Математическая статистика

13 июня 2020 г.

Содержание

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 $X : \Omega \to \mathbb{R}^n$.

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

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

Для произвольного A.

Определение. Пусть

$$\overline{X} = \langle X_1, \ldots, X_n \rangle.$$

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 $\hat{g}: \mathcal{X} \to \Theta$. 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 проеркой гипотезы назовем построение *теста* (*критерия*), то есть отображения
$$\varphi: \mathcal{X} \to N.$$
  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

$$\overline{\xi}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_n = \frac{k_n}{n}.$$

Тогда

$$\overline{\xi}_n \underset{n\to+\infty}{\rightrightarrows} p.$$

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

$$Y_n = \frac{\overline{X}_n - E\left(\overline{X}_n\right)}{\sigma\left(\overline{X}_n\right)}.$$

Верно:

$$F_{Y_n} \underset{\mathbb{R}}{\rightrightarrows} F_{N(0,1)}.$$

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:

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

Это буквально означает:

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

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

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

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

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

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 являтся случайной величиной.

Определение. *Риском* называется функция

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

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

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

Соответствующий риск будет вероятностью отклонения:

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

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

$$\forall \delta > 0 \; R^\delta(\hat{g}_n, \theta) = P_{n,\theta}(\|\hat{g}_n - g(\theta)\| \geqslant \delta) \xrightarrow[n \to +\infty]{} 0.$$

Или, что то же самое:

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

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

Доказательство.

$$\begin{aligned}\forall \delta > 0 \; R^\delta(\hat{g}_n, \theta) = P(\|\hat{g}_n - g(\theta)\| \geqslant \delta) = P(\|\hat{g}_n - g(\theta)\|^2 \geqslant \delta^2)\\ \leqslant \frac{E_\theta(\|\hat{g}_n - g(\theta)\|^2)}{\delta^2} = \frac{R_2(\hat{g}_n, \theta)}{\delta^2} \xrightarrow[n\to+\infty]{} 0.\end{aligned}$$

■

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:

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

5 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 $\widetilde{G}_\gamma \subseteq G = g(\Theta)$, обладающая свойством:

$$\forall \theta \in \Theta \; P_\theta(g(\theta) \in \widetilde{G}_\gamma) \geqslant \gamma.$$

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* $\widetilde{G}_{n,\gamma}$ задается условием:

$$\forall \theta \in \Theta \;\; \lim P_{n,\theta}(g(\theta) \in \widetilde{G}_{n,\gamma}) \geqslant \gamma.$$

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

6 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:

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

Это буквально означает:

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

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

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

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