19. **Выборочная средняя, её свойства.**

Вы́борочное (эмпири́ческое) сре́днее — это приближение теоретического среднего распределения, основанное на выборке из него.

Определение: Пусть $X_1, \ldots, X_n$ — выборка из распределения вероятности, определённая на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathbb{P})$. Тогда её выборочным средним называется случайная величина. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Свойства выборочного среднего :

Пусть $\hat{F}(x)$ — выборочная функция распределения данной выборки. Тогда для любого фиксированного $\omega \in \Omega$ функция $\hat{F}(\omega, x)$ является (неслучайной) функцией дискретного распределения. Тогда математическое ожидание этого распределения равно $\bar{X}(\omega)$

Выборочное среднее — несмещённая оценка теоретического среднего:

$$\mathbb{E}\left[\bar{X}\right] = \mathbb{E}[X_i], \quad i = 1, \ldots, n.$$

Выборочное среднее — сильно состоятельная оценка теоретического среднего:

$\bar{X} \to \mathbb{E}[X_i]$ почти наверное при $n \to \infty$.

Выборочное среднее — асимптотически нормальная оценка. Пусть дисперсия случайных величин X_i конечна и ненулевая, то есть $\mathrm{D}[X_i] = \sigma^2 < \infty, \sigma^2 \neq 0,\ i = 1, \ldots, n$.

Тогда

$\sqrt{n}\left(\bar{X} - \mathbb{E}[X_1]\right) \to \mathrm{N}(0, \sigma^2)$ по распределению при $n \to \infty$,

где $\mathrm{N}(0, \sigma^2)$ — нормальное распределение со средним 0 и дисперсией σ^2.

Выборочное среднее из нормальной выборки — эффективная оценка её среднего

20. Выборочная дисперсия, её свойства.

Выборочная дисперсия в математической статистике — это оценка теоретической дисперсии распределения на основе выборки. Различают выборочную дисперсию и несмещённую, или исправленную, выборочные дисперсии.

Определения

Пусть $X_1, \ldots, X_n, \ldots$ — выборка из распределения вероятности. Тогда

Выборочная дисперсия — это случайная величина

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2,$$

где символ $\bar{X}$ обозначает выборочное среднее.

Несмещённая (исправленная) дисперсия — это случайная величина

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

Замечание

Очевидно,

$$S^2 = \frac{n}{n-1} S_n^2.$$

Свойства выборочных дисперсий

Выборочная дисперсия является теоретической дисперсией выборочного распределения. Более точно, пусть $\hat{F}(x)$ — выборочная функция распределения данной выборки. Тогда для любого фиксированного $\omega \in \Omega$ функция $\hat{F}(\omega, x)$ является (неслучайной) функцией дискретного распределения. Дисперсия этого распределения равна $S_n^2(\omega)$.

Обе выборочные дисперсии являются состоятельными оценками теоретической дисперсии. Если $\mathrm{D}[X_i] = \sigma^2 < \infty$, то $S_n^2 \xrightarrow{\mathbb{P}} \sigma^2$ и $S^2 \xrightarrow{\mathbb{P}} \sigma^2$,

где $\xrightarrow{\mathbb{P}}$ обозначает сходимость по вероятности.

Выборочная дисперсия является смещённой оценкой теоретической дисперсии, а исправленная выборочная дисперсия несмещённой:

$$\mathbb{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2,$$

и

$$\mathbb{E}\left[S^2\right] = \sigma^2.$$

Выборочная дисперсия нормального распределения имеет распределение хи-квадрат.
Пусть $X_i \sim \mathrm{N}(\mu, \sigma^2),\ i = 1, 2, \ldots$. Тогда

$$(n-1)\frac{S^2}{\sigma^2} \equiv n\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1)$$

---

21. Статистические оценки: несмещенные, эффективные, состоятельные

Состоятельной называют такую точечную статистическую оценку, которая при n стрем к бесконечн стремится по вероятности к оцениваемому параметру. В частности, если дисперсия несмещенной оценки при n стр к беск стремится к нулю, то такая оценка оказывается и состоятельной.

Рассмотрим оценку θ_n числового параметра θ, определенную при n = 1, 2, … Оценка θ_nназывается *состоятельной*, если она сходится по вероятности к значению оцениваемого параметра θ при безграничном возрастании объема выборки. Выразим сказанное более подробно. Статистика θ_n является состоятельной оценкой параметра θ тогда и только тогда, когда для любого положительного числа ε справедливо предельное соотношение

$$\lim_{n\to\infty} P\{|\theta_n - \theta| > \varepsilon\} = 0.$$

Пример 3. Из закона больших чисел следует, что $\theta_n = \bar{x}$ является состоятельной оценкой θ = *M(X)* (в приведенной выше теореме Чебышёва предполагалось существование дисперсии *D(X);* однако, как доказал А.Я. Хинчин [6], достаточно выполнения более слабого условия – существования математического ожидания *M(X)*).

Пример 4. Все указанные выше оценки параметров нормального распределения являются состоятельными.

Вообще, все (за редчайшими исключениями) оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются состоятельными.

Пример 5. Так, согласно теореме В.И. Гливенко, эмпирическая функция распределения$F_n(x)$ является состоятельной оценкой функции распределения результатов наблюдений$F(x)$

*Несмещенной называют такую точечную статистическую оценку Q*математическое ожидание которой равно оцениваемому параметру: M(Q*)=Q*

Второе важное свойство оценок – *несмещенность*. Несмещенная оценка θ_n – это оценка параметра θ, математическое ожидание которой равно значению оцениваемого параметра: $M(\theta_n) = \theta$.

Пример 6. Из приведенных выше результатов следует, что $\bar{x}$ и s_0^2 являются несмещенными оценками параметров m и σ^2 нормального распределения. Поскольку М($\tilde{x}$) = М(m**) = m, то выборочная медиана $\tilde{x}$ и полусумма крайних членов вариационного ряда m** - также несмещенные оценки математического ожидания mнормального распределения. Однако

$$M(s^2) \neq \sigma^2, \quad M[(\sigma^2)^{**}] \neq \sigma^2,$$

поэтому оценки s^2 и (σ^2)** не являются состоятельными оценками дисперсии σ^2нормального распределения.

Оценки, для которых соотношение $M(\theta_n) = \theta$ неверно, называются смещенными. При этом разность между математическим ожиданием оценки θ_n и оцениваемым параметром θ, т.е. $M(\theta_n) - \theta$, называется смещением оценки.

Пример 7. Для оценки s^2, как следует из сказанного выше, смещение равно

$$M(s^2) - \sigma^2 = - \sigma^2/n.$$

Смещение оценки s^2 стремится к 0 при $n \to \infty$.

Оценка, для которой смещение стремится к 0, когда объем выборки стремится к бесконечности, называется *асимптотически несмещенной*. В примере 7 показано, что оценка s^2 является асимптотически несмещенной.

Практически все оценки параметров, используемые в вероятностно-статистических методах принятия решений, являются либо несмещенными, либо асимптотически несмещенными. Для несмещенных оценок показателем точности оценки служит дисперсия – чем дисперсия меньше, тем оценка лучше. Для смещенных оценок показателем точности служит математическое ожидание квадрата оценки $M(\theta_n - \theta)^2$. Как следует из основных свойств математического ожидания и дисперсии,

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] = D(\theta_n) + (M(\theta_n) - \theta)^2, \quad (3)$$

т.е. математическое ожидание квадрата ошибки складывается из дисперсии оценки и квадрата ее смещения.

Для подавляющего большинства оценок параметров, используемых в вероятностно-статистических методах принятия решений, дисперсия имеет порядок $1/n$, а смещение – не более чем $1/n$, где n – объем выборки. Для таких оценок при больших n второе слагаемое в правой части (3) пренебрежимо мало по сравнению с первым, и для них справедливо приближенное равенство

$$d_n(\theta_n) = M[(\theta_n - \theta)^2] \approx D(\theta_n) \approx \frac{c}{n}, \quad c = c(\theta_n, \theta), \quad (4)$$

где c – число, определяемое методом вычисления оценок θ_n и истинным значением оцениваемого параметра θ.

Эффективной называют такую точечную статистическую оценку, которая при фиксированном n имеет наименьшую дисперсию.

С дисперсией оценки связано третье важное свойство метода оценивания –*эффективность*. Эффективная оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра.

Доказано [11], что $\bar{x}$ и s_0^2 являются эффективными оценками параметров m и σ^2нормального распределения. В то же время для выборочной медианы $\tilde{x}$ справедливо предельное соотношение

$$\lim_{n\to\infty} \frac{D(\bar{x})}{D(\tilde{x})} = \frac{2}{\pi} \approx 0{,}637.$$

Другими словами, эффективность выборочной медианы, т.е. отношение дисперсии эффективной оценки $\bar{x}$ параметра m к дисперсии несмещенной оценки $\tilde{x}$ этого параметра при больших n близка к 0,637. Именно из-за сравнительно низкой эффективности выборочной медианы в качестве оценки математического ожидания нормального распределения обычно используют выборочное среднее арифметическое.

Понятие эффективности вводится для несмещенных оценок, для которых $M(\theta_n) = \theta$ для всех возможных значений параметра θ. Если не требовать несмещенности, то можно указать оценки, при некоторых θ имеющие меньшую дисперсию и средний квадрат ошибки, чем эффективные.

Пример 8. Рассмотрим «оценку» математического ожидания $m_1 \equiv 0$. Тогда $D(m_1) = 0$, т.е. всегда меньше дисперсии $D(\bar{x})$ эффективной оценки $\bar{x}$. Математическое ожидание среднего квадрата ошибки $d_n(m_1) = m^2$, т.е. при $|m| < \sigma/\sqrt{n}$ имеем $d_n(m_1) < d_n(\bar{x})$. Ясно, однако, что статистику $m_1 \equiv 0$ бессмысленно рассматривать в качестве оценки математического ожидания m.

Пример 9. Более интересный пример рассмотрен американским математиком Дж. Ходжесом:

$$T_n = \begin{cases} \bar{x}, & |\bar{x}| > n^{-1/4}, \\ 0{,}5\bar{x}, & |\bar{x}| \le n^{-1/4}. \end{cases}$$

Ясно, что T_n – состоятельная, асимптотически несмещенная оценка математического ожидания m, при этом, как нетрудно вычислить,

$$\lim_{n\to\infty} n d_n(T_n) = \begin{cases} \sigma^2, & m \ne 0, \\ \dfrac{\sigma^2}{4}, & m = 0. \end{cases}$$

Последняя формула показывает, что при $m \ne 0$ оценка T_n не хуже $\bar{x}$ (при сравнении по среднему квадрату ошибки d_n), а при $m = 0$ – в четыре раза лучше.

Подавляющее большинство оценок θ_n, используемых в вероятностно-статистических методах, являются асимптотически нормальными, т.е. для них справедливы предельные соотношения:

$$\lim_{n\to\infty} P\left\{\frac{\theta_n - M(\theta_n)}{\sqrt{D(\theta_n)}} < x\right\} = \Phi(x)$$

для любого x, где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Это означает, что для больших объемов выборок (практически - несколько десятков или сотен наблюдений) распределения оценок полностью описываются их математическими ожиданиями и дисперсиями, а качество оценок – значениями средних квадратов ошибок $d_n(\theta_n)$.

22. Точечные и интервальные оценки.

Точечной оценкой неизвестного параметра называют число (точку на числовой оси), которое приблизительно равно оцениваемому параметру и может заменить его с достаточной степенью точности в статистических расчетах.

Для того чтобы точечные статистические оценки обеспечивали “хорошие” приближения неизвестных параметров, они должны быть несмещенными, состоятельными и эффективными.

Определение :Пусть $X_1, \ldots, X_n, \ldots$ — случайная выборка из распределения, зависящего от параметра $\theta \in \Theta$. Тогда статистику $\hat{\theta}(X_1, \ldots, X_n)$, принимающую значения в Θ, называют точечной оценкой параметра θЗамечание

Формально статистика $\hat{\theta}$ может не иметь ничего общего с интересующим нас значением параметра θ. Её полезность для получения практически приемлемых оценок вытекает из дополнительных свойств, которыми она обладает или не обладает.

Свойства точечных оценок

Оценка $\hat{\theta} = \hat{\theta}(X)$ называется несмещённой, если её математическое ожидание равно оцениваемому параметру генеральной совокупности:

$$\mathbb{E}_\theta\left[\hat{\theta}\right] = \theta, \quad \forall\theta \in \Theta,$$

где $\mathbb{E}_\theta$ обозначает математическое ожидание в предположении, что θ — истинное значение параметра (распределения выборки X).

Оценка $\hat{\theta}$ называется эффективной, если она обладает минимальной дисперсией среди всех возможных несмещенных точечных оценок.

Оценка $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ называется состоятельной, если она по вероятности с увеличением объема выборки n стремится к параметру генеральной совокупности: $\forall\theta \in \Theta$,

$\hat{\theta}_n \to \theta$ по вероятности при $n \to \infty$.

Оценка $\hat{\theta}_n$ называется сильно состоятельной, если $\forall\theta \in \Theta$,

$\hat{\theta}_n \to \theta$ почти наверное при $n \to \infty$.

Надо отметить, что проверить на опыте сходимость «почти наверное» не представляется возможным, поэтому с точки зрения прикладной статистики имеет смысл говорить только о сходимости по вероятности.

Интервальной называют оценку, которая определяется двумя числами – концами отрезка.

Интервальные оценки – характеризуют не единственно возможную ситуацию, а их множественность. Этот вид экспертных оценок широко распространен. Одним из определяющих свойств интервальной оценки является то, что на множестве задано бинарное отношение МЕЖДУ.

Определение

Пусть θ - неизвестный параметр генеральной совокупности. По сделанной выборке по определенным правилам находятся числа θ_1 и θ_2, такие чтобы выполнялось неравенство:

$$P\{\theta_1 < \theta < \theta_2\} = 1 - \alpha.$$

Интервал (θ_1, θ_2) является доверительным интервалом для параметра θ, а число $1-\alpha$ - *доверительной вероятностью* или *надежностью* сделанной оценки. Обычно надежность задается заранее, причем выбираются числа близкие к 1 (0.95, 0.99 или 0.999).

Примеры интервальных оценок

Пример 1. Доверительное оценивание по вариационному ряду.

Пусть задана выборка $X^n = (x_1, \dots, x_n)$ некоторой случайной величины X. Построим вариационный ряд выборки $x^{(1)} < \dots < x^{(n)}$:

Очевидно, что вероятность попасть в любой из $(n+1)$ - го интервалов значений случайной величины X одинакова и равна $\frac{1}{n+1}$. Тогда вероятность того, что случайная величина X приняла значение из интервала $(x^{(k)}, x^{(l)})$, где $l > k$ будет равна:

$$P_{X^n, x}\left\{x \in (x^{(k)}, x^{(l)})\right\} = \frac{l-k}{n+1}.$$

Вопрос: чему должен быть равен размер выборки n, чтобы вероятность попасть в интервал $(\min_i(x_i), \max_i(x_i))$ составила 95%.

Подставляя значение для доверительной вероятности в формулу выше, получим:

$$0.95 = P_{X^n, x}\left\{x \in (x^{(1)}, x^{(n)})\right\} = \frac{n-1}{n+1},$$

откуда $n = 39$.

Таким образом, при достаточном для заданной доверительной вероятности числе измерений случайной величины X по набору ее порядковых статистикможет быть оценен диапазон принимаемых ею значений.

Пример 2. Доверительный интервал для медианы.

Пусть задана выборка $X^n = (x_1, \cdots, x_n)$ некоторой случайной величины X.

При $n > 50$ доверительный интервал для медианы $\tilde{x}$ определяется порядковыми статистиками

$$x_k \le \tilde{x} \le x_{n-k+1},$$

где

$k = \frac{1}{2}(n - 1.64\sqrt{n} - 1)$ при $\alpha = 0.1$;

$k = \frac{1}{2}(n - 1.96\sqrt{n} - 1)$ при $\alpha = 0.05$;

$k = \frac{1}{2}(n - 2.58\sqrt{n} - 1)$ при $\alpha = 0.01$.

Для значений $n \le 50$ номера порядковых статистик, заключающих в себе медиану, при $\alpha = 0.05$ и $\alpha = 0.01$ приведены в таблице 1, взятой из [3].

Пример 3. Доверительный интервал для математического ожидания.

Пусть задана выборка $X^n = (x_1, \cdots, x_n)$ некоторой случайной величины X, арактеристики которой (дисперсия D и математическое ожидание M) неизвестны. Эти параметры оценим так:

$$M^* = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$D^* = \frac{\sum_{i=1}^{n} (x_i - M^*)^2}{n-1}$$ - несмещенная оценка дисперсии.

Величину $\sqrt{D^*}$ называют оценкой среднего квадратического отклонения. Воспользуемся тем, что величина M^* представляет собой сумму n независимых случайных величин, и, согласно центральной предельной теореме, при достаточно большом n ее закон близок к нормальному. Поэтому будем считать, что величина M^* распределена по нормальному закону. Характеристики этого закона - математическое ожидание и дисперсия - равны соответственно M (настоящее МО случайной величины X) и $\frac{D}{n}$.

Найдем такую величину δ, для которой $P(|M^* - M| < \delta) = \alpha$. Перепишем это в эквивалентном виде $P(\frac{|M^* - M|}{\sqrt{D/n}} < \frac{\delta}{\sqrt{D/n}}) = \alpha$ и скажем, что случайная величина перед знаком неравенства есть модуль от стандартной нормальной. Получаем, что $2\Phi(\frac{\delta}{\sqrt{D/n}}) - 1 = \alpha$, и $\delta = \sqrt{\frac{D}{n}} * \Phi^{-1}\left(\frac{\alpha+1}{2}\right)$. В случае неизвестной дисперсии ее можно заменить на оценку D^*.

Например, выбирая $\alpha = 0.05$, получаем коэффициент $\Phi^{-1}\left(\frac{\alpha+1}{2}\right) = 1.96$

Окончательно: с вероятностью α можно сказать, что $M \in \left(M^* - \Phi^{-1}\left(\frac{\alpha+1}{2}\right)\frac{D^*}{\sqrt{n}}, M^* + \Phi^{-1}\left(\frac{\alpha+1}{2}\right)\frac{D^*}{\sqrt{n}}\right)$

23. Точность и надежность оценки, доверительный интервал.