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

Школа Анализа Данных

весна 2019

Задача 1 [5 баллов]

Пусть $\boldsymbol{X}^n = \{X_1, \ldots, X_n\} \sim \mathcal{N}(\theta, 1)$. Определим случайную величину Y, зависящую от $X \sim \mathcal{N}(\theta, 1)$ следующим образом.

$$Y = \begin{cases} 1, & \text{если } X > 0; \\ 0, & \text{если } X \leq 0. \end{cases}$$

Далее мы наблюдаем выборку $\boldsymbol{Y}^n$, по которой требуется оценить параметр $\psi = \mathsf{P}(Y = 1)$ распределения случайной величины Y.

a) Записать MLE-оценку ψ_{MLE} для ψ, в предположении знания выборки $\boldsymbol{X}^n$ и семейства породившего ее распределения $\mathcal{N}(\theta, 1)$.

b) Найти приближенный 95% доверительный интервал для ψ, воспользовавшись дельта-методом.

c) Пусть $\hat{\psi} = \langle \boldsymbol{Y}^n \rangle = n^{-1} \sum_{i=1}^n Y_i$. Доказать, что $\hat{\psi}$ является состоятельной оценкой для ψ.

d) Подсчитать асимптотическую относительную эффективность оценки $\hat{\psi}$ по сравнению с оценкой ψ_{MLE}. Стандартную ошибку оценки максимума правдоподобия для ψ_{MLE} предлагается взять из пункта b), где для ее нахождения использовался дельта-метод. После чего надо подсчитать стандартное отклонение величины $\hat{\psi}$.

e) Допустим, что случайные величины $X_1, \ldots, X_n$ на самом деле не распределены нормально. Показать, что в таком случае ψ_{MLE} не является состоятельной оценкой. Будет ли, и если ответ "да", то к чему, сходится при $n \to \infty$ оценка ψ_{MLE} в смысле какой-нибудь сходимости?

Задача 2 [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 3 [3 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \dots, X_n\} \sim \text{Uniform}(\theta, \theta+1)$. Необходимо протестировать гипотезу $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. В данном случае нельзя использовать тест Вальда, так как оценки θ при $n \to \infty$ не сходятся к нормальному распределению. Будем использовать следующее правило: гипотеза H_0 отвергается, если $X_{(n)} \geq 1$ или $X_{(1)} \geq c$, где c — некоторая константа, $X_{(1)} = \min\{X_1, \dots, X_n\}$, $X_{(n)} = \max\{X_1, \dots, X_n\}$.

(a) Найдите функцию мощности для данного теста.

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

(c) Найдите такое $n \geq 1$, что при $\theta = 0.1$ и размере теста 0.05 мощность критерия не меньше 0.8.

Задача 4 [2 балла]

Пусть $\{X_1, \dots, X_n\}$ — 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 5 [2 балла]

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

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

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

b) Используя дельта-метод найдите асимптотическое распределение для $\hat{\theta}$ при известном ν.

Задача 6 [2 балла]

В десятичной записи числа π среди первых 10002 знаков после запятой цифры $0, 1, \dots, 9$ встречаются соответственно 968, 1026, 1021, 974, 1014, 1046, 1021, 970, 948, 1014 раз. Можно ли при уровне значимости 0.05 считать эти цифры случайными? При каком уровне значимости эта гипотеза отвергается?

Задача 7 [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 можно сделать вывод исходя из найденных значений?
- Используйте критерий перестановок. Каково в этом случае значение p-value. Какой можно сделать вывод?

Задача 8 [2 балла]

Говорят, Джордж Р.Р. Мартин, автор цикла “Песнь Льда и Огня”, истребляет Старков: чаще “убивает” персонажей, относящихся к этому дому, чем персонажей других домов. В таблице 1 приведено количество персонажей, относящихся к тому или иному дому, упомянутых за первые 4 книги, а так же количество погибших персонажей.

Дом	Упомянутые персонажи	Погибшие персонажи
House Stark	72	18
House Lannister	49	11
House Greyjoy	41	12
Night’s Watch	105	41

Таблица 1: Данные взяты из датасета по ссылке

(a) Предлагается протестировать отличие уровня смертности дома Старков от уровня смертности каждого из других домов на 5% уровне значимости. Необходимо привести значения оценок вероятностей смертельных исходов для всех домов, а также найти p-value для каждого их трех тестов.

(b) Требуется протестировать множественную гипотезу “смертность дома Старков не отличается от смертности ни одного из домов”. Для этого предлагается воспользоваться методом Бонферрони. Также требуется построить тест по методу Benjamini-Hochberg обеспечивающий FDR не больше 5%.

Задача 9 [2 балла]

Девочка каждый будний день путешествует в метро от станции A до станции B. Со станции A составы идут в двух направлениях: до станции B и до станции C. Если приходит поезд до станции C, Девочке приходится дожидаться следующего поезда. Девочке кажется, что ей очень везёт с поездами до станции B и что поезда до станции B ходят чаще, чем поезда до станции C, поэтому на протяжении двух месяцев записывает, до какой станции идут поезда, которые она успела увидеть, спустившись на станцию (таблица `trains.csv`). Необходимо проверить, ходят ли поезда до станции B значимо чаще, чем до станции C.

Обозначим реальную частоту поездов до станции B через p и будем считать, что поезда приходят случайно и независимо друг от друга. Необходимо:

(a) Построить тест на основе критерия отношения правдоподобий для различения гипотез $H_0\colon p = p_0$ vs. $H_1\colon p \neq p_0$.

(b) Пусть $p_0 = \frac{1}{2}$. Сравнить (как аналитически, так и экспериментально) полученный тест с тестом Вальда для различения этих гипотез.

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

Разбаловка

- Задача 1. **5 балла**.
  - **1 балл**. Найдена MLE-оценку ψ_{MLE} для ψ.
  - **1 балл**. Найти приближенный 95% доверительный интервал для ψ.
  - **1 балл**. Доказано, что $\hat{\psi}$ является состоятельной оценкой для ψ.
  - **1 балл**. Подсчитана асимптотическая относительная эффективность (ARE) оценки $\hat{\psi}$ по сравнению с оценкой ψ_{MLE}.
  - **1 балл**.
    - **0.5 балла** Показано, что если $\boldsymbol{X}^n \not\sim \mathcal{N}(\theta, 1)$, то ψ_{MLE} не является состоятельной оценкой.
    - **0.5 балла** за обоснованный ответ на вопрос «Будет ли, и если ответ «да», то к чему, сходится при $n \to \infty$ оценка ψ_{MLE} в смысле какой-нибудь сходимости?».

- Задача 2. **4 балла**.
  - **1 балл**. Найдена MLE-оценку ψ_{MLE}.
  - **1 балл**. Вычислена информационная матрица Фишера $I(p_1, p_2)$.
  - **1 балл**. С помощью многопараметрического дельта-метода найдена асимптотическая ошибка для ψ_{MLE}.
  - **1 балл**. Найдены 90%-ые доверительные интервалы для ψ, с помощью многопараметрического дельта-метода и параметрического бутстрепа.
- Задача 3. **3 балла**.
  - **1 балл**. Найдите функцию мощности.
  - **1 балл**. Найдено критическое значение $c(\alpha)$ для $\alpha = 0.05$.
  - **1 балл**. Верно найдено n.
- Задача 4. **2 балла**.
  - **1 балл**. Построена статистика LRT.
  - **1 балл**. Указано название распределения для LRT.
- Задача 5. **2 балла**.
  - **1 балл**. Найдены MLE оценки.
  - **1 балл**. Найдено распределение $\widehat{\theta}$.
- Задача 6. **2 балла**.
  - **1 балл**. «Можно ли при уровне значимости 0.05 считать эти цифры случайными?»
  - **1 балл**. «При каком уровне значимости эта гипотеза отвергается?»
- Задача 7. **2 балла**.
  - **1 балл**. Найдено p-value для критерия Вальда, и построен 95%-ый доверительный интервал. Сделан вывод исходя из найденных значений.
  - **1 балл**. Найдено p-value для критерия перестановок. Сделан вывод исходя из найденного значения.
- Задача 8. **2 балла**.
  - **1 балл**. Сделан вывод об "предрасположенности" Д. Мартина у убийству персонажей из дома Старков на 5% уровне значимости. Найдены оценки вероятностей избавиться от персонажей различных домов.
  - **1 балл**. Проведено множественное тестирование с помощью методов Бонферрони (**0.5 балла**) и Benjamini-Hochberg (**0.5 балла**).
- Задача 9. **2 балла**.
  - **1 балл**. Построен тест для различения гипотез $H_0\colon p = p_0$ vs. $H_1\colon p \neq p_0$.
  - **1 балл**. Проведено сравнение полученного теста с тестом Вальда для различения гипотез H_0 и H_1.