1 Статистика.

Постановка задачи. Пусть задана *генеральная совокупность* $x_1, \ldots, x_N$, которую следует понимать как набор данных, например, среднюю ежедневную температуру или значение индексов фондового рынка. Из генеральной совокупности отбирают n значений. Тем самым, образуется *выборка* объёма n. Выборка бывает *повторная* или *безповторная* в зависимости от того возвращается ли выбранный элемент в генеральную совокупность. Задача заключается в том, чтобы по выборке определить свойства генеральной совокупности.

Генеральную совокупность можно интепретировать как случайную величину X, которая принимает значения $x_1, \ldots, x_N$ с вероятностями, равными доле соответствующего значения. Например, если значение x_1 встречается в генеральной совокупности дважды, то $\mathsf{P}\{X = x_1\} = 2/N$. Таким образом, сформулированная выше задача формулируется как нахождение свойств неизвестного распределения по выборке объёма n. В такой формулировке не имеет значения, неизвестное распределение является дискретным или непрерывным.

Для генеральной совокупности (неизвестного распределения) определены среднее (математическое ожидание) и дисперсия. Обозначим их μ и σ^2.

Аналогично генеральной совокупности определяется случайная величина X^*, имеющая распределение выборки: $\mathsf{P}\{X^* = x_k\} = 1/n$. Естественно значения x_k могут повторяться.

Определение 1. *Среднее* и *дисперсия* выборки равны математическому ожиданию и дисперсии случайной величины X^*. Обобщая это определение, получим, что момент выборки порядка ν равен соответствующему моменту случайной величины X^*.

Определение 2. *Функция распределения выборки* — это функция распределения случайной величины X^*.

Пример. При десяти бросаниях игральной кости выпали числа: 3, 1, 2, 5, 2, 4, 1, 4, 3, 2. Тогда случайная величина X^* определяется таблицей

X^*	1	2	3	4	5
P	0.2	0.3	0.2	0.2	0.1

Среднее выборки равно 2.5.

1.1 Оценки

Выборке соответствует последовательность случайных величин X_1, ..., X_n, имеющих неизвестное распределение X. В случае повторной выборки случайные величины X_1, ..., X_n взаимно независимы.

Пусть $g()$ — произвольная функция n переменных. Тогда $g(x_1, \dots, x_n)$ определяет некоторую *характеристику* выборки. Распределение случайной величины $g(X_1, \dots, X_n)$ называется *выборочным распределением* характеристики $g(x_1, \dots, x_n)$.

Пример. Среднее и дисперсия выборки:

$$g(x_1, \dots, x_n) = \bar{x} = (x_1 + \ldots + x_n)/n,$$
$$g(x_1, \dots, x_n) = s^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

Для оценок характеристик генеральной совокупности используются различные функции $g(x_1, \dots, x_n)$. Например, для оценки среднего генеральной совокупности можно взять функцию $\bar{x}$. Можно ли взять другую функцию для оценки среднего генеральной совокупности? Да. Можно взять любую функцию, например, $g(x_1, \dots, x_n) = x_1$. Но эта функция в отличие от предыдущей представляет собой плохую оценку. Теперь пришло время обсудить, что такое “хорошая” и “плохая” оценки.

Определение 3. Оценка $\theta^* = g(x_1, \dots, x_n)$ некоторого параметра θ генеральной совокупности называется *состоятельной*, если случайная величина $g(X_1, \dots, X_n)$ сходится по вероятности к θ:

$$\mathsf{P}\left\{|g(X_1, \dots, X_n) - \theta| < \varepsilon\right\} \to 1$$

при $n \to \infty$ для любого $\varepsilon > 0$.

Теорема 1. *Оценка $s^2 = (1/n) \sum (x_i - \bar{x})^2$ является состоятельной.*

Теорема 2. *Пусть x_1, ..., x_n — выборка из распределения со средним μ и дисперсией σ^2. Тогда оценка $m = (1/n) \sum x_i$ среднего значения μ является состоятельной.*

Доказательство. Нужно показать, что

$$\mathsf{P}\left\{\left|\frac{1}{n} \sum X_i - \mu\right| < \varepsilon\right\} \to 1 \quad \forall \varepsilon > 0.$$

Однако это неравенство в точности является законом большим чисел для последовательности взаимно независимых случайных величин с одинаковым распределением. □

1.2 Несмещённая оценка

Определение 4. Оценка $g(x_1, \dots, x_n)$ параметра θ неизвестного распределения называется несмещённой, если математическое ожидание соответствующей случайной величины $g(X_1, \dots, X_n)$ равно θ:

$$\mathbf{M}\big(g(X_1, \dots, X_n)\big) = \theta.$$

Пример. Оценка $m = (1/n)\sum x_i$ среднего значения μ неизвестного распределения является несмещённой. Действительно,

$$\mathbf{M}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n}\sum \mathbf{M}X_i = \frac{1}{n}(n\mu) = \mu.$$

Пример. Дана выборка x_1, x_2 неизвестного распределения X. Рассмотрим в качестве оценки дисперсии σ^2 функцию $s^2 = 0.5(x_1 - \bar{x})^2 + 0.5(x_2 - \bar{x})^2$. Проверим является ли эта оценка несмещённой. Итак,

$$s^2 = 0.5\left(x_1 - \frac{x_1 + x_2}{2}\right)^2 + 0.5\left(x_2 - \frac{x_1 + x_2}{2}\right)^2 = \left(\frac{x_1 - x_2}{2}\right)^2.$$

Чтобы проверить несмещённость оценки требуется перейти к независимым случайным величинам X_1 и X_2, имеющих неизвестное распределение X.

$$\mathbf{M}\left(\frac{X_1 - X_2}{2}\right)^2 = \frac{\mathbf{M}X_1^2 - 2\,\mathbf{M}X_1\,\mathbf{M}X_2 + \mathbf{M}X_2^2}{4} = \\ = \frac{\mathbf{M}X^2 - (\mathbf{M}X)^2}{2} = \frac{\mathbf{D}X}{2} \neq \mathbf{D}X.$$

Значит, оценка не является несмещённой.

Теорема 3. *Пусть x_1, ..., x_n — выборка из распределения X с дисперсией σ^2. Тогда оценка*

$$\frac{n}{n-1}s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

является несмещённой оценкой дисперсии.

Доказательство. Сначала преобразуем выражение для s^2:

$$\sum(x_i - \bar{x})^2 = \sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \\ = \sum x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2.$$

В соответствии с определением заменим выборочные значения x_i на взаимно независимые случайные величины X_i, имеющие неизвестное распределение X и вычислим математическое ожидание случайной величины $\sum X_i^2 - n\big((1/n)\sum X_i\big)^2$:

$$\mathbf{M}\Big(\sum X_i^2 - n\big((1/n)\sum X_i\big)^2\Big) = \mathbf{M}\sum X_i^2 - \frac{1}{n}\mathbf{M}\Big(\sum X_i\Big)^2 = \\ = \mathbf{M}\sum X_i^2 - \frac{1}{n}\left(\mathbf{D}\Big(\sum X_i\Big) + \Big(\mathbf{M}\Big(\sum X_i\Big)\Big)^2\right).$$

Математическое ожидание суммы случайных величин равно сумме математических ожиданий. Аналогичное утверждение справедливо для дисперсий взаимно независимых случайных величин. Следовательно,

$$\mathbf{M}\Big(\sum X_i^2 - n\big((1/n)\sum X_i\big)^2\Big) = n\,\mathbf{M}X^2 - \frac{1}{n}\big(n\,\mathbf{D}X + (n\,\mathbf{M}X)^2\big) = \\ = n\big(\sigma^2 + \mu^2\big) - \frac{1}{n}\big(n\sigma^2 + n^2\mu^2\big) = (n-1)\sigma^2.$$

$\square$

Пример. Выборка $x_1, \ldots, x_n$ предназначена для оценки вероятности p выпадения единицы при бросании кости. Рассмотрим оценку p^* параметра p, равную доле тех значений выборки, для которых $x_i = 1$. Покажем, что оценка p^* несмещённая. В соответствии с определением нужно рассмотреть последовательность взаимно независимых случайных величин $X_1, \ldots, X_n$, имеющих одинаковое распределение $\mathsf{P}\{X_i = 1\} = p$, определить новую случайную величину Y как долю X_i, принимающих значение 1, и проверить равенство $\mathbf{M}Y = p$.

Положим $Y_i = 1$, если $X_i = 1$ и $Y_i = 0$, если $X_i \neq 1$. Тогда $\mathbf{M}Y_i = p$ и

$$Y = \frac{1}{n}\big(Y_1 + \ldots + Y_n\big).$$

По свойствам математического ожидания $\mathbf{M}Y = (1/n)(np) = p$.

1.3 Эффективные оценки

Пример. Рассмотрим выборку x_1, x_2 неизвестного распределения X. Пусть $m_1 = 0.2x_1 + 0.8x_2$ и $m_2 = 0.4x_1 + 0.6x_2$ — две оценки математического ожидания μ распределения X. Легко проверить, что обе оценки являются несмещёнными. Чтобы определить, какая из оценок "лучше", вычислим их дисперсии:

$$\mathbf{D}\big(0.2X_1 + 0.8X_2\big) = 0.2^2\sigma^2 + 0.8^2\sigma^2 = 0.68\sigma^2.$$

Аналогично
$$\mathbf{D}\big(0.4X_1 + 0.6X_2\big) = 0.52\sigma^2.$$
Дисперсия оценки m_2 меньше дисперсии m_1, и в это смысле оценка m_2 "лучше".

В общем случае мы хотим назвать оценку *эффективной*, если на ней реализуется наименьшая дисперсия среди всех возможных оценок.

Формальное определение. Оценка $\alpha^* = g(x_1, \ldots, x_n)$ параметра α называется эффективной, если дисперсия $\mathbf{D}\big(g(X_1, \ldots, X_n)\big)$ наименьшая среди дисперсий произвольных оценок $\mathbf{D}\big(h(X_1, \ldots, X_n)\big)$.

Приведенное определение сложно проверять на практике. Для практического определения эффективности потребуются две следующие теоремы.

Теорема 4. *Пусть α — оцениваемый параметр распределения X, имеющего плотность $f(x, \alpha)$. Тогда*
$$\mathbf{D}^2\big(g(X_1, \ldots, X_n)\big) \geqslant \frac{1}{n\int_{-\infty}^{\infty}\left(\frac{\partial \ln f}{\partial \alpha}\right)^2 f(x,\alpha)dx} \quad . \tag{1}$$

Теорема 5. *Оценка эффективна тогда и только тогда, когда неравенство (1) становится равенством.*

Замечание. Неравенство (1) можно уточнить:
$$\mathbf{M}\big(g(X_1, \ldots, X_n) - \alpha\big)^2 \geqslant \frac{1}{n\int_{-\infty}^{\infty}\left(\frac{\partial \ln f}{\partial \alpha}\right)^2 f(x,\alpha)dx} \quad ,$$
где α — истинное неизвестное значение параметра. Отсюда следует, что только несмещённые оценки эффективны.

Пример. Покажем, что оценка среднего нормального распределения
$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$
является эффективной. Действительно, плотность нормального распределения определяется формулой
$$f(x, \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$
Тогда $\ln f = \left(\ln \frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x-\mu)^2}{2\sigma^2}$. Вычислим производную:
$$\frac{\partial \ln f}{\partial \mu} = -\frac{x-\mu}{\sigma^2}.$$

Пользуясь полученным выражением, найдём знаменатель неравенства (1).

$$\int_{-\infty}^{\infty} \left(\frac{\partial \ln f}{\partial \mu}\right)^2 f dx = \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma^2}\right)^2 f dx = \frac{1}{\sigma^2}.$$

Последнее равенство доказано при выводе формулы дисперсии для нормального распределения.

Итак, неравенство (1) утверждает, что для любой оценки μ^* выполнено $\mathbf{D}^2(\mu^*) \geqslant \sigma^2/n$. Однако вычисления показывают, что для оценки $\bar{x}$ дисперсия в точности равна $(1/n^2)\sum \mathbf{D}X_i = \sigma^2/n$, что и доказывает эффективность оценки в соответствии с теоремой 5.

Замечание. Несмещённая оценка дисперсии нормальной совокупности $s^2 = (1/(n-1))\sum(x_i - \bar{x})^2$ не является эффективной, в то время как при известном математическом ожидании μ оценка $s_0^2 = (1/n)\sum(x_i - m)^2$ оказывается эффективной.

1.4 Выборочная доля

Пусть $x_1, \ldots, x_n$ — выборка из дискретного распределения X. Требуется оценить вероятность $p = \mathsf{P}\{X \in I\}$, где I — некоторый интервал значений случайной величины X. Например, требуется оценить вероятность выпадения единицы и двойки при бросании игральной кости. Естественная точечная оценка p^* параметра p состоит в вычислении доли значений выборки, принимающей значения из интервала I.

Рассмотрим последовательность взаимно независимых случайных величин $X_1, \ldots, X_n$, имеющих неизвестное распределение X. Назовём успехом событие заключающееся в том, что значение случайной величины X_i лежит на интервале I. Пусть S_n — общее количество успехов. Тогда распределением выборочной доли является распределение случайной величины S_n/n. Легко проверить, что

1. $\mathsf{P}\{|(S_n/n) - p| < \varepsilon\} \to 1$ (следует из закона больших чисел), поэтому точечная оценка p^* состоятельная;

2. $\mathbf{M}(S_n/n) = p$, поэтому оценка p^* несмещённая;

3. $\mathbf{D}(S_n/n) = p(1-p)/n$, поэтому оценка p^* асимптотически эффективная (обладает наименьшей дисперсией при $n \to \infty$).

1.5 Интервальная оценка

Центральная предельная теорема утверждает, что

$$\mathsf{P}\left\{\left|\frac{S_n}{n}-p\right|<\frac{\varepsilon\sigma}{\sqrt{n}}\right\}\to 2\Phi(\varepsilon),$$

где $\Phi(\varepsilon)=\int_0^\varepsilon \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$. Если дисперсия σ известна, то интервал

$$\left[p^*-\frac{\varepsilon\sigma}{\sqrt{n}}, p^*+\frac{\varepsilon\sigma}{\sqrt{n}}\right]$$

является доверительным для p с надёжностью $2\Phi(\varepsilon)$. Если дисперсия σ^2 неизвестна, то нужно оценить σ оценкой

$$\sigma^*=\sqrt{p^*(1-p^*)}.$$

Тогда доверительным интервалом для p с надёжностью $2\Phi(\varepsilon)$ оказывается интервал

$$\left[p^*-\varepsilon\sqrt{\frac{p^*(1-p^*)}{n}}, p^*+\varepsilon\sqrt{\frac{p^*(1-p^*)}{n}}\right]. \tag{2}$$

Пример. Среди тысячи деталей, отобранных из генеральной совокупности, оказалось 10 бракованных деталей. Нужно найти доверительный интервал доли бракованных деталей в генеральной совокупности с надёжностью 0.99.

По формуле (2)

$$\mathsf{P}\left\{|p^*-0.01|\leqslant\varepsilon\sqrt{\frac{0.01\cdot 0.99}{1000}}\right\}\approx 2\Phi(\varepsilon)=0.99.$$

По таблице функции $\Phi(\varepsilon)$ находим, что $\varepsilon\approx 2.6$. Тогда

$$|p^*-0.01|\leqslant 2.6\sqrt{\frac{0.01\cdot 0.99}{1000}}\approx 0.0035.$$

Следовательно, $0.0065\leqslant p^*\leqslant 0.0135$.

1.6 Дисперсия выборочной доли среднего бесповторной выборки

Бесповторная выборка x_1, …, x_n по прежнему рассматривается как значения случайных величин X_1, …, X_n. Однако случайные величины не являются независимыми. В силу симметрии распределения случайных величин одинаковы. А вычислить распределение проще всего для случайной величины X_1. Пусть ξ_1, …, ξ_k — различные значения генеральной совокупности. Положим N_i — количество значений ξ_i в выборке, $N = N_1 + \ldots + N_k$. Тогда распределение X_1 определяется таблицей

X	ξ_1	ξ_2	$\ldots$	ξ_k
P	N_1/N	N_2/N	$\ldots$	N_k/N

Рассмотрим оценку среднего:

$$m = g(x_1, \ldots, x_n) = \frac{x_1 + \ldots + x_n}{n}.$$

Математическое ожидание соответствующей выборочной характеристики

$$\mathbf{M}g(X_1, \ldots, X_n) = \frac{1}{n}\big(\mathbf{M}X_1 + \ldots + \mathbf{M}X_n\big) = \mathbf{M}X = \mu.$$

Следовательно, оценка является несмещённой.

Вычислим дисперсию $g(X_1, \ldots, X_n)$, пользуясь формулой дисперсии суммы случайных величин:

$$\mathbf{D}g(X_1, \ldots, X_n) = \frac{1}{n^2}\left(n\,\mathbf{D}X + n(n-1)\,\mathbf{Cov}(X_1, X_2)\right).$$

Вновь из-за симметрии совместное распределение любых пар (X_i, X_j) одинаковое. Оно задаётся формулой:

$$\mathsf{P}\left\{X_1 = \xi_i, X_2 = \xi_j\right\} = \begin{cases} \dfrac{N_i}{N}\dfrac{N_j}{N-1}, & \text{если } i \neq j \\ \dfrac{N_i}{N}\dfrac{N_i - 1}{N-1}, & \text{если } i = j. \end{cases}$$

Преобразуя вторую строчку определения, получим

$$\mathsf{P}\left\{X_1 = \xi_i, X_2 = \xi_j\right\} = \begin{cases} \dfrac{N_i}{N}\dfrac{N_j}{N-1}, & \text{если } i \neq j \\ \dfrac{N_i^2}{N(N-1)} - \dfrac{N_i}{N(N-1)}, & \text{если } i = j. \end{cases}$$

Напомним, что ковариация определяется соотношением: $\mathbf{Cov}(X_1, X_2) = \mathbf{M}\big(X_1X_2\big) - \mathbf{M}X_1\,\mathbf{M}X_2$. Так как $\mathbf{M}X_1 = \mathbf{M}X_2 = \sum \xi_i N_i/N$, то

$$\mathbf{M}X_1\,\mathbf{M}X_2 = \left(\sum \xi_i \frac{N_i}{N}\right)\left(\sum \xi_j \frac{N_j}{N}\right) = \sum \xi_i \xi_j \frac{N_i N_j}{N^2}. \tag{3}$$

Для вычисления математического ожидания произведения случайных величин X_1 и X_2 воспользуемся формулой их совместного распределения:

$$\mathbf{M}(X_1X_2) = \sum_{i,j=1}^{n} \mathsf{P}\left\{X_1 = \xi_i, X_2 = \xi_j\right\}\xi_i\xi_j =$$

$$= \sum \xi_i\xi_j \frac{N_iN_j}{N(N-1)} - \sum_i \xi_i^2 \frac{N_i}{N(N-1)}.$$

Во второй сумме $1/(N-1)$ можно вынести за знак суммы. Тогда под знаком суммы остаётся математическое ожидание $\mathbf{M}X^2$. Подставив полученные выражения в формулу для ковариации, получим:

$$\mathbf{Cov}(X_1, X_2) = \sum \xi_i\xi_j \frac{N_iN_j}{N}\left(\frac{1}{N-1} - \frac{1}{N}\right) - \frac{1}{N-1}\mathbf{M}X^2.$$

Непосредственные вычисления и формула (3) показывают, что

$$\mathbf{Cov}(X_1, X_2) = \sum \xi_i\xi_j \frac{N_iN_j}{N^2}\frac{1}{N-1} - \frac{1}{N-1}\mathbf{M}X^2 =$$

$$= \frac{1}{N-1}(\mathbf{M}X)^2 - \frac{1}{N-1}\mathbf{M}X^2 = -\frac{\sigma^2}{N-1}.$$

Тогда

$$\mathbf{D}\big(g(X_1, \ldots, X_n)\big) = \frac{1}{n}\left(\sigma^2 - \frac{n-1}{N-1}\sigma^2\right) = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

Контрольные вопросы

1. Объясните, что такое генеральная совокупность, повторная и бесповторная выборка.

2. Дайте определение среднего, дисперсии выборки и выборочной функции распределения. Вычислите их для выборки: 2, 5, 2, 1, 2. Дайте определение моментов порядка ν. Напишите формулу коэффициентов ассиметрии и эксцесса. Объясните с помощью графика, что означает положительный коэффициент ассиметрии.

3. Как вводятся основные характеристики выборки: среднее, дисперсия, центральные моменты высших порядков, коэффициенты ассиметрии и эксцесса? Какие из перечисленных характеристик остаются неизменными при линейных преобразованиях $x \to ax + b$?

4. Докажите, что оценка $m = (1/n)\sum x_i$ среднего значения μ неизвестного распределения является состоятельной.

5. Дана выборка x_1, x_2, x_3. Является ли оценка $0.25x_1 + 0.25x_2 + 0.25x_3$ несмещённой оценкой среднего генеральной совокупности? Ответ обоснуйте.

6. Пусть x_1, …, x_n — выборка из распределения с дисперсией σ^2. Докажите, что $s^2 = \big(1/(n-1)\big)\sum(x_i - \bar{x})^2$ является несмещённой оценкой дисперсии.

7. Даны две оценки $m_1 = 0.1x_1 + 0.9x_2$ и $m_2 = 0.5x_1 + 0.5x_2$ среднего значения μ неизвестного распределения. Являются ли эти оценки несмещёнными? У какой из оценок меньше диспресия?

8. Найдите значения параметра a, лежащие на отрезке $[0, 1]$, при которых оценка $ax_1 + (1-a)x_2$ среднего значения μ неизвестного распределения является несмещённой. Среди этих оценок найдите оценку с наименьшей дисперсией.

9. Докажите, что оценка $(1/n)\sum_{i=1}^{n} x_i$ математического ожидания нормально распределённой случайной величины является эффективной.

10. Вычислите математическое ожидание оценки $\bar{x}$ среднего бесповоторной выборки.

11. Докажите, что дисперсия выборочной доли равна $\frac{N-n}{N-1}\frac{\sigma^2}{n}$, где σ^2 — дисперсия генеральной совокупности.