Contents

1 Descriptive Statistics

Hello everyone! You've already reviewed the mathematical models of random events and processes that can be studied using probability theory. You've also learned how to construct them. To sum up, the central problem of probability theory is making predictions. For example, we are trying to find the probability of some event, the distribution of a random variable, or a system of random variables, the characteristics thereof, as well as the distribution of a random variable process changing over time. It all looks good in theory, but putting it into practice seems to be a struggle.

In real life, we often don't know what to expect. Usually, we observe only the forms of some regularity after an experiment, but the regularity itself (or its probability model) remains hidden. We can measure the height of each child in kindergarten or volleyball team. We can obtain data on dollar exchange rate fluctuations from the bank. We can get the team stats for specific games or championships. But why do we need it? What for?

Well, we want to look in the future. The height of children in kindergarten can help us to choose the right-size furniture. If the tables or chairs are too high, children will not be able to climb them. When we analyze the dollar exchange rate, we want to know when to exchange the local currency, or when to sell it. The reason is clear. We want to benefit from it. The team statistics can help us in betting with friends. There are many similar examples.

What do we rely on? Of course, we are looking for patterns, regularities, and other things that are somehow predetermined. The well-known principle of physics claims that if you conduct the experiment several times, you will get the same result. We hope that the considered events have some patterns, even if these patterns are hidden. We can do a lot more with the known patterns because the apparatus of probability theory is well studied. We can find event probabilities, compare them, and calculate characteristics, for example, mean and spread. In other words, we can predict and analyze.

Finding patterns (or revealing the absence thereof) is one of the main problems of mathematical statistics. The next modules will review different methods of mathematical statistics.

1.1 Sample and Population

Let's start with the worst-case scenario when we know nothing. Well, we have a collected dataset but nothing else. What can we do? How to interpret it? Let's look at the example.

Example 1.1.1 *Pete collected the data about time (in minutes) when Donna was late for their dates. The data is a set of numbers separated by commas (in the form of a numeric vector). The set is designated by X:*

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

Pete assumes that Donna's late arrivals (although they are random) can be described by random variable ξ having some (unknown) distribution.

So, what do we have? As in the previous module, we are dealing with a sample from population ξ. What Pete wants is to find out how available sample X affects population ξ, so that he can make predictions and conclusions about Donna's lateness.

Recall that, a sample is a set of n numbers $X = (x_1, x_2, ..., x_n)$. In this example, the sample consists of 15 elements. Pete collected the data for that sample after 15 dates with Donna. If he had zeroed the counter, he would have obtained another sample. However, Donna would be the same. What statistics can we obtain when the data is constantly changing?

The thing is that the introduced sample definition is the so-called sample after the experiment. For statistical purposes, it is wise to accept the following definition of a sample.

Definition 1.1.1 *Let ξ be a random variable of interest. Sample $X = (X_1, X_2, ..., X_n)$ is a set of n independent random variables having the same distribution as ξ.*

When speaking about a sample before a particular experiment, we will think of a set of independent random variables and designate it by $X = (X_1, X_2, ..., X_n)$. If an experiment has been conducted, and we have obtained a set of specific population values, a sample will be numeric vector $X = (x_1, x_2, ..., x_n)$. Further on, we will not be specific about sample interpretation (a numeric vector or random variable vector). It will be clear from the context.

1.2 Empirical Distribution

So, our task is to learn more about population ξ. Since ξ is a random variable, it is logical to construct a sample-based random variable, which distribution will approximate true distribution of ξ (therefore, it will also approximate its true characteristics, including expected value, variance, and so on). As we noted in the previous module, a good approximation candidate is the so-called empirical random variable ξ^*. For convenience and further applications, we will introduce the following important concept.

Definition 1.2.1 *Assume that we have sample $X = (X_1, X_2, ..., X_n)$. If we sort the sample elements in ascending order, a new set of random variables satisfying the inequalities*

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

is called a variational series.

Obviously, for example, $X_{(1)} = \min\{X_1, ..., X_n\}$, $X_{(n)} = \max\{X_1, ..., X_n\}$. Please also note that kth term of the variational series is often called order the kth order statistic.

Let's return to our example. Based on the sample of Donna's late arrivals, we can construct the following variational series:

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

The variational series well visualizes a small amount of data and gives us some kind of statistics, because we can immediately conclude that Donna arrived on time just once, she was late twice for two minutes and once for three minutes, and so on.

Variational series for a particular sample are not constructed manually. Any data processing tool can do it for you using an order function. For example, Excel offers the sorting function in the Data tab.

Based on the variational series, it is handy to construct empirical distribution of random variable ξ^*. By grouping the same values and assigning them probabilities proportional to the value frequency in the sample, we obtain

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

.

The empirical distribution can help us a lot. For example, based on the empirical distribution of random variable ξ^*, we can construct the empirical distribution function, that is, distribution function ξ^*. In our case, it's defined as the ratio:

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

and its graph is shown on the screen.

Now we can estimate and predict. For example, how to estimate the probability of the event that a croissant and coffee that Pete has bought for Donna will not get cold when Donna finally arrives? In the language of probabilities, what is the probability of the event that she will be late for less than or equal to 5 minutes? We know how to calculate it:

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

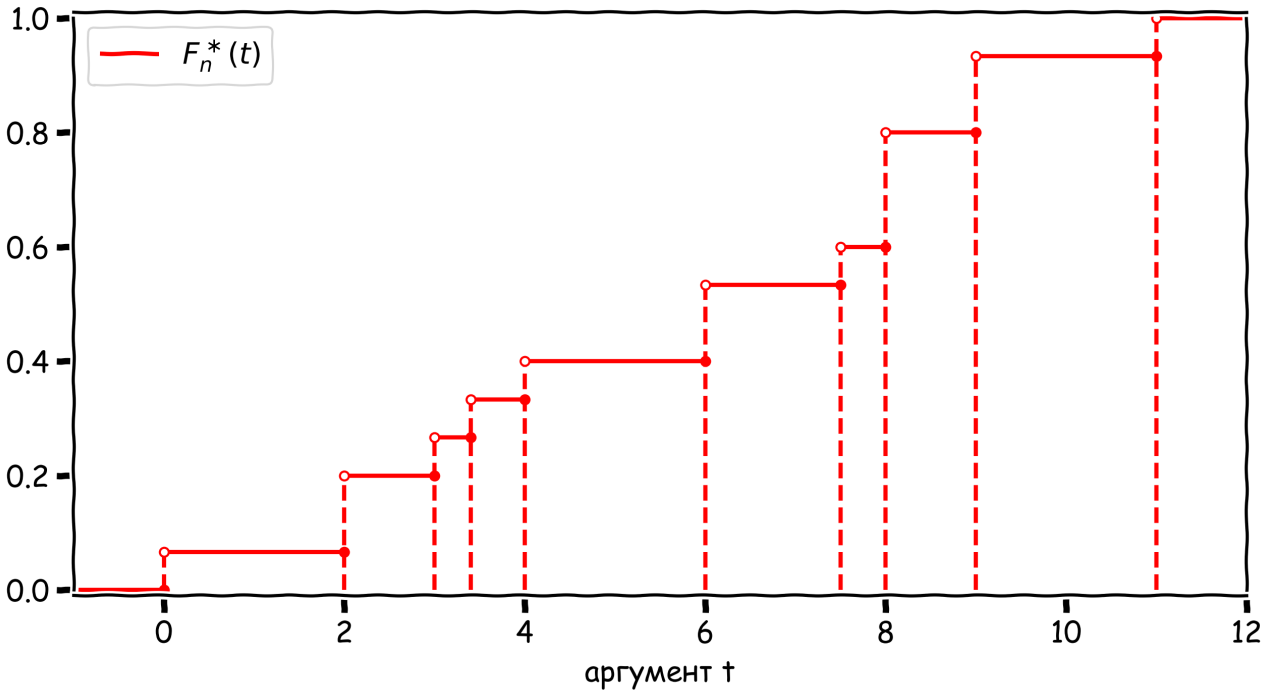

Figure 1: Empirical distribution function based on the sample.

well, not that much, actually. Anyway, it would be better to buy flowers or, at least, a hot travel mug.

It is often suffice to estimate some global characteristics of population ξ (also called measures of central tendency), including expected value $\mathsf{E}\xi$, variance $\mathsf{D}\xi$, standard deviation σ_ξ, and so on. Well, let's do it!

1.3 Sample Moments

It's logical to assume that if the distribution of empirical random variable ξ^* approximates the true distribution of population ξ, then the expected value, variance, and other characteristics of the empirical random variable are good counterparts of population characteristics.

In the general case, distribution ξ^* is given by the table:

ξ^*	X_1	X_2	X_3	...	X_n
P	$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$	...	$\frac{1}{n}$

,

where the value of each sample element is equally likely (since we don't give preference to any of them).

$$\mathsf{E}\xi^* = X_1 \cdot \frac{1}{n} + X_2 \cdot \frac{1}{n} + X_3 \cdot \frac{1}{n} + ... + X_n \cdot \frac{1}{n} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n}.$$

What did we get? Well, the expected value of the constructed random variable is nothing but the arithmetic mean of sample elements. That's why this characteristic has a unique name.

Definition 1.3.1 *Assume that we have sample* $X = (X_1, X_2, ..., X_n)$. *A value equal to the arithmetic mean of sample elements is called sample mean and denoted by* $\overline{X}$. *To put it differently,*

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

So, once again, what is the point of using sample mean $\overline{X}$? Since distribution ξ^* approximates true distribution ξ, then, expected value $\mathsf{E}\xi^*$ (being sample mean $\overline{X}$) approximates true distribution $\mathsf{E}\xi$. The meaning of the latter is simple. It's a mean probability value of random variable ξ.

Let's return to the example of Donna's late arrivals. The sample mean is calculated as follows:

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

What does it mean? It means that, on average, Pete was waiting for Donna for almost 6 minutes. The blue points are the sample elements. The red point corresponds to sample mean $\overline{X}$. The red point does not coincide with any of the sample elements, and it is located somewhere in the middle.

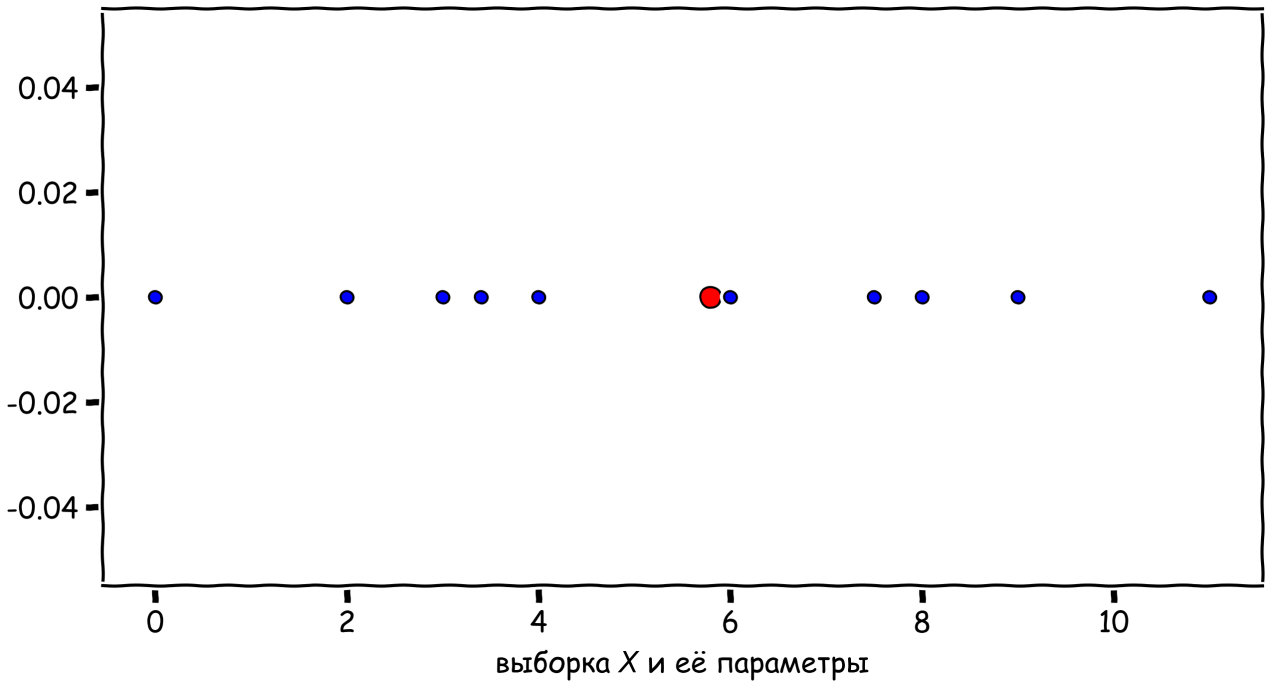

Figure 2: Sample X and its mean $\overline{X}$

You don't have to calculate sample means manually. Excel offers the AVERAGE() function that returns the average of numbers provided as arguments.

Due to similar reasons, the variance of empirical random variable ξ^* should well approximate true variance $\mathsf{D}\xi$. According to the definition of variance and taking into account that $\mathsf{E}\xi^* = \overline{X}$, we get

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

The variance of the empirical random variable is considered a special case.

Definition 1.3.2 *Assume that we have sample $X = (X_1, X_2, ..., X_n)$. The value equal to variance $\mathsf{D}\xi^*$ of empirical random variable ξ^* constructed for sample X is called sample variance. It is denoted by S^2. To put it differently,*

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

Let's take a closer look at sample variance S^2. Since distribution ξ^* approximates true distribution ξ, then variance $\mathsf{D}\xi^*$ (being sample variance S^2) approximates true distribution $\mathsf{D}\xi$. The meaning of the latter is simple. It's a mean square of the spread of the values of random variable ξ of its expected value $\mathsf{E}\xi$.

Let's consider the same sample

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

We can estimate how different Donna's arrivals are. It is easy to understand that $S^2 \approx 9.625$. The sample elements are represented by the blue points. The red point is the sample mean. The red line between the green points shows the interval.

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

What is S? It's the value of the mean spread of values of the empirical random variable with respect to the mean. We can consider it as the estimator of the standard deviation σ_ξ of random variable ξ.

In our case, $S \approx 3.1$, therefore, the spread (being approximately equal to 5.8 with respect to the mean) is quite large. What conclusion can be made? Well, for Pete, it is better not to be late for more than two minutes because Donna's arrivals differ in time, and Pete can come later than Donna. On the other side, if Donna is late for more than 9 minutes, Pete can start to worry about her.

Remark 1.3.1 *Note that statistics often use not only sample variance S^2 introduced earlier but also unbiased sample variance S_0^2 that is defined by sample*

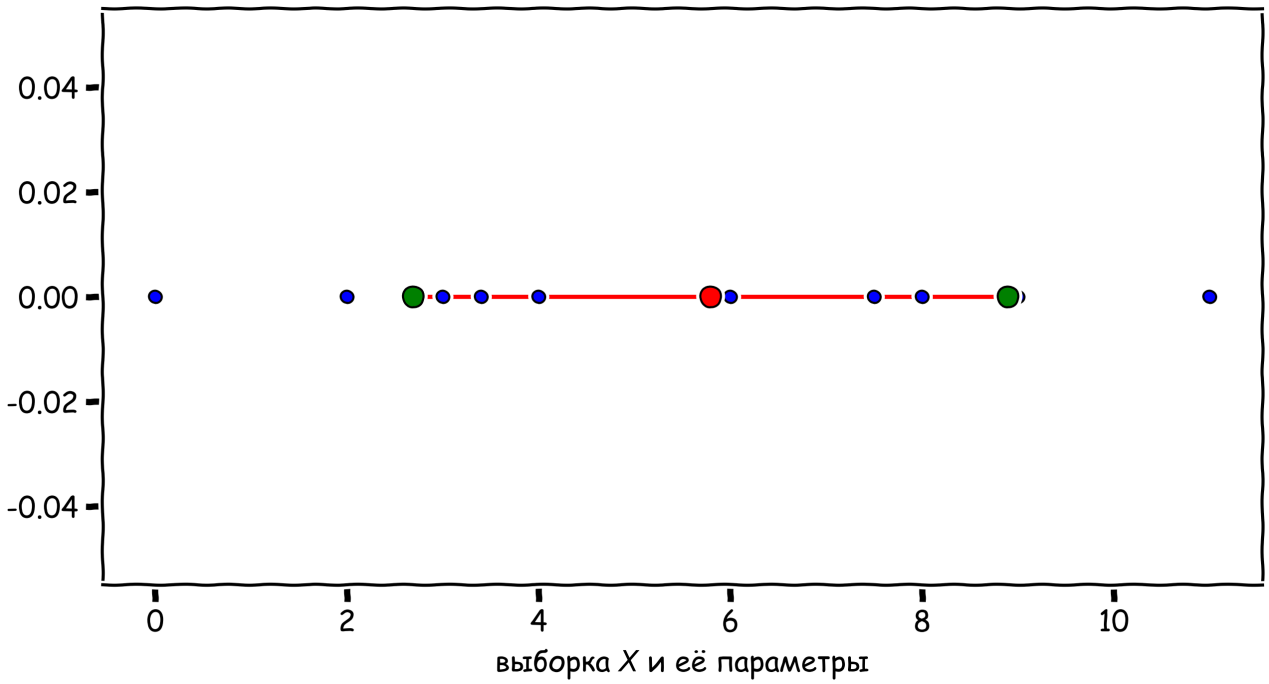

Figure 3: Sample X, its sample means $\overline{X}$, and S

$X = (X_1, X_2, ..., X_n)$ *as*

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

Unbiased sample variance S_0^2 *is algebraically different from sample variance* S^2 *in factor before the sum:* $\frac{1}{n}$ *changes to* $\frac{1}{n-1}$. *The observation shows that*

$$S_0^2 = \frac{n}{n-1} S^2$$

and, given $n \to +\infty$

$$\lim_{n \to +\infty} \frac{n}{n-1} = 1,$$

the samples S_0^2 *and* S^2 *are almost the same and can be equally used in case of large volumes. When the sample volume is small,* S_0^2 *is preferred because it is more precise. We will explain why it is so and what unbiasedness means a little bit later.*

In our example, $S_0^2 \approx 10.312$, and this value is very different from S^2 (almost by 0.8).

Based on the arguments, according to which estimator S_0^2 can be considered a reasonable estimator of the variance of population ξ given large n, value S_0 can also be considered a reasonable estimator of standard deviation σ_ξ. In our case, $S_0 \approx 3.21$.

Example 1.3.1 *Now we can move to another example showing the difference between* S^2 *and* S_0^2 *on synthetic samples from the population having the known variance equal to* 9.

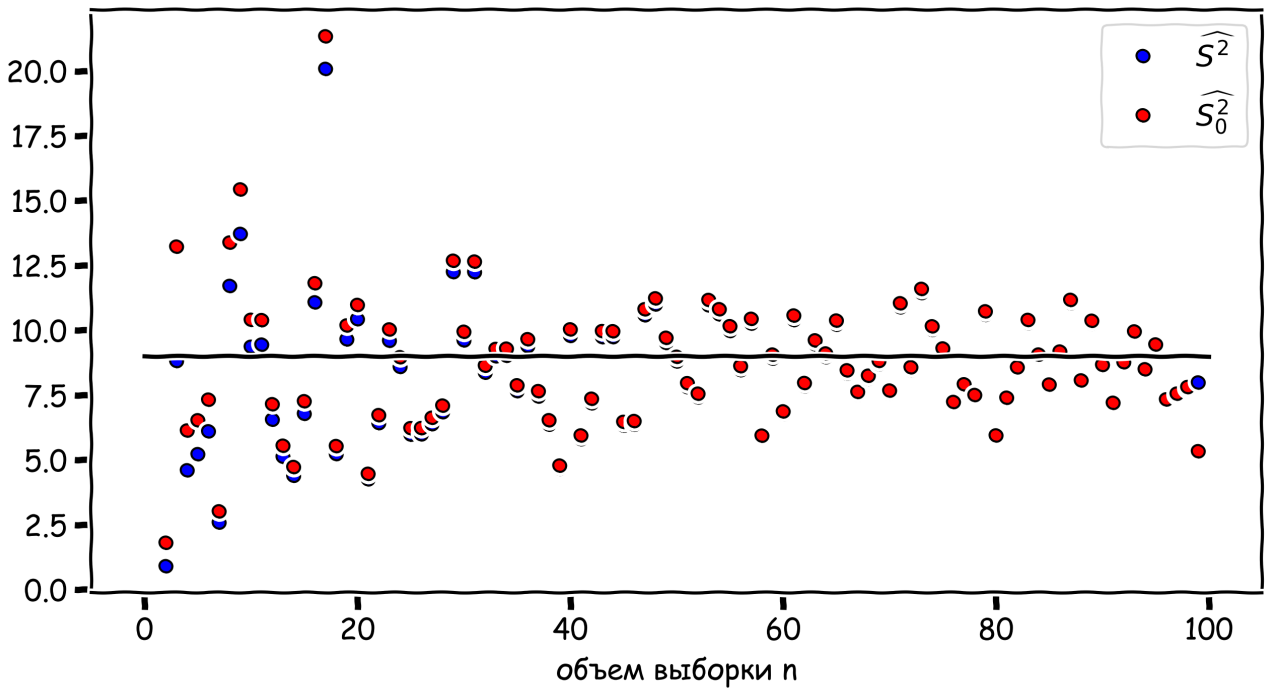

Figure 4: The difference between S^2 and S_0^2

When the sample volume is small, unbiased variance S_0^2 *is on average closer to the true variance value. However, when the number of sample elements is larger, the differences between* S^2 *and* S_0^2 *are becoming smaller, and the points basically merge.*

Like many other characteristics, sample variance is built in most data analysis packages. Excel provides the VAR.P or VARPA function. VARPA assumes that its arguments are the numeric data used to calculate sample variance S^2. Note that similar functions VAR.S and VARA calculate unbiased sample variance S_0^2.

1.4 Median. Sample Median.

We will leave statistics for a while and immerse ourselves in the apparatus of probability theory. In particular, we will study a probability characteristic of a random variable called the median. Let's start with the strict definition of a new concept, then expand its meaning and compare it to the expected value.

Definition 1.4.1 *Number* $\mathsf{med}\ \xi$ *is called a median of random variable* ξ *if*

$$\mathsf{P}(\xi \le \mathsf{med}\ \xi) \ge \frac{1}{2} \text{ and } \mathsf{P}(\xi \ge \mathsf{med}\ \xi) \ge \frac{1}{2}.$$

The median is such a number that a random variable is not greater or less than it with probability $\frac{1}{2}$. It's similar to the expected value, isn't it? Well, it's not that simple. For example, a median always exists, but it is not always unique. Why? Let's think of it and give an example.

Example 1.4.1 *The experiment is tossing a fair coin. The distribution of random variable ξ, which will be one if we have heads and zero if tails, is given by the following table:*

ξ	0	1
P	$\frac{1}{2}$	$\frac{1}{2}$

Any number $\mathsf{med}\ \xi$ *within the range* $[0, 1]$ *serves as a median of random variable* ξ *because*

$$\mathsf{P}(\xi \le \mathsf{med}\ \xi) \ge \frac{1}{2} \text{ and } \mathsf{P}(\xi \ge \mathsf{med}\ \xi) \ge \frac{1}{2}.$$

Why is it so? In this case, the outcomes are equipossible, and (if we may say so) the median does not know what to choose.

What is the expected value in this case? We can easily prove that it equals

$$\mathsf{E}\xi = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}.$$

Often, if a median is not unique (therefore, there's a line segment with such medians), the middle of the line segment is a median. In our case, $\mathsf{med}\ \xi$ coincides with the expected value and equals $\frac{1}{2}$. However, it is not always like this.

Example 1.4.2 *Let's consider a simple example. Assume that some company has* 100 *employees, and one of them is an executive. The executive's salary is 101 thousand dollars a month, and the salary of an employee is 1 thousand dollars a month. Let random variable ξ (or population) be the employee's salary. Then, the empirical distribution constructed based on the sample can be given as follows:*

ξ^*	1000	101.000
P	$\frac{99}{100}$	$\frac{1}{100}$

It is easy to understand that the expected value of random variable ξ^* equals

$$\mathsf{E}\xi^* = 1000 \cdot \frac{99}{100} + 101.000 \cdot \frac{1}{100} = 2000,$$

thus, the average salary is 2 thousand dollars a month. At the same time, median $\mathsf{med}\ \xi^*$ *equals 1 thousand dollars, and median salary equals 1 thousand dollars.*

Due to the outlier value (the salary of the boss), a more appropriate estimator of the average salary is the median and not the expected value.

Recalling the arguments provided earlier, we would like to add that median $\mathsf{med}\ \xi^*$ should well approximate true median $\mathsf{med}\ \xi$. Therefore, we will accept the following definition.

Definition 1.4.2 *Assume that we have sample $X = (X_1, X_2, ..., X_n)$. Sample median $\widehat{\mathsf{med}\ \xi}$ of population ξ is the median of the sample random variable, that is,*

$$\widehat{\mathsf{med}\ \xi} = \mathsf{med}\ \xi^*.$$

Given sample $X = (X_1, X_2, ..., X_n)$ and a variational series based on it,

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

the sample median can be easily found from the ratios:

$$\widehat{\mathsf{med}\ \xi} = \begin{cases} X_{([n/2]+1)}, & n \text{ odd}, \\ \frac{X_{([n/2])} + X_{([n/2]+1)}}{2}, & n \text{ even}, \end{cases}$$

where $[x]$ denote the integer part of x. For example, $[2.7] = 2$, $[4.2] = 4$.

How to find a median based on the sample (or variational series)? Let's begin with the sample median based on the variational series. It equals the central element of the variational series if it is unambiguously defined (that is, if n is odd) and half of the sum of central elements if n is even.

Let's return to the example of Donna's late arrivals. According to the variational series,

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

the sample median equals $X_{([15/2]+1)} = X_{(8)} = 6$. It means that Donna is late with the same probabilities for less than or more than 6 minutes. Recall that sample mean $\overline{X}$ was approximately equal to 5.8, so, in our example, the obtained parameters are close. Hence, the assumed distribution of population ξ is sufficiently symmetric with respect to the expected value and the median close to it.

Note that the median is resistant to outliers if we compare it to sample means. Assume that Pete has mistakenly added an extra zero to an observation and recorded 50 minutes instead of 5.

Thus, a new variational series will contain 16 elements:

$$(0, 2, 2, 3, 3.4, 4, 6, 6, 7.5, 8, 8, 8, 9, 9, 11, 50),$$

sample mean $\overline{X}$ will be approximately equal to $\overline{X} = 8.566$ and differ from the previous value by approximately 2.8. At the same time, a new sample median $\widehat{\mathsf{med}\ \xi}$ will be equal $\widehat{\mathsf{med}\ \xi} = 6.75$ and different from the previous value by 0.75.

Remark 1.4.1 *Outliers in data are fairly common in practice. A sensor might be misbehaving, or anomalies may occur (like Pete's mistakes). Many events may affect the experiment. All of these negatively impact predictions if they have a regular pattern. Sustainable (or robust) estimators of characteristics become very useful.*

As for calculations in practice, many data analysis packages can calculate a median. Excel has the MEDIAN function. This function assumes that its arguments are the entire sample.

1.5 Histogram as Density Estimation

Recall that our task is to learn more about population ξ. Additionally, random variable ξ can be continuous. What do we want to know about a continuous random variable? Perhaps, its density. An empirical analog of density is the so-called histogram.

Making a histogram is simple. First, we define a set of sample values. Given a variational series, this set of values will be the line segment

$$A = [X_{(1)}, X_{(n)}].$$

Next, this line segment is divided into several disjoint line segments, intervals, half-intervals. Then, we count the number of values that fall into such an interval.

Let $A_1, ..., A_k$ be these line segments of the same length l. Let ν_j be the number of sample elements that fall into line segment A_j, $n = \sum\limits_{j=1}^{k} \nu_j$. We substitute the true density within interval A_j of length l for the rectangle of height $h_j = \frac{\nu_j}{nl}$. Note that the the sum of the areas of all the rectangles equals 1. Therefore, the obtained non-negative function can be interpreted as the distribution density of a random variable. Area S_j of jth rectangle equals

$$S_j = h_j \cdot l = \frac{\nu_j}{nl} \cdot l = \frac{\nu_j}{n},$$

hence, the total area of the constructed rectangles equals

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

Thus, the constructed stepwise shape formed by consolidated rectangles is called a histogram.

Example 1.5.1 *Let's take one more look at the sample that Pete has collected. Recall that the variational series is as follows:*

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

First, we divide line segment $[0, 11]$ *into* $n = 4$ *parts of equal length* ($l = 2.75$) *to obtain the sets*

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

3 sample elements fall into set A_1. *Thus,* $\nu_1 = 3$. *Similarly, we get that* $\nu_2 = 3$, $\nu_3 = 6$, $\nu_4 = 3$. *The histogram corresponding to such distribution is shown on the screen. When* $n = 5$, *the overall picture changes.*

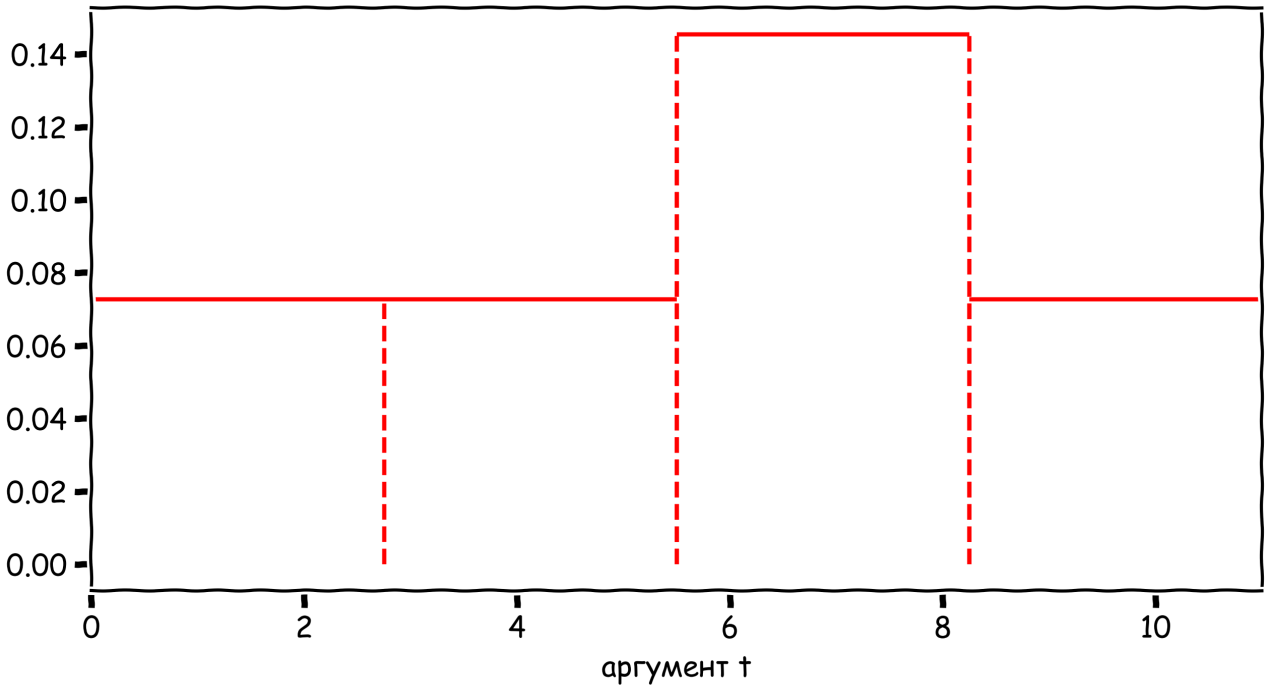

Figure 5: The histogram when $n = 4$

Remark 1.5.1 *Most data analysis packages can construct histograms. In particular, you can do it in* Excel *using a data analysis package or the FREQUENCY function.*

1.6 Multivariate Sampling. Sample Correlation

Let's assume that we observe not one but several random variables. This case is also very common. One of the most important questions, is there a relationship between the observed sample values? In probability theory, covariance and correlation coefficient indicate whether variables are related. Let's determine their sample counterparts, but, first, we will introduce the concept of multivariate sampling.

Well, let the population be a random vector $\vec{\xi} = (\xi_1, \xi_2)$ consisting of two components (for simplicity). Thus, as in the univariate case, we can reasonably introduce the concept.

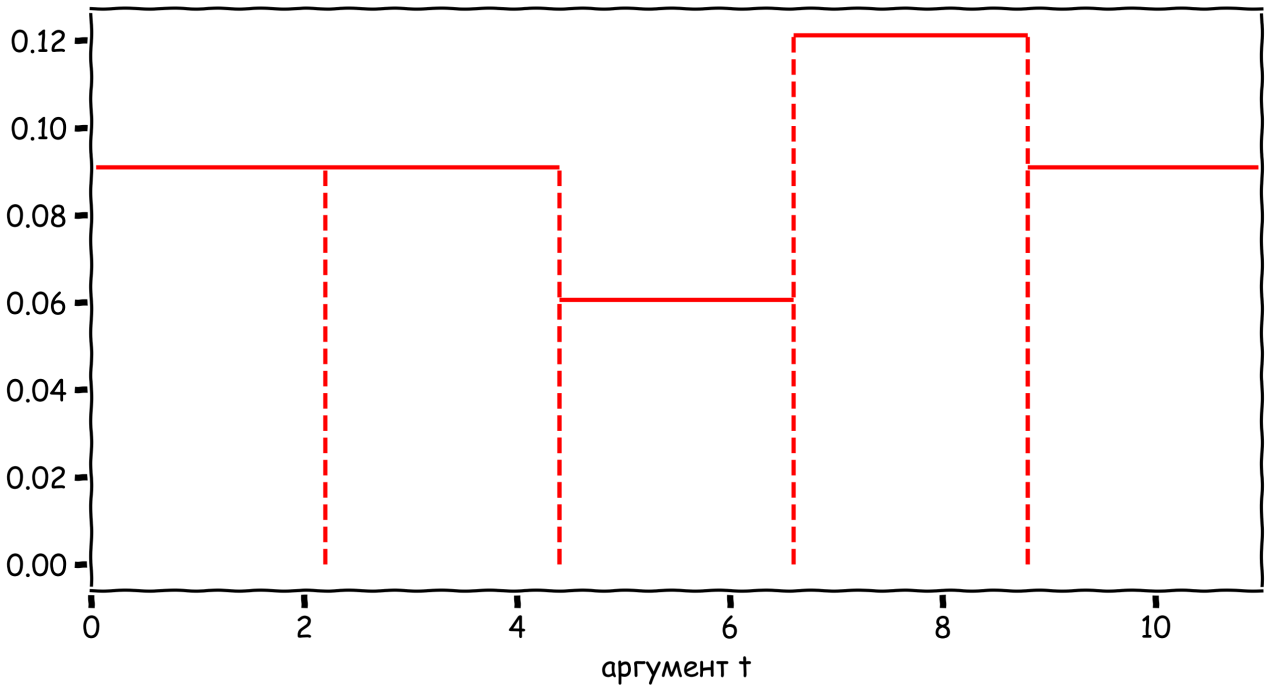

Figure 6: The histogram when $n = 5$

Definition 1.6.1 *Bivariate sample* $(X, Y) = ((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ *of size* n *is a set of* n *independent identically distributed pairs of random variables* (X_i, Y_i)*, each of which has the same joint distribution as pair* $\vec{\xi} = (\xi_1, \xi_2)$*.*

Recall that, since the covariance of two random variables ξ_1 and ξ_2 is defined as

$$\mathsf{cov}(\xi_1, \xi_2) = \mathsf{E}(\xi_1 - \mathsf{E}\xi_1)(\xi_2 - \mathsf{E}\xi_2) = \mathsf{E}(\xi_1\xi_2) - \mathsf{E}\xi_1\mathsf{E}\xi_2,$$

it is logical to accept the following definition.

Definition 1.6.2 *The sample covariance constructed for sample* $(X, Y) = ((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ *is*

$$k(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

We often consider the unbiased sample covariance, which is defined by the ratio

$$k_0(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

The justification of the unbiased sample covariance remains the same as that of the unbiased sample variance. If n are high, they behave the same. If n are low, the error of the unbiased estimator is less. We will discuss it further in detail.

Excel offers two functions to calculate the covariance. COVARIANCE.S is used for unbiased covariance, and COVARIANCE.P is for the biased one.

For considering the relationship, it is useful to obtain the so-called sample correlation. Since the correlation between ξ_1 and ξ_2 is defined as

$$\rho(\xi_1, \xi_2) = \frac{\mathsf{cov}(\xi_1, \xi_2)}{\sigma_{\xi_1}\sigma_{\xi_2}},$$

the sample correlation is defined as follows.

Definition 1.6.3 *Sample correlation based on sample* $(X, Y) = ((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ *is the value*

$$r(X, Y) = \frac{k(X, Y)}{S_X S_Y} = \frac{k_0(X, Y)}{S_{0X} S_{0Y}},$$

where S_X^2, S_Y^2 *are biased sample variances, and* S_{0X}^2, S_{0Y}^2 *are unbiased sample variances based on samples* $X = (X_1, X_2, ..., X_n)$ *and* $Y = (Y_1, Y_2, ..., Y_n)$ *respectively.*

Example 1.6.1 *The average monthly salary (in thousands of rubles) in the Leningrad Oblast in 2010-2011 is shown on the screen.*

	Utilities	*Healthcare*	*Education*	*IT*
2010	13.1	9.2	10.2	23.6
2011	16.2	11.6	13	28.8

.

Let's find the unbiased covariance using Excel. *We will use the COVARIANCE.S function.*

$$k_0(X, Y) \approx 51.6.$$

We will calculate the correlation coefficient using the CORREL function.

$$r(X, Y) \approx 0.99$$

2 Estimating the Parameters of Probability Model

2.1 Suggestive Example

The case we considered (when we have only the sample) is probably the worst but not the only one possible. Let's consider a model example.

Assume that a gun-range frequenter makes 10 shots from the same gun. Here's the sample, where 1 corresponds to hitting the target, and 0 to missing:

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

Of course, we can use only this available sample, but it is only one of the options. Each trial (out of 10 in total) has two possible outcomes (hitting or missing). The probability of a favorable outcome in each trial is the same and equal to p (since a gun-range frequenter should be good at shooting). We can assume that the random variable showing the result of the trial has Bernoulli distribution B_p with an unknown parameter p.

What use will we make of it? Let's see. If we can evaluate p using the given sample, we will obtain a probabilistic model of the experiment, which is a lot as we said earlier. But how to estimate p? It turns out that we have already prepared everything for it.

Let $\xi \sim \mathsf{B}_p$, then, $\mathsf{E}\xi = p$. A good estimator of expected value $\mathsf{E}\xi$ is a sample mean $\overline{X}$. Therefore, it can be considered an estimator $\widehat{p}$ of parameter p. Thus,

$$\widehat{p} = \overline{X} = 0.5$$

Now we can make predictions.

2.2 Parameter Estimators of Some Standard Distributions

So, let's briefly introduce the parameter estimators for some standard distributions. To highlight that the value calculated based on the sample is an estimator, we will use the circumflex sign, which is often called hat.

1. Let's consider Bernoulli distribution B_p. Since expected value $\mathsf{E}\xi$ of random variable ξ having a Bernoulli distribution equals p,

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

Let's conduct a numerical experiment based on synthetic samples of different sizes from the Bernoulli distribution with parameter $p = 0.6$. When n increases, the sample mean better approximates the true value of the parameter $p = 0.6$

2. Let's consider binomial distribution $\mathsf{Bin}(m, p)$ with unknown parameters m, p. We know that expected value $\mathsf{E}\xi$ of a random variable having a binomial distribution equals mp, and variance $\mathsf{D}\xi = mp(1-p)$. Then, taking into account that an expected value estimator is $\overline{X}$, and variance estimator is S^2, we solve a system of equations,

$$\begin{cases} mp = \overline{X} \\ mp(1-p) = S^2 \end{cases}$$

we obtain estimators

$$\widehat{p} = 1 - \frac{S^2}{\overline{X}}, \quad \widehat{m} = \frac{\overline{X}^2}{\overline{X} - S^2}.$$

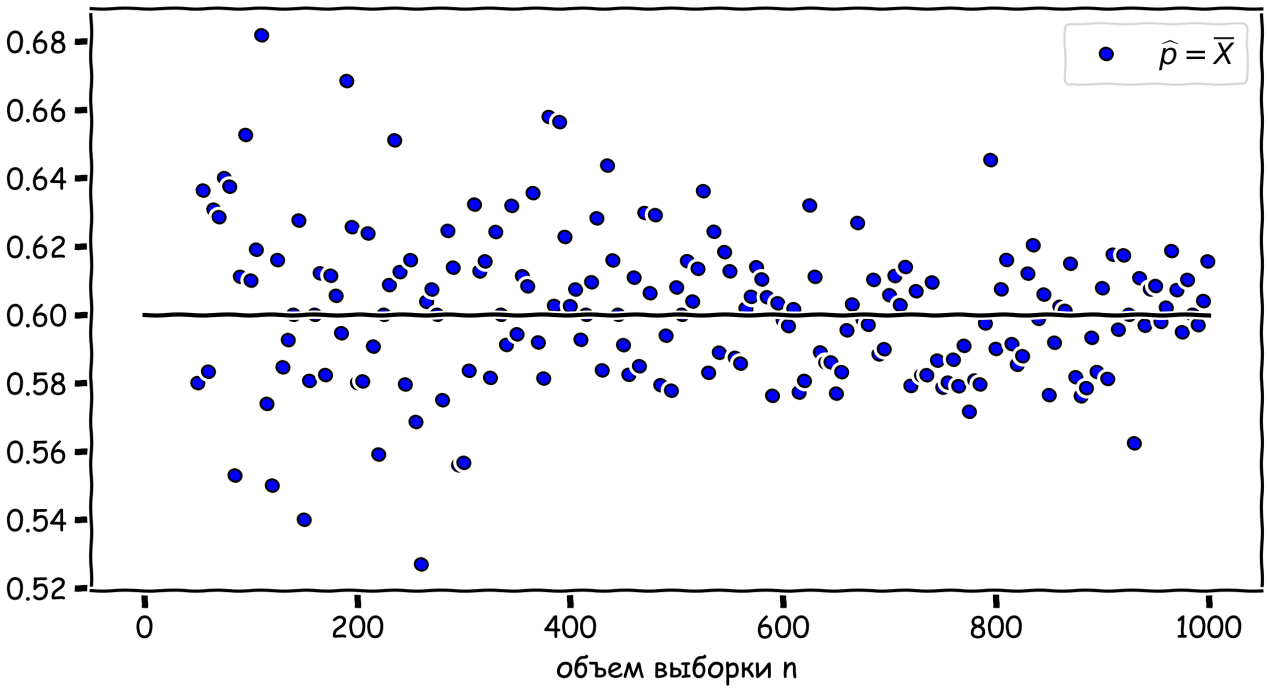

Figure 7: Relationship between $\overline{X}$ and sample size

If we use unbiased sample variance S_0^2 to estimate the variance,

$$\widehat{p} = 1 - \frac{S_0^2}{\overline{X}}, \quad \widehat{m} = \frac{\overline{X}^2}{\overline{X} - S_0^2}.$$

Since m is a natural number, the estimate should be the integer nearest to that obtained using the formulas that we discussed earlier.

If m is known, the formulas are simplified, and

$$\widehat{p} = \frac{\overline{X}}{m}.$$

If p is known,

$$\widehat{m} = \frac{\overline{X}}{p}$$

the nearest integer should be taken as an estimate.

The example of p estimators based on samples from a binomial distribution with parameter $p = 0.6$ is shown on the screen. If parameter m is known, parameter estimator p is more accurate than in the case when nothing is known.

3. Let's consider Poisson distribution Π_λ with unknown parameter $\lambda > 0$. Since expected value $\mathsf{E}\xi$ of a random variable having a Poisson distribution equals

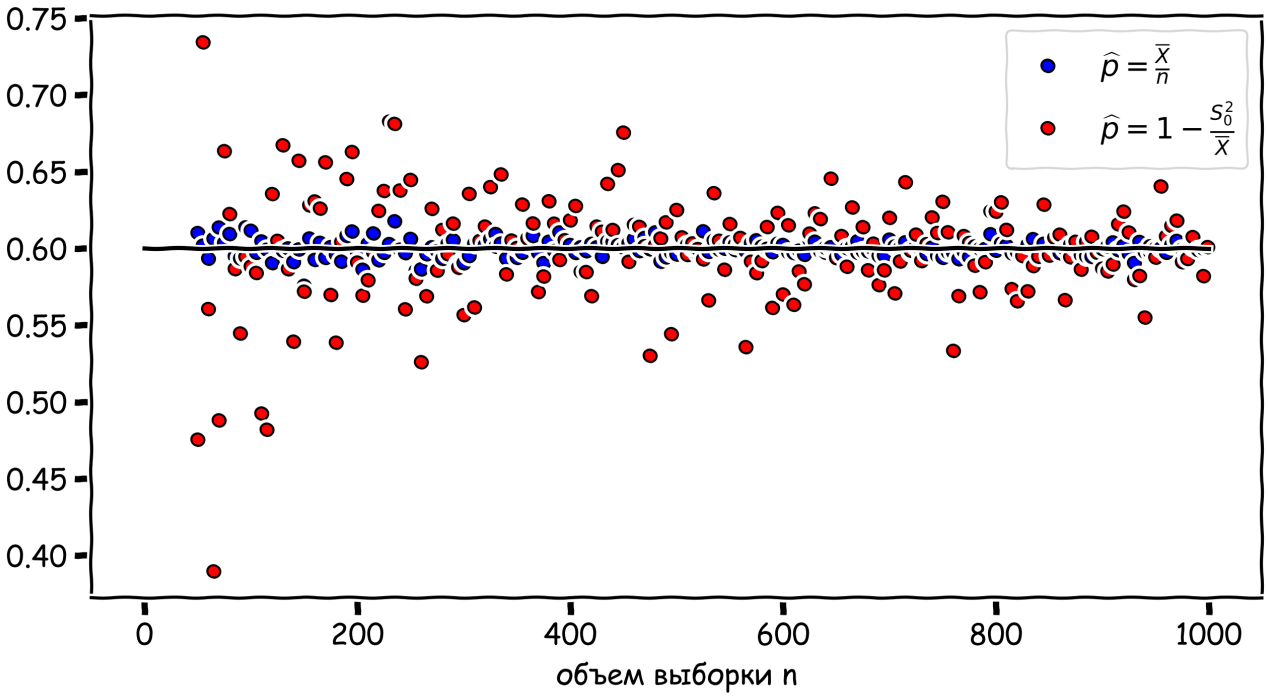

Figure 8: Relationship between $\widehat{p}$ and sample size

λ,

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

The example of λ estimators based on the samples from the Poisson distribution with parameter $\lambda = 1$ is shown on the screen.

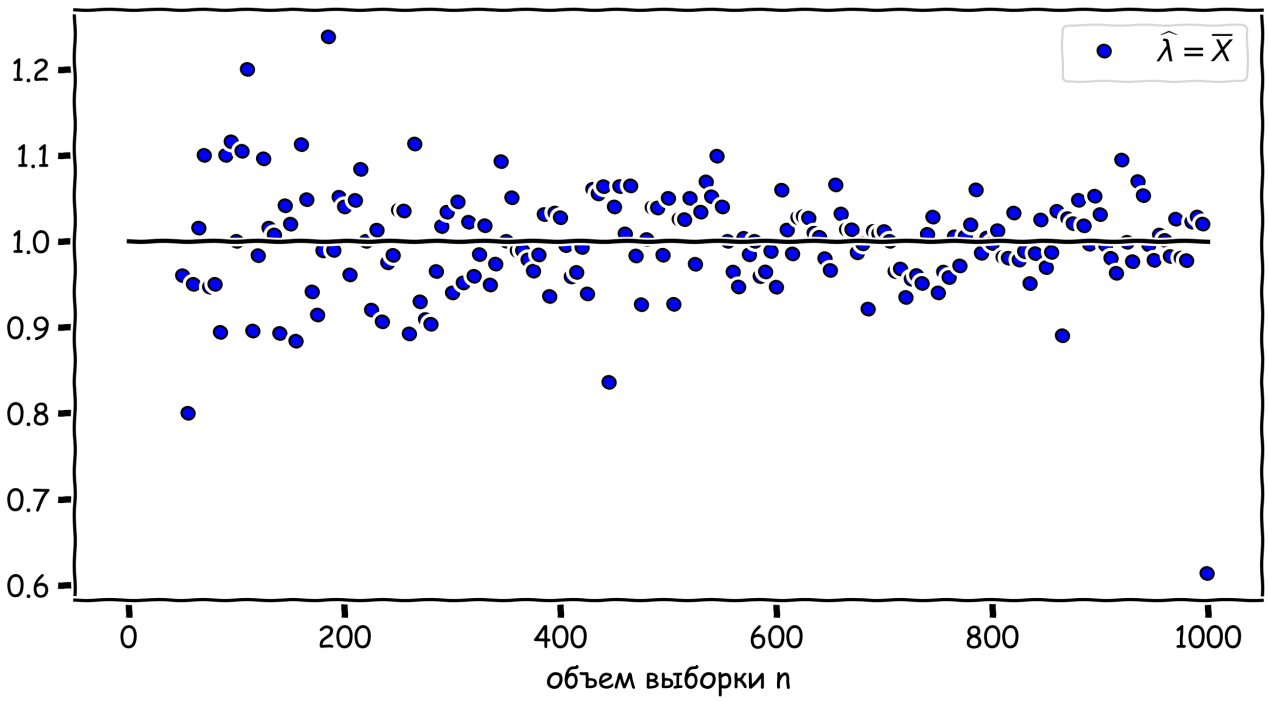

Figure 9: Relationship between $\widehat{\lambda}$ and sample size

4. Let's consider uniform distribution $\mathsf{U}_{a,b}$ with unknown parameters $a < b$. It turns out that these parameters are better estimated using 1st and nth order statistics:

$$\widehat{a} = X_{(1)}, \quad \widehat{b} = X_{(n)}.$$

The example of a, b estimators based on samples from a uniform distribution with parameters $a = 0, b = 10$ is shown on the screen. You can notice that the estimators are coping well with the task.

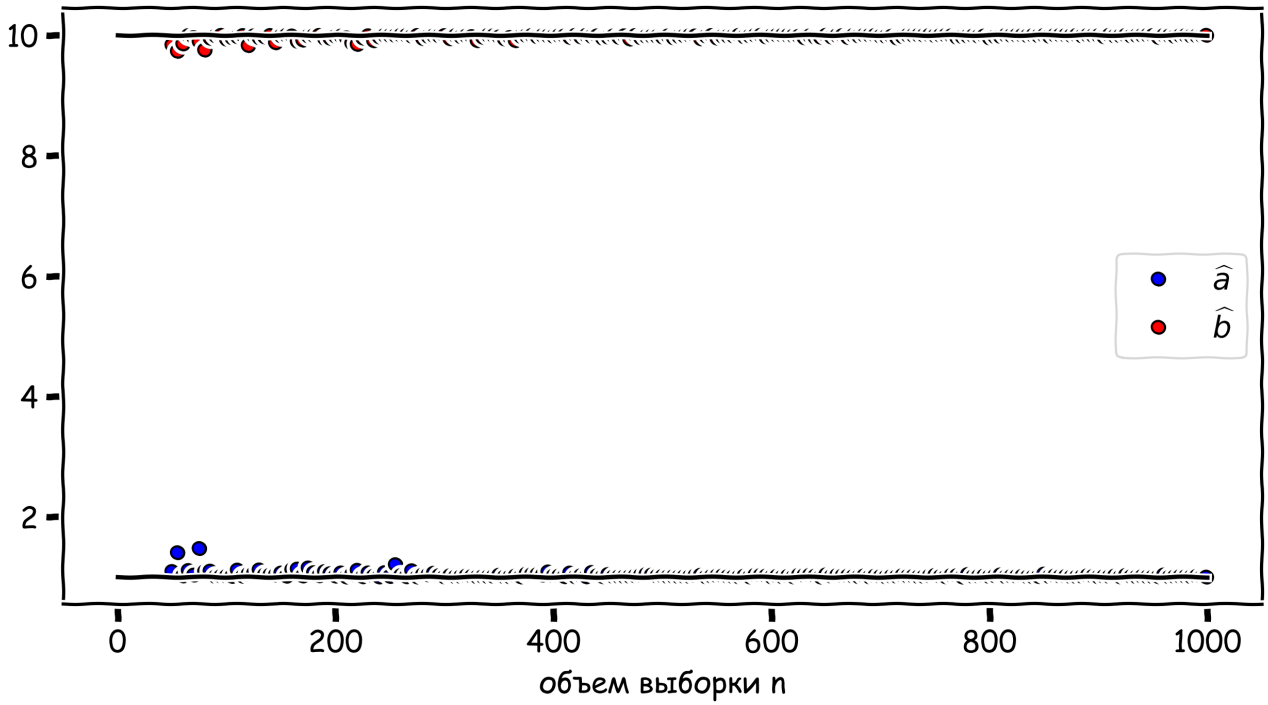

Figure 10: Relationship between $\widehat{a}, \widehat{b}$ and sample size

5. Let's consider exponential distribution Exp_λ with unknown parameter $\lambda > 0$. Since expected value $\mathsf{E}\xi$ of a random variable having an exponential distribution equals $\frac{1}{\lambda}$,

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

The example of λ estimators based on samples from a exponential distribution with parameter $\lambda = \frac{1}{3}$ is shown on the screen.

6. Let's consider normal distribution N_{a,σ^2} with unknown parameters a, σ^2. Since expected value $\mathsf{E}\xi$ of a random variable having a normal distribution equals a, and variance $\mathsf{D}\xi$ equals σ^2,

$$\widehat{a} = \overline{X}, \quad \widehat{\sigma^2} = S^2 \text{ or } \widehat{\sigma^2} = S_0^2.$$

The example of a, σ^2 estimators based on samples from a normal distribution with parameters $a = 2, \sigma^2 = 9$ is shown on the screen.

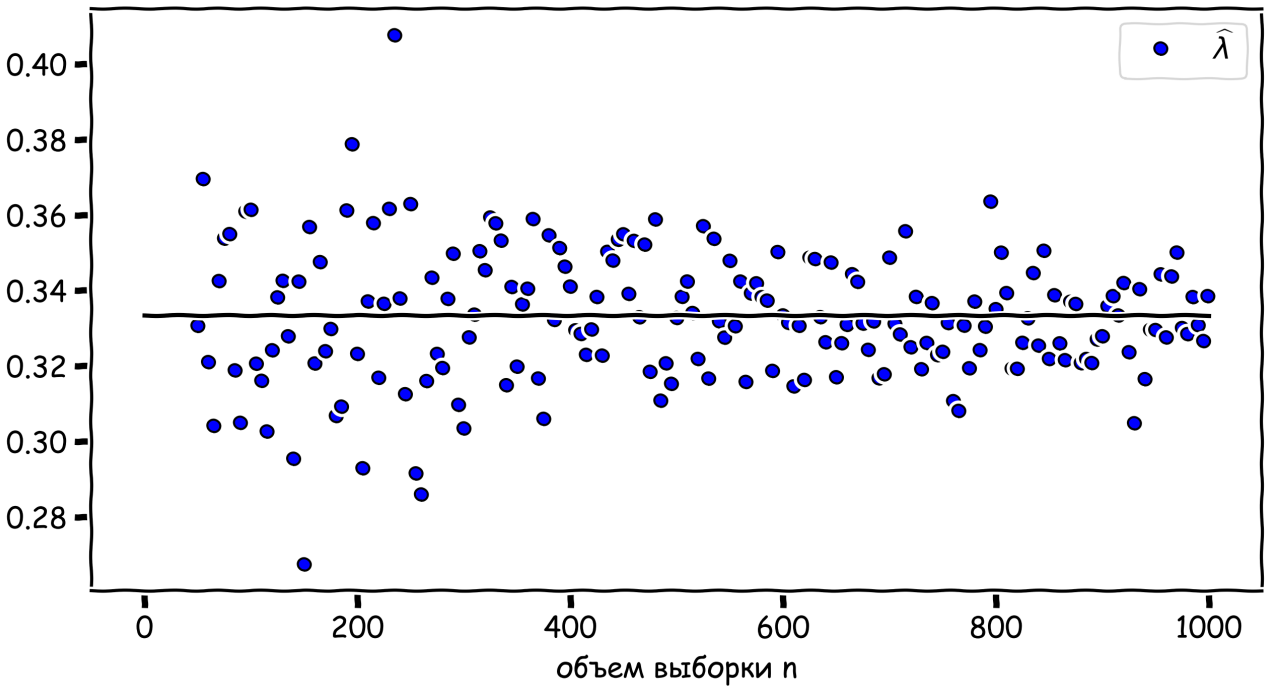

Figure 11: Relationship between $\widehat{\lambda}$ and sample size

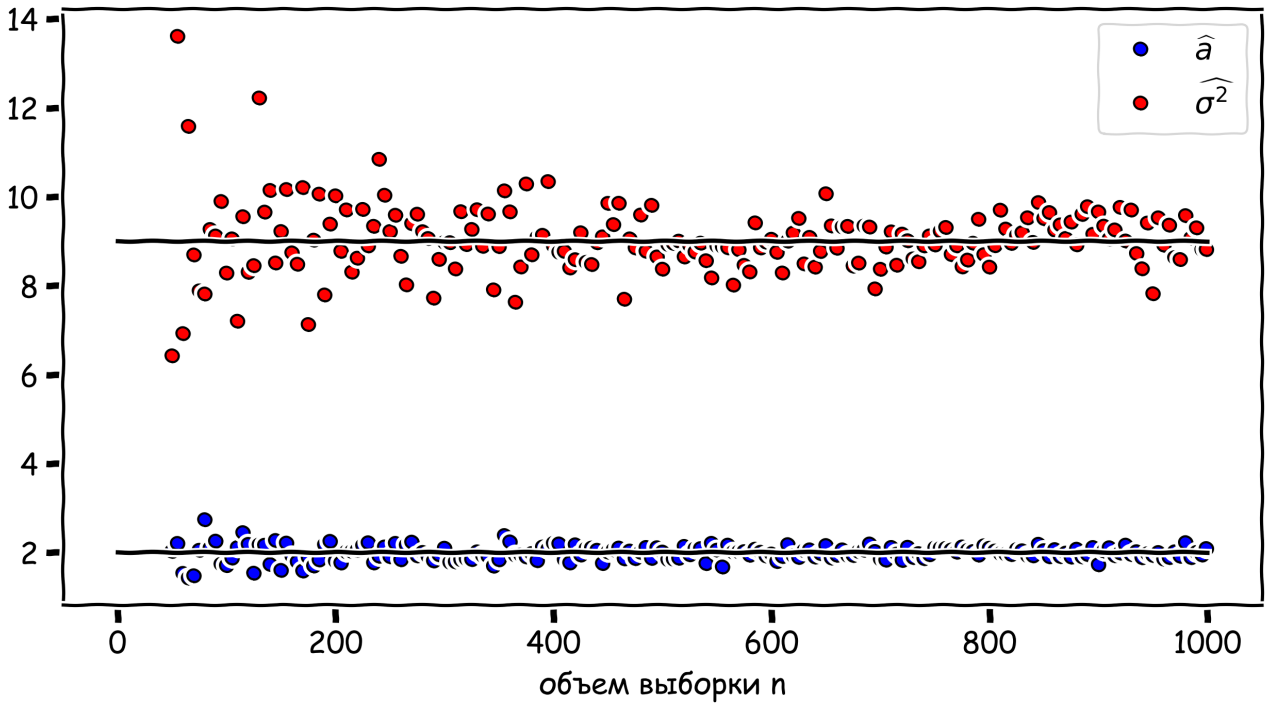

Figure 12: Relationship between $\widehat{a}, \widehat{\sigma^2}$ and sample size

The example is shown on the screen.

Data analysis packages can generate samples from known distributions having set parameters. For example, you can do this in Excel using the Data Analysis feature.

3 Law of Large Numbers and Estimator Properties

We've learned the practical details, but we haven't given any proof that everything we discussed works. We looked at the estimators from different angles, although we didn't introduce the corresponding definition. What is an estimator? When we said that a sample characteristic approximated the true one, we didn't provide explanations. How does it approximate? Let's try to figure things out!

3.1 What is an Estimator?

Let's start with the definition of an estimator. Assume that we have sample $X = (X_1, X_2, ..., X_n)$ from statistical population ξ, and θ is a parameter that describes the distribution of random variable ξ. θ can be expected value $\mathsf{E}\xi$, variance $\mathsf{D}\xi$, median $\mathsf{med}\ \xi$, and so on. θ can be a general distribution parameter. We can consider a family of distributions $\mathsf{U}_{0,\theta}$. Thus, θ is not the most obvious characteristic of the population (in fact, $\theta = 2\mathsf{E}\xi$, you can think why).

Definition 3.1.1 *An estimator of parameter θ is arbitrary function $\widehat{\theta}$ based on sample X, that is,*

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n).$$

Note that $\widehat{\theta}$ is a random variable because it is a function of random variables $(X_1, X_2, ..., X_n)$.

All estimators that we introduced earlier were the functions of the sample. Look:
1. Sample mean $\overline{X}$ is an arithmetic mean of sample elements:

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

2. Sample variance S^2 and unbiased sample variance S_0^2 are more complicated functions, but they also depend on the sample only:

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

etc.

However, not every function of the sample is a good estimator of the considered parameter. For example, we can take a 0-ary function as estimator $\widehat{\theta}$: $\widehat{\theta} \equiv 0$. This function satisfies the definition, but the question is whether it's reasonable.

A reasonable estimator is the estimator that approximates the true value of the parameter. What does it mean to approximate? To define it, we need to introduce the concept of convergence in probability.

3.2 Consistency and Convergence in Probability

Well, let's consider a sequence of random variables ξ_n.

Definition 3.2.1 *A sequence of random variables ξ_n is said to converge in probability to random variable ξ if*

$$\forall\ \varepsilon > 0\ \Rightarrow\ \mathsf{P}\left(|\xi_n - \xi| \geq \varepsilon\right) \xrightarrow[n\to+\infty]{} 0.$$

Convergence in probability is written as

$$\xi_n \xrightarrow[n\to+\infty]{\mathsf{P}} \xi.$$

How to explain this definition? The definition is not trivial, that's why it looks so cumbersome. We can formulate the statement given in the definition as follows. When n grows larger, the probability of the event that ξ_n will differ from ξ tends to zero.

It is reasonable to pose one more question. What all of these have to do with the sequence? Any estimator $\widehat{\theta}$ is a sequence because it depends on the sample size n:

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n).$$

Look at the sample mean $\overline{X}$:

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

In the sample mean expression, both denominator and terms in the numerator depend on n. Have you guessed what estimator will be considered a good one? Of course, the estimator that will be converging to the true value of the parameter. Such estimators are called consistent.

Definition 3.2.2 *Estimator $\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$ is a consistent estimator of parameter θ if*

$$\widehat{\theta} \xrightarrow[n\to+\infty]{\mathsf{P}} \theta.$$

Well, consistency is easy to understand. If the estimator values do not approximate the true value of the parameter when n grows larger (in other words, when the sample size increases), the estimator will be inconsistent. In the real world, people are looking for consistency. We invite you to think about the essence of this concept.

Well, OK then, but how to show that an estimator is consistent? The law of large numbers can help us with this.

3.3 Law of Large Numbers

Let's consider a random experiment and some random variable ξ observed within this experiment. Assume that the expected value of this random variable equals a.

If we repeat this experiment many times and observe ξ, we will obtain a sequence of independent identically distributed random variables $\xi_1, \xi_2, \ldots$ with expected value a. We cannot predict the exact values that random variables ξ_i will take within the experiments because these values vary on a case-by-case basis. In practice, an arithmetic mean of random variables $\xi_1, \xi_2, \ldots, \xi_n$ obtained as a result of n experiments approximates a when n grows larger. This empirical fact says that the behavior of a sum of random variables has a pattern.

In math, this phenomenon is called the law of large numbers.

Theorem 3.3.1 (Law of large numbers in the Khinchin's form) *Let us have sequence $\xi_1, \xi_2, ..., \xi_n, ...$ of independent identically distributed random variables with expected value a. Then,*

$$\forall \varepsilon > 0 \quad \mathsf{P}\left(\left|\frac{\xi_1 + \xi_2 + ... + \xi_n}{n} - a\right| \geq \varepsilon\right) \xrightarrow[n \to +\infty]{} 0.$$

The law of large numbers (LLN) states that, as for independent and identically distributed random variables, the probability that their arithmetic mean will deviate from the expected value tends to zero when n grows larger.

We can write this using the introduced notations as follows:

$$\frac{\xi_1 + \xi_2 + ... + \xi_n}{n} \xrightarrow[n \to +\infty]{\mathsf{P}} a = \mathsf{E}\xi_i = \mathsf{E}\xi_1,$$

since all random variables are identically distributed, they have the same expected value.

Statistical techniques for estimating unknown distribution parameters are based on the law of large numbers.

Theorem 3.3.2 (Sample mean consistency) *Let there be the expected value of population $\mathsf{E}\xi$. Then, $\overline{X}$ is a consistent estimator of $\mathsf{E}\xi$.*

Proof. Let $X = (X_1, X_2, ..., X_n)$ be a sample from population ξ. Then, $X_1, X_2, ...$ is a sequence of independent random variables having the same distribution as ξ. The condition says that there is expected value $\mathsf{E}\xi$. Thus, according to the law of large numbers in the Khinchin's form,

$$\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow[n \to +\infty]{\mathsf{P}} \mathsf{E}\xi.$$

□

It turns out that all other estimators that we introduced, including sample variance, unbiased sample variance, sample median, and sample correlation, are consistent estimators of the corresponding characteristics of population ξ. Therefore, they can be used.

3.4 Unbiased Estimators

All we've got left is to define the concept of estimator unbiasedness.

Definition 3.4.1 *Estimator $\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$ is an unbiased estimator of parameter θ if*

$$\mathsf{E}\widehat{\theta} = \theta.$$

In other words, unbiasedness means that, on average, the estimator coincides with the estimated parameter, or, to be more specific, its values, on average, coincide with the value of the estimated parameter.

Theorem 3.4.1 *Let there be the expected value of population $\mathsf{E}\xi$. Then, $\overline{X}$ is an unbiased estimator $\mathsf{E}\xi$.*

Proof. Let $X = (X_1, X_2, ..., X_n)$ be a sample from population ξ. Then, $X_1, X_2, ...$ is a sequence of independent random variables having the same distribution as ξ. Then, according to the property of expected value,

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

□

It turns out that unbiased sample variance, unbiased sample covariance, and correlation are unbiased estimators.

The sample variance is a biased estimator. We can show that, assuming that there is a variance of population $\mathsf{D}\xi$,

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

Unbiased estimators are rare in practice. We are often dealing with the so-called asymptotically unbiased estimators.

Definition 3.4.2 *Estimator $\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$ is an asymptotically unbiased estimator of parameter θ if*

$$\lim_{n\to+\infty}\mathsf{E}\widehat{\theta} = \theta.$$

The sample variance is obviously an asymptotically unbiased estimator of the variance of population $\mathsf{D}\xi$, since

$$\lim_{n\to+\infty} \frac{n-1}{n} = 1.$$

This also leads us to the conclusions stated earlier. Biased but asymptotically unbiased estimators work well for large samples. However, they can make significant errors for small samples. In such cases, unbiased estimators are preferable.

4 Summary

In this module, you've learned how to estimate the expected value, variance and density of the population using a sample, as well as how to calculate sample covariance and estimate distribution parameters using the obtained estimates. Moreover, we supported all empirical considerations with some theory. We reviewed the importance of consistent estimators, learned the difference between biased and unbiased estimators, as well as the principles of their application. Meanwhile, there remains one outstanding question that we will answer in the next module. How to understand that the obtained estimate is close to the true parameter value? For example, when we obtain $\overline{X} = 5$ as an estimate of the expected value, can we be sure it is enough? If we change a sample, we will get another value of $\overline{X}$. Which one is better? How far from them is the true one?

We will introduce interval estimation to answer all these questions in the next module.