Лекция №9

Точечные оценки и их свойства

Построение выборки, удовлетворяющей заданным критериям, и получение информации о попавших в выборку объектах является первым шагом выборочного исследования. Для выборочных значений можно рассчитать обобщающие показатели, такие как среднее арифметическое, дисперсия или среднее квадратическое отклонение.

Основной вопрос, который при этом возникает, - как соотносятся рассчитанные по выборке показатели с соответствующими показателями генеральной совокупности?

Именно ответу на этот вопрос, *оцениванию* показателей генеральной совокупности по выборочным значениям и посвящена данная лекция.

Рассчитанные по выборке и используемые для оценивания величины называются *оценками.*

В первом пункте будет рассмотрен ряд вопросов о том, как соотносятся друг с другом и с истинным генеральным значением показателя оценки, полученные для *нескольких выборок.* Эти вопросы рассматриваются более детально в математической статистике. Далее на основе полученных знаний мы сможем перейти к ситуации реального исследования, где в нашем распоряжении будет только *одна выборка.*

Так как нам придется одновременно рассматривать показатели и генеральной, и выборочной совокупностей, следует ввести для них раздельную систему обозначений, табл. 1.

Таблица 1. Система обозначений выборочных и генеральных показателей

Показатель	*Совокупность*	
	генеральна	*выборочна*
Число единиц	N	n
Среднее арифметическое	m	$\bar{x}$
Дисперсия	D	s^2
Среднее квадратическое	σ	s
Параметр биномиального	p	w

Выборочная средняя и ее распределение

Если для заданной выборки по обычной формуле рассчитать ее среднее арифметическое *x*, то полученное значение будет отличаться от генерального среднего. Кроме того, значение средней будет отличаться и от одной выборки к другой. Предположим, что исследователь выбрал из большой генеральной совокупности 100 различных выборок одной и той же длины n и для каждой рассчитал выборочную среднюю $\overline{x}_1, \overline{x}_2, \overline{x}_3, \ldots, \overline{x}_{100}$. Совокупность полученных значений представляет собой *распределение выборочной средней.* Отклонение выборочных средних от генеральной средней m называется ошибкой выборки и связано с тем, что выборка включает меньше значений и не является совершенной «копией в миниатюре» для генеральной совокупности.

Распределение выборочных средних обладает двумя важными свойствами:

Свойство 1. Среднее значение выборочных средних $m_{\overline{x}}$ совпадает с генеральной средней m, $m = m_{\overline{x}}$

Свойство 2. Среднее квадратическое отклонение выборочных средних $\sigma_{\overline{x}}$ будет меньше, чем среднее квадратическое отклонение генеральной совокупности σ, и будет равняться генеральному отклонению, деленному на корень квадратный из размера выборки, $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$.

Понятие распределения выборочной средней и свойства этого распределения проиллюстрируем на следующем примере.

Пример. Пусть генеральной совокупностью является 4 студента, а наблюдаемым признаком - полученное ими за тестовое задание число баллов: 2; 4; 6; 8. Генеральное среднее m этой совокупности и среднее квадратическое отклонение σ равны:

$$m = \frac{2+4+6+8}{4} = 5;$$

$$\sigma = \sqrt{\frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4}} = 2.236$$

Рассмотрим все возможные выборки из двух элементов для этой генеральной совокупности и для каждой из них найдем величину выборочной средней. Выборка делается случайным образом с повторениями, то есть одно и то же значение может оказаться выбранным дважды, табл. 2.

Таблица 2. Выборки и выборочные средние

Выборка	*Среднее* $\overline{x}$	*Выборка*	*Среднее* $\overline{x}$
2; 2	*(2+2)12=2*	6; 2	*4*
2; 4	3	6; 4	*5*
2; 6	4	6; 6	6
2; 8	5	6; 8	7
4; 2	3	8; 2	5
4; 4	4	8; 4	б
4; 6	5	8; 6	*1*
4; 8	6	8; 8	8

Общее число выборок равно 16. Как видно из таблицы, в зависимости от попавших в выборку значений выборочное среднее может принимать значения от 2 до 8, то есть в некоторых случаях оно совпадает с генеральным средним $m = 5$, но может быть также и меньше, и больше его. Построим для полученных значений таблицу частот, табл. 3, и гистограмму распределения, рис. 1. Обратим внимание на то, что гистограмма распределения выборочных средних имеет форму колокола с осью симметрии при $\overline{x} = 5$.

Таблица 3. Частотное распределение выборочных средних

Выборочное среднее $\overline{x}$	Частота повторений f
2	1
3	2
4	3
5	4

6	3
7	2
8	1

Рис. 1 Гистограмма распределения выборочных средних

В заключение вычислим среднее арифметическое выборочных средних:

$$m_{\bar{x}} = \frac{2+3+4+5+3+4+5+6+4+5+6+7+5+6+7+8}{16} = 5;$$

и их среднее квадратическое отклонение:

$$\sigma_{\bar{x}} = \sqrt{\frac{(2-5)^2+(3-5)^2+(4-5)^2+(5-5)^2+..+(7-5)^2+(8-5)^2}{16}} = 1.581 = \frac{2.236}{\sqrt{2}}$$

Таким образом, $m = m_{\bar{x}}$, $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{2}}$: оба свойства выборочного среднего выполнены. Величина $\sigma_{\bar{x}}$ представляет собой среднее отклонение выборочной средней от генеральной и называется *средней ошибкой выборки.*

Третье свойство выборочной средней относится к асимптотическому поведению распределения этой величины. Представим теперь, что рассматривается выборка значений непрерывно меняющегося признака большой генеральной совокупности.

Свойство 3. При увеличении размера выборки распределение выборочных средних из генеральной совокупности со средним арифметическим m и средним квадратическим отклонением σ приближается к нормальному распределению с параметрами m и $\frac{\sigma}{\sqrt{n}}$, $\bar{x} : N\left(m; \frac{\sigma}{\sqrt{n}}\right)$ независимо от того, какое распределение имела сама генеральная совокупность.

Свойство 3 представляет собой одну из формулировок центральной предельной теоремой. Ее можно использовать для ответа на ряд вопросов относительно значений выборочной средней аналогично тому, как нормальное распределение использовалось для ответа на вопросы об индивидуальных значениях.

Пример. Средняя величина покупки в магазине составляет 212 рублей при среднем квадратическом отклонении 45 рублей. Произвольно выберем 25 покупателей и определим среднюю величину покупки. Какова вероятность, что она превысит 230 рублей?

Решение. В данном случае генеральной совокупностью являются все покупки, совершаемые в магазине. По условию задачи генеральная совокупность имеет среднее $m = 212$ и $\sigma = 45$. Далее, из генеральной совокупности формируется выборка - берутся 25 покупателей, $n = 25$, и для них находится конкретное значение средней, выборочная средняя $\bar{x}$. Нужно найти вероятность, что эта величина больше 230 рублей, $P(\bar{x} > 230)$. Согласно свойству 3, $\bar{x}$ будет приближенно подчиняться нормальному распределению с параметрами:

$$m_{\bar{x}} = 212,\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{45}{\sqrt{25}} = 9;\ \bar{x} : N(212;9)$$

Для расчета вероятности перейдем от общего к стандартному нормальному распределению: $z = \dfrac{\bar{x} - 212}{9} : N(0;1)$. Тогда:

$$P(\bar{x} > 230) = P\left(z > \frac{230 - 212}{9}\right) = P(z > 2) = 0.5 - P(\le 0z \le 2) = 0.5 - 0.4772 = 0.0228$$

Вероятность, что средняя покупка 25 покупателей превысит 230 р., составляет 2.28%.

Средняя ошибку выборки для бесповторной схемы отбора

Рассмотренное соотношение для средней ошибки выборки $\dfrac{\sigma}{\sqrt{n}}$ выполняется в тех случаях, когда используется случайная схема отбора с повторениями, что в действительности встречается достаточно редко. При использовании бесповторной случайной схемы в выражение для средней ошибки выборки вводится поправочный множитель $\sqrt{\dfrac{N-n}{N-1}}$, где N - число элементов генеральной совокупности, n — размер выборки:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

В принципе, эту поправку можно использовать для любого соотношения размеров выборки и генеральной совокупности, но для случая, когда генеральная совокупность велика или выборка мала, множитель не используют, так как он будет почти совпадать с единицей. Согласно статистической практике поправочный фактор учитывается, если длина выборки составляет более 5% от размера генеральной совокупности.

Для практического применения формулу часто несколько упрощают:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \approx \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}} \approx \frac{\sigma}{\sqrt{n}} \times \left(1 - \frac{1}{2}\frac{n}{N}\right)$$

где первое приближенное равенство основано на соотношении $N-1 \approx N$, второе - на разложении корня квадратного в ряд Тейлора при условии $\frac{n}{N} \ll 1$. Полученный результат означает, что поправка по величине примерно равна половине от отношения длины выборки и генерального среднего. При условии $\frac{n}{N} < 5\%$ поправка составит менее 2.5%.

В случае, если размер генеральной совокупности не определен (например, обследование действия лекарства на этапе разработки), поправочный множитель не применяется.

Пример. Рассмотрим состоящую из 50 студентов выборку, сформированную из генеральной совокупности 500 студентов. Среднее квадратическое отклонение для возраста студента составляет $\sigma = 5$ лет. Вычислить среднюю ошибку выборки, определить, нужно ли учитывать конечность выборки, и если да, то вычислить поправочный множитель.

Решение. Средняя ошибка выборки равна: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{50}} \approx 0.7$.

Размер выборки составляет $\frac{50}{500} \times 100\% = 10\%$ от размера генеральной совокупности, превышает 5% и, следовательно, нужно рассчитывать поправочный множитель: $\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{500-50}{500-1}} = 0.9496$

Свойства точечных оценок

В предшествующем разделе рассмотрены свойства выборочного среднего $\bar{x}$, используемого в качестве оценки генеральной средней m. Аналогичная задача - оценить значение генерального показателя с помощью результатов выборочного наблюдения возникает и по отношению к другим величинам. Например, для генеральной дисперсии D используется следующая оценка:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Эта формула отличается от формулы расчета дисперсии, так как в знаменателе вместо n стоит выражение $n-1$. Почему следует использовать именно такое выражение, или, в более общем случае, какими качествами должна обладать "хорошая" оценка? Качество оценки характеризуется тремя свойствами.

Представим себе, что O^* - неизвестная оценка параметра O генеральной совокупности. Допустим, что по выборке из n единиц рассчитано значение $O_1^* = O_1^*\left(x_1, x_2, \ldots, x_n\right)$. Поскольку оценка использует лишь часть значений генеральной совокупности, она лишь случайно может совпасть с оцениваемым параметром. В других случаях в зависимости от выборки отклонение может быть большим или маленьким, положительным или отрицательным. Повторим опыт, то есть извлечем из генеральной совокупности другую выборку того же объема и по ее данным найдем значение оценки O_2^*. Повторяя опыт многократно, получим различные числа $O_1^*, O_2^*, \ldots, O_k^*$, каким-то образом распределенные относительно истинного значения O и характеризуемые некоторой плотностью распределения. f

Определение. Оценка называется *несмещенной,* если среднее значение оценок (математическое ожидание) для большого числа выборок совпадает со значением оцениваемой характеристики в генеральной совокупности: $M\left(O^*\right) = O$.

Как показано выше, оценка $\overline{x}$ является несмещенной оценкой для среднего арифметического m, так как $m_{\overline{x}} = m$. Выражение s^2 является несмещенной оценкой для дисперсии D. С другой стороны, если при оценивании дисперсии мы воспользовались бы формулой $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$, то получили бы смещенную оценку, которая в среднем на большом количестве выборок дала бы заниженное значение дисперсии. Гипотетическая плотность распределения несмещенной оценок O^* и смещенной O^{**} оценок параметра O представлена на рис. 2.2.

Определение. *Эффективной* называют оценку, которая при заданном объеме выборки n имеет наименьшую возможную дисперсию.

Это определение позволяет также сравнить эффективность двух оценок O^* и O^{**}. Например, если $D\left(O^*\right) < D\left(O^{**}\right)$, то O^* более эффективна. Графики плотности распределения двух оценок с разной эффективностью представлены на рис. 2.3. Более эффективная оценка имеет “более узкое” распределение, то есть для большинства выборок значения оценки лежат ближе к истинному (генеральному) значению.

Определение. Оценка называется *состоятельной,* если с увеличением размера выборки значения оценки приближаются к значению характеристики в генеральной совокупности и в пределе (по вероятности) совпадают с ним.

Свойство состоятельности удобно проиллюстрировать на примере оценки $\bar{x}$. Ее дисперсия равна $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$, поэтому при $n \to +\infty$ значение $\bar{x}$ приближается к m. Характерные графики плотности распределения состоятельной оценки O^* при увеличивающемся n показаны на рис. 2.4. В пределе для состоятельной и несмещенной оценки график плотности совпадет с прямой $x = O$.

Оценки, не обладающие тремя перечисленными свойствами, встречается намного реже, в основном на малых выборках, и их применение требует более высокой квалификации исследователя.

Плотность распределения оценок

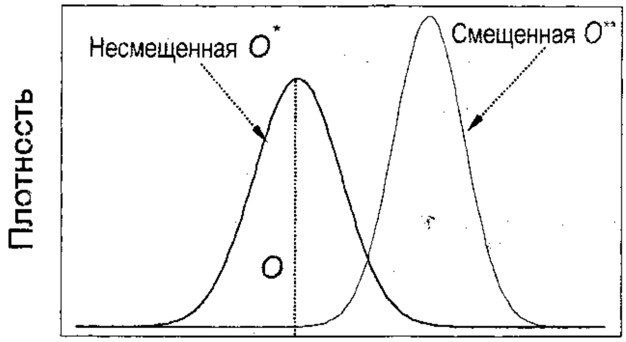

Рис. 2.2. Несмещенная О и смещенная О** оценки*

Плотность распределения оценок

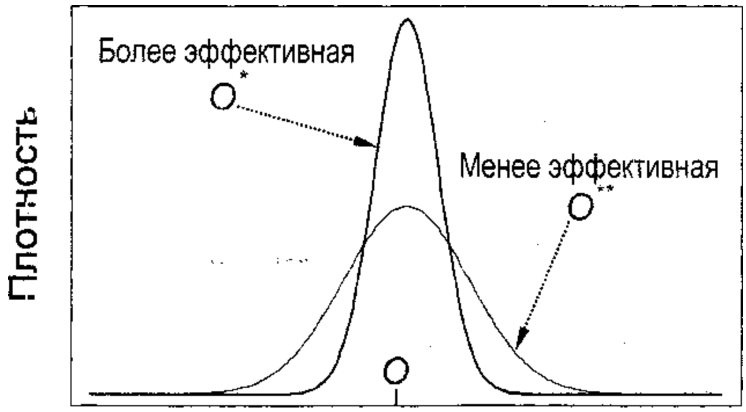

Рис. 2.3. Более эффективная О и менее эффективная О** оценки*

Плотность распределения оценок

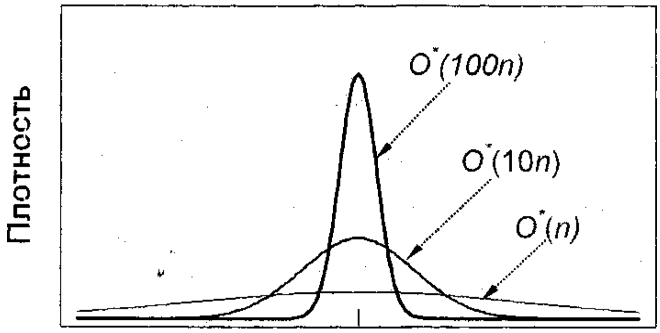

Рис. 2.4. Распределение состоятельной оценки О* *при увеличении длины выборки n*