Общие сведения

Теория производит тем большее впечатление, чем проще её предпосылки, чем разнообразнее предметы, которые она связывает, и чем шире область её приложения

Альберт Эйнштейн

Задачи математической статистики

Математическая статистика — это прикладная математическая дисциплина, базирующаяся на понятиях и методах *теории вероятностей*, имеющая, однако, свои задачи и методы.

Пусть исследуются исходы некоторого опыта, причем вероятности исходов известны. Задача теории вероятностей состоит в разработке методов нахождения вероятностей различных сложных событий, исходя из известных вероятностей более простых событий.

Теория вероятностей, другими словами, занимается разработкой и исследованием *вероятностных моделей* случайных экспериментов.

На практике вероятности элементарных событий (или законы распределения случайных величин) редко бывают известны. Часто известно лишь то, что опыт можно описать в рамках какой-либо *вероятностно-статистической модели*, имеющей некоторую неопределенность в задании вероятности P событий или закона распределения случайных величин.

Задача математической статистики состоит в том, чтобы уменьшить эту неопределенность (восстановить закон распределения исследуемой случайной величины), используя информацию, полученную из эксперимента (*статистические данные*).

В определенном смысле, математическая статистика решает задачи, *обратные теории вероятностей*: она уточняет структуру статистических моделей по результатам проводимых наблюдений.

Математическая статистика является также наукой о *статистических выводах*: зачастую на основании статистических данных нам приходится делать выбор одного из нескольких, противоречащих друг другу, предположений (*гипотез*) относительно законов распределения случайных величин или о значениях параметров распределений.

В силу своего прикладного характера, математическая статистика занимается также *разработкой методов получения, описания и обработки опытных данных для изучения закономерностей случайных массовых явлений.*

Особенность идей и методов математической статистики — универсальность, возможность использования в различных приложениях.

Рассмотрим некоторые конкретные задачи, решаемые математической статистикой.

- Оценка на основании измерений неизвестной функции распределения.

  Дано: множество значений случайной величины

  $$X = x_i,\ i\ =\ 1,\ \ldots,\ n.$$

  Найти функцию распределения случайной величины X.

- Оценка неизвестных параметров распределения.

  Дано: случайная величина X имеет функцию распределения вида

  $$F(x, \vartheta_1, \vartheta_2, \ldots, \vartheta_n),$$

  где ϑ_1, ϑ_2, $\ldots$, ϑ_n — неизвестные параметры.

  Найти оценки этих параметров.

- Проверка статистических гипотез.

  Например, гипотеза о виде распределения:

  Дано: Предполагаем, что функция распределения случайной величины есть $F(x)$. Имеем данные

  $$X:\ x_{1,}x_{2,}\ldots,x_N.$$

  Спрашивается: совместимы ли значения X с гипотезой о том, что случайная величина имеет распределение $F(x)$?

Глава 1.

Основные понятия математической статистики

1.1. Первичные данные и их представление

Пусть исследуется некоторая совокупность объектов, каждому из которых ставится в соответствие некоторая числовая функция — случайная величина X, распределенная по некоторому неизвестному закону $L(\xi)$. Множество (конечное или бесконечное) всех объектов (всех значений случайной величины X) называют *генеральной совокупностью*.

На практике мы имеем дело с конечным набором данных, полученным в результате проведения n измерений или наблюдений. Эту конечную совокупность экспериментальных данных

$$x : x_1, x_2, \ldots, x_n \tag{1.1.}$$

называют *выборкой* объема n из генеральной совокупности.

Выборка является первичной формой представления экспериментального материала.

Каждой реализации $x: \ x_1, x_2, \ldots, x_n$ можно поставить в соответствие упорядоченную последовательность [1]

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

которую называют *вариационным рядом* выборки.

$x_{(i)}$ — порядковые *статистики*, предполагается, что все они *различны*. [2]

$x_{(1)}$, $x_{(k)}$ — экстремальные значения выборки.

Если в выборке есть повторяющиеся значения ($k < n$), то выборка представляется в виде *статистического ряда*– это таблица, в которой указаны все различные значения (варианты) вариационного ряда и числа, показывающие их количество в выборке.

---

[1] То есть мы располагаем значения $\mathbf{x}$ выборки в порядке возрастания.

[2] Статистикой называют любую функцию от выборки не содержащую неизвестных параметров.

x_i	x_1	x_2	...	x_k
m_i	m_1	m_2	...	m_k

Таблица 1.1. Статистический ряд

Представление выборки в виде статистического ряда естественно для дискретных распределения генеральной совокупности. Если генеральная совокупность имеет непрерывный закон распределения, то представление выборок в виде статистического ряда обусловлено двумя причинами:

- округлением результатов – ограничением числа верных значащих цифр в представлении результатов измерений, в результате чего в выборке неизбежно появляются повторяющиеся значения;
- большим объемом выборок, что вызывает необходимость предварительной группировки данных.

При группировке данных область значений случайной величины (в выборке) разбивается на k непересекающихся интервалов, необязательно равной длины, подсчитывается число элементов выборки, попавших в каждый интервал. В качестве значения, представляющего каждый интервал берется либо (чаще всего) середина интервала, либо среднее арифметическое значений точек, принадлежащих данному интервалу. Количество точек, попавших в интервал служит *весом* представительской точки в полученном таким образом *статистическом ряде*. В таблице группированных данных часто указываются не сами представительные точки, а границы соответствующих интервалов.

1.2. Математическая модель выборки. Эмпирическая функция распределения

Если мы повторим еще раз серию из n экспериментов, то получим новый набор чисел $x_1^{\langle 2\rangle}, x_2^{\langle 2\rangle}, \ldots, x_n^{\langle 2\rangle}$. Повторяя раз за разом эти серии экспериментов, мы всякий раз будем получать новые наборы чисел $\{x_i^{\langle k\rangle}\}$, т.е. «результат i-го измерения в $k-$й серии» рассматривается как реализация случайной величины с тем же законом распределения, что и исходная случайная величина.

Таким образом, выборку (до опыта!) можно интерпретировать как *систему n независимых, одинаково распределенных случайных величин.*

Плотность распределения системы случайных величин для непрерывного распределения имеет вид:

$$\begin{aligned} L(X,\Theta) =& f(X_1,\Theta)\cdot f(X_2,\Theta)\cdot\ldots\cdot f(X_n,\Theta) = \\ =& \prod_{i=1}^{n} f(X_i,\Theta). \end{aligned} \tag{1.3.}$$

Здесь Θ – вектор параметров распределения.

После опыта выборка рассматривается как *реализация* либо одной случайной величины, либо – как *реализацию* случайного вектора $X = X_1, X_2, \ldots, X_n$, где X_i— независимые, одинаково распределенные случайные величины. [3]

Определим для каждого действительного x случайную величину $\mu_n(x)$, равную числу элементов выборки, значения которых меньше x:

$$\mu_n(x) = \sum_{i=1}^{n} I(x_i < x), \tag{1.4.}$$

где $I(A)$ — индикатор события A [4] .

Положим

$$F_n(x) = \frac{\mu_n(x)}{n}. \tag{1.5.}$$

Функция (1.5.) называется *эмпирической* (опытной, статистической) *функцией распределения* (ЭФР), соответствующей выборке X.

По своему определению ЭФР — случайная величина, принимающая дискретные значения $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1$.

Поскольку

$$P\left(F_n(x) = \frac{k}{n}\right) = P(\mu_n(x) = k),$$

то, как следует из определения $\mu_n(x)$, она подчиняется биномиальному распределению с параметром

$$p = \mathbf{P}(\xi < x) = F(x).$$

Итак, ЭФР (как и вариационный ряд) — некоторая сводная характеристика выборки. Для каждой реализации x выборки X функция $F_n(x)$ однозначно определена и обладает всеми свойствами функции распределения: изменяется от 0 до 1, не убывает и *непрерывна слева*. При этом она кусочно-постоянна и возрастает только в точках последовательности (1.5.). Если в вариационном ряду (1.5.) нет одинаковых значений, то

$$F_n(x) = \begin{cases} 0, & \text{при } x \leqslant x_{(1)}, \\ k/n & \text{при } x_{(k)} < x \leqslant x_{(k+1)},\ k = 1, \ldots, n-1, \\ 1 & \text{при } x > x_n \end{cases}$$

т. е. в этом случае величина всех скачков равна $1/n$ и типичный график функции $F_n(x)$ имеет вид, изображенный на рис. 1.1.

[3] Предполагается. что существует, по крайней мере гипотетически, возможность неограниченное число раз воспроизводить серии независимых испытаний.

[4] $I(A)$ равен единице, если событие A имеет место, и равен нулю в противном случае.

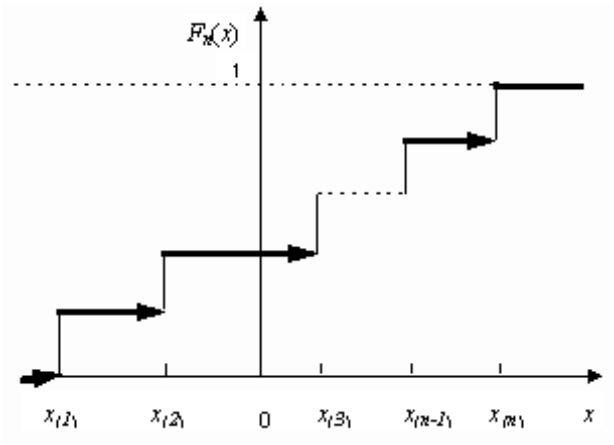

Рис. 1.1.

Согласно *теореме Бернулли*, ЭФР $F_n(x)$ при $n \to \infty$ сходится по вероятности к теоретической функции распределения $F(x)$.

Замечание. Процедуру получения выборки можно представить как выбор с возвращением из урны (*генеральной совокупности*) шаров (значений случайной величины). Чтобы выборка правильно представляла распределение генеральной совокупности, необходимо обеспечить случайность выборки, т.е., чтобы вероятность выбора любого элемента из генеральной совокупности была одинакова. В этом случае говорят. что выборка должна быть *репрезентативной*, т.е. представительной.

1.3. Гистограмма и полигон

Кроме эмпирической функции распределения существуют и другие способы наглядного представления статистических данных. Так, если наблюдаемая случайная величина принимает дискретные значения $a_1, a_2, \ldots$, то более наглядное представление о законе распределения случайной величины ξ дадут частоты $\dfrac{\nu_r}{n}$, где ν_r — число элементов выборки $X = (X_1, \ldots, X_n)$, принявших значение a_r: $\nu_r = \sum\limits_{j=1}^{n} I\left(X_j = a_r\right)$. В этом случае, по теореме Бернулли, частоты$\dfrac{\nu_r}{n}$ сходятся по вероятности к вероятностям соответствующих событий $P(\xi) = a_r$.

Если случайная величина непрерывна, то данную методику приспосабливают для оценивания неизвестной плотности распределения следующим образом: область возможных значений ξ разбиваем точками на k непересекающихся интервалов; подсчитываем число точек m_r, попавших в каждый $r-$й интервал, вероятность попадания в некоторый $r-$й интервал оцениваем величиной $\dfrac{m_r}{n}$.

С другой стороны, эту же вероятность можно выразить через интеграл от плотности по данному интервалу ε_r:

$$\frac{m_r}{n} \approx \int\limits_{\varepsilon_r} f(x)\,dx \approx |\varepsilon_r| \cdot f(x_r),$$

где x_r — некоторая внутренняя точка интервала, [5] в качестве которой можно взять середину интервала.

Отсюда мы получаем оценку плотности распределения:

$$f_n(x_r) = \frac{m_r}{n \cdot |\varepsilon_r|}. \tag{1.6.}$$

График кусочно-постоянной функции ((1.6.)) называется *гистограммой* (см. рис. 1.2). Если соединим точки $\mathrm{M}(x_r, f_n(x_r))$ отрезками прямых линий, то получим кусочно-линейный график, также являющийся статистическим аналогом плотности распределения, который называется *полигон частот*.

Более плавную кривую можно получить, используя *локальную аппроксимацию кривой полиномами* или *сглаживание результатов*.

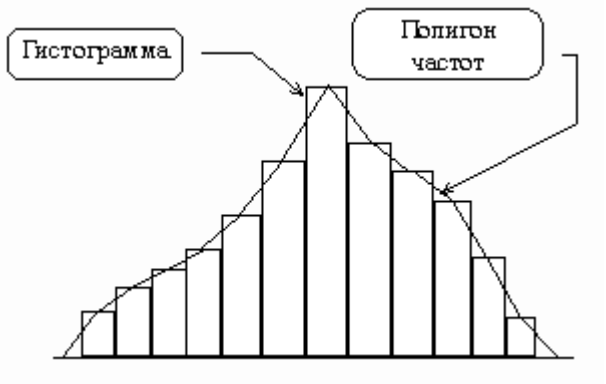

Рис. 1.2. Гистограмма и полигон частот

Замечание. При группировке данных приходится учитывать два взаимно исключающих обстоятельства. С одной стороны, число интервалов группировки должно быть достаточно велико, чтобы детально описать поведение плотности распределения, но, с другой стороны, число точек, попадающих в интервал, также должно быть достаточным, чтобы надежно представлять данный интервал. Если интервалов будет много. то некоторые из них могут оказаться пустыми, а плотность распределения – изрезанной, многолепестковой.

В литературе имеется много рекомендаций по выбору оптимального числа интервалов группировки. Приведем только две формулы:

[5] Здесь мы применили теорему о среднем.

Формула Старджеса:

$$K = 1 + \lfloor \log_2 n \rfloor, \tag{1.7.}$$

где $\lfloor x \rfloor$ есть целая часть, не превосходящая x.

Формула Брукса и Каррузера: $K = \lfloor \sqrt{n} \rfloor$.

При малых выборках ($n \leq 20$) некоторые интервалы могут оказаться пустыми. В таком случае надо уменьшить их количество. Считается приемлемым выбирать число интервалов так, чтобы выполнялось $n_j \leq 5$ (оптимисты допускают $n_j \leq 3$) для каждого интервала.