

Математическая статистика

14 июня 2020 г.

Содержание

1	Постановка задач математической статистики	2
1.1	Задачи теории вероятностей	2
1.2	Задачи математической статистики	2
2	Частота как оценка вероятности события и её свойства. Построение доверительного интервала для вероятности события на базе асимптотической нормальности частоты.	4
3	Постановка выборочной статистической модели. Точечная оценка параметра и характеристики.	6
4	Функции потерь и функции риска, состоятельность оценки характеристики, достаточное условие для состоятельности оценки.	7
5	Вид квадратичного риска в случае одномерной характеристики.	9
6	Постановка задачи доверительного оценивания, доверительный интервал.	10
7	Определение несмещенности и асимптотической нормальности оценки характеристики. Построение доверительного интервала для характеристики на базе асимптотической нормальности ее оценки.	11
8	Постановка задачи проверки гипотез	12
9	Ошибки первого и второго рода и их вероятности как критерий качества критерия (теста) проверки гипотез. Подход Неймана-Пирсона.	13
10	Эмпирическая функция распределения (ЭФР). Построение, свойства ЭФР при фиксированном значении аргумента (использовать свойства частоты).	14

1 Постановка задач математической статистики

Сравним задачи теории вероятностей и математической статистики

1.1 Задачи теории вероятностей

Заданы:

- Вероятностное пространство $\langle \Omega, \Sigma, P \rangle$.
- Случайная величина $X : \Omega \rightarrow \mathbb{R}^n$.

Требуется получить различного рода характеристики величины X и величин, получающихся из X .

1.2 Задачи математической статистики

Определение. Статистическим экспериментом называется четверка

$$\langle \mathcal{X}, \mathcal{A}, P_\theta, \Theta \rangle.$$

Здесь:

- \mathcal{X} – множество наблюдений.
- \mathcal{A} – σ -алгебра подмножеств \mathcal{X} .
- P_θ – известная с точностью до неизвестного параметра θ вероятностная мера – закон распределения наблюдаемых данных.
- Θ – множество допустимых значений неизвестного параметра, то есть $\theta \in \Theta$.

Задачей математической статистики является получение той или иной информации о законе распределения наблюдаемых данных $P = P_\theta$.

Определение. Статистикой называется измеримая функция

$$f : \mathcal{X} \rightarrow A.$$

Для произвольного A .

Определение. Пусть

$$\bar{X} = \langle X_1, \dots, X_n \rangle.$$

Где $X_i \sim X$ – одинаково распределенные случайные величины. Соответствующая модель называется моделью независимой однородной выборки.

Определение. Гипотезой H называется подмножество Θ :

$$H \subseteq \Theta.$$

Перечислим некоторые задачи математической статистики.

- Оценивание параметра θ или какой-либо функции $g(\theta)$, то есть построение статистики $\hat{g}: \mathcal{X} \rightarrow \Theta$. Оценивание может быть:
 - *точечным*, то есть указание численной оценки $g(\theta)$
 - *длительным*, то есть указание множества, с фиксированной вероятностью содержащего $g(\theta)$
- Проверка гипотез. Пусть имеется разбиение Θ на гипотезы: $\Theta = \bigsqcup_{n \in N} H_n$. Тогда проверкой гипотезы назовем построение *теста (критерия)*, то есть отображения

$$\varphi: \mathcal{X} \rightarrow N.$$

Которое по наблюдению выдает номер гипотезы, которому это наблюдение “соответствует”.

Естественно, перечисленные задачи можно оценивать с точки зрения качества. В этом смысле всегда требуется с точки зрения какой-либо метрики построить “лучшую” оценку.

2 Частота как оценка вероятности события и её свойства. Построение доверительного интервала для вероятности события на базе асимптотической нормальности частоты.

Теорема 2.1. (Яков, Бернулли)

Пусть имеется $\xi_i \sim \xi$ – последовательность одинаково распределенных и попарно независимых случайных величин. Пусть

$$\bar{\xi}_n = \frac{1}{n} \sum_{i=1}^n \xi_i = \frac{k_n}{n}.$$

Тогда

$$\bar{\xi}_n \xrightarrow[n \rightarrow +\infty]{} p.$$

Теорема 2.2. (Центральная предельная теорема, простейший вариант)

Пусть случайные величины $X_i \sim X$ независимы и одинаково распределены, причем $\exists E(X), D(X)$. Тогда для случайной величины

$$Y_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sigma(\bar{X}_n)}.$$

Верно:

$$F_{Y_n} \xrightarrow[\mathbb{R}]{} F_{N(0,1)}.$$

Теорема 2.3. (Свойства частоты как оценки p)

Пусть $\xi \sim B(p)$. Тогда

$$\hat{p} = \frac{k_n}{n}$$

Является несмещенной асимптотически нормальной оценкой p , то есть

$$E(\hat{p}) = p,$$

$$\sqrt{n} \cdot (\hat{p} - p) = Y_n \xrightarrow{P_{n,\theta}} Y \sim N(0, \Delta^2(p)), \Delta^2(p) = p(1-p).$$

Доказательство.

- Покажем несмещенность:

$$E(\hat{p}) = E\left(\frac{k_n}{n}\right) = \frac{1}{n}np = p.$$

- Асимптотическая нормальность с нормирующим множителем $\Delta^2(p) = p(1-p)$ следует непосредственно из центральной предельной теоремы.

■

На базе асимптотической нормальности можно построить доверительный интервал. Проделаем это на примере частоты. Выпишем определение асимптотической нормальности:

$$Y_n = \frac{\sqrt{n} \cdot (\hat{p} - p)}{\sqrt{p(1-p)}} \rightarrow N(0, 1).$$

Это буквально означает:

$$P_{n,\theta}(Y_n < t) \rightarrow F_{N(0,1)}(t).$$

Раскроем определение Y_n , возьмем его по модулю и воспользуемся квантилью:

$$P_{n,\theta} \left(\left| \frac{\sqrt{n} \cdot (\hat{p} - p)}{\sqrt{p(1-p)}} \right| < t_\gamma \right) \rightarrow \gamma \iff P_{n,\theta} \left(\frac{\sqrt{p(1-p)}}{\sqrt{n}} t_\gamma + \hat{p} > p > -\frac{\sqrt{p(1-p)}}{\sqrt{n}} t_\gamma + \hat{p} \right) \rightarrow \gamma.$$

Здесь $\gamma = P(|\xi| < t_\gamma)$, $\xi \sim N(0, 1)$. Построим д

3 Постановка выборочной статистической модели. Точечная оценка параметра и характеристики.

Определение. Напомним, что *точечной оценкой* параметра θ или какой-либо функции $g(\theta)$ называют численную оценку этой величины.

Пусть \hat{g}_n является некоторой точечной оценкой $g = g(\theta)$.

Определение. \hat{g}_n называется *несмещенной*, если $E(\hat{g}_n) = g(\theta)$.

Определение. \hat{g}_n называется *состоятельной*, если $\hat{g}_n \xrightarrow{p} g(\theta)$ при $n \rightarrow \infty$.

Определение. \hat{g}_n называется *асимптотически нормальной*, если

$$\frac{\sqrt{n}(\hat{g}_n - g(\theta))}{\sigma(g(\theta))} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

Определение. \hat{g}_n называется *эффективной* в классе оценок K , если для любой другой оценки $\hat{g}_n^* \in K$ имеет место неравенство:

$$E(\hat{g}_n - g(\theta))^2 \leq E(\hat{g}_n^* - g(\theta))^2.$$

4 Функции потерь и функции риска, состоятельность оценки характеристики, достаточное условие для состоятельности оценки.

Определение. Оценкой $g(\theta)$ называется статистика вида

$$\hat{g}: \mathcal{X} \rightarrow g(\Theta).$$

Определение. Пусть $\hat{g}(\theta)$ – оценка $g(\theta)$. Тогда функцией потерь называется неотрицательная функция $l(\hat{g}, g(\theta))$, характеризующая “близость” оценки к настоящему значению.

Замечание. Обычно в качестве функции потерь рассматривают функцию вида

$$l(\hat{g}, g(\theta)) = \omega(\|\hat{g}, g(\theta)\|).$$

Здесь ω – неотрицательная монотонно возрастающая функция, $\omega(0) = 0$.

Замечание. l является случайной величиной.

Определение. Риском называется функция

$$R(\hat{g}, \theta) \stackrel{\text{def}}{=} E_{\theta}(l(\hat{g}, g(\theta))).$$

Замечание. Риск – функция параметра θ и способа оценивания \hat{g} .

Опишем самые важные для нас виды функции потерь и риска.

Определение. Определим функцию потерь индикатором отклонений:

$$l^{\delta}(\hat{g}, g(\theta)) = \omega^{\delta}(\|\hat{g}, g(\theta)\|).$$

Где

$$\omega(t) = \mathbb{1}_{\delta}(t) = \begin{cases} 0, & t < \delta \\ 1, & t \geq \delta \end{cases}.$$

Соответствующий риск будет вероятностью отклонения:

$$R^{\delta}(\hat{g}, \theta) = E_{\theta}(l^{\delta}(\hat{g}, g(\theta))) = 0 \cdot P_{\theta}(\|\hat{g} - g(\theta)\| < \delta) + 1 \cdot P_{\theta}(\|\hat{g} - g(\theta)\| \geq \delta) = P_{\theta}(\|\hat{g} - g(\theta)\| \geq \delta).$$

Определение. При асимптотическом подходе оценка называется *состоятельной*, если

$$\forall \delta > 0 \quad R^{\delta}(\hat{g}_n, \theta) = P_{n,\theta}(\|\hat{g}_n - g(\theta)\| \geq \delta) \xrightarrow{n \rightarrow +\infty} 0.$$

Или, что то же самое:

$$\hat{g}_n \xrightarrow[n \rightarrow +\infty]{P_{n,\theta}} g(\theta).$$

Определение. Квадратичной функцией потерь называется функция

$$l_2(\hat{g}, g(\theta)) = \|\hat{g} - g(\theta)\|^2.$$

Соответствующий ей риск называется квадратичным:

$$R_2(\hat{g}, \theta) = E_\theta(\|\hat{g} - g(\theta)\|^2).$$

Теорема 4.1. (Достаточное условие для состоятельности оценки)

$R_2(\hat{g}_n, \theta) \xrightarrow[n \rightarrow +\infty]{} 0 \implies$ оценка состоятельна.

Доказательство.

$$\begin{aligned} \forall \delta > 0 \quad R^\delta(\hat{g}_n, \theta) &= P(\|\hat{g}_n - g(\theta)\| \geq \delta) = P(\|\hat{g}_n - g(\theta)\|^2 \geq \delta^2) \\ &\leq \frac{E_\theta(\|\hat{g}_n - g(\theta)\|^2)}{\delta^2} = \frac{R_2(\hat{g}_n, \theta)}{\delta^2} \xrightarrow[n \rightarrow +\infty]{} 0. \end{aligned}$$

■

5 Вид квадратичного риска в случае одномерной характеристики.

Определение. Смещением оценки называется величина

$$b(\hat{g}, \theta) = g(\theta) - E_{\theta}(\hat{g}).$$

Определение. Оценка называется несмещенной, если $b(\hat{g}, \theta) = 0$.

Теорема 5.1. $R_2(\hat{g}, \theta) = D_{\theta}(\hat{g}) + b^2(\hat{g}, \theta)$.

Доказательство.

$$\begin{aligned} R_2(\hat{g}, \theta) &= E_{\theta}(\|\hat{g} - g(\theta)\|^2) = E_{\theta}(\hat{g} - E_{\theta}(\hat{g}) - (g(\theta) - E_{\theta}(\hat{g})))^2 \\ &= E_{\theta}(\hat{g} - E_{\theta}(\hat{g}))^2 + (g(\theta) - E_{\theta}(\hat{g}))^2 - \underbrace{2(g(\theta) - E_{\theta}(\hat{g}))(E_{\theta}\hat{g} - E_{\theta}\hat{g})}_0 \\ &= D_{\theta}(\hat{g}) + b^2(\hat{g}, \theta). \end{aligned}$$

■

Следствие 5.2. Для одномерных несмещенных оценок квадратичный риск в точности равен дисперсии оценки:

$$R_2(\hat{g}, \theta) = D_{\theta}(\hat{g}).$$

6 Постановка задачи доверительного оценивания, доверительный интервал.

При оценивании параметров или характеристик распределений мы в качестве результата получаем числовое значение $\hat{g}(X) \in g(\Theta)$. Такой способ оценивания мы называем *точечной оценкой*. Заранее не понятно, насколько результат соответствует действительности. Для того, чтобы можно было оценивать качество результата, нужно предъявлять не точку, а подмножество в $g(\Theta)$, содержащее в некотором смысле наиболее подходящие значения.

Задача доверительного оценивания ставится следующим образом: задана величина $\gamma \in (0, 1)$, называемая *уровнем надежности*. По заданному наблюдению X и значению надежности требуется построить доверительную область надежности.

Определение. Доверительной областью надежности называется $\tilde{G}_\gamma \subseteq G = g(\Theta)$, обладающая свойством:

$$\forall \theta \in \Theta P_\theta(g(\theta) \in \tilde{G}_\gamma) \geq \gamma.$$

То есть множество, с достаточной вероятностью содержащее оцениваемую величину.

Определение. В случае одномерной оценки чаще всего доверительные области надежности выбирают в виде промежутков, которые называются *доверительными интервалами*.

Определение. В асимптотическом случае (когда имеется последовательность оценок и статистических экспериментов) последовательность *асимптотических областей надежности* $\tilde{G}_{n,\gamma}$ задается условием:

$$\forall \theta \in \Theta \lim P_{n,\theta}(g(\theta) \in \tilde{G}_{n,\gamma}) \geq \gamma.$$

Определение. Аналогично задается последовательность асимптотических доверительных интервалов в случае одномерной характеристики.

7 Определение несмещенности и асимптотической нормальности оценки характеристики. Построение доверительного интервала для характеристики на базе асимптотической нормальности ее оценки.

Определение. Напомним, оценка называется *несмещенной*, если

$$b(\hat{g}, \theta) = g(\theta) - E_{\theta}(\hat{g}) = 0.$$

Определение. Последовательность оценок \hat{g}_n называется *асимптотически нормальной*, если

$$\sqrt{n} \cdot (\hat{g}_n - g(\theta)) = Y_n \xrightarrow{P_{n,\theta}} Y \sim N(0, \Delta^2(\theta)).$$

Определение. Величина $\Delta(\theta)$ из определения асимптотически нормальной оценки называется *нормирующим множителем*.

Замечание. Определение асимптотически нормальной оценки можно переписать так:

$$\frac{\sqrt{n} \cdot (\hat{g}_n - g(\theta))}{\Delta(\theta)} \xrightarrow{P_{n,\theta}} Y \sim N(0, 1).$$

На базе асимптотической нормальности можно построить доверительный интервал. Выпишем определение асимптотической нормальности:

$$Y_n = \frac{\sqrt{n} \cdot (\hat{g} - g(\theta))}{\Delta(\theta)} \rightarrow N(0, 1).$$

Это буквально означает:

$$P_{n,\theta}(Y_n < t) \rightarrow F_{N(0,1)}(t).$$

Раскроем определение Y_n , возьмем его по модулю и воспользуемся квантилью:

$$P_{n,\theta}\left(\left|\frac{\sqrt{n} \cdot (\hat{g} - g(\theta))}{\Delta(\theta)}\right| < t_{\gamma}\right) \rightarrow \gamma \iff P_{n,\theta}\left(\frac{\Delta(\theta)}{\sqrt{n}}t_{\gamma} + \hat{g} > g(\theta) > -\frac{\Delta(\theta)}{\sqrt{n}}t_{\gamma} + \hat{g}\right) \rightarrow \gamma.$$

Здесь $\gamma = P(|\xi| < t_{\gamma})$, $\xi \sim N(0, 1)$.

8 Постановка задачи проверки гипотез

Определение. *Гипотезой* называется множество предполагаемых зафиксированных значений некоторого подмножества неизвестных параметров:

$$H : \theta \in \Theta_H \subseteq \Theta.$$

Определение. Гипотезу называют *простой*, если $|H| = 1$.

Определение. Гипотезу называют *сложной*, если $|H| > 1$.

Определение. Гипотезами *согласия* называют набор из двух гипотез: основной H_0 и альтернативы H_1 , причем $H_0 = \overline{H_1}$.

Определение. Правило принятия или отклонения основной гипотезы H_0 называют *тестом (критерием)* проверки гипотезы:

$$\varphi(X) : X_n \rightarrow \{0, 1\}.$$

При этом:

- $X_{n,0}$ называют *допустимым множеством*.
- $X_{n,1}$ называют *критическим множеством*.
- $X_{n,0} \sqcup X_{n,1} = X_n$.

Определение. Случайная величина $L(\overline{X}) : X_n \rightarrow \mathbb{R}$ называется *тестовой статистикой*, если она служит порогом для правила принятия или отклонения основной гипотезы:

$$\varphi(\overline{X}) = \begin{cases} 0, & L(\overline{X}) < T(H_0) \\ 1, & L(\overline{X}) \geq T(H_1) \end{cases}.$$

Где T называют *порогом принятия решения*.

9 Ошибки первого и второго рода и их вероятности как критерий качества критерия (теста) проверки гипотез. Подход Неймана-Пирсона.

Определение. *Ошибкой I рода* называют отклонение основной гипотезы, в то время как она была верна.

Определение. *Ошибкой II рода* называют принятие основной гипотезы, в то время как она не была верна.

Определение. α называют вероятностью ошибки I рода:

$$\alpha(\varphi, \theta) \stackrel{\text{def}}{=} P_{\theta}(X_{n,1}), \quad \theta \in \Theta_{H_0}.$$

Определение. *Уровнем значимости теста* называют верхнюю границу вероятности ошибки I рода по всем возможным наблюдаемым значениям неизвестных параметров, отвечающих основной гипотезе:

$$\alpha(\varphi) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta_{H_0}} \alpha(\varphi, \theta).$$

Определение. β называют вероятностью ошибки II рода:

$$\beta(\varphi, \theta) \stackrel{\text{def}}{=} P_{\theta}(X_{n,0}), \quad \theta \in \Theta_{H_1}.$$

Определение. *Мощностью теста* называют следующую величину:

$$\gamma(\varphi, \theta) \stackrel{\text{def}}{=} 1 - \beta(\varphi, \theta).$$

Подход Неймана-Пирсона. Зафиксируем $\alpha \in (0, 1)$ (обычно выбирают малое значение). Будем считать это значение минимальной допустимой величиной ошибки I рода (*допустимый уровень значимости*).

Рассмотрим множество всех тестов таких, что:

$$\bar{\Phi}_{\alpha} = \{\varphi = \varphi(x) \mid \alpha(\varphi) \leq \alpha\}.$$

Среди этих тестов выбирается тест с минимальным значением β .

В асимптотических задачах ограничения накладываются на предельные значения.

10 Эмпирическая функция распределения (ЭФР). Построение, свойства ЭФР при фиксированном значении аргумента (использовать свойства частоты).

Определение. Эмпирической функцией распределения (ЭФР) называют следующую оценку функции распределения генеральной совокупности:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}.$$

Иными словами, значение ЭФР в точке t равно отношению числа наблюдений, меньших t , к их общему числу n .

Свойства ЭФР:

1. ЭФР кусочно-постоянна.
2. Скачки ЭФР имеют вид $\frac{k}{n}$ для некоторого $k \in (1; n)$.
3. Область принимаемых значений: $[0; 1]$.
4. Частота может служить как оценка функции распределения генеральной совокупности. При фиксированном $t = t_0$:

$$F_x(t_0) \approx F_n(t_0) = \xi_1 + \dots + \xi_n = \frac{k_n}{n} - \text{частота}.$$

5. $F_n(t)$ является состоятельной оценкой:

$$F_n(t_0) = \bar{\xi}_n : F_n(t_0) \xrightarrow{p=1} F_x.$$

6. $F_n(t)$ является асимптотически нормальной оценкой.