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

Ч А С Т Ь **II**

МОСКВА **2004** год

В части I данного пособия были указаны исходные соображения, использованные при разработке программы преподавания курса "Универсальные и специализированные математические пакетов программ для решения задач математической статистики" для студентов кафедры МСС. Там же была приведена в развернутой форме сама эта программ.

Рассмотрим теперь подробнее те разделы программы, которые не входят в стандартный курс теории вероятностей и математической статистики. Речь идет о более подробном рассмотрении теории корреляции и регрессии, а также использование компьютерных программ разного уровня сложности для решения некоторых задач математической статистики.

ЛЕКЦИИ

Лекция 6. Программа EXCEL и его применение в математической статистике

а). Структура данных, преобразование данных.

Ячейки, диапазон ячеек.

Ячейки нумеруются одновременным заданием имени столбца и номером элемента в этом столбце. Например, A3 – это третья сверху ячейка, находящаяся в первом столбце.

Нужные данные для обработки задаются указанием их верхней правой и нижней левой границ в массиве всех данных. Например, оператор (A1:C8) выделит первые восемь элементов в столбцах A,B,C.

Формулы, способы их задания.

Для задания формул, по которых будут проводиться вычисления с элементами массивов, имеется специальная строка (в верхней половине окна, она начинается сразу после символа f_x). В EXCEL существуют различные группы формул, для наших целей наиболее важны такие:

МАТЕМАТИЧЕСКИЕ,

МАТ. и ТРИГОНОМЕТРИЯ,
ССЫЛКИ И МАССИВЫ,
а также
СТАТИСТИЧЕСКИЕ,
о которых подробнее будет сказано ниже.

Результат вычислений функции помещается в заранее выбранную ячейку.

Статистические формулы.

Список реализованных в EXCEL статистических команд можно получить, нажав значок f_x и выбрав там пункт СТАТИСТИЧЕСКИЕ ФУНКЦИИ. Выделив нужную статистическую функцию, можно получить по ней справку (для этого – имеется предложение в нижней левой части появившегося окна - СПРАВКА ПО ЭТОЙ ФУНКЦИИ).

б). Основные распределения.

Команды НОРМРАСП и НОРМОБР вычисляют значения для нормального распределения и для обратной ему функции. Например, команда НОРМОБР(0.12,2,0.5) дает значение, соответствующее значению вероятности 0.12 для нормального закона распределения со средним значением (математическим ожиданием), равным 2 и среднеквадратичным отклонением, равным 0.5.

Некоторые другие распределения – это ПУАССОН, СТЬЮРАСП, БИНОМРАСП.

в). Вычисление оценок. Перечислим некоторые нужные нам основные команды

СРЗНАЧ – вычисление среднего арифметического заданного массива данных (строки, столбца, матрицы и др.). Дает оценку для математического ожидания.

ДИСП (несмещенная оценка дисперсии, ее допустимо применять только для больших наборов данных – не менее 30, в противном случае нужно применять команду ДИСПР).

Пример: ДИСП(A1:A30) – вычисление несмещенной оценки дисперсии для 30 чисел из столбца A.

СТАНДОТКЛ, СРОТКЛ (несмещенная оценка) – это фактически корни квадратные √ДИСП и √ДИСП из соответствующих оценок дисперсии.

г). Построение гистограммы.

Гистограмма дает возможность приближенно определить вид графика плотности распределения изучаемой случайной величины. Например, если гистограмма напоминает собой гауссову кривую (хоть и состоит из горизонтальных отрезков), то можно предполагать, что изучаемая случайная величина распределена нормально (для более точного исследования на нормальность имеются специальные методы).

Для построения гистограммы можно использовать специальную кнопку в основном меню: «Мастер диаграмм» (гистограмма тут рассматривается как одна из многих диаграмм в EXCEL). Шаг за шагом, отвечая на задаваемые «Мастером диаграмм» вопросы, получим гистограмму.

д). Корреляция. Простая регрессия. Основные команды:

Команда ПИРСОН – вычисляет коэффициент корреляции (здесь он фигурирует как коэффициент Пирсона). Например, команда ПИРСОН(A1:A10;B1:B10) дает коэффициент корреляции для двух столбцов A и B. КВПИРСОН - дает квадрат коэффициента корреляции. Команда КОРРЕЛ – дает тот же коэффициент корреляции (непонятно правда, зачем было одну величину вычислять под разными именами).

НАКЛОН, ОТРЕЗОК, ЛИНЕЙН – команды для построения прямой регрессии и анализа ее точности.

Для уравнения линейной регрессии y=ax+b команда НАКЛОН дает коэффициент a, ОТРЕЗОК дает коэффициент b. Формат этих команд одинаков, например, НАКЛОН(A1:A10;B1:B10) дает наклон прямой регрессии столбца A зависимых значений на столбец независимых значений B.

Команда ЛИНЕЙН позволяет не только находить линейную регрессию, но и вычислять различные дополнительные параметры для ее анализа, а также проводить и кратную регрессию.

е). Кратная регрессия – производится с помощью команды ЛИНЕЙН. Например, ЛИНЕЙН(Y,X) решает задачу для массивов Y - зависимых данных (столбец) и X - независимых данных (один или несколько столбцов), выделяемых стандартным для EXCEL методом. Дополнительно можно задавать некоторые специальные опции. Мы не будем здесь на этом останавливаться, так как намного более эффективные исследование можно производить с помощью специализированных программ – математических (MathCad) и статистических (Stadia, StatGraph и др.), о чем будет рассказано ниже.

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

а). Система Help. Resource Center. Некоторые Quicksheets (Data Analysis, Statistics).

б). Ввод и редактирование величин и формул.

Выделение элементов формул, редактирование (замена, стирание). Быстрый ввод некоторых символов (корня, оператора присваивания, числа пи и др.).
Таблицы данных, их создание и редактирование.

в). Математические величины, вычисления.

Точность вычислений, ее регулирование (через команду FORMAT).

Палитры (Toolbars), их виды и использование.

г). Матрицы, их задание (через Insert и через палитру) и преобразования (слияние, выделение, редактирование). Матричные операции (сложение, умножение, обращение матриц, определитель – через палитру MATRIX).

д). Импорт и экспорт данных (из файлов разной структуры).

Импорт из файлов разной структуры (текстовых разного вида, EXCEL и др.). Форматы файлов данных, их

учет и модификация производятся через INSERT (Input Table, File Reading).

е). Простейшие распределения, вычисления с ними. Операторы pbinom, pnorm, pt, punif и обратные к ним операторы, оператор rnd.

Случайные данные, операторы rbinom, rnorm, rt, runif, rnd.

Биномиальное распределение pbinom(k,n,p)
dbinom(k,n,p)

Команда pbinom дает значение функции распределения биномиального закона, для которого параметр p - вероятность однократного наступления события A, n – число испытаний, k – число наступлений события A.

Команда dbinom – дает вероятность значения k для биномиального распределения.

Равномерное распределение punif(x,a,b)
dunif(x,a,b)

Тут команда punif дает значение в точке x функцию распределения для равномерно распределенной случайной величины с параметрами a,b.

Нормальное распределение pnorm(x,a,sigma)
dnorm(x,a,sigma)

и обратное к нему qnorm(p,a,sigma)

Тут (и ниже – для распределения Стьюдента) команда, имя которой начинается буквой p, дает значение функции распределения (при заданных значениях параметров распределения – для нормального закона это a и sigma). Начинающаяся с буквы d команда дает плотность распределения в точке x, а если команда начинается буквой q – то она дает функцию, обратную функции распределения.

Распределение Стьюдента pt(x,d)
dt(x,d)

и обратное к нему qt(x,d),

где d – число степеней свободы (этот параметр вычисляется, исходя из объема выборки, и показывает, сколько независимых параметров остается после учета специфики решаемой задачи).

Функции со сходными приведенным выше названиями, но начинающиеся с буквы r, дают случайные числа с соответствующими распределениями. Например, функция rnorm(m,a,sigma) выдает m чисел, распределенных по нормальному закону с параметрами a, sigma. Для получения одного случайного числа, равномерно распределенного в отрезке [0,x], используется и специальная функция rnd(x).

ж). Числовые характеристики случайных величин.

Операторы mean (оценка математического ожидания), var, stdev (средне-квадратичное отклонение – смещенная оценка).

Например, если данные оформлены в виде массива X, то формат выполнения этих команд следующий:

```
mean(X)=
stdev(X)=
```

з). Линейная регрессия.

Для оценки коэффициента корреляции используется функция corr(X,Y). Для проверки значимости отличия коэффициента корреляции от нуля в MathCad не предусмотрено специального оператора, поэтому эта задача решается с помощью использования других функций программы. Делается это так.

Пусть известна оценка r для коэффициента корреляции, а общее числа пар данных равно n. Тогда вычисляется следующая величина:

$$T = r\ \mathrm{sqrt}((n-2)/(1-r^2)).$$

Полученное значение сравнивается с критическим, которое находится по заданному n и выбранному уровню вероятности p с помощью функции qt(p,n) (обратная функция для распределения Стьюдента).

Если оказывается, что это критическое значение больше вычисленного выше числа T, то гипотезу о равенстве нулю коэффициента корреляции отвергать нет основания. Это означает, что оснований для подбора линейной зависимости между исследуемыми переменными нет. Возможно, стоит поискать зависимости другого рода, преобразовав для этого исследуемые переменные.

Такого рода рассмотрения в MathCad достаточно трудоемки, однако более совершенные, специализированные статистические программы позволяют решать такого рода задачи без труда (как это описано ниже для программы StatGraph).

Если же критическое значение больше вычисленного T, то на заданном уровне вероятности гипотезу о равенстве нулю коэффициента корреляции следует отвергнуть (рискуя совершить ошибку с вероятностью 1-p). В этом случае имеет переходить к следующему этапы исследования – построению уравнения линейной регрессии.

Операторы slope, intercept,

Формат этих команд (аналогичных командам НАКЛОН и ОТРЕЗОК в EXCEL) одинаков. Например, slope(X,Y) дает коэффициент a в уравнении линейной регрессии y=ax+b.

Для оценки погрешности линейного уравнения регрессии применяется команда stderr, вычисляющая сумму квадратов отклонений заданных точек от лежащих на прямой линии корреляции (с теми же аргументами). Формат команды stderr(X,Y)= .

Примеры организации вычислений (на примере вычисления постоянной Хаббла из астрофизики).

```
a:= slope(Y,X)
b:= intercept(Y,X)
a= 2.823
b= -1.053
stderr(Y,X)= 18.325
```

Полезно провести сравнение двух уравнений регрессии. Для этого на одном поле строятся две прямые - два графика – линейные зависимости Y от X и X от Y. Это две различные

функции отличаются наклоном их графиков (чем больше коэффициент корреляции отличается от нуля, тем меньше различаются между собой две эти прямые).

и). Полиномиальная регрессия.

Операторы regress(X,Y,n) пользуются для построения полиномиальной регрессии степени n между переменными X Y. Несмотря на то, что программа позволяет находить решения для практически произвольных значений степени n (лишь бы число данных в массивах X,Y было бы достаточно велико), на самом деле использование регрессий большой степени n не представляется эффективным. На практике имеет смысл ограничиться значениями степени n регрессии не более чем 3-4.

Функция regress выдает ответ в виде столбца, в котором в первых строках расположены вспомогательные коэффициенты, а вот искомые коэффициенты регрессионного полинома расположены в последних строках (последним тут идет коэффициент при старшей степени многочлена).

к). Кратная регрессия (оператор regress). Методика выбора размерности задачи. Частные регрессии.

л). Функции linfit, genfit.

Функция linfit строит уравнение регрессии в виде линейной комбинации нескольких заданных функций. При таком подходе линейная регрессия исходит из функций 1,x, а полиномиальная степени n – из функций $1,x,\ldots,x^{n-1}$. Формат команды linfit(X,Y,f), где X,Y – массивы данных, а f – столбец, составленный из заданных функций. Например, тригонометрическая регрессия сможет исходить из столбца, образованного функциями 1, sin(x), cos(x).

Функция genfit допускает еще большую свободу при выборе вида уравнения регрессии. Здесь можно искать уравнение регрессии в виде произвольной функции, зависящей от нескольких параметров. Подбор (методом наименьших квадратов) этих параметров и осуществляется программой. Фактически здесь производится решение некоторой системы нелинейных уравнений. Поэтому, как и при решении произвольных систем нелинейных уравнений, тут

необходимо предварительно задать некоторое начальное приближения для разыскиваемых параметров.

Выбор этих начальных значений может оказать на результат вычисления определяющее влияние. Дело в том, что гарантированного нахождения решения системы нелинейных уравнений не может обеспечить никакая программа.

Формат команды **genfit(X,Y,V,F)**. Здесь **X,Y** массивы исходных данных, **F** – заданные в виде столбца исходная зависимость и частные производные по всем входящим в нее параметрам (так как символическое вычисление производных этой процедурой не предусмотрено), а **V** – массив начальных значений для параметров.

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

а). Ввод данных и их модификация. Форматы файлов данных.

При вводе данных нужно обратить внимание на все вопросы, которые задаются программой в процессе импорта данных, так как при игнорирование особенностей системы ввода часть данных может быть потеряна или же введена неправильно. Например, при вводе данных из текстового файла знаком десятичного разделения в числах по умолчанию является запятая. Поэтому использование разделителя-точки в исходном файле (как это теперь иногда принято) в качестве десятичного разделителя приведет к тому, что вводимые данные не будут рассматриваться как числовые (и программа не сможет производить с ними статистические операции).

в). Окно **Describe** (Графические и табличные операторы).

Numeric data -> One-Variable Analysis. Тут вводятся данные для исследования.

Далее можно выбрать через две специальные кнопки различные табличные и графические опции. Среди них – **Summary Statistics,** где вычисляются основные

статистические характеристики одномерного массива (оценки для математического ожидания, дисперсии), а StatAdvisor высказывает предварительные суждения о нормальности рассматриваемого распределения. Графические опции позволяют по-разному изобразить исследуемые данные (особенно полезна гистограмма – она задается здесь как Frequency Plot). Среди табличных опций выделим проверку гипотез о среднем значении (равенство и неравенство нулю для заданного уровня вероятности).

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

Имеется множество графических форматов для вывода данных. Поначалу можно ограничиться простейшим из них – гистограммой, так как она изучается в курсе теории вероятностей.

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

(i). Простая (simple).
(ii). Полиномиальная регрессия (polynomial regression). По умолчанию обычно строится квадратичная регрессия (хотя можно вычислять и регрессию n-ой степени (см. однако замечания выше о выборе степени регрессии).
(iii). Кратная регрессия (multiple regression).

Лекция 10. Обзор универсальных и специальных статистико-математических пакетов

Обзор специальных задач математической статистики.

Сравнительный анализ универсальных пакетов (StatGraph, SPSS, Statistica, S-Plus), а также некоторых отечественных специализированных пакетов (Мезозавр, Эвриста и др.). Рекомендуется использовать литературу, перечисленную ниже, а также данные сравнительного анализа, которые можно найти в сети Интернет (например http://www.cemi.rssi.ru/rus/publicat/e-pubs/)

ЛАБОРАТОРНЫЕ РАБОТЫ

Переходим к некоторым методическим рекомендациям по проведению лабораторных работ. Вначале – основные темы лабораторных работ.

Лабораторная работа 1.

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

Лабораторная работа 2.

Работа с данными в MathCad. Статистические вычисления в MathCad.

Лабораторная работа 3.

Корреляция и регрессия в MathCad.

Лабораторная работа 4.

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

Лабораторная работа 5.

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

Лабораторная работа 6.

Исследование результатов статистических вычислений.

ВАРИАНТЫ УСЛОВИЙ ЗАДАЧ КУРСОВОЙ РАБОТЫ И МЕТОДИКА ВЫПОЛНЕНИЯ ЗАДАНИЯ.

Предполагается, что конкретные статистические данные для выполнения курсовых работ студенты находят самостоятельно – в известной им технической литературе, а также проводя самостоятельный сбор данных, в том числе и не обязательно технического характера. Ниже приводятся

примеры тем курсовых работ именно такого «вольного» характера (по материалам реальных курсовых работ, выполненных студентами МАТИ).

1. В отделении ГАИ провели исследования, сравнивая число водителей в нетрезвом состоянии, задержанных за сутки и число дорожно-транспортных происшествий за те же сутки. Нужно выяснить, если связь между этими двумя величинами, а если она имеется – то найти ее. Нужно также сформулировать результаты исследования в терминах, понятных работникам ГАИ.

   Число водителей в нетрезвом состоянии, задержанных за сутки:

   4, 8, 7, 9, 13, 2, 3, 12, 8, 4, 7, 9, 6, 4, 14, 12, 8, 4, 11, 9, 3, 5, 16, 10,14.

   Количество дорожно-транспортных происшествий за сутки:

   10, 19, 15, 21, 28, 10, 15, 25, 13, 18, 19, 23, 17, 12, 21, 19, 20, 9, 28, 18, 11, 9, 25, 26, 33.

2. Работником метрополитена могут заинтересовать данные, полученные одной из студенток. Она в течение 20 дней, исключая воскресенья, подсчитывала число людей, выходивших из одной станции метро с 7 час. 45 мин. до 7 час. 55 мин. и число входивших людей в то же время на ту же станцию. Нужно выяснить, нет ли связи между этими двумя величинами. Также нужно сформулировать ответ в Форме, доступной для работников метрополитена.

ДАТА	ЧИСЛО ВХОДЯЩИХ	ЧИСЛО ВЫХОДЯЩИХ
3 (Сб.)	72	110
5 (Пн.)	431	280
6 (Вт.)	450	307
7 (Ср.)	442	324
8 (Чт.)	457	295
9 (Пт.)	433	309
10 (Сб.)	51	155
12 (Пн.)	390	317

13 (Вт.)	430	277
14 (Ср.)	394	299
15 (Чт.)	401	269
16 (Пт.)	450	285
17 (Сб.)	63	101
19 (Пн.)	443	311
20 (Вт.)	448	287
21 (Ср.)	439	303
22 (Чт.)	433	325
23 (Пт.)	457	291
24 (Сб.)	190	147
26 (Пт.)	455	274

Интересно отметить, что это – реальные данные (студентка действительно стояла и считала пассажиров метро!).

Оказалось, что коэффициент корреляции тут весьма велик (0.83). Поэтому эти данные действительно могут пригодиться работникам метрополитена.

3. Экологическое исследование – о связи между содержанием азота в реке Дон и среднегодовым уловом рыбы.

Среднегодовое содержание азота в р. Дон (тыс. тонн):

20.0, 19.2, 16.44, 16.44, 14.0, 18.0, 16.44, 18.05, 18.16, 15.0, 16.80, 14.0, 15.09, 14.75, 15.0, 16.44, 16.0, 15.77, 14.73, 14.27.

Среднегодовой улов рыбы в р. Дон (тыс. тонн):

39.26, 34.22, 32.30, 35.80, 37.23, 39.0, 35.8, 36.42, 37.23, 37.45, 35.58, 32.80, 37.30, 38.58, 39.26, 38.55, 35.58, 35.58, 35.55, 33.22, 33.22

4. Анализ работы ЖЭКа.

В течение месяца записывалось число поступивших в ЖЭК жалоб и число рассмотренных (удовлетворенных). Выяснить наличие закономерностей между числом

поступивших жалоб и числом рассмотренных. Результат можно использовать для совершенствования работы ЖЭКа, а также для оценки ее работы.

Число жалоб (по числам с 1-го по 30-е)

5,9,3,14,21,6,4,32,6,4,32,11,19,19,36,25,20,16,9,10, 7,4,15,33,37,26,25,40,16,13,8,31,28.

Число рассмотренных жалоб

2,4,0,6,10,1,0,17,7,3,7,13,6,3,11,9,4,3,0,0,9,24,17, 12,10,13,2,6,16,11.

5. Анализ работы офтальмологической клиники по заключению контрактов на операции.

   В течение месяца записывалось число больных, обратившихся в офтальмологическую клинику и число заключенных контрактов на обслуживание (операцию). Нужно выяснить, имеется здесь связь между этими величинами. Это можно использовать для совершенствования работы клиники и для ее развития.

6. В одном учебном институте были собраны данные о средней месячной зарплате в зависимости от стажа. Отсчет велся от 1 сентября (поэтому много целых чисел). Нужно выяснить, есть ли закономерность между этими величинами. Это может оказаться полезным для работы отдела кадров.

Стаж работы:

1.0, 6.5, 9.2, 4.5, 6.0, 2.5, 2.7, 16.0, 13.2, 14.0, 11.0, 12.0, 10.5, 1.0, 9.0, 5.0, 6.0, 10.2, 5.0, 5.4, 7.5, 8.0, 8.5

Месячная зарплата (средняя):

150, 162, 195, 164, 170, 152, 162, 218, 204, 210, 200, 196, 188, 155, 187, 182, 165, 190, 178, 175, 185, 190, 198.