Федеральное государственное автономное образовательное учреждение высшего образования «Национальный Исследовательский Университет ИТМО»

ЛАБОРАТОРНАЯ РАБОТА №2
ПРЕДМЕТ «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»
ТЕМА «ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ»

Вариант 1, 1

Преподаватель: Лимар И. А.
Студент: Румянцев А. А.
Поток: Мат Стат 31.2

Факультет: СУиР
Группа: R3341

Санкт-Петербург
2024

Содержание

1 Задание 1

1.1 Условие

Предъявите доверительный интервал уровня $1-\alpha$ для указанного параметра при данных предположениях (с математическими обоснованиями). Сгенерируйте 2 выборки объёма объёма 25 и посчитайте доверительный интервал. Повторить 1000 раз. Посчитайте, сколько раз 95-процентный доверительный интервал покрывает реальное значение параметра. То же самое сделайте для объема выборки 10000. Как изменился результат? Как объяснить? Что изменяется при росте объемов выборок?

Даны две независимые выборки X_1, X_2 из нормальных распределений $\mathcal{N}(\mu_1, \sigma_1^2)$, $\mathcal{N}(\mu_2, \sigma_2^2)$ объемов n_1, n_2 соответственно. Сначала указывается оцениваемая функция, потом данные об остальных параметрах, затем параметры эксперимента и подсказки.

$\tau = \mu_1 - \mu_2$; σ_1^2, σ_2^2 известны; $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$; воспользуйтесь функцией

$$\frac{\overline{X_1} - \overline{X_2} - \tau}{\sigma}, \quad \sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

1.2 Выполнение

Пусть $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ – независимая выборка из нормального распределения, $\overline{X}$ – выборочное среднее. Тогда, по теореме Фишера, среднее выборочное также имеет нормальное распределение

$$Z = \sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

По условию задания имеем неизвестный параметр μ – математическое ожидание (генеральное среднее) случайной величины X. В качестве точечной оценки параметра μ возьмем выборочное среднее $\hat{\mu} = \overline{X}$. Для уточнения приближенного равенства $\mu \approx \overline{X}$ построим доверительный интервал, накрывающий параметр μ с заданной доверительной вероятностью

$$\gamma = 1 - \alpha,$$

при этом статистика

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

имеет нормальное распределение с параметрами

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Обозначим границы интервалов через квантили порядков $\frac{\alpha}{2}$ и $1 - \frac{\alpha}{2}$ соответственно

$$r_1 = x_{\frac{\alpha}{2}}, \ r_2 = x_{1-\frac{\alpha}{2}},$$

где r_1 – нижняя граница доверительного интервала, r_2 – верхняя. Тогда, доверительная вероятность γ удовлетворяет соотношению

$$\mathbb{P}\left(r_1 \le \sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma} \le r_2\right) = \gamma = 1 - \alpha,$$

что соответствует рис. 1

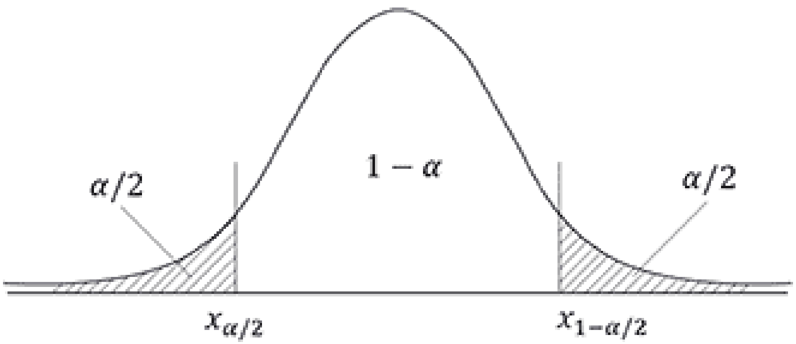

Рис. 1: Двусторонняя критическая область

Пользуясь свойством симметричности нормального распределения

$$r_1 = -r_2 = -x_{1-\frac{\alpha}{2}}.$$

Таким образом, исходя из приведенного ранее соотношения и симметричности границ интервала, необходимо выразить интервал для μ из выражения

$$\left|\sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma}\right| \leq x_{1-\frac{\alpha}{2}}$$

В нашем случае имеем две выборки из нормальных распределений. Исходя из предыдущих рассуждений, обе эти выборки также будут иметь нормальные распределения с параметрами

$$\overline{X_1} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \overline{X_2} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Так как выборочное среднее и математическое ожидание обладают свойством линейности, то разность выборок с нормальными распределениями даст выборку с нормальным распределением с параметрами

$$\overline{X_1} - \overline{X_2} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

что с учетом наших замен можно записать как

$$\overline{X_1} - \overline{X_2} \sim \mathcal{N}\left(\tau, \sigma^2\right).$$

Преобразуем к такому виду, чтобы справа осталось стандартное нормальное распределение $\mathcal{N}(0,1)$. Вычтем из левой и правой части τ

$$\overline{X_1} - \overline{X_2} - \tau \sim \mathcal{N}\left(0, \sigma^2\right),$$

теперь поделим правую часть на σ^2, а левую на $\sqrt{\sigma^2}$

$$\frac{\overline{X_1} - \overline{X_2} - \tau}{\sigma} \sim \mathcal{N}(0,1).$$

Таким образом слева получили данное в условии задания выражение, которое имеет стандартное нормальное распределение. Преобразуем дисперсию генеральной совокупности так, чтобы при подстановке в выражение выше получить искомое выражение

$$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{n_2\cdot\sigma_1^2 + n_1\cdot\sigma_2^2}{n_1\cdot n_2}} = \frac{\sqrt{n_2\cdot\sigma_1^2 + n_1\cdot\sigma_2^2}}{\sqrt{n_1\cdot n_2}}.$$

Подставим преобразованную дисперсию в данное в задаче выражение

$$\frac{\overline{X_1} - \overline{X_2} - \tau}{\frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}}} = \sqrt{n_1 \cdot n_2} \cdot \frac{\overline{X_1} - \overline{X_2} - \tau}{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}$$

Мы получили выражение, похожее на искомое. Теперь определеим доверительный интервал для τ. Выражения под корнями будут всегда положительны, так как в любой выборке должен быть хотя бы один элемент, дисперсия в четной степени и сумма положительных значений не может быть отрицательной. Следовательно, вынесем их из под модуля

$$\left| \sqrt{n_1 \cdot n_2} \cdot \frac{\overline{X_1} - \overline{X_2} - \tau}{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}} \right| \le x_{1-\frac{\alpha}{2}} \Rightarrow \left| \overline{X_1} - \overline{X_2} - \tau \right| \le x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}}$$

Раскроем модуль и преобразуем неравенство так, чтобы в его рамках осталась только τ

$$-x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}} \le \overline{X_1} - \overline{X_2} - \tau \le x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}},$$

$$-\left(\overline{X_1} - \overline{X_2}\right) - x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}} \le -\tau \le -\left(\overline{X_1} - \overline{X_2}\right) + x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}},$$

$$\left(\overline{X_1} - \overline{X_2}\right) - x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}} \le \tau \le \left(\overline{X_1} - \overline{X_2}\right) + x_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{n_2 \cdot \sigma_1^2 + n_1 \cdot \sigma_2^2}}{\sqrt{n_1 \cdot n_2}}$$

Теперь свернем дисперсию к изначальному виду и запишем полученный доверительный интервал

$$\left(\overline{X_1} - \overline{X_2}\right) - x_{1-\frac{\alpha}{2}} \cdot \sigma \le \tau \le \left(\overline{X_1} - \overline{X_2}\right) + x_{1-\frac{\alpha}{2}} \cdot \sigma$$

Проведем эксперимент. Для начала импортируем необходимые библиотеки

```
import numpy as np
import scipy.stats as st
```

Листинг 1: Импортирование необходимых библиотек

Запишем в переменные данные в условии значения параметров. Зададим стандартный уровень значимости $\alpha = 5\% = 0.05$

```
mu_1, mu_2 = 2, 1
tau = mu_1 - mu_2
sigma2_1, sigma2_2 = 1, 0.5
n_1, n_2 = 25, 25
alpha = 0.05
it = 1000
```

Листинг 2: Задаем данные по условию

Реализуем основной алгоритм – в цикле итерируемся it раз и считаем n доверительных интервалов по выведенной формуле. Проверим, попадает ли реальное значение параметра в доверительный интервал. Если True, то увеличим счетчик на 1. В конце выведем количество попавших в интервал τ и отношение относительно общего количества итераций

```
count = 0
for i in range(it):
    X_1 = np.random.normal(mu_1, sigma2_1, n_1)
    X_2 = np.random.normal(mu_2, sigma2_2, n_2)
```

```
        X_1_mean = np.mean(X_1)
        X_2_mean = np.mean(X_2)

        sigma = np.sqrt(sigma2_1 / n_1 + sigma2_2 / n_2)

        z = st.norm.ppf(1 - alpha / 2)

        lower_bound = (X_1_mean - X_2_mean) - z * sigma
        upper_bound = (X_1_mean - X_2_mean) + z * sigma

        print(f'{lower_bound:.4f}<={tau}<={upper_bound:.4f}')

        if lower_bound <= tau <= upper_bound:
            count += 1

    print(f'covers_tau_count={count}, ratio={count / it}')
```

Листинг 3: Код для подсчета доверительных интервалов и кол-ва попаданий

Выведем доверительный интервал для $n = 25$, $it = 1$

```
0.5934<=1<=1.5536
```

Листинг 4: Посчитанный доверительный интервал

Выведем количество попавших в интервал τ и отношение относительно общего количества итераций для $n = 25$, $it = 1000$. Сами доверительные интервалы можно посмотреть в приложении 1

```
covers_tau_count=970, ratio=0.97
```

Листинг 5: 95-% доверительный интервал для $n = 25$

Сделаем то же самое для $n = 10000$

```
covers_tau_count=960, ratio=0.96
```

Листинг 6: 95-% доверительный интервал для $n = 10000$

При увеличении объема выборки отношение количества попавших в доверительный интервал τ к общему количеству проверок на попадание в доверительный интервал стремится к уровню доверия $\gamma = 1 - \alpha = 1 - 0.05 = 0.95$. Доверительные интервалы становятся точнее, так как стандартная ошибка уменьшается. Однако большой объем выборки не гарантирует результат ровно в 95% – из-за вариации в выборках значение будет чаще всего либо больше, либо меньше 0.95.

2 Задание 2

2.1 Условие

Постройте асимптотический доверительный интервал уровня $1 - \alpha$ для указанного параметра. Проведите эксперимент по схеме, аналогичной первой задаче.

Сначала указывается класс распределений (однопараметрический), затем параметры эксперимента и подсказки.

$$\mathrm{Exp}(\lambda);\ \text{медиана};\ \lambda = 1$$

2.2 Выполнение

Экспоненциальное распределение задается следующим образом

$$f_X(x) = \begin{cases} \lambda \cdot \exp\{-\lambda x\}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Предположим, что некоторое распределение зависит от единственного параметра, для которого построен доверительный интервал. Известно, что при увеличении объема выборки оценка этого параметра асимптотически нормальна

$$\hat{\theta}_n \xrightarrow{n\to\infty} \mathcal{N}\left(\theta; \frac{\sigma^2(\theta)}{n}\right),$$

тогда, чтобы построить доверительный интервал, нужно воспользоваться одним из двух подходов:

1. Нормировать величину θ
2. Найти функцию преобразования

В нашем случае я буду использовать первый подход. Так как мы снова работаем с нормальным распределением, то рассуждения относительно построения доверительного интервала для параметра θ аналогичны. Обозначим $u_{1-\frac{\alpha}{2}}$ квантиль порядка $1-\frac{\alpha}{2}$

$$\left|\sqrt{n} \cdot \frac{\hat{\theta}_n - \theta}{\sigma(\theta)}\right| \sim \mathcal{N}(0,1), \quad \left|\sqrt{n} \cdot \frac{\hat{\theta}_n - \theta}{\sigma(\theta)}\right| \leq u_{1-\frac{\alpha}{2}}$$

$$\hat{\theta}_n - \frac{\sigma(\theta)}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \leq \theta \leq \hat{\theta}_n + \frac{\sigma(\theta)}{\sqrt{n}} u_{1-\frac{\alpha}{2}}$$

$$\hat{\theta}_n - SE\left[\hat{\theta}_n\right] u_{1-\frac{\alpha}{2}} \leq \theta \leq \hat{\theta}_n + SE\left[\hat{\theta}_n\right] u_{1-\frac{\alpha}{2}}$$

$$SE\left[\hat{\theta}_n\right] = \sqrt{\mathrm{Var}\left[\hat{\theta}_n\right]}$$

$$\hat{m} - SE[\hat{m}]\, u_{1-\frac{\alpha}{2}} \leq m \leq \hat{m} + SE[\hat{m}]\, u_{1-\frac{\alpha}{2}}$$

$$L(\lambda, x) = \prod_{i=1}^{n} f_X(x) = \lambda^n \cdot \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\}$$

$$\ln L(\lambda, x) = \ln\left(\lambda^n \cdot \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\}\right) = \ln \lambda^n + \ln \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\},$$

$$\ln \lambda^n = n \cdot \ln \lambda, \quad \ln \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\} = -\lambda \sum_{i=1}^{n} x_i \cdot \ln e = -\lambda \sum_{i=1}^{n} x_i,$$

$$\ln L(\lambda, x) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \mathrm{argmax}\left(L(\lambda, x)\right) \Rightarrow \frac{\partial \ln L(\lambda, x)}{\partial \lambda} = 0,$$

$$\frac{\partial \ln L(\lambda, x)}{\partial \lambda} = \left(n \ln \lambda - \lambda \sum_{i=1}^{n} x_i\right)'_{\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \frac{n}{\lambda} = \sum_{i=1}^{n} x_i \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i},$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \sum_{i=1}^{n} x_i = n \cdot \overline{X} \Rightarrow \hat{\lambda} = \frac{n}{n \cdot \overline{X}} = \frac{1}{\overline{X}}$$

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

$$F_X(x) = \begin{cases} 1 - \exp\{-\lambda x\}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$F_X(x) \in [0; 1] \Rightarrow \mathrm{Med}[x] = 0.5$$

$$F_X(x = m) = 1 - \exp\{-\lambda m\} = 0.5 \Rightarrow \exp\{-\lambda m\} = 0.5 \Rightarrow \ln e^{-\lambda m} = \ln 0.5,$$

$$\ln e^{-\lambda m} = -\lambda m \cdot \ln e = -\lambda m, \quad \ln 0.5 = \ln 2^{-1} = -\ln 2,$$

$$-\lambda m = -\ln 2 \Rightarrow \lambda m = \ln 2 \Rightarrow m = \frac{\ln 2}{\lambda} \Rightarrow \hat{m} = \frac{\ln 2}{\hat{\lambda}} = \ln 2 \cdot \overline{X}$$

$$\hat{m} = \overline{X} \ln 2$$

$$\mathrm{Var}[\hat{m}] = \mathrm{Var}\left[\overline{X} \ln 2\right] = \mathrm{Var}\left[\frac{\ln 2}{n}\sum_{i=1}^{n} x_i\right] = \left(\frac{\ln 2}{n}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} x_i\right],$$

$$\mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] = /\text{случ. вел. независимы}/ = \sum_{i=1}^{n} \mathrm{Var}[x_i] = n \cdot \mathrm{Var}[x],$$

$$\mathrm{Var}[x] = /\text{wiki}/ = \frac{1}{\lambda^2} \Rightarrow n \cdot \mathrm{Var}[x] = \frac{n}{\lambda^2},$$

$$\left(\frac{\ln 2}{n}\right)^2 \mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] = \left(\frac{\ln 2}{n}\right)^2 \cdot \frac{n}{\lambda^2} = \frac{(\ln 2)^2}{\lambda^2 n}$$

$$\mathrm{Var}[\hat{m}] = \frac{(\ln 2)^2}{\lambda^2 n}$$

$$SE[\hat{m}] = \sqrt{\mathrm{Var}[\hat{m}]} = \sqrt{\frac{(\ln 2)^2}{\lambda^2 n}} = \frac{\ln 2}{\lambda\sqrt{n}}$$

$$\overline{X}\ln 2 - \frac{\ln 2}{\lambda\sqrt{n}} u_{1-\frac{\alpha}{2}} \leq m \leq \overline{X}\ln 2 + \frac{\ln 2}{\lambda\sqrt{n}} u_{1-\frac{\alpha}{2}}$$

```python
import numpy as np
import scipy.stats as st
```

Листинг 7: import

```python
_lambda = 1
med = np.log(2) / _lambda
n = 25
alpha = 0.05
it = 1000
```

Листинг 8: pars

```
count = 0
for i in range(it):
    X = np.random.exponential(scale = 1 / _lambda, size=n)

    X_mean = np.mean(X)
    hat_m = np.log(2) * X_mean

    sigma = np.log(2) / (_lambda * np.sqrt(n))

    z = st.norm.ppf(1 - alpha / 2)

    lower_bound = hat_m - z * sigma
    upper_bound = hat_m + z * sigma

    print(f'{lower_bound:.4f}<={med:.4f}<={upper_bound:.4f}')

    if lower_bound <= med <= upper_bound:
        count += 1

print(f'covers_med_count={count}, ratio={count / it}')
```

Листинг 9: code2

$it = 1$

```
0.2781<=0.6931<=0.8216
```

Листинг 10: $n = 25$

$it = 1000$

```
covers_med_count=950, ratio=0.95
```

Листинг 11: $n = 25$

```
covers_med_count=948, ratio=0.948
```

Листинг 12: $n = 10000$

Тут будут выводы (аналогичные)

3 Приложения

3.1 Приложение 1

Доверительные интервалы для 1000 итераций для первого и второго заданий можно посмотреть в прикрепленном файле `intervals.txt`