Тема 1. Основные определения и задачи математической статистики.

1. Основные задачи математической статистики.

Математическая статистика – это раздел математики, в котором рассматриваются задачи восстановления «вероятностной структуры» исследуемых объектов, явлений или процессов на основе ряда наблюдений над этими объектами, явлениями или процессами.

Пусть, например, имеется монета с неизвестной вероятностью выпадения герба p. В данном случае вероятность p определяет «вероятностную структуру» монеты. Представим, что с данной монетой выполняется эксперимент, который заключается в том, что монета подбрасывается 100 раз и по результатам бросаний подсчитывается количество выпавших гербов. Предположим, что при выполнении эксперимента герб выпал 74 раза. Что можно сказать о монете и неизвестной вероятности p? Из проведенного над монетой наблюдения следует, что монета, скорее всего, не является симметричной, а неизвестная вероятность p больше, чем 0.5. Сделанные выводы являются интуитивными и неформальными, в то время как математическая статистика располагает формальными и эффективными методами решения целого ряда практических задач в достаточно общей постановке.

Применительно к рассматриваемому примеру с монетой методы математической статистики позволяют решить следующие основные задачи:

1) построить оценку (найти метод приближенного вычисления) неизвестной вероятности p (задача построения точечной оценки);

2) построить интервал (p_1, p_2), в котором с большой вероятностью находится вероятность p (задача построения доверительного интервала);

3) определить можно ли с большой долей уверенности считать, что монета является симметричной, то есть, считать, что вероятность $p = 0.5$ (задача проверки статистической гипотезы);

4) определить какое из двух утверждений «$p = 0.5$» либо «$p = 0.7$» является «более правдоподобным» (задача различения двух простых гипотез).

2. Основные определения.

В задачах математической статистики основу исходных данных образует серия наблюдений, которая в простых задачах может иметь вид вектора или последовательности случайных величин:

$$\xi_1(\omega),\ \xi_2(\omega),\ \ldots,\ \xi_n(\omega),\ \ldots$$

где ω – элементарное событие (исход) из множества всех элементарных событий Ω.

В наиболее простых и хорошо изученных задачах серия наблюдений представляет собой выборку [Боровков].

Определение 1.1.

Выборкой называется вектор случайных величин $(\xi_1,\ldots,\xi_n)$, в котором:

1) все случайные величины $\xi_1, \ldots, \xi_n$ независимы в совокупности,

2) все случайные величины $\xi_1, \ldots, \xi_n$ имеют одинаковую функцию распределения.

Если функция распределения $F_\xi(x)$, участвующая в определении выборки, известна (либо известен параметрический вид функции распределения $F_\xi(x\,|\,\theta)$), то коротко говорят «$(\xi_1,\ldots,\xi_n)$ – выборка из распределения $F_\xi(x)$». Поскольку всякая функция распределения $F_\xi(x)$ задает некоторую случайную величину ξ, то иногда говорят «$(\xi_1,\ldots,\xi_n)$ – выборка из распределения случайной величины ξ».

Определение 1.2.

Число n в определении выборки 1.1 называется *объемом выборки*.

Проведение статистического эксперимента фактически эквивалентно фиксированию некоторого элементарного события $\omega^* \in \Omega$ из множества всех элементарных событий Ω. Формально, каждая случайная величина ξ_i выборки является функцией $\xi_i(\omega)$ определенной на множестве элементарных исходов Ω, поэтому при фиксированном ω^* случайные величины $\xi_1(\omega)$, …, $\xi_n(\omega)$ выборки принимают определенные числовые значения $x_i = \xi_i(\omega^*)$ $(i = \overline{1,n})$.

Определение 1.3.

Реализациями выборки называются числовые векторы $(x_1,\dots, x_n)$, такие что:

$$(x_1, x_2,\dots, x_n) = (\xi_1(\omega^*), \xi_2(\omega^*),\dots, \xi_n(\omega^*)),$$

для некоторого $\omega^* \in \Omega$.

С каждой выборкой непосредственно связаны две основных характеристики: вариационный ряд и эмпирическая функция распределения.

Рассмотрим определение вариационного ряда. Представим, что проведен эксперимент, в результате которого реализовалось элементарное событие $\omega^* \in \Omega$ и функции $\xi_1(\omega)$, …, $\xi_n(\omega)$ приняли определенные числовые значения x_1, …, x_n:

$$x_i = \xi_i(\omega^*),\ i = \overline{1,n}.$$

Числа x_1, …, x_n могут быть упорядочены по возрастанию следующим образом: сперва найдем наименьшее из всех чисел $\{x_1,\dots, x_n\}$ и обозначим его $x_{(1)}$. Затем найденное число $x_{(1)}$ исключим из множества всех чисел $\{x_1,\dots, x_n\}$, найдем наименьшее из оставшихся чисел $\{x_1,\dots, x_n\} \setminus \{x_{(1)}\}$ и обозначим его $x_{(2)}$. Далее аналогичным образом продолжим процедуру упорядочивания и нахождения чисел $x_{(i)}$ до тех пор, пока не будет найдено число $x_{(n)}$, являющееся наибольшим из всех чисел $\{x_1,\dots, x_n\}$. Заметим, что подобная процедура упорядочивания и нахождения чисел $x_{(1)}$, …, $x_{(n)}$ может быть проделана при каждом фиксированном ω^*. Таким образом, при каждом $\omega \in \Omega$ однозначно определены все числа $x_{(1)}$, …, $x_{(n)}$ и имеет смысл определение функций $\xi_{(1)}(\omega)$, …, $\xi_{(n)}(\omega)$ таких, что каждая функция $\xi_{(i)}$ при фиксированном ω равна числовому значению $x_{(i)}$, полученному в процессе упорядочивания чисел x_1, …, x_n:

$$\xi_{(i)}(\omega) = x_{(i)},\ i = \overline{1,n}.$$

Определение 1.4.

Вариационным рядом выборки $\xi = (\xi_1,\dots, \xi_n)$ называется упорядоченная совокупность случайных величин $\xi_{(1)}, \xi_{(2)}, \dots, \xi_{(n)}$:

$$\xi_{(1)} \le \xi_{(2)} \le \dots \le \xi_{(n)},$$

в которой $\xi_{(1)}$ принимает наименьшее значение из ξ_1, …, ξ_n, $\xi_{(2)}$ – значение следующее по величине за $\xi_{(1)}$, и так далее, а $\xi_{(n)}$ – наибольшее из значений ξ_1, …, ξ_n.

Определение 1.5.

Случайные величины $\xi_{(1)}$, …, $\xi_{(n)}$ вариационного ряда называются *порядковыми статистиками*. Случайная величина $\xi_{(k)}$ называется *k-ой порядковой статистикой*. Случайные величины $\xi_{(1)}$ и $\xi_{(n)}$ называются *экстремальными значениями выборки*.

Фактически, простое аналитическое выражение имеют только экстремальные значения выборки:

$$\xi_{(1)} = \min\{\xi_1,\dots,\xi_n\},\ \xi_{(n)} = \max\{\xi_1,\dots,\xi_n\}.$$

Рассмотрим определение эмпирической функции распределения. Согласно определению выборки (определение 1.1) все случайные величины выборки ξ_i имеют одинаковую функцию распределения, которую обозначим $F_\xi(x)$. Функция $F_\xi(x)$ играет очень важную роль при решении многих задач, поскольку полностью определяет совместное распределение величин выборки $(\xi_1,..., \xi_n)$. Однако функция распределения $F_\xi(x)$ в большинстве задач либо неизвестна, либо известна с точность до параметра (известен параметрический вид функции распределения $F_\xi(x)$). Отсюда происходит вполне естественное стремление на основе имеющейся выборки $(\xi_1,..., \xi_n)$ для неизвестной функции $F_\xi(x)$ построить известное «приближение», которое затем использовать при решении задач. Таким «приближением» в математической статистике является эмпирическая функция распределения $F_\xi^*(x;\xi_1,..., \xi_n)$.

Пусть вектор $(\xi_1(\omega),..., \xi_n(\omega))$ является выборкой, определим случайную функцию $\mu_n(x;\xi_1,..., \xi_n)$ так, что при фиксированных x и ω функция $\mu_n(x;\xi_1,..., \xi_n)$ равна количеству значений из $\xi_1(\omega)$, …, $\xi_n(\omega)$ меньших x:

$$\mu_n(x;\xi_1,..., \xi_n) = \left|\{ j : \xi_j(\omega) < x\}\right|.$$

Заметим, что при каждом фиксированном x величина $\mu_n(x;\xi_1,..., \xi_n)$ является случайной, поскольку сами величины $\xi_1(\omega)$, …, $\xi_n(\omega)$ являются случайными и, следовательно, количество тех значений среди них, которые окажутся меньше x, также случайно.

Определение 1.6.

Эмпирической функцией распределения называется случайная функция $F_n^*(x;\xi_1,..., \xi_n)$:

$$F_n^*(x;\xi_1,..., \xi_n) = \frac{\mu_n(x;\xi_1,..., \xi_n)}{n},$$

где функция $\mu_n(x;\xi_1,..., \xi_n)$ равна количеству случайных величин выборки $(\xi_1(\omega),..., \xi_n(\omega))$ меньших x.

Представление о том насколько «хорошим» приближением к функции $F_\xi(x)$ является эмпирическая функция распределения $F_n^*(x;\xi_1,..., \xi_n)$ дают следующие теоремы.

Теорема (сходимость по вероятности)

Пусть $F_n^*(x;\xi_1,..., \xi_n)$ является эмпирической функцией распределения, построенной по выборке $(\xi_1,..., \xi_n)$ из распределения $F_\xi(x)$, тогда при всяком фиксированном x^* случайная величина $F_n^*(x^*;\xi_1,..., \xi_n)$ сходится по вероятности к $F_\xi(x)$ при $n \to \infty$:

$$F_n^*(x^*;\xi_1,..., \xi_n) \xrightarrow{P} F_\xi(x^*),\ \text{при}\ n \to \infty.$$

Теорема (равномерная сходимость по вероятности)

Пусть $F_n^*(x;\xi_1,..., \xi_n)$ является эмпирической функцией распределения, построенной по выборке $(\xi_1,..., \xi_n)$ из распределения $F_\xi(x)$, тогда последовательность случайных величин $\sup\limits_{-\infty<x<\infty} \left|F_n^*(x;\xi_1,..., \xi_n) - F_\xi(x)\right|$ сходится к нулю по вероятности при $n \to \infty$:

$$\sup_{-\infty<x<\infty} \left|F_n^*(x;\xi_1,..., \xi_n) - F_\xi(x)\right| \xrightarrow{P} 0,\ \text{при}\ n \to \infty.$$

Теорема (Гливенко, сходимость с вероятностью 1)

Пусть $F_n^*(x;\xi_1,..., \xi_n)$ является эмпирической функцией распределения, построенной по выборке $(\xi_1,..., \xi_n)$ из распределения $F_\xi(x)$, тогда последовательность случайных величин

$\sup\limits_{-\infty<x<\infty}\left|F_n^*(x;\xi_1,...,\xi_n)-F_\xi(x)\right|$ сходится к нулю с вероятностью 1 («почти наверное») при $n\to\infty$:

$$\sup_{-\infty<x<\infty}\left|F_n^*(x;\xi_1,...,\xi_n)-F_\xi(x)\right|\xrightarrow{п.н.}0,\ \text{при } n\to\infty.$$

На практике в результате проведения эксперимента будет получен вектор числовых значений $(x_1,..., x_n)$ (в результате проведения эксперимента происходит элементарное событие ω^*, так что все случайные величины выборки $\xi_i(\omega)$ принимают определенные числовые значения $x_i=\xi_i(\omega^*)$). Функция $\mu_n(x;x_1,..., x_n)$ оказывается равной количеству числовых значений в векторе $(x_1,..., x_n)$ меньших заданного x, а эмпирическая функция распределения $F_n^*(x;x_1,...,x_n)$ равна количеству значений меньших x, отнесенному к общему количеству числовых значений n. Типичный график реализации функции $F_n^*(x;x_1,...,x_n)$ как функции переменной x, представлен на рисунке 1.1.

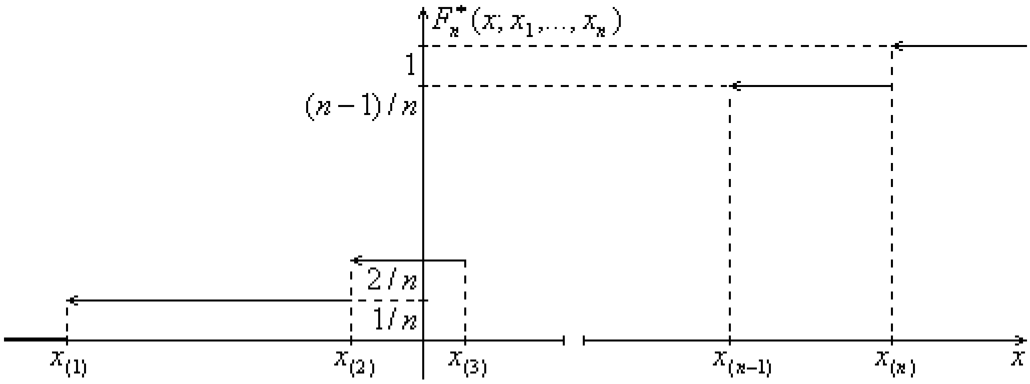

Рисунок 1.1. График реализации эмпирической функции распределения $F_n^*(x;x_1,...,x_n)$ *как функции* x.