ГОСУДАРСТВЕННЫЙ КОМИТЕТ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПО ВЫСШЕМУ ОБРАЗОВАНИЮ

МАТИ им. К.Э.Циолковского - ГОСУДАРСТВЕННЫЙ
АВИАЦИОННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра высшей математики

Горбацевич В.В.

Математическая статистика для студента с компьютером

Методическое пособие

к чтению лекций и проведению практических и
лабораторных работ по математической статистике
с использованием универсальных и специализированных
математических пакетов программ

Ч А С Т Ь **I**

МОСКВА **2004** год

ВВЕДЕНИЕ

Компьютеры в последние годы все интенсивнее вторгаются в и производство и в нашу обыденную жизнь. При этом неудивительно, что вторжение их в производство вызывает необходимость изменения методики преподавания многих дисциплин в инженерных вузах. Более того, именно такого рода изменения и могут дать возможность успешно использовать огромные возможности современных компьютеров для повышения эффективности и результативности производства на всех его этапах - от предварительного анализа проектов и прикидочного проектирования до выходного контроля, испытания и учета продукции.

Особую роль в процессах производства сейчас играют статистические методы. Именно они дают возможность давать обоснованные рекомендации там, где развитие современной теории не позволяет пока получать однозначные результаты и приходится анализировать события случайного характера. С увеличением сложности изготовляемой продукции число таких факторов, не поддающихся однозначному определению, неуклонно растет. А потому все более важными становятся методы теории вероятностей и математической статистики, которые как раз и приспособлены для работы с такого рода информацией.

До недавнего времени преподавание теории вероятности и математической статистики даже в прикладных, инженерных ВУЗах было довольно сильно удалено от решения реальных прикладных, практических задач. Студенты, даже успешно сдавшие экзамен по этим разделам математики, почти всегда оказывались беспомощными при решении реальных прикладных задач. Это было вызвано, в частности, тем, что на практических занятиях решались задачи по теории вероятностей и математической статистике, по формулировкам как бы приближенные к практике, но на самом деле находящихся от этой практики очень далеко. И дело тут не только (и даже не столько) в нежелании преподавателей математики рассматривать задачи чисто прикладного характера. Более серьезная причина - такого рода задачи требуют для своего решения значительных вычислений, которые при ограниченной в самом недавнем

прошлом вычислительной базе - в лучшем случае это был программируемый калькулятор - решить за приемлемое для учебного занятия время не удается. Более того, занятия по математической статистике вообще страдали от почти полного отсутствия задач, решение которых можно было бы провести до конца на занятии или хотя бы превратить в домашнее задание студента без риска загрузить его часами рутинных вычислений.

В результате получалось, что до самого последнего времени изучение теории вероятностей и математической статистики носило с точки зрения практики совершенно неудовлетворительной характер, что, правда, не мешало некоторым выпускникам на основе полученных за время обучения в ВУЗе теоретических знаний все же получать на производстве возможность освоить достаточно современные методы использования математической статистики в нужном ему направлении.

В самые последние годы ситуация с преподаванием теории вероятностей и математической статистики в ВУЗах кардинально начинает меняться. Во-первых, это вызвано тем, что штампование большого числа дипломированных, но праздношатающихся на производстве людей стало просто экономически невыгодным (хотя не всеми вузами это еще осознано в одинаковой степени). Во-вторых, все большее число семей имеют дома персональные компьютеры. По опросам (например, студентов МАТИ) видно, что практически все современные студенты имеют непосредственный доступ к компьютерам, а немалая часть их – и к сети Интернет. Компьютерная грамотность нашего населения растет стремительными темпами. По уровню освоения сложных программ наша страна до сих пор находится на передовых позициях в мире (даже в сравнении с самыми развитыми странами). Скорость, с которой наши студенты могут осваивать самые сложные программы, если они видят этому реальные применения, просто поразительна.

Преподавание теории вероятностей и математической статистики при ориентации на использование компьютеров должна претерпеть значительные изменения. Годами установившиеся программы занятий должны быть пересмотрены в сторону приближения их к нуждам практических применений и вычислений, доводящих до конкретного результата (в той или иной форме).

На кафедре высшей математики МАТИ несколько лет ведется работа в направлении внедрения прикладных компьютерных методов в реальный учебный процесс. Одним из результатов этой работы является разработка программы курса "Универсальные и специализированные математические пакеты для решения задач математической статистики", предназначенного для специальности МСС.

Исходными положениями при разработке этой программы были следующие:

1. Необходимо сохранить (даже при самом минимальном числе выделяемых на курс учебных часов) преподавание основ теории вероятностей. В противном случае работа с универсальными программами превращается в бездумный автоматический процесс, дающий сбой при первом же нестандартном результате. Тенденция сильно уменьшать курс теории вероятностей за счет расширения курса математической статистики и применения универсальных пакетов наблюдается сейчас в ряде стран Запада и Японии, и уже видны первые результаты такого перекоса - при необходимости проанализировать нестандартный ответ в этих странах часто рассчитывают на помощь находящих поблизости выпускников советских или российских вузов.

2. Преподавание математической статистики нужно вести на самом современном уровне (а не на уровне простейших, элементарных методов, которые в основном и описываются в стандартных учебниках), так как именно такие методы и реализованы в современных математических программах.

3. Значительное внимание нужно уделить процессу подготовки статистических данных и их вводу или импорту в ту или иную программу (возможно, с преобразованием формата данных). Это связано с необходимостью автоматизировать максимальное число самых различных рутинных операций при проведении статистических исследований.

4. Так как вся вычислительная работа в пакетах выполняется практически мгновенно, то ныне появляется возможность позволить себе строить не одну модель, а несколько моделей разного рода, в том числе и включающих довольно большое число параметров (это в прошлом было недостижимой мечтой большого числа специалистов-статистиков).

5. Особое внимание при перестройке процесса преподавания нужно уделить этапу анализа полученных результатов. Нужно учить студентов не только правильно ставить задачу для статистического исследования, но и правильно истолковывать тот объем информации, который практически без усилий он получает в результате работы программы.

6. Не стоит ограничивать изучение статистических пакетов какой-либо одной программой (даже если она и обладает совершенно выдающимися на данный момент времени характеристиками). Развитие программного обеспечения сейчас идет очень быстро, программы непрерывно совершенствуются и потому будущий специалист должен освоить не столько какую-то определенную программу, сколько саму методику использования компьютера в статистическом исследовании.

7. При рассмотрении конкретных примеров решения практических задач методами математической статистики следует выбирать такие задачи, решение которых представляется хоть какой-то интерес для студентов. Если речь идет о студентах старших курсов, то нужно просто брать задачи, близкие к их специальности (войдя в контакт с представителями соответствующей профилирующей кафедры). Для студентов же первого-второго курсов можно брать задачи общенаучного характера (с элементами занимательности или познавательности) или же просто познавательные задачи, взятые, например, из справочников или из повседневной жизни студента.

8. Необходимо ознакомить студентов с современными стандартами в обработке данных и с используемой при этом терминологией (в том числе и англоязычной, которая весьма существенно отличается от русскоязычной).

9. Полезно подсказать студентом, как можно использовать полученные знания при выполнении курсовых и дипломных работ на профилирующих кафедрах. Возможно, что эти подходы окажутся новыми, непривычными, но интересными и полезными, и для некоторых преподавателей этих кафедр, так что польза от их использования может быть обоюдная. Полезно также расширить кругозор студента и обрисовать ему перспективы изучаемых методов математической статистики за пределами его нынешней специализации, тем

самым давая ему возможность более легкой адаптации к меняющимся социальным и производственным условиям труда.

В соответствии с этими соображениями и была разработана приведенная ниже программа курса. Этот курс можно рассматривать как минимально возможный для обучения общим сведениям о математической статистике и одной специальной теме - методике нахождения и анализа закономерностей между двумя и большим числом переменных. Именно эта специальная тема была сформулирована выпускающей кафедрой МСС как максимально необходимая для студентов их специальности. Для других специальностей и специализаций темы лекций следует, конечно, видоизменить. Однако общую схему современного построения курса теории вероятностей и математической статистики нужно, как нам кажется, реализовывать именно в таком (или близком к такому) виде, как в прилагаемой программе.

В качестве универсальных пакетов, содержащих основные методы математической статистики, при проведения практических занятий были выбраны Excel (составляющая стандартного пакета Microsoft Office), а также широко распространенный математический пакет MathCad. Из множества специализированных статистических пакетов использовались StatGraph (на первых этапах преподавания) и отечественный пакет Stadia. Последний пакет имеет ряд преимуществ по сравнению с зарубежными образцами – он достаточно компактен, содержит широкий набор статистических процедур и полностью русифицирован, что избавляет (возможно, что и напрасно) от острой необходимости параллельно с изучением курса математической статистики дополнительно изучать и английский язык. Было признано желательным знакомить студентов с несколькими разными пакетами программ (в том числе и не русифицированными), так как в полученные навыки освоения различных программ могут пригодиться ему в будущем при освоении новых, ныне еще неизвестных пакетов программ. Ниже приводится расширенный указанием на многие конкретные темы и подтемы текст программы курса. Комментарии к некоторым разделам программы, не нашедшие своего отражения в учебной литературе, будут приведены во второй части данного пособия.

ПРОГРАММА

курса "Универсальные и специализированные математические пакетов программ для решения задач математической статистики" для студентов кафедры МСС

Всего учебных часов 32,
из них лекции - 20 часов,
практические занятия (в дисплейном классе) - 12 часов

Лекции

1. а). Примеры задач математической статистики (в технике, науке, общественной жизни и в быту). Возможности использования PC и современных компьютерных программ.
Эволюция практического использования методов математической статистики.

б). Основы теории вероятностей (обзор изученного ранее курса).

Понятие события, условие повторяемости. Алгебра событий, случайные события. Несовместные события. Вероятность как мера случайности события, ее свойства. Равновероятные события, классическое определение вероятности. Геометрические вероятности.

Случайные величины, дискретные и непрерывные. Способы задания случайных величин (закон распределения, функция распределения, плотность распределения), их свойства. Зависимость и независимость случайных событий, случайных величин. Меры зависимости.

Примеры законов распределения (биномиальный, Пуассона) и плотностей (равномерная, нормальная, Стьюдента, хи-квадрат, Фишера-Снедекора), связи между ними. Особенности их графиков. Обратные функции.

Числовые характеристики случайных величин – математическое ожидание, дисперсия и среднеквадратичное отклонение, асимметрия, эксцесс

(русскоязычная и англоязычная терминология). Свойства математического ожидания и дисперсии. Асимметрия и эксцесс. Квантиль, медиана.

Понятие о предельных теоремах теории вероятностей.

2. Основы математической статистики (в основном в виде обзора изученных ранее разделов).

а). Понятия генеральной совокупности и выборки. Способы записи данных эксперимента (табличные, интервальные, графические и др.).

б). Выборочные оценки, понятия о смещенности и состоятельности. Оценки для математического ожидания, для дисперсии (смещенная и несмещенная). Оценки точечные и интервальные. Доверительный интервал.

в). Гистограмма, методы ее построения. Методика практического использования.

г). Псевдослучайные числа и векторы (с заданным распределением координат).

б). Понятие статистической гипотезы и ее проверки. Понятие статистики. Доверительная вероятность, методика ее задания. Гипотезы о нормальности, равномерности распределения.

3. Совместное изучение нескольких случайных величин.

а). Постановка задач о нескольких случайных величинах. Случайные векторы. Ковариация, коэффициент корреляции, их свойства. Оценки. Ранговая корреляция.

б). Проверка гипотезы о равенстве нулю коэффициента корреляции.

в). Линейная регрессия. Метод наименьших квадратов. Уравнения регрессии. Исследование коэффициента регрессии.

4. Нелинейная регрессия - сведение ее к линейной, виды графиков некоторых нелинейных функций. Полиномиальная регрессия, линейное разложение по заданной системе функций.

Множественная регрессия - линейная, мультипликативная.

5. Дополнительные исследования случайных величин и векторов.

а). Проверка на нормальность. Критерии согласия (хи-квадрат и др.).

б). Отсеивание выбросов (методы двух и трех сигм и более эффективные методы).

в). Исследование остатков регрессии.

6. Программа Excel и его применение в математической статистике.

а). Структура данных, преобразование данных.

б). Основные распределения. Команды НОРМРАСП, НОРМСТОБР, БИНОМРАСП, ПУАССОН, ХИ2РАСП, ХИ2ОБР и др.) Мастер диаграмм и его использование для построения гистограммы.

в). Вычисление основных оценок. Команды СРЗНАЧ, КВАДРОТКЛ, ДИСП, ДИСПР, СТАНДОТКЛ, СРОТКЛ.

г). Корреляция. Простая регрессия. Команды КОРРЕЛ, НАКЛОН, ОТРЕЗОК, КВАДРОТКЛ.

д). Кратная регрессия, оператор ЛИНЕЙН. Пакет «Анализ данных» (в новых версиях Excel) и его использование.

7. Математический пакет MathCad и его использования для решения задач математической статистики

а). Система Help. Resource Center. Quicksheets.

б). Ввод и редактирование величин и формул.

в). Математические величины, вычисления. Точность вычислений, ее регулирование. Палитры и их использование.

г). Матрицы, их задание и преобразования. Матричные операции.

д). Импорт и экспорт данных.

е). Простейшие распределения, вычисления с ними. Операторы **pbinom, pnorm, pt, punif** и обратные к ним. Оператор **rnd**. Случайные данные, операторы **rbinom, rnorm, rt, runif, rnd**. Функция построения гистограммы **hist** (для заданной системы интервалов).

ж). Числовые характеристики случайных величин. Операторы **mean, var, stdev**.

з). Линейная регрессия. Операторы **regress (X,Y,k), slope(X,Y), intercept(X,Y), stderr(X,Y)**. Исследование коэффициента корреляции. Простейшие нелинейные регрессии.

и). Полиномиальная регрессия. Операторы **regress, pwrfit**.

к). Кратная регрессия, функция **regress(MXY,VZ,k)**.

л). Функции **linfit(X,Y,f), genfit**.

8-9. Математический пакет **StatGraph** и его использования для решения задач математической статистики

а). Ввод данных и их модификация.

б). **StatAdvisor**, его использование.

в). Окно **Describe** (Графические и табличные операторы). Гистограмма. Гипотезы о равномерности и нормальности распределения **(Hypothesis Tests)**.

г). Построение гистограмм в окне **Plot** (раздел **Exploratory Plots**, подраздел **Frequency**

Histogram).

д). Окно Relate (Графические и табличные операторы).

Simple regression. Линейная регрессия. Сравнение разных видов простых регрессий. Графики. Выделение и сохранение остатков. Исследования Unusual residuals, Influential Points. Goodness of fit (исследование остатков на нормальность).

Полиномиальная регрессия.

Кратная регрессия. Методы построение регрессии (в том числе правой клавишей мыши – Backward Selection и Forward Selection).

Выделение случайных выбросов и наиболее влиятельных точек.

ИЛИ ЖЕ

8-9. Математический пакет Stadia и его использования для решения задач математической статистики.

а). Система помощи, ее использования. Выбор статистического метода для исследования.

б). Ввод данных и их модификация. Экспорт и импорт данных. Окна «Файл», «Преобр.»

в). Графическое изображение данных и результатов.

г). Окно «Статист.» («Статистические методы»). Разделы Гистограмма, Корреляция, корреляция / независимость, Регрессионный анализ (простая и множественная корреляция).

Сравнение коэффициента корреляции Пирсона (раздел «Корреляция») с оценками ранговой корреляции Спирмена и Кендалла (из раздела «Корреляция/независимость»).

Сравнение линий регрессии для двух экспериментальных зависимостей – в разделе «Регрессионный анализ» окна

«Статист.» подраздел «Сравнение 2-х регрессий».

10. Обзор универсальных и специальных статистико-математических пакетов

Обзор специальных задач математической статистики.

- дисперсионный анализ, планирование эксперимента, анализ временных рядов, проверка статистических гипотез и др.

Сравнительный анализ универсальных пакетов (StatGraph, SPSS, Statistica, S-Plus), а также некоторых отечественных специализированных пакетов.

ЛАБОРАТОРНЫЕ РАБОТЫ (по 2 часа каждая)

1. Работа в Windows. Работа с данными в Excel. Простейшие статистические вычисления в Excel. Корреляция.

2. Работа с данными в MathCad. Статистические вычисления в MathCad.

3. Корреляция и регрессия в MathCad.

4. Обзор пакета StatGraph (или Stadia). Работа с данными.

5. Числовые характеристики случайных величин и векторов. Корреляционный анализ (простой и кратный).

6. Исследование результатов статистических вычислений.

КУРСОВАЯ РАБОТА

Наличие курсовой работы при изучении рассматриваемого нами здесь раздела математики представляется совершенно необходимым, так как только самостоятельная работа над индивидуальным заданием позволяет студенты освоить практику работы со статистическими пакетами, а преподавателю затем эффективно проконтролировать степень усвоения материала.

ЗАДАНИЕ

Сформулировать (желательно, самостоятельно) конкретные прикладные задачи как задачи на применение методов математической статистики (поиск и анализ связей между двумя и несколькими величинами), выполнить расчеты и проинтерпретировать полученные результаты.

1. Постановка задачи – исходные данные, цель исследования. Используемые методы решения задачи.

1. Полиномиальная регрессия (степеней 1,2,3), объем выборки не менее 10.

   а). С использованием Excel и MathCad.

   б). С использованием Stadia (Statgraphics, SPSS, Statistica).

   в). Варианты простой регрессии (пакеты Stadia, StatGraphics). Подбор подходящей нелинейной зависимости.

2. Кратная линейная или мультипликативная регрессия (объем выборки не менее 16) с использованием Stadia (Statgraphics, SPSS, Statistica).

3. Анализ и интерпретация результатов.

При реализации этой программы предполагается, что ранее студенты прослушали стандартный курс теории вероятностей и математической статистики. Поэтому здесь на первой лекции просто дается обзор той части этого курса, которая касается основных понятий теории вероятностей. Математическая статистика в таких стандартных курсах рассказывается менее подробно, поэтому Лекции 2 и 3 посвящены изложению некоторых необходимых для дальнейшего статистических тпонятий. Далее в соответствии со специализацией студентов предполагается рассказать о некоторых более специальных разделах математической статистики. В выбранном нами варианте программы (на основе реальных лекций, проводимых для студентов кафедры МСС) здесь рассматриваются вопросы

теории корреляции и регрессии и их практических применениях. Более подробно о некоторых деталях реализации указанной выше программы (в том числе и не отраженных в учебной литературе) будет рассказано во второй части данного пособия. Там же будут подробнее рассмотрены и вопросы выполнения курсовой работы.

ЛИТЕРАТУРА

1. Дюк В.А. Обработка данных на ПК в примерах. СПб, 1997.

2. Кулаичев А.П. Методы и средства анализа данных в среде Windows. Stadia 6.0. Москва, 1998 и др. годы издания.

3. Очков В.Ф. MathCad для студентов и инженеров (разные издания).

4. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. М., 1997.