Гришин, С. И. Компьютерный анализ данных. Модели, алгоритмы, программы [Текст] : монография / С. И. Гришин, В. Х. Кириллов, А. К. Ширшков ; под ред. В. Х. Кириллова. - Одесса : ВМВ, 2014. – 304 с. : табл., рис. - Библиогр.: с. 302-303. - ISBN 978-966-413-468-9.

Монография представляет собой обзор теоретических и компьютерных методов по математической и прикладной статистике для технологических, технических и экономических вузов. Цель монографии — теоретические и справочные сведения по решению основных задач статистического анализа данных, используя компьютерные среды MatLab и SPSS. Для каждой из рассмотренных в книге задач даны: краткие теоретические сведения, описание математического метода решения, описание порядка и этапов выполнения задания в компьютерных средах.

Рассмотрены основы работы в операционных средах MatLab и SPSS и проводится обзор основных задач прикладной статистики и теории вероятностей по обработке данных. Подробно обсуждаются следующие основные задачи обработки экспериментальных данных: интерполяция и аппроксимация данных, основы проверки статистических гипотез, корреляционный и регрессионный анализ, дисперсный анализ, анализ временных рядов. На конкретных примерах проводится подробное описание порядка решения упомянутых задач в математическом пакете MatLab и статистическом пакете SPSS.

Монография адресована широкому кругу читателей: студентам, бакалаврам и магистрантам инженерных и экономических специальностей, преподавателям естественных дисциплин, научным работникам, пользователям компьютеров, применяющим статистические методы в практической работе.

Введение

Все мы хорошо знакомы с закономерными явлениями и закономерными изменениями, они составляют главный объект научных исследований. Например, по астрономическим наблюдениям Тихо Браге Кеплер установил законы движения планет, а по этим законам, путём обобщения экспериментальных данных проведенных Галилеем для свободного падения тел, Ньютон установил знаменитый закон Всемирного тяготения. Знание этих

законов позволяет точно вычислять траектории спутников и комет, корректировать траекторию полёта снаряда и ракеты или предсказать когда и в какой местности произойдёт солнечное затмение.

Но отнюдь не во всех ситуациях интересующих нас результат полностью и жёстко определяется влияющими на него известными нам факторами. Например, мы не можем указать, сколько часов будет светить электрическая лампочка, или как долго будет служить жёсткий диск или DVD дисковод. Невозможно предвидеть число посетителей магазина и количество товаров, которое они купят, количество автомобилей, проходящих по улице за определённый промежуток времени, - и т. д. Явления, в которых результат испытания полностью определяется влияющими на него известными факторами, называются *детерминированными* или *закономерными,* а те, в которых это не выполняется — *недетерминированными* или *стохастическими.*

В большинстве явлений особенно в процессе измерений при проведении эксперимента присутствуют оба вида изменчивости — и закономерная и случайная (стохастическая), и для нахождения закономерностей нам приходится «отсеивать» возмущающие случайные факторы. Например, при одних и тех же начальных условиях стрельбы имеет место определённое рассеивание при падении снарядов, это рассеивание обусловлено многочисленными второстепенными факторами, влияющими на полёт снаряда в атмосфере (ветер, изменяющаяся плотность воздуха, турбулентность и др.). Другой пример, автомобиль движется равномерно по прямолинейному участку дороги, практически для того, чтобы обеспечить такое движение, водитель должен постоянно корректировать это движение с помощью рулевой колонки и педалью акселератора. Неизбежность корректировки связана с действием на автомобиль многочисленных возмущающих факторов (неровность полотна дороги, боковой ветер, турбулентность обтекаемого воздуха и др.).

Если требуется обработать большой и сложный набор данных, состоящий из небольших порций информации, статистика поможет классифицировать и проанализировать ситуацию, предоставив полезный обзор и резюме основных характеристик этих данных. Если данных недостаточно, статистика поможет собрать их и обеспечить получение ответов на заданные вопросы.

Вероятность — средство работы с риском и неопределенностью. Вероятность показывает возможность наступления в будущем каждого из потенциальных событий, рассчитанную на основе информации о некоторой ситуации. В то время как статистика позволяет переходить от наблюдений к обобщениям относительно рассматриваемой ситуации, вероятность имеет обратную направленность — исходя из характеристики можно выяснить, какие данные, скорее всего, удастся получить, и возможность получения каждого из вариантов таких данных.

С развитием и ростом сложности аппаратного и программного обеспечения все больше и больше внимания уделяется развитию методов прикладной статистики, резко ослабляющих допущения, на которых строятся используемые математические модели. Разработкой в этом секторе рынка программного обеспечения заняты как известные лидеры, так и новые развивающиеся компании. Инструменты прикладной статистики могут быть представлены либо как самостоятельное приложение, либо как дополнения к основному продукту.

Определенным импульсом для привлечения новых пользователей прикладной статистики явилось включение служб анализа данных в функциональность СУБД их поставщиками: это Microsoft (Microsoft SQL Server), Oracle, IBM (IBM Intelligent Miner for Data). Исследования показывают, что основным аргументом при выборе этих систем является то, что они поставляются совместно с сервером баз данных. Интеграция с системами оперативной обработки информации позволяет собрать оригинальные входные данные для анализа, а не «правильные» отчеты, подготовленные в отделах и филиалах корпорации.

Не только исследователи, но и менеджеры получили реальную возможность использовать в целях верного принятия решений очень сложный математический аппарат анализа данных, и знание этого аппарата является для них весьма существенным. По этой же причине люди, непосредственно использующие анализ своей деятельности (менеджеры и аналитики), должны понимать опасность выбора неадекватных методов анализа, так как некорректные результаты могут привести к принятию неверных решений.

Материалы по рассматриваемой теме входят в содержание обязательных дисциплин, преподаваемых студентам, обучающимся по направлениям областей знаний «Информатика и вычислительная техника» и «Системные науки и кибернетика». Однако наименования дисциплин и рекомендуемый объем материала, приведенные в стандартах высшего образования Украины для различных направлений подготовки, отличаются.

Наиболее полно материалы прикладной статистики представлены в направлении «Компьютерные науки», где они нашли отражение в дисциплинах «Теория вероятностей, вероятностные процессы и математическая статистика» и «Интеллектуальный анализ данных».

Достаточно подробно материалы по рассматриваемой теме освещены в направлении «Прикладная математика» - в дисциплинах «Теория вероятностей», «Математическая статистика» и «Анализ данных», а также в направлении «Информатика» - в дисциплинах «Теория вероятностей и математическая статистика» и «Распределенные информационно-аналитические системы». Из отсутствующих материалов в этих направлениях подготовки хотелось бы отметить популярную в последние годы технологию Data Mining.

Оглавление