случае с кумулятой непрерывной случайной величины — интервалы, в которые может попасть случайная величина. Во втором случае кумулята соединяет начало первого интервала с суммарными вероятностями, соответствующими концам интервалов.

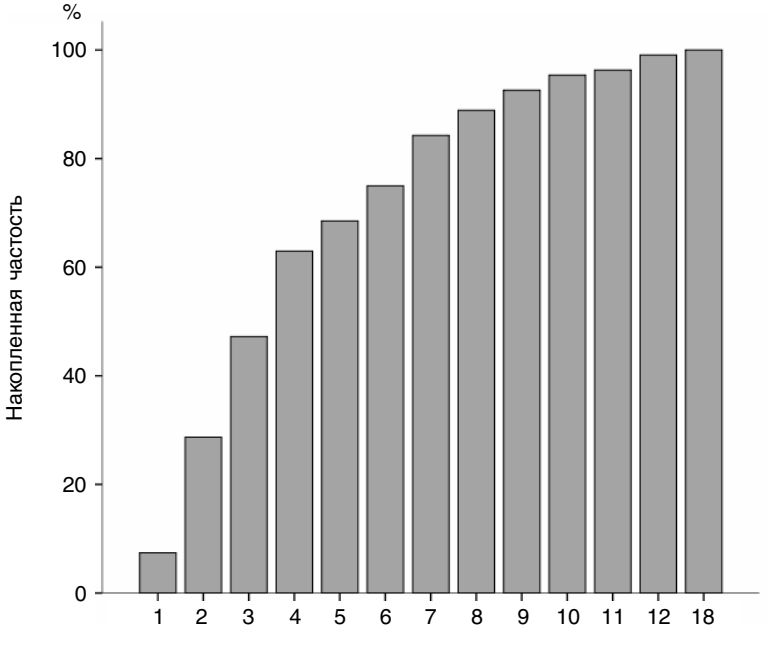

Рис. 1.3. Кумулята

Данные о частотах и частостях случайной величины, полученные на основе выборочной реализации случайной величины, называются вариационным рядом. Построение полигона, гистограммы, кумуляты на основании вариационного ряда позволяет считать

полигон и гистограмму выборочными аналогами функции плотности распределения, а кумуляту — выборочным аналогом функции распределения случайной величины.

2. Основные параметры одномерного и двумерного распределения. Если при построении кумуляты отложить по оси ординат не накопленную частоту, а накопленную частость (относительную накопленную частоту — отношение накопленной частоты к объему совокупности), можно сделать вывод о таких параметрах распределения, как квантили — значения случайной величины, при которых функция ее распределения принимает определенное значение. Квантили позволяют описать случайную величину, разбив функцию распределения на несколько одинаковых отрезков. Например, децили представляют собой квантили, разбивающие функцию распределения на десять равных частей таким образом, что первый дециль — это значение случайной величины, при котором функция распределения равна 0,1; второй дециль — значение, при котором функция распределения равна 0,2, и т.д.

В случае с непрерывными случайными величинами для расчета квантиля уровня a используется следующая формула:

$$q_a = x_0 + \delta \frac{an - n_H}{n_a}, \tag{1.1}$$

где x_0 — начало (нижняя граница) интервала, которому принадлежит квантиль; δ — величина этого интервала; n — объем совокупности (или 100%, или 1); n_H — частота (или частость, в процентах, либо в долях), накопленная до квантильного интервала; n_a — частота (или частость, в процентах, либо в долях) квантильного интервала.

Основными параметрами одномерного распределения, помимо квантилей, являются меры средней тенденции (математическое ожидание, мода, медиана) и вариации (дисперсия, среднее квадратическое отклонение, размах варьирования). Для оценки математического ожидания (MX) по выборочным данным используется среднее арифметическое: $\bar{x} = \theta(MX)$. Среднее арифметическое рассчитывается по формуле

$$\bar{x} = \frac{\sum x_i}{n}, \tag{1.2}$$

где x_i — наблюдаемые значения случайной величины; n — объем совокупности. Мода находится по формуле

$$Mo = x_0 + \delta \frac{n_{Mo} - n^-}{2n_{Mo} - n^- - n^+}, \tag{1.3}$$

где x_0 — начало (нижняя граница) модального интервала; δ — величина модального интервала; n_{Mo} — частота модального интервала; n^- — частота интервала, предшествующего модальному; n^+ — частота интервала, следующего за модальным. Если мода случайной величины оценивается на основании выборочных данных, используется обозначение $\overline{Mo}$, которое указывает на то, что речь идет о выборочной оценке параметра: $\overline{Mo} = \theta(Mo)$.

Медиана является квантилем уровня 0,5 (т.е. пятым децилем) и для непрерывной случайной величины определяется по формуле

$$Me = x_0 + \delta \frac{\frac{1}{2}n - n_H}{n_{Me}}, \tag{1.4}$$

где x_0 — начало (нижняя граница) медианного интервала; δ — величина медианного интервала; n — объем совокупности (или 100%, или 1); n_H — частота (или частость, в процентах, либо в долях), накопленная до медианного интервала; n_{Me} — частота (или относительная частота, в процентах, либо в долях) медианного интервала. Аналогично моде, выборочная оценка медианы обозначается с помощью черты: $\overline{Me} = \theta(Me)$.

Помимо этого, существует возможность определить медиану графически, с помощью кумуляты распределения. Для этого из точки, соответствующей на графике накопленной частости 50%, на ось абсцисс опускается перпендикуляр, который и указывает значение медианы. При этом следует отметить, что когда значения признака разбиты на интервалы, частость, соответствующая интервалу, может откладываться от любой точки, принадлежащей этому интервалу. Действительно, ведь у нас часто нет никакой информации о том, каким образом значения распределены внутри интервала. И несмотря на то что обычно в качестве точки, от которой откладывается частость, выступает начало интервала, эту роль мо-

жет также играть, например, середина интервала — и тогда начало координат будет находиться в точке, соответствующей середине первого интервала и нулевой частости. Соответственно значение медианы будет также зависеть от принятого способа изображения частости интервала. Аналогичные рассуждения применимы не только к расчету медианы, но и ко всем случаям, когда необходимо определить параметры случайной величины, значения которой разделены на интервалы. Эти параметры и их содержательная интерпретация будут зависеть от способа изображения интервала[1].

Для дискретной случайной величины медиана равна значению случайной величины, на которое приходится накопленная частость, равная 50%. Если в рассматриваемой совокупности задано четное число членов, возможен вариант, когда накопленной частости 50% не соответствует ни одно значение случайной величины и она находится как бы между двумя значениями. В этом случае медианой дискретной случайной величины будет считаться полусумма этих значений.

Следует заметить, что медиана и мода менее чувствительны к форме функции плотности распределения вероятности, нежели среднее арифметическое. Значение медианы не изменится, если форма распределения будет меняться только справа или только слева от медианного значения; значение моды не изменится, если изменения формы распределения не будут влиять на то, какое значение обладает наибольшей частотой. Среднее арифметическое более чувствительно к форме распределения, однако обладает наибольшей эффективностью при расчете средней тенденции на нормальном распределении (т.е., требует наименьшего объема выборки для получения точной оценки). Таким образом, в зависимости от формы распределения, значения мер средней тенденции могут как совпадать, так и различаться. Это означает, что выбор меры для определения средней тенденции непосредственно влияет как на формальный результат исследования (числовой результат), так и на его содержательный результат (интерпретацию данных).

---

[1] Дополнительная сложность возникает, когда последний интервал остается открытым, т.е. его верхняя граница не задается. В таких случаях величину этого интервала обычно (если не имеется дополнительной информации) считают равной величине предпоследнего интервала.

Дисперсия, среднее квадратическое отклонение и размах варьирования демонстрируют вариацию случайной величины, т.е. рассеяние наблюдений вокруг средних значений. Дисперсия определяется по формуле

$$DX = \sigma^2 = \frac{\sum (x_i - MX)^2}{n}. \tag{1.5}$$

Выборочную оценку дисперсии обозначают как $s^2 = \theta(\sigma^2)$; для получения несмещенной оценки в знаменатель вместо n подставляют $(n-1)$:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}. \tag{1.6}$$

Среднее квадратическое отклонение определяется как

$$\sigma = \sqrt{\frac{\sum (x_i - MX)^2}{n}}, \tag{1.7}$$

а для выборочной оценки как

$$s = \theta(\sigma) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}. \tag{1.8}$$

Размах варьирования определяется как разница между наибольшим и наименьшим значениями, которые принимает случайная величина в рассматриваемой совокупности:

$$R = x_{\max} - x_{\min}. \tag{1.9}$$

Как и в случае с мерами средней тенденции, в зависимости от выбора меры вариации результаты исследования могут существенно изменяться.

Основными параметрами для определения связи между случайными величинами (двумерное распределение) выступают ковариация и корреляция. Для расчета ковариации используется формула

$$\operatorname{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}. \tag{1.10}$$

Корреляция является результатом стандартизации ковариации на произведение средних квадратических отклонений случайных величин:

$$\rho = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}. \tag{1.11}$$

Для выборочной оценки ковариации существует более точная (несмещенная) формула

$$\operatorname{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}, \tag{1.12}$$

однако при больших выборках вместо знаменателя $(n - 1)$ используется n.

Аналогично, для выборочной оценки корреляции обычно используется выборочный коэффициент корреляции (коэффициент корреляции Пирсона)

$$r = \theta(\rho) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x \sigma_y}. \tag{1.13}$$

Коэффициент корреляции Пирсона применяют только для оценки линейной связи между случайными величинами. Этот коэффициент варьирует в пределах $x \in [-1;1]$. Абсолютное значение коэффициента показывает силу связи, а знак — ее направление. В ситуации $r = 0$ можно говорить об отсутствии связи. Если $r = 1$, имеет место строгая линейная прямая связь, если $r = -1$, речь идет о строгой обратной линейной связи. В социологии величина r редко приближается к границам интервала варьирования; обычно связь считают сильной, если $|r| > 0,3$, однако строгих правил здесь существовать не может. Но более корректным способом проверки наличия связи является проверка статистической гипотезы о равенстве нулю коэффициента корреляции (подробнее см. главу 7).

Задачи

Компетенции, на выработку которых направлены задачи:

- умение определять, что следует рассматривать в качестве случайной величины в данных условиях;

• умение строить полигон, гистограмму, кумуляту для дискретных, непрерывных признаков, а также для признаков, разбитых на интервалы;

• умение находить среднее арифметическое и моду;

• умение находить медиану разными способами;

• умение группировать значения случайной величины в интервалы, понимание зависимости формы и параметров распределения от группировки данных в интервалы и агрегирования данных;

• умение рассчитывать квантили, в том числе децили и децильный коэффициент;

• умение рассчитывать дисперсию, среднее квадратическое отклонение и размах признака;

• умение рассчитывать ковариацию и корреляцию признаков;

• общее представление об ограничениях применения выборочного коэффициента корреляции;

• умение интерпретировать результаты анализа связи между признаками;

• понимание различий содержательной интерпретации, порождаемых применением разных оценок средней тенденции, разброса и связи;

• умение выбирать наиболее адекватную формальным и содержательным условиям задачи оценку;

• умение соотносить различные виды представления исходных данных (частотные таблицы, матрицы объект/признак, матрицы признак/признак);

• общее представление о различии между параметрами и их выборочными оценками.

1.1. Одноклассники собрались спустя некоторое время после окончания школы и решили выяснить, каких карьерных успехов добился их класс. Чтобы оценить карьерные успехи, они решили собрать данные о своих заработках и получили следующее распределение:

Заработок, руб.	**Количество человек, *N***
До 1000	6
1001–1500	4
1501–2000	8

Заработок, руб.	**Количество человек, *N***
2001–2500	10
2501–3000	5
3001–4000	3
Более 4000	1

Определите значения математического ожидания, медианы и моды. Для нахождения медианы воспользуйтесь тремя известными вам способами.

Решение. По условию задачи, мы не знаем, идет ли речь о непрерывном или дискретном признаке. В самом деле, если данные о заработке были собраны с точностью, превосходящей имеющуюся группировку (например, с точностью до 1 руб.), а затем сгруппированы в интервалы, то признак имеет смысл считать непрерывным. Если же одноклассники просто указывали, к какому интервалу из предложенных относится их заработок, признак можно считать дискретным.

Рассчитаем медиану для разных вариантов.

Способ 1. Если мы считаем признак дискретным, то поскольку медианный интервал соответствует накопленной частоте 37/2 = 18,5, медиана находится в интервале 2001–2500. В этом случае, *Me* = 2001–2500.

Способ 2. Если рассматриваемый признак считается непрерывным, то накопленная частота для интервала, предшествующего медианному, равна 18, что соответствует 18/37 = 48%. Частость для медианного интервала равна 10/37 = 27%. Следовательно,

$$Me = x_{n-1} + \Delta_n \frac{50\% - P_{n-1}}{w_n} = 2000 + 500\frac{50-48}{27} =$$
$$= 2000 + 500 \cdot 0{,}07 = 2035.$$

Способ 3. Для графического определения медианы постройте кумуляту. Обратите внимание на способ изображения частости, соответствующей интервалу, — от какой точки интервала откладывается частость и соответственно какая точка первого интервала служит началом координат.

Рассчитаем моду. Модальный интервал: 2001–2500 руб. Следовательно,

$$Mo = x_0 + \delta \frac{n_{Mo} - n^-}{2n_{Mo} - n^- - n^+} = 2000 + 500 \frac{27 - 22}{2 \cdot 27 - 22 - 14} =$$
$$= 2000 + 500 \cdot 0{,}28 \approx 2140.$$

Для расчета математического ожидания также потребуется ввести предположение о том, какую точку интервала можно считать представлением всего интервала. Предположим, такой точкой выступает середина интервала. Тогда математическое ожидание составит:

$$MX = \frac{500 \cdot 6 + 1250 \cdot 4 + 1750 \cdot 8 + 2250 \cdot 10 + 2750 \cdot 5 + 3500 \cdot 3 + 4500}{37} = \\ = 1980.$$

1.2. Имеются следующие данные о трудовом стаже персонала предприятия.

Стаж работника, лет	**Менее 1 года**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9 и более лет**
Количество человек, *N*	112	214	192	136	94	73	72	53	32	22

Укажите, что в данной задаче выступает в качестве случайной величины. Постройте полигон плотности распределения случайной величины и выделите участки, на которых плотность распределения растет и падает. Сделайте выводы о кадровом составе предприятия.

1.3. В ходе исследования отношения населения к реформе ЖКХ были получены следующие результаты.

Поддерживаете ли вы проведение реформы	**Количество человек, *N***	**% ответивших**
Определенно да	30	12
Скорее да	65	26
Затрудняюсь ответить	118	47
Скорее нет	37	15
Определенно нет	0	0

Покажите, что в данной задаче выступает в качестве случайной величины. Постройте полигон распределения случайной величины. Сделайте выводы об отношении населения к реформе.

1.4. Магазин проводит опрос покупателей с целью построения социально-демографического портрета клиента. В результате исследования получены следующие данные о среднемесячном доходе респондентов.

Доход респондентов, руб.	**Количество человек, *N***
Менее 7000	254
7001–9000	312
9001–11 000	354
11 001–13 000	462
13 001–15 000	715
15 001–17 000	822
Более 17 000	611

Укажите, что в данной задаче выступает в качестве случайной величины. Постройте полигон распределения случайной величины и гистограмму распределения. Сделайте выводы о профиле покупателей по доходу.

1.5. Вы оказались в незнакомой молодежной компании. Не имея возможности прямо поинтересоваться, кем являются ваши компаньоны, вы поинтересовались их возрастом, чтобы на этом основании определить, к какой категории они принадлежат — школьники, студенты или выпускники вузов. Вы получили следующее распределение по возрасту.

Возраст респондента	**15**	**16**	**17**	**18**	**19**	**20**	**24**	**26**	**28**	**29**
Количество человек, *N*	1	2	2	3	3	4	5	4	3	2

Постройте полигон распределения исходных данных. Затем сгруппируйте возраст в интервалы 15–16 лет (школьники), 17–23 года (студенты), 24–29 лет (выпускники) и постройте полигон, а также гистограмму распределения для интервальной формы представления данных. Что вы можете сказать о компании, в которой вы находитесь?

1.6. По данным задачи 1.4 рассчитайте среднее, медиану, моду, дисперсию и среднеквадратическое отклонение для возраста

как исходных данных, так и данных, сгруппированных в интервалы.

1.7. Используя данные задачи 1.4, постройте полигон распределения и гистограмму при разбиении шкалы возраста на отрезки до 21 года и 21 год и более.

Вычислите среднее, медиану, моду, дисперсию и среднеквадратическое отклонение для возраста по новому разбиению.

По гистограмме определите вероятность того, что возраст респондента будет находиться между 21 и 30 годами.

1.8. Производитель мобильных телефонов проводит исследование, чтобы выяснить, как часто абоненты меняют аппараты. Полученные данные выглядят следующим образом.

Число мобильных телефонов, сменившихся у человека за последние 10 лет	**Число абонентов, *N***
1	120
2	323
3	781
4	1220
5	665
6	234
7	11

Полученные данные необходимо агрегировать. Изобразите полигоны распределения для следующих вариантов разбиения на интервалы: I) до 2 аппаратов; 3–4 аппарата; 5–6 аппаратов; более 6; II) до 3 аппаратов; 4 аппарата; 5 аппаратов; 6 и более аппаратов.

Укажите модальный интервал для каждого случая.

1.9. Среди курильщиков проводилось исследование, в рамках которого их попросили посчитать количество сигарет, выкуриваемых в день. Полученные данные описываются следующей частотной таблицей.

Номер распределения, *X*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**17**	**18**	**24**	**25**	**30**	**36**	**42**
Частота	2	10	9	3	4	5	3	1	1	1	2	6	1	8	16	2	6	3	1

Разделите совокупность на три группы потребителей сигарет: активные, средние и неактивные. При этом объедините значения в интервалы так, чтобы соблюдались следующие условия:

а) полигон распределения был представлен горизонтальной линией;

б) интервалы были равны.

Сравните полученные группировки и укажите, в каком случае доля активных потребителей больше.

1.10. Компания проводит сегментацию рынка по доходу. Для этого были собраны следующие данные о доходах в изучаемой совокупности потребителей.

Доход, руб.	**Количество человек, *N***
До 500	126
501–1000	245
1001–1500	511
1501–2000	468
2001–2500	126
Более 2500	24

Для того чтобы ограничить целевую аудиторию, необходимо определить порог дохода для верхних 10%, 30 и 50% потребителей. Укажите интервалы, к которым относятся соответствующие квантили, а также рассчитайте значения квантилей.

1.11. Оценку новой концепции автомобиля по 5-балльной шкале дали 30 экспертов.

Результаты приведены в следующей таблице.

Оценка	**Число экспертов**
1	2
2	6
3	9
4	10
5	3

Решение о выводе новой марки автомобиля на рынок принимается, если средняя оценка марки не ниже 4 баллов. Какое реше-

ние будет принято, если в качестве меры средней тенденции будет избрана мода? Если будет избрана медиана? Сделайте выводы.

1.12. Оператор сотовой связи оценивает затраты своих абонентов на сотовую связь. По итогам исследования получены следующие данные.

Среднемесячный счет за сотовую связь, руб.	Количество человек, *N*
Менее 300	810
301–450	1234
451–600	2620
601–750	1985
751–900	1800
Более 900	2150

Рассчитайте средние затраты на сотовую связь с помощью разных мер средней тенденции. Постройте гистограмму, кумуляту и получите медианное значение графическим путем. Какая мера средней тенденции, с вашей точки зрения, более адекватна для полученных данных?

1.13. Бывшие одноклассники собрались и решили выяснить, сколько у них в среднем детей. В результате получено следующее распределение.

Количество детей	Количество чел., *N*
0	5
1	8
2	9
3	3
4	2

К какому результату пришли одноклассники? Оцените результат с помощью разных мер средней тенденции.

1.14. Используется прогрессивная шкала налога на землю, при которой для каждой следующей категории землевладельцев ставка налога увеличивается на 1%.

Распределение землевладельцев выглядит следующим образом.

Площадь земли во владении, сотка	**Тыс. человек**
До 6	23
6	45
7	89
8	123
9	80
10	66
11	40
12	26
Более 12	22

В результате реформы пороговые значения площади для категорий землевладельцев были изменены следующим образом.

Категории землевладельцев	**Площадь земли во владении, сотка**	
	До реформы	**После реформы**
1	До 6	До 8
2	От 6 до 8	От 8 до 9
3	От 8 до 10	От 9 до 11
4	От 10 до 12	Более 11
5	Более 12	

Постройте полигоны распределения, используя группировку до и после реформы. Для обоих случаев рассчитайте среднюю площадь земли, приходящейся на одного землевладельца, с помощью математического ожидания и медианы.

1.15. При измерении некоторого признака на выборке, состоящей из 100 объектов, установлено, что размах варьирования равен 18, а медиана — 12. Также известно, что наименьшее значение встречается реже — в 2% случаев, а чаще (в 50% случаев) — значение 24. Укажите с максимально возможной точностью интервал, в котором будет лежать среднее арифметическое.

1.16. Какие содержательные выводы можно сделать по данным приведенной ниже таблицы? Для ответа используйте известные вам статистики.

Вас устраивает жизнь, которую вы ведете	Частота	%	Накопленный процент
Вполне устраивает	39	1,6	1,6
По большей части устраивает	132	5,5	7,1
Отчасти устраивает, отчасти нет	624	25,9	33,0
По большей части не устраивает	749	31,1	64,1
Совершенно не устраивает	816	33,9	98,0
Затрудняюсь ответить	48	2,0	100,0
Всего	2409	100,0	

1.17. Известно, что если заработная плата работников, выполняющих сходные функции в одном отделе, существенно различается, это негативно влияет на климат в коллективе и эффективность работы сотрудников. Будем считать, что условие приблизительного равенства доходов в рамках отдела соблюдается тогда, когда разброс в заработках (дисперсия) не превышает 4500.

Сотрудники отдела, который состоит из 10 человек, зарабатывают соответственно 240, 256, 334, 176, 254, 219, 277, 414, 215, 366 усл. единиц.

Докажите, что условие равенства доходов не соблюдается. Будет ли эта проблема решена, если перевести самого высокооплачиваемого работника в другой отдел?

1.18. На основе данных таблицы рассчитать коэффициент экономического расслоения в обществе — децильный коэффициент (отношение девятого дециля к первому децилю):

Доход, руб.	Менее 1000	1000–5000	5000–10 000	10 000–30 000	30 000–90 000	90 000 и более
Тыс. человек	160	174	40	20	4	2

1.19. В компании приятелей возник спор о том, растут ли доходы населения страны с возрастом. В конце концов, спорящие

решили выяснить, какова сила связи среди них между возрастом и доходом.

Номер респондента	Возраст, лет	Среднемесячный доход, тыс. руб.
1	16	26
2	22	15
3	25	22
4	40	30
5	22	16
6	16	30

Оцените силу линейной связи между признаками.

1.20. Используя данные таблицы, ответьте на вопросы.

а) Связаны ли показатели возраста и дохода?

б) Сравните данные этой задачи с данными задачи 1.19. В чем отличие?

Среднемесячный доход, тыс. руб.	Возраст			
	16	22	25	40
15		1		
16		1		
22			1	
26	1			
30	1			1

1.21. В ходе проверки качества работы продавцов магазина была составлена следующая таблица, отражающая связь стажа работы продавца с количеством зафиксированных в работе продавца нарушений стандартов обслуживания:

Количество нарушений стандартов обслуживания	Стаж работы продавца		
	Менее 6 месяцев	6–12 месяцев	Более 12 месяцев
0	8	16	27
1	13	19	16
2	24	13	3

Рассчитайте силу линейной связи между имеющимися показателями, сделайте выводы.

1.22. Для оценки электоральной активности населения воспользуемся индексом $IE = v/e$, где v — количество посещений выборов данным индивидом в течение последних 10 лет; e — общее количество выборов, в которых он мог принимать участие, за тот же период. Предположим, данные по этому индексу собраны в двух городах.

Город 1

Номер наблюдения	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Возраст	23	36	18	24	56	71	44	50	35	77	67	51	40	29	62
Индекс электоральной активности, *IE*	0	30	100	31	36	79	51	50	30	94	74	42	35	23	56

Город 2

Номер наблюдения	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Возраст	31	41	19	80	24	37	46	21	54	52	61	66	72	50
Индекс электоральной активности, *IE*	75	20	50	42	0	23	58	67	34	52	42	30	73	35

Сравните связь между возрастом и индексом электоральной активности в этих городах с помощью показателя ковариации. Коэффициент корреляции, рассчитанный по данным двух городов, составляет 0,31 и указывает на наличие связи между признаками. Можно ли распространить этот вывод на оба города?

1.23. Спортивные клубы заинтересованы в увеличении числа постоянных клиентов, приобретающих сезонные абонементы. Со-

гласно одной из гипотез, большие спортклубы успешнее, и поэтому способны привлечь большее количество постоянных клиентов не только в абсолютном, но и в относительном выражении — т.е. доля владельцев абонементов будет расти с ростом численности контингента. Для проверки этой гипотезы собраны данные по 16 спортклубам четырех городов. Данные агрегированы по городам. С помощью коэффициента корреляции Пирсона проверьте эту гипотезу.

Город	**Среднее число посетителей спортклуба, человек**	**Средняя доля владельцев сезонных абонементов, %**
A	5792	22
B	3667	8
C	4070	13
D	4393	15

Решите ту же задачу, используя неагрегированные данные по тем же спортклубам. Объясните, чем вызваны различия в результатах.

Город	***A***			***B***			***C***			***D***		
Спортклуб	1	2	3	1	2	3	1	2	3	1	2	3
Число посетителей	855	1290	15231	6587	3409	1005	5679	4260	2271	7520	1416	4243
Доля владельцев сезонных абонементов, %	27	21	18	19	3	2	12	19	8	7	19	19

1.24. По итогам обучения в университете часть выпускников получает за учебу красные дипломы. Можно предположить, что вузы, стремясь повысить свой статус, выдают приблизительно одинаковый процент красных дипломов, независимо от численности студентов. Имеются данные по числу студентов и доле красных дипломов в семи университетах.

Университет	Число студентов в вузе	Доля красных дипломов, %
A	2712	2,3
B	8200	3,3
C	4563	2,6
D	5322	2,7
E	1717	2,1
F	6340	2,9
G	5956	2,8

Рассчитайте коэффициент корреляции между этими признаками. Позволяет ли коэффициент корреляции сделать вывод, что доля красных дипломов различается в вузах с разным числом студентов?

1.25. Страховая компания на основе 20 наблюдений хочет узнать, существует ли связь между стажем вождения и число страховых случаев за последний год.

Номер наблюдения	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Стаж вождения (число полных лет)	0	2	1	3	4	5	0	1	4	4	5	3	2	2	4	1	3	1	1	2
Число страховых случаев за последний год	1	0	1	1	2	0	2	0	0	0	5	1	0	2	0	3	2	2	1	0

Определите силу линейной связи между признаками. Как изменится показатель, если не учитывать данные наблюдения 11?

1.26. Для того чтобы оценить компьютерную грамотность населения, исследовательская компания провела опрос, в рамках которого просила респондентов оценить по 5-балльной шкале (где 1 — «совершенно не владею», 5 — «владею в совершенстве») свое владение рядом программ. Ниже приводится элемент полученного массива данных, в котором содержатся данные об ответах пяти респондентов.

Респондент	**1**	**2**	**3**	**4**	**5**
Текстовые редакторы	4	5	4	5	3
Графические редакторы	2	2	1	5	5
Почтовые клиенты	5	5	4	4	3
Программы для поиска в сети Интернет	5	5	4	4	3
Программы для создания презентаций	3	4	4	4	2

Для изучения компьютерной грамотности формируется шкала Лайкерта, суммирующая ответы каждого респондента по всем пяти программам. Однако чтобы шкала Лайкерта обладала свойством внутренней валидности, требуется, чтобы ее элементы имели высокую корреляцию с самой шкалой (не менее 0,7) — в противном случае появится сомнение в том, что все элементы шкалы измеряют один и тот же индикатор.

На основе приведенных данных выясните, какие элементы оправданно включить в состав шкалы Лайкерта. Постройте полигон распределения полученной шкалы.

М34 **Математическая** статистика для социологов: задачник [Текст] : учеб. пособие для вузов / Ю. Н. Толстова, А. А. Куликова, А. В. Рыжова, Б. Г. Юдин ; отв. ред. Ю. Н. Толстова ; Гос. ун-т — Высшая школа экономики. — М.: Изд. дом Гос. ун-та — Высшей школы экономики, 2010. — 185 , [3] с. — 2000 экз. — ISBN 978-5-7598-0626-4 (в обл.).

Задачник «Математическая статистика для социологов» является приложением к учебному пособию «Математико-статистические модели в социологии (математическая статистика для социологов)» Ю.Н. Толстовой. В нем приводится более 170 задач, направленных на развитие навыков владения базовыми математико-статистическими моделями, используемыми при анализе социологических данных. Тематические разделы предваряются основными теоретическими положениями, которые необходимо знать для решения предлагаемых задач. Все задачи снабжены ответами. Содержательно задачи ориентированы на специфические для данной профессиональной области случаи использования математико-статистических моделей. Формулировка условия задач и предлагаемые образцы решений требуют от студента умения делать грамотные выводы, сравнивать результаты применения различных методов и использовать весь широкий арсенал математико-статистических средств для получения максимума информации из имеющихся данных.

Для студентов вузов, обучающихся по направлениям «социология», «политология», «управление», «экономика», «культурология». Задачник может быть использован при преподавании студентам данных специальностей таких дисциплин, как математическая статистика, методология социологических исследований, анализ социологических данных.

УДК 519.2(07)
ББК 22.172я7

Учебное издание

Толстова Юлиана Николаевна
Куликова Алина Алексеевна
Рыжова Анастасия Валентиновна
Юдин Григорий Борисович

Математическая статистика для социологов: задачник

Зав. редакцией *Е.А. Бережнова*
Редактор *Л.И. Кузнецова*
Художественный редактор *А.М. Павлов*
Компьютерная верстка: *А.В. Плотников*
Корректор *Е.Е. Андреева*

Подписано в печать 20.05.2010
Формат 60×88 1/16. Бумага офсетная № 1. Гарнитура Newton C. Печать офсетная
Усл.-печ. л. 11,4. Уч.-изд. л. 9,8. Тираж 2000 экз. Изд. № 986

Государственный университет — Высшая школа экономики
125319, Москва, Кочновский проезд, д. 3
Тел./факс: (495) 772-95-71

ISBN 978-5-7598-0626-4