ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Выборочные характеристики

Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Введение

В предыдущих лекциях мы познакомились с теми задачами, которые решает теория вероятностей, а также обсудили вопросы, которыми занимается статистика. Резюмируя, наверное можно сказать, что основная задача теории вероятностей – это прогноз, а именно: по построенной модели определить вероятность интересующего события; распределение той или иной случайной величины, системы случайных величин, их характеристики; распределение процесса случайных величин, изменяющегося во времени и проч. Основной задачей математической статистики можно считать обратную задачу – задачу выявления закономерностей в случайных данных.

1.1 Основные понятия и задачи математической статистики

В предыдущей лекции мы уже мимоходом вводили такие понятия, как выборка, генеральная совокупность и многие другие. Давайте теперь остановимся на них подробнее и несколько структурируем полученные знания. Вообще, в предыдущей лекции мы мельком обсудили все те вопросы, которые мы будем обсуждать в оставшихся лекциях курса. Но позволим себе несколько повториться и начать с базового понятия – понятия выборки.

Предположим, что некоторый эксперимент повторяется, скажем, n раз, и в результате эксперимента мы измеряем какую-то числовую характеристику. Можно считать, что в качестве результата эксперимента мы наблюдаем значение некоторой случайной величины $\xi : \Omega \to \mathbb{R}$. Вероятностное пространство будем считать заданным, оно нас интересовать не будет. Итак, проведя испытание n раз, мы получили набор чисел $X_1, X_2, ..., X_n$ – набор значений нашей случайной величины ξ в первом, втором и так далее испытаниях, соответственно.

На самом деле, как мы уже отмечали, величины $X_1, ..., X_n$ называть числами не очень-то хорошо. Дело в том, что проведя испытание еще n раз, мы получим уже другой набор чисел $X_1, ..., X_n$ – другие значения случайной величины ξ. Значит, до проведения конкретного испытания величины X_i (одинаково распределенные с ξ) резонно называть случайными величинами, а после испытания – числами, или значениями случайной величины ξ в конкретном наборе испытаний.

Определение 1.1.1 *Выборкой $X = (X_1, ..., X_n)$ объема n называется набор из n независимых одинаково распределенных случайных величин $X_1, ..., X_n$, имеющих распределение такое же, как и у случайной величины ξ.*

Замечание 1.1.1 *Рассматриваемая случайная величина ξ часто называется генеральной совокупностью.*

Как уже отмечалось, в теории вероятностей мы, как правило, работали с известными распределениями случайных величин. Зная их, мы строили различные характеристики случайных величин, говорили об их зависимости, независимости и др. В статистике мы будем исходить из выборки $X = (X_1, ..., X_n)$, которая получена в результате n-кратного повторения случайного эксперимента. Нашей задачей будет «узнать как можно больше» о распределении и различных параметрах рассматриваемой случайной величины.

Как мы уже говорили, традиционно в статистике рассматривается три типа задач. Первый тип задач – это задачи оценивания неизвестного параметра (задачи точечного оценивания). Мы достаточно подробно говорили о точечных оценках в предыдущей лекции: о выборочном среднем $\overline{X}$, как об оценке математического ожидания $\mathsf{E}\xi$ генеральной совокупности ξ, а также о выборочной медиане $\widehat{\mathsf{med}\ \xi}$, как оценке математического ожидания случайной величины ξ, имеющей распределение N_{a,σ^2}.

Второй тип задач, тесно примыкающий к первому, – это задачи нахождения интервала, в котором с большой вероятностью находится неизвестный параметр (задачи интервального оценивания).

Часто бывает возможно высказать некоторые предположения о наблюдаемом распределении или о его свойствах. В этом случае на основе опытных данных требуется подтвердить или опровергнуть эти предположения, называемые гипотезами. Тем самым, третий тип задач – задачи проверки гипотез о неизвестном распределении выборки, или о его параметрах. О каждом типе задач мы будем подробно говорить в дальнейших лекциях.

2 Выборочные характеристики

Давайте начнем с, наверное, самой сложной, но практически значимой ситуации – ситуации, когда мы ничего не знаем, кроме некоторой выборки (с так называемой описательной статистики). Как можно, имея эту выборку, сделать выводы о каких-то характеристиках распределения? Ведь само по себе распределение характеризуется, например, своей функцией распределения, значит неплохо бы научиться ее оценивать. Кроме того, распределение может быть охарактеризовано своей плотностью или таблицей распределения. Однако и это – не все характеристики. Как мы знаем, математическое ожидание, дисперсия и моменты старших порядков тоже играют важную роль, а значит и их хотелось бы оценить. Начнем по порядку.

2.1 Выборочное распределение

Пусть выборка реализуется на конкретном элементарном исходе ω_0, а в нашем распоряжении оказывается n чисел $X_1 = X_1(\omega_0)$, ..., $X_n = X_n(\omega_0)$ (числа могут и совпадать). Пусть теперь вспомогательная случайная величина ξ^* принимает значения $X_1, ..., X_n$ с равными вероятностями (например, случайная величина ξ^* реализуется с помощью правильного n-гранного кубика). Наша новая случайная величина определена, вообще говоря, на другом вероятностном пространстве, нежели исходная случайная величина ξ (на пространстве, связанным с бросанием кубика), поэтому вероятностную меру на этом пространстве будем обозначать $\widetilde{\mathsf{P}}$, а моменты $\widetilde{\mathsf{E}}$, $\widetilde{\mathsf{D}}$ и так далее.

Таблица распределения случайной величины ξ^* имеет очень простой вид

ξ^*	X_1	X_2	$\dots$	X_n
$\widetilde{\mathsf{P}}$	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

.

Напомним, что, как было уже сказано во вводной лекции, распределение разыгранной случайной величины называют эмпирическим, а при увеличении объема выборки оно сходится (по вероятности) к неизвестному истинному распределению. Значит, все вероятностные характеристики случайной величины ξ^* разумно использовать для оценки соответствующих характеристики генеральной совокупности.

Напомним, что функция распределения эмпирической случайной определяется соотношением

$$F_n^*(t) = \frac{\text{количество } X_i \in (-\infty, t)}{n} = \sum_{i:\ X_i<t} \frac{1}{n},$$

и ее разумно рассматривать, как оценку истинной функции распределения.

По аналогичным причинам с эмпирической случайной величиной связывают так называемые выборочные характеристики – ее собственные вероятностные характеристики такие, как: математическое ожидание, дисперсия, моменты старших порядков и другие. Математическое ожидание величины ξ^* равно

$$\widetilde{\mathsf{E}}\xi^* = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n},$$

дисперсия $\widetilde{\mathsf{D}}\xi^*$ же, как обычно, равна $\widetilde{\mathsf{E}}(\xi^* - \widetilde{\mathsf{E}}\xi^*)^2$ или

$$\widetilde{\mathsf{D}}\xi^* = S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widetilde{\mathsf{E}}\xi^*\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

момент k-ого порядка равен

$$\widetilde{\mathsf{E}}(\xi^*)^k = \overline{X^k} = \frac{X_1^k + ... + X_n^k}{n},$$

центральный момент k-ого порядка равен

$$\widetilde{\mathsf{E}}(\xi^* - \widetilde{\mathsf{E}}\xi^*)^k = \widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widetilde{\mathsf{E}}\xi^*\right)^k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k,$$

и вообще, математическое ожидание функции g от случайной величины ξ^* равно

$$\widetilde{\mathsf{E}}g(\xi^*) = \overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Кстати, дисперсия может быть вычислена и как

$$\mathsf{D}\xi^* = S^2 = \overline{X^2} - \left(\overline{X}\right)^2$$

Если теперь позволить исходу ω_0 меняться, то все введенные выше характеристики: функция распределения $F_n^*(t)$, математическое ожидание $\overline{X}$, дисперсия S^2, моменты k-ого порядка $\overline{X^k}$, центральные моменты k-ого порядка m_k, а также математическое ожидание функции от случайной величины $\overline{g(X)}$, станут случайными величинами, так как теперь они будут функциями от случайных величин $X_1, X_2, ..., X_n$.

Еще раз повторим, что введенные характеристики выборочного распределения используют для оценки по конкретной реализации выборки соответствующих параметров нам неизвестного истинного распределения. Причиной служит закон больших чисел, ведь все построенные нами выборочные характеристики являются ни чем иным, как средним арифметическим независимых и одинаково распределенных случайных величин, то есть с ростом объема выборки n сходятся (по вероятности) к истинным характеристикам случайной величины ξ: ее функции распределения, математическому ожиданию, дисперсии и проч. Давайте теперь будем разбираться с каждой из оценок, как и с проблемами, по одиночке.

2.2 Эмпирическая функция распределения

Перед тем, как перейти к основным понятиям данного пункта, напомним еще одно важное понятие, связываемое с выборкой – понятие вариационного ряда.

Определение 2.2.1 *Пусть имеется выборка $X = (X_1, ..., X_n)$. Если элементы выборки упорядочить по возрастанию, то новый набор случайных величин, удовлетворяющий неравенствам*

$$X_{(1)} \le X_{(2)} \le ... \le X_{(n)},$$

называется вариационным рядом.

Пример 2.2.1 *Рассмотрим такую синтетическую выборку, собранную молодым человеком Петей, и показывающую время (в минутах), на которое его девушка Даша опаздывала на свиданиях:*

$$X = (0, 11, 2, 3, 9, 2, 8, 6, 3.4, 8, 7.5, 9, 4, 8, 6).$$

Тогда вариационный ряд, построенный по этой выборке, таков:

$$(0, 2, 2, 3, 3.4, 4, 6, 6, 7.5, 8, 8, 8, 9, 9, 11).$$

Можно сразу сделать вывод, что не опоздала Даша всего раз, дважды опоздала на две минуты, один раз на три ну и так далее. По написанному вариационному ряду удобно строить эмпирическое распределение случайной величины ξ^*. Группируя одинаковые значения, получаем

ξ^*	0	2	3	3.4	4	6	7.5	8	9	11
$\widetilde{\mathsf{P}}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

.

Понятно, что $X_{(1)} = \min\{X_1, ..., X_n\}$, $X_{(n)} = \max\{X_1, ..., X_n\}$. Напомним также, что k-ый член вариационного ряда часто называют k-ой порядковой статистикой.

Как мы уже неоднократно отмечали, одной из основных характеристик случайной величины является ее функция распределения. Напомним, что она может быть вычислена по формуле

$$F_n^*(t) = \frac{\text{количество } X_i \in (-\infty, t)}{n} = \sum_{i:\ X_i<t} \frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{I}(X_i < t),$$

где функция

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

называется индикатором события $\{X_i < t\}$, при каждом значении t является случайной величиной и имеет распределение Бернулли B_p с параметром $p = \mathsf{P}(X_i < t) = F_\xi(t)$.

Пример 2.2.2 *Вернемся к только что рассмотренному примеру. Напомним, что выборочное распределение, построенное по обсуждаемой ранее выборке, имеет вид*

ξ^*	0	2	3	3.4	4	6	7.5	8	9	11
$\widetilde{\mathsf{P}}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{3}{15}$	$\frac{2}{15}$	$\frac{1}{15}$

.

Тогда эмпирическая функция распределения равна

$$F_n^*(t) = \begin{cases} 0, & t \le 0 \\ 1/15, & 0 < t \le 2, \\ 3/15, & 2 < t \le 3, \\ 4/15, & 3 < t \le 3.4, \\ \dots & \dots, \\ 14/15, & 9 < t \le 11, \\ 1, & t > 11. \end{cases}$$

а ее график представлен на рисунке 1.

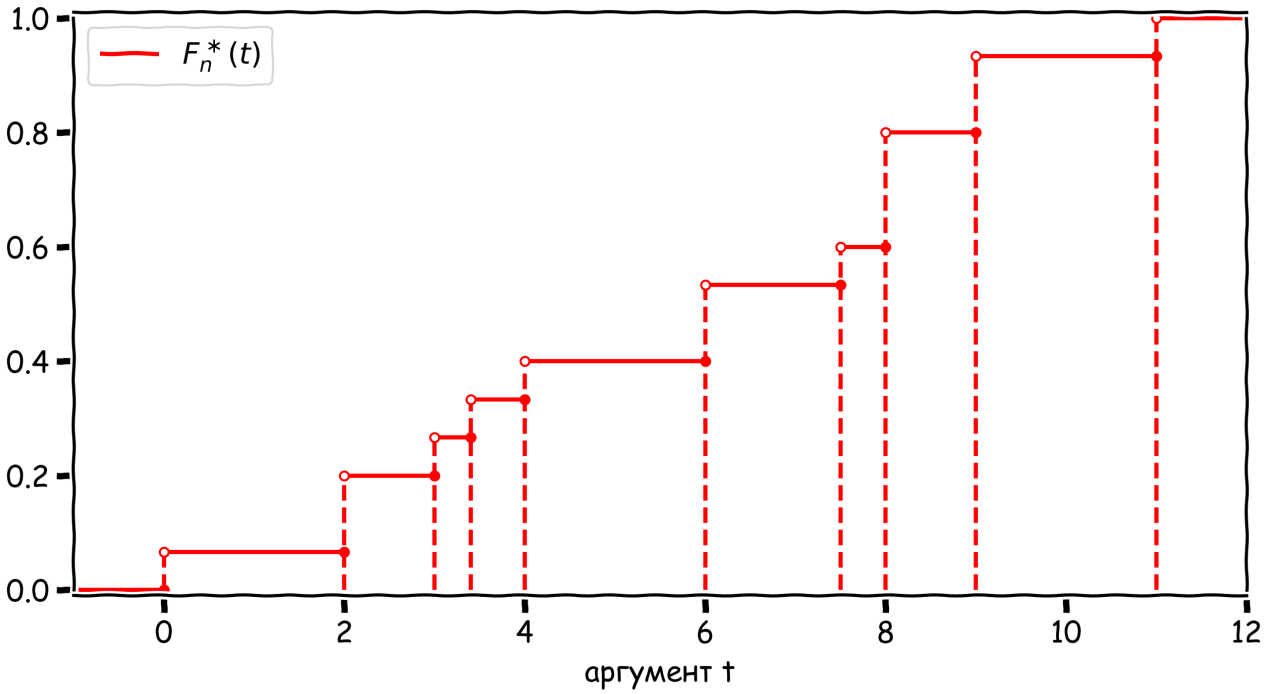

Рис. 1: Эмпирическая функция распределения, построенная по выборке

Скажем, как оценить вероятность события, что купленные для Даши круассан и стаканчик кофе не остынут? Иными словами, на языке вероятностей, какова вероятность события, что девушка опоздает не более, чем на 5 минут? Это мы умеем вычислять:

$$\widetilde{\mathsf{P}}(\xi^* \le 5) = F_n^*(5+0) = \frac{6}{15} = 0.4,$$

что, кстати, не так уж и много: надежнее, быть может, купить букет цветов, чтобы уж точно порадовать подругу, ну или прихватить с собой термосумку.

Давайте посмотрим на то, как ведет себя эмпирическая функция распределения на больших выборках из известного распределния. На рисунках 2, 3,

4 и 5 (объемы выборки 20, 40, 300 и 1000, соответсвенно) представлены графики эмпирических функций распределения, построенных по сгенерированному набору данных из равномерного распределения на отрезке $[1, 4]$ с помощью датчика псевдослучайных чисел. Синяя кривая – это истинная функция распределения. Видно, что с ростом объема выборки эмпирическая функция

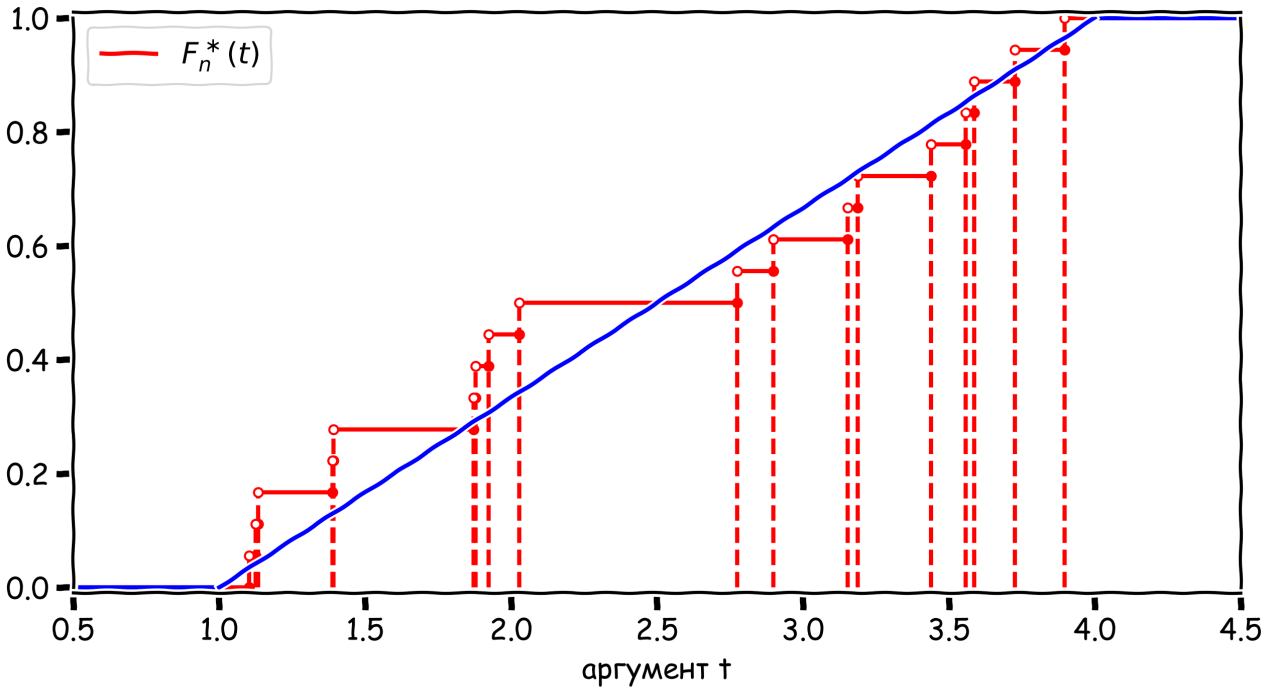

Рис. 2: Эмпирическая функция распределения, построенная по выборке объема 20 из равномерного распределения

распределения все ближе приближается к истинной.

Давайте возвратимся к некоторым теоретическим выводам. Легко понять, что эмпирическая функция распределения имеет скачки в точках вариационного ряда $X_{(i)}$, величина скачка равна $\frac{m}{n}$, где m – количество членов выборки, совпадающих с $X_{(i)}$. Вообще, по вариационному ряду ЭФР строится по следующему очевидному правилу

$$F_n^*(t) = \begin{cases} 0, & t \le X_{(1)}, \\ \frac{m}{n}, & X_{(m)} < t \le X_{(m+1)}, \\ 1, & t > X_{(n)}. \end{cases}$$

Естественным образом возникает вопрос: насколько хорошо введенная нами эмпирическая функция распределения приближает истинную? Ответим на этот вопрос следующей теоремой.

Теорема 2.2.1 (Основные свойства ЭФР) *Пусть $X_1, ..., X_n$ – выборка из генеральной совокупности ξ с функцией распределения F_ξ. Тогда эмпирическая функция распределения $F_n^*(t)$, построенная по этой выборке, удовлетворяет следующим свойствам:*

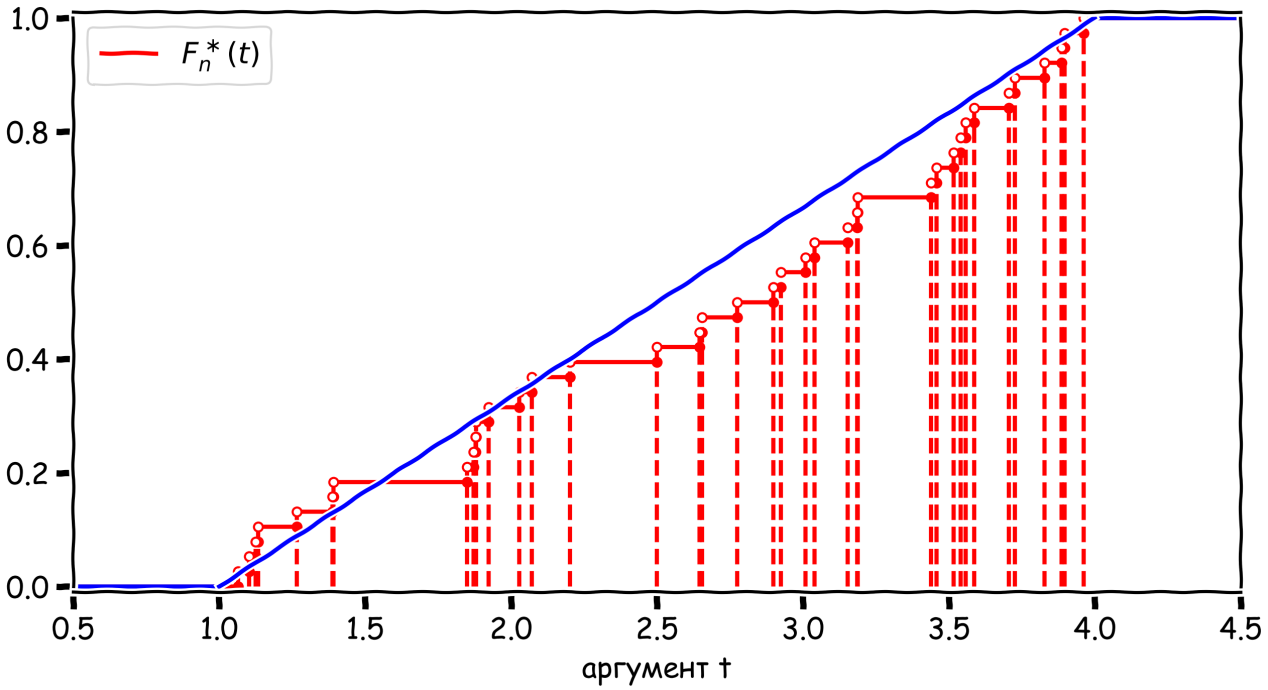

Рис. 3: Эмпирическая функция распределения, построенная по выборке объема 40 из равномерного распределения

1. Эмпирическая функция распределения является состоятельной оценкой F_ξ*, то есть*

$$F_n^*(t) \xrightarrow[n\to\infty]{\mathsf{P}} F_\xi(t), \quad \forall t \in \mathbb{R}.$$

2. Эмпирическая функция распределения является несмещенной оценкой F_ξ*, то есть*

$$\mathsf{E}F_n^*(t) = F_\xi(t).$$

3. Дисперсия эмпирической функции распределения равна

$$\mathsf{D}F_n^*(t) = \frac{F_\xi(t)(1-F_\xi(t))}{n}.$$

4. Эмпирическая функция распределения при $F_\xi(t) \neq 0,\ F_\xi(t) \neq 1$ *является асимптотически нормальной оценкой* F_ξ*, то есть*

$$Y_n = \sqrt{n}\frac{F_n^*(t) - F_\xi(t)}{\sqrt{\mathsf{D}F_n^*}} \xrightarrow[n\to\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Доказательство. 1. Напомним, что

$$F_n^*(t) = \frac{\sum\limits_{i=1}^{n} \mathsf{I}(X_i < t)}{n},$$

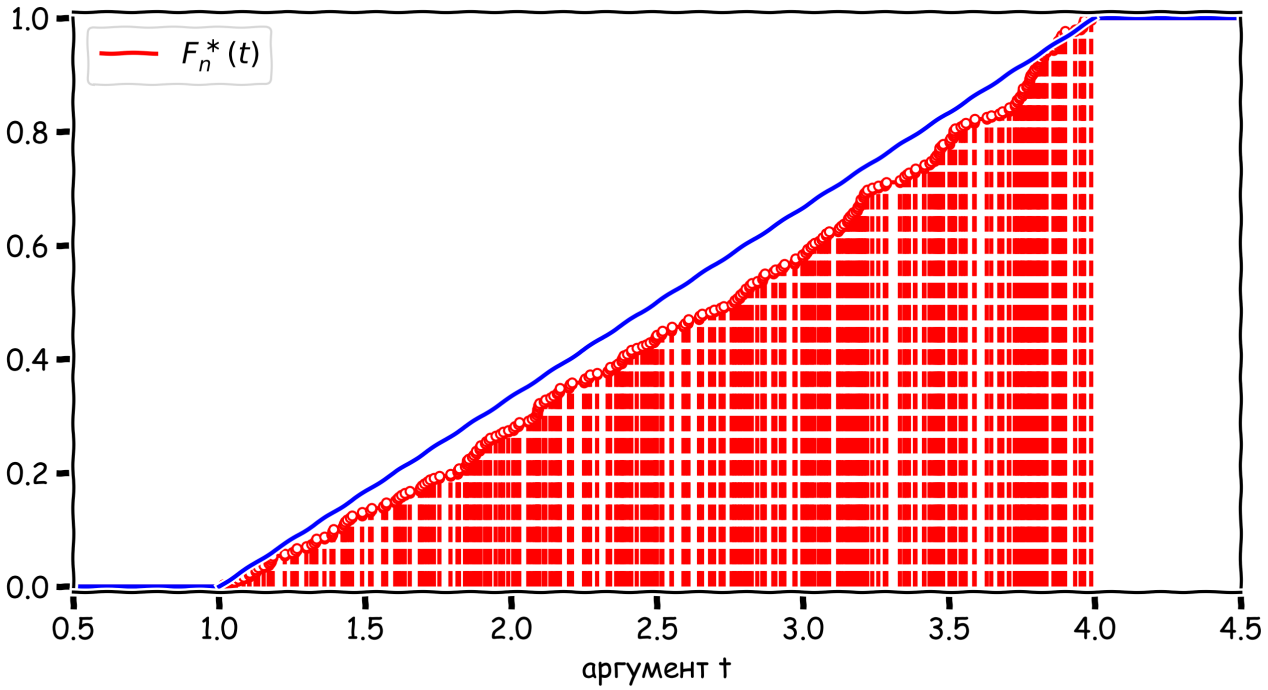

Рис. 4: Эмпирическая функция распределения, построенная по выборке объема 300 из равномерного распределения

и случайные величины $\mathsf{I}(X_i < t)$ одинаково распределены, независимы и имеют распределение Бернулли $\mathsf{B}_{F_\xi(t)}$. Тогда

$$\mathsf{E}\left(\mathsf{I}(X_i < t)\right) = F_\xi(t) < +\infty$$

и можно применить ЗБЧ в форме Хинчина из которого следует, что

$$F_n^*(t) = \frac{\sum\limits_{i=1}^{n} \mathsf{I}(X_i < t)}{n} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}\left(\mathsf{I}(X_1 < t)\right) = F_\xi(t).$$

2. Опираясь на рассуждения предыдущего пункта, имеем

$$\mathsf{E}F_n^*(t) = \mathsf{E}\left(\frac{\sum\limits_{i=1}^{n} \mathsf{I}(X_i < t)}{n}\right) = \frac{n\mathsf{E}\left(\mathsf{I}(X_1 < t)\right)}{n} = \mathsf{E}\left(\mathsf{I}(X_1 < t)\right) = F_\xi(t).$$

3. Случайные величины $\mathsf{I}(X_i < t)$ одинаково распределены, независимы и имеют распределение Бернулли, откуда

$$\mathsf{D}F_n^*(t) = \mathsf{D}\left(\frac{\sum\limits_{i=1}^{n} \mathsf{I}(X_i < t)}{n}\right) = \frac{\sum\limits_{i=1}^{n} \mathsf{D}\left(\mathsf{I}(X_i < t)\right)}{n^2} =$$

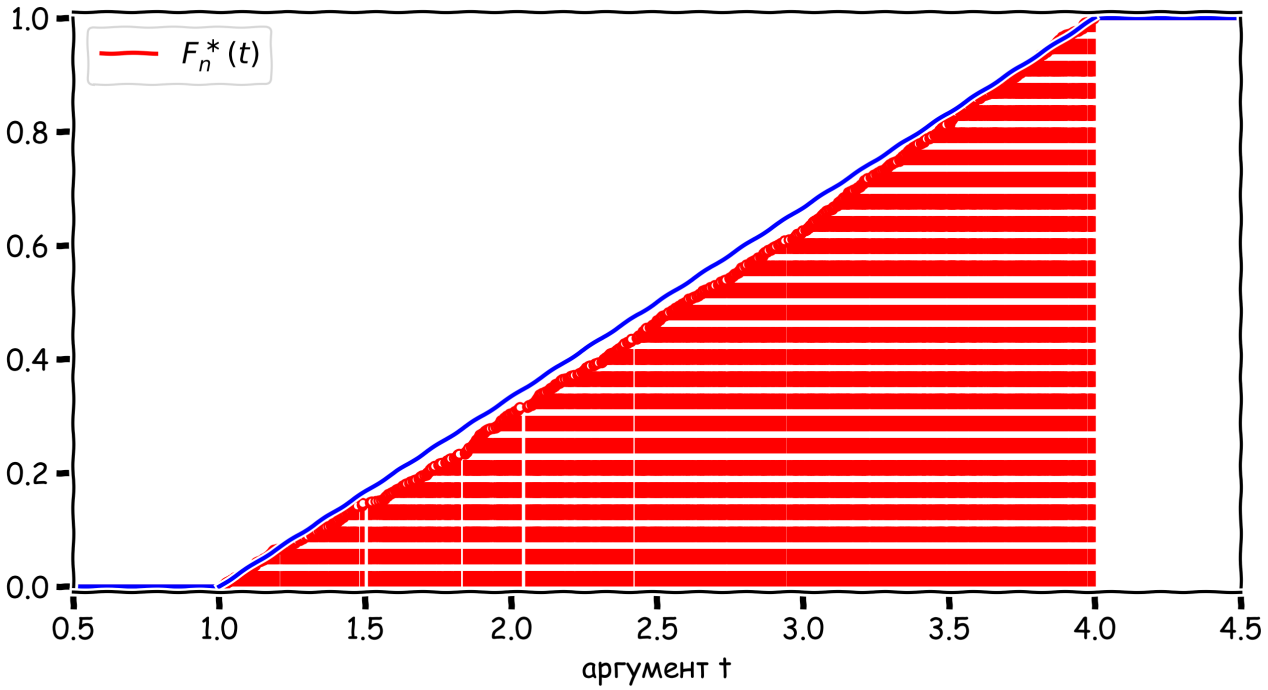

Рис. 5: Эмпирическая функция распределения, построенная по выборке объема 1000 из равномерного распределения

$$\frac{\mathsf{D}\left(\mathsf{I}(X_1 < t)\right)}{n} = \frac{F_\xi(t)(1 - F_\xi(t))}{n}.$$

4. Снова воспользуемся тем, что

$$F_n^*(t) = \frac{\sum\limits_{i=1}^{n} \mathsf{I}(X_i < t)}{n},$$

и, кроме того тем, что $\mathsf{E}\left(\mathsf{I}(X_i < t)\right) = F_\xi(t)$ и $\mathsf{D}\left(\mathsf{I}(X_i < t)\right) = F_\xi(t)(1 - F_\xi(t))$. Подставим это в левую часть доказываемого равенства, тогда

$$Y_n = \frac{\sum\limits_{i=1}^{n} \mathsf{I}(X_i < t) - n\mathsf{E}\left(\mathsf{I}(X_1 < t)\right)}{\sqrt{\mathsf{D}\left(\mathsf{I}(X_1 < t)\right)}}.$$

Осталось применить к рассматриваемой дроби центральную предельную теорему. □

Итак, первое свойство означает, что эмпирическая функция распределения действительно приближает истинную с ростом n, а значит использование ее в качестве оценки резонно. Второе свойство говорит о несмещенности рассматриваемой оценки, а последнее – об асимптотической нормальности (это нам пригодится для построения асимптотических доверительных интервалов

в дальнейшем). Мы вернемся к этому вопросу в последующих лекциях, когда строго будем обсуждать понятие и методы построения доверительных интервалов.

На самом деле, свойство состоятельности эмпирической функции распределения можно усилить. Сходимость в описанной теореме имеет даже (почти) равномерный характер, а именно справедлива следующая теорема Гливенко-Кантелли.

Теорема 2.2.2 (Гливенко-Кантелли) *Пусть $X_1, ..., X_n$ – выборка из генеральной совокупности ξ с функцией распределения F_ξ. Тогда эмпирическая функция распределения $F_n^*(t)$, построенная по этой выборке, сходится к истинной «почти равномерно», то есть*

$$\sup_{t\in\mathbb{R}} |F_n^*(t) - F_\xi(t)| \xrightarrow[n\to+\infty]{\mathsf{P}} 0.$$

Строгое доказательство этой теоремы требует обсуждения свойства непрерывности вероятностной меры, которое мы опустили в самом начале, поэтому и доказательство этой теоремы мы не приводим.

Еще одно важное свойства эмпирической функции распределения (которое мы будем использовать в дальнейшем) дает так называемая теорема Колмогорова.

Теорема 2.2.3 (Колмогорова) *Пусть $X_1, ..., X_n$ – выборка из генеральной совокупности ξ с непрерывной функцией распределения F_ξ. Тогда для эмпирической функция распределения $F_n^*(t)$ выполняется*

$$Y_n = \sqrt{n} \cdot \sup_{t\in\mathbb{R}} |F_n^*(t) - F_\xi(t)| \xrightarrow[n\to+\infty]{\mathsf{d}} Y,$$

где случайная величина Y имеет распределение Колмогорова с функцией распределения

$$F_Y(t) = \begin{cases} \sum\limits_{i=-\infty}^{+\infty} (-1)^i e^{-2i^2t^2}, & t \geq 0 \\ 0, & t < 0 \end{cases}.$$

На самом деле из этой теоремы следует то, что можно использовать уже сейчас, а именно: скорость сходимости эмпирической функции распределения к истинной в случае, когда последняя непрерывна, имеет порядок $\frac{1}{\sqrt{n}}$. Так, увеличивая объем выборки в 100 раз, погрешность уменьшается в примерно 10 раз.

2.3 Гистограмма

Еще одной важной характеристикой случайной величины наряду с функцией распределения является таблица распределения (для дискретных случайных величин) и плотность (для абсолютно непрерывных случайных величин). Зная плотность и ряд распределения, можно вычислять вероятности попадания в произвольные промежутки, различные числовые характеристики случайной величины, да много-много чего – мы все это подробно видели и «щупали» в лекциях по теории вероятностей. Кроме того, можно проверять: однородно или нет распределение, не рассматриваем ли мы какую-то смесь из разных распределений, корректно ли вообще работать с полученной выборкой и оценивать те или иные ее характеристики?

Эмпирическим аналогом этих объектов, то есть плотности и таблицы, является так называемая гистограмма. Гистограмма строится по группированным данным по следующему достаточно естественному алгоритму. В начале каким-то образом определяется множество значений случайной величины ξ. Эту область значений делят на некоторое количество непересекающихся отрезков, интервалов, полуинтервалов (все эти множества могут иметь разную длину). Пусть $A_1, ..., A_k$ – эти отрезки. Пусть ν_j – количество элементов выборки, попавших в отрезок A_j, $n = \sum\limits_{j=1}^{k} \nu_j$. Заменим истинную плотность на промежутке A_j длины l_j прямоугольником высоты $h_j = \frac{\nu_j}{nl_j}$. Заметим, что сумма площадей всех прямоугольников равна 1, а это значит, что полученная неотрицательная функция может трактоваться, как плотность распределения некоторой случайной величины. Действительно, площадь S_j j-ого прямоугольника равна

$$S_j = h_j \cdot l_j = \frac{\nu_j}{nl_j} \cdot l_j = \frac{\nu_j}{n},$$

а тогда суммарная площадь построенных прямоугольников равна

$$\sum_{j=1}^{k} S_j = \sum_{j=1}^{k} \frac{\nu_j}{n} = \frac{1}{n} \sum_{j=1}^{k} \nu_j = 1.$$

Таким образом построенная ступенчатая фигура, являющаяся объединением прямоугольников, называется гистограммой.

Пример 2.3.1 *Снова рассмотрим выборку, собранную Петей, из первого примера. Напомним, что вариационный ряд имеет вид*

$$(0, 2, 2, 3, 3.4, 4, 6, 6, 7.5, 8, 8, 8, 9, 9, 11).$$

Разобьем для начала отрезок $[0, 11]$ *на* $n = 4$ *части одинаковой длины (*$l = 2.75$*), тем самым получим множества*

$$A_1 = [0, 2.75),\ A_2 = [2.75, 5.5),\ A_3 = [5.5, 8.25),\ A_4 = [8.25, 11].$$

В множество A_1 *попадает 3 элемента выборки, значит* $\nu_1 = 3$*. Аналогичными рассуждениями получаем, что* $\nu_2 = 3$*,* $\nu_3 = 6$*,* $\nu_4 = 3$*. Гистограмма, отвечающая такому разбиению, представлена на рисунке 6. Если взять*

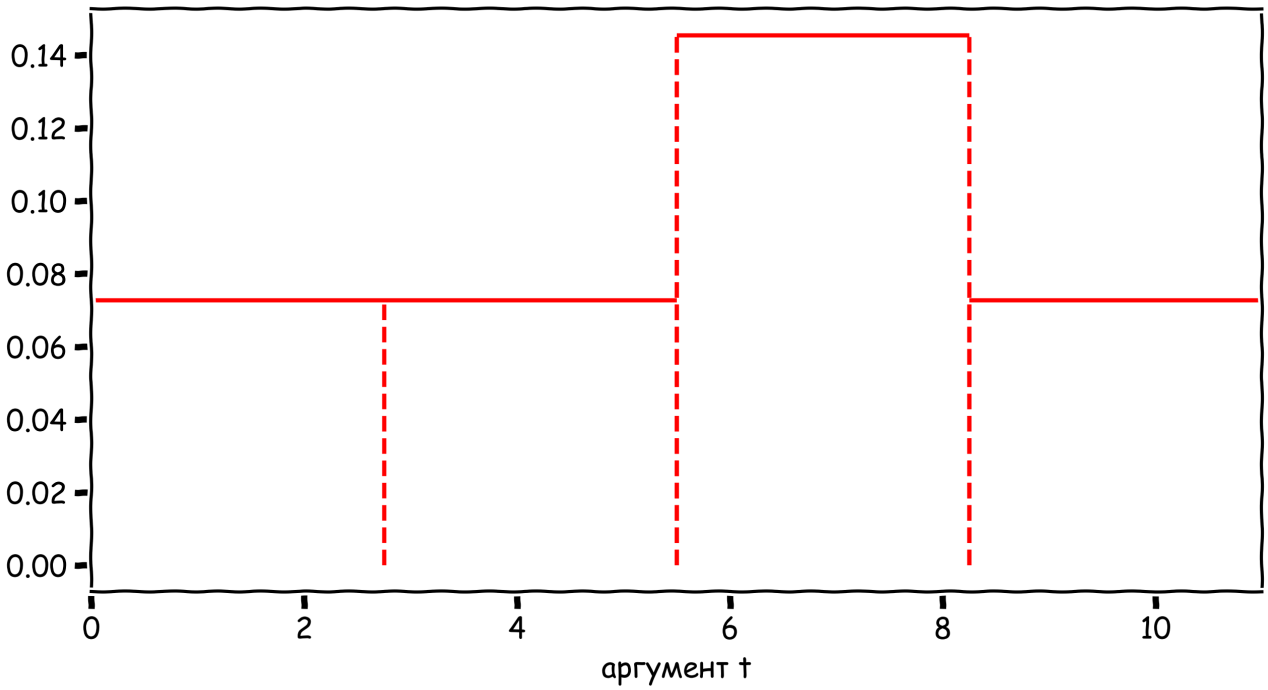

Рис. 6: Гистограмма при $n = 4$

$n = 5$*, то картина меняется, см. рисунок 7.*

По гистограмме можно четко увидеть и то, что распределение рассматриваемой генеральной совокупности не унимодально и является смесью. Например, на рисунке 8 четко видно, что рассматривается смесь трех распределений. Похоже, даже легко догадаться каких, не так ли? Видимо, двух нормальных и равномерного. Потому гистограмма может помочь определить хотя бы примерно вид распределения, вот так вот.

Ясно, что в какой-то степени можно утверждать, что чем больше интервалов группировки, тем лучше (можно провести аналогию с интегралом Римана и интегральными суммами Римана). С другой стороны, весьма опрометчиво брать число интервалов очень большим, ведь тогда на каждый интервал попадет в среднем по одной точке, и гистограмма не будет приближаться к истиной плотности с ростом n. Оказывается, справедливо следующее утверждение:

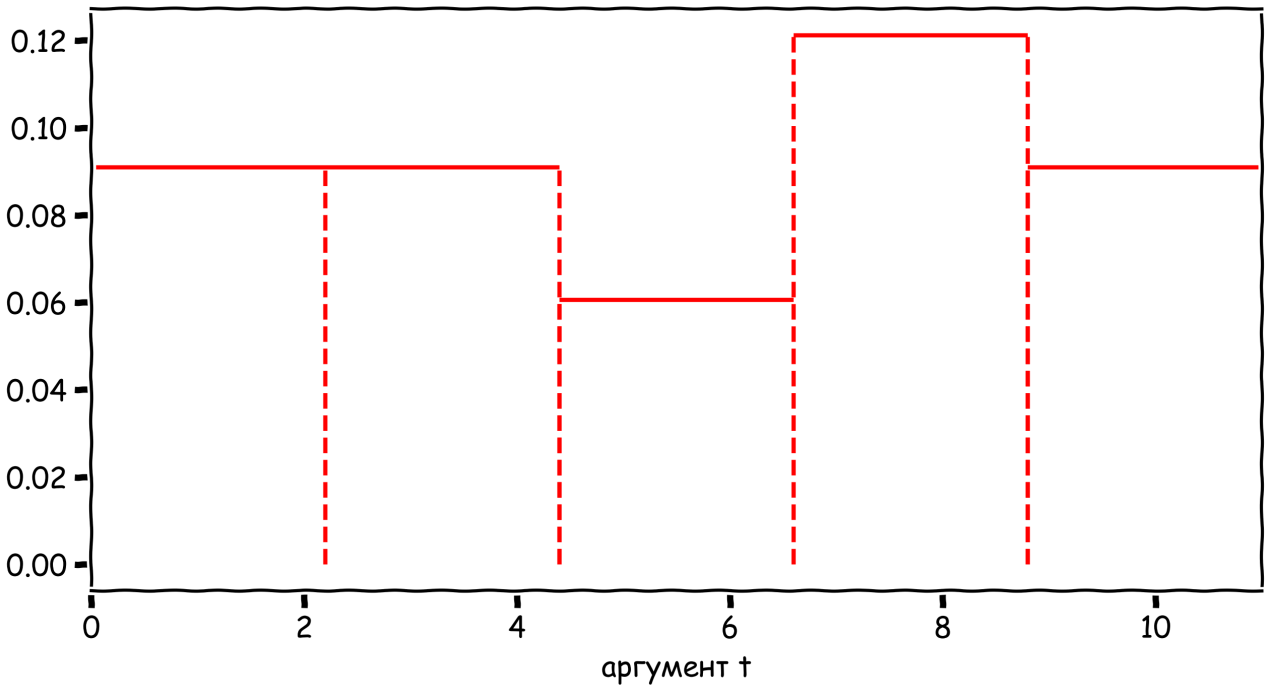

Рис. 7: Гистограмма при $n = 5$

Лемма 2.3.1 *Пусть плотность распределения f_ξ генеральной совокупности ξ – непрерывная функция, $k(n)$ – количество интервалов группировки. Если $k(n) \xrightarrow[n\to\infty]{} +\infty$ так, что*

$$\frac{k(n)}{n} \xrightarrow[n\to\infty]{} 0$$

и длина каждого из интервалов группировки стремится к нулю, то гистограмма в каждой точке сходится к истинной плотности f_ξ по вероятности.

Часто число интервалов группировки берут пропорционально $\sqrt[3]{n}$.

Отметим важное свойство гистограммы, которое показывает, что построенная нами оценка и правда приближает теоретическую плотность.

Теорема 2.3.1 (Основное свойство гистограммы) *Предположим, что случайная величина ξ имеет абсолютно непрерывное распределение с плотностью f_ξ. Для любого $j \in \{1, 2, ..., k\}$ при $n \to \infty$ имеет место сходимость по вероятности*

$$\frac{\nu_j}{n} \xrightarrow[n\to\infty]{\mathsf{P}} \int\limits_{A_j} f_\xi(x)dx.$$

Доказательство. Легко понять, что

$$\nu_j = \sum_{i=1}^{n} \mathsf{I}(X_i \in A_j).$$

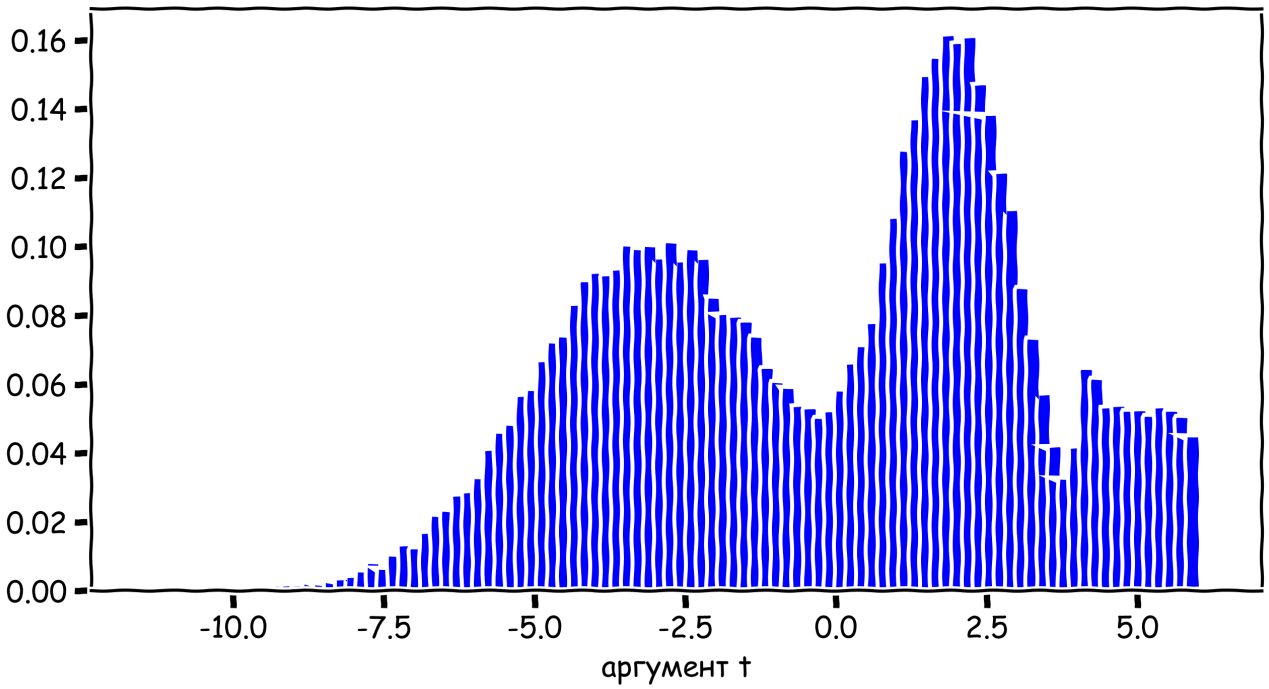

Рис. 8: Смесь

Согласно закону больших чисел, имеет место сходимость по вероятности

$$\frac{\nu_j}{n} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}\left(\mathsf{I}(X_1 \in A_j)\right) = \int\limits_{A_j} f_\xi(x)dx.$$

□

Итак, последняя теорема утверждает, что при увеличении объема выборки площадь j-ого прямоугольника приближается к площади под графиком плотности f_ξ, то есть к искомой вероятности. Это хорошо видно на рисунках 9, 10, 11.

2.4 Выборочные моменты

Напомним, что математическое ожидание случайной величины ξ показывает так называемое среднее вероятностное значение случайной величины, дисперсия же характеризует квадрат ее разброса. Эти характеристики полезны для получения каких-то выводов о генеральной совокупности. В этом пункте мы изучим теоретические свойства моментов, которые в дальнейшем будем использовать и при построении доверительных интервалов, и при решении задачи проверки гипотез.

Определение 2.4.1 *Выборочным средним называется случайная величина*

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

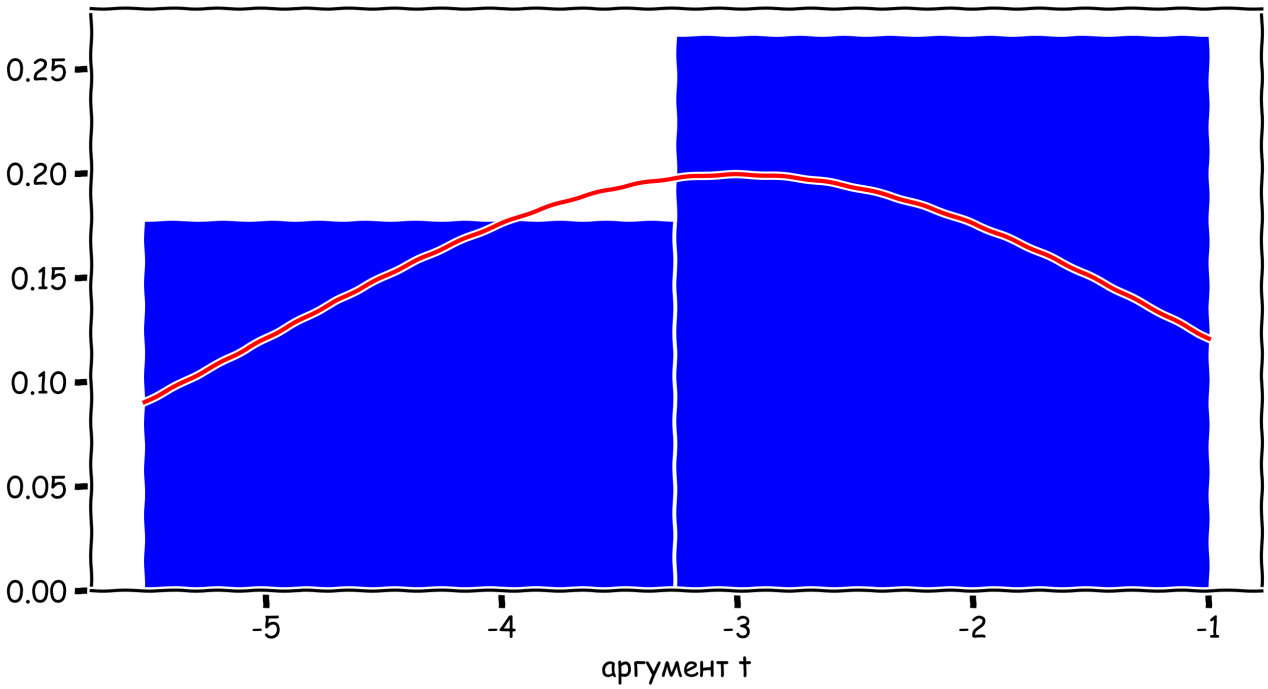

Рис. 9: Гистограмма для выборки объема $n = 10$

Еще раз полезно заметить, что введенное определение есть не что иное, как оценка неизвестного истинного математического ожидания математическим ожиданием эмпирической случайной величины ξ^*, которое можно посчитать.

Пример 2.4.1 *Снова рассмотрим выборку*

$$X = (0, 11, 2, 3, 9, 2, 8, 6, 3.4, 8, 7.5, 9, 4, 8, 6).$$

Тогда

$$\overline{X} = \frac{0 + 11 + 2 + 3 + 9 + 2 + 8 + 6 + 3.4 + 8 + 7.5 + 9 + 4 + 8 + 6}{15} \approx 5.793.$$

Что это значит? Это значит, что в среднем Пете приходится ждать Дашу почти что 6 минут. На рисунке 12 синие точки – это элементы выборки. Красная точка отвечает выборочному среднему $\overline{X}$.

Ясно, что выборочное среднее логично считать оценкой для математического ожидания истинного распределения, так как выборочное распределение при больших объемах выборки сходится по вероятности к истинному. Покажем, что это и правда так.

Теорема 2.4.1 (Свойства выборочного среднего) *Выборочное среднее обладает следующими свойствами:*
1. Если $\mathsf{E}|X_1| < +\infty$, то $\mathsf{E}\overline{X} = \mathsf{E}X_1$. Итак, выборочное среднее является несмещенной оценкой математического ожидания.

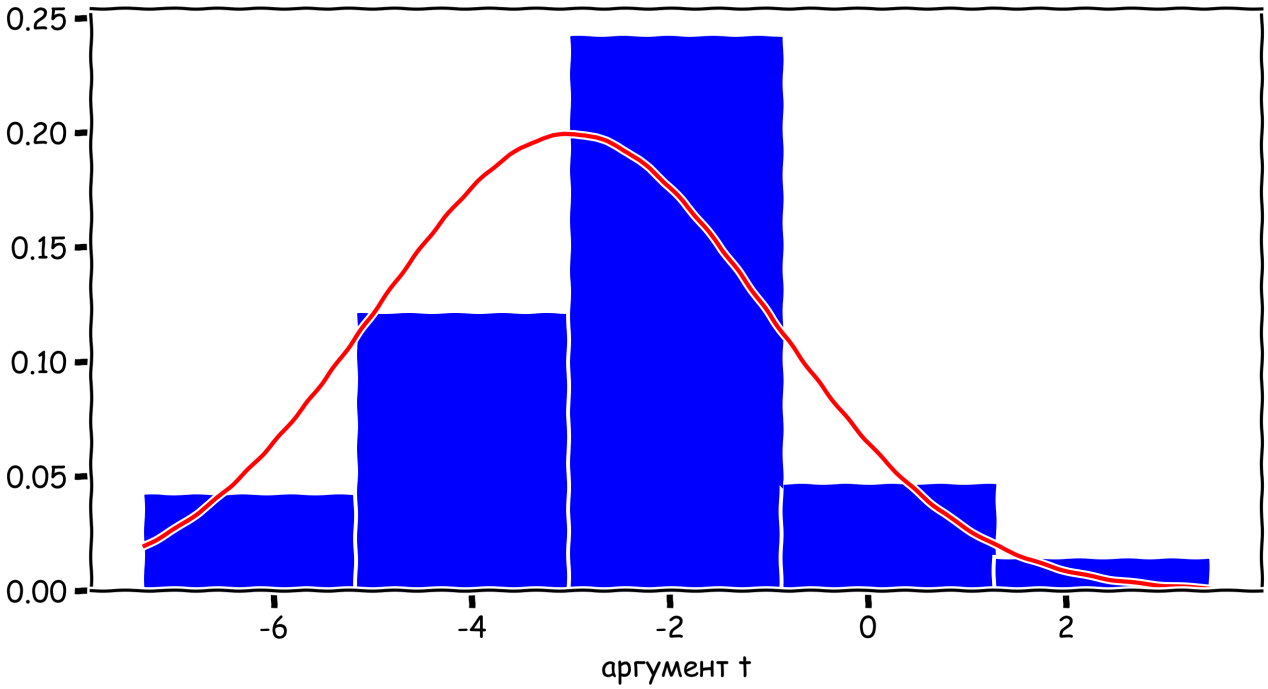

Рис. 10: Гистограмма для выборки объема $n = 100$

2. Если $\mathsf{E}|X_1| < +\infty$, *то* $\overline{X} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}X_1$, *то есть выборочное среднее является состоятельной оценкой математического ожидания.*
3. Если $\mathsf{D}X_1 \in (0, +\infty)$, *то*

$$Y_n = \sqrt{n}\frac{\overline{X} - \mathsf{E}X_1}{\sqrt{\mathsf{D}X_1}} \xrightarrow[n\to\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1},$$

что означает, что выборочное среднее является асимптотически нормальной оценкой математического ожидания.

Доказательство. 1. Это свойство следует из свойств математического ожидания, так как

$$\mathsf{E}\overline{X} = \mathsf{E}\frac{X_1 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}X_i = \frac{n\mathsf{E}X_1}{n} = \mathsf{E}X_1.$$

2. Это свойство немедленно следует из закона больших чисел в форме Хинчина.
3. Это свойство – прямое следствие центральной предельной теоремы, ведь если обозначить $S_n = X_1 + X_2 + ... + X_n$, то

$$\sqrt{n}\frac{\overline{X} - \mathsf{E}X_1}{\sqrt{\mathsf{D}X_1}} = \frac{S_n - n\mathsf{E}X_1}{\sqrt{n\mathsf{D}X_1}}.$$

Осталось проверить, что выполнены все условия ЦПТ. □

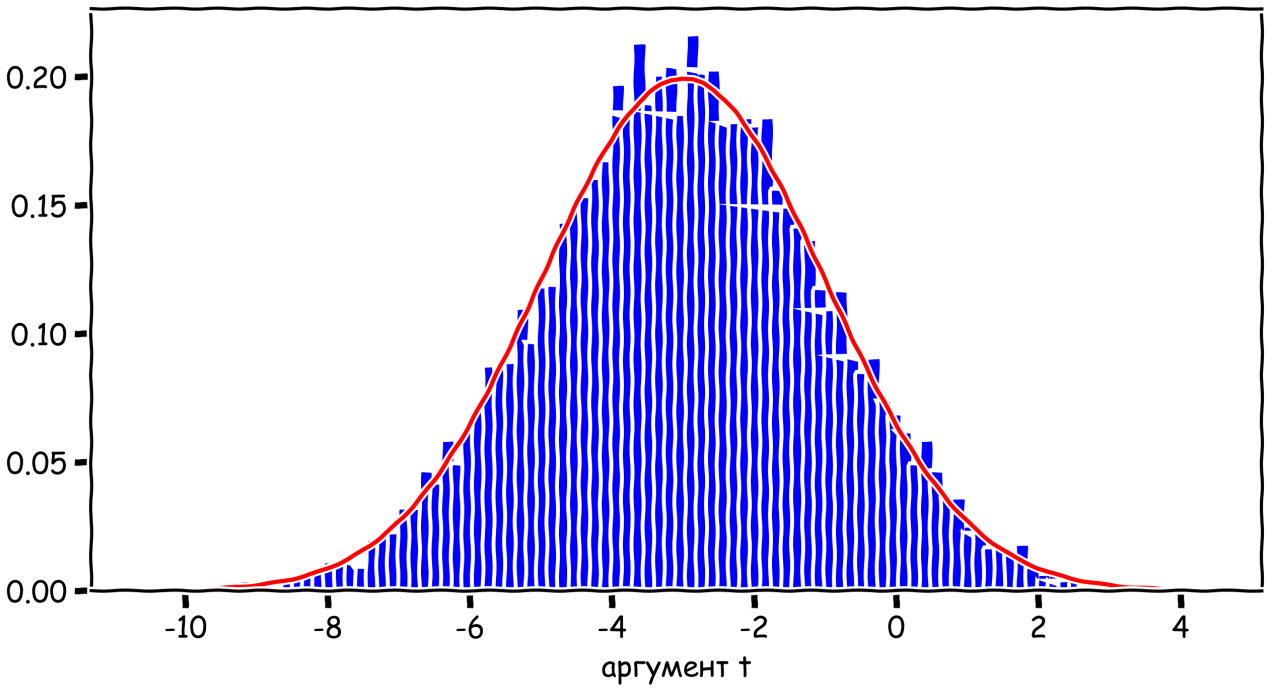

Рис. 11: Гистограмма для выборки объема $n = 10000$

Абсолютно аналогично тому, как мы рассмотрели момент первого порядка, рассмотрим момент k-ого порядка.

Определение 2.4.2 *Выборочным моментом k-ого порядка называется случайная величина*

$$\overline{X^k} = \frac{X_1^k + X_2^k + ... + X_n^k}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Свойства выборочных моментов k-ого порядка дает следующая теорема.

Теорема 2.4.2 (Свойства выборочных моментов k-ого порядка) *Выборочные моменты k-ого порядка являются несмещенными, состоятельными и асимптотически нормальными оценками для истинного k–ого момента, а именно:*

1. Если $\mathsf{E}|X_1|^k < +\infty$, *то* $\mathsf{E}\overline{X^k} = \mathsf{E}X_1^k$.

2. Если $\mathsf{E}|X_1|^k < +\infty$, *то* $\overline{X^k} \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}X_1^k$.

3. Если $\mathsf{D}X_1^k \in (0, +\infty)$, *то*

$$Y_n = \sqrt{n}\frac{\overline{X^k} - \mathsf{E}X_1^k}{\sqrt{\mathsf{D}X_1^k}} \xrightarrow[n\to\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Эта теорема доказывается абсолютно аналогично только что доказанной. Проделайте это самостоятельно!

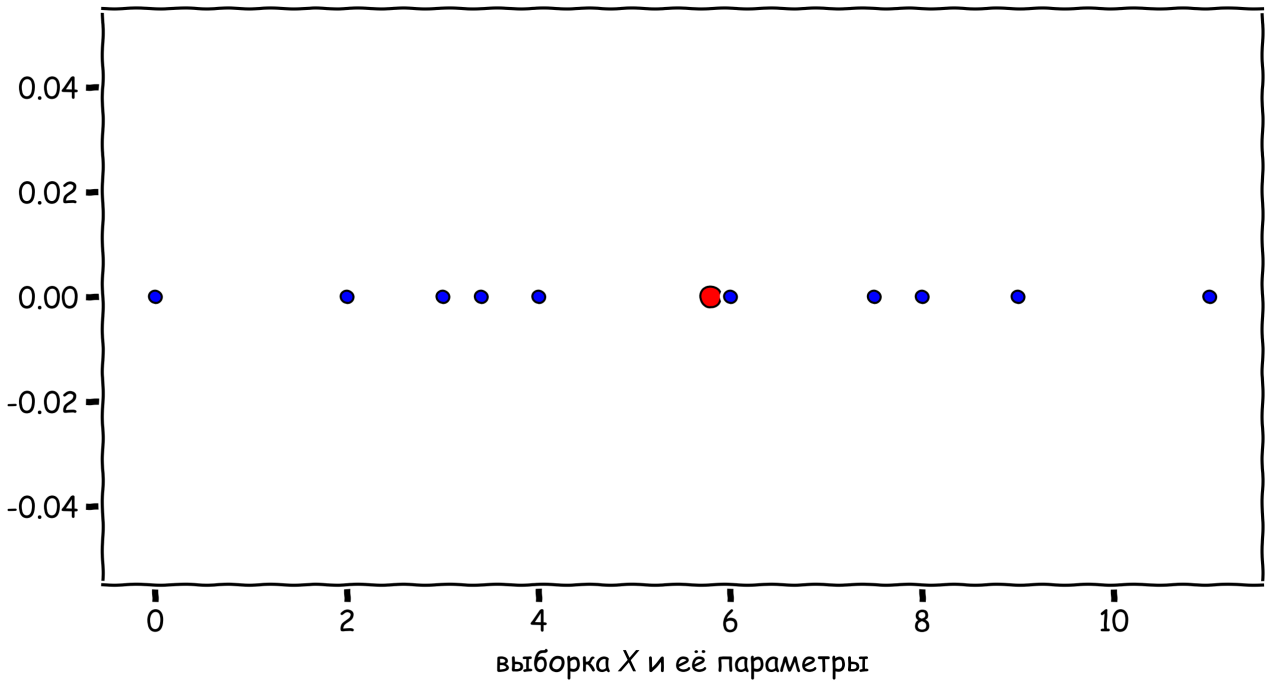

Рис. 12: Выборка X и ее среднее $\overline{X}$

Следующая важная характеристика – выборочная дисперсия. Она вводится по аналогии с обычным определением, ведь

$$\mathsf{D}\xi = \mathsf{E}(\xi - \mathsf{E}\xi)^2,$$

только истинный момент заменяется на выборочный. Итак:

Определение 2.4.3 *Выборочной дисперсией называется случайная величина*

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

Пример 2.4.2 *Рассмотрим все ту же выборку*

$$X = (0, 11, 2, 3, 9, 2, 8, 6, 3.4, 8, 7.5, 9, 4, 8, 6).$$

Давайте посмотрим, насколько сильно отличаются в среднем опоздания Даши. Легко понять, что $S^2 \approx 9.625$. *На рисунке 13 элементы выборки изображены синими точками, красная точка – выборочное среднее, а красная линия между зелеными точками показывает интервал*

$$\left(\overline{X} - S, \overline{X} + S\right).$$

Что такое S*? Это величина среднего разброса значений эмпирической случайной величины от среднего. Как легко понять,* $S \approx 3.1$, *так что разброс от среднего, равного, примерно,* 5.8, *достаточно велик. Какой отсюда*

можно сделать вывод? А такой, что Пете самому не стоит опаздывать больше, чем на где-то две минуты, так как опоздания Даши носят весьма «разный» характер, и он может опоздать больше, чем она. Ну а если Даша опаздывает больше, чем на 9 минут, то можно начинать волноваться.

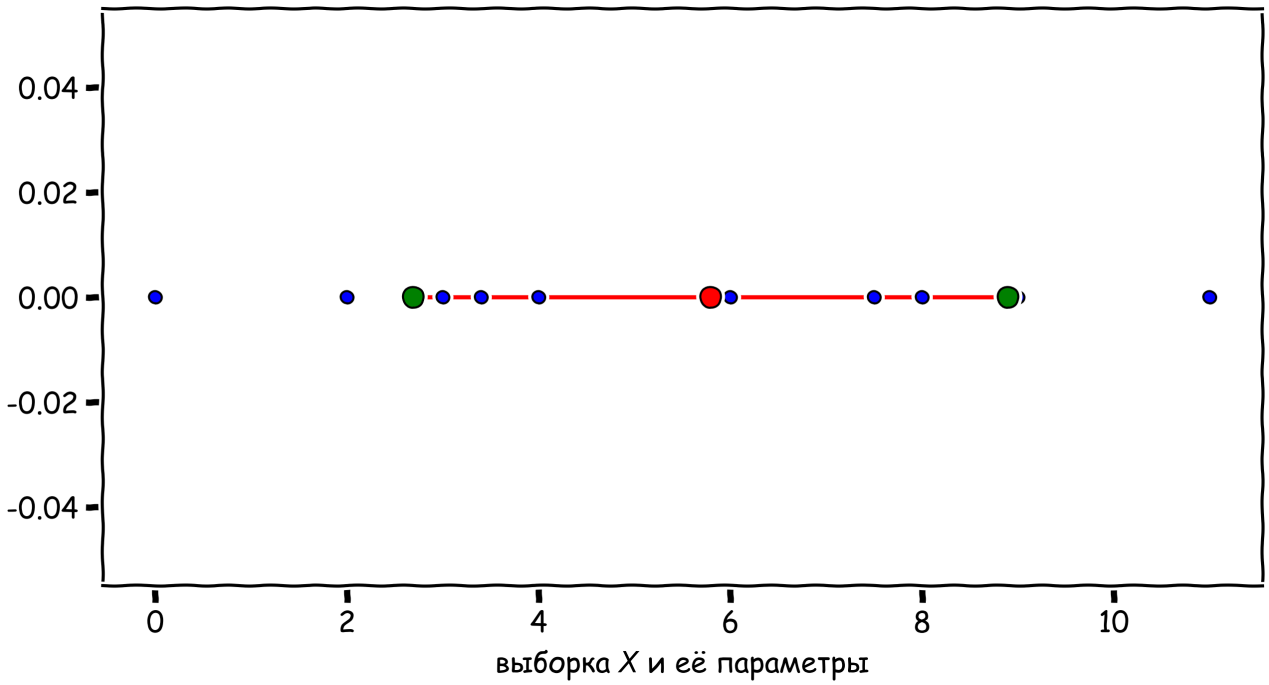

Рис. 13: Выборка X, ее выборочное среднее $\overline{X}$ и S

Убедимся, что справедлива следующая формула, являющаяся прямым аналогом хорошо известного соотношения для дисперсии случайной величины ξ:

$$\mathsf{D}\xi = \mathsf{E}\xi^2 - (\mathsf{E}\xi)^2.$$

Теорема 2.4.3 *Для выборочной дисперсии справедливы соотношения*

$$S^2 = \overline{X^2} - \overline{X}^2$$

В общем-то, убеждаться тут не в чем: так как выборочная дисперсия является дисперсией случайной величины, имеющей эмпирическое распределение, то для нее (дисперсии) выполнены все требуемые свойства, в частности то, что дисперсия равна математическому ожиданию квадрата минус квадрат математического ожидания.

Важно отметить, что выборочная дисперсия является смещенной, состоятельной и асимптотически нормальной оценкой для теоретической дисперсии, а именно:

Теорема 2.4.4 *Выборочная дисперсия обладает следующими свойствами:*
1. Если $\mathsf{D}X_1 < +\infty$*, то* S^2 *– смещенная оценка дисперсии, а именно*

$$\mathsf{E}S^2 = \frac{n-1}{n}\mathsf{D}X_1.$$

2. Если $\mathsf{D}X_1 < +\infty$*, то* S^2 *– состоятельная оценка дисперсии, а именно*

$$S^2 \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{D}X_1.$$

3. Если $0 < \mathsf{D}\left(X_1 - \mathsf{E}X_1\right)^2 < +\infty$*, то* S^2 *– асимптотически нормальная оценка дисперсии, а именно*

$$Y_n = \sqrt{n}\frac{S^2 - \mathsf{D}X_1}{\sqrt{\mathsf{D}(X_1 - \mathsf{E}X_1)^2}} \xrightarrow[n\to\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Доказательство. 1. В первой лекции, говоря о качестве оценок, мы получили формулу, справедливую для любой случайной величины ξ, имеющей второй момент, а именно:

$$\mathsf{E}(\xi - a)^2 = \mathsf{D}\xi + \left(\mathsf{E}\xi - a\right)^2.$$

Подставим вместо ξ эмпирическую величину ξ^*, а вместо a истинное значение $\mathsf{E}X_1$, тогда получим

$$\widetilde{\mathsf{E}}(\xi^* - \mathsf{E}X_1)^2 = \widetilde{\mathsf{D}}\xi + \left(\widetilde{\mathsf{E}}\xi^* - \mathsf{E}X_1\right)^2 = S^2 + \left(\overline{X} - \mathsf{E}X_1\right)^2,$$

откуда

$$S^2 = \widetilde{\mathsf{E}}(\xi^* - \mathsf{E}X_1)^2 - \left(\overline{X} - \mathsf{E}X_1\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\mathsf{E}X_1 \sum_{i=1}^{n} X_i + (\mathsf{E}X_1)^2 - \left(\overline{X} - \mathsf{E}X_1\right)^2.$$

Ну а тогда

$$\mathsf{E}S^2 = \mathsf{D}X_1 - \mathsf{E}\left(\overline{X} - \mathsf{E}X_1\right)^2 = \mathsf{D}X_1 - \mathsf{D}\overline{X} = \mathsf{D}X_1 - \frac{1}{n}\mathsf{D}X_1 = \frac{n-1}{n}\mathsf{D}X_1.$$

2. Воспользуемся тем, что $S^2 = \overline{X^2} - \overline{X}^2$ и тем, что выборочные моменты первого и второго порядков состоятельны, а также свойствами сходимости по вероятности получаем, что

$$S^2 = \overline{X^2} - \overline{X}^2 \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}X_1^2 - (\mathsf{E}X_1)^2 = \mathsf{D}X_1.$$

3. Введем в рассмотрение вспомогательные случайные величины $Y_i = X_i - \mathsf{E}X_1$. Тогда $\mathsf{E}Y_i = 0$, $\mathsf{E}Y_i^2 = \mathsf{D}X_i = \mathsf{D}X_1$. Значит,

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathsf{E}X_1 - (\overline{X} - \mathsf{E}X_1))^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Последнее равенство есть не что иное, как выборочная дисперсия выборки $Y_1, Y_2, ..., Y_n$, а значит

$$S^2 = \overline{Y^2} - (\overline{Y})^2.$$

Тогда

$$\sqrt{n}\frac{S^2 - \mathsf{D}X_1}{\sqrt{\mathsf{D}(X_1 - \mathsf{E}X_1)^2}} = \sqrt{n}\frac{\overline{Y^2} - (\overline{Y})^2 - \mathsf{E}Y_1^2}{\sqrt{\mathsf{D}Y_1^2}} = \sqrt{n}\frac{\overline{Y^2} - \mathsf{E}Y_1^2}{\sqrt{\mathsf{D}Y_1^2}} - \sqrt{n}\frac{(\overline{Y})^2}{\sqrt{\mathsf{D}Y_1^2}}.$$

Первое слагаемое последней суммы как раз, согласно центральной предельной теореме, сходится к случайной величине, имеющей стандартное нормальное распределение, ведь

$$\sqrt{n}\frac{\overline{Y^2} - \mathsf{E}Y_1^2}{\sqrt{\mathsf{D}Y_1^2}} = \frac{\sum\limits_{i=1}^{n} Y_i^2 - n\mathsf{E}Y_1^2}{\sqrt{n\mathsf{D}Y_1^2}} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Второе же слагаемое слабо сходится к нулю, так как

$$\sqrt{n}\frac{(\overline{Y})^2}{\sqrt{\mathsf{D}Y_1^2}} = \frac{\sqrt{\mathsf{D}Y_1}}{\sqrt{\mathsf{D}Y_1^2}}\overline{Y} \cdot \frac{\overline{Y}}{\sqrt{\mathsf{D}Y_1}},$$

и первое слагаемое сходится к нулю (по вероятности), а второе сходится слабо к случайной величине, имеющей стандартное нормальное распределение (проверьте!).

□

Итак, как мы уже отметили, оценка S^2 является смещенной оценкой истинной дисперсии. В то же время, так как $\mathsf{E}S^2 = \frac{n-1}{n}\mathsf{D}X_1$, то

$$\lim_{n\to\infty} \mathsf{E}S^2 = \mathsf{D}X_1.$$

Это свойство называется асимптотической несмещенностью, и оценка S^2 им обладает.

Из свойств моментов легко понять, как нужно исправить оценку дисперсии, чтобы она стала несмещенной.

Определение 2.4.4 *Несмещенной выборочной дисперсией называется случайная величина*

$$S_0^2 = \frac{n}{n-1} S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 .$$

Несмещенная выборочная дисперсия оказывается несмещенной, асимптотически нормальной и состоятельной оценкой теоретической дисперсии, тем самым справедлива следующая теорема

Теорема 2.4.5 *Несмещенная выборочная дисперсия обладает следующими свойствами:*
1. Если $\mathsf{D}X_1 < +\infty$*, то* S_0^2 *– несмещенная оценка дисперсии, а именно*

$$\mathsf{E}S_0^2 = \mathsf{D}X_1.$$

2. Если $\mathsf{D}X_1 < +\infty$*, то* S_0^2 *– состоятельная оценка дисперсии, а именно*

$$S_0^2 \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{D}X_1.$$

3. Если $0 < \mathsf{D}(X_1 - \mathsf{E}X_1)^2 < +\infty$*, то* S_0^2 *– асимптотически нормальная оценка дисперсии, а именно*

$$Y_n = \sqrt{n} \frac{S_0^2 - \mathsf{D}X_1}{\sqrt{\mathsf{D}(X_1 - \mathsf{E}X_1)^2}} \xrightarrow[n\to\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Доказательства проводятся аналогично тому, как это было сделано в предыдущей теореме. В доказательствах 2 и 3 пунктов используется то, что $\lim\limits_{n\to\infty} \frac{n}{n-1} = 1$.

На центральных моментах более высокого порядка мы в данной лекции останавливаться не будем.

2.5 Выборочные квантили

Понятие квантили вам уже встречалось в курсе хранения и обработки данных. Мы не будем останавливаться подробно на ее свойствах, а лишь дадим определения, которые будем использовать в дальнейшем. Напомним основное определение.

Определение 2.5.1 *Пусть фиксировано число* $\alpha \in (0, 1)$*. Квантилью уровня* α *называется такое число* x_α*, что*

$$\mathsf{P}(\xi \le x_\alpha) \ge \alpha, \quad \mathsf{P}(\xi \ge x_\alpha) \ge 1 - \alpha$$

Последнее определение можно дать и в терминах функции распределения случайной величины ξ, ведь

$$\mathsf{P}(\xi \leq x_\alpha) = F_\xi(x_\alpha + 0) \geq \alpha,$$

а

$$\mathsf{P}(\xi \geq x_\alpha) = 1 - \mathsf{P}(\xi < x_\alpha) = 1 - F_\xi(x_\alpha),$$

а значит второе неравенство равносильно тому, что

$$F_\xi(x_\alpha) \leq \alpha.$$

Итого, эквивалентным образом квантиль уровня α можно определить, как такое число x_α, что

$$F_\xi(x_\alpha) \leq \alpha, \quad F_\xi(x_\alpha + 0) \geq \alpha.$$

Как искать квантиль? Если функция распределения F_ξ непрерывна и строго монотонна, то для каждого $\alpha \in (0, 1)$ существует единственная квантиль уровня α, которую можно найти из уравнения $F_\xi(x_\alpha) = \alpha$, то есть

$$x_\alpha = F_\xi^{-1}(\alpha).$$

Возможны и две другие ситуации. Первая – когда это уравнение не имеет решений. Тогда квантиль снова единственна и находится, как число x_α, при котором

$$F_\xi(x_\alpha) < \alpha, \quad F_\xi(x_\alpha + 0) \geq \alpha.$$

Последняя же ситуация – ситуация, когда уравнение имеет несколько решений. В качестве квантили может быть выбрано любое из решений.

Определение 2.5.2 *Пусть $X_1, X_2, ..., X_n$ – выборка из генеральной совокупности ξ. Число*

$$\widehat{x}_\alpha = \begin{cases} X_{([n\alpha]+1)}, & \text{если число } n\alpha \notin \mathbb{Z} \\ \frac{X_{(n\alpha)} + X_{(n\alpha+1)}}{2}, & \text{если число } n\alpha \in \mathbb{Z} \end{cases}$$

называется выборочной квантилью уровня α для выборки $X_1, X_2, ..., X_n$.

Квантиль уровня 0.5 называется медианой. Давайте на примере нашей выборки найдем выборочную медиану.

Пример 2.5.1 *Рассмотрим уже знакомую выборку*

$$X = (0, 11, 2, 3, 9, 2, 8, 6, 3.4, 8, 7.5, 9, 4, 8, 6)$$

и построенный по ней вариационный ряд

$$(0, 2, 2, 3, 3.4, 4, 6, 6, 7.5, 8, 8, 8, 9, 9, 11).$$

Объем нашей выборки $n = 15$*, а значит* $n\alpha = 15 \cdot 0.5 = 7.5$ *и выборочная медиана* $\widehat{x}_{0.5} = X_{(8)} = 6$*. Медиана очень близка к выборочному среднему.*

Вообще, резюмируя, можно сказать, что в случае, когда объем выборки – нечетное число, то выборочная медиана – это «серединный» элемент вариационного ряда. Если же объем выборки – число четное, то выборочная медиана – это полусумма «серединных» элементов вариационного ряда. Смысл выборочной медианы должен быть прозрачен: половина элементов выборки не превосходит выборочной медианы, а половина элементов выборки не меньше ее.