Точечное оценивание

Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Точечное оценивание

В предыдущей лекции мы научились строить по выборке общие характеристики распределения генеральной совокупности ξ. В этой же лекции мы займемся так называемым параметрическим оцениванием, но для начала аккуратно введем такие понятия, косвенно используемые ранее, как: оценка, состоятельность, несмещенность и проч. Кроме того, поставим общую задачу оценивания.

Часто бывает, что распределение случайной величины нам известно с точностью до каких-то числовых параметров. Например, распределение роста юношей, как оказывается, имеет нормальное распределение N_{a,σ^2} с неизвестными средним a и дисперсией σ^2, а количество покупателей в магазине в течение часа (не «часа пик») – распределение Пуассона Π_λ с неизвестной «интенсивностью» – параметром $\lambda > 0$. Целью данной лекции является оценивание одного или нескольких параметров рассматриваемого распределения.

Итак, пусть имеется выборка $X_1, ..., X_n$, извлеченная из распределения $\mathcal{P}_\theta$, известным образом зависящего от скалярного (или векторного) параметра θ, причем параметр θ может принимать значения из некоторого множества Θ. О чем говорит последнее предложение? Наверное, проще всего его пояснить на примерах.

Пример 1.0.1 *Пусть генеральная совокупность ξ (а значит и все случайные величины X_i, $i \in \{1, ..., n\}$) имеют распределение Пуассона Π_λ с неизвестным параметром $\lambda > 0$. Тогда $\mathcal{P}_\theta = \Pi_\lambda$, $\theta = \lambda$, $\Theta = (0, +\infty)$.*

Так, говоря о магазине, параметр λ показывает среднее число покупателей. Кроме того, знание параметра распределения позволяет вычислять какие угодно вероятности, а это, как мы уже удостоверились, очень ценно. Скажем, владельцу магазина полезно задаваться вопросом: а когда резонно открывать дополнительную кассу? Или на языке вероятностей: в какие временные промежутки вероятность того, что число покупателей, стоящих в очереди на кассе, будет больше, например, трех, будет больше, чем 0.7?

Пример 1.0.2 *Пусть генеральная совокупность ξ имеет распределение Бернулли B_p с неизвестным параметром $p \in (0, 1)$. Тогда $\mathcal{P}_\theta = \mathsf{B}_p$, $\theta = p$, $\Theta = (0, 1)$.*

Примеров, иллюстрирующих такую задачу, море: это большинство ситуаций, в которых возможно выделение лишь двух исходов: «успех» и «неудача». Например, стрельба в тире по мишени, вытягивание билета на экзамене с заранее подготовленным вопросом и многое другое.

Пример 1.0.3 *Пусть генеральная совокупность ξ имеет равномерное распределение $\mathsf{U}_{a,b}$ с неизвестными параметрами $a < b$. Тогда $\mathcal{P}_\theta = \mathsf{U}_{a,b}$, $\theta = (a, b)$, $\Theta = \{(a, b) \in \mathbb{R}^2 : \ a < b\}$.*

Пример 1.0.4 *Пусть генеральная совокупность ξ имеет нормальное распределение $\mathsf{N}_{a,1}$ с неизвестным параметром a. Тогда $\mathcal{P}_\theta = \mathsf{N}_{a,1}$, $\theta = a$, $\Theta = \mathbb{R}$.*

Итак, мы видим, что параметры могут быть самыми разнообразными. Распределение от этих параметров может зависеть самым хитрым образом, но важно то, что оно полностью определяется этими параметрами.

Отметим еще один важный момент. Предположим, что $X_1, ..., X_n$ – выборка объема n из распределения $\mathcal{P}_\theta$. Тогда характеристики случайных величин X_i (математическое ожидание, дисперсия и проч.) зависят от параметра θ. Например, если генеральная совокупность ξ имеет распределение Бернулли с параметром $p = 0.3$, то $\mathsf{E}X_1 = 0.3$, а если параметр $p = 0.9$, то $\mathsf{E}X_1 = 0.9$. В дальнейшем нам будет важно отражать эту зависимость, поэтому мы будем наделять математическое ожидание, дисперсию и прочие параметры индексом θ и писать $\mathsf{E}_\theta X_1$, $\mathsf{D}_\theta X_1$ и так далее.

1.1 Точечные оценки и их свойства

Так как мы говорим о точечном оценивании, об оценивании неизвестных параметров, то следует для начала определить понятие оценки. Последнюю часто называют статистикой.

Определение 1.1.1 *Пусть $X_1, ..., X_n$ – выборка объема n из семейства распределений $\mathcal{P}_\theta$, $\theta \in \Theta$. Статистикой (или оценкой) параметра θ называется произвольная функция*

$$\widehat{\theta} = \widehat{\theta}(X_1, ..., X_n),$$

являющаяся случайной величиной.

Итак, статистика – это функция от выборки. Теоретически, конечно, не произвольная, так как она должна оставаться случайной величиной, но на практике – любая. Далее мы этот момент никогда отдельно оговаривать не будем, но будем держать в голове. Еще раз повторим важный момент: статистика зависит только от выборки, а оцениваемый параметр θ никаким образом в функцию не вписывается, в ней не участвует и в нее не заложен. Последнее не должно удивлять, ведь этот самый параметр мы пытаемся «оценить».

Теперь давайте дадим (или сформулируем в общем виде) те «хорошие» свойства оценок, которые уже были нами разобраны при изучении выборочных характеристик.

Определение 1.1.2 (Несмещенность оценки) *Статистика* $\widehat{\theta} = \widehat{\theta}(X_1, ..., X_n)$ *называется несмещенной оценкой параметра* θ*, если для любого* $\theta \in \Theta$ *справедливо равенство* $\mathsf{E}_\theta \widehat{\theta} = \theta$.

Итак, как и ранее, несмещенность означает, что математическое ожидание случайной величины $\widehat{\theta}$ совпадает с θ. Иными словами, «в среднем» наша статистика всегда равна оцениваемому параметру θ. Несмещенность – хорошее свойство, но в природе встречающееся редко. Часто удается выудить лишь ослабленное свойство, так называемую асимптотическую несмещенность.

Определение 1.1.3 (Асимптотическая несмещенность оценки) *Статистика* $\widehat{\theta} = \widehat{\theta}(X_1, ..., X_n)$ *называется асимптотически несмещенной оценкой параметра* θ*, если для любого* $\theta \in \Theta$ *выполняется* $\mathsf{E}_\theta \widehat{\theta} \xrightarrow[n\to\infty]{} \theta$.

Асимптотическая несмещенность – это несмещенность в пределе, при неограниченном возрастании объема выборки. Конечно, несмещенность влечет асимптотическую несмещенность. Обратное, вообще говоря, неверно.

Определение 1.1.4 (Состоятельность оценки) *Статистика* $\widehat{\theta} = \widehat{\theta}(X_1, ..., X_n)$ *называется состоятельной оценкой параметра* θ*, если для любого* $\theta \in \Theta$ *выполняется* $\widehat{\theta} \xrightarrow[n\to\infty]{\mathsf{P}} \theta$.

Состоятельность означает, что последовательность оценок (зависящая от объема выборки) приближается к неизвестному параметру (по вероятности) с ростом n. Как уже отмечалось, если бы это выполнено не было, то оценка была бы уж совсем «не состоятельна».

Пример 1.1.1 *Пусть* $X_1, ..., X_n$ *– выборка объема* n *из нормального распределения* $\mathcal{P}_\theta = \mathsf{N}_{a,\sigma^2}$. *Ясно, что* $\theta = (a, \sigma^2)$ *– векторный параметр, причем* $\Theta = \mathbb{R} \times (0, +\infty)$. *Как оценить параметры распределения:* a *– математическое ожидание и* σ^2 *– дисперсию?*

В предыдущей лекции мы выяснили, что «хорошей» оценкой для математического ожидания является выборочное среднее $\widehat{a} = \overline{X}$. *Что значит «хорошей»? А ровно то, что она является несмещенной и состоятельной оценкой параметра* a.

Для дисперсии σ^2 *мы получили даже две «неплохих» оценки:*

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

Обе эти оценки, как было показано, оказались состоятельными. Вторая, к тому же, несмещенной, а первая – асимптотически несмещенной.

Пример 1.1.2 *Пусть $X_1, ..., X_n$ – выборка объема n из показательного распределения $\mathcal{P}_\theta = \mathsf{Exp}_\lambda$. Ясно, что $\theta = \lambda$, причем $\Theta = (0, +\infty)$. Напомним, что математическое ожидание случайной величины, имеющей показательное распределение, равно $\frac{1}{\lambda}$, а дисперсия равна $\frac{1}{\lambda^2}$. Тогда для оценки параметра λ можно выбрать оценку*

$$\widehat{\lambda} = \frac{1}{\overline{X}}.$$

С другой стороны, можно оценить параметр λ, используя оценку дисперсии, тем самым

$$\widehat{\lambda} = \frac{1}{S} = \sqrt{\frac{n}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}} \text{ или } \widehat{\lambda} = \frac{1}{S_0} = \sqrt{\frac{n-1}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}.$$

Какая оценка лучше? Пока теоретического ответа на этот вопрос мы дать не можем (сравнением оценок мы займемся в следующей лекции), но давайте протестируем то, что есть, на конкретных числах. Пусть у нас есть выборка из показательного распределения с параметром $\lambda = 2$ объема $n = 10$:

$$(0.83854048, 0.12604081, 0.47325527, 0.29251617, 0.03712543, 3.4183022,$$

$$0.43344993, 0.32445651, 1.62311045, 0.27511339).$$

Выборка приводится скорее «просто чтобы показать», нежели еще для каких целей. Какие значения нам дают наши оценки на такой выборке?

$$\widehat{\lambda_1} = \frac{1}{\overline{X}} \approx 1.275, \text{ а значит абсолютная погрешность } |\lambda - \widehat{\lambda_1}| \approx 0.725,$$

$$\widehat{\lambda_2} = \frac{1}{S} = \sqrt{\frac{n}{\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2}} \approx 1.022, \quad |\lambda - \widehat{\lambda_2}| \approx 0.978.$$

И что мы видим: первая оценка лучше? Конечно, какие-либо выводы делать рано, ведь выборка очень мала. Да и вообще, на другой выборке такого же объема ситуация может измениться.

Давайте обсудим предыдущий пример детальнее: а что нас, собственно, интересует? Нас интересует вопрос: а как ведут себя оценки с ростом n? На рисунке 1 представлена зависимость $\widehat{\lambda_1}$ (синим) и $\widehat{\lambda_2}$ (красным) от объема выборки n из генеральной совокупности, имеющей показательное распределение Exp_λ с параметром $\lambda = 2$. Рисунок 2 иллюстрирует отклонение $|\widehat{\lambda_i} - 2|$,

$i \in \{1, 2\}$ в зависимости от n. На первом рисунке отчетливо видна тенденция к сгущению вблизи значения 2 с ростом объема выборки n. Второй же рисунок показывает, что вроде бы у «команды синих», отвечающих за оценку $\widehat{\lambda_1}$, как минимум отклонение, но вроде и «средняя точность» лучше, чем у красных. Иными словами, опираясь на численный эксперимент, можно сделать вывод, что оценка $\widehat{\lambda_1}$ является точнее, нежели оценка $\widehat{\lambda_2}$.

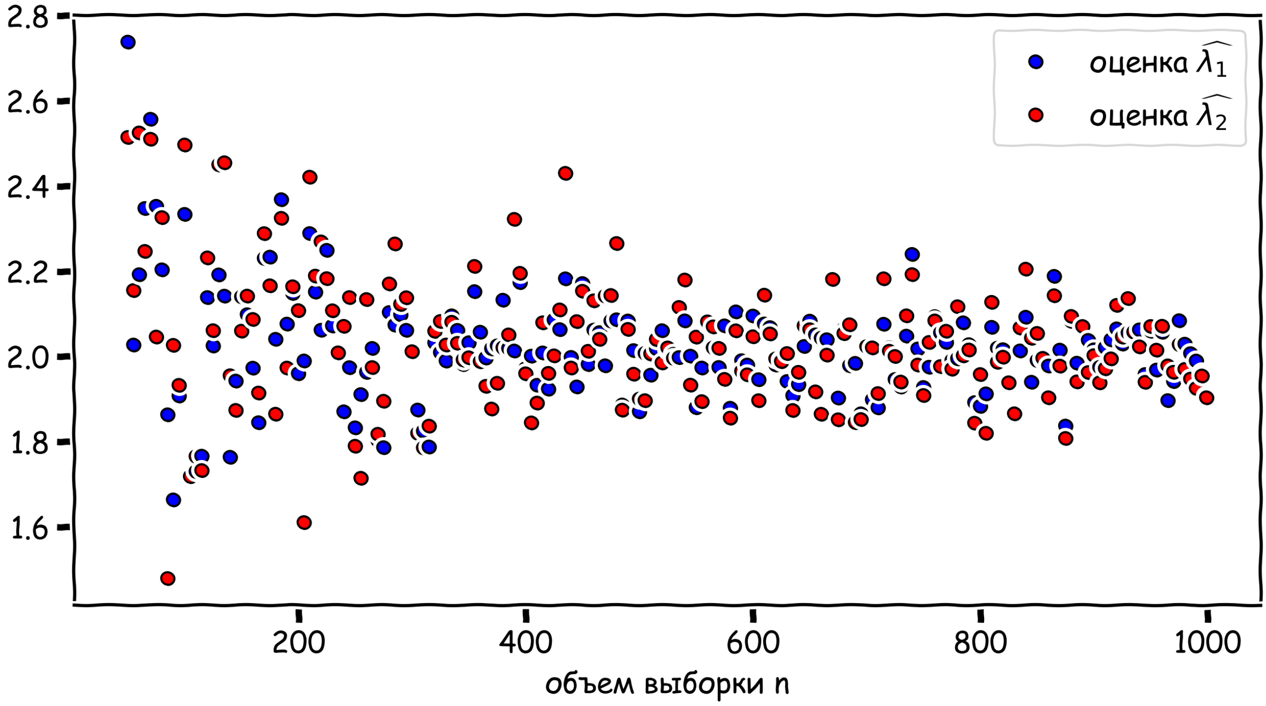

Рис. 1: Зависимость $\widehat{\lambda_1}$ и $\widehat{\lambda_2}$ от объема выборки n

В разобранных примерах так удачно сложилось, что параметры распределения связаны с моментами, которые мы научились оценивать в предыдущей лекции. Рассмотрим некоторые общие способы получения оценок: метод моментов и метод максимального правдоподобия.

2 Метод моментов

2.1 Метод моментов для одномерного параметра

Метод моментов в своей основе опирается вот на какую идею. Как мы уже отмечали, моменты случайной величины, имеющей распределение $\mathcal{P}_\theta$, зависят (функционально) от самого параметра θ. Иными словами, моменты – функции от θ. Но тогда и сам параметр θ может оказаться функцией от

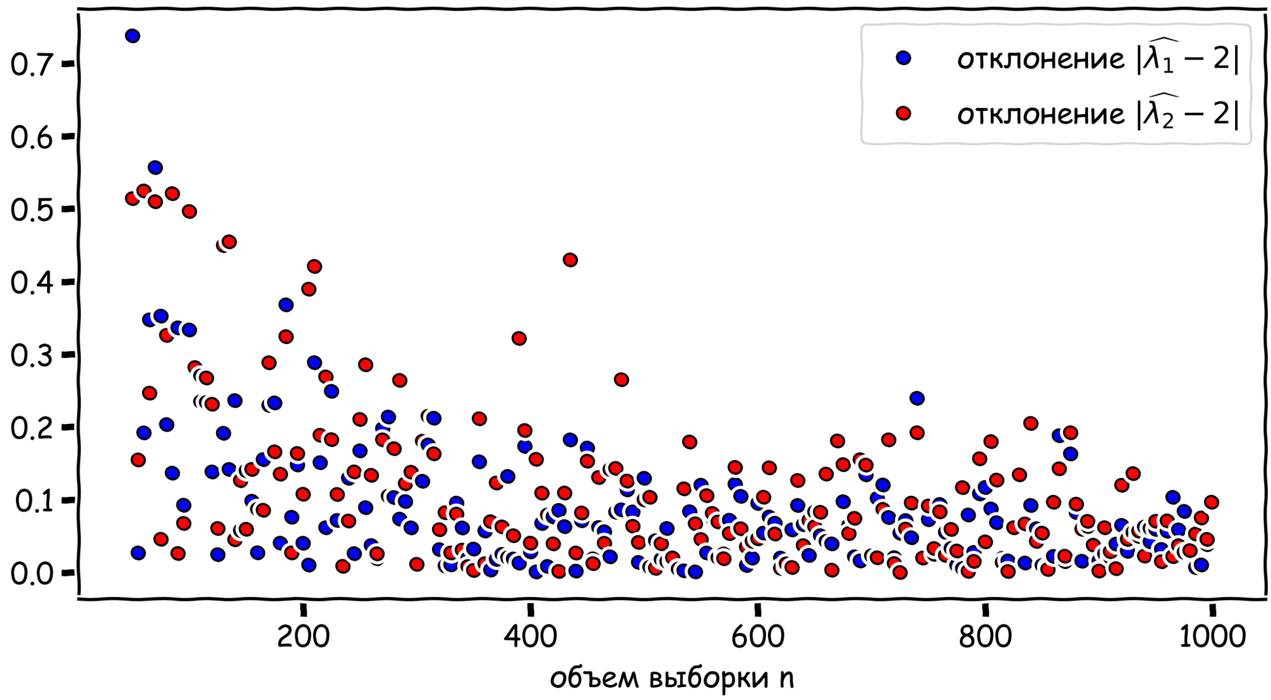

Рис. 2: Графики отклонений $|\widehat{\lambda}_i - 2|$, $i \in \{1, 2\}$

истинного момента. Подставив вместо этого истинного (теоретического) момента его выборочный аналог, получим вместо истинного параметра θ его оценку $\widehat{\theta}$.

Опишем метод более формально. Пусть $X_1, ..., X_n$ – выборка объема n из параметрического семейства распределений $\mathcal{P}_\theta$, $\theta \in \Theta \subset \mathbb{R}$. Пусть функция $g : \mathbb{R} \to \mathbb{R}$ такова, что существует момент

$$\mathsf{E}_\theta g(X_1) = f(\theta),$$

а функция f обратима на множестве Θ (напомним, что нам нужна обратимость, чтобы выразить θ через $\mathsf{E}_\theta g(X_1)$). Разрешим выписанное соотношение относительно θ и получим

$$\theta = f^{-1}\left(\mathsf{E}_\theta g(X_1)\right).$$

Подставим вместо теоретического момента $\mathsf{E}_\theta g(X_1)$ его выборочный аналог $\overline{g(X)}$, откуда получим оценку $\widehat{\theta}$ параметра θ вида

$$\widehat{\theta} = f^{-1}\left(\overline{g(X)}\right) = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right).$$

Определение 2.1.1 *Оценка $\widehat{\theta}$ параметра θ, полученная по описанной выше схеме, называется оценкой метода моментов (ОММ) для параметра θ.*

Чаще всего рассматривают моменты k-ого порядка, а потому берут функции $g(t) = t^k$, $k \in \{1, 2, ...\}$. В этом случае

$$\mathsf{E}_\theta X_1^k = f(\theta),$$

и если функция $f(\theta)$ обратима, то, согласно описанной выше идее, мы приходим к соотношению

$$\widehat{\theta} = f^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^k\right).$$

Пример 2.1.1 (ОММ для распределения Пуассона) *Пусть $X_1, ..., X_n$ – выборка из распределения Пуассона Π_θ с параметром $\theta > 0$. Пусть $g(t) = t$, тогда, так как математическое ожидание случайной величины, имеющей распределение Пуассона, равно θ, имеем*

$$\mathsf{E}_\theta X_1 = \theta.$$

В нашей ситуации все очень просто, так как $\theta = \mathsf{E}_\theta X_1$. Заменяя истинное математическое ожидание на выборочное, получаем оценку

$$\widehat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Давайте рассмотрим функцию $g(t) = t^2$. Воспользуемся тем, что для случайной величины ξ с конечным вторым моментом справедливо равенство $\mathsf{D}_\theta\xi = \mathsf{E}_\theta\xi^2 - (\mathsf{E}_\theta\xi)^2$, откуда

$$\mathsf{E}_\theta X_1^2 = \mathsf{D}_\theta X_1 + (\mathsf{E}_\theta X_1)^2 = \theta^2 + \theta.$$

Функция $\theta^2+\theta$ строго возрастает при $\theta > 0$, а потому обратима. Выразим θ через $\mathsf{E}_\theta X_1^2$, для этого решим уравнение (квадратное относительно θ):

$$\theta^2 + \theta - \mathsf{E}_\theta X_1^2 = 0.$$

Из уравнения находим

$$\theta = \frac{-1 \pm \sqrt{1 + 4\mathsf{E}_\theta X_1^2}}{2}.$$

Так как обращение происходило на множестве $\theta > 0$, то нам подходит только «правый» корень

$$\theta = \frac{-1 + \sqrt{1 + 4\mathsf{E}_\theta X_1^2}}{2}.$$

Так как выборочной оценкой для $\mathsf{E}_\theta X_1^2$ *является* $\overline{X^2} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2$, *то получаем итоговое соотношение*

$$\widehat{\theta} = \frac{-1+\sqrt{1+4\overline{X^2}}}{2}.$$

Как и в примере с показательным распределением, сравним две полученные оценки. Симуляция проводилась на выборке из распределения Пуассона с параметром $\lambda = 3$. Рисунок 3 показывает, как меняются оценки

$$\widehat{\theta_1} = \overline{X} \quad \text{и} \quad \widehat{\theta_2} = \frac{-1+\sqrt{1+4\overline{X^2}}}{2}$$

с ростом n. Во первых очевидно приближение обеих оценок к истинному значению 3 с ростом объема выборки n. Рисунок 4 показывает зависимость величины отклонений $|\widehat{\theta_i} - 3|$, $i \in \{1, 2\}$ оценок от истинного параметра в зависимости от объема выборки n. В отличие от случая с показательным распределением, в этом случае «на глаз» нельзя отдать предпочтение ни одной из двух оценок.

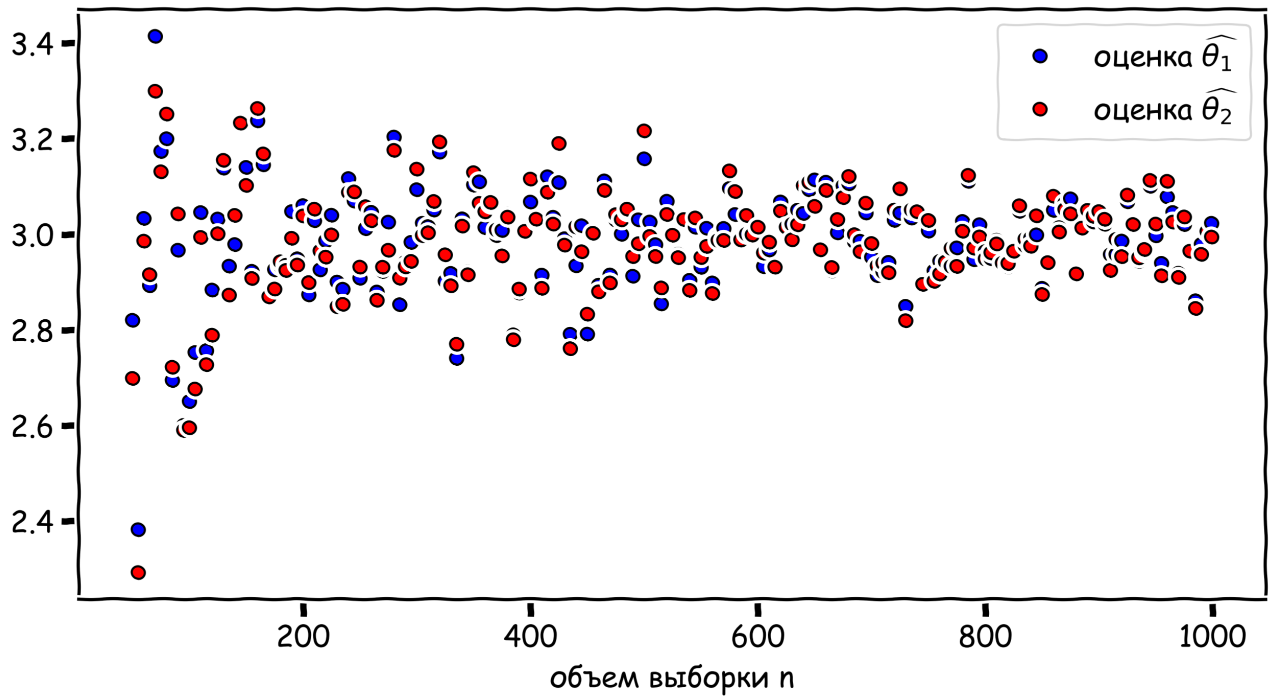

Рис. 3: Зависимость $\widehat{\theta_1}$ и $\widehat{\theta_2}$ от объема выборки n

Пример 2.1.2 (ОММ для равномерного распределения $\mathsf{U}_{0,\theta}$)
Пусть $X_1, \ldots, X_n$ *– выборка из равномерного распределения* $\mathsf{U}_{0,\theta}$, *где* $\theta \in \Theta = (0, +\infty)$. *Оценим параметр* θ *методом моментов. Пусть* $g(t) = t^k$.

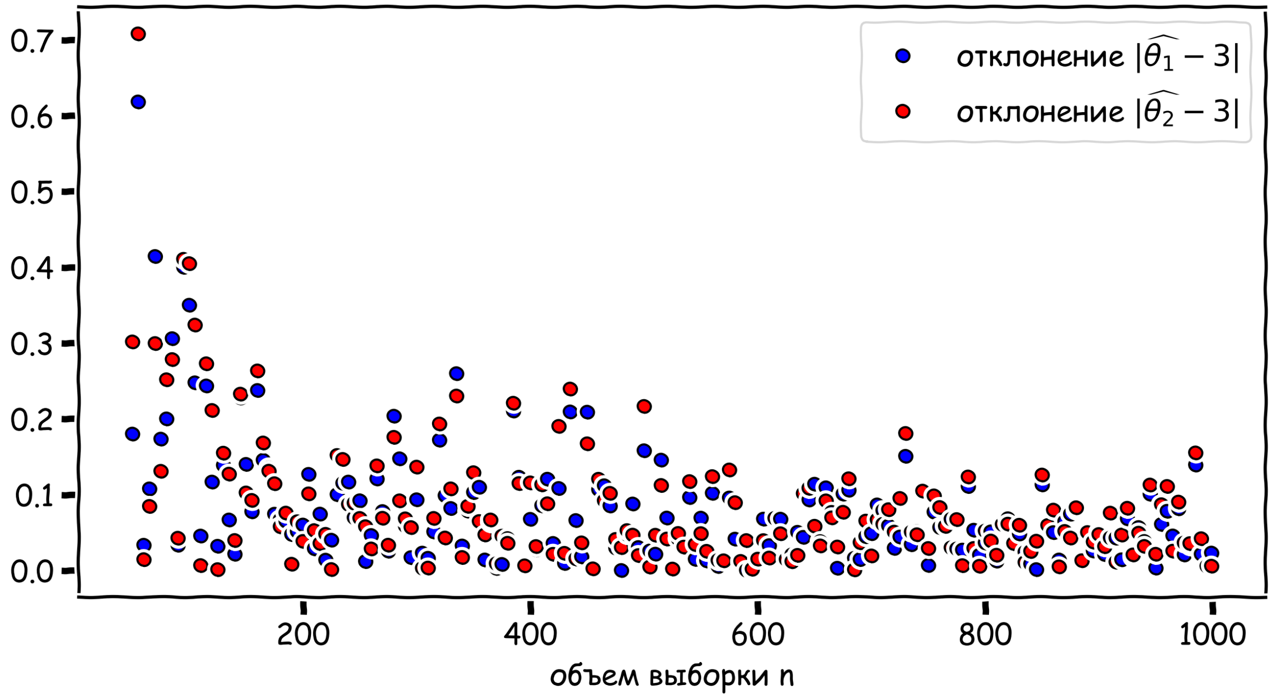

Рис. 4: Графики отклонений $|\widehat{\theta}_i - 2|$, $i \in \{1, 2\}$ от n

Так как генеральная совокупность ξ имеет равномерное распределение $\mathsf{U}_{0,\theta}$, то ее плотность задается соотношением

$$f_\xi(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & x \notin [0, \theta] \end{cases}$$

Тогда момент k-ого порядка вычисляется следующим образом:

$$\mathsf{E}_\theta X_1^k = \int\limits_0^\theta x^k \frac{1}{\theta} dx = \frac{\theta^k}{k+1}.$$

Функция, стоящая в правой части равенства, обратима при $\theta > 0$, причем

$$\theta = \sqrt[k]{(k+1)\mathsf{E}_\theta X_1^k}.$$

Меняя теоретический момент на выборочный, получаем оценку параметра θ

$$\widehat{\theta} = \sqrt[k]{(k+1)\overline{X^k}} = \sqrt[k]{\frac{k+1}{n} \sum_{i=1}^n X_i^k}.$$

Разберем еще один пример, полезный в дальнейшем.

Пример 2.1.3 (ОММ для равномерного распределения $\mathsf{U}_{\theta,\theta+1}$)
Пусть $X_1, ..., X_n$ – выборка из равномерного распределения $\mathsf{U}_{\theta,\theta+1}$, где $\theta \in \Theta = \mathbb{R}$. Оценим параметр θ методом моментов. Плотность генеральной совокупности ξ равна

$$f_\xi(x) = \begin{cases} 1, & x \in [\theta, \theta+1] \\ 0, & x \notin [\theta, \theta+1] \end{cases}$$

Момент 1-ого порядка вычисляется, как

$$\mathsf{E}X_1 = \int\limits_\theta^{\theta+1} x dx = \frac{(\theta+1)^2 - \theta^2}{2} = \frac{2\theta+1}{2}.$$

Значит,

$$\theta = \frac{2\mathsf{E}X_1 - 1}{2},$$

откуда

$$\widehat{\theta} = \frac{2\overline{X} - 1}{2}.$$

2.2 Состоятельность оценки метода моментов

Как мы видим, с помощью метода моментов мы можем получать сразу множество оценок одного и того же параметра. Конечно интересно, обладают ли полученные оценки как минимум свойством состоятельности? Конечно, если бы это было не так, то метод был бы, скажем так, провальным.

Теорема 2.2.1 (Состоятельность ОММ) *Пусть $\widehat{\theta} = f^{-1}\left(\overline{g(X)}\right)$ – оценка параметра θ, полученная по методу моментов, описанному выше, причем функция f^{-1} непрерывна. Тогда оценка $\widehat{\theta}$ состоятельна.*

Доказательство. Как мы знаем, согласно ЗБЧ в форме Хинчина,

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^n g(X_i) \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}_\theta g(X_1) = f(\theta).$$

Коль скоро функция f^{-1} непрерывна, то и

$$\widehat{\theta} = f^{-1}\left(\overline{g(X)}\right) \xrightarrow[n\to\infty]{\mathsf{P}} f^{-1}\left(\mathsf{E}_\theta g(X_1)\right) = f^{-1}\left(f(\theta)\right) = \theta,$$

а значит оценка состоятельна. $\square$

Для практического применения последней теоремы полезно помнить следующий элементарный факт из математического анализа: для обратимой функции $f : \mathbb{R} \to \mathbb{R}$ непрерывность f и непрерывность f^{-1} суть понятия равносильные. В частности из этого замечания следует, что все ОММ, полученные нами выше, как для распределения Пуассона, так и для равномерного распределения, состоятельны.

В то же время, несмещенность – свойство, скорее, случайное, и гарантировать его в общем случае не представляется возможным.

2.3 Метод моментов для многомерного параметра

Предположим, что теперь $\theta \in \Theta \subset \mathbb{R}^p$. Идея метода совершенно аналогична, только теперь вместо одной функции $g : \mathbb{R} \to \mathbb{R}$ введем p таких функций и по каждой из них построим, соответственно, функцию f_i:

$$\mathsf{E}_\theta g_i(X_1) = f_i(\theta) = f_i(\theta_1, ..., \theta_p), \quad i \in \{1, ..., p\}$$

Если полученная система из p уравнений разрешима, то есть мы можем выразить компоненты параметра θ через $\mathsf{E}_\theta g_i(X_1)$ так, что

$$\theta_i = \theta_i(\mathsf{E}_\theta g_1(X_1), ..., \mathsf{E}_\theta g_p(X_1)),$$

то снова, заменив истинные моменты выборочными, получим

$$\widehat{\theta_i} = \theta_i(\overline{g_1(X_1)}, ..., \overline{g_p(X_1)}),\ i \in \{1, ..., p\},$$

и $\widehat{\theta} = (\widehat{\theta_1}, ..., \widehat{\theta_p})$ – вектор оценок. Подобно одномерному случаю вводят следующее определение.

Определение 2.3.1 *Оценка $\widehat{\theta}$ параметра θ, полученная по описанной выше схеме, называется оценкой метода моментов (ОММ) для параметра θ.*

Поясним все высказанное на следующем достаточно простом примере.

Пример 2.3.1 (ОММ нормального распределения) *Пусть $X_1, ..., X_n$ – выборка из нормального распределения $\mathsf{N}_{\theta_1,\theta_2}$, где $\theta_1 \in \mathbb{R}$, $\theta_2 > 0$, то есть $\Theta = \mathbb{R} \times (0, +\infty)$. Взяв функцию $g_1(t) = t$, получим*

$$\mathsf{E}_\theta X_1 = \theta_1,$$

так как первый параметр нормального распределения отвечает за математическое ожидание. Теперь рассмотрим $g_2(t) = t^2$. Абсолютно аналогично случаю с распределением Пуассона заметим, что $\mathsf{D}_\theta X_1 = \mathsf{E}_\theta X_1^2 - (\mathsf{E}_\theta X_1)^2$, откуда

$$\mathsf{E}_\theta X_1^2 = \mathsf{D}_\theta X_1 + (\mathsf{E}_\theta X_1)^2 = \theta_2 + \theta_1^2.$$

Заменяя истинные моменты на выборочные, приходим к системе уравнений

$$\begin{cases} \overline{X} = \widehat{\theta}_1 \\ \overline{X^2} = \widehat{\theta}_2 + \widehat{\theta}_1^2 \end{cases} \Leftrightarrow \begin{cases} \widehat{\theta}_1 = \overline{X} \\ \widehat{\theta}_2 = \overline{X^2} - \overline{X}^2 \end{cases}$$

В итоге, метод моментов дает нам уже известные оценки для математического ожидания (выборочное среднее $\overline{X}$) и дисперсии (смещенная выборочная дисперсия S^2).

Замечание 2.3.1 *Может так случиться, что оценка $\widehat{\theta}$ не принадлежит множеству Θ. Тогда оценку несколько корректируют. Например, пусть $X_1, ..., X_n$ – выборка объема n из нормального распределения $\mathsf{N}_{\theta,1}$, причем $\Theta = [0, +\infty)$. Как мы знаем, выборочное среднее, конечно, является оценкой для θ, то есть можно считать, что*

$$\widehat{\theta} = \overline{X}.$$

В то же время на реализации конкретной выборки значение $\overline{X}$ может оказаться отрицательным, что плохо, ведь $\theta \geq 0$. Тогда резонно поправить оценку, например, следующим образом:

$$\widehat{\theta} = \max(0, \overline{X}).$$

Вопрос об исправлении оценки решается каждый раз по-своему и индивидуально. Мы не будем касаться этих вопросов детально.

3 Метод максимального правдоподобия

Еще один важный метод нахождения оценок – так называемый метод максимального правдоподобия. Пусть $X_1, ..., X_n$ – выборка из семейства распределений $\mathcal{P}_\theta$, $\theta \in \Theta$. Метод состоит в том, что «наиболее правдоподобным» значением параметра θ нарекается то, при котором вероятность получить данную выборку $(X_1, ..., X_n)$ максимальна. Что это означает? Давайте попробуем понять сказанное на следующем модельном примере.

3.1 Модельный пример

Предположим, что эксперимент заключается в стрельбе по мишени из пистолета. Всего проведено 10 выстрелов, из которых 4 раза по мишени мы попали, а 6 раз промахнулись, и выборка такова (1 отвечает попаданию, 0 – промаху):

$$(1, 0, 0, 1, 1, 0, 0, 0, 1, 0).$$

Как оценить вероятность попадания в мишень? В такой постановке задача достаточно простая, и ответ напрашивается сразу, не так ли? Давайте попробуем применить (а заодно и объяснить) идею метода максимального правдоподобия (ММП).

Генеральная совокупность ξ имеет распределение Бернулли $\xi \in \mathsf{B}_\theta$ с параметром $\theta \in \Theta = (0,1)$. Мы хотим максимизировать вероятность нашей выборки, то есть максимизировать функцию

$$f_\theta(\vec{X}) = \mathsf{P}_\theta\left(X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1, ..., X_{10} = 0\right).$$

Выборка – это последовательность независимых и одинаково распределенных случайных величин, а значит

$$f_\theta(\vec{X}) = \mathsf{P}_\theta(X_1 = 1) \cdot \mathsf{P}_\theta(X_2 = 0) \cdot \mathsf{P}_\theta(X_3 = 0) \cdot \mathsf{P}_\theta(X_4 = 1) \cdot ... \cdot \mathsf{P}_\theta(X_{10} = 0).$$

Полученная нами функция в общем случае называется функцией правдоподобия. Заметим, что в нашем случае, $\mathsf{P}_\theta(X_i = 0) = 1 - \theta$, $\mathsf{P}_\theta(X_i = 1) = \theta$. Значит, введенная функция преобразуется к виду

$$f_\theta(\vec{X}) = \theta^4(1-\theta)^6,$$

а нашей целью становится нахождение максимума этой функции. Это можно пытаться делать «в лоб», используя аппарат дифференциального исчисления, но с произведением часто (хотя и не в этом случае) бывает «все сложно», поэтому предлагается логарифмировать (чтобы получить сумму). Логарифм – монотонная функция, поэтому экстремумы переходят в экстремумы, новых экстремумов не появляется, а старые – не теряются. Посему приходим к понятию логарифмической функции правдоподобия

$$L_\theta(\vec{X}) = \ln f_\theta(\vec{X}) = 4\ln\theta + 6\ln(1-\theta).$$

Полученную функцию и будем максимизировать, ведь теперь производная берется очень просто:

$$\left(L_\theta(\vec{X})\right)'_\theta = \frac{4}{\theta} - \frac{6}{1-\theta} = \frac{4-10\theta}{\theta(1-\theta)}.$$

В области $\Theta = (0,1)$ лежит только одна точка, подозрительная на экстремум – это точка $\frac{4}{10} = 0.4$. Используя достаточное условие экстремума стандартно проверяется, что это точка максимума. Значит, вероятность «нашей выборке случиться» максимальна при $\theta = 0.4$. Удивительно? Да вовсе нет, это было понятно и с самого начала, без кучи проделанных вычислений. К сожалению, так просто бывает не всегда.

3.2 Общее описание метода максимального правдоподобия

Перейдем теперь к общему описанию метода максимального правдоподобия. На самом деле мы проделаем ровно те же выкладки, что и в примере выше, только в общем случае. Пусть $(X_1, ..., X_n)$ – выборка из параметрического семейства распределений $\mathcal{P}_\theta$. Чтобы не обсуждать дискретный и непрерывный случаи отдельно, положим

$$f_\theta(t) = \begin{cases} \text{плотность } f_\theta(t), & \text{если распределение } \mathcal{P}_\theta \text{ абсолютно непрерывно} \\ \mathsf{P}_\theta(X_1 = t), & \text{если распределение } \mathcal{P}_\theta \text{ дискретно} \end{cases}$$

Последнее обуславливается тем, что в абсолютно непрерывном случае вероятность попадания случайной величины ξ в точку (конечно равна 0) «почти равна» значению плотности в этой точке, точнее

$$\mathsf{P}(\xi \in (t, t + \Delta t)) = \int\limits_{\Delta t} f_\xi(t)dt \approx f_\xi(t)\Delta t.$$

Последнюю запись можно подвергнуть большой критике, но мы надеемся, что оно прольет больше понимания, а использовать мы эту «формулу» все равно не будем.

Итак, как обсуждалось ранее, мы хотим максимизировать вероятность получения нашей выборки. Так как выборка – последовательность независимых одинаково распределенных случайных величин, то возникает так называемая функция правдоподобия.

Определение 3.2.1 *Функция*

$$f_\theta(\vec{X}) = f_\theta(X_1) \cdot f_\theta(X_2) \cdot ... \cdot f_\theta(X_n) = \prod_{i=1}^{n} f_\theta(X_i)$$

называется функцией правдоподобия.

Замечание 3.2.1 *Увидьте (!), что в случае, когда генеральная совокупность имеет дискретное распределение, функция правдоподобия переписывается в виде*

$$f_\theta(\vec{X}) = \mathsf{P}_\theta(X_1 = x_1) \cdot \mathsf{P}_\theta(X_2 = x_2) \cdot ... \cdot \mathsf{P}_\theta(X_n = x_n).$$

Иными словами, в общем случае мы вторим той же аргументации, что у нас возникала ранее в примере.

Введенную функцию правдоподобия нам нужно максимизировать. Для этого, например, можно пытаться использовать дифференциальное исчисление функции одной (если параметр θ – скаляр) или нескольких (если θ – вектор) переменных. Что же называется оценкой оцениваемого параметра θ?

Определение 3.2.2 *Оценкой максимального правдоподобия (ОМП) $\widehat{\theta}$ параметра θ называется такое значение $\widehat{\theta} \in \Theta$, при котором функция $f_{\widehat{\theta}}(\vec{X})$ достигает локального максимума.*

Вместе с функцией правдоподобия рассмотрим так называемую логарифмическую функцию правдоподобия.

Определение 3.2.3 *Функция*

$$L_\theta(\vec{X}) = \ln f_\theta(\vec{X}) = \sum_{i=1}^{n} \ln f_\theta(X_i)$$

называется логарифмической функцией правдоподобия.

Как и ранее отметим, что поскольку функция $\ln t$ монотонна, то точки экстремума функций $f_\theta(\vec{X})$ и $L_\theta(\vec{X})$ совпадают. Поэтому оценкой максимального правдоподобия можно назвать не только точку максимума функции правдоподобия, но и точку максимума логарифмической функции правдоподобия.

Еще раз акцентируем внимание на том, что мы ищем не максимум функции, а аргумент, при котором исследуемая функция имеет максимум. По этим соображениям для ОМП часто пишут следующее определение:

$$\widehat{\theta} = \arg\max_{\theta} f_\theta(\vec{X}) = \arg\max_{\theta} L_\theta(\vec{X}).$$

3.3 Примеры применения метода максимального правдоподобия

Покажем на примерах, как можно применить метод максимального правдоподобия для решения различных задач, а также сравним полученные результаты с методом моментов.

Пример 3.3.1 (ОМП для распределения Пуассона) *Пусть $X_1, ..., X_n$ – выборка объема n из распределения Пуассона Π_θ, $\theta \in \Theta = (0, +\infty)$. В нашем случае*

$$f_\theta(t) = \mathsf{P}(X_1 = t) = \frac{\theta^t}{t!} e^{-\theta}, \; t \in \{0, 1, ...\}.$$

Тогда функция правдоподобия принимает вид

$$f_\theta(\vec{X}) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} \frac{\theta^{X_i}}{X_i!} e^{-\theta} = \frac{\theta^{X_1+X_2+...+X_n}}{X_1! \cdot X_2! \cdot ... \cdot X_n!} e^{-n\theta}.$$

Для нахождения точек максимума, рассмотрим логарифмическую функцию правдоподобия:

$$L_\theta(\vec{X}) = \ln f_\theta(\vec{X}) = \ln\left(\frac{\theta^{X_1+X_2+\ldots+X_n}}{X_1!\cdot X_2!\cdot\ldots\cdot X_n!}e^{-n\theta}\right).$$

По свойствам логарифма, последнее равенство перепишем в виде

$$L_\theta(\vec{X}) = \ln\theta^{X_1+X_2+\ldots+X_n} + \ln e^{-n\theta} - \ln\left(X_1!\cdot X_2!\cdot\ldots\cdot X_n!\right) =$$

$$= (X_1+X_2+\ldots+X_n)\ln\theta - n\theta - \ln\prod_{i=1}^{n} X_i!.$$

Рассматриваемая нами функция дифференцируема, а значит необходимым условием экстремума является равенство нулю производной по θ. Значит, нужно решить уравнение

$$\left(L_\theta(\vec{X})\right)'_\theta = 0 \text{ или } \frac{X_1+X_2+\ldots+X_n}{\theta} - n = 0,$$

откуда

$$\widehat{\theta} = \frac{X_1+X_2+\ldots+X_n}{n} = \overline{X}.$$

Конечно, «крышка» навешена параметру θ пока что неправомочно, ведь мы не показали, что найденная точка – это точка максимума. Это можно сделать, например, воспользовавшись часто называемым в литературе «вторым достаточным условием экстремума». Так как

$$\left(L_\theta(\vec{X})\right)''_\theta = -\frac{X_1+X_2+\ldots+X_n}{\theta^2} < 0,$$

то можно утверждать, что $\widehat{\theta}$ и правда является точкой максимума, а значит и крышка навешена заслуженно.

Замечание 3.3.1 *На всякий случай для тех, кто забыл: схема поиска экстремумов функции одной или нескольких переменных подробно описана в дополнительных материалах к данной лекции.*

Замечание 3.3.2 *То, что полученная точка является точкой максимума, можно непосредственно понять из вида функции $L_\theta(\vec{X})$ – она выпукла вверх. У выпуклой вверх функции экстремумами могут быть только максимумы.*

В рассмотренном примере оценка ММП совпала с оценкой метода моментов, полученной нами ранее. При этом ранее такую же оценку мы получили в разы легче. Однако оценки ММП вовсе не всегда совпадают с ОММ.

Пример 3.3.2 (ОМП для равномерного распределения) *Пусть* $X_1, ..., X_n$ *– выборка из равномерного распределения* $\mathsf{U}_{0,\theta}$, *где* $\theta \in \Theta = (0, +\infty)$. *Оценим параметр* θ *методом максимального правдоподобия. Так как генеральная совокупность* ξ *имеет равномерное распределение* $\mathsf{U}_{0,\theta}$, *то ее плотность задается соотношением*

$$f_\xi(t) = \begin{cases} \frac{1}{\theta}, & t \in [0, \theta] \\ 0, & t \notin [0, \theta] \end{cases}.$$

Но тогда функция правдоподобия имеет вид

$$f_\theta(\vec{X}) = \begin{cases} \frac{1}{\theta^n}, & \text{если } 0 \le X_i \le \theta, \quad i \in \{1, ..., n\} \\ 0, & \text{иначе} \end{cases}$$

или, что то же самое,

$$f_\theta(\vec{X}) = \begin{cases} \frac{1}{\theta^n}, & \max(X_1, ..., X_n) = X_{(n)} \le \theta \\ 0, & \text{иначе} \end{cases}$$

Из вида последней функции ясно, что ее максимум достигается при $\theta = X_{(n)} = \widehat{\theta}$.

Рассмотренный пример показывает, что, во-первых, при поиске максимума вовсе не всегда нужно «бежать дифференцировать». Кроме того, переход к логарифмической функции правдоподобия в приведенном примере тоже неуместен. И, наконец, последнее. Полученная ОМП не совпадает ни с одной оценкой, полученной методом моментов (а похожа ли на нее?). Напомним, что согласно методу моментов, мы получили набор оценок

$$\widehat{\theta_{k,n}} = \sqrt[k]{(k+1)\overline{X^k}} = \sqrt[k]{\frac{k+1}{n} \sum_{i=1}^{n} X_i^k}.$$

Индексами k, n специально обозначена зависимость оценки как от k, так и от n. Отметим, что все-таки оценки ММ связаны с оценками МП, а именно: при фиксированном n

$$\lim_{k \to +\infty} \widehat{\theta_{k,n}} = X_{(n)},$$

то есть ОММ в пределе (при $k \to +\infty$) сходятся к ОМП.

Давайте сравним полученные оценки по аналогии с тем, как мы это делали ранее. Выборка берется из генеральной совокупности, имеющей распределение $\mathsf{U}_{0,\theta}$ при $\theta = 4$. Оценку метода максимального правдоподобия будем обозначать $\widehat{\theta}$, $\widehat{\theta}_2$ – оценка метода моментов по второму моменту, $\widehat{\theta}_7$ – оценка метода моментов по седьмому моменту. Из рисунка 5 видно, что все оценки при увеличении объема выборки n приближаются к истинному параметру 4. Кроме того, рисунок 6 показывает, что оценка МП намного лучше оценок ММ. При этом оценка ММ по старшему моменту (седьмому) лучше, чем по младшему (второму).

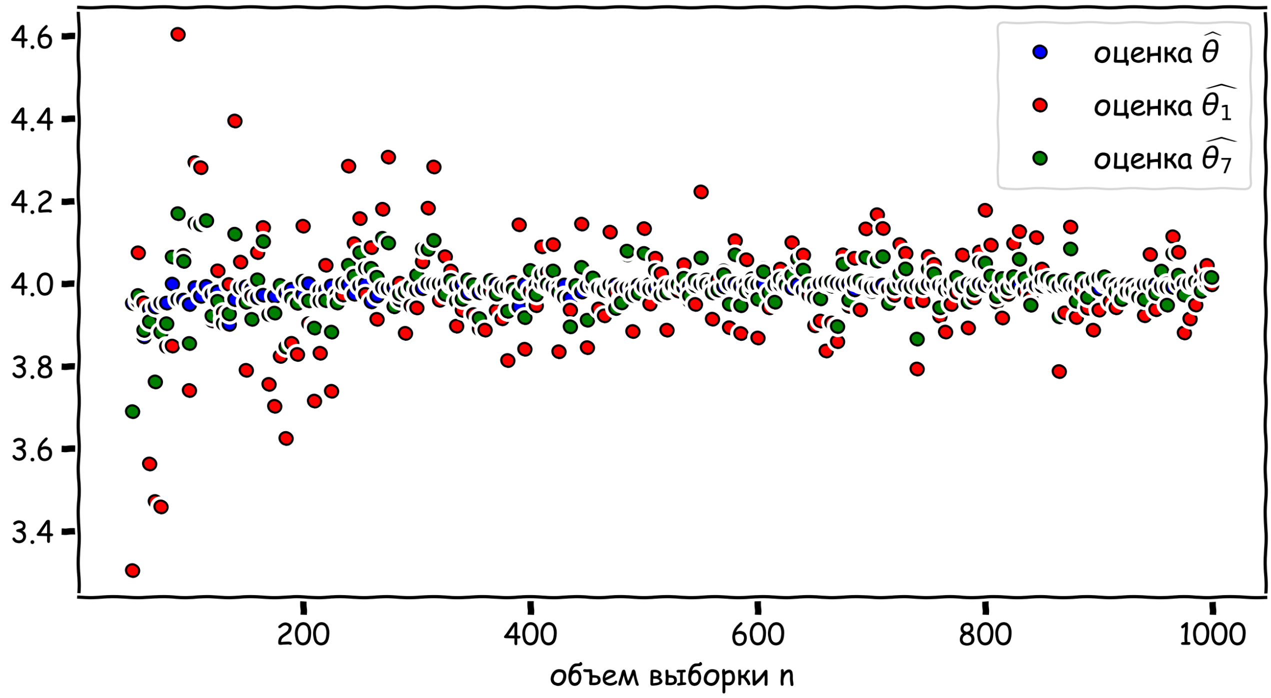

Рис. 5: Зависимость оценок $\widehat{\theta}$, $\widehat{\theta}_2$ и $\widehat{\theta}_7$ от объема выборки n

Замечание 3.3.3 *С точки зрения качества оценки (а что это такое мы пока с вами не знаем) можно отметить, что оценки МП, как правило, предпочтительнее оценок ММ, так как являются более точными. Может быть это сглаживает тот факт, что их, как правило, сложнее получать. Мы в этом убедились на практике, рассматривая приведенные примеры.*

Покажем, как можно использовать ММП для получения оценок параметра θ, являющегося векторным.

Пример 3.3.3 (ОМП для нормального распределения) *Пусть* $X_1, ..., X_n$ *– выборка из нормального распределения* $\mathsf{N}_{\theta_1,\theta_2}$*, где* $\theta \in \Theta = \mathbb{R} \times (0, +\infty)$*. Оценим параметр* $\theta = (\theta_1, \theta_2)$ *методом максимального правдоподобия. Так как генеральная совокупность* ξ *имеет нормальное распределение* $\mathsf{N}_{\theta_1,\theta_2}$*, то ее плотность задается соотношением*

$$f_\xi(t) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(t-\theta_1)^2}{2\theta_2}}.$$

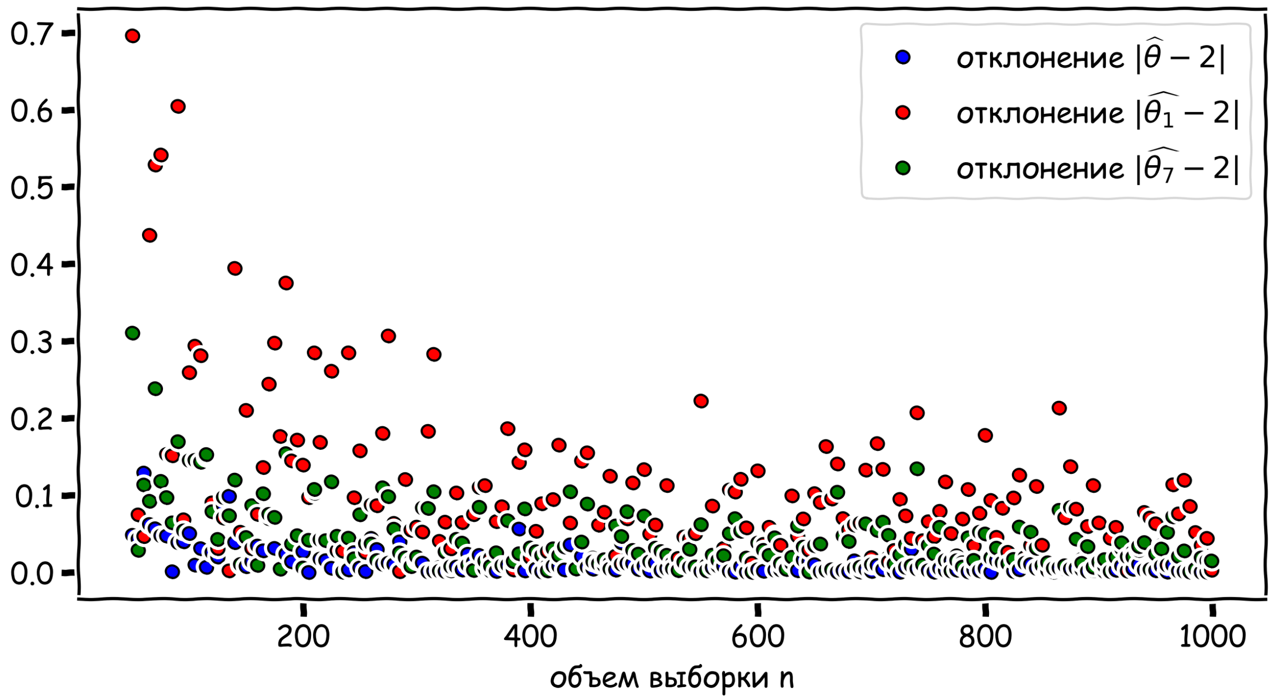

Рис. 6: Графики отклонений $|\widehat{\theta}-4|$ и $|\widehat{\theta}_i-4|$, $i\in\{2,7\}$ от объема n

Функция максимального правдоподобия задается выражением

$$f_\theta(\vec{X})=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\theta_2}}e^{-\frac{(X_i-\theta_1)^2}{2\theta_2}}=\frac{1}{(2\pi\theta_2)^{n/2}}e^{-\frac{\sum\limits_{i=1}^{n}(X_i-\theta_1)^2}{2\theta_2}}.$$

После некоторых преобразований, логарифмическая функция правдоподобия примет вид

$$L_\theta(\vec{X})=-\frac{\sum\limits_{i=1}^{n}(X_i-\theta_1)^2}{2\theta_2}-\frac{n}{2}\ln\left(2\pi\theta_2\right).$$

Необходимым условием экстремума дифференцируемой функции двух переменных является равенство нулю частных производных по всем переменным. В данном случае,

$$\begin{cases}\frac{\partial L_\theta}{\partial\theta_1}=0\\ \frac{\partial L_\theta}{\partial\theta_2}=0\end{cases}\Leftrightarrow\begin{cases}\frac{\sum\limits_{i=1}^{n}(X_i-\theta_1)}{\theta_2}=0\\ \frac{\sum\limits_{i=1}^{n}(X_i-\theta_1)^2}{2\theta_2^2}-\frac{n}{2\theta_2}=0\end{cases}\Rightarrow\begin{cases}\theta_1=\frac{1}{n}\sum\limits_{i=1}^{n}X_i=\overline{X}\\ \theta_2=\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2=S^2\end{cases}.$$

Итак, если мы докажем, что в рассматриваемой точке – максимум, то можно утверждать, что $\widehat{\theta}=(\widehat{\theta}_1,\widehat{\theta}_2)=(\overline{X},S^2)$ *– оценка максимального правдоподобия. Докажем это в замечании к данному примеру, чтобы не нагромождать идейные моменты вычислениями.*

В итоге, оценки максимального правдоподобия для параметров нормального распределения совпали с оценками метода моментов. В этом случае использование метода моментов было куда проще, чем ММП.

Замечание 3.3.4 (Обоснование предыдущего примера) *Для определения того, является ли полученная точка, подозрительная на экстремум, экстремумом, воспользуемся стандартным достаточным условием. Вычислим вторые производные функции $L_\theta(\vec{X})$:*

$$\frac{\partial^2 L_\theta}{\partial \theta_1^2} = -\frac{n}{\theta_2} < 0,$$

$$\frac{\partial^2 L_\theta}{\partial \theta_1 \partial \theta_2} = -\frac{\sum\limits_{i=1}^{n}(X_i - \theta_1)}{\theta_2^2} = \left|\theta_1 = \overline{X}\right| = -\frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})}{\theta_2^2} = 0,$$

$$\frac{\partial^2 L_\theta}{\partial \theta_2^2} = -\frac{\sum\limits_{i=1}^{n}(X_i - \theta_1)^2}{\theta_2^3} + \frac{n}{2\theta_2^2}.$$

Осталось определить знаки главных миноров матрицы, составленной из вторых производных функции L_θ:

$$\frac{\partial^2 L_\theta}{\partial \theta_1^2} \cdot \frac{\partial^2 L_\theta}{\partial \theta_2^2} = \frac{n}{\theta_2^4} \sum_{i=1}^{n}(X_i - \theta_1)^2 - \frac{n^2}{2\theta_2^3} = \left| \begin{array}{c} \theta_1 = \overline{X} \\ \theta_2 = S^2 \end{array} \right| =$$

$$= \frac{n}{S^8} \sum_{i=1}^{n}(X_i - \overline{X})^2 - \frac{n^2}{2S^6} = \frac{n^2}{S^8} S^2 - \frac{n^2}{2S^6} = \frac{n^2}{2S^6} > 0.$$

Тем самым, вторая квадратичная форма определена отрицательно, а значит (θ_1, θ_2) является точкой максимума функции L_θ.

Отметим важные детали. Во-первых, оценка максимального правдоподобия не всегда существует. Кроме того, если она и существует, она вовсе не обязана быть единственной.

Пример 3.3.4 *Пусть $X_1, ..., X_n$ – выборка из равномерного распределения $\mathsf{U}_{\theta,\theta+1}$, где $\theta \in \Theta = \mathbb{R}$. Найдем ОМП для параметра θ. Ясно, что*

$$f_\xi(t) = \begin{cases} 1, & t \in [\theta, \theta+1] \\ 0, & t \notin [\theta, \theta+1] \end{cases}.$$

Тогда функция правдоподобия имеет вид

$$f_\theta(\vec{X}) = \begin{cases} 1, & \textit{все } X_i \in [\theta, \theta+1] \\ 0, & \textit{иначе} \end{cases} = \begin{cases} 1, & \theta \le X_{(1)} \le X_{(n)} \le \theta + 1 \\ 0, & \textit{иначе} \end{cases}.$$

Функция правдоподобия достигает максимума при всех $\theta \in [X_{(n)} - 1, X_{(1)}]$. Значит, любое число $\widehat{\theta} \in [X_{(n)} - 1, X_{(1)}]$ является ОМП параметра θ.

Проведем моделирование для последнего примера. Пусть $\widehat{\theta_1} = X_{(1)}$, $\widehat{\theta_2} = X_{(n)} - 1$ и выборка берется из равномерного распределения $\mathsf{U}_{2.5,3.5}$, то есть оценивается параметр $\theta = 2.5$. На рисунке 7 показана зависимость оценок от объема выборки n. Очевидно приближение значений с ростом n к оцениваемому значению. Рисунок 8 показывает отклонение оценок от истинных значений параметров. Снова «на глаз» определить, какая оценка лучше, не представляется возможным. Кстати, достаточно очевидно, что в этом случае обе оценки являются смещенными (нет одинакового разброса от истинного значения)

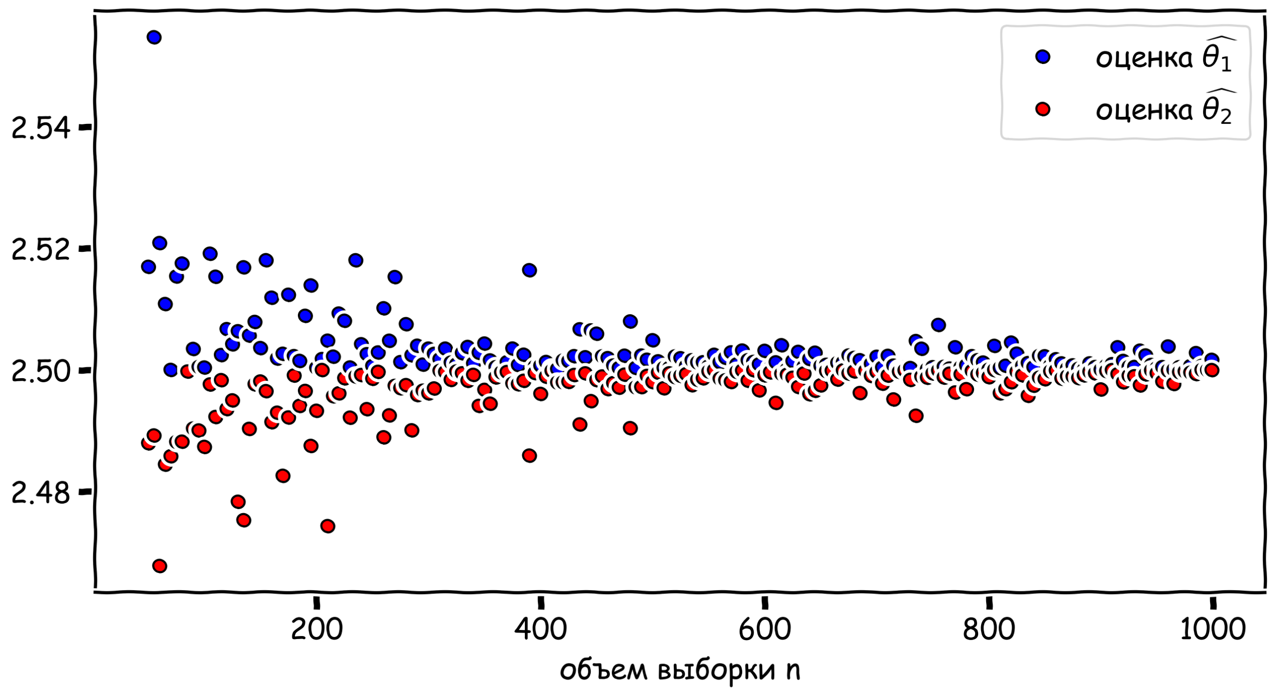

Рис. 7: Зависимость оценок $\widehat{\theta_1}$ и $\widehat{\theta_2}$ от объема выборки n

3.4 Состоятельность оценки метода максимального правдоподобия

Как мы уже не раз отмечали, одно из важнейших (и, наверное, необходимое) требование к оценке – это ее состоятельность. Оценки МП при достаточно общих условиях являются состоятельными. Сформулируем без доказательства теорему, относящуюся к случаю скалярного параметра θ. Условия этой теоремы несколько излишни, но они дадут нам в дальнейшем и другие свойства ОМП.

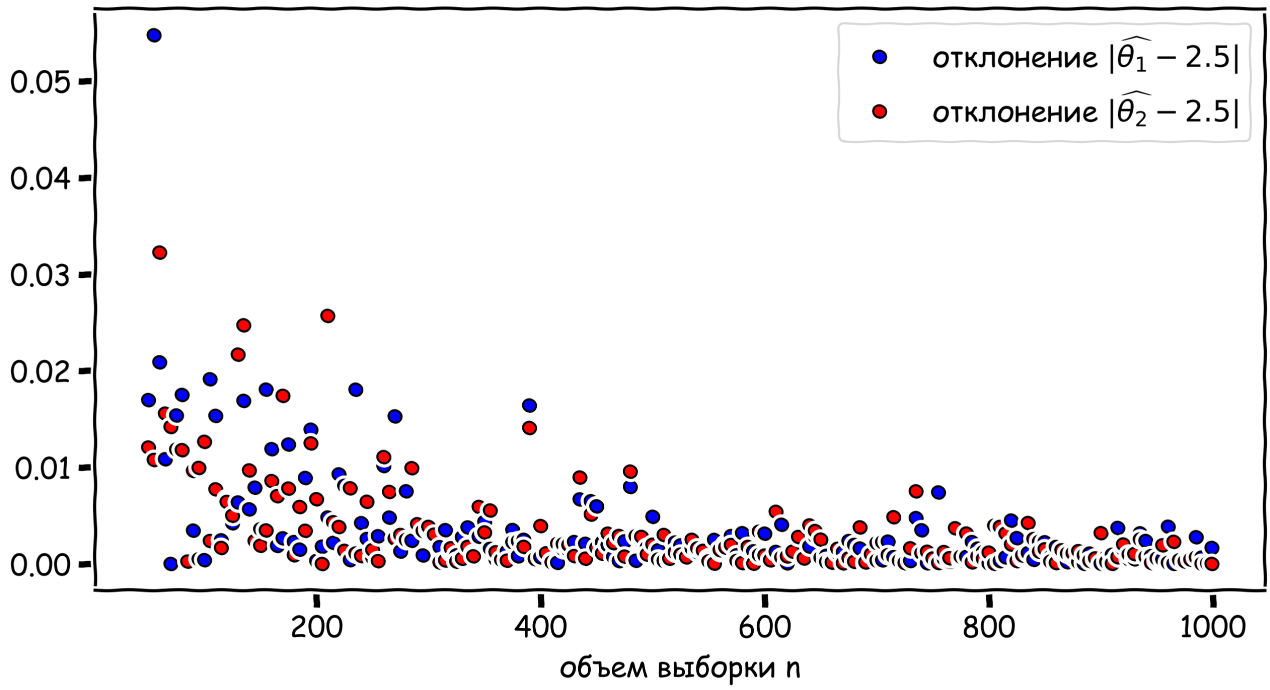

Рис. 8: Графики отклонений $|\widehat{\theta_i} - 2.5|$, $i \in \{1, 2\}$ от объема n

Теорема 3.4.1 (Состоятельность ОМП) *Пусть $\mathcal{P}_\theta$ – семейство распределений с плотностью $f_\theta(t)$, Θ – открытый интервал, возможно бесконечный, причем:*

1. *При каждом $\theta \in \Theta$ при всех $t \in \mathbb{R}$ существуют частные производные*
$$\frac{\partial \ln f_\theta(t)}{\partial \theta}, \quad \frac{\partial^2 \ln f_\theta(t)}{\partial \theta^2}, \quad \frac{\partial^3 \ln f_\theta(t)}{\partial \theta^3};$$
2. *При каждом $\theta \in \Theta$ выполняются соотношения*
$$\left|\frac{\partial f_\theta(t)}{\partial \theta}\right| \le F_1(t), \quad \left|\frac{\partial^2 f_\theta(t)}{\partial \theta^2}\right| \le F_2(t), \quad \left|\frac{\partial^3 \ln f_\theta(t)}{\partial \theta^3}\right| \le H(t),$$
где функции F_1 и F_2 интегрируемы на $\mathbb{R}$, а $\mathsf{E}_\theta H(\xi) < M$, где M не зависит от θ;
3. *При каждом $\theta \in \Theta$*
$$0 < \int\limits_{-\infty}^{+\infty} \left(\frac{\partial \log f_\theta(t)}{\partial \theta}\right)^2 f_\theta(t) dt < +\infty.$$

Тогда уравнение правдоподобия $\frac{\partial L_\theta}{\partial \theta} = 0$ имеет решение $\widehat{\theta}$, причем $\widehat{\theta}$ – состоятельная оценка параметра θ.

Отметим, что эта теорема с очевидными видоизменениями справедлива и в случае, когда семейство распределений дискретно. Можно проверить, что почти все рассмотренные нами семейства укладываются в данную теорему, а значит полученные оценки МП являются состоятельными.