Содержание

1 Задача интервального оценивания

Здравствуйте, уважаемые слушатели. В предыдущей лекции мы по выборке научились строить оценки различных характеристик генеральной совокупности, а также применять их для оценивания параметров семейств некоторых известных распределений. Выяснили, какими качествами должны обладать «хорошие» оценки: это обязательное качество состоятельности и желательное – несмещенности, или хотя бы асимптотической несмещенности. Еще мы узнали как получать выборочный аналог плотности распределения – гистограммы, научились работать с многомерными выборками и находить выборочные ковариацию и корреляцию.

Приобретенные знания и навыки уже являются солидным набором инструментов для всестороннего анализа данных и позволяют строить по выборке приближения распределений истинных случайных величин, оценивать вероятности интересующих событий, выявлять закономерности, строить прогнозы и многое другое.

В то же время разработанный нами аппарат имеет достаточно большой недостаток: нам совершенно неизвестно, насколько та или иная числовая оценка, полученная по выборке, близка к истинной характеристике генеральной совокупности. Так, по одной выборке мы можем получить выборочное среднее, равное, например, 3, а по другой (из того же распределения!) – 5. И какая из полученных оценок математического ожидания генеральной совокупности лучше? Пока на этот вопрос у нас ответа нет. И вот еще загвоздка: снова поменяв выборку, скорее всего мы опять получим другое значение выборочного среднего. Как же поступать? В этой лекции мы и займемся построением так называемых доверительных интервалов, помогающих в некотором смысле «оценить погрешность» между истинным значением параметра и его выборочным аналогом.

1.1 Центральная предельная теорема

Определение границ интервалов, в которых с заданной наперед вероятностью окажется некоторая характеристика случайной величины, помогает решить достаточно широкий спектр прикладных задач: сколько лекарств завозить в аптеку при известном спросе, чтобы не возникло дефицита, какой спред (разницу между покупкой и продажей валюты) установить банку на выходные, когда не совершается торгов на валютной бирже, сколько произвести телефонов, чтобы удовлетворить спрос на старте продаж, – на все эти вопросы помогает ответить интервальное оценивание. Давайте рассмотрим конкретный пример.

Пример 1.1.1 *Каждый год муниципальное образование проводит новогодний утренник, на котором, помимо спектакля, детишки-посетители полу-*

чают в качестве новогоднего подарка набор конфет от Деда Мороза. Праздник организуется из расчета на 1000 *посетителей, приглашения печатаются и раздаются заранее. Конфеты для подарков закупаются осенью, когда нет предновогоднего ажиотажа, и цена на конфеты ниже. В прошлом году из* 1000 *человек, получивших приглашение, на праздник явилось только* 753 *человека, и* 247 *подарков оказались невостребованными. Поэтому организаторы решили оптимизировать затраты, учитывая статистику, собранную за предыдущий год. Но, если закупить ровно* 753 *подарка (то есть столько же, сколько в прошлом году), а в этом году придет больше людей, то недостающие наборы придется закупать по завышенной цене и в спешке оформлять, пока идет спектакль. Поэтому резонно задуматься: а в каком диапазоне (ну хотя бы с некоторой высокой вероятностью) окажется число пришедших на праздник? Тогда можно будет оценить количество подарков, которое желательно закупить, чтобы, с одной стороны, не было дефицита, а с другой стороны – не осталось много лишних и никому ненужных подарков. На этот вопрос мы сможем ответить, если будем знать вероятность, что человек придет на праздник, если получил пригласительный билет.*

Для решения поставленной задачи, нам снова нужно ненадолго погрузиться в аппарат теории вероятностей. В прошлой лекции мы познакомились с понятием сходимости по вероятности, с помощью которого формулировали закон больших чисел. Однако часто нам интересно не то, к какой случайной величине сойдется рассматриваемая последовательность случайных величин, а лишь ее (предельной случайной величины) функция распределения. Все потому, что с помощью функции распределения мы и так сможем узнать все, что хотим про предельную случайную величину: вероятности попадания в те или иные промежутки, различные характеристики и проч. Поэтому разумно принять следующее определение.

Определение 1.1.1 (Сходимость по распределению) *Говорят, что последовательность случайных величин ξ_n сходится к случайной величине ξ по распределению (или слабо), если*

$$\lim_{n\to\infty} F_{\xi_n}(x) = F_\xi(x)$$

во всех точках x, в которых предельная функция $F_\xi(x)$ непрерывна.

Сходимость по распределению часто обозначают

$$\xi_n \xrightarrow[n\to+\infty]{\mathsf{d}} \xi,$$

где индекс d несет в себе сокращение от слова $\mathsf{distribution}$.

1.2 Немного о стандартизации

Перед тем, как сформулировать основной результат теории вероятностей, к которому мы ведем, – центральную предельную теорему, обсудим полезную в анализе данных процедуру стандартизации, которая, как мы надеемся, прольет на ЦПТ некоторый свет.

Определение 1.2.1 *Рассмотрим случайную величину ξ и предположим, что $\mathsf{E}\xi = a$, $\mathsf{D}\xi = \sigma^2$. Тогда случайная величина*

$$\eta = \frac{\xi - a}{\sigma}$$

называется стандартизированной случайной величиной.

Почему стандартизированной? Да потому, что согласно свойствам математического ожидания,

$$\mathsf{E}\eta = \frac{1}{\sigma}\left(\mathsf{E}\xi - a\right) = \frac{1}{\sigma}\left(a - a\right) = 0,$$

а согласно свойствам дисперсии,

$$\mathsf{D}\eta = \frac{1}{\sigma^2}\mathsf{D}\xi = \frac{\sigma^2}{\sigma^2} = 1.$$

Итак, у новой случайной величины η математическое ожидание равно нулю, а дисперсия – единице. На что похоже? На случайную величину, имеющую стандартное нормальное распределение: у нее, как вы, надеемся, помните, математическое ожидание тоже равно 0, а дисперсия – 1.

Смысл сделанного преобразования (стандартизации), наверное, угадывается – это, в некотором смысле, нормировка. Новая случайная величина имеет нулевое среднее и единичную дисперсию, ее значения безразмерны. В первичной обработке данных вы уже сталкивались с так называемой линейной нормировкой; приведенный же выше способ – хорошая ей альтернатива. В большинстве методов анализа данных нормировка оказывается чрезвычайно важна: это и метод k ближайших соседей, и k-средних, и логистическая регрессия, и многое-многое другое. Какую же нормировку выбирать часто зависит от задачи, этот вопрос каждый раз требует отдельного обсуждения.

Давайте теперь проясним следующий вопрос: а какое отношение все эти разговоры имеют к выборке и построению интервалов? Рассмотрим выборочное среднее (тоже случайную величину) $\overline{X}$, построенное по выборке $X = (X_1, X_2, ..., X_n)$ из генеральной совокупности ξ с математическим ожиданием $\mathsf{E}\xi = a$ и дисперсией $\mathsf{D}\xi = \sigma^2$. Тогда, как мы знаем,

$$\mathsf{E}\overline{X} = \mathsf{E}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}X_i = \frac{na}{n} = a,$$

а дисперсия

$$\mathsf{D}\overline{X} = \mathsf{D}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{D}X_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Выполнив стандартизацию, получим последовательность случайных величин

$$Y_n = \frac{\overline{X} - \mathsf{E}\overline{X}}{\sqrt{\mathsf{D}\overline{X}}} = \sqrt{n}\frac{\overline{X} - a}{\sigma},$$

у которых математическое ожидание равно нулю, а дисперсия – единице. Оказывается, что с ростом n эта случайная величина (по распределению) приближается к нормальной. Это наблюдение и носит название центральной предельной теоремы.

1.3 Центральная предельная теорема

Итак, давайте сформулируем один из самых замечательных и важных результатов теории вероятностей, на котором основана добрая часть методов и приемов математической статистики.

Теорема 1.3.1 (Центральная предельная теорема) *Пусть $X_1, X_2, \ldots, X_n$ – независимые, одинаково распределенные случайные величины, математическое ожидание которых равно a, а дисперсия σ^2 отлична от нуля. Тогда имеет место слабая сходимость*

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

Итак, как мы сказали только что, каждый член последовательности Y_n и так уже имеет математическое ожидание, равное 0, и дисперсию, равную 1. Но при неограниченном увеличении n можно сказать куда больше: распределение рассматриваемой нами последовательности случайных величин сходится к нормальному!

Используя определение сходимости по распределению, переформулируем эту теорему в том виде, в котором мы ее и будем дальше применять:

$$\mathsf{P}\left(A \le \sqrt{n}\frac{\overline{X} - a}{\sigma} \le B\right) \xrightarrow[n\to+\infty]{} \Phi_{0,1}(B) - \Phi_{0,1}(A),$$

так как функция распределения $\Phi_{0,1}$ случайной величины Y, имеющей стандартное нормальное распределения $\mathsf{N}_{0,1}$, всюду непрерывна. Несмотря на то, что аналитическое выражение для $\Phi_{0,1}(x)$ нам и не понадобится, мы его приведем, так как оно очень часто возникает в различных приложениях:

$$\Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}}dt.$$

Во многих задачах требуется найти не значение функции $\Phi_{0,1}$ в заданной точке, а выполнить обратную задачу, то есть найти аргумент по известному значению функции. Конечно, таблицу значений функции $\Phi_{0,1}$ можно найти в интернете, но нужно быть осторожным. Часто, если нет понимания о чем идет речь, можно нарваться на таблицу для функции ошибок $\mathsf{erf}(x)$. Так что лучше использовать проверенные встроенные функции. Большинство пакетов по обработке данных поддерживают построение таблиц значений функции $\Phi_{0,1}$. Рассмотрим построение такой таблицы в Excel.

Пример 1.3.1 *Рассмотрим построение таблицы значений функции* $\Phi_{0,1}$ *средствами* Excel. *Для этого в ячейки* $(A2 : A31)$ *поместим числа от* 0 *с шагом в* 0.1. *В ячейки* $(B1 : K1)$ *запишем числа от* 0 *с шагом* 0.01. *Используя функцию НОРМСТРАСП(), получим таблицу значений функции* $\Phi_{0,1}$ *(рисунок 1).*

G18 =НОРМСТРАСП($A18+G$1)

	A	B	C	D	E	F	G	H	I	J	K
1	Ф	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
2	0	0,5	0,50398936	0,50797831	0,51196647	0,51595344	0,51993881	0,52392218	0,52790317	0,53188137	0,53585639
3	0,1	0,53982784	0,54379531	0,54775843	0,55171679	0,55567	0,55961769	0,56355946	0,56749493	0,57142372	0,57534543
4	0,2	0,57925971	0,58316616	0,58706442	0,59095412	0,59483487	0,59870633	0,60256811	0,60641987	0,61026125	0,61409188
5	0,3	0,61791142	0,62171952	0,62551583	0,62930002	0,63307174	0,63683065	0,64057643	0,64430875	0,64802729	0,65173173
6	0,4	0,65542174	0,65909703	0,66275727	0,66640218	0,67003145	0,67364478	0,67724189	0,68082249	0,6843863	0,68793305
7	0,5	0,69146246	0,69497427	0,69846821	0,70194403	0,70540148	0,70884031	0,71226028	0,71566115	0,71904269	0,72240468
8	0,6	0,72574688	0,7290691	0,73237111	0,73565271	0,7389137	0,74215389	0,74537309	0,7485711	0,75174777	0,75490291
9	0,7	0,75803635	0,76114793	0,7642375	0,76730491	0,77035	0,77337265	0,77637271	0,77935005	0,78230456	0,78523612
10	0,8	0,7881446	0,79102991	0,79389195	0,79673061	0,79954581	0,80233746	0,80510548	0,8078498	0,81057035	0,81326706
11	0,9	0,81593987	0,81858875	0,82121362	0,82381446	0,82639122	0,82894387	0,83147239	0,83397675	0,83645694	0,83891294
12	1	0,84134475	0,84375235	0,84613577	0,848495	0,85083005	0,85314094	0,8554277	0,85769035	0,85992891	0,86214343
13	1,1	0,86433394	0,86650049	0,86864312	0,87076189	0,87285685	0,87492806	0,8769756	0,87899952	0,88099989	0,8829768
14	1,2	0,88493033	0,88686055	0,88876756	0,89065145	0,8925123	0,89435023	0,89616532	0,89795768	0,89972743	0,90147467
15	1,3	0,90319952	0,90490208	0,90658249	0,90824086	0,90987733	0,91149201	0,91308504	0,91465655	0,91620668	0,91773556
16	1,4	0,91924334	0,92073016	0,92219616	0,92364149	0,9250663	0,92647074	0,92785496	0,92921912	0,93056338	0,93188788
17	1,5	0,9331928	0,93447829	0,93574451	0,93699164	0,93821982	0,93942924	0,94062006	0,94179244	0,94294657	0,9440826
18	1,6	0,94520071	0,94630107	0,94738386	0,94844925	0,94949742	0,95052853	0,95154277	0,95254032	0,95352134	0,95448602
19	1,7	0,95543454	0,95636706	0,95728378	0,95818486	0,95907049	0,95994084	0,9607961	0,96163643	0,96246202	0,96327304
20	1,8	0,96406968	0,96485211	0,9656205	0,96637503	0,96711588	0,96784323	0,96855724	0,96925809	0,96994596	0,97062102
21	1,9	0,97128344	0,97193339	0,97257105	0,97319658	0,97381016	0,97441194	0,9750021	0,97558081	0,97614824	0,97670453
22	2	0,97724987	0,97778441	0,97830831	0,97882173	0,97932484	0,97981778	0,98030073	0,98077383	0,98123723	0,9816911
23	2,1	0,98213558	0,98257082	0,98299698	0,98341419	0,98382262	0,98422239	0,98461367	0,98499658	0,98537127	0,98573788
24	2,2	0,98609655	0,98644742	0,98679062	0,98712628	0,98745454	0,98777553	0,98808937	0,98839621	0,98869616	0,98898934
25	2,3	0,98927589	0,98955592	0,98982956	0,99009692	0,99035813	0,99061329	0,99086253	0,99110596	0,99134368	0,99157581
26	2,4	0,99180246	0,99202374	0,99223975	0,99245059	0,99265637	0,99285719	0,99305315	0,99324435	0,99343088	0,99361285
27	2,5	0,99379033	0,99396344	0,99413226	0,99429687	0,99445738	0,99461385	0,99476639	0,99491507	0,99505998	0,9952012
28	2,6	0,99533881	0,99547289	0,99560351	0,99573076	0,9958547	0,99597541	0,99609297	0,99620744	0,99631889	0,9964274
29	2,7	0,99653303	0,99663584	0,9967359	0,99683328	0,99692804	0,99702024	0,99710993	0,99719719	0,99728206	0,9973646
30	2,8	0,99744487	0,99752293	0,99759882	0,9976726	0,99774432	0,99781404	0,99788179	0,99794764	0,99801162	0,99807379
31	2,9	0,99813419	0,99819286	0,99824984	0,99830519	0,99835894	0,99841113	0,9984618	0,998511	0,99855876	0,99860511
32											

Рис. 1: Таблица значений функции $\Phi_{0,1}$

Как пользоваться этой таблицей? Давайте, например, найдем аргумент, при котором функция $\Phi_{0,1}$ принимает значение 0.95. Для этого в таблице

находим значение максимально близкое к 0.95 и получаем соответствующее значение аргумента:

$$1.6 + 0.05 = 1.65.$$

Удобно ввести следующее определение, которое вам уже встречалось в курсе обработки и анализа данных.

Определение 1.3.1 *Пусть фиксировано число $\alpha \in (0, 1)$ и функция распределения $F_\xi(x)$ строго возрастает . Квантилью уровня α распределения случайной величины ξ называется такое число x_α, что*

$$F_\xi(x_\alpha) = \alpha$$

В нашем случае 1.65 – это и есть примерное значение квантили уровня 0.95 для случайной величины, имеющей стандартное нормальное распределение. В виду частоты использования последнего, его квантили обозначают буквой τ, итак

$$\tau_{0.95} \approx 1.65.$$

Пример 1.3.2 *Вернемся к нашему примеру и повторим условие задачи . Согласно данным прошлого года, из 1000 приглашенных на праздник пришло лишь 753 человека. Опираясь на собранную статистику и предполагая, что она подчиняется какой-то вероятностной закономерности ξ, мы хотим построить интервал, в котором с некоторой, вообще говоря достаточно большой, вероятностью (например, 0.9) окажется истинное количество пришедших на праздник людей, чтобы сформировать необходимое число подарков. А для этого оценим вероятность, того, что человек придет на праздник, если имеет билет.*

Воспользуемся центральной предельной теоремой в предположении, что существует и отлична от нуля дисперсия случайной величины ξ, равная σ^2 , а математическое ожидание ξ равно a. Согласно ЦПТ и свойствам функции $\Phi_{0,1}$,

$$\mathsf{P}\left(-c \le \sqrt{n}\frac{\overline{X} - a}{\sigma} \le c\right) \xrightarrow[n\to+\infty]{} \Phi_{0,1}(c) - \Phi_{0,1}(-c) = 2\Phi_{0,1}(c) - 1.$$

Мы хотим, чтобы последняя вероятность была равна 0.9, тогда

$$2\Phi_{0,1}(c) - 1 = 0.9 \Leftrightarrow \Phi_{0,1}(c) = 0.95$$

и $c = \tau_{0.95}$ – квантиль уровня 0.95 стандартного нормального распределения.

Осталось разрешить неравенство под знаком вероятности, в котором в нашем случае $c = \tau_{0.95}$, относительно a. Получаем

$$-c \leq \sqrt{n}\frac{\overline{X} - a}{\sigma} \leq c \Leftrightarrow -\tau_{0.95} \leq \sqrt{n}\frac{\overline{X} - a}{\sigma} \leq \tau_{0.95}$$

и

$$\overline{X} - \tau_{0.95}\frac{\sigma}{\sqrt{n}} < a < \overline{X} + \tau_{0.95}\frac{\sigma}{\sqrt{n}}.$$

На этом моменте стоит остановиться и задуматься еще раз: а что мы получили? Мы получили интервал, в котором, с вероятностью 0.9*, находится истинное математическое ожидание случайной величины ξ, при $n \to +\infty$. Математическое ожидание случайной величины ξ в условиях сформулированной задачи можно интерпретировать, как вероятность прийти конкретному человеку на праздник (как в схеме Бернулли).*

Отлично, а при чем тут статистика? Как нам применить собранные нами данные? Собранные нами данные состоят из 753 *единиц и* 247 *нулей , а значит $\overline{X} = 0.753$. А что с σ, оно же неизвестно? Для его оценки можно использовать S_0, ведь S_0, как мы знаем, состоятельная оценка дисперсии, а значит при больших n*

$$\sigma \approx S_0 \approx 0.431.$$

Используя теперь то, что $\tau_{0.95} \approx 1.65$, получаем, что

$$0.753 - 1.65\frac{0.431}{\sqrt{1000}} \leq a \leq 0.753 + 1.65\frac{0.431}{\sqrt{1000}} \Leftrightarrow 0.730 \leq a \leq 0.776.$$

Иными словами, вероятность, что человек придет на праздник, имея билет, лежит в интервале от 0.730 *до* 0.776*. Напомним, что пригласительные билеты получили* 1000 *человек. Значит, в среднем стоит ожидать не менее* 730 *и не более* 776 *человек, то есть имеет смысл приобрести* 776 *подарков.*

Отметим несколько моментов. Во-первых, никто не гарантирует, даже с вероятностью 0.9, что найденный интервал содержит истинное число пришедших на праздник, так как наш интервал построен при конкретном (хоть и большом) n. Но то, что n достаточно большое дает нам надежду, что интервал близок к теоретическому. Во-вторых, мы использовали не точное значение σ, а приближенное, но, снова, на выборке большого объема. Все это делает наши предположения и оценки оправданными и пригодными для использования и прогнозирования.

Какую же выгоду получит муниципалитет? Во-первых, он не купит огромное количество ненужных подарков, тем самым сэкономит бюджет.

Кроме того, даже если придется покупать дополнительные подарки (что маловероятно), их количество будет не большим и суммарные траты будут меньше, чем в случае покупки ровно 1000 подарков с самого начала.

Как можно помочь муниципалитету еще? Можно было бы взять вероятность, равную, не 0.9, а, скажем, 0.95. Интервал бы, конечно, стал шире, но и вероятность «угадать» – куда больше. Проделайте это сами и проверьте себя в опросах!

Отметим также полезное для практики замечание.

Замечание 1.3.1 *Значения квантилей стандартного нормального распределения в* Excel *можно вычислить как описанным ранее способом при помощи таблицы, так и при помощи функции НОРМ.СТ.ОБР(), где в качестве аргумента указывается уровень необходимой квантили. Например,*

$$НОРМ.СТ.ОБР(0.95) = 1.6448\ldots$$

1.4 Асимптотические доверительные интервалы

Оказывается, тот интервал, что мы построили в разобранном нами примере с подарками к новому году, носит специальное название – это так называемый асимптотический доверительный интервал для математического ожидания генеральной совокупности при неизвестной дисперсии. А что, строго говоря, такое «асимптотический доверительный интервал», и какими они, эти интервалы, бывают?

У нас все готово для того , чтобы формально определить понятие асимптотического доверительного интервала. Пусть $X_1, X_2, ..., X_n$ – выборка из некоторого распределения, которое каким-то образом зависит от неизвестного параметра θ из некоторого множества Θ.

Определение 1.4.1 *Пусть* $0 < \varepsilon < 1$. *Интервал*

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

где θ^- *и* θ^+ *– это функции как от выборки (то есть оценки), так и от* ε, *называется асимптотическим доверительным интервалом уровня доверия (или надежности)* $1 - \varepsilon$, *если для любого* $\theta \in \Theta$ *выполняется*

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

Когда мы говорим о надежности $1 - \varepsilon$, мы подразумеваем что как минимум с такой вероятностью оцениваемая величина будет находиться внутри этого интервала – на меньшее мы не согласны (это и подчеркивает знак $\geq$ в определении)! При этом понятно, что чем меньше ε мы возьмем, тем больше

мы будем уверены, что оцениваемая величина действительно попадет внутрь этого интервала, а значит, скорее всего, длина интервала вырастет . А что, если взять в качестве ε число 0? На деле мы просто получим интервал, содержащий все возможные значения параметра θ, то есть интервал, содержащий все множество Θ. Как вы понимаете, толку от такого интервала мало: мы не получили никакой новой информации, а, скорее, внесли неопределенность.

Замечание 1.4.1 *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.

Найдем асимптотические доверительные интервалы для оценки параметров ранее рассмотренных распределений.

1.4.1 Асимптотический доверительный интервал для Exp_θ

Построим асимптотический доверительный интервал уровня доверия $(1-\varepsilon)$ для показательного распределения Exp_θ с параметром $\theta > 0$. Пусть имеется выборка $X = (X_1, X_2, ..., X_n)$ из этого распределения. Напомним, что $\mathsf{E}_\theta X_1 = a = \frac{1}{\theta}$, $\mathsf{D}_\theta X_1 = \frac{1}{\theta^2}$, а значит $\sigma = \frac{1}{\theta}$. Вспомнив центральную предельную теорему, получим

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

Значит, согласно определению слабой сходимости,

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

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

Последнее уравнение переписывается в виде

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

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

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

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

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

Видно, что с ростом n его длина со скоростью порядка $n^{-1/2}$ стремится к нулю. Давайте протестируем полученный интервал на примере.

Пример 1.4.1 *Пусть имеется выборка из показательного распределения* Exp_θ *с истинным параметром* $\theta = \frac{1}{3}$. *Требуется построить асимптотический доверительный интервал уровня доверия* 0.95 *(то есть при* $\varepsilon = 0.05$*).*

Мы уже знаем как найти квантиль требуемого уровня. При $\varepsilon = 0.05$, *это будет* $\tau_{1-0.05/2} = \tau_{0.975} \approx 1.96$. *На рисунке 2 изображены верхняя граница доверительного интервала (синим), нижняя (красным) и истинное значение параметра (зеленым) при разных объемах выборки* n. *Видно, что в начале (при достаточно малых* n*) ошибок куда больше, чем при больших.*

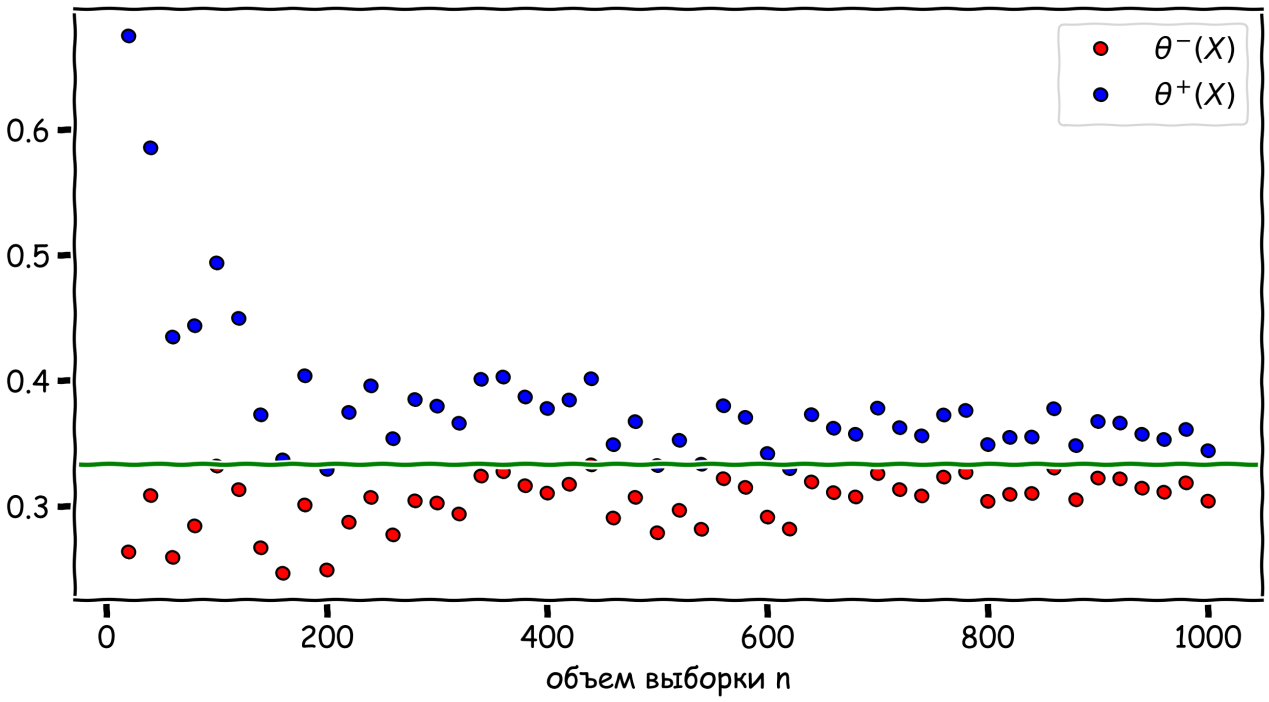

Рис. 2: Построение доверительных интервалов при разных n

1.4.2 Асимптотический доверительный интервал для B_θ

Аналогично предыдущему, построим асимптотический доверительный интервал для параметра θ распределения Бернулли B_θ. Так как $\mathsf{E}_\theta X_1 = a = \theta$, $\mathsf{D}_\theta X_1 = \theta(1-\theta)$, $\sigma = \sqrt{\theta(1-\theta)}$, то, используя центральную предельную теорему, получим

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

Аналогично предыдущему примеру приходим к неравенству вида

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

где $\tau_{1-\varepsilon/2}$ – квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения $\mathsf{N}_{0,1}$. Можно заметить, что рассматриваемый пример имеет существенное отличие от предыдущего. Дело в том, что разрешить полученное неравенство относительно параметра θ – задача не из легких. Как быть в таком случае? Вполне вероятно, что аналогично тому, как мы поступали в примере с новогодними подарками: там мы заменяли неизвестное среднеквадратическое отклонение на его состоятельный аналог S_0.

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 параметр θ на $\overline{X}$. Тогда дробь

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

заменится на

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

Тогда, для получения искомого интервала, нам достаточно решить неравенство

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

относительно θ, откуда асимптотический доверительный интервал имеет вид

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

Пример 1.4.2 *Посмотрим на численные расчеты при истинном значении параметра, равном* 0.6. *Будем строить асимптотический доверительный интервал уровня доверия* 0.95. *На рисунке 3 видно, что почти всегда зеленая линия, отвечающая истинному параметру, попадает в построенный асимптотический доверительный интервал.*

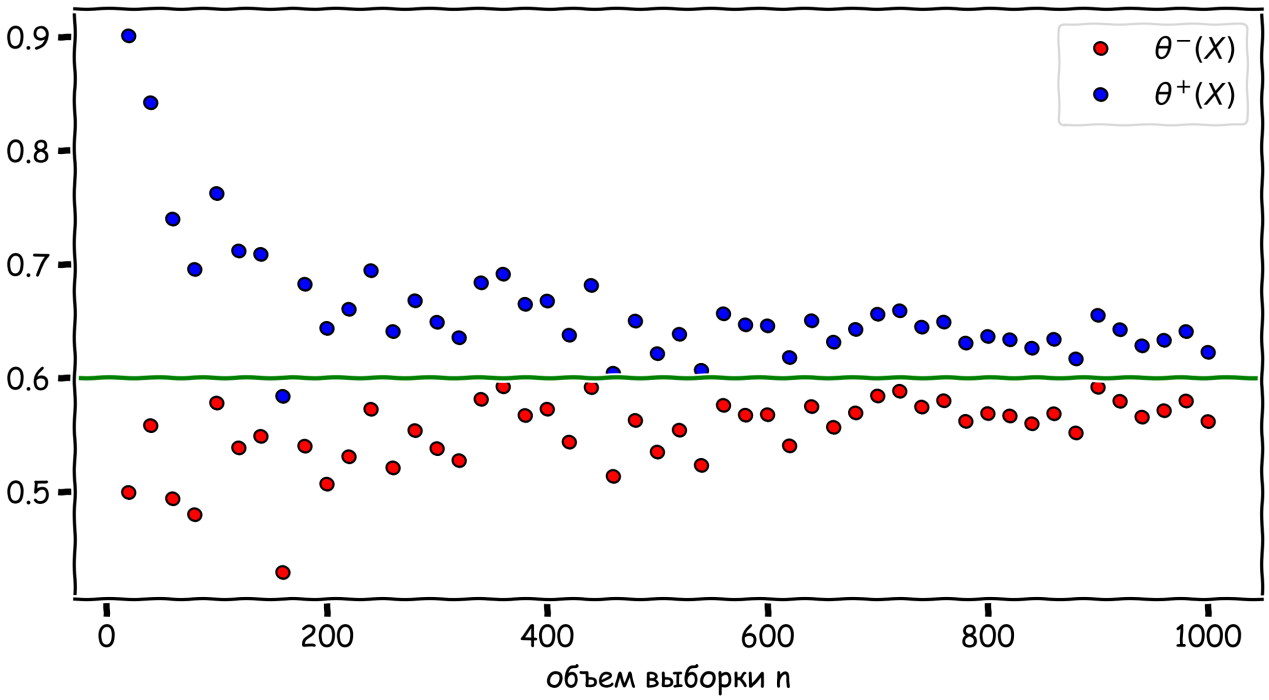

Рис. 3: Построение доверительных интервалов при разных n

1.4.3 Асимптотический доверительный интервал для $\mathsf{Bin}(\theta_1, \theta_2)$

Мы рассмотрели 2 распределения, которые зависят от единственного параметра θ, однако, как мы знаем, такое бывает не всегда. Например, биномиальное распределение зависит от двух параметров . Как тогда строить доверительный интервал? Тут возможны варианты. Если какой-то из параметров известен (очевидно , для него не нужен никакой интервал), то его можно использовать при построении доверительного интервала для другого. Если же неизвестен ни один из параметров, то для построения одного можно использовать точечную оценку (состоятельную) другого.

Если у биномиального распределения $\mathsf{Bin}(\theta_1, \theta_2)$ известен параметр θ_2, то, используя центральную предельную теорему, получим доверительный интервал для параметра θ_1 вида

$$\left(\theta_1^-,\ \theta_1^+\right) = \left(\frac{\overline{X}}{\theta_2} - \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}(1-\frac{\overline{X}}{\widehat{\theta}_1})}}{\sqrt{n}\theta_2}, \frac{\overline{X}}{\theta_2} + \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}(1-\frac{\overline{X}}{\widehat{\theta}_1})}}{\sqrt{n}\theta_2} \right),$$

где

$$\widehat{\theta}_1 = \frac{\overline{X}^2}{\overline{X} - S_0^2},$$

округленное до ближайшего целого числа.

Если же неизвестен и параметр θ_2, то вместо него в выражении для доверительного интервала имеет смысл подставить его состоятельную оценку

$$\widehat{\theta}_2 = \frac{\overline{X}}{\widehat{\theta}_1} = 1 - \frac{S_0^2}{\overline{X}}$$

Если у биномиального распределения $\mathsf{Bin}(\theta_1, \theta_2)$ известен параметр θ_1, то, используя центральную предельную теорему, получим доверительный интервал для параметра θ_2 вида

$$\left(\theta_2^-,\ \theta_2^+\right) = \left(\frac{\overline{X}}{\theta_1} - \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}(1-\frac{\overline{X}}{\theta_1})}}{\sqrt{n}\theta_1}, \frac{\overline{X}}{\theta_1} + \frac{\tau_{1-\varepsilon/2}\sqrt{\overline{X}(1-\frac{\overline{X}}{\theta_1})}}{\sqrt{n}\theta_1} \right).$$

Если неизвестен и θ_1, то имеет смысл использовать его состоятельную оценку

$$\widehat{\theta}_1 = \frac{\overline{X}^2}{\overline{X} - S_0^2},$$

округленную до ближайшего целого числа.

Во всех написанных соотношениях $\tau_{1-\varepsilon/2}$ – квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения $\mathsf{N}_{0,1}$.

Пример 1.4.3 *Рассмотрим пример численных расчетов для выборок из распределения* $\mathsf{Bin}(20, 0.8)$. *Построим доверительные интервалы для параметра* θ_2 *в случае известного параметра* $\theta_1 = 20$ *(рисунок 4) и в случае использования соответствующей оценки (рисунок 5).*

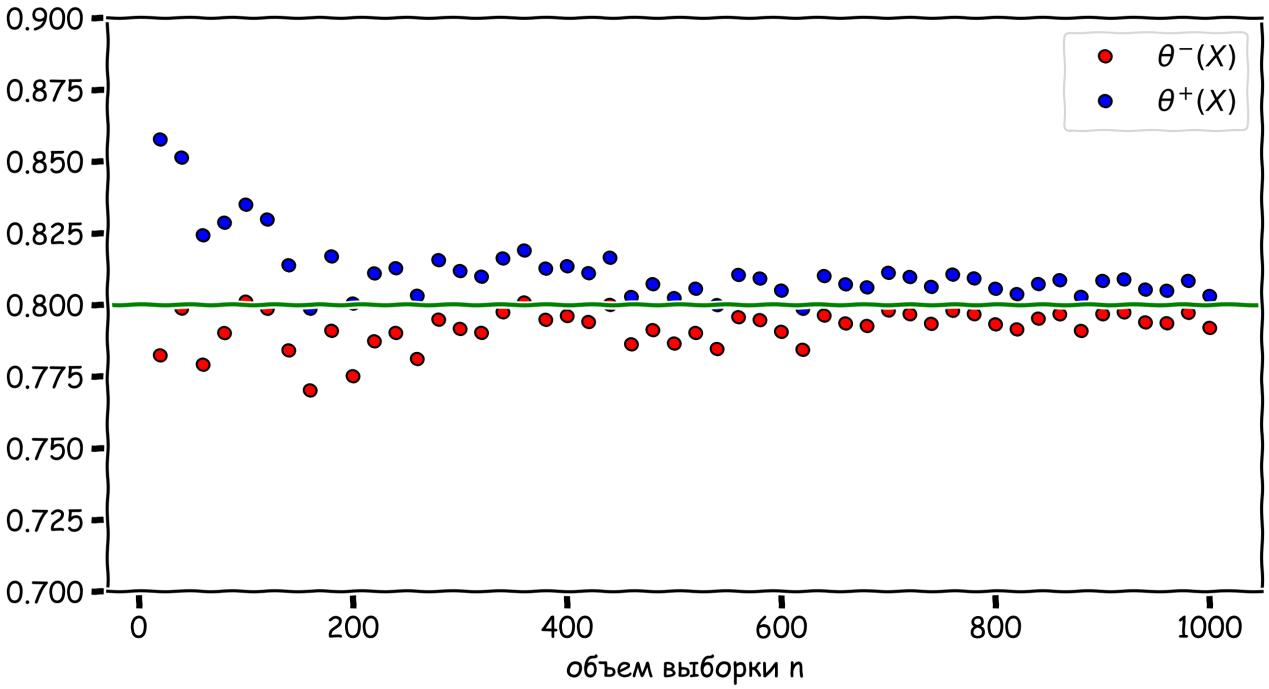

Рис. 4: Построение доверительных интервалов при разных n

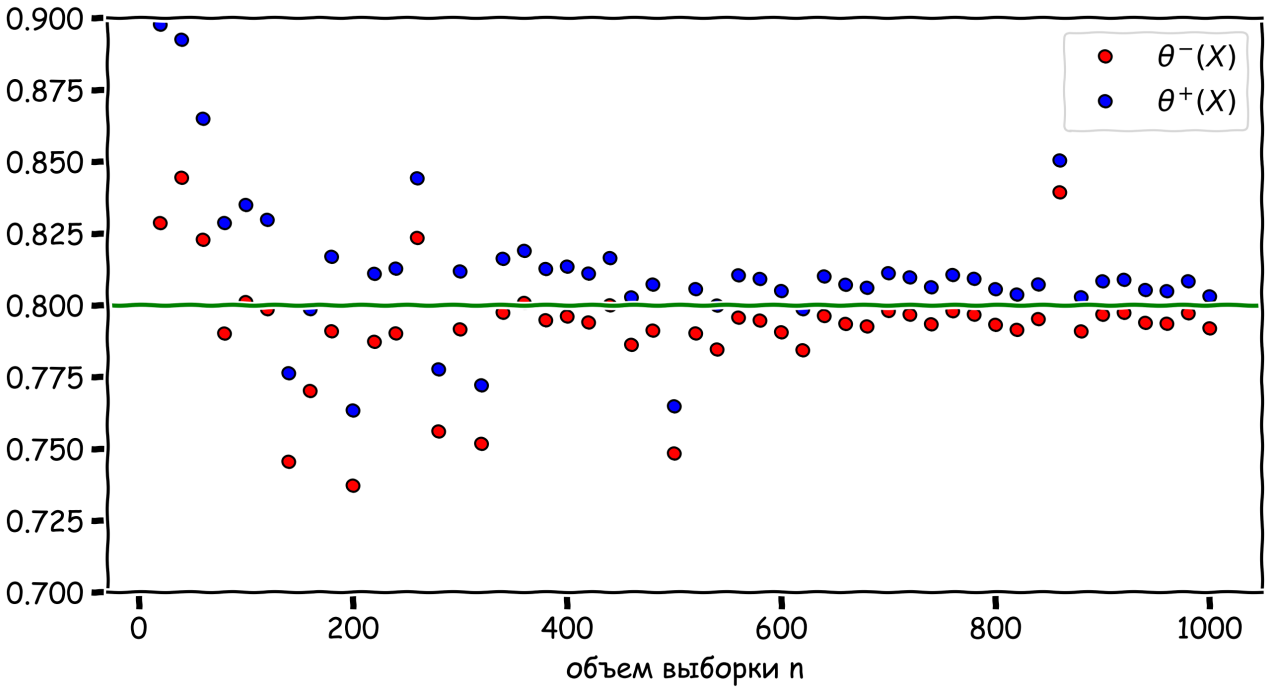

Рис. 5: Построение доверительных интервалов при разных n

На представленных рисунках отчетливо видно, что в случае известного параметра θ_1, доверительный интервал для θ_2 даже на выборках относительно малого объема содержит в себе значение истинного параметра.

В случае же, когда вместо истинного параметра θ_1 приходится использовать его оценку, хоть и состоятельную, ситуация на выборках малого объема достаточно сильно отличается. В частности, на относительно небольших объемах выборки доверительный интервал зачастую вообще не содержит в себе истинного параметра.

С ростом n оба интервала ведут себя практически одинаково и быстро сближаются к истинному значению параметра , однако «промахи» все равно случаются, что объясняется тем, что доверительный интервал содержит в себе истинный параметр с некоторой вероятностью. Напомним, что заданный уровень доверия (в нашем случае 0.95*) достигается лишь при* $n \longrightarrow +\infty$.

Отыскание доверительных интервалов для остальных распределений, рассмотренных в прошлых лекция (кроме нормального), выполняется аналогичным образом, так что для краткости приведем лишь финальные результаты.

1.4.4 Асимптотический доверительный интервал для $\mathsf{\Pi}_\theta$

Вспомним, что для распределения Пуассона $\mathsf{E}_\theta X_1 = a = \theta$, $\mathsf{D}_\theta X_1 = \theta$, $\sigma = \sqrt{\theta}$. Используя центральную предельную теорему, получим следующий доверительный интервал уровня доверия $(1 - \varepsilon)$

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

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

Пример 1.4.4 *Численные расчеты для выборок из распределения Пуассона с параметром* $\theta = 3$ *представлены на рисунке 6.*

Как и прежде можно убедиться, что с ростом объема выборки, доверительный интервал точнее «накрывает» истинное значение параметра.

1.4.5 Асимптотический доверительный интервал для G_θ

Для геометрического распределения распределения $\mathsf{E}_\theta X_1 = a = \frac{1}{\theta}$, $\mathsf{D}_\theta X_1 = \frac{1-\theta}{\theta^2}$, $\sigma = \sqrt{\frac{1-\theta}{\theta^2}}$. Используя центральную предельную теорему, получим следующий доверительный интервал уровня доверия $(1 - \varepsilon)$:

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

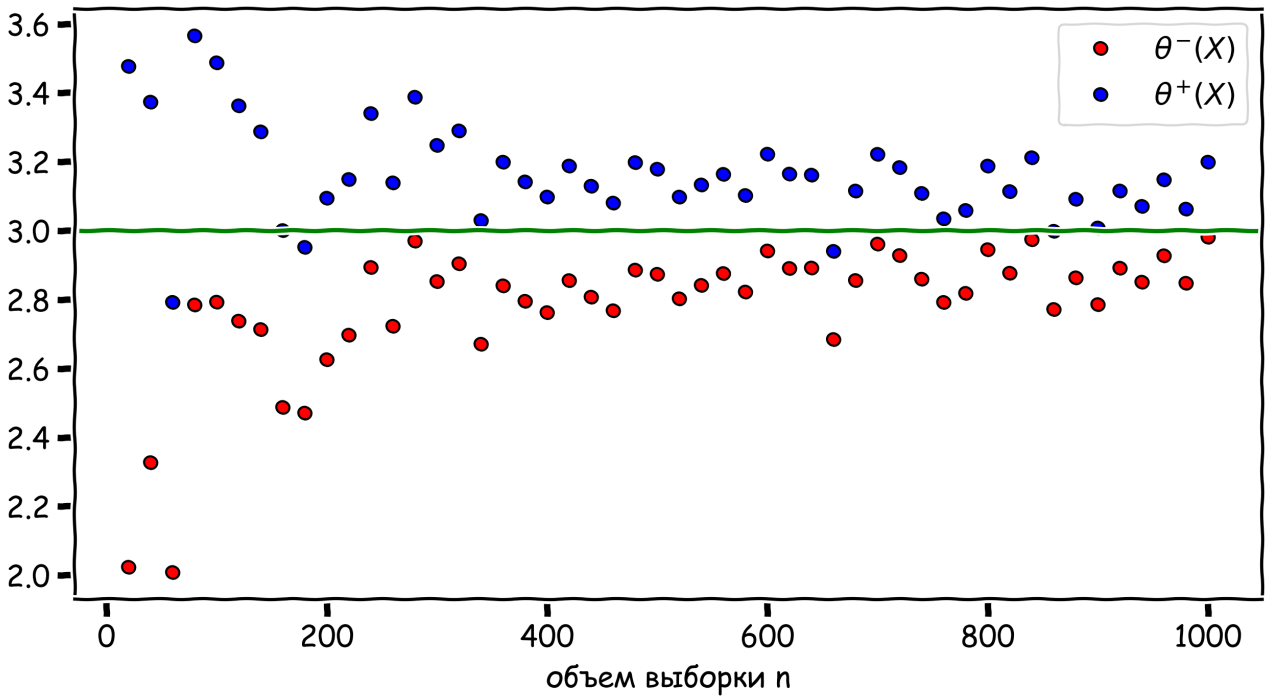

Рис. 6: Построение доверительных интервалов при разных n

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

Пример 1.4.5 *Численный пример зависимости границ доверительного интервала для различных выборок из геометрического распределения с параметром* 0.8 *представлен на рисунке 7.*

1.4.6 Асимптотический доверительный интервал для $\mathsf{U}_{\theta_1,\theta_2}$

Для равномерного распределения $\mathsf{E}_\theta X_1 = a = \frac{\theta_1+\theta_2}{2}$, $\mathsf{D}_\theta X_1 = \frac{(\theta_2-\theta_1)^2}{12}$, $\sigma = \sqrt{\frac{(\theta_2-\theta_1)^2}{12}}$. Используя центральную предельную теорему, получим следующие асимптотические доверительные интервалы уровня доверия $(1-\varepsilon)$:

$$\left(\theta_1^-,\ \theta_1^+\right) = \left(2\overline{X} - \frac{\tau_{1-\varepsilon/2}(\theta_2 - X_{(1)})}{\sqrt{3n}} - \theta_2, 2\overline{X} + \frac{\tau_{1-\varepsilon/2}(\theta_2 - X_{(1)})}{\sqrt{3n}} - \theta_2\right),$$

$$\left(\theta_2^-,\ \theta_2^+\right) = \left(2\overline{X} - \frac{\tau_{1-\varepsilon/2}(X_{(n)} - \theta_1)}{\sqrt{3n}} - \theta_1, 2\overline{X} + \frac{\tau_{1-\varepsilon/2}(X_{(n)} - \theta_1)}{\sqrt{3n}} - \theta_1\right).$$

Замечание 1.4.2 *В случае, если какой-то параметр известен, то разумнее всего использовать его. Если какой-то параметр не известен, то вместо него можно использовать соответствующие состоятельные оценки:*

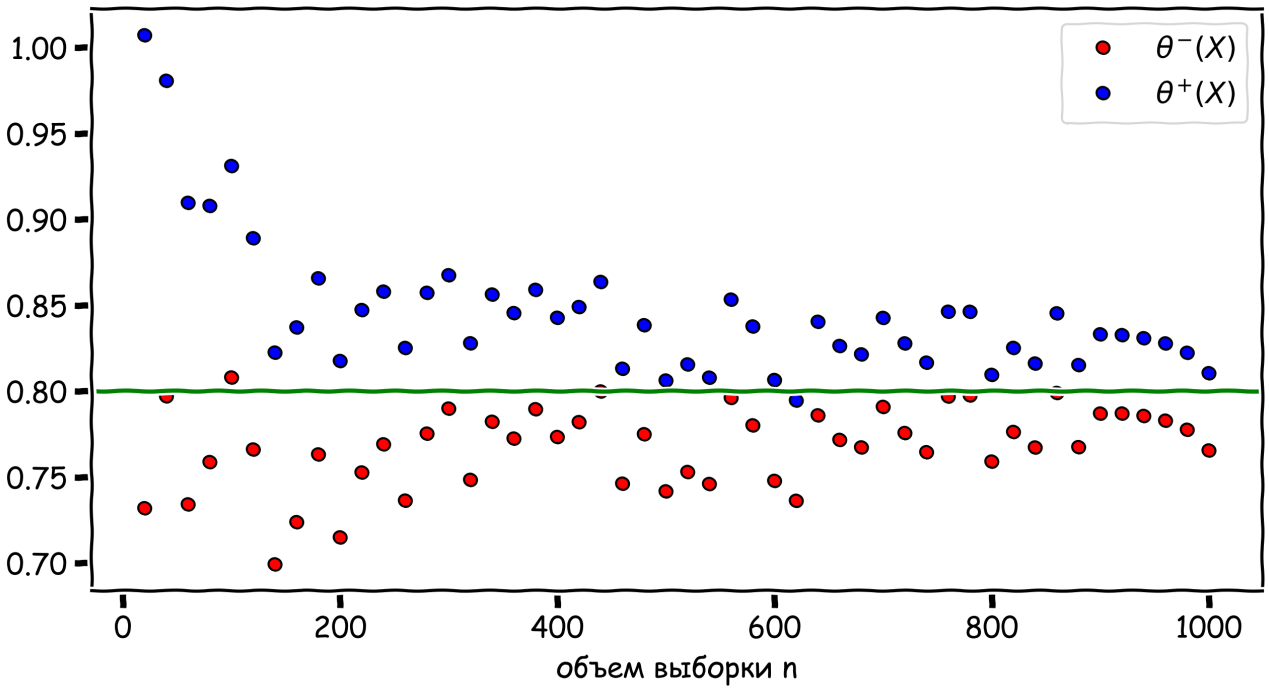

Рис. 7: Построение доверительных интервалов при разных n

$$\widehat{\theta_1} = X_{(1)},$$
$$\widehat{\theta_2} = X_{(n)},$$

где $X_{(1)}, X_{(n)}$ *– первая и* n*-ая порядковые статистики, соответственно.*

Пример 1.4.6 *Приведем пример численного расчета для равномерного распределения с параметрами* $\theta_1 = 3, \theta_2 = 7$*. Построим доверительный интервал, например, для параметра* θ_2*. На рисунках 8 и 9 представлены графики границ доверительного интервала для неизвестного параметра* θ_2 *в случаях неизвестного и известного параметра* θ_1*, соответственно.*

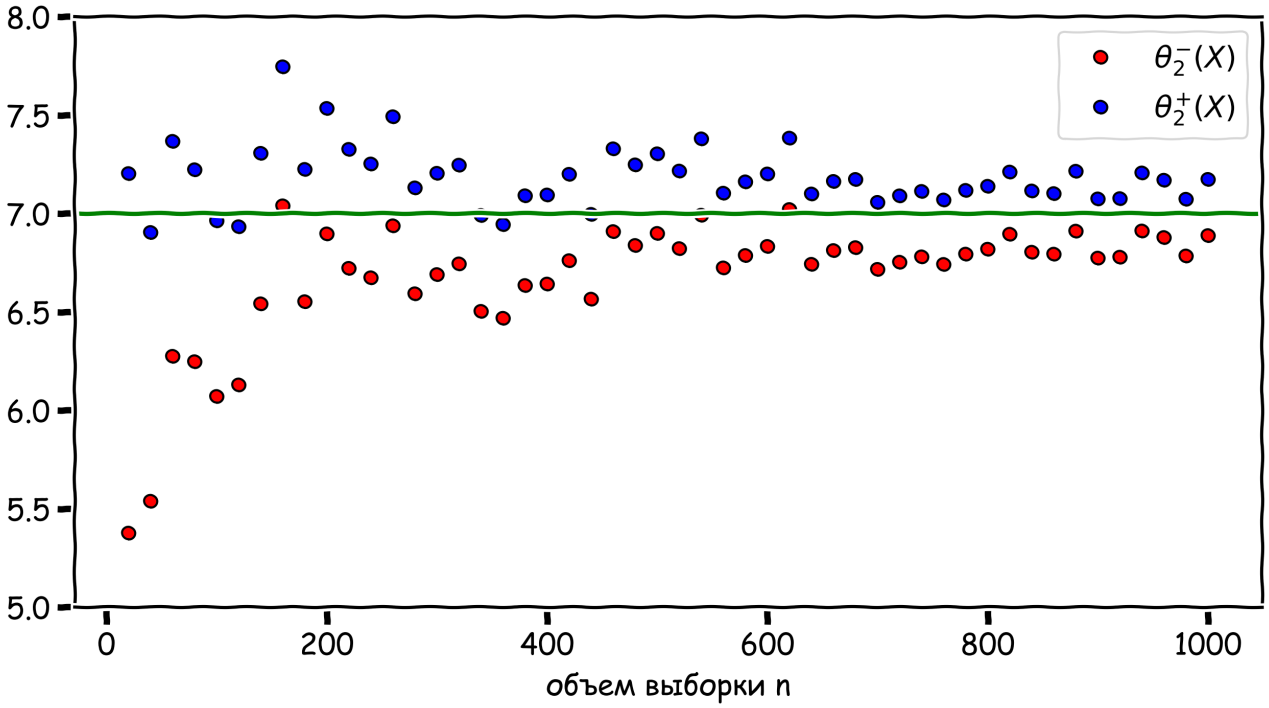

Рис. 8: Параметр θ_1 неизвестен.

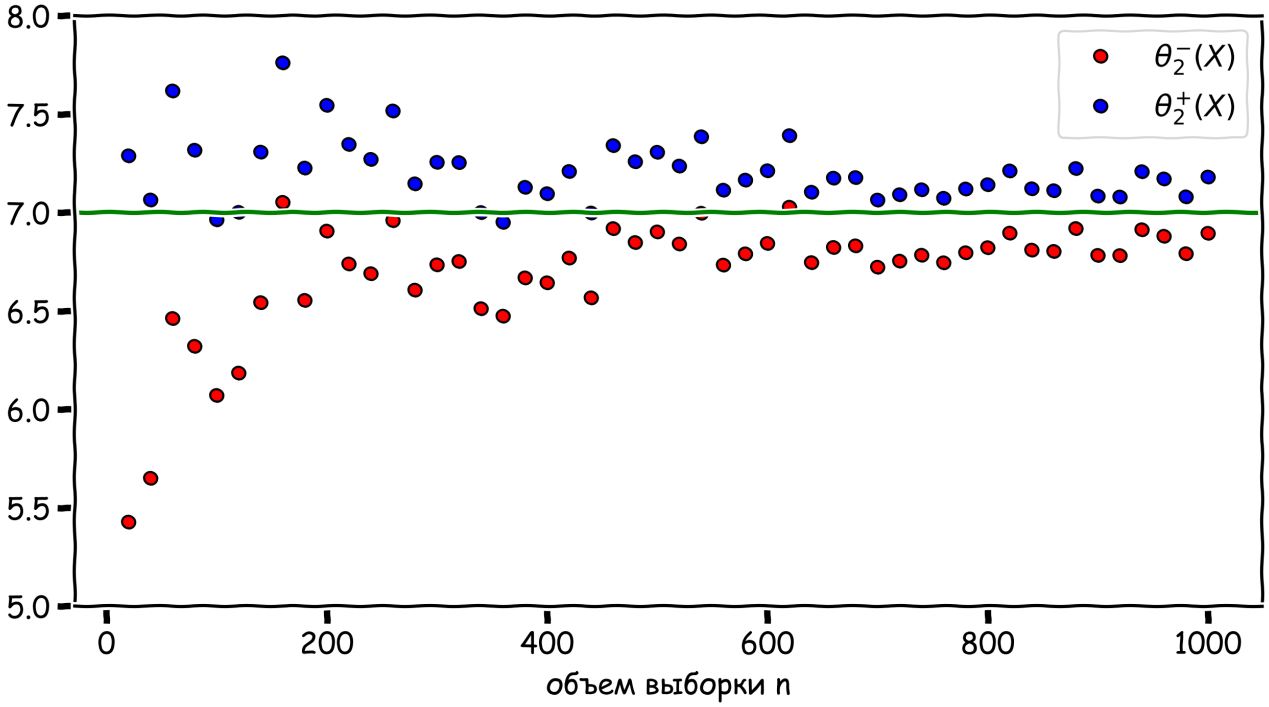

Рис. 9: Параметр θ_1 известен.

Как можно заметить, различимые отличия наблюдаются только на достаточно малых объемах выборки (примерно до 200). Далее границы интервалов в случае известного и неизвестного параметра θ_1 становятся очень похожими. Это объясняется тем, что оценка $X_{(1)}$ является очень хорошей (она

очень быстро сходится к неизвестному значению параметра), и ее использование вместо θ_1 для поиска границ, в которых лежит параметр θ_2, практически не сказывается на точности.

1.4.7 Доверительный интервал для математического ожидания в непараметрической модели

Всюду в предыдущих примерах мы пользовались параметрической моделью: считали, что наблюдаемая нами генеральная совокупность имеет некоторое конкретное распределение из заранее известного семейства. Такое бывает не всегда.

Предположим, что $X = (X_1, X_2, ..., X_n)$ – выборка из генеральной совокупности ξ, про которую известно, что дисперсия σ^2 существует и отлична от нуля, но сама дисперсия может и не быть известной. Как мы уже говорили (в частном случае) в рассмотренном в самом начале примере, асимптотический доверительный интервал для математического ожидания при известной дисперсии σ^2 уровня доверия $(1 - \varepsilon)$ имеет вид:

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

Если же дисперсия неизвестна, то, как обычно, разумно использовать ее состоятельную оценку S_0^2, и выражение для асимптотического доверительного интервала уровня доверия $(1 - \varepsilon)$ примет вид

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

1.5 Точные доверительные интервалы

Асимптотические доверительные интервалы при достаточно больших объемах выборки позволяют весьма точно оценивать границы, в которых находится истинное значение рассматриваемого параметра. Однако, как мы убедились, если объемы выборки не велики, распределение зависит более чем от одного параметра (которые также не известны), то асимптотические доверительные интервалы далеко не всегда показывают себя наилучшим образом. А есть ли какой-то способ строить интервалы, в которых лежит истинное значение параметра с заданной наперед вероятностью, вне зависимости от объема выборки? Оказывается , есть.

Если в уже рассмотренном определении асимптотического интервала убрать предел, то получим определение доверительного интервала уровня доверия $(1 - \varepsilon)$.

Определение 1.5.1 *Пусть* $0 < \varepsilon < 1$. *Интервал*

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

где θ^-, θ^+ *– это функции как от выборки (то есть оценки), так и от* ε, *называется доверительным интервалом уровня доверия (или надежности)* $1 - \varepsilon$, *если для любого* $\theta \in \Theta$ *выполняется*

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

В случае, когда в последнем выражении вместо неравенства возникает равенство, доверительный интервал называется точным.

При построении асимптотических доверительных интервалов мы использовали нормальное распределение, но доверительные интервалы для его параметров так и не построили. Пора исправить эту несправедливость.

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

Пусть $X = (X_1, X_2, ..., X_n)$ – выборка из распределения N_{a,σ^2}, где параметр a неизвестен, а дисперсия σ^2 известна. Что представляет собой параметр a в рассматриваемом случае? А это математическое ожидание случайной величины, имеющей нормальное распределение. Для математического ожидания мы уже строили доверительный интервал (хоть тогда еще и не знали, что он так называется) в задаче про подарки для новогоднего праздника, а также в конце предыдущего пункта .

Как вы, надеемся, помните, в самом начале этой лекции мы говорили, что стандартизированная случайная величина

$$Y_n = \sqrt{n}\frac{\overline{X} - a}{\sigma}$$

имеет математическое ожидание, равное нулю, и дисперсию, равную единице. Оказывается, что в случае, когда выборка берется из нормального распределения N_{a,σ^2}, то эта случайная величина имеет стандартное нормальное распределение, то есть $Y_n \sim \mathsf{N}_{0,1}$. Вспомните, для произвольно распределенной генеральной совокупности такой факт можно гарантировать только в пределе, о чем и говорит ЦПТ!

Дальше схема практически аналогична тому, что мы делали ранее. Так как $Y_n \sim \mathsf{N}_{0,1}$, то

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

Приравнивая последнее выражение к $1-\varepsilon$, мы снова получаем, что $c = \tau_{1-\varepsilon/2}$ – квантиль уровня $(1-\varepsilon/2)$ стандартного нормального распределения.

Разрешив неравенство

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

относительно a , получим точный доверительный интервал уровня доверия $(1-\varepsilon)$:

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

Замечание 1.5.1 *Заметим, что длина доверительного интервала с ростом объема выборки n уменьшается со скоростью порядка $n^{-1/2}$.*

Пример 1.5.1 *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. Мы уже использовали $\tau_{0.975} \approx 1.96$. 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.

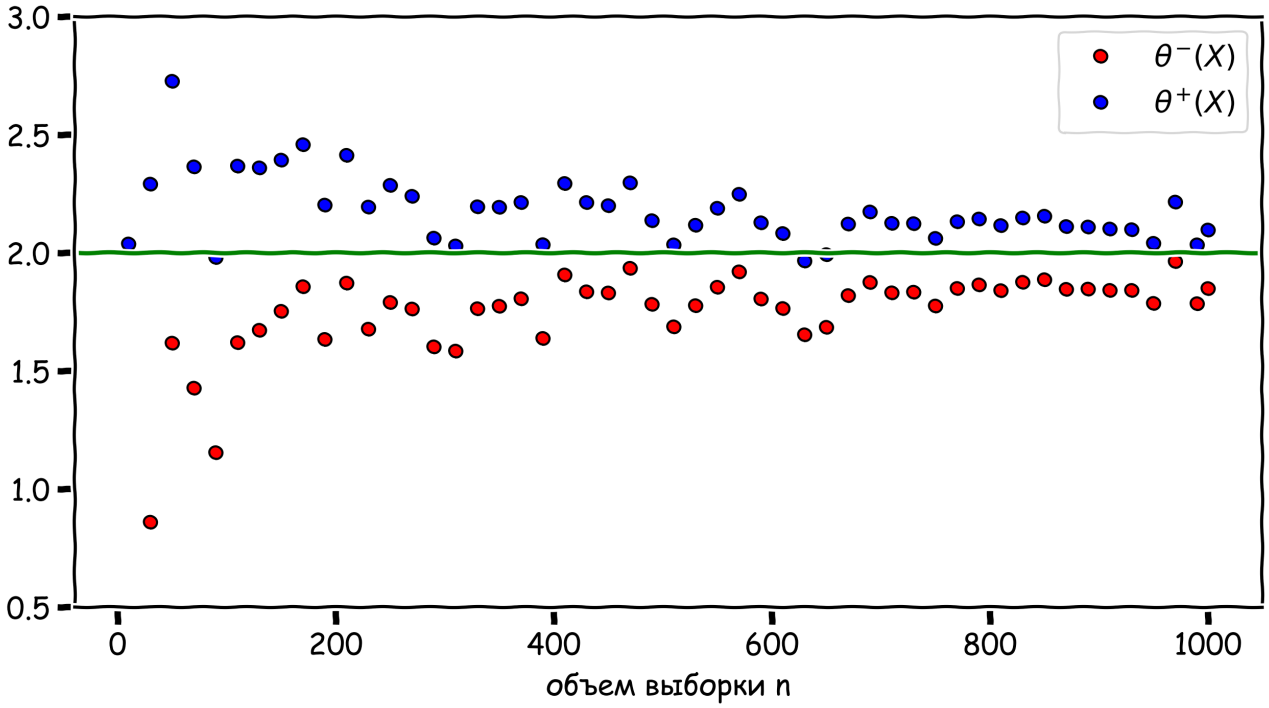

Рис. 10: Построение доверительных интервалов при разных n

Протестируем доверительный интервал на синтетических выборках большего объема и построим доверительные интервалы того же уровня доверия 0.95. На рисунке 10 , как обычно, красными точками обозначены нижние границы доверительных интервалов $\theta^-(X)$, а синими – верхние $\theta^+(X)$. Из рисунка также видно, что зеленая линия (истинное значение среднего, равное двум) не всегда попадает в доверительный интервал. Однако, в основном попадает. Кроме того, хорошо видно, что длина доверительного интервала убывает с ростом n.

1.5.2 Точный доверительный интервал для N_{a,σ^2} при неизвестной дисперсии

В реальных условиях (имея дело непосредственно с выборкой), истинное значение дисперсии может быть неизвестным. 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 не стандартное нормальное распределение, а так называемое распределение Стьюдента T_{n-1}. 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 $1 - \varepsilon$.

Замечание 1.5.2 *Значения квантилей распределения Стьюдента доступны в соответствующих таблицах. Кроме того большинство пакетов для анализа данных имеют соответствующую функцию. В частности, в* Excel *для этих целей можно использовать функцию СТЬЮДЕНТ.ОБР(*$1 - \varepsilon/2; n - 1$*).*

Проведем численный эксперимент при $\varepsilon = 0.05$. Пусть выборка берется из нормального распределения $\mathsf{N}_{3,4}$. На рисунке 11 видны соответсвующие доверительные интервалы.

Давайте сравним, насколько влияет знание дисперсии на качество доверительного интервала. Снова $\varepsilon = 0.05$ и выборка берется из нормального распределения $\mathsf{N}_{3,4}$. На рисунке 12 изображены границы доверительных интервалов: красным – при известной дисперсии, синим – при неизвестной.

Из рисунка видно, что на малых объемах выборок знание истинного значения дисперсии оказывает заметный эффект, в то время как с ростом n этот эффект снижается.

1.5.3 Доверительный интервал для σ^2 при известном a

Построим точный доверительный интервал для параметра σ^2 при известном a. Оказывается, что случайная величина

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

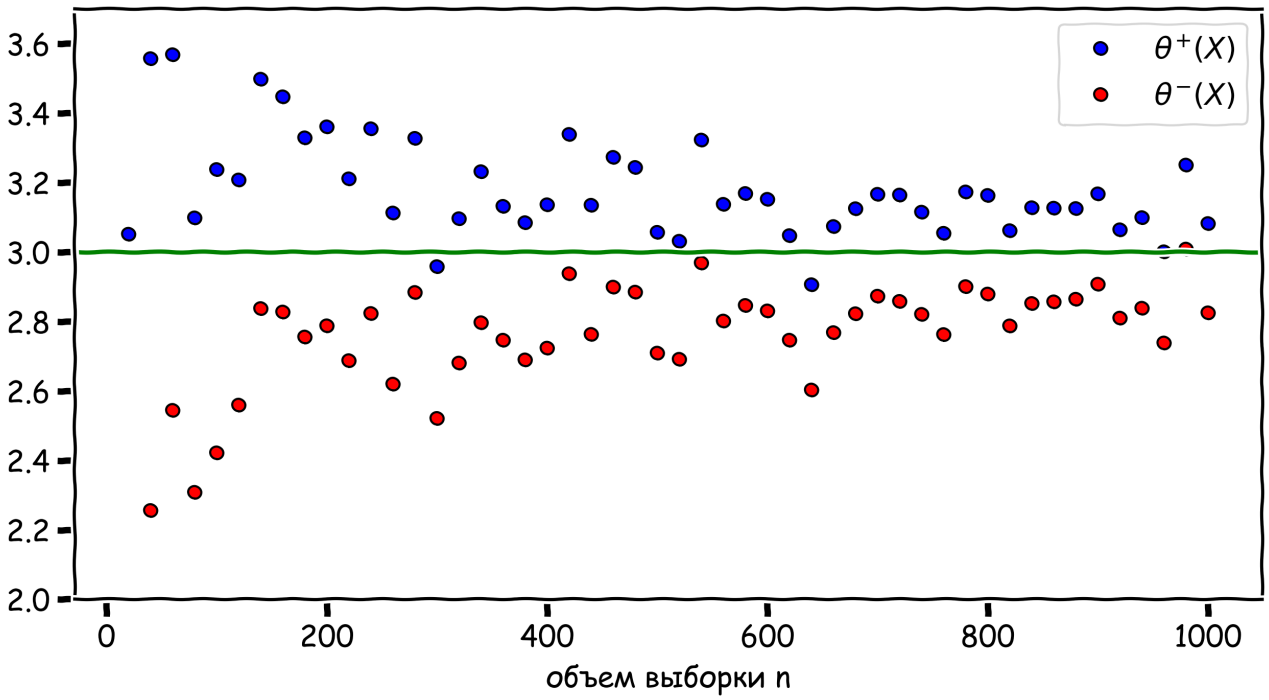

Рис. 11: Доверительный интервал для a при неизвестном σ^2

имеет так называемое распределение хи-квадрат с n степенями свободы H_n. Пусть $c_{\varepsilon/2}$ – квантиль распределения H_n уровня $\varepsilon/2$, а $c_{1-\varepsilon/2}$ – квантиль распределения H_n уровня $1-\varepsilon/2$, тогда доверительный интервал для σ будет иметь вид

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

и являться точным доверительным интервалом уровня доверия $1-\varepsilon$.

Замечание 1.5.3 *Найти значение нужной квантили распределения хи-квадрат* H_n *можно в таблицах. В* Excel *для этого можно использовать функцию ХИ2.ОБР(уровень;n), где в качестве первого аргумента нужно указать уровень квантили, а в качестве второго – n.*

1.5.4 Доверительный интервал для σ^2 при неизвестном a

При неизвестном a можно рассмотреть случайную величину

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

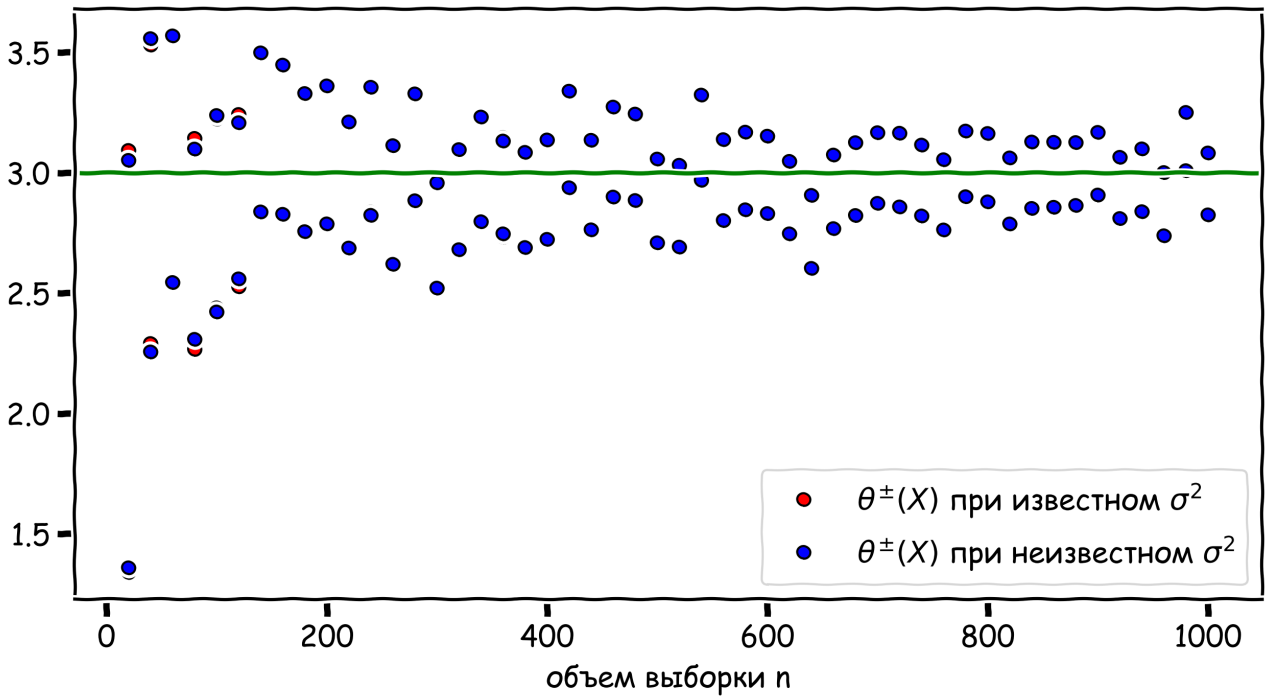

Рис. 12: Сравнение доверительных интервалов

которая имеет распределение хи-квадрат с $n-1$ степенью свободы H_{n-1}. Пусть c_1 – квантиль распределения хи-квадрат H_{n-1} уровня $\varepsilon/2$, а c_2 – квантиль распределения хи-квадрат H_{n-1} уровня $1-\varepsilon/2$, тогда точный доверительный интервал уровня доверия $1-\varepsilon$ для σ при неизвестном a будет иметь следующий вид

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

Замечание 1.5.4 *Обратите внимание, что в случае неизвестного параметра a используется распределение хи-квадрат H_{n-1}. Тогда для поиска значений нужных квантилей средствами* Excel *используется та же функция ХИ2.ОБР(), однако второй аргумент нужно указать равным $n-1$.*

Проведем численный эксперимент при $\varepsilon = 0.05$. Пусть выборка берется из нормального распределения $\mathsf{N}_{3,4}$. Логично сравнить, сильно ли влияет на ширину доверительных интервалов информация о параметре a. На рисунке 13 приведены границы интервалов. Синим – при неизвестном a, красным – при известном. Как видно, границы практически сливаются, особенно при больших объемах выборки.

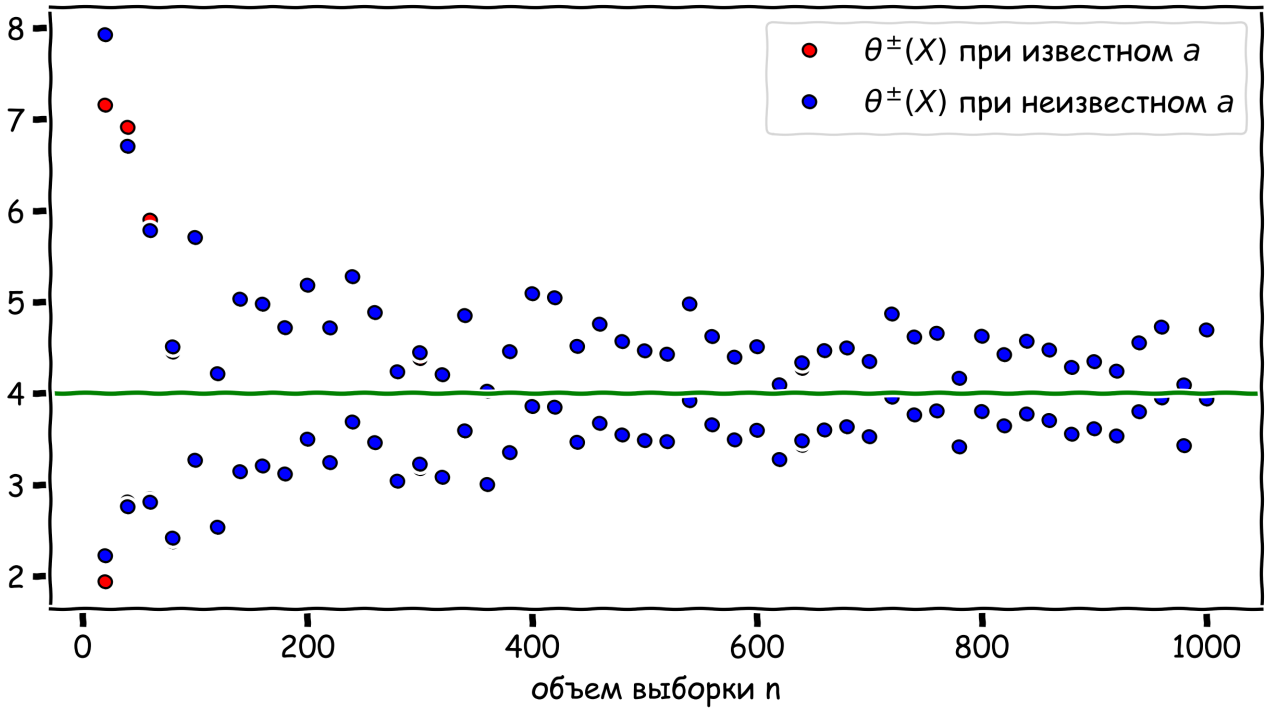

Рис. 13: Сравнение доверительных интервалов

1.6 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