

Лекция 11. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Основная цель теории вероятностей - исследование имеющейся вероятностной модели. Математическая статистика ставит перед собой обратную задачу – построение вероятностно-статистической модели по результатам наблюдений.

Из истории математической статистики. Как наука математическая статистика возникла в XXв., но отдельные задачи рассматривались еще в XVIIв. Математическая статистика развивалась параллельно с теорией вероятностей. Дальнейшее развитие математической статистики (вторая половина XIX – начало XX вв.) связано с именами Чебышева, Маркова, Ляпунова, а также Гаусса, Кетле, Гальтона, К. Пирсона и др. В XXв. большой вклад в развитие математической статистики внесли советские математики Романовский, Слуцкий, Колмогоров, Смирнов, а также английские Стьюдент, Р. Фишер, Э. Пирсон и американские Нейман, Вальд.

Основные задачи математической статистики. Основная задача математической статистики – получение выводов о массовых явлениях и процессах по данным наблюдений над ними или экспериментов. Эти выводы относятся не к отдельным испытаниям, а являются утверждениями об общих характеристиках изучаемого явления – вероятностях, законах распределения и их параметрах и др. – в предположении постоянства условий, порождающих явление.

Содержание математической статистики составляет *разработка приемов статистического наблюдения и анализа статистических данных*. Исходный материал для статистического исследования реального явления - набор результатов наблюдений над ним или результатов специальных испытаний. Укажем некоторые основные вопросы, которые при этом возникают.

1. Оценка неизвестной вероятности случайного события.
2. Определение неизвестной функции распределения. В результате n независимых испытаний над случайной величиной ξ получены ее значения $x_1, x_2, \dots, x_n$. Требуется определить, хотя бы приближенно, неизвестную функцию распределения $F_\xi(x)$ величины ξ .
3. Определение неизвестных параметров распределения. Случайная величина ξ имеет функцию распределения определенного вида, зависящую от k параметров, значения которых неизвестны. На основании последовательных наблюдений величины ξ нужно найти значения этих параметров.
4. Проверка статистических гипотез. На основании некоторых соображений можно считать, что функция распределения случайной величины ξ есть $F_\xi(x)$; спрашивается, совместимы ли наблюдаемые значения с гипотезой, что ξ действительно имеет распределение $F_\xi(x)$?

Задачами 1-4 не исчерпываются основные проблемы математической статистики. Современную математическую статистику определяют как *науку о принятии решений в условиях неопределенности*.

Генеральная совокупность и выборка. В практике статистических наблюдений различают два основных вида наблюдений: сплошное, когда изучаются все элементы интересующей исследователя совокупности, и выборочное, когда изучается часть элементов. Вся подлежащая изучению совокупность называется *генеральной совокупностью*. Та часть объектов, которая отобрана для непосредственного изучения из генеральной совокупности, называется *выборкой*.

В математической статистике генеральная совокупность - это *совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном*

комплексе условий и аналогично понятию случайной величины, так как полностью обусловлено комплексом условий.

О п р е д е л е н и е. Все множество значений подлежащей исследованию случайной величины называется *генеральной совокупностью*.

О п р е д е л е н и е. Последовательность наблюдений $x_1, x_2, \dots, x_n$ называется (*случайной*) *выборкой объема n*, если $x_1, x_2, \dots, x_n$ получены как независимые реализации некоторой случайной величины ξ с функцией распределения $F_\xi(x)$.

При этом также говорят, что $x_1, x_2, \dots, x_n$ есть выборка из генеральной совокупности ξ . С теоретико-вероятностной точки зрения случайная выборка $x_1, x_2, \dots, x_n$ может рассматриваться как последовательность независимых случайных величин, имеющих одну и ту же функцию распределения $F_\xi(x)$.

О п р е д е л е н и е. Последовательность выборочных значений, записанных в порядке возрастания, называется *вариационным рядом*.

Пусть из генеральной совокупности извлечена выборка объема n :

$$\underbrace{Z_1, Z_1, \dots, Z_1}_{n_1 \text{ раз}}; \underbrace{Z_2, Z_2, \dots, Z_2}_{n_2 \text{ раз}}; \dots; \underbrace{Z_k, Z_k, \dots, Z_k}_{n_k \text{ раз}} \quad (n_1 + n_2 + \dots + n_k = n).$$

О п р е д е л е н и е. Различные выборочные значения $z_1, z_2, \dots, z_k$ называются *вариантами*, числа $n_1, n_2, \dots, n_k$ – их *частотами* (сумма частот равна объему выборки), а числа $w_1 = n_1/n, w_2 = n_2/n, \dots, w_k = n_k/n$ – их *относительными частотами* (сумма относительных частот равна единице).

z_i	z_1	z_2	$\dots$	z_k
n_i	n_1	n_2	$\dots$	n_k

называется *статистическим рядом*.

Иногда статистическим рядом (или рядом относительных частот) называют таблицу вида

z_i	z_1	z_2	$\dots$	z_k
w_i	w_1	w_2	$\dots$	w_k

З а м е ч а н и е. Последний статистический ряд в математической статистике – аналог ряда распределения дискретной случайной величины генеральной совокупности), где вместо вероятностей p_i значений случайной величины стоят относительные частоты w_i вариантов.

О п р е д е л е н и е. Функцию $F_n(x) = m/n$, где n – объем выборки, а m – число значений x_i в выборке, не превосходящих x , называют *эмпирической функцией распределения*.

В отличие от $F_n(x)$, полученной на основании выборочных значений генеральной совокупности, функцию распределения всей генеральной совокупности $F_\xi(x)$ называют *теоретической функцией распределения*. Различие этих функций в том, что $F_\xi(x)$ есть вероятность события $\{\xi \leq x\}$, а $F_n(x)$ – относительная частота того же события. Из теоремы Бернулли следует, что при больших n числа $F_n(x)$ и $F_\xi(x)$ мало отличаются друг от друга в том смысле, что для любого положительного числа ε выполняется: $P\{|F_n(x) - F_\xi(x)| < \varepsilon\} \rightarrow 1$ при $n \rightarrow \infty$. Это подтверждается и тем, что эмпирическая функция распределения обладает свойствами, аналогичными свойствам теоретической. Некоторые из них:

- 1) $0 \leq F_n(x) \leq 1$.
- 2) $F_n(x)$ – неубывающая функция, ее график имеет ступенчатый вид.
- 3) Если z_1 – наименьшая варианта, то $F_n(x) = 0$ при $x < z_1$; если z_k – наибольшая варианта, то $F_n(x) = 1$ при $x \geq z_k$.

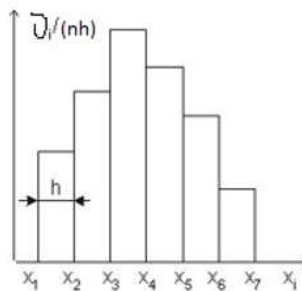
Одно из основных понятий математической статистики – оценка, или статистика.

О п р е д е л е н и е. Пусть $x_1, x_2, \dots, x_n$ – выборка. Любая функция от выборочных значений $x_1, x_2, \dots, x_n$ называется *оценкой* (той или иной характеристики генеральной совокупности).

Итак, эмпирическая функция распределения $F_n(x)$ может служить оценкой по выборке для (неизвестной) теоретической функции распределения $F_\xi(x)$ генеральной совокупности.

Если изучаемая случайная величина ξ (генеральная совокупность) непрерывна, то для оценки ее неизвестной плотности вероятности $p_\xi(x)$ целесообразно использовать т.н. *гистограмму*, которая строится следующим образом.

- 1) Упорядоченная по возрастанию выборка разбивается на интервалы (их число N можно определить, напр., по формуле Стерджесса $N \approx 1 + 3.322 \lg n$, где n - объем выборки, эта формула дает хорошие результаты, если n достаточно велико, распределение близко к нормальному, и используются равные интервалы).
- 2) Длина h каждого интервала (в случае их одинаковых длин) находится как $h = (x_{\max} - x_{\min})/N$, где $x_{\max}$ ($x_{\min}$) - наибольшее (наименьшее) выборочное значение.
- 3) По оси x откладываются границы интервалов $x_{\min}, x_{\min}+h; x_{\min}+2h; \dots; x_{\min}+Nh=x_{\max}$, а над самими интервалами строятся прямоугольники так, чтобы площадь каждого из них представляла собой относительную частоту распределения: $h l_i = n_i/n$, где l_i - высота i -го прямоугольника, n_i - частота попадания выборочных значений в i -й интервал (если выборочное значение попадает на границу i -го и $(i+1)$ -го интервалов, то считается, что оно принадлежит $(i+1)$ -му интервалу. Имеем: $l_i = n_i/(nh)$).



Заметим (аналог нормировки плотности: $\int_{-\infty}^{+\infty} p_\xi(x) dx = 1$), что общая площадь прямоугольников составляет 100% распределения: $S = \sum_{i=1}^N h n_i / (nh) = 1$.

Выборочное среднее и выборочная дисперсия. Пусть имеется (повторная) выборка объема n из генеральной совокупности, т.е. N независимых одинаково распределенных случайных величин $x_1, x_2, \dots, x_n$ с функцией распределения $F_\xi(x)$.

О п р е д е л е н и е. *Выборочным средним* называется число $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (для статистического ряда $\bar{x} = \frac{1}{n} \sum_{i=1}^k z_i n_i$).

Запишем последнюю формулу в виде $\bar{x} = z_1 n_1/n + z_2 n_2/n + \dots + z_k n_k/n = z_1 w_1 + z_2 w_2 + \dots + z_k w_k$. При больших n (в силу теоремы Бернулли) относительные частоты значений z_i близки к их вероятностям p_i : $w_i \approx p_i$, и мы получаем: $\bar{x} \approx z_1 p_1 + z_2 p_2 + \dots + z_k p_k = M\xi$, где ξ - (дискретная) случайная величина, представленная выборкой. Итак, в качестве оценки математического ожидания $M\xi$ генеральной совокупности можно взять выборочное среднее $\bar{x}$.

Выборочное среднее, найденное по данным одной выборки, есть определенное число. Если извлекать другие выборки того же объема из генеральной совокупности, то выборочное среднее будет меняться от выборки к выборке. Таким образом, выборочное среднее можно рассматривать как случайную величину, а следовательно, говорить о ее распределении (теоретическом и эмпирическом) и их числовых характеристиках.

В качестве оценки для дисперсии генеральной совокупности по выборочным данным берут т. н. выборочную дисперсию.

О п р е д е л е н и е. *Выборочной дисперсией* называется число $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Если выборка записана в виде статистического ряда, то выборочная дисперсия вычисляется как $s^2 = \frac{1}{n} \sum_{i=1}^k (z_i - \bar{x})^2 n_i$

Эта формула – аналог теоретико-вероятностной формулы для вычисления дисперсии (дискретной) случайной величины.

Часто бывает легче вычислять выборочную дисперсию по формуле $s^2 = \overline{x^2} - \bar{x}^2$. Действительно, $s^2 = \frac{1}{n} \sum_{i=1}^k (z_i - \bar{x})^2 n_i = \frac{1}{n} \sum_{i=1}^k z_i^2 n_i - 2 \bar{x} \frac{1}{n} \sum_{i=1}^k z_i n_i + \bar{x}^2 \frac{1}{n} \sum_{i=1}^k n_i = \overline{x^2} - 2 \bar{x}^2 + \bar{x}^2 = \overline{x^2} - \bar{x}^2$.

З а м е ч а н и е. Обычно (ниже будет показано, почему) в практике статистических вычислений вместо s^2 используют оценку $s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ или (для статистического ряда) $s_1^2 = \frac{1}{n-1} \sum_{i=1}^k (z_i - \bar{x})^2 n_i$, называемую *исправленной выборочной дисперсией*.

О п р е д е л е н и е. *Выборочным среднеквадратическим отклонением* (стандартным отклонением) называется число $s = \sqrt{s^2}$ ($s_1 = \sqrt{s_1^2}$).

Выборочная ковариация и выборочный коэффициент корреляции. Пусть при проведении некоторого опыта наблюдаются две случайные величины ξ и η . Тогда n независимых повторений опыта дадут n пар наблюдавшихся значений $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (x_i – выборочные значения ξ , y_i – выборочные значения η).

О п р е д е л е н и е. *Выборочной ковариацией* случайных величин ξ и η называется число

$$S_{\xi, \eta} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \frac{1}{n} \sum_{i=1}^n y_i.$$

Выборочная ковариация служит оценкой по выборке ковариации случайных величин ξ и η (генеральной ковариации) и является выборочной характеристикой их линейной связи.

О п р е д е л е н и е. *Выборочным коэффициентом корреляции* случайных величин ξ и η называется число $r_{\xi, \eta} = S_{\xi, \eta} / (s_{\xi} s_{\eta})$, где s_{ξ}, s_{η} – выборочные среднеквадратические отклонения случайных величин ξ и η (соответственно).

Выборочный коэффициент корреляции служит оценкой по выборке коэффициента корреляции случайных величин ξ и η (генерального коэффициента корреляции) и является выборочной характеристикой направления и тесноты стохастически-линейной линейной связи между случайными величинами ξ и η .