Домашнее задание № 3

Тема: Доверительные интервалы. Общая теория построения статистических тестов.

Крайний срок сдачи: 15 ноября 2020 г. (до конца дня).

1

T1 (i) (1.5 балла) Пусть даны i.i.d. случайные величины $X_1, ...X_n$ с непрерывной функцией распределения $F(x)$ (возможно, зависящей от параметров). Докажите, что

$$-\sum_{i=1}^{n} \log\bigl(1 - F(X_i)\bigr) \sim \Gamma(n, 1).$$

Пользуясь этим результатом, постройте точный доверительный интервал методом центральной функции для параметра λ из распределения Вейбулла с функцией распределения

$$F(x) = 1 - e^{-(x/\lambda)^\tau}, \qquad x > 0.$$

Параметр τ предполагается известным.

(ii) (1.5 балла) Для случая $\tau = 1$ (экспоненциальное распределение) постройте асимтотические доверительные интервалы для параметра λ, используя асимптотическую нормальность статистик

$$\text{(iiia)}\ \ \frac{1}{n}\sum_{i=1}^{n} X_i; \qquad \text{(iiib)}\ \ \sqrt{\frac{1}{2n}\sum_{i=1}^{n} X_i^2}.$$

Выясните, какой из этих двух асимптотических интервалов имеет меньшую длину при больши́х n (для ответа на этот вопрос можно использовать компьютер).

(iii) (1 балл) Для этого же интервала постройте непараметрическую оценку параметра λ, используя зависимость медианы распределения от параметра λ.

N1 (1 балл) Промоделируйте выборку длины $n = 1000$ из экспоненциального распределения с фиксированным параметром λ. Расширяя подвыборку от 50 до 1000 наблюдений, вычислите длины доверительных интервалов, построенных в п. (iii) и (iv) (при этом зафиксируйте $\alpha = 0.05$) и постройте графики, выражающие зависимость длины интервала от размера выборки (отдельно для каждого метода). Отобразите зависимость длины интервала от α (для этого зафиксируйте параметр $n = 1000$).

2

T2 (1 балл) Проводится опрос общественного мнения с целью понять различия в предпочтениях между жителями городской и сельской местности. Задавался единственный вопрос - "Поддерживаете ли Вы действия правительства" , возможные ответы - да или нет. Было решено опросить равное количество респондентов - n человек в городе и n человек в сельской местности. Нужно проводить опрос до тех пор, пока разность между вероятностями поддержки в городе и силе не будет оценена с точностью 0.05. Решение принимается на основе асимптотического доверительного интервала с уровнем доверия 0.95. Определите, какое минимальное количество респондентов $2n$ нужно опросить.

T3 (1 балл) Количество респондентов для опроса из [T2] вычисляется исходя тестирования гипотезы "уровни поддержки в городе и селе совпадают" против альтернативы "уровень поддержки в городе превышает уровень поддержки в сельской местности на 3%". По-прежнему в городе и селе опрашивается равное количество респондентов. Исходя из проведённых ранее опросов предполагается, что уровень поддержки в городе примерно равен 14%, а в сельской местности - 9%.

Определите, какое минимальное количество респондентов нужно опросить, чтобы допустимая вероятность ошибки первого составила 0.05, а допустимая вероятность ошибки второго рода - 0.04.

Указание. Условие "Исходя из проведённых ранее опросов предполагается, что уровень поддержки в городе примерно равен 14%, а в сельской местности - 9%" следует использовать для оценивания асимптотической дисперсии (метод подстановки параметра).

3

T4 (1.5 балла) Дана выборка длины n из распределения с функцией плотности

$$p(x,\theta) = \theta(1-x)^{\theta-1}, \qquad x \in [0,1]$$

с неизвестным параметром $\theta > 0$. Опишите равномерно наиболее мощный тест для проверки гипотезы $\theta = \theta_0$ против альтернативы $\theta > \theta_0$. Докажите, что мощность этого теста равна

$$W(\theta) = F_{(n,\theta_0)}\Big(\frac{\theta}{\theta_0} \cdot q_{(n,\theta_0)}(\alpha)\Big),$$

где $F_{(n,\theta_0)}, q_{(n,\theta_0)}(\alpha)$ - это функция распределения и α-квантиль гамма-распределения, имеющего плотность

$$p(x) = \frac{\theta_0^n}{\Gamma(n)} x^{n-1} e^{-\theta_0 x}, \qquad x > 0.$$

T5 (1.5 балла) Наблюдаемые величины $X_1, ..., X_n$ имеют нормальное распределение с неизвестным средним μ и известной дисперсией σ^2. Для тестирования гипотезы $H_0 : \mu = \mu_0$ против альтернативы $H_1 : \mu = \mu_1 > \mu_0$ используются тесты вида $\{\bar{X} \geq c\}$, где $\bar{X} = (X_1 + .. + X_n)/n$. Докажите, что при выборе $c = (\mu_1 + \mu_0)/2$ получается тест, у которого сумма ошибок первого и второго рода минимальна среди всех возможных тестов такого вида.

4

Т6* (2 балла) Пусть задана выборка $x_1, .., x_n$ из экспоненциального семейства распределений с плотностью

$$p(x, \theta) = g(x)e^{x\theta - d(\theta)},$$

где $g : \mathbb{R} \to \mathbb{R}_+$ и $d : \mathbb{R} \to \mathbb{R}$. Для тестирования гипотезы $H_0 : \theta = \theta_0$ против альтернативы $\theta > \theta_0$ используется LR (likelihood ratio) тест вида

$$\Big\{\Lambda(\vec{x}) := \frac{\max_{\theta > \theta_0} L(\vec{x}, \theta)}{L(\vec{x}, \theta_0)} > c_\alpha\Big\}, \tag{1}$$

где $L(\vec{x}, \theta) = \prod_{i=1}^n p(x_i, \theta)$ - функция правдоподобия, и пороговый уровень c_α подбирается из условия, что уровень значимости теста равен $\alpha \in (0, 1)$.

(i) Докажите, что

$$\log(\Lambda(\vec{x})) = \begin{cases} nK(\hat{\theta}, \theta_0), & \text{если } \hat{\theta} > \theta_0, \\ 0, & \text{иначе,} \end{cases}$$

где $\hat{\theta}$ - оценка максимального правдоподобия (вычисленная по выборке $x_1, .., x_n$), а $K-$ расстояние Кульбака-Ляйблера, определяемое как

$$K(\hat{\theta}, \theta_0) = \int \log\Big(\frac{p(x, \hat{\theta})}{p(x, \theta_0)}\Big) p(x, \theta_0) dx.$$

(ii) Докажите, что LR - тест (1) в данной модели может быть записан в виде

$$\Big\{\hat{\theta} > \theta_0 + \tilde{c}_\alpha\Big\},$$

где $\tilde{c}_\alpha > 0$.