Математическая статистика.

Основной задачей математической статистики является разработка методов получения научно обоснованных выводов о массовых явлениях и процессах из данных наблюдений и экспериментов. Эти выводы и заключения относятся не к отдельным испытаниям, из повторения которых складывается данное массовое явление, а представляют собой утверждения об общих вероятностных характеристиках данного процесса, то есть о вероятностях, законах распределения, математических ожиданиях, дисперсиях и т. д. Такое использование фактических данных как раз и является отличительной чертой статистического метода.

Пусть мы располагаем сведениями (обычно довольно ограниченными), например, о числе дефектных изделий в изготовленной в определенных условиях продукции или о результатах испытаний материалов на разрушение и т. п. Собранные нами данные могут представлять непосредственный интерес в смысле информации о качестве той или иной партии продукции. Статистические же проблемы возникают тогда, когда мы на основе той же информации начинаем делать выводы относительно более широкого круга явлений. Так например нас может интересовать качество технологического процесса, для чего мы оцениваем вероятность получения в нем дефектного изделия или среднюю долговечность изделия. В этом случае мы рассматриваем собранный материал не ради его самого, а лишь как некую пробную группу или выборку, представляющую только серии из возможных результатов, которые мы могли бы встретить при продолжении наблюдений массового процесса в данной обстановке. Выводы и оценки, основанные на материале наблюдений, отражают случайный состав пробной группы и поэтому считаются приблизительными оценками вероятностного характера. Во многих случаях теория указывает, как наилучшим способом использовать имеющуюся информацию для получения по возможности более точных и надежных характеристик, указывая при этом степень надежности выводов, объясняющуюся ограниченностью запаса сведений.

В математической статистике рассматриваются две основные категории задач: оценивание и статистическая проверка гипотез. Первая задача разделяется на точечное оценивание и интервальное оценивание параметров распределения. Например может возникнуть необходимость по наблюдениям получить точечные оценки параметров $M\xi$ и $D\xi$. Если мы хотим получить некоторый интервал, с той или иной степенью достоверности содержащий истинное значение параметра, то это задача интервального оценивания.

Вторая задача – проверка гипотез – заключается в том, что мы делаем предположение о распределении вероятностей случайной величины (например, о значении одного или нескольких параметров функции распределения) и решаем, согласуются ли в некотором смысле эти значения параметров с полученными результатами наблюдений.

Выборочный метод.

Пусть нам нужно обследовать количественный признак в партии экземпляров некоторого товара. Проверку партии можно проводить двумя способами:

1) провести сплошной контроль всей партии;

2) провести контроль только части партии.

Первый способ не всегда осуществим, например, из–за большого числа экземпляров в партии, из–за дороговизны проведения операции контроля, из–за того, что контроль связан с разрушением экземпляра (проверка электролампы на долговечность ее работы).

При втором способе множество случайным образом отобранных объектов называется **выборочной совокупностью** или **выборкой**. Все множество объектов, из которого производится выборка, называется **генеральной совокупностью**. Число объектов в выборке называется **объемом выборки**. Обычно будем считать, что объем генеральной совокупности бесконечен.

Выборки разделяются на **повторные** (с возвращением) и **бесповторные** (без возвращения).

Обычно осуществляются бесповторные выборки, но благодаря большому (бесконечному) объему генеральной совокупности ведутся расчеты и делаются выводы, справедливые лишь для повторных выборок.

Выборка должна достаточно полно отражать особенности всех объектов генеральной совокупности, иначе говоря, выборка должна быть **репрезентативной** (представительной).

Выборки различаются по способу отбора.

1. Простой случайный отбор.

Все элементы генеральной совокупности нумеруются и из таблицы случайных чисел берут, например, последовательность любых 30-ти идущих подряд чисел. Элементы с выпавшими номерами и входят в выборку.

2. Типический отбор.

Такой отбор производится в том случае, если генеральную совокупность можно представить в виде объединения подмножеств, объекты которых однородны по какому–то признаку, хотя вся совокупность такой однородности не имеет (партия товара состоит из нескольких групп, произведенных на разных предприятиях). Тогда по каждому подмножеству проводят простой случайный отбор, и в выборку объединяются все полученные объекты.

3. Механический отбор.

Отбирают каждый двадцатый (сотый) экземпляр.

4. Серийный отбор.

В выборку подбираются экземпляры, произведенные на каком–то производстве в определенный промежуток времени.

В дальнейшем под генеральной совокупностью мы будем подразумевать не само множество объектов, а множество значений случайной величины, принимающей числовое значение на каждом из объектов. В действительности генеральной совокупности как множества объектов может и не существовать. Например имеет смысл говорить о множестве деталей, которые **можно произвести**, используя данный технологический процесс. Используя какие–то известные нам характеристики данного процесса, мы можем оценивать параметры этого

несуществующего множества деталей. Размер детали – это случайная величина, значение которой определяется воздействием множества факторов, составляющих технологический процесс. Нас, например, может интересовать вероятность, с которой эта случайная величина принимает значение, принадлежащее некоторому интервалу. На этот вопрос можно ответить, зная закон распределения этой случайной величины, а также ее параметры, такие как $M\xi$ и $D\xi$.

Итак, отвлекаясь от понятия генеральной совокупности как множества объектов, обладающих некоторым признаком, будем рассматривать генеральную совокупность как случайную величину ξ, закон распределения и параметры которой определяются с помощью выборочного метода.

Рассмотрим выборку объема n, представляющую данную генеральную совокупность. Первое выборочное значение x_1 будем рассматривать как реализацию, как одно из возможных значений случайной величины ξ_1, имеющей тот же закон распределения с теми же параметрами, что и случайная величина ξ. Второе выборочное значение x_2 – одно из возможных значений случайной величины ξ_2 с тем же законом распределения, что и случайна величина ξ. То же самое можно сказать о значениях x_3, x_4,..., x_n .

Таким образом на выборку будем смотреть как на совокупность независимых случайных величин ξ_1, ξ_2, ..., ξ_n, распределенных так же, как и случайная величина ξ, представляющая генеральную совокупность. **Выборочные значения x_1, x_2, ..., x_n – это значения, которые приняли эти случайные величины в результате 1-го, 2-го, ..., n-го эксперимента**.

Вариационный ряд.

Пусть для объектов генеральной совокупности определен некоторый признак или числовая характеристика, которую можно замерить (размер детали, удельное количество нитратов в дыне, шум работы двигателя). Эта характеристика – случайная величина ξ, принимающая на каждом

объекте определенное числовое значение. Из выборки объема n получаем значения этой случайной величины в виде ряда из n чисел:

$$x_1, x_2,..., x_n. \qquad (*)$$

Эти числа называются значениями признака.

Среди чисел ряда (*) могут быть одинаковые числа. Если значения признака упорядочить, то есть расположить в порядке возрастания или убывания, написав каждое значение лишь один раз, а затем под каждым значением x_i признака написать число m_i, показывающее сколько раз данное значение встречается в ряду (*):

x_1	x_2	x_3	...	x_k
m_1	m_2	m_3	...	m_k

то получится таблица, называемая **дискретным вариационным рядом**. Число m_i называется частотой i-го значения признака.

Очевидно, что x_i в ряду (*) может не совпадать с x_i в вариационном ряду. Очевидна также справедливость равенства

$$\sum_{i=1}^{k} m_i = n\,.$$

Если промежуток между наименьшим и наибольшим значениями признака в выборке разбить на несколько интервалов одинаковой длины, каждому интервалу поставить в соответствие число выборочных значений признака, попавших в этот интервал, то получим **интервальный вариационный ряд**. Если признак может принимать любые значения из некоторого промежутка, то есть является непрерывной случайной величиной, приходится выборку представлять именно таким рядом. Если в вариационном интервальном ряду каждый интервал $[\alpha_i;\ \alpha_{i+1})$ заменить лежащим в его середине числом $(\alpha_i+\alpha_{i+1})/2$, то получим дискретный вариационный ряд. Такая замена вполне естественна, так как, например, при измерении размера детали с точностью до одного миллиметра всем размерам из промежутка [49,5; 50,5), будет соответствовать одно число, равное 50.

Точечные оценки параметров генеральной совокупности.

Во многих случаях мы располагаем информацией о виде закона распределения случайной величины (нормальный, бернуллиевский, равномерный и т. п.), но не знаем параметров этого распределения, таких как $M\xi$, $D\xi$. Для определения этих параметров применяется выборочный метод.

Пусть выборка объема n представлена в виде вариационного ряда. Назовем **выборочной средней** величину

$$\overline{x} = \frac{x_1 m_1 + x_2 m_2 + \ldots + x_k m_k}{n} = x_1 \frac{m_1}{n} + x_2 \frac{m_2}{n} + \ldots + \frac{m_k}{n}$$

Величина $\omega_i = \frac{m_i}{n}$ называется **относительной частотой** значения признака x_i. Если значения признака, полученные из выборки не группировать и не представлять в виде вариационного ряда, то для вычисления выборочной средней нужно пользоваться формулой

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Естественно считать величину $\overline{x}$ выборочной оценкой параметра $M\xi$. Выборочная оценка параметра, представляющая собой число, называется **точечной оценкой**.

Выборочную дисперсию

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \overline{x})^2 \omega_i = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

можно считать точечной оценкой дисперсии $D\xi$ генеральной совокупности.

Приведем еще один пример точечной оценки. Пусть каждый объект генеральной совокупности характеризуется двумя количественными признаками x и y. Например деталь может иметь два размера – длину и ширину. Можно в различных районах измерять концентрацию вредных веществ в воздухе и фиксировать количество легочных заболеваний населения в месяц. Можно через равные промежутки времени

сопоставлять доходность акций данной корпорации с каким-либо индексом, характеризующим среднюю доходность всего рынка акций. В этом случае генеральная совокупность представляет собой двумерную случайную величину ξ, η. Эта случайная величина принимает значения *x*, *y* на множестве объектов генеральной совокупности. Не зная закона совместного распределения случайных величин ξ и η, мы не можем говорить о наличии или глубине корреляционной связи между ними, однако некоторые выводы можно сделать, используя выборочный метод.

Выборку объема *n* в этом случае представим в виде таблицы, где *i*-тый отобранный объект (*i*= 1,2,...*n*) представлен парой чисел x_i, y_i :

x_1	x_2	...	x_n
y_1	y_2	...	y_n

Выборочный коэффициент корреляции рассчитывается по формуле

$$r_{xy} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}$$

Здесь

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \ , \ \sigma_x = \sqrt{\sigma_x^{\,2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \ ,$$

$$\sigma_y = \sqrt{\sigma_y^{\,2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} \ .$$

Выборочный коэффициент корреляции можно рассматривать как точечную оценку коэффициента корреляции $\rho_{\xi\eta}$, характеризующего генеральную совокупность.

Выборочные параметры $\overline{x}, s_x, r_{xy}$ или любые другие зависят от того, какие объекты генеральной совокупности попали в выборку и различаются от выборки к выборке. Поэтому они сами являются случайными величинами.

Пусть выборочный параметр δ рассматривается как выборочная оценка параметра Δ генеральной совокупности и при этом выполняется равенство

$$M\delta = \Delta.$$

Такая выборочная оценка называется **несмещенной.**

Для доказательства несмещённости некоторых точечных оценок будем рассматривать выборку объема n как систему n независимых случайных величин $\xi_1, \xi_2, \ldots \xi_n$, каждая из которых имеет тот же закон распределения с теми же параметрами, что и случайная величина ξ, представляющая генеральную совокупность. При таком подходе становятся очевидными равенства: $Mx_i = M\xi_i = M\xi$; $Dx_i = D\xi_i = D\xi$ для всех $k = 1,2,\ldots n$.

Теперь можно показать, что выборочная средняя $\bar{x}$ есть несмещенная оценка средней генеральной совокупности или , что то же самое, математического ожидания интересующей нас случайной величины ξ:

$$M\bar{x} = M\,\frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\left(M\xi_1 + M\xi_2 + \ldots + M\xi_n\right) = \frac{1}{n}nM\xi = M\xi.$$

Выведем формулу для дисперсии выборочной средней:

$$D\bar{x} = D\,\frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n^2}\left(D\xi_1 + D\xi_2 + \ldots + D\xi_{n1}\right) = \frac{1}{n^2}nD\xi = \frac{D\xi}{n}.$$

Найдем теперь, чему равно математическое ожидание выборочной дисперсии σ^2. Сначала преобразуем σ^2 следующим образом:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - M\xi + M\xi - \bar{x}\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - M\xi\right)^2 - 2\left(x_i - M\xi\right)\left(\bar{x} - M\xi\right) + \left(\bar{x} - M\xi\right)^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - M\xi)^2 - (\bar{x} - M\xi)^2$$

Здесь использовано преобразование:

$$\sum_{i=1}^{n} 2(x_i - M\xi)(\bar{x} - M\xi) = 2(\bar{x} - M\xi)\sum_{i=1}^{n}(x_i - M\xi) =$$

$$= 2(\bar{x} - M\xi)\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} M\xi\right) = 2(\bar{x} - M\xi)(n\bar{x} - nM\xi) = 2n(\bar{x} - M\xi)^2$$

Теперь, используя полученное выше выражение для величины σ^2, найдем ее математическое ожидание.

$$M\sigma^2 = M\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - M\xi)^2 - (\bar{x} - M\xi)^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} M(x_i - M\xi)^2 - M(\bar{x} - M\xi)^2 = \frac{1}{n} nD\xi - D\bar{x} =$$

$$= D\xi - \frac{D\xi}{n} = \frac{n-1}{n} D\xi .$$

Так как $M\sigma^2 \neq D\xi$, **выборочная дисперсия не является несмещенной оценкой дисперсии генеральной совокупности**.

Чтобы получить несмещенную оценку дисперсии генеральной совокупности, нужно умножить выборочную дисперсию на $\frac{n}{n-1}$. Тогда получится величина $s^2 = \frac{n}{n-1}\sigma^2$, называемая **исправленной выборочной дисперсией**.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Пусть имеется ряд несмещенных точечных оценок одного и того же параметра генеральной совокупности. Та оценка, которая имеет наименьшую дисперсию называется **эффективной**.

Полученная из выборки объема n точечная оценка δ_n параметра Δ генеральной совокупности называется **состоятельной**, если она сходится по вероятности к Δ. Это означает, что для любых положительных чисел ε и γ найдется такое число $n_{\varepsilon\gamma}$, что для всех чисел n, удовлетворяющих неравенству $n > n_{\varepsilon\gamma}$ выполняется условие $P\left(\left|\delta_n - \Delta\right| < \varepsilon\right) > 1 - \gamma$.

$\bar{x}$ и s^2 являются несмещёнными, состоятельными и эффективными оценками величин $M\xi$ и $D\xi$.