Практическое занятие ПЗ-9

**Тема:Методы статистической обработки данных**

При проведении опытных исследований различают пассивный или активный эксперименты. Методология пассивного эксперимента предполагает проведение большой серии опытных исследований с поочередным варьированием значений входных переменных и анализом результатов измерений выходной переменной у. К пассивному эксперименту принято относить также сбор и обработку опытных данных в процессе эксплуатации электроэнергетических установок. Обработка результатов пассивного эксперимента проводится методами регрессионного и корреляционного анализа, а также выбора эмпирической модели, то есть установления закона распределения исследуемой случайной величины.

Законы распределения случайных величин практически могут быть получены двумя путями: путем аналитического исследования и путем обработки экспериментальных данных. В первом случае законы распределения находят, проводя анализ природы какого-либо явления или процесса и определенных математических операций. Во втором случае производится сбор необходимых экспериментальных данных. Данные могут быть получены либо в результате специально поставленного эксперимента, либо в результате наблюдений. В процессе проведения эксперимента накапливается большой объем разнообразных данных, которые на первый взгляд могут выглядеть весьма хаотично. От их правильной обработки во многом зависит интерпретация результатов эксперимента.

Статистическая обработка накопленной информации позволяет получить аналитическую зависимость искомого закона распределения вероятностей и вычислить параметры распределения.

Экспериментальное определение закона распределения случайной величины играет особую роль. Как бы глубоко и тщательно не было проведено аналитическое исследование, в результате которого получен закон распределения, окончательное заключение может быть сделано только базируясь на эксперименте.

Специальная дисциплина «Математическая статистика» занимается разработкой методов регистрации, описания и анализа статистических экспериментальных данных, получаемых в результате наблюдения массовых случайных явлений.

Она решает следующие основные задачи:

* построение статистической функции распределения случайной величины;
* нахождение статистических числовых характеристик;
* определение статистических параметров закона распределения, если тип закона известен;
* определение типа закона распределения;
* статистическая проверка гипотез.

Генеральная совокупность и случайная выборка. На практике исследователь обычно располагает лишь ограниченным числом значений случайной величины, представляющим собой некоторую выборку из генеральной совокупности. В силу закона случая какие-то величины, определенные по малому числу наблюдений, в общем случае могут не совпадать с теми же величинами, вычисленными по большому числу наблюдений, выполненных в тех же условиях. Поэтому в математической статистике вводят абстрактное понятие генеральной совокупности, состоящей из всех допустимых значений случайной величины и выборочной совокупности, представляющей собой ограниченное число значений, полученных в результате опытов. Иными словами, выборка представляет собой отбор значений случайной величины из генеральной совокупности.

В соответствии с этим различают выборочные характеристики случайной величины, найденные по ограниченному числу наблюдений и зависящие от этого числа, и соответствующие им характеристики генеральной совокупности. При этом выборочные характеристики рассматриваются как оценки соответствующих характеристик генеральной совокупности. Выборки могут быть повторными и бесповторными. В первом случае отобранный объект возвращается в генеральную совокупность, во второмслучае – не возвращается. Обычно применяются бесповторные выборки. Выборка называется репрезентативной (представительной), если она
дает достаточное представление об особенностях генеральной совокупности.

Случайная функция распределения. Исходными данными для статистического исследования случайной величины чаще всего является набор наблюдений ее реализаций (реализация случайной величины – значение, принимаемое ею в опыте). Этот набор называется простой статистической совокупностью. Обычно простая статистическая совокупность оформляется в виде таблицы с одним входом, в первом столбце которой стоит номер опыта, а во втором – наблюденное значение случайной величины.

*Методами статистической обработки результатов экспери­мента* называются математические приемы, формулы, способы количественных расчетов, с помощью которых показатели, по­лучаемые в ходе эксперимента, можно обобщать, приводить в си­стему, выявляя скрытые в них закономерности.

Речь идет о та­ких закономерностях статистического характера, которые су­ществуют между изучаемыми в эксперименте переменными ве­личинами.

*Данные* – это основные элементы, подлежащие классифицированию или разбитые на категории с целью обработки.

Некоторые из методов математико-статистического анализа позволяют вычислять так называемые элементарные математические статистики, характеризующие выборочное распреде­ление данных, например:

- выборочное среднее,

- выборочная диспер­сия,

- мода,

- медиана и ряд других.

Иные методы математической статистики позволяют судить о динамике изменения отдельных статис­тик выборки, например,:

- дисперсионный анализ,

- регрессионный ана­лиз.

С помощью третьей группы методов выборочных данных, можно достоверно судить о статистических связях, существующих между переменными величинами, кото­рые исследуют в данном эксперименте:

- кор­реляционного анализа;

- факторного анализа;

- методов сравнения.

Все методы математико-статистического анализа условно де­лятся на первичные и вторичные.

Первичными называют мето­ды, с помощью которых можно получить показатели, непосред­ственно отражающие результаты производимых в эксперимен­те измерений.

Вторичными называются методы статистической обработки, с помощью которых на базе первичных данных выявляют скры­тые в них статистические закономерности.

К первичным методам статистической обработки относят, на­пример:

- определение выборочной средней величины;

- выбороч­ной дисперсии;

- выборочной моды;

- выборочной медианы.

В чис­ло вторичных методов обычно включают:

- корреляционный ана­лиз;

- регрессионный анализ;

- методы сравнения первичных ста­тистик у двух или нескольких выборок.

Рассмотрим методы вычисления элементарных математичес­ких статистик, начав с выборочного среднего.

*Среднее арифметическое значение* ***–*** это отношение суммы всех значений данных к числу слагаемых.

Среднее значение как статистический показатель представляет собой среднюю оценку изучаемого в эксперименте психологического качества.

Эта оценка характеризует степень его развития в целом у той группы испытуемых, которая была под­вергнута психодиагностическому обследованию. Сравнивая не­посредственно средние значения двух или нескольких выборок, мы можем судить об относительной степени развития у людей, составляющих эти выборки, оцениваемого качества.

Выборочное среднее определяется при помощи следующей формулы:

 (1)

где хср —выборочная средняя величина или среднее арифметичес­кое значение по выборке;

п — количество испытуемых в выбор­ке или частных психодиагностических показателей, на основе ко­торых вычисляется средняя величина;

xk — частные значения по­казателей у отдельных испытуемых. Всего таких показателей п, поэтому индекс k данной переменной принимает значения от 1 до п;

∑ — принятый в математике знак суммирования величин тех переменных, которые находятся справа от этого знака.

*Дисперсия* – это мера разброса данных относительно среднего значения30.

Чем больше дисперсия, тем больше отклонения или разброс данных. Ее определяют для того, чтобы можно было отличать друг от друга величины, име­ющие одинаковую среднюю, но разный разброс.

Дисперсия определяется по следую­щей формуле:

 (2)

где — выборочная дисперсия, или просто дисперсия;

— выражение, означающее, что для всех xk от перво­го до последнего в данной выборке необходимо вычислить раз­ности между частными и средними значениями, возвести эти раз­ности в квадрат и просуммировать;

п — количество испытуемых в выборке или первичных зна­чений, по которым вычисляется дисперсия.

*Медианой* называется значение изучаемого признака, кото­рое делит выборку, упорядоченную по величине данного призна­ка, пополам.

Знание медианы полезно для того, чтобы установить, явля­ется ли распределение частных значений изученного признака симметричным и приближающимся к так называемому нормаль­ному распределению. Средняя и медиана для нормального рас­пределения обычно совпадают или очень мало отличаются друг от друга.

Если выборочное распределение признаков нормаль­но, то к нему можно применять методы вторичных статистичес­ких расчетов, основанные на нормальном распределении данных. В противном случае этого делать нельзя, так как в расчеты могут вкрасться серьезные ошибки.

*Мода* еще одна элементар­ная математическая статистика и характеристика распределе­ния опытных данных. Модой называют количественное зна­чение исследуемого признака, наиболее часто встречающееся в выборке.

Для симметричных распределений признаков, в том числе для нормального распределения, значения моды совпадают со значений среднего и медианы. Для других типов распре­делений, несимметричных, это не характерно.

Метод вторичной статистической обработки, по­средством которого выясняется связь или прямая зависимость между двумя рядами экспериментальных данных, носит назва­ние *метод корреляционного анализа.* Он показывает, каким образом одно яв­ление влияет на другое или связано с ним в своей динамике. По­добного рода зависимости существуют, к примеру, между вели­чинами, находящимися в причинно-следственных связях друг с другом. Если выясняется, что два явления статистически досто­верно коррелируют друг с другом и если при этом есть уверен­ность в том, что одно из них может выступать в качестве причи­ны другого явления, то отсюда определенно следует вывод о на­личии между ними причинно-следственной зависимости.

Имеется несколько разновидностей данного метода:

Линейный корреля­ционный анализ позволяет устанавливать прямые связи между переменными величинами по их абсолютным значениям. Эти связи графически выражаются прямой линией, отсюда название «линейный».

Коэффициент линейной корреляции определяется при по­мощи следующей формулы:

 (3)

где rxy *—* коэффициент линейной корреляции;

*х, у -* средние выборочные значения сравниваемых величин;

*хi,уi —* частные выборочные значения сравниваемых величин;

*п —* общее число величин в сравниваемых рядах показателей;

— дисперсии, отклонения сравниваемых величин от средних значений.

Ранговая корреляция определяет зависимость не между абсолютными значениями переменных, а между поряд­ковыми местами, или рангами, занимаемыми ими в упорядочен­ном по величине ряду. Формула коэффициента ран­говой корреляции следующая32:

 (4)

где Rs — коэффициент ранговой корреляции по Спирмену;

*di —* разница между рангами показателей одних и тех же ис­пытуемых в упорядоченных рядах;

*п —* число испытуемых или цифровых данных (рангов) в кор­релируемых рядах.