Специальная математика и основы статистики

Математическая статистика

Вопрос 1. Предмет математической статистики и ее основные задачи

В теории вероятностей, если мы изучаем случайную величину X, ее закон распределения считается заданным, и мы можем достоверно ответить на любой вопрос, касающийся данной случайной величины В математической
статистике ситуация прямо противоположная – мы ничего не знаем о законе распределения изучаемой случайной величины X. У нас имеются только некоторые ее наблюдения или измерения. Понятно, что по конечному числу наблюдений невозможно достоверно сделать какие-либо выводы об изучаемой случайной величине. Ясно также, что чем больше таких наблюдений, тем более надежными будут наши приближенные выводы
В этом состоит основная особенность математической статистики – она не определяет достоверно закономерности поведения изучаемых случайной явлений, а оценивает их с той или иной степенью достоверности. Но при неограниченном увеличении числа наблюдений выводы математической статистики становятся практически достоверными.
Поэтому содержание этой дисциплины – как и сколько сделать наблюдений и как их обработать, чтобы ответить на интересующий нас вопрос о случайном явлении с требуемой степенью достоверности.
Итак, установление закономерностей, которым подчинены массовые случайные явления основано на изучении статистических данных – результатах наблюдений.
Математическая статистика решает две главные задачи:
1) указать способы сбора и группировки (если данных очень много) статистических сведений (результатов наблюдений);
2) разработать методы анализа собранных статистических данных в зависимости от целей исследования.

Вопрос 2. Выборка. Статистический ряд. Эмпирический закон распределения. Полигон и гистограмма

Пусть требуется изучить совокупность однородных объектов относительно качественного или количественного признака, характеризующего эти объекты.

Можно производить этот контроль сплошным обследованием, то есть измерять каждый из объектов совокупности. Но на практике сплошное обследование применяется редко:

а) из-за очень большого числа объектов;

б) из-за того, что иногда обследование заключается в физическом уничтожении, например, проверяем взрываемость гранат или проверяем на
крепость произведеннуюпосуду и т.д.

В таких случаях производится случайный отбор ограниченного (небольшого) числа объектов, которые и подвергают изучению.

Выборочной совокупностью (выборкой) называется совокупность случайно отобранных однородных объектов.

Генеральной совокупностью (ГС) называется совокупность всех однородных объектов, из которых производится выборка.

Объемом совокупности (выборочной или генеральной) называется число объектов этой совокупности.

При наборе выборки можно поступать двояко: после того, как объект отобран и над ним произведено наблюдение, он может быть возвращен либо не возвращен в генеральную совокупность. В связи с этим выборки подразделяются на ***повторные*** и ***бесповторные***.

Для того, чтобы по данным выборки можно было достаточно уверенно судить об интересующем нас признаке ГС, необходимо, чтобы объекты выборки правильно его представляли. Это требование коротко формулируется так: выборка должна быть ***репрезентативной (представительной)***.

Способы отбора выборки:

1. Отбор, не требующий расчленения ГС на части:

а) простой случайный бесповторный;

б) простой случайный повторный.

2. Отбор, при котором ГС разбивается на части (если объем ГС слишком большой):

а) типический отбор. Объекты отбираются не из всей ГС, а из ее «типичных» частей. Например, цех из тридцати станков производит одну и ту же деталь. Тогда отбор делается по одной или по две детали с каждого станка в случайные моменты времени;

б) механический отбор. Например, если нужно выбрать 5% деталей, то выбирают не случайно, а каждую двадцатую деталь;

в) серийный отбор. Объекты выбирают не по одному, а сериями.

Итак, пусть из ГС значений некоторого количественного признака произведена вьборка объема N: $X = \{x_1, \dots, x_N\}$.

Таблица вида

№	1	2	…	N
x	x_1	x_2	…	x_N

называется ***простым статистическим рядом***, являющимся первичной формой представления статистического материала.

Если данные таблицы, называемые ***вариантами***, упорядочены по возрастанию, то выборка $X = \{x_1, \dots, x_N\}$ называется ***вариационным рядом***

Размах выборки – это длина основного интервала $[x_{min}; x_{max}]$, в который попадают все значения выборки. Вычисляется размах выборки следующим образом: $d = x_{max} - x_{min}$. Затем по формуле $k = 1 + 3{,}322 \cdot lgN$, определяется число k . Данное число задает количество подынтервалов (классов), на которые разбиваем основной интервал. Длины h подынтервалов и их границы a_i $(i = \overline{0,k})$ вычисляются по формулам: $h = \frac{d}{k}$, $a_0 = x_{min}, a_1 = a_0 + h, \dots, a_i = a_{i-1} + h$

Далее находятся ***частоты*** m_j $(j = \overline{1,k})$ и ***относительные частоты*** $\mu_j = \frac{m_j}{N}$ $(j = \overline{1,k})$ попадания значений выборки X в j-й подынтервал. Причем для частот должно выполняться равенство $\sum_{j=1}^{k} m_j = N$, а для относительных частот соответственно $\sum_{j=1}^{k} \mu_j = 1$.

Относительные частоты иногда называют ***эмпирическими вероятностями***.

Результаты проведенных расчетов сводятся в таблицу, которая называется ***интервальным вариационным рядом***

x	$[a_0;a_1)$	$[a_1;a_2)$	…	$[a_{k-1};a_k]$
m	m_1	m_2	…	m_k
μ	μ_1	μ_2	…	μ_k

Далее находятся середины подынтервалов: $\overline{x_i} = \frac{a_i + a_{i-1}}{2}$ и составляется таблица, которая называется ***дискретный вариационный ряд***:

x	$\overline{x_1}$	$\overline{x_2}$	…	$\overline{x_k}$
m	m_1	m_2	…	m_k
μ	μ_1	μ_2	…	μ_k

Вариационный ряд распределения является оценкой теоретического ряда распределения и сходится к нему по вероятности. Поскольку ряд распределения является одной из форм задания закона распределения дискретной случайной величины, то мы получили ***эмпирический закон распределения*** исследуемой дискретной случайной величины

Сгруппированные данные вариационного ряда несут в себе меньше информации, чем выборочные, так как в них теряется информация о

порядке следования выборочных значений. При группировке также фактически происходит округление наблюдаемых значений выборки внутри j-го класса
(подынтервала) до значения x_j, что приводит к потере информации о распределении исследуемой случайной величины внутри каждого класса. Это распределение в дальнейшем предполагается равномерным. Преимуществом же сгруппированных данных является их компактность и большая наглядность.

В целях визуального изучения полученных вариационных рядов пользуются различными способами их графического изображения. К ним относятся гистограмма и полигон.

Гистограммой относительных частот называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные
подынтервалы длины h , а высоты равны числам $\frac{\mu_j}{h}$ (плотности вероятностей) $(j = \overline{1,k})$.

Площадь гистограммы относительных частот равна единице. Аналогичным образом строится гистограмма частот. Площадь гистограммычастот равна объему выборки.

Полигоном относительных частот называется ломаная, отрезки которой соединяют точки $(\overline{x_1}; \mu_1), \dots, (\overline{x_k}; \mu_k)$. Полигон относительных частот есть визуальное представление эмпирического закона распределения выборки.

Любая функция выборки $\varphi = \varphi(x_1, \dots, x_N)$ называется ***статистикой***. Статистика является случайной величиной, так как на различных реализациях выборки она получает различные наблюдаемые значения. Статистиками являются: частоты m_j , границы классов a_j и их середины $\overline{x}_j$, размах. Вариационный ряд распределения также является статистикой. Из определения статистики следует, что любая функция от статистик также является статистикой, поэтому статистикой является любая функция от сгруппированных данных.

Статистики служат для оценки любых характеристик изучаемой случайной величины: вероятностей случайных событий, связанных с изучаемой величиной, ее числовых характеристик, параметров закона распределения и так далее. Изучение статистик на основе теории вероятностей есть теоретическое ядро математической статистики.

Вопрос 3. Статистические оценки генеральных параметров. Точечные и интервальные оценки

Пусть $\Theta_Г$ – некоторый параметр ГС, который невозможно вычислить. Но знать его значение (хотя бы приближенное, оценочное) надо! Поэтому по выборочным данным производят расчет статистических оценок данного генерального параметра.

Оценки параметров подразделяются на точечные и интервальные.

Точечной называется статистическая оценка генерального параметра $\Theta_Г$, которая определяется одним числом $\widetilde{\Theta}_в$.

Интервальной называется оценка генерального параметра $\Theta_Г$, которая определяется двумя числами $\widetilde{\Theta}_в$ и $\widetilde{\widetilde{\Theta}}_в$ – концами интервала, покрывающего оцениваемый генеральный параметр $\Theta_Г$.

Для того, чтобы точечная оценка давала «хорошие» приближения оцениваемого параметра, она должна быть: несмещенной, эффективной, состоятельной.

Несмещенной называют такую точечную оценку $\widetilde{\Theta}_в$, математическое ожидание которой равно оцениваемому генеральному параметру при любом

объеме выборки: $M(\theta_в) = \theta_Г$, в противном случае оценка называется ***смещенной***.

Эффективной называется точечная оценка $\widetilde{\Theta}_в$, которая (при заданном объеме выборки) имеет наименьшую возможную дисперсию: $M[(\theta_в - \theta_Г)^2] \to min$.

Состоятельной называется точечная оценка $\widetilde{\Theta}_в$, которая (с увеличением объема выборки) стремится по вероятности к оцениваемому параметру $\Theta_Г$: $\forall\delta > 0 \quad \lim\limits_{N\to\infty} P(|\theta_Г - \theta_в| < \delta) = 1$.

Несмещенной оценкой генеральной средней (генерального математического ожидания $M_Г[x]$) служит выборочная средняя ***(выборочное математическое ожидание)***: $M_в[x] = \sum_{j=1}^{k} \overline{x_j} \cdot \mu_j$, где $\overline{x_j}$ и μ_j – данные дискретного вариационного ряда.

Кроме того, $M_в[x]$ является состоятельной оценкой. Если случайная величина X подчинена нормальному закону распределения, то $M_в[x]$ является и эффективной оценкой.

Смещенной оценкой генеральной дисперсии $D_Г[x]$ служит ***выборочная дисперсия***: $D_в[x] = \sum_{j=1}^{k} \left(\overline{x_j} - M_в[x]\right)^2 \cdot \mu_j$.

Поскольку $D_в[x]$ является смещенной оценкой, то ее «исправляют» следующим образом: $\sigma_в^2 = \frac{N}{N-1} \cdot D_в[x]$.

Полученная оценка $\sigma_в^2$ – это состоятельная несмещенная выборочная дисперсия, а $\sigma_в$ – ***выборочное среднее квадратичное отклонение***.

При выборке малого объема (N<30) точечная оценка может значительно отличаться от оцениваемого генерального параметра, то есть приводить к грубым ошибкам. Поэтому при небольшом объеме выборки следует пользоваться интервальными оценками.

Пусть найденная (по данным выборки) статистическая оценка $\Theta_в$ является оценкой неизвестного генерального параметра $\Theta_Г$. Ясно, что $\Theta_в$ тем точнее определяет $\Theta_Г$, чем меньше значение разности $|\Theta_Г - \Theta_в|$. То есть при $|\Theta_Г - \Theta_в| < \delta$ $(\delta > 0)$ чем меньше δ, тем оценка $\Theta_Г$ точнее. Значит, положительное число δ характеризует ***точность*** оценки.

Надежностью (доверительной вероятностью) оценки $\Theta_в$ называется вероятность γ, с которой осуществляется событие $|\Theta_Г - \Theta_в| < \delta$, то есть $\gamma = P(|\Theta_Г - \Theta_в| < \delta)$.

Обычно надежность оценки (доверительная вероятность γ) задается. Причем в качестве γ берут число, близкое к единице (0,95; 0,99; 0,999).

Доверительным называется интервал, который с заданной надежностью γ покрывает оцениваемый генеральный параметр.

Из формулы, определяющей γ, если раскрыть модуль, получается $P(-\delta < \Theta_Г - \Theta_в < \delta) = \gamma$ или $P(\Theta_в - \delta < \Theta_Г < \Theta_в + \delta) = \gamma$.

Тогда интервал $(\Theta_в - \delta; \Theta_в + \delta)$ и есть доверительный интервал. Из общих соображений ясно, что длина доверительного интервала будет зависеть от объема выборки N и доверительной вероятности γ.

Для оценки математического ожидания $M_Г[x]$ нормально распределенной ГС X по выборочной средней $M_в[x]$ при <u>известном</u> среднем квадратическом отклонении $\sigma_Г = \sqrt{D_Г[x]}$ служит доверительный интервал: $M_в[x] - t \cdot \frac{\sigma_Г}{\sqrt{N}} < M_Г[x] < M_в[x] + t \cdot \frac{\sigma_Г}{\sqrt{N}}$, где $t \cdot \frac{\sigma_Г}{\sqrt{N}} = \delta$ – точность оценки; N – объем выборки; t – такое значение функции Лапласа Ф(t), при котором $\Phi(t) = \frac{\gamma}{2}$.

Для оценки математического ожидания $M_Г[x]$ нормально распределенной ГС X по выборочной средней $M_в[x]$ при <u>неизвестном</u> среднем квадратическом отклонении $\sigma_Г = \sqrt{D_Г[x]}$ (при объеме выборки N>30) служит доверительный интервал: $M_в[x] - t_\gamma \cdot \frac{\sigma_в}{\sqrt{N}} < M_Г[x] < M_в[x] + t_\gamma \cdot \frac{\sigma_в}{\sqrt{N}}$, где t_γ такое значение функции Лапласа Ф(t), при котором $\Phi(t) = \gamma$.

Для оценки среднего квадратического отклонения $\sigma_Г$ нормально распределенной ГС X с доверительной вероятностью γ служат доверительные интервалы: $\begin{matrix} \sigma_в(1-q) < \sigma_Г < \sigma_в(1+q) \ \ (\text{при } q < 1) \\ 0 < \sigma_Г < \sigma_в(1+q) \ \ (\text{при } q > 1) \end{matrix}$,

где q – находится из таблицы при заданном N и γ.

<u>Замечание 1.</u> Оценку $|\Theta_Г - \Theta_в| < t \cdot \frac{\sigma_Г}{\sqrt{N}}$ называют классической. Из формулы $\delta = t \cdot \frac{\sigma_Г}{\sqrt{N}}$, определяющей точность классической оценки, можно сделать выводы

1) при возрастании объема выборки N число δ убывает, значит, точность оценки увеличивается;

2) увеличение надежности оценки γ приводит к увеличению значения t (так как $\gamma = 2\Phi(t)$ и функция $\Phi(t)$ – возрастающая), значит, к возрастанию δ, то есть к уменьшению точности оценки.

Замечание 2. В связи со второй частью замечания 1 обычно предлагается строить доверительные интервалы для двух значений вероятности γ=0,95 и γ=0,99. Затем провести анализ, как меняются границы интервалов с увеличением доверительной вероятности.