АКАДЕМИЯ УПРАВЛЕНИЯ МВД РОССИИ

Б. А. Торопов, Ш. Х. Гонов

СТАТИСТИЧЕСКИЕ МЕТОДЫ ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ

Сборник задач
(задачник)

Москва • 2019

УДК 1:332.025.15:35
ББК 67.5в6
Т 61

Одобрено редакционно-издательским советом Академии управления МВД России

Рецензенты: И. А. Кубасов, начальник вычислительного центра ФКУ «ГИАЦ МВД России», доктор технических наук; Г. Г. Плотников, доцент кафедры информационной безопасности учебно-научного комплекса информационных технологий Московского университета МВД России им. В. Я. Кикотя, кандидат технических наук.

Т 61

Торопов Б. А., Гонов Ш. Х.
Статистические методы принятия управленческих решений : сборник задач (задачник). – М. : Академия управления МВД России, 2019. – 76 с.

ISBN 978-5-906942-84-5

Основное назначение сборника задач «Статистические методы принятия управленческих решений» – методическое обеспечение одноименной учебной дисциплины, преподаваемой на факультете подготовки научных и научно-педагогических кадров Академии управления МВД России. Однако изложенный в сборнике материал подходит для широкого круга обучающихся с различным уровнем начальных знаний в области теории принятия решений и начальных навыков применения статистических методов. Задачи могут быть легко адаптированы к различным сторонам управленческой и научно-исследовательской деятельности, как в системе МВД России, так и за ее пределами.

УДК 31:332.025.15:35
ББК 67.5в6

ISBN 978-5-906942-84-5

Оглавление

Введение

Теория принятия решений – это комплексное научное направление, характеризующееся единством цели, решаемых задач, но при этом чрезвычайным разнообразием методов и подходов к их решению.

Цель всегда одна – выбрать наилучшее управленческое решение из имеющихся альтернатив, а методы для ее достижения происходят из различных сфер человеческого знания. Процесс принятия решений далеко не всегда формализован и формализуем, однако отдельные классы решений вполне позволяют вводить строгие критерии выбора альтернатив. Как правило, это происходит тогда, когда в распоряжении лица, принимающего решение (далее – ЛПР), имеются некоторые сведения о массовых статистических показателях за некоторые предыдущие периоды, характеризующих ту сферу деятельности, в которой решение принимается.

В современной статистической науке существует большое разнообразие методов анализа данных и критериев принятия тех или иных выводов на основе этих данных.

Задачи, вошедшие в сборник, показывают, каким образом можно формировать статистические выборки для дальнейшего принятия решений на их основе, каким образом группировать статистические данные и оценивать достигнутые значения по различным показателям. Рассматриваются также методы построения линейных моделей в задачах, где имеются взаимообусловленные показатели, и методы принятия решений на основе вывода об однородности или неоднородности групп данных. Внимание уделяется также формализованным моделям принятия решений в условиях риска и неопределенности.

Книга устроена следующим образом: в начале каждого подраздела присутствует очень сжатый теоретический материал, проиллюстрированный одним или несколькими примерами. После разбора решения типовых примеров следуют относящиеся к этому же типу задачи для самостоятельного решения.

К обучающимся не предъявляется требований по наличию углубленных знаний в предметной области. Однако для решения отдельных задач необходимо программное средство MicrosoftExcel. В соответствующих примерах разбирается решение именно при помощи функций этой программы, несмотря на то, что разбор решения, как правило, достаточно подробный, к обучающимся предъявляется требование владения некоторыми базовыми навыками работы в Excel.

1. Формирование статистических выборок и проблема репрезентативности

Прежде чем приступать к рассмотрению методов принятия решений на основе статистических процедур, следует рассмотреть некоторые ключевые понятия в области математической статистики и теории вероятностей. Некоторые расчетные примеры, приведенные ниже, проиллюстрируют важные концепции, которые будут полезны как сами по себе, так и в свете решения задач из последующих разделов.

В первую очередь рассмотрим понятия генеральной совокупности и выборки, а также некоторые связанные с ними определения.

Статистическая совокупность – набор некоторых объектов, которым присуща массовость и качественная однородность, но вместе с тем – наличие вариации (отличий между объектами по одному или нескольким признакам).

Объекты, образующие статистическую совокупность – это реально существующие единицы (сотрудники, организации, подразделения, регионы, районы и т. п.), в отношении которых проводится статистическое исследование.

Единица совокупности – каждый отдельный объект статистической совокупности.

Признак – это свойство, характерная черта единиц статистической совокупности, которая может быть зафиксирована. Признаки делятся на количественные и качественные. Многообразие и изменчивость величины признака у отдельных единиц совокупности называется вариацией.

Качественная однородность – сходство всех единиц совокупности по некоторому признаку и различие по всем остальным.

В статистической совокупности отличия одной единицы совокупности от другой чаще имеют количественную природу. Количественные изменения значений признака разных единиц совокупности называются вариацией.

Атрибутивные (качественные) признаки не поддаются числовому выражению (состав населения по полу). Количественные признаки имеют числовое выражение (состав населения по возрасту).

Показатель – это обобщающая количественно-качественная характеристика какого-либо свойства единиц совокупности в целом в конкретных условиях времени и места.

Система показателей – это совокупность показателей всесторонне отражающих изучаемое явление.

Любое статистическое исследование основывается на данных, полученных в результате измерения одного или нескольких признаков для каждого из исследуемых объектов. При этом наблюдаемая совокупность объектов, статистически представленная рядом наблюдений случайной величины или нескольких величин, является выборкой, а полная совокупность этих объектов (домысливаемая, гипотетически существующая) – генеральной совокупностью. Генеральная совокупность может быть как конечной ($N = const$), так и бесконечной ($N = \infty$), выборка же – это всегда результат ограниченного ряда наблюдений. Количество объектов, вошедших в выборку, называется ее объемом. В случае, когда объем выборки достаточно велик ($n \to \infty$), она считается большой, в противном случае она называется выборкой ограниченного объема.

Неизбежно возникает вопрос: можно ли на основе изучения ограниченной выборки экстраполировать полученные выводы на всю генеральную совокупность? Чтобы это было возможно, выборка должна обладать свойством репрезентативности.

Репрезентативность – это степень соответствия характеристик выборки характеристикам генеральной совокупности. Только данные по репрезентативным выборкам можно экстраполировать на генеральную совокупность.

Существуют различные способы достижения репрезентативности выборки. Основной – это случайная выборка, когда все объекты генеральной совокупности нумеруются и при помощи генератора случайных чисел из нее извлекаются n случайных элементов. При достаточном n выборка окажется репрезентативной, поскольку у каждого элемента генеральной совокупности равные шансы быть отобранным, а значит, распределение объектов внутри выборки будет стремиться к пропорциональному от генеральной совокупности. Как правило, случайный отбор на практике не осуществим, но есть методы, позволяющие приблизить выборку к истинно случайной или другими способами обеспечить ее близость к репрезентативной. Например, при проведении социологических опросов известны следующие способы формирования выборки: стратифицированная выборка, механический отбор, серийная выборка, метод снежного кома, стихийная выборка и др.

Правильно сформированная выборка позволяет прийти к достоверным статистическим выводам о генеральной совокупности, следовательно, и к более выверенным и адекватным управленческим решениям.

1.1. Стратифицированные выборки

Далее рассмотрим несколько задач, связанных со стратифицированной выборкой, для формирования которой генеральная совокупность разделяется на категории по некоторой характеристике или нескольким характеристикам, а затем из каждой категории случайно отбираются объекты в количестве, пропорциональном объему данной категории.

Пример 1.1

Дано: генеральная совокупность для некоторого исследования – это весь списочный состав сотрудников и работников МВД России, где из 900 000 единиц 700 000 – это аттестованные сотрудники, 20 000 – федеральные госслужащие, 180 000 – вольнонаемные работники.

Требуется: выяснить, какое количество сотрудников каждой категории войдут в выборку из 100 человек.

Решение

Для того, чтобы выборка оказалась репрезентативной по виду службы, в нее должны войти различные категории в долях, пропорциональных генеральной совокупности. Так, если в выборку войдут 100 человек, то среди них должно быть 700 000*100/900 000 ≈ 78 аттестованных сотрудников, 20 000*100/900 000 ≈ 2 федеральных госслужащих и 180 000*100/900 000 ≈ 20 вольнонаемных работников.

Задача 1.1

Дано: проводится исследование того, каким образом сотрудники организации относятся к курению. Известно, что количество сотрудников организации 680 человек, из которых порядка 120 человек – курящие, среди некурящих 290 человек курили, но бросили, при этом 55 от их количества бросили в течение года. Для проводимого исследования представляют интерес следующие категории сотрудников: курящие, бросившие в течение года, бросившие раньше, никогда не курившие.

Требуется: определить, какое количество сотрудников каждой категории войдут в выборку объемом 100 человек.

Задача 1.2

Дано: в целях выявления проблем регистрационно-учетной работы центральный аппарат МВД планирует внезапные выезды контрольных проверок в районные отделения полиции 4-х регионов. В регионах различное количество отделений полиции: 1) 34; 2) 124; 3) 40; 4) 67. При этом центр может осуществить только 20 выездов проверок.

Требуется: определить, какое количество отделений полиции будет проверено в каждом из 4-х субъектов, если результаты должны быть репрезентативны относительно величины регионов.

Задача 1.3

Дано: количество сотрудников организации 450 000 человек, из которых 115 000 – мужчины, при этом 23 000 человек имеют стаж службы более 25 лет, еще 120 000 человек – от 16 до 25 лет, 240 000 – от 6 до 15 лет, остальные имеют стаж службы до 5 лет.

Требуется:

а) для формирования выборки, репрезентативной по полу, определить, какое количество мужчин и женщин войдут в выборку из 200 человек;

б) для формирования выборки, репрезентативной по стажу службы, определить, какое количество сотрудников из различных категорий по стажу службы войдут в выборку из 300 человек.

Задача 1.4

Дано: сотрудники научной организации делятся по ученому званию на профессоров, доцентов и сотрудников без ученого звания; по ученой степени – на докторов наук, кандидатов наук и сотрудников без ученой степени; по возрасту в организации выделяется 3 возрастные группы – до 30 лет, от 30 до 50 лет, старше 50 лет. Известно, что в организации нет ни одного доцента без ученой степени, нет ни одного профессора, который не был бы доктором наук, и нет ни одного доктора наук младше 30 лет.

Требуется: определить количество категорий для формирования выборки, если для проводимого исследования важны и ученое звание, и ученая степень, и возрастная группа.

1.2. Необходимый объем выборки

Далее рассмотрим вопрос определения необходимого объема n выборки, при котором ее можно считать репрезентативной. В общем случае необходимый объем выборки зависит от двух параметров: изменчивости исследуемого признака и планируемой достоверности результатов исследования.

Формула для расчета объема следующая:

$$n = \frac{t^2 p(100 - p)}{\Delta^2}, \qquad (1.1)$$

где: t – доверительный уровень, статистическая величина, значение которой для исследований в социально-правовой сфере принято 1,96 (при 95 % точности статистического вывода);

p – % объектов, у которых предположительно проявляется признак, важный для проводимого исследования;

Δ – допустимая ошибка в % задается произвольно при планировании исследования.

Пример 1.2

Дано: проводится исследование уровня стрессоустойчивости среди сотрудников. В ходе исследования ключевым фактором является наличие или отсутствие у сотрудников опыта службы в горячих точках. Предполагаемая доля сотрудников, несших службу в горячих точках, – 10 %. Допустимая ошибка при планировании исследования принята за 3 %.

Требуется: определить достаточный объем выборки.

Решение

Пользуясь формулой 1.1, получаем необходимый объем выборки:

$$n = \frac{1,96^2 \cdot 10 \cdot (100 - 10)}{3^2} = 384,16$$

Таким образом, для проведения исследования с заданной допустимой ошибкой необходимо опросить не менее 385 человек (округляем 384,16 до целого в большую сторону).

Как видно, объем выборки увеличивается, во-первых, с увеличением вариативности признака, самый большой объем выборки потребуется при вариативности в 50 %. Во-вторых, с уменьшением допустимой ошибки объем выборки также растет.

Задача 1.5

Дано: проводится исследование взглядов городского и сельского населения России на некоторые политические вопросы. По данным ФСГС на начало 2018 г. в России насчитывалось 109 179 631 человек городского и 37 662 771 человек сельского населения.

Требуется: рассчитать объем выборки, репрезентативной по принадлежности респондентов к селу и к городу, если допустимая ошибка установлена в 5 %.

Задача 1.6

Дано: проводится исследование различий во взглядах мужчин и женщин на некоторые политические вопросы. По данным ФСГС на начало 2018 г. в России насчитывалось порядка 68 044 000 мужчин и 78 760 000 женщин.

Требуется: рассчитать объем выборки, репрезентативной по полу респондентов, если допустимая ошибка установлена в 2 %.

Задача 1.7

Дано: проводится исследование аудитории российского Интернета по вопросам знакомства с ресурсами невидимого Интернета и использования анонимных прокси-серверов. При этом среднедневная аудитория Интернета в России составляет порядка 45 млн человек, а выходов в Интернет через систему анонимных прокси-серверов Tor насчитывается порядка 250 тыс. в день.

Требуется: рассчитать объем выборки, репрезентативной по использованию системы Tor при работе в Интернете, если допустимая ошибка установлена в 0,2 %.

1.3. Расчет объема необходимой выборки на основе известного объема генеральной совокупности

Иногда, если объем генеральной совокупности точно известен, например – это все работники определенной организации или все автомобили определенной марки и года выпуска, то возможно еще сильнее снизить необходимый объем выборки, пользуясь для его расчета следующей формулой:

$$n = \frac{t^2 pN(100-p)}{\Delta^2 N + t^2 p(100-p)}, \qquad (1.2)$$

где: N – объем генеральной совокупности.

Задача 1.8

Дано: проводится исследование отношения сотрудников компании к нововведениям в области социальных выплат. Важным критерием при проведении этого исследования является наличие кредитных обязательств у сотрудников. Известно, что 218 сотрудников из 540 имеют кредиты. Допустимая ошибка установлена на уровне 7 %.

Требуется:

а) рассчитать объем выборки n с учетом известной вариативности исследуемого признака (по формуле 1.1);

б) рассчитать объем выборки n с учетом известного объема генеральной совокупности (по формуле 1.2);

в) сравнить два полученных результата.

Понижающий коэффициент

Обычно, если доступная для исследования выборка составляет менее 5 % от генеральной совокупности, то эта совокупность считается большой, и расчеты проводятся по вышеприведенным правилам. Но в задаче 1.8 расчетный объем выборки составляет достаточно большую долю от генеральной совокупности. Если объем доступной выборки превышает 5 % от генеральной совокупности, то в объем выборки, рассчитанный по формулам 1.1 или 1.2 вводится понижающий коэффициент:

$$n_0 = n * \sqrt{\frac{N-n}{N-1}}, \qquad (1.3)$$

Задача 1.9

Для данных из задачи 1.8 на основе результата, полученного в задании б), рассчитайте объем выборки n_0 с учетом понижающего коэффициента по формуле 1.3.

2. Группировка данных

Часто для принятия управленческих решений необходимо сгруппировать объекты по некоторому признаку. Например, разделить районы города на группы с высоким и низким уровнями преступности или разделить сотрудников на группы по возрасту и т. п. Эти задачи решаются в целях формирования оценочных показателей и в целях организации контрольно-инспекционных мероприятий, оказания методической помощи отстающим и поощрения преуспевающих. Зачастую управленческое решение содержательно состоит в том, чтобы выделить группу тех управляемых объектов, которые по определенному признаку являются отстающими или наоборот опережающими.

Разбиение на группы также необходимо для визуального отображения статистических данных, которое позволяет наглядно продемонстрировать объекты с каким уровнем определенного показателя встречаются с какой частотой.

Рассмотрим примеры формирования групп объектов, опирающиеся исключительно на количество этих объектов и не учитывающие вариативность наблюдаемых параметров.

2.1. Группировка данных на основе количества наблюдений

Пример 2.1

Дано: имеется 60 обучающихся, каждый из которых набрал некоторое количество баллов по результатам прохождения теста.

Требуется: определить, на какое количество групп следует разбить обучающихся.

Решение

В нашем распоряжении нет никаких данных о том, какие результаты показали обучающиеся. В этом случае можно воспользоваться правилом Стерджеса, предполагающим, что количество групп будет равно единице плюс логарифм от количества объектов по основанию 2:

$$k = 1 + log_2 n \text{ ,} \qquad (2.1)$$

где n – количество объектов в выборке;
k – количество формируемых групп.

Логарифм показывает, в какую степень нужно возвести основание, чтобы получить *n*. В какую степень необходимо возвести двойку, чтобы получить число не менее 60? 2 в степени 6 дает 64, таким образом, по правилу Стерджеса количество групп обучающихся будет равно 7.

И если в нашем распоряжении все же есть результаты теста, то эти данные можно визуализировать в виде диаграммы плотности распределения, например, так:

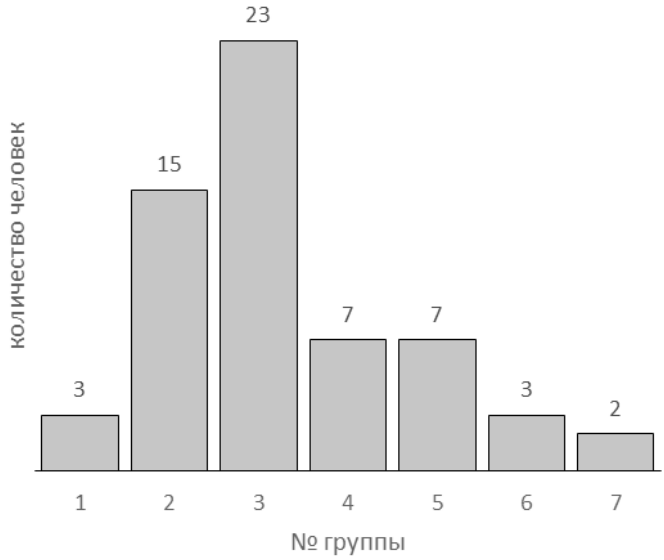


Рис. 2.1. Диаграмма плотности распределения обучающихся по группам

Границы между группами устанавливаются таким образом: вычисляется размах выборки L по исследуемому показателю, он равен разности между максимальным и минимальным значениями.

$$L = x_{max} - x_{min}, \tag{2.2}$$

где x_{max} и x_{min} – это максимальное и минимальное значения показателя.

Например, если наименьший из имеющихся результатов теста равен 15, а наибольший 100, то размах равняется L = 85. Далее рассчитаем длину интервала для группы:

$$l = L/k \qquad (2.3)$$

Разделим 85 на количество групп: 85/7 ≈ 12,14.

Интервалы групп будут такими:

1: от x_{min} до $x_{min} + l$

2: от $x_{min} + l$ до $x_{min} + 2*l$

3: от $x_{min} + 2*l$ до $x_{min} + 3*l$

4: от $x_{min} + 3*l$ до $x_{min} + 4*l$

....

k: от $x_{min} + (k-1)*l$ до $x_{min} + k*l$

В первую группу попадут обучающиеся, набравшие баллы от 15 до 27,14; во вторую – от 27,14 до 39,28; в седьмую – от 87,86 до 100.

Задача 2.1

Дано: результаты тестирования испытуемых:

Таблица 2.1

Ф.И.О.	**Результат тестирования**
Бабкин А. В.	87
Галимов А. О.	70
Гномов Н. Л.	45
Граблина А. Ф.	39
Гришин Н. Н.	92
Елкина О. Т.	82
Иванова И. И.	59
Лютов В. П.	64
Матюшенко Л. Р.	73
Наумова А. Ю.	75
Ратова П. П.	60
Саблин Е. С.	89
Семушкин Г. И.	90
Филипов В. Д.	99

Требуется: на основе правила Стерджеса рассчитать количество групп, на которые следует разбить выборку, рассчитать интервалы значений показателя для каждой группы и количество человек, попадающих в каждую из групп.

Другой способ расчета количества интервалов группировки данных, также основанный на количестве наблюдений, заключается в извлечении квадратного корня из *n*:

$$k = \sqrt{n}\,, \tag{2.4}$$

Задача 2.2

К данным из задачи 2.1 примените корень из *n* для расчета количества групп.

2.2. Группировка данных на основе их распределения внутри выборки

Другие способы расчета количества интервалов группировки данных связаны не с объемом выборки, а с распределением значений наблюдаемого параметра.

Например, можно рассчитать стандартное отклонение (среднеквадратичное отклонение) для выборки:

$$\sigma = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\,, \tag{2.5}$$

где: x_i – отдельное наблюдение;

$\bar{x}$ – среднеарифметическое по выборке.

Теперь можно для каждого наблюдения рассчитать нормированное значение:

$$x_i^* = \frac{(x_i - \bar{x})}{\sigma}\,, \tag{2.6}$$

Пример 2.2

Сгруппируем данные из задачи 2.1 на основе стандартного отклонения. Сначала рассчитаем, какую долю от стандартного отклонения по выборке составляет расстояние от среднего в каждом наблюдении. Формула 2.6 в Excel примет вид:

	Ф.И.О.	Результат тестирования	в стандартных отклонениях
2	Бабкин А.В.	87	=(B2-СРЗНАЧ($B$2:$B$15))/СТАНДОТКЛОН.В($B$2:$B$15)
3	Галимов А.О.	70	-0,17
4	Гномов Н.Л.	45	-1,56
5	Граблина А.Ф.	39	-1,89
6	Гришин Н.Н.	92	1,04
7	Елкина О.Т.	82	0,49
8	Иванова И.И.	59	-0,78
9	Лютов В.П.	64	-0,51
10	Матюшенко Л.Р.	73	-0,01
11	Наумова А.Ю.	75	0,10
12	Ратова П.П.	60	-0,73
13	Саблин Е.С.	89	0,88

Рис. 2.2. Расчет нормированных значений по выборке на основе стандартного отклонения

Теперь на листе Excel ниже полученной таблицы введем диапазоны, в которых данные будут сгруппированы и воспользуемся функцией «ЧАСТОТА» для определения того, какое количество наблюдений оказалось в каждом интервале. Обратите внимание, что данная функция Excel предназначена для расчета массива значений. Перед вводом функции задайте диапазон ячеек, а функцию вводите при помощи сочетания клавиш Ctrl+Shift+Enter.

-3	=ЧАСТОТА($C$2:$C$15;$B$18:$B$24)
-2	0,00
-1	2,00
0	5,00
1	5,00
2	2,00
3	0,00

Рис. 2.3. Расчет количества наблюдений в каждом интервале

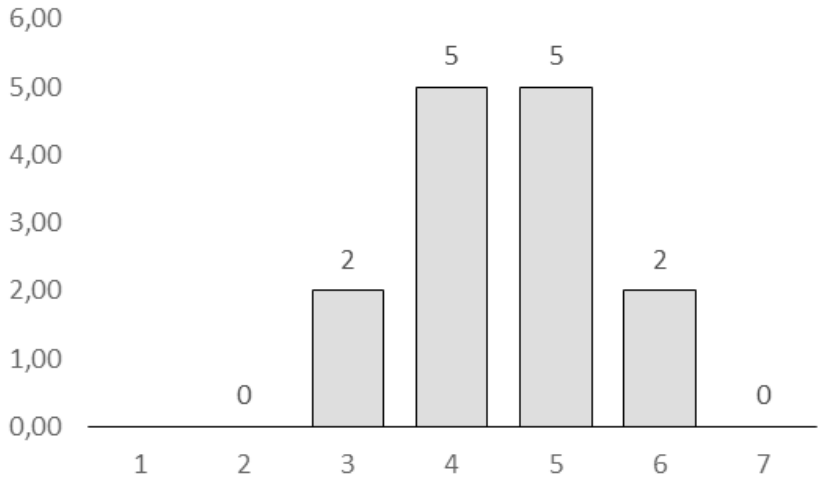


Рис. 2.4. Диаграмма

Задача 2.3

Попытайтесь самостоятельно воспроизвести расчеты, приведенные в примере 2.2, но задав следующие интервалы:

Таблица 2.2

-2,5	-2	-1,5	-1	-0,5	0	0,5	1	1,5	2	2,5

2.3. Стратифицированные выборки на основе списочных сведений

Если проводимое исследование касается конкретной общности объектов, например, сотрудников определенной организации, то часто в распоряжении исследователя имеются списочные данные, содержащие сведения о всей генеральной совокупности исследуемых объектов. На основе списочных сведений можно легко стратифицировать объекты инструментами MicrosoftExcel.

Пример 2.3

Дано: пусть имеются списочные данные, представленные в таблице:

Таблица 2.3

Ф.И.О.	ученое звание	ученая степень	возрастная группа
Андреев П.Г.	б/з	б/с	3
Андреева У.М.	б/з	б/с	1
Андреева Ю.Г.	б/з	б/с	1
Антонов Г.О.	доцент	кандидат	2
Антонова В.П.	доцент	кандидат	4
Березин Т.Л.	доцент	кандидат	2
Борисов Е.Д.	б/з	б/с	3
Боровская Г.Е.	доцент	доктор	2
Быкова Б.Б.	профессор	доктор	4
Венедиктова М.К.	б/з	кандидат	2
Вероломов Ю.О.	б/з	кандидат	4
Виленкина Е.Д.	профессор	доктор	4
Владимиров И.К.	б/з	кандидат	3
Владимирова Р.Р.	профессор	доктор	3
Воробьева Р.В.	доцент	кандидат	4
Воронов С.Р.	б/з	б/с	2
Воронов Т.Н.	профессор	доктор	2
Вячеславов П.Н.	б/з	кандидат	4
Геннадьев Ю.У.	б/з	доктор	4
Горин К.Н.	б/з	б/с	2
Горшков Л.О.	доцент	доктор	2
Грачев Б.Е.	б/з	б/с	1
Грачев У.Т.	профессор	доктор	4
Грибов Г.Б.	доцент	кандидат	2
Григ В.М.	доцент	кандидат	2
Сидоров Л.С.	б/з	б/с	1

Требуется: стратифицировать сотрудников по категориям, учитывая ученые звания, степени и возрастные группы.

Решение

Данные из таблицы 2.3 скопируем в новый файл MicrosoftExcel на пустой лист.

Мышью необходимо выделить любую ячейку с данными.

Далее откроем вкладку Excel «Вставка» и нажмем кнопку «Сводная таблица».

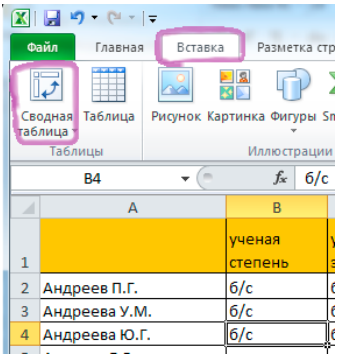


Рис. 2.5. Расположение инструмента Excel «Сводная таблица»

В появившемся окошке важно обратить внимание, чтобы поле «Таблица или диапазон» содержало диапазон всех имеющихся данных. Нажмем ОК.

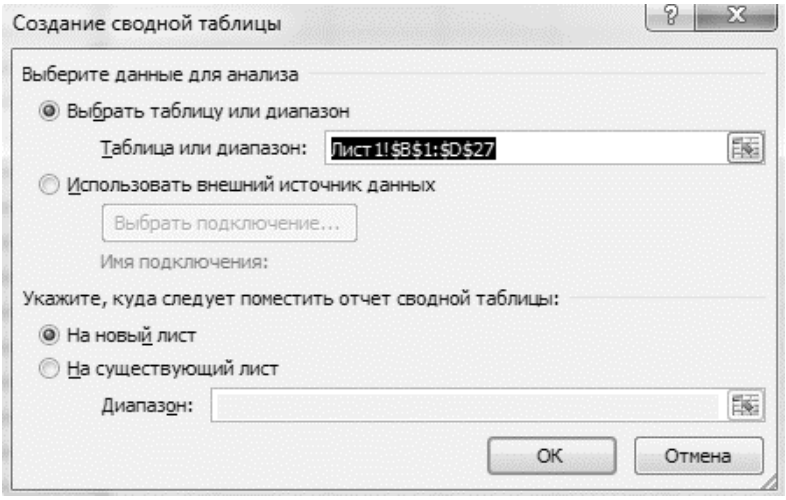


Рис. 2.6. Создание сводной таблицы

На новом листе Excel появится заготовка сводной таблицы.

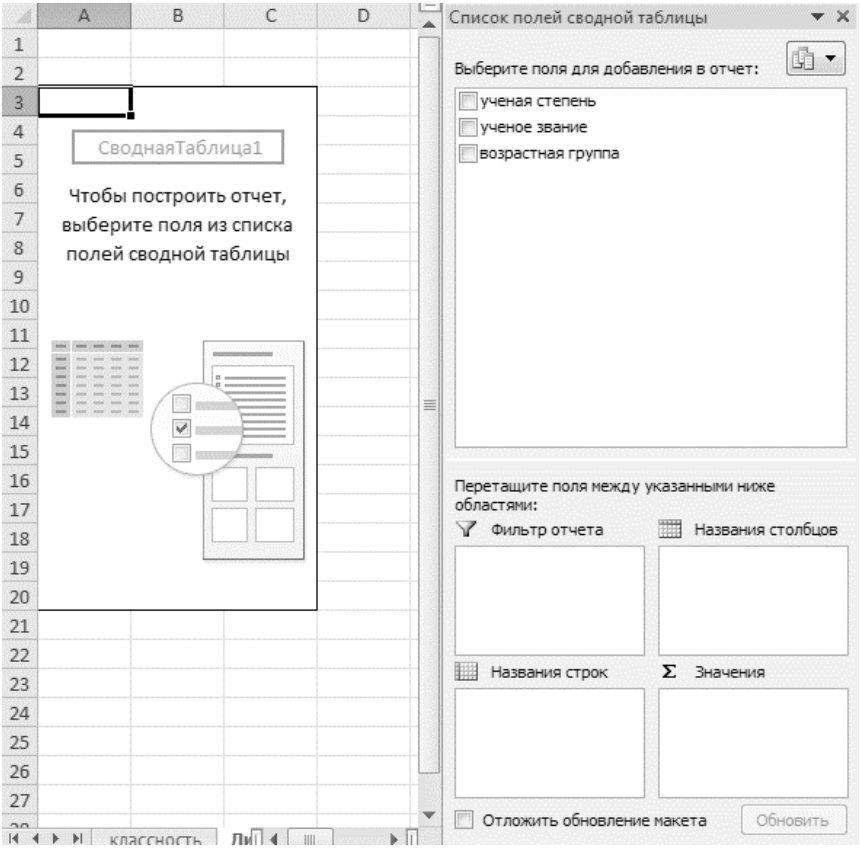


Рис. 2.7. Заготовка сводной таблицы

Теперь перетащим названия отдельных категорий в области сводной таблицы как показано на рисунке 2.8 в его правой части. В область «названия строк» поместим «ученую степень» и «ученое звание», в область «названия столбцов» – «возрастную группу». В область «значения» снова перетащим «ученую степень». Важно обратить внимание, чтобы в области «значения» отобразилось количество по полю «ученая степень», но ни в коем случае не сумма или что-либо другое.

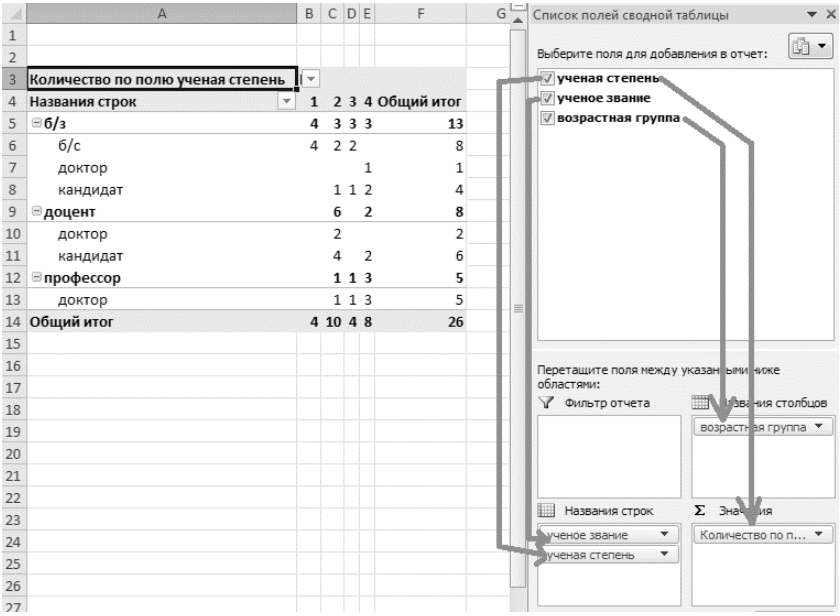


Рис. 2.8. Формирование сводной таблицы

Результат представлен на рисунке 2.8 в его левой части. Сводная таблица позволила перейти от списочных данных к массовым статистическим данным по отдельным категориям. Например, видим, что среди сотрудников без ученого звания имеется 13 человек. Среди них 8 человек не имеют ученой степени. 4 из этих 8 относятся к 1 возрастной группе, по 2 человека находятся во второй и третьей возрастных группах. Ни одного человека без степени и звания нет в четвертой возрастной группе, то есть эта категория пуста.

Задача 2.4

Дано: списочные сведения о классности сотрудников подразделения полиции:

Таблица 2.3

№ п/п	Дата назначения на должность	Классность	специальное звание	возрастная группа
1	03.09.2016	нет	ст. лейтенант	2
2	15.12.2015	2	капитан	4
3	05.03.2017	1	лейтенант	2
4	26.04.2017	2	ст. лейтенант	1
5	04.09.2016	1	подполковник	3
6	29.04.2016	1	ст. лейтенант	2
7	25.11.2014	3	подполковник	4
8	25.07.2018	2	ст. лейтенант	3
9	25.03.2014	2	капитан	2
10	12.05.2015	2	ст. лейтенант	2
11	13.06.2015	2	капитан	2
12	31.10.2014	2	лейтенант	1
13	11.08.2017	нет	капитан	3
14	03.03.2016	3	ст. лейтенант	2
15	24.11.2017	2	подполковник	5
16	25.05.2016	3	ст. лейтенант	3
17	11.01.2015	2	лейтенант	2
18	29.10.2018	1	ст. лейтенант	1
19	07.11.2015	1	ст. лейтенант	2
20	31.08.2018	1	лейтенант	1

Требуется:

а) определить количество категорий сотрудников, если для проводимого исследования важна классность сотрудника, специальное звание и возрастная группа;

б) определить, какое количество сотрудников войдет в каждую категорию из задания (а);

в) определить количество категорий сотрудников по классности и специальному званию, из числа тех, кто назначен на должность раньше, чем три года назад;

г) определить, какое количество сотрудников войдет в каждую категорию из задания (в).

3. Однородность данных

3.1. Проверка однородности групп данных на основе критерия Фишера

Большой класс управленческих решений связан с определением того, являются ли объекты, вошедшие в выборку однородными, или, напротив, разделяются на разнородные категории по некоторому параметру.

В статистике существуют строгие способы сделать вывод об однородности или неоднородности данных. Ниже для учебного примера возьмем небольшую выборку, хотя рассматриваемый метод дает более точные результаты при наличии большого объема данных.

Пример 3.1

Дано: исходные данные по условному показателю *X* такие:

Таблица 3.1

№ п/п	X
1	400
2	1003
3	2360
4	1784
5	2510
6	3080
7	7740
8	6400
9	7461
10	7957
11	9035
12	9490

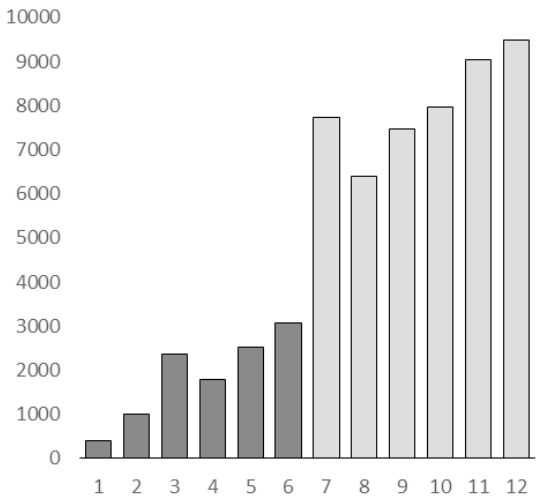


Рис. 3.1. Значения X

Диаграмма (рис. 3.1) показывает, что наблюдения с 7 по 12 обладают более высоким уровнем показателя *X*. Таким образом, выборку можно разделить на две части. Проверим это. Сделать строгий вывод

о неоднородности поможет критерий Фишера. Сначала проведем подготовку к его применению. Данные скопируем на новый лист Excel.

Рассчитаем среднее значение X по всей выборке целиком. Это позволяет сделать функция из категории «Статистические» «СРЗНАЧ», в графу Число 1 аргументов этой функции занесем значения из столбца X находящиеся в диапазоне ячеек.

Теперь рассчитаем средние значения в обеих подвыборках X_1 и X_2, неоднородность которых проверяется. На рисунке ниже первый диапазон B2:B7, второй – B8:B13.

	A	B	C
1	№ п/п	X	
2	1	400	
3	2	1003	
4	3	2360	
5	4	1784	
6	5	2510	
7	6	3080	
8	7	7740	
9	8	6400	
10	9	7461	
11	10	7957	
12	11	9035	
13	12	9490	
14			
15	среднее по X		
16	4935		
17	среднее по X1		
18	1856,167		
19	среднее по X2		
20	=СРЗНАЧ(B8:B13)		

Рис. 3.2. Расчет средних значений

Сумма квадратов отклонений X от среднего SS складывается из двух компонент: суммы квадратов отклонений внутри подвыборок SSW и суммы квадратов отклонений между подвыборками SSB:

$$SS = SSW + SSB \quad (3.1)$$

Определим, какую долю в общей сумме квадратов отклонений *SS* занимает сумма квадратов отклонений внутри подвыборок *SSW*.

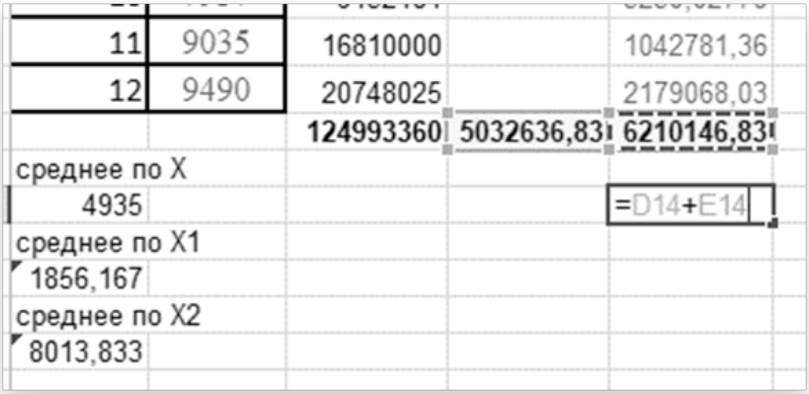
Если больше половины, значит различия между двумя подвыборками невелики и данные можно считать однородными.

Рассчитаем квадраты отклонений значений *X* от среднего как показано на рис. 3.3. Рассчитаем таким же способом квадраты отклонений внутри подвыборок от подвыборочных средних значений, после чего найдем суммы квадратов отклонений по всем трем сформировавшимся столбцам: сумма квадратов отклонений по всей выборке *X*; по первой подвыборке; по второй подвыборке.

	A	B	C
1	№ п/п	X	кв. откл. X
2	1	400	=(B2-$A$16)^2
3	2	1003	
4	3	2360	
5	4	1784	
6	5	2510	
7	6	3080	
8	7	7740	
9	8	6400	
10	9	7461	
11	10	7957	
12	11	9035	
13	12	9490	
14			
15	среднее по X		
16	4935		
17	среднее по X1		
18	1856,167		
19	среднее по X2		
20	8013,833		

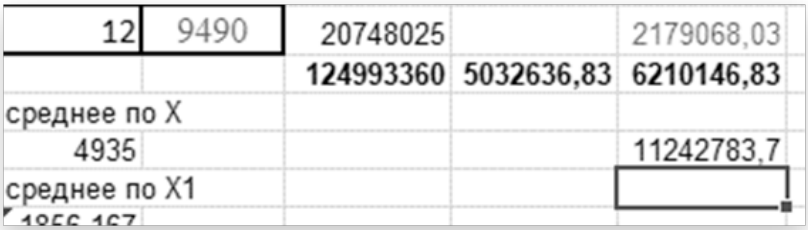
Рис. 3.3. Расчет квадратов отклонений от среднего

Теперь рассчитаем сумму квадратов отклонений для обеих подвыборок вместе:



11	9035	16810000		1042781,36
12	9490	20748025		2179068,03
		124993360	5032636,83	6210146,83
среднее по X				
4935				=D14+E14
среднее по X1				
1856,167				
среднее по X2				
8013,833				

Рис. 3.5. Расчет сумм квадратов отклонений подвыборок



12	9490	20748025		2179068,03
		124993360	5032636,83	6210146,83
среднее по X				
4935				11242783,7
среднее по X1				
1856,167				

Рис. 3.6. Результат

Получен результат SSW = 11 242 784.

Ранее мы выяснили, что SS = 124 993 360. Это на порядок больше, чем *SSW*. Таким образом, сумма квадратов отклонений *SS* в основном состоит из отклонений между подвыборками SSB = SS – SSW= 124 993 360 – 11 242 784 = 113 750 576, а не внутри них *SSW*. То есть данные можно считать неоднородными.

Применим строгий критерий – критерий Фишера.

Сначала рассчитаем эмпирическое значение *F* для имеющихся данных по формуле:

$$F = \frac{SSB/(m-1)}{SSW/(N-m)}, \qquad (3.2)$$

где m – это количество подвыборок;

N – количество наблюдений.

Результат такой:

$$F = \frac{113750576/(2-1)}{11242784/(12-2)} \approx 101$$

Теперь рассчитаем критическое значение *F*, причем вообще-то это значение уже рассчитано Фишером. Оно – табличное, и Excel позволяет его определить при помощи функции «F.ОБР.ПХ». У этой функции три аргумента: 1) вероятность – это вероятность получить значение *F* выше критического случайно, при том что выборка на самом деле однородна (в исследовании социальных систем принято использовать вероятность 5 %); 2) степени свободы 1 – это значение m-1, уже рассчитанное нами выше (m–1 = 1); 3) степени свободы 2 – это N – m (в рассматриваемом случае N–m = 12–2 = 10). Таким образом аргументы функции «F.ОБР.ПХ» следует заполнить так:

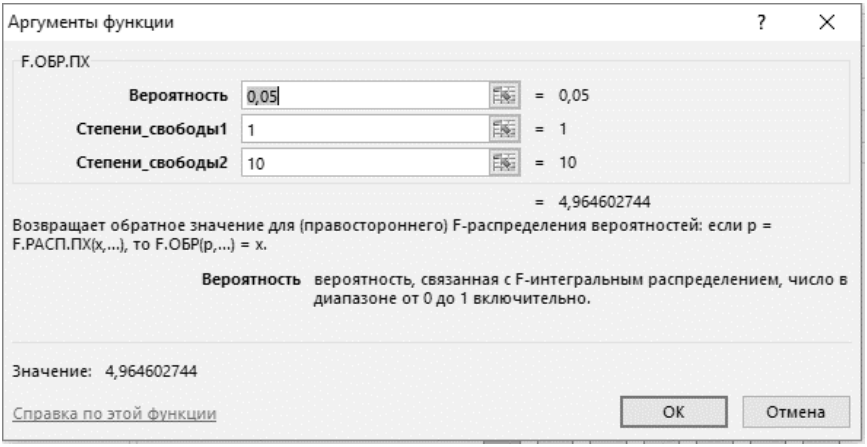


Рис. 3.7. Аргументы функции «F.ОБР.ПХ»

Результат расчета F ≈ 5. Эмпирическое значение 101 значительно превышает критическое, что позволяет сделать вывод о неоднородности исходной выборки.

Задача 3.1

Дано: исходные данные характеризуют преступность несовершеннолетних в условных районах некоторой области, в 7 из которых введена новая система профилактики преступности несовершеннолетних, в остальных 12 районах – нет. В таблице ниже данные о количестве преступлений несовершеннолетних на 10 000 населения.

Таблица 3.2

Районы	**Новая система профилактики**	**Количество преступлений несовершеннолетних на 10 000 населения**
Соколовский	введена	0,23
Воронинский	введена	0,32
Орловский	введена	0,2
Сорокинский	введена	0,17
Воробьевский	введена	0,23
Синички	введена	0,24
Снегиревский	введена	0,1
Совиный	нет	0,05
Галкинский	нет	0,65
Ястребовский	нет	0,39
Гусево	нет	0,41
Уткино	нет	0,32
Дятловский	нет	0,59
Лебединый	нет	0,29
Журавлево	нет	0,47
Глухаринский	нет	0,52
Тетеревка	нет	0,59
Голубевский	нет	0,24
Филино	нет	0,42

Требуется: опираясь на критерий Фишера, принять решение о том, стоит ли признать новую профилактическую систему эффективной и применять в остальных районах.

Задача 3.2

Дано: исходные данные характеризуют длину жизненного цикла внутренних управляющих документов, циркулирующих в подразделениях организации. В 10 подразделениях введена инновационная схема согласования документов, направленная на ускорение прохождения жизненного цикла документов, в других 8 – нет.

Таблица 3.3

Подразделения	Инновационная система согласования	Средний жизненный цикл документа
Подразделение 1	введена	12,2
Подразделение 2	введена	7,6
Подразделение 3	введена	15,3
Подразделение 4	введена	10,3
Подразделение 5	введена	6,2
Подразделение 6	введена	14,2
Подразделение 7	введена	4,1
Подразделение 8	введена	17,5
Подразделение 9	введена	10,6
Подразделение 10	введена	11,9
Подразделение 11	нет	12,1
Подразделение 12	нет	5,2
Подразделение 13	нет	6,5
Подразделение 14	нет	12,2
Подразделение 15	нет	18,7
Подразделение 16	нет	4,2
Подразделение 17	нет	7,9
Подразделение 18	нет	11,4

Требуется: опираясь на критерий Фишера, принять решение о том, стоит ли признать инновационную систему эффективной и тиражировать этот опыт далее.

Задача 3.3

Дано: три группы сотрудников полиции, каждая из которых проходила огневую подготовку по своей схеме: 1) тренировалась в электронном тире; 2) тренировалась в обычном тире; 3) тренировалась в обычном тире под руководством мастеров «Динамо».

Таблица 3.4

Сотрудник № п/п	Вид подготовки	Результаты стрельб (количество попаданий)
1	1	87
2	1	89
3	1	80
4	1	92
5	1	87
6	1	90
7	2	88
8	2	93
9	2	86
10	2	90
11	2	79
12	2	85
13	2	99
14	3	98
15	3	99
16	3	95
17	3	89
18	3	90

Требуется: на основе критерия Фишера определить, имеется ли статистически значимая разница между группами, и если да, то какой вид подготовки дает наилучшие результаты.

3.2. Принятие решений об однородности на основе ресемплинга

Как уже отмечалось, метод определения однородности на основе критерия Фишера и другие похожие на него методы лучше подходят для больших выборок.

Для случаев, когда исследователь вынужден довольствоваться маленькими наборами данных, существует подход, задействующий вместо теоретических положений вычислительные мощности компьютера и именуемый – ресемплинг. Этот подход

позволяет сгенерировать произвольное количество выборок еще меньшего объема (сэмплов), чем реально имеющаяся. Для полученных выборок измеряются массовые статистические показатели, которые затем сравниваются с полученными в исходной выборке, на основе чего и формируется статистический вывод.

К методам ресемплинга относится так называемый бутстрэп анализ, состоящий в многократном извлечении выборок меньшего объема из реально имеющейся небольшой выборки. Для оценки любых параметров можно сформировать тысячи повторных бутстрэп-выборок (обычно 500–10 000), каждая из которых содержит 2/3 исходных объектов.

Рассмотрим пример использования бутстрэп-анализа в принятии управленческого решения.

Пример 3.2

Дано: пусть в некоторой области имеется 20 отделений ГИБДД, в которых граждане получают водительские права и регистрируют транспортные средства. Нагрузка по количеству обращающихся за правами человек во всех 20 отделениях примерно равна. При этом в 10 пилотных отделениях внедрена новая информационная система, позволяющая оптимизировать процесс обслуживания граждан, в других 10 отделениях работа ведется без ее использования.

В распоряжении ЛПР имеются данные о максимальной пропускной способности в единицах человек в день для каждого отделения.

Таблица 3.5

Пилотные отделения	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9	№ 10
Количество человек	57	46	29	65	56	49	38	49	54	45

Обычные отделения	№ 11	№ 12	№13	№ 14	№ 15	№ 16	№ 17	№ 18	№ 19	№ 20
Количество человек	43	27	30	46	32	51	47	36	23	20

Требуется: выяснить, эффективна ли новая информационная система и следует ли внедрять ее по всей области.

Решение

Рассчитываем t-нулевое – это различие между средними арифметическими по двум группам. Получается 13,3, то есть в группе пилотных отделений пропускная способность в целом повысилась. Однако необходимо определить, не находится ли это значение в рамках статистической ошибки. Данных для принятия статистического вывода недостаточно.

Теперь предположим, что внедрение новой ИС никак не связано с пропускной способностью отделений. Если это так, то неважно, к какой из двух групп относится каждое наблюдение. Случайно перемешиваем участников и рассчитываем разницу между средними для 1 000 пар смешанных выборок по 7 отделений в каждой (примерно 2/3 от 10).

Для этого перенесите имеющиеся данные в Excel, расположив их в одну строку.

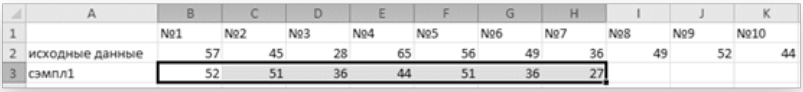
	A	B	C	D	E	F		T	U
1		№1	№2	№3	№4	№5	• • •	№19	№20
2	исходные данные	57	45	28	65	56		23	20
3	сэмпл 1								
4	сэмпл 2								

Рис. 3.8. Подготовка к ресэмплингу

Строки ниже пронумеруем при помощи автозаполнения как «сэмпл1», «сэмпл2» и т. д. до «сэмпл 1000». Составление подвыборок осуществим при помощи функций Excel «СМЕЩ» и «СЛУЧМЕЖДУ».

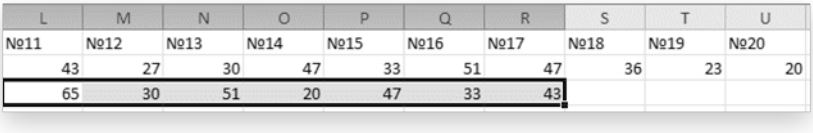
	A	B	C	D	E
1		№1	№2	№3	№4
2	исходные данные	57	45	28	65
3	сэмпл1	=СМЕЩ($A$2;0;СЛУЧМЕЖДУ(1;20))			

Рис. 3.9. Формула для отбора наблюдений,
которые войдут в подвыборку



	A	B	C	D	E	F	G	H	I	J	K
1		№1	№2	№3	№4	№5	№6	№7	№8	№9	№10
2	исходные данные	57	45	28	65	56	49	36	49	52	44
3	сэмпл1	52	51	36	44	51	36	27			

Рис. 3.10. Генерация 1-й подвыборки из семи элементов



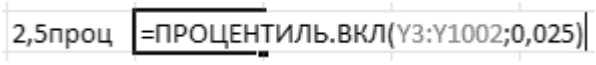
L	M	N	O	P	Q	R	S	T	U
№11	№12	№13	№14	№15	№16	№17	№18	№19	№20
43	27	30	47	33	51	47	36	23	20
65	30	51	20	47	33	43			

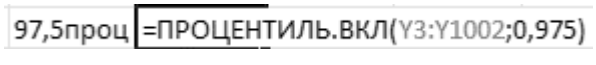
Рис. 3.11. Генерация 2-й подвыборки из семи элементов

Каждая из двух получившихся подвыборок состоит из элементов исходных данных, причем взятых случайным образом вне зависимости от принадлежности к одной из двух групп. Кроме того, эти элементы могут повторяться.

Когда две первые подвыборки сформированы, можно применить автозаполнение и сформировать все 1 000 пар подвыборок. Затем для каждой пары рассчитать разность средних *ti*.

Полученные разности и будут той базой, которая позволит сделать статистический вывод. Отсортируем 1 000 полученных *t* по возрастанию. Теперь найдем 2,5 процентиль и 97,5 процентиль от полученных разностей. Для этого можно воспользоваться функцией Excel «процентиль»:





У этой функции 2 аргумента: массив (выбирается весь диапазон, содержащий разности) и доля (для 2,5 процентиля – 0,025 и для 97,5 процентиля – 0,975).

Если t-нулевое, рассчитанное ранее, не входит в центральные 95 % значений эмпирического распределения, то у нас есть все основания отвергнуть нулевую гипотезу о равенстве средних значений в двух группах. Собственно интерес представляет 97,5 процентиль

полученных разностей. Если t-нулевое больше этого значения, то внедренную ИС можно признать эффективной и принять решение о ее дальнейшем распространении.

Само решение может отличаться от раза к разу, поскольку случайно генерируемые подвыборки каждый раз будут иметь разные средние значения. Чем больше подвыборок генерируется, тем более стабильное решение будет получаться.

Так, при подготовке этого раздела расчетное значение 97,5 процентиля было получено 12,57 чел. То есть самые большие 2,5 % разностей между двумя случайно сформированными подвыборками превышали 12,57 чел. Вспомним, что разница между средними для исходных данных равнялась 13,3 чел., что позволяет сделать вывод о наличии существенной разницы в пропускной способности отделений, где использовалась новая ИС и теми, где она не использовалась.

Задача 3.4

Дано: в 10 из 20 районов внедрена новая система видеонаблюдения на улицах. Имеются данные за отчетный период о количестве зарегистрированных преступлений, которые совершены на улицах в этих 20 районах.

Требуется: при помощи ресэмплинга сделать вывод об эффективности или неэффективности внедренной системы видеонаблюдения.

Таблица 3.6

Пилотные районы	№ 1	№ 2	№ 3	№ 4	№ 5	№ 6	№ 7	№ 8	№ 9	№ 10
Количество преступлений	17	27	11	12	15	10	7	20	14	13

Обычные районы	11	12	13	14	15	16	17	18	19	20
Количество преступлений	21	22	23	24	25	26	27	28	29	30

Задача 3.5

На основе данных из задачи 3.1 и метода ресэмплинга определите является ли эффективной новая система профилактики преступлений несовершеннолетних.

4. Линейные модели в принятии решений

Некоторые управленческие решения имеют параметрический вид, то есть задают количественное значение некоторого показателя (штатной численности сотрудников, пропускной способности некоторого канала, выделяемых ресурсов и т. п.). Если этот параметр обусловливает некоторую другую статистическую величину, управляемую таким образом опосредованно, то эффект управления может быть задан линейной регрессионной моделью.

4.1. Парные линейные регрессионные модели

Рассмотрим парные линейные регрессионные модели. Такие модели имеют вид:

$$y = \beta_0 + \beta_1 * x, \qquad (4.1)$$

где: y – зависимая переменная, то есть та самая опосредованно управляемая величина;

x – независимая переменная, тот параметр, который выражает управленческое решение;

β_0 – значение y, в случае нулевого x, в графическом отображении линейной модели – это координата по оси y, где линия модели пересекается с осью y (на рисунке ниже это координата 1);

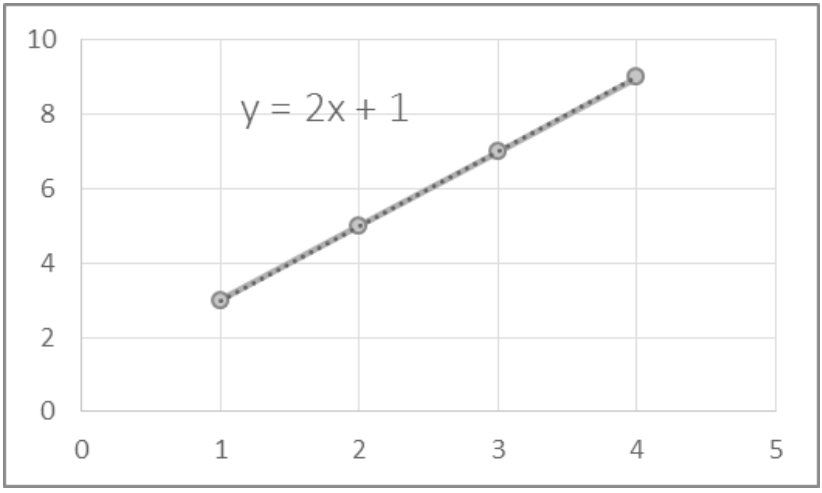


Рис. 4.1. Графическое отображение линейной модели

β_1 – коэффициент при независимой переменной x, показывает, на сколько в среднем прирастает y при увеличении x на единицу, графически – это наклон линии модели.

Здесь не будем приводить методик расчета коэффициентов. В расчетном примере используем встроенный инструмент Excel «Регрессия», позволяющий сделать это автоматически.

Пример 4.1

Дано: необходимо собрать представительное собрание собственников жилья. Данные о предыдущих таких собраниях показывают, что доля пришедших на собрание связана с долей обзвоненных собственников от общего их количества. Данные такие.

Таблица 4.1

№ набл.	1	2	3	4	5	6	7	8	9	10	11	12
Обзвон, %	50	42	20	50	60	60	70	0	30	12	20	25
Явка, %	15	21	12	23	32	33	24	8	11	8	11	10

Требуется: определить, какую долю собственников необходимо обзвонить для обеспечения явки порядка 20 %.

Решение

Данные занесем в Excel, расположив в столбик.

	A	B	C
1	№ набл.	Обзвон, %	Явка, %
2	1	50	15
3	2	42	21
4	3	20	12
5	4	50	23
6	5	60	32
7	6	60	33
8	7	70	24
9	8	0	8
10	9	30	11
11	10	12	8
12	11	20	11
13	12	25	10

Рис. 4.2. Данные

Во вкладке «Данные» выберем «Анализ данных».

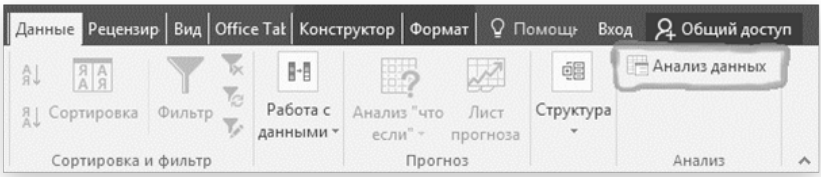


Рис. 4.3. Вкладка «Данные» – «Анализ данных»

В появившемся меню инструментов Excel выберем «Регрессия».

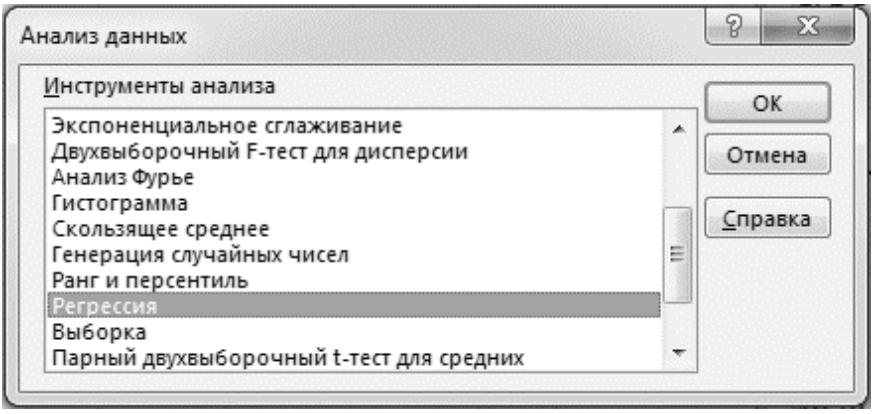


Рис. 4.4. «Анализ данных» – «Регрессия»

В появившемся окне необходимо задать «Входной интервал Y» – столбец данных, отражающих % явки, «Входной интервал X» – столбец данных, отражающих % обзвона собственников. Выходной интервал можно расположить как на новом листе, так и на текущем.

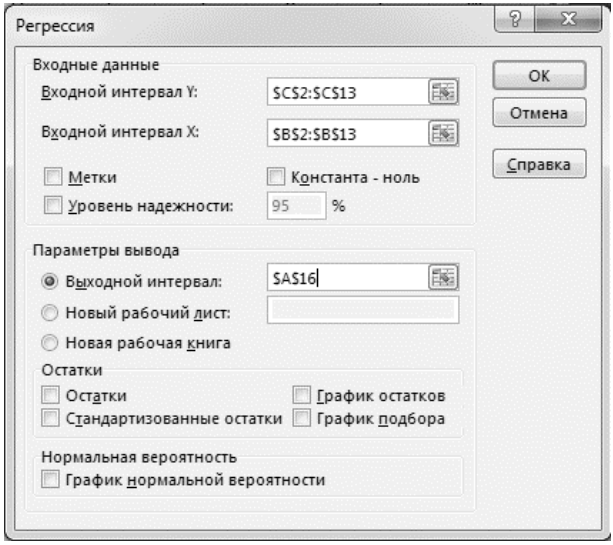


Рис. 4.5. Параметры инструмента «Регрессия»

Результат расчета линейной регрессионной модели выглядит так:

Регрессионная статистика				
Множественный R	0,870533			
R-квадрат	**0,75783**			
Нормированный R-кв	0,73361			
Стандартная ошибка	4,649508			
Наблюдения	12			
Дисперсионный анализ				
	df	*SS*	*MS*	*F*
Регрессия	1	676,4874174	676,4874	31,29289
Остаток	10	216,1792493	21,61792	
Итого	11	892,6666667		
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	**4,22982**	2,699708612	1,56677	0,148238
Переменная X 1	**0,35818**	**0,064029712**	5,594005	**0,00023**

Рис. 4.6. Результат расчета модели

Прокомментируем некоторые важные значения, полученные в результате расчета:

Y-пересечение ≈ 4,23 – это коэффициент β0, такой процент собственников явится на собрание, если никого не обзванивать (в среднем согласно модели);

Переменная X 1 ≈ 0,36 это коэффициент β1, на такой процент повысится явка с увеличением процента обзвоненных собственников на 1 % (опять же, в среднем согласно модели);

Таким образом, уравнение полученной модели (с учетом округления):

$$y = 4{,}23 + 0{,}36 * x$$

R–квадрат ≈ 0,76 – доля объясненной дисперсии переменной *Y*, можно сказать, что полученная модель объясняет состояние переменной *Y* на 76 %.

Стандартная ошибка ≈ 0,064 – характеризует точность расчетных значений, в среднем они отклоняются от реально наблюдаемых в прошлом на 6,4 %.

p-значение – показывает уровень значимости коэффициента β_1, характеризуя вероятность того, что на самом деле *x* не оказывает значимого влияния на *y*, в рассматриваемом случае это примерно 0,00023, то есть эта вероятность составляет 0,23 %. При том что в статистике при изучении социально-правовых явлений принято допустимым p-значение, не превышающее 5 %, признаем коэффициент β_1 значимым.

По условию требовалось найти такой процент обзвоненных собственников, чтобы явка составила порядка 20 %. Тогда 20 = 4,23+ 0,36*x. Соответственно:

$$x = (20 - 4{,}23)/0{,}36 \approx 43{,}81 \quad .$$

Такой уровень обзвона позволит обеспечить явку в среднем равную 20 %.

Задача 4.1

Дано: имеются данные о плотности нарядов полиции на 1 площади (км^2) обслуживаемой территории и о количестве уличных преступлений, зарегистрированных на этой территории.

Таблица 4.2

№ наб.	Плотн., ед./км^2	Прест., ед.	№ наб.	Плотн., ед./км^2	Прест., ед.
1	0,32	5	8	0,15	8
2	0,12	12	9	0,11	11
3	0,15	9	10	0,38	8
4	0,08	11	11	0,45	2
5	0,4	7	12	0,5	1
6	0,32	6	13	0,38	2
7	0,21	7	14	0,04	12

Требуется: определить плотность нарядов, необходимую для поддержания уровня преступности на улицах порядка 3.

Задача 4.2

Дано: данные ниже характеризуют среднедушевые доходы населения в регионах Центрального федерального округа России и количество выявленных несовершеннолетних, совершивших преступления на 10 000 населения.

Таблица 4.3

Регион	Средне-душевые денежные доходы (в месяц), руб.	Выявлено несовершеннолетних, совершивших преступления на 10 т.н.
Белгородская область	28 327	1,879923
Брянская область	25 375	3,641702
Владимирская область	23 732	3,962684
Воронежская область	30 109	2,505205
Ивановская область	22 560	3,703314
Калужская область	27 550	4,007972
Костромская область	22 466	4,599704

Регион	Средне-душевые денежные доходы (в месяц), руб.	Выявлено несовершеннолетних, совершивших преступления на 10 т.н.
Курская область	25 814	2,720655
Липецкая область	27657	2,400971
Московская область	37 622	1,620784
Орловская область	22 840	3,397667
Рязанская область	24 219	2,228215
Смоленская область	24 763	3,35824
Тамбовская область	25 076	3,407312
Тверская область	23 450	3,786868
Тульская область	26 286	2,504014
Ярославская область	27 369	3,491584

Требуется: построить регрессионную линейную модель зависимости количества несовершеннолетних, совершивших преступления на 10 000 населения от денежных доходов населения, определить, на сколько в среднем изменяется количество несовершеннолетних преступников при увеличении доходов на 1 000 рублей.

4.2. Однородность регрессионных моделей

В разделе 3 нами уже рассматривались методы установления того, однородны ли данные по 1 показателю. Теперь рассмотрим случаи, когда вывод об однородности или неоднородности данных делается на основе совокупности двух показателей.

Пример 4.2

Имеются данные по двум условным показателям, причем каждое наблюдение относится к одному из двух типов. Точечная диаграмма показывает, что данные первого и второго типов хоть и показывают примерно одинаковую тенденцию к спаду показателя Y с увеличением показателя X, однако явно различаются.

Таблица 4.4

№	тип	X	Y
1	1	4328,94	9 450
2	1	7779,41	9 045
3	1	12 410	7 897
4	1	18 480	7 521
6	1	22 950	7 690
5	1	23 020	7 500
7	2	25 610	3 180
8	2	28 340	2 210
9	2	31 500	1 484
10	2	32 910	2 260
11	2	33 950	450
12	2	34 930	1 403

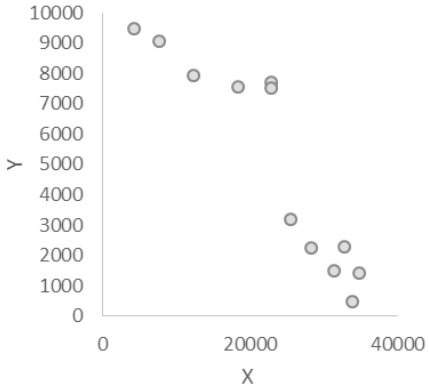


Рис. 4.7. Точечная диаграмма

Требуется: выяснить однородны ли данные. Можно ли изучать их в рамках одной линейной модели или лучше для каждой группы построить свою?

Решение

Построим общую регрессионную модель y = β0 + β1∗x, как это уже делалось нами ранее. Результат таков:

рессионная статистика			
Множеств	0,912714		
R-квадрат	0,833046		
Нормиров	0,816351		
Стандартн	1467,58		
Наблюден	12		
Дисперсионный анализ			
	df	*SS*	*MS*
Регрессия	1	107467361	1,07E+08
Остаток	10	**21537903,98**	2153790
Итого	11	129005265	
	Коэффициент	*тандартная ошиб*	*татисти*
Y-пересеч	11951,24	1070,414417	11,16506
Переменн	-0,30167	0,042707234	-7,06377

Рис. 4.8. Результат расчета общей регрессионной модели

Здесь нас будет интересовать характеристика остатков модели, это значение на рисунке выделено жирным, оно находится в столбце SS блока «Дисперсионный анализ».

Теперь построим две частные регрессионные модели, отдельно для данных первого типа $y_{т1} = \beta0 + \beta1 * x_{т1}$, отдельно для второго $y_{т2} = \beta0 + \beta1 * _{хт2}$.

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	1	1830227,317	1830227	4,188931	0,110104
Остаток	4	**1747679,5**	436919,9		
Итого	5	3577906,833			

Рис. 4.9. Результат расчета регрессионной модели
для данных первого типа

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	1	2804894,077	2804894	7,222859	0,054798
Остаток	4	**1553342,8**	388335,7		
Итого	5	4358236,833			

Рис. 4.10. Результат расчета регрессионной модели
для данных второго типа

Таким образом, нами получены характеристики остатков по общей модели и двум частным:

SSобщ.= 21537903,98

SS_1= 1747679,5

SS_2= 1553342,8

Сделаем вывод об однородности данных на основе теста Грегори Чоу.

Рассчитаем эмпирическое значение по формуле:

$$F = \frac{(SS_{общ.} - SS_1 - SS_2)/(k+1)}{(SS_1 + SS_2)/(n - 2 * k - 2)}, \quad (4.2)$$

где: n – количество наблюдений, в нашем случае 14;

k – количество независимых переменных, в нашем случае 1.

Имеем следующий результат:

$$F = \frac{(21537903{,}98 - 1747679{,}5 - 1553342{,}8)/(1+1)}{(1747679{,}5 + 1553342{,}8)/(14 - 2*1 - 2)} \approx 22{,}1$$

Теперь выясним критическое табличное значение *F*, как это уже делалось нами ранее при помощи функции Excel «F.ОБР.ПХ».

Заполним три аргумента функции: 1) вероятность – 0,05; 2) степени свободы 1 – это значение k+1, уже рассчитанное нами выше (k+1 = 2); 3) степени свободы 2 – это n – 2*k-2 (в рассматриваемом случае 8). Таким образом, аргументы функции «F.ОБР.ПХ» следует заполнить так:

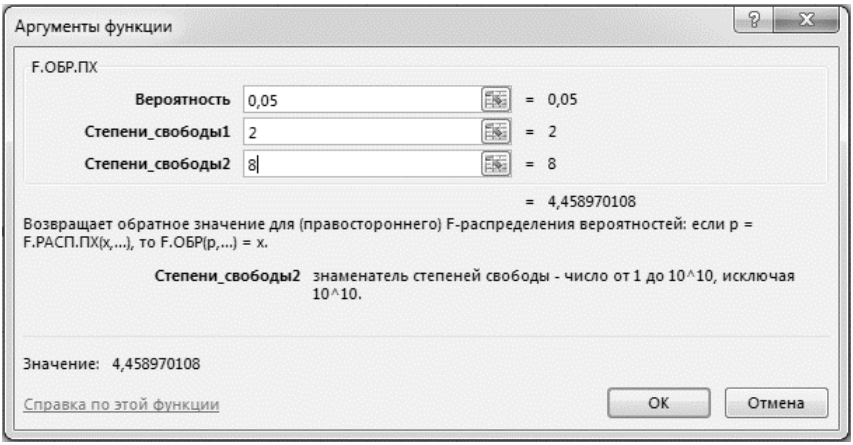


Рис. 4.11. Аргументы функции «F.ОБР.ПХ» для теста Г. Чоу

Получим результат *F крит.* ≈ 4,46.

Расчетное значение существенно превышает табличное, значит данные неоднородны.

Задача 4.3

Дано: данные двух типов по двум показателям: характеризуют деятельность дорожной полиции в двух регионах: *X* – среднее количество экипажей дорожной полиции; *Y* – среднее количество выписанных штрафов. Тип 1 – первый регион, тип 2 – второй.

Таблица 4.5

№	тип	X	Y
1	1	15	90
2	1	16	125
3	1	17	140
4	1	17	132
5	1	19	133
6	1	20	150
7	2	12	45
8	2	14	44
9	2	15	54
10	2	17	71
11	2	18	78
12	2	19	67

Требуется: при помощи теста Г. Чоу определить однородны ли данные, то есть одинаково ли изменяется в регионах 1 и 2 количество штрафов в зависимости от количества полицейских экипажей.

5. Статистические методы в оценке результатов деятельности

Формирование системы оценки некоторой деятельности – это важнейшая управленческая задача. Система оценки должна стремиться к объективному учету всех реально достигнутых объектами оценивания результатов, при этом быть непротиворечивой, а также в наименьшей степени способствовать фальсификации, искусственному завышению или занижению достигаемых показателей.

В этом разделе рассматриваются статистические методы и процедуры, позволяющие осуществлять оценивание. Приводимые примеры на этот раз будут связаны с деятельностью научных организаций.

5.1. Парето-оптимальные состояния достигнутых результатов

Оптимальность по Парето – такое состояние некоторого объекта (системы), при котором значение каждого характеризующего этот объект показателя не может быть улучшено без ухудшения как минимум одного другого показателя.

Оптимальное по Парето множество представляет собой множество объектов, обладающих следующим свойством: любой из объектов, входящих во множество Парето, хотя бы по одному показателю лучше любого другого объекта, входящего в это множество.

Управление научной деятельностью в организации помимо прочего предполагает оценивание ее результатов, достигнутых как отдельными сотрудниками, так и структурными подразделениями. Действительно, оценка необходима для выработки и реализации стимулирующей политики в сфере научной деятельности. Но каким образом построить непротиворечивую и объективную систему формирования оценочных показателей? Здесь возникает ряд сложностей, ведь результаты научной деятельности разнообразны, а объективного подхода к установлению их значимости, скорее всего, просто не существует.

Пример 5.1

Научную деятельность в организации осуществляли 3 сотрудника, каждый из которых в течение года мог достигнуть различных результатов по четырем показателям.

Таблица 5.1

Ф.И.О.	**Опубликовано научных статей**	**Опубликовано учебной литературы**	**Завершено научно-исследовательских работ**	**Получено свидетельств об интеллектуальной собственности**
Земляникин С.С.	14	2	2	0
Апельсинов П.А.	9	3	2	2
Вишневский А.В.	10	4	4	0
Подсолнухов Р.О.	9	3	2	1

Земляникин опубликовал больше всех статей, Апельсинов единственный получил свидетельства об интеллектуальной собственности, а Вишневский опубликовал наибольшее количество учебной литературы. Таким образом, по всей совокупности показателей среди этих трех сотрудников нельзя выделить лидеров и отстающих. Очевидно лишь, что результаты Подсолнухова не превосходят результатов Апельсинова.

В этом примере результаты Земляникина, Апельсинова и Вишневского образуют оптимальное по Парето множество. Они составляют три таких состояния достигнутых научных результатов, переход среди которых из одного состояния в другое с целью увеличения любого показателя невозможен без ухудшения как минимум одного либо другого показателя.

Действительно, если бы Земляникин работал так же как Апельсинов, то он получил бы на 2 свидетельства больше и на 1 больше публикацию учебной литературы, но проиграл бы в количестве опубликованных статей. Если бы примеру Апельсинова последовал бы Вишневский, то он тоже получил бы больше свидетельств, проиграв при этом по всем остальным показателям. Другими словами, проведя все возможные попарные сравнения среди этих троих нельзя выделить лидера, который выиграет хотя бы одно сравнение.

Если же сравнить результаты Подсолнухова с каждым из троих других, то получим следующий результат:

Подсолнухов не выигрывает и не проигрывает сравнения с Земляникиным, опережая его по двум показателям и уступая в одном.

Подсолнухов не выигрывает и не проигрывает сравнения с Вишневским, опережая его по одному показателю и уступая в трех других.

Но Подсолнухов проигрывает сравнение с Апельсиновым, поскольку, не уступая ему в трех показателях, по одному все же отстает.

Задача 5.1

Дано: список научных сотрудников подразделения и достигнутых ими научных результатов.

Таблица 5.2

Ф.И.О.	**Принято участие в конференциях**	**Опубликовано монографий**	**Внедрено результатов научной деятельности**
Зайцев А. А.	4	4	4
Волков В. Н.	5	2	6
Лисицина Е. Г.	2	7	3
Барсукова М. М.	2	3	3

Требуется: определить научных сотрудников, составляющих оптимальное по Парето множество.

5.2. Свертка показателей и суперкритерий

Свертка показателей – переход от множества показателей к одному, на основе которого рассчитывается суперкритерий.

В предложенных выше примерах оценивалось небольшое количество сотрудников по небольшому же набору показателей. В реальной организации, ведущей научную деятельность, как правило, будет больше сотрудников и больше видов результатов их труда. В этом случае, во-первых, провести все попарные сравнения будет затруднительно. Во-вторых, Парето-оптимальное множество может оказаться довольно большим и при этом все равно придется

разрабатывать механизмы сравнения и оценки сотрудников, попавших в него.

Наиболее простым способом формирования оценочных результатов по множеству показателей является их свертка к одному единственному показателю, то есть формирование *суперкритерия*.

Каким образом можно от четырех показателей, приведенных в таблице 1, перейти к одному? Необходимо взять от них какую-то функцию. Очевидно, что самые простые варианты – это просуммировать их, или, например, перемножить друг на друга для каждого сотрудника. Получим следующие результаты:

Таблица 5.3

Ф.И.О.	**Сумма показателей**	**Произведение показателей**
Земляникин С.С.	18	0
Апельсинов П.А.	16	108
Вишневский А.В.	18	0
Подсолнухов Р.О.	15	54

В случае суммирования лидируют Земляникин и Вишневский, в случае умножения на первом месте Апельсинов. Однако видим, что объективность обоих способов оценки, особенно второго, вызывает вопросы.

Во втором случае, примененное нами умножение, слишком сурово наказало Земляникина и Вишневского за отсутствие свидетельств, нулевые значения не позволили им соперничать с оставшимися сотрудниками.

При суммировании же определяющее значение сыграли статьи, поскольку этот показатель имеет самую большую размерность среди прочих, колеблясь от 9 до 14, а вот свидетельства напротив практически потеряли свою значимость по причине малых количеств, от 0 до 2. То есть размерность показателей определила их значимость относительно друг друга. Преодолеть эту проблему позволяет предварительная нормировка или нормализация значений показателей.

Нормализация – преобразование значений показателей, выражаемых в различных единицах, к безразмерному виду с целью их сопоставления и сравнительной оценки.

Методы нормализации должны обладать следующими свойствами:

1. Учитывать значимые параметры, значения которых изменяются в небольшом диапазоне по сравнению с другими.

2. Результирующие (нормированные) величины должны находится в ограниченном интервале.

Способов нормировать данные может быть довольно много. Например, в органах внутренних дел принята методика оценивания территориальных подразделений на региональном уровне, предполагающая так называемое минимаксное нормирование значений показателей. Алгоритм нормирования значений по показателю следующий:

1. Определяются минимальное x_{min} и максимальное $x_{\max}$ значения среди всех имеющихся.

2. Нормированное значение по показателю для территориального органа *i*, обозначим его как x_i вычисляется по формуле:

$$\text{оценка}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} * 100, \quad (5.1)$$

если этот показатель положительный, и по формуле:

$$\text{оценка}_i = \frac{x_{max} - x_i}{x_{max} - x_{min}} * 100, \quad (5.2)$$

если показатель отрицательный.

Пример 5.2

Дано: воспользуемся данными из таблицы 5.2.

Требуется: рассчитать нормированные значения трех показателей, достигнутых учеными.

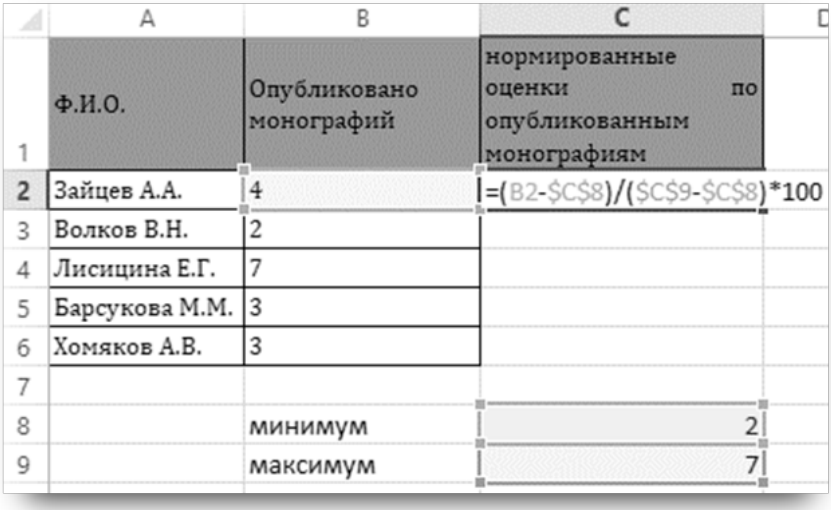
Решение

Приведем пример нормировки по показателю «Опубликовано монографий». Сначала для имеющихся данных найдем минимальное и максимальное значения. Для этого можно использовать функции Excel «МИН» и «МАКС» соответственно.

	A	B	C
1	Ф.И.О.	Опубликовано монографий	
2	Зайцев А.А.	4	
3	Волков В.Н.	2	
4	Лисицина Е.Г.	7	
5	Барсукова М.М.	3	
6	Хомяков А.В.	3	
7			
8		минимум	=МИН(B2:B6)
9		максимум	7

Рис. 5.1. Расчет минимального и максимального значений по выборке

Теперь при помощи формулы (5.1) рассчитаем нормированные оценки.



	A	B	C	D
1	Ф.И.О.	Опубликовано монографий	нормированные оценки по опубликованным монографиям	
2	Зайцев А.А.	4	=(B2-$C$8)/($C$9-$C$8)*100	
3	Волков В.Н.	2		
4	Лисицина Е.Г.	7		
5	Барсукова М.М.	3		
6	Хомяков А.В.	3		
7				
8		минимум	2	
9		максимум	7	

Рис. 5.2. Расчет нормированных оценок показателя

	A	B	C
1	Ф.И.О.	Опубликовано монографий	нормированные оценки по опубликованным монографиям
2	Зайцев А.А.	4	40
3	Волков В.Н.	2	0
4	Лисицина Е.Г.	7	100
5	Барсукова М.М.	3	20
6	Хомяков А.В.	3	20
7			
8		минимум	2
9		максимум	7

Рис. 5.3. Результат

Видим, что Лисицина, опубликовав максимальное количество монографий, получила и максимально возможный балл – 100, Волков получил 0, остальные оценки расположились на 100-балльной шкале пропорционально достигнутым результатам.

Переход от разнородных показателей к нормированным оценкам позволяет рассчитать итоговую оценку на основе аддитивной свертки. Например так, как показано на рисунке 5.4. Все три оценки просуммированы и полученный результат разделен на их количество – 3. Отметим, что полученные итоговые оценки также лежат в 100-балльной шкале.

	A	B	C	D	E	F	G	H
1	Ф.И.О.	Принято участие в конференциях	Опубликовано монографий	Внедрено результатов научной деятельности	нормированные оценки по участию в конференциях	нормированные оценки по опубликованным монографиям	нормированные оценки по внедрению	ИТОГОВАЯ ОЦЕНКА
2	Зайцев А.А.	4	4	4	40	40	40	=СУММ(E2:G2)/3
3	Волков В.Н.	5	2	6	60	0	80	46,67
4	Лисицина Е.Г.	2	7	3	0	100	20	40,00
5	Барсукова М.М.	2	3	3	0	20	20	13,33
6	Хомяков А.В.	5	3	6	60	20	80	53,33
7								
8	минимум	2	2	3				
9	максимум	5	7	6				

Рис. 5.4. Расчет итоговой оценки

Задача 5.3

Требуется: воспроизвести расчет нормированных оценок результатов работы ученых по всем трем показателям из таблицы 5.2, а также рассчитать итоговую оценку.

В рассматриваемых нами примерах все показатели научной деятельности положительные, хотя в принципе возможно и введение отрицательных. Например, количество полученных отрицательных рецензий на научные работы, или количество НИР отклоненных заказчиком, или количество выявленных некорректных заимствований. Тогда для расчета оценки используется формула (5.2).

Задача 5.4

Дано: сведения о результатах научной деятельности коллектива ученых:

Таблица 5.4

Ф.И.О.	**Количество опубликованных статей**	**Количество выступлений на конференциях с опубликованными докладами**	**Количество полученных отрицательных рецензий**
Нетесов А. В.	12	16	0
Грачев О. Р.	14	22	0
Смекалкин А. Д.	24	19	0
Неелова Б. М.	12	12	0
Римман Е. Т.	19	16	2
Лейбниц Е. М.	16	4	0
Венедиктов Е. Ф.	10	30	0
Мусоргский Р. Г.	26	21	0
Павлинов Г. Ф.	15	17	0
Свечков Е. С.	16	7	0
Шапиро Ф. Г.	16	0	0
Горин Б. Ф.	16	12	0
Игорев П.Ф.	22	13	1
Олегов О. В.	23	0	0
Воронов Н. С.	15	17	0
Рыжков П. О.	6	8	0
Партков Р. С.	4	16	1
Амбаров Р. Д.	21	14	0

Требуется: рассчитать нормированные оценки по всем трем показателям (с учетом того, являются они положительными или же негативными), рассчитать итоговые оценки ученых. Определить на основе оценок пять лучших ученых для поощрения их премиями.

6. Элементы теории массового обслуживания в принятии решений

Процесс принятия решений может описываться с позиций различных теорий, известных в математике, статистике, экономике. Одна из таких теорий – теория массового обслуживания, лежащая на стыке указанных предметных областей.

Массовое обслуживание возникает в тех случаях, когда в некоторой системе имеется входящий поток заявок на получение каких-либо услуг и имеется аппарат отработки этих заявок с ограниченными возможностями по одновременному количеству обслуживаемых заявок. Модели теории массового обслуживания позволяют принимать решения по организации систем обслуживания, необходимой пропускной способности и скорости обработки одной заявки.

6.1. Система массового обслуживания с отказами

Пример 6.1

Дано: система массового обслуживания представляет собой одноканальную телефонную линию. Система работает с отказами, так как вызов, поступающий в момент, когда линия занята, не обрабатывается, то есть получает отказ. Интенсивность потока $\lambda = 0{,}2$ вызова в минуту. Средняя продолжительность разговора $t = 2$ мин.

Требуется: определить вероятность отказа, а также то, какое количество каналов должно работать параллельно, чтобы вероятность отказа была не более 2 %?

Решение

Определим исходные данные. Интенсивность обслуживания равняется $\mu = 1 / t = 1 / 2 = 0{,}5$ вызова в минуту. При этом система может находиться в двух возможных состояниях: 1) телефонная линия свободна; 2) телефонная линия занята. Входящий вызов переводит систему из первого состояния во второе с интенсивностью $\lambda = 0{,}2$, обратно система возвращается с интенсивностью $\mu = 0{,}5$. Интенсивность нагрузки в системе выражается как:

$$\rho = \frac{\lambda}{\mu} \qquad (6.1)$$

$$\rho = \frac{0,2}{0,5} \approx 0,4$$

– интенсивность нагрузки в разбираемом примере.

Теперь определим вероятности нахождения системы в каждом из двух состояний. Для одноканальной системы с отказами они вычисляются следующим образом:

p_0 – вероятность того, что система свободна (телефонная линия свободна, заявок нет):

$$p_0 = \frac{\mu}{\mu + \lambda}, \qquad (6.2)$$

$$p_0 = \frac{0,5}{0,5 + 0,2} \approx 0,71$$

p_1 – вероятность того, что заявка в системе (телефонная линия занята). Этот же показатель – вероятность отказа в обслуживании, ведь канал всего один:

$$p_1 = \frac{\lambda}{\mu + \lambda} \qquad (6.3)$$

$$p_1 = \frac{0,2}{0,5 + 0,2} \approx 0,29$$

Как видим, при одноканальной линии вероятность отказа равна 29 %, а по условию требуется не более 2 %.

Определим, сколько телефонов должно работать параллельно, чтобы вероятность отказа вышла на проектный уровень, то есть подберем такое количество каналов обслуживания k, при котором $p_1 \leq 0,02$.

Для многоканальной системы с отказами многоканальной системы массового обслуживания с отказами формула вычисления вероятности свободного состояния системы (хотя бы один канал свободен) следующая:

$$p_0 = \left(\sum_{i=0}^{k} \frac{\rho^i}{i!} \right)^{-1}, \qquad (6.4)$$

где k – это количество каналов обслуживания.

Вероятность того, что система с k каналами занята (нет ни одного свободного канала):

$$p_k = p_0 * \frac{\rho^k}{k!}, \qquad (6.5)$$

Теперь предположим, что $k = 2$, тогда пересчитаем вероятность нахождения системы в свободном состоянии по формуле (6.4):

$$p_0 = 1/\left(\frac{0{,}4^0}{0!} + \frac{0{,}4^1}{1!} + \frac{0{,}4^2}{2!} \right) \approx 0{,}68,$$

Вероятность того, что оба канала заняты, равна:

$$p_1 = p_0 * \frac{\rho^i}{i!} = 0{,}68 * \frac{0{,}4^2}{2!} \approx 0{,}05$$

Вероятность отказа при двухканальной системе равняется 5 %, что больше, чем требовалось. Значит, двух каналов недостаточно для того, чтобы вероятность отказа была не более 2 %. Попробуем предположить, что $k = 3$, и подсчитать те же показатели.

$$p_0 = \left(\sum_{i=0}^{k} \frac{\rho^i}{i!} \right)^{-1} = \left(\frac{0{,}4^0}{0!} + \frac{0{,}4^1}{1!} + \frac{0{,}4^2}{2!} + \frac{0{,}4^3}{3!} \right)^{-1} \approx 0{,}67$$

$$p_3 = 0{,}61 * \frac{0{,}4^3}{3!} \approx 0{,}007$$

Таким образом, вероятность отказа при трех каналах равна 0,007, что меньше 2 %. Значит трех каналов достаточно.

Задача 6.1

Дано: система массового обслуживания представляет собой трехканальную телефонную линию. Система работает с отказами, так как вызов, поступающий в момент, когда три линии заняты, не обрабатывается. Интенсивность потока $\lambda = 0{,}95$ вызова в минуту. Средняя продолжительность разговора $t = 1$ мин. Требуется: определить вероятность отказа.

Задача 6.2

Дано: система массового обслуживания представляет собой двухканальную телефонную линию. Система работает с отказами, так как вызов, поступающий в момент, когда обе линии заняты, не обрабатывается. Интенсивность потока $\lambda = 0{,}75$ вызова в минуту. Средняя продолжительность разговора $t = 2$ мин. Требуется: определить, какое количество каналов должно работать параллельно, чтобы вероятность отказа была не более 5 %?

6.2. Одноканальная система массового обслуживания с очередью

В предыдущем примере система просто сбрасывала поступающий вызов, если он поступал при ее занятом состоянии. Во многих реальных системах обслуживания заявки не сбрасываются, а образуют очередь. В этом случае управленческие решения могут быть направлены на управление параметрами очереди путем установления определенных параметров обслуживания.

Пример 6.2

Дано: в организации имеется окно выдачи справок. Среднее время обработки запроса (подготовки справки) составляет 7 мин. Среднее количество поступающих заявок составляет 9 человек в час.

Требуется: определить, справляется ли система с обработкой заявок (не образуется ли бесконечная очередь), если не справляется, подобрать время обслуживания заявки так, чтобы отладить работу системы, после чего рассчитать параметры очереди.

Решение

Вновь, как и ранее выясним интенсивность входящего потока $\lambda = 9/60 = 0{,}15$ человек в минуту, а также интенсивность обработки запросов $\mu = 1/7 \approx 0{,}14$. Интенсивность нагрузки в системе массового обслуживания вычислим по формуле (10):

$$\rho = \frac{0{,}15}{0{,}14} = 1{,}07$$

Поскольку $\rho \geq 1$, то система не справляется с поступающими заявками, образуя бесконечную очередь. Чтобы этого не происходило, нужно уменьшить время обслуживания одной заявки и обеспечить $\rho < 1$. Найдем μ^* – достаточную интенсивность обслуживания заявок:

$$\rho = \frac{0{,}15}{\mu^*} < 1 \ \rightarrow \ \mu^* > 0{,}15,$$

здесь 0,15 – это количество заявок, обрабатываемых в минуту. В час необходимо обрабатывать более 0,15∗60 = 9 заявок, значит на одну заявку необходимо отводить не более, чем 60/9 ≈ 6,6 (6) мин.

Установим время обслуживания одной заявки 6 мин.

Тогда в случае одноканальной системы обслуживания параметры очереди будут описываться нижеследующими выражениями.

Вероятность того, что система свободна:

$$p_0 = 1 - \rho, \qquad (6.6)$$

Вероятность того, что в системе находится *m* заявок:

$$p_m = \rho^m * p_0, \qquad (6.7)$$

Очевидно, что при $\rho < 1$ эти вероятности p_0, p_1, $p_2 \ldots p_m \ldots$ образуют убывающую последовательность, и соблюдается условие:

$$p_{m-1} > p_m > p_{m+1},$$

значит, более длинная очередь всегда менее вероятна, чем той же длины, что уже есть или более короткая. Наиболее вероятное состояние системы – свободное.

Среднее количество заявок в системе составит:

$$L_{\text{сист.}} = \frac{\rho}{1-\rho} \qquad (6.8)$$

Среднее количество заявок в очереди:

$$L_{\text{оч.}} = L_{\text{сист.}} - L_{\text{об.}},$$

где $L_{\text{об.}}$ – количество заявок в обработке, равное вероятности, что канал занят 1- p_0.

Тогда количество заявок в очереди можно рассчитать согласно выражению:

$$L_{\text{оч.}} = L_{\text{сист.}} - L_{\text{об.}} = \frac{\rho}{1-\rho} - \rho = \frac{\rho}{1-\rho} - \frac{\rho-\rho^2}{1-\rho} = \frac{\rho^2}{1-\rho}$$

$$L_{\text{оч.}} = \frac{\rho^2}{1-\rho} \qquad (6.9)$$

Среднее время пребывания заявки в системе:

$$T_{\text{сист.}} = \frac{1}{\lambda} * L_{\text{сист.}} = \frac{\rho}{\lambda * (1-\rho)} \qquad (6.10)$$

Среднее время пребывания заявки в очереди:

$$T_{\text{оч.}} = \frac{1}{\lambda} * L_{\text{оч.}} = \frac{\rho^2}{\lambda * (1-\rho)} \qquad (6.11)$$

В рассматриваемом примере параметры системы и очереди следующие. Интенсивность нагрузки в системе:

$$\rho = \frac{0{,}15}{1/6} = 0{,}9,$$

вероятность того, что система свободна:

$$p_0 = 1 - \rho = 1 - 0{,}9 = 0{,}1,$$

вероятность того, что в системе одна заявка:

$$p_1 = \rho^1 * p_0 = 0{,}9^1 * 0{,}1 = 0{,}09,$$

вероятность того, что в системе две заявки:

$$p_2 = \rho^2 * p_0 = 0{,}9^2 * 0{,}1 = 0{,}081,$$

вероятность того, что в системе m заявок:

$$p_m = \rho^m * p_0 = 0{,}9^m * 0{,}1,$$

среднее количество заявок в системе:

$$L_{\text{сист.}} = \frac{\rho}{1-\rho} = \frac{0{,}9}{1-0{,}9} = 9,$$

среднее количество заявок в очереди:

$$L_{\text{оч.}} = \frac{\rho^2}{1-\rho} = \frac{0{,}9^2}{1-0{,}9} = 8{,}1,$$

среднее время пребывания заявки в системе:

$$T_{\text{сист.}} = \frac{\rho}{\lambda * (1-\rho)} = \frac{0{,}9}{0{,}15 * (1-0{,}9)} = 60,$$

среднее время пребывания заявки в очереди:

$$T_{\text{оч.}} = \frac{\rho^2}{\lambda * (1-\rho)} = \frac{0{,}9^2}{0{,}15 * (1-0{,}9)} = 54$$

Задача 6.3

Требуется: на основе данных из предыдущего примера определить, каким образом необходимо изменить время обслуживания одной заявки, чтобы среднее время пребывания заявителя в организации не превышало 15 минут.

Задача 6.4

Дано: в пункт техосмотра поступают заявки (прибываю автомобили), в среднем 2 в час. На осмотр одного автомобиля уходит 20 минут.

Требуется: рассчитать среднюю длину очереди и среднее время пребывания в ней.

6.3. Многоканальная система массового обслуживания с очередью

Если в системе массового обслуживания k каналов, то приведенные выше формулы изменяются следующим образом. Вероятность того, что система свободна:

$$p_0 = \left(\sum_{i=0}^{k} \frac{\rho^i}{i!} + \frac{\rho^{k+1}}{k!\,(k-\rho)} \right)^{-1}, \qquad (6.12)$$

вероятность того, что в системе m заявок:

$$p_m = p_0 * \frac{\rho^m}{m!}, \qquad (6.13)$$

вероятность того, что новая заявка попадет не на обслуживание, а в очередь:

$$p_{\text{оч.}} = p_0 * \frac{\rho^{k+1}}{k!\,(k-\rho)}, \tag{6.14}$$

среднее количество занятых каналов:

$$\bar{k} = \rho, \tag{6.15}$$

среднее число заявок в очереди:

$$L_{\text{оч.}} = \frac{\rho^{k+1} * p_0}{k * k!\,(k-\frac{\rho}{k})^2}, \tag{6.16}$$

среднее количество заявок в системе:

$$L_{\text{сист.}} = L_{\text{оч.}} + \rho \tag{6.17}$$

Среднее время пребывания заявки в системе и в очереди рассчитывается, как и ранее по 6.10 и 6.11 соответственно.

Задача 6.5

Дано: в организации имеется окно выдачи справок. Среднее время обработки запроса (подготовки справки) составляет 15 мин. Среднее количество поступающих заявок составляет 3 в час.

Требуется: рассчитать параметры системы, определить вероятность нахождения в очереди более 5 человек.

6.4. Ошибки первого и второго рода

Следующие несколько задач к теории массового обслуживания напрямую не относятся, однако связаны с возникновением ошибок в системах массового обслуживания, где происходит идентификация объектов.

Например, в правоохранительной деятельности существует множество задач, связанных с идентификацией различных объектов (лиц, предметов вооружения, транспортных средств, следов и т. п.). Эффективность решения этих задач описывается статистически. Основные характеристики, общие для всех систем идентификации, – это частота возникновения ошибок первого рода (FAR –

FalseAcceptanceRate, ошибка ложного отождествления) и второго рода (FRR – FalseRejectRate, ошибка ложного неотождествления).

Параметры FAR и FRR рассчитываются так:

$$FAR = \frac{(\text{кол–во ошибок 1 рода})}{(\text{кол–во проверок})} \qquad (6.18)$$

$$FRR = \frac{(\text{кол–во ошибок 2 рода})}{(\text{кол–во проверок})} \qquad (6.19)$$

Ниже рассмотрим управленческую задачу о выборе системы управления контролем доступа для подразделения.

Пример 6.3

Дано: пусть в подразделении работает N = 600 сотрудников, биометрические параметры каждого из которых занесены в базу данных. Система контроля доступа при проходе в здание каждого из сотрудников проводит его отождествление по этой базе данных.

Требуется: определить параметр FAR (частота ошибки ложного отождествления), которым должна обладать система контроля управления доступом, если ложные отождествления недопустимы.

Решение

В худшем случае система сравнит предъявленные каждым отдельным сотрудником биометрические параметры с каждой из 600 записей в базе. Таким образом, максимально возможное количество сравнений при приходе на работу всех 600 сотрудников равняется N2 = 360 000. Если при этом допускается одно ложное отождествление, то параметр FAR будет равен:

FAR = 1/360000 ≈ 0,0000028 = 0,00028 %.

Соответственно, если не допускается ни одного ложного отождествления, то FAR устанавливаемой системы должна быть меньше, чем 0,00028 %.

Задача 6.6

Дано: на режимном объекте работают 1 500 человек. Перед руководством стоит задача закупки системы идентификации, которая будет использоваться для управления контролем доступа на объект.

Требуется: определить, какой частотой ошибки первого рода должна обладать система идентификации прибывающих на объект сотрудников.

Задача 6.7

Дано: разработана система идентификации для контроля управления доступом в здание. Ее частота ошибок первого рода FAR = 0,004 %.

Требуется: определить, какое максимальное количество сотрудников допустимо на объекте, где будет эксплуатироваться эта система.

Задача 6.8

Дано: экспертное подразделение осуществляет деятельность по идентификации лиц на основе биометрических данных. Решение об идентичности предъявленных данных и данных, содержащихся в базе данных, принимает эксперт. В подразделение закупается новая система идентификации для поддержки принятия решений экспертом, которая при объеме базы данных в 700 000 записей давала бы не более 40 ошибок первого рода при идентификации одного человека.

Требуется: определить допустимую частоту ошибок первого рода для закупаемой системы.

Задача 6.9

Дано: у организации имеются два режимных объекта. На первом ежедневно работают 450 человек, на втором – 130 человек. К системе контроля доступа предъявляется требование о том, что в течение рабочего дня не допустимо ложных отождествлений.

Рассматривается несколько альтернативных систем, параметры которых приведены в таблице:

Таблица 6.1

№	Название	FAR	FRR	Стоимость, у.е.	Стоимость эксплуатации в год, у.е.
1	Юпитер	0,0002	0,00001	7 000 000	100 000
2	Нептун	0,00004	0,0003	2 500 000	120 000
3	Орион	0,00009	0,00001	5 300 000	50 000
4	Сатурн	0,000003	0,007	3 900 000	150 000
5	Марс	0,000007	0,0004	6 700 000	90 000
6	Меркурий	0,0000045	0,001	5 000 000	90 000
7	Плутон	0,000004	0,002	4 600 000	95 000

Требуется: выбрать наиболее подходящую систему контроля доступа для первого объекта и отдельно для второго объекта. Выбор должен быть аргументированным.

7. Принятие решений в условиях риска и неопределенности

На практике встречаются задачи принятия решения, когда ЛПР вынуждено принимать решения, не обладая полнотой информации об условиях, в которых решение будет реализовано, при том что эти условия оказывают влияние на достигаемый результат.

Такие задачи относятся либо к классу принятия решений в условиях риска, либо к классу принятия решений в условиях неопределенности.

В задачах принятия решения в условиях риска и неопределенности учитывается внешняя среда (природа), состояния которой считаются случайными, поэтому задачи такого типа иногда называют играми с природой.

7.1. Принятие решений в условиях риска

Введем следующие обозначения:

d_i, $i \in [1; m]$ – возможные варианты решения;

s_j, $j \in [1; n]$ – возможные варианты условий внешней среды;

p_j – вероятность реализации условий sj;

u_{ij} – полезность решения di, реализованного при условиях внешней среды s_j.

Обладая этими данными, можно представить их в виде матрицы полезности:

Таблица 7.1

	$\mathbf{s_1}$	$\mathbf{s_2}$	…	$\mathbf{s_n}$
	$\mathbf{p_1}$	$\mathbf{p_2}$	…	$\mathbf{p_n}$
$\mathbf{d_1}$	u_{11}	u_{21}	…	u_{1n}
$\mathbf{d_2}$	u_{21}	u_{22}	…	u_{2n}
…	…	…	…	…
$\mathbf{d_m}$	u_{m1}	u_{m2}	…	u_{mn}

Теперь для каждой альтернативы решения можно рассчитать его ожидаемую полезность:

$$M(u_i) = \sum_{j=1}^{n} p_j * u_{ij} \qquad (7.1)$$

Критерий Байеса-Лапласа

Выбрать решение теперь можно по критерию Байеса-Лапласа – критерию ожидаемой полезности:

$$\max_i M(u_i) = \max_i \sum_{j=1}^{n} p_j * u_{ij}) \qquad (7.2)$$

Выбираем ту альтернативу решения, которая обладает наибольшей ожидаемой полезностью.

В случае, когда решения различаются не по достигаемому эффекту, а по затратам v_{ij}, которые придется понести в зависимости от условий среды, то критерий модифицируется:

$$\min_i M(v_i) = \min_i \sum_{j=1}^{n} p_j * v_{ij}) \qquad (7.3)$$

Пример 7.1

Дано: имеются три альтернативные системы информационной безопасности, которые организация может закупить. Также имеются экспертные оценки по возможному ущербу для информационной инфраструктуры организации при различных типах атак на нее и при использовании каждой из систем безопасности. Кроме того, имеется оценка вероятностей осуществления атак различного типа.

Таблица 7.2

	атака 1	**атака 2**	**атака 3**	**атака 4**
	p_1 =0,60	**p_2=0,20**	**p_3=0,15**	**p_4=0,05**
система 1	2	1	6	16
система 2	3	3	6	0
система 3	2	4	8	5

Решение

Рассчитаем ожидаемые ущербы при использовании каждого варианта системы безопасности:

$$M(v_1) = 0{,}6 * 2 + 0{,}2 * 1 + 0{,}15 * 6 + 0{,}05 * 16 = 3{,}1$$

$$M(v_2) = 0{,}6 * 3 + 0{,}2 * 3 + 0{,}15 * 2 + 0{,}05 * 0 = 3{,}3$$

$$M(v_3) = 0{,}6 * 2 + 0{,}2 * 4 + 0{,}15 * 8 + 0{,}05 * 5 = 3{,}45$$

Минимальный ожидаемый ущерб 3,1 ожидается при использовании первой системы безопасности.

Задача 7.1

Дано: выбирается один из пяти альтернативных планов строительства жилого комплекса. Однако не известно качественное состояние экономики на момент завершения работ и соответственно, какую прибыль удастся извлечь от реализации объекта. Известно, что экономическое состояние может ухудшиться, улучшиться или остаться без изменений.

Таблица 7.3

	ухудшение	**без изменений**	**улучшение**
	p1 =0,30	**p2=0,65**	**p3=0,05**
план 1	340	300	270
план 2	280	320	400
план 3	120	190	980
план 4	500	210	200
план 5	280	320	350

Требуется: на основе критерия ожидаемой полезности выбрать план строительства.

Задача 7.2

Дано: выбирается один из трех альтернативных планов размещения стационарных постов охраны на территории. Однако не известно качественное состояние окружающей инфраструктуры, которая

может измениться разными способами. Известно, что рядом с охраняемой территорией может появиться транспортный узел, либо может быть разбит парк, либо ничего из перечисленного. В зависимости от этого придется понести дополнительные затраты на оптимизацию постов охраны.

Таблица 7.4

	транспортный узел	**без изменений**	**парк**
	p1 =0,35	**p2=0,25**	**p3=0,40**
план 1	240	0	20
план 2	180	10	100
план 3	220	20	80

Требуется: на основе критерия Байеса-Лапласа выбрать план строительства постов охраны на территории.

Критерий максимальной устойчивости

Но что, если рассматриваемые альтернативы имеют равные или очень близкие по значению ожидаемые полезности?

В этом случае после отбора нескольких подходящих альтернатив на основе критерия Байеса-Ла пласа можно дополнительно применить критерий максимальной устойчивости и найти наименее дисперсионную по ожидаемой полезности альтернативу, дающую наиболее устойчивое решение в любых условиях внешней среды.

Дисперсия полезности альтернативы *i* определяется по формуле:

$$D(v_i) = \sum_{j=1}^{n} p_j * (v_{ij} - M(v_i))^2, \qquad (7.4)$$

где n – количество вариантов условий внешней среды.

Критерий максимальной устойчивости (минимальной дисперсии полезности):

$$\min_i D(v_i) = \min_i \sum_{j=1}^{n} p_j * (v_{ij} - M(v_i))^2, \qquad (7.5)$$

Пример 7.2

Имеются альтернативы управленческих решений d_1, d_2, d_3. Имеются варианты условий реализации данных решений s_1, s_2, s_3. Построена матрица полезности, каждая ячейка которой содержит значение выигрыша u_{ij}.

Таблица 7.5

	s1	**s2**	**s3**
	p1 = 0,16	**p2 = 0,5**	**p3 = 0,34**
d1	50	12	24
d2	89	49	14
d3	27	30	71

На основе критерия Байеса-Лапласа получим ожидаемые полезности для трех альтернативных решений:

$$M(u_1) = 0{,}16 * 50 + 0{,}5 * 12 + 0{,}34 * 24 = 22{,}16$$

$$M(u_2) = 0{,}16 * 89 + 0{,}5 * 49 + 0{,}34 * 14 = 43{,}5$$

$$M(u_3) = 0{,}16 * 27 + 0{,}5 * 30 + 0{,}34 * 71 = 43{,}46$$

Как видим, два решения – второе и третье – чрезвычайно близки по ожидаемой полезности.

Найдем наиболее устойчивое из них:

$$D(u_2) = 0{,}16 * (89 - 43{,}5)^2 + 0{,}5 * (49 - 43{,}5)^2 + 0{,}34 * (14 - 43{,}5)^2 = 642{,}25$$

$$D(u_3) = 0{,}16 * (27 - 43{,}46)^2 + 0{,}5 * (30 - 43{,}46)^2 + 0{,}34 * (71 - 43{,}46)^2 \approx 391{,}81$$

Третье решение показывает дисперсию меньше, чем второе, а значит более устойчиво.

Задача 7.3

Дано: имеются три альтернативные системы вооружения. По каждой из них имеются экспертные оценки эффективности в зависимости от климатических условий. Эксперты выделили 4 типа климатических условий и вероятности использования закупаемой системы вооружения в этих условиях.

Таблица 7.6

	к. усл. 1	к. усл. 2	к. усл. 3	к. усл. 4
	p_1 =0,60	p_2=0,20	p_3=0,15	p_4=0,05
вооруж. 1	2	1	7	16
вооруж. 2	3	3	6	0
вооруж. 3	2	4	8	5

Можете убедиться, что критерий Байеса-Лапласа указывает на близость ожидаемых полезностей трех систем вооружений.

Требуется: на основе критерия максимальной устойчивости найти лучшую альтернативу.

7.2. Принятие решений в условиях неопределенности

Принятие решений в условиях неопределенности характеризуется тем, что вероятности наступления тех или иных состояний внешней среды неизвестны. В распоряжении ЛПР по-прежнему имеются данные о том, какими могут быть эти состояния, и какую полезность принесет то или иное решение в этих состояниях внешней среды.

Критерий крайнего оптимизма

Самый простой критерий, опираясь на который ЛПР рассчитывает на наиболее благоприятное по отношению к выбранному решению состояние внешней среды. Проще говоря, выбирается то решение, которое содержит при одном из вариантов состояния среды наибольшее значение полезности среди всех прочих:

$$\max_i \max_j v_{ij} \qquad (7.6)$$

Пример 7.3

Таблица 7.7

	s1	s2	s3
d1	0	15	0
d2	13	14	7
d3	11	8	12

Требуется на основе имеющихся данных и критерия крайнего оптимизма выбрать решение.

Решение

В рассматриваемом примере наилучшая возможная полезность у первого решения – d_1, реализуемого в состоянии среды s_2.

Очевиден и недостаток этого подхода – легко заметить, что решение d_1 дает наименее устойчивую полезность из всех имеющихся.

Критерий Вальда (крайнего пессимизма)

В противоположность критерию крайнего оптимизма, критерий Вальда или крайнего пессимизма предполагает сначала нахождение наихудших исходов для каждой альтернативы, а затем выбор из этих наихудших вариантов самого лучшего:

$$\max_i \min_j v_{ij} \tag{7.6}$$

Пример 7.4

На тех же данных, что и в предыдущем примере выберем лучшее решение по критерию Вальда.

Решение

Минимально возможные исходы при выборе решения 1 следующие:
первое решение: 0;
второе решение: 7;
третье решение: 8.

Среди наихудших исходов третье решение дает наилучший, его и следует выбрать по критерию Вальда.

Критерий Гурвица (оптимизма – пессимизма)

Критерий пессимизма – оптимизма Гурвица при выборе альтернативы предполагает, что она будет реализовываться либо при самом благоприятном состоянии среды, либо при самом неблагоприятном, причем важность первого α и второго $(1-\alpha)$ задается экспертно:

$$\max_i(\alpha * \arg\max_j v_{ij} - (1-\alpha)\arg\min_j v_{ij}) \tag{7.7}$$

Пример 7.5

На тех же данных, что и в предыдущих примерах (таблица 7.7) выберем лучшее решение по критерию Гурвица с коэффициентом $\alpha = 0{,}5$ (одинаково значимы, и максимально, и минимально возможные полезности).

Решение

Разность максимальной и минимальной полезностей с коэффициентами α и (1 - α) для каждого решения следующая:

первое решение: 0,5*15+0,5*0 = 7,5;

второе решение: 0,5*14 + 0,5*7 = 10,5;

третье решение: 0,5*12 + 0,5*8 = 10.

Наилучшее решение по критерию Гурвица – второе.

Критерий Сэвиджа (наименьших сожалений)

Согласно критерию Сэвиджа, при выборе решения важна упущенная выгода, поэтому для определения лучшего решения сначала строится матрица упущенной выгоды (матрица сожалений). Для каждого из альтернативных решений в каждом варианте среды рассчитывается разность с наилучшей полезностью, которая могла бы быть получена в этом варианте среды.

$$v_{ij}^{*}$$

Каждый элемент матрицы сожалений находится как разность полезности решения i в состоянии среды j и максимально возможной полезностью в этом состоянии среды.

Формальная запись такая:

$$v_{ij}^{*} = v_{ij} - \underset{i}{argmax}\, v_{ij} \qquad (7.8)$$

Для данных из таблицы 7.7 эта матрица примет вид:

Таблица 7.8

	s1	**s2**	**s3**
d1	13 - 0	15 - 15	12 – 0
d2	13 – 13	15 – 14	12 – 7
d3	13 – 11	15 – 8	12 – 12

Наилучшее решение найдем по наименьшей сумме сожалений, наступающих во всех возможных состояниях среды:

$$\min_i \sum_{j=1}^{n} v_{ij}^* \qquad (7.9)$$

Пример 7.5

На тех же данных, что и в предыдущих примерах выберем лучшее решение по критерию Сэвиджа.

Решение

Суммарные сожаления для каждого решения следующие:

первое решение: 13 + 0 + 12 = 25;

второе решение: 0 + 1 + 5 = 6;

третье решение: 2 + 7 + 0 = 9.

Наилучшее решение по критерию Сэвиджа – второе.

Задача 7.4

Имеются данные о полезности 4-х управленческих решений, реализуемых в 4-х возможных состояниях внешней среды. Вероятности наступления того или иного состояния среды неизвестны:

Таблица 7.9

	s1	**s2**	**s3**	
d1	4	6	2	3
d2	1.1	1.2	3	7
d3	3	6.5	0	7,5
d4	2.5	2.5	3	4

Требуется:

выбрать наилучшее решение, руководствуясь:

а) критерием крайнего оптимизма;

б) критерием Вальда;

в) критерием Гурвица (с коэффициентом $\alpha = 0{,}4$);

г) критерием Сэвиджа.

Список рекомендованной литературы:

Васильева Э. К., Лялин В. С. Статистика: учебник. М.: Юнити, 2012. 399 с.

Гнеденко Б. В., Коваленко И. Н. Введение в теорию массового обслуживания. М.: КомКнига, 2005. 400 с.

Использование многомерных статистических методов анализа и прогнозирования результатов деятельности ОВД на региональном уровне: методические рекомендации / [Горошко И. В. и др.]. М.: Академия управления МВД России, 2015. 44 с.

Кравченко Ю. А. Работа полицейского в MS Excel 2013: практическое пособие. М.: ИнтерКрим-пресс, 2016. 320 с.

Кремер Н. Ш. Теория вероятностей и математическая статистика: учебник. М.: ЮНИТИ-ДАНА, 2001. 551 с.

Орлов А. И. Теория принятия решений: учебное пособие. М.: Март, 2004. 656 с.

Торопов Б. А., Апульцин В. А. Технологии многокритериального оценивания результатов деятельности территориальных органов МВД России на региональном уровне: учебное пособие. М.: Академия управления МВД России, 2016. 112 с.

Шашков В. Б. Прикладной регрессионный анализ. Многофакторная регрессия: учебное пособие. Оренбург: ГОУ ВПО ОГУ, 2003. 363 с.

Учебное издание

Торопов Б. А.,
Гонов Ш. Х.

Статистические методы
принятия управленческих решений

Сборник задач
(задачник)

Редактор: *А. А. Уварова*
Верстка: *С. Х. Аминов*

Подписано в печать 12.08.2019. Формат $60 \times 84\ ^{1}/_{16}$.
Уч.-изд. л. 2,00. Усл. печ. л. 4,42. Тираж 66 экз. Заказ № _____
Отделение полиграфической и оперативной печати РИО
Академии управления МВД России.
125993, Москва, ул. Зои и Александра Космодемьянских, д. 8

ISBN 978-5-906942-84-5