Contents

1 Problem of Interval Estimation

Hello everyone! In the previous module, we've learned to construct estimators of population characteristics, as well as to apply them to estimate the parameters of some well-known distribution families. We've also found out what good estimators are. They are consistent and, preferably, unbiased, or at least asymptotically unbiased. We've also figured out how to make a histogram, a sample analog of distribution density. We've dealt with multivariate samples and found sample covariance and correlation.

The acquired knowledge and skills constitute a well-grounded basis for comprehensive data analysis and allow approximating distributions of random variables based on a sample, estimating the probability of the events of interest, finding patterns, making predictions, and so on.

On the other side, the developed apparatus has one sufficient disadvantage. We don't know how well the estimate derived from the sample approximates the true characteristic of the population. One sample can give the sample mean equal, for example, 3, and another can give 5, even though the distribution is the same. What estimate of the expected value of the population is better? We have no answer to this question as for now. Here's a hitch. Another change of the sample will give another value of a sample mean. So, what can we do? In this module, we will construct so-called confidence intervals that help estimate the inaccuracy between the true value and its sample analog.

1.1 Central Limit Theorem (CLT)

We can solve a wide range of applied problems by defining interval boundaries, in which a random variable characteristic will occur with the set probability. We need interval estimation, for example, to decide on a pharmaceutical supply, production volume of new products, or a weekend spread (the difference between the selling rate and buying rate for a currency pair). Let's consider the example.

Example 1.1.1 *Every year, a theater shows a Christmas play for kids. After the play, children receive sweet gift boxes from Santa Claus. Everything is prepared in advance. The invitations for* 1000 *persons are printed and sent beforehand. Sweet gifts are purchased during the fall when there's no rush and prices are lower. Last year,* 1000 *persons received the invitations, but only* 753 *came to see the play, and* 247 *gift boxes were left. The theater managers decided to optimize the costs, considering the statistics for the previous year. However, if they buy only* 753 *gifts (the same number of gifted sweet boxes as in the previous year), and more people are coming, the managers will have to buy additional gift boxes while the play is ongoing. Therefore, it is reasonable to pose a question. How many people will come to see the play? For at least, we want to find the probability. Thus, we* *can decide on the number of gifts to buy while making sure that everyone gets the gift, and no unneeded gifts are left behind. To answer this question, we need to know the probability that an invitee will come to see the play.*

To solve this problem, we need the apparatus of probability theory. In the previous module, we introduced the concept of convergence in probability to formulate the law of large numbers. However, we are often interested not in a random variable to which the sequence of random variables will converge but rather in its distribution function (a bounded random variable). The reason is simple. We can use the distribution function to find everything we need to learn about the bounded random variable, including probabilities of falling in certain intervals, different characteristics, and so on. Therefore, it is logical to accept the following definition.

Definition 1.1.1 (Convergence in Distribution) *Sequence of random variables ξ_n is said to converge in distribution (or converge weakly) to random variable ξ if*

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

for every point x, at which $F_\xi(x)$ is continuous.

Convergence in distribution is often denoted as

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

where d is short for $\mathsf{distribution}$.

1.2 Standardization

Before we formulate a general result of probability theory (the central limit theorem), let's discuss standardization that is useful for data analysis and can shed the light on CLT.

Definition 1.2.1 *Let's consider random variable ξ and assume that $\mathsf{E}\xi = a$, $\mathsf{D}\xi = \sigma^2$. Then, random variable*

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

is a standardized random variable.

Why standardized? As follows from the properties of the expected value,

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

and by the variance properties,

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

The expected value of the new random variable η equals zero, and variance equals one. Does it remind you of anything? Well, it is like the random variable with the standard normal distribution. As you know, its expected value also equals zero, and variance equals one.

The need for a standard transformation (standardization) is probably clear. In a sense, it's a type of normalization. A new random variable has zero mean and unit variance. Its values are dimensionless. In the Data Storage and Processing course, you've come across linear normalization. The way that we have discussed is a good alternative to it. Normalization plays a decisive role in most of the methods, including k-nearest neighbors algorithm, k-means, logistic regression, and so on.

The choice of normalization often depends on a problem and deserves a separate discussion.

Here's an interesting question. How do these talks relate to sampling and constructing intervals? Let's consider sample mean $\overline{X}$ (it is also a random variable) derived from sample $X = (X_1, X_2, ..., X_n)$ taken from population ξ with expected value $\mathsf{E}\xi = a$ and variance $\mathsf{D}\xi = \sigma^2$. Hence, as we know,

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

and variance

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

After standardization, we obtain a sequence of random variables

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

which expected value equals zero, and variance equals one. It turns out that, when n grows larger, the random variable converges (in distribution) to a normal distribution. This observation constitutes the basis for the central limit theorem.

1.3 Central Limit Theorem (CLT)

Let's formulate one of the most remarkable and significant results of probability theory, which provides a solid foundation for many methods and techniques of mathematical statistics.

Theorem 1.3.1 (Central Limit Theorem) *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables having an expected value equal to a and non-zero variance σ^2. Then, there is weak convergence*

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

As we have said, each element of sequence Y_n has 0 expected value and variance equal to 1. However, with the unbounded increase of n, we can say even more. The distribution of the considered sequence of random variables converges to a normal distribution.

Using the definition of convergence in distribution, we can reformulate the theorem so that we can apply it further:

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

because distribution function $\Phi_{0,1}$ of random variable Y with the standard normal distribution $\mathsf{N}_{0,1}$ is everywhere continuous. Despite that the analytical expression for $\Phi_{0,1}(x)$ is not necessary, we will consider it because it is often needed in different applications:

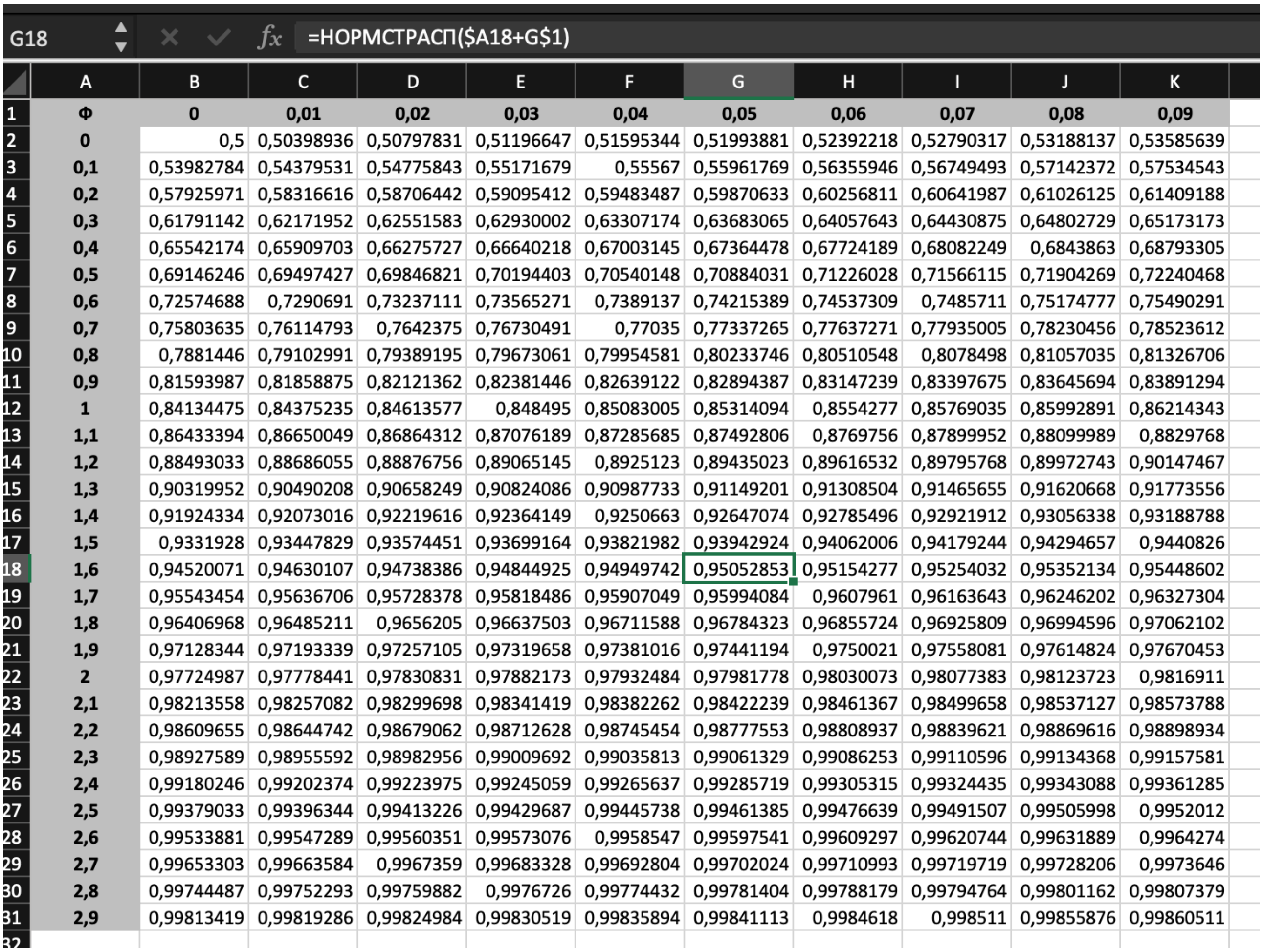
$$\Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

In many cases, we need to find not the value of function $\Phi_{0,1}$ at a given point, but rather solve an inverse problem. In other words, we aim to find an argument based on the known function value. You can find the table of values for function $\Phi_{0,1}$ on the Internet but make sure that you are choosing the right one. Those who don't have understanding can mistakenly use the error function table $\mathsf{erf}(x)$. Thus, it's better to use proven built-in functions. Most data analysis packages can create tables of values for function $\Phi_{0,1}$. Let's create this table in Excel.

Example 1.3.1 *It's easy to make a table of values for function $\Phi_{0,1}$ in* Excel. *We paste the numbers starting from* 0 *in cells* $(A2 : A31)$ *with the increment of* 0.1. *Then, we type the numbers starting from* 0 *in cells* $(B1 : K1)$ *with the increment of* 0.01. *Using the NORMSDIST function, we obtain the table of values for function* $\Phi_{0,1}$.

How to use this table? For example, let's find an argument at which function $\Phi_{0,1}$ takes the value 0.95. To do so, we look at the table to find the value closest to 0.95 and obtain the corresponding value of the argument:

$$1.6 + 0.05 = 1.65.$$



G18 =НОРМСТРАСП($A18+G$1)

	A	B	C	D	E	F	G	H	I	J	K
1	Φ	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

Figure 1: The table of values for function $\Phi_{0,1}$

For convenience, we can introduce the definition you've learned in the Data Processing and Analysis course.

Definition 1.3.1 *Let number $\alpha \in (0,1)$ be fixed and distribution function $F_\xi(x)$ be strictly increasing. The quantile at level α of the distribution function of random variable ξ is such a value x_α that*

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

In our case, 1.65 is an approximate value of the quantile at the level of 0.95 for the random variable having the standard normal distribution. Due to the frequent usage of the latter, its quantiles are designated by τ, so

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

Example 1.3.2 *Let's return to our example and recall the problem constraints. According to the last-year data, only 753 persons out of 1000 invitees came to see the play. Based on the collected statistics, and assuming that the statistics* *have a probability pattern ξ, we want to construct the interval at which the real number of visitors will fall with a high probability (for example, 0.9). This will help us to prepare the necessary number of gifts. Let's estimate the probability of the invitee's arrival to the play.*

We can use the central limit theorem and assume that there is a non-zero variance equal to σ^2 of random variable ξ, given that expected value ξ equals a. By CLT and distribution function $\Phi_{0,1}$,

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

We want the last probability to be equal to 0.9, then,

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

and $c = \tau_{0.95}$ is the quantile at level 0.95 of the standard normal distribution.

All we've got left is to solve the inequality under the probability sign. In our case, $c = \tau_{0.95}$ with respect to a. We get

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

and

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

At this point, we should stop for a moment to think about what we got. We've obtained the interval at which the true expected value of random variable ξ will fall with probability 0.9 given that $n \to +\infty$. We can interpret the expected value of random variable ξ in the problem setting as the probability of a certain person's arrival to the play (as in the Bernoulli process).

OK then, but what about statistics? How to process the collected data? The collected data consists of 753 ones and 247 zeros. Therefore, $\overline{X} = 0.753$. What's about σ? Is it unknown? To estimate it, we can use S_0, since S_0 is a consistent estimator of variance. Thus, given big n

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

Given that $\tau_{0.95} \approx 1.65$, we obtain that

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

To put it differently, the probability that an invitee will attend the play lies in the interval from 0.730 to 0.776. Recall that 1000 persons have received the invitations. Therefore, we should expect that not less than 730 and not more than 776 persons will attend the play. As a result, it makes sense to buy 776 gifts.

It's important to note several things at this point. First, there's no guarantee that the found interval will reflect the real number of visitors even though probability 0.9 is high and n is defined. Considering that n is big, we can assume that the interval approaches the theoretical one. Secondly, we used not the exact value of σ but rather the approximate one based on a larger size sample. Hence, our assumptions and estimators are justified and suitable for processing and predicting.

So what's the profit? Buying less unnecessary gifts will save the budget. Additionally, even if more gifts are needed (which is highly unlikely), the number will be small, and the overall costs will be less than in case of buying 1000 gifts from the very beginning.

What else can we do? We can take the probability equal to 0.95 instead of 0.9. The interval would become wider, but the probability to guess would be higher too. Challenge yourself and take the test!

Well, here's an important note.

Remark 1.3.1 *Quantile values from the standard normal distribution can be calculated in* Excel *either using the table or the NORM.S.INV function, which has a quantile level as the argument. For example,*

$$\mathit{NORM.S.INV}(0.95) = 1.6448\ldots$$

1.4 Asymptotic Confidence Intervals

The interval we constructed in the gift example has a special name. It is the so-called asymptotic confidence interval for the expected value of the population given unknown variance. What is an asymptotic confidence interval? What is it alike?

Everything is ready for us to formally define the concept of an asymptotic confidence interval.

Let $X_1, X_2, ..., X_n$ be the sample from a distribution that depends on unknown parameter θ from set Θ.

Definition 1.4.1 *Assume that* $0 < \varepsilon < 1$. *The interval*

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

where θ^- *and* θ^+ *are the functions of the sample (the estimators) and of* ε, *is an asymptotic confidence interval of confidence level* $1 - \varepsilon$ *(or reliability) if, for any* $\theta \in \Theta$, *the following is true*

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

When speaking about reliability $1 - \varepsilon$, we assume that, at least with this probability, the estimated value will be in this interval. We would not be content with less ($\geq$ in the definition emphasizes it). The less ε we take, the more certain we will be that the estimated value falls within the interval. Therefore, the interval will be longer. What if we take 0 as ε? We will obtain the interval that contains all possible values of parameter θ, that is, the interval that contains the entire set Θ. As you may guess, such an interval is not much use. We didn't get new data but rather added uncertainty.

Remark 1.4.1 *One more important fact. The narrower the confidence interval, the better. The construction is justified only if*

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

that is, if the interval length tends to zero (by probability) when n grows larger.

Let's find the asymptotic confidence intervals to estimate the parameters of the distributions considered earlier.

1.4.1 Asymptotic Confidence Interval for Exp_θ

We will construct the asymptotic confidence interval of confidence level $(1-\varepsilon)$ for exponential distribution Exp_θ with parameter $\theta > 0$. Let there be sample $X = (X_1, X_2, ..., X_n)$ from the distribution. Recall that $\mathsf{E}_\theta X_1 = a = \frac{1}{\theta}$, $\mathsf{D}_\theta X_1 = \frac{1}{\theta^2}$, thus, $\sigma = \frac{1}{\theta}$. By CLT, we obtain

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

Hence, according to the definition of weak convergence,

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

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

The last equation is rewritten in the following form

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

thus, $c = \tau_{1-\varepsilon/2}$ is the quantile at level $1-\varepsilon/2$ of standard normal distribution $\mathsf{N}_{0,1}$.

We're almost done. Let's solve the inequality with respect to θ to obtain

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

hence,

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

The asymptotic confidence interval of confidence level $(1-\varepsilon)$ takes the following form:

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

When n grows larger, the interval length tends to zero at the speed of $n^{-1/2}$. Let's test the obtained interval based on the example.

Example 1.4.1 *Let there be a sample from exponential distribution* Exp_θ *with true parameter* $\theta = \frac{1}{3}$. *We need to construct the asymptotic confidence interval of confidence level* 0.95 *(that is, when* $\varepsilon = 0.05$*).*

We also know how to find the quantile of the required level. Given that $\varepsilon = 0.05$, *it will be* $\tau_{1-0.05/2} = \tau_{0.975} \approx 1.96$. *The upper boundary of the confidence interval is shown in blue. The lower boundary is shown in red. The true value of the parameter is shown in green given different sample sizes* n. *As you can see, there are more errors when* n *are small than when* n *are big.*

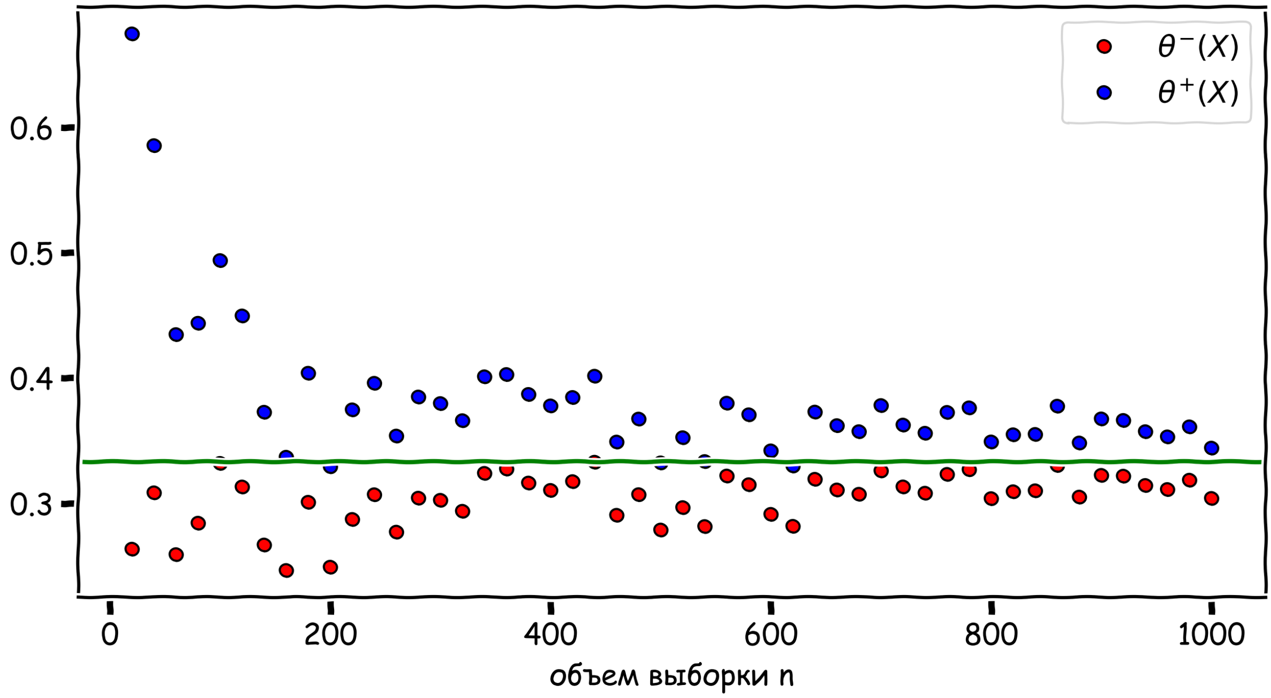


Figure 2: Constructing confidence intervals given different n

1.4.2 Asymptotic Confidence Interval for B_θ

In the same way, we can construct the asymptotic confidence interval for parameter θ of Bernoulli distribution B_θ. Since $\mathsf{E}_\theta X_1 = a = \theta$, $\mathsf{D}_\theta X_1 = \theta(1-\theta)$, $\sigma = \sqrt{\theta(1-\theta)}$, by CLT, we obtain

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

As in the previous example, we get the following inequality:

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

where $\tau_{1-\varepsilon/2}$ is the quantile at level $1-\varepsilon/2$ of standard normal distribution $\mathsf{N}_{0,1}$. You may notice that the discussed example is significantly different from the previous one. The thing is that solving the obtained inequality with respect to θ is a difficult task. What should we do? Probably, we should do the same as in the gift example, when we substituted the unknown standard deviation for its consistent analog S_0.

A sample mean is a consistent estimator for expected value, that is, $\overline{X} \xrightarrow[n\to+\infty]{\mathsf{P}} \mathsf{E}_\theta X_1 = \theta$. We substitute parameter θ in the denominator for $\overline{X}$. Then, the fraction

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

turns into

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

Then, to find the interval of interest, we need to solve the inequality

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

with respect to θ. The asymptotic confidence interval takes the following form:

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

Example 1.4.2 *Look at the numerical calculations given the true value of the parameter being equal to* 0.6. *Next, we are going to construct the asymptotic confidence interval of confidence level* 0.95.

The green line corresponding to the true parameter almost always falls within the constructed asymptotic confidence interval.

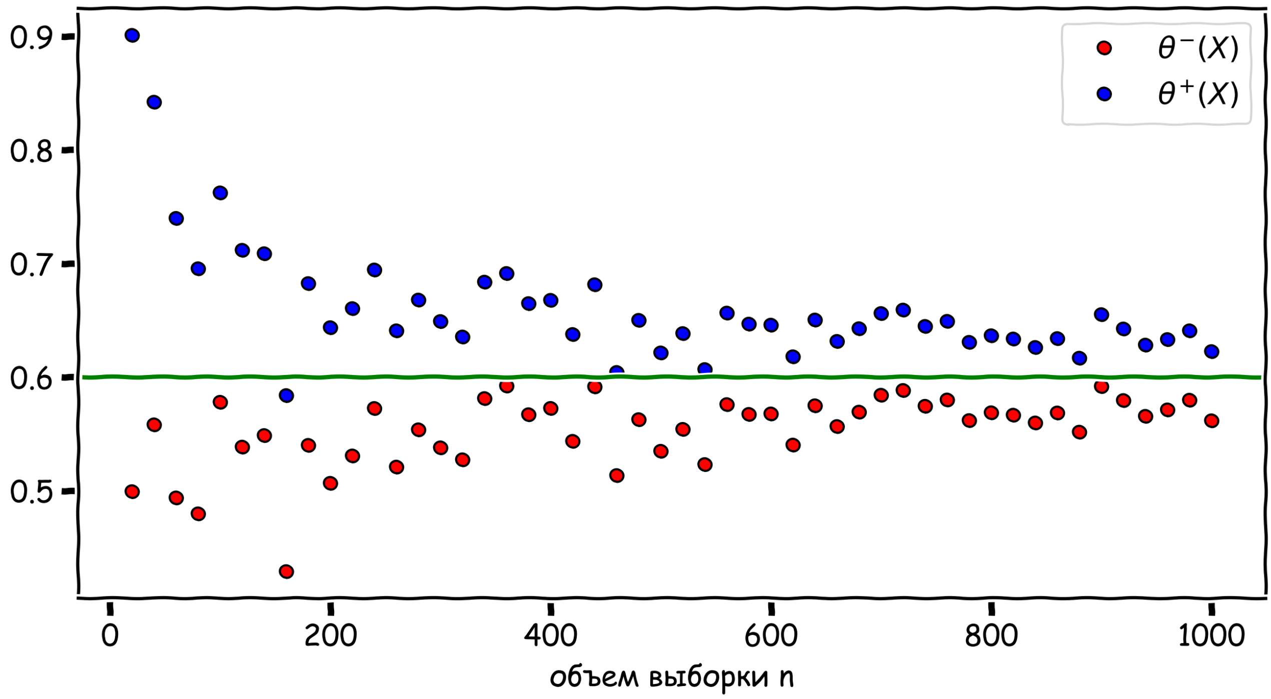


Figure 3: Constructing confidence intervals given different n

1.4.3 Asymptotic Confidence Interval for $\mathsf{Bin}(\theta_1, \theta_2)$

We've considered 2 distributions that depend on the one parameter θ. However, it is not always the case. For example, binomial distribution depends on two parameters. So, how to construct a confidence interval? There are many options. If a parameter is known (obviously, it needs no interval), this known parameter can be used to construct a confidence interval for another. If none of the parameters is known, to construct an interval, we can use the point estimator (consistent) of another.

If parameter θ_2 is known for binomial distribution $\mathsf{Bin}(\theta_1, \theta_2)$, by the central limit theorem, we will obtain the confidence interval for parameter θ_1 of the following form

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

where

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

rounded to the nearest integer.

If parameter θ_2 is also unknown, it makes sense to substitute it in the expression for the confidence interval for its consistent estimator:

$$\widehat{\theta_2} = \frac{\overline{X}}{\widehat{\theta_1}} = 1 - \frac{S_0^2}{\overline{X}}$$

If parameter θ_1 is known for binomial distribution $\mathsf{Bin}(\theta_1, \theta_2)$, by the central limit theorem, we will obtain the confidence interval for parameter θ_2 of the following form

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

If θ_1 is also unknown, it makes sense to use its consistent estimator:

$$\widehat{\theta_1} = \frac{\overline{X}^2}{\overline{X} - S_0^2},$$

rounded to the nearest integer.

In all the ratios, $\tau_{1-\varepsilon/2}$ is the quantile at level $1 - \varepsilon/2$ of standard normal distribution $\mathsf{N}_{0,1}$.

Example 1.4.3 *Let's consider an example of the numerical calculations for the samples from distribution* $\mathsf{Bin}(20, 0.8)$. *Let's construct the confidence intervals for parameter* θ_2 *when parameter* $\theta_1 = 20$ *is known and when the corresponding estimator is used.*

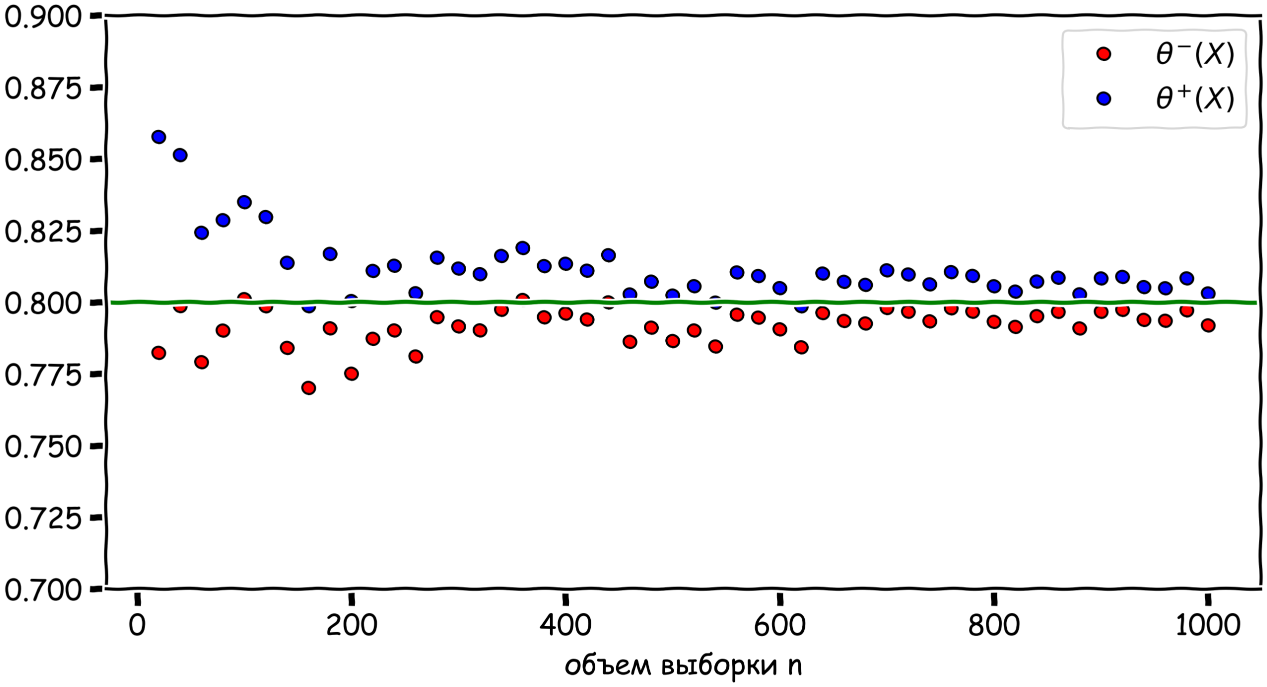


Figure 4: Constructing confidence intervals given different n

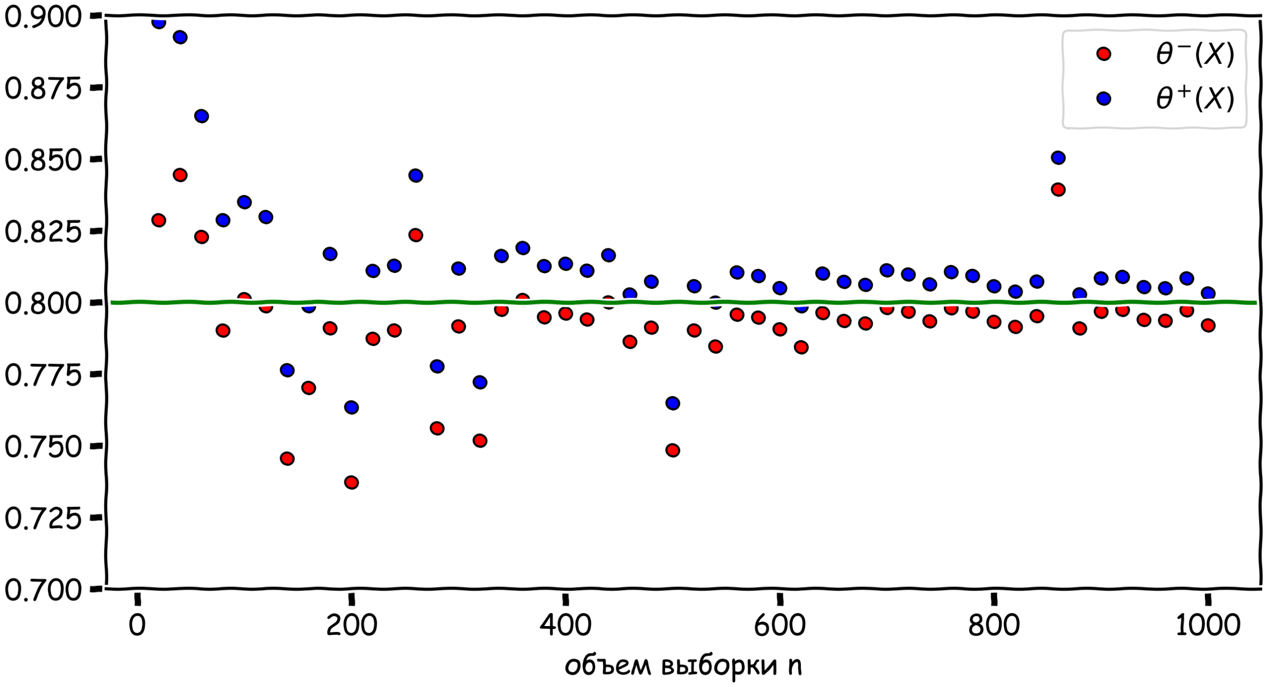


Figure 5: Constructing confidence intervals given different n

You can see that, when parameter θ_1 is known, the confidence interval for θ_2 contains the true value of the parameter even based on the small samples. When, instead of the true value of θ_1, we need to use its estimator, even though it's consistent, the situation differs drastically if the samples are small. In particular, *the confidence interval does not contain the true value of the parameter when samples are small.*

When n grows larger, the intervals behave almost the same and quickly approach the true value of the parameter. However, it isn't going entirely smoothly because the confidence interval contains the true parameter only with a certain probability. Recall that the set confidence level (0.95 in our case) is achieved only when $n \longrightarrow +\infty$.

In the same way, you can find confidence intervals for other types of distribution (except the normal one). For brevity, we will show you only the results.

1.4.4 Asymptotic Confidence Interval for Π_θ

Recall that, for a Poisson distribution, $\mathsf{E}_\theta X_1 = a = \theta$, $\mathsf{D}_\theta X_1 = \theta$, $\sigma = \sqrt{\theta}$. By CLT, we obtain the following confidence interval of the confidence level $(1-\varepsilon)$

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

where $\tau_{1-\varepsilon/2}$ is the quantile at level $1-\varepsilon/2$ of standard normal distribution $\mathsf{N}_{0,1}$.

Example 1.4.4 *The numerical calculations for the samples from the Poisson distribution with parameter $\theta = 3$ are shown on the screen.*

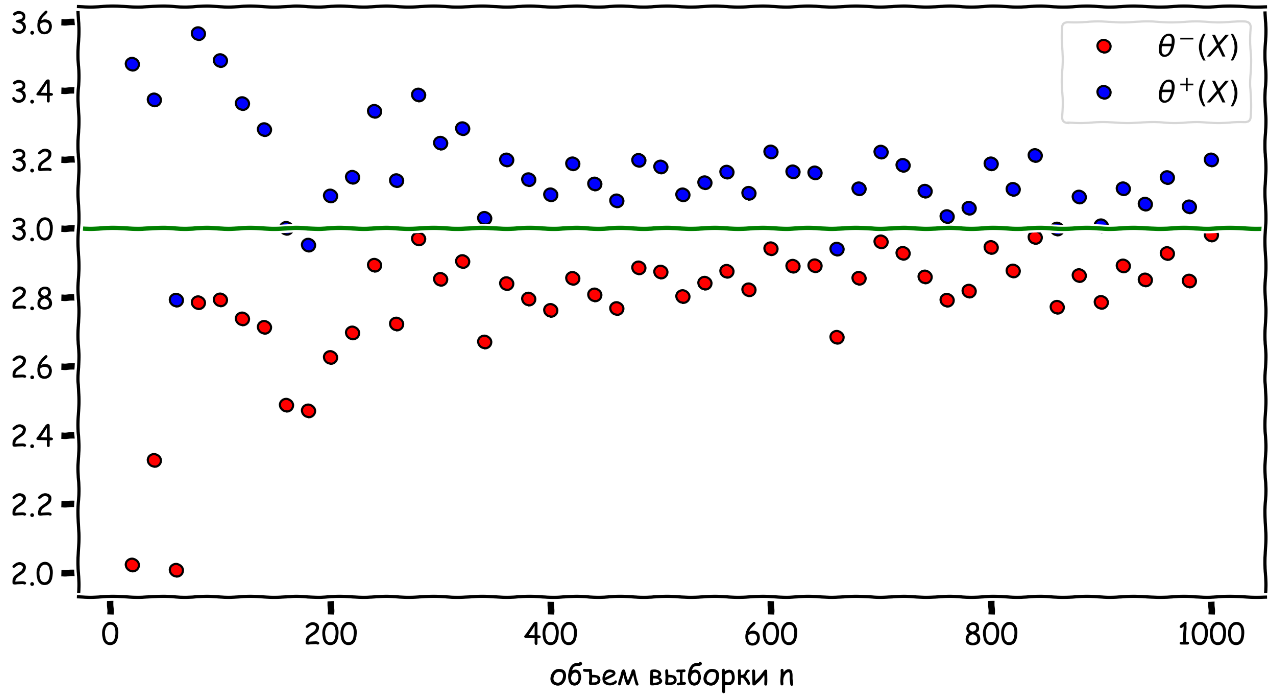


Figure 6: Constructing confidence intervals given different n

As before, you can see that, when the sample size grows, the confidence interval more accurately covers the true value of the parameter.

1.4.5 Asymptotic Confidence Interval for G_θ

For a geometric distribution, $\mathsf{E}_\theta X_1 = a = \frac{1}{\theta}$, $\mathsf{D}_\theta X_1 = \frac{1-\theta}{\theta^2}$, $\sigma = \sqrt{\frac{1-\theta}{\theta^2}}$. By CLT, we obtain the following confidence interval of confidence level $(1-\varepsilon)$:

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

where $\tau_{1-\varepsilon/2}$ is the quantile at level $1-\varepsilon/2$ of standard normal distribution $\mathsf{N}_{0,1}$.

Example 1.4.5 *The numerical example of the relationship between the boundaries of the confidence interval for different samples from the geometric distribution with parameter* 0.8 *is shown on the screen.*

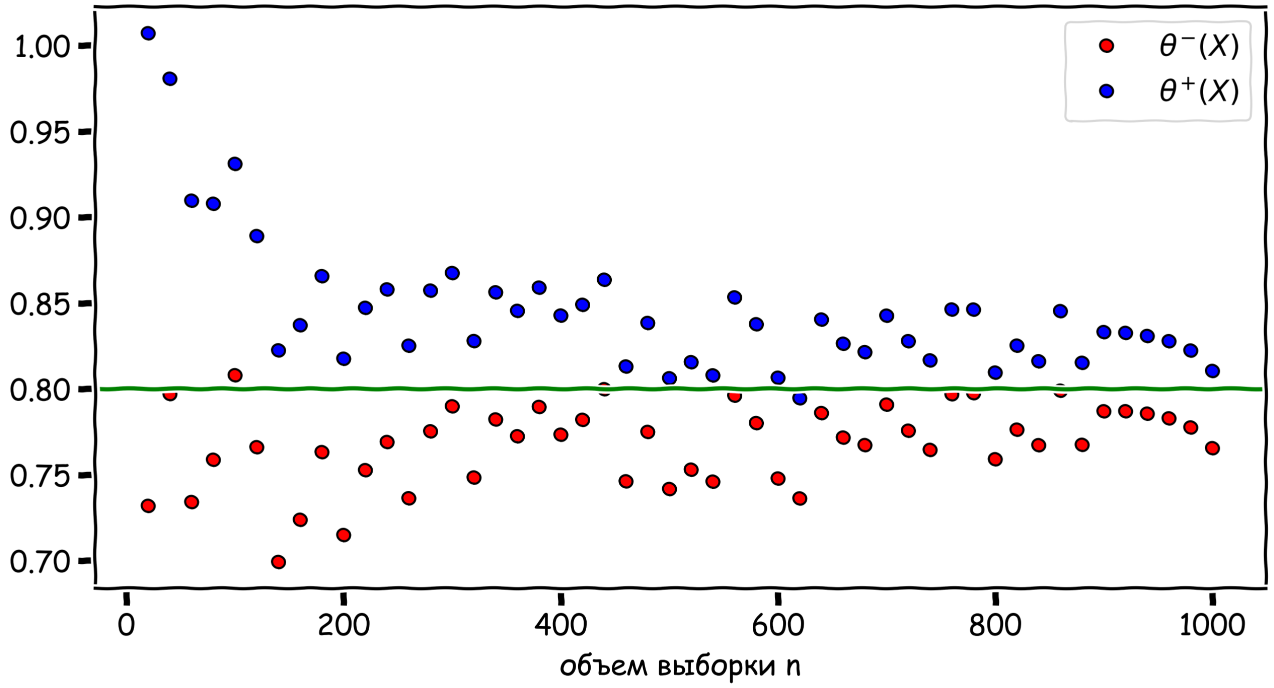


Figure 7: Constructing confidence intervals given different n

1.4.6 Asymptotic Confidence Interval for $\mathsf{U}_{\theta_1,\theta_2}$

For a uniform distribution, $\mathsf{E}_\theta X_1 = a = \frac{\theta_1+\theta_2}{2}$, $\mathsf{D}_\theta X_1 = \frac{(\theta_2-\theta_1)^2}{12}$, $\sigma = \sqrt{\frac{(\theta_2-\theta_1)^2}{12}}$. By CLT, we obtain the following asymptotic confidence intervals of the confidence level: $(1-\varepsilon)$:

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

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

Remark 1.4.2 *If the parameter is known, it's better to use it. If the parameter is unknown, the corresponding consistent estimators can be used instead:*

$$\widehat{\theta}_1 = X_{(1)},$$
$$\widehat{\theta}_2 = X_{(n)},$$

where $X_{(1)}, X_{(n)}$ is the first and nth order statistics respectively.

Example 1.4.6 *We can consider the example of the numerical calculations for a uniform distribution with parameters $\theta_1 = 3, \theta_2 = 7$. Let's construct the confidence interval, for example, for parameter θ_2. Look at the graphs of the confidence interval boundaries for unknown parameter θ_2 when parameter θ_1 is unknown or known respectively.*

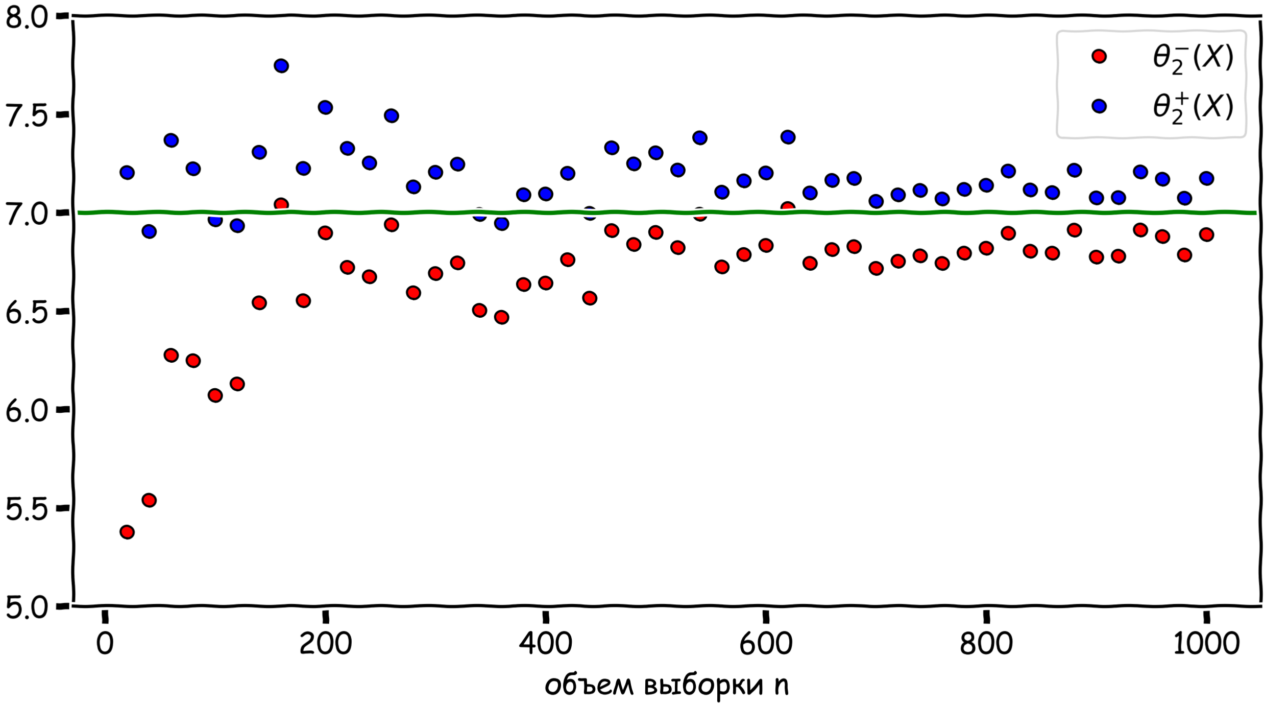


Figure 8: Parameter θ_1 is unknown.

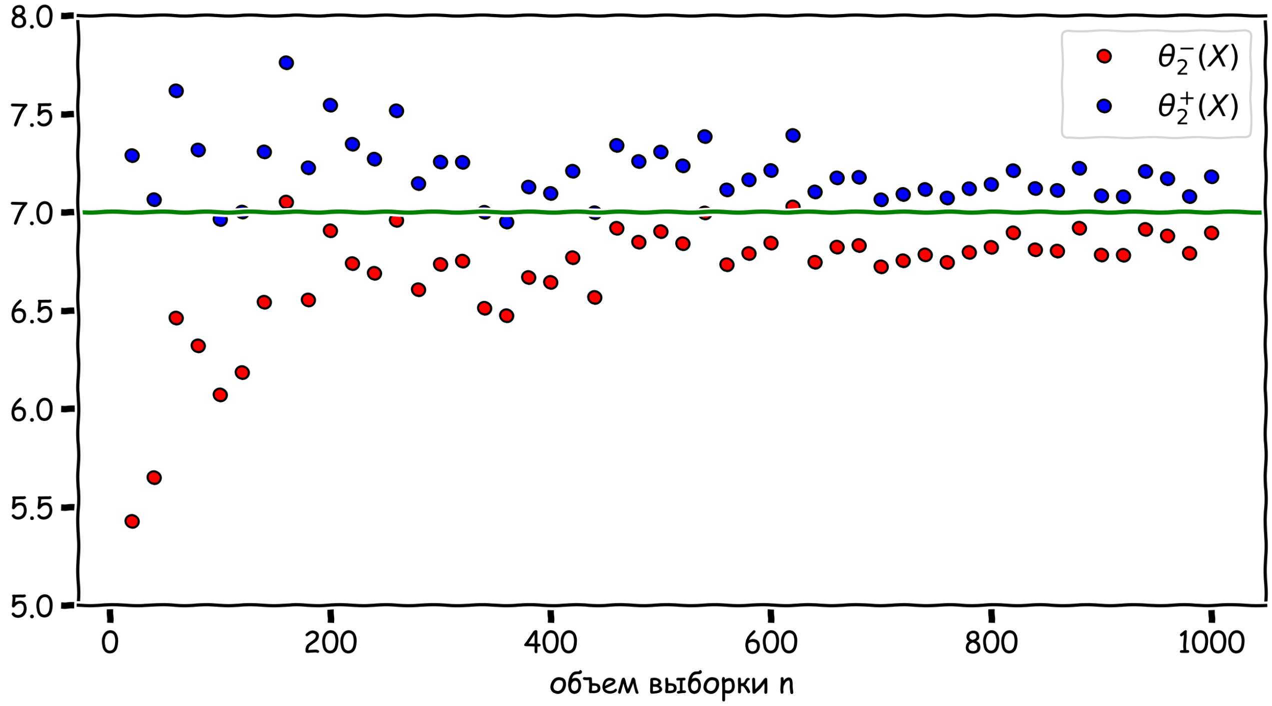


Figure 9: Parameter θ_1 is known.

These differences are distinguishable only with small sample sizes (approximately up to 200). In case of known and unknown parameter θ_1, interval boundaries become very similar. It is so because estimate $X_{(1)}$ is very good (it quickly converges to an unknown parameter value), and its use instead of θ_1 for searching for the boundaries, in which parameter θ_2 lies, does not practically affect how exact the interval is.

1.4.7 Confidence Interval for Expected Value in Non-Parametric Model

In previous examples, we used the parametric model. We assumed that the observed population has a particular distribution from the predefined family. However, it is not always like this.

Suppose that $X = (X_1, X_2, ..., X_n)$ is a sample from population ξ with existing variance σ^2 different from zero. However, the variance may be unknown. In one of the first examples, the asymptotic confidence interval for the expected value with known variance σ^2 of confidence level $(1 - \varepsilon)$ has the following form:

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

If the variance is unknown, it is reasonable to use its consistent estimator S_0^2, and the expression for the asymptotic confidence interval of confidence level $(1 - \varepsilon)$

will take the form:

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

1.5 Exact Confidence Intervals

When sample sizes are sufficiently large, asymptotic confidence intervals make it possible to exactly estimate the boundaries within which the true value of the parameter lies. However, if the sample sizes are small, the distribution will depend on several parameters (also unknown), and asymptotic confidence intervals will not always be the best choice. Can we construct intervals, in which the true value of the parameter lies with a given probability despite the sample size? It turns out we can.

If we remove the limit from the definition of the asymptotic interval, we will obtain the definition of the confidence interval of confidence level $(1-\varepsilon)$.

Definition 1.5.1 *Assume that* $0 < \varepsilon < 1$. *The interval*

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

where θ^- *and* θ^+ *are the functions of the sample (the estimators) and of* ε, *is a confidence interval of confidence level* $1-\varepsilon$ *(or reliability) if, for any* $\theta \in \Theta$, *the following is true*

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

When the equality appears instead of the inequality in the last expression, the confidence interval is called exact.

We used the normal distribution to construct the asymptotic confidence intervals, but we did not construct confidence intervals for its parameters. It's time to correct this injustice.

1.5.1 Exact Confidence Interval for N_{a,σ^2} Given Known Variance

Let $X = (X_1, X_2, ..., X_n)$ be a sample from distribution N_{a,σ^2}, where parameter a is unknown, and variance σ^2 is known. What is a in the considered case? It's the expected value of the random variable having a normal distribution. We have already constructed the confidence interval for the expected value (although we didn't know the confidence interval as a term), for example, when solving the gift problem.

As you know, standardized random variable

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

has the expected value equal to zero and variance equal to one. When a sample is taken from normal distribution N_{a,σ^2}, the random variable has a standard normal distribution, that is, $Y_n \sim \mathsf{N}_{0,1}$. Recall that this can only be guaranteed by the limit for a randomly distributed population. That's is what CLT is all about!

As to the next steps, the scheme is practically the same as before. Since $Y_n \sim \mathsf{N}_{0,1}$,

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

By equating the last expression to $1-\varepsilon$, we obtain that $c = \tau_{1-\varepsilon/2}$ is the quantile at level $(1-\varepsilon/2)$ of the standard normal distribution.

After solving the inequality

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

with respect to a, we obtain the exact confidence interval of confidence level $(1-\varepsilon)$:

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

Remark 1.5.1 *Note that the length of the confidence interval with the growth of sample size n decreases at the speed of $n^{-1/2}$.*

Example 1.5.1 *On a specific day in November, average temperature ξ in St. Petersburg is known to have the normal distribution with unknown mean a and known variance $\sigma^2 = 4$. Sample X represents the observed data in degrees Celsius:*

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

We are going to find the confidence interval of confidence level 0.95 for the estimate of expected value θ of population ξ.

Based on the sample, we find $\overline{X} = 0.7987$. Since $\varepsilon = 0.05$, we need to find quantile $\tau_{0.975}$ at level 0.975 of the standard normal distribution. We've already used $\tau_{0.975} \approx 1.96$. We plug everything we obtained in the expression for the confidence interval:

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

we obtain

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

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

In this example, the sample was taken from distribution $\mathsf{N}_{2,4}$*. Thus, true value* θ *that is equal to* 2 *falls within the confidence interval.*

We can test the confidence interval based on larger synthetic samples and construct confidence intervals of the same confidence level 0.95. As usual, red dots stand for the lower boundaries of confidence intervals $\theta^-(X)$, and blue for the upper $\theta^+(X)$. The green line (the true mean value that equals two) does not always fall within the confidence interval. But usually, it does. Moreover, the length of the confidence interval decreases with the growth of n.

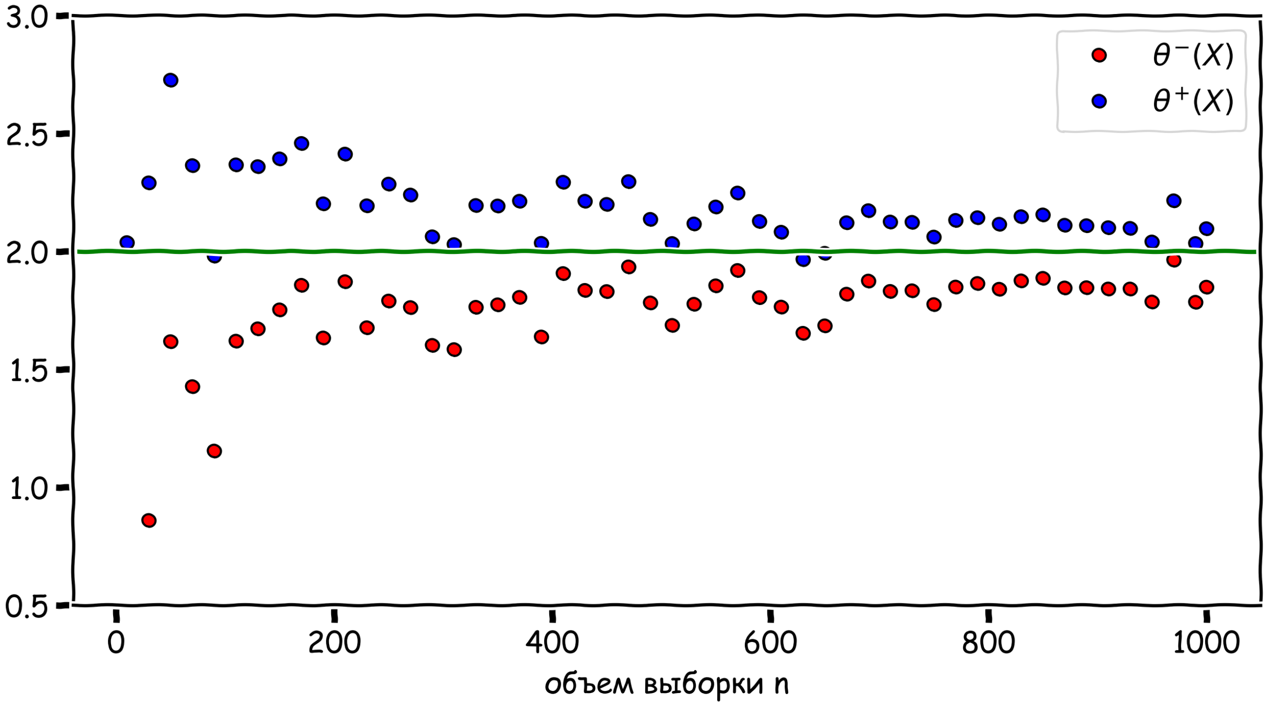


Figure 10: Constructing confidence intervals given different n

1.5.2 Exact Confidence Interval for N_{a,σ^2} Given Unknown Variance

When dealing with samples in practice, the true variance value may be unknown. Let's construct the exact confidence interval for parameter a given unknown variance σ^2. It turns out that the random variable

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

has Student's t-distribution T_{n-1} instead of a standard normal distribution. Let t_1 be a quantile of Student's t-distribution T_{n-1} at level $\varepsilon/2$ and t_2 be a quantile of Student's t-distribution T_{n-1} at level $1-\varepsilon/2$. Since the Student's t-distribution is symmetric, $t_1 = -t_2$. Hence, if $F_{t_{n-1}}$ is a distribution function of random variable t_{n-1},

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

All we've got left to do is to express a. We get

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

hence,

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

is the exact confidence interval of confidence level $1-\varepsilon$.

Remark 1.5.2 *The values of the quantiles of Student's t-distribution are given in the corresponding tables. The corresponding function is also built in most data analysis packages. In* Excel*, you can do this using T.INV(*$1-\varepsilon/2; n-1$*).*

Let's conduct a numerical experiment given that $\varepsilon = 0.05$. Assume that we have a sample taken from normal distribution $\mathsf{N}_{3,4}$. Look at the corresponding confidence intervals.

First, let's compare how the known variance affects the quality of the confidence interval. $\varepsilon = 0.05$, and the sample is also taken from normal distribution $\mathsf{N}_{3,4}$. Look at the confidence interval boundaries. The red boundaries are constructed when the variance is known, and blue when it is unknown.

When sample sizes are small, a known true variance value is significant, because this effect decreases as n grows larger.

1.5.3 Confidence Interval for σ^2 Given Known a

Let's construct the exact confidence interval for parameter σ^2 given known a. It turns out that the random variable

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

has the chi-square distribution with n degrees of freedom H_n. Let $c_{\varepsilon/2}$ be a quantile of distribution H_n at level $\varepsilon/2$ and $c_{1-\varepsilon/2}$ be a quantile of distribution H_n at level $1-\varepsilon/2$. Thus, the confidence interval for σ takes the form:

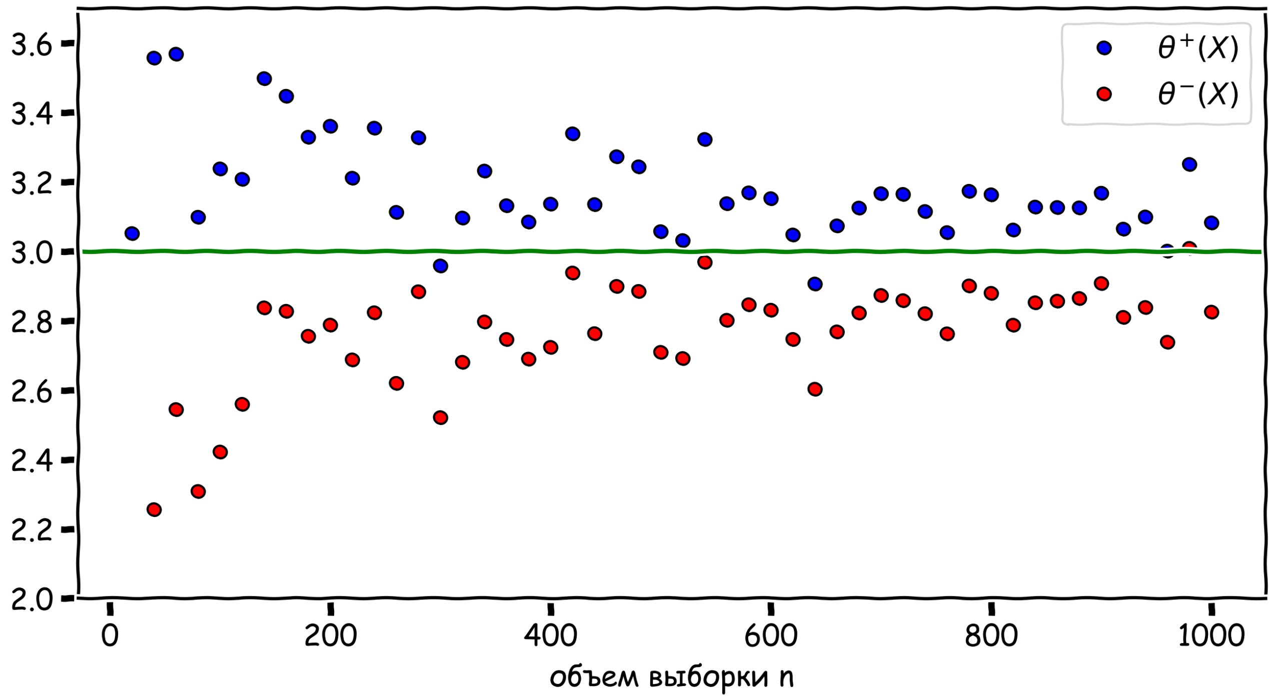


Figure 11: Confidence interval for a given unknown σ^2

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

and becomes the exact confidence interval of confidence level $1 - \varepsilon$.

Remark 1.5.3 *You can find the value of a certain quantile of chi-square distribution* H_n *in the tables. In* Excel*, you can use CHISQ.INV(level;n) where the first argument is a quantile level, and the second argument is* n*.*

1.5.4 Confidence Interval for σ^2 Given Unknown a

When a is unknown, we can consider the random variable

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

that has the chi-square distribution with $n-1$ degrees of freedom H_{n-1}. Let c_1 be a quantile of chi-square distribution H_{n-1} at level $\varepsilon/2$ and c_2 be a quantile of chi-square distribution H_{n-1} at level $1 - \varepsilon/2$. Thus, the exact confidence interval

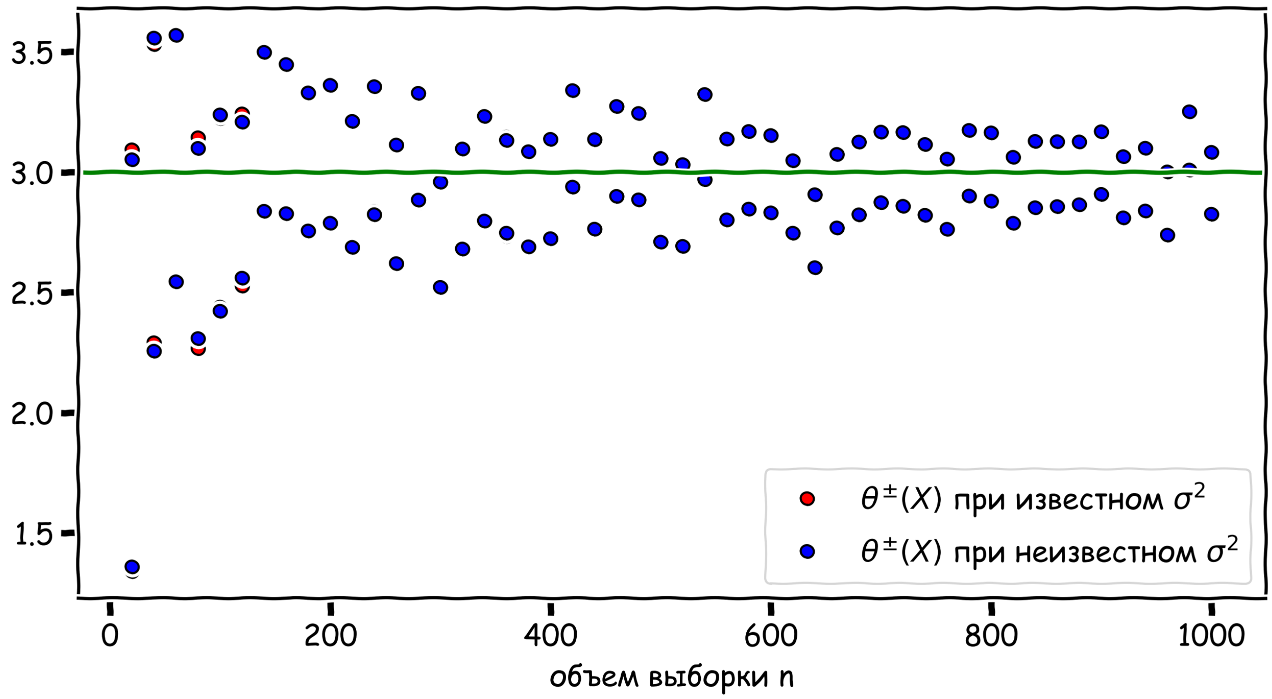


Figure 12: Comparing confidence intervals

of confidence level $1-\varepsilon$ for σ given unknown a takes the following form:

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

Remark 1.5.4 *Note that chi-square distribution H_{n-1} is used in case of unknown parameter a. To find the quantiles of interest in* Excel*, we use the CHISQ.INV function. However, the second argument should be equal to $n-1$.*

Let's conduct a numerical experiment given that $\varepsilon = 0.05$. Assume that we have a sample taken from normal distribution $\mathsf{N}_{3,4}$. It's logical to compare how the information about parameter a affects the width of the confidence intervals. Look at the interval boundaries. The red boundaries are constructed when a is known, and blue when it is unknown. The boundaries almost merge, especially when sample sizes are large.

1.6 Summary

To recapitulate briefly, in this module, we've learned to construct confidence intervals and asymptotic confidence intervals for parameters of different distributions. The confidence interval covers the true parameter with a given probability. In a sense, it even estimates the absolute error (of course, with a given confidence

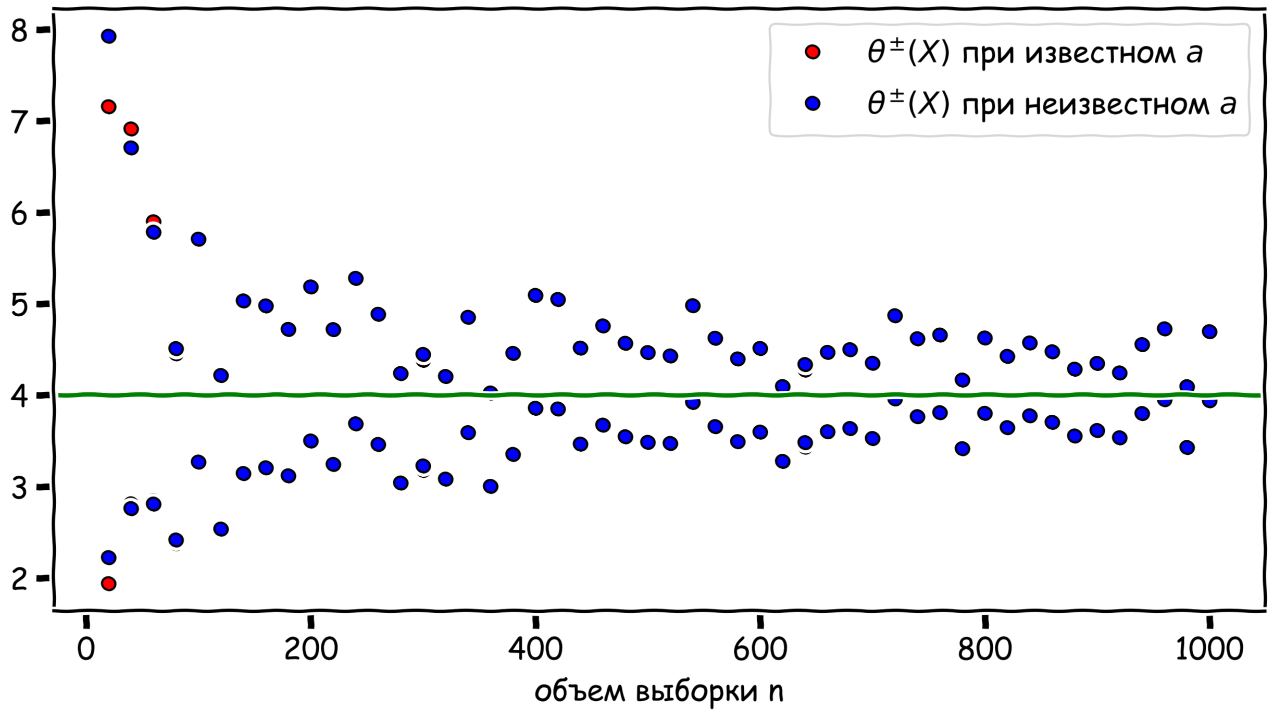


Figure 13: Comparing confidence intervals

level) of a particular value with respect to the true value of the parameter. Moreover, it often shows the error of a given point estimate (especially located right in the middle of the confidence interval) that approximates the true value. Well, we haven't discussed the remaining problem of mathematical statistics — hypothesis testing. We will cover it in the next module.