**Лекция 10.**

**Статистическое исследование зависимостей:** *Стохастичесепя зависимость, регрессии и корреляции, регрессионный анализ.*

          Теория регрессионного анализа описана и расчетные формулы даны в специальной литературе [1-6] силлабуса. В этой теории разработаны методы точечного и интервального оценивания параметров, задающих функциональную зависимость, а также непараметрические методы оценивания этой зависимости, методы проверки различных гипотез, связанных с регрессионными зависимостями. Выбор планов эксперимента, т.е. точек *xi*, в которых будут проводиться эксперименты по наблюдению *yi* – предмет теории планирования эксперимента.

          Дисперсионный анализ применяют для изучения влияния качественных признаков на количественную переменную. Например, пусть имеются *k* выборок результатов измерений количественного показателя качества единиц продукции, выпущенных на *k* станках, т.е. набор чисел (*x*1(*j*), *x*2(*j*), … , *x*n(*j*)), где*j* – номер станка,*j* = 1, 2, …, *k*, а *n* – объем выборки. В распространенной постановке дисперсионного анализа предполагают, что результаты измерений независимы и в каждой выборке имеют нормальное распределение *N*(*m*(*j*), σ2) с одной и той же дисперсией. Хорошо разработаны и непараметрические постановки.

          Проверка однородности качества продукции, т.е. отсутствия влияния номера станка на качество продукции, сводится к проверке гипотезы *H*0: *m*(1) = *m*(2) = … = *m*(*k*).

В дисперсионном анализе разработаны методы проверки подобных гипотез. Теория дисперсионного анализа и расчетные формулы рассмотрены в специальной литературе, предложенной в силлабусе.

Гипотезу *Н*0 проверяют против альтернативной гипотезы *Н*1, согласно которой хотя бы одно из указанных равенств не выполнено. Проверка этой гипотезы основана на следующем «разложении дисперсий», указанном Р.А.Фишером:

 

где *s*2 – выборочная дисперсия в объединенной выборке, т.е.



Далее, *s*2(*j*) – выборочная дисперсия в *j*-ой группе,



Таким образом, первое слагаемое в правой части формулы (7) отражает внутригрупповую дисперсию. Наконец,  - межгрупповая дисперсия,



          Область прикладной статистики, связанную с разложениями дисперсии типа формулы (7), называют дисперсионным анализом. В качестве примера задачи дисперсионного анализа рассмотрим проверку приведенной выше гипотезы *Н*0 в предположении, что результаты измерений независимы и в каждой выборке имеют нормальное распределение *N*(*m*(*j*), σ2) с одной и той же дисперсией. При справедливости *Н*0 первое слагаемое в правой части формулы (7), деленное на σ2, имеет распределение хи-квадрат с *k*(*n*-1) степенями свободы, а второе слагаемое, деленное на σ2, также имеет распределение хи-квадрат, но с (*k*-1) степенями свободы, причем первое и второе слагаемые независимы как случайные величины. Поэтому случайная величина



имеет распределение Фишера с (*k*-1) степенями свободы числителя и *k*(*n*-1) степенями свободы знаменателя. Гипотеза *Н*0 принимается, если *F* < *F*1-α, и отвергается в противном случае, где *F*1-α – квантиль порядка 1-α распределения Фишера с указанными числами степеней свободы.  Такой выбор критической области определяется тем, что при *Н*1 величина *F* безгранично увеличивается при росте объема выборок *n*. Значения *F*1-α берут из соответствующих таблиц.

Разработаны непараметрические методы решения классических задач дисперсионного анализа, в частности, проверки гипотезы *Н*0.

Следующий тип задач многомерного статистического анализа – задачи классификации. Они делятся на три принципиально различных вида – дискриминантный анализ, кластер-анализ, задачи группировки.

Задача дискриминантного анализа состоит в нахождении правила отнесения наблюдаемого объекта к одному из ранее описанных классов. При этом объекты описывают в математической модели с помощью векторов, координаты которых – результаты наблюдения ряда признаков у каждого объекта. Классы описывают либо непосредственно в математических терминах, либо с помощью обучающих выборок. Обучающая выборка – это выборка, для каждого элемента которой указано, к какому классу он относится.

Рассмотрим пример применения дискриминантного анализа для принятия решений в технической диагностике. Пусть по результатам измерения ряда параметров продукции необходимо установить наличие или отсутствие дефектов. В этом случае для элементов обучающей выборки указаны дефекты, обнаруженные в ходе дополнительного исследования, например, проведенного после определенного периода эксплуатации. Дискриминантный анализ позволяет сократить объем контроля, а также предсказать будущее поведение продукции. Дискриминантный анализ сходен с регрессионным – первый позволяет предсказывать значение качественного признака, а второй – количественного. В статистике объектов нечисловой природы разработана математическая схема, частными случаями которой являются регрессионный и дискриминантный анализы.

Кластерный анализ применяют, когда по статистическим данным необходимо разделить элементы выборки на группы. Причем два элемента группы из одной и той же группы должны быть «близкими» по совокупности значений измеренных у них признаков, а два элемента из разных групп должны быть «далекими» в том же смысле. В отличие от дискриминантного анализа в кластер-анализе классы не заданы, а формируются в процессе обработки статистических данных. Например, кластер-анализ может быть применен для разбиения совокупности марок стали (или марок холодильников) на группы сходных между собой.

Другой вид кластер-анализа – разбиение признаков на группы близких между собой. Показателем близости признаков может служить выборочный коэффициент корреляции. Цель кластер-анализа признаков может состоять в уменьшении числа контролируемых параметров, что позволяет существенно сократить затраты на контроль. Для этого из группы тесно связанных между собой признаков (у которых коэффициент корреляции близок к 1 – своему максимальному значению) измеряют значение одного, а значения остальных рассчитывают с помощью регрессионного анализа.

Задачи группировки решают тогда, когда классы заранее не заданы и не обязаны быть «далекими» друг от друга. Примером является группировка студентов по учебным группам. В технике решением задачи группировки часто является параметрический ряд – возможные типоразмеры группируются согласно элементам параметрического ряда. В литературе, нормативно-технических и инструктивно-методических документах по прикладной статистике также иногда используется группировка результатов наблюдений (например, при построении гистограмм).

Задачи классификации решают не только в многомерном статистическом анализе, но и тогда, когда результатами наблюдений являются числа, функции или объекты нечисловой природы. Так, многие алгоритмы кластер-анализа используют только расстояния между объектами. Поэтому их можно применять и для классификации объектов нечисловой природы, лишь бы были заданы расстояния между ними. Простейшая задача классификации такова: даны две независимые выборки, требуется определить, представляют они два класса или один. В одномерной статистике эта задача сводится к проверке гипотезы однородности.

Третий раздел многомерного статистического анализа – задачи снижения размерности (сжатия информации). Цель их решения состоит в определении набора производных показателей, полученных преобразованием исходных признаков, такого, что число производных показателей значительно меньше числа исходных признаков, но они содержат возможно большую часть информации, имеющейся в исходных статистических данных. Задачи снижения размерности решают с помощью методов многомерного шкалирования, главных компонент, факторного анализа и др. Например, в простейшей модели многомерного шкалирования исходные данные – попарные расстояния  между *k* объектами, а цель расчетов состоит в представлении объектов точками на плоскости. Это дает возможность в буквальном смысле слова увидеть, как объекты соотносятся между собой. Для достижения этой цели необходимо каждому объекту поставить в соответствие точку на плоскости так, чтобы попарные расстояния *sij* между точками, соответствующими объектам с номерами *i* и *j*, возможно точнее воспроизводили расстояния ρ*ij* между этими объектами. Согласно основной идее метода наименьших квадратов находят точки на плоскости так, чтобы величина



достигала своего наименьшего значения. Есть и многие другие постановки задач снижения размерности и визуализации данных.