Математическая статистика

Михайлов Максим

11 сентября 2021 г.

Оглавление

Лекция 1

6 сентября

1 Организационные вопросы

Большая часть баллов зарабатывается индивидуальными заданиями, выполняемыми в Excel — 30 баллов. Тест с большим числом вопросов — 20 или 25 баллов.

2 Введение

Теория вероятности состоит в следующем: исследуется случайная величина с заданным распределением. Математическая статистика занимается обратным — даны данные, нужно приближенно найти числовые характеристики случайной величины и с некоторой уверенностью найти вид распределения. Матстатистика также исследует связанность случайных величин, их корреляцию. В идеале есть цель построить модель, которая по значениям одних случайных величин предсказывает другие.

Пусть проводится некоторое количество экспериментов, в ходе которых появились некоторые данные.

Определение. **Генеральная совокупность** — набор всех исходов проведенных экспериментов.

В реальности наблюдается некоторая выборка генеральной совокупности, ибо рассматривать всю совокупность нецелесообразно.

Определение. **Выборочная совокупность** — исходы наблюдаемых экспериментов.

Определение. Выборка называется **репрезентативной**, если её распределение совпадает с распределением генеральной совокупности.

Выборка может быть нерепрезентативной, как в примере с ошибкой выжившего. Мы считаем, что таких ошибок у нас нет и все выборки репрезентативны, ибо исправление этих ошибок — задача конкретной области, в которой используется матстатистика.

Определение (после опыта). Пусть проведено n наблюдаемых независимых экспериментов, в которых случайная величина приняла значение $X_1, X_2 \dots X_n$. Набор[1] этих данных называется **выборкой объема** n.

Определение (до опыта). **Выборкой объема** n называется набор из n независимых одинаково распределенных случайных величин.

Пусть имеется выборка в смысле "после опыта" объема n. Её можно интерпретировать как следующую дискретную случайную величину:

X_i	X_1	X_2	$\dots$	X_n	$\sum$
p_i	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$	1

Средневыборочное:

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

Выборочная дисперсия:

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

2.1 Выборочная функция распределения

$$F_n^*(z) \coloneqq \frac{1}{n} \sum_{i=1}^{n} I(X_i < z) = \frac{\text{число } X_i \in (-\infty, z)}{n}$$

Примечание. I — индикатор:

$$I(X_i < z) = \begin{cases} 1, & X_i < z \\ 0, & X_i \geq z \end{cases}$$

Теорема 1.

$$\forall x \in \mathbb{R} \quad F_n^*(z) \xrightarrow[n \to \infty]{P} F(z)$$

Доказательство. Заметим, что

$$\mathbb{E} I(X_1 < z) = 1 \cdot P(X_1 < z) + 0 \cdot P(X_1 \geq z) = P(X_1 < z) = F(z)$$

[1] Или вектор.

, где $F(z)$ — функция распределения X_1. Заметим, что $F(z) \leq 1 < \infty$, следовательно применим ЗБЧ Хинчина:

$$F_n^*(z) = \frac{\sum_{i=1}^n I(X_i < z)}{n} \xrightarrow{P} \mathbb{E}I(X_1 < z) = F(z)$$

□

Примечание. На самом деле имеется даже равномерная сходимость по вероятности — это теорема Гливенко-Кантелли:

$$\sup_{z\in\mathbb{R}} |F_n^*(z) - F(z)| \xrightarrow[n\to\infty]{P} 0$$

3 Первоначальная обработка статданных

Если отсортировать данные, то получим **вариационный ряд**: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Если учесть повторяющиеся экземпляры, то получим **частотный вариационный ряд**:

$X_{(i)}$	$X_{(1)}$	$X_{(2)}$	$\dots$	$X_{(k)}$	$\sum$
n_i	n_1	n_2	$\dots$	n_k	n
p_i^*	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\dots$	$\frac{n_k}{n}$	1

Определение. $h \coloneqq X_{\max} - X_{\min}$ — **размах выборки**

Допустим, что разбили интервал $(X_{\min}, X_{\max})$ на k интервалов, чаще всего одинаковой длины.[2] Тогда $l_i = \frac{h}{k}$ — длина каждого интервала и интервальный ряд можно заменить интервальным вариационным рядом.

i	l_1	l_2	$\dots$	l_k	$\sum$
m_i	m_1	m_2	$\dots$	m_k	n
$\frac{m_i}{n}$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\dots$	$\frac{m_k}{n}$	1

m_i — число попавших в i-тый интервал данных.

По такой таблице можно построить **гистограмму**. На координатной плоскости построим прямоугольники с основаниями l_i и высотами $\frac{m_i}{nl_i}$. В результате получаем ступенчатую фигуру площади 1, которая и называется гистограммой.

Теорема 2. При $n \to \infty, k(n) \to \infty$, причем $\frac{k(n)}{n} \to 0$, гистограмма будет стремиться к плотности распределения:

$$\frac{m_i}{n} \xrightarrow{P} P(X_i \in l_i) = \int_{l_i} f(x)dx$$

[2] Применяются и другие разбиения, например равнонаполненное.

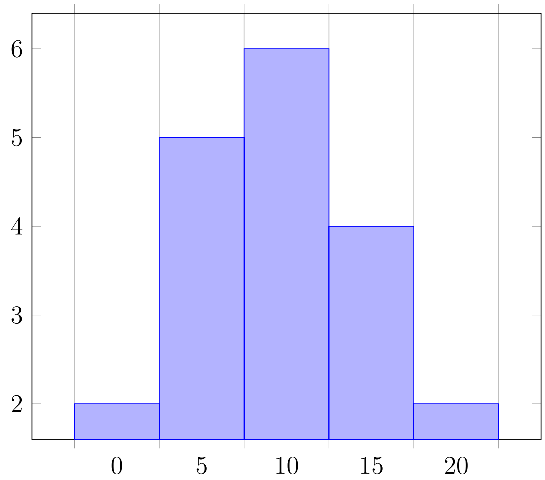

Рис. 1.1: Пример гистограммы

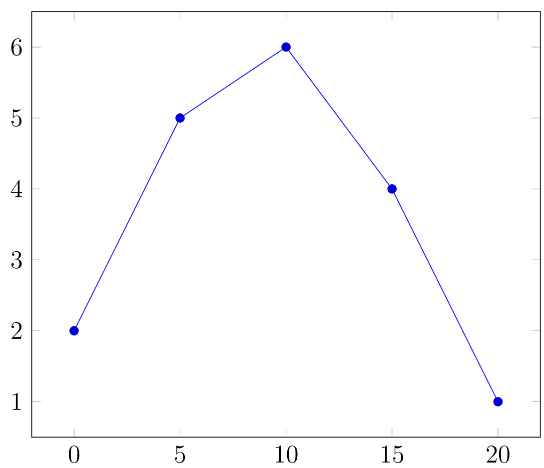

Рис. 1.2: Пример полигона

Чаще всего число интервалов берется по формуле Стёрджесса: $k \approx 1 + \log_2 n$. Иногда $k \approx \sqrt[3]{n}$.

Примечание. Иногда выборка изображается в виде **полигона**: отображаются точки, соответствующие серединам интервалов и ставим точки на высоте $\frac{m_i}{n}$.