Основы математической статистики

Лекция 3

Задачи математической статистики

- Установление закономерностей, которым подчинены массовые случайные явления, основано на изучении методами теории вероятностей статистических данных — результатов наблюдений.

- Первая задача математической статистики — указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.
- Вторая задача математической статистики—разработать методы анализа статистических данных в зависимости от целей исследования.

- Задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Генеральная и выборочная совокупности

- *Выборочной совокупностью* или просто выборкой называют совокупность случайно отобранных объектов.
- *Генеральной совокупностью* называют совокупность объектов, из которых производится выборка.
- *Объемом совокупности* (выборочной или генеральной) называют число объектов этой совокупности.

Пример

- Если из 1000 деталей отобрано для обследования 100 деталей, то объем генеральной совокупности $N = 1000$, а объем выборки $n = 100$.

Повторная и бесповторная выборки. Репрезентативная выборка

- *Повторной* называют выборку, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.
- *Бесповторной* называют выборку, при которой отобранный объект в генеральную совокупность не возвращается. На практике обычно пользуются бесповторным случайным отбором.

- Свойства объектов выборки должны правильно отражать свойства объектов генеральной совокупности, или, как говорят, выборка должна быть *репрезентативной* (представительной). Считается, что выборка репрезентативна, если все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку, т. е. выбор осуществляется *случайно*.

Статистическое распределение выборки

- Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 — n_2 раз, ..., x_k —n_k раз и $n_1 + n_2 + \ldots + n_k = n$— объем выборки. Наблюдаемые значения x_1, x_2, … x_k называются *вариантами*, а последовательность вариант, записанная в возрастающем порядке — *вариационным* рядом.

- Числа наблюдений n_1, n_2, … n_k называют *частотами*, а их отношения к объему выборки $\frac{n_1}{n} = p_1^*, \frac{n_2}{n} = p_2^*, \ \dots \ , \frac{n_k}{n} = p_k^*$

— *относительными частотами*. Отметим, что сумма относительных частот равна единице:

$$p_1^* + p_2^* + \dots + p_k^* = \frac{n_1}{n} + \frac{n_2}{n} + \dots + \frac{n_k}{n} = \frac{n_1 + n_2 + \dots + n_k}{n} = \frac{n}{n} = 1$$

- *Статистическим распределением* выборки называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (*непрерывное распределение*). В качестве частоты, соответствующей интервалу, принимают сумму частот вариант, попавших в этот интервал.

Пример

- Дана выборка:
- 12,2,6,2,6,12,6,2,6,12,6,6,12,12,12,6,6,12,6,6
- Составить вариационный ряд и статистическое распределение

Решение

- Составим вариационный ряд:
- 2,2,2,6,6,6,6,6,6,6,6,6,6,12,12,12,12,12,12,12
- Создадим статистическое распределение:

Варианта x_i	2	6	12
Частота n_i	3	10	7

Пример

- Задано распределение частот выборки объема n — 20:

Варианта x_i	2	6	12
Частота n_i	3	10	7

- Написать распределение относительных частот.

Решение

- Найдем относительные частоты, для чего разделим частоты на объем выборки:

$$p_1^* = \frac{3}{20} = 0{,}15; \quad p_2^* = \frac{10}{20} = 0{,}50; \quad p_3^* = \frac{7}{20} = 0{,}35$$

- Поэтому получаем следующее распределение:

Варианта x_i	2	6	12
Относительная частота n_i	0,15	0,50	0,35

Полигон и гистограмма

- Для графического изображения статистического распределения используются *полигоны* и *гистограммы*.
- *Полигоном частот* называют ломаную, отрезки которой соединяют точки

$(x_1; n_1), (x_2; n_2), \ldots, (x_k; n_k)$.

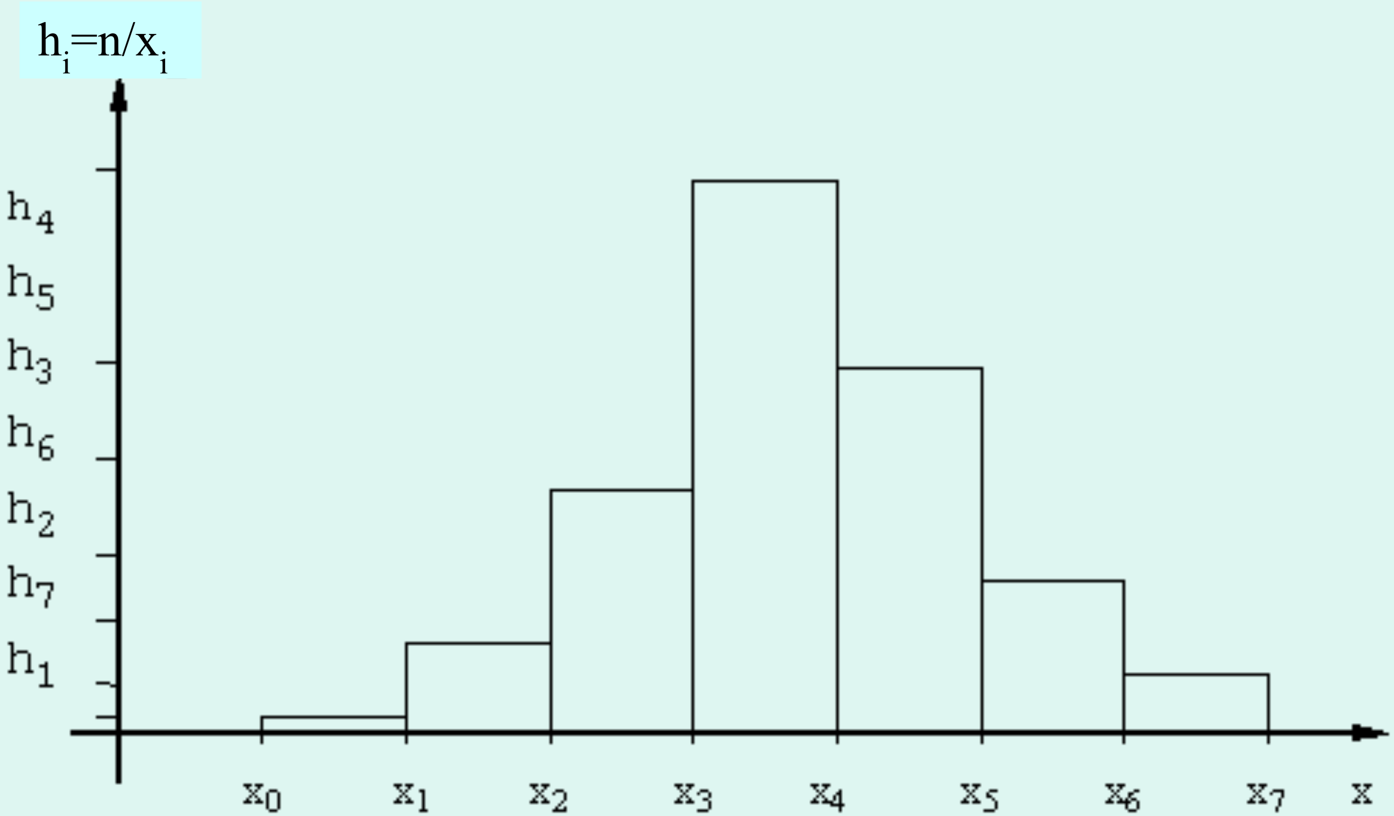
- *Полигоном относительных частот* называют ломаную, отрезки которой соединяют точки

$$\left(x_1; p_1^*\right), \ \left(x_2; p_2^*\right), \ \ldots \ \left(x_k; p_k^*\right)$$

- *Гистограммой частот* называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$(плотность частоты).
- Площадь i-го частичного прямоугольника равна

  $\frac{n_i}{h} \cdot h = n_i$ —сумме частот вариант i-го интервала; следовательно, площадь гистограммы частот равна сумме всех частот, т. е. объему выборки.

Пример гистограммы



- Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{p_i^*}{h}$ (плотность относительной частоты).
- Площадь i-го частичного прямоугольника равна $\frac{p_i^*}{h} \cdot h = p_i^*$ - относительной частоте вариант, попавших в i-й интервал, следовательно, площадь гистограммы относительных частот равна сумме всех относительных частот, т. е. единице.

Статистические оценки параметров распределения

- Статистической оценкой неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин.

Генеральная и выборочная средние. Методы их расчета .

- Пусть изучается дискретная генеральная совокупность объема N относительно количественного признака X.
- *Генеральной средней* $\bar{x}_{Г}$ называют среднее арифметическое значений признака генеральной совокупности.

- Если все значения x_1, x_2, … x_{N} признака генеральной совокупности объема N различны, то

$$\bar{x}_{\Gamma} = \frac{1}{N}\left(x_1 + x_2 + ... + x_N\right)$$

- Если же значения признака x_1, x_2, … x_{k} имеют соответственно частоты N_1, N_2, ..., N_{k}, причем $N_1 + N_2 + ... + N_{\mathrm{k}} = N$, то

$$\bar{x}_{\Gamma} = \frac{1}{N}\left(x_1 N_1 + x_2 N_2 + ... + x_k N_k\right)$$

или

$$\bar{x}_{\Gamma} = \frac{1}{N}\sum_{i=1}^{k} x_i N_i$$

- Пусть для изучения генеральной совокупности относительно количественного признака X произведена выборка объема n.
- *Выборочной средней* $\overline{x}_B$ называют среднее арифметическое значение признака выборочной совокупности.

- Если все значения x_1, x_2, … x_k признака выборки объема n различны, то

$$\bar{x}_B = \frac{1}{n}\left(x_1 + x_2 + ... + x_k\right)$$

- Если же значения признака x_1, x_2, … x_k имеют соответственно частоты n_1, n_2 … n_k причем $n_1 + n_2 + \ldots + n_k = n$, то

$$\bar{x}_B = \frac{1}{n}\left(x_1 n_1 + x_2 n_2 + \ldots + x_k n_k\right)$$

или

$$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

Генеральная и выборочная дисперсии.

- Для того чтобы охарактеризовать рассеяние значений количественного признака X генеральной совокупности вокруг своего среднего значения, вводят сводную характеристику — генеральную дисперсию.

- Генеральной дисперсией $D_{\text{Г}}$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\bar{x}_{\tilde{A}}$.
- Если все значения x_1, x_2, … x_N признака генеральной совокупности объема N различны, то

$$D_{Г} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}_{Г}\right)^2$$

- Если же значения признака $x_1, x_2, \dots x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причем $N_1 + N_2 + \ldots + N_k = N$, то

$$D_{\Gamma} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}_{\Gamma}\right)^2 N_i$$

- Генеральным средним квадратическим отклонением (стандартом) называют квадратный корень из генеральной дисперсии:

$$\sigma_{Г} = \sqrt{D_{Г}}$$

- Выборочной дисперсией $D_{\text{В}}$ называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\bar{x}_{\hat{A}}$.
- Если все значения x_1, x_2, … x_{k} признака выборки объема n различны, то

$$D_B = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_B\right)^2$$

- Если же значения признака $x_1, x_2, \ldots x_{\mathrm{n}}$ имеют соответственно частоты $n_1, n_2 \ldots n_{\mathrm{k}}$ причем $n_1 + n_2 + \ldots + n_{\mathrm{k}} = n$, то

$$D_B = \frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \bar{x}_B \right)^2$$

- Выборочным средним квадратическим отклонением (стандартом) называют квадратный корень из выборочной дисперсии: $\sigma_B = \sqrt{D_B}$

- Выборочная дисперсия является смещенной оценкой генеральной дисперсии, т.е. математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии, а равно

$$M\left[D_B\right]=\frac{n-1}{n}D_{\Gamma}$$

- Легко «исправить» выборочную дисперсию так, чтобы ее математическое ожидание было равно генеральной дисперсии. Достаточно для этого умножить $D_{\text{В}}$ на дробь $\dfrac{n}{n-1}$. Итак, в качестве оценки генеральной дисперсии принимают исправленную дисперсию

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}_B\right)^2$$

- *Точечной* называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше, — *точечные*. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться интервальными оценками.

- *Интервальной* называют оценку, которая определяется двумя числами — концами интервала. Интервальные оценки позволяют установить точность и надежность оценок.

- Пусть X — оцениваемый параметр генеральной совокупности (генеральное среднее, генеральная дисперсия и т.д.), а — $\overline{X}$ его точечная оценка, полученная по выборке.
- Интервальной оценкой параметра X называется оценка вида

$$P\left(\overline{X} - \varepsilon < X < \overline{X} + \varepsilon\right) = P_{\alpha}$$

- При этом интервал $\left(\overline{X}-\varepsilon, \overline{X}+\varepsilon\right)$ называется доверительным интервалом.
- Величина ε характеризует точность оценки, а величина Pα равна вероятности того, что параметр $\overline{X}$ лежит в указанных пределах.

- Пусть рассматриваемый признак распределен в генеральной совокупности по нормальному закону с неизвестными значениями генерального среднего и дисперсии. Для того чтобы получить интервальную оценку генерального среднего $X_г$, поступают следующим образом.

- Для выборки некоторого объема *n* вычисляют выборочное среднее и исправленное среднее квадратическое отклонение .
- Выбирают доверительную вероятность P_{α} (обычно 0,95 или 0,99) и для нее по таблице распределения Стьюдента находят параметр t.
- Рассчитывают полуширину доверительного интервала ε:

$$\varepsilon = \frac{t \cdot s}{\sqrt{n}}$$

- Получают интервальную оценку генеральной средней с выбранной доверительной вероятностью:

$$\overline{X}_{B} - \varepsilon < \overline{X}_{\Gamma} < \overline{X}_{B} + \varepsilon$$