Математическая статистика

Основные понятия

Предметом математической статистики является изучение свойств и характеристик случайных величин по результатам наблюдений. Последовательность значений изучаемой случайной величины, полученных в ходе ряда независимых испытаний, называется ***случайной выборкой.*** Элементы выборки часто называют ***вариантами***. Заметим, что выборку можно рассматривать как совокупность независимых случайных величин с одинаковыми распределениями. ***Генеральная совокупность*** – совокупность всех значений изучаемой случайной величины с частотами, равными их вероятностям. Фактически, понятие генеральной совокупности тождественно понятию случайной величины. Выборка является реализацией этой случайной величины. Задачи математической статистики состоят в получении информации о законе распределения генеральной совокупности по данным о случайной выборке. Количество элементов в генеральной совокупности и в выборке называются их ***объемами***. Для того, чтобы по выборке можно было адекватно судить о генеральной совокупности надо, чтобы элементы генеральной совокупности отбирались в выборку равновероятным образом. Такая выборка называется ***репрезентативной***. Вопрос о репрезентативности выборки может быть решен только в рамках предметной области, на основе анализа содержательной, а не формальной постановки задачи.

Пусть исследуется одномерная случайная величина. Тогда выборка объема n представляет собой совокупность n чисел. ***Ранжированием*** выборки называется упорядочение элементов выборки по возрастанию.

Для дискретной случайной величины, каждому ее значению x можно поставить в соответствие количество испытаний $n(x)$, в которых было получено это значение. Отношение $w = n(x)/N$ указанного количества испытаний к объему выборки N называется ***частотой варианта*** x в выборке. Набор пар $\{(x_i, w_i)\}$, $i = 1,\dots n$, в котором варианты ранжированы, называется ***вариационным рядом***. Естественно, $n \le N$. Графическое изображение вариационного ряда называется ***полигоном распределения***. Он представляет собой ломаную, соединяющие точки с координатами (x_i, w_i).

Для непрерывной случайной величины вводится аналогичное понятие ***интервального вариационного ряда*** $\{(I_i, w_i)\}$. Для этого все множество значений разбивается на несколько интервалов I_i (обычно, равной длины) и для каждого интервала вычисляется частота w_i попадания в этот интервал. Графическое изображение интервального вариационного ряда называется ***гистограммой***. Иногда строится и полигон, который в этом случае соединяет точки (x_i, w_i), где I_i – середина интервала I_i. При построении интервального вариационного ряда и гистограммы, рекомендуется число интервалов выбирать по эмпирической формуле Стерджерса $k \approx 1 + 1.4 \ln n$, где n – число наблюдений.

Можно определить функции на множестве случайных выборок. Такие функции называются ***статистическими характеристиками*** или коротко – ***статистиками***. Заметим, что статистика является случайной величиной. Примеры статистик для вариационного ряда $\{(x_i, w_i)\}$, $i = 1,\dots n$: выборочное среднее $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{n} x_i w_i$, выборочная дисперсия: $S^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right) = \sum_{i=1}^{n}\left(x_i - \bar{x}\right) w_i$. Для интервального вариационного ряда эти статистики имеют вид $\bar{x} = \sum_{i=1}^{n} x_i w_i$ и $S^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right) w_i$, где x_i – середина интервала I_i.

Точечные оценки

Пусть имеется некоторая случайная величина (генеральная совокупность), закон распределения которой зависит от некоторого параметра θ. ***Точечной оценкой*** параметра θ закона распределения случайной величины называется произвольная статистика $\bar{\theta}_n = \bar{\theta}_n(x_1, \dots, x_n)$.

Определение. Оценка $\bar{\theta}_n$ параметра θ называется ***несмещенной***, если $\mathbf{E}\bar{\theta}_n = \theta$

Определение. Оценка $\overline{\theta}_n$ параметра θ называется ***состоятельной***, если эта оценка стремится по вероятности к истинному значению параметра с ростом объема выборки, т.е. для любого $\varepsilon > 0$ $\lim\limits_{n\to\infty} p\left(\left|\overline{\theta}_n - \theta\right| < \varepsilon\right) = 1$

Определение. Несмещенная оценка $\overline{\theta}_n$ параметра θ называется ***эффективной***, если ее дисперсия $\mathbf{D}\overline{\theta}_n = \mathbf{E}\left(\overline{\theta}_n - \theta\right)^2$ является наименьшей среди дисперсий всех возможных несмещенных оценок параметра θ, вычисленных по выборке одного и того же объема.

Замечание 1: Несмещенность оценки гарантирует отсутствие систематических ошибок при вычислении параметра по данной статистической оценке. Чем меньше дисперсия оценки, тем меньше будет отклонение оценки, вычисленной по конкретной выборке от истинного значения. Поэтому, эффективность оценки характеризует ее близость к оцениваемому параметру, т.е. малость случайных ошибок. Наконец, состоятельность оценки говорит о том, что ростом числа измерений позволяет улучшить точность оценки. На практике, не всегда удается выполнить все три требования одновременно.

Замечание 2: Обычно, вопрос об эффективности оценки является наиболее трудным при исследовании свойств этой оценки. Для случая непрерывной случайной величины с плотностью распределения $\rho(x,\theta)$ для несмещенной оценки $\overline{\theta}_n$ параметра θ справедливо неравенство Рао-Крамера $\mathbf{D}\overline{\theta}_n \geq \dfrac{1}{n\mathbf{E}\left(\partial \ln \rho(x,\theta)/\partial\theta\right)^2}$. Поэтому, если в этом неравенстве справедливо точное равенство, то рассматриваемая оценка является эффективной.

Теорема 1. *Выборочное среднее* $\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$, вычисленное по выборке $\{x_1,\ldots,x_n\}$, является несмещенной и состоятельной оценкой для математического ожидания $\mathbf{E}x$.

Доказательство. Поскольку каждая из величин $x_1,\ldots,x_n$ имеет математическое ожидание $\mathbf{E}x$, математическое ожидание выборочного среднего $\mathbf{E}\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n}\mathbf{E}x_i = \mathbf{E}x$, т.е. выборочное среднее является несмещенной оценкой. В силу независимости величин $x_1,\ldots,x_n$, имеем $\mathbf{D}\overline{x} = \dfrac{1}{n^2}\sum\limits_{i=1}^{n}\mathbf{D}x_i = \dfrac{\mathbf{D}x}{n}$. Для доказательства состоятельности, запишем неравенство Чебышева для выборочного среднего с учетом несмещенности: $p(|\overline{x} - \mathbf{E}x| \geq \varepsilon) \leq \dfrac{\mathbf{D}\overline{x}}{\varepsilon^2} = \dfrac{\mathbf{D}x}{n\varepsilon^2}$ для всякого $\varepsilon > 0$. Поэтому $\lim\limits_{n\to\infty} p(|\overline{x}_n - \mathbf{E}x| \geq \varepsilon) = 0$. Отсюда $\lim\limits_{n\to\infty} p(|\overline{x} - \mathbf{E}x| < \varepsilon) = 1 - \lim\limits_{n\to\infty} p(|\overline{x} - \mathbf{E}x| \geq \varepsilon) = 1$.

Замечание. Приведенное здесь доказательство состоятельности выборочного среднего легко обобщается на любую несмещенную оценку, дисперсия которой стремится к нулю с ростом объема выборки.

Теорема 2. Выборочное среднее является эффективной оценкой для математического ожидания нормально распределенной случайной величины

Доказательство. Воспользуемся неравенством Рао-Крамера для оценки математического ожидания нормально распределенной случайной величины с плотностью вероятности $\rho(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$. Имеем $\mathbf{D}\overline{x}_n \geq \dfrac{1}{n\mathbf{E}\left(\partial \ln \rho(x,\mu)/\partial\mu\right)^2} = \dfrac{1}{n\mathbf{E}\left(\dfrac{x-\mu}{\sigma^2}\right)^2} = \dfrac{\sigma^2}{n}$. Как было показано выше, дисперсия выборочного среднего также равна $\mathbf{D}\overline{x} = \dfrac{\sigma^2}{n}$, что и доказывает его эффективность.

Теорема 3. *Выборочная дисперсия* $s^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2$, где $\overline{x}$ – выборочное среднее, не является несмещенной оценкой для дисперсии $\mathbf{D}x$.

Доказательство. Преобразуя формулу для выборочной дисперсии, получим: $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\left(x_i - \mathbf{E}x\right) - \left(\bar{x} - \mathbf{E}x\right)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbf{E}x\right)^2 - \frac{2}{n}\left(\bar{x} - \mathbf{E}x\right)\sum_{i=1}^{n}\left(x_i - \mathbf{E}x\right) + \left(\bar{x} - \mathbf{E}x\right)^2 =$

$= \left\{\sum_{i=1}^{n}\left(x_i - \mathbf{E}x\right) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\mathbf{E}x = n\cdot\left(\bar{x} - \mathbf{E}x\right)\right\} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbf{E}x\right)^2 - \left(\bar{x} - \mathbf{E}x\right)^2$. Отсюда ее математическое ожидание

$\mathbf{E}s^2 = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mathbf{E}x\right)^2\right] - \mathbf{E}\left(\bar{x} - \mathbf{E}x\right)^2 = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\cdot\mathbf{E}x + \left(\mathbf{E}x\right)^2\right] - \mathbf{D}\bar{x} =$

$= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{E}\left(x_i^2\right) - \left(\mathbf{E}x\right)^2\right] - \mathbf{D}\bar{x} = \mathbf{D}x - \mathbf{D}\bar{x} = \frac{n-1}{n}\mathbf{D}x \neq \mathbf{D}x$.

Замечание. Из доказательства теоремы 3 следует, что *исправленная выборочная дисперсия* $\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ является несмещенной оценкой для дисперсии $\mathbf{D}x$. Опираясь на замечание после теоремы 1, можно показать, что обе оценки s^2 и $\hat{s}^2$ являются состоятельными оценками дисперсии. Однако, эффективной оценка $\hat{s}^2$, вообще говоря, не является.

Метод максимального правдоподобия

Пусть $x_i, i = 1,\ldots,n$ – результаты n независимых наблюдений над случайной величиной x с плотностью вероятности (вероятностью значений для дискретной величины) $f(x,\theta)$, где θ – неизвестный параметр. ***Правдоподобием*** называется функция $L(\theta) = \prod_{i=1}^{n} f(x_i,\theta)$. Метод максимального правдоподобия состоит в том, что в качестве оценки параметра θ выбирается значение $\tilde{\theta}$, при котором $L(\tilde{\theta})$ достигает своего максимума.

Оценки, полученные методом максимального правдоподобия, всегда являются состоятельными, асимптотически (т.е. при объеме выборки стремящемся к бесконечности) несмещенными и асимптотически эффективными. Если эффективная оценка существует, то оценка, полученная методом максимального правдоподобия, будет эффективной.

Метод наименьших квадратов

Пусть известно, что измеряемые величины x и y связаны соотношением (*уравнение регрессии*), зависящим от вектора неизвестных параметров $\Theta = (\theta_1,\ldots,\theta_k)$: $y = f(x;\Theta) = f(x;\theta_1,\ldots,\theta_k)$. Рассмотрим эксперимент, в котором задается значение управляющей переменной x и измеряется соответствующее значение y. Будем предполагать, что ошибки эксперимента, т.е. случайная величина $\xi = y - f(x;\Theta)$, распределены по нормальному закону с нулевым математическим ожиданием. Результаты эксперимента представляют случайную выбору $(y_1,\ldots,y_n)$. Построим оценку для неизвестных параметров $\Theta = (\theta_1,\ldots,\theta_k)$. Функция правдоподобия равна $L(\theta_1,\ldots,\theta_k) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - f(x_i;\Theta)\right)^2\right]$. Очевидно, что максимизация этой функции эквивалентна минимизации суммы квадратов уклонений $\sum_{i=1}^{n}\left(y_i - f(x_i;\Theta)\right)^2$. Для минимизации этой величины надо решать систему уравнений $\frac{\partial}{\partial\theta_i}\sum_{i=1}^{n}\left(y_i - f(x_i;\Theta)\right)^2 = 0$

В качестве примера рассмотрим распространенный случай линейной регрессии $y = ax + b$ с искомыми параметрами $\Theta = (a,b)$. Система уравнений минимизации дает

$$\begin{cases} \sum_{i=1}^{n} (y_i - ax_i - b)\, x_i = 0 \\ \sum_{i=1}^{n} (y_i - ax_i - b) = 0 \end{cases}$$

Отсюда $a = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$, $b = \dfrac{\bar{y} \cdot \overline{x^2} - \overline{xy} \cdot \bar{x}}{\overline{x^2} - (\bar{x})^2}$

Интервальные оценки

Определение. ***Интервальной оценкой*** или ***доверительным интервалом*** для неизвестного параметра θ с ***уровнем надежности*** γ называется интервал (α, β), такой что вероятность $p(\alpha < \theta < \beta) = \gamma$. Здесь нижняя и верхняя границы доверительного интервала являются функциями выборки и уровня надежности: $\alpha = \alpha(x_1, \dots x_n, \gamma)$ и $\beta = \beta(x_1, \dots x_n, \gamma)$.

Лемма. Пусть независимые случайные величины $x_1, \dots x_n$ распределены нормально с нулевыми математическим ожиданиями и одинаковыми дисперсиями, а матрица $\mathbf{C}$ – *ортогональна*, т.е. обладает свойством $\mathbf{C}^T\mathbf{C} = \mathbf{E}$. Тогда компоненты случайные вектора $(y_1, \dots y_n) = (x_1, \dots x_n) \cdot \mathbf{C}$ также независимы и распределены по нормальным законам с одинаковыми параметрами.

Следствие. Произвольная линейная комбинация независимых случайных величин, распределенных по нормальному закону с одинаковыми параметрами, имеет нормальное распределение.

Теорема 4. Интервал $\left(\bar{x} - C_\gamma \dfrac{\sigma}{\sqrt{n}}; \bar{x} + C_\gamma \dfrac{\sigma}{\sqrt{n}}\right)$ является доверительным интервалом для математического ожидания нормально распределенной случайной величины с дисперсией σ^2 при уровне надежности γ и объеме выборки n. Здесь $\bar{x}$ – выборочное среднее, а C_γ – квантиль стандартного нормального распределения порядка γ.

Доказательство. Обозначим математическое ожидание по генеральной совокупности через μ. Выборочное среднее $\bar{x}$ является линейной комбинацией n независимых нормально распределенных случайных величин. Согласно следствию из леммы, такая линейная комбинация также распределена по нормальному закону. Как было показано ранее, ее математическое ожидание $\mathbf{E}\bar{x} = \mu$, а дисперсия $\mathbf{D}\bar{x} = \sigma^2/n$. Поэтому статистика $\eta = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ распределена по стандартному нормальному закону. Вероятность $p(|\eta| < \alpha) = \Phi(\alpha)$. При этом, по определению, если эта вероятность равна $\Phi(\alpha) = \gamma$, то $\alpha = C_\gamma$. Так что $p\left(\bar{x} - C_\gamma \dfrac{\sigma}{\sqrt{n}} < \mu < \bar{x} + C_\gamma \dfrac{\sigma}{\sqrt{n}}\right) = \gamma$, а это и есть утверждение теоремы.

Замечание. Важные значения квантилей: $C_{0.95} \approx 1.96$, $C_{0.99} \approx 2.58$

Теорема 5. Интервал $\left(\bar{x} - t_{n-1,p} \dfrac{\hat{s}}{\sqrt{n}}; \bar{x} + t_{n-1,p} \dfrac{\hat{s}}{\sqrt{n}}\right)$ является доверительным интервалом для математического ожидания нормально распределенной случайной величины с неизвестной дисперсией при уровне надежности p и объеме выборки n. Здесь $\bar{x}$ – выборочное среднее, $\hat{s}$ – исправленная выборочная дисперсия, а $t_{n-1,p}$ – квантиль порядка p распределения Стьюдента с $n-1$ степенями свободы.

Доказательство. Достаточно показать, что статистика $\eta = \dfrac{\bar{x} - \mu}{\hat{s}/\sqrt{n}}$ распределена по закону Стьюдента с $n-1$ степенями свободы. После этого доказательство теоремы 5 в точности повторяет доказательство теоремы 4. Введем величины $(y_1, \dots y_n)$, полученные из $(x_1 - \mu, \dots x_n - \mu)$ ортогональным преобразованием $\mathbf{C}$, причем

$y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu)$. Согласно лемме, величины $(y_1, \dots y_n)$ независимы и имеют нормальное распределение. Тогда для исправленной выборочной дисперсии получим

$$\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left((x_i - \mu) - (\bar{x} - \mu) \right)^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2 \right] = \frac{1}{n-1} \left[\sum_{i=1}^{n} y_i^2 - y_1^2 \right] = \frac{\sum_{i=2}^{n} y_i^2}{n-1}$$

Поэтому, рассматриваемую статистику можно записать в виде $\eta = \left[\frac{1}{n-1} \sum_{i=2}^{n} y_i^2 \right]^{-1/2} y_1$, а эта величина имеет распределение Стьюдента с $n-1$ степенями свободы.

Замечание. При числе степеней свободы стремящемся к бесконечности, распределение Стьюдента приближается к нормальному. Поэтому при числе наблюдений порядка 50 и более обычно используют квантили нормального распределения вместо распределении Стьюдента.

Теорема 6. Интервал $\left(\frac{(n-1)\hat{s}^2}{\chi^2_{n-1,(1+\gamma)/2}}; \frac{(n-1)\hat{s}^2}{\chi^2_{n-1,(1-\gamma)/2}} \right)$ является доверительным интервалом для дисперсии нормально распределенной случайной величины при уровне надежности γ и объеме выборки n. Здесь $\hat{s}$ – исправленная выборочная дисперсия, а $\chi^2_{n-1,p}$ $t_{n-1,p}$ – квантиль порядка p распределения χ^2 с $n-1$ степенями свободы.

Доказательство. Как и в теореме 5, доказательство сразу следует из того, что статистика $\frac{(n-1)\hat{s}^2}{\sigma^2}$ распределена по закону χ^2 с $n-1$ степенями свободы. А это в свою очередь следует из определения распределения χ^2.

Задание по теме «Статистические оценки»

1. В магазине за день было продано 45 пар обуви. Приводится выборка значений размеров проданных пар: 39, 41, 40, 42, 41, 40, 42, 44, 40, 43, 42, 41, 43, 39, 42, 41, 42, 39, 41, 37, 43, 41, 38, 43, 42, 41, 40, 41, 38, 44, 40, 39, 41, 40, 42, 40, 41, 42, 40, 43, 38, 39, 41, 41, 42. Построить дискретный вариационный ряд и полигон. Вычислить выборочные среднее и дисперсию, а также дисперсию выборочного среднего.
2. Результаты измерения отклонений от номинала диаметров 50 подшипников дали следующие численные значения:

−1,752	−0,291	−0,933	−0,450	0,512
−1,256	1,701	0,634	0,720	0,490
1,531	−0,433	1,409	1,730	−0,266
−0,058	0,248	−0,095	−1,488	−0,361
0,415	−1,382	0,129	−0,361	−0,087
−0,329	0,086	0,130	−0,244	−0,882
0,318	−1,087	0,899	1,028	−1,304
0,349	−0,293	−0,883	−0,056	0,757
−0,059	−0,539	−0,078	0,229	0,194
−1,084	0,318	0,367	−0,992	0,529

Построить интервальный вариационный ряд и гистограмму, рассчитав необходимое число интервалов по формуле Стерджерса. Вычислить выборочные среднее и дисперсию, а также дисперсию выборочного среднего.

3. Для определения среднего процентного содержания белка в зернах пшеницы было отобрано 626 зерен, исследование которых дало среднее выборочное $\bar{x} = 16.8\%$ и выборочную дисперсию $\hat{s}^2 = 4$. Указать доверительный интервал для искомой величины с уровнем надежности 0.988.
4. Выборочное среднее квадратичное отклонение (корень из выборочной дисперсии) некоторой величины, определенное по 9 измерениям, оказалось равно 10 см. Найти предельную ошибку измерения с надежностью 0.6.
5. В 10 квартирах одного дома расход электроэнергии за месяц составил (в $кВт \cdot ч$) 125, 78, 102, 140, 90, 45, 125, 115, 112. Определить с надежностью 95% доверительный интервал для оценки среднего расхода электроэнергии по квартирам этого дома.
6. Каков должен быть объем выборки, чтобы она позволила определить среднее значение с предельной ошибкой 0.1σ при уровне надежности 95%.
7. Пользуясь методом максимального правдоподобия получить оценки для среднего и дисперсии нормально распределенной случайной величины.
8. Методом наименьших квадратов построить линейное уравнение регрессии для данных, представленных таблицей

x	11.0	11.5	12.0	12.5	13.0	13.5
y	1.5	1.5	1.6	1.7	1.9	1.9

9. Методом наименьших квадратов подобрать квадратичную функцию $y = ax^2 + bx + c$ для данных, представленных таблицей

x	7	8	9	10	11	12	13
y	7.4	8.4	9.1	9.4	9.5	9.5	9.4

10. Методом наименьших квадратов подобрать степенную функцию $y = Ax^q$ для данных, представленных таблицей

x	1	2	3	4	5
y	7.1	27.8	62.1	110	161

11. Методом наименьших квадратов подобрать показательную функцию $y = Ae^{sx}$ для данных, представленных таблицей

x	0	2	4	6	8	10	12
y	1280	635	324	162	76	43	19