Московский физико-технический институт

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

II ЗАДАНИЕ

Автор: Яфаров Руслан,
Б13-202

весна 2024

1. Байесовские оценки

1. $X_1 \sim Exp(\theta)$, θ имеет сопряженное априорное распределение $\Gamma(\alpha, \beta)$. Проверьте оценку $\theta^* = \frac{n+\beta}{\alpha+\sum_{i=1}^n X_i}$ на состоятельность.

Решение.

2. По выборке $X_1, \ldots, X_n$ из пуассоновского распределения с параметром θ, где $\theta \sim \Gamma(\alpha, \lambda)$, постройте наилучшую оценку в байесовском подходе с квадратичной функцией потерь.

Решение.

3. Пусть $X_1, \ldots, X_n$ - выборка из нормального распредения с параметрами $(\theta, 1)$. Найдите байесовскую оценку параметра θ, если априорное распределение θ есть $Bin(1, p)$. Будет ли полученная оценка состоятельной оценкой параметра θ?

Решение.

4. Пусть $X_1, \ldots, X_n$ - выборка из равномерного распределения на отрезке $[0, \theta]$. Найдите байесовскую оценку параметра θ, если θ имеет априорное распределение а) равномерно на отрезке $[0, 1]$, б) с плотностью $q(t) = 1/t^2$ при $t \geq 1$. Проверьте полученные оценки на состоятельность.

Решение.

5. Пусть $X_1, \ldots, X_n$ - выборка из нормального распределения с параметрами $(\theta, 1)$. Найдите байесовскую оценку параметра θ, если априорное распределение θ есть $\mathcal{N}(b, \sigma^2)$.

Решение.

2. Проверка гипотез и доверительное оценивание

6. $X_1, \ldots, X_n$ - выборка из распределения с плотностью

$$p_\theta(x) = \frac{3x^2}{8\theta^3} I_{[0,2\theta]}(x)$$

С помощью статистики $X_{(1)}$ постройте точный доверительный интервал уровня доверия γ для параметра θ.

Решение.

7. $X_1, \ldots, X_n$ - выборка, $X_1 = \xi + \eta$, где ξ, η - независимые случайные величины, $\xi \sim R[0, \theta], \eta \sim Bin(1, \theta)$. Постройте доверительный интервал для θ уровня доверия $1 - \alpha$ с помощью неравенства Чебышева.

Решение.

8. Пусть $X_1, \ldots, X_n$ - выборка из гамма-распределения с параметрами (θ, λ). Постройте асимптотический доверительный интервал для θ уровня доверия α, если а) λ известно, б) λ неизвестно.

Решение.

9. Имеется X_1 - выборка объема 1. Основная гипотеза H_0 состоит в том, что X_1 имеет равномерное распределение на отрезке $[0, 1]$, альтернатива - в том, что X_1 имеет показательное распределение с параметром 1. Постройте наиболее мощный критерий уровня значимости α для различения этих гипотез и вычислите его мощность.

Решение.

10. * Пусть $X_1, \ldots, X_n$ - выборка из равномерного распределения на отрезке $[0, \theta], \theta > 0$. Постройте р.н.м.к. уровня значимости α для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta \neq \theta_0$ в виде

$$S(X_1, \ldots, X_n) = \{X_{(n)} \leq c\theta_0\} \cup \{X_{(n)} > \theta_0\}$$

Решение.

11. Пусть $\{P_\theta, \theta \in \Theta\}$ - семейство с невозрастающим отношением правдоподобия по статистике $T(X)$, а $\alpha < 1$ - некоторое положительное число. Постройте р.н.м.к. уровня значимости α для проверки гипотезы H_0 против альтернативы H_1 , где а) $H_0 : \theta \leq \theta_0$ (или $\theta = \theta_0$), $H_1 : \theta > \theta_0$; б) $H_0 : \theta \geq \theta_0$ (или $\theta = \theta_0$), $H_1 : \theta < \theta_0$.

Решение.

12. * Показать, что любой равномерно наиболее мощный несмещённый (т.е. $\inf_{\theta \in \Theta_0} \beta(\theta, S) \geq \sup_{\theta \in \Theta_1} \beta(\theta, S)$) критерий S является допустимым, т.е. не существует другого критерия R, который был бы не менее мощен, чем S, при всех альтернативах и более мощен хотя бы при одной из альтернатив.

Решение.

13. Докажите, что в предположении гипотезы $H_0 : F = F_0$ для любого $x \in R$ выполнено

$$F_n^*(x) \xrightarrow[n \to \infty]{\text{п. н.}} F_0(x)$$

Решение.

Теорема 1 (А. Колмогорова, 12.2 из С2). В предположении верности гипотезы $H_0 : F = F_0$ имеет место равенство

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq t) = K(t)$$

14. С помощью теоремы 1 докажите состоятельность критерия Колмогорова.

Решение.

15. Докажите, что при условии $0 \leq X_{(1)} \leq X_{(n)} \leq 1$ справедливо равенство

$$\int_0^1 (F_n^*(y) - y)^2 dy = \frac{1}{12n^2} + \frac{1}{n}\sum_{k=1}^{n}(X_{(k)} - (2k-1)/2n)^2$$

(с помощью этого представления часто вычисляется значение статистики ω^2 , которая используется в критерии Крамера–Мизеса–Смирнова).

Решение.

16. Цифры $0, 1, 2, \ldots, 9$ среди 800 первых десятичных знаков числа π появились $74, 92, 83, 79, 80, 73, 77, 75, 76, 91$ раз соответственно. С помощью хи-квадрат критерия проверьте гипотезу о согласии этих данных с законом равномерного распределения на множестве $\{0, 1, \ldots, 9\}$ на уровне значимости а) 0.05, б) 0.5, в) 0.8.

Решение.

17. Среди 5000 семей, имеющих трех детей, есть ровно 1010 семей с тремя мальчиками, 2200 семей с двумя мальчиками и одной девочкой, 950 семей с одним мальчиком и двумя девочками (во всех остальных семьях все дети - девочки). Можно ли с уровнем значимости $\alpha = 0.02$ считать, что количество мальчиков ξ в семье с тремя детьми имеет следующее распределение

$$P(\xi = 0) = \theta, P(\xi = 1) = \theta,$$

$$P(\xi = 2) = 2\theta, P(\xi = 3) = 1 - 4\theta,$$

где $\theta \in (0, 1/4)$?

Решение.

18. Пусть $X_1, \ldots, X_n$ - выборка из распределения: а)$Bern(\theta)$, б)$Pois(\theta)$, в)$\mathcal{N}(\theta, 1)$, г)$\exp(\theta)$. Ростроить доверительный интервал для параметра θ.

Решение.

19. Рассмотрим распределения Коши с плотностью $p_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}$. С помощью выборочной медианы построить доверительный интервал для θ^2.

Решение.

20. Пусть $X_1, \dots, X_n$ - выборка из экспоненциального распределения с параметром θ. Построить равномерно наиболее мощный критерий для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативы: а) $H_1 : \theta > \theta_0$; б) $H_1 : \theta < \theta_0$

Решение.

21. Пусть $X_1, \dots, X_n$ - выборка из $Bern(\theta)$. Проверить гипотезу $H_0 : \theta \le \theta_0$ против альтернативы $H_1 : \theta > \theta_0$

Решение.

3. Линейная регрессия. Проверка линейных гипотез.

22. Пусть

$$X_i = \beta_1 + i\beta_2 + \varepsilon_0 + \dots + \varepsilon_i,$$

$i = 0, 1, \dots, n$, где β_1, β_2 - неизвестные параметры, а $\varepsilon_0, \dots, \varepsilon_n$ - независимые, распределенные по закону $\mathcal{N}(0, \sigma^2)$ случайные величины. Сведите задачу к линейной модели и найдите оценки наименьших квадратов для β_1 и β_2 , а также несмещенную оценку для σ^2.

Решение. Заметим, что $X_{i+1} - X_i = \beta_2 + \varepsilon_i$ Запишем модель линейной регрессии:

$$\begin{pmatrix} X_0 \\ X_1 - X_0 \\ \vdots \\ X_n - X_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon \Rightarrow$$

$$Z^T Z = \begin{pmatrix} 1 & 0 \\ 0 & n \end{pmatrix} \Rightarrow (Z^T Z)^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{n} \end{pmatrix} \Rightarrow \hat{\beta} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{n} \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ X_1 - X_0 \\ \vdots \\ X_n - X_{n-1} \end{pmatrix} =$$

$$\begin{pmatrix} X_0 \\ \frac{X_n - X_0}{n} \end{pmatrix}, \hat{\sigma^2} = \tfrac{1}{n+1-2} \left\| \begin{pmatrix} X_0 \\ X_1 - X_0 \\ \vdots \\ X_n - X_{n-1} \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_0 \\ \frac{X_n - X_0}{n} \end{pmatrix} \right\|^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - X_{i-1} - \frac{X_n - X_0}{n} \right)^2 =$$

$$\frac{1}{n-1} \sum_{i=1}^{n} \left((X_i - X_{i-1})^2 - 2(X_i - X_{i-1}) \frac{X_n - X_0}{n} + \left(\frac{X_n - X_0}{n} \right)^2 \right) =$$

$$\frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i-X_{i-1})^2-\frac{(X_n-X_0)^2}{n}\right)$$

23. Пусть $X_1,\dots,X_n$- выборка из нормального распределения с параметрами (a,σ^2). Докажите, что статистики X и

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

независимы и вычислите распределение статистики nS^2.

Решение.

24. Пусть $X_1,\dots,X_n$ - выборка из $\mathcal{N}\left(a,\sigma^2\right)$ (оба параметра неизвестны). Постройте точные доверительные интервалы для каждого из параметров a,σ^2.

Решение.

25. Взвешивание трех грузов массами a и b производится следующим образом: n_1 раз взвешивается первый груз (все ошибки измерения имеют распределение $\mathcal{N}\left(0,\sigma^2\right)$), n_2 раза взвешивается второй груз на тех же самых весах, затем n_3 раза на других весах взвешиваются первый и второй груз вместе, все ошибки измерения на которых имеют распределение $\mathcal{N}\left(0,3\sigma^2\right)$. Сведите задачу к линейной модели и найдите оценки наименьших квадратов для a и b, а также оптимальную оценку для σ^2.

Решение.

$$\tilde{X}=\begin{pmatrix} X_1\\ \vdots\\ X_{n_1}\\ X_{n_1+1}\\ \vdots\\ X_{n_1+n_2}\\ \frac{X_{n_1+n_2+1}}{\sqrt{3}}\\ \vdots\\ \frac{X_{n_1+n_2+n_3}}{\sqrt{3}}\end{pmatrix}=\begin{pmatrix} 1 & 0\\ \vdots & \vdots\\ 1 & 0\\ 0 & 1\\ \vdots & \vdots\\ 0 & 1\\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\\ \vdots & \vdots\\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}}\end{pmatrix}\begin{pmatrix} a\\ b\end{pmatrix}+\varepsilon\Rightarrow Z^TZ=\begin{pmatrix} n_1+\frac{n_3}{3} & \frac{n_3}{3}\\ \frac{n_3}{3} & n_2+\frac{n_3}{3}\end{pmatrix}\Rightarrow$$

$$\det Z^TZ=n_1n_2+\frac{n_3}{3}(n_1+n_2)\Rightarrow (Z^TZ)^{-1}=\frac{1}{n_1n_2+\frac{n_3}{3}(n_1+n_2)}\begin{pmatrix} n_2+\frac{n_3}{3} & -\frac{n_3}{3}\\ -\frac{n_3}{3} & n_1+\frac{n_3}{3}\end{pmatrix},$$

$$Z^T\tilde{X}=\begin{pmatrix} \displaystyle\sum_{i=1}^{n_1}X_i+\frac{1}{3}\sum_{j=n_1+n_2+1}^{n_1+n_2+n_3}X_j\\ \displaystyle\sum_{k=n_1+1}^{n_1+n_2}X_k+\frac{1}{3}\sum_{j=n_1+n_2+1}^{n_1+n_2+n_3}X_j\end{pmatrix},\ \text{Пусть } S_1=\sum_{i=1}^{n_1}X_i, S_2=\sum_{k=n_1+1}^{n_1+n_2}X_k, S_3=\sum_{j=n_1+n_2+1}^{n_1+n_2+n_3}X_j\Rightarrow$$

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \frac{1}{n_1 n_2 + \frac{n_3}{3}(n_1+n_2)} \begin{pmatrix} n_2(S_1 + \frac{S_3}{3}) + \frac{n_3}{3}(S_1 - S_2) \\ n_1(S_2 + \frac{S_3}{3}) + \frac{n_3}{3}(S_2 - S_1) \end{pmatrix}$$

26. Пусть $X_i, i \in \{1, \ldots, n\}$ - независимые случайные величины, распределенные по нормальному закону с параметрами $(a + bi, \sigma^2)$. Постройте точные доверительные интервалы для параметров a, b, σ^2.

Решение.

27. $X_1, \ldots, X_n$ - выборка из распределения $\mathcal{N}(a_1, \sigma^2), Y_1, \ldots, Y_m$ - выборка из распределения $\mathcal{N}(a_2, \sigma^2), Z_1, \ldots, Z_k$ - выборка из распределения $\mathcal{N}(a_3, \sigma^2)$. Постройте F -критерий размера α для проверки гипотезы а) $H_0 : a_1 = a_2$ и $a_1 + a_2 = a_3$, б) $H_0 : a_1 = 2a_2$ и $a_1 + 3a_2 = a_3$.

Решение.

28. Пусть $X_i \sim \mathcal{N}(a, i\sigma^2)$, $i = 1, \ldots, n$, $Y_j \sim \mathcal{N}(jb, \sigma^2), j == 1, \ldots, m$, - независимые случайные величины, где a, b, σ^2 - неизвестные параметры. Сведите задачу к линейной модели и постройте F -критерий размера α для проверки гипотезы $H_0 : a + b = 1$.

Решение.

$$\begin{pmatrix} \frac{X_1}{\sqrt{1}} \\ \vdots \\ \frac{X_n}{\sqrt{n}} \\ \frac{Y_1}{\sqrt{1}} \\ \vdots \\ \frac{Y_m}{\sqrt{m}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1}} & 0 \\ \vdots & \vdots \\ \frac{1}{\sqrt{n}} & 0 \\ 0 & \frac{1}{\sqrt{1}} \\ \vdots & \vdots \\ 0 & \frac{1}{\sqrt{m}} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \varepsilon$$

29. Используя метод линейной регрессии, постройте приближение функции $f(x)$ многочленом третьей степени по следующим данным:

$f(x_i)$	3.9	5.0	5.7	6.5	7.1	7.6	7.8	8.1	8.4
x_i	4.0	5.2	6.1	7.0	7.9	8.6	8.9	9.5	9.9

Решение. Пусть искомый многочлен равен $ax^3 + bx^2 + cx + d$. Тогда модель линейной регрессии выглядит следующим образом:

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_9) \end{pmatrix} = \begin{pmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_9^3 & x_9^2 & x_9 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} + \varepsilon$$

Если считать это руками, то уйдет миллион лет и, кстати, если вычилять точно в виде дробей, оценивая $a/1000, b/100, c/10, d$, то получится такая матрица $(Z^T Z)^{-1}$

Input

$$\begin{pmatrix} 3014244005669 & 33620996231 & 381630821 & 4436327 \\ 33620996231 & 381630821 & 4436327 & 53309 \\ 381630821 & 4436327 & 53309 & 671 \\ 4436327 & 53309 & 671 & 9 \end{pmatrix}^{-1}$$

(matrix inverse)

Enlarge

Result

Enlarge | Data | Customize | Plain Text

$$\frac{1}{63889272833421192} \begin{pmatrix} 271735636 & -56787546343 & 3778438005837 & -79283636974068 \\ -56787546343 & 11939857168089 & -799801239296084 & 16899212322135336 \\ 3778438005837 & -799801239296084 & 53990447527625285 & -1150374729835829442 \\ -79283636974068 & 16899212322135336 & -1150374729835829442 & 24757040297579835354 \end{pmatrix}$$

Поэтому приведем программу на python:

```python
import numpy as np

X = [4.0, 5.2, 6.1, 7.0, 7.9, 8.6, 8.9, 9.5, 9.9]
F = [3.9, 5.0, 5.7, 6.5, 7.1, 7.6, 7.8, 8.1, 8.4]
Z = list(map(lambda x: [x**3, x**2, x, 1], X))
Z_np = np.array(Z)
F_np = np.array(F)
theta = np.linalg.inv(Z_np.T @ Z_np) @ Z_np.T @ F_np
a, b, c, d = theta
print(a, b, c, d)
```

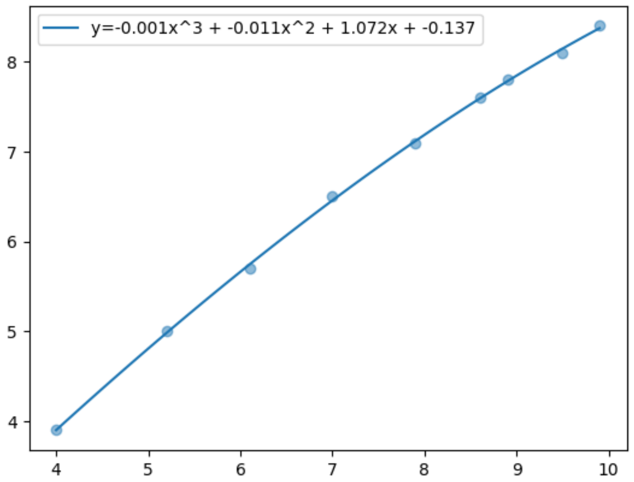
и получим

$$(a, b, c, d) = (-0.00100145912636040, -0.0114849959500949, 1.07154905548792, -0.137289441132785)$$

Вроде неплохо:

```python
import matplotlib.pyplot as plt

plt.scatter(X, F, alpha=0.5)
xseq = np.linspace(min(X), max(X), num=100)
plt.plot(xseq, a * xseq**3 + b * xseq**2 + c * xseq + d,
         label=f'y={round(a, 3)}x^3 + {round(b, 3)}x^2 + {round(c, 3)}x +
                                                {round(d, 3)}')
plt.legend()
plt.show()
```



30. Убедиться в том, что наиболее мощный критерий для различения двух простых гипотез о симметричном относитлтеьно нуля распределении набдюдаемой случайной величины ξ H_0 $\mathcal{L}(\xi) = R[-a, a]$ и H_1 $\mathcal{L}(\xi) = \mathcal{N}(0, \sigma^2)$ (a и σ известны) имеет для больших выборок следующую асимптотическую форму

$$\mathfrak{X}^*_{1,\mathfrak{a}} = \left\{ \sum_{i=1}^{n} X_i^2 \le \frac{n}{3} a^2 + \zeta_a \frac{2a^2}{3} \sqrt{\frac{n}{5}} \right\}, \Phi(\zeta_a) = a$$

Указание. Воспользоваться центральной предельной теоремой при отыскании распределения тестовой статистики.

Решение.

31. В последовательности независимых испытаний с двумя исходами вероятсноть “успеха” равна p. Построить критей проверки гипотезы H_0 $p = 0$ против альтернативы H_1 $p = 0.01$ и определить наименьший объем выборки, при котором вероятности ошибок 1-го и 2-го родов не превышают 0.01.

Решение.

32. Имеется 2 гирьки с весами θ_1 и θ_2. На одних и тех же весах сначала взвесили первую гирьку, затем вторую, а потом обе сразу. Найти оценку наименьших квадратов для θ_1 и θ_2 и несмещенную оценку дисперсии ошибки измерений. Проверьте гипотезы: а) $H_0 : \theta_1 = \theta_2$; б) $H_0 : 2\theta_1 = 3\theta_2$.

Решение.