ЦИФРОВАЯ КУЛЬТУРА
УНИВЕРСИТЕТ ИТМО

Точные и асимптотические доверительные интервалы

Высшая школа цифровой культуры
Университет ИТМО
dc@itmo.ru

Содержание

1 Интервальное оценивание

В предыдущих лекциях мы поговорили про так называемое точечное оценивание неизвестных параметров генеральной совокупности. Нами были изучены методы моментов и максимального правдоподобия для получения этих оценок, а также было сказано несколько слов про их свойства: несмещенность, состоятельность. Однако наличие точечной оценки параметра ничего не говорит о близости этой оценки к истинному значению. Так, поменяв выборку, мы, вообще говоря, получим другую точечную оценку. Но какая из этих оценок лучше и точнее – остается совершенно неизвестным. Нельзя ли каким-то образом, скажем так, оценить погрешность, то есть указать некоторый интервал, в котором находится истинное значение параметра? Ну конечно можно, скажете Вы, это множество – множество Θ (в случае, когда рассматривается параметрическая модель). Но отвечать на вопрос таким образом – это ровным счетом никак на него не отвечать, ведь множество Θ может иметь бесконечные «размеры», и где тут конкретика?

Тогда может быть поступить иначе? Может быть, можно задаться некоторой, вообще говоря большой, вероятностью, и искать интервал, в который истинное значение попадает с этой вероятностью? Конечно, ошибки не исключены, но вдруг свойства таких интервалов окажутся весьма привлекательными? Вдруг с ростом объема выборки длины этих интервалов будут стремиться к нулю, тем самым будет сужаться и возможный (с большой вероятностью) диапазон значений для искомого параметра?

Оказывается, такой подход хорошо работает, и этот подход называется интервальным оцениванием. Давайте к нему и приступим.

1.1 Точные (и не очень) доверительные интервалы

Пусть у нас, как обычно, имеется выборка $X_1, X_1, ..., X_n$ из семейства распределений $\mathcal{P}_\theta$, которое известным образом зависит от неизвестного параметра θ из некоторого множества Θ. На самом деле вовсе не обязательно говорить о параметрической постановке задачи. Можно говорить о доверительном интервале для некоторой числовой характеристики генеральной совокупности ξ, об этом мы еще поговорим позднее.

Определение 1.1.1 *Пусть* $0 < \varepsilon < 1$. *Интервал*

$$(\theta^-, \theta^+) = (\theta^-(X, \varepsilon), \theta^+(X, \varepsilon)),$$

где θ^-, θ^+ *– статистики, называется доверительным интервалом уровня доверия (или надежности)* $1 - \varepsilon$, *если для любого* $\theta \in \Theta$ *выполняется*

$$\mathsf{P}_\theta\left(\theta^- < \theta < \theta^+\right) \geq 1 - \varepsilon.$$

Итак, доверительный интервал уровня доверия $(1-\varepsilon)$ – это интервал, который с вероятностью не меньше, чем $(1-\varepsilon)$, накрывает (содержит) интересующий нас параметр. Концы этого интервала – функции от выборки, на теоретическом уровне – случайные величины.

Отметим сразу несколько замечаний.

Замечание 1.1.1 *Почему берется неравенство $\geq$, а не просто равенство $=$? Действительно, для абсолютно непрерывных распределений вопрос о вероятности, равной $(1-\varepsilon)$, имеет место для любого $\varepsilon \in (0,1)$, так как каждая такая вероятность достигается хотя бы потому, что функция распределения абсолютно непрерывного распределения непрерывна.*

Ситуация совершенно меняется в случае, если рассматривается дискретное распределение. Скажем, для случайной величины ξ с распределением Бернулли $\mathsf{B}_{0.5}$ с параметром $p = 0.5$, равенство

$$\mathsf{P}\left(a < \xi < b\right) = 0.75$$

невозможно в принципе ни для каких $a, b \in \mathbb{R}$. При этом если заменить равенство на неравенство $\geq$, все становится вполне осмысленным.

Определение 1.1.2 *Если в определении доверительного интервала вместо неравенства достигается равенство, то есть*

$$\mathsf{P}_\theta\left(\theta^- < \theta < \theta^+\right) = 1-\varepsilon,$$

то доверительный интервал называется точным.

Естественен еще один сразу напрашивающийся (довольно наивный) вопрос: а зачем нам этот ε? Почему нельзя положить его равным нулю и искать доверительный интервал уровня доверия 1? Можно, конечно, только этот интервал есть не что иное, как все пространство Θ. Хорошо это или плохо?

Замечание 1.1.2 *Отметим еще одно довольно простое замечание. Ясно, что доверительный интервал тем лучше, чем он уже. Да и вообще, конструкция оправдана, если только*

$$\theta^+ - \theta^- \xrightarrow[n\to+\infty]{\mathsf{P}} 0,$$

то есть если длина интервала стремится к нулю (по вероятности) с ростом значений n.

Приведем несколько примеров, иллюстрирующих общую идею построения описанных конструкций.

1.1.1 Точный доверительный интервал для N_{a,σ^2} при известной дисперсии

Пусть $X_1, X_2, ..., X_n$ – выборка из распределения $\mathsf{N}_{\theta,\sigma^2}$, где параметр θ неизвестен, $\theta \in \Theta = \mathbb{R}$, а дисперсия σ^2 известна. В лекции про точечное оценивание для оценки параметра θ удачным (несмещенным, состоятельным и асимптотически нормальным) кандидатом было выборочное среднее $\widehat{\theta} = \overline{X}$. Теперь построим доверительный интервал для этого параметра.

Вспомним из лекции по теории вероятностей, что нормальное распределение устойчиво по суммированию, то есть если $\xi_1 \sim \mathsf{N}_{a_1,\sigma_1^2}$, $\xi_2 \sim \mathsf{N}_{a_2,\sigma_2^2}$, то

$$\xi = \xi_1 + \xi_2 \sim \mathsf{N}_{a_1+a_2,\sigma_1^2+\sigma_2^2}.$$

В частности, отсюда получается, что случайная величина $X_1 + X_2 + ... + X_n$, построенная по выборке из распределения $\mathsf{N}_{\theta,\sigma^2}$ имеет нормальное распределение с какими параметрами? Правильно, с параметрами $n\theta$ и $n\sigma^2$, иными словами

$$\sum_{i=1}^{n} X_i \sim \mathsf{N}_{n\theta,n\sigma^2}.$$

Но тогда по свойствам линейных преобразований от случайных величин,

$$\sum_{i=1}^{n} X_i - n\theta \sim \mathsf{N}_{0,n\sigma^2}, \quad \frac{\sum\limits_{i=1}^{n} X_i - n\theta}{\sqrt{n}\sigma} \sim \mathsf{N}_{0,1}.$$

Последняя случайная величина имеет стандартное нормальное распределение, и может быть переписана в виде

$$\frac{\sum\limits_{i=1}^{n} X_i - n\theta}{\sqrt{n}\sigma} = \sqrt{n}\frac{\overline{X} - \theta}{\sigma}.$$

Итак, полученная нами случайная величина имеет стандартное нормальное распределение. Рассмотрим симметричный интервал $(-c, c)$ и найдем вероятность попадания в него нашей случайной величины постров, тем самым, точный доверительный интервал надежности $(1 - \varepsilon)$. Как обычно, $\Phi_{0,1}$ обозначает функцию распределения случайной величины со стандартным нормальным распределением:

$$\mathsf{P}_\theta\left(-c < \sqrt{n}\frac{\overline{X} - \theta}{\sigma} < c\right) = \Phi_{0,1}(c) - \Phi_{0,1}(-c) = 2\Phi_{0,1}(c) - 1 = 1 - \varepsilon.$$

В итоге мы приходим к уравнению

$$2\Phi_{0,1}(c) = 2 - \varepsilon \Leftrightarrow \Phi_{0,1}(c) = 1 - \frac{\varepsilon}{2}.$$

Так как распределение абсолютно непрерывно, то нам нужно найти квантиль $c = \tau_{1-\varepsilon/2}$ уровня $1 - \varepsilon/2$.

Разрешим неравенство под знаком вероятности относительно θ и получим искомый доверительный интервал, итак

$$-\tau_{1-\varepsilon/2} < \sqrt{n}\frac{\overline{X} - \theta}{\sigma} < \tau_{1-\varepsilon/2} \Leftrightarrow -\tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}} < \overline{X} - \theta < \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}}$$

и в итоге

$$\overline{X} - \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}}.$$

В общих обозначениях имеем

$$\theta^- = \overline{X} - \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}}, \quad \theta^+ = \overline{X} + \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}}.$$

Итак, мы получили наш точный доверительный интервал.

Замечание 1.1.3 *Заметим, что длина доверительного интервала с ростом объема выборки n уменьшается со скоростью порядка $n^{-1/2}$.*

Пример 1.1.1 *Известно, что в конкретный день ноября средняя температура ξ в Санкт-Петербурге имеет нормальное распределение с неизвестным средним a и известной дисперсией $\sigma^2 = 4$. Данные наблюдений представлены следующей выборкой X в градусах Цельсия:*

$$X = (-1.579, 0.759, -0.342, 2.297, 3.787, -1.15, 1.423, 1.695, 0.451, 0.646).$$

Найти доверительный интервал уровня доверия 0.95 для оценки математического ожидания θ генеральной совокупности ξ.

По выборке находим $\overline{X} = 0.7987$. Так как $\varepsilon = 0.05$, то нужно найти квантиль $\tau_{0.975}$ уровня 0.975 стандартного нормального распределения. Пользуясь таблицами получим $\tau_{0.975} \approx 1.96$. Подставим все в полученное нами выражение для доверительного интервала:

$$\left(\overline{X} - \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}},\ \overline{X} + \tau_{1-\varepsilon/2}\frac{\sigma}{\sqrt{n}}\right),$$

получим

$$\big(\theta^-(X,\varepsilon),\ \theta^+(X,\varepsilon)\big) = \left(0.7987 - 1.96 \cdot \frac{2}{\sqrt{10}},\ 0.7987 + 1.96 \cdot \frac{2}{\sqrt{10}}\right) =$$

$$= (-0.4409,\ 2.0383) \approx (-0.45, 2.04).$$

В данном примере выборка бралась из распределения $\mathsf{N}_{2,4}$, так что истинное значение θ равно 2 и оно попадает в доверительный интервал.

Протестируем доверительный интервал на синтетических выборках большего объема. Рассматриваются выборки из того же распределения и строятся доверительные интервалы того же уровня доверия 0.95. На рисунке 1 красными точками обозначены границы доверительных интервалов $\theta^{\pm}(X)$, а синими, для удобства, центры доверительных интервалов $\overline{X}$. Из рисунка 1 видно, что

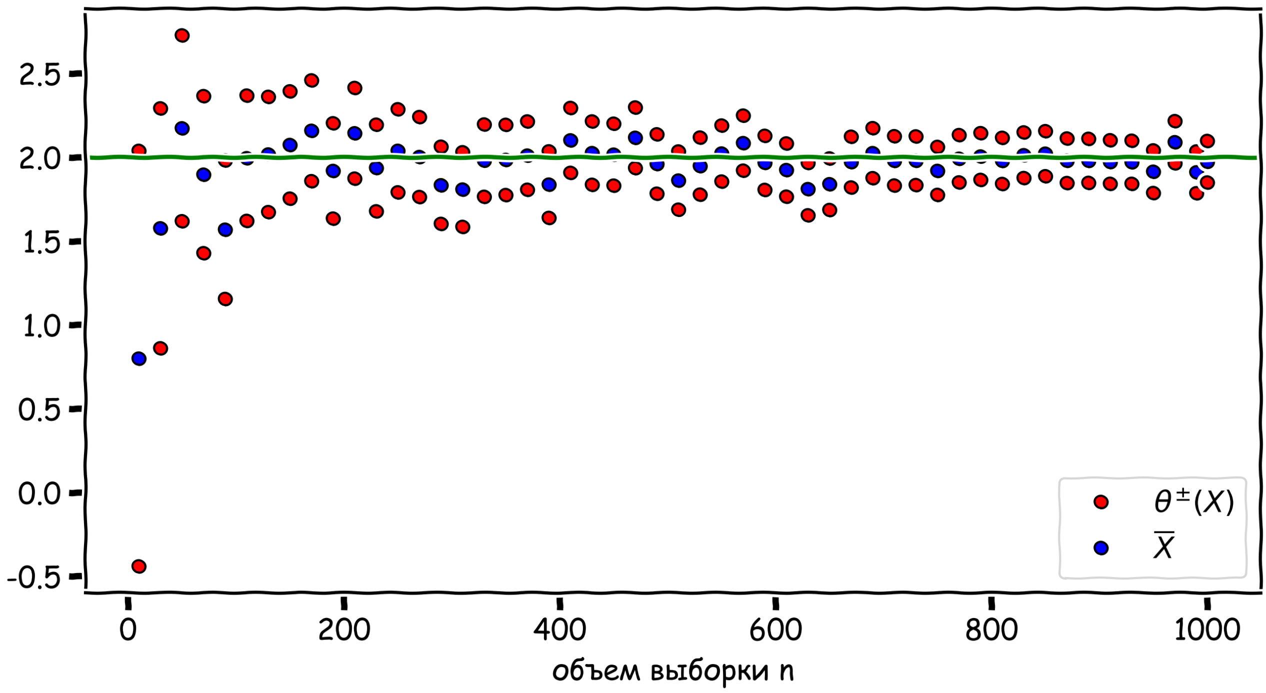

Рис. 1: Построение доверительных интервалов при разных n

зеленая линия (истинное значение среднего, равное двум) не всегда проходит между двумя красными точками, то есть не всегда попадает в доверительный интервал. Однако, в основном попадает. Кроме того хорошо видно, что длина доверительного интервала убывает с ростом n.

1.1.2 Доверительный интервал для a при неизвестном σ^2

Может оказаться так, что как a, так и σ^2 неизвестны. Построим точный доверительный интервал для параметра a при неизвестной дисперсии σ^2. Оказывается, что случайная величина

$$\sqrt{n}\frac{\overline{X}-a}{\sqrt{S_0^2}} = \sqrt{n}\frac{\overline{X}-a}{S_0}, \quad S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

имеет распределение Стьюдента T_{n-1} (справку о распределении Стьюдента, а также обоснование этого факта можно найти в дополнительных материалах). Пусть t_1 – квантиль распределения Стьюдента T_{n-1} уровня $\varepsilon/2$, а t_2 –

квантиль распределения Стюдента T_{n-1} уровня $1-\varepsilon/2$. Так как распределение Стьюдента симметрично, то $t_1 = -t_2$, а значит, если $F_{t_{n-1}}$ – функция распределения случайной величины t_{n-1}, то

$$\mathsf{P}_{a,\sigma^2}\left(-t_2 < \sqrt{n}\frac{\overline{X}-a}{S_0} < t_2\right) = F_{t_{n-1}}(t_2) - F_{t_{n-1}}(-t_2) =$$

$$= 1-\varepsilon/2-\varepsilon/2 = 1-\varepsilon.$$

Осталось выразить a, получим

$$-t_2 < \sqrt{n}\frac{\overline{X}-a}{S_0} < t_2 \Leftrightarrow \overline{X} - t_2\frac{S_0}{\sqrt{n}} < a < \overline{X} + t_2\frac{S_0}{\sqrt{n}},$$

откуда

$$\left(\theta^-,\ \theta^+\right) = \left(\overline{X} - t_2\frac{S_0}{\sqrt{n}},\ \overline{X} + t_2\frac{S_0}{\sqrt{n}}\right)$$

искомый точный доверительный интервал уровня доверия $1-\varepsilon$.

Проведем численный эксперимент при $\varepsilon = 0.05$. Пусть выборка берется из нормального распределения $\mathsf{N}_{3,4}$. На рисунке 2 видны соответсвующие доверительные интервалы: их границы нарисованы красным, середины – синим, а истинное значение $a = 3$ – зеленым.

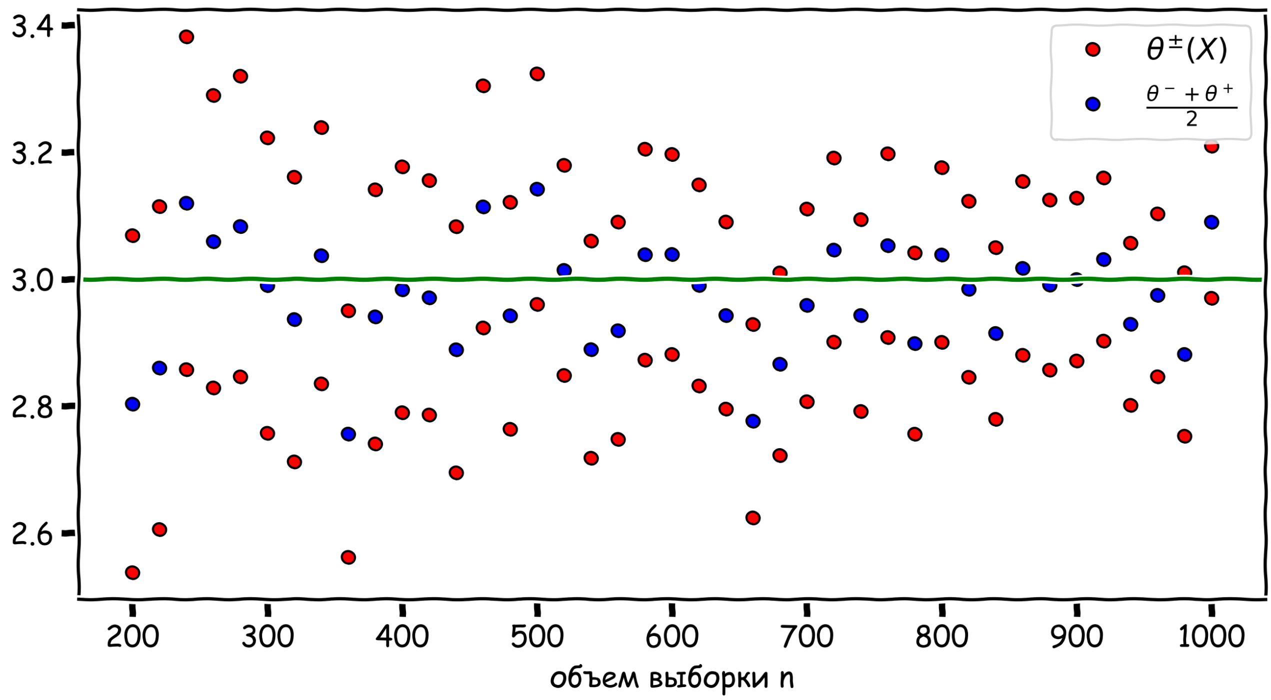

Рис. 2: Доверительный интервал для a при неизвестном σ^2

Давайте сравним, насколько влияет знание дисперсии на качество доверительного интервала. Снова $\varepsilon = 0.05$ и выборка берется из нормального

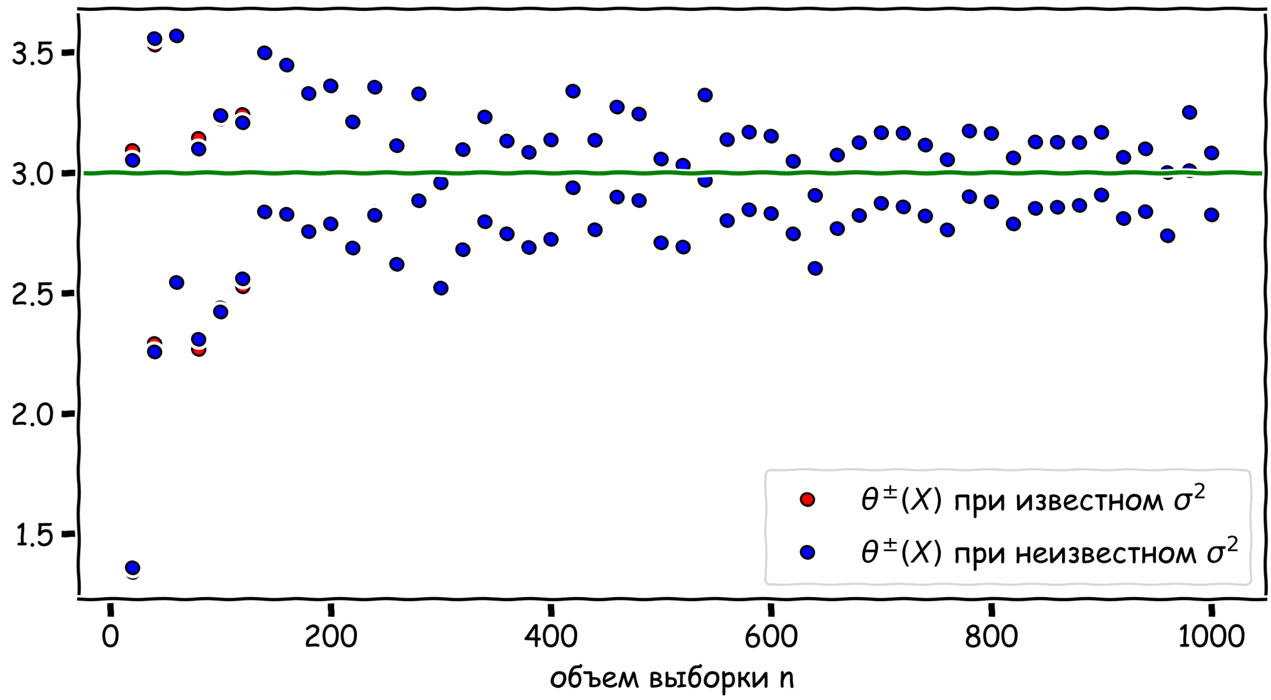

Рис. 3: Сравнение доверительных интервалов

распределения $\mathsf{N}_{3,4}$. На рисунке 3 изображены границы доверительных интервалов: красным – при известной дисперсии, синим – при неизвестной.

Важно ответить себе на вопрос, а почему этот доверительный интервал хорош? Почему его длина стремится к нулю с ростом n? В данном случае это не сразу очевидно, ведь во-первых квантиль зависит от n, а во-вторых есть сомножитель S_0, находящийся в числителе, который содержит сумму, зависящую от n.

Что же, первый вопрос нам помогает разрешить свойство распределения Стьюдента T_k, которое можно найти в дополнительных материалах: оно слабо сходится к стандартному нормальному $\mathsf{N}_{0,1}$ при $k \to +\infty$. Значит, квантили распределения Стьюдента асимптотически (!) не зависят от n. Ответ же на второй вопрос следует из состоятельности несмещенной дисперсии.

1.1.3 Доверительный интервал для σ^2 при известном a

Построим точный доверительный интервал для параметра σ^2 при известном a. Из дополнительных материалов следует, что случайная величина

$$\sum_{i=1}^{n} \left(\frac{X_i - a}{\sigma} \right)^2$$

имеет распределение Пирсона H_n (с информацией о нем можно ознакомиться в дополнительных материалах). Пусть c_1 – квантиль распределения H_n уровня $\varepsilon/2$, а c_2 – квантиль распределения H_n уровня $1-\varepsilon/2$, $F_{\chi^2_n}$ – функция распределения случайной величины χ^2_n, тогда

$$\mathsf{P}_{a,\sigma^2}\left(c_1 < \sum_{i=1}^{n}\left(\frac{X_i-a}{\sigma}\right)^2 < c_2\right) = F_{\chi^2_n}(c_2) - F_{\chi^2_n}(c_1) = 1-\frac{\varepsilon}{2}-\frac{\varepsilon}{2} = 1-\varepsilon.$$

Осталось выразить σ^2 и посмотреть, получилось ли что-то приличное:

$$c_1 < \sum_{i=1}^{n}\left(\frac{X_i-a}{\sigma}\right)^2 < c_2 \Leftrightarrow \frac{\sum\limits_{i=1}^{n}(X_i-a)^2}{c_2} < \sigma^2 < \frac{\sum\limits_{i=1}^{n}(X_i-a)^2}{c_1}.$$

Итак, интервал

$$\left(\theta^-,\ \theta^+\right) = \left(\frac{\sum\limits_{i=1}^{n}(X_i-a)^2}{c_2},\ \frac{\sum\limits_{i=1}^{n}(X_i-a)^2}{c_1}\right)$$

является точным доверительным интервалом уровня доверия $1-\varepsilon$.

Проведем численный эксперимент. Пусть выборка берется из распределения $\mathsf{N}_{3,4}$ и $\varepsilon = 0.05$. На рисунке 4 изображены границы доверительных интервалов (красными точками), их центры – синими и зеленой линией истинное значение параметра σ^2.

Здесь снова стоит задаться вопросом, а за счет чего рассматриваемый доверительный интервал стремится к нулю? Здесь вообще нет какого-либо убывающего сомножителя, а числитель, будучи умноженной на $(n-1)$ несмещенной выборочной дисперсией, в виду состоятельности последней стремится к бесконечности со скоростью порядка n! Корректный ответ на этот вопрос требует дополнительных сведений, а их, как и сам ответ, можно найти в приложенных материалах. Мы лишь отметим, что длина построенного доверительного интервала стремится к нулю со скоростью порядка $n^{-1/2}$.

1.1.4 Доверительный интервал для σ^2 при неизвестном a

Построим теперь точный доверительный интервал для параметра σ^2 при неизвестном a. Снова можно показать (и это сделано в дополнительных материалах), что случайная величина

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

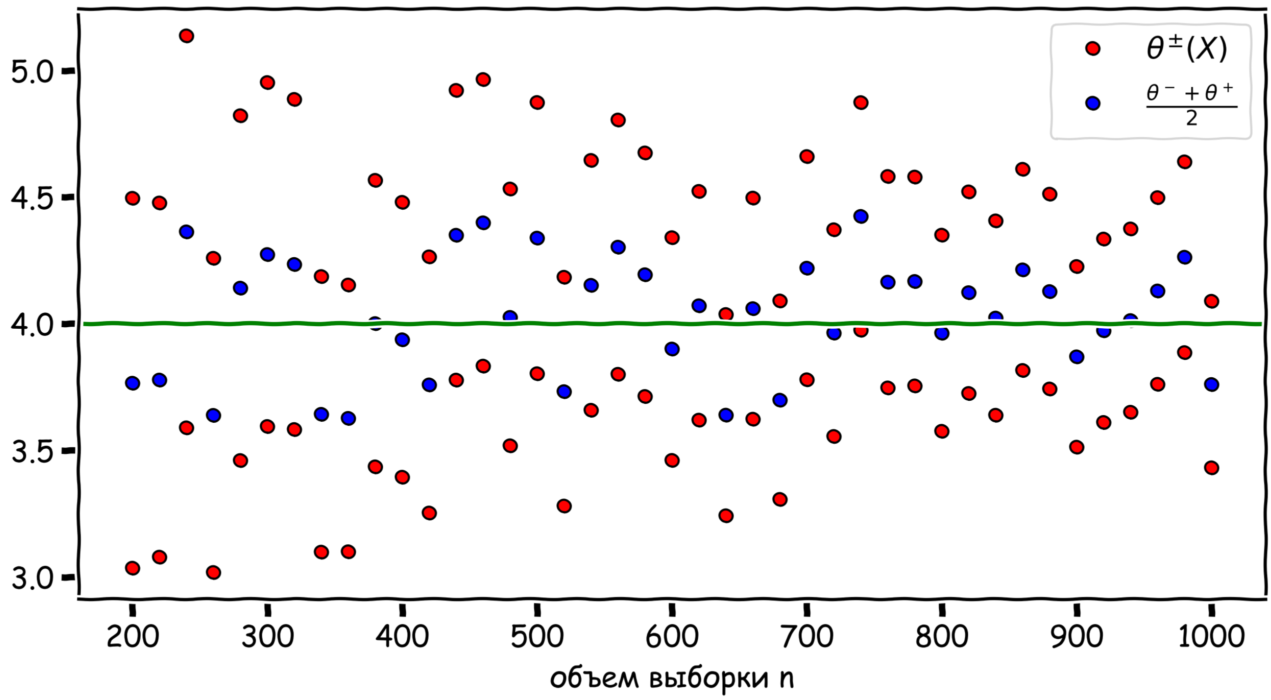

Рис. 4: Доверительный интервал для σ^2 при известном a

имеет распределение Пирсона H_{n-1}. Пусть c_1 – квантиль распределения Пирсона H_{n-1} уровня $\varepsilon/2$, а c_2 – квантиль распределения Пирсона H_{n-1} уровня $1-\varepsilon/2$, $F_{\chi^2_{n-1}}$ – функция распределения случайной величины χ^2_{n-1}, тогда

$$\mathsf{P}_{a,\sigma^2}\left(c_1 < \frac{n-1}{\sigma^2}S_0^2 < c_2\right) = F_{\chi^2_{n-1}}(c_2) - F_{\chi^2_{n-1}}(c_1) = 1 - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} = 1-\varepsilon.$$

Выразим σ^2, тогда получим

$$c_1 < \frac{n-1}{\sigma^2}S_0^2 < c_2 \Leftrightarrow \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{c_2} < \sigma^2 < \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{c_1},$$

откуда интервал

$$\left(\theta^-,\ \theta^+\right) = \left(\frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{c_2},\ \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{c_1}\right)$$

является точным доверительным интервалом уровня доверия $1-\varepsilon$. Стремление к нулю написанного интервала обосновывается ровно так же, как и в предыдущем пункте.

Проведем численный эксперимент при $\varepsilon = 0.05$. Пусть выборка берется из нормального распределения $\mathsf{N}_{3,4}$. На рисунке 5 показаны границы доверительного интервала (красным), его середина (синим) и истинное значение параметра $\sigma^2 = 4$ (зеленым). Видно, что интервалы уменьшаются с ростом n, а почему?

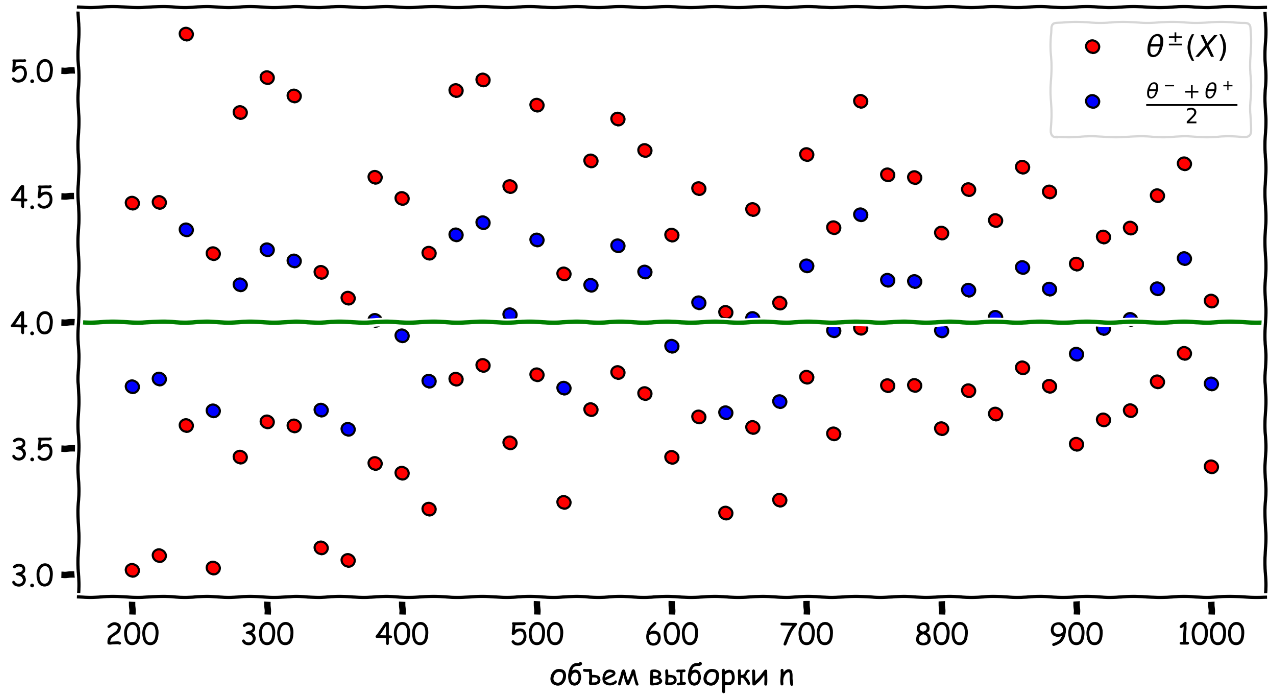

Рис. 5: Доверительный интервал для σ^2 при неизвестном a

Логично сравнить, сильно ли влияет на ширину доверительных интервалов информация о параметра a. На рисунке 6 приведены границы интервалов по выборке из нормального распределения $\mathsf{N}_{3,4}$. Синим – при неизвестном a, красным – при известном. Как видно, границы практически сливаются, особенно при больших объемах выборки.

1.1.5 Общий принцип построения доверительных интервалов

Рассмотрев несколько примеров, можно вычленить и общий способ построения доверительных интервалов. Итак, общий принцип построения точных доверительных интервалов можно сформулировать в следующем виде:
1. Составляется случайная величина $T(X, \theta)$, распределение которой не зависит от параметра θ, и которая обратима по θ при фиксированной выборке X;
2. Находятся квантили t_1 и t_2 распределения случайной величины T так, чтобы выполнялось равенство

$$\mathsf{P}_\theta\left(t_1 < T(X, \theta) < t_2\right) = 1 - \varepsilon;$$

3. Неравенство $t_1 < T(X, \theta) < t_2$ разрешается относительно θ, откуда и получается интересующий нас доверительный интервал.

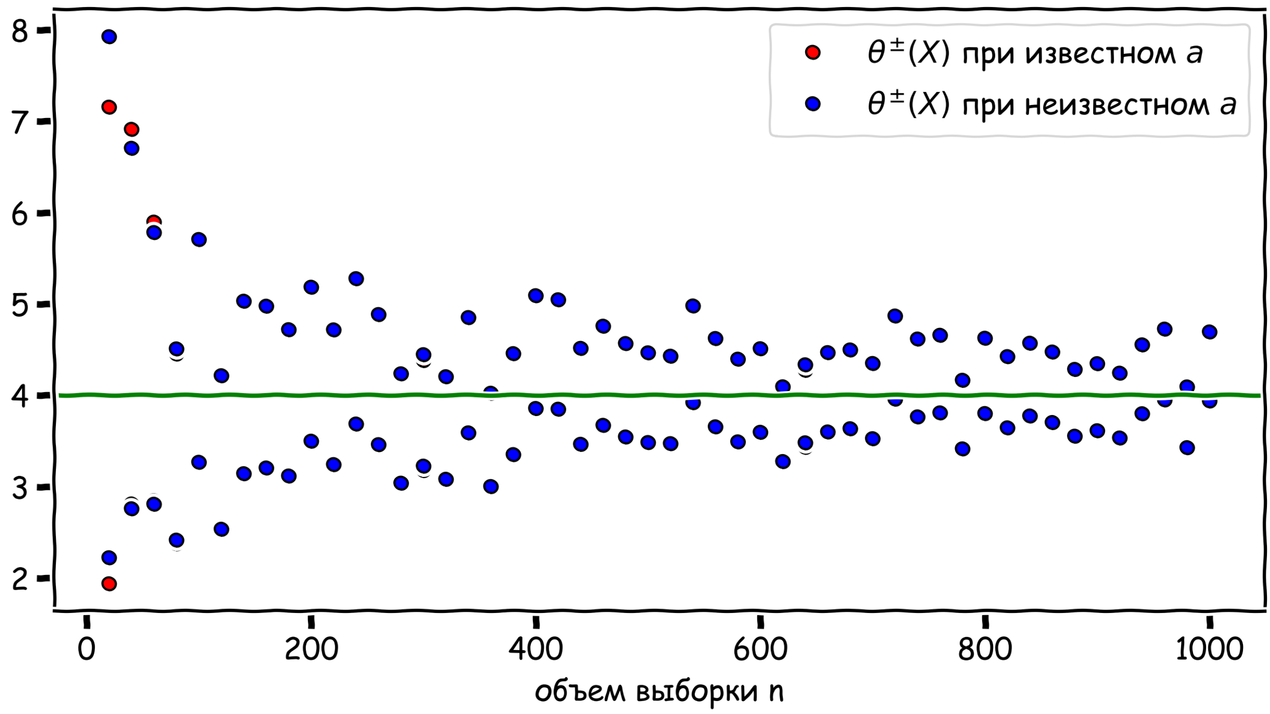

Рис. 6: Сравнение доверительных интервалов

В примерах, разобранных нами ранее, мы, по сути, поступали ровно-таки по схеме.

Несмотря на четкий алгоритм действий, сложность описанного подхода достаточно очевидна: совершенно непонятно, как найти «достойную» функцию $T(X, \theta)$, удовлетворяющую описанным условиям. В частности, как ее заставить не зависеть от θ? Именно поэтому, а еще из-за предельных теорем, чаще строят так называемые асимптотические доверительные интервалы, о которых и пойдет речь далее.

1.2 Асимптотические доверительные интервалы

Определение вводимого объекта напрашивается уже из названия. Раз асимптотический, значит, в пределе. Давайте сформулируем точнее, итак.

Определение 1.2.1 *Пусть* $0 < \varepsilon < 1$. *Интервал*

$$(\theta^-, \theta^+) = (\theta^-(X, \varepsilon), \theta^+(X, \varepsilon)),$$

где θ^-, θ^+ *– статистики, называется асимптотическим доверительным интервалом уровня доверия (или надежности)* $1 - \varepsilon$, *если для любого* $\theta \in \Theta$ *выполняется*

$$\liminf_{n \to +\infty} \mathsf{P}_\theta \left(\theta^- < \theta < \theta^+\right) \geq 1 - \varepsilon.$$

Замечание 1.2.1 *Как мы видим, определение отличается наличием предела :) На самом деле, правильнее писать, не просто* θ^- *и* θ^+, *а* θ^-_n, θ^+_n, *так* *как в определении речь идет не об одном интервале, а о последовательности интервалов.*

Замечание 1.2.2 *Для тех, кого пугают слова или обозначения «нижнего предела» могут считать, что это совершенно классический предел. Больших неприятностей при этом, как правило, не возникает.*

Замечание 1.2.3 *Знак неравенства $\geq$ в определении асимптотического доверительного интервала объясняется ровно также, как и в случае доверительного интервала: для дискретных распределений знак равенства часто оказывается бессмысленным. Кстати, немедленно объясните себе, а почему нельзя (или можно?) взять $\varepsilon = 0$? Каким в этом случае окажется асимптотический доверительный интервал? Плохо ли это?*

Ну и, наконец, каково же главное отличие? А главное отличие заключается в том, что теперь написанное равенство (точнее, неравенство) справедливо лишь в пределе. Так как в жизни бесконечности быть не может, то асимптотические доверительные интервалы имеет смысл рассматривать лишь при достаточно больших объемах выборки. Все это мы проиллюстрируем на примерах, описанных ниже.

1.2.1 Асимптотический доверительный интервал для Exp_θ

Построим асимптотический доверительный интервал уровня доверия $(1-\varepsilon)$ для показательного распределения Exp_θ с параметром $\theta > 0$. Напомним, что $\mathsf{E}_\theta X_1 = \frac{1}{\theta}$, $\mathsf{D}_\theta X_1 = \frac{1}{\theta^2}$. Вспомнив центральную предельную теорему, получим

$$Y_n = \frac{\sum\limits_{i=1}^{n} X_i - n\mathsf{E}_\theta X_1}{\sqrt{n\mathsf{D}_\theta X_1}} = \sqrt{n}\frac{\overline{X} - \frac{1}{\theta}}{\frac{1}{\theta}} = \sqrt{n}\left(\theta\overline{X} - 1\right) \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Значит, согласно определению слабой сходимости,

$$\mathsf{P}_\theta\left(-c < Y_n < c\right) = \mathsf{P}_\theta\left(-c < \sqrt{n}\left(\theta\overline{X} - 1\right) < c\right) \xrightarrow[n\to+\infty]{} \mathsf{P}_\theta\left(-c < Y < c\right) =$$

$$= \Phi_{0,1}(c) - \Phi_{0,1}(-c) = 2\Phi_{0,1}(c) - 1 = 1 - \varepsilon,$$

откуда $c = \tau_{1-\varepsilon/2}$ – квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения $\mathsf{N}_{0,1}$.

Осталось разрешить наше неравенство относительно θ, получим

$$-\tau_{1-\varepsilon/2} < \sqrt{n}\left(\theta\overline{X} - 1\right) < \tau_{1-\varepsilon/2} \Leftrightarrow -\frac{\tau_{1-\varepsilon/2}}{\sqrt{n}} < \theta\overline{X} - 1 < \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}},$$

откуда

$$\frac{1}{\overline{\overline{X}}} - \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\overline{\overline{X}}} < \theta < \frac{1}{\overline{\overline{X}}} + \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\overline{\overline{X}}}.$$

В итоге, асимптотический доверительный интервал уровня доверия $(1-\varepsilon)$ имеет вид:

$$\left(\theta^-,\ \theta^+\right) = \left(\frac{1}{\overline{\overline{X}}} - \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\overline{\overline{X}}},\ \frac{1}{\overline{\overline{X}}} + \frac{\tau_{1-\varepsilon/2}}{\sqrt{n}\overline{\overline{X}}}\right).$$

Видно, что с ростом n его длина со скоростью порядка $n^{-1/2}$ стремится к нулю. Давайте протестируем полученный интервал на примере. Пусть имеется выборка из показательного распределения Exp_θ с истинным параметром $\theta = 1.5$. Требуется построить асимптотический доверительный интервал уровня доверия 0.9 (то есть при $\varepsilon = 0.1$).

Для начала вычислим квантиль $\tau_{1-0.5/2} = \tau_{0.95}$. Пользуясь таблицами для нормального распределения, она равна $\tau_{0.95} \approx 1.65$. На рисунке 7 изображены границы доверительных интервалов (красным), их середины (синим) и истинное значение параметра (зеленым) при разных объемах выборки n. Видно, что в начале (при достаточно малых n) ошибок куда больше, чем при достаточно больших.

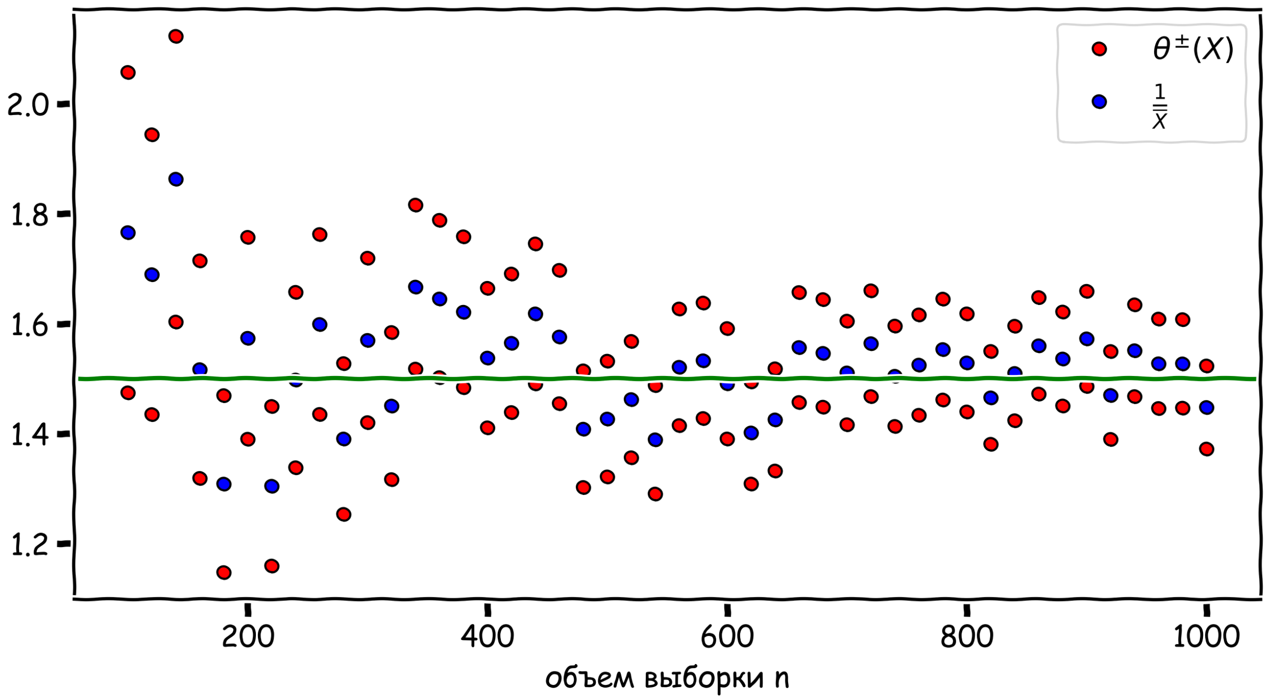

Рис. 7: Построение доверительных интервалов при разных n

1.2.2 Асимптотический доверительный интервал для B_θ

Построим асимптотический доверительный интервал для параметра θ распределения Бернулли B_θ. Вспомните, что мы это уже делали в самой первой лекции по статистике, и столкнулись с некоторыми трудностями, которые придется решать и сейчас. Так как $\mathsf{E}_\theta X_1 = \theta$, $\mathsf{D}_\theta X_1 = \theta(1-\theta)$, то, используя центральную предельную теорему, получим

$$Y_n = \frac{\sum\limits_{i=1}^{n} X_i - n\mathsf{E}_\theta X_1}{\sqrt{n\mathsf{D}_\theta X_1}} = \sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

Аналогично предыдущему примеру, отсюда получается, что, согласно определению слабой сходимости,

$$\mathsf{P}_\theta\left(-c < Y_n < c\right) = \mathsf{P}_\theta\left(-c < \sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\theta(1-\theta)}} < c\right) \xrightarrow[n\to+\infty]{} \mathsf{P}_\theta\left(-c < Y < c\right) =$$

$$= \Phi_{0,1}(c) - \Phi_{0,1}(-c) = 2\Phi_{0,1}(c) - 1 = 1-\varepsilon,$$

откуда $c = \tau_{1-\varepsilon/2}$ – квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения $\mathsf{N}_{0,1}$.

Рассматриваемый нами пример имеет существенное отличие от предыдущего. Дело в том, что разрешить неравенство

$$-\tau_{1-\varepsilon/2} < \sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\theta(1-\theta)}} < \tau_{1-\varepsilon/2}$$

относительно параметра θ – задача не из легких. Можно решить этот вопрос достаточно грубо, а именно можно рассмотреть эквивалентное неравенство

$$-\tau_{1-\varepsilon/2}\sqrt{\theta(1-\theta)} < \sqrt{n}\left(\overline{X}-\theta\right) < \tau_{1-\varepsilon/2}\sqrt{\theta(1-\theta)}$$

и, так как

$$\theta(1-\theta) = \frac{1}{4} - \left(\theta - \frac{1}{2}\right)^2 \leq \frac{1}{4}$$

заменить $\sqrt{\theta(1-\theta)}$ на 0.5, после чего решить неравенство и прийти к асимптотическому доверительному интервалу

$$\overline{X} - \frac{\tau_{1-\varepsilon/2}}{2\sqrt{n}} < \theta < \overline{X} + \frac{\tau_{1-\varepsilon/2}}{2\sqrt{n}}.$$

Как можно поступить иначе? Мы знаем, что выборочное среднее – это состоятельная оценка для математического ожидания, то есть $\overline{X} \xrightarrow[n\to+\infty]{\mathsf{P}} \mathsf{E}_\theta X_1 = \theta$. Заменим в знаменателе дроби

$$\sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\theta(1-\theta)}}$$

параметр θ на $\overline{X}$. Естественно возникает вопрос: не изменится ли сходимость? Но по свойствам слабой сходимости, если $\xi_n \xrightarrow[n\to+\infty]{\mathsf{P}} 1$ и $\eta_n \xrightarrow[n\to+\infty]{\mathsf{d}} \eta$, то

$$\xi_n \cdot \eta_n \xrightarrow[n\to+\infty]{\mathsf{d}} \eta,$$

а тогда

$$\sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\theta(1-\theta)}} = \sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\overline{X}(1-\overline{X})}} \cdot \sqrt{\frac{\overline{X}}{\theta} \cdot \frac{1-\overline{X}}{1-\theta}},$$

причем, в силу, как уже отмечалось, состоятельности выборочного среднего, последний корень по вероятности стремится к единице, а значит

$$Y_n = \sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\overline{X}(1-\overline{X})}} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}$$

и нам достаточно решить неравенство

$$-\tau_{1-\varepsilon/2} < \sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\overline{X}(1-\overline{X})}} < \tau_{1-\varepsilon/2},$$

откуда асимптотический доверительный интервал имеет вид

$$\left(\theta^-,\ \theta^+\right) = \left(\overline{X} - \tau_{1-\varepsilon/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}},\ \overline{X} + \tau_{1-\varepsilon/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right).$$

Видно, что длина асимптотического доверительного интервала стремится к нулю с ростом n со скоростью порядка $n^{-1/2}$. Посмотрим на численные расчеты при истинном значении параметра, равном 0.9. Будем строить асимптотический доверительный интервал уровня доверия 0.95. Как мы нашли в начале лекции, $\tau_{0.95} \approx 1.96$. На рисунке 8 видно, что почти всегда зеленая линия, как обычно отвечающая истинному параметру, попадает в построенный асимптотический доверительный интервал.

Сравним между собой два подхода: использованный ранее и использованный теперь (на выборках объема от 200, чтобы были меньше начальные выбросы, и корректнее масштаб). Центры интервалов совпадают, а отступы от центров отличаются. Сильно ли грубой была наша оценка? На рисунке 9 красным изображены концы доверительных интервалов, построенных по формулам

$$(\theta_1^-,\ \theta_1^+) = \left(\overline{X} - \tau_{1-\varepsilon/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}},\ \overline{X} + \tau_{1-\varepsilon/2}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right),$$

а синим – построенных по формулам

$$(\theta_2^-,\ \theta_2^+) = \left(\overline{X} - \frac{\tau_{1-\varepsilon/2}}{2\sqrt{n}},\ \overline{X} + \frac{\tau_{1-\varepsilon/2}}{2\sqrt{n}}\right).$$

Красные оказываются уже: хорошо ли это? Посмотрите, а нет ли при таком подходе дополнительных ошибок? Интересно, что если истинное значение

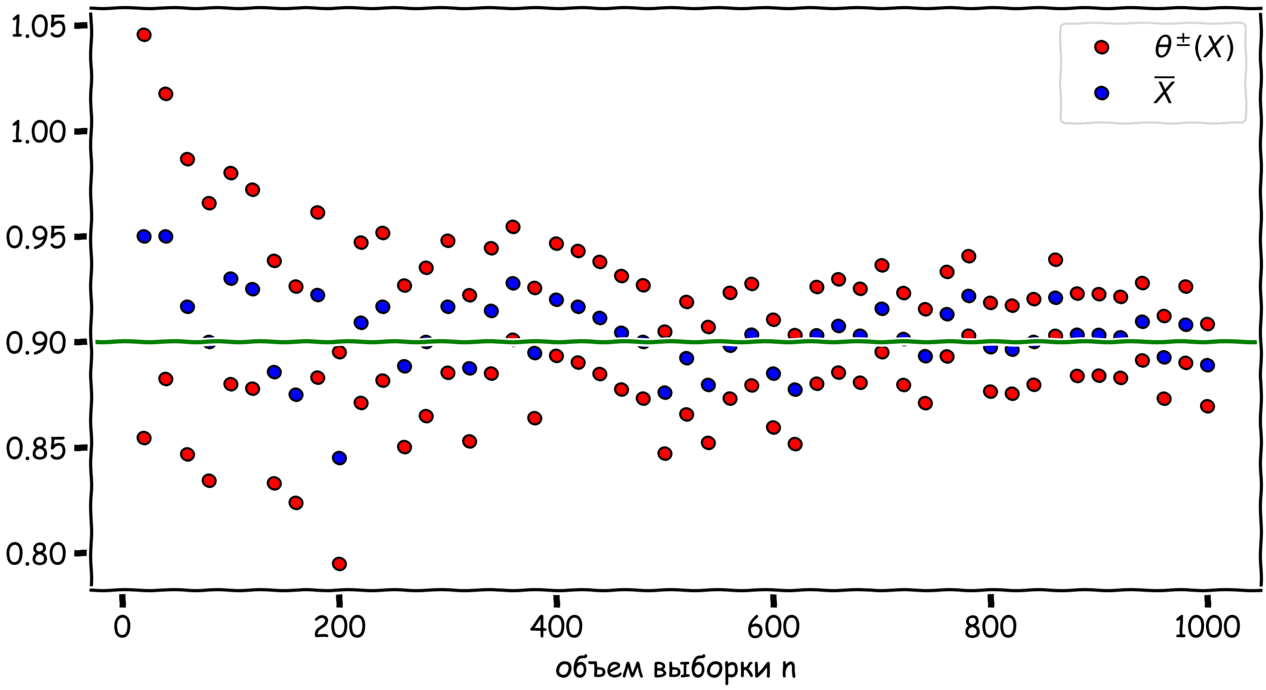

Рис. 8: Построение доверительных интервалов при разных n

близко к 0.5, то разницы почти нет. Например, на рисунке 10 показаны совсем слившиеся (на самом деле в самом начале видно, что расстояние между красными точками уже, они вылезают из-под синих) интервалы (истинное значение $\theta = 0.4$). Как думаете, почему так произошло?

А вот почему. Если вспомнить, то при оценке корня $\sqrt{p(1-p)}$ сверху, мы воспользовались тем, что $p(1-p) \leq 0.25$. Это число есть не что иное, как

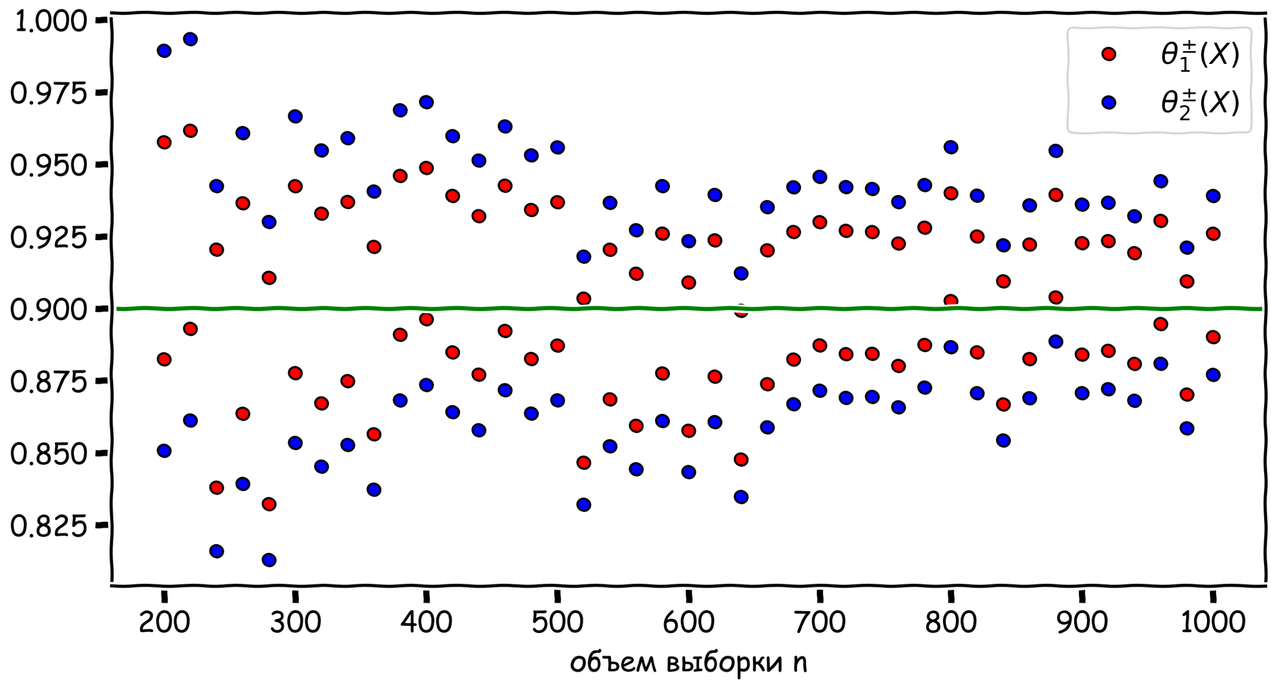

Рис. 9: Построение доверительных интервалов при разных n

значение рассматриваемой параболы в ее вершине, которая расположена в точке $p = 0.5$. Отсюда и из вида функции понятно, что чем ближе истинное значение к половине, тем точнее построенный нами ранее интервал.

Рассмотрев достаточно много частных примеров построения асимптотических доверительных интервалов, можно сформулировать, как мы это делали и в доверительных интервалах, общий подход, итак.

1.2.3 Общий принцип построения асимптотических доверительных интервалов

Общий принцип построения асимптотических доверительных интервалов может быть описан следующим алгоритмом:
1. Составляется случайная величина $T(X, \theta)$, которая слабо сходится к распределению, не зависящему от параметра θ, и которая обратима по θ при фиксированной выборке X;
2. Находятся квантили t_1 и t_2 предельного распределения случайной величины T так, чтобы выполнялось равенство

$$\mathsf{P}_\theta\left(t_1 < T(X, \theta) < t_2\right) = 1 - \varepsilon;$$

3. Неравенство $t_1 < T(X, \theta) < t_2$ разрешается относительно θ, откуда и получается интересующий нас асимптотический доверительный интервал.

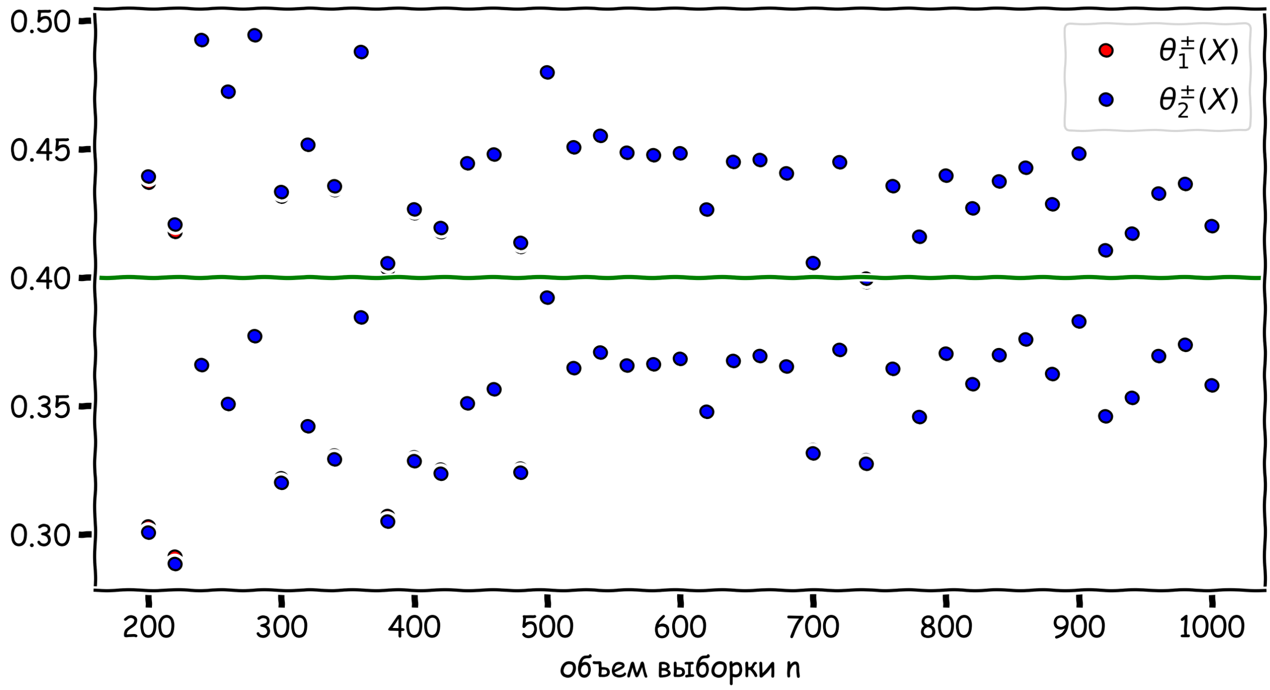

Рис. 10: Построение доверительных интервалов при разных n

Как мы видели в примерах с семействами распределений Exp_θ, B_θ, предельное распределение для рассматриваемой статистики $T(X,\theta)$ было $\mathsf{N}_{0,1}$ и выдавалось центральной предельной теоремой. Это – довольно общий прием, до которого внимательный слушатель наверняка уже догадался, и который мы формально осветим в следующем, завершающем пункте данной лекции.

1.2.4 Асимптотические интервалы в случае АНО

Оказывается, построение асимптотических доверительных интервалов тесно связано с понятием асимптотически нормальной оценки.

Определение 1.2.2 *Статистика $\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$ называется асимптотически нормальной оценкой (АНО) параметра θ с коэффициентом $\sigma^2(\theta)$, если имеет место слабая сходимость*

$$Y_n = \sqrt{n}\frac{\widehat{\theta} - \theta}{\sigma(\theta)} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

С примерами асимптотически нормальных оценок мы встречались: это и выборочное среднее, и выборочные моменты высших порядков, и выборочная дисперсия. Само определение асимптотической нормальности, вид предельного распределения и схема, описанная нами в предыдущем пункте, подсказывает, что для асимптотически нормальных оценок доверительные интервалы можно строить по готовым формулам, не так ли? И правда, справедлива следующая теорема

Теорема 1.2.1 *Пусть $\widehat{\theta}$ – асимптотически нормальная оценка для параметра θ с коэффициентом $\sigma^2(\theta)$, причем $\sigma(\theta) \in C(\Theta)$. Тогда интервал*

$$\left(\theta^-,\ \theta^+\right) = \left(\widehat{\theta} - \tau_{1-\varepsilon/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}},\ \widehat{\theta} + \tau_{1-\varepsilon/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}}\right),$$

где $\tau_{1-\varepsilon/2}$ – квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения, является асимптотическим доверительным интервалом для параметра θ уровня доверия $1-\varepsilon$.

Доказательство. Согласно определению АНО, имеем

$$Y_n = \sqrt{n}\frac{\widehat{\theta} - \theta}{\sigma(\theta)} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1}.$$

В силу состоятельности АНО, выполняется $\widehat{\theta} \xrightarrow[n\to+\infty]{\mathsf{P}} \theta$, а в силу непрерывности $\sigma^2(\theta)$ и свойств сходимости по вероятности получаем, что $\sigma(\widehat{\theta}) \xrightarrow[n\to+\infty]{\mathsf{P}} \sigma(\theta)$. Тогда в качестве статистики $T(X,\theta)$ резонно взять

$$T(X,\theta) = Y_n = \sqrt{n}\frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})} = \frac{\sigma(\theta)}{\sigma(\widehat{\theta})} \cdot \sqrt{n}\frac{\widehat{\theta} - \theta}{\sigma(\theta)} \xrightarrow[n\to+\infty]{\mathsf{d}} Y \sim \mathsf{N}_{0,1},$$

так как первый сомножитель по вероятности стремится к 1.

$$\mathsf{P}_\theta\left(-c < T(X,\theta) < c\right) \xrightarrow[n\to+\infty]{\mathsf{d}} \mathsf{P}_\theta\left(-c < Y < c\right) = \Phi_{0,1}(c) - \Phi_{0,1}(-c) = 2\Phi_{0,1}(c) - 1.$$

Приравняв последнее выражение к $1-\varepsilon$, получим

$$2\Phi_{0,1}(c) - 1 = 1 - \varepsilon \Leftrightarrow \Phi_{0,1}(c) = 1 - \frac{\varepsilon}{2},$$

а значит в качестве c удобно брать квантиль $\tau_{1-\varepsilon/2}$ уровня $1-\varepsilon/2$ стандартного нормального распределения $\mathsf{N}_{0,1}$. Значит, осталось разрешить неравенство

$$-\tau_{1-\varepsilon/2} < T(X,\theta) < \tau_{1-\varepsilon/2} \Leftrightarrow \widehat{\theta} - \tau_{1-\varepsilon/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}} < \theta < \widehat{\theta} + \tau_{1-\varepsilon/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}}.$$

В итоге получаем асимптотический доверительный интервал

$$\left(\theta^-,\ \theta^+\right) = \left(\widehat{\theta} - \tau_{1-\varepsilon/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}},\ \widehat{\theta} + \tau_{1-\varepsilon/2}\frac{\sigma(\widehat{\theta})}{\sqrt{n}}\right).$$

□

Легко понять, что асимптотические доверительные интервалы, построенные нами в этой лекции, подходят под только что доказанную теорему, и могли быть получены как элементарное следствие.

На самом деле, асимптотический доверительный интервал для математического ожидания можно строить и без модели, в предположении конечности второго момента. Мы знаем, что в этом случае

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

Так как состоятельными оценками дисперсии являются как S^2, так и S_0^2, то доверительным интервалом для математического ожидания уровня доверия $1-\varepsilon$ является как

$$\left(\theta^-,\ \theta^+\right) = \left(\overline{X} - \tau_{1-\varepsilon/2}\frac{S}{\sqrt{n}},\ \overline{X} + \tau_{1-\varepsilon/2}\frac{S}{\sqrt{n}}\right),$$

так и

$$\left(\theta^-,\ \theta^+\right) = \left(\overline{X} - \tau_{1-\varepsilon/2}\frac{S_0}{\sqrt{n}},\ \overline{X} + \tau_{1-\varepsilon/2}\frac{S_0}{\sqrt{n}}\right).$$

1.2.5 Некоторое резюме

Давайте резюмируем. В этой лекции мы научились строить как доверительные, так и асимптотические доверительные интервалы для параметров различных распределений. Доверительный интервал накрывает истинный параметр с заданной вероятностью. Можно утверждать, что в некотором смысле он даже оценивает абсолютную погрешность (конечно, с заданным уровнем доверия) конкретного значения над истинным значением параметра. Кроме того, он часто показывает и погрешность, с которой заданная точечная оценка (особенно, когда она является серединой интервала) приближает истинное значение. Перед нами осталась одна задача математической статистики, которая все еще не освещена – задача проверки гипотез. Ее мы и рассмотрим в последней лекции