Разбор демо-версии контрольной работы.

Андрей Тищенко @AndrewTGk

2024/2025

Задача 1

Выборка $X = (X_1, \ldots,\ X_n)$, порождённая случайной величиной $\xi : f_\xi(x) = \begin{cases} \frac{2x}{\theta}\exp\left(-\frac{x^2}{\theta}\right), & x > 0 \\ 0, & x \leqslant 0 \end{cases}$.
Найти оценку максимального правдоподобия, доказать её несмещённость и состоятельность.

Нахождение ОМП

Найти оценку максимального правдоподобия дляч параметра θ, доказать её несмещённость и состоятельность.
Необходимо построить функцию правдоподобия $L(x,\ \theta)$ (здесь x — вектор-реализация X).
По определению функцией правдоподобия вектора X с реализацией x, порождённого случайной величиной ξ называется:

$$L(x,\ \theta) = \begin{cases} \prod\limits_{i=1}^{n} f_\xi(x_i,\ \theta), & \xi - \text{непрерывная} \\ \prod\limits_{i=1}^{n} P(\xi = x_i,\ \theta), & \xi - \text{дискретная} \end{cases}$$

В нашем случае ξ непрерывна:

$$L(x,\ \theta) = \prod_{i=1}^{n} \frac{2x_i}{\theta}\exp\left(-\frac{x_i^2}{\theta}\right) = \left(\frac{2}{\theta}\right)^n \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} x_i^2\right)\prod_{i=1}^{n} x_i$$

Теперь наша задача — найти такую θ, при которой $L(x,\ \theta)$ достигает своего максимума (при этом вектор x фиксирован, значит в нашей власти только менять значение θ), сделав это, мы получим некое $\hat{\theta}$, при котором максимальна вероятность получения именно вектора x среди всех реализаций.

Однако сама функция выглядит очень страшно и искать максимум произведения я точно не буду, поэтому вспоминаем, что логарифм является монотонной функцией (значит его максимум будет достигаться при том же аргументе, что и у функции под логарифмом), позволяющей избавляться от произведений. То есть мы можем записать:

$$\underset{\theta}{\operatorname{argmax}}\big(L(x,\ \theta)\big) = \underset{\theta}{\operatorname{argmax}}\Big(\ln\big(L(x,\ \theta)\big)\Big)$$

Нагло воспользуемся данным фактом и логарифмируем функцию правдоподобия:

$$\ln\big(L(x,\ \theta)\big) = n\ln\left(\frac{2}{\theta}\right) - \frac{1}{\theta}\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\ln x_i = n\ln\left(\frac{2}{\theta}\right) - \frac{n}{\theta}\hat{\mu}_2 + \sum_{i=1}^{n}\ln x_i$$

Теперь, чтобы найти максимум по θ, стоит взять частную производную по θ:

$$\frac{\partial}{\partial\theta}\ln\big(L(x,\ \theta)\big) = n\frac{\theta}{2}\left(\frac{\partial}{\partial\theta}\frac{2}{\theta}\right) + \frac{n}{\theta^2}\hat{\mu}_2 + 0 = -\frac{n}{\theta} + \frac{n}{\theta^2}\hat{\mu}_2$$

Точки экстремума находятся в точках, где производная принимает значение 0, значит $\hat{\theta}$ удовлетворяет равенству:

$$-\frac{n}{\hat{\theta}} + \frac{n}{\hat{\theta}^2}\hat{\mu}_2 = 0 \Leftrightarrow \frac{\hat{\mu}_2}{\hat{\theta}^2} - \frac{\hat{\theta}}{\hat{\theta}^2} = 0 \overset{\hat{\theta}\neq 0}{\Leftrightarrow} \hat{\theta} = \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Итак, оценка максимального правдоподобия получилась $\hat{\theta} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i^2$

Проверка несмещённости

Оценка $\hat{\theta}$ параметра θ называется несмещённой, если выполняется равенство:

$$E\hat{\theta} = \theta$$

Давайте это проверим:

$$E\hat{\theta} = E\frac{1}{n}\sum_{i=1}^{n} x_i^2 = E\left(x_1^2\right) = E(\xi^2)$$

На всякий случай напомню как нужно считать математическое ожидание случайной функции:

$$E(g(\xi)) = \int\limits_{-\infty}^{+\infty} g(x) \cdot f_\xi(x)\, dx$$

В нашем случае $g(\xi) = \xi^2 \Rightarrow$ нужно посчитать интеграл:

$$\int\limits_{-\infty}^{+\infty} x^2 f_\xi(x)\, dx$$

Поскольку $f_\xi(x)$ равняется нулю при $x \leqslant 0$, можем сузить область интегрирования до $(0,\ +\infty)$.

$$\int\limits_{0}^{+\infty} x^2 f_\xi(x)\, dx = \int\limits_{0}^{+\infty} \frac{2x^3}{\theta} \exp\left(-\frac{x^2}{\theta}\right)\, dx = \left\langle \begin{aligned} a &= \frac{x^2}{\theta} \\ da &= \frac{2x}{\theta} dx \end{aligned} \right\rangle = \int\limits_{0}^{+\infty} \theta \cdot a \cdot e^{-a}\, da =$$

$$= \theta \underbrace{\int\limits_{0}^{+\infty} a \cdot e^{-a}\, da}_{=1} = \theta$$

$\int\limits_{0}^{+\infty} a \cdot e^{-a}\, da$ равняется 1, потому что это математическое ожидание экспоненциального распределения с параметром 1 (то есть $1^{-1} = 1$).
Итак, получили $E\hat{\theta} = E\left(\xi^2\right) = \theta \Rightarrow$ оценка несмещённая.

Проверим состоятельность

Оценка называется <u>состоятельной</u>, если для неё выполняется:

$$\hat{\theta} \underset{p}{\to} \theta$$

Оценка называется <u>сильно состоятельной</u>, если для неё выполняется:

$$\hat{\theta} \xrightarrow[\text{п. н.}]{} \theta$$

Для независимых, одинаково распределённых случайных величин $\eta_1, \ldots,\ \eta_n$ с конечным математическим ожиданием по теореме Колмогорова выполняется усиленный закон больших чисел, то есть:

$$\frac{1}{n}\sum_{i=1}^{n} \eta_i \xrightarrow[\text{п. н.}]{} E\eta_1$$

В нашем случае $\forall i \quad \eta_i \sim \xi^2 \Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \xi_i^2 \xrightarrow[\text{п. н.}]{} E\xi^2 = \theta$, то есть полученная нами оценка является сильно состоятельной (значит состоятельной также является).

Задача 2

Выборка $X_1, \ldots,\ X_n$ соответствует распределению Пуассона $\Pi(\theta)$ (пусть она порождается случайной величиной $\xi \sim \Pi(\theta)$). Пользуясь критерием эффективности, построить эффективную по Рао-Крамеру оценку параметра θ.

Решение

Чтобы воспользоваться критерием эффективности, стоит вспомнить определение функции вклада:

$$U(x,\ \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(x_i,\ \theta)$$

Теперь запишем сам критерий:

$$\theta \text{ — R-эффективная оценка } \theta \Leftrightarrow \hat{\theta} - \theta = \left(\mathcal{D}\hat{\theta}\right) \cdot U(x,\ \theta)$$

Как всем известно (я только что подсмотрел в лекцию):

$$\xi \sim \Pi(\theta) \Rightarrow P(\xi = k) = \frac{e^{-\theta}\theta^k}{k!}$$

Теперь возникает проблема: у дискретного распределения, каковым является $\Pi(\theta)$, не существует плотности распределения. В лекциях вроде не проговаривалось (однако в одном из ДЗ была на это задача), но в таких случаях стоит заменять плотность распределения на вероятность конкретного значения. То есть функция вклада будет такой:

$$U(x,\ \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln P(\xi = x_i,\ \theta)$$

Теперь нужно посчитать частную производную логарифма. Начнём с логарифмирования:

$$\ln P(\xi = x_i,\ \theta) = \ln\left(\frac{e^{-\theta}\theta^{x_i}}{x_i!}\right) = -\theta + x_i \ln\theta - \ln(x_i!)$$

Теперь возьмём частную производную:

$$\frac{\partial}{\partial\theta} \ln P(\xi = x_i,\ \theta) = -1 + \frac{x_i}{\theta} + 0 = \frac{x_i - \theta}{\theta}$$

Итак, функция вклада для распределения $\Pi(\theta)$:

$$U(x,\ \theta) = \sum_{i=1}^{n} \frac{x_i - \theta}{\theta}$$

Теперь мы хотим привести $\displaystyle\sum_{i=1}^{n} \frac{x_i - \theta}{\theta}$ к виду $\frac{1}{\mathcal{D}\hat{\theta}}\left(\hat{\theta} - \theta\right)$, делаем преобразования:

$$\sum_{i=1}^{n} \frac{x_i - \theta}{\theta} = \frac{1}{\theta}\left(n\overline{X} - n\theta\right) = \frac{n}{\theta}\left(\overline{X} - \theta\right)$$

Итак, получаем $\frac{n}{\theta} = \frac{1}{\mathcal{D}\hat{\theta}} \Rightarrow \mathcal{D}\hat{\theta} = \frac{\theta}{n},\ \hat{\theta} = \overline{X}$.
На всякий случай сделаем проверку:

$$\mathcal{D}\hat{\theta} = \mathcal{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\left(n\mathcal{D}x_1\right) = \frac{\theta}{n}$$

Действительно получили разложение $\mathcal{D}\hat{\theta} \cdot U(x,\ \theta) = \hat{\theta} - \theta \Rightarrow \hat{\theta} = \overline{X}$ является ответом к этой задаче (то есть эта оценка обладает наименьшей дисперсией среди всех несмещённых оценок).

Задача 3

В округе A было опрошено $n_1 = 100$ избирателей возраста 60+ и $n_2 = 200$ избирателей возраста $30 - 40$ лет.
Среди избирателей старшего возраста $N_1 = 62$ будут голосовать за NN, а среди избирателей среднего возраста за NN проголосуют $N_2 = 105$ человек. Построить асимптотический доверительный интервал уровня надёжности 0.95 для разности вероятностей поддержки кандидата NN среди избирателей среднего и старшего возраста.

Решение

Итак, есть две выборки:

$$\begin{cases} X = (X_1, \ldots,\ X_{n_1}), & \forall \xi \in X : \xi \sim Be(p_1) \\ Y = (Y_1, \ldots,\ Y_{n_2}), & \forall \eta \in Y : \eta \sim Be(p_2) \end{cases}$$

Нас просят найти интервал в который попадает $p_2 - p_1$ с вероятностью 0.95.
Поскольку p_1 и p_2 мы не знаем, их следует оценить. Лучшей оценкой параметра распределения Бернулли в нашем случае будут частоты:

$$\begin{cases} \hat{p}_1 = \frac{N_1}{n_1} \\ \hat{p}_2 = \frac{N_2}{n_2} \end{cases}$$

Замечание: конкретно в моей записи $\hat{p}_1,\ \hat{p}_2$ совпадают с $\overline{X},\ \overline{Y}$ соответственно.
На данный момент мы имеем оценку разности вероятностей: $\hat{p}_2 - \hat{p}_1$. Осталось её центрировать и нормировать, после чего результат будет асимптотически нормальным (там Муавры-Лапласы всякие). Для осуществления этих операций нужно посчитать математическое ожидание и дисперсию нашей оценки:

$$E(\hat{p}_2 - \hat{p}_1) = E(\hat{p}_2) - E(\hat{p}_1) = E\left(\frac{1}{n_2}\sum_{i=1}^{n_2} y_i\right) - E\left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_i\right) = p_2 - p_1$$

$$\mathcal{D}(\hat{p}_2 - \hat{p}_1) = \mathcal{D}\hat{p}_2 + \mathcal{D}\hat{p}_1 - \underbrace{2\operatorname{cov}(\hat{p}_2 - \hat{p}_1)}_{=0} = \mathcal{D}\left(\frac{1}{n_2}\sum_{i=1}^{n_2} y_i\right) + \mathcal{D}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_i\right) =$$

$$= \frac{1}{n_2^2}(n_2 \mathcal{D} y_1) + \frac{1}{n_1^2}(n_1 \mathcal{D} x_1) = \frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}$$

В $\mathcal{D}(\hat{p}_2 - \hat{p}_1)$ участвуют $p_1,\ p_2$, которые мы не знаем, там их тоже заменяем на оценку:

$$\hat{\mathcal{D}}(\hat{p}_2 - \hat{p}_1) = \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}$$

Теперь мы можем составить очень крутую случайную величину:

$$\frac{\hat{p}_2 - \hat{p}_1 - (p_2 - p_1)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_2 - \hat{p}_1)}}$$

Её математическое ожидание равно 0, а дисперсия равна 1, а ещё она асимптотически нормальная (можно поиграться с ЦПТ). Нам нужен доверительный интервал уровня 0.95, делаем:

$$P\left(Z_{0.025} < \frac{\hat{p}_2 - \hat{p}_1 - (p_2 - p_1)}{\sqrt{\hat{\mathcal{D}}(\hat{p}_2 - \hat{p}_1)}} < Z_{0.975}\right) = 0.95$$

$Z_{0.975} = -Z_{0.025} = 1.96$

$$P\left(\hat{p}_2 - \hat{p}_1 - Z_{0.975}\sqrt{\hat{\mathcal{D}}(\hat{p}_2 - \hat{p}_1)} < p_2 - p_1 < \hat{p}_2 - \hat{p}_1 + Z_{0.975}\sqrt{\hat{\mathcal{D}}(\hat{p}_2 - \hat{p}_1)}\right) = 0.95$$

Посчитал на калькуляторе:

$$P\left(-0.213 < p_2 - p_1 < 0.023\right) = 0.95$$

Искомый доверительный интервал: $(-0.213,\ 0.023)$

Задача 4

Среднее время сборки изделия $m_0 = 90$ минут. Предложили новый метод сборки, за $n = 6$ испытаний время сборки составило $X = (79,\ 74,\ 102,\ 95,\ 70,\ 90)$ минут. Можно ли считать, что время сборки в среднем сократилось? Предполагается, что время сборки имеет нормальное распределение. Уровень значимости 0.05.

Решение

Имеется два распределения $N(m,\ \sigma^2)$. Про дисперсию нам ничего не сказано, поэтому считаем её неизвестной.
Задача на проверку гипотез, значит сначала надо их сформулировать:

$$H_0 : m = m_0 = 90, \text{ против } H_1 : m < 90$$

Принятие H_0 даёт отрицательный ответ на вопрос задачи. Сделали так, чтобы сделать основную гипотезу простой (чтобы далее было проще выписывать статистику при условии этой самой H_0).
Теперь нужно составить статистику (некую функцию, которая при справедливости H_0 окажется центрированной и нормированной).
Осталось оценить m. Для оценки математического ожидания лучше всего подходит выборочное среднее, то есть m мы оценим $\overline{X}$. Значит наша статистика должна выглядеть вот так:

$$T(x) = \frac{\overline{X} - m_0}{\sqrt{\mathcal{D}\left(\overline{X}\right)}} = \frac{\sqrt{n}\left(\overline{X} - m_0\right)}{\sigma}$$

σ мы не знаем, поэтому стоит заменить на несмещённую выборочную дисперсию:

$$\tilde{S} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}$$

Настала пора считать:

$$\overline{X} = \frac{79 + 74 + 102 + 95 + 70 + 90}{6} = 85$$
$$\tilde{S} = \sqrt{\frac{1}{5}\sum_{i=1}^{6}(x_i - 85)^2} \approx 12.617$$
$$\sqrt{n} = \sqrt{6} \approx 2.449$$

Итого, получаем статистику:

$$T(x) = \frac{\sqrt{n}(\overline{X} - m_0)}{\tilde{S}}\Big|_{H_0} \sim t(n-1) = t(5)$$

Определим критическую область. H_1 хочет маленьких значений m. Если m уменьшается, значит уменьшается и $\overline{X} \Rightarrow$ значения статистики будут сильно отрицательными. Получается, что критической областью является $(-\infty,\ t_{5,\ 0.05}) = (-\infty,\ -2.015)$. Посчитаем статистику:

$$T(x) = \frac{2.449 \cdot (85 - 90)}{12.617} \approx -0.971$$

Попали в доверительную область, значит принимаем H_0.

Ответ

Так считать нельзя.

Задача 5

В тесте по английскому было $n = 100$ вопросов с 4 вариантами ответа (1 из них правильный). Некий студент правильно ответил на $N = 30$ вопросов. Можно ли при $\alpha = 0.05$ считать, что этот студент не знает предмет?

Решение

Дана выборка $X = (X_1, \ldots,\ X_{100})$, $\forall \xi \in X \quad \xi \sim Be(p)$, количество правильных ответов тогда имеет распределение $Bi(n,\ p)$ (через него будем решать)
Опять задача на проверку гипотез, значит их надо сформулировать.

$$H_0 : p = p_0 = 0.25, \text{ против } H_1 : p > 0.25$$

H_0 — гипотеза о том, что студент угадывал ответы, то есть принятие её даст положительный ответ на задачу (студент не знает предмет).
Возьмём такую статистику:

$$T(x) = \frac{N - np_0}{\sqrt{np_0(1 - p_0)}} = \frac{30 - 25}{\sqrt{25 \cdot 0.75}} \approx 1.155$$

При справедливости H_0 статистика является асимптотически нормальной, критическая область $(Z_{0.95},\ +\infty)$
$Z_{0.95} = 1.64$, значит $T(x)$ попала в доверительную область, то есть мы принимаем H_0

Ответ

Можно так считать.