МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ МЕТОДОВ ПРЕДВАРИТЕЛЬНОГО ПРЕДСТАВЛЕНИЯ БИОМЕДИЦИНСКИХ ДАННЫХ

Методические указания к лабораторной работе

САМАРА 2012

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ
БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

ИССЛЕДОВАНИЕ МЕТОДОВ ПРЕДВАРИТЕЛЬНОГО ПРЕДСТАВЛЕНИЯ БИОМЕДИЦИНСКИХ ДАННЫХ

САМАРА 2012

УДК 519.688

Составитель: В.Н. Конюхов

Исследование методов предварительного представления биомедицинских данных. Метод. указания к лабораторной работе/ Самар. гос. аэрокосм. ун-т; Сост. В.Н. Конюхов, Самара, 2012. 31с.

В методических указаниях изложены сведения об основных методах предварительного представления данных, таких как описательные статистики и графические методы анализа данных. Показаны области применения, назначение и ограничения. Рассмотрены некоторые пакеты прикладных программ, позволяющие решать задачи предварительного представления данных. Приведено краткое описание пакета MATLAB и его основных функций.

Методические указания предназначены для бакалавров, обучающихся по направлению подготовки 201000.62 (Биотехнические системы и технологии) и выполняющих лабораторные работы по дисциплине «Автоматизация обработки биомедицинской информации» на кафедре радиотехники и медицинских диагностических систем.

Печатаются по решению редакционно-издательского совета Самарского государственного аэрокосмического университета им. академика С.П. Королёва

Рецензент: проф. Гречишников В.М.

Цель работы: изучение методов предварительного представления биомедицинских данных, описательных статистик, способов графического представления данных, а также основных задач, решаемых с помощью этих методов и ограничений, накладываемых на них.

1. КРАТКИЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1.1. ОСНОВНЫЕ ПОЛОЖЕНИЯ.

Главной задачей любого исследования или научного анализа является установление взаимосвязей между различными явлениями природы. Эти взаимосвязи есть ничто иное, как зависимости одних переменных, количественных или качественных, от других. Для установления зависимостей необходимо измерить значения переменных, причем измерения могут проводиться в различных шкалах- номинальной, порядковой (ранговой), интервальной, относительной[1]. Таким образом, по сути, определить, являются ли переменные зависимыми или нет, какой вид имеет зависимость, насколько она сильна и какова ее достоверность, можно лишь обрабатывая данные, полученные в результате опыта. Выбор методов обработки во многом определяется характером полученной совокупности данных.

Одними из основных особенностей данных, описывающих биологический объект, является их существенная многомерность и большое разнообразие типов. Переменные, характеризующие состояние биологического объекта, могут быть получены из различных источников: первичных документов, например истории болезни, лабораторных анализов, с помощью технических средств мониторинга состояния и т.п. Следствием такого положения является то, что количество переменных, описывающих биообъект, может достигать десятков и сотен, причем измеренных в различных шкалах. При этом общее число данных может достигать значительного числа.

---

[1] ***Номинальные переменные*** используются только для качественной классификации. Это означает, что данные переменные могут быть измерены только в терминах принадлежности к некоторым, существенно различным классам; при этом вы не сможете определить количество или упорядочить эти классы. Например, вы сможете сказать, что 2 индивидуума различимы в терминах переменной А (например, индивидуумы принадлежат к разным национальностям). Типичные примеры номинальных переменных - пол, национальность, цвет, город и т.д. Часто номинальные переменные называют категориальными.

Порядковые переменные позволяют ранжировать (упорядочить) объекты, указав какие из них в большей или меньшей степени обладают качеством, выраженным данной переменной. Однако они не позволяют сказать "на сколько больше" или "на сколько меньше". Порядковые переменные иногда также называют ординальными. Типичный пример порядковой переменной - социоэкономический статус семьи. Мы понимаем, что верхний средний уровень выше среднего уровня, однако сказать, что разница между ними равна, скажем, 18% мы не сможем. Само расположение шкал в следующем порядке: номинальная, порядковая, интервальная является хорошим примером порядковой шкалы.

Интервальные переменные позволяют не только упорядочивать объекты измерения, но и численно выразить и сравнить различия между ними. Например, температура, измеренная в градусах Фаренгейта или Цельсия, образует интервальную шкалу. Вы можете не только сказать, что температура 40 градусов выше, чем температура 30 градусов, но и что увеличение температуры с 20 до 40 градусов вдвое больше увеличения температуры от 30 до 40 градусов.

Относительные переменные очень похожи на интервальные переменные. В дополнение ко всем свойствам переменных, измеренных в интервальной шкале, их характерной чертой является наличие определенной точки абсолютного нуля, таким образом, для этих переменных являются обоснованными предложения типа: х в два раза больше, чем у. Типичными примерами шкал отношений являются измерения времени или пространства. Например, температура по Кельвину образует шкалу отношения, и вы можете не только утверждать, что температура 200 градусов выше, чем 100 градусов, но и что она вдвое выше. Интервальные шкалы (например, шкала Цельсия) не обладают данным свойством шкалы отношения. Заметим, что в большинстве статистических процедур не делается различия между свойствами интервальных шкал и шкал отношения (данные с сайта statsoft.ru, электронный учебник).

В такой ситуации, выявление полезной информации, раскрывающей структурно-функциональные связи и закономерности исследуемой системы, требует использование различных методов обработки данных, направленных на сжатие исходной информации и выделение наиболее информативных параметров, которые необходимы для решения поставленной задачи исследования.

К таким методам относятся методы предварительного представления данных, основная цель которых заключается в компактном описании массива исходных данных, выявлении общих тенденций и особенностей в поведении объекта исследования. Получение компактного описания позволяет быстро интерпретировать данные, выдвинуть гипотезы относительно их структуры и взаимосвязи и, при необходимости, принять решение о корректировке исследования, например с целью получения дополнительных значений переменных состояния, повышения точности измерений и т.д.

Основными способами решения задач предварительного представления данных являются сжатие и преобразование данных, их сглаживание и построение различных диаграмм и графиков. Выбор того или иного способа определяется целями проведения исследования.

Учитывая, что биомедицинские данные являются в общем случае случайными величинами, то разумно применять способы обработки этих данных основанные на методах математической статистики.

1.2. СЖАТИЕ ДАННЫХ.

В общем случае биомедицинские данные можно представить с помощью матрицы данных:

$$\begin{pmatrix} X_{11}, X_{12} \cdots X_{1n} \\ X_{21}, X_{22} \cdots X_{2n} \\ \cdots\cdots\cdots\cdots\cdots \\ X_{N1}, X_{N2} \cdots X_{Nn} \end{pmatrix},$$

где X_{ij} - значение j-го показателя относящегося к i-ому объекту, N- количество объектов наблюдения, n- количество измеряемых показателей. В простейшем случае, если объект один или показатель один, то матрица данных трансформируется в вектор-столбец или вектор-строку. Элементы матрицы данных представляют собой случайные величины, выбранные из генеральной совокупности. Таким образом, каждый столбец представляет собой выборку из генеральной совокупности. **Выборкой** $x_1, \ldots, x_n$ объема n из совокупности, распределенной по $F(x)$, называется n независимых наблюдений над случайной величиной ξ с функцией распределения $F(x)$.

При решении конкретных прикладных задач выборка, зачастую, обладает достаточно большой избыточностью, что затрудняет анализ данных, маскируя основные закономерности в их поведении.

Сжатие данных позволяет представить данные или их характеристики в более компактном и наглядном виде. Наиболее часто на практике для сжатия данных исходную выборку представляют вариационным рядом, интервальным рядом распределения (гистограммой), группой обобщенных выборочных характеристик.

Вариационным рядом $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ называется выборка из N элементов, записанная в порядке возрастания ее элементов.

В случае большого количества различных значений, или когда измеряемая характеристика принимает непрерывный ряд значений, строят **интервальный ряд распределения (гистограмму)**:

X	x_1, x_2	.....	x_i, x_{i+1}	.....	x_{n-1}, x_n
m	m_1	.....	m_i	.....	m_n
P=m/N	p_1	.....	p_i	.....	p_n

Здесь m_i - частота попадания в i-ый интервал, p_i - вероятность попадания в i-ый интервал.

По гистограмме можно визуально дать оценку форме распределения случайной величины, и, если наложить на гистограмму график нормального распределения, сравнить, насколько они близки[2] (рис.1).

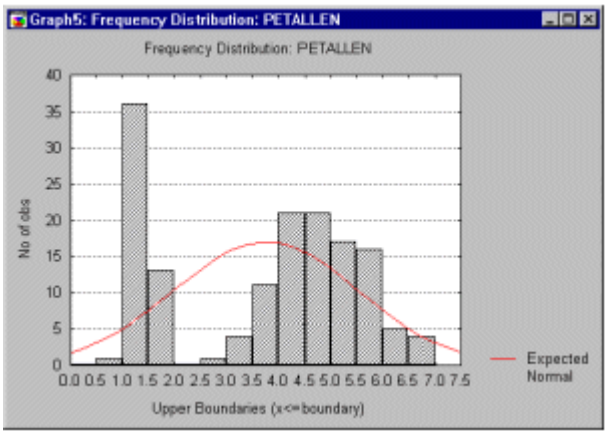

Рисунок 1- Гистограмма распределения случайной величины.

При построении гистограммы важно выбрать оптимальную величину интервала группирования $h = x_{i+1} - x_i$, т.к. неправильный выбор существенно искажает форму распределения. Например, если взять интервал равным размаху экспериментальных данных $h = x_n - x_1$, то каким бы ни было

[2] Если распределение нормальное, то существенно упрощаются дальнейшие математические преобразования данных, описание всего массива данных можно свести к двум параметрам- математическому ожиданию и дисперсии. Кроме того, для нормального распределения существует много решенных практических задач. Однако, для того чтобы результаты анализа были корректными, только визуального анализа недостаточно. Необходимо проверить гипотезу о нормальности с помощью соответствующих критериев согласия.

исходное распределение, гистограмма будет соответствовать равномерному распределению. Если число интервалов разбиения будет равно трем, то любое колоколообразное распределение сведется к треугольному распределению.

Существуют различные походы к выбору интервала группирования. Не углубляясь в этот специфический вопрос, приведем формулу Стерджеса, как одну из исторически первых, для оценки оптимальной величины интервала h. В соответствии с ней

$$h \approx \frac{x_{\max} - x_{\min}}{1 + 2{,}322 \lg(n)},$$

где n - объем выборки. Начало первого интервала $x_1 = x_{\min} - 0{,}5 \cdot h$.

Вариационные и интервальные ряды дают наглядное представление об изменении того или иного количественного признака. Однако они не позволяют количественно сопоставить различные группы данных.

1.3. ОПИСАТЕЛЬНЫЕ СТАТИСТИКИ.

Для количественного сопоставления используются обобщенные выборочные характеристики данных представляющие собой различные статистики. **Статистикой** называется любая функция наблюдаемых случайных переменных $x_1, x_2 \cdots x_n$.

Если есть основания предполагать, что выборка получена из генеральной совокупности с известным параметрическим распределением, то для ее описания обычно достаточно несколько параметров, представляющих собой статистики наблюдаемых переменных[3]. Например, для описания нормального распределения достаточно двух параметров- математического ожидания и дисперсии. Тогда, если эти параметры одинаковы для двух выборок, то можно говорить, что выборки получены из одной и той же генеральной совокупности (в предположении, что распределение нормальное)[4].

На практике в качестве обобщенных характеристик наиболее часто используются выборочные среднее, дисперсия, среднеквадратичное отклонение, коэффициенты асимметрии и эксцесса, медиана, мода и ряд других.

Выборочное среднее (математическое ожидание) определяется как:

В некоторых случаях гипотезу о нормальности распределения можно отвергнуть на основе визуального анализа. Так, на рис.1 гистограмма явно не соответствует требованиям нормальности распределения. В частности, гистограмма бимодальна.

[3] В параметрической статистике рассматриваются семейства параметрических распределений, такие как нормальное распределение, логарифмически нормальное, экспоненциальное, гамма-распределение, распределение Вейбулла-Гнеденко и др., которые зависят от одного, двух или трех параметров. Поэтому для полного описания распределения достаточно знать или оценить одно, два или три числа.

[4] Следует отметить, что выборочные характеристики являются случайными величинами и имеют свое собственное распределение. Поэтому отличие значений выборочных характеристик, как и их равенство, полученных из двух различных выборок, не может однозначно истолковываться в пользу совпадения или несовпадения генеральных совокупностей, из которых они получены. Необходимо дополнительно проверять статистические гипотезы, с помощью соответствующих статистических критериев, о достоверности различий выборочных характеристик.

$$\widetilde{m}_x = \frac{1}{n}\sum_{i=1}^{n} x_i \ ,$$

и представляет собой центр тяжести выборочного распределения.

Дисперсия и среднеквадратичное отклонение являются мерой разброса данных относительно среднего значения и определяются, соответственно, как:

$$\widetilde{D}_x = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widetilde{m}_x\right)^2 \ ,$$

$$\widetilde{\sigma}_x = \sqrt{\widetilde{D}_x} \ .$$

Выборочный коэффициент асимметрии представляет собой количественную меру несимметричности распределения:

$$\widetilde{a}_x = \frac{1}{(n-1)\widetilde{\sigma}_x^3}\sum_{i=1}^{n}\left(x_i - \widetilde{m}_x\right)^3 \ .$$

Например, если коэффициент асимметрии существенно отличается от 0, то распределение несимметрично, в то время как нормальное распределение абсолютно симметрично и коэффициент асимметрии равен 0. Коэффициент асимметрии распределения с длинным правым хвостом положителен. Если распределение имеет длинный левый хвост, то его коэффициент асимметрии отрицателен.

Выборочный коэффициент эксцесса является количественной мерой крутизны распределения по сравнению с нормальным распределением:

$$\vec{e}_x = \frac{1}{(n-1)\widetilde{\sigma}_x^4}\sum_{i=1}^{n}\left(x_i - \widetilde{m}_x\right)^4 - 3 \ .$$

Если коэффициент эксцесса существенно отличен от 0, то распределение имеет или более закругленный пик, чем нормальное, или, напротив, имеет более острый пик (возможно, имеется несколько пиков). Обычно, если коэффициент эксцесса положителен, то пик заострен, если отрицательный, то пик закруглен. Эксцесс нормального распределения равен 0.

В том случае, если нет достаточных оснований выдвигать гипотезу о пренадлежности данных какому либо параметрическому распределению, в том числе и из-за малого объема данных, или переменные измеренны в номинальной или порядковой шкале, определение обычных описательных статистик не дает надежной и достаточной информации[5]. В такой ситуации используются непараметрические описательные статистики, которые являются различными мерами

---

[5] Например, в психометрии хорошо известно, что воспринимаемая интенсивность стимулов (например, воспринимаемая яркость света) представляет собой логарифмическую функцию реальной интенсивности (яркости, измеренной в объективных единицах - люксах). В данном примере, обычная оценка среднего (сумма значений, деленная на число стимулов) не дает верного представления о среднем значении действительной интенсивности стимула (данные с сайта statsoft.ru, электронный учебник)/

положения и рассеяния. К таковым, например, относятся медиана, мода, минимальное и максимальное значение выборки, размах выборки, квантили распределения и т.д.

Медиана выборки- это значение, которое разбивает выборку на две равные части. Половина наблюдений лежит ниже медианы, и половина наблюдений лежит выше медианы. Медиана вычисляется следующим образом. Изучаемая выборка упорядочивается в порядке возрастания. Получаемая последовательность x_i, где i=1…2*m+1 называется вариационным рядом. Если число наблюдений нечетно, то медиана оценивается как: x_{m+1}. Если число наблюдений четно, то медиана оценивается как:

$$m = \frac{x_{m+1} + x_m}{2}.$$

Выборочная мода- это значение x_i, которое наиболее часто встречается в выборке (на рис.1 мода лежит в диапазоне 1-1.5) .

Размах выборки определяется как $R = x_{\max} - x_{\min}$.

Квантиль распределения- такое значение x_p, что p-часть всех значение выборки меньше или равна x_p. Некоторые квантили имеют собственное название. Так $x_{0.25}$ называется нижней квартилью распределения, $x_{0.1}$ - нижней децелью, $x_{0.5}$ - медиана.

1.4. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

Для наглядного представления о характере изучаемой совокупности данных используются различные графики и диаграммы. Существует множество типов графиков и диаграмм для представления данных, которые можно классифицировать по ряду признаков:

- по форме построения графического образа (рис.2);
- по геометрическим знакам, отображающим статистические показатели;
- по задачам, решаемым с помощью графического изображения и т.д.

Рисунок 2- Классификация статистических графиков по форме графического образа

Прежде чем отображать результаты измерений в графической форме, нужно ясно представить себе цель подобного отображения. Таких целей обычно бывает две: либо подчеркнуть общую тенденцию, сгладив второстепенные детали, либо сконцентрировать внимание на особенностях поведения изучаемой характеристики. В зависимости от этого и выбирается такой способ графического отображения, который способствует наглядности восприятия и полезен для последующего анализа результатов.

Ряд значений, полученный в результате измерения, может быть отображен графически различными способами: в виде эмпирической функции распределения вероятностей, полигона частот или гистограммы частот. Они представляют собой наглядное изображение вероятностных характеристик случайных величин: функции распределения вероятности и функции распределения плотности вероятности. Наиболее часто для графического представления используются гистограммы относительных частот, т.к. в этом случае по виду полученной фигуры можно предста ю вить вид функции распределения плотности вероятности. Отображать данные можно в виде двух- и трехмерных гистограмм непрерывных или целочисленных данных. Основная проблема при построении гистограмм заключается в выборе оптимальных длин интервалов разбиения, т.к. при большом числе интервалов вид гистограммы становится чувствителен к случайным ошибкам измерений, а уменьшение числа интервалов сглаживает особенности распределения плотности вероятности изучаемой величины. При сравнении двух или нескольких выборок с различными объемами на одном графике необходимо переходить к относительным частотам, т.к. в противном случа е отображаемые результаты будут несопоставимы между собой.

Представление данных в виде схематических диаграмм используется, если требуется сравнивать между собой несколько выборок или объектов, характеризующихся одним и тем же набором параметров. Для каждого вида схематических диаграмм рассчитываются обобщенные числовые характеристики выборки и представляются в виде схемы определенного вида. Удобство их заключается в том, что происходит концентрация внимания на самом важном и устранение второстепенных деталей.

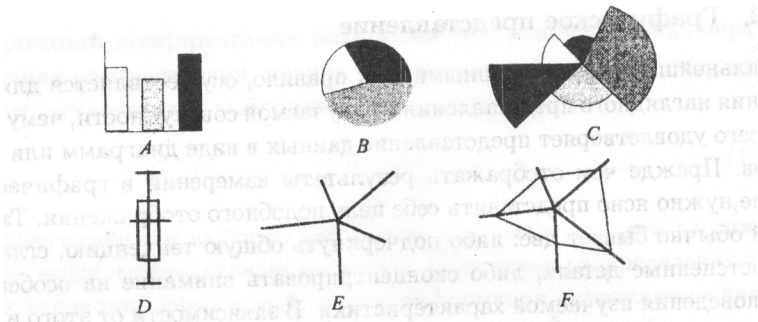

Рисунок3- Типы диаграмм. A-столбиковая, B-круговая, C- секторно-полярная, D-“ящик с усами”, F-звездная, E- солнечная.

Существует большое число видов схематических диаграмм. Наряду с общеупотребительными круговыми и столбиковыми диаграммами, которые позволяют отображать данные, классифицированные по какому-либо фактору (как правило, выраженные в процентах), часто используются следующие виды диаграмм (рис. 3):

• Секторно-полярные диаграммы. Удобно использовать для отображения сезонных, временных изменений показателя. Площадь всей диаграммы пропорциональна значению показателя, а площадь секторов - значению показателя в соответствующие сезоны.

• Диаграммы в виде «ящика с усами» (Box-and-Whisker plot). Выборка изображается в виде прямоугольника, расположенного вокруг срединного значения (среднего значения, медианы и т.п.), причем длина прямоугольника представляет стандартизованный размах выборки (стандартная ошибка, стандартное отклонение, интерквартильный размах и т.п.). Усы снаружи ящика представляют фиксированный процент от стандартизированного размаха. Иногда, если каждое значение выборки имеет смысловой эквивалент (например, название болезни, местности в которой проводилось измерение и т.п.), бывает нелишним указывать на диаграмме ряд крайних значений в виде отдельных точек. По виду такой диаграммы бывает удобно судить о центрировании изучаемой величины, а также сравнивать между собой различные выборки.

Строиться «ящика с усами» следующим образом. Горизонтальная линия соответствует тому значению первой (интервальной) переменной, которое является медианой. Верхняя и нижняя границы прямоугольника соответствуют соответственно 75-му и 25-му процентилям , таким образом "внутрь" прямоугольника попадает 50% наблюдений. Из ящика вверх и вниз "торчат" усы, которые заканчиваются в наибольшем и наименьшем значениях, не являющихся выбросами. Выбросами считаются значения, лежащие в диапазонах, больших, чем в 1.5 раза, чем высота ящика от его верхней и нижней границы соответственно. Выбросы на графике обозначаются кружочками и звездочками.

• Звездные диаграммы. Изображение каждого объекта представляется в виде звездчатой фигуры, вершины которой соответствуют измеряемым величинам, а длины радиус-векторов вершин пропорциональны значениям соответствующих переменных у данного объекта.

• Солнечные диаграммы. Изображение каждого объекта представляется в виде звездчатой фигуры, лучи которой соответствуют переменным, длины лучей- стандартным отклонениям, точки, в которых контур фигуры пересекает лучи,- отклонениям значений переменной на данном объекте от среднего значения.

2. ПАКЕТЫ ПРИКЛАДНЫХ ПРОГРАММ ДЛЯ АНАЛИЗА БИОМЕДИЦИНСКИХ ДАННЫХ

Все пакеты прикладных программ, которые могут быть использованы для статистической обработки данных, можно разделить на универсальные, специализированные и узкоспециализированные пакеты.

Универсальные пакеты предназначены для выполнения различных инженерных и научных расчетов, в том числе и статистического анализа данных. Преимуществом данных пакетов является то, что с помощью них можно не только проводить анализ данных, но и комплексно решать поставленную задачу, например, моделировать систему, проводить оптимизацию и т.д. Как правило, подобные пакеты имеют встроенный язык программирования и позволяют создавать законченные пользовательские приложения, позволяющие решать конкретную задачу. Примерами таких пакетов программ являются MATLAB, Excel, Mathcad, Maple и ряд других.

Наиболее известной программой для анализа данных является приложение MS Excel из пакета офисных программ компании Microsoft MS Office. Это обусловлено широким распространением данного пакета офисных программ в России, наличием русскоязычной версии, интеграцией с MS Word и PowerPoint. Однако в MS Excel набор статистических функций достаточно мал, плохо развита графика. MS Excel хорошо подходит для накопления данных, промежуточного преобразования, предварительных статистических прикидок, для построения некоторых видов диаграмм.

Специализированные пакеты ориентируются исключительно на статистическую обработку различных данных. Преимуществом подобных программ является наличие большого числа специализированных функций и процедур для анализа данных, охватывающих современные методы анализа. К таким пакетам программ относятся SPSS (Statistical Package for Social Science), STATA, STATISTICA и другие.

SPSS- это пакет статистической обработки данных с более чем 30-и летней историей. Отличается гибкостью, мощностью, применим для всех видов статистических расчетов применяемых в биомедицине. Существует полностью русифицированная версия SPSS 12.0.2 для Windows. Есть учебник на русском языке, позволяющий шаг за шагом освоить возможности SPSS, репетитор по статистике на русском языке, помогающий в выборе нужной статистической или графической процедуры для конкретных данных и задач.

STATA- это профессиональный статистический программный пакет, который может применятся для биомедицинских целей. Один из самых популярных в образовательных и научных учреждениях США наряду с SPSS. Программа хорошо документирована, издается специальный журнал для пользователей системы.

STATISTICA включает большое количество методов статистического анализа (более 250 встроенных функций) объединенных следующими специализированными статистическими модулями: основные статистики и таблицы, непараметрическая статистика, дисперсионный анализ, множественная регрессия, нелинейное оценивание, анализ временных рядов и прогнозирование, кластерный анализ, факторный анализ, дискриминантный функциональный анализ, анализ длительностей жизни, каноническая корреляция, многомерное шкалирование, моделирование структурными уравнениями и др. Несложный в освоении этот статистический пакет может быть рекомендован для биомедицинских исследований любой сложности. Российское представительство компании (http://www.statsoft.ru/) предлагает полностью русифицированную 6-ю версию программы. Сайт компании содержит много информации по статистической обработке медицинских данных, учебник по статистике на русском языке.

Узкоспециализированные пакеты предназначены для обработки данных, в какой либо узкой области- экономике, биомедицине и т.д. Преимуществом подобных программ является наиболее полный учет особенностей применения статистического анализа в конкретной области знаний. Примером подобного пакета программ является PRISM. Эта программа создавалась специально для биомедицинских целей. Интуитивно понятный интерфейс позволяет в считанные минуты проанализировать данные и построить качественные графики. Программа содержит основные часто применяемые статистические функции, которых в большинстве исследований будет достаточно. Однако, как отмечают сами разработчики, программа не может полностью заменить серьезных статистических пакетов.

3.КРАТКОЕ ОПИСАНИЕ ПАКЕТА MATLAB

MATLAB, как язык программирования, был разработан в конце 1970-х годов. Целью разработки служила задача дать студентам возможность использования программных библиотек Linpack и EISPACK без необходимости изучения Фортрана. Вскоре новый язык распространился среди университетов, и был с большим интересом встречен учёными, работающими в области прикладной математики. В дальнейшем создатели языка переписали MATLAB на C и основали в 1984 компанию The MathWorks для дальнейшего развития. Эти переписанные на C библиотеки долгое время были известны под именем JACKPAC. Первоначально MATLAB предназначался для проектирования систем управления, но быстро завоевал популярность во многих других научных и инженерных областях. Он также широко использовался в образовании, в частности для преподавания линейной алгебры и численных методов.

Язык MATLAB является высокоуровневым языком программирования, включающим основанные на матрицах структуры данных, широкий спектр функций, интегрированную среду разработки, объектно-ориентированные возможности и интерфейсы к программам, написанным на других языках программирования.

Программы, написанные на MATLAB, бывают двух типов — функции и скрипты. Функции

имеют входные и выходные аргументы, а также собственное рабочее пространство для хранения промежуточных результатов вычислений и переменных. Скрипты же используют общее рабочее пространство. Как скрипты, так и функции не интерпретируются в машинный код и сохраняются в виде текстовых файлов. Существует также возможность сохранять так называемые pre-parsed программы — функции и скрипты, обработанные в вид, удобный для машинного исполнения. В общем случае такие программы выполняются быстрее.

Основной особенностью языка MATLAB является его широкие возможности по работе с матрицами. MATLAB предоставляет пользователю большое количество (несколько сотен) функций для анализа данных, покрывающие практически все области математики, в частности.

MATLAB предоставляет удобные средства для разработки алгоритмов, включая высокоуровневые с использованием концепций объектно-ориентированного программирования. В нём имеются все необходимые средства интегрированной среды разработки, включая отладчик и профайлер. Функции для работы с целыми типами данных облегчают создание алгоритмов для микроконтроллеров и других приложений, где это необходимо.

В составе пакета MATLAB имеется большое количество функций для построения графиков, в том числе трёхмерных, визуального анализа данных и создания анимированных роликов.

Встроенная среда разработки позволяет создавать графические интерфейсы пользователя с различными элементами управления, такими как кнопки, поля ввода и другими. С помощью компонента MATLAB Compiler эти графические интерфейсы могут быть преобразованы в самостоятельные приложения.

Пакет MATLAB включает различные интерфейсы для получения доступа к внешним подпрограммам, написанным на других языках программирования, данным, клиентам и серверам, общающимся через технологии Component Object Model или Dynamic Data Exchange, а также периферийным устройствам, которые взаимодействуют напрямую с MATLAB. Многие из этих возможностей известны под названием MATLAB API.

Пакет MATLAB предоставляет доступ к функциям, позволяющим создавать, манипулировать и удалять COM-объекты (как клиенты, так и сервера). Поддерживается также технология ActiveX. Все COM-объекты принадлежат к специальному COM-классу пакета MATLAB.

Пакет MATLAB содержит функции, которые позволяют ему получать доступ к другим приложениям среды Windows, равно как и этим приложениям получать доступ к данным MATLAB, посредством технологии динамического обмена данными (DDE). Каждое приложение, которое может быть DDE-сервером, имеет своё уникальное идентификационное имя. Для MATLAB это имя-MATLAB.

В MATLAB существует возможность вызывать методы веб-сервисов. Специальная функция создаёт класс, основываясь на методах API веб-сервиса.

MATLAB взаимодействует с клиентом веб-сервиса с помощью принятия от него посылок, их обработки и посылок ответа. Поддерживаются следующие технологии: Simple Object Access

Protocol (SOAP) и Web Services Description Language (WSDL).

Интерфейс для последовательного порта пакета MATLAB обеспечивает прямой доступ к периферийным устройствам, таким как модемы, принтеры и научное оборудование, подключающееся к компьютеру через последовательный порт (COM-порт). Интерфейс работает путём создания объекта специального класса для последовательного порта. Имеющиеся методы этого класса позволяют считывать и записывать данные в последовательный порт, использовать события и обработчики событий, а также записывать информацию на диск компьютера в режиме реального времени. Это бывает необходимо при проведении экспериментов, симуляции систем реального времени и для других приложений.

Пакет MATLAB включает интерфейс взаимодействия с внешними приложениями, написанными на языках C и Фортран. Осуществляется это взаимодействие через MEX-файлы. Существует возможность вызова подпрограмм, написанных на C или Фортране из MATLAB, как будто это встроенные функции пакета. MEX-файлы представляют собой динамически подключаемые библиотеки, которые могут быть загружены и исполнены интерпретатором, встроенным в MATLAB.

Для MATLAB имеется возможность создавать специальные наборы инструментов (англ. toolbox), расширяющих его функциональность. Наборы инструментов представляют собой коллекции функций, написанных на языке MATLAB для решения определённого класса задач. Компания Mathworks поставляет наборы инструментов, которые используются во многих областях, включая следующие:

Цифровая обработка сигналов, изображений и данных: DSP Toolbox, Image Processing Toolbox, Wavelet Toolbox, Communication Toolbox, Filter Design Toolbox — наборы функций, позволяющих решать широкий спектр задач обработки сигналов, изображений, проектирования цифровых фильтров и систем связи.

Системы управления: Control Systems Toolbox, μ-Analysis and Synthesis Toolbox, Robust Control Toolbox, System Identification Toolbox, LMI Control Toolbox, Model Predictive Control Toolbox, Model-Based Calibration Toolbox — наборы функций, облегчающих анализ и синтез динамических систем, проектирование, моделирование и идентификацию систем управления, включая современные алгоритмы управления, такие как робастное управление, H∞-управление, ЛМН-синтез, μ-синтез и другие.

Финансовый анализ: GARCH Toolbox, Fixed-Income Toolbox, Financial Time Series Toolbox, Financial Derivatives Toolbox, Financial Toolbox, Datafeed Toolbox — наборы функций, позволяющие быстро и эффективно собирать, обрабатывать и передавать различную финансовую информацию.

Анализ и синтез географических карт, включая трёхмерные: Mapping Toolbox.

Сбор и анализ экспериментальных данных: Data Acquisition Toolbox, Image Acquisition Toolbox, Instrument Control Toolbox, Link for Code Composer Studio — наборы функций, позволяющих сохранять и обрабатывать данные, полученные в ходе экспериментов, в том числе в реальном

времени. Поддерживается широкий спектр научного и инженерного измерительного оборудования.

Визуализация и представление данных: Virtual Reality Toolbox — позволяет создавать интерактивные миры и визуализировать научную информацию с помощью технологий виртуальной реальности и языка VRML.

Средства разработки: MATLAB Builder for COM, MATLAB Builder for Excel, MATLAB Compiler, Filter Design HDL Coder — наборы функций, позволяющих создавать независимые приложения из среды MATLAB.

Взаимодействие с внешними программными продуктами: MATLAB Report Generator, Excel Link, Database Toolbox, MATLAB Web Server, Link for ModelSim — наборы функций, позволяющие сохранять данные различных видов таким образом, чтобы другие программы могли с ними работать.

Базы данных: Database Toolbox — инструменты работы с базами данных.

Научные и математические пакеты: Bioinformatics Toolbox, Curve Fitting Toolbox, Fixed-Point Toolbox, Fuzzy Logic Toolbox, Genetic Algorithm and Direct Search Toolbox, OPC Toolbox, Optimization Toolbox, Partial Differential Equation Toolbox, Spline Toolbox, Statistic Toolbox, RF Toolbox — наборы специализированных математических функций, позволяющие решать широкий спектр научных и инженерных задач, включая разработку генетических алгоритмов и другие.

Нейронные сети: Neural Network Toolbox — инструменты для синтеза и анализ нейронных сетей.

Символьные вычисления: Symbolic Math Toolbox — инструменты для символьных вычислений с возможностью взаимодействия с символьным процессором программы Maple.

Помимо вышеперечисленных, существуют тысячи других наборов инструментов для MATLAB, написанных другими компаниями и энтузиастами.

Информация по пакету MATLAB, а также по его применениям в различных областях науки техники, широко представлена и легко доступна в сети Интернет. На сайте **matlab.exponenta.ru** Вы можете найти подробную информацию по пакету MATLAB. Для первоначального ознакомления с пакетом рекомендуется прочитать файлы «MatLab. Руководство для начинающих» и Intro_Matlab. **Список статистических функций, необходимых для выполнения лабораторной работы, приведен в приложении А. Справку по использованию той или иной функции можно получить, набрав в командной строке help «имя функции». Например, команда help hist выведет следующий текст:**

```
HIST  Histogram.
   N = HIST(Y) bins the elements of Y into 10 equally spaced containers
   and returns the number of elements in each container.  If Y is a
   matrix, HIST works down the columns.
   N = HIST(Y,M), where M is a scalar, uses M bins.
   N = HIST(Y,X), where X is a vector, returns the distribution of Y
   among bins with centers specified by X.  Note: Use HISTC if it is
```

more natural to specify bin edges instead.

[N,X] = HIST(...) also returns the position of the bin centers in X.

HIST(...) without output arguments produces a histogram bar plot of the results.

See also HISTC.

4.ОПИСАНИЕ ДАННЫХ И ЭКСПОРТ ДАННЫХ В ПАКЕТ MATLAB

Данные содержаться в файле rus_dasl16.xsl в формате электронных таблиц Excel. Данные включают в себя 4 переменных: var1 (липопротеин высокой плотности), var2 (общий холестерин), var3 (курение), var4 (индекс массы тела).

Экспорт данных в пакет MATLAB можно осуществить следующим образом. Пометьте требуемые данные в таблице Excel и скопируете их в буфер обмена. Создайте документ Ms Word и вставьте данные из буфера обмена. Далее сохраните документ Ms Word в формате текст MS DOS с форматированием. В результате должен получиться файл с расширением «asc». Измените расширение на «txt». Файл данных готов для импорта в MATLAB.

Примечание. Данные должны быть расположены в столбик. В качестве разделителя целой и десятичной части использоваться точка.

Далее в пакете MATLAB в меню «файл» выберите «Import Data» и загрузите данные из файла в рабочую среду MATLAB. Данным будет автоматически присвоено имя переменной, и станут возможные все операции с этой переменной, определенные в среде MATLAB.

Примечание. Если включена надстройка обмена данными с Excel, то возможен непосредственный экспорт данных из электронных таблиц Excel.

5. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

1. Запустите пакет MATLAB и ознакомьтесь с элементами управления. Набрав команду demo в командной строке, ознакомьтесь с основными возможностями пакета.
2. Загрузите в Excel файл rus_dasl16.xsl.
3. Создайте шесть выборок переменной var2 объемом n=100 для индексов массы тела в диапазоне от 18 до 24.5 и в диапазоне от 26 и более по следующим правилам. Первые две выборки var2 для индексов массы тела в диапазоне от 18 до 24.5 и в диапазоне от 26 и более без учета переменной var3- одна для индекса массы тела в диапазоне от 18 до 24.5, другая для индекса массы тела в диапазоне от 26 и более. Аналогично следующие две выборки var2 для курящих для индексов массы тела в диапазоне от 18 до 24.5 и в диапазоне от 26 и более и две выборки var2 для некурящих для индексов массы тела в диапазоне от 18 до 24.5 и в диапазоне от 26 и более. Импортируйте полученные выборки в пакет MATLAB. Сформируйте для каждой выборки

вариационный ряд.

4. Вычислите для каждой выборки выборочные среднюю, дисперсию, среднеквадратичное отклонение, коэффициенты асимметрии и эксцесса, медиану, моду.
5. Постройте для каждой выборки гистограмму распределения. **При построении гистограммы определите оптимальный (по Стерджесу) интервал группирования и построение гистограммы осуществите в соответствии с вычисленным оптимальным интервалом.**
6. Постройте диаграмму в виде «ящика с усами» для каждой выборки.

6. СОДЕРЖАНИЕ ОТЧЕТА

1. Наименование и цель работы.
2. Исходные выборки для расчетов и соответствующие им вариационные ряды.
3. Вычисленные выборочные значения для каждой выборки.
4. Гистограммы распределения и диаграммы полученные в п.3 и п.4.
5. Выводы.

Примечание. ***Выводы должны быть сделаны относительно полученных результатов и в соответствии с целями лабораторной работы, а именно- существуют ли различия по группам, какие методы предварительного представления данных позволяют наиболее четко заметить эти различия, влияют ли выбросы в данных на полученные результаты и .т.д. Не допускаются выводы описывающие процедуру выполнения лабораторной работы, например, такие как «Были сформированы вариационные ряды», «В результате выполнения лабораторной работы построены гистограммы распределения» и т.п.***

7. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Перечислите шкалы, в которых могут проводиться измерения переменных. В чем их отличие?
2. Какие основные особенности данных, описывающих биологический объект?
3. В чем заключается основная цель методов предварительного представления данных?
4. Перечислите основные способы решения задачи представления данных.
5. Что такое выборка и матрица данных?
6. Что такое вариационный и интервальный ряд?
7. Как влияет выбор интервала группирования на вид гистограммы?
8. Приведите формулу Стерджеса.
9. Что такое статистика?
10. Приведите выражения для наиболее часто используемых выборочных обобщенных характеристик.
11. Что характеризует коэффициент асимметрии?

12. Что характеризует коэффициент эксцесса?
13. В каких случаях используются непараметрические описательные статистики?
14. Приведите примеры непараметрических описательных статистик.
15. Что такое квантиль распределения?
16. Для каких целей используется графическое представление данных?
17. По каким признакам можно классифицировать различные типы диаграмм и графиков?
18. Поясните, как строятся основные типы диаграмм.
19. На какие группы можно разделить пакеты прикладных программ для статистической обработки данных? В чем заключаются отличия между ними? Приведите примеры.
20. Покажите, как можно экспортировать данные в среду MATLAB.
21. Каким образом можно вызвать справку по конкретной функции в среде MATLAB?
22. Покажите, как можно построить гистограмму средствами MATLAB.
23. Покажите, как можно построить диаграмму «ящик с усами» средствами MATLAB.

СПИСОК ЛИТЕРАТУРЫ

1. Половко А. М., Бутусов П. Н. MATLAB для студента. — СПб.: БХВ-Петербург, 2005. -320 с.

2. О.Ю. Реброва. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. М., МедиаСфера, 2003. 312с.

3. Новиков Д.А., Новочадов В.В. Статистические методы в медико-биологическом эксперименте (типовые случаи). Волгоград: Издательство ВолГМУ, 2005. – 84 с.

ПРИЛОЖЕНИЕ А

Список функций Statistics Toolbox

Оценка параметров закона распределения по экспериментальным данным

- **betafit** - Оценка параметров бета распределения
- **binofit** - Оценка параметров биномиального распределения
- **nbinfit** - Оценка параметров отрицательного биномиального распределения
- **expfit** - Оценка параметров экспоненциального распределения
- **gamfit** - Оценка параметров гамма распределения
- **normfit** - Оценка параметров нормального распределения
- **poissfit** - Оценка параметров распределения Пуассона
- **raylfit** - Оценка параметров распределения Релея
- **unifit** - Оценка параметров равномерного распределения

- **weibfit** - Оценка параметров распределения Вейбулла
- **mle** - Расчет функции максимального правдоподобия

Законы распределения случайных величин

- **betacdf** - Бета распределение
- **binocdf** - Биномиальное распределение
- **cdf** - Параметризованная функция распределения
- **chi2cdf** - Функция распределения хи-квадрат
- **expcdf** - Экспоненциальное распределение
- **ecdf** - Эмпирическая функция распределения (на основе оценки Каплана-Мейера)
- **fcdf** - Распределение Фишера
- **gamcdf** - Гамма распределение
- **geocdf** - Геометрическое распределение
- **hygecdf** - Гипергеометрическое распределение
- **logncdf** - Логнормальное распределение
- **nbincdf** - Отрицательное биномиальное распределение
- **ncfcdf** - Смещенное распределение Фишера
- **nctcdf** - Смещенное распределение Стьюдента
- **ncx2cdf** - Смещенное хи-квадрат распределение
- **normcdf** - Нормальное распределение
- **poisscdf** - Распределение Пуассона
- **raylcdf** - Распределение Релея
- **tcdf** - Распределение Стьюдента
- **unidcdf** - Дискретное равномерное распределение
- **unifcdf** - Непрерывное равномерное распределение
- **weibcdf** - Распределение Вейбулла

Функции плотности распределения случайных величин

- **betapdf** - Бета распределение
- **binopdf** - Биномиальное распределение
- **chi2pdf** - Функция распределения хи-квадрат
- **exppdf** - Экспоненциальное распределение
- **fpdf** - Распределение Фишера
- **gampdf** - Гамма распределение

- **geopdf** - Геометрическое распределение
- **hygepdf** - Гипергеометрическое распределение
- **lognpdf** - Логнормальное распределение
- **nbinpdf** - Отрицательное биномиальное распределение
- **ncfpdf** - Смещенное распределение Фишера
- **nctpdf** - Смещенное распределение Стьюдента
- **ncx2pdf** - Смещенное хи-квадрат распределение
- **normpdf** - Нормальное распределение
- **poisspdf** - Распределение Пуассона
- **mvnpdf** - Функция плотности вероятности многомерного нормального распределения
- **raylpdf** - Распределение Релея
- **pdf** - Параметризованная функция плотности распределения
- **tpdf** - Распределение Стьюдента
- **unidpdf** - Дискретное равномерное распределение
- **unifpdf** - Непрерывное равномерное распределение
- **weibpdf** - Распределение Вейбулла

Обратные функции распределения случайных величин

- **betainv** - Бета распределение
- **binoinv** - Биномиальное распределение
- **chi2inv** - Функция распределения хи-квадрат
- **expinv** - Экспоненциальное распределение
- **finv** - Распределение Фишера
- **gaminv** - Гамма распределение
- **geoinv** - Геометрическое распределение
- **hygeinv** - Гипергеометрическое распределение
- **icdf** - Параметризованная обратная функция распределения
- **logninv** - Логнормальное распределение
- **nbininv** - Отрицательное биномиальное распределение
- **ncfinv** - Смещенное распределение Фишера
- **nctinv** - Смещенное распределение Стьюдента
- **ncx2inv** - Смещенное хи-квадрат распределение
- **norminv** - Нормальное распределение

- **poissinv** - Распределение Пуассона
- **raylinv** - Распределение Релея
- **tinv** - Распределение Стьюдента
- **unidinv** - Дискретное равномерное распределение
- **unifinv** - Непрерывное равномерное распределение
- **weibinv** - Распределение Вейбулла

Генерация псевдослучайных чисел по заданному закону распределения

- **betarnd** - Бета распределение
- **binornd** - Биномиальное распределение
- **chi2rnd** - Функция распределения хи-квадрат
- **exprnd** - Экспоненциальное распределение
- **frnd** - Распределение Фишера
- **gamrnd** - Гамма распределение
- **geornd** - Геометрическое распределение
- **hygernd** - Гипергеометрическое распределение
- **iwishrnd** - Обратная матрица случайных чисел распределения Уишарта
- **lognrnd** - Логнормальное распределение
- **mvnrnd** - Многомерное нормальное распределение
- **mvtrnd** - Многомерное распределение Стьюдента
- **nbinrnd** - Отрицательное биномиальное распределение
- **ncfrnd** - Смещенное распределение Фишера
- **nctrnd** - Смещенное распределение Стьюдента
- **ncx2rnd** - Смещенное хи-квадрат распределение
- **normrnd** - Нормальное распределение
- **poissrnd** - Распределение Пуассона
- **random** - Параметризованная функция генерации псевдослучайных чисел
- **raylrnd** - Распределение Релея
- **trnd** - Распределение Стьюдента
- **unidrnd** - Дискретное равномерное распределение
- **unifrnd** - Непрерывное равномерное распределение
- **weibrnd** - Распределение Вейбулла
- **wishrnd** - Матрица случайных чисел распределения Уишарта

Оценка математического ожидания и дисперсии по заданному закону распределения и его параметрам

- **betastat** - Бета распределение
- **binostat** - Биномиальное распределение
- **chi2stat** - Функция распределения хи-квадрат
- **expstat** - Экспоненциальное распределение
- **fstat** - Распределение Фишера
- **gamstat** - Гамма распределение
- **geostat** - Геометрическое распределение
- **hygestat** - Гипергеометрическое распределение
- **lognstat** - Логнормальное распределение
- **nbinstat** - Отрицательное биномиальное распределение
- **ncfstat** - Смещенное распределение Фишера
- **nctstat** - Смещенное распределение Стьюдента
- **ncx2stat** - Смещенное хи-квадрат распределение
- **normstat** - Нормальное распределение
- **poisstat** - Распределение Пуассона
- **raylstat** - Распределение Релея
- **tstat** - Распределение Стьюдента
- **unidstat** - Дискретное равномерное распределение
- **unifstat** - Непрерывное равномерное распределение
- **weibstat** - Распределение Вейбулла

Расчет логарифма функции максимального правдоподобия

- **betalike** - Расчет логарифма функции максимального правдоподобия бета распределения
- **gamlike** - Расчет логарифма функции максимального правдоподобия гамма распределения
- **normlike** - Расчет логарифма функции максимального правдоподобия нормального распределения
- **weiblike** - Расчет логарифма функции максимального правдоподобия распределения Вейбулла
- **nbinlike** - Расчет логарифма функции максимального правдоподобия отрицательного биномиального распределения

Функции описательной статистики

- **bootstrp** - Бутстреп оценки. Оценка статистик для данных с дополненным объемом выборки посредством математического моделирования
- **corrcoef** - Оценка коэффициента корреляции (функция MATLAB)

- **cov** - Оценка матрицы ковариаций (функция MATLAB)
- **crosstab** - Кросстабуляция для нескольких векторов с положительными целыми элементами
- **geomean** - Среднее геометрическое
- **grpstats** - Сводные статистики по группам
- **harmmean** - Среднее гармоническое
- **iqr** - Разность между 75% и 25% квантилями или между 3-й и 1-ой квартилями
- **kurtosis** - Оценка коэффициента эксцесса (в отечественной литературе коэффициент эксцесса определяется как b_2=kurtosis-3)
- **mad** - Среднее абсолютное отклонение от среднего значения
- **mean** - Среднее арифметическое (функция MATLAB)
- **median** - Медиана (функция MATLAB)
- **moment** - Оценка центрального момента. Порядок момента задается как аргумент функции
- **nanmax** - Максимальное значение в выборке. Нечисловые значения в выборке игнорируются
- **nanmean** - Среднее арифметическое выборки. Нечисловые значения в выборке игнорируются
- **nanmedian** - Медиана выборки. Нечисловые значения в выборке игнорируются
- **nanmin** - Минимальное значение в выборке. Нечисловые значения в выборке игнорируются
- **nanstd** - Оценка среднего квадратического отклонения выборки. Нечисловые значения в выборке игнорируются
- **nansum** - Сумма элементов выборки. Нечисловые значения в выборке игнорируются
- **prctile** - Выборочная процентная точка (процентиль)
- **range** - Размах выборки
- **skewness** - Оценка коэффициента асимметрии
- **std** - Оценка среднего квадратического отклонения (функция MATLAB)
- **tabulate** - Определение частот целых положительных элементов вектора случайных значений
- **trimmean** - Оценка среднего арифметического значения, находимая с игнорированием заданного процента минимальных и максимальных элементов в выборки
- **var** - Оценка дисперсии

Функции статистических графиков

- **boxplot** - График "Ящик с усами". График 0%, 25%, 50%, 75%, 100% процентилей выборки
- **cdfplot** - График кумулятивной кривой по эмпирическим данным
- **fsurfht** - Контурный график заданной функции. Операция построения графика выполняется интерактивно.
- **gline** - Операция прорисовки прямой линии в текущем графике
- **gname** - Нанесение меток на график

- **gplotmatrix** - Матрица графиков рассеяния группированных по общей переменной
- **gscatter** - График рассеяния двух переменных группированных по значениям третьей переменной
- **lsline** - График рассеяния двух переменных с линией регрессии по методу наименьших квадратов
- **normplot** - Нормальный вероятностный график
- **qqplot** - График "квантиль-квантиль" для двух выборок
- **refcurve** - Построение полиномиальной кривой на текущий график
- **refline** - Построение прямой на текущий график
- **surfht** - Контурный график по матрице данных
- **weibplot** - Вероятностный график Вейбулла

Функции статистического контроля качества

- **capable** - Расчет индексов воспроизводимости процесса Cp, Cpk
- **capaplot** - График воспроизводимости процесса
- **ewmaplot** - Контрольная карта экспоненциально взвешенного среднего
- **histfit** - Гистограмма по негруппированным экспериментальным данным с наложенной на нее кривой функции плотности распределения нормального закона
- **normspec** - График функции плотности нормального закона с наложенными границами допусков контролируемого параметра
- **schart** - Контрольная карта среднего квадратического отклонения
- **xbarplot** - Контрольная карта среднего арифметического

Функции линейного регрессионного анализа

- **anova1** - Однофакторный дисперсионный анализ
- **anova2** - Двухфакторный дисперсионный анализ
- **anovan** - Многофакторный дисперсионный анализ
- **aoctool** - Однофакторный анализ ковариационных моделей. Выходными параметрами функции являются:
  - Интерактивный график исходных данных линейных математических моделей
  - Таблица однофакторного дисперсионного анализа ·
  - Таблица с оценками параметров математических моделей
- **dummyvar** - Условное кодирование переменных. Функция возвращает матрицу единиц и нулей содержащую число колонок равное сумме чисел возможных значений в столбцах исходной матрицы. Единицы и нули характеризуют отсутствие или наличие определенного значения в каждой колонки исходной матрицы.
- **friedman** - Тест Фридмана (непараметрический двухфакторный дисперсионный анализ Фридмана)
- **glmfit** - Определение параметров обобщенной линейной модели

- **glmval** - Прогнозирование с использованием обобщенной линейной модели
- **kruskalwallis** - Тест Краскала-Уоллиса (непараметрический однофакторный дисперсионный анализ)
- **leverage** - Оценка степени влияния отдельных наблюдений в исходном многомерном множестве данных на значения параметров линии регрессии.
- **lscov** - Линейная регрессия (метод наименьших квадратов) при заданной матрице ковариаций (встроенная функция MATLAB)
- **manova1** - Однофакторный многомерный дисперсионный анализ
- **manovacluster** - Дендрограмма, показывающая группировку исходных данных в кластеры по средним значениям. В качестве исходных данных используются выходные данные однофакторного многомерного дисперсионного анализа (manova1)
- **multcompare** - Множественной сравнение оценок средних, параметров линии регрессии и т.д. В качестве входных параметров используются выходные параметры функций anova1, anova2, anovan, aoctool, friedman, kruskalwallis.
- **polyconf** - Определение доверительных интервалов для линии регрессии
- **polyfit** - Полиномиальная регрессия (встроенная функция MATLAB)
- **polyval** - Прогноз с использованием полиномиальной регрессии (встроенная функция MATLAB)
- **rcoplot** - График остатков
- **regress** - Множественная линейная регрессия
- **regstats** - Функция диагностирования линейной множественной модели. Графический интерфейс.
- **ridge** - Линейная регрессия с применением гребневых оценок (ридж-регрессия)
- **rstool** - Интерактивный подбор и визуализация поверхности отклика
- **robustfit** - Робастная оценка параметров регрессионной модели
- **stepwise** - Пошаговая регрессия (графический интерфейс пользователя)

Функции нелинейного регрессионного анализа

- **lsqnonneg** - Функция реализует метод наименьших квадратов и возвращает только неотрицательные значения параметров модели (встроенная функция MATLAB)
- **nlinfit** - Нелинейный метод наименьших квадратов (метод Гаусса-Ньютона)
- **nlintool** - График прогнозируемых значений
- **nlparci** - Вектор доверительных интервалов для параметров модели
- **nlpredci** - Прогнозируемые значения и их доверительные интервалы

Функции планирования эксперимента

- **bbdesign** - Планы Бокса-Бенкена
- **candexch** - D-оптимальный план (на основе алгоритма перестановки строк для формирования множества возможных значений)

- **candgen** - Генерирует множество возможных сочетаний факторов соответствующих D-оптимальному плану
- **ccdesign** - Центральный композиционный план
- **cordexch** - Функция для определения точного D-оптимального плана эксперимента на основе алгорима обмена координатами
- **daugment** - Определение матрица плана дополняющую матрицу заданного плана до D-оптимального
- **dcovary** - Функция для построения D-оптимального блочного плана
- **ff2n** - Определение плана полного факторного эксперимента для факторов имеющих 2 уровня
- **fracfact** - Функция для формирования двухуровнего дробного факторного плана
- **fullfact** - Функция формирования плана полного факторного эксперимента для числа уровней факторов задаваемых пользователем
- **hadamard** - Матрица Адамара. Матрица Адамара соответствует плану дробного факторного эксперимента для факторов, каждый из которых задан на отрезке [-1 1]. И служит для построения линейной регрессионной модели. (Встроенная функция MATLAB)
- **lhsdesign** - План на основе латинских квадратов
- **lhsnorm** - Латинские квадраты для многомерной нормальной выборки
- **rowexch** - Функция для определения точного D-оптимального плана на основе алгоритма обмена строк

Функции кластерного анализа

- **cluster** - Деление иерархического дерева кластеров (группировка выходных данных функции linkage) на отдельные кластеры
- **clusterdata** - Группировка матрицы исходных данных в кластеры
- **cophenet** - Расчет коэффициента качества разбиения исходных данных на кластеры (этот коэффициент можно рассматривать как аналог коэффициента корреляции, чем его значение ближе к 1, тем лучше выполнено разбиение на кластеры)
- **dendrogram** - Дендрограмма кластеров
- **inconsistent** - Расчет коэффициентов несовместимости для каждой связи в иерархическом дереве кластеров и может использоваться как оценка качества разбиения на кластеры
- **kmeans** - Кластеризация на основе внутригрупповых средних
- **linkage** - Формирование иерархического дерева бинарных кластеров
- **pdist** - Расчет парных расстояний между объектами (векторами) в исходном множестве данных
- **silhouette** - График силуэта кластеров
- **squareform** - Преобразование вектора выходных данных функции pdist в симметричную квадратную матрицу

Функции снижения размерности задачи

- **factoran** - Факторный анализ
- **pcacov** - Функция служит для реализации метода главных компонент по заданной в качестве входного параметра матрице ковариаций
- **pcares** - Функция служит для определения остатка после удаления заданного количества главных компонент
- **princomp** - Функция служит для реализации метода главных компонент по заданной в качестве входного параметра матрице исходных значений

Функции анализа многомерных случайных величин

- **barttest** - Тест Бартлета
- **canoncorr** - Канонический корреляционный анализ
- **cmdscale** - Классическое многомерное шкалирование
- **classify** - Линейный дискриминантый анализ
- **mahal** - Функция определяет расстояния Махаланобиса между строками двух матриц, являющихся входными параметрами.
- **manova1** - Однофакторный многомерный дисперсионный анализ
- **procrustes** - Ортогональное вращение, позволяющее поставить в прямое соответствие одно множество точек другому

Функции нелинейного регрессионного анализа на основе графа возможных решений

- **treedisp -** Отображает граф возможных решений
- **treefit -** Построение графа возможных решений на основе исходных данных
- **treeprune -** Исключение незначимых решений в графе возможных решений
- **treetest -** Оценка погрешности узлов графа возможных решений
- **treeval -** Оценка параметров регрессионной модели с использованием графа возможных решений

Статистическая проверка гипотез

- **ranksum** - Ранговый тест Вилкоксона для проверки однородности двух генеральных совокупностей
- **signrank** - Знаковый тест Вилкоксона для проверки гипотезы о равенстве медиан двух выборок
- **signtest** - Знаковый тест для проверки гипотезы о равенстве медиан двух выборок
- **ttest** - t-test для одной выборки. Проверка гипотезы о равенстве (или неравенстве) математического ожидания выборки заданному значению при условии, что величина дисперсии неизвестна. Закон распределения выборки нормальный.
- **ttest2** - t-test для двух выборок. Проверка гипотезы о равенстве (или неравенстве) математических ожиданий двух выборок при условии, что величины дисперсий выборок неизвестны и равны. Закон распределения выборки нормальный.

- **ztest** - Z-тест. Проверка гипотезы о равенстве (или неравенстве) математического ожидания выборки заданному значению при условии, что известна величина дисперсии. Закон распределения выборки нормальный.

Проверка статистических гипотез о согласии распределения экспериментальным данным

- **jbtest** - Тест на соответствие выборки нормальному распределению с неопределенными параметрами нормального распределения. Этот тест является асимптотическим и не может быть использован на малых выборках. Для проверки гипотезы о соответствии выборки нормальному распределению на малых выборках необходимо использовать функцию lillietest.
- **kstest** - Тест Колмогорова-Смирнова на соответствие выборки заданному распределению
- **kstest2** - Тест Колмогорова-Смирнова на соответствие распределений двух выборок
- **lillietest** - Тест на соответствие выборки нормальному распределению. Параметры нормального распределения рассчитываются исходя из значений элементов в выборке.

Проверка непараметрических гипотез

- **friedman** - Тест Фридмана (непараметрический двухфакторный дисперсионный анализ Фридмана)
- **kruskalwallis** - Тест Краскала-Уоллиса (непараметрический однофакторный дисперсионный анализ)
- **ksdensity** - Подгонка функции плотности вероятности по экспериментальным данным
- **ranksum** - Ранговый тест Вилкоксона для проверки однородности двух генеральных совокупностей
- **signrank** - Знаковый тест Вилкоксона для проверки гипотезы о равенстве медиан двух выборок
- **signtest** - Знаковый тест для проверки гипотезы о равенстве медиан двух выборок

Запись и чтение данных из файлов

- **caseread** - Функция для чтения данных из текстового файла. Возвращает матрицу символов из текстового файла
- **casewrite** - Функция для записи строковой матрицу в текстовый файл
- **tblread** - Функция для чтения табличных данных из текстового файла
- **tblwrite** - Функция для записи табличных данных из текстового файла
- **tdfread** - Функция для чтения табличных данных разделенных знаком табуляции в строке из текстового файла

Таблица демонстрационных примеров

- **aoctool** - Интерактивное средство ковариационного анализа
- **disttool** - Интерактивное средство для исследования функций распределения случайных величин
- **glmdemo** - Пример использования обобщенной линейной модели
- **randtool** - Интерактивное средство для генерации псевдослучайных чисел

- **polytool** - Интерактивное определение параметров полиномиальной модели
- **rsmdemo** - Интерактивное моделирование химическое реакции и нелинейный регрессионный анализ
- **robustdemo** - Интерактивное средство для сравнения методов МНК и робастной регрессии

Таблица вспомогательных функций

- **combnk** - Вычисляет количество комбинаций которыми можно выбрать k объектов из n
- **grp2idx** - Преобразование группирующей переменной в индексы массива
- **hougen** - Функция прогнозирования для модели Хогена
- **tiedrank** - Расчет ранга выборки с учетом ее объема
- **zscore** - Выполняет нормализацию матрицы по колонкам. Приводит значения по колонкам матрицы к нормальным с 0 математическим ожиданием и единичной дисперсией.

Учебное издание

ИССЛЕДОВАНИЕ МЕТОДОВ ПРЕДВАРИТЕЛЬНОГО ПРЕДСТАВЛЕНИЯ БИОМЕДИЦИНСКИХ ДАННЫХ

Методические указания

Составитель: Конюхов Вадим Николаевич

Самарский государственный аэрокосмический университет
имени академика С.П. Королёва.
443086 Самара, Московское шоссе, 34