Домашнее задание №2 по курсу «Математическая Статистика в Машинном Обучении»

Школа Анализа Данных

Задачи

Задача 1 [4 балла]

Пусть n_1 — количество людей, которые получили лечение по методике 1, а n_2 — количество людей, которые получили лечение по методике 2. Обозначим через X_1 — количество людей, получивших лечение по методике 1, на которых эта методика повлияла положительно. Аналогично, обозначим через X_2 — количество людей, получивших лечение по методике 2, на которых эта методика повлияла положительно. Предположим, что $X_1 \sim \text{Binomial}(n_1, p_1)$ и $X_2 \sim \text{Binomial}(n_2, p_2)$. Положим $\psi = p_1 - p_2$.

(a) Найдите MLE-оценку ψ_{MLE} для параметра ψ.

(b) Найдите информационную матрицу Фишера $I(p_1, p_2)$.

(c) Используя многопараметрический дельта-метод найдите асимптотическую стандартную ошибку для ψ_{MLE}.

(d) Допустим, что $n_1 = n_2 = 200$, и конкретные значения случайных величин X_1 и X_2 равны 160 и 148 соответственно. Чему в этом случае равна оценка ψ_{MLE}. Найдите приблизительный (асимптотический) 90%-ый доверительный интервал для ψ, используя (а) многопараметрический дельта-метод и (б) параметрический бутстреп.

Задача 2 [2 балла]

Пусть $\boldsymbol{X} = \{X_1, \ldots, X_n\} \sim \text{Poisson}(\lambda)$.

- Постройте оценки $\tilde{\lambda}$ параметра λ с помощью метода моментов с использованием пробных функций $g_1(x) = x$ и $g_2(x) = x^2$.
- Постройте оценку $\hat{\lambda}$ параметра λ с помощью метода максимального правдоподобия. Найдите информацию Фишера $I_X(\lambda)$. Является ли оценка $\hat{\lambda}$ эффективной?

Задача 3 [4 балла]

Пусть $\boldsymbol{X} = \{X_1, \ldots, X_n\} \sim \text{Pareto}(\theta, \nu)$, $\theta > 0$, $\nu > 0$, с функцией плотности

$$f_{\theta,\nu}(x) = \begin{cases} \frac{\theta \nu^\theta}{x^{\theta+1}}, & x \geq \nu, \\ 0, & x < \nu \end{cases}$$

a) Найдите MLE-оценки $\hat{\theta}$ и $\hat{\nu}$ для параметров θ и ν.

c) Пусть параметр ν известен. Найдите истинные значения $\mathbb{E}_\theta[\hat{\theta}]$ и $\mathbb{V}_\theta[\hat{\theta}]$ как функции параметров θ, ν и размера выборки n. *Подсказка: следует использовать тот факт, что логарифм от случайной величины с распределением Парето, имеет экспоненциальное распределение.*

b) Пусть параметр ν известен. Найдите асимптотическое распределение оценки $\hat{\theta}$ с помощью дельта-метода.

d) Пусть параметр ν известен. Найдите информацию Фишера $I_X(\theta)$. Является ли MLE-оценка параметра $\hat{\theta}$ эффективной?

Задача 4 [4 балла]

Пусть $\boldsymbol{X} = \{X_1, \ldots, X_n\} \sim \text{Uniform}(0, \theta)$, $Y = \max\{X_1, \ldots, X_n\}$. Необходимо протестировать основную гипотезу $H_0 : \theta = 1/2$ против альтернативы $H_1 : \theta > 1/2$. В данном случае нельзя использовать тест Вальда, так как Y при $n \to \infty$ не сходится к нормальному распределению. Допустим, что мы будем использовать следующее правило: гипотеза H_0 отвергается, если $Y > c$.

(a) Найдите функцию мощности для данного теста.

(b) При каком значении параметра c размер теста будет равен 0.05?

(c) Каково значение p-value, если размер выборки $n = 20$ и $Y = 0.48$? Что можно сказать о гипотезе H_0?

(d) Каково значение p-value, если размер выборки $n = 20$ и $Y = 0.52$? Что можно сказать о гипотезе H_0?

Задача 5 [1 балл]

Пусть $\boldsymbol{X} = \{X_1, \dots, X_n\} \sim \text{Exp}(\theta)$. Постройте критерий отношения правдоподобий для проверки гипотезы $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$.

Задача 6 [3 балла]

Пусть $\boldsymbol{X} = \{X_1, \dots, X_n\} \sim \mathcal{N}(\mu, \sigma^2)$, где параметр μ известен. Требуется протестировать гипотезу $H_0 \colon \sigma = \sigma_0$ против альтернативы $H_1 \colon \sigma \neq \sigma_0$.

- Постройте критерий отношения правдоподобий для различения гипотез H_0 и H_1.
- Постройте критерий Вальда для различения гипотез H_0 и H_1.
- Сравните аналитически полученные критерии.

Примечание. Аналитическое сравнение тестов подразумевает доказательство их (асимптотической) эквивалентности или неэквивалентности, где под эквивалентностью понимается идентичность выносимых тестами решений.

Задача 7 [2 балла]

Пусть $\boldsymbol{X} = \{X_1, \dots, X_n\}$ — выборка н.о.р. с.в. со следующей функцией плотности:

$$f(x, \theta) = \begin{cases} c(\theta) d(x), & a \leqslant x \leqslant b(\theta) \\ 0, & \text{иначе} \end{cases}$$

где $b(\theta)$ — монотонно возрастающая функция одного аргумента.

(a) Построить статистику отношения правдоподобий λ для тестирования гипотезы $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$

(b) Найти распределение статистики λ при выполнении H_0 для следующей функции плотности:

$$f(x, \theta) = \begin{cases} \frac{2x}{\theta^2}, & 0 \leqslant x \leqslant \theta \\ 0, & \text{иначе} \end{cases}$$

Задача 8 [2 балла]

Найдите наилучшую критическую область (НКО) для проверки гипотезы $H_0 \colon \text{Uniform}[-a, a]$ против гипотезы $H_1 \colon \mathcal{N}(0, \sigma^2)$ по одному наблюдению ($n = 1$) при уровне значимости $\alpha = 0.1$. Найдите мощность полученного критерия.

Задача 9 [2 балла]

Проверяются гипотезы о плотности f распределения наблюдений $\boldsymbol{X} = \{X_1, \dots, X_n\}$: гипотеза $H_0 \colon f = f_0$ против альтернативы $H_1 \colon f = f_1$, где

$$f_1(x) = \begin{cases} 1, & x \in [0, 1], \\ 0, & x \notin [0, 1], \end{cases} \qquad f_2(x) = \begin{cases} 2x, & x \in [0, 1], \\ 0, & x \notin [0, 1]. \end{cases}$$

Построить наиболее мощный критерий размера α при $n = 1$ и $n = 2$.

Задача 10 [2 балла]

В процессе настольной игры у игроков возникло подозрение, что два кубика, которые шли в комплекте с игрой, несимметричны. Поэтому, начиная с некоторого момента, они начали записывать результаты бросков. В каждом броске участвуют оба кубика. Результаты приведены в таблице.

Сумма очков	2	3	4	5	6	7	8	9	10	11	12
Количество бросков	2	4	20	18	34	41	32	26	16	9	12

Проверьте гипотезу о том, что оба кубика симметричны на уровне значимости $\alpha = 0.05$. Найдите p-value.

Задача 11 [2 балла]

Предположим, что у нас есть 10 статей, написанных автором, скрывающемся под псевдонимом. Мы подозреваем, что эти статьи на самом деле написаны некоторым известным писателем. Чтобы проверить эту гипотезу, мы подсчитали доли четырехбуквенных слов в 8-и сочинениях подозреваемого нами автора:

$$.224 \ .261 \ .216 \ .239 \ .229 \ .228 \ .234 \ .216$$

В 10 сочинениях, опубликованных под псевдонимом, доли четырехбуквенных слов равны

$$.207\ .204\ .195\ .209\ .201\ .206\ .223\ .222\ .219\ .200$$

- Используйте критерий Вальда. Найдите p-value и 95%-ый доверительный интервал для разницы средних значений. Какой вывод можно сделать исходя из найденных значений?
- Используйте критерий перестановок. Каково в этом случае значение p-value. Какой вывод можно сделать?

Задача 12 [2 балла]

Маршрут грузового состава начинается в пункте A и последовательно проходит через пункты B_0, B_1 и т.д. По прибытии в очередной пункт те составы, которые направлялись в этот пункт, отцепляются. Очередной состав из 500 грузовых вагонов отправился из пункта A вдоль пунктов B_0, B_1, ... В таблице приведено количество отцепленных составов в каждом из пунктов (последним пунктом в данном случае оказался пункт B_9).

Пункт	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8	B_9
Количество составов	15	55	126	110	113	49	20	9	2	1

Возникло предположение, что распределение грузовых составов по пунктам назначения можно описать некоторым дискретным распределением, где $P(X = B_i)$ — вероятность того, что состав направляется в пункт B_i. В рамках данного предположения требуется провести проверку следующих гипотез на уровне значимости $\alpha = 0.05$ и найти p-value:

1. $\boldsymbol{X} \sim \text{Poisson}(\theta)$, т.е. $P(X = B_j) = e^{-\theta}\frac{\theta^j}{j!}$, где $j \geq 0$.
2. $\boldsymbol{X} \sim \text{Binomial}(m, p)$, т.е. $P(X = B_j) = C_m^j p^j (1-p)^{m-j}$, где $j \in \{0, \ldots, 9\}$ и $m = 9$.

Подсказка. Воспользуйтесь параметрическим критерием хи-квадрат.