Министерство образования и науки Российской федерации
Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ» (ТУСУР)

Н.Н.Несмелова

Статистическая обработка данных

(методические указания по практическим занятиям
и самостоятельной работе студентов,
обучающихся по направлениям
022000.62 «Экология и природопользование»,
280700.62 «Техносферная безопасность»)

2013

Аннотация

Методические указания для студентов, обучающихся по направлениям 022000.62 «Экология и природопользование» и 280700.62 «Техносферная безопасность» содержат описания практических занятий по дисциплине «Статистическая обработка данных» и задания для самостоятельной работы. Дополнительно приведены вопросы для подготовки к итоговой аттестации по дисциплине и тесты, которые могут использоваться как на занятиях, в качестве промежуточного контроля знаний, так и для самопроверки студентов в ходе самостоятельной работы. Методические указания сопровождаются списком литературы для самоподготовки. Методические указания могут использоваться студентами при подготовке отчетов по научно-исследовательской работе студентов, отчетов по групповому проектному обучению, выпускной квалификационной работы, а также преподавателями дисциплины «Статистическая обработка данных» для подготовки к занятиям.

Оглавление

1. Создание и редактирование файлов данных в программе «STATISTICA»

STATISTICA работает с четырьмя типами документов, которые выводятся в собственном окне рабочей области системы:

1) электронная таблица для ввода исходных данных и их преобразования (файлы с расширением *sta)*;
2) электронная таблица для вывода численных и текстовых результатов (файлы с расширение *scr*);
3) график – документ в специальном графическом формате для визуализации и графического представления численной информации (файлы с расширение *stg*);
4) отчёт – документ в расширенном текстовом формате для вывода текстовой и графической информации (файлы с расширение RTF);

STATISTICA может работать как с числовыми, так и с текстовыми данными. В частности, электронные таблицы данных могут содержать и числовую, и текстовую информацию и поддерживают различные типы операций с данными:

- операции с использованием *буфера обмена Windows;*
- операции с выделенными блоками значений;
- автозаполнение блоков и т. д.

Электронная таблица данных состоит из строк и столбцов, которые имеют разные смысловые значения. Столбцы электронной таблицы данных называются *Variables – Переменные,* а строки *Cases – Случаи.* В качестве переменных обычно выступают исследуемые величины, а случаи – это значения, которые принимают переменные и которые измеряются в процессе наблюдения. Максимальное число переменных в таблице – 4092, число случаев до 2 000 000.

Способы ввода данных в электронную таблицу:

- с клавиатуры;
- вычислить новые данные на основе уже введенных с помощью формул, при написании которых можно пользоваться библиотекой математических и статистических функций, а также использовать логические операторы;
- перенести в STATISTICA данные из других приложений путем операций копирования или импорта данных.

Приведём пример табличной организации данных в системе STATISTICA.

Пример. Предположим, что наблюдается температура в течение нескольких дней. Данные таких измерений могут быть занесены в таблицу, имеющую следующую структуру:

Номер наблюдения	Дата	Температура
1	1 – янв – 91	-20.5
2	2 – янв – 91	-19.3
3	3 – янв - 91	-23.7

Задание: создайте файл данных Statistica и внесите в него данные из таблицы.

Полученная таблица содержит две переменные: «Дата» и «Температура», измеренные в трех случаях с номерами 1, 2 и 3.

Нажмите кнопку Vars и выберите в открывшемся меню пункт Current Specs. Откроется окно для работы с текущей переменной. Для каждой переменной в таблице данных можно задать:

- формат отображения данных;
- код, который приписывается пропущенным данным – пустым ячейкам в электронной таблице;

- длинные имена переменных и комментарии к переменным;
- метки для текстовых значений переменных, содержащие длинные текстовые значения;
- формулы для перекодировки или преобразования значений переменных.
-

Основные операции над переменными (кнопка Vars)

Команда	Действие
Add Variables	Добавление переменных.
Move Variables	Перемещение переменных.
Copy Variables	Копирование переменных.
Delete Variables	Удаление переменных.
Current Specs	Открытие диалогового окна, позволяющего задать спецификации текущей переменной.
All Specs	Просмотр и редактирование спецификаций всех переменных в таблице данных.
Text Values	Открытие диалогового окна, в котором может быть установлено или изменено соответствие между текстовыми и числовыми значениями переменной.
Date Values	Основные операции с датами: позволяет создать дату из нескольких переменных или разбить дату на несколько переменных. Можно также перевести дату в текстовые значения или наоборот.
Recalculate Variables	Пересчет значений переменных, которые связаны формулами.
Shift (Lag) Variables	Сдвиг значений переменной на несколько случаев вперед или назад.
Rank Variables	Ранжирование значений переменной.
Recode Variables	Перекодировка значений переменной.

Основные операции над случаями (кнопка Cases)

Команда	Действие
Add Cases	Добавление случаев (строк) в таблицу.
Move Cases	Перемещение строк.
Copy Cases	Копирование строк.
Delete Cases	Удаление строк.
Case Names	Задание имен случаев.

Задание 1.

1. Создайте файл Valeo, содержащий данные валеологического обследования группы томичей.
2. Добавьте еще одну переменную – ВИК (вегетативный индекс Кердо) и введите значения этой переменной в электронную таблицу используя формулу: ВИК=(1-АДД/Пульс)*100. Используйте команды Add Variables и Current Specs.
3. Введите в переменную пол текстовые значения: женский (код 0) и мужской (код 1). Используйте команду Text Values.
4. Введите длинные имена переменных: АДС – артериальное давление систолическое; АДД – артериальное давление диастолическое. Используйте команду All Specs.
5. Создайте новую переменную «Возр_гр» (длинное название – возрастная группа). Введите значения этой переменной, используя формулу с логическими условиями: если возраст до 30 лет – группа 1 (молодость); если возраст от 30 до 55 лет – группа 2 (зрелость), если возраст больше 55 лет – группа 3 (пожилые). Используйте команду Recode Variables.
6. Определите тип каждой переменной в соответствии с измерительными шкалами.

№	ПОЛ	ВОЗРАСТ	ВЕС	РОСТ	АДС	АДД	ПУЛЬС
1	0	43	50	157	110	80	88
2	1	63	76	180	155	80	76
3	0	21	59	165	120	80	92
4	1	40	89	175	140	105	88
5	0	43	89	158	120	70	60
6	1	45	89	176	115	75	72
7	1	17	93	180	118	85	88
8	0	20	58	167	120	78	72
9	0	20	64	169	105	80	72
10	1	48	67	167	100	75	88
11	0	41	50	161	110	65	80
12	0	17	60	176	100	65	80
13	1	17	57	182	105	68	72
14	1	20	69	176	120	80	66
15	1	19	74	179	135	75	96
16	0	54	73	156	125	75	86
17	1	59	62	168	110	75	64
18	0	37	60	168	120	80	80
19	1	50	74	177	120	75	64
20	0	18	50	168	120	95	60
21	0	49	58	168	110	80	112
22	1	42	104	181	130	80	80
23	0	50	63	164	118	85	72
24	1	48	73	173	120	90	52
25	0	18	69	166	95	55	76
26	0	52	75	146	135	95	84
27	0	41	92	177	138	80	80
28	0	50	75	149	148	90	74
29	0	24	47	159	90	60	72
30	0	50	58	155	120	80	64
31	0	51	74	166	135	78	78
32	0	43	76	177	120	80	72
33	0	21	74	173	94	60	80
34	0	18	57	169	100	60	76
35	0	55	75	172	120	80	72
36	0	46	50	155	118	74	88
37	0	18	47	162	120	80	80
38	0	30	56	158	118	65	90
39	1	47	95	175	155	92	72
40	0	31	54	162	120	80	92

2. Методы визуализации и графического анализа данных

Для современных компьютерных средств анализа данных характерно наличие всесторонней графической поддержки. Графические средства используются для визуализации как исходных данных, так и результатов статистического анализа.

Программа «Statistica» включает в себя большое количество разнообразных типов двумерных и трехмерных графиков, причем графические средства доступны на любом шаге статистического анализа и в любом модуле. Каждый график выводится в своем собственном окне, его можно редактировать, копировать, печатать, вставлять в документы, которые созданы в других программах (например, в текстовые документы Word), а также сохранять на жестком диске или на дискете, как файлы с расширением *.stg.

Рассмотрим некоторые приемы графического анализа данных, доступные в программе «Statistica».

Категоризованные графики

Одним из наиболее мощных аналитических методов исследования является разделение данных на группы для сравнения структуры получившихся подмножеств. Эти методы широко применяются как в разведочном анализе данных, так и при проверке гипотез и известны под разными названиями (классификация, группировка, категоризация, разбиение, расслоение и пр.). Для количественного описания различий между группами наблюдений разработаны специальные методы, такие как, например, дисперсионный анализ. Однако графические средства позволяют выявить закономерности, которые трудно обнаружить с помощью вычислительных процедур.

Термин "категоризованные графики" впервые был использован в программе STATISTICA в 1990 году. Эти графики представляют собой наборы двумерных, трехмерных, тернарных или n-мерных графиков (таких как гистограммы, диаграммы рассеяния, линейные графики, поверхности, тернарные диаграммы рассеяния и пр.), по одному графику для каждой выбранной категории (подмножества) наблюдений. Эти "входящие" графики располагаются последовательно в одном графическом окне, позволяя сравнивать структуру данных для каждой из указанных подгрупп.

Методы категоризации. Существует пять основных методов категоризации значений: целые числа, категории, границы, коды и сложные подгруппы.

1. Целые числа. При использовании этого режима для определения категорий будут использованы целые значения выбранной группирующей переменной, и для всех наблюдений, принадлежащих каждой категории будет построено по одному графику. Если выбранная группирующая переменная содержит не целочисленные значения, то программа автоматически округлит каждое значение выделенной переменной до целого числа.
2. Категории. В этом режиме категоризации нужно указать желаемое число категорий. Программа разделит весь диапазон значений выбранной группирующей переменной (от минимального до максимального) на указанное число интервалов равной длины.
3. Границы. Метод границ также представляет собой интервальную категоризацию, однако в этом случае интервалы могут иметь произвольную (например, различную) длину, определяемую пользователем (например, "меньше -10", "больше или равно -10, но меньше 0", "больше или равно 0, но меньше 10" и "больше или равно 10").
4. Коды. Этот метод следует использовать в том случае, если выбранная группирующая переменная содержит "коды " (т.е. особые смысловые значения, такие как «Мужчина», «Женщина»), по которым можно разбить данные на категории.
5. Сложные подгруппы. Этот метод дает возможность пользователю использовать для выделения подгрупп более одной переменной. Например, можно указать шесть категорий,

задаваемых комбинациями значений трех переменных «Пол», «Возраст» и «Образование».

Гистограммы

Гистограммы используются для изучения распределений частот значений переменных. Такое распределение показывает, какие именно конкретные значения или диапазоны значений исследуемой переменной встречаются наиболее часто, насколько различаются эти значения, расположено ли большинство наблюдений около среднего значения, является распределение симметричным или асимметричным, полимодальным или одномодальным и т.д. Гистограммы также используются для сравнения наблюдаемых и теоретических распределений.

Частотные распределения могут представлять интерес по двум основным причинам:

- по форме распределения можно судить о природе исследуемой переменной (например, бимодальное распределение позволяет предположить, что выборка не является однородной и содержит наблюдения, принадлежащие двум различным множествам, которые в свою очередь нормально распределены).
- многие статистики основываются на определенных предположениях о распределениях анализируемых переменных; гистограммы позволяют проверить, выполняются ли эти предположения.

Как правило, работа с новым набором данных начинается с построения гистограмм всех переменных.

Задание № 1.

1. Откройте файл Valeo. С помощью программы Statistica постройте гистограммы всех переменных из этого файла, поместите их в ваш отчет по лабораторной работе. Проанализируйте полученные гистограммы и ответьте на вопросы (для каждой переменной):

- **Является ли выборка однородной или она представляет собой смесь из нескольких выборок?**
- **Имеются ли в выборке аномальные объекты, выбросы?**
- **Подчиняется ли характер распределения нормальному закону? Какие свойства гистограммы позволяют ответить на данный вопрос?**

2. Постройте для переменных «Рост» и «Вес» категоризированные гистограммы по переменной «Пол». Проведите анализ этих гистограмм.

3. Используя методы категоризация «граница» постройте гистограммы переменных «САД» и «Пульс» отдельно для молодых (моложе 30 лет) и для лиц среднего возраста. Проведите анализ этих гистограмм.

Диаграммы рассеяния

Двумерные диаграммы рассеяния используются для визуализации взаимосвязей между двумя переменными X и Y (например, весом и ростом). На этих диаграммах отдельные точки данных представлены маркерами на плоскости, где оси соответствуют переменным. Две координаты (X и Y), определяющие положение точки, соответствуют значениям переменных. Если между переменными существует сильная взаимосвязь, то точки на графике образуют упорядоченную структуру (например, прямую линию или характерную кривую). Если переменные не взаимосвязаны, то точки образуют "облако".

С помощью диаграмм рассеяния можно исследовать и нелинейные взаимосвязи между переменными. При этом не существует каких-либо "автоматических" или простых способов оценки нелинейности. Стандартный коэффициент корреляции Пирсона r позволяет оценить только линейность связи, а некоторые непараметрические корреляции, например, Спирмена R, дают возможность оценить нелинейность, но только для

монотонных зависимостей. На диаграммах рассеяния можно изучить структуру взаимосвязей, чтобы затем с помощью преобразования привести данные к линейному виду или выбрать подходящую нелинейную подгонку.

Задание № 2.

1. **Используя диаграммы рассеяния, изучите взаимосвязи между перменными файла Valeo. Сформулируйте свои гипотезы о характере и возможных причинах выявленных взаимосвязей.**
2. **Постройте и проанализируйте категоризированные диаграммы рассеяния для переменных «САД» и «ДАД»; «Рост» и «Вес»; «Рост» и «Возраст», проведя категоризацию по переменным «Пол» и «Возраст».**

Диаграммы размаха

На диаграммах размаха (этот термин был впервые использован Тьюки в 1970 году) представлены диапазоны значений выбранной переменной (или переменных) для отдельных групп наблюдений. Для выделения этих групп используются от одной до трех категориальных (группирующих) переменных или набор логических условий выбора подгрупп. Для каждой группы наблюдений вычисляется центральная тенденция (медиана или среднее), а также размах или изменчивость (квартили, стандартные ошибки или стандартные отклонения). Выбранные параметры отображаются на графике одним из пяти способов (Прямоугольники-Отрезки, Отрезки, Прямоугольники, Столбцы или Верхние-нижние засечки). На этом графике можно показать и выбросы.

Можно выделить два основных направления использования диаграмм размаха:

a) отображение диапазонов значений отдельных элементов, наблюдений или выборок (например, типичные минимаксные графики цен на акции или товары или графики агрегированных данных с диапазонами);
b) отображение изменения значений в отдельных группах или выборках (например, когда точкой внутри прямоугольника представлено среднее значение для каждой выборки, сам прямоугольник соответствует значениям стандартной ошибки, а меньший прямоугольник или пара "отрезков" обозначает стандартное отклонение от среднего).

На этих графиках можно изобразить и так называемые усеченные средние (этот термин был впервые использован Тьюки в 1962 году), которые вычисляются после исключения заданного пользователем процента наблюдений с концов (хвостов) распределения.

Задание № 3.

1. **Используйте диаграммы размаха для сравнения значений переменных «Рост» и «Вес» в группах мужчин и женщин.**
2. **Считая, что 10% наблюдаемых значений переменной «Рост» представляют собой «засорения», определите усеченное среднее значение этой переменной, используя диаграмму размаха.**

Линейные графики

На линейных графиках отдельные точки данных соединяются линиями. Это простой способ визуального представления последовательности значений (например, цены на фондовом рынке за несколько дней торгов).

Круговые диаграммы

Одним из наиболее широко используемых типов графического представления данных являются круговые диаграммы, на которых показаны пропорции или сами значения переменных. Категоризованные графики этого типа состоят из нескольких круговых диаграмм, где данные разделены по группам с помощью одной или нескольких

группирующих переменных (например, «пол») или категоризованы согласно логическим условиям выбора подгрупп.

Задание № 4. Постройте круговую диаграмму, иллюстрирующую распределение обследуемых лиц по возрастам и категоризированную круговую диаграмму распределения испытуемых по весу с учетом пола.

Трехмерные (3М) графики

Трехмерные графики в координатах XYZ отображают взаимосвязи между тремя переменными. С помощью различных способов категоризации можно исследовать эти зависимости при различных условиях (т.е. в разных группах). Основная задача этих графиков - упростить сравнение взаимосвязей между тремя и более переменными для различных групп или категорий наблюдений.

Задание № 5. Постройте трехмерную диаграмму рассеяния переменных «Возраст», «Рост» и «Вес». Проанализируйте взаимосвязи между этими переменными.

Пиктографики

На пиктографиках каждое наблюдение представлено в виде многомерного символа, что позволяет использовать эти типы графического представления данных в качестве не очень простого, но мощного исследовательского инструмента. Главная идея такого метода анализа основана на человеческой способности "автоматически" фиксировать сложные связи между многими переменными, если они проявляются в последовательности элементов (в данном случае "пиктограмм"). Иногда понимание (или "чувство") того, что некоторые элементы "чем-то похожи" друг на друга, приходит раньше, чем наблюдатель (аналитик) может объяснить, какие именно переменные обусловливают это сходство. Конкретную природу проявившихся взаимосвязей между переменными позволяет выявить уже последующий анализ данных, основанный на изучении этого интуитивно обнаруженного сходства.

Основная идея пиктографиков заключается в представлении элементарных наблюдений как отдельных графических объектов, где значения переменных соответствуют определенным чертам или размерам объекта (обычно одно наблюдение = одному объекту). Это соответствие устанавливается таким образом, чтобы общий вид объекта менялся в зависимости от конфигурации значений.

Таким образом, объекты имеют определенный "внешний вид", который уникален для каждой конфигурации значений и может быть идентифицирован наблюдателем. Изучение таких пиктограмм помогает выявить как простые связи, так и сложные взаимодействия между переменными.

Целесообразно проводить анализ пиктографиков в пять этапов:

1. Выберите порядок анализируемых переменных. На этом этапе можно дать только один универсальный совет: прежде чем использовать какие-либо сложные методы, попробуйте случайную последовательность переменных.
2. Попробуйте обнаружить какие-либо закономерности, например, сходства между группами пиктограмм, выбросы или определенные связи между элементами (например, "если первые два луча звезды длинные, то как правило, с другой стороны есть один или два коротких луча"). На этом этапе лучше использовать пиктографики кругового типа.
3. При обнаружении закономерностей постарайтесь сформулировать их в терминах конкретных переменных.
4. Измените соответствие переменных и элементов пиктограмм (или переключитесь на один из последовательных пиктографиков), чтобы проверить обнаруженную

структуру взаимосвязей (например, попробуйте переместить ближе друг к другу элементы, между которыми обнаружена связь). В некоторых случаях в конце этого этапа целесообразно исключить из рассмотрения те переменные, которые не вносят явного вклада в обнаруженную структуру.

5. И наконец, используйте один из численных методов (таких как регрессионный анализ, нелинейное оценивание, дискриминантный или кластерный анализ), чтобы проверить и попытаться количественно оценить обнаруженные закономерности или хотя бы их часть.

Большинство пиктографиков можно отнести к одной из двух групп: круговые и последовательные. Круговые пиктографики (звезды, лучи, многоугольники) имеют вид "велосипедного колеса", на них значения переменных представлены расстояниями между центром пиктограммы ("втулкой") и их концами. Такие графики могут помочь в обнаружении связей между переменными, которые проявляются в общей структуре пиктограмм и зависят от конфигурации значений самих переменных.

Последовательные пиктографики (столбцы, профили, линии) представляют собой набор картинок с маленькими последовательными графиками (различных типов). Значения переменных представлены здесь расстояниями между основанием пиктограммы и последовательными точками (например, высотами показанных выше столбцов). Эти графики менее эффективны на начальной стадии разведочного анализа, поскольку пиктограммы очень похожи между собой. Однако, такое представление может быть весьма полезным для проверки уже сформулированной гипотезы.

Как правило, при построении пиктографиков значения переменных должны быть стандартизованы, чтобы их можно было сравнивать в пределах одной пиктограммы. Исключения составляют те случаи, когда на пиктограммах необходимо отобразить глобальные различия диапазонов выбранных переменных. Поскольку масштаб пиктограммы определяется наибольшим значением, то на пиктограмме могут отсутствовать те переменные, которые имеют значения другого порядка малости, например, на пиктограмме звезды некоторые лучи могут оказаться настолько короткими, что совсем не будут видны..

Пиктографики обычно используются: (1) для обнаружения структур или кластеров наблюдений и (2) для исследования сложных взаимосвязей между несколькими переменными.

Существуют различные типы пиктографиков:

1. "Лица Чернова". Для каждого наблюдения рисуется отдельное "лицо"; при этом относительные значения выбранных переменных соответствуют форме и размерам определенных его черт (например, длине носа, изгибу бровей, ширине лица).
2. Звезды. Это пиктографики кругового типа. Для каждого наблюдения рисуется пиктограмма в виде звезды; относительные значения выбранных переменных соответствуют относительным длинам лучей каждой звезды (по часовой стрелке, начиная с 12:00). Концы лучей соединены линиями.
3. Лучи. Эти пиктографики также относятся к круговому типу. Для каждого наблюдения строится одна пиктограмма. Каждый луч соответствует одной из выбранных переменных (по часовой стрелке, начиная с 12:00), и на нем отложено значение соответствующей переменной. Эти значения соединены линиями.
4. Многоугольники. Это пиктографикикругового типа. Для каждого наблюдения рисуется отдельный многоугольник; относительные значения выбранных переменных соответствуют расстояниям вершин от центра многоугольника (по часовой стрелке, начиная с 12:00).
5. Круговые диаграммы. Это пиктографики кругового типа. Для каждого наблюдения рисуется круговая диаграмма; относительные значения выбранных переменных соответствуют размерам сегментов диаграммы (по часовой стрелке, начиная с 12:00).

6. Столбцы. Это пиктографики последовательного типа. Для каждого наблюдения строится столбчатая диаграмма; относительные значения выбранных переменных соответствуют высотам последовательных столбцов.
7. Линии. Это пиктографики последовательного типа. Для каждого наблюдения строится линейный график; относительные значения выбранных переменных соответствуют расстояниям точек излома линии от основания графика.
8. Профили. Это пиктографики последовательного типа. Для каждого наблюдения строится зонный график; относительные значения выбранных переменных соответствуют расстояниям последовательных пиков сечения над линией основания.

Задание № 6. Изучите различные типы пиктографиков в программе «Statistica". Зарисуйте примеры пиктографиков нескольких типов (по вашему выбору).

Вопросы для проверки:
1. Что такое гистограмма? Построение и анализ гистограмм.
2. Что такое диаграмма рассеяния? Для чего используются такие диаграммы?
3. Как построить круговую диаграмму?
4. Для чего используются диаграммы размаха?
5. Что такое линейна диаграмма?
6. Что такое пиктографики? Типы пиктографиков. Назначение и анализ пиктографиков.

3. Первичная обработка данных, проверка статистических гипотез

Для решения задач используются возможности модулей Basic Statistics и Nonparametrics. При выборе метода для определения уровня значимости различий между группами следует учитывать:
1) характер распределения переменных;
2) количество наблюдений;
3) наличие «выпадающих» значений.

Параметрические критерии (модуль Basic Statistics) предпочтительнее, если характер распределения переменных не отличается существенно от нормального, объемы выборок не меньше 25-30 наблюдений и при отсутствии "выпадающих" значений. Если эти условия нарушаются, для проверки статистических гипотез следует использовать непараметрические критерии (модуль Nonparametrics).

Задача № 1. На двух делянках селекционной станции выращивали два новых сорта пшеницы. С каждой из этих делянок одновременно перед сбором урожая были взяты по 30 проб зерна, в каждой из которых находилось по 10 зерен пшеницы, взятых с одного колоса. Даны результаты взвешивания каждой пробы (табл. 1).

Задания:
1. Для каждого сорта пшеницы построить вариационный ряд, ряд распределения и гистограмму.
2. Для каждой выборки определить основные меры положения (среднее, моду, медиану, квартили).
3. Для каждой выборки определить основные меры рассеяния, размах, дисперсию, стандартное отклонение, коэффициент вариации.
4. По выборочным данным построить точечные и интервальные оценки параметров генеральной совокупности (среднее, стандартное отклонение, асимметрия, эксцесс).
5. Проверить гипотезу о соответствии характера распределения этих генеральных совокупностей предположению о нормальности.
6. Проверить гипотезу о том, что обе выборки взяты из одной генеральной совокупности (равенство средних и дисперсий).

Таблица 1. Результаты взвешивания проб зерна двух сортов пшеницы (г)

Номер пробы	Сорт А	Сорт Б	Номер пробы	Сорт А	Сорт Б
1	4,5	3,5	16	3,5	4,5
2	3,5	4,7	17	5,6	4,6
3	3,0	5,6	18	5,1	3,7
4	5,5	5,9	19	4,3	5,5
5	6,0	6,5	20	6,3	6,2
6	4,2	2,9	21	2,3	4,1
7	2,5	2,2	22	4,8	4,3
8	2,0	1,6	23	5,2	2,5
9	4,0	4,0	24	3,8	6,3
10	5,0	5,4	25	3,4	5,2
11	6,2	4,2	26	2,5	1,6
12	2,2	3	27	5,7	4,7
13	2,8	2,4	28	3,9	3,9
14	4,7	1,9	29	3,7	3,1
15	3,7	1,8	30	2,6	2,6

Задача № 2. Даны результаты эколого-аналитического контроля содержания сульфатов в сточных водах, сбрасываемых в Братское водохранилище в 1997 году (млн.т.) промышленными предприятиями трех групп: с восточной, юго-западной и северо-западной стороны (табл. 2). Определить, различаются ли эти группы предприятий по содержанию сульфатов в сточных водах.

Таблица 2.Содержание сульфата в сточных водах промышленных предприятий, сбрасываемых в Братское водохранилище с восточной (1), юго-западной (2) и северо-восточной (3) стороны

Группа	1	2	3
Номер			
1	55	33	15
2	72	31	11
3	66	81	30
4	69	60	37
5	92	38	15
6	50	27	36
7	65	20	45
8	85	75	52
9	70	87	61
10	75	28	88
11	83	85	13
12	80	42	35
13	64	32	40
14	93	59	28
15	65	78	50
16	78	22	32
17	43	55	17
18	90	75	26
19	91	72	30
20	48	63	28
21	50	40	12

22	76	82	20
23	96	15	21
24	88	58	10
25	25	66	23
26	100	29	8

Задача № 3. Чарльз Дарвин поставил опыт с целью проверки предположения о том, что способ получения семян (перекрестное опыление или самоопыление) влияет на рост и развитие растений, полученных из этих семян. Для этого он выращивал в 15 одинаковых горшочках 15 семян, полученных разными способами: в каждом горшочке находилось по 1 семени из каждой группы. Через определенное время для каждой пары фиксировалась разница по высоте между растениями из первой и второй группы (табл. 3). Подтверждают ли эти данные предположение Дарвина?

Табл.3.Разница в росте растений, полученных из семян разных групп

№ горшочка	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Разница (мм)	49	-67	8	16	6	23	28	41	14	29	56	24	75	60	-40

Задача № 4. Даны сведения об ежедневном обороте фирмы за 15 дней до и за 15 дней после публикации рекламы (табл.4). Оценить эффективность рекламной компании.

Таблица 4. Ежедневный оборот фирмы до и после публикации рекламы (тыс.руб.)

До	101	102	81	106	97	88	110	102	98	90	121	113	78	98	97
После	116	100	99	121	102	122	117	114	101	116	111	96	122	91	114

Задача № 5. В результате анализа газа из двух разных источников были получены следующие данные о содержании метана (мольный процент):
источник 1 - 64, 65, 75, 67, 64.5, 74, 75;
источник 2 - 69, 69, 61.5, 67.5, 64.
Определить, различается ли содержание метана в этих источниках.

Задача № 6. Охраняемые природные территории – это территории, в пределах которых обеспечивается их охрана от традиционного хозяйственного использования и поддержание их естественного состояния для сохранения экологического равновесия, а также в научных, учебно-просветительных и культурно-эстетических целях. Доля площади охраняемых природных территорий существенно отличается в разных странах (табл. 5). Используя данные таблицы, проведите сравнительный анализ ситуации с охраной природных территорий в странах Европы и Америки. Для этого:

- постройте гистограммы: для всех данных таблицы; для Европы; для Америки;
- определите характер распределения, наличие выбросов, шкалу измерений;
- определите основные описательные статистики для каждой группы стран и для всей выборки (среднее, размах, минимум, максимум, стандартная ошибка среднего, дисперсия);
- оцените достоверность межгрупповых различий по среднему, дисперсиям, характеру распределения.

Таблица 5. Доля площади охраняемых природных территорий в разных странах мира
(Вронский В.А., Прикладная экология, 1996. – С.341)

Страна	Доля (%)	Страна	Доля (%)

Европа		Северная и Южная Америка	
Австрия	15,08	Коста-Рика	11,1
Чехословакия	10,34	Панама	8,64
Норвегия	9,2	Венесуэла	8,4
Исландия	8,05	Эквадор	7,35
Великобритания	6,11	Боливия	3,96
Финляндия	2,85	Колумбия	3,47
Венгрия	2,82	Перу	3,34
Югославия	2,68	США	3,33
Швеция	2,61	Парагвай	3,04
Нидерланды	2,35	Канада	1,45
Италия	1,12	Бразилия	1,25
Швейцария	0,82	Аргентина	0,96
Болгария	0,75	Мексика	0,28
Греция	0,74	Уругвай	0,16
Франция	0,70	Никарагуа	0,12

4. Исследование взаимосвязей и линейный регрессионный анализ

Определение корреляции. Корреляция представляет собой меру зависимости переменных. Наиболее известна корреляция Пирсона. При вычислении корреляции Пирсона предполагается, что переменные измерены, как минимум, в интервальной шкале. Некоторые другие коэффициенты корреляции могут быть вычислены для менее информативных шкал. Коэффициенты корреляции изменяются в пределах от -1.00 до +1.00. Обратите внимание на крайние значения коэффициента корреляции. Значение -1.00 означает, что переменные имеют строгую отрицательную корреляцию. Значение +1.00 означает, что переменные имеют строгую положительную корреляцию. Отметим, что значение 0.00 означает отсутствие корреляции. Наиболее часто используемый коэффициент корреляции Пирсона r называется также линейной корреляцией, т.к. измеряет степень линейных связей между переменными.

Простая линейная корреляция (Пирсона r). Корреляция Пирсона предполагает, что две рассматриваемые переменные измерены, по крайней мере, в интервальной шкале. Корреляция высокая, если на графике зависимость "можно представить" прямой линией (с положительным или отрицательным углом наклона). Проведенная прямая называется прямой регрессии или прямой, построенной методом наименьших квадратов. Последний термин связан с тем, что сумма квадратов расстояний (вычисленных по оси Y) от наблюдаемых точек до прямой является минимальной. Заметим, что использование квадратов расстояний приводит к тому, что оценки параметров прямой сильно реагируют на выбросы.

Как интерпретировать значения корреляций. Коэффициент корреляции Пирсона (r) представляет собой меру линейной зависимости двух переменных. Если возвести его в квадрат, то полученное значение коэффициента детерминации R^2 представляет долю вариации, общую для двух переменных (иными словами, "степень" зависимости или связанности двух переменных). Чтобы оценить зависимость между переменными, нужно знать как "величину" корреляции, так и ее значимость.

Значимость корреляций. Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Значимость определенного коэффициента корреляции зависит от объема выборок. Критерий значимости основывается на предположении, что распределение остатков (т.е. отклонений наблюдений от регрессионной прямой) для зависимой переменной y является нормальным (с постоянной дисперсией для всех значений независимой переменной x). Исследования методом Монте-Карло показали, что нарушение этих условий не является абсолютно критичным, если размеры выборки не слишком малы, а отклонения от нормальности не очень большие.

Выбросы. По определению, выбросы являются нетипичными, резко выделяющимися наблюдениями. Так как при построении прямой регрессии используется сумма квадратов расстояний наблюдаемых точек до прямой, то выбросы могут существенно повлиять на наклон прямой и, следовательно, на значение коэффициента корреляции. Поэтому единичный выброс (значение которого возводится в квадрат) способен существенно изменить наклон прямой и, следовательно, значение корреляции. Если размер выборки относительно мал, то добавление или исключение некоторых данных способно оказать существенное влияние на прямую регресии (и коэффициент корреляции).

Во многих задачах, возникающих на практике, мы имеем измерения лишь в порядковой шкале. Для переменных, измеренных в порядковой шкале, имеются свои типы корреляции, позволяющие оценить зависимости.

R Спирмена. Статистику R Спирмена можно интерпретировать так же, как и корреляцию Пирсона (r Пирсона) в терминах объясненной доли дисперсии (имея, однако,

в виду, что статистика Спирмена вычислена по рангам). Предполагается, что переменные измерены как минимум в порядковой шкале.

Тау Кендалла. Статистика тау Кендалла эквивалентна R Спирмена при выполнении некоторых основных предположений. Также эквивалентны их мощности. Однако обычно значения R Спирмена и тау Кендалла различны, потому что они отличаются как своей внутренней логикой, так и способом вычисления.

Гамма-статистика. Если в данных имеется много совпадающих значений, статистика гамма предпочтительнее R Спирмена или тау Кендалла. С точки зрения основных предположений, статистика гамма эквивалентна статистике R Спирмена или тау Кендалла. Ее интерпретация и вычисления более похожи на статистику тау Кендалла, чем на статистику R Спирмена. Говоря кратко, гамма представляет собой также вероятность; точнее, разность между вероятностью того, что ранговый порядок двух переменных совпадает, минус вероятность того, что он не совпадает, деленную на единицу минус вероятность совпадений. Таким образом, статистика гамма в основном эквивалентна тау Кендалла, за исключением того, что совпадения явно учитываются в нормировке.

Задание 1. Даны результаты наблюдений за динамикой биомассы (грамм/гектар) мелких млекопитающих по данным многолетних отловов на опытной площадке в районе Томского нефтехимического комбината.

Виды/годы	1991	1992	1993	1994	1996	1997	1998	1999	2000	2001
C.rutilus	223	451,4	0	158,2	158,4	314	402,5	290,1	102,5	98,4
C.glareolus	537,2	380,8	1288	550,4	150	666	391	259	260,4	43,4
C.rufocanus	50,4	0	0	661	0	714,4	42,9	0	0	7,4
M.oeconomus	32,8	234,2	0	300,8	0	251,2	78,9	0	280,8	17,2
M.gregalis	0	52,3	0	0	0	0	0	0	17,5	0
A.agrarius	0	367,2	0	126,7	0	201	0	0	26,9	0
A.peninsulae	141,2	0	0	0	0	128,4	0	0	0	12,8
Insectivora	911,2	281,3	294,4	69,4	0	725,2	416,6	78,3	230,4	12,9
прочие виды	0	0	0	0	0	222,5	0	0	0	13,2

Для каждого вида, входящего в состав сообщества, а также для сообщества в целом, необходимо:

1) определить среднюю многолетнюю биомассу (среднее значение биомассы период наблюдений) и ошибку среднего;
2) определить объем выборки, минимальные и максимальные значения биомассы, размах, стандартное отклонение, асимметрию и эксцесс;
3) рассчитать парные корреляции между изменениями биомассы разных вилов мелких млекопитающих;
4) сравнить среднюю многолетнюю биомассу разных видов в сообществе и определить уровни значимости различий;
5) построить линейные графики динамики биомассы для всего сообщества и для 3 доминирующих видов в сообществе за период наблюдений.
6) Исследовать временные ряды с помощью линейного регрессионного анализа. Определить виды мелких млекопитающих, биомасса которых может использоваться для наиболее точной оценки суммарной биомассы сообщества.

Задание 2. «Экологическая бумага» – это бумага, произведенная на 100% из вторичного сырья (макулатуры). Увеличение доли такой бумаги в общем объеме ее производства и потребления может сократить вырубку лесов. Потребление бумаги и доля «экологической бумаги» существенно различаются в разных странах мира (табл.). Основываясь на приведенных в таблице данных, исследуйте взаимосвязи между объемом потребления

бумаги на душу населения, долей «экологической бумаги» и объемом потребления «экологической бумаги.

Таблица

Потребление бумаги и картона на душу населения (конец 80-х годов)
(Вронский В.А., Прикладная экология, 1996. – С.377)

Страна / регион	Годовой объем потребления (кг)	Доля «экологической бумаги» в общем потреблении (%)
США	317	29
Швеция	311	40
Канада	247	20
Япония	204	50
Норвегия	151	27
Бывший СССР	35	19
Латинская Америка	25	32
КНР	12	21
Африка	5	17
Индия	2	26

5. Дисперсионный анализ

Основной целью дисперсионного анализа является исследование значимости различия между средними в нескольких выборках. Если сравниваются средние в двух выборках, дисперсионный анализ даст такой же результат, как t-критерий.

Название «Дисперсионный анализ» связано с тем, что при исследовании статистической значимости различия между средними двух (или нескольких) групп, мы на самом деле сравниваем (т.е. анализируем) выборочные дисперсии. Фундаментальная концепция дисперсионного анализа предложена Фишером в 1920 году

Задание 1.

Откройте файл данных по сбросам промышленных предприятий в Братское водохранилище. Проведите дисперсионный анализ этих данных о определите, одинаковое ли количество сульфатов сбрасывается в водохранилище с тремя группами предприяти, которые с с восточной (1), юго-западной (2) и северо-восточной (3) стороны

Задание 2.

Создайте файл данных mous.sta, содержащий данные по биомассе мелких млекопитающих на трех опытных площадках Томского района. Используйте дисперсионный анализ для сравнения средних значений биомассы разных видов на этих площадках. Для видов, биомасса которых достоверно различается, постройте диаграммы сравнения типа «Ящики с усами». С чем могут быть связаны выявленные различия? Дайте экологическую интерпретацию результатов дисперсионного анализа.

Протопопово	1980	1981	1982	1983	1984	1985	1986	1987	1988
C.rutilus	2164,8	0	606,8	468,7	998,4	1896	873,1	1400	561
C.glareolus	91,2	0	6,9	14,6	178	119	108,4	85,3	41,8
C.rufocanus	0	0	36,4	159,6	46	0	34,4	0	72,8
M.oeconomus	1490,4	0	18,4	145,1	305,6	0	0	0	0
M.gregalis	0	0	0	0	0	0	0	0	0
A.agrarius	0	0	0	0	0	0	0	0	0
A.peninsulae	553,6	0	197,3	0	366	116,4	72,6	128,4	318
Insectivora	32,4	0	17,1	115,3	126,9	0	0	0	210,6
прочие виды	0	0	0	371,3	0	0	0	0	0

Манатка, годы:	1983	1984	1986	1987	1988	1989	1990	1991	1992
C.rutilus	844	1252,8	174,9	290,6	104,7	159	411,7	85,2	67,9
C.glareolus	1608,2	1024,4	134,6	207,3	56,5	120	245,3	71,7	64,8
C.rufocanus	0	0	42,8	172,5	11,7	69,2	270,4	99,8	0
M.oeconomus	0	110	35,7	104,1	9,5	16,4	44,8	131,5	0
M.gregalis	0	0	0	0	0	0	0	0	0
A.agrarius	0	0	8,9	5,7	0	5,8	132	54,9	0
A.peninsulae	0	0	11,5	4,4	5,2	0	3,6	0	0
Insectivora	0	0	0	0	0	0	0	0	0
прочие виды	0	0	0	0	0	0	0	0	0
Заварзино, годы:	1981	1982	1983	1984	1985	1986	1987	1988	
C.rutilus	843,8	557,2	434,7	1306,8	436,5	543,4	472	255,6	
C.glareolus	0	42	304,2	74	240	0	68,4	399,6	
C.rufocanus	136,8	0	128,4	444,8	0	78,8	76	0	
M.oeconomus	569,2	49,8	95,8	0	0	0	84,8	0	
M.gregalis	0	0	0	0	0	0	0	0	
A.agrarius	0	0	0	0	0	0	0	0	
A.peninsulae	218	93,6	123,4	69,6	35,6	300	37,8	166,4	
Insectivora	1010	68,4	98	231	105,2	28,4	44,8	86,4	
прочие виды	0	19	231	0	0	0	0	0	

6. Факторный анализ

Факторный анализ используют для выявления скрытых общих факторов, объясняющих связи между наблюдаемыми признаками объекта. Переход от анализа большого числа признаков к рассмотрению нескольких факторов или главных компонент позволяет не только лаконично описать структуру данных, но и вскрыть непосредственно не наблюдаемые закономерности и свойства экологических систем. Математической моделью, на которой основываются методы факторного анализа, является многомерное нормальное распределение.

Главными целями факторного анализа являются:

1) сокращение числа переменных (редукция данных);
2) определение структуры взаимосвязей между переменными, т.е. классификация переменных.

В **Statistica** факторный анализ проводится с использованием модуля **Factor Analysis**. В окне диалога этого модуля следует указать тип исходного файла:

- Correlation Matrix (корреляционная матрица);
- Raw Data (матрица объект-свойство).

Вычисление корреляционной матрицы, если она не задается сразу, первый этап факторного анализа.

После выбора переменных для анализа следует перейти в окно **Define Method of Factor Extraction,** где производится выбор метода выделения факторов, определяется максимальное количество факторов и минимальное собственное значение. Факторы, которые характеризуются меньшими собственными, значениями будут проигнорированы.

На практике для выбора оптимального числа факторов используют несколько процедур:

1) критерий Кайзера – расматриваются только те факторы, собственное значение которых превышает 1;
2) критерий каменистой осыпи – графический метод, в котором используется линейный график зависимости собственных значений от номера фактора.

В верхней части окна результатов факторного анализа дается следующая информация: число анализируемых переменных, метод анализа, число выделенных факторов, собственные значения. В нижней части окна находятся функциональные кнопки, позволяющие всесторонне посмотреть результаты анализа численно и графически.

Кнопка Factor Rotation позволяет выбрать метод и провести вращение факторов. Важно найти такое решение, которое возможно содержательно интерпретировать.

Кнопка Factor позволяет посмотреть значения факторных нагрузок в электронной таблице, а кнопка Plot of Loadings – на графике.

Задание. Исследуйте данные валеологического обследования студентов (файл valeo.sta) с помощью факторного анализа. С помощью критерия каменистой осыпи определите оптимальное количество факторов. Рассмотрите проекцию данных на плоскость двух первых факторов, охарактеризуйте выборку по признакам однородности, наличия выбросов, наличия подгрупп. Проведите вращение факторов, найдите факторную структуру, позволяющую дать содержательную интерпретацию. Опишите полученные факторы.

7. Канонический корреляционный анализ

Теоретическая часть. Во многих модулях STATISTICA можно вычислить парные коэффициенты корреляции для выражения зависимости между двумя переменными. Можно также вычислить матрицы парных коэффициентов корреляции. Например, коэффициент корреляции Пирсона (Pearson r) показывает степень линейной зависимости между двумя переменными, измеренными в интервальной шкале. Модуль «Непараметрическая статистика и распределения» («Nonparametrics and Distributions») предлагает различные статистики, основанные на рангах исследуемых переменных. Модуль «Множественная регрессия» («Multiple Regression») позволяет оценить зависимость между зависимой переменной (откликом) и множеством предикторных переменных.

Модуль **«Каноническая корреляция»** («Canonical Correlation») предназначен для анализа зависимостей между списками переменными, он позволяет исследовать зависимость между двумя множествами переменных.

Например, исследователь в сфере образования может оценить зависимость между навыками по трем учебным дисциплинам и оценками по пяти школьным предметам. Социолог может исследовать зависимость между прогнозами социальных изменений, печатаемыми в двух газетах, и реальными изменениями, оцененными с помощью четырех различных статистических признаков. Медик может изучить зависимость между различными неблагоприятными факторами и появлением определенной группы симптомов заболевания. Эколог может исследовать зависимость между содержанием в почве ряда химических элементов и составом растительного сообщества в экосистемах. Во всех этих случаях нас интересует зависимость между двумя множествами переменных.

Задание № 1. Придумать и записать три экологических задачи, в которых необходимо оценить зависимость между двумя или более множествами переменных.

Обыкновенная и множественная корреляции являются специальными случаями канонической корреляции, при которых один или оба набора содержат единственную переменную. Рассмотрим возможности модуля на примере.

Пример. **У 20 мужчин среднего возраста, посещающих клубы здоровья, были измерены три физиологических переменных: вес (фунты), обхват талиии (дюймы) и**

частота пульса (удары в минуту), и три переменных, характеризующих нагрузку: количество подтягиваний, приседаний и прыжков. Полученные данные приведены в таблице 1. Анализ канонических корреляций можно использовать, чтобы определить, связаны ли каким-либо образом физиологические переменные с переменными, характеризующими нагрузку.

Примечание: 1 дюйм=2,54 см; 1 фунт – 453,59 граммов.

Табл.1

№	Weight	Waist	Pulse	Chins	Situps	Jumps
1	191	36	50	5	162	60
2	189	37	52	2	110	60
3	193	38	58	12	101	101
4	162	35	62	12	105	37
5	189	35	46	13	155	58
6	182	36	56	4	101	42
7	211	38	56	8	101	38
8	167	34	60	6	125	40
9	176	31	74	15	200	40
10	154	33	56	17	251	250
11	169	34	50	17	120	38
12	166	33	52	13	210	115
13	154	34	64	14	215	105
14	247	46	50	1	50	50
15	193	36	46	6	70	31
16	202	37	62	12	210	120
17	176	37	54	4	60	25
18	157	32	52	11	230	80
19	156	33	54	15	225	73
20	138	33	68	2	110	43

Задание № 2. Используйте модуль «Basic Statistics» для определения описательных статистик этих переменных (среднее, стандартное отклонение, дисперсия, минимум, максимум, размах, асимметрия и эксцесс). Постройте в тетрадях таблицу, содержащую основные статистики для каждой переменной.

Описательные статистики показывают, что переменные Waist (обхват талии) и Jumps (прыжки) характеризуются высокими значениями эксцесса. Это говорит о возможных отклонениях характера распределения указанных признаков от нормального.

Задание № 3. Проверить характер распределения анализируемых признаков и сделать выводы о том, соответствует ли характер распределения нормальному закону.

Поскольку выборка мала и не все переменные распределены по нормальному закону, результаты канонического корреляционного анализа следует интерпретировать с осторожностью.

Задание № 4. Рассчитать парные корреляции между переменными. Построить матрицу корреляций, указать коэффициенты корреляций, значимые на уровни 0,05 и 0,001.

Корреляции между физиологическими и нагрузочными переменными умеренные, наибольшая между объемом талии (Waist) и количеством приседаний (Situps). В то же время, существуют большие внутримножественные корреляции между весом и объемом талии, между количеством подтягиваний и приседаний, между количеством приседаний и прыжков.

Подчеркните указанные корреляции в таблице.

Задание № 5. Используя модуль **«Каноническая корреляция»** («Canonical Correlation») программы STATISTICA проведите канонический корреляционный анализ между группами физиологических и нагрузочных переменных. В первом диалоговом окне модуля укажите все переменные, которые будут использованы в анализе, в качестве входного файла укажите «Исходные данные» (Raw data), выберите способ обработки пропущенных значений – удаление наблюдения (Casewise). Нажмите кнопку OK. Во втором диалоговом окне укажите состав каждой группы переменных (физиологические и нагрузочные) и нажмите OK.

Первая каноническая корреляция равна 0,7956. Это значение существенно больше, чем любая парная корреляция между физиологическими и нагрузочными переменными. Уровень значимости для нулевой гипотезы о том, что первая каноническая корреляция отлична от нуля, равен всего лишь 0,62, так что уверенных выводов сделать нельзя. Оставшиеся канонические корреляции не стоит и рассматривать, так как их уровни значимости всегда больше, чем у первой пары канонических переменных.

Убедитесь в этом, нажав на кнопку «Chi-square test of can. roots» (критерий хи-квадрат для канонических корреляций) и выпишите уровни значимости коэффициентов корреляции для первой .и последующих пар канонических переменных.

Вспомните, что такое коэффициент детерминации. Нажмите кнопку «Eigenvalues» и выпишите в тетрадь коэффициенты детерминации для всех пар канонических переменных. О чем говорят эти коэффициенты?

Теперь рассмотрим стандартизованные канонические коэффициенты изучаемых переменных, для этого следует нажать кнопку «Canonical Weights for Left and Right Set» (канонические веса для левого и правого наборов).

Первая каноническая переменная из левого набора (для физиологических признаков) равна:

КП1=1,58*Waist-0,78*Weight-0,06*Pulse

Перепишите в тетрадь формулу для расчета КП1Л. Используя таблицу стандартизованных коэффициентов канонических переменных, запишите формулу для первой канонической переменной из правого набора (для нагрузочных переменных).

КП1П=

Общая интерпретация первой пары канонических переменных состоит в том, что вес и количество прыжков действуют как переменные супрессоры и увеличивают корреляцию между обхватом талии и количеством приседаний. Эта каноническая корреляция может представлять практический интерес, но объем выборки недостаточен, чтобы сделать уверенные выводы.

Канонический анализ избыточности (кнопка «Factor Structures and Redundancies» в окне результатов) показывает, что ни одна из первой пары канонических переменных не является хорошим общим предикатором другого набора переменных. Доли объяненной дисперсии равны всего лишь 0,2854 и 0,2584. Вторая и третья канонические переменные не добавляют практически ничего, так как накопленные доли объясненной дисперсии равны 0,2969 и 0,2767 (кнопка «Results Summary»).

Изучите самостоятельно графические возможности модуля (кнопки и правой части панели окна «Вывод результатов анализа». Какой из графиков кажеться вам наиболее полезным и почему?

Задание № 6. Проверьте высказанные предположения путем анализа частных корреляции в модуле “Множественная регрессия”. Запишите в тетрадь парные и частные корреляции между всеми количеством и физиологическими переменными (вес и обхват талии). Сравните парные и частные коэффициенты и сформулируйте свои выводы о характере взаимосвязей между переменными.

Исследуйте коэффициенты множественных корреляций между:

а) нагрузочными переменными и набором физиологических переменных;

б) физиологическими переменными и набором нагрузочных переменных.

В каких случаях множественные коэффициенты корреляции превышают парные коэффициенты корреляции? Целесообразно ли использовать физиологические переменные для определения допустимой нагрузки?

Задание № 7. Повторите канонический корреляционный анализ физиологических и нагрузочных переменных, исключив из анализа переменную Pulse, вклад которой в формирование первой канонической переменной при первом исследовании оказался незначительным. Сравните результаты двух вариантов анализа. Сделайте выводы.

8. Многомерное шкалирование

Теоретические сведения.

Для повышения объективности при описании систем используются строгие статистические методы. Выявлять сходные системы можно с помощью методов ординации (**многомерное шкалирование**). Этот метод представляет собой математическую обработку данных, в результате которой точки, представляющие изучаемые объекты или системы, располагаются на графике (обычно это плоскость или трехмерное пространство) таким образом, что расстояние между точками пропорционально степени различия между объектами. На следующем этапе анализа исследователь пытается дать содержательную интерпретацию выделенных осей, то есть определить, изменению каких факторов соответствует сдвиг положения точек (объектов) вдоль каждой из осей.

В методе многомерного шкалирования (МНШ) качестве исходных данных можно использовать произвольный тип матрицы сходства объектов. То есть, на входе всех алгоритмов МНШ используется матрица, элемент которой на пересечении ее i-й строки и j-го столбца, содержит сведения о попарном сходстве анализируемых объектов (объекта [i] и объекта [j]). На выходе алгоритма МНШ получаются числовые значения координат, которые приписываются каждому объекту в некоторой новой системе координат (во "вспомогательных шкалах", связанных с латентными переменными, откуда и название МНШ), причем размерность нового пространства признаков существенно меньше размерности исходного (за это собственно и идет борьба).

Логику МНШ можно проиллюстрировать на следующем простом примере. Предположим, что имеется матрица попарных расстояний (т.е. сходства некоторых признаков) между крупными американскими городами. Анализируя матрицу, стремятся расположить точки с координатами городов в двумерном пространстве (на плоскости), максимально сохранив реальные расстояния между ними. Полученное размещение точек на плоскости впоследствии можно использовать в качестве приближенной географической карты США.

В общем случае метод МНШ позволяет таким образом расположить "объекты" (города в нашем примере) в пространстве некоторой небольшой размерности (в данном случае она равна двум), чтобы достаточно адекватно воспроизвести наблюдаемые расстояния между ними. В результате можно "измерить" эти расстояния в терминах найденных латентных переменных. Так, в нашем примере можно объяснить расстояния в терминах пары географических координат Север/Юг и Восток/Запад.

МНШ - это не просто определенная процедура, а скорее способ наиболее эффективного размещения объектов, приближенно сохраняющий наблюдаемые между ними расстояния. Другими словами, МНШ размещает объекты в пространстве заданной размерности и проверяет, насколько точно полученная конфигурация сохраняет расстояния между объектами. Говоря более техническим языком, МНШ использует алгоритм минимизации некоторой функции, оценивающей качество получаемых вариантов отображения.

Стресс. Мерой, наиболее часто используемой для оценки качества подгонки модели (отображения), измеряемого по степени воспроизведения исходной матрицы

сходств, является так называемый **стресс.** Обычно используется одна из несколько похожих мер сходства. Тем не менее, большинство из них сводится к вычислению суммы квадратов отклонений наблюдаемых расстояний (либо их некоторого монотонного преобразования) от воспроизведенных расстояний. Таким образом, чем меньше значение стресса, тем лучше матрица исходных расстояний согласуется с матрицей результирующих расстояний.

Диаграмма Шепарда. Можно построить для текущей конфигурации точек график зависимости воспроизведенных расстояния от исходных расстояний. Такая диаграмма рассеяния называется диаграммой Шепарда. По оси ординат OY показываются воспроизведенные расстояния (сходства), а по оси OX откладываются истинные сходства (расстояния) между объектами (отсюда обычно получается отрицательный наклон). На этом график также строится график ступенчатой функции. Ее линия представляет так называемые величины «D с крышечкой», то есть, результат монотонного преобразования исходных данных. Если бы все воспроизведенные результирующие расстояния легли на эту ступенчатую линию, то ранги наблюдаемых расстояний (сходств) был бы в точности воспроизведен полученным решением (пространственной моделью). Отклонения от этой линии показывают на ухудшение качества согласия (т.е. качества подгонки модели).

Вообще говоря, чем больше размерность пространства, используемого для воспроизведения расстояний, тем лучше согласие воспроизведенной матрицы с исходной (меньше значение стресса). Если взять размерность пространства равной числу переменных, то возможно абсолютно точное воспроизведение исходной матрицы расстояний. Однако нашей целью является упрощение решаемой задачи, с тем, чтобы объяснить матрицу сходства (расстояний) в терминах лишь нескольких важнейших факторов (латентных переменных или вспомогательных шкал). Возвращаясь к нашему примеру с расстояниями между городами, если получена двумерная карта, намного проще представить себе расположение городов и планировать передвижение между ними, чем если бы имелась только матрица попарных расстояний.

Приложения. "Красота" метода МНШ в том, что вы можете анализировать произвольный тип матрицы расстояний или сходства. Эти сходства могут представлять собой оценки экспертов относительно сходства данных объектов, результаты измерения расстояний в некоторой метрике, процент согласия между судьями по поводу принимаемого решения, количество раз, когда субъект затрудняется различить стимулы и мн.др. Например, методы МНШ весьма популярны в психологическом исследовании восприятия личности. В этом исследовании анализируются сходства между определенными чертами характера с целью выявления основополагающими личностных качеств. Также они популярны в маркетинговых исследованиях, где их используют для выявления числа и сущности латентных переменных (факторов), например, с целью с целью изучения отношения людей к товарам известных торговых марок. В экологии методы ординации используются, например, при изучении влияния факторов среды на структуру сообществ. При этом анализу подвергается матрица сходства сообществ, а полученные оси интерпретируются как экологические факторы, влияющие на структру сообщества.

Суммируя вышесказанное, можно сказать, что методы МНШ применимы к широкому классу исследовательских задач.

Вопросы (запишите ответы в тетрадь):

1) В чем суть МНШ?
2) Какие исходные данные могут быть использованы для МНШ?
3) Что получается в результате МНШ?
4) Как оценить успешность полученной модели?
5) Что показывает диаграмма Шепарда?
6) Приведите пример использования МНШ в экологических исследованиях (Бигон и др. Экология, т.2, С.129.).

Задание.
Даны географические расстояния в километрах между точками, которые необходимо посетить в ходе экспедиции по местам геологических обнажений с остатками ископаемых организмов. Создайте файл данных Statistica, содержащий эту матрицу данных.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18
T1	0,0	,4	,4	1,2	1,5	4,7	1,9	2,0	2,6	2,8	2,8	3,3	3,6	3,9	5,4	5,6	5,9	6,0
T2	,4	0,0	,3	,8	1,1	4,6	1,6	1,7	2,3	2,4	2,5	3,0	3,2	3,5	5,1	5,3	5,5	5,7
T3	,4	,3	0,0	,9	1,3	4,3	1,9	2,0	2,4	2,5	2,6	3,0	3,2	3,5	5,1	5,3	5,5	5,7
T4	1,2	,8	,9	0,0	,3	4,1	1,4	1,3	1,5	1,6	1,7	2,2	2,4	2,7	4,3	4,5	4,7	4,9
T5	1,5	1,1	1,3	,3	0,0	3,9	1,5	1,4	1,4	1,4	1,4	1,9	2,1	2,4	4,0	4,2	4,4	4,6
T6	4,7	4,6	4,3	4,1	3,9	0,0	5,3	5,1	4,5	4,3	4,0	3,0	3,2	3,4	3,8	3,8	4,0	4,2
T7	1,9	1,6	1,9	1,4	1,5	5,3	0,0	,3	1,3	1,6	1,8	2,8	2,9	3,1	4,6	4,8	5,0	5,2
T8	2,0	1,7	2,0	1,3	1,4	5,1	,3	0,0	1,0	1,3	1,6	2,6	2,7	2,9	4,3	4,6	4,7	4,9
T9	2,6	2,3	2,4	1,5	1,4	4,5	1,3	1,0	0,0	,3	,6	1,5	1,7	1,8	3,3	3,5	3,7	3,8
T10	2,8	2,4	2,5	1,6	1,4	4,3	1,6	1,3	,3	0,0	,3	1,4	1,4	1,6	3,0	3,2	3,4	3,6
T11	2,8	2,5	2,6	1,7	1,4	4,0	1,8	1,6	,6	,3	0,0	1,1	1,1	1,3	2,8	3,0	3,3	3,4
T12	3,3	3,0	3,0	2,2	1,9	3,0	2,8	2,6	1,5	1,4	1,1	0,0	,3	,7	2,1	2,3	2,5	2,8
T13	3,6	3,2	3,2	2,4	2,1	3,2	2,9	2,7	1,7	1,4	1,1	,3	0,0	,3	1,8	2,0	2,3	2,5
T14	3,9	3,5	3,5	2,7	2,4	3,4	3,1	2,9	1,8	1,6	1,3	,7	,3	0,0	1,6	1,8	2,0	2,2
T15	5,4	5,1	5,1	4,3	4,0	3,8	4,6	4,3	3,3	3,0	2,8	2,1	1,8	1,6	0,0	,3	,4	,6
T16	5,6	5,3	5,3	4,5	4,2	3,8	4,8	4,6	3,5	3,2	3,0	2,3	2,0	1,8	,3	0,0	,2	,4
T17	5,9	5,5	5,5	4,7	4,4	4,0	5,0	4,7	3,7	3,4	3,3	2,5	2,3	2,0	,4	,2	0,0	,2
T18	6,0	5,7	5,7	4,9	4,6	4,2	5,2	4,9	3,8	3,6	3,4	2,8	2,5	2,2	,6	,4	,2	0,0

Добавьте к этой матрице 4 пустые строки снизу и назовите эти строки:
Means
Std.Dev.
No.Cases
Matrix

1. В модуле «Многомерное шкалирование» постройте «Карту расстояний» проекцию точек на плоскость. Считая, что полученные шкалы можно расматривать как координатные, спланируйте оптимальный маршрут экскурсии.
2. Чему равен показатель стресса?
3. Теперь постройте проекцию точек в трехмерное пространство. Изменился ли показатель стресса и если да, то как? О чем говорит это изменение? Какая модель (двумерная или трехмерная) в большей степени соответствует исходным данным? Рассмотрите расположение точек в трехмерном пространстве. Попытайтесь дать содержательную интерпретацию осям. Спланируйте оптимальный маршрут, считая, что одна из осей отражает высоту точки над уровнем моря, а две другие – долготу и широту места.

Составьте таблицу, в которую занесите показатели стресса при проекции точек на прямую, на плоскость и в пространство размерностью от 3 до 5. Постройте график «Каменистой осыпи». Какое количество осей следует использовать в модели?
Запишите в тетради ход анализа и ваши выводы.

9. Кластерный и дискриминантный анализ в Statistica

Кластерный анализ объединяет различные процедуры, используемые для классификации наблюдений или признаков. Под кластером обычно понимают группу объектов, которая характеризуется определенными плотностью, дисперсией, формой и размером, а также отделяемую от других кластеров.

Перед началом классификации данные обычно стандартизуются, путем вычитания из каждого значения переменной ее среднего и деления на корень квадратный из дисперсии. В результате получаются стандартизованные переменные, имеющие нулевое среднее и единичную дисперсию.

Для кластеризации используются агломеративные и дивизивные методы. Агломеративные методы основаны на последовательном объединении наиболее близких объектов в один кластер, которое можно показать на графике в виде дендрограммы.

Расстояние между объектами в кластерном анализе определяется с помощью различных мер сходства. Если объект описывается с помощью числовых признаков, в качестве меры сходства обычно используют евклидову метрику.

Для объединения объектов в кластеры могут быть использованы различные алгоритмы. Метод одиночной связи основан на включении в кластер на каждом шаге объекта, обладающего максимальной степенью сходства с одним из членов кластера. Недостатком этого метода является образование «цепочек»: длинных продолговатых кластеров. Метод полных связей, в котором мера сходства между вновь включаемым объектом и всеми членами кластера не может быть меньше порогового значения, устраняет указанный недостаток. При использовании метода Уорда на каждом шаге кластеризации проводится объединение, дающеее минимальное приращение внутригрупповой суммы квадратов отклонений, при этом образуются кластеры примерно равных размеров и гиперсферической формы.

Целью дискриминантного анализа является классификация объекта в одну из нескольких заданных групп. При этом среди известных характеристик объекта производится отбор признаков, на основе которых конструируются дискриминатные функции, позволяющие провести оптимальную классификацию. Линейный дискриминатный анализ предполагает, что характеристики объекта имеют нормальное распределение и что переменные в разных классах имеют равные дисперсии и ковариации, отличаясь только по средним значениям. Однако, умеренные отклонения от этих предположений допустимы, и наиболее важным критерием правильности построенного классификатора является его практическая работоспособность. Качество работы классификатора можно проверить путем переклассификации наблюдений, принадлежность которых к одной из групп заранее известна. Такая переклассификация проводится путем скользящего экзамена или методом разбиения выборки на «обучающую» и «контрольную» группы. Оценкой качества классификатора служит процент ошибочной классификации.

Технологию выполнения кластерного и дискриминантного анализов в пакете STATISTICA рассмотрим на примерах.

Пример 1. Кластерный анализ автомобилей разных марок

Выберите в переключателе модулей название модуля **Cluster analysis** и нажмите кнопку Switch.

Создайте файл car.sta и внесите в таблицу характеристики автомобилей разных марок. Все характеристики автомобилей в предлагаемой таблице уже стандартизованы. Задача состоит в том, чтобы выделить в предлагаемой совокупности группы машин, сходных друг с другом не по одному параметру, а в целом по всем расматриваемым характеристикам.

Марки машин	Цена ($)	Технические характеристики			Расход горючего (мили/галлон)
		ACCELER	BRAKING	HANDLING	
Acura	-,521	,477	-,007	,382	2,079
Audi	,866	,208	,319	-,091	-,677
BMW	,496	-,802	,192	-,091	-,154
Buick	-,614	1,689	,933	-,210	-,154
Corvette	1,235	-1,811	-,494	,973	-,677
Chrysler	-,614	,073	,427	-,210	-,154
Dodge	-,706	-,196	,481	,145	-,154
Eagle	-,614	1,218	-4,199	-,210	-,677
Ford	-,706	-1,542	,987	,145	-1,724
Honda	-,429	,410	-,007	,027	,369
Isuzu	-,798	,410	-,061	-4,230	1,067
Mazda	,126	,679	-,133	,500	-1,724
Mercedes	1,051	,006	,120	-,091	-,154
Mitsub.	-,614	-1,003	,084	,382	,718
Nissan	-,429	,073	-,007	,263	,997
Olds	-,614	-,734	,409	,382	2,114
Pontiac	-,614	,679	,536	,145	,195
Porsche	3,454	-2,215	-,296	,618	-1,026
Saab	,588	,679	,246	,263	,021
Toyota	-,059	1,218	,228	,736	-,851
VW	-,706	-,128	,102	,382	,195
Volvo	,219	,612	,138	-,210	,369

В строке меню рабочего окна модуля выберите пункт **Analysis**, в выпадающем меню выберите строку **Startup panel.** На экране появиться стартовая панель модуля **Кластерный анализ.** В списке методов высветите **k-means** (к-средних) и нажмите кнопку OK. На экране появится диалоговое окно метода **k-means.**

Воспользуйтесь для этого кнопкой **Variables,** чтобы выбрать переменные для анализа.

В поле **Cluster** выбирете **Cases**, если хотите провести классификацию случаев. Если вас интересует класификация переменных, выберите **Variables.**

В поле **Number of Cluster** следует определить количество групп, которые мы хотим получить в результате анализа. Запишите в это поле число 3.

Сделав необходимые установки, нажмите кнопку OK. Появится окно результатов анализа. В верхней части окна записана следующая информация: количество переменных и наблюдений, метод кластеризации, количество кластеров, а также количество проведенных итераций. В нижней части окна имеются следующие кнопки:

- **Analysis of Variance;** позволяет посмотреть результаты дисперсионного анализа исходных переменных с группирующей переменной «Кластер», то есть определить, имеются ли достоверные различия между полученными группами по исходным переменным;
- **Cluster Means and Euclidean Distances;** выводит таблицы средних по исходным переменным для каждого кластера и евклидовых растояний между кластерами;
- **Graph of Means;** позволяет посмотреть средние значения для каждого кластера на линейном графике;
- **Descriptive Statistics for each Cluster**; открывает электронную таблицу с опитсательными статистиками для каждого кластера;

- **Save Classifications and Distances;** позволяет сохранить результаты кластерного анализа для дальнейшего исследования полученных кластеров.

Задание.

1) Изучите результаты кластерного анализа автомобилей, опишите характерные признаки полученных групп. Для каких практических целей может быть использована подобная классификация?
2) Откройте файл Valeo и проведите кластерный анализ данных из этого файла в два этапа: сначала используйте метод иерархической кластеризации **(Joining clustering)**, изучите дерево кластеризации и выберите оптимальное количество групп; затем используйте анализ к-средних для более подробного исследования выбранной классификации.
3)

Пример 2. Дискриминантный анализ. Классификация цветов ириса

Задача состоит в том, чтобы по результатам измерения длины и ширины чашелистиков и лепестков цветков ириса отнести ирис к одному из трех типов: SETOSA, VERSICOL, VIRGINIC.

Данные для этого примера имеются в файле Irisdat. sta. В файле содержатся результаты измерений 150 цветков ириса, по 50 каждого типа.

Шаг 1. Из переключателя модулей STATISTICA откройте стартовую панель модуля **Discriminant function analysis** (**Дискриминантный анализ**).

Шаг 2. Нажмите кнопку **Open Data (Открыть данные**) и откройте файл данных *Irisdat. sta* из каталога *Examples*.

Шаг 3. Нажмите кнопку **Variables (Переменные**) и выберите переменные для анализа.

В качестве **Группирующей переменной – Grouping variable** выберите переменную **IRISTYPE – ТИПИРИСА.**

В качестве **Независимых переменных – Independent variables** выберите переменные **SEPALLEN – ДЛИНА ЧАШЕЛИСТИКА, SEPALWD – ШИРИНА ЧАШЕЛИСТИКА, PETALLEN – ДЛИНА ПЕСТИКА, PETALWD – ШИРИНА ПЕСТИКА.**

Шаг 4. Нажмите кнопку **OK** и откройте диалоговое окно **Model Definition (Определ ение модели).**

Нажмите кнопку **OK** и запустите вычислительную процедуру, реализующую пошаговый метод включения.

Шаг 5. Всесторонне просмотрите итоги в диалоговом окне **Результаты дискриминантного анализа.**

Информационная часть окна сообщает, что использован

- Stepwise analysis – Пошаговый анализ, Step 4 (Final step) – Шаг 4 (Заключительный шаг);
- Number of variables in the model –Число переменных в модели 4;
- Last variable entered – Последняя включенная переменная : ДЛЧАШЕЛИ, соответствующее значение статистики F – критерия F(2,144) = -4.72, уровень значимости p<0.0103;
- Wilks lambda –Значение лямбды Уилкса : 0.0234;
- Approx. F(2,292) = 199.1454 – Приближенное значение F – статистики, связанной с лямбдой Уилкса;
- P – уровень значимости F – критерия для значения 199.1454.

Значения статистики лямбда Уилкса лежат в интервале (0,1). Значения статистики Уилкса, лежащие около 0, свидетельствуют о хорошей дискриминации. Значения статистики Уилкса, лежащие около 1, свидетельствуют о плохой дискриминации. Иными словами, это можно выразить следующим образом : если значения лямбда Уилкса близки

к 0, то мощность дискриминации (мощность = 1 – вероятность ошибки) близка к 1, если лямбда Уилкса близка к 1, то мощность близка к 0.

Нажмите кнопку **Variables in the model (**Переменные, включенные в модель**).**

Просмотрите разделение групп на графике. Для этого инициируйте кнопку **Canonical analysis & graphs (**Канонический анализ и графики**).** В появившемся диалоговом окне **Canonical Analysis (**Канонический анализ**)** нажмите кнопку **Scatterplot of canonical scores** (Диаграмма рассеяния канонических значений). На экране появится график.

Задание.

1) **Используйте дискриминантный анализ для того, чтобы проверить качество классификации автомобилей разных марок, полученной при изучении кластерного анализа. Определите достоверность разграничения групп и проценты ошибочной классификации для каждой группы. Запишите дискриминантные функции.**
2) **Используйте дискриминантный анализ для того, чтобы проверить качество классификации людей по валеологическим показателям.**

10. Вероятностный калькулятор в Statistica

Задача № 1. Масса зерен пшеницы сорта А нормально распределена, средняя масса зерна равна 0,44 грамма, стандартное отклонение – 0,11 граммов. Определить вероятность того, что случайное взятое зерно будет иметь массу
а) превышающую 0,5 грамма;
б) в пределах от 0,35 до 0,55 граммов;
в) что каждое из 10 случайно взятых зерен будет иметь массу не более 0,6.

Задача № 2. Пусть бросается правильная симметричная монета. Какова вероятность того, что при бросании 100 раз равно в 50 случаях выпадет “орел”.
(Подсказка: эта задача на биномиальное распределение. Следует создать в файле Statistica переменную, содержащую вероятности выпадения «орла» 1, 2, 3, 4 ….. 100 раз при 100 испытаниях. Используйте функцию binom(x;p;n), где x – количество успехов в независимых испытаниях; p – вероятность успеха в одном испытании; n – количество испытаний).

Задача № 3. Семь коров пасутся на лугу площадью 10000 м^2, равномерно покрытому травой. Коровы перемещаются независимо друг от друга и каждая из них с равной вероятностью может оказаться в любой точке луга. Найти вероятности того, что
а) ровно 2 коровы окажутся в данном квадрате со стороной 50 м^2;
б) только одна корова окажется в данном квадрате;
в) все семь коров окажутся в данном квадрате.

Задача № 4 (задача азартного игрока). Некогда один англичанин по имени Пейпас послал Ньютону письмо, в котором спрашивал, на что лучше ставить
- на выпадение хотя бы одной «6» при бросании кости 6 раз;
- на выпадение хотя бы двух «6» при бросании кости 12 раз;
- на выпадение хотя бы трех «6» при бросании кости 18 раз;
- на выпадение хотя бы четырех «6» при бросании кости 24 раза.

Определить вероятности и дать обоснованный ответ.

Задания для самостоятельной работы студентов

1. Подготовить сообщение с презентацией на тему «Статистический анализ экологических проблем». Для подготовки реферата использовать данные и описания результатов статистического анализа, опубликованные в научных экологических журналах, монографиях, учебниках. Приветствуется использование собственных данных, полученных в ходе учебной или производственной практики, при подготовке курсовых работ, при работе в группах проектного обучения (ГПО). В сообщении должны быть отражены все основные этапы статистического исследования: постановка задачи, формирование выборки, способы получения изучаемых характеристик, методы статистической обработки, результаты обработки их интерпретация и выводы. При использовании литературных данных обязательны ссылки на источники.
2. Подготовить реферат по одной из предлагаемых тем:
   - История использования статистических методов в биологии и в экологии.
   - Измерительные шкалы и их свойства. Измерения в экологии.
   - Возможности использования методов статистического контроля качества для управления экологическими процессами.
   - Планирование экспериментов в экологии.

- Параметрические и непараметрические критерии сравнения выборок и их использование в экологических задачах.
- Корреляционный анализ и примеры его использования в экологических исследованиях.
- Дисперсионный анализ: применение в экологических исследованиях.
- Регрессионные модели: применение в экологии.
- Многомерное шкалирование экологических данных.
- Методы снижения размерности (факторный анализ, метод главных компонент) и их использование в экологии.
- Анализ временных рядов в экологических исследованиях.
- Группировка экологических данных с использованием кластерного анализа.
- Дискриминантный анализ и возможности его использования в экологии.

Тестовые задания

1. **В чем причина случайной изменчивости показателей состояния окружающей среды?**
   а) состояние окружающей среды формируется под влиянием большого числа воздействий, эффект каждого из которых по отдельности незначителен;
   б) состояние окружающей среды формируется под воздействием жизнедеятельности организмов;
   в) состояние окружающей среды в данный момент зависит от ее состояние в предшествующие моменты времени.
2. **Какая прикладная наука дает возможность разграничить закономерные и случайные изменения показателей состояния окружающей среды?**
   а) кибернетика; б) информатика; в) статистика.
3. **Что такое генеральная совокупность?**
   а) множество всех возможных наблюдений, которые в принципе могли бы быть сделаны при заданных условиях;
   б) совокупность доступных для изучения объектов определенного типа;
   в) множество объектов, которые не вошли в изучаемую выборку.
4. **Какой принцип следует соблюдать при отборе данных для получения репрезентативной выборки?**
   а) принцип одновременного отбора;
   б) принцип последовательного отбора; в) принцип случайного отбора.
5. **Какой тип измерительных шкал позволяет сравнивать и упорядочивать объекты по изучаемому признаку, но не дает информации о степени различия между ними?**
   а) номинальные б) ранговые в) количественные.
6. **Какой тип измерительных шкал позволяет судить только о принадлежности объекта к одной из нескольких групп?**
   а) номинальные б) ранговые в) количественные.
7. **Какой тип графика представляет собой ряд прямоугольных столбиков, основание которых соответствует диапазону изменения значений признака, а высота – количеству объектов, характеризующихся значениями признака в этом диапазоне?**
   а) диаграмма рассеяния б) линейная диаграмма в) гистограмма.
8. **Какой график целесообразно использовать для визуальной оценки наличия и формы связи между двумя признаками, измеренными в количественной или ранговой шкале?**
   а) диаграмма рассеяния б) линейная диаграмма в) гистограмма.

9. **Какое из перечисленных свойств характерно для нормального распределения?**
   а) асимметричность;б) полимодальность;
   в) равенство моды и медианы.
10. **Какой из перечисленных критериев можно использовать для проверки предположения о нормальном распределении изучаемого признака в генеральной совокупности?**
   а) критерий Стьюдента; б) критерий Колмогорова-Смирнова;
   в) критерий Манна-Уитни.
11. **Какое свойство изучаемого показателя следует проверять для корректного использования параметрических методов статистического анализа?**
   а) полимодальность;б) значение эксцесса;
   в) характер распределения.
12. **Какой из перечисленных критериев является параметрическим?**
   а) критерий Вилкоксона; б) критерий Стьюдента;
   в) критерий Манна-Уитни.
13. **Какой статистический метод используется для исследования линейных связей между признаками?**
   а) дисперсионный анализ; б) кластерный анализ;
   в) корреляционный анализ.
14. **Какой коэффициент корреляции следует использовать для изучения связей между признаками, измеренными в порядковой шкале?**
   а) коэффициент корреляции Пирсона;
   б) коэффициент ранговой корреляции Спирмена;
   в) коэффициент канонической корреляции.
15. **Какой статистический метод позволяет построить линейную модель, отражающую зависимость выходной переменной от одного или нескольких факторов (например, зависимость скорости фотосинтеза от освещенности и температуры воздуха)?**
   а) регрессионный анализ; б) дискриминатнтный анализ;
   в) корреляционный анализ.
16. **Для чего используется корреляционный анализ?**
   а) для сравнения параметров двух или нескольких выборок;
   б) для оценки степени взаимосвязи между переменными;
   в) для изучения характера распределения переменных в выборке.
17. **Для чего используется анализ остатков в регрессионном анализе?**
   а) для оценки качества регрессионной модели;
   б) для получения более точного прогноза на основе построенной модели;
   в) для уточнения параметров модели.
18. **Для чего используется метод кластерного анализа?**
   а) для классификации объектов в заданные группы;
   б) для выделения среди множества объектов однородных групп;
   в) для выявления латентных факторов.
19. **Какой статистический метод позволяет классифицировать наблюдения в заданные группы?**
   а) дисперсионный анализ; б) кластерный анализ;
   в) дискриминантный анализ.
20. **Какой статистический метод позволяет путем анализа наблюдаемых признаков обнаружить и исследовать внутренние (латентные) свойства системы?**
   а) факторный анализ; б) дискриминатный анализ;
   в) многомерное шкалирование.

Вопросы к зачету

1. Прикладная статистика как наука.
2. История развития прикладной статистики.
3. Пакеты и методы анализа данных.
4. Статистические данные (типы матриц, измерительные шкалы).
5. Выборочный метод анализа данных. Репрезентативность выборки.
6. Измерительные шкалы.
7. Алгоритм статистического анализа данных.
8. Ошибки в данных, их природа и устранение.
9. Графические методы анализа данных.
10. Построение и анализ гистограмм.
11. Нормальное распределение, его свойства и значение.
12. Теоремы Чебышева (закон больших чисел и центральная предельная теорема).
13. Описательные статистики.
14. Статистические методы контроля качества.
15. Параметрические и непараметрические методы прикладной статистики.
16. Проверка статистических гипотез.
17. Методы статистического исследования взаимосвязей.
18. Корреляционный анализ.
19. Анализ таблиц сопряженности.
20. Дисперсионный анализ.
21. Регрессионный анализ.
22. Оценка качества регрессионной модели.
23. Факторный анализ, метод главных компонент.
24. Использование метода главных компонент в задачах классификации.
25. Метод главных компонент в анализе динамических моделей.
26. Эвристические методы снижения размерности.
27. Кластерный анализ: меры сходства.
28. Кластерный анализ: иерархические методы.
29. Кластерный анализ: неиерархические методы.
30. Дискриминантный анализ.

Литература для самоподготовки

1. Несмелова Н.Н., Незнамова Е.Г., Смирнов Г.В. Многомерные методы исследования биологических систем. - Томск: ТУСУР, 2007. - 178 с. (ауд. – 37 экз.)
2. Боровиков В.П. Statistica. Искусство анализа данных на компьютере.–С.П-б.: Питер, 2001. – 656 с.
3. Дюк В. Обработка данных на ПК в примерах.– С.П-б.: Питер, 2001.– 656 с.
4. Зайцев В.М., Лифляндский В.Г., Маринкин В.И. Прикладная медицинская статистика: Учебное пособие. – С.-Пб.: «Фолиант», 2006. – 432 с.
5. Халафян А.А. Statistica 6. Статистический анализ данных. Учебник. – М.: «Бином-Пресс», 2007. – 512 с.