Теория вероятностей и статистика, МИРЭК, 2021-2022

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

Почты семинаристов, на которые следует отправлять домашние задания:

1. Погорелова Полина Вячеславовна – tvis.we.2021@gmail.com (группы 202 и 203)
2. Потанин Богдан Станиславович – studypotanin@gmail.com (группа 201)
3. Слаболицкий Илья Сергеевич – tvis.fweia.hse@gmail.com (группы 204, 205 и 206)

Почты ассистентов, на которые следует продублировать домашнее задание (поставить в копию при отправке):

1. Романова Дарья Юрьевна – dyuromanova_1@edu.hse.ru (группа 201)
2. Афонина Ангелина Геннадьевна – agafonina@edu.hse.ru (группа 202)
3. Макаров Антон Андреевич – aamakarov_5@edu.hse.ru (группа 203)
4. Атласов Александр Александрович – aaatlasov@edu.hse.ru (группа 204)
5. Костромина Алина Максимовна – amkostromina@edu.hse.ru (группа 205)
6. Краевский Артем Андреевич – aakraevskiy@edu.hse.ru (группа 206)

Домашнее задание должно быть отправлено на указанные почты в **pdf** формате до **15.01.2022, 8.00 (утра)** включительно (по московскому времени). Тема письма должна иметь следующий формат: "МИРЭК Фамилия Имя Группа Номер ДЗ", например, "МИРЭК Потанин Богдан 200 ДЗ 4".
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 №4

Новогодние статистические оценки

Задание №1. Предновогодняя суета. (20 баллов)

В преддверии Нового года Санта-Клаус поручил своим помощникам эльфам наполнить подарками мешок и погрузить его в сани. Вес каждого подарка — случайная величина X — измеряется в килограммах, не зависит от веса других подарков и имеет следующую плотность:

$$f_X(t) = \begin{cases} \frac{3\alpha^3}{t^4}, \text{ если } t \geq \alpha \\ 0, \text{ иначе} \end{cases} \quad \text{, где } \alpha > 1$$

Всего эльфам нужно погрузить n подарков. Санта-Клаус не знает статистику, поэтому безосновательно предположил следующие оценки для неизвестного параметра α: $\hat{\alpha}_1 = 2022\bar{X}_n$ и $\hat{\alpha}_2 = \frac{2}{3}\bar{X}_n$.

1. Определите, является ли оценка $\hat{\alpha}_1$ несмещенной оценкой параметра α. Если нет, то преобразуйте оценку $\hat{\alpha}_1$ так, чтобы она стала несмещенной оценкой параметра α. **(5 баллов)**

2. Проверьте, является ли оценка $\hat{\alpha}_2$ состоятельной оценкой параметра α. **(5 баллов)**

3. Найдите состоятельную оценку для вероятности того, что вес случайного подарка превысит 5 килограмм. **(10 баллов)**

Решение:

1. Оценка $\hat{\alpha}_1$ не является несмещенной, посокольку:

$$E(\hat{\alpha}_1) = 2022E(\overline{X}_n) = 2022E(X_1) = 2022\int\limits_{\alpha}^{\infty} t\frac{3\alpha^3}{t^4}dt = 2022 \times 1.5\alpha = 3033\alpha \neq \alpha$$

   В результате получаем, что несмещенной оценкой параметра α будет $\frac{\hat{\alpha}_1}{3033}$.

2. Обратим внимание, что оценка $\hat{\alpha}_2$ является несмещенной, а значит и асимптотически несмещенной, поскольку:

$$E(\hat{\alpha}_2) = \frac{2}{3}E(X_1) = \frac{2}{3} \times \frac{3}{2}\alpha = \alpha$$

   Оценка является состоятельной, поскольку также соблюдается условие сходимости дисперсии к нулю:

$$\lim_{i=1}^{n} Var(\hat{\alpha}_2) = \lim_{i=1}^{n} \frac{2^2}{3^2}\frac{Var(X_1)}{n} = 0$$

3. Рассчитаем соответствующую вероятность:

$$P(X_1 > 5) = \begin{cases} \int\limits_{5}^{\infty} \frac{3a^3}{t^4}dt = \left(\frac{a}{5}\right)^3, \text{ если } \alpha \leq 5 \\ 1, \text{ в противном случае} \end{cases}$$

Покажем, что следующая оценка будет состоятельной[1]:

$$\hat{P}(X_1 > 5) = \begin{cases} \left(\frac{\hat{a}_2}{5}\right)^3 = \frac{8\overline{X}_n^3}{3375}, \text{ если } \hat{\alpha}_2 \leq 5 \\ 1, \text{ если } \hat{\alpha}_2 > 5 \end{cases} = I(\hat{\alpha}_2 \leq 5)\left(\frac{\hat{a}_2}{5}\right)^3 + (1 - I(\hat{\alpha}_2 \leq 5))$$

Сперва обратим внимание, что $\left(\frac{a}{5}\right)^3$ является функцией, непрерывной по α. Причем $\hat{\alpha}_2$ является состоятельной оценкой α. Следовательно, по теореме Манна-Вальда:

$$\left(\frac{\hat{a}_2}{5}\right)^3 \xrightarrow{p} \left(\frac{a}{5}\right)^3$$

Далее рассмотрим два случая. Во-первых, при $\alpha \leq 5$ можно показать[2], что:

$$I(\hat{\alpha}_2 \leq 5) \xrightarrow{p} 1$$

Поэтому, дважды применяя теорему Слуцкого получаем, что при $1 < \alpha \leq 5$:

$$I(\hat{\alpha}_2 \leq 5)\left(\frac{\hat{a}_2}{5}\right)^3 + (1 - I(\hat{\alpha}_2 \leq 5)) \xrightarrow{p} 1 \times \left(\frac{a}{5}\right)^3 + (1-1) = \left(\frac{a}{5}\right)^3 = P(X_1 > 5)$$

Во-вторых, при $\alpha > 5$:

$$I(\hat{\alpha}_2 \leq 5) \xrightarrow{p} 0$$

По аналогии дважды применяя теорему Слуцкого имеем:

$$I(\hat{\alpha}_2 \leq 5)\left(\frac{\hat{a}_2}{5}\right)^3 + (1 - I(\hat{\alpha}_2 \leq 5)) \xrightarrow{p} 0 \times \left(\frac{a}{5}\right)^3 + (1-0) = 1 = P(X_1 > 5)$$

Проверка в R:

```
cdf <- function(t, a)
{
  if (t < a)
  {
    return(0)
  }
  return(1 - a ^ 3 / t ^ 3)
}
icdf <- function(p, a)
{
  return(a / (1 - p) ^ (1 / 3))
}
n <- 10000000
a <- 3
```

[1] Для получения полного балла за задание достаточно записать оценку, доказывать состоятельность не обязательно.

[2] Доказательство опускается и не требуется, поскольку предполагает применение дельта-метода. Сперва дельта-метод применяется для нахождения распределения $\left(\frac{\hat{a}_2}{5}\right)^3$. Затем, исходя из найденного асимптотического распределения находится асимптотическая вероятность $P(\hat{\alpha}_2 > 5)$, которая применяется для нахождения асимптотического распределения $I(\hat{\alpha}_2 > 5)$.

```
x <- icdf(runif(n), a)
# пункт 2
a.est <- (2 / 3) * mean(x)
# пункт 3
prob.est <- ifelse(a.est <= 5, yes = a.est ^ 3 / 125, no = 1)
cbind(est = prob.est, true = 1 - cdf(5, a))
```

Задание №2. Сказочное поручение. (10 баллов)

Каждый раз Санта-Клаусу удается пробраться в дом и оставить подарок под елкой незамеченным с вероятностью $p \in (0, 1)$. Соответствующая вероятность не зависит от того, был ли Санта-Клаус обнаружен в предыдущие разы. С целью оценки целесообразности увеличения расходов на скрытность Санта-Клаус поручил аналитическому отделу (с дедлайном в новогоднюю ночь) по выборке из 5-ти наблюдений найти оценку параметра p с дисперсией, не превышающей $\frac{p^2-p^3}{10}$. Несчастные эльфы из аналитического отдела потратили целый год, но так и не смогли выполнить поручение Санта-Клауса. Отчаявшись, в канун новогодней ночи они просят вас о помощи. Используя знания в области математической статистики помогите эльфам выполнить поручение Санта-Клауса или обосновать невозможность его реализации.

Решение:

Обратим внимание, что информация Фишера о параметре p составляет:

$$I_5(p) = \frac{5}{p(1-p)}$$

Обозначим через $\hat{p}_5$ оценку параметра p. полученную по выборке из 5-ти наблюдений. Покажем, что выполнить поручение Санта-Клауса невозможно, поскольку оно вступает в противоречие с неравенством Рао-Крамера:

$$Var(\hat{p}_5) = \frac{p^2-p^3}{10} < \frac{p^2(1-p)}{5} < \frac{p(1-p)}{5} = \frac{1}{I_5(p)}$$

Задание №3. Оценка для Снеговика. (40 баллов)

Температура в новогоднюю ночь в случайно выбранном городе является случайной величиной со следующей функцией плотности:

$$f_{X_1}(t) = \begin{cases} \frac{3t^2}{\theta^3}, \text{ при } t \in [-\theta, 0] \\ 0, \text{ в противном случае} \end{cases} \quad \text{, где } \theta > 0$$

Допустим, что температуры в городах независимы. Из соображений мониторинга ситуации с глобальным потеплением Снеговик Почтовик просит вас оценить параметр θ по выборке из новогодних температур в n городах.

1. Найдите оценку параметра θ при помощи метода моментов. **(5 баллов)**

2. Рассчитайте эффективность найденной вами в предыдущем пункте оценки. **(5 баллов)**

3. Найдите оценку параметра θ при помощи метода максимального правдоподобия. **(5 баллов)**

4. Рассчитайте информацию Фишера или обоснуйте, почему в данном случае она не определена. **(5 балла)**

5. Проверьте, является ли найденная вами оценка несмещенной. Если нет, то попытайтесь скорректировать ее таким образом, чтобы она стала несмещенной. **(5 баллов)**

6. Вычислите эффективность ММП оценки. **(10 баллов)**

7. Найдите оценку, которая будет более эффективна, чем ММП оценка. Покажите, что она действительно является более эффективной. **(5 балла)**

Рекомендация: подумайте, как бы решалась данная задача, если бы носитель наблюдений равнялся $[-\theta, \theta]$.

Решение:

1. Для нахождения оценки воспользуемся первым начальным моментом:

$$E(X_1) = \int_{-\theta}^{0} t \times \frac{3t^2}{\theta^3} dt = -0.75\theta \implies \hat{\theta}_n^{MM} = -\frac{4}{3}\overline{X_n}$$

2. Обратим внимание, что найденная оценка является несмещенной:

$$E(\hat{\theta}_n^{MM}) = E\left(-\frac{4}{3}\overline{X_n}\right) = -\frac{4}{3} \times \left(-\frac{3}{4}\theta\right) = \theta$$

Следовательно, эффективность рассматриваемой оценки совпадает с ее дисперсией:

$$E(X_1^2) = \int_{-\theta}^{0} t^2 \frac{3t^2}{\theta^3} dt = 0.6\theta^2$$

$$Var(\hat{\theta}_n^{MM}) = Var\left(\left(-\frac{4}{3}\right)^2 \overline{X_n}\right) = \frac{16}{9}\frac{Var(X_1^2)}{n} = \frac{16}{9}\frac{0.6\theta^2 - (-0.75\theta)^2}{n} = \frac{\theta^2}{15n}$$

3. Обратим внимание, что если $-\theta > \min(x_1, ..., x_n)$, то функция правдоподобия обращается в ноль. Следовательно, следует максимизировать функцию правдоподобия при ограничении $-\theta \le \min(x_1, ..., x_n)$. При данном ограничении функция правдоподобия принимает следующий вид:

$$L(\theta; x) = \prod_{i=1}^{n} \frac{3x_i^2}{\theta^3} = 3^n \theta^{-3n} \prod_{i=1}^{n} x_i^2$$

Поскольку правдоподобие строго убывает по θ, то в силу наложенного ограничения получаем точку максимума $\theta^* = -\min(x_1, ..., x_n)$. В результате ММП оценка принимает вид:

$$\hat{\theta}_n = -\min(X_1, ..., X_n) = \max(-X_1, ..., -X_n)$$

4. Поскольку в данном случае носитель распределения зависит от параметра, то информация Фишера не существует.

5. Найдем функцию распределения наблюдения при $t \in [-\theta, 0]$:

$$F_{X_1}(t) = \int\limits_{-\theta}^{t} \frac{3t^2}{\theta^3} dt = 1 + \frac{t^3}{\theta^3}$$

Далее, найдем функцию распределения $-X_1$ при $t \in [0, \theta]$:

$$F_{-X_1}(t) = P(-X_1 \leq t) = P(X_1 \geq -t) = 1 - F_{X_1}(-t) = -\frac{(-t)^3}{\theta^3} = \frac{t^3}{\theta^3}$$

В результате получаем выражение для функции распределения нашей оценки при $t \in [0, \theta]$:

$$F_{\hat{\theta}_n} = (F_{-X_1}(t))^n = \frac{t^{3n}}{\theta^{3n}}$$

Дифференцируя полученный результат получаем, что при $t \in [0, \theta]$ функция плотности примет вид:

$$f_{\hat{\theta}_n}(t) = \frac{t^{3n}}{\theta^{3n}} = \frac{3nt^{3n-1}}{\theta^{3n}}$$

Оценка является смещенной, поскольку:

$$E(\hat{\theta}_n) = \int\limits_{0}^{\theta} t \times \frac{3nt^{3n-1}}{\theta^{3n}} dt = \frac{3n}{3n+1}\theta$$

Для того, чтобы нивелировать смещение, введем новую, несмещенную оценку:

$$\hat{\theta}^* = \frac{3n+1}{3n}\hat{\theta}_n = \frac{3n+1}{3n}\max(-X_1, ..., -X_n)$$

6. Рассчитаем дисперсию ММП оценки:

$$E(\hat{\theta}_n^2) = \int\limits_{0}^{\theta} t^2 \times \frac{3nt^{3n-1}}{\theta^{3n}} dt = \frac{3n}{3n+2}\theta^2$$

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

В результате получаем эффективность:

$$MSE(\hat{\theta}_n) = \frac{3n}{(3n+1)^2(3n+2)}\theta^2 + \left(\frac{3n}{3n+1}\theta - \theta\right)^2 = \frac{2\theta^2}{9n^2+9n+2}$$

7. Рассчитаем эффективность найденной ранее несмещенной оценки:

$$MSE(\hat{\theta}_n^*) = Var\left(\hat{\theta}_n{}^*\right) = Var\left(\frac{3n+1}{3n}\hat{\theta}_n\right) =$$

$$= \left(\frac{3n+1}{3n}\right)^2 \frac{3n}{(3n+1)^2(3n+2)}\theta^2 = \frac{\theta^2}{9n^2+6n}$$

Данная оценка более эффективна, чем ММП оценка, поскольку ее среднеквадратическое отклонение меньше при любом возможном значении параметра:

$$MSE(\hat{\theta}_n^*) - MSE(\hat{\theta}_n) = \frac{\theta^2}{9n^2+6n} - \frac{2\theta^2}{9n^2+9n+2} = \frac{1-3n}{3n(3n+1)(3n+2)}\theta^2 < 0$$

Задание №4. Дифференциация новогодних подарков. (30 баллов)

Дед Мороз раздает подарки разного качества в зависимости от поведения детей. Он делит детей на послушных, средних и непослушных, кодируя соответствующие категории как 1, 0 и -1 соответственно. Дети попадают в различные категории независимо друг от друга. Категория случайно взятого ребенка описывается следующим распределением:

t	-1	0	1
$P(X_1 = t)$	p	$1-2p$	p

, где $p \in (0, 0.5)$

Дед Мороз собирается раздать подарки n детям. Для того, чтобы на будущее оптимизировать число подготавливаемых подарков разного качества, он хочет оценить параметр p.

1. Оцените параметр p при помощи произведения второго и четвертого начальных моментов. **(5 баллов)**

2. Оцените параметр p при помощи метода максимального правдоподобия. **(5 баллов)**
   Подсказка: вспомните функцию правдоподобия для выборки из распределения Бернулли, а также обратите внимание, что $P(X_1 = -1) = P(X_1 = 1) = p$.

3. Проверьте, является ли найденная вами оценка эффективной. **(5 баллов)**

4. Найдите ММП оценку для выборки из распределения со следующей функцией плотности:

t	-1	0	1
$P(X_1 = t)$	p	$1-3p$	$2p$

   , где $p \in (0, 1/3)$ **(5 баллов)**

   Подсказка: обратите внимание, что при выборке из распределения Бернулли выражение $\prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$ можно заменить на $\prod_{i:X_i=1} p \prod_{i:X_i=0}(1-p)$.

5. Найдите асимптотическую дисперсию найденной вами в предыдущем пункте оценки. **(5 баллов)**

6. Известно, что $p = 0.15$. Рассчитайте вероятность, с которой ММП оценка (из двух предыдущих пунктов) отклонится от истинного значения более, чем на 5%, если Дед Мороз использует выборку из 100 детей. **(5 балла)**

Решение:

1. Обратим внимание, что:

$$E(X_1) = (-1) \times p + 0 \times (1-2p) + 1 \times p = 0$$

$$E(X_1^2) = (-1)^2 \times p + 0^2 \times (1-2p) + 1^2 \times p = 2p$$

$$E(X_1^4) = (-1)^4 \times p + 0^2 \times (1-2p) + 1^4 \times p = 2p$$

В результате получаем:

$$4p^2 = E(X_1^2)E(X_1^4) \implies p = 0.5\sqrt{E(X_1^2)E(X_1^4)} \implies \hat{p}^{MM} = 0.5\sqrt{\overline{X_n^2} \times \overline{X_n^4}} = 0.5\overline{X_n^2}$$

2. Запишем функцию правдоподобия:

$$L(p; x) = \prod_{i=1}^{n} p^{x_i^2}(1-2p)^{1-x_i^2}$$

Найдем ее логарифм:

$$\ln L(p; x) = \ln(p)\sum_{i=1}^{n} x_i^2 + \ln(1-2p)\sum_{i=1}^{n}(1-x_i^2)$$

Рассмотрим условие первого порядка:

$$\frac{d \ln L(p; x)}{dp} = \frac{1}{p}\sum_{i=1}^{n} x_i^2 - \frac{2}{1-2p}\sum_{i=1}^{n}(1-x_i^2) = 0 \implies$$

$$\implies \frac{1-2p}{2p} = \frac{\sum\limits_{i=1}^{n}(1-x_i^2)}{\sum\limits_{i=1}^{n} x_i^2} \implies p^* = \frac{\overline{x_n^2}}{2} \implies \hat{p}_n = \frac{\overline{X_n^2}}{2}$$

Убедимся, что мы нашли максимум, показав, что функция правдоподобия является вогнутой:

$$\frac{d^2 \ln L(p; x)}{d^2 p} = -\frac{1}{p^2}\sum_{i=1}^{n} x_i^2 - \frac{4}{(1-2p)^2}\sum_{i=1}^{n}(1-x_i^2) < 0$$

3. Убедимся, что оценка является несмещенной:

$$E(\hat{p}_n) = E(0.5\overline{X_n^2}) = 0.5E(X_1^2) = 0.5 \times 2p = p$$

Найдем информацию Фишера:

$$i_n(p) = -E\left(-\frac{1}{p^2}X_1^2 - \frac{4}{(1-2p)^2}(1-X_1^2)\right) = \frac{2p}{p^2} + \frac{4(1-2p)}{(1-2p)^2} = \frac{2}{p(1-2p)}$$

$$I_n(p) = ni_n(p) = \frac{2n}{p(1-2p)}$$

В данном случае ММП оценка является эффективной, поскольку совпадает с границей неравенства Рао-Крамера:

$$Var(\hat{p}_n) = Var(0.5\overline{X_n^2}) = \frac{Var(X_1^2)}{4n} = \frac{E((X_1^2)^2) - E(X_1^2)^2}{4n} =$$
$$= \frac{E(X_1^4) - E(X_1^2)^2}{4n} = \frac{2p - (2p)^2}{4n} = \frac{p(1-2p)}{2n} = \frac{1}{I_n(p)}$$

4. Запишем функцию правдоподобия:

$$L(p; x) = \prod_{i:x_i=-1} p^{x_i} \prod_{i:x_i=0} (1-3p)^{x_i} \prod_{i:x_i=1} (2p)^{x_i}$$

Рассмотрим ее логарифм:

$$\ln L(p; x) = \ln(p) \sum_{i:x_i=-1} x_i + \ln(1-3p) \sum_{i:x_i=0} x_i + \ln(2p) \sum_{i:x_i=1} x_i =$$
$$= n_{-1}\ln(p) + n_0 \ln(1-3p) + n_1 \ln(2p)$$

В данном случае n_{-1}, n_0 и n_1 отражают число элементов в выборке, равняющихся -1, 0 и 1 соответственно. Дифференцируя получаем:

$$\frac{d\ln L(p; x)}{dp} = \frac{n_{-1}}{p} - \frac{3n_0}{1-3p} + \frac{2n_1}{2p} = 0$$

Решая, получаем точку, подозреваемую на максимум:

$$p^* = \frac{n_{-1} + n_1}{3(n_{-1} + n_0 + n_1)}$$

Убедимся, что найденная точка является максимумов, рассмотрев условия второго порядка:

$$\frac{d^2 \ln L(p; x)}{d^2 p} = -\frac{n_{-1}}{p^2} - \frac{9n_0}{(1-3p)^2} - \frac{n_1}{p^2} < 0$$

В итоге получаем ММП оценку:

$$\hat{p}_n = \frac{N_{-1} + N_1}{3(N_{-1} + N_0 + N_1)}$$

$$N_k = \sum_{i=1}^{N} I(X_i = k)$$

$$N_{-1} \sim B(n, p) \qquad N_0 \sim B(n, 1-3p) \qquad N_1 \sim B(n, 2p)$$

5. Найдем информацию Фишера:

$$I_n(p) = -E\left(-\frac{N_{-1}}{p^2} - \frac{9N_0}{(1-3p)^2} - \frac{N_1}{p^2}\right) = \frac{np}{p^2} + \frac{9n(1-3p)}{(1-3p)^2} + \frac{2np}{p^2} = \frac{3n}{p(1-3p)}$$

   В результате получаем асимптотическую дисперсию:

$$As.Var(\hat{p}_n) = \frac{p(1-3p)}{3n}$$

6. Обратим внимание, что поскольку $p = 0.1$, то:

$$As.Var(\hat{p}_n) = \frac{0.15(1-3\times 0.15)}{3\times 100} = 0.000275$$

   Пользуясь асимптотической нормальностью ММП оценок получаем, что:

$$(\hat{p}_n - p) \dot{\sim} \mathcal{N}\left(0, 0.000275\right)$$

   Найдем искомую вероятность:

$$P\left(|\hat{p}_n - p| \geq 0.05p\right) = 1 - P\left(|\hat{p}_n - p| < 0.05p\right) = 1 - P(-0.05p < \hat{p}_n - p < 0.05p) =$$

$$1 - P(-0.05\times 0.15 < \hat{p}_n - p < 0.05\times 0.15) = 1 - P(-0.0075 < \hat{p}_n - p < 0.0075)$$

$$= 1 - \left(\Phi\left(\frac{0.0075}{\sqrt{0.000275}}\right) - \Phi\left(\frac{-0.0075}{\sqrt{0.000275}}\right)\right) \approx 2 - 2\Phi(0.452267) \approx 0.651$$