

# ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

# ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

- *Генеральная совокупность* – совокупность всех объектов, подлежащих изучению.
- *Выборочная совокупность (выборка)* – часть объектов генеральной совокупности, отобранных для исследования.
- *Объем совокупности* (генеральной или выборочной) – число объектов этой совокупности.
- Генеральная совокупность может содержать конечное или бесконечное число элементов. Выборка всегда содержит конечное число элементов.

# ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

- Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем исследователя признаке генеральной совокупности, выборка должна быть *репрезентативной* (представительной). Репрезентативность выборки обеспечивается случайностью отбора ее элементов.

# Статистическое распределение

Пусть из генеральной совокупности извлечена выборка объема  $n$ , причем значение  $x_1$  наблюдалось  $n_1$  раз,  $x_2$  —  $n_2$  раз, ...,  $x_k$  —  $n_k$  раз. Наблюдаемые значения  $x_i$  называются *вариантами*. Последовательность вариантов, записанных в возрастающем порядке, называется *вариационным рядом*. Число  $n_i$  называется *частотой* варианты  $x_i$ , а отношение  $\omega_i = \frac{n_i}{n}$  называется *относительной частотой* варианты  $x_i$ .

*Статистическим распределением выборки* называется перечень вариантов вариационного ряда и соответствующих им частот (или относительных частот).

# Статистическое распределение

Статистическое распределение обычно задают в виде таблицы частот (или таблицы относительных частот).

|       |       |       |         |       |
|-------|-------|-------|---------|-------|
| $x_i$ | $x_1$ | $x_2$ | $\dots$ | $x_k$ |
| $n_i$ | $n_1$ | $n_2$ | $\dots$ | $n_k$ |

$$n_1 + n_2 + \dots + n_k = n$$

|            |            |            |         |            |
|------------|------------|------------|---------|------------|
| $x_i$      | $x_1$      | $x_2$      | $\dots$ | $x_k$      |
| $\omega_i$ | $\omega_1$ | $\omega_2$ | $\dots$ | $\omega_k$ |

$$\omega_1 + \omega_2 + \dots + \omega_k = 1$$

# Пример

**Пример** В результате исследования получена выборка: 2, 5, 0, 1, 6, 3, 0, 1, 5, 4, 0, 3, 3, 2, 1, 4, 0, 0, 2, 3, 6, 0, 3, 0, 1. Найти распределение частот и распределение относительных частот.

Объем этой выборки  $n = 25$ .

Запишем вариационный ряд, выбрав различные значения и расположив их в порядке возрастания, получим: 0, 1, 2, 3, 4, 5, 6.

Заметим, что число 0 встретилось в выборке 7 раз, число 1 – 4 раза, число 2 – 3 раза, число 3 – 5 раз, числа 4, 5 и 6 – по 2 раза.

О.....

# Продолжение примера

Следовательно, распределение частот имеет вид:

|       |   |   |   |   |   |   |   |
|-------|---|---|---|---|---|---|---|
| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| $n_i$ | 7 | 4 | 3 | 5 | 2 | 2 | 2 |

Сделаем проверку: сумма частот должна быть равна объему выборки. Действительно,  $7 + 4 + 3 + 5 + 2 + 2 + 2 = 25$ .

|            |      |      |      |     |      |      |      |
|------------|------|------|------|-----|------|------|------|
| $x_i$      | 0    | 1    | 2    | 3   | 4    | 5    | 6    |
| $\omega_i$ | 0,28 | 0,16 | 0,12 | 0,2 | 0,08 | 0,08 | 0,08 |

Выполним проверку: сумма относительных частот должна быть равна 1. Действительно,  $0,28 + 0,16 + 0,12 + 0,2 + 0,08 + 0,08 + 0,08 = 1$ .



# Большие выборки

В случае большого объема выборки ( $n > 40$ ) статистическое распределение принято задавать в виде *интервальной таблицы частот*. Длина частичного интервала находится по формуле Стерджеса:

$$h = \frac{x_{\max} - x_{\min}}{1 + 1,4 \ln n}, \text{ где } x_{\max}, x_{\min} - \text{максимальное и минимальное}$$

значения в выборке,  $n$  – объем выборки.

Для простоты вычислений полученное значение  $h$  принято округлять (например, если все числа в выборке целые, то  $h$  также удобно округлить до целого числа).



# Пример

**Пример 3.** В результате исследования получена выборка: 75, 76, 51, 40, 81, 72, 54, 53, 66, 44, 130, 100, 110, 113, 103, 112, 99, 114, 122, 115, 68, 111, 92, 124, 145, 118, 140, 117, 133, 120, 85, 88, 87, 94, 77, 146, 111, 102, 96, 81, 111, 92, 103, 98, 102, 80, 108, 82, 88, 129. Задать статистическое распределение выборки в виде интервальной таблицы частот и интервальной таблицы относительных частот.

В данном случае  $x_{\min} = 40, x_{\max} = 146, n = 50$ , следовательно, длина частичного интервала равна

$$h = \frac{146 - 40}{1 + 1,4 \ln 50} \approx 16,366 \approx 16.$$

# Продолжение примера

В качестве начала первого интервала берется число  $a_1 = x_{\min}$ .

Если  $a_i$  - начало  $i$ -го интервала, то  $a_{i+1} = a_i + h$ .

Значит  $a_1 = 40, a_2 = a_1 + h = 40 + 16 = 56, a_3 = a_2 + h = 56 + 16 = 72$  и т.д.

Полученные интервалы перечисляются во второй строке интервальной таблицы частот.

В третьей строке таблицы указываются частоты (относительные частоты) интервалов. Под частотой  $n_i$  интервала понимается количество значений в выборке, принадлежащих данному интервалу.

| Номер интервала  | 1          | 2          | 3          | 4           | 5            | 6            | 7            |
|------------------|------------|------------|------------|-------------|--------------|--------------|--------------|
| $[a_i; a_{i+1})$ | $[40; 56)$ | $[56; 72)$ | $[72; 88)$ | $[88; 104)$ | $[104; 120)$ | $[120; 136)$ | $[136; 152)$ |
| $n_i$            | 5          | 2          | 10         | 13          | 11           | 6            | 3            |

Проверка:  $n_1 + n_2 + \dots + n_7 = 5 + 2 + 10 + 13 + 11 + 6 + 3 = 50$ .

# Продолжение примера

Для построения интервальной таблицы относительных частот нужно найти для каждого интервала его относительную частоту по формуле  $\omega_i = \frac{n_i}{n}$ . Эта таблица будет иметь следующий вид:

| Номер интервала  | 1         | 2         | 3         | 4          | 5           | 6           | 7           |
|------------------|-----------|-----------|-----------|------------|-------------|-------------|-------------|
| $[a_i; a_{i+1})$ | $[40;56)$ | $[56;72)$ | $[72;88)$ | $[88;104)$ | $[104;120)$ | $[120;136)$ | $[136;152)$ |
| $\omega_i$       | 0,1       | 0,04      | 0,2       | 0,26       | 0,22        | 0,12        | 0,06        |

Проверка:  $\omega_1 + \omega_2 + \dots + \omega_7 = 0,1 + 0,04 + 0,2 + 0,26 + 0,22 + 0,12 + 0,06 = 1$ .



# Графическое изображение

  
*Полигоном частот* называется ломаная, отрезки которой последовательно соединяют точки  $(x_1; n_1), (x_2; n_2), \dots, (x_k; n_k)$ , где  $x_i$  – варианты выборки,  $n_i$  – соответствующие им частоты.

*Полигоном относительных частот* называется ломаная, отрезки которой последовательно соединяют точки  $(x_1; \omega_1), (x_2; \omega_2), \dots, (x_k; \omega_k)$ , где  $x_i$  – варианты выборки,  $\omega_i$  – соответствующие им относительные частоты.

*Гистограммой частот (относительных частот)* называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длины  $h$ , а высоты равны отношению  $\frac{n_i}{h}$  ( $\frac{\omega_i}{h}$ ).

# Графическое изображение

Площадь гистограммы частот равна объему выборки  $n$ .

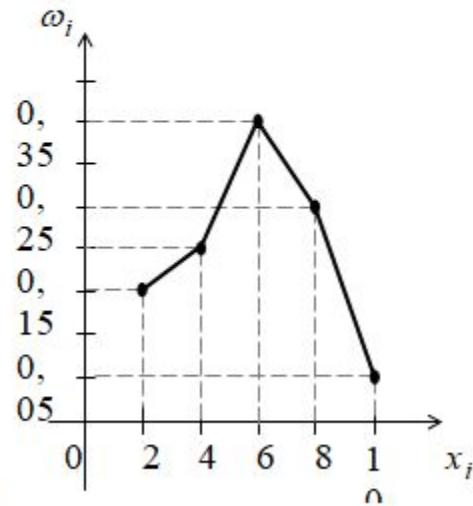
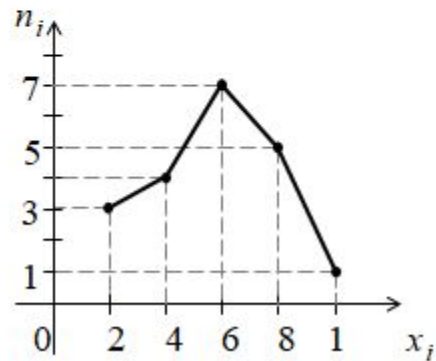
Площадь гистограммы относительных частот равна 1.

Гистограмма служит только для изображения статистического распределения, заданного в виде интервальной таблицы. Полигон можно изобразить для статистического распределения, заданного любым способом. Если статистическое распределение выборки задано в виде интервальной таблицы, то при построении полигона в качестве вариант  $x_i$  берут середины интервалов, а в качестве частот  $n_i$  (относительных частот  $\omega_i$ ) берут частоты (относительные частоты) соответствующих интервалов.

# Пример (полигон)

|       |   |   |   |   |    |
|-------|---|---|---|---|----|
| $x_i$ | 2 | 4 | 6 | 8 | 10 |
| $n_i$ | 3 | 4 | 7 | 5 | 1  |

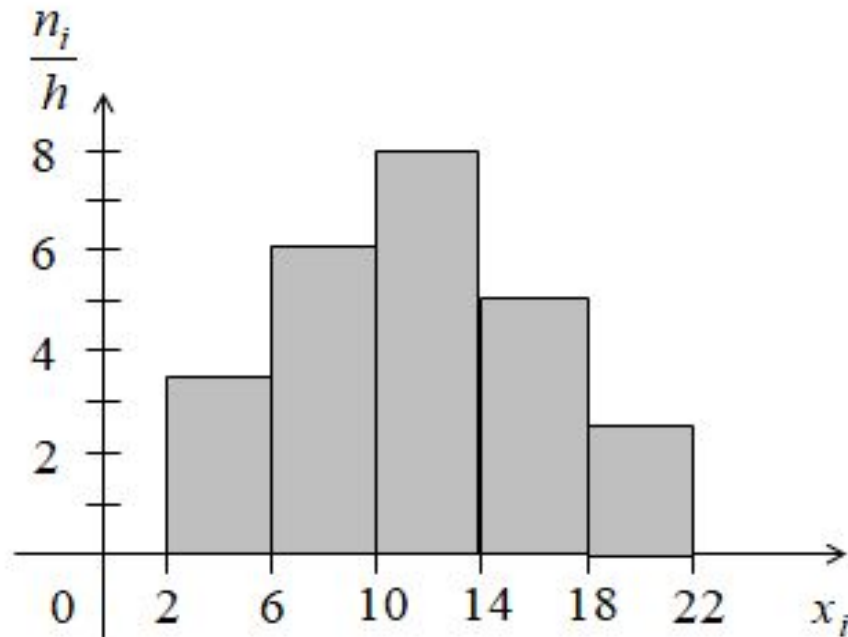
|            |      |     |      |      |      |
|------------|------|-----|------|------|------|
| $x_i$      | 2    | 4   | 6    | 8    | 10   |
| $\omega_i$ | 0,15 | 0,2 | 0,35 | 0,25 | 0,05 |





# Пример (гистограмма)

| Номер<br>интервала | 1       | 2        | 3         | 4         | 5         |
|--------------------|---------|----------|-----------|-----------|-----------|
| $[a_i; a_{i+1})$   | $[2;6)$ | $[6;10)$ | $[10;14)$ | $[14;18)$ | $[18;22)$ |
| $n_i$              | 14      | 24       | 32        | 20        | 10        |



# Эмпирическая функция распределения

Эмпирической функцией распределения (функцией распределения выборки) называется функция  $F^*(x)$ , которая определяет для каждого значения  $x$  относительную частоту события  $X < x$ :

$$F^*(x) = \frac{n_x}{n}, \text{ где } n_x \text{ — число выборочных значений, меньших } x; n \text{ —}$$

объем выборки.

Если изучаемый количественный признак  $X$  является дискретным, то график функции  $F^*(x)$  имеет ступенчатый вид, а если  $X$  является непрерывным, то график  $F^*(x)$  представляет собой непрерывную линию.

# Задача

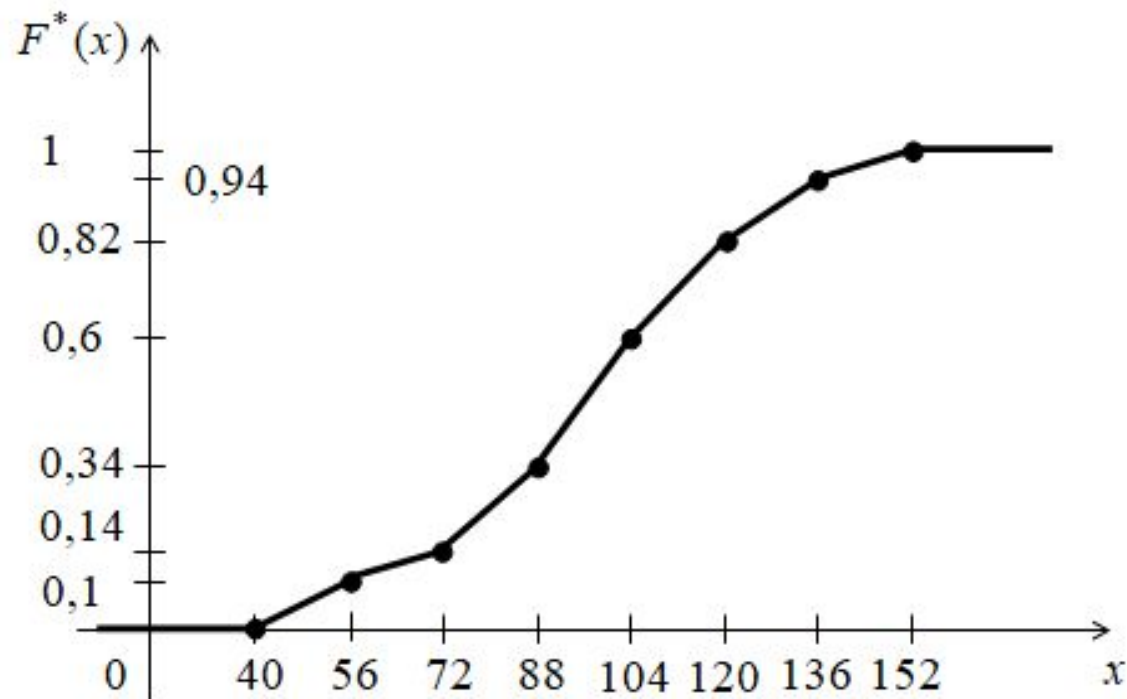
**Пример** . Статистическое распределение выборки задано в виде интервальной таблицы частот. Считая, что изучаемый количественный признак является непрерывным, найти эмпирическую функцию распределения и построить ее график.



| Номер интервала  | 1         | 2         | 3         | 4          | 5           | 6           | 7           |
|------------------|-----------|-----------|-----------|------------|-------------|-------------|-------------|
| $[a_i; a_{i+1})$ | $[40;56)$ | $[56;72)$ | $[72;88)$ | $[88;104)$ | $[104;120)$ | $[120;136)$ | $[136;152)$ |
| $n_i$            | 5         | 2         | 10        | 13         | 11          | 6           | 3           |

# Таблица и график для эмпирической функции распределения

|          |    |     |      |      |     |      |      |     |
|----------|----|-----|------|------|-----|------|------|-----|
| $x$      | 40 | 56  | 72   | 88   | 104 | 120  | 136  | 152 |
| $F^*(x)$ | 0  | 0,1 | 0,14 | 0,34 | 0,6 | 0,82 | 0,94 | 1   |



# Точечные статистические оценки числовых характеристик случайной величины

Пусть для изучения генеральной совокупности относительно количественного признака  $X$  извлечена выборка объема  $n$  и известно статистическое распределение выборки:

|       |       |       |         |       |
|-------|-------|-------|---------|-------|
| $x_i$ | $x_1$ | $x_2$ | $\dots$ | $x_k$ |
| $n_i$ | $n_1$ | $n_2$ | $\dots$ | $n_k$ |



# Точечные статистические оценки числовых характеристик случайной выборки

Выборочной средней называется число  $\bar{x}_B = \frac{\sum_{i=1}^k n_i \cdot x_i}{n}$ .

Выборочной дисперсией называется число  $D_B = \frac{\sum_{i=1}^k n_i \cdot (x_i - \bar{x}_B)^2}{n}$ .

Выборочным средним квадратическим отклонением называется число  $\sigma_B = \sqrt{D_B}$ .

Для вычисления выборочной дисперсии удобно использовать

формулу 
$$D_B = \frac{\sum_{i=1}^k n_i \cdot x_i^2}{n} - \left( \frac{\sum_{i=1}^k n_i \cdot x_i}{n} \right)^2 = \frac{\sum_{i=1}^k n_i \cdot x_i^2}{n} - (\bar{x}_B)^2.$$

# Точечные статистические оценки числовых характеристик случайной

## ВОПРОСЫ

Исправленной выборочной дисперсией называется число

$$s^2 = \frac{n}{n-1} \cdot D_B = \frac{\sum_{i=1}^k n_i \cdot (x_i - \bar{x}_B)^2}{n-1}.$$

Исправленным выборочным средним квадратическим отклонением  $s$  называется число, равное квадратному корню из исправленной дисперсии  $s^2$ .

На практике при малом объеме выборки ( $n < 30$ ) пользуются исправленной выборочной дисперсией  $s^2$ . При достаточно большом объеме выборки  $n$  числа  $D_B$  и  $s^2$  различаются мало, поэтому для оценки генеральной дисперсии можно использовать как  $s^2$ , так и  $D_B$ .

Чем больше объем выборки  $n$ , тем более верными будут точечные оценки математического ожидания и генеральной дисперсии.

# Пример

**Пример** . Найти выборочную среднюю, выборочную дисперсию и выборочное среднее квадратическое отклонение, если статистическое распределение выборки задано в виде интервальной таблицы частот:



|                  |           |           |           |            |             |             |             |
|------------------|-----------|-----------|-----------|------------|-------------|-------------|-------------|
| $[a_i; a_{i+1})$ | $[40;56)$ | $[56;72)$ | $[72;88)$ | $[88;104)$ | $[104;120)$ | $[120;136)$ | $[136;152)$ |
| $n_i$            | 5         | 2         | 10        | 13         | 11          | 6           | 3           |

Составим вспомогательную расчетную таблицу. Числа третьего столбца этой таблицы получаются в результате умножения чисел первого и второго столбцов, а числа четвертого столбца – умножения чисел первого и третьего столбцов. □

# Пример

| Середины<br>интервалов $x_i$ | Частоты $n_i$ | $n_i \cdot x_i$ | $n_i \cdot x_i^2$ |
|------------------------------|---------------|-----------------|-------------------|
| 48                           | 5             | 240             | 11520             |
| 64                           | 2             | 128             | 8192              |
| 80                           | 10            | 800             | 64000             |
| 96                           | 13            | 1248            | 119808            |
| 112                          | 11            | 1232            | 137984            |
| 128                          | 6             | 768             | 98304             |
| 144                          | 3             | 432             | 62208             |
| Сумма                        | 50            | 4848            | 502016            |

# Пример

Используя сумму, найденную в третьем столбце таблицы, вычислим выборочную среднюю:

$$\bar{x}_B = \frac{\sum_{i=1}^7 n_i \cdot x_i}{n} = \frac{4848}{50} = 96,96 .$$

Используя сумму значений четвертого столбца, найдем выборочную дисперсию:

$$D_B = \frac{\sum_{i=1}^7 n_i \cdot x_i^2}{n} - (\bar{x}_B)^2 = \frac{502016}{50} - (96,96)^2 \approx 639,08 .$$

Тогда выборочное среднее квадратическое отклонение

$$\sigma_B = \sqrt{D_B} = \sqrt{639,08} \approx 25,28 .$$



# Мода и медиана

*Выборочной модой  $M_0$*  называется варианта, которая имеет наибольшую частоту.

Пусть дана выборка объема  $n$ . Запишем элементы выборки в порядке возрастания:  $x_1, x_2, \dots, x_n$ . Если в этой выборке нечетное число элементов, то *выборочной медианой  $M_e$*  называется значение  $x_i$ , стоящее в центре ряда. Если в выборке четное число элементов, то есть  $n = 2k$ , то *выборочной медианой  $M_e$*  называется среднее арифметическое двух значений, стоящих в центре ряда, то есть

$$M_e = \frac{x_k + x_{k+1}}{2}.$$

# Интервальные оценки числовых характеристик случайной величины

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра. Так, выборочная средняя  $\bar{x}_B$ , найденная по выборке малого объема, может значительно отличаться от генеральной средней, то есть от математического ожидания случайной величины  $X$ . Это приведет к тому, что исследователь получит неверную информацию об изучаемой генеральной совокупности. Поэтому при небольших объемах выборки используются интервальные оценки числовых характеристик случайной величины.

# Интервальная оценка математического ожидания

Интервальной оценкой (с надежностью  $\gamma$ ) математического ожидания  $a$  нормально распределенного количественного признака  $X$  по выборочной средней  $\bar{x}_B$  при *неизвестном среднем квадратическом отклонении* генеральной совокупности (и объеме выборки  $n < 30$ ) является доверительный интервал:

$$\bar{x}_B - \frac{t_\gamma \cdot s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma \cdot s}{\sqrt{n}}, \text{ где}$$

$$\delta = \frac{t_\gamma \cdot s}{\sqrt{n}} \text{ — точность оценки; } s \text{ — исправленное выборочное}$$

среднее квадратическое отклонение;  $n$  — объем выборки;  $t_\gamma$  — находят по специальной таблице (приложение 3) по заданным  $n$  и  $\gamma$ .



# Интервальная оценка математического ожидания

Точность оценки  $\delta$  определяет «широту» доверительного интервала. Надежность оценки  $\gamma$  – это вероятность того, что неизвестное математическое ожидание  $a$  действительно находится в найденном доверительном интервале. Обычно надежность  $\gamma$  задается заранее, причем в качестве  $\gamma$  берут близкое к 1 число (чаще всего 0,95; 0,99; 0,999). При этом точность оценки зависит от заданной надежности: чем больше надежность, тем шире доверительный интервал (то есть тем «меньше» точность), и наоборот.

# Пример

**Пример** . Случайная величина  $X$  распределена нормально. По выборке объема  $n=20$  найдены выборочная средняя  $\bar{x}_B=15,4$  и исправленное выборочное среднее квадратическое отклонение  $s=0,9$ . Оценить неизвестное математическое ожидание при помощи доверительного интервала с надежностью 0,95.

$$\text{В данном случае } \bar{x}_B - \frac{t_\gamma \cdot s}{\sqrt{n}} < a < \bar{x}_B + \frac{t_\gamma \cdot s}{\sqrt{n}}.$$

Используя таблицу значений  $t_\gamma$ , зная  $\gamma=0,95$  и  $n=20$ , найдем  $t_\gamma = 2,093$ .

$$\text{Значит, точность оценки } \delta = \frac{2,093 \cdot 0,9}{\sqrt{20}} \approx 0,421.$$

Таким образом, доверительный интервал для математического ожидания можно записать в виде:

$$15,4 - 0,421 < a < 15,4 + 0,421 \text{ или } 14,979 < a < 15,821.$$



# Интервальная оценка среднего квадратического отклонения

Интервальной оценкой (с надежностью  $\gamma$ ) среднего квадратического отклонения  $\sigma$  нормально распределенного количественного признака  $X$  по исправленному выборочному среднему квадратическому отклонению  $s$  служит доверительный интервал:

$$s \cdot (1 - q) < \sigma < s \cdot (1 + q), \text{ если } q < 1,$$

$$0 < \sigma < s \cdot (1 + q), \text{ если } q > 1,$$

где  $q$  находят по специальной таблице (приложение 4) по заданным  $n$  и  $\gamma$ .

# Пример

**Пример** . Используя условие примера 2, найти доверительный интервал для оценки неизвестного среднего квадратического отклонения случайной величины  $X$ .

Так как  $n = 20$  и  $\gamma = 0,95$ , то по таблице значений  $q$  найдем  $q = 0,37 < 1$ .

Тогда искомый доверительный интервал будет иметь вид:

$$0,9 \cdot (1 - 0,37) < \sigma < 0,9 \cdot (1 + 0,37).$$

Окончательно получаем:  $0,567 < \sigma < 1,233$ .

Таким образом, с вероятностью  $\gamma = 0,95$  можно утверждать, что неизвестное среднее квадратическое отклонение  $\sigma$  заключено в интервале  $(0,567; 1,233)$ .