1. ОСНОВНЫЕ ПОНЯТИЯ ВЫБОРОЧНОЙ ТЕОРИИ

1.1. Генеральная совокупность. Выборка. Выборочные характеристики

Прежде чем ввести основные понятия математической статистики, рассмотрим пример. Некоторое стабильно работающее (т.е. работающее в одних и тех же условиях) предприятие изготавливает приборы, которые характеризуются некоторым количественным признаком. В силу влияния не поддающихся учету факторов значение количественного признака от прибора к прибору меняется. Например в случае, когда интерес представляет доля брака в производстве приборов, каждому изделию можно приписать значение 1, если прибор функционирует нормально, и значение 0, если прибор неисправен. Количественным признаком может быть также время бесперебойной работы прибора, точность измерительного прибора, чувствительность датчика и т.п.

В силу объективных причин обеспечивать контроль каждого прибора, как правило, не удается. Поэтому для контроля качества продукции поступают следующим образом. Выбирают наудачу некоторое количество n (конечное число) приборов и по их показателям судят о всей продукции в целом, например о доле бракованных изделий или о средней продолжительности бесперебойной работы прибора и т.д. В подобных ситуациях естественно предполагать, что наблюдения за контролируемым показателем (хотя бы мысленно) можно проводить сколько угодно раз. Результаты n наблюдений рассматриваются как значения случайной величины — рассматриваемого количественного признака. Эта случайная величина может быть как

дискретной, так и непрерывной. Например, она может принимать только два значения 0 и 1, если речь идет о проверке, является прибор бракованным или нет. В другой же ситуации, когда оценивается время бесперебойной работы прибора, естественно считать, что случайная величина может принимать любое неотрицательное значение и является непрерывной.

В математической статистике множество возможных значений случайной величины X называют ***генеральной совокупностью*** случайной величины X или просто генеральной совокупностью X. Под ***законом распределения*** (***распределением***) ***генеральной совокупности*** X будем понимать закон распределения вероятностей случайной величины X.

Исходным материалом для изучения свойств генеральной совокупности (т.е. некоторой случайной величины) являются ***экспериментальные*** (***статистические***) ***данные***, под которыми понимают значения случайной величины, полученные в результате повторений случайного эксперимента (наблюдений над случайной величиной).

Предполагаем, что эксперимент хотя бы теоретически может быть повторен сколько угодно раз в одних и тех же условиях. Под словами „в одних и тех же условиях“ будем понимать, что распределение случайной величины X_i, $i = 1, 2, \ldots$, заданной на множестве исходов i-го эксперимента, не зависит от номера испытания и совпадает с распределением генеральной совокупности X. В этом случае принято говорить о ***независимых повторных экспериментах*** (***испытаниях***) или о ***независимых повторных наблюдениях*** над случайной величиной.

Совокупность независимых случайных величин $X_1, \ldots, X_n$, каждая из которых имеет то же распределение, что и случайная величина X, будем называть ***случайной выборкой*** из генеральной совокупности X и записывать $\vec{X}_n = (X_1, \ldots, X_n)$ (иногда просто $X_1, \ldots, X_n$). При этом число n называют ***объемом случайной выборки***, а случайные величины X_i — ***элементами случайной выборки.***

Любое возможное значение $\vec{x_n} = (x_1, \ldots, x_n)$ случайной выборки $\vec{X_n}$ будем называть ***выборкой*** из генеральной совокупности X (также ***реализацией случайной выборки*** $\vec{X_n}$) Число n характеризует ***объем выборки***, а числа x_i, $i = \overline{1, n}$, представляют собой ***элементы выборки*** $\vec{x_n}$. Выборку $\vec{x_n}$ можно интерпретировать как совокупность n чисел $x_1, \ldots, x_n$, полученных в результате проведения n повторных независимых наблюдений над случайной величиной X.

Основой любых выводов о вероятностных свойствах генеральной совокупности X, т.е. ***статистических выводов***, является ***выборочный метод***, суть которого заключается в том, что свойства случайной величины X устанавливаются путем изучения тех же свойств на случайной выборке.

Множество возможных значений случайной выборки $\vec{X_n}$ содержит информацию о случайной величине, полученную в эксперименте. Это множество называют ***выборочным пространством*** и обозначают $\mathcal{X}_n$. Выборочным пространством может быть или *n-мерное линейное арифметическое пространство* $\mathbb{R}^n$, или его подмножество. Если X — дискретная случайная величина, то выборочное пространство — конечное или счетное.

Элементы X_i, $i = \overline{1, n}$, случайной выборки $\vec{X}_n$ независимы и имеют то же распределение, что и генеральная совокупность X. Таким образом, функция распределения $F_{\vec{X}}(t_1, \ldots, t_n)$ случайной выборки $\vec{X_n}$ имеет вид

$$F_{\vec{X}}(t_1, \ldots, t_n) = \mathbf{P}\{X_1 < t_1, \ldots, X_n < t_n\} = \prod_{i=1}^{n} \mathbf{P}\{X_i < t_i\} = \prod_{i=1}^{n} F(t_i), \quad (1.1)$$

где $F(t)$ — функция распределения случайной величины X (генеральной совокупности X).

О распределении случайной величины X в одних случаях у исследователя могут быть самые общие представления. Напри-

мер, X является непрерывной случайной величиной и только (о распределении практически ничего не известно!). В других случаях функция распределения (в случае непрерывной случайной величины — плотность распределения вероятностей) известна, но не известны параметры, от которых она зависит. Например, известно, что генеральная совокупность X имеет нормальный закон распределения

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

где μ и σ — неизвестные параметры.

Значит, можно говорить только о семействе (классе) $\mathcal{P}$ распределений случайной выборки, в котором содержится ***априорная информация*** (информация до опыта) исследователя.

Выборочное пространство, на котором задан класс распределений $\mathcal{P}$, назовем ***статистической моделью****.

В случае повторных независимых испытаний статистическую модель будем обозначать $\{F(x)\}$, поскольку она полностью определена функцией распределения $F(x)$ генеральной совокупности X.

Если функция распределения (плотность распределения) задана с точностью до неизвестного вектора параметров $\vec{\theta} = (\theta_1, \ldots, \theta_r)$ с множеством возможных значений Θ, т.е. $\vec{\theta} \in \Theta$, то статистическую модель называют ***параметрической моделью***. Параметрическую модель обозначают $\{F(x;\vec{\theta});\ \vec{\theta} \in \Theta\}$. Множество Θ называют ***параметрическим множеством***.

Следует отметить, что о параметрическом множестве исследователь может не иметь никакой априорной информации.

Статистическую модель называют ***непрерывной*** или ***дискретной***, если случайная величина X является, соответственно, непрерывной или дискретной. В дальнейшем мы будем

предполагать, что генеральная совокупность X с функцией распределения $F(x)$ является либо дискретной, либо непрерывной случайной величиной. В первом случае распределение X задают в виде таблицы (ряда распределений), а во втором — в виде плотности распределения $p_X(x)$. При этом будем использовать единое обозначение $p(x)$ (или $p(x;\theta)$ для параметрических моделей) как для плотности распределения случайной величины X, когда она непрерывная, так и для вероятности $\mathbf{P}\{X = x\}$ в случае дискретной случайной величины X.

Пример 1.1. Пусть известно, что генеральная совокупность случайной величины X распределена по нормальному закону с известной дисперсией и неизвестным средним θ. Тогда статистическая модель имеет вид $\{F(x;\theta);\ \theta \in \Theta = \mathbb{R}\}$ и может быть задана с помощью плотности распределения вероятностей

$$p(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Если неизвестны оба параметра: среднее значение θ_1 и среднее квадратичное отклонение θ_2, то статистическая модель имеет вид $\left\{F(x;\vec{\theta});\ \vec{\theta} = (\theta_1,\ \theta_2) \in \Theta\right\}$, где $\Theta \subset \mathbb{R}^2$ $(\theta_1 \in \mathbb{R},\ \theta_2 \in \mathbb{R}^+)$ и плотность распределения вероятностей содержит два неизвестных параметра:

$$p(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}}, \quad x \in \mathbb{R}.$$

Пример 1.2. Пусть случайная величина X имеет распределение Пуассона с неизвестным параметром. Тогда статистическая модель имеет вид $\{F(x;\theta);\ \theta \in \Theta = (0,\infty)\}$, где $F(x;\theta)$ определяется равенством

$$p(x;\theta) = PrbX = x = \frac{\theta^x}{x!} e^{-\theta}, \quad x = 0, 1, 2, \ldots$$

1.2. Основные задачи математической статистики

При решении любой задачи математической статистики исследователь располагает двумя источниками информации. Первый и наиболее определенный (явный) — это результаты наблюдений (эксперимента) в виде *выборки* из некоторой *генеральной совокупности* скалярной или векторной случайной величины. При этом *объем выборки* n может быть фиксирован, а может увеличиваться в ходе эксперимента (т.е. могут использоваться так называемые последовательные процедуры статистического анализа).

Второй источник — это вся *априорная информация* об интересующих исследователя свойствах изучаемого объекта, которая накоплена к текущему моменту. Формально объем априорной информации отражается в той исходной *статистической модели*, которую исследователь выбирает при решении своей задачи.

В математической статистике всегда в той или иной мере используют априорную информацию об исследуемом объекте, но степень обоснованности такого использования лежит на совести (или зависит от компетентности) конкретного исследователя.

Если есть сомнения в том или ином исходном допущении при решении конкретной задачи, то его нужно проверять и обосновывать, а при невозможности это сделать — отбросить и попытаться найти решение задачи без привлечения сомнительных допущений.

Перечислим некоторые задачи математической статистики, наиболее часто встречающиеся в ее приложениях.

Оценка неизвестных параметров. Задача оценивания неизвестных параметров возникает в тех случаях, когда функция распределения генеральной совокупности известна с точностью до параметра θ. В этом случае необходимо найти такую

статистику $\widehat{\theta}(\vec{X}_n)$, *выборочное значение* $\widehat{\theta} = \widehat{\theta}(\vec{x}_n)$ которой для рассматриваемой *реализации* $\vec{x}_n$ *случайной выборки* можно было бы считать приближенным значением параметра $\vec{\theta}$.

Статистику $\widehat{\theta}(\vec{X}_n)$, выборочное значение $\widehat{\theta}$ которой для любой реализации $\vec{x}_n$ принимают за приближенное значение неизвестного параметра θ, называют его ***точечной оценкой*** или просто ***оценкой***, а $\widehat{\theta}$ — ***значением точечной оценки*** (просто ***оценки***).

Понятно, что точечная оценка $\widehat{\theta}(\vec{X}_n)$ должна удовлетворять вполне определенным требованиям для того, чтобы ее выборочное значение $\widehat{\theta}$ соответствовало истинному значению параметра $\vec{\theta}$.

Для точечных оценок параметра θ будем использовать и другие обозначения, например $\widetilde{\theta}(\vec{X}_n)$, $\theta^*(\vec{X}_n)$.

Возможным является и иной подход к решению рассматриваемой задачи: найти такие статистики $\overline{\theta}(\vec{X}_n)$ и $\underline{\theta}(\vec{X}_n)$, чтобы с вероятностью γ выполнялось неравенство

$$\mathbf{P}\{\underline{\theta}(\vec{X}_n) \leqslant \theta \leqslant \overline{\theta}(\vec{X}_n)\} = \gamma.$$

В этом случае говорят об *интервальной оценке* для θ. Интервал

$$(\underline{\theta}(\vec{X}_n), \overline{\theta}(X_n))$$

называют *доверительным интервалом* для θ с *коэффициентом доверия* γ.

Проверка статистических гипотез. ***Статистической гипотезой*** называют любое предположение о распределении вероятностей наблюдаемой случайной величины — скалярной или векторной.

В некотором смысле задача проверки статистической гипо-тезы является обратной к задаче оценивания параметра. При оценивании параметра мы ничего не знаем о его истинном зна-чении. При проверке статистической гипотезы мы из каких-то

соображений предполагаем известным его значение и хотим по результатам эксперимента проверить наше предположение.

Примерами гипотез могут служить следующие предположения о вероятностных свойствах наблюдаемых случайных величин:

1) $\mu = \mu_0$, где μ — математическое ожидание случайной величины X (гипотеза о величине математического ожидания);

2) $\sigma_1^2 = \sigma_2^2$, где σ_1^2 и σ_2^2 — дисперсии случайных величин X_1 и X_2 (гипотеза об однородности дисперсий);

3) $F(x) = F_T(x)$, где $F(x)$ — неизвестная функция распределения наблюдаемой случайной величины X, а $F_T(x)$ — некоторая предполагаемая исследователем функция распределения (гипотеза о виде распределения).

Установление формы и степени связи между случайными величинами. Методы математической статистики, способствующие установлению формы и степени связи между случайными величинами, излагаются в таких разделах математической статистики, как корреляционный анализ, дисперсионный анализ, регрессионный анализ и др.

Смысл таких задач поясним на простом примере. Пусть Y — случайная величина, поведение которой мы хотели бы определять по значениям двух других случайных величин X_1 и X_2. Например, Y — это степень шума двигателя автомашины, а X_1 и X_2 — соответственно величина пробега автомобиля и вес груза в нем. Корреляционный и дисперсионный анализ позволяет нам ответить на вопрос: есть ли связь между X_1, X_2 и Y и насколько она существенна. На основе же регрессионного анализа мы можем построить так называемую регрессионную модель в виде зависимости

$$y = \varphi(x_1, x_2),$$

где y — среднее значение шума Y в зависимости от значений x_1 и x_2 случайных величин X_1 и X_2. Наличие такой модели

(которую строят, опираясь на результаты имеющихся *статистических данных* — результатов эксплуатации автомобилей) позволяет в дальнейшем выбрать наилучший режим эксплуатации и решать многие другие задачи.

1.3. Предварительная обработка результатов эксперимента

Прежде чем перейти к детальному анализу полученных в результате проведенного эксперимента *статистических данных*, обычно проводят их предварительную обработку. Иногда результаты такой обработки уже сами по себе дают ответы на многие вопросы. Но в большинстве случаев они служат исходным материалом для дальнейшего анализа.

Вариационный ряд. Одним из самых простых преобразований статистических данных является их упорядочивание по величине. Пусть $(x_1, \ldots, x_n)$ — *выборка объема* n из *генеральной совокупности* X. Ее можно упорядочить, расположив значения в неубывающем порядке:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(i)} \leqslant \ldots \leqslant x_{(n)}, \tag{1.2}$$

где $x_{(1)}$ — наименьший, $x_{(n)}$ — наибольший из *элементов выборки*.

Определение 1.1. Последовательность чисел

$$x_{(1)}, \quad x_{(2)}, \quad \ldots, x_{(i)}, \quad \ldots, \quad x_{(n)},$$

удовлетворяющих условию (1.2), называют ***вариационным рядом выборки***, или, для краткости, просто ***вариационным рядом***; число $x_{(i)}$, $i = \overline{1, n}$, называют i-м ***членом вариационного ряда***.

Обозначим $X_{(i)}$, $i = 1, n$, случайную величину, которая при каждой *реализации случайной выборки* $\vec{X}_n$ принимает значение, равное i-му члену вариационного ряда.

Определение 1.2. Последовательность случайных величин

$$X_{(1)}, \quad X_{(2)}, \quad \ldots, \quad X_{(i)}, \quad \ldots, \quad X_{(n)}$$

называют ***вариационным рядом случайной выборки.*** При этом $X_{(i)}$, $i = \overline{1, n}$, называют i-м ***членом вариационного ряда случайной выборки.***

Переход от случайной выборки $\vec{X}_n$ к ее вариационному ряду не приводит к потере информации, содержащейся в случайной выборке, поскольку их совместная функция распределения (1.1) остается одной и той же. Однако функция распределения каждой случайной величины $X_{(i)}$, $i = \overline{1, n}$, уже не совпадает с функцией распределения $F(x)$ генеральной совокупности X, хотя и может быть через нее выражена. Например, можно показать (см. пример 2.20), что для ***крайних членов вариационного ряда*** случайной выборки $X_{(1)}$ и $X_{(n)}$ их функции распределения имеют вид

$$\mathbf{P}\{X_{(1)} < x\} = 1 - \big(1 - F(x)\big)^n$$

и

$$\mathbf{P}\{X_{(n)} < x\} = F^n(x).$$

Эти соотношения позволяют находить неизвестную функцию распределения $F(x)$ генеральной совокупности X, имея в эксперименте лишь результаты измерений либо величины $X_{(1)}$, либо $X_{(n)}$.

Пример 1.3. В результате пяти повторных независимых наблюдений некоторой случайной величины X (например, X — давление в газовом баллоне, измеряемое в мегапаскалях) полу-