Министерство сельского хозяйства Российской Федерации
ФГОУ ВПО РГАУ–МСХА имени К.А. Тимирязева
Кафедра экономической кибернетики

Н.М. Светлов

ПРАКТИКУМ
по теории систем и системному анализу

для студентов бакалавриата по направлениям
«Прикладная информатика в экономике» и
«Математические методы в экономике»

МОСКВА 2009

С-27 Светлов Н.М. Практикум по теории систем и системному анализу для студентов бакалавриата по направлениям «Прикладная информатика в экономике» и «Математические методы в экономике» / Издательство ФГОУ ВПО РГАУ–МСХА имени К.А. Тимирязева. М., 2009. – 75 с.

Рецензенты: профессор **Е.В. Худякова** (МГАУ имени В.П. Горячкина);
профессор **А.А. Землянский** (РГАУ-МСХА имени К.А. Тимирязева).

Рекомендовано к изданию методической комиссией экономического факультета РГАУ-МСХА имени К.А. Тимирязева.
Протокол №__ от __ _____ 2008 г.
Председатель методической комиссии профессор **Ф.К. Шакиров**.

ВВЕДЕНИЕ

Данный практикум рекомендуется в качестве руководства для выполнения лабораторных работ по курсу «Теория систем и системный анализ» студентами, проходящими обучение в образовательных учреждениях высшего профессионального образования по направлениям 351400 – прикладная информатика в экономике и 061800 – математические методы в экономике. Он разработан с учётом действующих государственных образовательных стандартов высшего профессионального образования по данным направлениям.

Лабораторные работы, вошедшие в состав практикума, основаны на сквозной задаче, ежегодно решавшейся студентами в течение 1993...2007 гг. В течение этого периода задание совершенствовалось с целью повышения эффективности использования учебного времени и степени усвоения материала. Накопленный в течение 15 лет опыт нашёл отражение в данном практикуме. В данном издании цикл лабораторных работ дополнен рядом новых элементов:

- существенно переработана и пополнена теоретическая часть практикума с учётом имеющихся различий в степени освоения отдельных дисциплин (прежде всего статистики и математики) студентами различных специальностей, относящихся к вышеуказанным направлениям;
- пересмотрен набор вариантов заданий с ориентацией на системный анализ аграрных производственных систем национального уровня, что обеспечивает применимость практикума для решения более широких педагогических задач — в частности, для подготовки специалистов для любого уровня управления АПК и сельским хозяйством;
- в качестве рекомендуемой информационной базы практикума используются международные информационные ресурсы, представленные в сети Internet, причём поиск и отбор конкретных данных для анализа студентам предлагается выполнять самостоятельно;
- уточнены объём, содержание и методика выполнения ряда лабораторных работ;
- списки рекомендуемой литературы полностью обновлены и дополнены источниками на английском языке по тем вопросам, которые не нашли достаточного отражения в отечественных и переводных изданиях.

Особенностью настоящего практикума является то, что задания ориентированы на коллективное выполнение рабочими группами

студентов. Это, во-первых, позволяет решать учебную задачу той степени сложности, при которой удаётся сохранить содержательность предметной области в сочетании с необходимой степенью разнообразия используемых аналитических процедур, приёмов и методик. Во-вторых, в ходе выполнения заданий формируются начальные навыки координации и компетенции, необходимые для командного стиля работы.

Приступая к выполнению заданий лабораторного практикума, студент обязан внимательно изучить раздел «Постановка задачи» и при необходимости получить консультации у преподавателя по всем возникшим вопросам.

Постановка задачи и каждая изучаемая тема снабжены теоретическим материалом, минимально необходимым для понимания задания и его выполнения. Его наличие не освобождает от необходимости обращения к лекционному материалу, рекомендуемой литературе и ресурсам сети интернет для вовлечения в процесс решения учебной задачи самых современных методических подходов, адекватных специфике анализируемой системы.

Практическая часть каждой темы содержит формулировку цели работы, перечень необходимых приборов и материалов, задание для самостоятельного выполнения, методические указания по его выполнению, включающие рекомендации технического и организационного плана, облегчающие и ускоряющие выполнение работ, и перечень требований к отчёту, обязательных для выполнения. Отчёт принимается преподавателем только в печатном виде на листах формата А4 или А5, аккуратно оформленным. Небрежность в оформлении отчёта (включая ошибки компьютерного редактирования, непоименованные показатели, пропущенные единицы изменения, неправильные названия рисунков и таблиц) является достаточным основанием для повторного выполнения задания с самого начала по новому варианту во внеучебное время. Перед сдачей отчёта каждый участник рабочей группы ***обязан внимательно прочитать*** отчёты (или индивидуальные разделы коллективного отчёта) всех своих товарищей по группе, указать им на замеченные ошибки, неточности и опечатки в отчёте и проконтролировать их исправление.

Отзывы, замечания и предложения по совершенствованию практикума просьба направлять автору по адресу электронной почты svetlov@timacad.ru.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПРЕПОДАВАТЕЛЮ

Предлагаемый практикум рассчитан на 45 академических часов аудиторного времени для выполнения лабораторных работ и 15 академических часов самостоятельной работы на освоение теоретического материала, необходимого для их выполнения. Если рабочей программой курса «Теория систем и системный анализ» предусмотрен меньший объём лабораторного практикума либо если данный практикум сочетается с другими лабораторными или практическими заданиями, можно осуществить системный анализ одноуровневой системы. В этом случае выполняются задания, относящиеся к темам 1...4. Кроме того, не является обязательной тема 1: преподаватель может регламентировать не только выходную, но и множество входных переменных.

Цель практикума — освоение комплекса методических подходов к системному анализу производственных систем в условиях ограниченности априорных знаний об их структуре (на примере аграрно-промышленного комплекса).

Используемый для организации лабораторного практикума комплекс методических подходов обладает следующими характерными чертами:

♦ в прагматическом плане — пригодность к исследованию систем, плохо поддающихся структуризации, слабо изученных, при условии, что их переменные поддаются наблюдению (количество доступных наблюдений может быть ограниченным);

♦ в педагогическом плане — иллюстрация специфики предметных областей, требующих применения метода системного анализа, и сущности самого системного анализа как достаточно общего метода, предполагающего выбор и использование более конкретных методов исследования для решения частных задач.

Вследствие естественных условий проведения учебного практикума требования к детальности структуризации системы и информационной базе, предусматриваемыми заданиями к лабораторным работам, сообразуются с ограничениями, налагаемыми учебным процессом. Это приводит к ограниченной достоверности получаемого решения. Студенты должны иметь ясное представление об учебном характере задачи и о том, каким образом достигается требуемая достоверность результатов анализа при применении подобных методик для решения задач, имеющих научно-исследовательское или прикладное значение.

Для системного анализа предлагаются системы, структура которых хорошо изучена и известна студентам. Это даёт им возможность соотнести результаты формального подхода к структуризации с известными и проверенными практикой представлениями о структуре данных систем, а также сократить требуемый объём данных, опираясь на собственный опыт в данных предметных областях, накопленный при изучении соответствующих дисциплин и в ходе производственной практики.

Для решения одного варианта практического задания создаётся рабочая группа численностью 4...6 студентов. Как правило, функции каждого члена рабочей группы определяются студентами самостоятельно. Задание принимается только при услови и подготовки отчёта в соответствии с требованиями, приведёнными в практических заданиях по каждой теме. Студенты, не принимающие участие в работе группы либо выполняющие задание несвоевременно, исключаются из состава группы и работают самостоятельно по индивидуальным заданиям. При наличии уважительных причин они по решению деканата могут быть направлены на повторное прохождение курса «Теория систем и системный анализ», в противном случае на итоговой аттестации не могут претендовать на оценку выше удовлетворительной, а в случае, если отдельные задания практикума не выполнены до окончания учебного времени, выделенного на освоение данного курса согласно календарно-тематическому плану — не допускаются к ней.

На титульном листе отчёта указываются наименование темы, номер варианта задания, номер рабочей группы, фамилии и инициалы составителей и дата составления. В целях экономии бумаги допускается замена титульного листа заголовком, содержащим вышеуказанную информацию.

Формат бумаги, используемой для отчёта, — А4 или А5. Размер шрифта основного текста — 12 пунктов, межстрочный интервал — 12 пунктов (минимум). Разрешается двусторонняя печать. Отчёт рекомендуется печатать на принтере ПЭВМ, но допускаются и рукописные отчёты при условии выполнения их разборчивым почерком и без помарок. Страницы должны быть пронумерованы, листы надёжно скреплены или сшиты. Таблицы и рисунки оформляются в соответствии с ГОСТ 2.105-95.

В конце отчёта *обязательно* приводится список литературы, использованной при выполнении практического задания, оформленный в соответствии со стандартом библиографического описания ГОСТ 7.1-2003 (как в библиотечных карточках). В списке литературы не следует указывать настоящие методические указания и неопубликованные источники.

Ссылки на источники в сети Интернет допустимы при условии указания автора или составителя (в том числе коллективного), наименования документа, адреса (URL) и даты доступа. Адреса источников должны быть точными: адресуемый ресурс должен действительно содержать использованную в отчёте информацию (а не ссылки на неё).

При невыполнении требований, сформулированных выше, отчёт не принимается.

Отметка о принятии отчёта с указанием даты ставится преподавателем на титульном листе отчёта или на первой странице.

Оценка выполненного задания осуществляется по пятибалльной системе с учётом следующих факторов:

- степень владения теоретическим материалом;
- трудоёмкость выполнения задания[1];
- личный вклад студента в результат работы группы;
- своевременность выполнения задания;
- оригинальность предложенного решения.

Оценки за выполненные задания по каждой теме рекомендуется использовать в системе рейтинговой оценки знаний студентов по изучаемому курсу. Рекомендуется применять к оценке по каждой теме весовые коэффициенты, пропорциональные количеству часов, выделенных на изучение данной темы (аудиторной и самостоятельной работы в совокупности).

[1] Например, следует учитывать, что трудоёмкость предварительного статистического анализа числовой переменной значительно выше, чем нечисловой. Преподавателю рекомендуется контролировать равномерность распределения учебной нагрузки между студентами в рабочих группах, а при необходимости своевременно предупреждать студентов как о чрезмерности намеченного объёма работ, так и о его недостаточности для отличной (хорошей, удовлетворительной) рейтинговой оценки.

ПОСТАНОВКА ЗАДАЧИ

Теоретическая часть

Представим процесс производства, распределения обмена и (или) потребления, характеризующий аграрную или аграрно-промышленную систему, в форме системы, обладающей структурой ‹**x**, **q(x)**›, где **x** — вектор переменных системы (не обязательно числовых), **q(x)** — вектор отношений, упорядочивающих вектор **x**. Для многих приложений можно предположить, что вектор-функция **q(x)** отображает вектор **x** на вектор действительных чисел, а правило упорядочения представляет собой векторное уравнение **q(x) = 0**.

Предположим далее, что вектор-функция **q(x)** нам не известна, зато имеются данные наблюдений некоторых (возможно, не всех) компонентов вектора **x**, и в их числе того компонента, который характеризует цель управления данной системой.

Задача состоит в том, чтобы аппроксимировать реально существующую структуру ‹**x**, **q(x)**› некоторой другой структурой ‹**y**, **r(y)**›, обладающей следующими свойствами:

- она гомоморфна структуре ‹**x**, **q(x)**›, откуда, в частности, следует существование отношения, отображающего **x** на **y**;
- её можно синтезировать на основе имеющихся данных, пользуясь некоторой формализованной процедурой.

Аппроксимацию нужно выполнить таким образом, чтобы возможно полнее использовать информацию о структуре ‹**x**, **q(x)**›, содержащуюся в матрице **X**, в которой представлены все имеющиеся в распоряжении исследователя результаты наблюдений данной системы.

Если бы имело место следующее:

a) в распоряжении исследователя были сведения, достаточные для обоснованного выбора функциональной формы уравнения **r(y)=0**;

b) данные наблюдений представляли бы собой репрезентативную выборку;

c) компоненты вектора **y** представляли бы нормально распределённые случайные величины;

d) все они, кроме одного, были бы независимы между собой,

тогда можно было бы воспользоваться классическими методами регрессионного анализа.

Если бы выполнялось по крайней мере условие (а), существовала бы возможность воспользоваться специальными методами оценивания параметров корреляционных связей — например, методом максимальной энтропии. При подобных обстоятельствах необходимо, чтобы результат оценивания параметров уравнений регрессии в полном объёме сохранял неопределённость, объективно обусловленную недостаточностью, неполнотой, а подчас и недостоверностью имеющихся данных. Методы данного класса отвечают указанному требованию. Благодаря этому они обеспечивают использование информации, заключённой в теоретической модели исследуемого процесса и в имеющихся наблюдениях, в условиях, когда этой информации недостаточно для применения классических методов.

Но часто случается, что нет никаких оснований для того, чтобы предположить ту или иную функциональную форму. В этом случае постулирование функциональной формы приводит к систематическим ошибкам в принятии управленческих решений, подготавливаемых на основе результата системного анализа — модели ‹**y**, **r(y)**›. Причина в том, что предположение о форме функциональной связи, если только оно случайно не совпало с действительным законом, присущим системе ‹**x**, **q(x)**›, препятствует отражению действительной степени неопределённости исследуемой системы, создавая иллюзию более высокой управляемости исследуемой системы в сравнении с действительностью.

Методика, представленная в практикуме, используется (наряду с другими приёмами системного анализа) для формализации систем, структура которых изучена недостаточно. Она опирается на систему общенаучных и специальных методов, используемых в различных областях знания.

Цель методики — описать структуру исследуемой системы в форме таблиц условных вероятностей реализации возможных состояний её переменных.

Реализация данной методики обычно предполагает следующие этапы:

1. Выбор *выходной* переменной, отражающей полезный эффект функционирования изучаемой системы.

2. Выбор входных переменных, влияющих на выходную переменную.

3. Приведение действительных переменных (если таковые имеются) к дискретной форме.

4. Проверка существенности влияния входных переменных на выходную и взаимной независимости входных переменных.

5. Построение таблиц условных вероятностей и оценка достоверности значений условных вероятностей.

6. При необходимости — рассмотрение некоторых или всех переменных, отобранных на шаге 2, в качестве выходных переменных и выполнение для каждой из них шагов 2...6 данного алгоритма.

7. Проверка работоспособности модели.

Данная методика может применяться при выполнении следующих условий.

- Постановка задачи системного исследования должна включать спецификацию переменной, закон изменения значений которой требуется установить (далее — *выходной переменной*).
- Исследуемая система должна допускать декомпозицию на подсистемы, описываемые единственной выходной и произвольным числом входных переменных.
- Входные переменные каждой подсистемы должны быть взаимно независимыми или степень зависимости между ними должна быть пренебрежимой.
- Обусловленность значения выходной переменной каждой подсистемы значениями входных переменных должна быть достаточно высока, чтобы обеспечить необходимую точность его определения.

На тип переменных никаких ограничений не накладывается: допустимы как числовые, так и нечисловые (в частности, логические) переменные. Примеры переменных: норма внесения удобрений (ц действующего вещества на 1 га пашни), сорт культуры, наличие системы орошения, число полей в севообороте.

Этап 6 выполняется в тех случаях, когда не удаётся установить непосредственное влияние некоторых переменных на выходную переменную (нет соответствующих данных). Тогда, если возможно, изучают их влияние на другие входные переменные, зависимость от которых выходной переменной уже изучена, но которые на практике не могут использоваться для её оценивания[1].

Формализм условных вероятностей, применяемый для представления знаний о связях между переменными исследуемой системы, не требует предположений о форме функциональной связи. Он, в отличие, например, от метода наименьших квадратов, широко используемого для стати-

[1] Например, информация о них поступает лишь тогда, когда выходная переменная уже известна достоверно.

стического оценивания[1] параметров регрессионных зависимостей, не имеет теоретических ограничений по применению в случае малого количества наблюдений, на основании которых можно судить об исследуемых связях. Практические ограничения, связанные со снижением достоверности оценивания параметров связей, сохраняются: о том, достаточно ли достигнутой точности для принятия конкретного управленческого решения, судит лицо, принимающее данное решение.

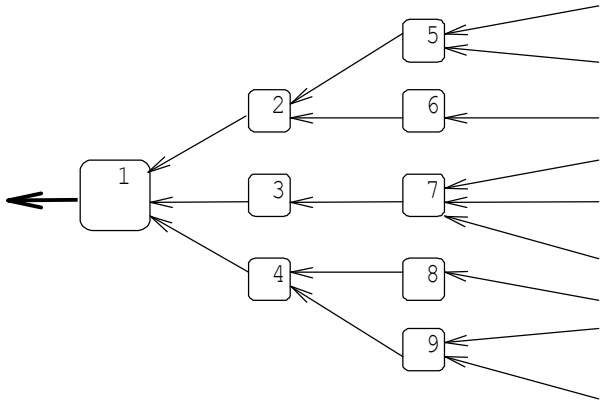


Рис. 1. Представление производственной системы после декомпозиции.

Декомпозиция позволяет представить исследуемую систему в виде дерева, подобного изображённому на рис. 1. Здесь (1) — подсистема первого уровня, (2)...(4) — подсистемы второго уровня, (5)...(9) — третьего. Стрелками обозначены переменные системы, в том числе жирной стрелкой — выходная переменная.

Число входных переменных каждой подсистемы и число уровней иерархии модели определяются:

- доступной информационной базой;
- требуемой точностью предсказания значения выходной переменной на основе информации о значениях входных переменных.

Кроме того, обычно необходимо, чтобы входные переменные терминальных подсистем (то есть подсистем низшего уровня) допускали непосредственное наблюдение либо поддавались управлению со стороны человека. Иначе их невозможно будет использовать для определения значения выходной переменной.

[1] В статистико-математических и эконометрических приложениях следует различать понятия «оценка» (estimate – англ.) — суждение о величине параметра, не поддающегося непосредственному наблюдению, на основе и «оценивание» (estimation– англ.) — процесс получения оценки.

Библиографический список

Городецкий В.И. Байесовский вывод. Л.: ЛИИАН, 1991.

Нейлор К. Как построить свою экспертную систему. М.: Энергоатомиздат, 1991.

Zellner, A. Bayesian analysis in econometrics and statistics. North-Holland publ., 1980.

Zellner, A. An introduction to Bayesian inference in econometrics. Wiley, 1971.

Задание

Описать структуру системы, определяющей значение выходной переменной, указанной в разделе «Варианты заданий для лабораторного практикума», в форме таблиц условных вероятностей. Оценить степень адекватности описания путём тестирования модели и сопоставления его результатов с фактическими данными.

Самостоятельно определить множество входных переменных, принимая во внимание следующие ограничения, обусловленные учебным характером задачи:

- число уровней — 2 (см. этап 6 последовательности реализации методики, с. 10);
- число переменных первого уровня — 4 или 5;
- число переменных в каждой модели второго уровня — 2;
- число моделей второго уровня — не менее 3 (остальные переменные первого уровня предполагаются поддающимися непосредственному наблюдению или управлению);
- число наблюдений, используемых для формулирования моделей первого уровня — от 45 до 60; для формулирования моделей второго уровня — от 20 до 60.

В процессе выполнения лабораторного практикума добиться возможно большей информативности модели по отношению к выходной переменной.

Проделанную работу отразить в письменных отчётах в соответствии с требованиями, сформулированными в практикуме.

Варианты заданий для лабораторного практикума

Наименование выходной переменной

1. Цена кукурузы, произведённой в странах Европы.
2. Производство кукурузы в странах Европы.
3. Потребление молока в странах Европы.
4. Урожайность пшеницы в странах Европы.
5. Производство яблок в странах Европы.
6. Импорт картофеля в страны Европы.
7. Производство хлопковолокна в странах мира.
8. Производство мяса птицы в странах Европы.
9. Поголовье овец в странах Европы.
10. Поголовье овец в странах Азии.
11. Производство куриных яиц в странах Европы.
12. Производство шерсти в странах Азии.
13. Мясная продуктивность свиней в странах Европы.

Примечание. Дополнительные варианты при необходимости могут быть получены выбором другой группы стран.

ТЕМА 1. СПЕЦИФИКАЦИЯ ПЕРВОГО УРОВНЯ АГРАРНОЙ ПРОИЗВОДСТВЕННОЙ СИСТЕМЫ

Теоретическая часть

Приступая к исследованию системы, структура которой неизвестна, прежде всего определяют множество переменных (как количественных, так и нечисловых), которыми можно описать её состояние и поведение. Выделив в их числе выходную переменную — ту, зависимость которой от других переменных необходимо определить для решения тех или иных задач управления, — и задавшись целью представить данную зависимость в форме таблиц условных вероятностей, полезно предварительно определить набор переменных, связь которых с выходной переменной наиболее существенна.

Теоретически для этой цели можно использовать все доступные для наблюдения переменные. Однако на практике такое решение приводит к неприемлемо высоким затратам труда на представление системы в требуемой форме. Поэтому обычно из всего множества доступных для наблюдения переменных отбирают те, которые, по мнению экспертов, накопивших большой опыт наблюдения исследуемой системы, сильнее других влияют на выходную переменную. В дальнейшем мнение экспертов подвергают проверке с помощью формализованных методов, которые будут рассмотрены в теме 3.

В практических приложениях число отобранных таким образом переменных имеет порядок сотен или тысяч. При этом в процессе оценивания выходной переменной участвуют лишь немногие из них, отбираемые на основе формализованных критериев (статистических оценок тесноты связи, показателей относительной информативности и т.д.).

Мнения одного эксперта относительно степени влияния переменных (факторов) на выходную переменную обычно бывает недостаточно. Если система сложна, каждый эксперт, как правило, располагает достаточными сведениями о зависимости выходной переменной лишь от части факторов, с которыми она связана в действительности. Чтобы повысить вероятность адекватного представления исследуемой системы, включающего наиболее существенные факторы, к оценке их значимости привлекают группы экспертов. При этом необходимо заботиться о том, чтобы мнение каждого

эксперта оставалось, по возможности, не зависимым от мнений его коллег. В противном случае внимание экспертов, как показывает практика, сосредоточивается на сравнительно узком круге факторов, и многие существенные переменные ускользают от их внимания.

Для организации коллективных экспертиз предложен ряд специальных методик, содействующих преодолению данной проблемы: метод мозгового штурма, метод Дельфи, метод провокаций, метод решающих матриц и др. В нашем случае целесообразно использовать форму организации коллективной экспертизы, в которой выделяются три этапа:

- ♦ идентификация факторов;
- ♦ согласование мнений экспертов о факторах;
- ♦ ранжирование факторов.

На этапах выявления и ранжирования факторов каждый эксперт работает самостоятельно, чем достигается его независимость от мнения других экспертов. Процедура согласования обладает характерными чертами метода комиссий: она представляет собой открытое обсуждение с целью уточнения смыслового содержания отобранных факторов и характера их влияния на выходную переменную. Последнее необходимо для того, чтобы эксперты, не знакомые со смысловым содержанием отдельных факторов, могли на третьем этапе экспертизы дать оценку их ранга в сравнении с другими факторами.

Для проведения первого этапа каждый эксперт получает задание в течение установленного времени (обычно 15-20 минут) указать (как правило, в письменном виде) как можно больше известных ему факторов, влияющих на заданную целевую переменную. На этом этапе взаимодействие между экспертами должно быть полностью исключено. Факторы могут иметь числовое или нечисловое выражение, но должны характеризоваться единственным значением — в частности, не могут выражаться векторами[1]. Эксперту вменяется в обязанность формулировать факторы таким образом, чтобы из формулировки однозначно вытекала процедура их измерения или оценивания. Тем самым, в частности, подразумевается следующее:

- ♦ каждому фактору, допускающему количественное выражение, должна сопоставляться единица измерения, а также процедура его измерения, если она не очевидна[2];

[1] Фактор, требующий представления в векторной форме, должен рассматриваться как набор факторов, соответствующих каждому компоненту вектора.

[2] Как правило, процедура измерения приводится в форме ссылки на источник, в котором она описана.

- ♦ если фактор отражается нечисловой переменной, эксперт должен однозначно указать множество его значений и процедуру определения конкретного значения данного фактора.

На данном этапе эксперт не должен принимать во внимание доступность фактора для наблюдения и измерения, сопряжённые с этим процессом затраты и другие возможные препятствия его использованию. Его задача состоит лишь в том, чтобы перечислить возможно больше факторов, информация о которых (если доступна) снимает неопределённость выходной переменной.

Цель второго этапа — формирование объединённого списка факторов. На этом этапе должны быть исключены повторяющиеся (возможно, под разными наименованиями) факторы, встречающиеся в индивидуальных списках, и достигнуто единообразное понимание смысла каждого фактора, названного каждым экспертом.

Работа осуществляется по процедуре, схожей с методом комиссий в том отношении, что решения принимаются по результатам открытого обсуждения (как правило, консенсусом). Отличие состоит в отсутствии заранее определённого списка дискутируемых положений: его функцию выполняет объединённый список факторов, названных каждым из экспертов. Комиссия (состоящая из тех же экспертов, которые работали на первом этапе) обладает правами:

- ♦ исключить фактор, названный каким-либо экспертом, только в том случае, если он в точности повторяет по смыслу и по процедуре измерения фактор, названный другим экспертом и уже включённый в объединённый список;
- ♦ уточнять наименование факторов, а также единицы их измерения либо множество их значений.

Продолжительность второго этапа, как правило, не регламентируется.

Неповторяющиеся факторы включаются в объединённый список даже в том случае, если эксперт, предложивший данный фактор, на втором этапе экспертизы отказывается от своего мнения, выраженного на первом этапе. Во избежание непродуктивных дискуссий не разрешается также включение в объединённый список факторов, не названных на первом этапе, но выявленных в процессе работы комиссии.

При необходимости по результатам второго этапа координатор экспертизы может вынести решение о повторении её первого этапа с целью пополнения полученного объединённого списка факторов. Такое решение

принимается в случае, если комиссия в процессе уточнения смысла уже названных факторов выявила отсутствие в результатах работы экспертов целых классов факторов, отражающих существенные аспекты формирования значения целевой переменной. Вновь предложенные факторы *пополняют* ранее полученный объединённый список.

На третьем этапе эксперты получают задание ранжировать объединённый список факторов, выработанный комиссией, ***по предполагаемой степени информативности*** для оценивания значения целевой переменной. Время, выделяемое на ранжирование, как правило, не регламентируется. От эксперта не требуется указание мотивов, по которым он присвоил показателю тот или иной ранг.

Технически этот этап поддерживается программным обеспечением, позволяющим эксперту визуально располагать факторы в определённой последовательности.

Лучше других зарекомендовала себя следующая процедура ранжирования. Вначале каждому фактору присваивается балльная оценка по пятибалльной шкале, отражающая мнение эксперта о его информативности для получения оценки выходной переменной, и производится ранжирование по баллам. Далее процедура повторяется для всех показателей, получивших одинаковый балл на предыдущем этапе, но высшую оценку (пять баллов) получает наиболее информативный, а низшую (один балл) — наименее информативный фактор из числа получивших одинаковый балл на предыдущем этапе. Новая балльная оценка приписывается к предыдущей в качестве разряда десятичной дроби.

Например, если некоторый фактор в группе факторов, оценённых в 4 балла, получил оценку, равную 2 баллам, то ему приписывается оценка 4,2. Когда все показатели получили двузначную балльную оценку, ранжирование всего списка повторяется.

Факторы, снова получившие одинаковый балл, могут быть вновь подвергнуты оценке по тому же принципу, что и выше, и таким образом приобретают трёхзначную оценку (например, 4,25); но если численность факторов, имеющих одинаковый балл, не более 4 или 5, то они могут быть упорядочены между собой непосредственно, без помощи присвоения баллов.

Характерная ошибка, допускаемая экспертами на этом этапе, — размещение целой группы факторов в ранжированном ряду как неделимого целого. Эксперты должны вполне уяснить, что ранг каждого фактора определяется безотносительно к смысловым связям данного фактора с другими

факторами. Он должен отражать только способность фактора информировать о вероятном значении целевой переменной.

Завершается процедура суммированием индивидуальных оценок рангов каждого фактора, полученных каждым экспертом, и повторным ранжированием списка факторов в порядке возрастания полученной суммы. Факторы, сумма номеров рангов которых оказалась одинаковой, упорядочиваются решением комиссии экспертов.

По завершении экспертизы в формализованное описание исследуемой системы включается заранее оговорённое число факторов, получивших наибольший ранг. Это число зависит, с одной стороны, от уровня приемлемых затрат труда и денежных средств на представление исследуемой системы, с другой — от требуемой точности предсказания значения выходной переменной. При этом допускается:

- заменять факторы, не доступные для наблюдения, их аппроксиматорами, если выполняются два условия: аппроксиматор поддаётся наблюдению и не встречается в ранжированном списке факторов;
- пропускать факторы, которые в принципе не поддаются наблюдению в сроки, обусловленные целью исследования системы, с помощью средств, имеющихся в распоряжении исследователей.

Библиографический список

Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. М.: Радио и связь, 1982.

Глушков В.М. О прогнозировании на основе экспертных оценок // Кибернетика, 1969. — №2.

Литвак Б.Г. Экспертные технологии в управлении. М.: Дело, 2004.

Нейлор К. Как построить свою экспертную систему. М.: Энергоатомиздат, 1991.

Практическая часть

Аудиторные занятия: 2 часа.

Цель работы

Овладеть приёмами спецификации входных и выходных переменных исследуемой подсистемы.

Закрепить теоретические знания по вопросам «предмет теории систем» и «методы организации сложных экспертиз».

Приборы и материалы

Компьютерный класс с доступом к сети Internet; программное обеспечение, автоматизирующее рутинные операции по ранжированию факторов[1]; информационный сайт Продовольственной и сельскохозяйственной организации ООН (FAO):
http://faostat.fao.org/DesktopDefault.aspx?PageID=567&lang=ru

Задание

Выполнить предварительную спецификацию входных переменных подсистемы первого уровня производственной системы, исследуемой в соответствии с индивидуальными вариантами задания, приведёнными на с.13 практикума. Для этого:

- составить ранжированный список факторов, влияющих на целевой показатель, соответствующий индивидуальному варианту задания;
- выбрать факторы для включения в модель исследуемой системы с учётом их положения в ранжированном списке, требования их взаимной независимости, имеющейся информационной базы.

Оформить отчёт.

Методические указания по выполнению задания

Задание выполняется коллективно рабочей группой численностью 4...6 чел., сформированной преподавателем.

Ранжированный список факторов составляется в соответствии с методиками, изложенными в теоретической части данной темы.

Учитывая учебный характер задачи, предлагается отобрать 4 (рекомендуется) или 5 входных переменных подсистемы первого уровня.

Каждый член рабочей группы индивидуально составляет список переменных, оказывающих, с его точки зрения, непосредственное влияние на выходную переменную.

Рабочая группа совместно производит объединение индивидуальных списков, устранение повторов, достигает соглашения о точных наименованиях переменных, исключает переменные, не связанные непосредственно с выходной.

[1] Всеми необходимыми возможностями для этого обладают табличные процессоры.

Отмечаются переменные, информация по которым не содержится в материалах производственной практики и не может быть предоставлена преподавателем.

Оставшиеся переменные тем или иным способом ранжируются по степени их влияния на выходную. Из наиболее существенных формируется список переменных для включения в модель. Среди них не должно быть заведомо зависящих друг от друга переменных.

Чтобы исключить возможную неоднозначность толкования переменных, каждая переменная должна иметь название, исчерпывающим образом характеризующее её смысл. Для каждой переменной должна быть указана единица её измерения или (если переменная нечисловая) возможные значения. Рекомендуется указывать источник, из которого можно получить значение переменной (коды документа, строки и столбца).

Требования к отчёту

Отчёт о выполнении практического задания состоит из коллективной и индивидуальных частей. Объём коллективной части не должен превышать 3 страниц[1], каждой индивидуальной — 1 страницы. При необходимости отчёт может быть дополнен приложениями.

<u>В индивидуальной части</u> должны быть представлены:

- краткая характеристика личного вклада студента в работу группы;
- список предложенных составителем переменных, из которых производился отбор входных переменных;
- список использованной литературы.

<u>В коллективной части</u> должны быть представлены:

- ранжированный список переменных, составленный рабочей группой;
- список выбранных входных переменных;
- краткие аргументы в пользу выбранных входных переменных;
- краткое описание использованных подходов к спецификации подсистемы первого уровня, отличающихся от рекомендуемых в методических указаниях (с указанием источника).

[1] Предполагается, что одна страница содержит не более 40 строк по 66 символов.

ТЕМА 2. ПРИВЕДЕНИЕ ЧИСЛОВЫХ ПЕРЕМЕННЫХ К ДИСКРЕТНОЙ ФОРМЕ

Теоретическая часть

Для приведения числовых переменных системы к дискретной форме проводится их статистический анализ, преследующий цели:

- снизить энтропию модели до уровня, обусловленного целями исследования;
- повысить достоверность определения вероятности каждого состояния модели.

Один из приёмов приведения числовых переменных к дискретной форме состоит в разбиении интервала вариации переменной на квантили — интервалы, обладающие тем свойством, что вероятности попадания значения переменной в каждый из них равны. На практике часто выделяют квартили приближённо, пользуясь непосредственно эмпирическими данными. Однако во многих (хотя не во всех) случаях использование теоретического знания о законе распределения исследуемой переменной в дополнение к имеющимся опытным данным (часто ограниченным и не всегда достоверным) позволяет несколько повысить точность разбиения, а значит, и достоверность результатов системного анализа. В этом случае следует:

- определить число наблюдений исследуемой переменной (N).
- разбить интервал вариации переменной на $\sqrt{N}$ аналитических интервалов, определить число наблюдений в каждом аналитическом интервале, выдвинуть гипотезу о характере статистического распределения вариации переменной и проверить её (см. Приложение 2);
- определить число квантилей, учитывая требования снижения энтропии модели и обеспечения достаточной точности её результатов;
- выделить квантили.

Для выделения квантилей используется алгоритм, приведённый ниже.

- Определить вероятность $p = 1 / Q$ того, что значение переменной принадлежит требуемой квантили (Q — число квантилей).
- Определить верхнюю границу x_1 первой квантили из уравнения

$$\int_{a}^{x_1} f(x)dx = p, \qquad (1)$$

где $f(x)$ — функция *плотности распределения* вероятностей значений переменной, a — нижняя граница области определения $f(x)$, x_1 — верхняя граница первой квантили. Если известны значения функции *распределения* вероятностей $F(x)$, то следует решить относительно x_1 уравнение $F(x_1) - F(a) = p$.

- Определить верхнюю границу следующей квантили из уравнения

$$\int_{x_a}^{x_b} f(x)\,dx = p, \qquad (2)$$

где x_a — верхняя граница предыдущей, x_b — искомая верхняя граница данной квантили.

Если определены границы $N-1$-й квантили, перейти следующему шагу; иначе повторить предыдущий.

- Убедиться, что имеет место равенство

$$\int_{x_a}^{\beta} f(x)\,dx = p \qquad (3)$$

(β — верхняя граница области определения $f(x)$). Расхождение, обусловленное ограниченной точностью численных методов, не должно быть слишком большим.

После разбивки интервала вариации на квантили каждое значение переменной заменяется номером квантили, которой оно соответствует. В результате получаем отображение непрерывного множества значений переменной на конечное дискретное множество значений. Это впоследствии обеспечит требуемую уровень грубости (робастности) модели анализируемой системы, обеспечивающую её работоспособность при ограниченной эмпирической базе для её разработки.

Библиографический список

Бронштейн И.Н., Семендяев К.А. Справочник по математике для инженеров и учащихся ВТУЗов. М, 1980.

Гатаулин А.М. Система прикладных статистико-математических методов обработки экспериментальных данных в сельском хозяйстве. М., 1992.

Искусственный интеллект: Справочник: в 3 книгах / Под ред. Э.В. Попова. М., 1990.

Орлов А.И. Прикладная статистика: Учебник. М.: «Экзамен», 2004.

Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. 4-е изд. М.: БИНОМ. Лаборатория знаний, 2006.

Практическая часть

Аудиторные занятия: 4 часа.

Самостоятельная работа: 2 часа.

Цель работы

Освоить приёмы приведения числовых переменных к дискретной форме.

Закрепить теоретические знания по вопросу «виды шкал».

Приборы и материалы

Компьютерный класс с доступом к сети Internet; программное обеспечение, реализующее аналитическую группировку, проверку статистических гипотез о характере распределения случайной величины и численные методы решения интегральных уравнений (рекомендуется MathCad; в его отстутствие задача может быть решена средствами Excel и VBA); информационный сайт Продовольственной и сельскохозяйственной организации ООН (FAO):
http://faostat.fao.org/DesktopDefault.aspx?PageID=567&lang=ru

Задание

Привести числовые переменные системы, специфицированной при выполнении предыдущего задания, к дискретной форме. Для этого:

- обоснованно выдвинуть и проверить гипотезу о характере статистического распределения факторных переменных;
- выделить квартили интервалов вариации факторных переменных.

Методические указания по выполнению задания

Чтобы избежать неоправданно большого объёма вычислений, в учебной задаче достаточно выделить четыре квантили для каждой числовой переменной, а нечисловые переменные не должны иметь более пяти вариантов.

Если переменная принимает только неотрицательные значения, её распределение не может быть нормальным[1]. Многие из неотрицательных экономических переменных имеют распределения, близкие к гамма-распределению или логнормальному распределению. Переменные, означающие численность редких событий (неисправностей сельхозтехники, заболеваний скота, отказов контрагентов от выполнения обязательств), распределены согласно закону Пуассона.

Если эмпирическое распределение многовершинное, совокупность наблюдений (при их достаточной численности) часто удаётся разделить на качественно различающиеся совокупности, в каждой из которых эмпирическое распределение одновершинное. Если наблюдений мало (до 20), в подобных случаях практически оправданно выдвигать гипотезу о равномерном распределении.

Требования к отчёту

Отчёты о выполнении практического задания составляются индивидуально. Объём каждого отчёта не должен превышать 6 страниц (не считая приложений).

В каждом отчёте должны присутствовать:

- характеристики распределения вероятности для каждой числовой переменной, исследованной составителем отчёта;
- расчёт и результаты проверки гипотезы о соответствии эмпирического распределения выбранному теоретическому распределению;
- краткое описание подходов к статистическому анализу, не описанных в методических указаниях, но использованных при выполнении практического задания (со ссылками на источники);
- границы квантилей числовых переменных;
- список использованной литературы.

[1] В ряде случаев для таких переменных гипотеза о нормальном распределении может быть приемлемой, если вероятность отрицательных значений согласно теоретическому распределению исчезающе мала.

ТЕМА 3. ПРЕДСТАВЛЕНИЕ ЗНАНИЙ О СТРУКТУРЕ СИСТЕМЫ В ФОРМЕ УСЛОВНЫХ ВЕРОЯТНОСТЕЙ. ПРОВЕРКА СУЩЕСТВЕННОСТИ И НЕЗАВИСИМОСТИ ПЕРЕМЕННЫХ

Теоретическая часть

Проверка существенности и независимости переменных

Входные переменные подсистем изучаемой производственной системы должны обладать свойствами независимости и существенности.

Свойство *независимости* состоит в том, что все входные переменные должны быть взаимно независимы либо связь между ними должна быть достаточно слабой, чтобы её можно было игнорировать.

Свойство *существенности* — в том, что выходная переменная должна зависеть от каждой из входных, причём после получения информации о значениях всех входных переменных энтропия выходной переменной должна быть как можно меньше.

О наличии этих свойств у переменных, включённых в модель, судят на основе статистических показателей тесноты связей, проверки статистических гипотез о независимости переменных, доли энтропии (относительной информативности) переменной, снимаемой информацией о значении другой переменной. Выбирая методы оценки тесноты связи, следует учитывать особенности их содержания. В частности:

♦ критерий χ^2 может быть использован применительно к дискретным переменным для проверки гипотез о независимости двух дискретных переменных на основании имеющихся наблюдений (см. Приложение 4), а также о том, не противоречит ли предполагаемая форма связи между переменными имеющимся данным;

♦ однофакторный дисперсионный анализ имеет целью проверку гипотезы о существовании статистически достоверной зависимости непрерывной нормально распределённой переменной[1] от дискретной (или приведённой к дискретной форме) переменной (см. Приложение 5);

[1] При гамма-распределении результаты оценки тесноты связи при посредстве дисперсионного анализа содержат ошибку, величина которой, однако, для большинства практических приложений не слишком велика.

♦ метод относительной информативности (см. Приложение 6) позволяет определить, какая доля энтропии одной дискретной переменной снимается другой дискретной переменной. Проверку тесноты связи по этому методу делают *после* построения таблиц условных вероятностей (см. ниже);

♦ корреляционный анализ оценивает тесноту связи между переменными непрерывными при условии, что связь между ними предполагается линейной. Если величина $r\sqrt{N-2}\big/\sqrt{1-r^2}$, где N — число наблюдений, а r — коэффициент парной корреляции по Пирсону, оказывается за пределами соответствующего выбранному уровню доверия квантиля распределения Стьюдента для числа степеней свободы $N-2$, гипотеза о независимости переменных отвергается[1]. Соответствующие вычисления можно выполнить по формуле Excel

```
=СТЬЮДРАСП(ABS(КоэфКор)/КОРЕНЬ(1-КоэфКор^2)*

КОРЕНЬ(СЧЁТ(Ряд1)-2);СЧЁТ(Ряд1)-2;2).
```

Здесь `КоэфКор` — имя ячейки, содержащей коэффициент парной корреляции по Пирсону, вычисляемый по формуле

```
=ПИРСОН(Ряд1;Ряд2),
```

`Ряд1` и `Ряд2` — имена диапазонов ячеек, содержащих наблюдаемые значения переменных, связь между которыми исследуется. В обоих рядах должно быть одинаковое количество ячеек, нечисловых значений и пустых ячеек быть не должно. В программе MathCad соответстствующие вычисления выглядят следующим образом:

$$=\mathrm{dt}\left(\frac{\mathrm{corr}(\text{Ряд1};\text{Ряд2})\cdot\sqrt{\mathrm{length}(\text{Ряд1})-2}}{\sqrt{1-\mathrm{corr}(\text{Ряд1};\text{Ряд2})^2}};\mathrm{length}(\text{Ряд1})-2\right),$$

[1] Если наблюдений больше 30 — можно использовать нормальное распределение, которое является пределом распределения Стьюдента при бесконечном числе наблюдений.

где `Ряд1` и `Ряд2` — имена векторов, содержащих наблюдения исследуемых переменных.

При исследовании систем принимают во внимание, что независимость некоторой переменной x_1 от каждой из остальных ($x_2 \ldots x_n$) ещё не означает, что x_1 не зависит от некоторой функции $f(x_2 \ldots x_n)$.

Входную (факторную) переменную исключают из модели в следующих случаях:

- отсутствие её связи с выходной переменной статистически достоверно;
- она тесно коррелирует с другой входной переменной, не исключаемой из модели, либо снимает существенную часть её энтропии.

Представление знаний о структуре системы в форме условных вероятностей

Числовая модель производственной системы в данном случае представляет собой систему количественных зависимостей выходных переменных от входных.

В данном случае в каждой подсистеме входные переменные предполагаются независимыми, сами переменные — дискретными, а связи между выходными и входными переменными — вероятностными. Следовательно, связи могут быть количественно охарактеризованы математическим ожиданием вероятностью значений входных переменных при заданном значении выходной переменной.

Такая количественная характеристика связей может быть построена на основе наблюдений моделируемых систем даже при полном отсутствии какого-либо априорного знания о характере связей. Однако её достоверность зависит от количества имеющихся наблюдений моделируемых систем и от точности выполнения условий применимости формализма условных вероятностей. Часто наличие априорного знания позволяет получить значительно более точные и достоверные количественные характеристики связей. В этом случае создание числовой модели требует более мощных формализмов для представления знаний о связях.

На основе наблюдений за поведением изучаемой системы нельзя сделать полностью достоверное заключение о вероятностях её состояний. Например, если 18 раз бросить игральную кость, то из того, что единица выпала шесть раз, не следует, что вероятность её выпадания равна ⅓.

Наблюдаемая частота некоторого значения переменной может быть обусловлена различной действительной вероятностью этого значения. Одна-

нако при разных действительных вероятностях шансы на то, чтобы наблюдать именно такую частоту, различны.

Располагая только ограниченным количеством наблюдений изучаемой дискретной переменной, исследователь не имеет никакой более обоснованной *оценки* вероятности её значений, нежели средняя взвешенная вероятностей данного значения, которые могли вызвать его реализацию n раз из N наблюдений. Эта величина называется *наиболее правдоподобной оценкой вероятности*.

Можно доказать, что наиболее правдоподобная оценка вероятностей, которые могли вызвать наблюдение некоторого значения дискретной переменной n раз из N наблюдений, равна $\frac{n+1}{N+k}$, где k — число возможных значений. Чем больше число наблюдений, тем меньше эта величина отличается от $\frac{n}{N}$.

Для полной характеристики стохастических связей дискретной выходной переменной от дискретных взаимно независимых входных переменных достаточно определить:

- оценки вероятности каждого значения всех переменных;
- оценки условной вероятности каждого значения всех *входных* переменных при заданном значении *выходной* переменной.

При отсутствии какой-либо другой информации математические ожидания условной вероятности рассчитываются на основе комбинационных таблиц (таблиц сопряжённости), включающих выходную и одну из входных переменных. Столбцы такой таблицы соответствуют дискретным значениям входной, а строки — выходной переменной. В клетках таблицы помещается число наблюдений, в которых наблюдались соответствующих значения обеих переменных.

При этом:

- вероятность выходной переменной оценивается по вышеприведённой формуле (в при выполнении заданий данного практикума этот способ применяется редко: см. ниже!);
- условные вероятности значений *входной* переменной при известных значениях *выходной* переменной (именно эти вероятности потребуются нам для модели) — по формуле

$$\frac{n_{ij}+1}{n_j+Q},$$

где n_{ij} — число наблюдений, при которых выходная переменная имела значение i, а входная — j; n_j — общее число наблюдений j-го значения входной переменной; Q — число квантилей выходной переменной. При правильном вычислении сумма всех условных вероятностей, имеющих одинаковый индекс j, должна быть равна единице.

Для вероятностей значений числовых переменных, *приведённых к дискретной форме* путём разбиения интервала вариации на Q квантилей, возможна лучшая оценка, чем вышеприведённая, поскольку, кроме данных, можно использовать знание закона распределения случайной величины, основанное на теоретическом представлении о причинах её вариации.

В этом случае вместо оценки вероятности по вышеприведённой формуле используется оценка, равная $1/Q$. Эта оценка надёжнее математического ожидания вероятности: ведь при выдвижении гипотезы о распределении вероятности значений данной переменной мы опирались не только на результаты наблюдения, но и на другие знания: экономическое содержание данной переменной, диапазон вариации, аналогию с другими экономическими переменными и др.

Библиографический список

Гатаулин А.М. Система прикладных статистико-математических методов обработки экспериментальных данных в сельском хозяйстве. М., 1992.

Искусственный интеллект: Справочник: в 3 книгах / Под ред. Э.В. Попова. М., 1990.

Красс М.С., Чупрынов Б.П. Математические методы и модели для магистрантов экономики: Учеб. пособие. СПБ.: Питер, 2006.

Нейлор К. Экспертные системы: принципы работы и примеры. М., 1987.

Орлов А.И. Теория принятия решений: Учеб. пособие. М.: Изд-во «Март», 2004.

Светлов Н.М. Обоснование весовых коэффициентов исходов в стохастических моделях сельскохозяйственного производства // Доклады ТСХА. М., 1995, вып. 266, с. 190-195.

Практическая часть

Аудиторные занятия: 2 часа.

Самостоятельная работа: 1 час.

Цель работы

Приобрести навыки количественного описания зависимостей между дискретными переменными средствами формализма условных вероятностей.

Научиться обосновывать взаимную независимость входных переменных системы и исследовать существенность их влияния на выходную.

Закрепить теоретические знания по вопросам «формы представления систем», «свойства систем», «метод системного анализа» и «связь теории систем с другими науками».

Приборы и материалы

Компьютерный класс с доступом к сети Internet; программное обеспечение, реализующее вычислительные процедуры проверки существенности и независимости переменных (рекомендуется MathCad; в его отстутствие задача может быть решена средствами Excel); информационный сайт Продовольственной и сельскохозяйственной организации ООН (FAO): http://faostat.fao.org/DesktopDefault.aspx?PageID=567&lang=ru

Задание

1. Проверить соответствие подсистемы первого уровня требованиям существенности и независимости входных переменных.

2. При необходимости пересмотреть набор входных переменных. Числовые переменные, вновь включённые в модель, привести к дискретной форме. Для каждой переменной, включённой в модель, рассчитать таблицы условных вероятностей.

3. Определить математические ожидания условной вероятности возможных значений каждой входной переменной при заданном значении выходной и построить таблицы условных вероятностей.

Замечание. Если для проверки существенности и независимости некоторых входных переменных рабочая группа решила применять метод относительной информативности, то для данных переменных последовательность выполнения задания меняется: сначала выполняется п.3, затем пп.1 и 2.

Методические указания по выполнению задания

При решении задач практического уровня сложности по мере возможности исследуются многофакторные зависимости. Для достижения це-

лей изучения данной темы с учётом естественных ограничений по времени и сложности выполнения задания достаточно исследовать ***только парные*** зависимости между переменными.

Для обеспечения достоверности анализа рекомендуется использовать не менее двух методов оценки тесноты связи для каждой пары переменных.

Если преподавателем не указано иначе, используйте следующие критерии исключения входной переменной из модели:

отсутствие статистически достоверной связи с выходной переменной при $\alpha = 0{,}1$ по подходящему статистическому критерию независимости;

снятие более 15% энтропии какой-либо выходной переменной, не исключаемой из модели, либо отклонение гипотезы об их независимости по подходящему статистическому критерию при $\alpha = 0{,}05$.

Исключённые переменные заменяют новыми переменными из ранжированного ряда, составленного при выполнении задания к теме 1, отдавая предпочтение переменным с наиболее высоким рангом. Для новых переменных повторяют процедуру проверки их существенности и независимости.

Если по результатам проверки существенности и независимости переменных не удаётся выбрать достаточное количество переменных для включения в модель, а также в случае возникновения сомнений относительно того, следует ли вносить изменения в модель подсистемы первого уровня, необходимо обратиться к преподавателю.

Требования к отчёту

Отчёт о выполнении практического задания состоит из коллективной и индивидуальных частей. Объём коллективной части — не более 2 страниц, индивидуальной — до 8 страниц (не считая приложений).

В коллективной части указываются переменные подсистемы первого уровня, исключённые из модели, и переменные, предложенные для включения в модель вместо исключённых. Изменения в модели должны быть обоснованы.

В каждой индивидуальной части должны быть приведены:

- комбинационные таблицы, построенные составителем;
- математические ожидания вероятности, рассчитанные составителем;

- использованные составителем методы анализа связей для каждой пары показателей, исследованной составителем отчёта;
- количественная оценка тесноты связей;
- заключение о тесноте связей;
- предложения по совершенствованию модели;
- результаты проверки гипотез о распределении вероятностей, границы квантилей и таблицы условных вероятностей для исследованных составителем отчёта переменных, введённых в модель взамен не отвечающих условиям существенности и независимости;
- список литературы, использованной при подготовке к практическому занятию.

ТЕМА 4. СПЕЦИФИКАЦИЯ ВТОРОГО УРОВНЯ АГРАРНОЙ ПРОИЗВОДСТВЕННОЙ СИСТЕМЫ

Теоретическая часть

При использовании формализма условных вероятностей модели второго уровня требуются в том случае, если данные о значении сооветствующей входной переменной первого уровня отсутствуют. Хотя данный формализм позволяет получить оценки распределения вероятностей выходной переменной, наилучшим образом согласующиеся с поступившей информацией о значениях входных переменных даже в тех случаях, когда значения некоторых переменных не поступили вовсе или известны лишь с некоторой вероятностью, необходимо принимать меры по получению информации о возможно большем количестве входных переменных, так как чем больше данных поступило, тем меньше неопределённость результата, обусловленная неопределённостью значений некоторых переменных.

Если наблюдать некоторые входные переменные первого уровня всё же не удаётся, имеется возможность оценить распределение вероятностей их значений, опираясь на наблюдения тех переменных, от которых они зависят, то есть входных переменных моделей второго уровня.

Процедура спецификации второго уровня аграрной производственной системы и используемые при её реализации методики отличаются от рассмотренных в предыдущих трёх темах лишь в деталях. В целом определение набора входных переменных второго уровня требует выполнения всё тех же этапов:

- предварительного отбора входных переменных второго уровня при посредстве построенного с помощью экспертных процедур ранжированного ряда переменных, влияющих на выходную переменную второго уровня (одновременно являющуюся входной переменной первого уровня);
- их дискретизации (если они непрерывные);
- проверки их существенности и независимости и, при необходимости, корректировки модели;
- формирования таблиц условных вероятностей.

Отличия состоят в том, что на практике спецификация систем второго уровня обыкновенно сталкивается с ещё большим недостатком эмпирических данных, чем это наблюдается при работе с первым уровнем. Чаще остаются неизвестными формы распределений вероятностей переменных второго уровня, и потому дискретизация чаще выполняется непосредственно по эмпирическим данным, а не по теоретическому распределению. Чаще используются переменные, значения которых для каждого наблюдения получены не путём статистического наблюдения или постановки опыта, а посредством экспертных оценок.

При практическом использовании формализма условных вероятностей для разработки интеллектуальных информационных систем часто используется подход, отличающийся от рассматриваемого в данном практикуме. Именно, входные переменные второго и нижележащих уровней выбираются по такой же или схожей процедуре, но таблицы условных вероятностей строятся для вероятностей значений входной переменной *второго (или более низкого)* уровня при условии заданного значения выходной переменной *первого* уровня. При этом, во избежание смещённой оценки выходной переменной из-за зависимости факторов, одновременно используемых в расчётах (ведь входные переменные первого уровня заведомо зависят от соответствующих входных переменных второго уровня, что обеспечивается процедурой их отбора), данные о значениях факторов низших уровней обрабатываются только при отсутствии данных о соответствующей переменной более высокого уровня.

Такой подход упрощает алгоритм работы формализма и сокращает объём вычислений, но у него есть существенный недостаток: не всегда существуют наблюдения, в которых зафиксированы значения выходной переменной первого уровня вместе со значениями входных переменных низших уровней. Многоуровневая модель даёт возможность использовать независимые источники данных для построения таблиц условных вероятностей для разных подсистем. В случае, если все таблицы условных вероятностей связывают входные переменные разных уровней с выходной переменной первого уровня, требуется, чтобы значения всех этих переменных фиксировались в одних и тех же наблюдениях.

Теоретически входная переменная некоторой подсистемы второго уровня не может одновременно быть входной переменной другой подсистемы второго уровня: если бы такое имело место, две выходных переменных второго уровня оказались бы зависимыми. То же касается и более низких уровней. На практике смещение оценки выходной переменной первого уровня, обусловленное подобными зависимостями, может оказаться неизбежным, так как полную независимость факторов обеспечить удаётся далеко не всегда. При недостатке данных с подобными явлениями прихо-

дится мириться, а в дальнейшем, по мере накопления опытных данных, возникающие в связи с этим проблемы неадекватности модели решаются либо путём замены парных таблиц условных вероятностей таблицами большей размерности (трёх- или четырёхмерными), либо обращением к более мощным формализмам. Поэтому на практике включение одной и той же входной переменной в две подсистемы второго уровня в исключительных случаях допускается. При этом связь её с соответствующими выходными переменными должна быть существенной, но слабой.

Библиографический список

Теория систем: Учеб. пособие / В.Н. Волкова, А.А. Денисов. М.: Высшая школа, 2006.

Франс Дж., Торнли Дж. Математические модели в сельском хозяйстве / Пер. с англ. М.: Агропромиздат, 1987.

Практическая часть

Аудиторные занятия: 4 часа.
Самостоятельная работа: 4 часа.

Цель работы

Приобрести навыки спецификации подсистем производственной системы.

Научиться решать задачи системного анализа в условиях ограниченной информационной базы, пользуясь экспертными оценками, картографическим материалом, справочными изданиями и другими источниками информации.

Закрепить теоретические знания по теме «структура систем».

Приборы и материалы

Компьютерный класс с доступом к сети Internet; программное обеспечение, автоматизирующее рутинные операции по ранжированию факторов, реализующее аналитическую группировку, проверку статистических гипотез о характере распределения случайной величины, численные методы решения интегральных уравнений, вычислительных процедур проверки существенности и независимости переменных (рекомендуются Excel и MathCad; в отстутствие MathCad задача может быть решена средствами Excel и VBA); информационный сайт Продовольственной и сельскохозяйственной организации ООН (FAO):
http://faostat.fao.org/DesktopDefault.aspx?PageID=567&lang=ru

Задание

1. Специфицировать входные переменные всех подсистем второго уровня исследуемой производственной системы.

2. Построить таблицы условных вероятностей для подсистем второго уровня.

3. Проверить соответствие переменных подсистем второго уровня требованиям независимости и существенности; при необходимости пересмотреть спецификацию подсистем.

Методические указания по выполнению задания

Для достижения целей практического задания общее число входных переменных подсистем второго уровня должно составить 7...10, из них не менее 5 числовых. Существенность и независимость переменных проверяются только одним методом, по возможности наименее трудоёмким. С разрешения преподавателя для некоторых переменных эту проверку можно опустить.

Учитывая ограниченность информационной базы, доступной для выполнения практического задания, замену входных переменных, для которых не выполняются требования существенности и независимости, производить не обязательно (в практических приложениях это делать необходимо!).

В процессе решения задачи предполагается использование электронных таблиц, алгоритмов и программ, разработанных при выполнении предыдущих заданий.

Требования к отчёту

Отчёт о выполнении практического задания включает коллективную и индивидуальные части. Объём коллективного раздела не должен превышать 1 страницы, индивидуального — 6 страниц (не считая приложений).

Коллективный раздел должен содержать схему модели производственной системы, отражающую все её переменные и связи между ними.

В индивидуальных разделах должны присутствовать:

♦ сведения о виде распределения и границах квантилей, а также таблицы условных вероятностей для каждой переменной, исследованной составителем отчёта;

♦ краткое описание методов, не описанных в методических указаниях, но использованных при выполнении практического задания;

♦ список использованной литературы.

ТЕМА 5. ТЕСТИРОВАНИЕ ДВУХУРОВНЕВОЙ МОДЕЛИ

Теоретическая часть

Вероятности значений входной переменной низшего уровня определяются путём последовательного использования формулы Байеса для учёта информации о состоянии каждой входной переменной.

Формула Байеса имеет вид

$$p(A_i / B_{gh}) = \frac{p(A_i)\ p(B_{gh} / A_i)}{\sum_{j=1}^{n} p(A_j)\ p(B_{gh} / A_j)}. \qquad (4)$$

Она позволяет перейти от вероятности $p(A_i)$ события A_i к вероятности $p(A_i / B_{gh})$ события A_i при условии, что имеет место событие B_{gh}. В нашем случае $p(A_i / B_{gh})$ — вероятность i-го значения выходной переменной при условии, что имеет место h-е значение входной переменной x_g; n — число возможных значений выходной переменной; $p(B_{gh} / A_i)$, $p(B_{gh} / A_j)$ — вероятность h-го значения входной переменной x_g при условии i-го (j-го) значения выходной переменной; $p(A_i)$, $p(A_j)$ — вероятность i-го (j-го) значения выходной переменной.

Если входные переменные независимы, можно вычислить вероятность i-го значения выходной переменной при условии, что известны значения некоторых или всех входных переменных. Например, предположим, что требуется определить вероятность $p(A_i / (B_{gh} \cup B_{qw}))$ события A_i при условии, что $x_g = h$ и $x_q = w$. Для этого можно использовать формулу Байеса в любой из нижеследующих форм:

$$\frac{p(A_i / B_{gh})\ p(B_{qw} / A_i)}{\sum_{j=1}^{n} p(A_j / B_{gh})\ p(B_{qw} / A_j)}, \qquad (5)$$

где значение $p(A_i / B_{gh})$ ранее определено по формуле (4), либо

$$\frac{p(A_i / B_{qw})\ p(B_{gh} / A_i)}{\sum_{j=1}^{n} p(A_j / B_{qw})\ p(B_{gh} / A_j)}, \qquad (6)$$

если ранее с помощью формулы, аналогичной (4), определено значение $p(A_i / B_{qw})$.

Рассмотрим использование формулы Байеса на упрощённом примере, в котором каждая переменная имеет по два дискретных значения, а входных переменных две. Положим, что вероятности значений выходной переменной x_0 до получения информации о входных переменных отличаются лишь немного: $p(x_0=1)=0{,}55$, $p(x_0=2)=0{,}45$. Заданы следующие условные вероятности (вертикальную черту следует читать как «при условии, что»):

- $p(x_1=1 \mid x_0=1) = 0{,}5$;
- $p(x_1=2 \mid x_0=1) = 0{,}5$;
- $p(x_1=1 \mid x_0=2) = 0{,}2$;
- $p(x_1=2 \mid x_0=2) = 0{,}8$;
- $p(x_2=1 \mid x_0=1) = 0{,}1$;
- $p(x_2=2 \mid x_0=1) = 0{,}9$;
- $p(x_2=1 \mid x_0=2) = 0{,}8$;
- $p(x_2=2 \mid x_0=2) = 0{,}2$.

Положим, что поступила информация о значении второй входной переменной: $x_2=1$. Согласно формуле (4), вероятность события $x_0=1 \mid x_2=1$ составит

$$\frac{0{,}55 \cdot 0{,}1}{0{,}55 \cdot 0{,}1 + 0{,}45 \cdot 0{,}8} = 0{,}1325.$$

Вероятность события $x_0=2 \mid x_2=1$ составит

$$\frac{0{,}45 \cdot 0{,}8}{0{,}55 \cdot 0{,}1 + 0{,}45 \cdot 0{,}8} = 0{,}8675.$$

Как и следует, сумма этих двух вероятностей равна единице.

Теперь положим, что в дополнение к уже имеющейся информации о второй переменной поступила информация ещё и о первой: $x_1=2$. Согласно формуле (5), вероятность события $x_0=1 \mid (x_2=1 \cup x_1=2)$ равна

$$\frac{0{,}1325 \cdot 0{,}5}{0{,}1325 \cdot 0{,}5 + 0{,}8675 \cdot 0{,}8} = 0{,}0871.$$

Вероятность события $x_0=2 \mid (x_2=1 \cup x_1=2)$ составит

$$\frac{0{,}8675 \cdot 0{,}8}{0{,}1325 \cdot 0{,}5 + 0{,}8675 \cdot 0{,}8} = 0{,}9129.$$

Как и следует, сумма этих двух вероятностей равна единице.

Итак, если энтропия переменной x_0 до получения информации составляла $-0{,}45 \cdot \log_2 0{,}45 - 0{,}55 \cdot \log_2 0{,}55 = 0{,}9928$ бит, то после получения первого сигнала она стала равна 0,5643 бит, а после второго сократилась до 0,4267 бит.

На практике поступление новой информации может не только снижать, но и увеличивать энтропию.

В общем случае для определения вероятности i-го значения выходной переменной формулу Байеса применяют ровно столько раз, сколько имеется известных значений входных переменных.

Вероятности значений выходных переменных более высоких уровней при заданных значениях переменных низшего уровня определяются по формуле средней взвешенной

$$p(C_k / B) = \sum_{i=1}^{m} p(C_k / D_i)\, p(D_i / B), \qquad (7)$$

где B — сочетание значений входных переменных *низшего* уровня; D_i — сочетание значений входных переменных *данного* уровня; $p(C_k / B)$ — вероятность k-го значения выходной переменной при условии, что имеет место сочетание B; $p(C_k / D_i)$ — вероятность k-го значения выходной переменной при условии, что имеет место сочетание D_i (эта вероятность определяется последовательным применением формулы Байеса); $p(D_i / B)$ — вероятность сочетания D_i при условии, что имеет место сочетание B (равна произведению вероятностей вошедших в сочетание D_i значений переменных данного уровня при условии, что имеет место сочетание B), m — число сочетаний значений входных переменных данного уровня.

Рассмотрим числовой пример применения этой формулы. Пусть в вышеприведённом примере информация об x_1 не поступила, и для её оценивания была использована модель второго уровня, которая дала следующие результаты: $p(x_1=1)=0{,}3$, $p(x_1=2)=0{,}7$. Тогда в соответствии с формулой (7) нам следует определить величину

$$p(x_0=1 \mid (x_2=1 \cup x_1=1)) \cdot 0{,}3 + p(x_0=1 \mid (x_2=1 \cup x_1=2)) \cdot 0{,}7$$

Второе слагаемое, как следует из расчётов величины $p(x_0=1 \mid (x_2=1 \cup x_1=2))$, проведённых выше, составляет $0{,}0871 \cdot 0{,}7$. Для первого слагаемого нужно заново вычислить $p(x_0=1 \mid (x_2=1 \cup x_1=1))$ по формуле Байеса, пользуясь уже определённым ранее значением $p(x_0=1 \mid x_2=1)$.

Получим 0,2763. Окончательно имеем 0,0871·0,7+0,2763·0,3, то есть 0,1439.

Аналогичным образом получим оценку вероятности для второго значения переменной x_0, то есть величину

$$p(x_0=2|(x_2=1 \cup x_1=1))\cdot 0{,}3 + p(x_0=2|(x_2=1 \cup x_1=2))\cdot 0{,}7.$$

Она составит 0,8561. Сумма вероятностей всех возможных значений переменной (в данном случае двух) равна единице — иное означало бы, что в расчётах допущена ошибка.

Библиографический список

Искусственный интеллект: Справочник: в 3 книгах / Под ред. Э.В. Попова. М., 1990.

Нейлор К. Экспертные системы: принципы работы и примеры. М., 1987.

Построение экспертных систем / Под ред. Ф. Хейеса-Рота. М., 1987.

Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. 4-е изд. М.: БИНОМ. Лаборатория знаний, 2006.

Практическая часть

Аудиторные занятия: 4 часа.

Самостоятельная работа: 4 часа или 8 часов (см. ниже).

Цель работы

Приобрести практические навыки применения алгоритмов вычисления вероятностей значений выходной переменной системы, описанной средствами формализма условных вероятностей.

Закрепить теоретические знания по темам «метод моделирования», «свобода систем».

Приборы и материалы

Отчёты о выполнении предыдущих лабораторных работ; ПЭВМ, табличные процессоры, трансляторы алгоритмических языков.

Задание

1. Разработать программное средство для определения вероятностей значений выходной переменной модели.

2. Определить вероятности значений выходной переменной и математическое ожидание её величины для десяти различных комбинаций значений входных переменных.

3. Дать оценку энтропии, снимаемой с выходной переменной поступающей информацией.

Методические указания по выполнению задания

Комбинации значений входных переменных формируются студентами самостоятельно. При этом должны выполняться следующие требования:

- среди комбинаций должны быть такие, по которым отсутствуют данные по входной переменной первого уровня и по всем соответствующим ей входным переменным второго уровня;
- не менее чем в половине вариантов должны использоваться данные о значениях переменных второго уровня;
- не менее чем в двух вариантах должны использоваться данные всех переменных второго уровня;
- варианты, в которых данные о значениях переменных второго уровня не используются, должны быть обязательно;
- для каждой входной переменной первого уровня должен найтись вариант, в котором она принимает любое из возможных для неё дискретных значений.

По результатам расчётов определяются:

- распределение вероятностей выходной переменной первого уровня;
- её математическое ожидание (для числовых переменных);
- её энтропия после получения информации о значениях входных переменных и размер снятой энтропии.

Математическое ожидание определяется:

- для переменных, приведённых к дискретному виду, — по формуле

$$\sum_{i=1}^{Q}\left(p_i \int_{x_{i-1}}^{x_i} f(x)dx \right),$$

где Q — число квантилей, p_i — оценка вероятности i-го дискретного значения переменной с учётом поступившей информации о входных переменных; $f(x)$ — функция плотности распределения вероятностей данной переменной, x_{i-1} и x_i — нижняя и верхняя границы квартили i;

♦ для остальных дискретных переменных — как сумма произведений их возможных значений на соответствующие оценки вероятностей, полученные с учётом поступившей информации о входных переменных.

Доказательство корректности расчётов, выполненных программой, производится посредством аннотированных расчётов по одному из вариантов, в котором используется наименьшее количество входных переменных второго уровня (рекомендуемое количество — одна), выполненных в электронных таблицах или вручную.

Если рабочей программой курса на выполение практического задания по данной теме выделено 2 часа, преподаватель не проверяет качество программных компонентов, разработанных для решения задачи, а только корректность расчётов.

Если рабочей программой курса на выполение практического задания по данной теме выделено 8 часов, к программному средству для выполнения расчётов согласно заданию к данной теме предъявляются следующие требования:

♦ оно должно принимать значение входных переменных как в дискретной форме (номер квантили), так и в непрерывной, с присущей данной переменной единицей измерения, за исключением тех переменных, которые являются дискретными по своей природе, если таковые имеются в модели;

♦ отсутствие данных о значении каких-либо переменных не должно препятствовать выполнению вычислений и их корректности;

♦ интерфейс пользователя должен предотвращать ввод данных по входным переменным второго уровня, если введено значение соответствующей переменной первого уровня;

♦ если значение входной переменной первого уровня неизвестно, но известны соответствующие значения переменных второго уровня (хотя бы одной), программой должны отображаться вероятности каждого дискретного значения входной переменной первого уровня, оценённые при помощи формулы Байеса;

♦ исполнение программы не должно сопровождаться ошибками, приводящими к её аварийному завершению или зависанию.

Требования к отчёту

Отчёты о выполнении практического задания составляются индивидуально. Объём каждого отчёта не должен превышать 3 страниц (не считая приложений).

В каждом отчёте должны присутствовать:

♦ значения входных переменных, для которых определяются вероятности значений выходной переменной;

♦ вероятности значений выходной переменной и её математическое ожидание, определённые составителем;

♦ доказательства корректности вычислений;

♦ результаты оценки энтропии, снимаемой с выходной переменной поступившей информацией;

♦ наименования инструментальных средств, использованных для выполнения расчётов;

♦ исходные тексты тех фрагментов программ для выполнения расчётов согласно заданию, которые *разработаны составителем отчёта*;

♦ список использованной литературы.

Если на выполнение самостоятельной работы выделяется 8 часов самостоятельной работы, неотъемлемой частью отчёта является исполняемый файл или дистрибутив, предоставленный преподавателю на съёмном носителе данных (возвращаемом студенту) либо посредством электронных коммуникаций.

ПРИЛОЖЕНИЯ

1. Основные статистические распределения

Нормальное распределение

Нормальное распределение (рис. 2) является теоретической моделью случайной величины, представляющей собой сумму константы с бесконечно большим количеством независимых случайных величин (помех), распределённых по произвольным законам на интервале $(-\infty; \infty)$. Данная константа равна математическому ожиданию нормально распределённой случайной величины.

Функция плотности вероятности нормального распределения:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где x — значение случайной величины, μ — её математическое ожидание, σ — среднее квадратическое отклонение, $e \approx 2{,}7182818$ — основание натурального логарифма.

Функция нормального распределения не выражается через элементарные функции и вычисляется с использованием численных методов интегрирования (например, метода трапеций). Математическая запись:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

В Excel плотность распределения вероятности нормального распределения для значения, хранящегося в ячейке `Значение`, вычисляется с помощью формулы

`=НОРМРАСП(Значение;Средняя;Корень(Дисперсия);0)`,

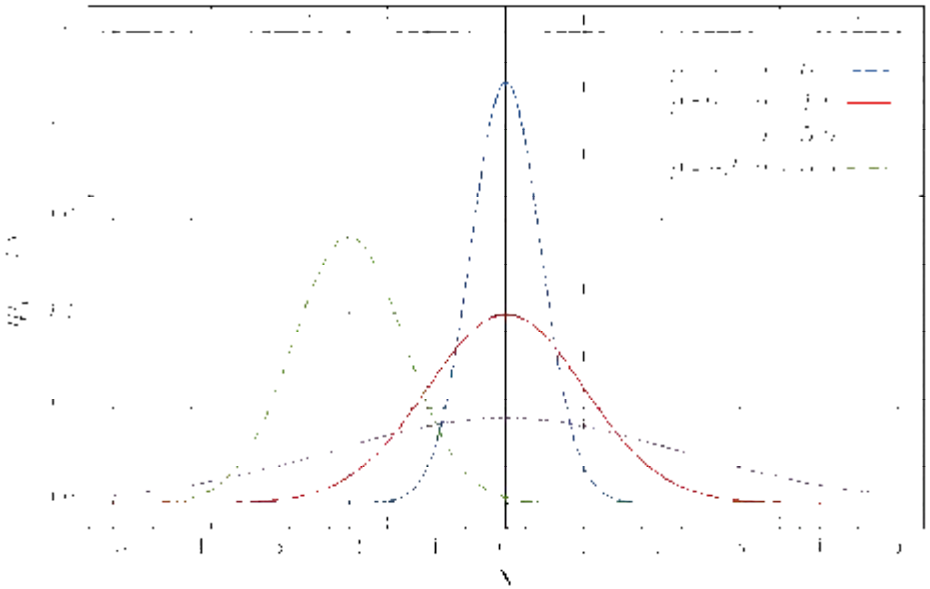
где `Средняя` и `Дисперсия` — имена ячеек, содержащих соответствующие значения. Значение функции нормального распределения (вероятности того, что нормально распределённое случайное значение не превысит указанную величину) вычисляется с помощью формулы

`=НОРМРАСП(Значение;Средняя;Корень(Дисперсия);1)`.

Определить величину, которую с заданной вероятностью не превысит нормально распределённое случайное значение, можно с помощью формулы

`=НОРМОБР(Вероятность;Средняя;Корень(Дисперсия))`,

где `Вероятность` — имя ячейки, содержащей требуемое значение вероятности.



Источник: http://ru.wikipedia.org

Рис. 2. Графики нормального распределения.

В программе MathCad те же вычисления могут быть выполнены с помощью формул

`dnorm(Значение;Средняя;`$\sqrt{}$`Дисперсия)`,

`pnorm(Значение;Средняя;`$\sqrt{}$`Дисперсия)`,

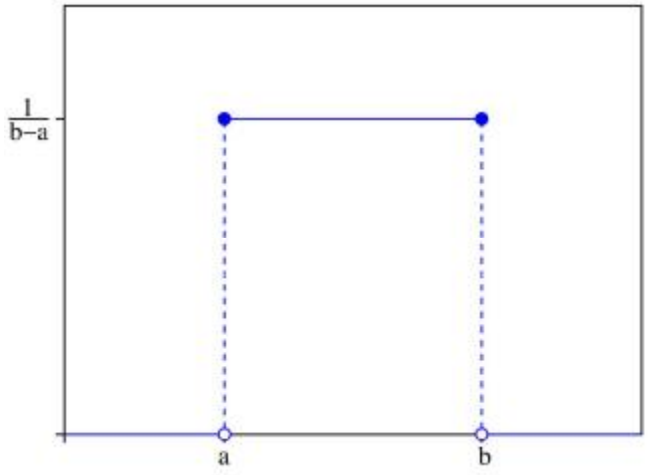
`qnorm(Вероятность;Средняя;`$\sqrt{}$`Дисперсия)`,

где `Значение`, `Средняя`, `Дисперсия` и `Вероятность` — имена соответствующих переменных.

Равномерное распределение

Равномерное распределение (рис. 3) не характерно для случайных величин, описывающих экономические, социальные и природные процессы[1]. Однако оно может оказаться подходящим приближением к реальному (неизвестному) распределению при следующих условиях:

- диапазон вариации случайной величины x заключён между значениями a и b, каждое из которых имеет интерпретацию в терминах исследуемого процесса (подобно тому, как температура воды при атмосферном давлении может быть распределена между 0 и 100°C);
- среднее и модальное значения отличаются от медианы $(a+b)/2$ несущественно;
- дисперсия исследуемой случайной величины отличается от величины $(b-a)^2/12$ несущественно;
- на гистограмме эмпирического распределения отсутствуют выраженные вершины.



Источник: http://ru.wikipedia.org

Рис. 3. График равномерного распределения.

Обычно равномерное распределение оказывается приемлемой моделью только при малом числе наблюдений случайной величины. Принятие гипотезы о равномерном распределении, как правило, означает недостаточную степень изученности моделируемой случайной величины, но может оказаться лучшей гипотезой из всех, которые не могут быть отвергнуты на имеющихся опытных данных.

[1] За исключением тех редких случаев, когда оно оказывается частным случаем бета-распределения.

Функция плотности вероятности равномерного распределения:

$$p(x)=\frac{1}{b-a},\ x\in[a;b],$$

где x — значение случайной величины, a и b — границы множества её значений.

Функция равномерного распределения:

$$F(x)=\frac{x-a}{b-a},\ x\in[a;b].$$

Математическое ожидание равномерно распределённой случайной величины равно $(a+b)/2$; дисперсия — $(b-a)^2/12$.

Треугольное распределение

Треугольное распределение (рис. 4) не характерно для случайных величин, описывающих экономические, социальные и природные процессы[1]. Однако оно может оказаться подходящим приближением к реальному распределению при следующих условиях:

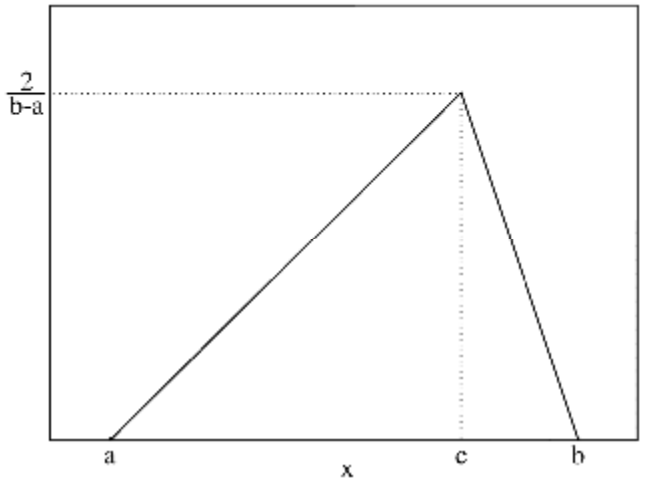
- диапазон вариации случайной величины x заключён между значениями a и b, каждое из которых имеет интерпретацию в терминах исследуемого процесса (подобно тому, как температура воды при атмосферном давлении может быть распределена между 0 и 100°C);
- есть основания считать, что при $x \to a$ и при $x \to b$ плотность вероятности стремится к нулю;
- известно модальное значение случайной величины, равное c;
- среднее значение отличается от величины $(a+b+c)/3$ несущественно;
- дисперсия исследуемой случайной величины отличается от величины

$$\frac{(a^2+b^2+c^2)-(ab+bc+ac)}{18}$$

несущественно.

[1] За исключением тех редких случаев, когда оно оказывается частным случаем бета-распределения.

Обычно треугольное распределение оказывается приемлемой моделью только при малом числе наблюдений случайной величины. Принятие гипотезы о треугольном распределении, как правило, означает недостаточную степень изученности моделируемой случайной величины, но может оказаться лучшей гипотезой из всех, которые не могут быть отвергнуты на имеющихся опытных данных.



Источник: http://en.wikipedia.org

Рис. 4. График треугольного распределения.

Функция плотности вероятности равномерного распределения:

$$p(x)=\begin{cases}\dfrac{2(x-a)}{(b-a)(c-a)}, x\in[a;c];\\ \dfrac{2(b-x)}{(b-a)(b-c)}, x\in(c;b],\end{cases}$$

где x — значение случайной величины, a и b — границы множества её значений, c — модальное (наиболее часто встречающееся) значение.

Функция треугольного распределения:

$$F(x)=\begin{cases}\dfrac{(x-a)^2}{(b-a)(c-a)}, x\in[a;c];\\ 1-\dfrac{(b-x)^2}{(b-a)(b-c)}, x\in(c;b].\end{cases}$$

Математическое ожидание случайной величины, распределённой по треугольному закону, равно $(a+b+c)/3$; дисперсия составляет

$$\frac{(a^2+b^2+c^2)-(ab+bc+ac)}{18}.$$

Экспоненциальное распределение

Экспоненциальное распределение (рис. 5) является теоретической моделью случайной величины, представляющей собой время, проходящее между независимыми однородными случайными событиями, вероятность наступления которых в единицу времени постоянна. Эта величина распределена на интервале $[0;\infty)$. Помимо области определения, признаком экспоненциального распределения является отсутствие существенного различия между средним значением случайной величины и её среднеквадратическим отклонением.

Экспоненциальное распределение является частным случаем гамма-распределения.

Функция плотности вероятности экспоненциального распределения:

$$p(x)=\frac{1}{\mu}e^{-\frac{x}{\mu}},$$

где x — значение случайной величины, μ — её математическое ожидание, $e\approx 2{,}7182818$ — основание натурального логарифма.

Функция экспоненциального распределения:

$$F(x)=1-e^{-\frac{x}{\mu}}.$$

В Excel плотность распределения вероятности экспоненциального распределения для значения, хранящегося в ячейке `Значение`, вычисляется с помощью формулы

```
=ЭКСПРАСП(Значение;1/Средняя;0),
```

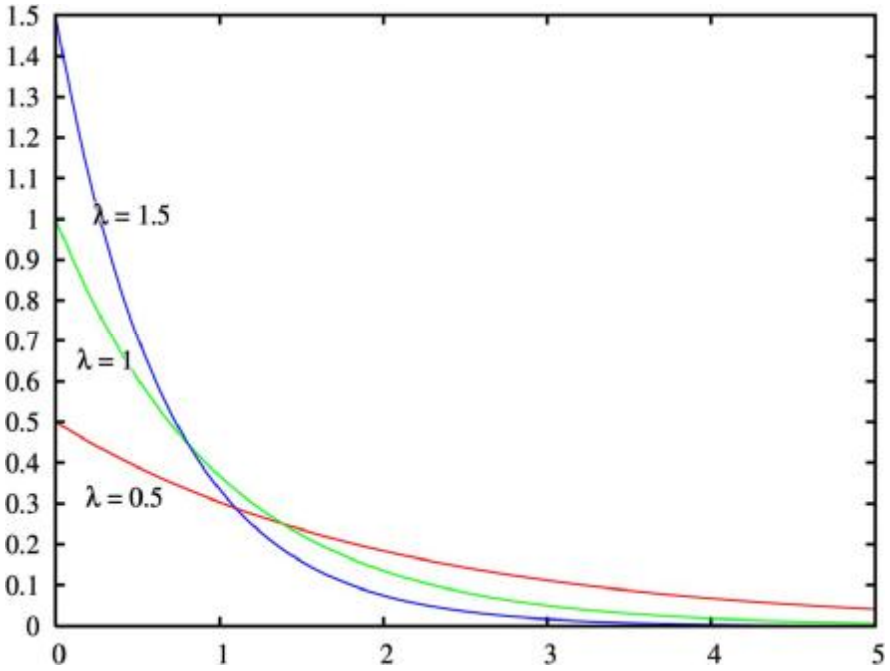
где `Средняя` и `Дисперсия` — имена ячеек, содержащих соответствующие значения. Значение функции экспоненциального распределения (вероятности того, что нормально распределённое случайное значение не превысит указанную величину) вычисляется с помощью формулы

```
=ЭКСПРАСП(Значение;1/Средняя;1).
```

Определить величину, которую с заданной вероятностью не превысит экспоненциально распределённое случайное значение, можно с помощью формулы

```
=ГАММАОБР(Вероятность;1;Средняя),
```

где `Вероятность` — имя ячейки, содержащей требуемое значение вероятности.



Источник: http://ru.wikipedia.org

Рис. 5. Графики экспоненциального распределения.

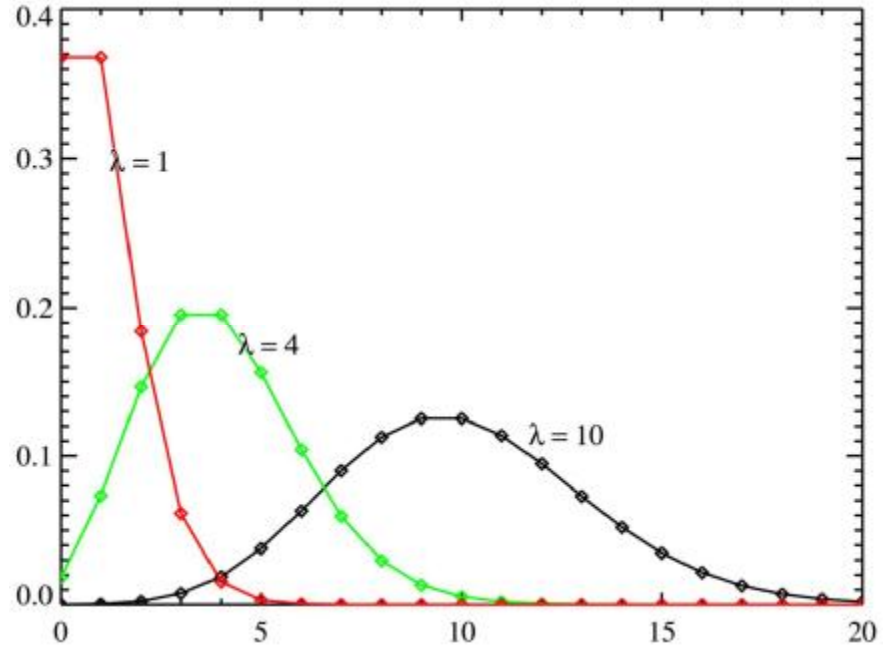
В программе MathCad те же вычисления могут быть выполнены с помощью формул

```
dexp(Значение,1/Средняя),

pexp(Значение,1/Средняя),

qexp(Вероятность,1/Средняя),
```


где `Значение`, `Средняя` и `Вероятность` — имена соответствующих переменных.

Распределение Пуассона

Распределение Пуассона (рис. 6) является дискретным распределением, моделирующим число независимых событий, происходящих в течение заданного промежутка времени, если вероятность наступления каждого из них в течение периода данной продолжительности одна и та же. Оно тесно связано с экспоненциальным распределением, моделирующим длительность промежутков времени между такими событиями.



Источник: http://ru.wikipedia.org

Рис. 6. Полигоны распределения Пуассона.

Областью определения распределения Пуассона является множество целых неотрицательных чисел. Если случайная величина принимает дробные или отрицательные значения, её заведомо нельзя моделировать распределением Пуассона. Характерным признаком применимости распределения Пуассона в качестве модели случайной величины с заданным эм-

пирическим распределением является отсутствие существенного различия между эмпирическими значениями средней и дисперсии.

В соответствии с распределением Пуассона вероятность наступления k событий в течение периода составляет

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

где λ — параметр распределения, одновременно равный математическому ожиданию величины k и её дисперсии. Вероятность наступления k событий или менее (включая отсутствие события) вычисляется по формуле

$$F(k) = \sum_{x=0}^{k} \frac{\lambda^x}{k!} e^{-\lambda}.$$

В Excel $p(k)$ вычисляется с помощью формулы

```
=ПУАССОН(ЧислоСобытий;Средняя;0),
```

а $F(k)$ — с помощью функции

```
=ПУАССОН(ЧислоСобытий;Средняя;1),
```

где в ячейках с именами `ЧислоСобытий` и `Средняя` хранятся значения k и λ. Функции для вычисления k по заданной вероятности в Excel не предусмотрено. Эту величину не составляет труда найти подбором либо написав соответствующую функцию на VBA.

В MathCad аналогичные вычисления производятся с помощью формул

```
dpois(ЧислоСобытий;Средняя),

ppois(ЧислоСобытий;Средняя),

qpois(Вероятность;Средняя),
```

где `ЧислоСобытий`, `Средняя` и `Вероятность` — имена соответствующих переменных.

Логнормальное распределение

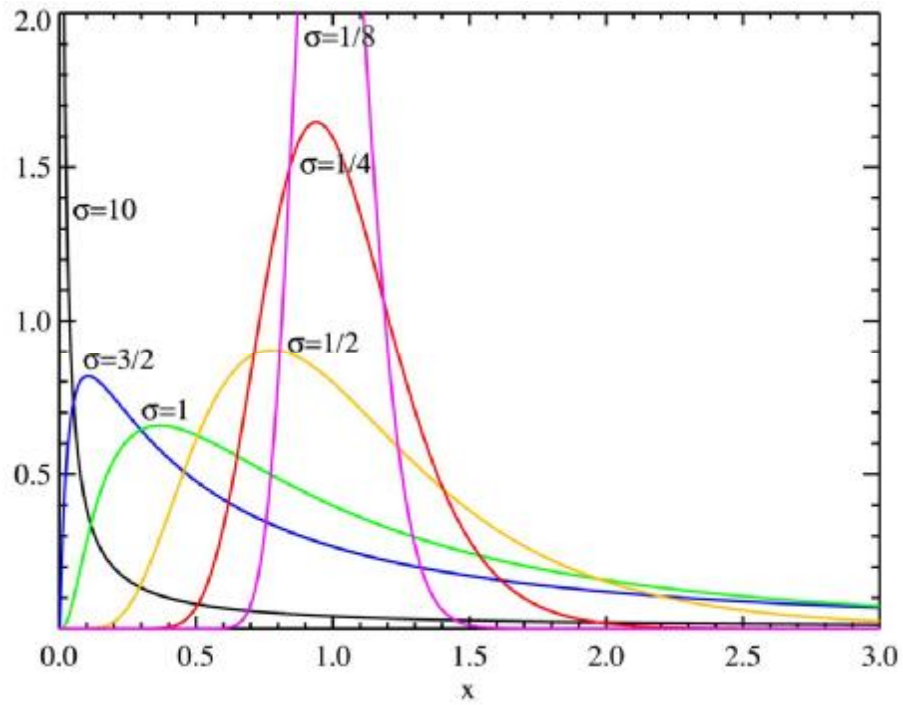
Логнормальное распределение (рис. 7) определено на интервале $(0;\infty)$. Если величина $\ln(x)$ подчиняется нормальному распределению, то x — логнормальному. Логнормальное распределение является теоретической моделью случайной величины, представляющей собой произведение

константы и стремящегося к бесконечности количества случайных величин (помех), распределённых по произвольным законам на интервале $(0; \infty)$.

Плотность вероятности логнормального распределения задаётся формулой

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2},\ x \in (0;\infty),$$

где μ — математическое ожидание величины $\ln(x)$, а σ — её среднеквадратическое отклонение. Математическое ожидание самой величины x составляет $e^{\mu+\sigma^2/2}$, а дисперсия — $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$.



Источник: http://ru.wikipedia.org

Рис. 7. Графики логнормального распределения при $\mu = 0$.

Функция логнормального распределения через элементарные функции не выражается. Она записывается следующим образом:

$$F(x) = \frac{1}{2} + \frac{1}{2} \cdot \mathrm{Erf}\left(\frac{\ln(x) - \mu}{\sigma\sqrt{2}} \right),$$

где

$$\mathrm{Erf}(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt.$$

Для вычисления функции плотности вероятности логнормального распределения в Excel при условии, что требуемое значение x хранится в ячейке под именем `Значение`, используйте формулу

```
=НОРМРАСП(LN(Значение);Средняя;СтандОткл;0),
```

где `Средняя` и `СтандОткл` — имена ячеек, содержащих значения μ и σ. Значение функции логнормального распределения (вероятности того, что нормально распределённое случайное значение не превысит указанную величину) вычисляется с помощью формулы

```
=НОРМРАСП(LN(Значение);Средняя;СтандОткл;1).
```

Определить величину, которую с заданной вероятностью не превысит нормально распределённое случайное значение, можно с помощью формулы

```
=EXP(НОРМОБР(Вероятность;Средняя;СтандОткл)),
```

где `Вероятность` — имя ячейки, содержащей требуемое значение вероятности.

В MathCad для аналогичных целей используйте формулы `dlnorm(x;μ;σ)`, `plnorm(x;μ;σ)` и `qlnorm(p;μ;σ)` соответственно, где используемые имена переменных имеют те же значения, что и в формуле плотности распределения.

Гамма-распределение

Гамма-распределение (рис. 8) описывает многие случайные величины, распределённые на интервале $[0; \infty)$. Оно представляет собой теоретическую модель суммы α независимых случайных величин, распределённых по экспоненциальному закону с одинаковым параметром, равным β. Функция плотности гамма-распределения:

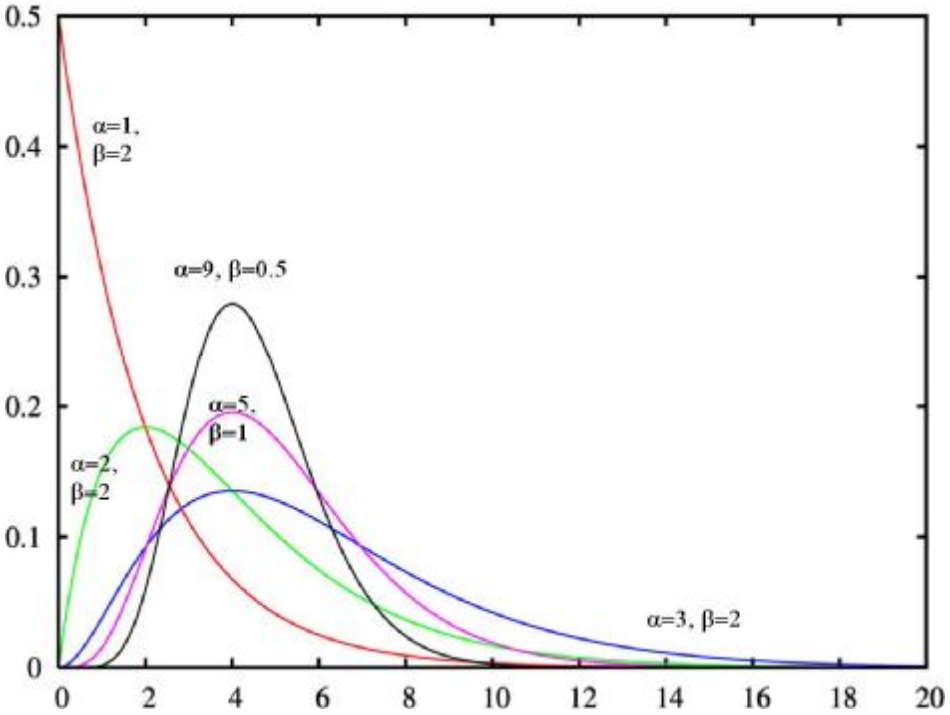
$$p(x) = x^{\alpha-1} \cdot \frac{e^{-\frac{x}{\beta}}}{\beta^{\alpha} \cdot \Gamma(\alpha)},$$

где

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx -$$

гамма-функция, значение которой для целых чисел равно факториалу её аргумента, уменьшенного на единицу; $e \approx 2{,}7182818$ — основание натурального логарифма; α и β — параметры, которые можно определить, зная математическое ожидание μ и дисперсию σ^2, по следующим формулам:

$$\alpha = \frac{\mu^2}{\sigma^2}; \; \beta = \frac{\sigma^2}{\mu}.$$



Источник: http://ru.wikipedia.org

Рис. 8. Графики гамма-распределения.

Частными случаями гамма-распределения являются экспоненциальное распределение (при $\alpha = 1$), распределение Эрланга (при натуральном α) и распределение χ^2 для n степеней свободы (при $\alpha = n/2$ и $\beta = 2$).

С помощью гамма-распределения можно (при наличии теоретических оснований) моделировать левоскошенные эмпирические распределения на интервалах $[c; \infty)$ и правоскошенные на интервалах $(-\infty; c]$, где c — произвольное действительное число. Для этого в формуле плотности распределения в первом случае x прибавляют к c, во втором — отнимают от c.

В Excel плотность распределения вероятности гамма-распределения для значения, хранящегося в ячейке `Значение`, вычисляется с помощью формулы

```
=ГАММАРАСП(Значение;
Средняя^2/Дисперсия;Дисперсия/Среднее;0),
```

где `Средняя` и `Дисперсия` — имена ячеек, содержащих соответствующие значения. Значение функции гамма-распределения (вероятности того, что случайное значение, распределённое по данному закону, не превысит указанную величину) вычисляется с помощью формулы

```
=ГАММАРАСП(Значение;
Средняя^2/Дисперсия;Дисперсия/Среднее;1),
```

Определить величину, которую с заданной вероятностью не превысит случайное значение, подчиняющееся гамма-распределению, можно с помощью формулы

```
=ГАММАОБР(Вероятность;
Средняя^2/Дисперсия;Дисперсия/Среднее),
```

где `Вероятность` — имя ячейки, содержащей требуемое значение вероятности.

В программе MathCad те же вычисления могут быть выполнены с помощью формул

$$\frac{\mu \cdot \mathrm{dgamma}\left(\frac{\mu x}{\sigma^2}; \frac{\mu^2}{\sigma^2}\right)}{\sigma^2},$$

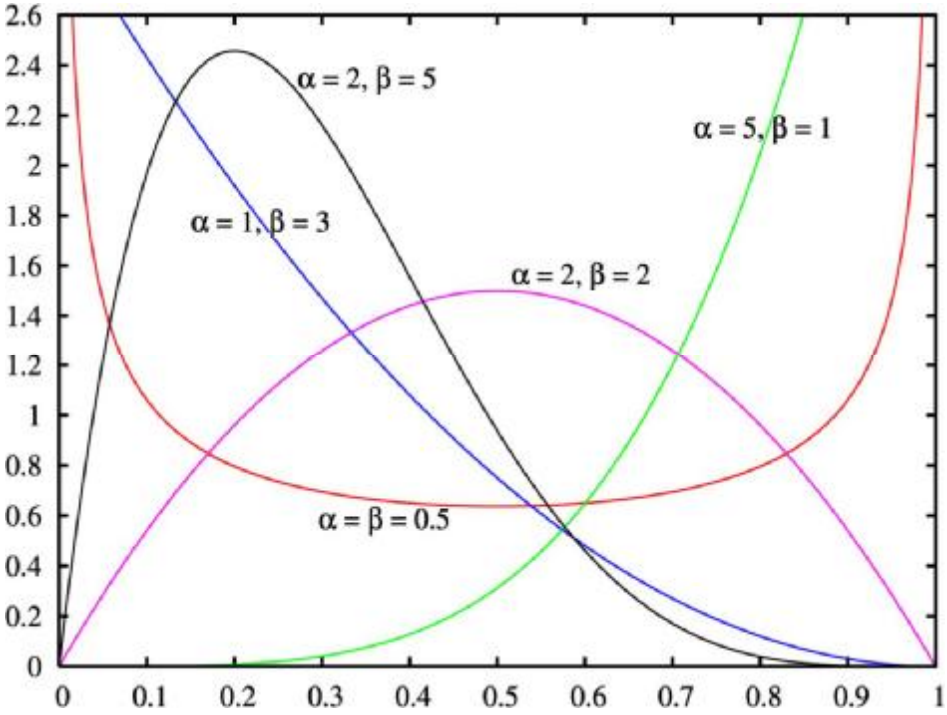
$$\mathrm{pgamma}\left(\frac{\mu x}{\sigma^2}; \frac{\mu^2}{\sigma^2}\right),$$

$$\frac{\sigma^2 \cdot \mathrm{qnorm}\left(p; \frac{\mu^2}{\sigma^2}\alpha\right)}{\mu},$$

где имена переменных соответствуют обозначениям в формуле плотности гамма-распределения.

Бета-распределение

Бета-распределение (рис. 9) определено на интервале [0; 1]. Оно является теоретической моделью случайной величины $A/(A+B)$, зависящей от двух других случайных величин A и B, каждая из которых подчиняется гамма-распределению. Часто бета-распределение является подходящей моделью для величины, представляющей собой долю (или процент) от целого — например, доли пашни в сельхозугодьях или степени использования производственного потенциала.



Источник: http://ru.wikipedia.org

Рис. 9. Графики бета-распределения.

Плотность бета-распределения задаётся функцией

$$p(x)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx},$$

где α и β — параметры, которые можно определить, зная математическое ожидание μ и дисперсию σ^2, по следующим формулам:

$$\alpha=\frac{\mu^2}{\sigma^2}-\frac{\mu^3}{\sigma^2}-\mu;\ \beta=\frac{\mu\cdot(\mu-1)^2}{\sigma^2}+\mu-1.$$

Равномерное распределение является частным случаем бета-распределения при α=1 и β=1.

Бета-распределение может быть использовано (при наличии теоретических оснований) для моделирования случайных величин, распределённых на произвольном отрезке [a; b], где a и b имеют содержательную интерпретацию[1]. Для этого нужно перенормировать исходную случайную величину y, распределённую на [a; b], по следующему правилу: $x = (y-a)/(b-a)$.

В Excel для вычисления плотности бета-распределения потребуется писать функцию на VBA. Функция бета-распределения может быть вычислена с помощью формулы

```
=БЕТАРАСП(Значение;Альфа;Бета;Начало;Конец),
```

где в ячейке под именем `Значение` хранится значение случайной величины y, в ячейке `Альфа` — параметр α, в ячейке `Бета` — параметр β, в ячейке `Начало` — значение a, в ячейке `Конец` — значение b. Перенормирование величины y производится автоматически.

Определить значение y по заданной веротяности того, что оно не будет превышено (предположим, оно записано в ячейку под именем `Вероятность`), можно с помощью формулы

```
=БЕТАОБР(Вероятность;Альфа;Бета;Начало;Конец).
```

[1] Например, если коровы массой менее 400 и более 520 кг выбраковываются из основного стада, то при проверке гипотезы о согласии распределения живой массы коров с бета-распределением значения a=400, b=520 будут приняты обоснованно. Если же верхняя граница массы для выбраковки не установлена, достаточных оснований для моделирования эмпирического распределения живой массы с помощью бета-распределения нет.

Встроенные функции MathCad не предусматривают перенормирование случайной величины — оно должно быть выполнено заранее. Плотность бета-распределения вычисляется с помощью формулы

```
dbeta(x;α;β),
```

где обозначения соответствуют использованным в формуле плотности распределения. Вероятность непревышения заданной величины определяется по формуле

```
pbeta(x;α;β),
```

а обратное вычисление —

```
qbeta(p;α;β),
```

где переменная `p` содержит пороговую вероятность. Поскольку результат представляет собой перенормированное значение, получить исходное значение y можно при помощи следующей формулы:

```
qbeta(p;α;β)·(b-a)+a,
```

полагая, что границы a и b хранятся в одноимённых переменных программы MathCad.

2. Проверка согласованности эмпирического и теоретического распределений с помощью критерия χ^2

Как правило, критерий χ^2 имеет практическое значение для совокупностей численностью не менее 40 наблюдений. Для применения данного критерия интервал вариации случайной величины разбивается на непересекающиеся классы. О согласии теоретического и эмпирического распределений судят по наблюдаемым различиям в частоте попадания наблюдений в каждый класс по сравнению с частотой, которая должна бы была иметь место, если бы распределение в точности соответствовало теоретическому. Если различия настолько велики, что с достаточно высокой вероятностью[1] (обычно в экономических исследованиях требуют, чтобы она

[1] Эту пороговую вероятность называют *уровнем доверия*, или *доверительной вероятностью*.

была не менее 95%, при остром недостатке данных — не менее 90%[1]) не могли бы возникнуть, если бы распределение случайной величины соответствовало предполагаемому закону, — гипотезу о согласии эмпирического распределения с выбранным теоретическим отвергают.

В противном случае считают, что расхождение с предлагаемой теоретической моделью не доказано с достаточной степенью надёжности; а значит, нет оснований ставить под сомнение те теоретические соображения, на основе которых выдвинута гипотеза о законе распределения — по крайней мере, до тех пор, пока новые, более полные, данные не придут в противоречие с нею.

Выдвигая гипотезу о распределении, принимают во внимание следующие сведения (в меру их доступности):

- область определения случайной величины;
- происхождение данной случайной величины;
- моменты распределения и их соотношение;
- форму гистограммы;
- результаты моделирования данной случайной величины, полученные другими исследователями;
- аналогии с другими случайными величинами, распределение которых установлено;
- численность наблюдений.

В качестве области определения случайной величины не следует принимать наблюдаемый диапазон вариации (иначе у нас никогда не оказалось бы оснований для использования нормального распределения). Её определяют исходя из сущности процесса или явления, отражаемого случайной величиной. Например, урожайность культуры не может быть ниже нуля; существует также её объективный верхний предел, зависящий от массы гумуса в почве. Поэтому для её моделирования может подойти какое-либо распределение, определённое на интервале $[0; b]$ — например, бета или (при недостатке данных) треугольное. При этом величину b, раз она неизвестна, можно определить подбором, добиваясь наилучшего согласия опытных данных с теоретическим распределением.

Можно ли использовать для моделирования урожайности, например, гамма-распределение? Очевидно, что в действительности урожайность не может соответствовать этому распределению, так как она в принципе не может быть сколь угодно большой. Но с некоторой степенью грубости

[1] В последнем случае результаты обычно требуют перепроверки с привлечением новых наблюдений.

гамма-распределение может оказаться *практически приемлемой* моделью, если оценённая по гамма-распределению (то есть теоретическая) вероятность значений урожайности, превышающих фактически наблюдаемые, пренебрежимо мала. То же касается нормального распределения, но тогда пренебрежимо мала должна быть также теоретическая вероятность отрицательных значений урожайности. Последнее часто не выполняется.

Если, кроме наблюдений, нет никаких оснований для выбора распределения, то следует отдавать предпочтение самым простым распределениям с наименьшим числом параметров. Если к тому же наблюдения малочисленны, лучше пользоваться такими распределениями, как равномерное и треугольное. Результаты, полученные при подобных обстоятельствах, требуют перепроверки в дальнейшем.

Параметры гипотетических распределений, если только они не известны заранее из теоретических соображений, определяют, когда возможно, на основе моментов эмпирического распределения (средней и дисперсии)[1], а когда невозможно — подбором.

После того, как гипотеза сформулирована, можно приступать к её проверке. Процедура проверки по критерию χ^2 предполагает следующие этапы:

- разбиение интервала вариации на непересекающиеся классы;
- определение численности наблюдений эмпирического распределения, приходящихся на каждый класс;
- определение теоретической численности наблюдений в соответствии с выбранной моделью случайной величины;
- расчёт значения критерия χ^2;
- определение критического уровня χ^2 для заданной доверительной вероятности;
- сравнение фактического и критического значений χ^2 и заключение о том, следует ли отвергнуть предложенную теоретическую модель распределения случайной величины.

Рассмотрим каждый из этих этапов.

Считается, что практически приемлемый компромисс между численностью классов и численностью наблюдений в каждом классе достигается, если число классов определять по формуле $\sqrt{N}$, где N — число наблюдений, а ширину классов принимают равной. Чтобы обеспечить приемлемую вероятность ошибки при расчёте значения χ^2, необходимо следить

[1] См. формулы для определения значений параметров распределений при известных средней и дисперсии в Приложении 1.

за тем, чтобы как фактическая, так и теоретическая численность наблюдений в каждом классе была не меньше 6...8. Если это не выполняется, малочисленные классы объединяют; при этом численность классов не должна оказаться меньше пяти. В случае невыполнимости этих требований критерию χ^2 доверять нельзя[1]. Если данная процедура порождает очень много пустых классов, а случайная величина строго положительна, то целесообразно перейти к исследованию распределения её логарифмов.

Численность наблюдений, относящихся к каждому классу, обычно определяется по ранжированному ряду наблюдаемых данных с помощью функции Excel `=СЧЁТЕСЛИ(Ряд,Условие)`.

Теоретическая численность наблюдений для каждого класса определяется как $(F(x_2) - F(x_1))\cdot N$, где $F(\cdot)$ — функция выбранного теоретического распределения, N — число имеющихся наблюдений, x_2 и x_1 — соответственно верхняя и нижняя границы класса.

Значение критерия χ^2 рассчитывается по формуле

$$\sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i},$$

где k — число классов, n_i — число фактических наблюдений в классе i, n'_i — теоретическая численность наблюдений в классе i. При различных разбиениях на классы значение χ^2 оказывается различным, но при выполнении требований к числу наблюдений всего и в каждом классе, сформулированных выше, вероятность статистически существенных различий невелика.

Критическое значение может быть определено с помощью формулы Excel

```
=ХИ2ОБР(1-УровеньДоверия;СтепениСвободы),
```

где в ячейке `УровеньДоверия` содержится требуемая доверительная вероятность (выраженная в долях, а не в процентах), а в ячейке `СтепениСвободы` — величина, равная числу классов за вычетом увеличенного на единицу числа параметров теоретического распределения, определённых с использованием эмпирических данных. В MathCad аналогичный расчёт выполняется с помощью формулы

[1] В учебных заданиях данного практикума разрешается смягчать эти требования в соответствии с указаниями преподавателя, обязательно отмечая в отчёте, что результат проверки гипотезы о согласии теоретического и эмпирического распределений недостоверен по причине недостаточной численности имеющихся наблюдений.


```
qchisq(1-УровеньДоверия;СтепениСвободы).
```

Если значение χ^2 превышает критическое, гипотезу о согласии распределений *отвергают* с выбранным уровнем доверия. В противном случае гипотеза *не отвергается* (что, разумеется, не означает её безусловной истинности: быть может, этот результат случаен, а может, действительное распределение мало отличается от гипотетического).

Расчёты по проверке согласованности теоретического и эмпирического распределений рекомендуется выполнять в таблице, строки которой (кроме итоговой) соответствуют классам, а столбцы — этапам вычислений. В частности, в ней должны быть представлены величины n_i, n'_i и $(n_i - n'_i)^2 / n'_i$.

3. Проверка статистических гипотез относительно многовершинных распределений

Многовершинность эмпирического распределения обычно свидетельствует о смешении совокупностей с разными качественными характеристиками. Строгий подход к исследованию таких совокупностей состоит в отыскании критерия, по которому наблюдения можно отнести к каждой из качественно различных совокупностей, которые затем исследуются отдельно. В частности, для каждой из них формулируется и проверяется отдельная гипотеза о распределении вероятностей значений исследуемых переменных.

Распределения наблюдений по качественно различающимся совокупностям необходимо выполнять всегда, когда имеется возможность для этого.

На этапе системного анализа часто отсутствуют данные, необходимые для выполнения такой процедуры. Возможны две ситуации: либо отсутствуют данные о показателях, необходимых для построения критерия отнесения наблюдения к различным совокупностям, либо наблюдений слишком мало, так что после классификации они вообще не будут поддаваться анализу.

В подобных случаях совокупность разбивают в точках минимума между вершинами, после чего для получившихся совокупностей выдвигают гипотезы о распределениях, не подвергая их проверке. В результате получают функции распределения $F_1(x)$, $F_2(x)$ и т.д.

Далее формулируют функцию вида

$$\frac{1}{N}\sum_{k=1}^{n} N_i F_i(x),$$

где N — число наблюдений всего, N_i — число наблюдений в совокупности i, n — число совокупностей (на одну меньше числа вершин).

Затем выдвигается гипотеза, что исследуемая случайная величина имеет данную функцию распределения. Затем она проверяется в обычном порядке по критерию χ^2, только для определения теоретических частот вместо обычной $F(x)$, соответствующей одному из известных распределений, используется данная функция, а при расчёте числа степеней свободы учитывается общее количество параметров, определённых на основе эмпирического распределения для всех $F_i(x)$.

4. Проверка независимости факторов с помощью критерия χ^2

Критерий χ^2 очень удобен для проверки независимости двух дискретных переменных. Если имеется набор наблюдений, в каждом из которых зафиксировано значение двух дискретных переменных, такой, что каждой паре значений дискретных переменных *теоретическая* частота, составляющая не менее 6-8 наблюдений, то с помощью данного критерия можно, не привлекая никаких других теоретических соображений, сделать заключение о том, проявляется ли *какая-либо* зависимость между этими переменными в имеющихся результатах наблюдений.

При достаточной численности наблюдений данный критерий наилучшим образом соответствует целям практического задания к теме 3 при проверке независимости переменных. Если гипотеза о независимости двух факторов отвергается, один из них должен быть исключён из модели и заменён другим. Если гипотеза о независимости результата от фактора не отвергается, фактор также следует исключить из модели, заменив его другим.

Процедура проверки предполагает следующие этапы:

- подсчёт числа наблюдений, для каждого сочетания значений двух переменных;
- подсчёт теоретической частоты n'_{ij} для каждого сочетания значений двух переменных, составляющей $n_{1i} \cdot n_{2j}/N$, где n_{1i} — число наблюдений

i-го значения первой переменной, n_{2j} — число наблюдений j-го значения второй переменной;

- расчёт значения критерия χ^2 по формуле

$$\sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\frac{(n_{ij}-n'_{ij})^2}{n'_{ij}},$$

где k_1 — число значений первой переменной; k_2 — число значений второй переменной; n_{ij} — фактическое число наблюдений, при которых первая переменная принимала значение i, а вторая — значение j; остальные обозначения прежние;

- определение критического уровня χ^2 для заданной доверительной вероятности и числа степеней свободы $(k_1-1)\cdot(k_2-1)$ — например, с помощью формулы Excel

```
=ХИ2ОБР(1-УровеньДоверия;(_k1-1)*(_k2-1)),
```

где в ячейке `УровеньДоверия` содержится требуемая доверительная вероятность (выраженная в долях, а не в процентах), в ячейках `_k1` и `_k2` — число значений соответствующих дискретных переменных. В MathCad аналогичный расчёт выполняется с помощью формулы

```
qchisq(1-УровеньДоверия;(k1-1)*(k2-1));
```

- сравнение фактического и критического значений χ^2 и заключение о том, следует ли отвергнуть предложенную теоретическую модель распределения случайной величины.

Если значение χ^2 превышает критическое, гипотезу о независимости факторов *отвергают* с выбранным уровнем доверия. В противном случае гипотеза *не отвергается* (что, разумеется, не означает её безусловной истинности: быть может, этот результат случаен).

Расчёты по проверке независимости факторов рекомендуется выполнять в таблице, строки которой (кроме итоговой) соответствуют комбинациям значений двух исследуемых переменных, а столбцы — этапам вычислений. В частности, в ней должны быть представлены величины n_{ij}, n'_{ij} и $(n_{ij} - n'_{ij})^2 / n'_{ij}$.

5. Проверка существенности связи между переменными с помощью однофакторного дисперсионного анализа

Однофакторный дисперсионный анализ проверяет гипотезу о равенстве дисперсий некоторой *нормально распределённой* переменной в нескольких выборках. Отклонение этой гипотезы указывает, что различие между выборками заведомо не случайно, и тем самым выявляет существование зависимости между признаком, по которому осуществлялись выборки, и данной переменной.

Таким образом, он может быть использован для проверки наличия существенной связи между двумя переменными, из которых по крайней мере одна дискретна, а другая подчиняется нормальному закону распределения. Практически приемлемые результаты достигаются также для случая гамма-распределения: доверять им можно тем в большей степени, чем меньше его асимметрия.

Для выполнения однофакторного дисперсионного анализа в Excel следует расположить значения нормально распределённой переменной (она может быть как непрерывной, так и дискретной, но, разумеется, числовой; следовательно, процедуру можно проводить как до, так и после дискретизации переменной, выступающей в качестве зависимой), соответствующие разным значениям дискретного влияющего фактора (он может быть как числовым, так и нечисловым), в соседних столбцах. Число значений переменной в разных столбцах может быть различным. Над каждым столбцом указывают соответствующее значение влияющего фактора.

Далее следует подключить надстройку «Анализ данных» (если она не подключена) и дать команду **Сервис → Анализ данных** либо **Данные → Анализ данных**, смотря по версии программы. В качестве входного нужно указать интервал, охватывающий все ячейки со значениями нормально распределённой переменной и притом не содержащий никаких других текстовых или числовых данных, кроме меток влияющего фактора в его первой строке. Переключатели **Группирование: по столбцам** и **Метки в первой строке** должны быть включены. Выходной интервал указывается таким образом, чтобы выводимые в него данные не перезаписали уже имеющиеся (рекомендуется выводить результаты на новый лист).

Если по результатам анализа *p*-значение (уровень значимости) оказалось ниже величины[1], дополняющей желаемый уровень доверия до единицы (например, меньше 0,05), то гипотеза о равенстве дисперсий переменной при разных значениях влияющего фактора отвергается, что означает наличие связи между ним и нормально распределённой зависимой переменной.

Применяя дисперсионный анализ в целях практикума, следует иметь в виду, что в качестве влияющей переменной всегда выбирается входная, а в качестве зависимой (нормально распределённой) может быть использована как входная, так и выходная переменная. Основаниями для исключения входной переменной из модели могут быть:

- невозможность отвергнуть гипотезу о равенстве дисперсий выходной переменной при разных значениях данной входной переменной[2];
- отвергнутая гипотеза о равенстве дисперсий одной входной переменной при разных значениях другой.

В процедурах системного анализа, выполняемого по данной методике, нет необходимости использовать многофакторный дисперсионный анализ, более требовательный к числу наблюдений, так как формализм условных вероятностей требует независимости входных переменных. При данных обстоятельствах процедура однофакторного дисперсионного анализа даёт достаточные основания для принятия решения о наборе переменных, включаемых в модель.

6. Процедура расчёта энтропии, снимаемой с переменной информацией о значении другой переменной

Полная энтропия зависимой дискретной переменной на основе имеющихся эмпирических данных рассчитывается следующим образом:

- если исходные данные по переменной дискретны — по формуле

[1] Алгоритм расчёта приведён, например, в издании: Красс М.С., Чупрынов Б.П. Математические методы и модели для магистрантов экономики: Учеб. пособие. СПБ.: Питер, 2006. — С. 171-172.

[2] При большом числе входных переменных влияние каждой из них может быть весьма слабым. В этом случае при использовании однофакторного дисперсионного анализа в целях определения набора входных переменных, включаемых в модель, следует использовать уровни доверия, очень близкие к единице.

$$H = \sum_{i=1}^{k} (-p_i \log_2 p_i),$$

где $p_i = (n_i+1)/(N+k)$ — оценка вероятности i-го дискретного значения зависимой переменной; k — число дискретных значений зависимой переменной; n_i — число наблюдений i-го дискретного значения зависимой переменной; N — общее число наблюдений;

♦ если проводилась дискретизация переменной путём разбиения на квантили — по формуле $\log_2 k$, где k — число квантилей.

Остаточная энтропия зависимой дискретной переменной при поступлении информации о j-м состоянии влияющей дискретной переменной вычисляется по формуле

$$H_j = \sum_{i=1}^{k} (-p_{ij} \log_2 p_{ij}),$$

где $p_{ij} = (n_{ij}+1)/(N_j+k)$ — оценка вероятности i-го дискретного значения зависимой переменной при j-м значении влияющей переменной; k — число дискретных значений зависимой переменной; n_{ij} — число наблюдений i-го дискретного значения зависимой переменной при j-м значении влияющей переменной; N_j — число наблюдений j-го значения влияющей переменной.

Средняя информативность влияющей переменной относительно данной зависимой переменной составляет

$$I = H - p_j \sum_{j=1}^{l} H_j,$$

где p_j — оценка вероятности j-го дискретного значения влияющей переменной, получаемая аналогично оценке для зависимой переменной.

Решение об исключени входной переменной из модели принимают в следующих случаях:

♦ если в качестве зависимой переменной принимается выходная — если величина I/H меньше величины α/Q, где Q — число входных переменных, а параметр надёжности α, не превышающий 1, выбирается субъективно[1]. Чем больше его значение, тем труднее выполнить требования к переменной, включаемой в модель;

♦ если в качестве зависимой переменной принимается входная — если величина I/H больше α.

7. Некоторые полезные статистические функции табличного процессора Microsoft Excel

`=ДИСП(Ряд)`

Вычисляет дисперсию выборочных данных, содержащихся в интервале `Ряд`.

`=ДИСПР(Ряд)`

Вычисляет дисперсию генеральной совокупности данных, содержащейся в интервале `Ряд`.

`=ДОВЕРИТ(Значимость;СтандОткл;ЧислоНаблюдений)`

Вычисляет одностороннюю предельную ошибку среднего для нормально распределённой совокупности данных для уровня доверия, равного `(1-Значимость)`, при заданных среднеквадратичном отклонении `СтандОткл` и численности наблюдений `ЧислоНаблюдений`.

`=КОРРЕЛ(Ряд1;Ряд2)`

Вычисляет коэффициент парной линейной корреляции по Пирсону для двух совокупностей данных, содержащихся в интервалах `Ряд1` и `Ряд2`. Число ячеек в обоих рядах должно быть одинаковым. Все они должны содержать числовые данные (пустые ячейки не допускаются).

`=МАКС(Ряд)`

Находит наибольшее значение среди данных, содержащихся в интервале `Ряд`.

`=МЕДИАНА(Ряд)`

Находит медиану совокупности данных, содержащихся в интервале `Ряд`.

`=МИН(Ряд)`

Находит наименьшее значение среди данных, содержащихся в интервале `Ряд`.

[1] Для целей данного практикума можно принять его равным 0,3.

=МОДА(Ряд)

Находит модальное значение совокупности данных, содержащихся в интервале Ряд, если таковое существует.

=НАИБОЛЬШИЙ(Ряд;Ранг)

Находит среди данных в интервале Ряд значение, имеющее порядковый номер Ранг, если значения пронумеровать в порядке убывания.

=НАИМЕНЬШИЙ(Ряд;Ранг)

Находит среди данных в интервале Ряд значение, имеющее порядковый номер Ранг, если значения пронумеровать в порядке возрастания.

=ПЕРСЕНТИЛЬ(Ряд;Персентиль)

Находит значение, которое вместе с другими не превышающими его значениями образует требуемую Персентиль (указываемую в долях) совокупности данных в интервале Ряд.

=РАНГ(Число;Ряд;Порядок)

Определяет ранг значения Число в совокупности данных, содержащейся в интервале Ряд, по возрастанию (если значение Порядок равно нулю либо опущено) или по убыванию (если значение Порядок указано и не равно нулю). Значение Число обязательно должно присутствовать в интервале Ряд.

=СКОС(Ряд)

Вычисляет коэффициент асимметрии для эмпирического распределения, представленного данными в интервале Ряд.

=СРЗНАЧ(Ряд)

Вычисляет среднее арифметическое по данным интервала Ряд.

=СРЗНАЧЕСЛИ(Ряд,Условие)

Вычисляет среднее арифметическое для данных интервала Ряд, отвечающих критерию Условие. Критерий представляет собой текст вида ">2", "<-3,14159", где число может быть произвольным, либо ссылку на ячейку, содержащую формулу, результатом вычисления которой является подобное текстовое значение.

=СРЗНАЧЕСЛИМН(Ряд,Условия)

Вычисляет среднее арифметическое для данных интервала Ряд, отвечающих одновременно всем критериям, хранящимся в интервале Условия. Каждый критерий представляет собой текст вида ">2", "<-3,14159", где число может быть произвольным. Поддерживается не всеми версиями Excel.

=СТАНДОТКЛОН(Ряд)

Вычисляет среднеквадратическое отклонение выборочных данных, содержащихся в интервале Ряд.

=СТАНДОТКЛОНП(Ряд)

Вычисляет среднеквадратическое отклонение данных генеральной совокупности, содержащейся в интервале Ряд.

=СЧЁТ(Ряд)

Определяет число значений в интервале Ряд.

=СЧЁТЕСЛИ(Ряд;Условие)

Определяет число значений в интервале Ряд, отвечающих критерию Условие. Критерий представляет собой текст вида ">2", "<-3,14159", где число может быть произвольным, либо ссылку на ячейку, содержащую формулу, результатом вычисления которой является подобное текстовое значение.

=СЧЁТЕСЛИМН(Ряд;Условия)

Определяет число значений в интервале Ряд, отвечающих одновременно всем критериям, хранящимся в интервале Условия. Каждый критерий представляет собой текст вида ">2", "<-3,14159", где число может быть произвольным. Поддерживается не всеми версиями Excel.

=ЧАСТОТА(РядДанных;Границы)

Вычисляет массив значений, каждое из которых означает число наблюдений из интервала РядДанных, относящихся к классу, задаваемому данными в интервале Границы.

Для использования функции следует выделить на одну ячейку больше, чем содержится их в интервале Границы, набрать содержащую её формулу и нажать сочетание клавиш [Ctrl]+[Shift]+[Enter]. В первой ячейке выделенного интервала отобразится число значений, которые не больше первого значения в интервале Границы; во второй — число значений между первым и вторым значениями в интервале Границы (исключая нижнюю границу и включая верхнюю) и т.д.; в последнем — значения, превышающие наибольшее значение в интервале Границы.

Значения в интервале Границы должны быть упорядочены по возрастанию. Пустые ячейки и текстовые значения игнорируются.

=ЭКСЦЕСС(Ряд)

Вычисляет коэффициент эксцесса для эмпирического распределения, представленного данными в интервале Ряд.

8. Численное интегрирование

Необходимость вычисления определённых интегралов при решении задач системного анализа по методике, положенной в основу настоящего практикума, возникает, например, при определении ошибки оценки вероятности события по результатам наблюдений, при отыскании квантилей либо (в некоторых случаях) при проверке гипотезы о законе распределения случайной величины.

Для вычисления определённых интегралов в MathCad достаточно ввести требуемый интеграл в виде формулы. Чтобы ввести знак интеграла, следует нажать клавишу [&]. Например, вычисление формулы

$$\int_{-\infty}^{10} \mathrm{dnorm}(x,5,2)\,dx$$

даст тот же результат, что и формулы pnorm(10,5,2), а именно 0,99379.

Excel не имеет встроенных возможностей численного интегрирования. Если лабораторные работы выполняются в Excel, вычисление определённых интегралов можно осуществлять любым известным методом, например, методом трапеций или методом Симпсона. Описание соответствующих алгоритмов можно найти в сети Интернет либо в учебной литературе по численным методам[1].

[1] Численные методы / Н.С. Бахвалов, Н.П. Жидков, Г.М. Кобельков. 4-е изд. М.: БИНОМ. Лаборатория знаний, 2006.

СОДЕРЖАНИЕ