

Теория вероятностей и статистика, МИРЭК, 2022-2023

Дедлайн: домашнее задание отправляется в **pdf** формате на почту семинариста. В копию письма необходимо поставить ассистента группы.

Почты, на которые следует отправлять домашние задания, в зависимости от вашего семинариста:

1. Погорелова Полина Вячеславовна – tvis.we.2021@gmail.com
2. Потанин Богдан Станиславович – tvismirec@gmail.com
3. Слаболицкий Илья Сергеевич – tvis.fweia.hse@gmail.com

Домашнее задание должно быть отправлено на указанные почты в **pdf** формате до конца дня **12.02.2023** включительно (по московскому времени). Тема письма должна иметь следующий формат: “МИРЭК Фамилия Имя Группа Номер ДЗ”, например, “МИРЭК Потанин Богдан 200 ДЗ 3”.

Оформление: первый лист задания должен быть титульным и содержать лишь информацию об имени и фамилии, а также о номере группы студента и сдаваемого домашнего задания. Если pdf файл содержит фотографии, то они должны быть разборчивыми и повернуты правильной стороной.

Санкции: домашние задания, не удовлетворяющие требованиям к оформлению, выполненные не самостоятельно или сданные позже срока получают 0 баллов.

Проверка: при оценивании каждого задания проверяется не ответ, а весь ход решения, который должен быть описан подробно и формально, с использованием надлежащих определений, обозначений, теорем и т.д.

Самостоятельность: задания выполняются самостоятельно. С целью проверки самостоятельности выполнения домашнего задания студент может быть вызван на устное собеседование, по результатам которого оценка может быть либо сохранена, либо обнулена.

Домашнее задание №3

Задание №1. Снеговики в холодильниках. (60 баллов)

До наступления очередной зимы снеговики ложатся спать в специальные холодильники. Температура в каждом холодильнике является случайной величиной и не зависит от температур в других холодильниках. Имеется выборка X_1, \dots, X_n из замеров температур в холодильниках. При этом известно, что:

$$f_{X_1}(t) = \begin{cases} \ln(\theta)\theta^t, & \text{если } t < 0 \\ 0, & \text{в противном случае} \end{cases}, \text{ где } \theta > 1.$$

По результатам замеров температур в 100 холодильниках оказалось, что их суммарная (не средняя) температура равна -62.1335 . Помогите снеговикам разобраться в распределении температур в холодильниках.

1. Оцените параметр θ при помощи метода моментов (используйте любой начальный момент). Посчитайте реализацию найденной вами оценки. (5 баллов)

Примечания: Для взятия интеграла можете воспользоваться программными средствами, например, wolframalpha. Реализацию оценки округлите до целой части.

2. Оцените параметр θ при помощи метода максимального правдоподобия (в данном пункте проверять соблюдение условий второго порядка не обязательно¹). (5 баллов)

3. Убедитесь, что вы действительно нашли оценку метода максимального правдоподобия, проверив соблюдение условий второго порядка. (5 баллов)

Подсказка: $\frac{d\left(\frac{1}{t \ln(t)}\right)}{dt} = -\frac{\frac{1}{\ln(t)} + \left(\frac{1}{\ln(t)}\right)^2}{t^2}$.

4. Найдите информацию Фишера о параметре θ , содержащуюся во всей выборке. (5 баллов)

5. Найдите асимптотическую дисперсию ММП оценки. (5 баллов)

6. Найдите оценку асимптотической дисперсии ММП оценки. Посчитайте реализацию этой оценки. (5 баллов)

7. Посчитайте, приблизительно, вероятность, с которой ММП оценка отклонится от истинного значения параметра не более, чем на 1.2. (10 баллов)

8. При помощи дельта метода найдите асимптотическую дисперсию ММП оценки первого начального момента X_1 . (5 баллов)

9. Оцените полученную в предыдущем пункте асимптотическую дисперсию. Посчитайте реализацию этой оценки. (5 баллов)

¹Но если в задаче или пункте задачи непосредственно не сказано, что проверять условия второго порядка не обязательно, то их следует проверить. В противном случае решение не будет засчитано. В данном случае проверка условий второго порядка достаточно затруднительна, вследствие чего вынесена в отдельный пункт в качестве задачи повышенной сложности.

10. Посчитайте, приблизительно, вероятность, с которой ММП оценка первого начального момента отклонится от истинного значения параметра не более, чем на 0.1. **(10 баллов)**

Решение:

1. Попробуем воспользоваться первым начальным моментом:

$$E(X_1) = \int_{-\infty}^0 t * \ln(\theta) \theta^t dt = -\frac{1}{\ln(\theta)}$$

Решим соответствующее равенство для $E(X_1)$:

$$E(X_1) = -\frac{1}{\ln(\theta)} \implies \ln(\theta) = -\frac{1}{E(X_1)} \implies \theta = e^{-\frac{1}{E(X_1)}}$$

Подставляя вместо $E(X_1)$ выборочное среднее \bar{X}_n получаем оценку параметра:

$$\hat{\theta}_n^{MM} = e^{-\frac{1}{\bar{X}_n}}$$

Наконец, вычислим реализацию данной оценки:

$$\hat{\theta}_n^{MM}(x) = e^{-\frac{1}{\bar{x}_n}} = e^{\frac{1}{(-62.1335)/100}} \approx 5$$

2. Запишем функцию правдоподобия:

$$L(\theta; x) = \prod_{i=1}^n \ln(\theta) \theta^{x_i}$$

Найдем логарифм функции правдоподобия:

$$\ln L(\theta; x) = n \ln(\ln(\theta)) + \ln(\theta) \sum_{i=1}^n x_i$$

Для того, чтобы максимизировать логарифм функции правдоподобия по θ , сперва, рассмотрим условия первого порядка:

$$\frac{d \ln L(\theta; x)}{d\theta} = \frac{n}{\theta \ln(\theta)} + \frac{\sum_{i=1}^n x_i}{\theta} = 0$$

Решая соответствующее равенство получаем точку θ^* , подозреваемую на максимум:

$$\theta^* = e^{\left(\frac{-n}{\sum_{i=1}^n x_i} \right)} = e^{-\frac{1}{\bar{x}_n}}$$

3. Убедимся, что мы нашли максимум, показав, что в соответствующей точке вторая производная отрицательна:

$$\begin{aligned}\frac{d^2 \ln L(\theta; x)}{d\theta^2} \Big|_{\theta=\theta^*} &= -n \frac{\frac{1}{\ln(\theta^*)} + \left(\frac{1}{\ln(\theta^*)}\right)^2}{(\theta^*)^2} - \frac{\sum_{i=1}^n x_i}{(\theta^*)^2} = \\ &= -n \frac{-\bar{x}_n + (-\bar{x}_n)^2}{(\theta^*)^2} - \frac{\sum_{i=1}^n x_i}{(\theta^*)^2} = \frac{\sum_{i=1}^n x_i - n(\bar{x}_n)^2}{(\theta^*)^2} - \frac{\sum_{i=1}^n x_i}{(\theta^*)^2} = -n \left(\frac{\bar{x}_n}{\theta^*}\right)^2 < 0\end{aligned}$$

Поскольку найденная ранее точка действительно является точкой максимума, то ММП оценка будет иметь вид:

$$\hat{\theta}_n = e^{-\frac{1}{\bar{x}_n}}$$

4. Найдем информацию Фишера:

$$\begin{aligned}I(\theta) &= -E \left(\frac{d^2 \ln L(\theta; X)}{d\theta^2} \right) = n \frac{\frac{1}{\ln(\theta)} + \frac{1}{\ln(\theta)^2}}{\theta^2} + \frac{\sum_{i=1}^n E(X_i)}{\theta^2} = \\ &= \frac{n(\ln(\theta) + 1)}{(\theta \ln(\theta))^2} - \frac{n}{\theta^2 \ln(\theta)} = \frac{n}{(\theta \ln(\theta))^2}\end{aligned}$$

5. Вычислим асимптотическую дисперсию:

$$\text{As.Var}(\hat{\theta}) = \frac{1}{I(\theta)} = \frac{(\theta \ln(\theta))^2}{n}$$

6. Оценим найденную ранее асимптотическую дисперсию:

$$\widehat{\text{As.Var}}(\hat{\theta}) = \frac{(\hat{\theta} \ln(\hat{\theta}))^2}{n}$$

Посчитаем реализацию соответствующей оценки:

$$\widehat{\text{As.Var}}(\hat{\theta})(x) = \frac{(5 \ln(5))^2}{100} \approx 0.648$$

7. В силу асимптотической нормальности ММП оценок предположим, что:

$$(\hat{\theta} - \theta) \sim \mathcal{N}(0, 0.648)$$

Отсюда получаем приближение искомой вероятности:

$$\begin{aligned}P(|\hat{\theta} - \theta| \leq 1.2) &\approx P(-1.2 \leq \hat{\theta} - \theta \leq 1.2) = \\ &= \Phi\left(\frac{1.2 - 0}{\sqrt{0.648}}\right) - \Phi\left(\frac{-1.2 - 0}{\sqrt{0.648}}\right) = 2\Phi\left(\frac{1.2 - 0}{\sqrt{0.648}}\right) - 1 \approx 0.864\end{aligned}$$

8. Найдем производную первого начального момента по параметру:

$$\frac{dE(X_1)}{d\theta} = \frac{d\frac{-1}{\ln(\theta)}}{d\theta} = \frac{1}{\theta(\ln(\theta))^2}$$

В итоге получаем асимптотическую дисперсию:

$$\text{As.Var}(\hat{E}(X_1)) = \text{As.Var}\left(\frac{-1}{\ln(\hat{\theta})}\right) = \left(\frac{1}{\theta(\ln(\theta))^2}\right)^2 \frac{n}{(\theta \ln(\theta))^2}$$

9. Для получения оценки этой асимптотической дисперсии достаточно вместо θ подставить $\hat{\theta}$:

$$\widehat{\text{As.Var}}(\hat{E}(X_1)) = \widehat{\text{As.Var}}\left(\frac{-1}{\ln(\hat{\theta})}\right) = \left(\frac{1}{\hat{\theta}(\ln(\hat{\theta}))^2}\right)^2 \frac{n}{(\hat{\theta} \ln(\hat{\theta}))^2}$$

Посчитаем реализацию найденной оценки:

$$\widehat{\text{As.Var}}(\hat{\theta})(x) = \left(\frac{1}{5(\ln(5))^2}\right)^2 \frac{100}{(5 \ln(5))^2} \approx 0.0092$$

10. В силу асимптотической нормальности ММП оценок предположим, что:

$$(\hat{E}(X_1) - E(X_1)) \sim \mathcal{N}(0, 0.0092)$$

Отсюда получаем приближение искомой вероятности:

$$\begin{aligned} P(|\hat{E}(X_1) - E(X_1)| \leq 0.1) &\approx P(-0.1 \leq \hat{E}(X_1) - E(X_1) \leq 0.1) = \\ &= \Phi\left(\frac{0.1 - 0}{\sqrt{0.0092}}\right) - \Phi\left(\frac{-0.1 - 0}{\sqrt{0.0092}}\right) = 2\Phi\left(\frac{0.1 - 0}{\sqrt{0.0092}}\right) - 1 \approx 0.703 \end{aligned}$$

Задание №2. Эксперимент Лаврентия. (40 баллов)

В компьютерной игре Лаврентий сражается с противником, который с равной вероятностью совершает прыжок вперед или назад. При этом длина прыжка является равномерно распределенной случайной величиной от 0 до θ , где $\theta > 0$. Для того, чтобы всегда держаться от противника на безопасном расстоянии, но при этом не отходить слишком далеко, Лаврентий хочет оценить максимальную длину прыжка противника θ . Для этого он собрал выборку X_1, \dots, X_n из равномерного распределения $U(-\theta, \theta)$, наблюдения которой соответствуют расстоянию между исходным положением противника и тем, в котором он оказался после прыжка (положительное расстояние соответствует прыжку вперед, а отрицательное – прыжку назад). Лаврентий хочет использовать следующую оценку параметра θ :

$$\hat{\theta}_n = 0.5(\max(X_1, \dots, X_n) - \min(X_1, \dots, X_n)) = 0.5(X_{(n)} - X_{(1)})$$

1. Найдите функцию распределения и функцию плотности экстремальных статистик $X_{(1)}$ и $X_{(n)}$. (5 баллов)
2. Проверьте, является ли оценка $\hat{\theta}_n$ несмещенной. (5 баллов)
3. Проверьте, является ли последовательность оценок $\hat{\theta}_n$ состоятельной. (5 баллов)
4. Оцените параметр θ при помощи метода максимального правдоподобия. Обозначьте полученную оценку как $\hat{\theta}_n^*$ (10 баллов)
5. Определите, какая из оценок $\hat{\theta}_n$ и $\hat{\theta}_n^*$ является более эффективной. (10 баллов)
6. Преобразуйте оценку $\hat{\theta}_n^*$ таким образом, чтобы она стала несмещенной и обозначьте полученную оценку как $\hat{\theta}_n^a$. Определите, при каких объемах выборки n оценка $\hat{\theta}_n^a$ окажется более эффективной, чем $\hat{\theta}_n^*$. (5 баллов)

Решение:

1. Сперва найдем распределение экстремальных статистик $X_{(1)}$ и $X_{(n)}$ на носителе $t \in [-\theta, \theta]$:

$$F_{X_{(n)}}(t) = \left(\frac{t + \theta}{2\theta}\right)^n$$
$$F_{X_{(1)}}(t) = 1 - \left(1 - \frac{t + \theta}{2\theta}\right)^n = 1 - \left(\frac{\theta - t}{2\theta}\right)^n$$

Дифференцируя полученные функции распределения находим функции плотности:

$$f_{X_{(n)}}(t) = \begin{cases} \frac{n(t+\theta)^{n-1}}{(2\theta)^n}, & \text{если } t \in [-\theta, \theta] \\ 0, & \text{в противном случае} \end{cases}$$
$$f_{X_{(1)}}(t) = \begin{cases} \frac{n(\theta-t)^{n-1}}{(2\theta)^n}, & \text{если } t \in [-\theta, \theta] \\ 0, & \text{в противном случае} \end{cases}$$

2. Найдем математические ожидания экстремальных статистик:

$$E(X_{(n)}) = \int_{-\theta}^{\theta} t \frac{n(t+\theta)^{n-1}}{(2\theta)^n} dt = \frac{n-1}{n+1}\theta$$

$$E(X_{(1)}) = \int_{-\theta}^{\theta} t \frac{n(\theta-t)^{n-1}}{(2\theta)^n} dt = -\frac{n-1}{n+1}\theta$$

Из полученного результата следует, что оценка является смещенной, поскольку:

$$E(\hat{\theta}_n) = 0.5(E(X_{(n)}) - E(X_{(1)})) = 0.5 \left(\frac{n-1}{n+1}\theta - \left(-\frac{n-1}{n+1}\theta \right) \right) = \frac{n-1}{n+1}\theta \neq \theta$$

3. Оценка является асимптотически несмещенной, поскольку:

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \lim_{n \rightarrow \infty} \frac{n-1}{n+1}\theta = \theta$$

Для проверки второго условия состоятельности сперва найдем дисперсии:

$$E(X_{(n)}^2) = \int_{-\theta}^{\theta} t^2 \frac{n(t+\theta)^{n-1}}{(2\theta)^n} dt = \frac{n^2 - n + 2}{n^2 + 3n + 2}\theta^2$$

$$E(X_{(1)}^2) = \int_{-\theta}^{\theta} t^2 \frac{n(\theta-t)^{n-1}}{(2\theta)^n} dt = \frac{n^2 - n + 2}{n^2 + 3n + 2}\theta^2$$

$$Var(X_{(n)}) = Var(X_{(1)}) = \frac{n^2 - n + 2}{n^2 + 3n + 2}\theta^2 - \left(\frac{n-1}{n+1}\theta \right)^2 = \frac{4n}{(n+1)^2(n+2)}\theta^2$$

Используя полученные выражения убеждаемся в состоятельности последовательности оценок:

$$\begin{aligned} \lim_{n \rightarrow \infty} Var(\hat{\theta}_n) &= \lim_{n \rightarrow \infty} 0.25Var(X_{(n)}) + 0.25Var(X_{(1)}) - 0.5Cov(X_{(1)}, X_{(n)}) \leq \\ &\leq \lim_{n \rightarrow \infty} 0.25Var(X_{(n)}) + 0.25Var(X_{(1)}) + 0.5\sqrt{Var(X_{(1)})Var(X_{(n)})} = \\ &= \lim_{n \rightarrow \infty} \frac{4n}{(n+1)^2(n+2)}\theta^2 = 0 \end{aligned}$$

Примечание: в данном и последующем пункте в случае введения предположения $Cov(X_{(1)}, X_{(n)}) = 0$ балл не снижается. Однако, в случае грамотного учета верхней границы дисперсии разницы с учетом наличия ковариации выставятся дополнительные 5 баллов.

4. Запишем функцию правдоподобия:

$$L(\theta; x) = \begin{cases} \frac{1}{(2\theta)^n}, & \text{если } \max(|x_1|, \dots, |x_n|) \leq \theta \\ 0, & \text{в противном случае} \end{cases}$$

Поскольку правдоподобие либо обнуляется, либо убывает по θ при наблюдении $\max(|x_1|, \dots, |x_n|) \leq \theta$, то:

$$\hat{\theta}_n^* = \max(|X_1|, \dots, |X_n|)$$

5. Обратим внимание, что:

$$Var(\hat{\theta}_n) = 0.25Var(X_{(n)}) + 0.25Var(X_{(1)}) - 0.5\rho\sqrt{Var(X_{(1)})Var(X_{(n)})}$$

Где $\rho = Corr(X_{(1)}, X_{(n)}) \in (-1, 1)$. Наименьшим значение $MSE(\hat{\theta}_n)$ будет, очевидно, при $\rho = 1$ (что достигается лишь при $n = 1$), поскольку в таком случае дисперсия оценки обнуляется. В результате получаем:

$$\begin{aligned} MSE(\hat{\theta}_n) &= Var(\hat{\theta}_n) + \left(E(\hat{\theta}_n)^2 - \theta\right) \geq \\ &\geq 0 + \left(\frac{n-1}{n+1}\theta - \theta\right)^2 = \frac{4}{(n+1)^2}\theta^2 \end{aligned}$$

Для того, чтобы посчитать среднеквадратическое отклонение второй (ММП) оценки, обратим внимание, что $|X_1| \sim U(0, \theta)$, откуда:

$$\begin{aligned} F_{\hat{\theta}_n^*}(t) &= \left(\frac{t}{\theta}\right)^n, \text{ при } t \in [-\theta, \theta] \\ f_{\hat{\theta}_n^*}(t) &= \frac{nt^{n-1}}{\theta^n}, \text{ при } t \in [-\theta, \theta] \\ E(\hat{\theta}_n^*) &= \int_0^\theta t \frac{nt^{n-1}}{\theta^n} dt = \frac{n}{n+1}\theta \\ E((\hat{\theta}_n^*)^2) &= \int_0^\theta t^2 \frac{nt^{n-1}}{\theta^n} dt = \frac{n}{n+2}\theta^2 \\ Var(\hat{\theta}_n^*) &= \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 \\ MSE(\hat{\theta}_n^*) &= \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{n}{n+1}\theta - \theta\right)^2 = \frac{2}{n^2+3n+2}\theta^2 \end{aligned}$$

Рассмотрим разницу среднеквадратических отклонений:

$$MSE(\hat{\theta}_n^*) - MSE(\hat{\theta}_n) \leq \frac{2}{n^2+3n+2}\theta^2 - \frac{4}{(n+1)^2}\theta^2 = -\frac{2(n+3)}{(n+1)^2(n+2)} < 0$$

Поскольку $MSE(\hat{\theta}_n^*) < MSE(\hat{\theta}_n)$, то оценка $\hat{\theta}_n^*$ более эффективна.

Примечание: при допущении $\rho = 0$, за которое балл не снижается, должен был получиться тот же вывод о соотношении эффективностей оценок, а также:

$$\begin{aligned} MSE(\hat{\theta}_n) &= Var(\hat{\theta}_n) + \left(E(\hat{\theta}_n)^2 - \theta\right) = \\ &= \frac{4n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{n-1}{n+1}\theta - \theta\right)^2 = \frac{8}{n^2+3n+2}\theta^2 \end{aligned}$$

6. Очевидно, что у оценки $\hat{\theta}^*$ можно убрать смещение за счет умножения на $\frac{n+1}{n}$, поэтому введем новую оценку $\hat{\theta}^a = \frac{n+1}{n}\hat{\theta}^*$. В данном случае увеличение дисперсии будет компенсировано снижением смещения оценки, в результате среднеквадратическое отклонение примет вид:

$$\begin{aligned} MSE(\hat{\theta}_n^a) &= Var\left(\frac{n+1}{n}\hat{\theta}_n^*\right) + \left(E\left(\frac{n+1}{n}\hat{\theta}_n^*\right) - \theta\right)^2 = \\ &= \frac{(n+1)^2}{n^2} \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{n+1}{n} \frac{n}{n+1}\theta - \theta\right)^2 = \frac{1}{n^2 + 2n}\theta^2 \end{aligned}$$

дисперсия растет в $((n+1)/n)^2$ раз смещение падает до нуля

Убедимся, что среднеквадратическое отклонение полученной оценки меньше, чем у ММП при $n \geq 2$:

$$MSE(\hat{\theta}_n^a) - MSE(\hat{\theta}_n^*) = \frac{1}{n^2 + 2n}\theta^2 - \frac{2}{n^2 + 3n + 2}\theta^2 = -\frac{n-1}{n(n+1)(n+2)}\theta^2 > 0$$

Таким образом оценка $\hat{\theta}_n^a$ более эффективна чем $\hat{\theta}_n^*$ при $n \geq 2$. Если же $n = 1$, то обе оценки обладают равной эффективностью.

P.S. Нетрудно заметить, что при $n \rightarrow \infty$ разница в MSE стремится к нулю, то есть при больших объемах выборки разница в эффективности между двумя оценками окажется несущественной.