АКАДЕМИЯ НАУК СССР
ACADEMY OF SCIENCES USSR

ОТДЕЛЕНИЕ ТЕХНИЧЕСКИХ НАУК
ГРУППА ТЕХНИЧЕСКОЙ МЕХАНИКИ

DEPARTMENT OF TECHNICAL SCIENCES
SECTION OF TECHNICAL MECHANICS

ПРИКЛАДНАЯ МАТЕМАТИКА И МЕХАНИКА

APPLIED MATHEMATICS AND MECHANICS

Т. II, в. 4 1939 V. II, № 4

Е. Н. БУХМАН

НОМОГРАММЫ ЧИСЛА НАБЛЮДЕНИЙ, НЕОБХОДИМОГО ПРИ ОПРЕДЕЛЕНИИ ВЫБОРОЧНЫМ ПУТЕМ СРЕДНЕЙ ИЛИ ДОЛИ (ЧАСТОСТИ)

(Москва)

При организации выборочного обследования, в какой бы области оно ни проводилось, одним из основных вопросов является установление минимума наблюдений, обеспечивающего нужную точность. Для уже произведенного обследования не менее важно определить надежность полученных результатов.

Теория выборочного метода, опираясь на теорию вероятностей, дает вполне определенное решение обеих этих задач. Для практического использования важно представить это решение в такой форме, которая позволила бы просто и быстро получать нужный ответ. В этом отношении номограммы, подобные приводимым здесь, достаточно удобны.

Определяемая выборочным обследованием характеристика обычно является или средней или долей (частостью). Исходные данные расчета: а) требуемая *степень точности результата*, определяемая величиной наибольшей допускаемой ошибки (удобнее всего ее выражать в процентах к определяемой характеристике); б) *дисперсия данного вариационного ряда* (с ростом которой растет и необходимое число наблюдений). К этим двум величинам присоединяется иногда еще и третья: в) *объем генеральной совокупности* — в том случае когда объем этот не очень велик по сравнению с объемом выборочной совокупности.

К исходным данным, вообще говоря, относится и вероятность того, что фактическая ошибка не превзойдет допускаемой величины. Однако эту последнюю характеристику мы будем считать заранее фиксированной числом 0.99.

1. Расчет числа наблюдений или величины ошибки в случае, когда выборка охватывает малую долю генеральной совокупности

Основанием расчета здесь является теорема Ляпунова. Она позволяет установить, что с вероятностью

$$P_t \underset{N\to\infty}{\longrightarrow} \sqrt{\frac{2}{\pi}} \int_0^t e^{-\frac{x^2}{2}} dx \tag{1}$$

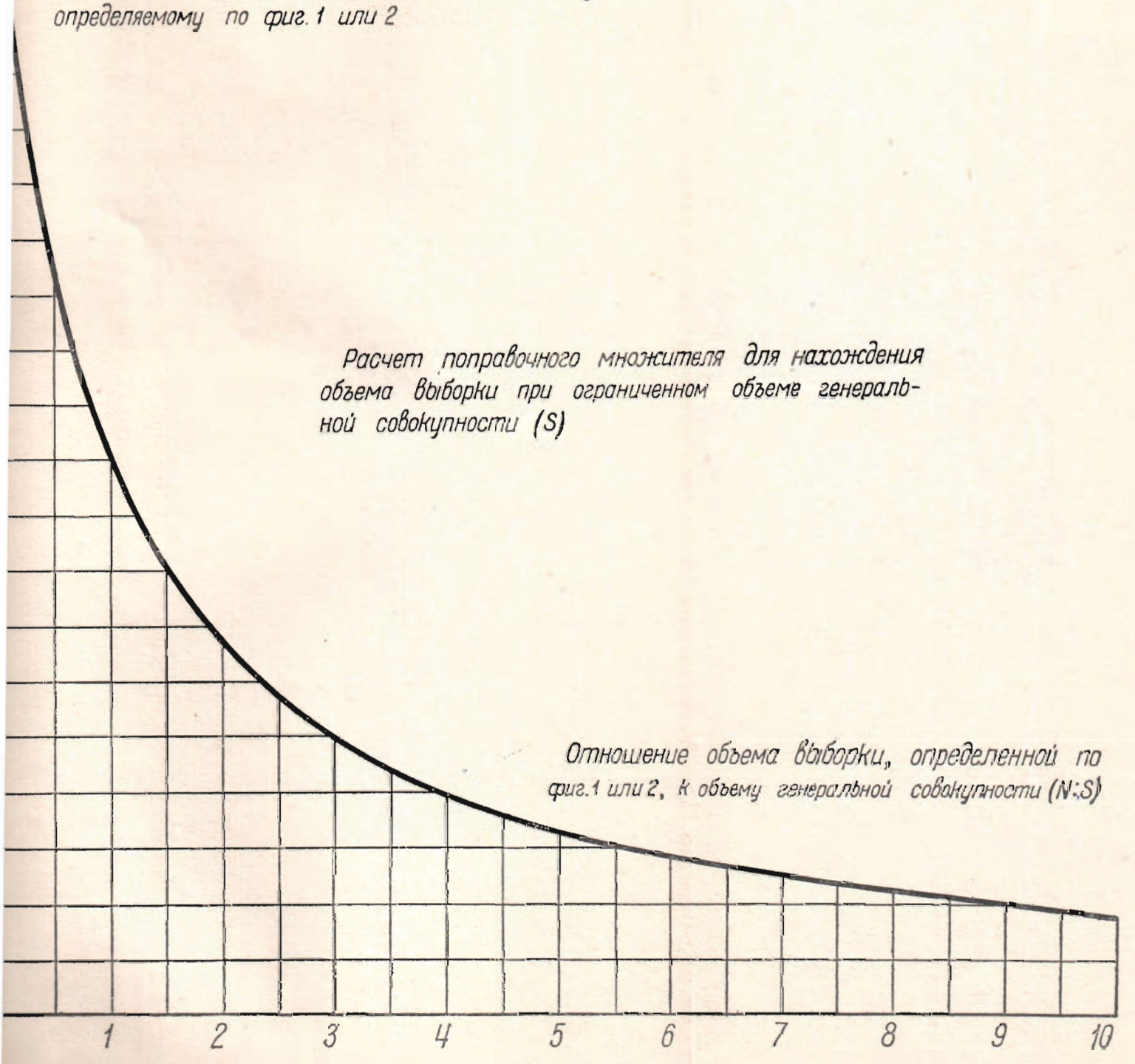
выборочная средняя (доля) будет отличаться от генеральной средней (доли) не более, чем на величину

$$|\Delta| = \frac{t\sigma}{\sqrt{N}}. \tag{2}$$

Здесь σ — среднее квадратическое отклонение в соответствующем вариационном ряду; N — число наблюдений; t — произвольный коэффициент, выбор которого определяет, с одной стороны, величину ошибки Δ, а с другой стороны, вероятность того, что фактическая ошибка не окажется большей, чем Δ. Во многих приложениях вполне достаточно принять величину этой вероятности равной 0.99; это означает практически, что в 99 из 100 (в среднем) случаях осуществления выборки ошибка окажется не больше предусмотренной величины ее. Исходя из этой величины вероятности и составлена данная номограмма, для чего принято $t = 2.6$.

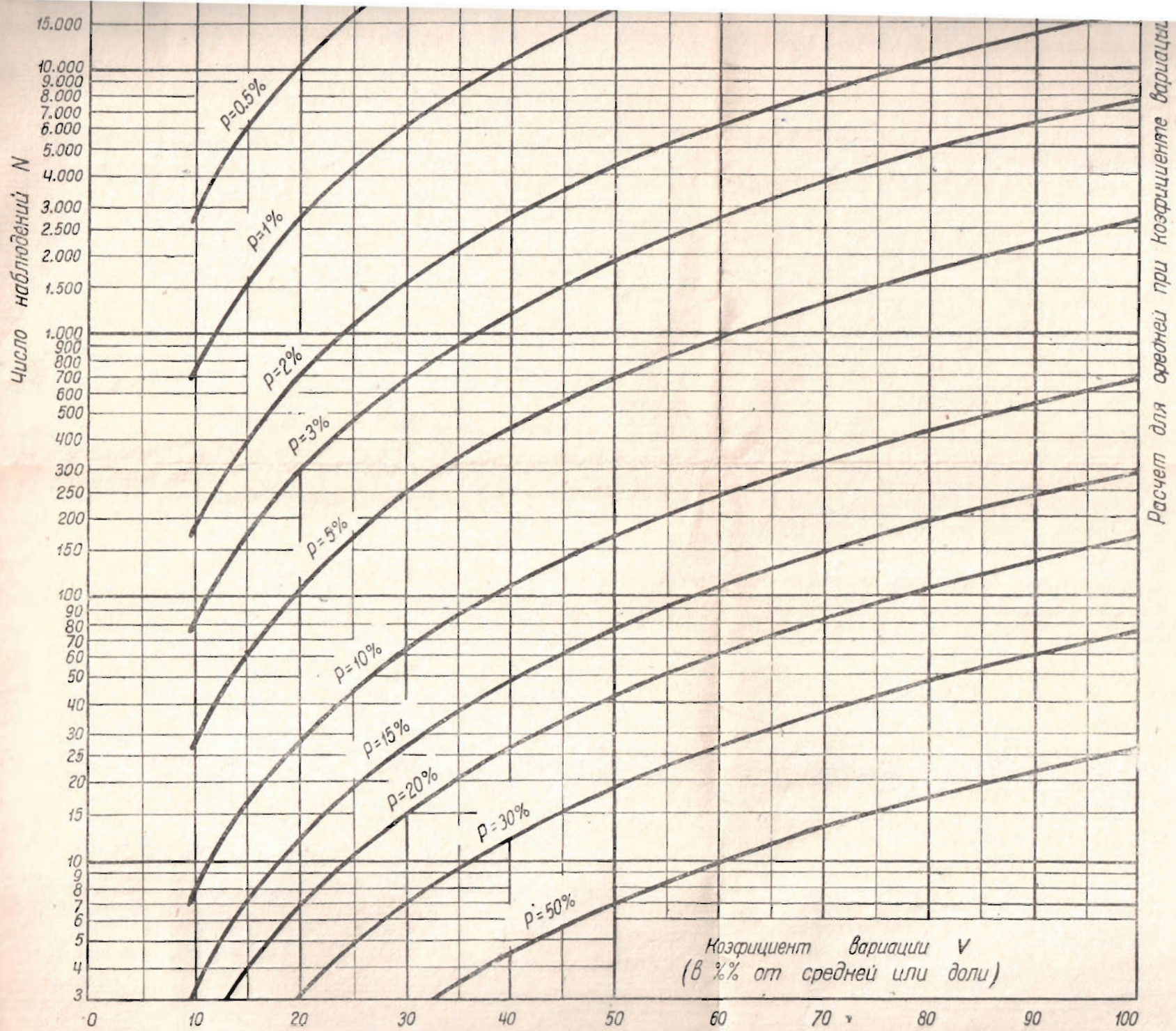
Величина поправочных множителей к числу N, определяемому по фиг. 1 или 2

Расчет поправочного множителя для нахождения объема выборки при ограниченном объеме генеральной совокупности (S)

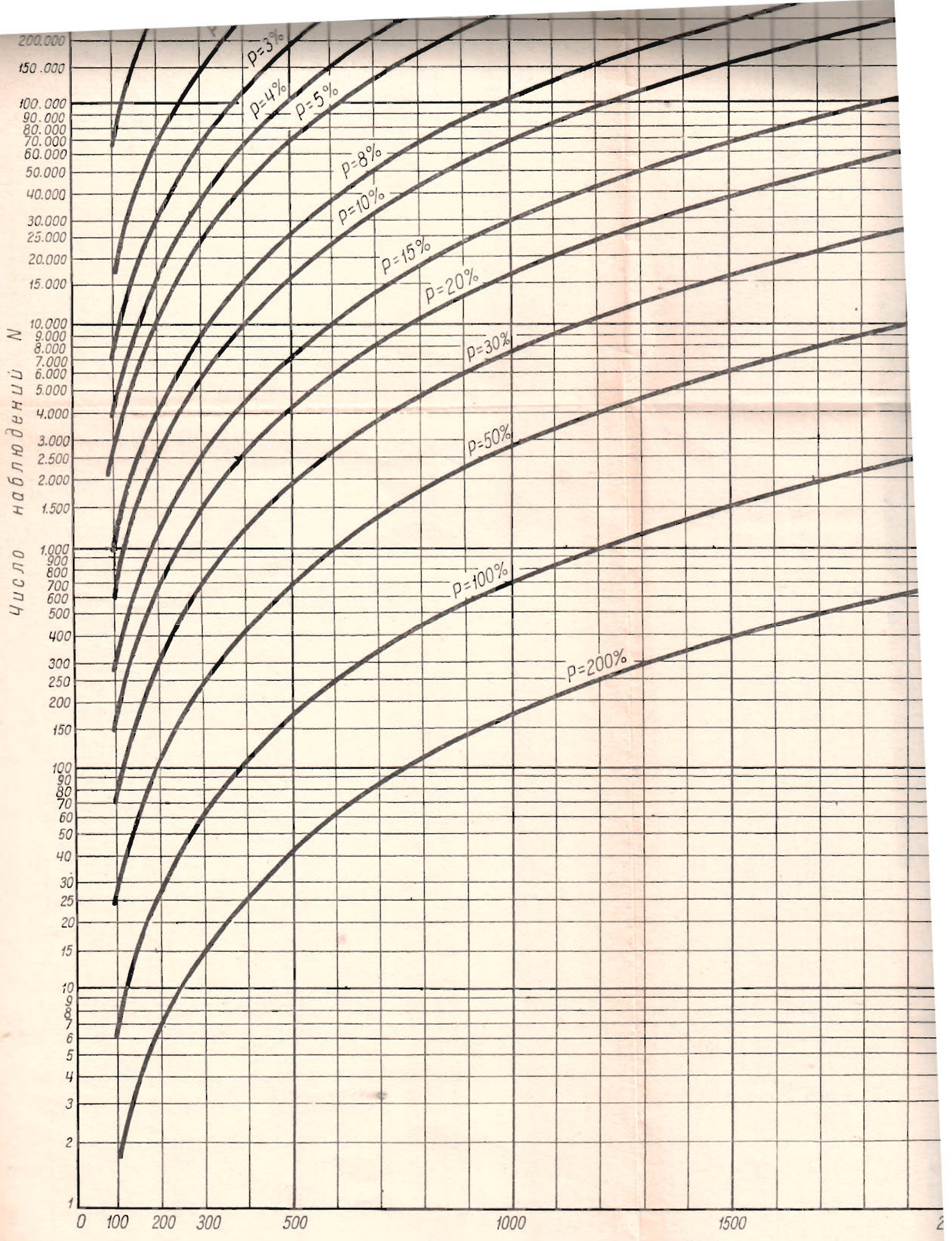


Фиг. 3.

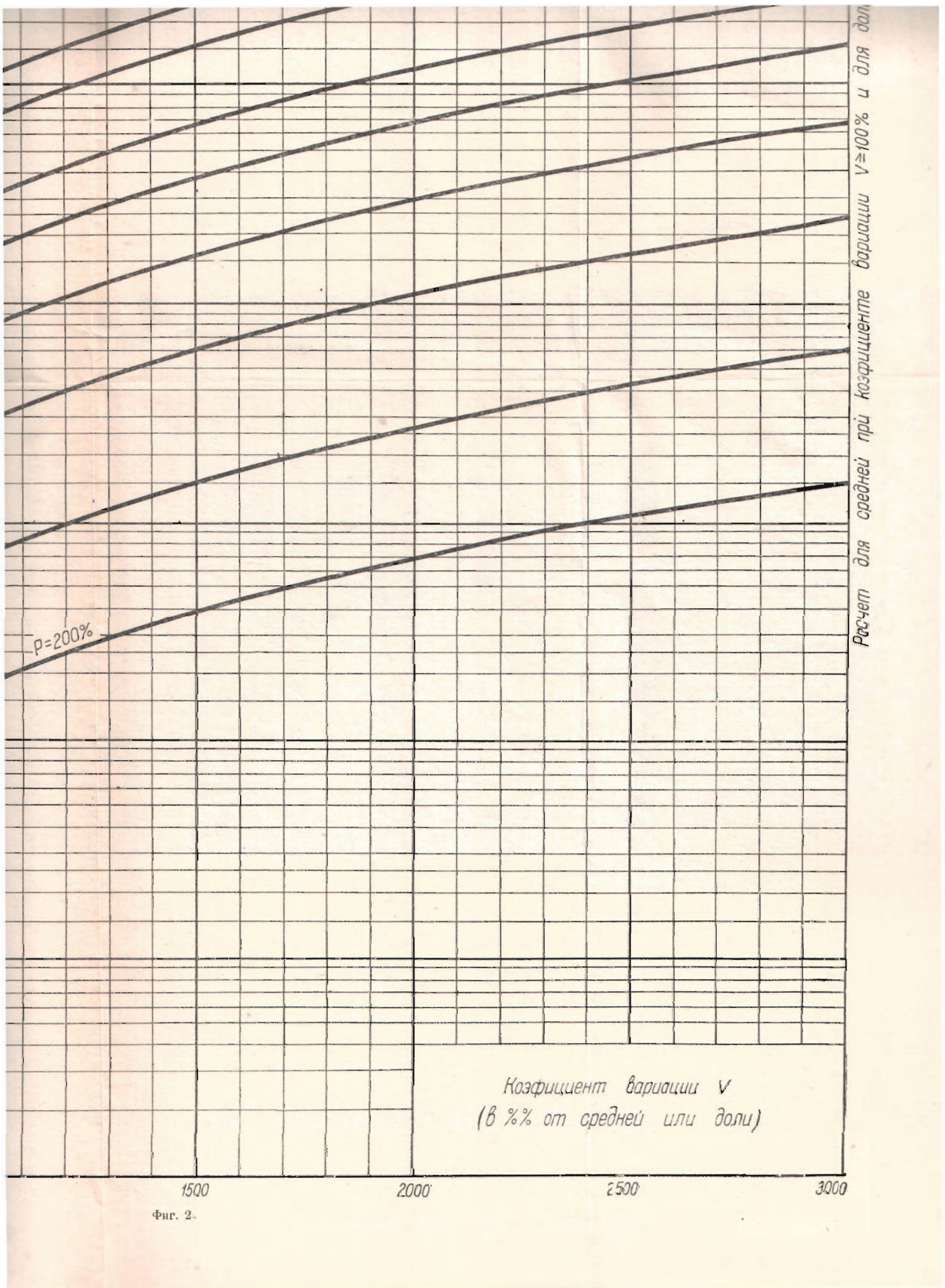
едней при коэфициенте вариации $V \leq 100\%$, для доли $p \geq 0.5$


Число наблюдений N
15.000
10.000
9.000
8.000
7.000
6.000
5.000
4.000
3.000
2.500
2.000
1.500
1.000
900
800
700
600
500
400
300
250
200
150
100
90
80
70
60
50
40
30
25
20
15
10
9
8
7
6
5
4
3
p=0.5%
p=1%
p=2%
p=3%
p=5%
p=10%
p=15%
p=20%
p=30%
p=50%
0
10
20
30
40
50
60
70
80
90
100
v
Коэфициент вариации V
(в %% от средней или доли)
Расчет для средней при коэфициенте вариации

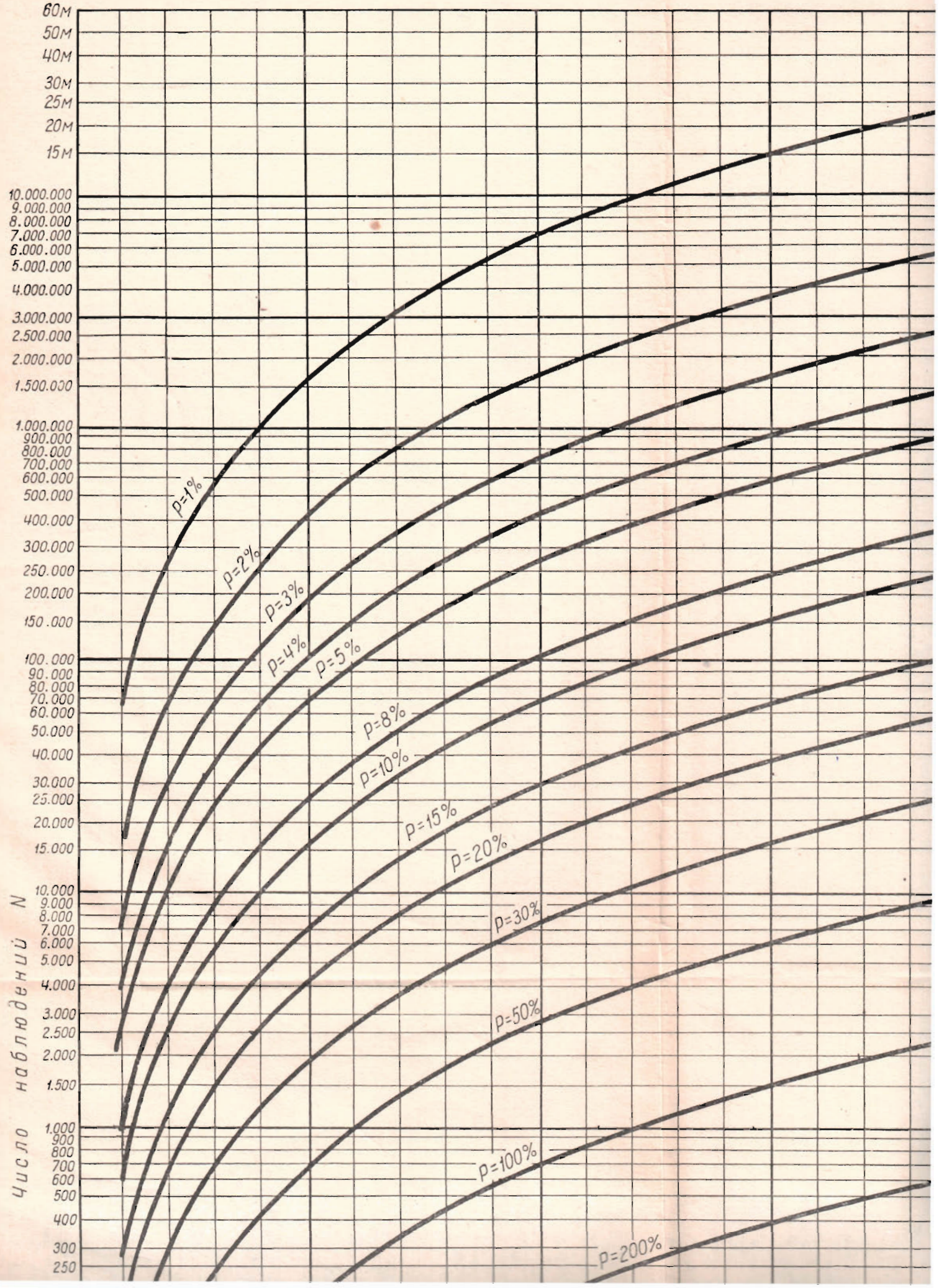
Фиг. 1

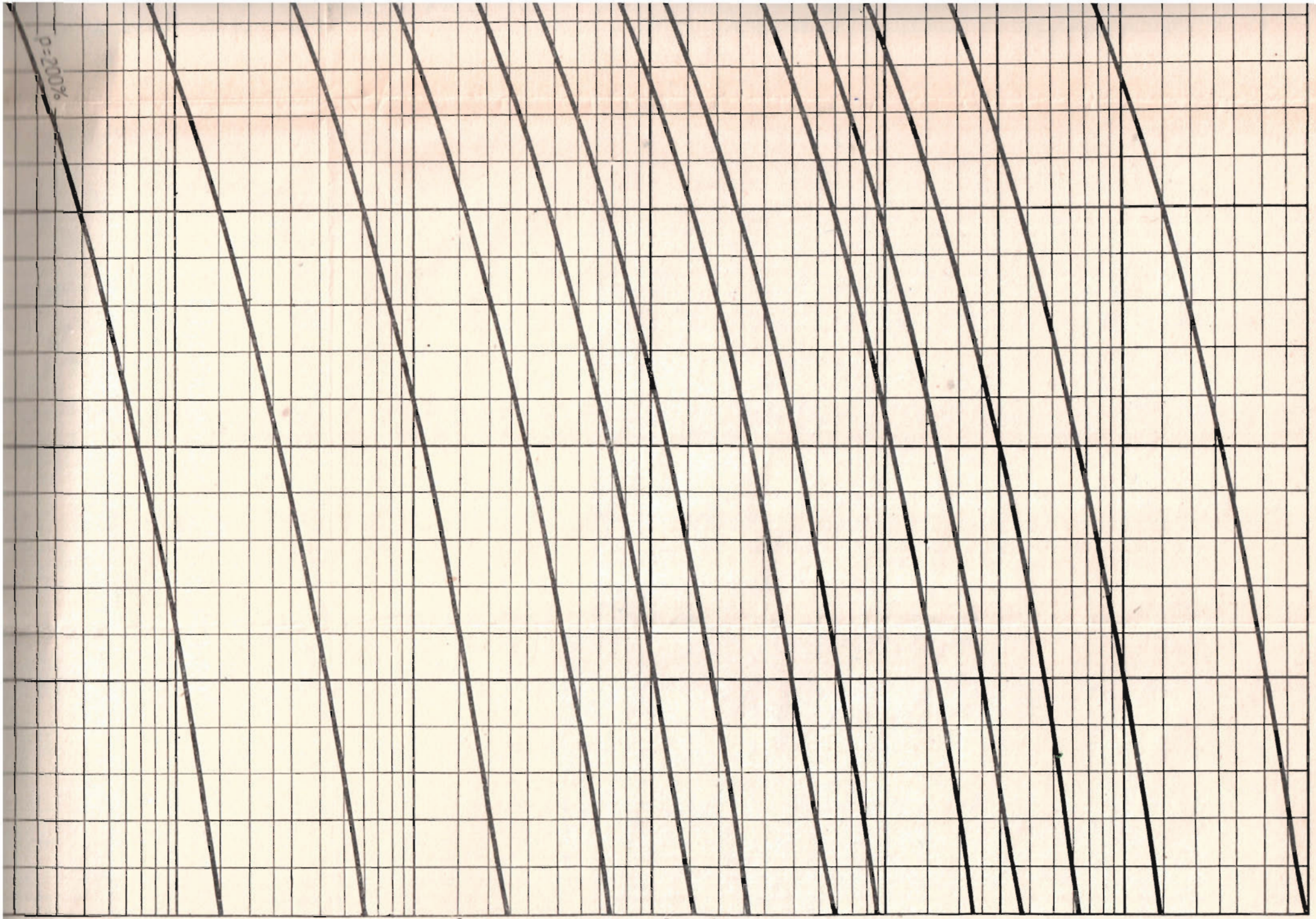

Число наблюдений N
200.000
150.000
100.000
90.000
80.000
70.000
60.000
50.000
40.000
30.000
25.000
20.000
15.000
10.000
9.000
8.000
7.000
6.000
5.000
4.000
3.000
2.500
2.000
1.500
1.000
900
800
700
600
500
400
300
250
200
150
100
90
80
70
60
50
40
30
25
20
15
10
9
8
7
6
5
4
3
2
1
0
100
200
300
500
1000
1500
P=3%
P=4%
P=5%
P=8%
P=10%
P=15%
P=20%
P=30%
P=50%
P=100%
P=200%

Фиг. 2.


Расчет для средней при коэфициенте вариации V≥100% и для доп
P=200%
1500
2000
2500
3000
Коэфициент вариации V
(в %% от средней или доли)

Фиг. 2.


60м
50м
40м
30м
25м
20м
15м
10.000.000
9.000.000
8.000.000
7.000.000
6.000.000
5.000.000
4.000.000
3.000.000
2.500.000
2.000.000
1.500.000
1.000.000
900.000
800.000
700.000
600.000
500.000
400.000
300.000
250.000
200.000
150.000
100.000
90.000
80.000
70.000
60.000
50.000
40.000
30.000
25.000
20.000
15.000
10.000
9.000
8.000
7.000
6.000
5.000
4.000
3.000
2.500
2.000
1.500
1.000
900
800
700
600
500
400
300
250
Число наблюдений N
P=1%
P=2%
P=3%
P=4%
P=5%
P=8%
P=10%
P=15%
P=20%
P=30%
P=50%
P=100%
P=200%



Расчет для средней при коэфициенте вариации $V \geq 100\%$ и для доли $p \leq 0,5$

В этом мы отступаем от достаточно распространенного так называемого „правила 3σ", рекомендующего выбирать $t=3$, в силу чего вероятность $P_t=0.997$.

Выбор P_t равным как 0.997, так и 0.99 является в одинаковой мере условным. 99 шансов из 100 составляют обычно достаточно высокую гарантию, чтобы говорить о том, что соответствующее событие осуществляется „почти наверное". Вместе с тем это более определенная характеристика, чем „997 шансов из 1000", отвечающее „правилу трех сигм". В данном случае практически важна та экономия в числе измерений, которую дает выбор $t=2.6$, вместо $t=3$. Эта экономия составляет около 25%, так как число наблюдений пропорционально t^2, и соотношение t^2 в том и другом случаях составляет $6.76:9.00=0.75$.

Непосредственно из (2) имеем, что

$$N=\frac{t^2\sigma^2}{\Delta^2}.$$

Измеряя ошибку Δ в процентном отношении к определяемой средней M (или доле), что делает расчет более универсальным, имеем:

$$N=\frac{t^2V^2}{p^2}, \tag{3}$$

где V — отношение σ/M (так называемый коэффициент вариации), обычно выражаемое в процентах, и p — допускаемая ошибка, также в процентном отношении к определяемой характеристике M.

Составленная на основе (3) номограмма состоит из двух частей: фиг. 1 — для значений коэффициента вариации V до 100% и фиг. 2 — для значений V свыше 100%, с которыми приходится иметь дело при определении доли, меньшей 0.5.

Основное условие применимости номограммы — *абсолютная случайность выборки, гарантирующая от всякой предвзятости в отборе.*

При определении доли (частости) величина коэффициента вариации V находится по формуле:

$$V=100\sqrt{\frac{1-P}{P}} \quad \text{(в процентах)}, \tag{4}$$

где P есть значение подлежащей определению доли, практически устанавливаемое приближенно на основе предшествующих обследований.

При определении средней коэффициент вариации

$$V=100\frac{\sigma}{M} \quad \text{(в процентах)}$$

определяется на основе соответствующего вариационного ряда. Если обозначим отдельные варианты через x_i и соответствующие частоты через m_i, причем i изменяется от 1 до k, то средняя M и среднее квадратическое отклонение σ, как известно, вычисляются по формулам:

$$M=\frac{1}{n}\sum_{i=1}^{k}x_i m_i \quad \text{и} \quad \sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i-M)^2 m_i},$$

где n есть объем совокупности, охваченной данным рядом, т. е.

$$n=\sum_{i=1}^{k}m_i.$$

Следует отметить важное практически упрощение техники вычисления величин M и σ применением метода сумм, возможного в случае равных интервалов вариационного ряда. В этом случае предварительно на основе исходного ряда частостей составляются два производных ряда. Первый из них получается в процессе последовательного суммирования снизу вверх частостей m_i путем записывания нарастающего итога, получающегося по прибавлении каждого следующего члена, как это выполнено для ряда Σ_1 (см. табл. 1). Второй ряд Σ_2 получается посредством подобной же операции, проводимой с рядом Σ_1. Обозначив итоги рядов m_i, Σ_1 и Σ_2 соответственно через s_0, s_1 и s_2, можем воспользоваться для определения M и σ следующими формулами:

Таблица 1

Варианты x_i	Частоты m_i	Производные ряды	
		Σ_1	Σ_2
1.1	3	419	1 296
1.4	79	416	877
1.7	237	337	461
2.0	79	100	124
2.3	18	21	24
2.6	3	3	3
	$s_0 = 419$	$s_1 = 1\,296$	$s_2 = 2\,785$

$$M = d\frac{s_1}{s_0} + x_0, \quad \sigma = d\sqrt{\frac{2s_2 - s_1\left(\frac{s_1}{s_0} + 1\right)}{s_0}}. \tag{5}$$

Здесь d — величина интервала между последовательными вариантами (в данном примере $d = 0.3$), x_0 — варианта, предшествующая первой, встречающейся в данном ряду (в данном примере $x_0 = 1.1 - 0.3 = 0.8$).

В частности, для приведенного примера имеем:

$$M = 0.3 \times \frac{1296}{419} + 0.8 = 1.73,$$

$$\sigma = 0.3\sqrt{\frac{5570 - 1296\,(1296 : 419 + 1)}{419}} = 0.24$$

и

$$V = \frac{0.24}{1.73}100\% = 14\%.$$

Примеры использования номограмм

Пример 1. Требуется найти число наблюдений, обеспечивающее (с вероятностью 0.99) ошибку средней не свыше 2% при коэффициенте вариации $V = 14\%$.

По фиг. 1 номограммы для точки $V = 14$ горизонтальной оси находим, двигаясь по вертикали, соответствующую точку кривой $p = 2\%$. Для этой точки, пользуясь вертикальным масштабом, отсчитываем $N = 320$ (так как найденная точка лежит между 300 и 400 на расстоянии примерно 1/5 этого промежутка от 300).

Пример 2. Требуется найти число наблюдений, обеспечивающее нахождение доли с ошибкой не свыше 5%, если известно, по предыдущим наблюдениям, что величина этой доли P близка к 0.1 (т. е. 10%).

Предварительно рассчитываем величину V:

$$V = 100\sqrt{\frac{1-P}{P}} = 100\sqrt{\frac{0.9}{0.1}} = 300 \quad \text{(в процентах)}.$$

После этого по фиг. 2 номограммы для $V = 300$ находим, двигаясь по вертикали, соответствующую точку кривой $p = 5\%$, для которой затем по вертикальному масштабу отсчитываем $N = 25.000$.

Пример 3. На основе $N = 1000$ наблюдений определена средняя $M = 6$ при коэффициенте вариации $V = 14\%$. Требуется дать характеристику надежности найденной средней.

По фиг. 1 номограммы находим точку, отвечающую значениям $V=14$ и $N=1000$. Установив, что эта точка лежит между кривыми $p=1\%$ и $p=2\%$ на расстоянии 1/5 промежутка между ними от первой кривой, определяем эту ошибку в 1.2% от 6, т. е. в 0.07. Согласно условию, принятому при построении номограммы, вероятность того, что фактическая ошибка не превышает по абсолютной величине 0.07, равна 0.99.

Пример 4. На основе $N=10000$ наблюдений определена доля, равная 0.1 (т. е. 10%). Требуется дать характеристику точности этой цифры.

В данном случае коэффициент вариации

$$V=\sqrt{\frac{0.9}{0.1}}\,100=300 \quad \text{(в процентах)}.$$

По фиг. 2 номограммы находим точку, отвечающую значениям $V=300$ и $N=10000$. После того, установив, что эта точка лежит примерно на кривой $p=8\%$, заключаем, что ошибка найденной доли составляет около 8%, т. е. 0.008.

2. Расчет числа наблюдений или величины ошибки в случае, когда выборка охватывает значительную долю генеральной совокупности

Для данного случая величина ошибки:

$$|\Delta|=\frac{t\sigma}{\sqrt{N'}}\sqrt{1-\frac{N'}{S}}, \tag{6}$$

где через N' обозначено число наблюдений при ограниченном объеме генеральной совокупности (в отличие от N — числа наблюдений при неограниченном ее объеме) и S — объем генеральной совокупности (общее число составляющих ее единиц). Вероятность иметь такую ошибку попрежнему определяется формулой (1). Из равенства (6) можно определить:

$$N'=\frac{1}{\frac{1}{S}+\frac{\Delta^2}{t^2\sigma^2}}.$$

Заменив выражение $\frac{t^2\sigma^2}{\Delta^2}$ через N, получим:

$$N'=N\frac{1}{1+\frac{N}{S}}. \tag{7}$$

Это выражение связывает число наблюдений данной задачи с числом наблюдений, даваемым прежней же номограммой, через умножение последнего на множитель $\frac{1}{1+N/S}$, не зависящий ни от характеристики колеблемости V, ни от задаваемой степени точности p.

Значения множителя $\frac{1}{1+N/S}$ для различных значений отношений N/S представлены графически (фиг. 3). Таким образом для расчета числа наблюдений при ограниченном объеме генеральной совокупности первоначально определяется соответствующее число N для случая неограниченного объема совокупности, отсюда определяется величина отношения N/S, затем по фиг. 3 определяется добавочный множитель и на него умножается найденное ранее число N.

При определении надежности результата выборочного обследования в случае ограниченного объема генеральной совокупности заданным является именно число N'. Для этого случая, пользуясь (7), находим, что

$$N=\frac{N'}{1-\frac{N'}{S}}. \tag{8}$$

Величина $1-N'/S$ представляет необследованную долю генеральной совокупности. Найдя по формуле (8) число N, отвечающее фактическому объему выборки N', решаем

далее задачу, так же как и для случая неограниченного объема генеральной совокупности, пользуясь попрежнему номограммами фиг. 1 и 2.

Пример 5. При коэффициенте вариации $V = 14\%$ рассчитать число наблюдений N' для определения средней с ошибкой не свыше 2%, если объем генеральной совокупности составляет $S = 300$ единиц.

Находим по фиг. 1, как и прежде (пример 1), $N = 320$, т. е. отношение $N/S = 1.1$. По фиг. 3 находим, что этому значению N/S отвечает множитель, равный 0.48. Следовательно, искомое число наблюдений $N' = 320 \times 0.48 = 154$.

Пример 6. На основе выборки, охватившей $N' = 600$ единиц из $S = 3000$, найдена средняя, равная 10, при коэффициенте вариации $V = 30\%$. Требуется определить степень надежности найденной средней.

Необследованная часть генеральной совокупности $1 - N'/S$ составляет в данном случае 4/5. Разделив $N' = 600$ на 4/5, имеем $N = 750$. После того по фиг. 1 находим точку, отвечающую $V = 30$, $N = 750$. Установив, что эта точка лежит между кривыми $p = \%$ и $p = 2\%$, ближе к первой, заключаем, что ошибка средней определяется в данном случае не более чем в 3%, точнее 2.8% (с вероятностью 0.99).

Поступило в редакцию 10 V 1938.

DIAGRAMS FOR THE NUMBER OF OBSERVATIONS REQUIRED FOR DETERMINING AVERAGE VALUES IN STATISTICAL INVESTIGATIONS

E. N. BOUCHMAN

(Summary)

In organizing a statistical work by method of selection at random it is necessary to have the minimum number of experiments in order to obtain the satisfactory accuracy in the result. This result usually is an average value or a relative part. When the statistical work is accomplished, it is no less important to know the value of the probable error of the above-mentioned average result.

The Liapunov theorem works both these problems out when the characteristic of the variating series is known.

For the same purpose it is convenient to use the diagrams which are offered in the present work. These diagrams are constructed on the basis of Liapunov's theorem.

Fig. 1 and 2 permit us to calculate the number of the experiments selected at random or the size of error in the case when the general totality is unlimited.

We use Fig. 1 for determination the average value

a) when it is known that the coefficient of variation does not exceed than 100%;

b) when the relative part of the general totality is not more than 0.5.

We use fig. 2 in the opposite cases. Fig. 3 permits us to find the correction for the case when the general totality is limited.

The procedure is illustrated by the numerical examples.