Лекция 1.

Теория вероятности изучает характеристику случайных величин, тогда как математическая статистика решает обратную задачу
Допустим, что у нас есть случайная величина, по ней мы можем найти матожидание, моменты и оценить, какое распределение имеет случайная величина.

Выборки

Def. Выборка - набор данных, полученных в ходе экспериментов. Тогда количество экспериментов n - объем Выборки
Def. Генеральной совокупностью называются все результаты проведенных экспериментов
Def. Выборочной совокупностью называются наблюдаемые данные экспериментов
Не все данные экспериментов мы можем наблюдать, например, выборы, тогда опросы голосовавших - выборочная совокупность, а результаты выборов - генеральная. Очевидно, что выборочная и генеральная совокупности могут иметь различные распределения.
Def. Выборка называется **репрезентативной**, если ее распределение близко к распределению генеральной совокупностью
Пример - ошибка выжившего. Во время Второй Мировой стал вопрос, в каких местах стоит бронировать корпус самолета. Самолеты возвращались с пулевыми отверстиям, и интуитивно казалось, что стоит бронировать те места, которые больше всего пострадали. Однако не были учтены те самолеты, которые не вернулись, а те, которые выжили, выжили благодаря тому, что были прострелены в нелетальных местах, поэтому было принято решение бронировать фюзеляж в менее пострадавших местах
В дальнейшем считаем, что все выборки репрезентативны

Def. 1. Выборкой объема n называется набор из n эксперментаных данных $\vec{X} = (x_1, x_2, \dots, x_n)$ (апостериорное определение)
Def. 2. Выборкой объема n называется набор из n независимых одинаково распределенных случайных величин $\vec{X} = (X_1, X_2, \dots, X_n)$ (априорное определение)

Выборочные характеристики

Можно выборку рассматривать как дискретную случайную величину с одинаковыми вероятностями $p_i = \dfrac{1}{n}$ и вычислить для нее математическое ожидание, дисперсию и функцию распределения

Def. Выборочным средним $\overline{x}$ называется величина $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

Def. Выборочной дисперсией D^* называется величина $D^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{x})^2$ (или $D^* = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{x}^2$)

По закону больших чисел выборочное среднее будет сходиться к матожиданию

Def. Исправленной дисперсией называется величина $S^2 = \frac{n}{n-1}D^* = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$

Def. Выборочной функцией распределения $F^*(x)$ называется функция $F^*(x) = \dfrac{\text{число данных } x_i < x}{n}$

Th. Выборочная функция распределения поточечно сходится к теоретической функции распределения:

$$\forall y \in \mathbb{R} F^*(y) \xrightarrow{p} F(y)$$

$F(y) = P(X < y)$

$F_y^* = \frac{1}{n}\sum_{i=1}^{n} I(X_i < y) \xrightarrow[\text{по ЗБЧ}]{p} EI(X_i < y) = P(X_i < y) = P(X_1 < y) = F_{X_1}(y)$

Усилим теорему

Th. Гливенко-Кантелли. $\sup\limits_{x\in\mathbb{R}} |F^*(x) - F(x)| \xrightarrow{p} 0$

Th. Колмогорова. $\sqrt{n}\sup\limits_{x\in\mathbb{R}} |F^*(x) - F(x)| \rightrightarrows K$ - распределение Колмогорова с функцией распределения $F_K(x) = \sum\limits_{j=-\infty}^{\infty}(-1)^j e^{-2j^2x^2},\ x \in [0;\infty)$

Начальная обработка статданных

1. Ранжирование данных - упорядочиваем выборки по возрастанию. В результате получаем вариационный ряд $\vec{X} = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$

   $X_{(1)} = \min X_i;\quad X_{(n)} = \max X_i$

   $X_{(i)} = i$-ая порядковая статистика
2. Объединим повторяющиеся данные - получаем т.н. частотный вариационный ряд

X_i	$X_{(1)}$	...	$X_{(r)}$	$\sum$
n_i	n_1	...	n_r	n

   Иногда часть данных отбрасывается сверху и снизу (по 5, по 10, по 5% и так далее), чтобы сделать выборку репрезентативной

   Тогда $\overline{x} = \frac{1}{n}\sum X_i n_i,\ D^* = \frac{1}{n}\sum(X_i - \overline{x})^2 n_i$

3. Чтобы уменьшить количество вычислений или сделать гистограмму, делают интервальный вариационный ряд: разбиваем данные на интервалы и считаем, сколько данных n_i попало в интервал.

   Тогда n_i - частота интервала A_i

   Есть два основные способа разбиения на интервалы:

   (a) Интервалы одинаковой длины
   (b) Равнонаполненные интервалы (в каждом интервале примерно одинаковое количество данных)

   Число интервалов K такое, что $\dfrac{K(n)}{n} \longrightarrow 0$ и $K(n) \underset{n\to\infty}{\longrightarrow} \infty$

   Обычно применяют формулу Стерджесса $K \approx 1 + \log_2 n$ или $K \approx \sqrt[3]{n}$

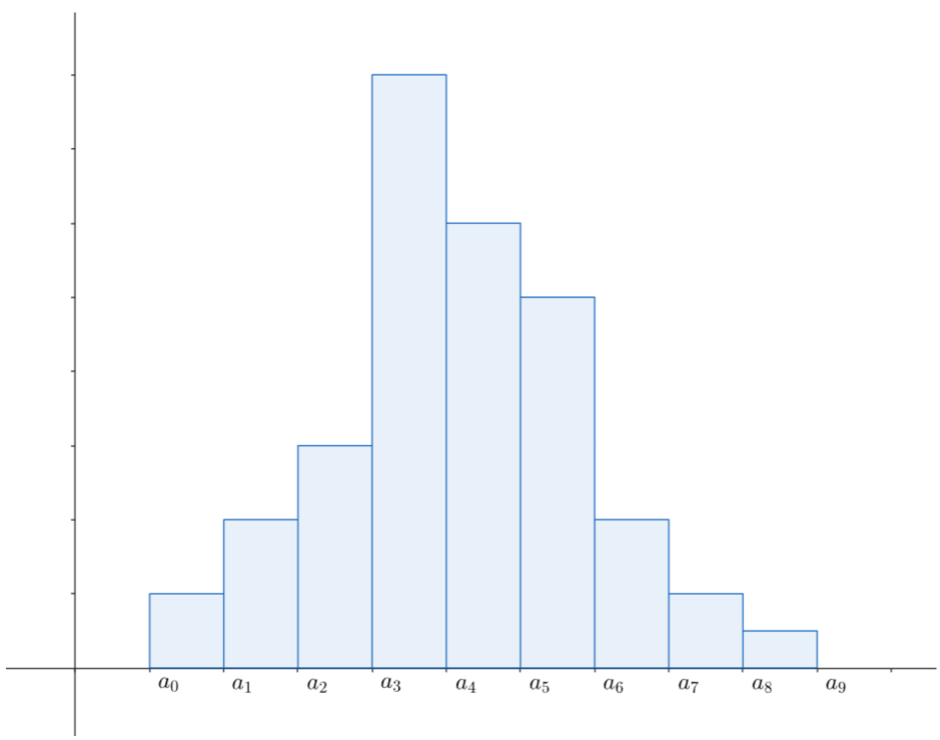
   Пусть получили интервальный вариационный ряд

интервалы	$[a_0; a_1)$	$[a_1; a_2)$	$\dots$	$[a_{K-1}; a_K]$	$\sum$
частоты	n_1	n_2	$\dots$	n_K	n

Геометрическая интерпретация данных

- Гистограмма

  Строится ступенчатая фигура из прямоугольников, основание i-ого прямоугольника - интервал, высота прямоугольника - $\dfrac{n_i}{nl_i}$, где l_i - длина интервала

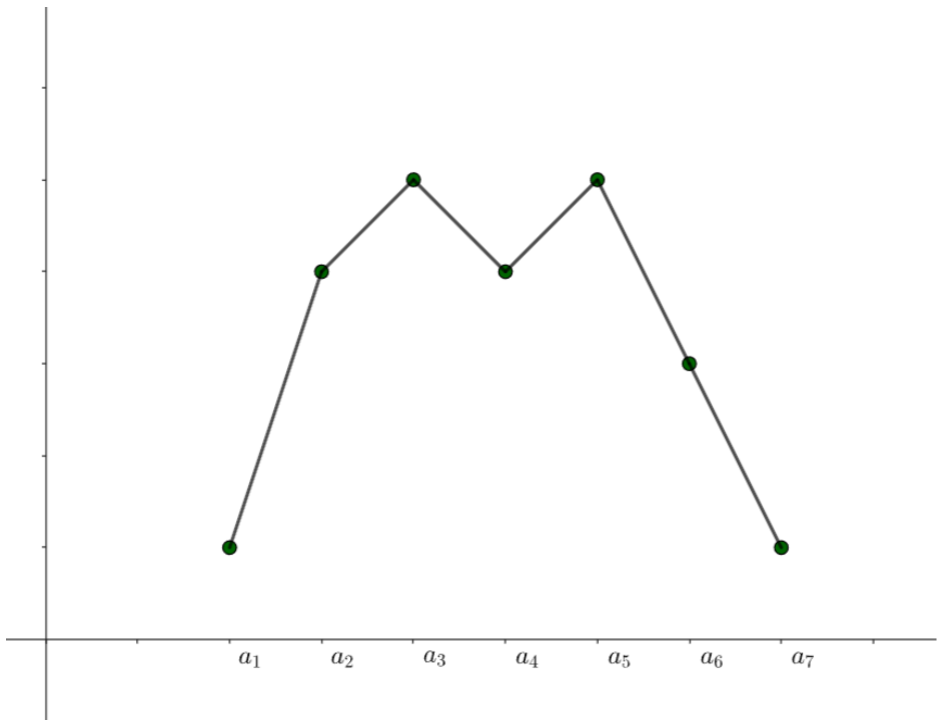


Визуально можно сделать гипотезу, как ведет себя распределение.

Th. Гистограмма поточечно сходится к теоретической плотности

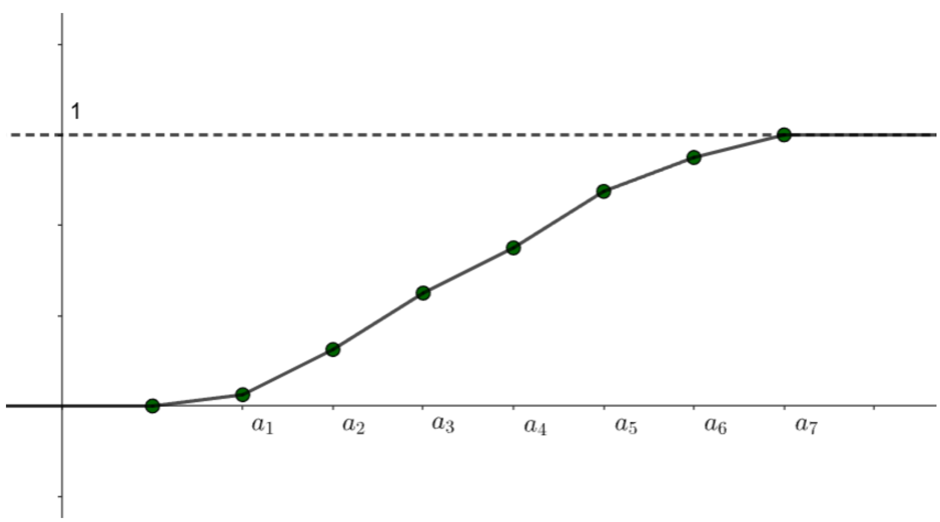
- Полигон

  На оси абсцисс отмечаем значения частотного вариационного ряда, по оси ординат - их частоты. Получившиеся точки соединяем отрезками



- Выборочная функция распределения

  На основе таблицы строится график функции распределения



Она может быть ступенчатой, ломаной или соединена по усмотрению