# 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 **27.02.2022, 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 5”.

**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 №5

### Доверительные интервалы и статистические гипотезы

**Задание №1. Предприятие (20 баллов)**

Прибыль фирмы за месяц (в миллионах рублей) является нормально распределенной случайной величиной. Каждый месяц прибыль фирмы не зависит от прибылей в предыдущие месяцы. Суммарная прибыль за год составила 120 миллионов рублей, а реализация исправленной выборочной дисперсии – 22. Найдите реализацию двухстороннего 90%-го доверительного интервала для:

1. Математического ожидания прибыли фирмы за месяц **(3 балла)**.
2. Дисперсии прибыли фирмы за месяц **(2 балла)**.
3. Дисперсии прибыли фирмы за год **(5 балла)**.
4. Математического ожидания прибыли фирмы за месяц, если известно, что еще за один месяц прибыль фирмы составила 10 миллионов рублей и (истинная) дисперсия прибыли фирмы за месяц равняется 25 **(5 балла)**.
5. Повторите предыдущий пункт предполагая, что дисперсия неизвестна. **(5 балла)**.

**Решение**:

1. Имеется выборка $X = (X_1, ..., X_{12})$ из нормального распределения, то есть $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, где $\mu$ и $\sigma^2$ – неизвестны.

   Из условия известно, что $n = 12$, $\overline{x}_{12} = 120/12 = 10$ и $\hat{\sigma}^2_{12}(x) = 22$. Требуется построить доверительный интервал для математического ожидания наблюдения из нормального распределения с неизвестной дисперсией. Поскольку речь идет о 90%-м доверительном интервале, то $1 - \gamma = 0.9$, а значит $\gamma = 0.1$, откуда получаем необходимую квантиль $t_{12-1,1-0.1/2} = t_{11,0.95} \approx 1.796$. В результате получаем искомую реализацию:

   $$\left[10 - 1.796\sqrt{\frac{22}{12}}, 10 + 1.796\sqrt{\frac{22}{12}}\right] \approx [7.568, 12.432]$$

2. Необходимо построить доверительный интервал для дисперсии наблюдения из нормального распределения с неизвестным математическим ожиданием. Для этого воспользуемся квантилями Хи-квадрат распределения $\chi^2_{12-1,0.05} \approx 4.57$ и $\chi^2_{12-1,0.95} \approx 19.68$, откуда искомая реализация принимает вид:

   $$\left[\frac{22}{(12-1) \times 19.68}, \frac{22}{(12-1) \times 4.57}\right] \approx [0.102, 0.438]$$

3. Дисперсия прибыли за год имеет вид:

   $$Var(X_1 + ... + X_{12}) = 12Var(X_1) = 12\sigma^2$$

   Нетрудно догадаться, что искомая реализация будет иметь вид:

   $$\left[12 \times \frac{22}{(12-1) \times 19.68}, 12 \times \frac{22}{(12-1) \times 4.57}\right] \approx [1.22, 5.25]$$

4. Обратим внимание, что реализация выборочного среднего не поменялась:

$$\overline{x}_{13} = \frac{(12\overline{x}_{12} + x_{13})}{13} = \frac{12 \times 10 + 10}{13} = 10$$

Поскольку дисперсия известна, то воспользуемся квантилью стандартного нормального распределения $z_{0.95} \approx 1.65$, откуда реализация принимает вид:

$$\left[10 - 1.65\sqrt{\frac{25}{13}}, 10 + 1.65\sqrt{\frac{25}{13}}\right] \approx [7.71, 12.29]$$

5. Обратим внимание, что:

$$\hat{\sigma}^2_{13}(x) = \frac{(12-1)\hat{\sigma}^2_{12}(x) + (10-10)^2}{13-1} = \frac{(12-1) \times 22 + (10-10)^2}{13-1} = \frac{121}{6} \approx 20.167$$

Отсюда получаем искомую реализацию:

$$\left[10 - 1.796\sqrt{\frac{\frac{121}{6}}{13}}, 10 + 1.796\sqrt{\frac{\frac{121}{6}}{13}}\right] \approx [7.76, 12.24]$$

**Задание №2. Звезды (15 баллов)**

В далекой-далекой галактике средняя температура случайно выбранной звезды (измеренная в градусах Лаврентия) является нормально распределенной случайной величиной, у которой математическое ожидание совпадет со стандартным отклонением. У первой звезды средняя температура составила 100 градусов, у второй – 120 градусов, а у третьей – 80 градусов.

1. Используя обозначенную в условии информацию об ограничении на параметры постройте 80%-й двухсторонний доверительный интервал для математического ожидания средней температуры случайно выбранной звезды (**5 баллов**).

2. Найдите реализацию построенного вами доверительного интервала (**5 баллов**).

3. Запишите реализацию 80%-го доверительного интервала для математического ожидания произведения температур трех случайно взятых звезд (**5 баллов**).

**Решение**:

1. В данном случае имеется выборка из трех наблюдений $X = (X_1, X_2, X_3)$, причем, в силу наложенного ограничения на равенство математического ожидания и дисперсии, $X_1 \sim \mathcal{N}(\mu, \mu^2)$, где $\mu > 0$. Рассмотрим следующую центральную статистику:

$$\frac{\overline{X}_n - \mu}{\frac{\mu}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

Используя данную центральную статистику, а также принимая во внимание, что $z_{0.9}/\sqrt{3} < 1$, получаем:

$$P\left(-z_{0.9} \leq \frac{\overline{X}_3 - \mu}{\frac{\mu}{\sqrt{3}}} \leq z_{0.9}\right) = P\left(\frac{\overline{X}_n}{1 - z_{0.9}/\sqrt{3}} \leq \mu \leq \frac{\overline{X}_3}{1 + z_{0.9}/\sqrt{3}}\right) = 0.8$$

В результате получаем искомый доверительный интервал:

$$\left[\frac{\overline{X}_3}{1 + z_{0.9}/\sqrt{3}}, \frac{\overline{X}_3}{1 - z_{0.9}/\sqrt{3}}\right]$$

2. Поскольку $z_{0.9} \approx 1.28$ и $\overline{x}_3 = \frac{100+120+80}{3} = 100$, то искомая реализация принимает вид:

$$\left[\frac{100}{1-1.28/\sqrt{3}}, \frac{100}{1+1.28/\sqrt{3}}\right] \approx [57.5, 383.2]$$

3. Получим выражение, для которого необходимо построить доверительный интервал, пользуясь независимостью элементов выборки:

$$E(X_1X_2X_3) = E(X_1)E(X_2)E(X_3) = \mu^3$$

Обратим внимание, что:

$$P\left(\frac{\overline{X}_3}{1-z_{0.9}/\sqrt{3}} \le \mu \le \frac{\overline{X}_3}{1+z_{0.9}/\sqrt{3}}\right) = P\left(\left(\frac{\overline{X}_3}{1-z_{0.9}/\sqrt{3}}\right)^3 \le \mu^3 \le \left(\frac{\overline{X}_3}{1+z_{0.9}/\sqrt{3}}\right)^3\right)$$

Таким образом, искомый доверительный интервал имеет вид:

$$\left[\left(\frac{\overline{X}_3}{1-z_{0.9}/\sqrt{3}}\right)^3, \left(\frac{\overline{X}_3}{1+z_{0.9}/\sqrt{3}}\right)^3\right]$$

Рассчитаем его реализацию:

$$\left[\left(\frac{100}{1-1.28/\sqrt{3}}\right)^3, \left(\frac{100}{1+1.28/\sqrt{3}}\right)^3\right] \approx [1.9 \times 10^5, 5.6 \times 10^7]$$

**Проверка в R:**

```
n.sim <- 100000
n <- 3
mu <- 10
l <- rep(NA, n.sim)
r <- rep(NA, n.sim)
l2 <- rep(NA, n.sim)
r2 <- rep(NA, n.sim)
z <- qnorm(0.9)
for (i in 1:n.sim)
{
  x <- rnorm(n, mu, mu)
  l[i] <- mean(x) / (1 + z / sqrt(n))
  r[i] <- mean(x) / (1 - z / sqrt(n))
}
mean((l < mu) & (r > mu))
```

**Задание №3. Банки (25 баллов)**

Число звонков за час в банки **Случайный вклад** и **Доверительный процент** (работающие круглосуточно и ежедневно) являются независимыми Пуассоновскими случайными величинами с параметрами $\lambda_1$ и $\lambda_2$ соответственно. За неделю в первый банк позвонило 840 клиентов, а во второй – 2520. Реализации исправленных выборочных дисперсий для данных банков оказались равны 6 и 12 соответственно. Постройте 70%-й асимптотический доверительный интервал (и найдите его реализацию) для:

1. Математического ожидания числа ежедневных звонков, поступающих в банк Случайный вклад (без использования ММП оценки) **(5 балла)**.

2. Математического ожидания числа ежедневных звонков, поступающих в банк Случайный вклад (с использованием ММП оценки) **(5 балла)**.

3. Вероятности того, что в банк Случайный вклад за **день** (не час) не поступит ни одного звонка **(4 балла)**.

4. Математического ожидания разницы числа звонков, ежедневно поступающих в соответствующие банки. **(3 балла)**.

5. Математического ожидания суммы числа звонков, ежедневно поступающих в соответствующие банки. **(3 балла)**.

6. Математического ожидания числа ежедневных звонков, поступающих в банк Случайный вклад, не используя ни информацию о реализации выборочной дисперсии, ни ММП оценку (воспользуйтесь ЦПТ и теоремой Слуцкого) **(5 баллов)**.

**Решение**:

1. Обозначим через $X = (X_1, .., X_n)$ выборку из звонков, ежечасно поступавших в банк Случайный вклад на протяжении недели. Очевидно, что $n = 7 \times 24 = 168$. Кроме того, из условия известно, что $\sum\limits_{i=1}^{168} x_i = 840$, откуда $\overline{x}_{168} = 840/168 = 5$. Наконец, по условию $\hat{\sigma}^2_{168}(x) = 5$.

   Обратим внимание, что речь идет о построении асимптотического доверительного интервала для математического ожидания. Кроме того, $1-\gamma = 0.7$, откуда $\gamma = 0.3$, а значит необходимо воспользоваться квантилью $z_{0.85} \approx 1.036$. В результате получаем искомую реализацию:

$$\left[5 - 1.036\sqrt{\frac{6}{168}}, 5 + 1.036\sqrt{\frac{6}{168}}\right] \approx [4.8, 5.2]$$

2. Нетрудно показать, что $\hat{\lambda}_1(x) = 5$ и $i(\hat{\lambda}(x)) = 1/5 = 0.2$, откуда искомая реализация принимает вид:

$$\left[5 - 1.036\sqrt{\frac{1}{0.2 \times 168}}, 5 + 1.036\sqrt{\frac{1}{0.2 \times 168}}\right] \approx [4.82, 5.18]$$

3. Пользуясь свойством воспроизводимости независимых Пуассоновских случайных величин запишем соответствующую вероятность:

$$P(X_1 + ... + X_{24}) = e^{-24\lambda_1}$$

   Дифференцируя получаем:

$$P'(X_1 + ... + X_{24}) = -24e^{-24\lambda_1}$$

   В результате получаем искомую реализацию:

$$\left[e^{-24\times5} - \sqrt{\frac{\left(-24e^{-24\times5}\right)^2}{0.2 \times 168}}, e^{-24\times5} + \sqrt{\frac{\left(-24e^{-24\times5}\right)^2}{0.2 \times 168}}\right]$$

4. Через $Y = (Y_1, ...Y_{168})$ обозначим выборку из числа звонков, поступавших в банк Доверительный вклад. Нетрудно догадаться, что $\overline{y}_{168} = 2520/168 = 15$, откуда получаем реализацию:

$$\left[5 - 15 - 1.036\sqrt{\frac{6+12}{168}}, 5 - 15 + 1.036\sqrt{\frac{6+12}{168}}\right] \approx [-10.34, -9.66]$$

5. Используя ЦПТ и теорему Слуцкого можно показать, что:

$$\frac{\overline{X}_n + \overline{Y}_n - (\lambda_1 + \lambda_2)}{\sqrt{\frac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2}{n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

Применяя соответствующую центральную статистику нетрудно построить необходимый доверительный интервал и найти его реализацию:

$$\left[5 + 15 - 1.036\sqrt{\frac{6+12}{168}}, 5 + 15 + 1.036\sqrt{\frac{6+12}{168}}\right] \approx [19.66, 20.34]$$

6. Используя ЦПТ получаем, что:

$$\frac{\sqrt{n}\left(\overline{X}_n - \lambda\right)}{\sqrt{\lambda}} \xrightarrow{d} \mathcal{N}(0,1)$$

Нетрудно показать, что $\overline{X}_n \xrightarrow{p} \lambda$, а значит по теореме Манна-Вальда $\sqrt{\frac{\lambda}{\overline{X}_n}} \xrightarrow{d} 1$, откуда, по теореме Слуцкого получаем:

$$\frac{\sqrt{n}\left(\overline{X}_n - \lambda\right)}{\sqrt{\lambda}}\sqrt{\frac{\lambda}{\overline{X}_n}} \xrightarrow{d} \mathcal{N}(0,1) \times 1 \implies \frac{\sqrt{n}\left(\overline{X}_n - \lambda\right)}{\sqrt{\overline{X}_n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Исходя из соответствующей тестовой статистики нетрудно построить соответствующий доверительный интервал (фактически он совпадает с тем, что был получен с помощью ММП оценки) и найти его реализацию:

$$\left[5 - 1.036\sqrt{\frac{5}{168}}, 5 + 1.036\sqrt{\frac{5}{168}}\right] \approx [4.82, 5.18]$$

**Задание №4. Волшебная палочка (25 баллов)**

Сила заклинания, произнесенного с помощью волшебной палочки, является нормальной случайной величиной с единичной дисперсией. У хороших волшебных палочек математическое ожидание силы заклинания равняется 10, а у плохих – 7. Лаврентий взял случайную волшебную палочку и хочет проверить, является ли она хорошей. Для этого он три раза произносит заклинание и считает волшебную палочку хорошей, только если сила даже самого слабого из произнесенных заклинаний оказалась больше 8.

**Подсказка:** вспомните вариационный ряд и распределение экстремальных статистик.

1. Сформулируйте нулевую и альтернативную гипотезы теста Лаврентия в параметрическом виде (возможны два верных варианта ответа, получаемые за счет перестановки нулевой и альтернативной гипотез) **(5 баллов)**.

2. Найдите уровень значимости теста Лаврентия **(5 баллов)**.

3. Вычислите мощность теста Лаврентия **(5 баллов)**.

4. **Бонусный пункт**: посчитайте p-value теста Лаврентия, если сила самого слабого заклинания оказалась равна 9 и Лаврентий предполагает, что палочка плохая, если сила самого слабого заклинания не превысила некоторое значение. **(5 баллов)**.

5. Используя лемму Неймана-Пирсона предложите Лаврентию альтернативный критерий тестирования гипотезы и рассчитайте его мощность при 10%-м уровне значимости. **(5 баллов)**.

**Решение**:

1. Обозначим через $X = (X_1, ..., X_5)$ выборку из сил заклинаний, произнесенных Лаврентием и использованием соответствующей волшебной палочки. Нулевая гипотеза о том, что палочка является хорошей, отвергается, если $X_{(1)} < 8$, где $X_{(1)} = \min(X)$. Поскольку $X_1 \sim \mathcal{N}(\mu, 1)$, то гипотезы могут быть сформулированы как $H_0 : \mu = 10$ и $H_1 : \mu = 7$.

2. Рассмотрим тестовую статистику $T(X) = X_{(1)}$ и найдем ее функцию распределения:
$$F_{X_{(1)}}(t) = 1 - (1 - \Phi(t - \mu))^3$$
В результате в зависимости от верности нулевой или альтернативной гипотез получаем:
$$F_{X_{(1)}|H_0}(t) = 1 - (1 - \Phi(t - 10))^3$$
$$F_{X_{(1)}|H_1}(t) = 1 - (1 - \Phi(t - 7))^3$$
Найдем уровень значимости теста рассчитав вероятность ошибки первого рода:
$$\alpha = P(X_{(1)} < 8|H_0) = F_{X_{(1)}|H_0}(8) = 1 - (1 - \Phi(8 - 10))^3 \approx 0.067$$

3. По аналогии рассчитаем мощность теста используя вероятность ошибки второго рода:
$$1 - \beta = 1 - P(X_{(1)} \geq 8|H_1) = P(X_{(1)} < 8|H_1) = 1 - (1 - \Phi(8 - 7))^3 \approx 0.996$$

4. Обратим внимание, что критическая область является левосторонней, откуда:
$$\text{p-value} = F_{X_{(1)}|H_0}(9) = 1 - (1 - \Phi(9 - 10))^3 \approx 0.4$$

5. Запишем тестовую статистику:
$$T_1(X) = \frac{L(X;7)}{L(X;10)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{\sum_{i=1}^{3}(X_i-7)^2}{2}} \Big/ \frac{1}{\sqrt{2\pi}} e^{-\frac{\sum_{i=1}^{3}(X_i-10)^2}{2}} = e^{\frac{\sum_{i=1}^{3}(X_i-10)^2-(X_i-7)^2}{2}}$$

Найти распределение соответствующей тестовой статистики весьма затруднительно. Поэтому, совершим ряд монотонных преобразований над тестовой статистикой. Сперва произведем логарифмирование и умножение на два:

$$T_2(X) = 2\ln\left(T_1(X)\right) = \sum_{i=1}^{3}(X_i - 10)^2 - (X_i - 7)^2 = \sum_{i=1}^{3} 51 - 6X_i$$

Осуществляя ряд тривиальных линейных преобразование получаем:

$$T_3(X) = -\overline{X}_3$$

В результате нулевая гипотеза отвергается на уровне значимости $\alpha$, если:

$$-\overline{X}_3 > c^*_{1-\alpha}$$

Умножая левую и правую части неравенства на минус единицу и осуществляя замену $c_\alpha = -c^*_{1-\alpha}$ получаем:

$$\overline{X}_3 < c_\alpha$$

Обратим внимание, что $c_\alpha$ является квантилью уровня $\alpha$ (при верной нулевой гипотезе) тестовой статистики $T(X) = \overline{X}_3 \sim \mathcal{N}(\mu, 1/3)$, причем $T(X)|H_0 \sim \mathcal{N}(10, 1/3)$. Отсюда нетрудно записать выражение для уровня значимости теста:

$$\alpha = \Phi\left(\sqrt{3}\,(c_\alpha - 10)\right)$$

Поэтому, при $\alpha = 0.1$ получаем:

$$0.1 = \Phi\left(\sqrt{3}\,(c_{0.1} - 10)\right) \implies \sqrt{3}\,(c_{0.1} - 10) \approx -1.281552 \implies c_{0.1} \approx 9.26$$

Пользуясь полученным результатом рассчитаем мощность теста:

$$1 - \beta = 1 - P\left(\overline{X}_3 \geq 9.26|H_1\right) = P\left(\overline{X}_3 < 9.26|H_1\right) = \Phi\left(\sqrt{3}\,(9.26 - 8)\right) \approx 0.985$$

**Задание №5. Гонщики (20 баллов)**

Время (в минутах), затрачиваемое случайно выбранным гонщиком на прохождение трассы, является экспоненциальной случайной величиной. По результатам 200 заездов оказалось, что среднее время, за которое один гонщик проходит трассу, составляет 10 минут, а реализация исправленной выборочной дисперсии – 100. На уровне значимости 0.1 протестируйте (против двухсторонней альтернативы) гипотезу о том, что:

1. Математическое ожидание времени прохождения трассы случайно выбранным гонщиком равняется 11 минутам (не используйте ММП оценку) **(10 баллов)**.

2. Дисперсия времени прохождения трассы случайно выбранным гонщиком равняется 121 (используйте ММП оценку) **(10 баллов)**.

**Решение**:

1. Обозначим через $X = (X_1, ..., X_{100})$ выборку из времени, затраченного гонщиками на прохождение трассы. Причем из условия известно, что $X_1 \sim EXP(\lambda)$. То есть выборка была получена из распределения с конечными математическим ожиданием и дисперсией.

   Тестируется гипотеза $H_0 : E(X_1) = 11$ против альтернативы $H_1 : E(X_1) \neq 11$.

   Рассчитаем реализацию тестовой статистики:

   $$T(x) = \frac{10 - 11}{\sqrt{\frac{100}{200}}} \approx -1.41$$

   Посчитаем p-value:

   $$\text{p-value} = 2\min\left(\Phi(-1.41), 1 - \Phi(-1.41)\right) \approx 2\min(0.08, 0.92) = 0.16$$

   Поскольку $0.16 > 0.1$, то нулевая гипотеза не отвергается на 10%-м уровне значимости.

   В качестве альтернативного решения можно записать критическую область:

   $$\mathcal{T}_{0.1} \approx (-\infty, -1.65) \cup (1.65, \infty)$$

   Поскольку $-1.41 \notin \mathcal{T}_{0.1}$, то мы (ожидаемо) вновь приходим к выводу о том, что нулевая гипотеза не отвергается на 10%-м уровне значимости.

2. Тестируется гипотеза $H_0 : Var(X_1) = 121$ против альтернативы $H_1 : Var(X_1) \neq 121$. Поскольку $Var(X_1) = \frac{1}{\lambda^2}$, то данные гипотезы можно переписать параметрически, как $H_0 : \lambda = \frac{1}{11}$ и $H_1 : \lambda \neq \frac{1}{11}$.

   Учитывая, что $\hat{\lambda}(x) = 1/10$ и $i(\hat{\lambda}(x)) = 1/(1/10)^2 = 100$, получаем реализацию тестовой статистики:

   $$T(x) = \frac{1/10 - 1/11}{\sqrt{\frac{1}{100 \times 200}}} \approx -1.29$$

   Посчитаем p-value:

   $$\text{p-value} = 2\min\left(\Phi(-1.29), 1 - \Phi(-1.29)\right) \approx 2\min(0.0985, 0.9015) = 0.197$$

   Поскольку $0.197 > 0.1$, то нулевая гипотеза не отвергается на 10%-м уровне значимости.