

Математическая статистика

13 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	Функции потерь и функции риска, состоятельность оценки характеристики, достаточное условие для состоятельности оценки.	6
4	Вид квадратичного риска в случае одномерной характеристики.	8

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

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

$$\sqrt{n} \cdot (\hat{p} - p) = Y_n \xrightarrow[n \rightarrow +\infty]{} Y \sim N(0, \Delta^2(p)), \Delta^2(p) = p(1 - p).$$

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

3 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 3.1. (Достаточное условие для состоятельности оценки)

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

Доказательство.

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

■

4 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 4.1. $R_2(\hat{g}, \theta) = D_{\theta}(\hat{g}) + b^2(\hat{g}, \theta)$.

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

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