Прикладная статистика. Введение. Определение.

Под методами математической статистики принято понимать лишь те методы статистической обработки исходных данных, разработка и использование которых апеллируют к вероятностной природе этих данных.

Широкий класс методов статистической переработки исходной информации, а именно вся совокупность тех методов, которые не опираются на вероятностную природу обрабатываемых данных (представителями методов такого типа являются, например, разнообразные методы кластерного анализа, многомерного шкалирования, теории измерений и др.), остается за рамками научной дисциплины «математическая статистика».

Прикладная статистика - самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации сбора, стандартной записи, систематизации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов.

Математическая статистика является по отношению к прикладной статистике разработчиком и поставщиком части используемого в последней математического аппарата, не участвует в следующих задачах прикладной статистики: «прилаживание» и доработка необходимого математического инструментария в соответствии с конкретной спецификой решаемой реальной задачи; разработка логико-алгебраических методов статистической обработки данных; преобразование разнообразных форм получаемой информации к стандартному виду исходных статистических данных, их удобное представление и подготовка к обработке; организация автоматизированной обработки данных на ЭВМ, создание необходимого программного обеспечения.

Прикладная статистика. Введение. Этапы статистической обработки данных.

Этап 1: исходный (предварительный) анализ исследуемой реальной системы.

В результате этого анализа определяются: а) основные цели исследования на неформализованном, содержательном уровне; б) совокупность единиц, представляющая предмет статистического исследования; в) перечень $(x^{(1)}, x^{(2)}, ..., x^{(p)})$ отобранных из представленного специалистами набора показателей, характеризующих состояние (поведение) каждого из обследуемых объектов, который предполагается использовать в данном исследовании; г) степень формализации соответствующих записей при сборе данных; д) общее время и трудозатраты, отведенные на планируемые работы, и коррелированные с ними временная протяженность и объем необходимого статистического обследования; е) моменты, требующие предварительной проверки перед составлением детального плана исследования (например, не всегда заранее ясна возможность идентификации единиц наблюдения, в медицинских исследованиях не всегда может быть получено согласие больного следовать определенным рекомендациям медперсонала и т. п.); ж) формализованная постановка задачи, по возможности включающая вероятностную модель изучаемого явления, и природа статистических выводов, к которым должен прийти исследователь в результате переработки массива исходных данных; з) формы, используемые для сбора первичной информации и для введения ее в ЭВМ.

Прикладная статистика. Введение. Этапы статистической обработки данных.

Этап 2: составление детального плана сбора исходной статистической информации.

При составлении этого плана необходимо, по возможности, учитывать полную схему дальнейшего статистического анализа. Априорное представление о том, как и для чего данные будут анализироваться, может оказать существенное влияние на их сбор. При планировании особого внимания заслуживают случаи, когда: а) используется аппарат общей теории выборочных обследований т. е. определяется, какой должна быть выборка — случайной, пропорциональной, расслоенной и др.; б) производится расчет «разрешающей силы» исследования заданного объема и продолжительности; в) хотя бы для части входных переменных эксперимент носит активный характер: переменные допускают фиксацию в каждом конкретном наблюдении на определенном уровне, и выбор плана обследования осуществляется с привлечением методов планирования (регрессионных) экспериментов.

Также этот этап называют **этапом «организационно-методической подготовки».**

Прикладная статистика. Введение. Этапы статистической обработки данных.

Этап 3: сбор исходных статистических данных и их введение в ЭВМ.

Независимо от того, производится ли исследователем выбор метода и плана статистического обследования или он уже располагал результатами пассивного эксперимента, к моменту определения основного инструментария статистического исследования исследователь в общем случае располагает в качестве массива исходных статистических данных временной последовательностью матриц наблюдений вида

$$X(t)=\begin{pmatrix} x_1^{(1)}(t) & x_2^{(1)}(t) & \dots & x_n^{(1)}(t) \\ x_1^{(2)}(t) & x_2^{(2)}(t) & \dots & x_n^{(2)}(t) \\ \dots & \dots & \dots & \dots \\ x_1^{(p)}(t) & x_2^{(p)}(t) & \dots & x_n^{(p)}(t) \end{pmatrix}, \ (t=t_1,\dots,t_N).$$

Прикладная статистика. Введение. Этапы статистической обработки данных.

Этап 4: первичная статистическая обработка данных.

В ходе первичной статистической обработки данных решаются следующие задачи: а) отображение переменных, описанных текстом, в номинальную (с предписанным числом градаций) или ординальную (порядковую) шкалу; б) статистическое описание исходных совокупностей с определением пределов варьирования переменных; в) анализ резко выделяющихся наблюдений; г) восстановление пропущенных наблюдений; д) проверка статистической независимости последовательности наблюдений, составляющих массив исходных данных; е) унификация типов переменных, когда с помощью различных приемов добиваются унифицированной записи всех переменных; ж) экспериментальный анализ закона распределения исследуемой генеральной совокупности и параметризация сведений о природе изучаемых распределений

Этот этап также называют **процессом составления сводки и группировки**.

Прикладная статистика. Введение. Этапы статистической обработки данных.

Этап 5: составление детального плана вычислительного анализа материала.

Этап начинается с составления справки по собранному материалу и результатам предварительного анализа. Определяются основные группы, для которых будет проводиться дальнейший анализ. Пополняется и уточняется тезаурус содержательных понятий. Четко описывается блок-схема анализа с указанием привлекаемых методов. Формулируется оптимизационный критерий, на основании которого выбирается один из альтернативных методов (или одно из альтернативных семейств методов) основной статистической обработки исходных данных.

Этап 6: вычислительная реализация основной части статистической обработки данных.

Основная забота исследователя на этом этапе — эффективное управление вы-вычислительным процессом путем формулировки задачи обработки и описания данных. Учитываются размерность задачи, алгоритмическая сложность вычислительного процесса, возможности используемого программного обеспечения и, наконец, особенности данных (степень обусловленности используемых при реализации линейных процедур матриц, надежность априорных оценок параметров и т. п.).

Этап 7: подведение итогов исследования.

Этап начинается с построения формального статистического отчета о проведенном исследовании. При интерпретации результатов применения статистических процедур (оценка параметров, проверка гипотез, отображения в пространство меньшей размерности, классификация и т. п.) учитывается как место этих процедур в блок-схеме анализа, так и соотношение объемов используемых выборок, размерности пространства наблюдений, числа и значений параметров.

Прикладная статистика. Введение. Исследование зависимостей между анализируемыми показателями.

Исследование характера и структуры взаимосвязей, существующих между анализируемыми показателями, характеризующими состояние или поведение статистически обследованных объектов (процессов), является сущностью и главной целью **многомерного статистического анализа**.

Многомерный статистический анализ включает методы регрессионного, корреляционного, дисперсионного и ковариационного анализа, методы экстремального планирования регрессионных экспериментов, методы анализа временных рядов, некоторые методы и модели зависимостей специального (например, марковского) типа.

Общая схема исследования:

Вектор статистически регистрируемых на исследуемой реальной системе показателей X подразделяется на два подвектора, один из которых интерпретируется как вектор характеристик условий функционирования (или состояния) исследуемой системы, а второй, интерпретируется как вектор результирующих показателей, характеризующих поведение или эффективность функционирования (качество) исследуемой системы.

$$X^{(1)} = \begin{pmatrix} x^{(1)} \\ \dots \\ x^{(m)} \end{pmatrix} \qquad X^{(2)} = \begin{pmatrix} x^{(m+1)} \\ \dots \\ x^{(p)} \end{pmatrix}$$

Прикладная статистика. Введение. Исследование зависимостей между анализируемыми показателями.

Проблема состоит в конструктивном объяснении поведения результирующих показателей за счет изменения факторов-аргументов, т. е. в определении такой векторной функции из класса допустимых решений *F*, которая давала бы наилучшую, в определенном смысле, аппроксимацию поведения вектора $X^{(2)}$ на множестве точек-наблюдений

$$f(X^{(1)}) = \begin{pmatrix} f_1(x^{(1)}) \\ \dots \\ f_m(x^{(m)}) \end{pmatrix}$$

Прикладная статистика. Введение. Классификация объектов и признаков.

Говоря о **классификации совокупности объектов**, мы будем подразумевать, что каждый из них задан соответствующим столбцом матрицы статистических данных, либо что геометрическая структура их попарных расстояний (связей) задана матрицей.

Аналогично интерпретируется исходная информация в задаче **классификации совокупности признаков**, с той лишь разницей, что каждый из признаков задается соответствующей строкой матрицы.

Общая постановка проблемы классификации объектов заключается в том, чтобы всю анализируемую совокупность объектов $O = \{O_i\}, i = 1,..., n$, статистически представленную в виде матрицы данных, разбить на сравнительно небольшое число однородных, в определенном смысле, групп или классов. Для формализации этой проблемы удобно интерпретировать анализируемые объекты в качестве точек в соответствующем факторном пространстве.

Проблема классификации состоит в разбиении анализируемой совокупности точек-наблюдений на сравнительно небольшое число — заранее известное или нет — сгустков (кластеров, скоплений, таксонов, образов), которые находятся на некотором расстоянии друг от друга (в смысле метрики, введенной в соответствующем пространстве *X*), но сами не разбиваются на столь же удаленные классы.

Для решения таких задач используются методы **дискриминантного анализа**, **расщепления смесей распределений**, **кластерного анализа**.

Прикладная статистика. Введение.
Снижение размерности факторного пространства.

Снижение размерности в исходных данных обусловлено следующими причинами:
во-первых, **дублирование информации**, доставляемой сильно взаимосвязанными признаками;
во-вторых, **неинформативность признаков**, мало меняющихся при переходе от одного объекта к другому (малая вариабельность признаков);
в-третьих, **возможность агрегирования**, т. е. простого или взвешенного суммирования, по некоторым признакам.

Для решения данной задачи используются **метод главных компонент, факторный анализ, экстремальная группировке параметров**.