Математическая статистика
Лекция 1

Математическая статистика

- **Математическая статистика** — наука, которая разрабатывает математические методы систематизации и применения статистических данных для практических и научных выводов

План

- Основные понятия
- Точечное оценивание
- Интервальное оценивание
- Тестирование гипотез

Понятие генеральная совокупности

- Совокупность всех наблюдений случайной величины *X*, которые могли бы быть сделаны при данном комплексе условий, называется **генеральной совокупностью** случайной величины *X*, или просто *генеральной совокупностью X*.
- Распределение случайной величины *X* называется *распределением генеральной совокупности*. Число элементов для конечной генеральной совокупности, называют *объёмом* генеральной совокупности. Генеральная совокупность может быть как конечным, так и бесконечным множеством.

Понятие случайной выборки

- Совокупность независимых случайных величин $X_1,…,X_n$, каждая из которых имеет то же распределение, что и наблюдаемая случайная величина X, называется **случайной выборкой** из генеральной совокупности X. При этом число n называют объёмом случайной выборки, а случайные величины $X_1,…,X_n$ – элементами случайной выборки. Любую реализацию $x_1, x_2, . . . , x_n$ случайной выборки $X_1,…,X_n$ будем называть **выборкой** из генеральной совокупности X, или выборочной совокупностью. Выборка из генеральной совокупности X представляет собой некоторое подмножество этой генеральной совокупности.

Пример

- Пример 1. Эксперимент состоит в подбрасывании правильной игральной кости. Случайная величина X – число очков, выпавшее на верхней грани, возможные значения случайной величины X: 1,…,6. В результате эксперимента получаем случайное число x – реализацию случайной величины X, . При повторении эксперимента n раз получаем выборку $x_1,\ldots,x_n$ наблюдений случайной величины X, или, что то же самое, единственное наблюдение случайной выборки $X_1,\ldots,X_n$ объёма n. Генеральная совокупность случайной величины X содержит бесконечное число значений 1,…,6 в равных пропорциях.

Способы формирования выборки

- Полностью случайный отбор
- Простой отбор с помощью регулярной процедуры
- Стратифицированный отбор
- Серийные выборки
- Комбинированный отбор

Выборочные характеристики

- 1) **Вариационный ряд.**

Элементы реализации выборки $x_1, x_2, \ldots, x_n$ упорядочим по возрастанию

$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n-1)} \leqslant x_{(n)}$

где $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$, $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$.

получится новый набор значений случайных величин, называемый вариационным рядом

Вариационный ряд выборки $x_1, x_2, \ldots, x_n$ можно рассматривать как реализацию вариационного ряда случайной выборки $X_1,\ldots,X_n$.

Случайную величину $X_{(k)}$ называется k -м членом вариационного ряда или k -й порядковой статистикой.

$X_{(1)}, X_{(n)}$ – экстремальные статистики

Выборочные характеристики

- **2) Эмпирическая функция распределения**.
- Пусть выборка $x_1, x_2, \ldots, x_n$ объёма n берется из распределения с функцией распределения $F(x) = P(X_i \leq x)$.

Эмпирической функцией распределения, построенной по случайной выборке $X_1, X_2, \ldots, X_n$ объёма n, называется случайная функция

$$\hat{F}_n(x) = \frac{\text{количество } X_i \epsilon(-\infty, x]}{n} = \frac{\sum_{i=1}^{n} I(\{X_i \leq x\})}{n}$$

где $I(\{X_i \leq x\}) = \begin{cases} 1, \text{если } X_i \leq x \\ 0, \text{если } X_i > x \end{cases}$

Пример

П р и м е р 1. Пусть дана числовая выборка

$$\vec{X} = (0;\ 2;\ 1;\ 2{,}6;\ 3{,}1;\ 4{,}6;\ 1;\ 4{,}6;\ 6;\ 2{,}6;\ 6;\ 7;\ 9;\ 9;\ 2{,}6).$$

Построим по ней вариационный ряд

$$(0;\ 1;\ 1;\ 2;\ 2{,}6;\ 2{,}6;\ 2{,}6;\ 3{,}1;\ 4{,}6;\ 4{,}6;\ 6;\ 6;\ 7;\ 9;\ 9)$$

и эмпирическую функцию распределения (рис. 1).

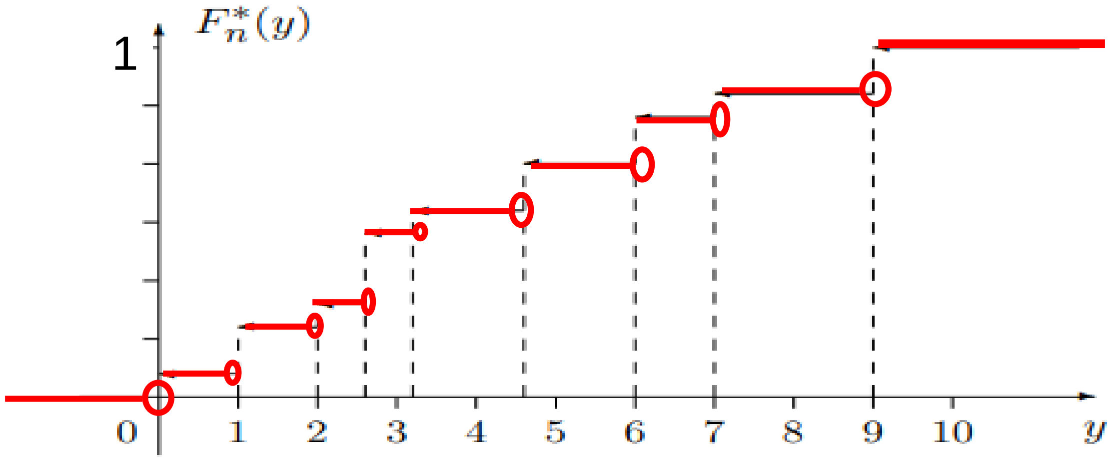

Рис. 1. Эмпирическая функция распределения

Эта функция является функцией распределения случайной величины, принимающей значение 0 с вероятностью $\frac{1}{15}$, значение 1 с вероятностью $\frac{2}{15}$, значение 2 с вероятностью $\frac{1}{15}$ и т. д.

Выборочные характеристики

- **3)Гистограмма** Эмпирическим аналогом таблицы или плотности распределения является гистограмма.
- Гистограмма строится по группированным данным. Предполагаемую область значений случайной величины X (или область выборочных данных) делят на некоторое количество не обязательно одинаковых интервалов. Пусть $A_1, \dots, A_k$ — интервалы на прямой, называемые интервалами группировки. Обозначим для $j = 1, \dots, k$ через v_j число элементов выборки, попавших в интервал A_j:

$$v_i = \sum_{i=1}^{n} I(\{X_i \in A_i\}) - \text{количество } X_i \in A_i$$

$$n = \sum_{j=1}^{k} v_j$$

На каждом из интервалов A_j строят прямоугольник, площадь которого пропорциональна v_i. Общая площадь всех прямоугольников должна равняться единице. Если l_j— длина интервала A_j, то высота f_j прямоугольника над этим интервалом равна

$$\hat{f}_j = \frac{v_j}{n \cdot l_j}$$

Полученная фигура, состоящая из объединения прямоугольников, называется гистограммой.

Пример

Пример 2. Имеется вариационный ряд из примера 1:

$$(0;\ 1;\ 1;\ 2;\ 2{,}6;\ 2{,}6;\ 2{,}6;\ 3{,}1;\ 4{,}6;\ 4{,}6;\ 6;\ 6;\ 7;\ 9;\ 9).$$

Разобьём отрезок $[0,\ 10]$ на четыре равных отрезка. Отрезку $[0,\ 2{,}5)$ принадлежат четыре элемента выборки, отрезку $[2{,}5,\ 5)$ — шесть, отрезку $[5,\ 7{,}5)$ — три, и отрезку $[7{,}5,\ 10]$ — два элемента выборки. Строим гистограмму (рис. 2). На рис. 3 — гистограмма для той же выборки, но при разбиении области на пять равных отрезков.

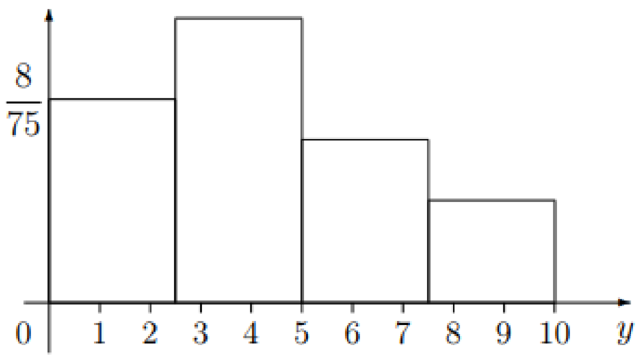

Рис. 2. Гистограмма при $k = 4$

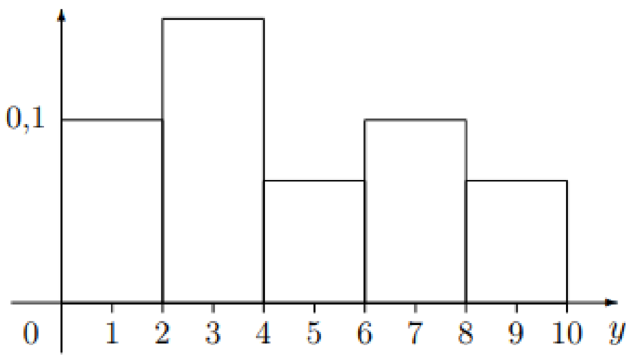

Рис. 3. Гистограмма при $k = 5$

Чем больше интервалов группировки, тем лучше: фигура, состоящая из более узких прямоугольников, точнее приближает истинную плотность распределения. С другой стороны, бессмысленно брать число интервалов $k(n)$ порядка n: тогда в каждый интервал попадёт в среднем по одной точке и гистограмма не будет приближаться к плотности с ростом n.

Выборочные характеристики

4) Выборочные моменты

Истинные моменты	Оценки
$E(X)=E(X_1)=\mu$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
$D(X)=D(X_1)=\sigma^2$	$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$
α_k	$\widehat{\alpha_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
β_k	$\widehat{\beta_k} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k$

Свойства выборочной функции

Рассмотрим случайную выборку $X_1,\dots,X_n$ - (независимые одинаково распределенные случайные величины) из распределения F

- 1) $E(\widehat{F}_n(x)) = F(x)$
- 2) $D(\widehat{F}_n(x)) = \frac{1}{n}F(x)(1 - F(x))$
- 3) $\widehat{F}_n(x) \overset{p}{\to} F(x), \text{при } n \to \infty$
- Доказательство (см лекцию)

- 4)$\sup\limits_x \left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{\text{п.н.}} 0, \text{при } n \to \infty$

(теорема Гливенко-Кантелли), сильная состоятельность

Свойства выборочной функции

- 5) $n\hat{F}_n(x) = \sum_{i=1}^{n} I(\{x_i \leq x\}) \sim B\big(n, F(x)\big)$ – биномиальное распределение
- 6) $\dfrac{\hat{F}_n(x) - F(x)}{\sqrt{\frac{1}{n}F(x)(1-F(x))}} \xrightarrow[n \to \infty]{d}$ z$\sim N(0,1)$

Напомним, что по ЦПТ $\dfrac{\bar{\xi}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{d} z\sim N(0,1),$

где $\bar{\xi}_n = \dfrac{S_n}{n}, E(\xi_1) = \mu, D(\xi_1) = \sigma^2$

Свойства выборочного среднего

$X_1,\ldots,X_n$ независимые одинаково распределенные случайные величины

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Пусть $E(X_1)<\infty$, $E(X_1)=\mu$, $D(X_1)=\sigma^2$

1) $E(\bar{X})=E(X_1)=\mu$, $D(\bar{X})=\frac{D(X_1)}{n} = \frac{\sigma^2}{n}$

2) $\bar{X} \xrightarrow[n \to \infty]{p} \mu$ (из ЗБЧ)

3) $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow[n \to \infty]{d} z \sim N(0,1)$ (из ЦПТ)

Свойства выборочной дисперсии

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1) $\mathrm{E}(S^2) \neq \sigma^2$, $\mathrm{E}(S^2) = \frac{n-1}{n}\sigma^2$

Исправленная выборочная дисперсия (несмещенная оценка)

$S_0^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

2) $S^2 \xrightarrow[n \to \infty]{\mathrm{p}} \sigma^2$, $S_0^2 \xrightarrow[n \to \infty]{\mathrm{p}} \sigma^2$

3) $\frac{S^2 - \sigma^2}{\sqrt{D(S^2)}} \xrightarrow[n \to \infty]{\mathrm{d}} z \sim \mathrm{N}(0,1)$, $\frac{S_0^2 - \sigma^2}{\sqrt{D(S_0^2)}} \xrightarrow[n \to \infty]{\mathrm{d}} z \sim \mathrm{N}(0,1)$

Конечный размер генеральной совокупности

$X_1, X_2, \ldots, X_n$ – одинаково распределенные, но не являются независимыми

$\mathrm{E}(X_i) = \mathrm{E}(X_1)$, т. е. $\mathrm{E}(\bar{X}) = E(X_1) = \mu$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Утв 1. $cov(X_i, X_j) = \frac{-\sigma^2}{N-1}$

Док-во. (смотри лекцию)

Утв 2. $D(\bar{X}) = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right)$

Док-во. (смотри лекцию)