Московский государственный технический университет
имени Н. Э. Баумана

Факультет: Информатика и системы управления
Кафедра: Программное обеспечение ЭВМ и информационные технологии

Математическая статистика
Лекции

Москва, 2015 г.

Содержание

1 Предельные теоремы теории вероятностей

1.1 Неравенства Чебышева

Теорема 1.1 (*первое неравенство господина Чебышева*).

$$\left.\begin{matrix} X \geq 0; \\ \exists MX. \end{matrix}\right\} \Rightarrow \forall \varepsilon > 0,\ \mathsf{P}\{X \geq \varepsilon\} \leq \frac{MX}{\varepsilon} \tag{1}$$

- X — случайная величина;
- $\mathsf{P}\{X \leq 0\} = 0$ так как $X \geq 0$.

Доказательство. Для непрерывной случайное величины X и зная, что при $X \geq 0 \Rightarrow f(x) = 0,\ x < 0$

$$MX = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \int_{0}^{+\infty} x\, f(x)\, dx = \underbrace{\int_{0}^{\varepsilon} x\, f(x)\, dx}_{\geq 0} + \int_{\varepsilon}^{+\infty} x\, f(x)\, dx$$

учитывая $x \geq \varepsilon$

$$\underbrace{\int_{0}^{\varepsilon} x\, f(x)\, dx}_{\geq 0} + \int_{\varepsilon}^{+\infty} x\, f(x)\, dx \geq \int_{\varepsilon}^{+\infty} x\, f(x)\, dx \geq \varepsilon \cdot \int_{\varepsilon}^{+\infty} f(x)\, dx$$

где

$$\varepsilon \cdot \int_{\varepsilon}^{+\infty} f(x)\, dx = \varepsilon \cdot \mathsf{P}\{X \geq \varepsilon\}$$

таким образом

$$MX \geq \varepsilon \cdot \mathsf{P}\{X \geq \varepsilon\} \Rightarrow \mathsf{P}\{X \geq \varepsilon\} \leq \frac{MX}{\varepsilon}$$

□

Теорема 1.2 (*второе неравенство лорда Чебышева*).

$$\exists MX, \exists DX \Rightarrow \forall \varepsilon > 0,\ \mathsf{P}\big\{|X - MX| \geq \varepsilon\big\} \leq \frac{DX}{\varepsilon^2} \tag{2}$$

- X — случайная величина.

Доказательство. Выпишем дисперсию

$$DX = M\big[(X - MX)^2\big]$$

Рассмотрим случайную величину $Y = (X - MX)^2$, где $Y \geq 0$. Тогда из *первого неравенства Чебышева* следует, что $\forall \delta \geq 0$, $MY \geq \delta\, \mathsf{P}\{Y \geq \delta\}$, где получается, что $\delta = \varepsilon^2$.

$$\Big[DX = M\big[(X - MX)^2\big]\Big] \geq \Big[\varepsilon^2 \cdot \mathsf{P}\big\{(X - MX)^2 \geq \varepsilon^2\big\} = \varepsilon^2 \cdot \mathsf{P}\big\{|X - MX| \geq \varepsilon\big\}\Big]$$

таким образом

$$DX \geq \varepsilon^2 \cdot \mathsf{P}\big\{|X - MX| \geq \varepsilon\big\} \Rightarrow \mathsf{P}\big\{|X - MX| \geq \varepsilon\big\} \leq \frac{DX}{\varepsilon^2}$$

□

Пример 1.1. Предельно допустимое давление в пневмосистеме ракеты равна 200 (Па). После проверки большого количество ракет было получено среднее значение давления 150 (Па). Оценить вероятность того, что давление в пневмосистеме очередной ракеты будет больше 200 (Па), если по результатам проверки ракет было получено среднеквадратичное отклонение 5 (Па).

Решение. Имеем следующее:

- случайная величина X — давление в пневмосистеме;
- $X \geq 0$;
- $MX = 150$ (Па);
- $DX = 25$ (Па);

Решим поставленную задачу с помощью *первого неравенства Чебышева*

$$\Big[\mathsf{P}\{X \geq \varepsilon\} = \mathsf{P}\{X \geq 200\}\Big] \leq \left[\frac{MX}{\varepsilon} = \frac{150}{200} = \frac{3}{4} = 0.75\right]$$
$$\mathsf{P}\{X \geq 200\} \leq 0.75$$

Поскольку нам известна дисперсия почему бы не воспользоваться *вторым неравенством Чебышева*? Действуем. Для начало рассмотрим вероятность следующего события

$$\mathsf{P}\{X \geq \varepsilon\} = \mathsf{P}\{X \geq 200\} = \mathsf{P}\{X - \underbrace{150}_{MX} \geq \underbrace{50}_{\varepsilon}\}$$

Остаётся построить вероятность, которая будет удовлетворять форме *второго неравенства Чебышева* (т. е. сделать модуль).

$$\mathsf{P}\{X - 150 \geq 50\} \leq \mathsf{P}\{X - 150 \geq 50\} + \mathsf{P}\{X - 150 \leq -50\}$$

Так как *события* $\{X - 150 \geq 50\}$ и $\{X - 150 \leq -50\}$ *несовместные*, то по *формуле сложения вероятностей несовместных событий* получаем

$$\mathsf{P}\{X - 150 \geq 50\} + \mathsf{P}\{X - 150 \leq -50\} = \\ = \mathsf{P}\big\{\{X - 150 \geq 50\} + \{X - 150 \leq -50\}\big\} = \mathsf{P}\big\{|X - 150| \geq 50\big\}$$

Таким образом применяем *второе неравенство Чебышева*

$$\Big[\mathsf{P}\big\{|X - MX| \geq \varepsilon\big\} = \mathsf{P}\big\{|X - 150| \geq 50\big\}\Big] \leq \left[\frac{DX}{\varepsilon^2} = \frac{25}{50^2} = \left(\frac{5}{50}\right)^2 = 0.01\right]$$
$$\mathsf{P}\big\{|X - 150| \geq 50\big\} \leq 0.01$$

Ответ:

- с использованием *первого неравенства Чебышева* $\mathsf{P} \leq 0.75$;
- с использованием *второго неравенства Чебышева* $\mathsf{P} \leq 0.01$.

Замечание. *Второе неравенство Чебышева* даёт более точную оценку, так как используется информация о дисперсии случайной величины.

Замечание. Использование *первого неравенства Чебышева* при $\varepsilon < MX$ и *второго неравенства Чебышева* при $\varepsilon < \sqrt{DX}$ даёт тривиальную оценку: $\mathsf{P} \leq 1$.

1.2 Сходимость последовательности случайных величин

Будем считать, что $X_1, \ldots, X_n, \ldots$ — *последовательность случайных величин*, заданных на одном вероятностном пространстве.

Определение 1.1. Последовательность случайных величин $X_1, \ldots, X_n, \ldots$ *сходится по вероятности* к случайной величине Z, если $\forall \varepsilon > 0$, $\mathsf{P}\big\{|X_n - Z| \geq \varepsilon\big\} \xrightarrow[n\to\infty]{} 0$
Обозначение:

$$X_n \xrightarrow[n\to\infty]{\mathsf{P}} Z$$

Определение 1.2. Последовательность случайных величин $X_1, \ldots, X_n, \ldots$ *слабо сходится* к случайной величине Z, если $\forall x \in \Re$ где F_z непрерывна в точке x, числовая последовательность $F_{X_1}(x), \ldots, F_{X_n}(x), \ldots$ сходится к $F_Z(x)$. Обозначение:

$$F_{X_n}(x) \underset{n\to\infty}{\Longrightarrow} F_Z(x)$$

Замечание. Данные виды сходимости *неэквивалентны.*

1.3 Закон больших чисел (ЗБЧ)

Будем считать, что

- $X_1, \ldots, X_n, \ldots$ — *последовательность случайных величин*;
- $\exists MX_i = m_i$, где $i = \overline{1, \infty}$.

Определение 1.3. Последовательность случайных величин $X_1, \ldots, X_n, \ldots$ удовлетворяет *закону больших чисел*, если

$$\forall \varepsilon > 0,\ \mathsf{P}\left\{\left|\frac{1}{n} \cdot \sum_{i=1}^{n} X_i - \frac{1}{n} \cdot \sum_{i=1}^{n} m_i\right| \geq \varepsilon\right\} \xrightarrow[n\to\infty]{} 0$$

Замечание. Выполнение *закона больших чисел* для последовательности $X_1, \ldots, X_n, \ldots$ означает, что при достаточно больших n величина

$$Y_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i - \frac{1}{n} \cdot \sum_{i=1}^{n} m_i$$

практически теряет случайный характер.

Теорема 1.3 (*Закон больших чисел в форме Чебышева* или *достаточное условие выполнимости для последовательности случайных величин*)**.** Последовательность случайных величин $X_1, \ldots, X_n, \ldots$ удовлетворяет *закону больших чисел* тогда и только тогда, когда выполняются следующие условия:

- случайные величины $X_1, \ldots, X_n, \ldots$ — независимы;
- $\exists MX_i = m_i$, $\exists DX_i = \sigma_i^2$, $i = 1, 2, \ldots$;
- Дисперсия случайных величин ограничена в совокупности т. е.

$$\exists C > 0 : \sigma_i^2 \leq C, \quad i = 1, 2, \ldots$$

Доказательство.

$$\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \quad n \in N$$
$$M\overline{X}_n = \frac{1}{n} \cdot \sum_{i=1}^{n} m_i$$
$$D\overline{X}_n = \frac{1}{n^2} \cdot D\left(\sum_{i=1}^{n} X_i\right) = \{X_i \text{ независимы}\} = \frac{1}{n^2} \cdot \sum_{i=1}^{n} DX_i = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \sigma_i^2$$

Используем *второе неравенство Чебышева*

$$\mathsf{P}\{|\overline{X}_n - M\overline{X}_n| \geq \varepsilon\} \leq \frac{D\overline{X}_n}{\varepsilon^2}$$

В нашем случае

$$\mathsf{P}\left\{\left|\frac{1}{n} \cdot \sum_{i=1}^{n} X_i - \frac{1}{n} \cdot \sum_{i=1}^{n} m_i\right| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \cdot \sum_{i=1}^{n} \sigma_i^2 \leq \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \cdot \sum_{i=1}^{n} C = \frac{1}{\varepsilon^2} \cdot \frac{1}{n^2} \cdot n\,C = \frac{C}{\varepsilon^2\,n}$$

таким образом

$$0 \leq \mathsf{P}\left\{\left|\frac{1}{n} \cdot \sum_{i=1}^{n} X_i - \frac{1}{n} \cdot \sum_{i=1}^{n} m_i\right| \geq \varepsilon\right\} \leq \frac{C}{\varepsilon^2\,n}$$

По *теореме о двух милиционерах*

$$\mathsf{P}\left\{\left|\frac{1}{n} \cdot \sum_{i=1}^{n} X_i - \frac{1}{n} \cdot \sum_{i=1}^{n} m_i\right| > \varepsilon\right\} \xrightarrow[n\to\infty]{} 0$$

□

Следствие. Пусть

- выполняется теорема о ЗБЧ в форме Чебышева
- X_i одинаково распределены т. е. $MX_i = m$, $DX_i = \sigma^2$, $i \in \aleph$

Тогда

$$\mathsf{P}\left\{\left|\frac{1}{n} \cdot \sum_{i=1}^{n} X_i - m\right| > \varepsilon\right\} \xrightarrow[n\to\infty]{} 0 \tag{3}$$

Следствие (Теорема Бернулли, ЗБЧ в форме Бернулли)**.** Пусть

- проводится серия испытаний по *схеме Бернулли*
  - с вероятностью успеха p;
  - с вероятностью неудачи $q = 1 - p$;
- наблюденная частота успеха $r_n = \{\text{число успехов в первых n испытаниях}\}/n$.

Тогда

$$r_n \xrightarrow[n\to\infty]{\mathsf{P}} p \tag{4}$$

Доказательство. Рассмотрим случайную величину

$$X_i = \begin{cases} 1, & \text{если в } i\text{-ом испытании успех;} \\ 0, & \text{иначе.} \end{cases}$$

тогда $MX_i = p$, $DX_i = p\,q$. Поскольку $\exists MX_i$, $\exists DX_i$ и X_i — независимы по определению *схемы Бернулли* $\Rightarrow$ выполняются все условия предыдущего следствия

$$\mathsf{P}\left\{ \left| \underbrace{\frac{1}{n} \cdot \sum_{i=1}^{n} X_i}_{r_n} - p \right| > \varepsilon \right\} \xrightarrow[n\to\infty]{} 0$$

Таким образом

$$r_n \xrightarrow[n\to\infty]{\mathsf{P}} p$$

□

1.4 Центральная предельная теорема (ЦПТ)

Пусть имеется:

- последовательность независимых случайных величин $X_1, \dots, X_n, \dots$;
- X_i одинаково распределены;
- $MX_i = m$, $DX_i = \sigma^2$, $i \in \aleph$

Составим последовательность:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad n \in \aleph; \qquad M\overline{X}_n = m; \qquad D\overline{X}_n = \frac{1}{n^2} \sum_{i=1}^{n} DX_i = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

Составим случайную величину $Y = \dfrac{\overline{X}_n - M\overline{X}_n}{\sqrt{D\overline{X}_n}} = \dfrac{1/n \cdot \sum_{i=1}^{n} X_i - m}{\sigma/\sqrt{n}}$. Тогда $MY_n = 0$, $DY_n = 1$.

Теорема 1.4. Пусть выполнены условия приведённые выше, тогда последовательность Y_n *слабо сходится* к случайной величине $Z \sim N(0,1)$ т. е.

$$\forall x \in \Re, \quad F_{Y_n}(x) \xrightarrow[n\to\infty]{} \Phi(x), \quad \text{где}$$

- $\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-t^2/2}\, dt$

Замечание. В этом случае говорят, что *случайная величина Y_n* имеет *асимптотически стандартное нормально распределение.*

Замечание. Для последовательности X_i удовлетворяющей *центральной предельной теореме* также выполнены условия *закона больших чисел в форме Чебышева.* Поэтому из закона больших чисел $\Rightarrow$ $\overline{X}_n \xrightarrow[n\to\infty]{} m$. *Центральная предельная теорема* уточняет характер этой сходимости.

Замечание. *Центральная предельная теорема* иллюстрирует особую роль *нормального распределения.* Все естественные процессы, протекание которых обусловлено многочисленными случайными факторами (независимыми) "в среднем" имеют нормальное распределение.

Теорема 1.5 (ЦПТ Муавра-Лапласа)**.** Пусть

- проводится $n \gg 1$ испытаний по *схеме Бернулли* с вероятностью успеха p ($q = 1-p$ — вероятность неудачи);
- k — число успехов из n испытаний.

Тогда $\mathsf{P}\{k_1 \leq k \leq k_2\} \approx \Phi_0(x_2) - \Phi(x_1)$, где

- $x_i = \frac{k_i - np}{\sqrt{npq}}$, $i = \overline{1,2}$;
- $\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt$.

Доказательство. Рассмотрим $X_i = \begin{cases} 1, \text{в } i\text{-ом испытании успех} \\ 0, \text{иначе} \end{cases}$ при этом:

- X_i, $i \in \aleph$ — независимы
- $MX_i = p$, $DX_i = p\,q$
- $\sum_{i=1}^n X_i$ — число успехов в серии из n испытаний

Тогда имеем

$$\mathsf{P}\{k_1 \leq k \leq k_2\} = \mathsf{P}\{k_1 \leq \sum_{i=1}^n X_i \leq k_2\} =$$

применяем *центральную предельную теорему* к $x_1, \ldots, x_n$ (проверить)

$$= \mathsf{P}\left\{\frac{1}{n}k_1 - m \leq \frac{1}{n}\sum_{i=1}^n X_i - m \leq \frac{1}{n}k_2 - m\right\} =$$

$$= \mathsf{P}\left\{\frac{1/n\ k_1 - m}{\sqrt{p\,q}/\sqrt{n}} \leq \frac{1/n\ \sum_{i=1}^n X_i - m}{\sqrt{p\,q}/\sqrt{n}} \leq \frac{1/n\ k_2 - m}{\sqrt{p\,q}/\sqrt{n}}\right\} =$$

$\dfrac{1/n\ \sum_{i=1}^n X_i - m}{\sqrt{p\,q}/\sqrt{n}} = Y_n \sim N(0,1)$ — приближённо, так как $n \gg 1$

$$= \mathsf{P}\{x_1 \leq Y_n \leq n_2\} \approx \Phi_0(x_2) - \Phi_0(x_1)$$

$$x_i = \frac{1/n\ k_i - m}{\sqrt{p\,q}/\sqrt{n}} = \frac{k_i - mp}{\sqrt{npq}}$$

$\square$

Пример 1.2. В эксперименте Пиреосса о подбрасывании монеты из 24000 бросков герб выпал 12012 раз. Какова вероятность того, что при повторном испытании отклонение относительной частоты успеха окажется таким же или больше.

Решение. Используем *схему Бернулли*

- $n = 24000$, $p = q = 1/2$

- $A = \{\text{отклонение окажется не меньше}\}$
- $\overline{A} = \{\text{отклонение окажется меньше}\}$

Тогда имеем

$$\mathsf{P}(A) = 1 - \mathsf{P}(\overline{A}) = 1 - \mathsf{P}\{12000 - 12 < k < 12000 + 12\} = 1 - \mathsf{P}\{11988 < k < 12012\} \approx$$

Муавра-Лапласа

$$\approx 1 - \big[\Phi_0(x_2) - \Phi_0(x_1)\big] = 1 - 2\Phi_0(0.155) \approx -0.877$$

$$x_1 = \frac{11988 - 12000 \cdot 1/2}{\sqrt{24000 \cdot 1/2 \cdot 1/2}} \approx -0.155; \qquad x_2 \approx 0.155$$

Математическая Статистика

2 Основные понятия выборочной теории

2.1 Основные определения

Теория вероятностей является одной из областей "чистой" математики, которая строится дедуктивно, исходя из вполне определённых аксиом.

Математическая статистика является разделом прикладной математики, которая строится индуктивно: от наблюдения к гипотезе, при этом аргументация основана на выводах теории вероятностей.

Типовая задача теории вероятностей При одном подбрасывании монеты вероятность выпадения герба равна p. Какова вероятность того, что при n подбрасываниях герб выпадет m раз?

Типовая задача математической статистики При n подбрасываниях монеты герб выпал m раз. Чему равна вероятность p выпадания герба при одном подбрасывании?

Основная задача математической статистики Разработка методов получения обоснованных выводов о массовых явлениях и процессах по результатом наблюдений или экспериментов. Эти выводы относятся не к результатам отдельных экспериментов, а предоставляют собой вероятностные характеристики случайных явлений.

"Общая" задача математической статистики X является случайной величиной, законы распределения которой не известны. Требуется по данным наблюдений (или экспериментов) за -0-0-0- *случайной величины* X сделать выводы о её законе распределения

Определение 2.1. Множество возможных значений случайной величины X называется *генеральной совокупностью*

Определение 2.2. *Случайной выборкой* из генеральной совокупности X называется *вектор*

$$\overrightarrow{X_n} = (X_1, \dots, X_n), \quad \text{где} \tag{5}$$

- X_i, $i \in \aleph$ — независимые (в совокупности) случайные величины, имеющие то же распределение, что и генеральная совокупность X.

Замечание. При этом n называется *объёмом случайной выборки* $\overrightarrow{X_n}$.

Замечание. Пусть $F(t)$ — *функция распределения случайной величины* X. Тогда *функция распределения случайной выборки* $\overrightarrow{X_n}$ имеет вид

$$F_{\overrightarrow{X_n}(x_1,\dots,x_n)} = \mathsf{P}\{X_1 < x_1, \dots, X_n < x_n\} = \\ = \mathsf{P}\{X_1 < x_1\} \cdot \ldots \cdot \mathsf{P}\{X_n < x_n\} = F(x_1) \cdot \ldots \cdot F(x_n) \tag{6}$$

Определение 2.3. Любую возможную реализацию $\vec{x}_n = (x_1, \dots, x_n)$ случайной выборки $\overrightarrow{X_n}$ называют *выборкой объёма* n для *случайной величины* X.

Замечание. При этом x_k — k-ый элемент выборки $\vec{x}_n$.

Определение 2.4. Множество всех возможных значений случайного вектора $\overrightarrow{X_n}$ называют *выборочным пространством*

Определение 2.5. Любую числовую функцию $g(\overrightarrow{X_n})$ будем называть *статистикой* (или *выборочной характеристикой*).

Замечание. Значение $g(\vec{x}_n)$ статистики $g(\overrightarrow{X_n})$ называют *выборочным значением статистики g.*

Замечание. Пусть $\vec{x}_n$ — реализация случайной выборки $\overrightarrow{X_n}$. Это позволяет моделировать случайную величину X (закон распределения которой не известен) дискретной случайной величиной, ряд распределения которой имеет вид

Значение	x_1	$\dots$	x_i	$\dots$	x_n
Вероятность	$^1/_n$	$\dots$	$^1/_n$	$\dots$	$^1/_n$

Математическое ожидание такой случайной величины

$$m = \sum_{i=1}^{n} \frac{1}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Дисперсия

$$\sigma^2 = \sum_{i=1}^{n} (x_i - m)^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2$$

Эти соображения приводят к следующему определениям.

Определение 2.6. *Выборочным средним (выборочным математическим ожиданием)* называют статистику

$$\dot{\mu}_1(\overrightarrow{X_n}) = \overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7}$$

Определение 2.7. *Выборочной дисперсией* называют статистику

$$\hat{\sigma}^2(\overrightarrow{X_n}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \tag{8}$$

Определение 2.8. *Выборочным моментом j-го порядка* называется статистика

$$\hat{\mu}_j(\overrightarrow{X_n}) = \frac{1}{n} \sum_{i=1}^{n} X_i^j \tag{9}$$

Определение 2.9. *выборочным центральным моментом j-го порядка* называется статистика

$$\hat{\nu}_j(\overrightarrow{X_n}) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^j$$

Замечание.

$$\hat{\sigma}^2 = \hat{\nu}_2 \tag{10}$$

2.2 Предварительная обработка результатов экспериментов

2.2.1 Вариационный ряд

Рассмотрим $\vec{x}_n = (x_1, \ldots, x_n)$ — реализация случайной выборки. Упорядочим значения $x_1, \ldots, x_n$, расположив их в порядке не убывания

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}; \quad x_{(1)} = \min\{x_1, \ldots, x_n\}; \quad x_{(n)} = \max\{x_1, \ldots, x_n\}.$$

Определение 2.10. Последовательность $x_{(1)}, \ldots, x_{(n)}$, которая удовлетворяет условию

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

называется *вариационным рядом выборки* $\vec{x}_n$.

Замечание. $x_{(i)}$ — i-ый член вариационного ряда.

Определение 2.11. *Вариационным рядом случайной выборки* $\overrightarrow{X_n}$ называется последовательность случайных величин $X_1, \ldots, X_n$, где случайная величина $X_{(i)}$ для каждой реализации $\vec{x}_n$ случайной выборки $\overrightarrow{X_n}$ принимает значение, равное значению $x_{(i)}$.

Замечание. $\mathsf{P}\{X_{(i)} \le X_{(i+1)}\} = 1$

Замечание. Пусть $F(x)$ — функция распределения случайной величины X. Тогда

$$F_{X_n}(x) = \mathsf{P}\{X_{(n)} < x\} = \mathsf{P}\big\{\{X_1 < x\} \cdot \ldots \cdot \{X_n < x\}\big\} =$$

так как события независимы

$$= \mathsf{P}\,\underbrace{\{X_1 < x\}}_{F_{X_1}(x)=F(x)} \cdot \ldots \cdot \mathsf{P}\{X_n < x\} = \big[F(x)\big]^n$$

$$\begin{aligned} F_{X_{(1)}}(x) &= \mathsf{P}\{X_{(1)} < x\} = 1 - \mathsf{P}\{X_{(1)} \ge x\} = 1 - \mathsf{P}\{X_1 \ge x, \ldots, X_n \ge x\} = \\ &= 1 - \mathsf{P}\{X_1 \ge x\} \cdot \ldots \cdot \mathsf{P}\{X_n \ge x\} = 1 - \big(1 - \mathsf{P}\{X_1 \le x\}\big) \cdot \ldots \cdot \big(1 - \mathsf{P}\{X_n \le x\}\big) = \\ &= 1 - \big(1 - F(x)\big)^n \end{aligned}$$

2.2.2 Статический ряд

Среди элементов выборки $\vec{x}_n = (x_1, \ldots, x_n)$ могут встретиться одинаковые. Это может иметь место, например, если генеральная совокупность X является дискретной случайной величиной или если X непрерывная случайная величина, но при измерениях имело место округление.

Предположим, что среди значений выборки $\vec{x}_n = (x_1, \ldots, x_n)$ выделены m попарно различных значений.

$$z_1 < z_2 < \ldots < z_m, \quad \text{так что} \quad \forall i \in \{1, \ldots, n\}, \quad \exists j \in \{1, \ldots, m\}, \quad x_i = z_{(j)}$$

Пусть среди компонент вектора x_n ровно n_j компонент приняли значение z_j, $j = \overline{1, m}$. Тогда таблицу

$z_{(1)}$	$\ldots$	$z_{(j)}$	$\ldots$	$z_{(m)}$
n_1	$\ldots$	n_j	$\ldots$	n_m

, $\quad \sum_{j=1}^{m} n_j = n$

называют *статическим рядом для выборки* $\vec{x}_n$. При этом n_j называют *частотой значения* $z_{(j)}$, а величина $\frac{n_j}{n}$ — *относительной частотой значения* $z_{(j)}$.

2.2.3 Эмпирическая функция распределения

- $\overrightarrow{X_n} = (X_1, \ldots, X_n)$ — случайная выборка
- $\vec{x}_n = (x_1, \ldots, x_n)$ — реализация случайной выборки $\overrightarrow{X_n}$
- $(X_1, \ldots, X_n)$ — независимые случайные величины;
- $n(x, \vec{x}_n)$ — количество элементов выборки $\vec{x}_n$, которые меньше x

Определение 2.12. *Эмпирической функцией распределения* называют функцию

$$F_n \colon \Re \to \Re, \quad F_n(x) = \frac{n(x, \vec{x}_n)}{n}.$$

Замечание. $F_n(x)$ обдалает всеми свойствами функции распределения. При этом она кусочно-постоянна и принимает значения $0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{(n-1)}{n}, 1$

Замечание. Если все элементы вектора $\vec{x}_n$ различны, то

$$F_n(x) = \begin{cases} 0, & x \le x_{(1)}; \\ \frac{i}{n}, & x_{(i)} < x \le x_{(i+1)},\ i = \overline{1, n-1}; \\ 1, & x > x_{(n)}. \end{cases}$$

Замечание. Эмпирическая функция распределения позволяет интерпретировать выборку $\vec{x}_n$ как реализацию дискретной случайной величины $\widetilde{X}$ ряд распределения которой

$\widetilde{X}$	$x_{(1)}$	$\ldots$	$x_{(n)}$
P	$1/n$	$\ldots$	$1/n$

В дальнейшем это позволит рассматривать числовые характеристики случайной величины $\widetilde{X}$ как приближённые значения числовых характеристик случайной величины X.

2.2.4 Выборочная функция распределения

- $n(x, \overrightarrow{X_n})$ — функция, которая для каждой реализации $\vec{x}_n$ случайной выборки $\overrightarrow{X_n}$ принимает значение, равное числу элементов $\vec{x}_n$ которые меньше x

Определение 2.13. *Выборочной функцией распределения* называют функцию

$$\widehat{F}(x, \overrightarrow{X_n}) = \frac{n(x, \overrightarrow{X_n})}{n} \tag{11}$$

Замечание. Зафиксируем некоторе значение $x \in \Re$. Тогда для этого x, $\widehat{F}(x, \overrightarrow{X_n})$ является *функцией случайной выборки* $\overrightarrow{X_n}$ $\Rightarrow$ является случайной величиной.

Замечание. При каждом фиксированном $x \in \Re$ случайная величина $\widehat{F}(x, \overrightarrow{X_n})$ может принимать значения

$$0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1.$$

Обозначим $p = \mathsf{P}\{X_i < x\}$ — не зависит от i, так как все X_i одинаково распределены. blablabla

$$\mathsf{P}\left\{\widehat{F}(x, \overrightarrow{X_n}) = \frac{k}{n}\right\} = \mathsf{P}\left\{\frac{n(x, \overrightarrow{X_n})}{n} = \frac{k}{n}\right\} = \mathsf{P}\big\{n(x, \overrightarrow{X_n}) = k\big\} =$$

В векторе $\overrightarrow{X_n}$ ровно k компонент принимающих значение $< x$, при этом $X_1, \ldots, X_n$ — независимы, $\mathsf{P}\{X_i, x\} - p$

bla bla bla

$$= C_n^k p^k (1-p)^{n-k}$$

bla bla bla

Схема Бернулли, $\Rightarrow$ $\hat{F}(x, \overrightarrow{X_n}) \sim B(n, p)$ — биноминальная случайная величина

Теорема 2.1. $\forall x \in \Re$ последовательность $\widehat{F}(x, \overrightarrow{X_n})$ сходится по вероятности к значению $F_X(x)$ bla bla bla то есть

$$\widehat{F}(x, \overrightarrow{X_n}) \xrightarrow[n\to\infty]{\mathsf{P}} F_X(x) \tag{12}$$

Доказательство. При каждом фиксированном $x \in \Re$ величина $\widehat{F}(x, \overrightarrow{X_n})$ равна относительной (наблюдённой) частоте реализации реализации успеха в серии из n испытаний по *схеме Бернулли* (успех осуществления событий $\{X < x\}$). В соответствии с *законом больших чисел в форме Бернулли*

$$\widehat{F}(x, \overrightarrow{X_n}) \xrightarrow[n\to\infty]{\mathsf{P}} F_X(x)$$

□

2.2.5 Интервальный статический ряд

Выше было введено понятие *статического ряда.* Однако если число наблюдений велико ($n > 50$), то их группируют не только в виде статического ряда, но и в виде интервального статического ряда. Для этого отрезок $J = [x_{(1)}, x_{(n)}]$ разбивают на m равновеликих интервалов

$$\begin{aligned}
&J_i = \big[x_{(1)} + (i-1)\Delta, x_{(i)} + i\Delta\big)\,, \quad i = \overline{1, m-1};\\
&J_m = \big[x_{(1)} + (m-1)\Delta, x_{(1)} + m\Delta\big]\,;\\
&\Delta = \frac{|J|}{m}.
\end{aligned}$$

Замечание. При выборе m обычно используют формулу

$$m = [\log_2 n] + 1, \quad \text{где} \tag{13}$$

- $[\alpha]$ — целая часть от α

Определение 2.14. *Интервальным статическим рядом* называют таблицу

J_1	...	J_m
n_1	...	n_m

- $n_1 + \cdots + n_m = n$,
- n_i — количество элементов выборки $\vec{x}_n$, принадлежащих множеству J_i, $i = \overline{1, m}$

2.2.6 Эмпирическая плотность

Пусть для данной выборки $\vec{x}_n$ построим интервальный статический ряд

Определение 2.15. *Эмпирической плотностью распределения случайной величины* X называют функцию

$$f_n(x) = \begin{cases} \frac{n_i}{n\Delta}, & x \in J_i; \\ 0, & \text{иначе.} \end{cases} \tag{14}$$

Определение 2.16. График функции $f_n(x)$ называют гистограммой. (гистограмма)

Замечание.

$$\int_{-\infty}^{+\infty} f_n(x)\,dx = \sum_{i=1}^{n} (\text{площадь прямоугольников}) = \sum_{i=1}^{n} \frac{n_i}{n\Delta}\Delta = \frac{1}{n}\sum_{i=1}^{n} n_i = \frac{n}{n} = 1$$

Замечание. Функция $f_n(x)$ является статическим аналогом плотности распределения случайной величины можно показать, что для непрерывной случайной величины X

$$\forall x \in \Re, \quad f_n(x) \xrightarrow[n\to\infty]{\mathsf{P}} f_X(x)$$

то есть при больших $n \;\Rightarrow\; f_n(x) \approx f(x)$

2.2.7 Полигональная частота

Пусть для данной выборки $\vec{x}_n$ построена гистограмма (рисунок)

Определение 2.17. *Полигоном частот* называют ломанную звенья которой соединяют середины верхних сторон прямоугольников гистограммы. (рисунок)