Анализ данных
Семинар 1 (НБ)

Свойства оценок

2.3. СВОЙСТВА ОЦЕНОК

Точность какой-либо оценки, полученной по выборке, зависит от двух факторов: от способа, которым оценка вычисляется по данным выборки, и от способа отбора. Для экономии места мы иногда будем писать «точность выборочного среднего» или «точность простого случайного отбора», не упоминая особо другой существенный фактор. Мы надеемся, что в таких случаях из контекста всегда будет ясно, какой из двух факторов пропущен. Знакомясь с какой-либо новой формулой, читатель должен быть уверен, что он понимает, для какого конкретного способа отбора и способа оценивания эта формула справедлива.

В этой книге способ оценивания называется *состоятельным*, если оценка становится в точности равной оцениваемому значению для совокупности при $n = N$, т. е. когда выборку составляет вся совокупность. Очевидно, что при простом случайном отборе $\bar{y}$ и $N\bar{y}$ представляют собой состоятельные оценки соответственно среднего и суммарного значений для совокупности. Состоятельность — желательное свойство оценок. Но может найти себе применение и несостоятельная оценка, если только она обеспечивает удовлетворительную точность, ког
n мало по сравнению с N. По-видимому, ее полезность этим случа
и ограничивается. Другое определение состоятельности для случ
конечной совокупности дают Хансен, Хервиц и Мэдоу (Hanse
Hurwitz and Madow, 1953).

Как мы уже говорили, метод оценивания называется *несмеще
ным*, если среднее значение оценки, взятое по всем возможным выбо
кам данного объема n, в точности равно истинному значению для сов
купности. Если метод называют несмещенным без всяких оговорок,
это утверждение должно быть справедливым для любой совокупнос
конечных значений y_i и для любого n. Чтобы установить, будет ли
несмещенной оценкой при простом случайном отборе, вычислим знач
ние $\bar{y}$ для всех C_N^n выборок и найдем их среднее. Символ E означае
что среднее берется по всем возможным выборкам.

Теорема 2.1. Выборочное среднее $\bar{y}$ есть несмещенная оценка $\bar{Y}$.

Доказательство. По определению

$$E\bar{y} = \frac{\Sigma\bar{y}}{C_N^n} = \frac{\Sigma\,(y_1 + y_2 + \ldots + y_n)}{n\,[N!/n!\,(N-n)!]}, \qquad (2.$$

где суммирование распространяется на все C_N^n выборок. Для подсче этой суммы найдем, во скольких выборках участвует каждое конкр ное значение y_i. Поскольку на остальных $(n-1)$ местах в выбор могут стоять любые из $(N-1)$ единиц, остающихся для отбора, чис выборок, содержащих y_i, равно:

$$C_{N-1}^{n-1} = \frac{(N-1)!}{(n-1)!\,(N-n)!}\,. \qquad (2.$$

Следовательно,

$$\Sigma\,(y_1 + y_2 + \dots + y_n) = \frac{(N-1)!}{(n-1)!\,(N-n)!}\,(y_1 + y_2 + \dots + y_N).$$

Отсюда, учитывая (2.2), получаем

$$E\bar{y} = \frac{(N-1)!}{(n-1)!\,(N-n)!}\,\frac{n!\,(N-n)!}{nN!}\,(y_1 + y_2 + \dots + y_N) =$$

$$= \frac{(y_1 + y_2 + \dots + y_N)}{N} = \bar{Y}. \qquad (2.$$

Следствие. $\hat{Y} = N\bar{y}$ есть несмещенная оценка суммарного значе для совокупности, Y.

Дисперсии оценок

2.4. ДИСПЕРСИИ ОЦЕНОК

Дисперсия y_i для конечной совокупности обычно определяется по формуле

$$\sigma^2 = \frac{\sum_{1}^{N} (y_i - \overline{Y})^2}{N}. \tag{2.6}$$

Мы будем пользоваться несколько иным выражением, в котором в знаменателе вместо N стоит $(N - 1)$. Положим

$$S^2 = \frac{\sum_{1}^{N} (y_i - \overline{Y})^2}{N-1}. \tag{2.7}$$

В таком виде формула применяется теми, кто рассматривает теорию выборочного метода с точки зрения дисперсионного анализа. Преимущество такой записи в том, что большинство результатов принимает несколько более простой вид. При условии, что одни и те же обозначения применяются постоянно, все результаты эквивалентны в любой из двух форм записи.

Рассмотрим теперь дисперсию $\overline{y}$. Под ней мы подразумеваем $E(\overline{y} - \overline{Y})^2$, взятое по всем C_N^n возможным выборкам.

Теорема 2.2. Дисперсия среднего $\overline{y}$ для простой случайной выборки равна:

$$V(\overline{y}) = E(\overline{y} - \overline{Y})^2 = \frac{S^2}{n}\,\frac{(N-n)}{N} = \frac{S^2}{n}(1-f), \tag{2.8}$$

где $f = n/N$ есть доля отбора.

Следствие 1. Стандартная ошибка $\overline{y}$ равна:

$$\sigma_{\overline{y}} = \frac{S}{\sqrt{n}}\sqrt{(N-n)/N} = \frac{S}{\sqrt{n}}\sqrt{1-f}. \tag{2.12}$$

Следствие 2. Дисперсия величины $\hat{Y} = N\overline{y}$, применяемой в качестве оценки суммарного значения для совокупности, Y, равна:

$$V(\hat{Y}) = E(\hat{Y} - Y)^2 = \frac{N^2 S^2}{n}\,\frac{(N-n)}{N} = \frac{N^2 S^2}{n}(1-f). \tag{2.13}$$

Следствие 3. Стандартная ошибка $\hat{Y}$ равна:

$$\sigma_{\hat{Y}} = \frac{NS}{\sqrt{n}}\sqrt{(N-n)/N} = \frac{NS}{\sqrt{n}}\sqrt{1-f}. \tag{2.14}$$

2.5. ПОПРАВКА НА КОНЕЧНОСТЬ СОВОКУПНОСТИ

Хорошо известно, что для случайной выборки объема n из беско нечной совокупности дисперсия среднего равна σ^2/n. Если совоку ность конечна, то единственное изменение в этом результате состо в том, что нужно ввести множитель $(N-n)/N$. Множители $(N-n)/$ для дисперсии и $\sqrt{(N-n)/N}$ для стандартной ошибки называютс *поправками на конечность совокупности* (пкс). Авторы, которые в

дут изложение в терминах σ, приводят их с делителем $(N-1)$ вместо N. Если доля отбора n/N остается низкой, то эти множители близки к единице и объем совокупности сам по себе не оказывает непосредственного влияния на стандартную ошибку выборочного среднего. Например, если для двух совокупностей S одинаково, то выборка объемом в 500 единиц из совокупности, насчитывающей 200 000, обеспечивает почти ту же точность оценки среднего для совокупности, что и выборка в 500 единиц из 10 000. Лица, незнакомые с выборочным методом, с трудом воспринимают этот, действительно замечательный, результат. Им кажется интуитивно очевидным, что выборочное среднее не может быть достаточно достоверной оценкой, если оно получено на основе сведений об очень небольшой части совокупности. Читателю

На практике пкс можно не учитывать, если доля отбора не превышает 5%, а для многих целей даже если она достигает 10%. Если поправка не учитывается, то это приводит к некоторому преувеличению стандартной ошибки оценки $\bar{y}$.

Оценивание стандартной ошибки по выборке

2.6. ОЦЕНИВАНИЕ СТАНДАРТНОЙ ОШИБКИ ПО ВЫБОРКЕ

Формулы стандартных ошибок оценок средних и суммарных значений для совокупности применяются в основном для трех целей: (1) сравнить точность, которую дает простой случайный отбор, с точностью других способов отбора, (2) оценить объем выборки, необходимый для предполагаемого обследования, (3) оценить точность, действительно достигнутую в проведенном обследовании. В эти формулы входит S^2, дисперсия для совокупности. В действительности она заранее не известна, но ее можно оценить по данным выборки. Соответствующее утверждение сформулировано в теореме 2.4.

Теорема 2.4. Для простой случайной выборки

$$s^2=\frac{\sum_1^n (y_i-\bar{y})^2}{n-1}$$

есть несмещенная оценка

$$S^2=\frac{\sum_1^N (y_i-\bar{Y})^2}{N-1}.$$

Следствие. Несмещенные оценки дисперсий $\bar{y}$ и $\hat{Y} = N\bar{y}$ есть

$$v(\bar{y})=s_{\bar{y}}^2=\frac{s^2}{n}\left(\frac{N-n}{N}\right)=\frac{s^2}{n}(1-f); \qquad (2.18)$$

$$v(\hat{Y})=s_{\hat{Y}}^2=\frac{N^2 s^2}{n}\left(\frac{N-n}{N}\right)=\frac{N^2 s^2}{n}(1-f). \qquad (2.19)$$

Для стандартных ошибок положим

$$s_{\bar{y}}=\frac{s}{\sqrt{n}}\sqrt{1-f}, \quad s_{\hat{Y}}=\frac{Ns}{\sqrt{n}}\sqrt{1-f}. \qquad (2.20)$$

Эти оценки несколько смещены. Для большинства приложений их смещение не играет роли.

Читателю следует обратить внимание на символы, применяемые для обозначения истинной и оцениваемой по данным выборки дисперсий оценок. Так, для $\bar{y}$ мы пишем:

истинная дисперсия: $V(\bar{y}) = \sigma_{\bar{y}}^2$,

оцениваемая дисперсия: $v(\bar{y}) = s_{\bar{y}}^2$.

ЗАДАЧИ

Упражнения

2.1. Для совокупности с $N = 6$ значения y_i равны 8, 3, 1, 11, 4 и 7. Вычислите выборочное среднее $\bar{y}$ для всех возможных простых случайных выборок объема 2. Проверьте, что $\bar{y}$ представляет собой несмещенную оценку $\bar{Y}$ и что ее дисперсия совпадает с указанной в теореме 2.2.

2.2. Для той же совокупности вычислите s^2 для всех простых случайных выборок объема 3 и проверьте, что $E(s^2) = S^2$.

2.3. Если из той же совокупности путем отбора с возвращением извлекаются выборки объема 2, покажите, найдя все возможные выборки, что $V(\bar{y})$ удовлетворяет уравнению

$$V(\bar{y}) = \frac{\sigma^2}{n} = \frac{S^2}{n} \cdot \frac{(N-1)}{N}.$$

Задача. В день через отделение таможни проходит в среднем 83 000 посылок. Для оценки общего веса груза была произведена простая случайная выборка 200 посылок. Получены следующие значения для их масс (в граммах):

$$\sum_{i=1}^{200} y_i = 9943; \sum_{i=1}^{200} y_i^2 = 500388.$$

Оцените общую массу грузов, обрабатываемых отделением за 1 день, и найдите стандартную ошибку этой оценки.