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 **29.11.2021, 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 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

Непрерывные случайные величины и асимптотические теоремы

Задание №1. Фермер. (20 баллов)
Выручка фермера является случайной величиной с функцией плотности:

$$f_X(x) = \begin{cases} 2x, \text{ если } x \in [0, \alpha] \\ 0, \text{ иначе} \end{cases}$$

Издержки фермера постоянны и составляют 0.1. В качестве налогов государство забирает 30% от выручки фермера. Прибыль фермера определяется как разница между выручкой (за вычетом налогов) и постоянными издержками.

1. Найдите параметр α. **(2 балла)**
2. Рассчитайте вероятность того, что выручка фермера превысит 0.5. **(2 балла)**
3. Посчитайте математическое ожидание и дисперсию выручки фермера. **(2 балла)**
4. Запишите функцию распределения выручки фермера. **(2 балла)**
5. Запишите квантиль уровня 0.64 выручки фермера. **(2 балла)**
6. Посчитайте условное математическое ожидание выручки фермера, если известно, что она превысила 0.5. **(3 балла)**
7. Найдите математическое ожидание и дисперсию прибыли фермера. **(2 балла)**
8. Зарплата доярки рассчитывается в зависимости от выручки фермера (и даже может превышать ее) как e^X. Запишите функцию плотности и математическое ожидание зарплаты доярки. **(5 баллов)**
   Подсказка: $\int ln(x)d(x) = x\ln(x) - x + C$

Решение:

1. Пользуясь свойствами функции плотности получаем:

$$\int\limits_0^{\alpha} 2xd(x) = \alpha^2 = 1 \implies \alpha = 1$$

2. Соответствующая вероятность составит:

$$P(X > 0.5) = \int\limits_{0.5}^{1} 2xd(x) = 0.75$$

3. Последовательно рассчитывая необходимые моменты получаем:

$$E(X) = \int\limits_0^1 x \times 2x d(x) = \frac{2}{3}$$

$$E(X^2) = \int\limits_0^1 x^2 \times 2x d(x) = 0.5$$

$$Var(X) = E(X^2) - E(X)^2 = 0.5 - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

4. Функция распределения будет иметь вид:

$$F_X(x) = \begin{cases} 0, \text{ если } x < 0 \\ \int\limits_0^x 2tdt, \text{ если } x \in [0, 1] \\ 1, \text{ если } x > 1 \end{cases} = \begin{cases} 0, \text{ если } x < 0 \\ x^2, \text{ если } x \in [0, 1] \\ 1, \text{ если } x > 1 \end{cases}$$

5. Обозначим искомую квантиль как $x_{0.64}$, откуда:

$$F_X(x_{0.64}) = 0.64 \implies x_{0.64}^2 = 0.64 \implies x_{0.64} = 0.8$$

6. Для начала найдем условную функцию плотности:

$$f_{X|X\geq 0.5}(x) = \begin{cases} \frac{2x}{P(X\geq 0.5)}, \text{ если } x \in [0.5, 1] \\ 0, \text{ если } x \notin [0.5, 1] \end{cases} = \begin{cases} \frac{8}{3}x, \text{ если } x \in [0.5, 1] \\ 0, \text{ если } x \notin [0.5, 1] \end{cases}$$

   С помощью условной функции плотности найдем условное математическое ожидание:

$$E(X|X \geq 0.5) = \int\limits_{0.5}^1 x \times \frac{8}{3} x d(x) \approx 0.778$$

7. Пользуясь свойствами математического ожидания и дисперсии получаем:

$$E((1 - 0.3) \times X - 0.1) = 0.7E(X) - 0.1 = 0.7 \times \frac{2}{3} - 0.1 = \frac{11}{30}$$

$$Var((1 - 0.3) \times X - 0.1) = 0.7^2 Var(X) = 0.49 \times \frac{1}{18} = \frac{49}{1800}$$

8. Поскольку функция e^x строго возрастает, то носитель зарплаты доярки будет иметь вид $\text{supp}(e^X) = [e^0, e^1] = [1, e]$. Найдем функцию плотности зарплаты доярки на носителе, то есть при $x \in [1, e]$:

$$F_{e^X}(x) = P(e^X \leq x) = P(X \leq \ln(x)) = F_X(\ln(x)) = \ln(x)^2$$

Отсюда находим функцию плотности:

$$f_{e^X}(x) = \frac{dF_{e^X}(x)}{dx} = \begin{cases} \frac{2\ln(x)}{x}, \text{ если } x \in [1, e] \\ 0, \text{ в противном случае} \end{cases}$$

Используя полученную функцию плотности найдем математическое ожидание зарплаты доярки:

$$E(e^X) = \int\limits_1^e x \times \frac{2\ln(x)}{x} d(x) = 2\int\limits_1^e \ln(x)d(x) = 2$$

Соответствующее математическое ожидание можно было бы найти и альтернативным образом:

$$\int\limits_0^1 e^x \times 2xd(x) = 2$$

Задание №2. Подозрительный аналитик. (20 баллов)

Цена акции является непрерывной случайной величиной с конечными математическим ожиданием и дисперсией. Проверьте, не ошибается ли аналитик, утверждающий, что в силу высокой волатильности на рынке с вероятностью 0.5 цена акции может:

1. Не менее чем втрое превысить ее математическое ожидание. **(5 баллов)**
2. Отклониться от математического ожидания более, чем на два стандартных отклонения. **(5 баллов)**
3. Превысить цену другой (второй) акции. При этом известно, что математическое ожидание первой акции на 1.5 меньше, чем у второй. Кроме того, максимально возможная разница в ценах акций (по абсолютному значению) не превышает 2. **(10 баллов)**

Решение

1. Обозначим цену акции как X. Поскольку цена не может быть отрицательной, то $P(X < 0) = 0$, а значит в данном случае применимо неравенство Маркова:

$$P\left(X \geq 3E(X)\right) \leq \frac{E(X)}{3E(X)} = \frac{1}{3}$$

Поскольку верхняя граница для соответствующей вероятности меньше 0.5, то утверждение аналитика явно является ошибочным.

2. Применяя неравенство Чебышева нетрудно показать ошибочность утверждения аналитика:

$$P\left(|X - E(X)| > 2\sqrt{Var(X)}\right) \leq \frac{1}{2^2} = \frac{1}{4}$$

3. Обозначим через Y цену второй акции. Известно, что $E(X - Y) = -1.5$ и $P(|X - Y| > 2) = 0$. В этих предположениях противоречия нет, поскольку соответствующим условиям удовлетворяют, например, случайные величины $X \sim U(0, 0.5)$ и $Y \sim U(1.5, 2)$.

   Обратим внимание, что:

$$P(|X-Y| > 2) = P(X-Y+2 < 0)+P(X-Y-2 > 0) = 0 \implies P(X-Y+2 < 0) = 0$$

   Следовательно, $X - Y + 2$ является неотрицательной случайной величиной, а значит в отношении нее применимо неравенство Маркова:

$$P(X - Y \geq 0) = P(X - Y + 2 \geq 2) \leq \frac{E(X - Y + 2)}{2} = \frac{-1.5 + 2}{2} = \frac{1}{4}$$

   Полученный результат демонстрирует ошибочность утверждения аналитика.

Задание №3. Закон Лаврентия. (20 баллов)

Лаврентий бесконечное число раз независимо друг от друга с равной вероятностью загадывает число 10, 100 или 1000. Порадуйте Лаврентия, определив, к чему стремится по вероятности:

1. Среднее арифметическое загаданных им чисел. **(3 балла)**
2. Геометрическое среднее загаданных им чисел. **(7 баллов)**
3. Отношение арифметического среднего к геометрическому среднему загаданных им чисел. **(5 баллов)**
4. Произведение загаданных им чисел, если теперь он каждый раз с равной вероятностью загадывает одно из целых чисел от 0 до 99999 включительно. **(5 баллов)**

Примечание: Ответ сопроводите формальным доказательством.

Подсказка: среднее геометрическое положительных чисел $x_1, ..., x_n$ может быть записано как:

$$\left(\prod_{i=1}^{n} x_i\right)^{1/n} = e^{\frac{\ln(x_1)+...+\ln(x_n)}{n}}$$

Решение

1. Через X_i обозначим случайную величину, отражающую i-е из загаданных Лаврентием чисел. Ее математическое ожидание будет равно:

$$E(X) = \frac{1}{3}(10 + 100 + 1000) = \frac{1110}{3} = 370$$

   Поскольку Лаврентий каждый раз загадывает числа в соответствии с одним и тем же законом, а также независимо друг от друга, то в данной ситуации применим закон больших чисел, вследствие которого:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} 370$$

2. Найдем математическое ожидание логарифма загаданного Лаврентием числа:

$$E\left(X_i\right) = \frac{1}{3}\left(\ln\left(10\right) + \ln\left(100\right) + \ln\left(1000\right)\right) = 2\ln\left(10\right)$$

   Применяя закон больших чисел получаем, что:

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left(X_i\right) \xrightarrow{p} E(\ln\left(X_i\right)) = 2\ln\left(10\right)$$

   В силу непрерывности функции e^x можно применить теорему Манна-Вальда:

$$e^{\frac{1}{n}\sum\limits_{i=1}^{n}\ln(X_i)} \xrightarrow{p} e^{2\ln(10)} = 100$$

3. Поскольку обе рассматриваемые последовательности сходятся по вероятности к некоторым константам, то по теореме Слуцкого получаем, что:

$$\frac{1}{n}\sum_{i=1}^{n} X_i/e^{\frac{1}{n}\sum\limits_{i=1}^{n}} \xrightarrow{p} 370/100 = 3.7$$

4. Интуиция подсказывает, что соответствующая последовательность стремится к нулю, поскольку при большом n по крайней мере один из X_i скорее всего примет нулевое значение. Покажем данный результат формально:

$$\lim_{n\to\infty} P(|X_1 \times ... \times X_n - 0| > \varepsilon) \leq \lim_{n\to\infty} P(X_1 \times ... \times X_n > 0) =$$
$$= \lim_{n\to\infty} P\left(X_1 > 0 \cap X_2 > 0 \cap ... \cap X_n > 0\right) = \lim_{n\to\infty}(0.99999)^n = 0$$

Проверка в R:

```
n.sim <- 10^6
numbers <- c(10, 10^2, 10^3)
lavrent <- sample(numbers, size = n.sim, replace = TRUE)
# пункт 1
mean(lavrent)
# пункт 2
exp(mean(log(lavrent)))
# пункт 3
mean(lavrent) / exp(mean(log(lavrent)))
# пункт 3
prod(sample(0:99999, size = n.sim, replace = TRUE))
```

Задание №4. Видеоролик. (10 баллов)

Число просмотров видеоролика является Пуассоновской случайной величиной с математическим ожиданием 10000. Используя центральную предельную теорему приблизительно посчитайте вероятность того, что ролик посмотрит не более 10100 человек.

Примечание: необходимо обосновать применимость ЦПТ в данном случае.

Подсказка: вспомните свойство воспроизводимости пуассоновских случайных величин.

Решение

Если $X \sim Pois(10000)$, то в силу свойства воспроизводимости распределения Пуассона его можно представить как сумму независимых пуассоновских случайных величин $X_i \sim Pois(1)$, то есть:

$$X = X_1 + ... + X_{10000}$$

Поскольку речь идет о сумме независимых, одинаково распределенных случайных величин, то можно воспользоваться центральной предельной теоремой:

$$\begin{cases} E(X_i) = 1 \\ Var(X_i) = 1 \end{cases} \implies (X_1 + ... + X_{10000}) \dot{\sim} \mathcal{N}\,(10000, 10000)$$

В результате получаем:

$$P(X \leq 10100) = P(X_1 + ... + X_{10000} \leq 10100) \approx \Phi\left(\frac{10100 - 10000}{\sqrt{10000}}\right) \approx 0.841$$

Проверка в R:

```
ppois(10100, lambda = 10000)
pnorm(10100, mean = 10000, sd = sqrt(10000))
```

Задание №5. Большой курс. (30 баллов)

Контрольную работу, включающую 3 задания, пишут 100 студентов. Они решают задания независимо друг от друга. Студент успешно решает каждое задание с вероятностью 0.5, независимо от того, смог ли он успешно решить другие задания. Найдите приблизительную вероятность того, что:

1. Студенты (суммарно) решат верно не более 160 заданий. **(10 балла)**
2. Процент студентов, не решивших ни одного задания, не превышает 10-ти. **(10 баллов)**
3. Общее число успешно решенных заданий (всеми студентами в сумме) окажется хотя бы в 1.7 раза больше числа студентов, успешно решивших по крайней мере одно задание. **(10 баллов)**

Решение

1. Через $X_i \sim B(3, 0.5)$ обозначим случайную величину, отражающую число задач, решенных i-м студентом.

   Обратим внимание, что поскольку студенты решают задачи независимо друг от друга и с равной вероятностью приходят к верному ответу, то можно воспользоваться ЦПТ:

$$E(X_i) = 3 \times 0.5 = 1.5$$
$$Var(X_i) = 3 \times 0.5 \times (1 - 0.5) = 0.75$$
$$(X_1 + ... + X_{100}) \dot{\sim} \mathcal{N}(1.5 \times 100, 0.75 \times 100) = \mathcal{N}(150, 75)$$

   В результате получаем:

$$P(X_1 + ... + X_{100} \leq 160) = \Phi\left(\frac{160 - 150}{\sqrt{75}}\right) \approx 0.876$$

2. Через Y_i обозначим случайную величину, принимающую значение 1, если студент решил хотя бы одну задачу и 0 – в противном случае. При этом обратим внимание, что $Y_i = 1$ когда $X_i \geq 1$ и $Y_i = 0$, если $X_i = 0$, откуда нетрудно показать, что $Y_i \sim Ber(1 - 0.5^3)$, а значит в силу ЦПТ:

$$E(Y_i) = 1 - 0.5^3 = 0.875$$
$$Var(Y_i) = 0.875 - 0.875^2 = 0.109375$$
$$(Y_1 + ... + Y_n) \dot{\sim} \mathcal{N}(0.875 \times 100, 0.109375 \times 100) = \mathcal{N}(87.5, 10.9375)$$

   Посчитаем искомую вероятность:

$$P(\frac{1}{100}(100 - (Y_1 + ... + Y_{100})) \leq 0.1) = P(Y_1 + ... + Y_{100} \geq 90) \approx$$
$$\approx 1 - \Phi\left(\frac{90 - 87.5}{\sqrt{10.9375}}\right) \approx 0.2248459$$

3. Необходимо рассчитать следующую вероятность:

$$P(X_1 + ... + X_{100} \geq 1.7\,(Y_1 + ... + Y_{100})) =$$
$$= P((X_1 - 1.7Y_1) + (X_2 - 1.7Y_2) + ... + (X_{100} - 1.7Y_{100}) \geq 0)$$

   Обратим внимание, что случайные величины $(X_i - 1.7Y_i)$ независимы и одинаково распределены, а значит в отношении их суммы применима центральная предельная теорема.

$$E(X_i - 1.7Y_i) = 1.5 - 1.7 \times 0.875 = 0.0125$$
$$E(X_iY_i) = E(X_i|Y_i = 1)P(Y_i = 1) + E(X_i|Y_i = 0)P(Y_i = 0) =$$
$$= E(X_i|X_i \geq 1)P(X_i \geq 1) + E(X_i|X_i = 0) \times P(X_i = 0) = E(X_i) = 1.5$$
$$Cov(X_i, Y_i) = 1.5 - 1.5 \times 0.875 = 0.1875$$

$$Var(X_i - 1.7Y_i) = 0.75 + 1.7^2 \times 0.109375 - 2 \times 1.7 \times 0.1875 = 0.42859375$$

В результате получаем:

$$(X_i - 1.7Y_i) \dot{\sim} \mathcal{N}\left(0.0125 \times 100, 0.42859375 \times 100\right) = \mathcal{N}\left(1.25, 42.859375\right)$$

Воспользуемся найденным приближением распределения для расчета искомой вероятности:

$$P((X_1 - 1.7Y_1) + (X_2 - 1.7Y_2) + ... + (X_{100} - 1.7Y_{100}) \geq 0) =$$
$$= 1 - \Phi\left(\frac{0 - 1.25}{\sqrt{42.859375}}\right) \approx 0.575712$$

Проверка в R:

```
n.sim <- 10000
n.student <- 100
yes <- rep(NA, n.sim)
works160 <- rep(NA, n.sim)
works90prc <- rep(NA, n.sim)
for (i in 1:n.sim)
{
  works <- rbinom(n = n.student, size = 3, prob = 0.5)
  works160[i] <- sum(works) <= 160
  works90prc[i] <- sum(works >= 1) >= 90
  mean(works >= 1)
  yes[i] <- sum(works) >= (1.7 * sum(works >= 1))
}
# пункт 1
mean(works160)
pnorm((160 - 150) / sqrt(75))
# пункт 2
mean(works90prc)
1 - pnorm((90 - 87.5) / sqrt(10.9375))
# пункт 3
mean(yes)
1 - pnorm(-1.25 / sqrt(42.859375))
```