Лекция № 13 . Математическая статистика. Понятие выборки.

Математическая статистика является частью общей прикладной математической дисциплины «Теория вероятностей и математическая статистика», однако задачи, решаемые ею, носят специфический характер. Если теория вероятностей исследует явления, полностью заданные их моделью, то в математической статистике вероятностная модель определена с точностью до неизвестных параметров. Отсутствие сведений о параметрах компенсируется «пробными» испытаниями, на основе которых и восстанавливается недостающая информация. Задачи математической статистики:

Первая задача математической статистики – определение способов сбора и группировки статистических информаций;

Вторая задача математической статистики – разработка методов анализа статистических данных, интерпретация и формирование выводов.

Изучение закономерностей объектов достаточно большой совокупности методами математической статистики основано на использовании статистических данных для некоторой конечной части рассматриваемых объектов.

Допустим, у нас есть некоторая совокупность однородных объектов и нас интересует некоторый количественный или качественный признак, характеризующий эти объекты, например, раз мер деталей, в магазине – вес расфасованных продуктов. Данный признак мы будем интерпретировать как случайную величину, значение которой меняется от объекта к объекту. Иногда проводят сплошное обследование – обследуют каждый объект совокупности относительно признака, которым интересуются. Но не всегда это возможно. Обычно из всей совокупности объектов случайным образом отбирают ограниченное число объектов, которые и подвергают изучению.

Генеральная совокупность – это все мыслимые значения (из мерения, наблюдения), описывающие поведение исследуемого объекта или явления.

Выборка – ограниченный набор реально наблюдаемых выборочных из генеральной совокупности значений, описывающих исследуемый объект или явление. Количество этих значений называют объемом выборки.

Понятие генеральной совокупности аналогично понятию случайной величины. Выборку можно рассматривать как некоторый эмпирический аналог генеральной совокупности.

Пример: Количество зарегистрированных малых предприятий торговли питания в городе Ташкенте равно 3 525. Для исследования предприятий по

объему товарооборота взято 150. В данном случае 3 525 – объем генеральной совокупности, а 150 объем выборки.

Сущность выборочного метода состоит в том, чтобы по некоторой части генеральной совокупности (т. е. по выборке) выносить данные о ее свойствах в целом.

Достоинства выборочного метода:

– позволяет существенно экономить затраты ресурсов;

– является единственно возможным в случае бесконечной генеральной совокупности;

– при тех же затратах ресурсов дает возможность проведения углубленного исследования за счет расширения программы исследования;

– позволяет снизить ошибки регистрации.

Недостатки выборочного метода:

– ошибки исследования, называемые ошибками репрезентативности.

Однако неизбежные ошибки могут быть и заранее оценены с помощью правильной организации выборки и сведены к практически незначительным величинам.

Чтобы по данным выборки иметь возможность судить о генеральной совокупности, она должна быть отобрана случайно.

Выборка называется репрезентативной (представительной), если она достаточно хорошо воспроизводит генеральную совокупность.

Различают несколько видов выборки:

x_i – значение признака (т. е. случайной величины X);

N , n – объемы генеральной совокупности и выборки соответственно.

Средние арифметические распределения признака в генеральной и выборочной совокупности называются генеральным и выборочным средним соответственно. Дисперсии – генеральной и выборочной дисперсией.

Важнейшей задачей выборочного метода является оценка параметров (характеристик) генеральной совокупности по данным выборки. Теоретическую основу применимости выборочного метода составляет закон больших чисел, согласно которому при неограниченном увеличении выборки практически достоверно, что случайные выборочные характеристики как угодно близко

приближаются по вероятности к ограниченным параметрам генеральной совокупности.

Оценка неизвестных параметров переменной происходит на основании анализа материала наблюдения. Однако прежде чем при ступить к оцениванию, производят предварительную обработку материала наблюдения – составляют вариационный ряд и рассчитывают некоторые описательные статистики этого ряда, которые будут анализироваться дальше.

Возникает вопрос: Зачем нужна математическая статистика? Рассмотрим несколько областей, в которых она применяется.

- Маркетинг. Изучение окупаемости рекламы, исследование рынка, исследование и анализ целевых аудиторий и потребительских предпочтений, выстраивание прогнозов спроса и предложения.

- Бизнес. При разработке бизнес-плана потребуется все детально рассчитывать: за сколько вы сможете купить или продать тот или иной товар или услугу. Необходимо построить (смоделировать) несколько сценариев: оптимистичный, средний, пессимистичный. Затем в каждом из сценариев есть разветвления. Таким образом, можно построить некое подобие дерева решений, где можно в каждой ситуации просчитать ожидаемую прибыль к тому или иному месяцу.

- Банковское дело. Построение разумной стратегии по выдаче кредитов. Возникает случайная величина: будет возвращен кредит или нет. Чтобы определить, кому выдать кредит, а кому – нет, банк анализирует статистическую информацию. Сюда входит и кредитная история самого человека, и процент вернувших кредит в срок и т. д. Этот анализ проводится методами теории вероятностей и математической статистики.

Пример: Банк выдает кредиты по 1 млн руб. сроком на 1 год. Известно, что в среднем вероятность невозврата кредита равна 1 %. Какую процентную ставку должен установить банк, чтобы в среднем иметь прибыль?

- Страхование. Наступление страхового случая или избежание такового – величина случайная. Страховая компания анализирует статистические данные по наступлению страхового случая, в котором они наступили. Таким образом, можно оценить вероятность наступления страхового случая и назначить для него страховой взнос.

Пример: Пусть страховая компания заключает договоры страхования сроком на 1 год на S рублей каждый. Страховой случай происходит с вероятностью p и не происходит с вероятностью $q = 1 - p$. Таким образом, имеем закон

распределения случайной величины X – количество страховых случаев у одного страхователя, нухна узнать количество страховых случаев у страхователей, в среднем страховая компания должна будет выплатить страховых возмещений, т. е. если с каждого брать страхового взноса, то у компании будет нулевой баланс. Реальная страховая ставка ставится в силу сбора данных.

Таким образом, нам нужно исследовать поведение тех или иных объектов или явлений. А оно осуществляется на основе изучения статистических данных – наблюдений и измерений.

Вариационные ряды и их характеристики Установление статистических закономерностей, присущих массовым случайным явлениям, основано на изучении статистических данных – сведений о том, какие значения принял в результате наблюдений интересующий нас признак X. Рассмотрим X – числовую характеристику совокупности объектов.

Пример. Необходимо изучить распределение размеров обуви, проданной в интересующем магазине, с целью обеспечить нужное количество обуви каждого размера. Получены следующие данные о размерах про данной в магазине за сутки обуви (женской): 35, 35, 36, 36, …, 42, 43. Всего продано 100 штук.

Рассмотрение и осмысление данных, представленных в таком виде, практически невозможно из-за обилия числовой информации. Поэтому проводят группировку представленной совокупности чисел.

Различные значения признака X, наблюдавшиеся у объектов, называются

вариантами, а их количество – *частотами*.

Сгруппированный ряд представляют в виде таблицы. За смену продано 100 пар обуви.

x_i	36	37	38	39	40	41	42	43
n_i	2	6	12	10	22	21	13	14
$\frac{n_i}{N}$	0,02	0,06	0,12	0,1	0,22	0,21	0,13	0,14

где $\sum_{i=1}^{k} n_i = N = 100$, $\sum_{i=1}^{k} \frac{n_i}{N} = 1$

Частоты показывают, сколько раз встречаются наблюдения, у которых значение признака X равно данной варианте.

Вариационным рядом называется ранжированный в порядке возрастания или убывания ряд вариант с соответствующими весами (частотами или относительными частотами).

Вариационный ряд можно определить для дискретных и непрерывных величин X. В последнем случае проводят интервальную группировку ряда.

Число интервалов рекомендуется брать по формуле Стерджеса:

$$n = 1 + 3{,}322 \ln N\,, \quad h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \ln N}$$

где $x_{\max} - x_{\min}$ разность между наибольшим и наименьшим значением признака. За начало первого интервала рекомендуют брать величину $x_0 = x_{\min} - \frac{h}{2}$.

Частота показывает число членов совокупности, у которых признак X принимает значения в границах интервалов.

Пример. Урожайность хлопка 50 хозяйств, с каждого гектара в центнерах дала такие результаты

$x_i - x_{i+1}$	8-12	12-16	16-20	20-24	25-32	32-40
n_i	5	8	9	11	7	10
$\frac{n_i}{N} = w_i$	0,1	0,16	0,18	0,22	0,14	0,2

В этом случае n_i – плотность распределения, а w_i – относительная плотность распределения вариационного ряда.

Полученный вариационный ряд позволяет выявить закономерности изменчивости признака, закономерности распределения обуви по размеру проданных пар и участков по урожайности, что сделать по первичным, не сгруппированным данным оказалось затруднительно.

Наряду с понятием частот и относительных частот для описания вариационного ряда используются накопленные частоты и на копленные относительные частоты.

Графическое представление вариационных рядов.

Представление вариационного ряда в виде таблицы не всегда удобно. Поэтому используют различные способы графического представления вариационных рядов.

Полигон частот – ломаная, соединяющая точки (x_i, n_i).

Полигон относительных частот – ломаная, соединяющая точ ки (x_i, w_i).

В случае непрерывного признака X целесообразно строить различные гистограммы.

Гистограмма частот называют ступенчатую фигуру, состоящую из прямоугольников, основанием которых служит интервалы длиной h, высотой – плотностью вариационного ряда, деленной на величину соответствующего интервала $\frac{n_i}{h}$. Площадь гистограммы равна сумме всех частот N.

Гистограмма относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основанием которых служит интервалы длиной h, высотой – относительной плотностью вариационного ряда, деленной на величину соответствующего интервала $\frac{w_i}{h}$. Площадь гистограммы равна 1.

Эмпирическая функция распределения.

Эмпирической функцией распределения называется функция $F^*(x)$, выражающая для каждого значения x относительную частоту события $X < x$:

$$F^*(x) = \frac{n_x}{N},$$

где n_x – число вариант, меньше x, N - объема выборки.

Вариационным рядом называется ранжированный в порядке возрастания (или убывания) ряд вариантов с соответствующими им весами (частотами или частостями).

Вариационный ряд является статистическим аналогом (реализацией) распределения признака (случайной величины X). В этом смысле полигон или гистограмма аналогичен кривой распределения, а эмпирическая функция распределения – функции распределения случайной величины X.

Вариационный ряд содержит достаточно полную информацию об изменчивости признака X. Однако на практике часто оказывается, что этого недостаточно и необходимо найти некоторые сводные характеристики вариационных рядов: средних, центральной тенденции и изменчивости (показателей вариации), расчет которых представляет собой следующий этап после группировки и об работки данных наблюдений.

Средние величины характеризуют значения признака, вокруг которого концентрируются наблюдения. Наиболее распространенной среди средних величин является средняя арифметическая.