Практика 2.7. Доверительные интервалы.

In [60]:
```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
from scipy.stats import norm
```

1. Пусть $X_1, \dots, X_n$ - выборка из распределения Бернулли с параметром p. $\hat{p} = \overline{X}$ - оценка параметра p. Постройте $(1-\alpha)$ доверительные интервалы для оценки p:
   - используя неравенство Чебышёва.
   - асимптотически нормальный доверительный интервал.

Решение: $\mathbb{E}\hat{p} = p$ (честно посчитаем мат. ожидание). Неравенство Чебышёва:

$$P(|\hat{p} - \mathbb{E}\hat{p}| \geq \varepsilon) = P(|\hat{p} - p| \geq \varepsilon) \leq \frac{\mathbb{V}\hat{p}}{\varepsilon^2} = \alpha$$

α - вероятность того, что истинное значение не попадёт в д.и. Таким образом:

$$\frac{\sqrt{\mathbb{V}\hat{p}}}{\sqrt{\alpha}} = \frac{\mathrm{se}\,(\hat{p})}{\sqrt{\alpha}} = \varepsilon$$

Посчитаем дисперсию $\hat{p}$: $\mathbb{V}\hat{p} = \frac{p(1-p)}{n}$. Поскольку значение p - неизвестно, лучшее, что можно сделать - подставить вместо p его оценку:

$$\hat{\mathbb{V}}\hat{p} = \frac{\hat{p}(1-\hat{p})}{n}$$

И получить, что

$$\hat{\varepsilon} = \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}}$$

Подставив это в неравенство Чебышёва, получим:

$$P(|\hat{p} - p| \geq \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}}) \leq \alpha$$
$$P(|\hat{p} - p| < \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}}) \geq 1 - \alpha$$

Тогда доверительный интервал выглядит так:

$$|\hat{p} - p| < \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}}$$
$$-\frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}} < \hat{p} - p < \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}}$$
$$\hat{p} - \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}} < p < \hat{p} + \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}{\sqrt{\alpha}}$$

Обратите внимание, д.и. строится именно для p, а не для $\hat{p}$ - мы пытаемся узнать, где находится настоящее значение, а не в какую точку попадёт её оценка по выборке.

Теперь построим асимптотически нормальный доверительный интервал. Рассмотрим следующее выражение:

$$\frac{\hat{p}-p}{\mathrm{se}\left(\hat{p}\right)}$$

Можно показать, что при $n \to +\infty$ эта величина сходится к стандартному нормальному распределению:

$$\frac{\hat{p}-p}{\mathrm{se}\left(\hat{p}\right)} \sim N(0,1)$$

Для удобства введём $\xi \sim N(0,1)$. Тогда $\forall x$:

$$P\left(\frac{\hat{p}-p}{\mathrm{se}\left(\hat{p}\right)} \leq x\right) \approx P\left(\xi \leq x\right) = \Phi(x)$$

$$\Phi(x) = \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$$

А теперь построим доверительный интервал:

$$P\left(\left|\frac{\hat{p}-p}{\mathrm{se}\left(\hat{p}\right)}\right| \leq x\right) \approx P\left(|\xi| \leq x\right) = \Phi(x) - \Phi(-x) = 1-\alpha$$

Т.е. значение лежит в промежутке от $-x$ до x с вероятностью $1-\alpha$. Этот x находится так:

$$x = \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

Теперь можно записать доверительный интервал:

$$\left|\frac{\hat{p}-p}{\mathrm{se}\left(\hat{p}\right)}\right| \leq \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

Проделав преобразования, подобно предыдущему пункту задачи, получим:

$$\hat{p} - \mathrm{se}\left(\hat{p}\right) \cdot \Phi^{-1}\left(\frac{1-\alpha}{2}\right) \leq p \leq \hat{p} + \mathrm{se}\left(\hat{p}\right) \cdot \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

Поскольку se $\left(\hat{p}\right)$ зависит от неизвестного p, вместо se подставим $\widehat{\mathrm{se}}$:

$$\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \Phi^{-1}\left(\frac{1-\alpha}{2}\right) \leq p \leq \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \cdot \Phi^{-1}\left(\frac{1-\alpha}{2}\right)$$

Это и есть асимптотически нормальный доверительный интервал.

1. При проверке электрических лампочек выяснилось, что из 400 штук 40 - бракованных. Пользуясь результатами предыдущей задачи, посчитайте 95% доверительные интервалы для вероятности брака. При подсчёте округляйте значения до 2 знаков.

In [61]:
```
alpha = 0.05
n = 400
est_p = 40 / n
z_alpha = np.around(norm.ppf(1 - alpha / 2, loc=0.0, scale=1.0), decimals=2)   # α-квантиль нормального распределения
est_se = np.around(np.sqrt(est_p * (1 - est_p) / n), decimals=2)

print(f'оценка p: {est_p}')
print(f'альфа-квантиль: {z_alpha}')
print(f'оценка стандартного отклонения оценки p: {est_se}')

left_bound = np.around(est_p - est_se / alpha ** 0.5, decimals=2)
right_bound = np.around(est_p + est_se / alpha ** 0.5, decimals=2)
print(f'Д.и. на основе неравенства Чебышёва: {left_bound}; {right_bound}')

left_bound = np.around(est_p - z_alpha * est_se, decimals=2)
right_bound = np.around(est_p + z_alpha * est_se, decimals=2)
print(f'Асимптотически нормальный д.и.: {left_bound}; {right_bound}')
```

```
оценка p: 0.1
альфа-квантиль: 1.96
оценка стандартного отклонения оценки p: 0.02
Д.и. на основе неравенства Чебышёва: 0.01; 0.19
Асимптотически нормальный д.и.: 0.06; 0.14
```

1. Дана выборка из экспоненциального распределения с параметром λ (Функция распределения: $F(x) = 1 - e^{-\lambda x}$ при $x > 0$, иначе 0). Постройте $(1-\alpha)$ доверительный интервал для лямбда с помощью оценки $\hat{\lambda} = \frac{1}{\overline{X}}$. Используйте 3 подхода на основе бутстрепа. В ответе запишите границы. Попадает ли истинное значение в доверительные интервалы? При вычислениях округляйте до 2 знаков. Исходные данные:

In [62]:
```
alpha = 0.05
print(f'α = {alpha}')
lambda_param = 0.1
n = 10
B = 5
seed = 0
rnd = np.random.RandomState(seed)
sample = np.around(rnd.exponential(1 / lambda_param, n))
bootstrap_sample = rnd.choice(sample, (B, n), replace=True)
print(f'Выборка: {sample}')
print(f'Бутстрепные выборки:\n{bootstrap_sample}')
```

```
α = 0.05
Выборка: [ 8. 13.  9.  8.  6. 10.  6. 22. 33.  5.]
Бутстрепные выборки:
[[ 6. 22. 22. 33. 13. 10.  5. 33.  5.  6.]
 [ 8.  8.  8. 10.  8.  9.  8. 33. 13.  8.]
 [ 8.  8. 22.  8. 13.  5.  5.  8.  6. 22.]
 [ 8.  9. 22.  9.  8.  8.  6. 10. 10.  6.]
 [33.  6. 13.  6.  5. 33. 13. 13. 22.  5.]]
```

Решение: Бутстреп - приём, позволяющий оценить что-нибудь по выборке, не прибегая к сложным рассчётам. В данном случае бутстреп позволит не считать интегралы, чтобы оценить математическое ожидание и дисперсию $\hat{\lambda}$.

Суть метода в следующем: настоящая функция распределения аппроксимируется эмпирической. Много раз производится выборка из эмпирической функции распределения, по каждой такой выборке (их называют бутстрепными) можно оценить то, что просят. Получится массив значений, с помощью которого можно оценить разброс значений и построить д.и.

Ещё раз то же самое в виде алгоритма:

- Посчитать значение функционала $\hat{T}$ от исходной выборки.
- Сгенерировать по выборке бутстрепные выборки.
- По каждой бутстрепной выборке посчитать T_i.
- Построить доверительный интервал, пользуясь одной из формул.

Подход первый - асимптотически нормальный д.и. Предположим, что оценка $\hat{\lambda}$ - асимптотически нормальная. Как и в первом задании, требуется оценить стандартное отклонение $\hat{\lambda}$. Вместо того, чтобы считать интегралы, сгенерируем бутстрепные выборки и получим значения: $\hat{\lambda}_1, \hat{\lambda}_2, \ldots$. Теперь стандартное отклонение $\hat{\lambda}$ оценим по формуле:

$$\hat{\text{se}}\left(\hat{\lambda}\right) = \frac{1}{n}\sum_{i=1}^{n} T_i$$

Для данного нам α: $\Phi^{-1}\left(\frac{1-\alpha}{2}\right) \approx 2$. Поэтому асимптотически нормальный доверительный интервал выглядит так:

$$\frac{1}{\overline{X}} - 2\hat{\text{se}}\left(\hat{\lambda}\right) \leq \lambda \leq \frac{1}{\overline{X}} + 2\hat{\text{se}}\left(\hat{\lambda}\right)$$

Второй подход - центральный доверительный интервал. По каждой бутстрепной выборке посчитаем оценки, получим из них массив. Из него выберем верхнюю и нижнюю $\frac{\alpha}{2}$ квантили, и запишем ответ по формуле:

$$2\hat{T} - \hat{F}_{T*}^{-1}\left(1 - \frac{\alpha}{2}\right) \leq \lambda \leq 2\hat{T} + \hat{F}_{T*}^{-1}\left(\frac{\alpha}{2}\right)$$

Здесь $\hat{F}_{T*}^{-1}\left(1 - \frac{\alpha}{2}\right)$ и $\hat{F}_{T*}^{-1}\left(\frac{\alpha}{2}\right)$ - квантили бутстрепных оценок.

Третий подход: квантильный доверительный интервал. Отличается от предыдущего формулой, и тем, что оценку по исходной выборке считать не надо. Вот формула, где для квантилей обозначения те же:

$$\hat{F}_{T*}^{-1}\left(\frac{\alpha}{2}\right) \leq \lambda \leq \hat{F}_{T*}^{-1}\left(1 - \frac{\alpha}{2}\right)$$

Обратите внимание, что квантили поменялись местами.

Зачем 3 подхода? Если все 3 дали примерно одинаковый результат, значит всё хорошо. Если нет, то или данных мало, или ошибка где-то, или не повезло (seed плохой). Кроме того, может оказаться так, что оценка вовсе не асимптотически нормальная, тогда первый подход выдаст неизвестно что.

Решение в виде кода:

In [63]:
```
print(f'Средние значения в бутстрепных выборках: {np.mean(bootstrap_sample, axis=1)}')
bootstrap_est_sample = np.around(np.power(np.mean(bootstrap_sample, axis=1), -1), decimals=2)
    # бутстрепные оценки lambda
print(f'Бутстрепные оценки лямбда: {bootstrap_est_sample}')

est_func = np.around(1 / np.mean(sample), decimals=2)
print(f'оценка по выборке: {est_func}')
estse_boot = np.around(bootstrap_est_sample.std(), decimals=2)  # стандартное отклонение оценки
print(f'стандартное отклонение оценки: {estse_boot}')
# асимптотически нормальный доверительный интервал
z_alpha = np.around(norm.ppf(1 - alpha / 2, loc=0.0, scale=1.0), decimals=2)
    # α-квантиль нормального распределения
print(f'α-квантиль: {z_alpha}')
left_bound = np.around(est_func - z_alpha * estse_boot, decimals=2)
right_bound = np.around(est_func + z_alpha * estse_boot, decimals=2)
print(f'асимптотически нормальный интервал '
      f'для α = {alpha}: ({left_bound}; {right_bound})')

higher_quantile, lower_quantile = np.percentile(
        bootstrap_est_sample, np.array([alpha / 2, 1 - alpha / 2]) * 100)  # квантили оценки

left_bound = np.around(2 * est_func - lower_quantile, decimals=2)
right_bound = np.around(2 * est_func - higher_quantile, decimals=2)
print(f'центральный доверительный интервал для α = {alpha}: ({left_bound}; {right_bound})')

left_bound = higher_quantile
right_bound = lower_quantile
print(f'квантильный доверительный интервал для α = {alpha}: ({left_bound}; {right_bound})')
```

```
Средние значения в бутстрепных выборках: [15.5 11.3 10.5  9.6 14.9]
Бутстрепные оценки лямбда: [0.06 0.09 0.1  0.1  0.07]
оценка по выборке: 0.08
стандартное отклонение оценки: 0.02
α-квантиль: 1.96
асимптотически нормальный интервал для α = 0.05: (0.04; 0.12)
центральный доверительный интервал для α = 0.05: (0.06; 0.1)
квантильный доверительный интервал для α = 0.05: (0.061; 0.1)
```