

ФКН МНаД: ТВиМС 2026.
Обзор решений листа задач #9.
Доверительные интервалы.

1. Предположим, что случайная выборка будет взята из нормального распределения с неизвестным средним μ и стандартным отклонением $\sigma = 2$. Какого размера должна быть случайная выборка, чтобы для любого возможного значения μ выполнялось следующее условие:

$$P(|\bar{X} - \mu| \leq 0.1) \geq 0.95.$$

Объяснение:

Пусть $X_1, \dots, X_n$ — выборка из $\mathcal{N}(\mu, \sigma^2)$, где $\sigma = 2$, μ неизвестно.

Тогда

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

Требование задачи:

$$P(|\bar{X} - \mu| \leq 0.1) = P(-0.1 \leq \bar{X} - \mu \leq 0.1).$$

Делим все части двойного неравенства на $2/\sqrt{n} > 0$:

$$P\left(\frac{-0.1}{\frac{2}{\sqrt{n}}} \leq \frac{\bar{X} - \mu}{\frac{2}{\sqrt{n}}} \leq \frac{0.1}{\frac{2}{\sqrt{n}}}\right) \geq 0.95.$$

В центральной части неравенства получаем

$$Z = \frac{\bar{X} - \mu}{2/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

И все неравенство теперь выглядит как:

$$P\left(-\frac{0.1\sqrt{n}}{2} \leq Z \leq \frac{0.1\sqrt{n}}{2}\right) \geq 0.95.$$

По симметрии плотности стандартного нормального распределения в пограничном случае хвосты суммарно дают 0.05, то есть по 0.025 на каждый хвост. Тогда

$$\frac{0.1\sqrt{n}}{2} \geq z_{0.975} = 1.96.$$

Следовательно,

$$\sqrt{n} \geq 39.2 \iff n \geq 1536.64.$$

Минимальный подходящий целый размер выборки: $n = 1537$.

2. Предположим, что случайная выборка будет взята из распределения Бернулли с параметром $p = 0.2$. Найдите приблизительно размер случайной выборки, который должен быть взят, чтобы выполнялось следующее условие:

$$P(|\hat{p} - p| \leq 0.1) \geq 0.95.$$

Напоминание: $\hat{p}$ — выборочная доля, случайная величина, показывающая долю единиц в случайной выборке. Также интерпретируется, как отношение числа успешных исходов и/или положительных ответов в случайной выборке к размеру выборки.

Объяснение:

Пусть $X_1, \dots, X_n$ — выборка из распределения Бернулли с $P(X_i = 1) = p = 0.2$.

Тогда $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ и при больших n (по ИТМЛ):

$$\hat{p} \approx \mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Требование задачи:

$$P(|\hat{p} - p| \leq 0.1) = P(-0.1 \leq \hat{p} - p \leq 0.1).$$

Делим все части двойного неравенства на $\sqrt{p(1-p)/n} > 0$:

$$P\left(\frac{-0.1}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq \frac{0.1}{\sqrt{p(1-p)/n}}\right) \geq 0.95.$$

В центральной части неравенства получаем

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx \mathcal{N}(0, 1).$$

При $p = 0.2$:

$$P\left(-\frac{0.1}{\sqrt{0.16/n}} \leq Z \leq \frac{0.1}{\sqrt{0.16/n}}\right) \geq 0.95.$$

По симметрии стандартной нормали в пограничном случае хвосты суммарно дают 0.05, то есть по 0.025 на каждый хвост. Тогда

$$\frac{0.1}{\sqrt{0.16/n}} \geq z_{0.975} = 1.96 \iff \frac{\sqrt{n}}{4} \geq 1.96 \iff n \geq (7.84)^2 \approx 61.47.$$

Минимально подходящий целый размер выборки: $n = 62$.

3. Для определения среднего возраста своих клиентов крупный производитель одежды провёл случайную выборку из 50 клиентов и получил $\bar{x} = 36$. Известно, что стандартное отклонение генеральной совокупности $\sigma = 12$:

- Постройте 98% доверительный интервал для среднего возраста μ всех клиентов.
- Предположим, что требуется, чтобы 92% доверительный интервал был строго равен $[\bar{X} - 2, \bar{X} + 2]$. Какой размер выборки для этого потребуется?

Объяснение:

Исследуемая случайная величина X — возраст клиента производителя одежды, $E[X] = \mu$. Истинная дисперсия X известна: $\sigma = 12$. Так как размер выборки $n = 50$, по ЦПТ можем сказать, что выборочное среднее $\bar{X}$ имеет приблизительно нормальное распределение.

$(1 - \alpha)100\%$ доверительный интервал для неизвестного математического ожидания μ при известной дисперсии:

$$\mu \in \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

- $n = 50$, $\bar{x} = 36$, $\sigma = 12$, $\alpha = 0.02$, $z_{0.01} = 2.326$:

$$\mu \in \left(36 - 2.326 \cdot \frac{12}{\sqrt{50}}, 36 + 2.326 \cdot \frac{12}{\sqrt{50}}\right) = (32.05, 39.95)$$

- Полуширина интервала $E = 2$, $\alpha = 0.08$, $z_{0.04} = 1.751$:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} = \frac{1.751^2 \cdot 12^2}{2^2} \approx 111$$

4. Менеджер ресторана хочет оценить среднюю сумму m , которую посетитель тратит на обед. По выборке из $n = 36$ посетителей получено выборочное среднее $\bar{x} = 3.60$ (в долларах). Известно, что стандартное отклонение расходов одного посетителя равно $\sigma = 0.72$.

Найдите уровень доверия, которому соответствует доверительный интервал $(3.5; 3.7)$ для m .

Объяснение:

Исследуемая случайная величина X — сумма, которую посетитель тратит на обед, $E[X] = m$. Истинное стандартное отклонение известно: $\sigma = 0.72$. По выборке $n = 36$ получено $\bar{x} = 3.60$.

При известной σ (и достаточно большом n) $(1 - \alpha)100\%$ доверительный интервал для m :

$$m \in \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right).$$

Полуширина интервала $E = 3.7 - \bar{x} = \bar{x} - 3.5 = 0.1$. Из формулы $E = z_{\alpha/2} \sigma / \sqrt{n}$:

$$z_{\alpha/2} = \frac{E\sqrt{n}}{\sigma} = \frac{0.1 \cdot 6}{0.72} = \frac{5}{6}.$$

Уровень доверия — это $1 - \alpha = P(|Z| \leq z_{\alpha/2})$ для $Z \sim \mathcal{N}(0, 1)$:

$$1 - \alpha = 2\Phi\left(\frac{5}{6}\right) - 1 \approx 0.594.$$

5. Проведена случайная выборка из 200 студентов. 30 из них говорят, что им "очень нравится" статистика.
- (а) Вычислите долю студентов в этой выборке, которые говорят, что им "очень нравится" статистика, и затем постройте 95% доверительный интервал для истинной доли.
 - (б) Теперь предположим, что вы решили спросить снова, но уже других студентов того же возраста. На этот раз в выборке 40 студентов, и 16 из них говорят, что им "очень нравится" статистика. Постройте 95% доверительный интервал для истинной доли в этом случае. Подумайте, почему два одинаковых по уровню доверия интервала для "поимки" одного и того же параметра могут отличаться.

Объяснение:

Исследуемая случайная величина X — бинарная случайная величина, показывающая, нравится ли студенту статистика ($X = 1$ если "очень нравится" $X = 0$ иначе), $E[X] = p$. Так как размер выборки $n = 200 > 30$, по ИТМЛ можем использовать нормальное приближение для выборочной доли $\hat{p}$.

- (а) $\hat{p} = \frac{30}{200} = 0.15$, $z_{0.025} = 1.96$:

$$p \in \left(0.15 - 1.96\sqrt{\frac{0.15 \cdot 0.85}{200}}, 0.15 + 1.96\sqrt{\frac{0.15 \cdot 0.85}{200}} \right) = (0.100, 0.200)$$

- (б) $\hat{p} = \frac{16}{40} = 0.40$, $z_{0.025} = 1.96$:

$$p \in \left(0.40 - 1.96\sqrt{\frac{0.40 \cdot 0.60}{40}}, 0.40 + 1.96\sqrt{\frac{0.40 \cdot 0.60}{40}} \right) = (0.248, 0.552)$$

Интервалы отличаются из-за того, что выборочная доля — случайная величина, и мы каждый раз получаем разные её реализации, но более того, при разных размерах выборок у этой случайной величины разные дисперсии. Чем меньше выборка, тем больше дисперсия, и тем шире интервал при прочих равных условиях.

6. Маркетинговая команда премиальной кофейной сети хочет понять предпочтения клиентов относительно нового сезонного напитка. Они провели опрос среди 400 случайно выбранных клиентов, и 156 из них выразили заинтересованность в новом напитке.
- (а) Постройте 90% доверительный интервал для истинной доли всех клиентов, которые были бы заинтересованы в новом сезонном напитке.
 - (б) Маркетинговая команда хочет быть уверена на 95%, что их оценка p находится в пределах ± 0.03 (3 процентных пункта) от истинного значения. Какой размер выборки им потребуется для достижения этого уровня точности?

- (с) Основываясь на доверительном интервале из пункта 1, порекомендовали бы вы запуск нового напитка, если компания требует, чтобы по крайней мере 35% клиентов были заинтересованы для того, чтобы запуск был прибыльным? Объясните ваши рассуждения.

Объяснение:

Пусть X — бернуллиевская случайная величина (клиент заинтересован / не заинтересован):

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p, \quad E[X] = p.$$

Для выборки размера $n = 400$ имеем

$$\tilde{p} = \frac{156}{400} = 0.39.$$

При большом n используем нормальное приближение:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right), \quad p \in \left(\tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\right).$$

- (а) Для 90% уровня доверия: $\alpha = 0.1$, $z_{0.05} = 1.645$.

$$p \in \left(0.39 \pm 1.645 \sqrt{\frac{0.39 \cdot 0.61}{400}}\right) \approx (0.35, 0.43).$$

- (б) Для требуемой точности $E = 0.03$ при доверии 95% ($z_{0.025} = 1.96$):

$$E = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \implies n \geq \frac{z_{\alpha/2}^2 p(1-p)}{E^2}.$$

Так как выборочная доля — случайная величина, и мы не знаем, какая она будет в новой выборке, то для оценки используем наихудший сценарий, который максимально увеличивает дисперсию величины Бернулли:

$$p(1-p) \leq \frac{1}{4}, \quad \text{максимум при } p = \frac{1}{2}.$$

Тогда гарантированно достаточно

$$n \geq \frac{1.96^2 \cdot 0.25}{0.03^2} \approx 1067.1,$$

то есть нужен минимум $n = 1068$.

- (с) Для условия прибыльности нужно обосновать $p \geq 0.35$. По интервалу из пункта 1 нижняя граница как раз равна 0.35, поэтому мы можем ожидать, что истинная доля выше 35%, что подходит для запуска нового напитка.

7. Рассмотрим случайную выборку размера 20 из распределения $\mathcal{N}(\mu, \sigma^2)$. Наблюдаемые значения выборочного среднего и выборочной дисперсии равны $\bar{x} = 81.2$ и $s^2 = 26.5$. Найдите соответственно 90%, 95% и 99% доверительные интервалы для среднего генеральной совокупности μ . Отметьте и прокомментируйте, как увеличивается ширина доверительных интервалов при увеличении уровня доверия.

Объяснение:

Исследуемая случайная величина X имеет нормальное распределение $\mathcal{N}(\mu, \sigma^2)$, где $\mu \equiv E[X]$ — неизвестное математическое ожидание. Истинная дисперсия σ^2 неизвестна. Так как известно, что X имеет нормальное распределение, то для случайной выборки размера $n = 20$ выполняется:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

$n = 20$, $\bar{x} = 81.2$, $s = \sqrt{26.5} \approx 5.148$, $t_{19,0.05} = 1.729$, $t_{19,0.025} = 2.093$, $t_{19,0.005} = 2.861$:

90% интервал:

$$\mu \in \left(81.2 - 1.729 \cdot \frac{5.148}{\sqrt{20}}, 81.2 + 1.729 \cdot \frac{5.148}{\sqrt{20}} \right) = (79.21, 83.19)$$

95% интервал:

$$\mu \in \left(81.2 - 2.093 \cdot \frac{5.148}{\sqrt{20}}, 81.2 + 2.093 \cdot \frac{5.148}{\sqrt{20}} \right) = (78.79, 83.61)$$

99% интервал:

$$\mu \in \left(81.2 - 2.861 \cdot \frac{5.148}{\sqrt{20}}, 81.2 + 2.861 \cdot \frac{5.148}{\sqrt{20}} \right) = (77.90, 84.50)$$

При увеличении уровня доверия критическая точка t -распределения сдвигается вправо, увеличивается в своем значении, что приводит к расширению интервала.

8. Есть опасения по поводу скорости автомобилей, проезжающих по определённой участку шоссе. Для случайной выборки из 7 автомобилей радар зафиксировал следующие скорости (в милях в час):

79 73 68 77 86 71 69

- (a) Вычислите реализации выборочного среднего и выборочной дисперсии. Вы здесь можете использовать любой инструмент для вычисления выборочной дисперсии. Если используете `numpy`, можно взять метод `numpy.std(sample, ddof=1)`, где параметр `ddof = 1` обеспечивает корректное деление на $(n - 1)$ в подсчете дисперсии и стандартного отклонения.
- (b) Указав все необходимые предположения, постройте 90% доверительный интервал для средней скорости всех автомобилей, проезжающих по этому участку шоссе.

Напоминание: реализация выборочной дисперсии считается по формуле: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Объяснение:

Исследуемая случайная величина X — скорость автомобиля, проезжающего по данному участку шоссе, $E[X] = \mu$. Предполагаем, что X имеет нормальное распределение, истинная дисперсия σ^2 неизвестна. Так как известно, что X имеет нормальное распределение, то для случайной выборки размера $n = 7$ выполняется:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

- (a) $\bar{x} = \frac{79+73+68+77+86+71+69}{7} \approx 74.71$
 $s^2 = \frac{1}{6} [(79 - 74.71)^2 + \dots + (69 - 74.71)^2] \approx 40.9$
 $s \approx 6.4$

- (b) Предположения: скорости распределены нормально, выборка случайная и независимая.
 $n = 7$, $t_{6,0.05} = 1.943$:

$$\mu \in \left(74.71 - 1.943 \cdot \frac{6.4}{\sqrt{7}}, 74.71 + 1.943 \cdot \frac{6.4}{\sqrt{7}} \right) = (70.012, 79.407)$$

9. Исследовательская группа анализирует эффективность новой методики обучения на успеваемость студентов. Они собрали результаты тестов у небольшой группы студентов, которые использовали новую методику. Данные следующие:

баллы: [78, 82, 85, 79, 83, 81, 84, 80, 82, 83]

Из предыдущих крупномасштабных исследований известно, что результаты теста имеют нормальное распределение, и истинное стандартное отклонение результатов теста равно $\sigma = 2.5$ балла.

Команда хочет оценить истинное среднее значение баллов всех студентов, которые будут использовать эту методику, с 95%-й уверенностью.

- Вычислите реализации выборочного среднего и выборочной дисперсии. Вы здесь можете использовать любой инструмент для вычисления выборочной дисперсии. Если используете `numpy`, можно взять метод `numpy.std(sample, ddof=1)`, где параметр `ddof = 1` обеспечивает корректное деление на $(n - 1)$ в подсчете дисперсии и стандартного отклонения.
- Постройте доверительный интервал с учетом известной истинной дисперсии случайной величины.
- Постройте доверительный интервал, основываясь только на предоставленных данных, *т.е.* если истинная дисперсия случайной величины предполагается неизвестной.
- Сравните длины обоих интервалов и объясните, что влияет на их различие.
- Если бы размер выборки был увеличен до 100 студентов (предполагая то же выборочное среднее, выборочное стандартное отклонение и уровень доверия), как бы изменилась разница между двумя интервалами? Объясните и посчитайте два новых интервала.

Объяснение:

Пусть X — тестовый балл студента после применения методики, $E[X] = \mu$. По условию X нормально распределена.

- По данным выборки размера $n = 10$:

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 81.7, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{10} (x_i - \bar{x})^2 = 4.9, \quad s = \sqrt{4.9} \approx 2.214.$$

- При известной истинной дисперсии ($\sigma = 2.5$) используем z -интервал:

$$\mu \in \left(\bar{x} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} \right), \quad z_{0.025} = 1.96.$$

$$\mu \in \left(81.7 \pm 1.96 \cdot \frac{2.5}{\sqrt{10}} \right) \approx (80.15, 83.25).$$

- При неизвестной дисперсии используем t -интервал:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} = t_9, \quad \mu \in \left(\bar{x} \pm t_{9,0.025} \frac{s}{\sqrt{n}} \right),$$

где $t_{9,0.025} \approx 2.262$.

$$\mu \in \left(81.7 \pm 2.262 \cdot \frac{2.214}{\sqrt{10}} \right) \approx (80.12, 83.28).$$

- Длина интервала:

$$L_z = 2z_{0.025} \frac{\sigma}{\sqrt{n}}, \quad L_t = 2t_{9,0.025} \frac{s}{\sqrt{n}}.$$

В этой задаче $L_t > L_z$ (примерно 3.17 против 3.10), так как при малом n критическая точка t больше, чем z , и даже несмотря на то, что реализация выборочной дисперсии оказалась меньше, чем истинная дисперсия, отличие в критической точке перевешивает.

- Если формально увеличить размер до $n = 100$ (с теми же $\bar{x}$ и s):

$$\mu_z \in \left(81.7 \pm 1.96 \cdot \frac{2.5}{10} \right) = (81.21, 82.19),$$

$$\mu_t \in \left(81.7 \pm t_{99,0.025} \cdot \frac{2.214}{10} \right), \quad t_{99,0.025} \approx 1.984,$$

$$\mu_t \approx (81.26, 82.14).$$

В этом случае так как реализация выборочной дисперсии меньше, чем истинная дисперсия, а критическая точка $t_{n-1,0.025} \rightarrow z_{0.025}$ при росте n , то интервал, построенный по t -распределению, оказывается даже уже, чем интервал, построенный по нормальному распределению.

10. Рассмотрим оригинальное клиническое исследование вакцины Солка от полиомиелита, проведённое в 1954 году. Случайным образом одна группа детей получила вакцину (группа лечения), а другая группа получила плацебо (контрольная группа). Пусть p_c и p_T обозначают истинные доли заболевших полиомиелитом в контрольной группе и группе лечения соответственно. Результаты исследования представлены в таблице:

Группа	Количество детей	Количество случаев полиомиелита
Лечение	200, 745	57
Контроль	201, 229	199

Постройте 95% доверительный интервал для разности $(p_c - p_T)$. (Не округляйте слишком сильно, оставьте хотя бы 4 – 5 знаков в дробной части). Что можно сказать об эффективности вакцины на основе построенного доверительного интервала?

Объяснение:

Исследуемые случайные величины:

X_c — случайная величина Бернулли для контрольной группы (заболел или нет), $E[X_c] = p_c$;

X_T — случайная величина Бернулли для группы лечения, $E[X_T] = p_T$.

Нас интересует разность $\theta = p_c - p_T$, хотим оценить её с помощью доверительного интервала.

$$\tilde{p}_c = \frac{199}{201229} \approx 0.0009889, \tilde{p}_T = \frac{57}{200745} \approx 0.0002839, z_{0.025} = 1.96:$$

$$p_c - p_T \in \left(\tilde{p}_c - \tilde{p}_T - z_{0.025} \sqrt{\frac{\tilde{p}_c(1 - \tilde{p}_c)}{n_c} + \frac{\tilde{p}_T(1 - \tilde{p}_T)}{n_T}}, \tilde{p}_c - \tilde{p}_T + z_{0.025} \sqrt{\frac{\tilde{p}_c(1 - \tilde{p}_c)}{n_c} + \frac{\tilde{p}_T(1 - \tilde{p}_T)}{n_T}} \right)$$

$$p_c - p_T \in (0.0007050 - 0.0001558, 0.0007050 + 0.0001558) = (0.0005491, 0.0008608)$$

Интервал не содержит нуля и полностью находится в положительной области, что означает статистически значимое снижение заболеваемости в группе лечения. Вакцина эффективна.

11. Сравнивают две рекламные кампании. В первой опросили $n = 30$ клиентов, из них 13 совершили покупку. Во второй — $m = 1.5n = 45$ клиентов, из них 15 совершили покупку. Построенный доверительный интервал для разности долей конверсии $p_1 - p_2$ равен $(-0.07; 0.27)$.

Найдите уровень доверия этого интервала.

Объяснение:

Для двух долей доверительный интервал обычно имеет вид

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{m}}.$$

Здесь

$$\tilde{p}_1 = \frac{13}{30}, \quad \tilde{p}_2 = \frac{15}{45}, \quad \tilde{p}_1 - \tilde{p}_2 = 0.1.$$

Данный интервал $(-0.07; 0.27)$ симметричен относительно 0.1, значит его полуширина

$$E = \frac{0.27 - (-0.07)}{2} = 0.17.$$

Стандартная ошибка:

$$SE = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{30} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{45}} \approx 0.1146.$$

Тогда

$$z_{\alpha/2} = \frac{E}{SE} \approx \frac{0.17}{0.1146} \approx 1.484.$$

Уровень доверия:

$$1 - \alpha = P(|Z| \leq z_{\alpha/2}) = 2\Phi(1.484) - 1 \approx 0.862.$$

Итого: уровень доверия примерно 86.2% (около 86%).

12. Производственная компания хочет сравнить среднюю производительность двух производственных линий. Известно, что стандартное отклонение производительности для первой линии составляет $\sigma_X = 12$ единиц продукции в час, а для второй линии — $\sigma_Y = 15$ единиц продукции в час.

Было проведено тестирование:

- Первая линия: выборка из $n = 40$ часов работы, средняя производительность $\bar{x} = 145$ единиц/час
- Вторая линия: выборка из $m = 35$ часов работы, средняя производительность $\bar{y} = 138$ единиц/час

- (а) Постройте 95% доверительный интервал для разности средних производительностей двух линий.
 (б) Можете ли вы сделать вывод о том, какая линия более производительна?

Объяснение:

Пусть X — производительность первой линии, Y — производительность второй линии. Интересующий параметр:

$$\theta = \mu_X - \mu_Y.$$

При известных σ_X, σ_Y и независимых выборках

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\theta, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right).$$

Тогда 95% доверительный интервал:

$$\theta \in \left((\bar{x} - \bar{y}) \pm z_{0.025} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right), \quad z_{0.025} = 1.96.$$

Подстановка данных:

$$\begin{aligned} \bar{x} - \bar{y} &= 145 - 138 = 7, & SE &= \sqrt{\frac{12^2}{40} + \frac{15^2}{35}} \approx 3.167. \\ \theta &\in (7 \pm 1.96 \cdot 3.167) \approx (0.79, 13.21). \end{aligned}$$

Интервал целиком выше нуля, значит на уровне доверия 95% получаем $\mu_X > \mu_Y$: первая линия в среднем производительнее.

13. Пусть $X_1, \dots, X_8$ и $Y_1, \dots, Y_{10}$ — независимые случайные выборки из распределений $\mathcal{N}(\mu_X, \sigma^2)$ и $\mathcal{N}(\mu_Y, 2\sigma^2)$ соответственно, где истинный параметр σ^2 известен и равен 4, а μ_X и μ_Y неизвестны.

Известны реализации выборочных средних: $\bar{x} = 5$, $\bar{y} = 3$.

- (а) Вычислите 90% доверительные интервалы для μ_X и μ_Y .
 (б) Получите теоретическую формулу для доверительного интервала для разницы $\mu_X - \mu_Y$, а потом на имеющихся данных вычислите реализацию 90% доверительного интервала для этой разницы.
 (с) Вопрос на подумать: почему интервал для разницы будет отличаться от того, что если просто вычесть граничные значения отдельных интервалов для μ_X и μ_Y ?

Объяснение:

Так как дисперсия известна, используем нормальные интервалы. Для уровня доверия 90%:

$$\alpha = 0.1, \quad z_{\alpha/2} = z_{0.05} = 1.645.$$

- (а) Для выборки $X_1, \dots, X_8 \sim \mathcal{N}(\mu_X, \sigma^2)$, где $\sigma^2 = 4$:

$$\bar{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma^2}{8}\right), \quad \mu_X \in \left(\bar{x} \pm z_{0.05} \sqrt{\frac{\sigma^2}{8}}\right).$$

Подстановка $\bar{x} = 5$:

$$\mu_X \in \left(5 \pm 1.645 \sqrt{\frac{1}{2}}\right) \approx (3.84, 6.16).$$

Для выборки $Y_1, \dots, Y_{10} \sim \mathcal{N}(\mu_Y, 2\sigma^2)$:

$$\bar{Y} \sim \mathcal{N}\left(\mu_Y, \frac{2\sigma^2}{10}\right), \quad \mu_Y \in \left(\bar{y} \pm z_{0.05} \sqrt{\frac{2\sigma^2}{10}}\right).$$

Подстановка $\bar{y} = 3$:

$$\mu_Y \in \left(3 \pm 1.645 \sqrt{\frac{4}{5}}\right) \approx (1.53, 4.47).$$

(b) Обозначим $\theta = \mu_X - \mu_Y$. Так как выборки независимы,

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{8} + \frac{2\sigma^2}{10}\right).$$

Ищем границы L, U такие, что

$$1 - \alpha = P(L < \theta < U) = P\left(\frac{\bar{X} - \bar{Y} - U}{\sqrt{13/10}} < \frac{\bar{X} - \bar{Y} - \theta}{\sqrt{13/10}} < \frac{\bar{X} - \bar{Y} - L}{\sqrt{13/10}}\right) = P(-z_{\alpha/2} < Z < z_{\alpha/2}),$$

где $Z \sim \mathcal{N}(0, 1)$. Отсюда

$$L = (\bar{x} - \bar{y}) - z_{\alpha/2} \sqrt{\frac{13}{10}}, \quad U = (\bar{x} - \bar{y}) + z_{\alpha/2} \sqrt{\frac{13}{10}}.$$

Для 90%: $\alpha = 0.1$, $z_{\alpha/2} = z_{0.05} = 1.645$, а $\bar{x} - \bar{y} = 2$, поэтому

$$\theta \in \left(2 \pm 1.645 \sqrt{\frac{13}{10}}\right) \approx (0.12, 3.88).$$

(c) Если просто вычесть границы интервалов для μ_X и μ_Y , получится полуширина

$$z_{0.05} \left(\sqrt{\frac{\sigma^2}{8}} + \sqrt{\frac{2\sigma^2}{10}} \right),$$

то есть сумма стандартных ошибок. Правильная же полуширина для разницы использует стандартную ошибку

$$\sqrt{\frac{\sigma^2}{8} + \frac{2\sigma^2}{10}},$$

то есть корень из суммы дисперсий. Поскольку $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$, «вычитание границ» даёт более широкий (консервативный) интервал и в общем случае не совпадает с интервалом для $\mu_X - \mu_Y$.