Генеральная средняя

Пусть изучается дискретная генеральная совокупность относительно количественного признака X.

Генеральной средней $\overline{x}_г$ называют среднее арифметическое значений признака генеральной совокупности.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N различны, то

$$\overline{x}_г = (x_1 + x_2 + \ldots + x_N)/N.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причем $N_1 + N_2 + \ldots + N_k = N$, то

$$\overline{x}_г = (x_1 N_1 + x_2 N_2 + \ldots + x_k N_k)/N,$$

т. е. генеральная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

Выборочная средняя

Пусть для изучения генеральной совокупности относительно количественного признака X извлечена выборка объема n.

Выборочной средней $\overline{x}_в$ называют среднее арифметическое значение признака выборочной совокупности.

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

$$\overline{x}_в = (x_1 + x_2 + \ldots + x_n)/n.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$, причем $n_1 + n_2 + \ldots + n_k = n$, то

$$\overline{x}_в = (n_1x_1 + n_2x_2 + \ldots + n_kx_k)/n,$$

или

$$\overline{x}_в = \left(\sum_{i=1}^{k} n_i x_i \right) / n,$$

Пример несмещенной оценки:

Оценка генеральной средней по выборочной средней

Доказано, что выборочная средняя есть несмещенная оценка генеральной средней

$$M(\overline{X}_в) = \overline{x}_г.$$

Итак, при увеличении объема выборки n выборочная средняя стремится по вероятности к генеральной средней, а это и означает, что выборочная средняя есть состоятельная оценка генеральной средней. Из сказанного следует также, что если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности будут найдены выборочные средние, то они будут приближенно равны между собой. В этом и состоит свойство *устойчивости выборочных средних*.

Генеральная дисперсия

Для того чтобы охарактеризовать рассеяние значений количественного признака X генеральной совокупности вокруг своего среднего значения, вводят сводную характеристику — генеральную дисперсию.

Генеральной дисперсией $D_г$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\overline{x}_г$.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N различны, то

$$D_г = \left(\sum_{i=1}^{N} (x_i - \overline{x}_г)^2\right) \Big/ N.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$, причем $N_1 + N_2 + \ldots + N_k = N$, то

$$D_г = \left(\sum_{i=1}^{k} N_i (x_i - \overline{x}_г)^2\right) \Big/ N,$$

т. е. генеральная дисперсия есть средняя взвешенная квадратов отклонений с весами, равными соответствующим частотам.

Пример. Генеральная совокупность задана таблицей распределения

x_i	2	4	5	6
N_i	8	9	10	3

Найти генеральную дисперсию.

Р е ш е н и е. Найдем генеральную среднюю (см. § 3):

$$\overline{x}_г = \frac{8\cdot 2+9\cdot 4+10\cdot 5+3\cdot 6}{8+9+10+3} = \frac{120}{30} = 4.$$

Найдем генеральную дисперсию;

$$D_г = \frac{8\cdot(2-4)^2+9\cdot(4-4)^2+10\cdot(5-4)^2+3\cdot(6-4)^2}{30} = 54/30 = 1{,}8.$$

Кроме дисперсии для характеристики рассеяния значений признака генеральной совокупности вокруг своего среднего значения пользуются сводной характеристикой — средним квадратическим отклонением.

Генеральным средним квадратическим отклонением (стандартом) называют квадратный корень из генеральной дисперсии:

$$\sigma_г = \sqrt{D_г}.$$

Выборочная дисперсия

Для того чтобы охарактеризовать рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения $\bar{x}_в$, вводят сводную характеристику — выборочную дисперсию.

Выборочной дисперсией $D_в$ называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\bar{x}_в$.

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

$$D_в = \left(\sum_{i=1}^{n} (x_i - \bar{x}_в)^2 \right) / n.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$, причем $n_1 + n_2 + \ldots + n_k = n$, то

$$D_в = \left(\sum_{i=1}^{k} n_i (x_i - \bar{x}_в)^2 \right) / n,$$

т. е. выборочная дисперсия есть средняя взвешенная квадратов отклонений с весами, равными соответствующим частотам.

Пример. Выборочная совокупность задана таблицей распределения

x_i	1	2	3	4
n_i	20	15	10	5

Найти выборочную дисперсню.

Р е ш е н и е. Найдем выборочиую среднюю (см. § 4):

$$\overline{x}_{\text{в}} = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{20 + 15 + 10 + 5} = \frac{100}{50} = 2.$$

Найдем выборочную дисперсию:

$$D_{\text{в}} = \frac{20\,(1-2)^2 + 15 \cdot (2-2)^2 + 10 \cdot (3-2)^2 + 5 \cdot (4-2)^2}{50} =$$

$$= 50/50 = 1.$$

Кроме дисперсии для характеристики рассеяния значений признака выборочной совокупности вокруг своего среднего значения пользуются сводной характеристикой — средним квадратическим отклонением.

Выборочным средним квадратическим отклонением (стандартом) называют квадратный корень из выборочной дисперсии:

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}}.$$

Формула для вычисления дисперсий

Вычисление дисперсии, безразлично — выборочной или генеральной, можно упростить, используя следующую теорему.

Теорема. *Дисперсия равна среднему квадратов значений признака минус квадрат общей средней:*

$$D = \bar{x}^2 - [\bar{x}]^2.$$

Пример. Найти дисперсию по данному распределению

x_i	1	2	3	4
n_i	20	15	10	5

Решение. Найдем общую среднюю:

$$\bar{x} = \frac{20 \cdot 1 + 15 \cdot 2 + 10 \cdot 3 + 5 \cdot 4}{20 + 15 + 10 + 5} = \frac{100}{50} = 2.$$

Найдем среднюю квадратов значений признака

$$\bar{x}^2 = \frac{20 \cdot 1^2 + 15 \cdot 2^2 + 10 \cdot 3^2 + 5 \cdot 4^2}{50} = 5.$$

Искомая дисперсия

$$D = \bar{x}^2 - [\bar{x}]^2 = 5 - 2^2 = 1.$$

Оценка генеральной дисперсии по исправленной выборочной

Пусть из генеральной совокупности в результате n независимых наблюдений над количественным признаком X извлечена повторная выборка объема n:

значения признака	$x_1\ x_2\ \ldots\ x_k$
частоты	$n_1\ n_2\ \ldots\ n_k$

При этом $n_1+n_2+\ldots+n_k=n$.

Требуется по данным выборки оценить (приближенно найти) неизвестную генеральную дисперсию $D_г$.

Математическое ожидание выборочной дисперсии равно

$$M[D_в]=\frac{n-1}{n}D_г.$$

Точность оценки. Доверительная вероятность (надежность). Доверительный интервал.

Точечной называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше, — точечные. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться интервальными оценками.

Интервальной называют оценку, которая определяется двумя числами — концами интервала. Интервальные оценки позволяют установить точность и надежность оценок (смысл этих понятий выясняется ниже).

Пусть найденная по данным выборки статистическая характеристика Θ^* служит оценкой неизвестного параметра Θ. Будем считать Θ постоянным числом (Θ может быть и случайной величиной). Ясно, что Θ^* тем точнее определяет параметр Θ, чем меньше абсолютная величина разности $|\Theta - \Theta^*|$. Другими словами, если $\delta > 0$ и $|\Theta - \Theta^*| < \delta$, то чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует *точность оценки*.

Однако статистические методы не позволяют категорически утверждать, что оценка Θ^* удовлетворяет неравенству $|\Theta - \Theta^*| < \delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надежностью (доверительной вероятностью) оценки Θ по Θ^* называют вероятность γ, с которой осуществляется неравенство $|\Theta - \Theta^*| < \delta$. Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, что $|\Theta - \Theta^*| < \delta$, равна γ:

$$P[|\Theta - \Theta^*| < \delta] = \gamma.$$

Заменив неравенство $|\Theta - \Theta^*| < \delta$ равносильным ему двойным неравенством $-\delta < \Theta - \Theta^* < \delta$, или $\Theta^* - \delta < \Theta < \Theta^* + \delta$, имеем

$$P[\Theta^* - \delta < \Theta < \Theta^* + \delta] = \gamma.$$

Это соотношение следует понимать так: вероятность того, что интервал $(\Theta^* - \delta, \Theta^* + \delta)$ заключает в себе (покрывает) неизвестный параметр Θ, равна γ.

Доверительным называют интервал $(\Theta^* - \delta, \Theta^* + \delta)$, который покрывает неизвестный параметр с заданной надежностью γ.

Метод доверительных интервалов разработал американский статистик Ю. Нейман, исходя из идей английского статистика Р. Фишера.

Доверительные интервалы для оценки математического ожидания нормального распределения при известном σ

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\overline{x}$. Поставим своей задачей найти доверительные интервалы, покрывающие параметр a с надежностью γ.

Будем рассматривать выборочную среднюю $\overline{x}$ как случайную величину $\overline{X}$ ($\overline{x}$ изменяется от выборки к выборке) и выборочные значения признака $x_1, x_2, \ldots, x_n$ — как одинаково распределенные независимые случайные величины $X_1, X_2, \ldots, X_n$ (эти числа также изменяются от выборки к выборке). Другими словами, математическое ожидание каждой из этих величин равно a и среднее квадратическое отклонение — σ.

Примем без доказательства, что если случайная величина X распределена нормально, то выборочная средняя $\overline{X}$, найденная по независимым наблюдениям, также распределена нормально. Параметры распределения $\overline{X}$ таковы

$$M(\overline{X}) = a, \quad \sigma(\overline{X}) = \sigma/\sqrt{n}.$$

Потребуем, чтобы выполнялось соотношение

$$P(|\overline{X}-a|<\delta)=\gamma,$$

где γ — заданная надежность.

Пользуясь формулой

$$P(|X-a|<\delta)=2\Phi(\delta/\sigma),$$

Приняв во внимание, что вероятность P задана и равна γ, окончательно имеем (чтобы получить рабочую формулу, выборочную среднюю вновь обозначим через $\overline{x}$)

$$P(\overline{x}-t\sigma/\sqrt{n}<a<\overline{x}+t\sigma/\sqrt{n})=2\Phi(t)=\gamma.$$

Смысл полученного соотношения таков: с надежностью γ можно утверждать, что доверительный интервал $(\overline{x}-t\sigma/\sqrt{n},\ \overline{x}+t\sigma/\sqrt{n})$ покрывает неизвестный параметр a; точность оценки $\delta=t\sigma/\sqrt{n}$.

Итак, поставленная выше задача полностью решена. Укажем еще, что число t определяется из равенства $2\Phi(t)=\gamma$, или $\Phi(t)=\gamma/2$; по таблице функции Лапласа (см. приложение 2) находят аргумент t, которому соответствует значение функции Лапласа, равное $\gamma/2$.

Пример. Случайная велнчина X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma=3$. Найти доверительные интервалы для оценки неизвестного математического ожндания a по выборочным средним $\overline{x}$, если объем выборки $n=36$ и задана надежность оценкн $\gamma=0{,}95$.

Р е ш е н и е. Найдем t. Из соотношения $2\Phi(t)=0{,}95$ получим $\Phi(t)=0{,}475$. По таблице приложения 2 находим $t=1{,}96$.

Найдем точность оценки:

$$\delta=t\sigma/\sqrt{n}=(1{,}96\cdot 3)/\sqrt{36}=0{,}98.$$

Доверительный интервал таков: $(\overline{x}-0{,}98;\ \overline{x}+0{,}98)$. Например, если $\overline{x}=4{,}1$, то доверительный интервал имеет следующие довернтельные границы:

$$\overline{x}-0{,}98=4{,}1-0{,}98=3{,}12;\qquad \overline{x}+0{,}98=4{,}1+0{,}98=5{,}08.$$

Другие характеристики вариационного ряда

Кроме выборочной средней и выборочной дисперсии применяются и другие характеристики вариационного ряда. Укажем главные из них.

Модой M_0 называют варианту, которая имеет наибольшую частоту. Например, для ряда

варианта	1	4	7	9
частота	5	1	20	6

мода равна 7.

Медианой m_e называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно, т. е. $n = 2k + 1$, то $m_e = x_{k+1}$; при четном $n = 2k$ медиана

$$m_e = (x_k + x_{k+1})/2.$$

Например, для ряда 2 3 5 6 7 медиана равна 5; для ряда 2 3 5 6 7 9 медиана равна $(5 + 6)/2 = 5{,}5$.

Размахом варьирования R называют разность между наибольшей и наименьшей вариантами:

$$R = x_{max} - x_{min}.$$

Например, для ряда 1 3 4 5 6 10 размах равен $10 - 1 = 9$.

Размах является простейшей характеристикой рассеяния вариационного ряда.

Средним абсолютным отклонением θ называют среднее арифметическое абсолютных отклонений:

$$\theta = (\sum n_i \mid x_i - \bar{x}_в \mid)/\sum n_i.$$

Например, для ряда

x_i	1	3	6	16
n_i	4	10	5	1

имеем:

$$\bar{x}_в = \frac{4 \cdot 1 + 10 \cdot 3 + 5 \cdot 6 + 1 \cdot 16}{4 + 10 + 5 + 1} = \frac{80}{20} = 4;$$

$$\theta = \frac{4 \cdot |1 - 4| + 10 \cdot |3 - 4| + 5 \cdot |6 - 4| + 1 \cdot |16 - 4|}{20} = 2{,}2.$$

Среднее абсолютное отклонение служит для характеристики рассеяния вариационного ряда.