ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Факультет – электрофизический
Направление – стандартизация и сертификация
Кафедра – компьютерных измерительных систем и метрологии

О.В. Стукач

Программные статистические комплексы

Лабораторный курс

2009

Подробно рассмотрена работа с универсальным пакетом «Statistica» по системному подходу к анализу данных: анализу закономерностей в данных, всестороннему и последовательному исследованию статистической информации, формированию статистических выводов. Усовершенствован метод экспертных оценок применительно к задачам управления предприятием. Материал позволяет по-новому взглянуть на методы статистического анализа процессов и использовать их как комплекс системных мероприятий по повышению качества управления.

Материал используется как учебно-методическое пособие по изучению использования статистических методов в промышленном управлении, для классификации и поиска максимально точной и прагматичной информации о структуре данных. Он адресуется всем, кому необходимо применять статистические методы в своей деятельности, руководителям и менеджерам предприятий, преподавателям и студентам.

Лабораторная работа № 1

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ В ПАКЕТЕ STATISTICA

1. ЦЕЛЬ РАБОТЫ

Целью работы является освоение пакета, знакомство с вероятностным калькулятором и решение с помощью него вероятностных задач. В работе исследуется графический анализ данных, проводится визуализация значений заданных переменных с использованием статистических графиков.

2. ЛАБОРАТОРНОЕ ЗАДАНИЕ

Задание 1. *Визуализация значений заданных переменных с использованием статистических графиков.*

Прочитать файл с исходными данными VARIANT0.STA. Построить гистограммы (команда Graphs/ Histograms) для двух переменных на одном графике в зависимости от номера вашего варианта N: VarN, VarN+1. Использовать опцию Multiple (несколько графиков на одной сетке) во вкладке Quick.

Для тех же переменных построить столбчатую диаграмму (Graphs/ 2D Graphs/ Bar Columns Plots).

Для переменной VarN построить круговую диаграмму (Pie Chart - Counts). Обратить внимание, как

строится график Pie Chart при изменении переменной Categories.

Для переменных VarN, VarN+1, VarN+2 построить 3D график (Graphs/ 3D XYZ Graphs/ Surface Plots). Обратить внимание, что трёхмерный график можно разворачивать на любой угол в подменю «свойства графика / все свойства». Настроить графики, подписав переменные и оси. Поместить графики в отчёт.

Задание 2. *Решить с помощью вероятностного калькулятора следующие задачи.*

Варианты 1,3,5,7... (нечётные): Известно, что в некоторой стране рост мужчин приближенно имеет нормальное распределение со средним 176 см и стандартным отклонением 7,63 см. Какова вероятность того, что рост случайно встреченного вами мужчины будет не менее 186 см?

Варианты 2,4,6... (чётные): Вы попали в страну, где рост мужчин приближенно имеет нормальное распределение со средним 173 см и стандартным отклонением 8,65 см. Какова вероятность того, что рост случайно встреченного мужчины будет не менее 195 см?

Задание 3. Исследование средних величин.

Конвертировать файл с данными в систему Statistica в соответствии с таблицей:

Номер варианта	Имя файла с данными	Используемые переменные
1...5	mmvb.txt	1 и 2 столбцы (ежедневное изменение индекса

		Московской межбанковской валютной биржи)
6...10	mmvb.txt	1 и 3 столбцы (ежедневное изменение стоимости чистых активов ПИФ ММВБ)
11...15	deposit.txt	1 и 2 столбцы (ежедневное изменение стоимости пая ПИФ депозитный)
16...20	deposit.txt	1 и 3 столбцы (ежедневное изменение стоимости чистых активов пая ПИФ депозитный)

Построить гистограмму для выбранной 2-й или 3-й переменной. Сравнить построение гистограммы для разных значений интервалов группировки, которые можно изменить в окне построения гистограммы. В отчёте привести две гистограммы – с правильным и неправильным значениями количества интервалов.

Задание 4. Составить новую таблицу, разделив выбранную переменную по месяцам. При этом каждая переменная новой таблицы будет соответствовать одному месяцу. С помощью модуля *Statistics/ Basic Statistics/*

Tables исследовать изменение среднего арифметического значения переменной и медианы. Усреднения проводить каждый месяц. Построить графики "ящик с усами" для всех средних значений и медиан (кнопка *Box & Whisker plot for all variables* окна *Descriptive statistics*). На графике соединить средние значения прямыми линиями. Сделать выводы.

Лабораторная работа № 2

ПРОСТЫЕ ИНСТРУМЕНТЫ ОЦЕНКИ КАЧЕСТВА В ПРОГРАММЕ STATISTICA

1. ЦЕЛЬ РАБОТЫ

Целью работы является изучение простых методов оценки качества с помощью программы STATISTICA: анализ Парето, диаграмма причин и результатов (Исикавы), стратификация.

2. ПРОГРАММА РАБОТЫ

Задание 1. Построить диаграмму причин и результатов с помощью модуля Statistics / Industrial Statistics & Six Sigma/ Process Analysis/ Cause–Effect [Ishikawa, Fishbone] diagrams. Студенты с чётными номерами вариантов строят диаграмму для процесса варки супа. Студенты с нечётными номерами вариантов строят диаграмму для процесса выращивания морковки.

Указание. Необходимо обратить внимание, как строится диаграмма. Пользователь может самостоятельно выбирать переменные, которые располагаются вверху и внизу центральной линии диаграммы. Кроме того, обратить внимание, в каком порядке располагаются факторы на диаграмме.

Факторы выше второго порядка добавляются на диаграмму с помощью панели рисования.

Задание 2. С помощью модуля Statistics/ Industrial Statistics & Six Sigma/ Quality Control Charts/ Pareto chart analysis построить диаграмму Парето. Причины и их число выбрать самостоятельно, можно любые, не относящиеся к какому-либо процессу. Пример составления таблицы с исходными данными для построения диаграммы Парето имеется в файле Primer.sta.

Построить диаграммы Парето:

- только для переменной 1 «причина», без учета значимости причины для общего вклада в качество анализируемого процесса. Для этого на вкладке Quick выбрать Codes (requires tabulation of data by codes) – установлен по умолчанию, а в окне выбора переменных указать одну переменную с перечнем причин;
- для переменных 1 «причина» и 2 «число», с учётом значимости причины для общего вклада в качество анализируемого процесса. Для этого на вкладке Quick выбрать Codes and counts (one variable with defect type, one variable with counts).

В чем разница между этими картами Парето?

Задание 3. С целью выяснения причин брака составлен контрольный листок в предположении, что причинами могут быть рабочий, станок или смена. Определить, кто виноват, если это возможно.

Распределение дефектов по рабочим, станкам и сменам.

Рабочий	Станок	1 смена	2 смена	3 смена	Число дефектов на станках	Сумма дефектов рабочего
А	А 1			●	1	24
	А 2	● ●	●		3	
	А 3	●	● ● ● ● ● ● ●	● ● ● ● ● ● ● ● ● ● ●	20	
Б	Б 1	● ●	● ● ● ● ● ● ●	● ● ● ● ● ●	15	45
	Б 2	●	●	● ● ●	5	
	Б 3	● ●	● ● ● ● ● ● ●	● ● ● ● ● ● ● ● ● ● ● ● ●	25	
В	В 1	●	●	● ●	4	52
	В 2	● ●	● ●	● ● ● ●	8	
	В 3	● ● ●	● ● ● ● ● ● ● ● ● ● ● ● ● ● ●	● ● ● ● ● ● ● ● ● ● ● ● ● ● ●	40	

Лабораторная работа № 3

КАРТЫ КОНТРОЛЯ КАЧЕСТВА В ПРОГРАММЕ STATISTICA

1. ЦЕЛЬ РАБОТЫ

Целью работы является изучение особенностей построения карт контроля качества в системе Statistica.

2. ПРОГРАММА РАБОТЫ

Задание 1. Прочитать файл с данными Variant2.sta. С помощью модуля Statistics / Industrial Statistics & Six Sigma/ Quality Control Charts/ Individuals & moving range построить контрольную X-R карту индивидуальных значений для одной переменной своего варианта (VarN, Time) и провести анализ качества производственного процесса. Проверить машинное построение карты расчетом средней линии.

Задание 2. С помощью модуля Statistics / Industrial Statistics & Six Sigma/ Quality Control Charts/ X-bar & R Chart for variables построить контрольную X-R карту процесса для одной переменной своего варианта (VarN) и временных интервалов (Time, Day) и провести анализ качества производственного процесса. Сравнить построение карты с результатами предыдущего задания. Чем отличается построение карты индивидуальных значений от карты процесса?

Задание 3. Построить контрольную C-карту индивидуальных значений для одной переменной своего варианта (VarN) и провести анализ качества процесса.

Лабораторная работа № 4

МОДЕЛИРОВАНИЕ РАСПРЕДЕЛЕНИЙ СЛУЧАЙНЫХ ВЕЛИЧИН В ПАКЕТЕ STATISTICA

1. ЦЕЛЬ РАБОТЫ

Целью работы является исследование центральной предельной теоремы и изменения плотности нормального распределения при изменении среднего значения и среднеквадратического отклонения.

2. ЛАБОРАТОРНОЕ ЗАДАНИЕ

Задание 1. Теорема Хинчина.

Теорема Хинчина утверждает, что среднее арифметическое

$$\overline{X} = \left| \frac{X_1 + X_2 + ... + X_n}{n} \right|$$

независимых случайных величин X_j, j=1, 2,..., n, имеющих одно и тоже распределение и конечное математическое ожидание m, сходится по вероятности при $n \to \infty$ к m. Таким образом, при заданном ε и достаточно большом n событие

$$\left| \frac{X_1 + X_2 + ... + X_n}{n} - m \right| < \varepsilon$$

можно считать практически достоверным.

Постепенно увеличивая n, с помощью пакета Statistica показать, выполняется теорема или нет для заданного закона распределения случайных чисел $X \in [0; 1]$. Закон распределения случайных чисел выбрать из таблицы по номеру своего варианта (табл. 1):

Таблица 1

Номер варианта	Распределение	Функция в пакете Statistica
1...5	равномерное	Rnd

6...10	нормальное	Rndnormal
11...15	Пуассона	Poisson
16...20	равномерное	Rnd
21...25	нормальное	Rndnormal
26...30	Пуассона	Poisson

Указание. Равномерно распределённые случайные числа генерируются программно. Для заполнения переменной случайными числами из интервала [0;1] необходимо щелкнуть правой кнопкой мыши на имени переменной и выбрать команды Fill/Standardize Block / Fill Random Values. Среднее значение вычисляется тем же способом командой Statistics of Block Data / Block Columns / Means.

Заполнить переменную вычисленными значениями можно, если дважды щёлкнуть левой кнопкой мыши по имени переменной и в появившемся окне «Long name (label or formula)» записать формулу. Например, формула =rndnormal(2) позволяет получить нормально распределённые случайные числа в интервале[0;2].

Задание 2. Сформировать набор данных для последующего анализа в программе Statistica, состоящий из 1 переменной и 150 наблюдений. Переменную заполнить числами из табл. 2. При этом N-му варианту соответствуют элементы выборки, расположенные в 15–ти следующих строчках таблицы, начиная с N–й (объем выборки при этом n=150).

Таблица 2 – варианты задания и переменные

N										
1	48	30	43	44	30	34	32	43	40	46
2	25	21	34	49	39	37	45	49	31	49
3	43	46	34	35	42	30	41	34	42	22
4	38	40	26	47	34	42	38	20	38	36

5	30	13	41	40	40	15	35	11	38	45
6	37	12	38	36	14	39	32	54	43	39
7	23	30	32	36	32	34	49	18	49	50
8	37	20	44	28	44	35	45	34	33	41
9	43	45	50	14	33	39	41	39	46	31
10	40	52	44	39	35	54	33	42	42	36
11	44	51	45	19	34	44	40	37	43	32
12	33	42	40	35	37	13	48	48	50	32
13	40	48	45	23	36	36	42	40	37	30
14	44	50	46	39	31	48	44	42	36	51
15	44	50	54	37	33	34	42	43	43	47
16	33	48	18	42	15	32	34	14	39	45
17	48	26	31	34	38	36	46	49	40	48
18	42	47	35	34	41	33	41	35	43	42
19	39	37	47	27	33	22	37	19	19	37
20	43	41	30	39	38	36	36	34	42	46
21	39	44	37	35	43	38	33	47	45	38
22	37	48	38	52	40	45	44	42	38	40
23	44	46	37	34	41	37	41	39	30	38
24	32	41	48	36	51	36	33	39	45	40
25	34	41	38	34	33	27	51	45	27	38
26	42	37	46	41	47	36	30	45	41	40
27	37	37	39	42	48	41	36	39	33	47
28	43	49	27	31	41	46	40	36	36	42
29	41	46	33	37	47	35	31	29	30	36
30	48	38	37	34	40	34	36	50	48	39
31	30	38	43	41	44	45	38	37	46	50
32	41	48	41	43	47	37	42	34	32	44
33	37	48	46	41	41	37	37	48	49	46

34	38	44	50	37	47	27	48	37	46	38
35	48	47	38	52	34	36	34	41	41	32
36	31	43	34	46	37	40	41	39	32	42
37	47	33	51	41	40	45	37	36	27	36
38	37	42	46	35	34	38	45	36	20	40
39	34	48	30	51	33	41	44	42	39	39
40	45	45	41	40	36	27	50	44	41	48
41	36	36	32	32	36	49	27	45	30	35
42	40	38	45	40	40	50	42	37	50	39
43	43	38	30	59	42	41	33	42	38	44
44	44	41	47	52	51	38	50	39	50	48
45	49	43	52	50	30	30	26	50	27	49
46	27	49	46	39	47	26	49	52	29	44
47	51	53	48	49	53	45	27	43	48	44

Примечание: жёлтым цветом выделены числа для третьего варианта.

По данной выборке объема n=150 построить статистический ряд

x_1	x_2	...	x_e
n_1	n_2	...	n_e

где $x_1 < x_2 < ... < x_e$ элементы выборки, записанные в порядке возрастания, n_i – частоты появления одинаковых значений случайной величины *y*.

Задание 3. На основе статистического ряда построить сгруппированную выборку. Для этого задается определенный отрезок [*a, в*], внутри которого расположены все элементы исследуемой выборки, число интервалов k, на которое делится этот отрезок. Находятся длины интервалов $h = \frac{b-a}{k}$, концы

интервалов $x_i = a + (i-1)h$, середины интервалов $z_i = \frac{1}{2}(x_i + x_{i+1})$ и соответствующие эмпирические частоты m_i (m_i – число элементов выборки, попавших в i – й интервал), i = 1, 2, ... k. Результаты вычислений заносятся в таблицу 3:

Таблица 3

номер интервала	границы интервала	середины интервалов	эмпирические частоты
i	x_i, x_{i+1}	z_i	m_i
1 2 . . . k			

Принять уровень значимости α = 0,05, отрезок [24,5; 54,5], число интервалов k = 10.

Задание 4. Построить график эмпирической функции распределения

$$F^*(x) = \begin{cases} 0 & x \le z_1 \\ \frac{1}{n}(m_1 + ... + m_i), & \text{при } z_i < x \le z_{i+1}, \\ 1 & x > z_k \end{cases}$$
$$(i = 1,2,...,k-1).$$

и гистограмму.

Известно, что гистограмма строится из прямоугольников с основаниями $[x_i, x_{i+1}]$ и высотами $\frac{m_i}{nh}$. Проверить, выполняются ли эти условия при построении гистограммы в программе Statistica.

Задание 5. Найти выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i z_i$, исправленную выборочную дисперсию $S^2 = \frac{1}{n-1}\sum_{i=1}^{k} m_i \left(z_i - \bar{x} \right)^2$; исправленное выборочное среднеквадратическое отклонение $S = \sqrt{S^2}$.

Проверить гипотезу о нормальном распределении случайной величины X с математическим ожиданием $a = \bar{x}$ и среднеквадратическим отклонением $\sigma = S$ с помощью критерия χ^2 Пирсона.

Для этого вычисляют теоретические частоты попадания случайной величины X в i-й интервал np_i,

где $p_i = p\{x_i \le X < x_{i+1}\} = \Phi\left(\frac{x_{i+-} - \bar{x}}{S} \right) - \Phi\left(\frac{x_i - \bar{x}}{S} \right)$.

Значения функции Лапласа $\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}} du$ находятся по таблице или с помощью вероятностного калькулятора.

Если при некотором i эмпирическая или теоретическая частота меньше 5, тогда этот интервал объединяют с соседним, при этом теоретические и эмпирические частоты суммируются. После объединения получают r интервалов ($r \le k$).

Составляется статистика χ^2 Пирсона

$\chi^2_{\text{набл.}} = \sum_{i=1}^{P} \frac{(m_i - np_i)^2}{np_i}$.

Затем по закону уровня значимости α и числу степеней свободы $\nu=r-3$ находится критическая точка $\chi^2_{\alpha,\nu}$ по таблице квантилей распределения χ^2. Если $\chi^2_{\text{набл.}} > \chi^2_{\alpha,\nu}$, то гипотеза отвергается. Если $\chi^2_{\text{набл.}} \le \chi^2_{\alpha,\nu}$, гипотеза принимается.

Задание 6. Построить график плотности вероятности $f(x)=\frac{1}{\sqrt{2\pi}S}e^{-\frac{(x-\bar{x})^2}{2S^2}}$ случайной величины *X*. Какой закон распределения *X* вы наблюдаете?

Задание 7. Проанализировать набор данных из задания 2, состоящий из 1 переменной и 150 наблюдений в программе Statistica. Целью задания является проверка гипотезы о нормальном распределении случайной величины по критерию χ^2 автоматически. Для этого выполнить следующие действия: *Statistics/ Distribution Fitting (подбор распределений)/ Continuous Distributions (непрерывные распределения)/ OK/ Normal (нормальное распределение)/ Variable/ Summary*. На экран выводится таблица для расчёта статистики критерия.

Для вычерчивания измеряемого и ожидаемого распределения нажимаем соответствующую кнопку (Plot of observed and expected distribution). Появится гистограмма, вверху которой написано рассчитанное значение χ^2 (Chi-Square test), число степеней свободы (*df*) и уровень значимости (*p*). Именно *p*-уровень представляет собой вероятность ошибки, связанной с распространением наблюдаемого результата на всю выборку.

Сравнить графически наблюдаемые (Observed Frequency) и ожидаемые частоты (Expected Frequency): записать соответствующие столбцы в отдельную таблицу и построить график

рассеяния (команды *Graphs/ Scatterplots/ Variables/ OK*). Переменные существенно различаются, если точки плохо укладываются на прямую линию. Если бы переменные были одинаковы, все наблюдения лежали бы на прямой с уравнением Var2=Var1.

Является ли исследуемая переменная нормально распределённой? Сравнить полученный результат с заданием 6 и сделать выводы.

Задание 8. Таблицу из восьми переменных (Var1...Var8) и 150 наблюдений заполнить случайными числами из интервала [0;1]. Найти Var9=Var1+Var2+...+Var8. Для этого необходимо дважды щёлкнуть левой кнопкой мыши по Var9 и в появившемся окне «Long name (label or formula)» записать формулу

=v1+v2+v3+v4+v5+v6+v7+v8

или

=sum(v1:v8).

Построить гистограммы для Var1 и Var9 отдельно. Какие распределения вы получили? Объяснить полученный результат.

Лабораторная работа № 5

РЕГРЕССИОННЫЙ АНАЛИЗ ДАННЫХ

1. ЦЕЛЬ РАБОТЫ

Целью работы является исследование статистических зависимостей между переменными с помощью нелинейного

регрессионного анализа, оценка ошибки регрессии с помощью модуля "Nonlinear Estimation" системы Statistica.

2. РЕГРЕССИОННЫЙ АНАЛИЗ

Задание 1. Прочитать файл с исходными данными Variant5.sta. Считать первый столбец независимой переменной (Argument), а остальные (Var-N) – зависимой переменной. Номер зависимой переменной соответствует вашему варианту по списку группы. Построить график зависимости второй переменной Var-N от первой: Graphs / 2D Scatterplot; во вкладке Advanced выбирайте Off, в опции «свойства графика» соединить точки линиями (см. рис. 1).

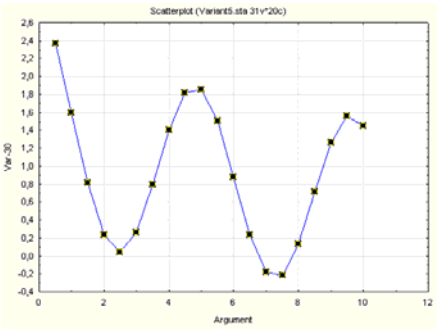

Рис. 1. Пример графика

Задание 4. Выбирая в настройках Advanced аппроксимирующую функцию, определить, какая функция является наилучшей аппроксимацией для предложенных данных.

Задание 5. В модуле "Nonlinear Estimation" – "Нелинейное оценивание" собраны процедуры, позволяющие оценить нелинейные зависимости между данными. Вы можете

выбрать различные модели зависимостей, задать собственную функцию, выбрать метод оценивания неизвестных параметров.

Провести нелинейный регрессионный анализ данных. Для этого воспользоваться модулем Statistics / Advanced Linear/Nonlinear models / Nonlinear estimation / user-specified regression, least squares. Кнопкой Function to be estimated ввести модельную функцию (см. таблицу 2). Кнопкой Variables ввести зависимую переменную.

В появившемся окне во вкладке Advanced задать критерий сходимости 10^{-3} (по умолчанию 10^{-6}), кнопкой Start values ввести произвольные начальные значения искомых коэффициентов (b1, b2, b3). Начальное приближение *желательно* выбирать из указанного в таблице 1 интервала, в противном случае сходимость решения к истинному не гарантируется.

Показать, что полученные коэффициенты нелинейной регрессии далеки от реальных при неудачном начальном приближении. Включить типичные графики в отчёт. Как зависит успех расчётов от хорошего начального приближения? Как это влияет на скорость сходимости? Сделать выводы.

Найти коэффициенты нелинейной регрессии. Провести анализ остатков и показать, что найдено наилучшее решение.

Задание 6. Ввести в формулу модельной функции точное значение для двух коэффициентов, найденных в предыдущем задании и вновь провести регрессионный анализ поиска только одного коэффициента, введя для него «хорошее» начальное приближение. Насколько уменьшилась ошибка регрессии в этом случае? Увеличилась ли скорость сходимости?

Задание 7.

Данные для анализа предоставлены торговой организацией. Смысл переменных в таблице: 1 – год, 2 – месяц, 3 – сколько товара продали, 4 – его себестоимость, 5 – почём продавали, 6 – доход (разность 5 и 4 переменной), 7 – торговая наценка в процентах.

Построить разумную с вашей точки зрения нелинейную регрессионную модель по предложенным данным. Оценить адекватность модели по остаткам.

Задание 9. Для этих же данных проверить гипотезу о нормальном распределении случайных величин Var1 Var9 по критерию χ^2 автоматически, как это делалось в задании 7. Сравнить с результатами Задания 8 и объяснить полученный результат.

Лабораторная работа № 6

КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ

1. ЦЕЛЬ РАБОТЫ

Целью работы является изучение и проведение кластерного анализа.

2. ВАЖНОЕ УКАЗАНИЕ

Изучите theory-8.doc, что поможет быстрому изучению теоретического материала.

3. КЛАСТЕРНЫЙ АНАЛИЗ

Задание

Исходные данные – файл wooden.sta представляют собой статистику отдела сбыта предприятия по производству стройматериалов. Для каждого вида продукции указан объём продаж по месяцам и выручка.

Провести кластерный анализ с целью:

- сегментировать продукцию по месяцам года;
- повышения выручки путём рационализации выпуска нужной продукции по месяцам.

Предположить существование трёх кластеров, на которые сегментируются выручка и объём продукции. По результатам анализа дать рекомендации для принятия управленческих решений: что необходимо сделать для улучшения качества производственного процесса.