ВШБ Бизнес-информатика: ТВиМС 2025.
Лист задач для самостоятельного решения #11.
Точечные оценки.
Метод моментов. Метод максимального правдоподобия.
Интервальные оценки.

Основные формулы

Свойства точечных оценок

- Несмещенность: $E[\hat{\theta}] = \theta$
- Смещение: $\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$
- Среднеквадратичная ошибка: $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$

Доверительные интервалы

- Среднее μ, дисперсия известна:

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

- Среднее μ, дисперсия неизвестна:

$$\mu \in \left(\bar{x} - t_{(n-1,\alpha/2)} \frac{s}{\sqrt{n}},\ \bar{x} + t_{(n-1,\alpha/2)} \frac{s}{\sqrt{n}} \right)$$

- Доля p:

$$p \in \left(\tilde{p} - z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}},\ \tilde{p} + z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \right)$$

- Разность долей $p_1 - p_2$:

$$p_1 - p_2 \in \left(\tilde{p}_1 - \tilde{p}_2 - z_{\alpha/2} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{m}},\ \tilde{p}_1 - \tilde{p}_2 + z_{\alpha/2} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{m}} \right)$$

- Разность средних $\mu_X - \mu_Y$, дисперсии σ_X^2, σ_Y^2 известны:

$$\mu_X - \mu_Y \in \left(\bar{x} - \bar{y} - z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}},\ \bar{x} - \bar{y} + z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right)$$

- Разность средних $\mu_X - \mu_Y$, дисперсии неизвестны, но предполагаются равными:

$$\mu_X - \mu_Y \in \left(\bar{x} - \bar{y} - t_{(n+m-2,\alpha/2)} s_p \sqrt{\frac{1}{n} + \frac{1}{m}},\ \bar{x} - \bar{y} + t_{(n+m-2,\alpha/2)} s_p \sqrt{\frac{1}{n} + \frac{1}{m}} \right)$$

где $s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$ — объединённая выборочная дисперсия.

1. Независимые случайные величины X_1, X_2 и X_3 имеют одинаковое матожидание μ но разные стандартные отклонения $\sigma, 2\sigma$ и 3σ соответственно. В качестве оценки матожидания мы рассматриваем три варианта: $\hat{\theta}_1 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$, $\hat{\theta}_2 = \frac{1}{6}X_1 + \frac{1}{5}X_2 + \frac{1}{4}X_3$, $\hat{\theta}_3 = \frac{1}{6}X_1 + \frac{1}{3}X_2 + \frac{1}{2}X_3$. Какая из этих оценок лучше? Указание - проверить несмещенность, у несмещенных сравнить дисперсии.

$$E[\hat{\theta}_1] = E\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \mu$$
$$E[\hat{\theta}_2] = E\left(\frac{1}{6}X_1 + \frac{1}{5}X_2 + \frac{1}{4}X_3\right) \neq \mu$$
$$E[\hat{\theta}_3] = E\left(\frac{1}{6}X_1 + \frac{1}{3}X_2 + \frac{1}{2}X_3\right) = \mu$$

$$Var[\hat{\theta}_1] = Var\left(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3\right) = \frac{1}{9}Var(X_1) + \frac{1}{9}Var(X_2) + \frac{1}{9}Var(X_3) = \frac{1}{9}(\sigma^2 + 4\sigma^2 + 9\sigma^2) = \frac{14}{9}\sigma^2$$
$$Var[\hat{\theta}_3] = Var\left(\frac{1}{6}X_1 + \frac{1}{3}X_2 + \frac{1}{2}X_3\right) = \frac{1}{36}Var(X_1) + \frac{1}{9}Var(X_2) + \frac{1}{4}Var(X_3) = \frac{1}{36}\sigma^2 + \frac{1}{9}4\sigma^2 + \frac{1}{4}9\sigma^2 = \frac{49}{18}\sigma^2$$

Отсюда получаем, что $\hat{\theta}_1$ лучше, так как имеет меньшую дисперсию. Это означает, что $\hat{\theta}_1$ будет более стабильна и будет меньше отклоняться от истинного значения.

2. Пусть X_1, X_2 — случайная выборка из распределения с математическим ожиданием μ и дисперсией σ^2.

   Рассмотрим следующую оценку дисперсии σ^2:

   $$\hat{\theta} = c(X_1 - X_2)^2.$$

   Найти константу c такую, что $\hat{\theta}$ является несмещенной оценкой для σ^2.

   Решение:

   Оценка будет несмещенной, если $E[\hat{\theta}] = \sigma^2$. Найдем $E[\hat{\theta}]$, используя свойства математического ожидания и не забывая, что X_1 и X_2 независимы.

   $$E[\hat{\theta}] = E[c(X_1 - X_2)^2] = cE[(X_1 - X_2)^2] = cE[X_1^2 - 2X_1X_2 + X_2^2] = c(\ E[X_1^2]\ - 2E[X_1X_2] + \ E[X_2^2]\ ) =$$
   $$= c\left(\ (Var(X_1) + (E[X_1])^2)\ - 2E[X_1]E[X_2] + \ (Var(X_2) + (E[X_2])^2)\ \right) = c\left((\sigma^2 + \mu^2) - 2\mu^2 + (\sigma^2 + \mu^2)\right) = 2c\sigma^2$$

   Отсюда получаем, что $c = \frac{1}{2}$, тогда получим $E[\hat{\theta}] = \sigma^2$.

3. Случайные величины X_1 и X_2 распределены по одному закону и независимы. Среди всех несмещенных оценок матожидания вида $c_1X_1 + c_2X_2$ найти оценку с наименьшей дисперсией.

Решение:

(a) Условие несмещенности: оценка $\hat{\theta} = c_1X_1 + c_2X_2$ должна быть несмещенной, т.е. $E[\hat{\theta}] = \mu$.

$$E[c_1X_1 + c_2X_2] = c_1E[X_1] + c_2E[X_2] = c_1\mu + c_2\mu = (c_1 + c_2)\mu$$

Для несмещенности необходимо: $c_1 + c_2 = 1$.

(b) Найдем дисперсию оценки:

$$Var[c_1X_1 + c_2X_2] = c_1^2Var[X_1] + c_2^2Var[X_2] = c_1^2\sigma^2 + c_2^2\sigma^2 = \sigma^2(c_1^2 + c_2^2)$$

(c) Задача оптимизации: минимизировать $Var[\hat{\theta}] = \sigma^2(c_1^2 + c_2^2)$ при условии $c_1 + c_2 = 1$.
Подставляя $c_2 = 1 - c_1$:

$$\sigma^2(c_1^2 + c_2^2) = \sigma^2(c_1^2 + (1 - c_1)^2) = \sigma^2(c_1^2 + 1 - 2c_1 + c_1^2) = \sigma^2(2c_1^2 - 2c_1 + 1)$$

Находим минимум, приравнивая производную к нулю:

$$\frac{d}{dc_1}[\sigma^2(2c_1^2 - 2c_1 + 1)] = \sigma^2(4c_1 - 2) = 0 \Rightarrow c_1 = \frac{1}{2}$$

Отсюда $c_2 = 1 - c_1 = \frac{1}{2}$.

Ответ: $c_1 = 0.5$, $c_2 = 0.5$.

4. Количество покупок, совершаемых клиентами в интернет-магазине за день, подчиняется распределению Пуассона. У нас есть выборка данных по количеству покупок за несколько дней, результаты записаны в таблице. Необходимо определить параметр λ по этой выборке, используя метод моментов.

Количество покупок за день x_i	0	1	2	3	4	5
Количество дней с количеством покупок x_i	10	37	38	22	12	6

Решение:

Для распределения Пуассона $E[X] = \lambda$. Выборочный момент первого порядка: $m_1 = \bar{x}$.

Приравниваем: $\bar{x} = \lambda$, откуда $\hat{\lambda}_{MM} = \bar{x}$.

Выборочное среднее: $\bar{x} = \frac{0\cdot10+1\cdot37+2\cdot38+3\cdot22+4\cdot12+5\cdot6}{10+37+38+22+12+6} = \frac{257}{125} = 2.056$

Ответ: $\hat{\lambda}_{MM} = 2.056$

5. Найти методом моментов по выборке $x_1, x_2, \ldots, x_n$ точечную оценку параметра p биномиального распределения.

Решение: Задача на самом деле не очень корректно сформулирована. Оказывается, мы можем её интерпретировать по-разному. Давайте разберемся.

Вариант 1: Интерпретируем выборку как последовательность экспериментов Бернулли длины n.

Выборка $x_1, x_2, \ldots, x_n$, где каждая величина X_i принимает значение 0 или 1 с вероятностью p. То есть $X_i \sim \text{Bernoulli}(p)$ для $i = 1, \ldots, n$.

Теоретический момент первого порядка: $\mu_1 = E[X_i] = p$.

Выборочный момент первого порядка: $m_1 = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ — это доля единиц в выборке, то есть количество успехов, делённое на n.

Приравниваем: $m_1 = \mu_1$, откуда $\frac{1}{n}\sum\limits_{i=1}^{n} x_i = p$, и получаем $\hat{p}_{MM} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$.

Вариант 2: Интерпретируем выборку как количество успехов в n разных сериях испытаний.

Выборка $x_1, x_2, \ldots, x_n$, где каждая величина X_i — это количество успехов в i-й серии из k последовательных испытаний Бернулли. То есть $X_i \sim \text{Bin}(k, p)$ для $i = 1, \ldots, n$, где k — длина каждой серии испытаний. Про k не было сказано в условии, предположим, что оно известно.

Теоретический момент первого порядка: $\mu_1 = E[X_i] = kp$.

Выборочный момент первого порядка: $m_1 = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$.

Приравниваем: $m_1 = \mu_1$, откуда $\frac{1}{n}\sum\limits_{i=1}^{n} x_i = kp$, и получаем $\hat{p}_{MM} = \frac{1}{kn}\sum\limits_{i=1}^{n} x_i$.

Ответ:

- Вариант 1: $\hat{p}_{MM} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$
- Вариант 2: $\hat{p}_{MM} = \frac{1}{kn}\sum\limits_{i=1}^{n} x_i$, где k — длина серии испытаний

6. При условии равномерного распределения случайной величины X произведена выборка:

3	5	7	9	11	13	15	17	19	21	21	16	15	26	22	14	21	22	18	25

Найти оценку параметров a и b по методу моментов.

Решение:

Для равномерного распределения $X \sim \text{Uniform}(a, b)$ теоретические моменты:

$$\mu_1 = E[X] = \frac{a+b}{2}$$

$$\mu_2 = E[X^2] = \frac{a^2+ab+b^2}{3}$$

Выборочные моменты:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Приравниваем теоретические моменты к выборочным:

$$m_1 = \mu_1 : \quad m_1 = \frac{a+b}{2} \Rightarrow a + b = 2m_1$$

$$m_2 = \mu_2 : \quad m_2 = \frac{a^2 + ab + b^2}{3}$$

Из первого уравнения: $b = 2m_1 - a$. Подставляем во второе:

$$m_2 = \frac{a^2 + a(2m_1 - a) + (2m_1 - a)^2}{3} = \frac{4m_1^2 - 2am_1 + a^2}{3}$$

Откуда $3m_2 = 4m_1^2 - 2am_1 + a^2$, или $a^2 - 2am_1 + 4m_1^2 - 3m_2 = 0$.

Подставляем выборочные моменты: $m_1 = 16$, $m_2 = 296.1$.

$$a^2 - 2a \cdot 16 + 4 \cdot 16^2 - 3 \cdot 296.1 = 0$$

$$a^2 - 32a + 1024 - 888.3 = 0$$

$$a^2 - 32a + 135.7 = 0$$

Решаем квадратное уравнение:

$$a = \frac{32 \pm \sqrt{32^2 - 4 \cdot 135.7}}{2} = \frac{32 \pm \sqrt{1024 - 542.8}}{2} = \frac{32 \pm \sqrt{481.2}}{2} = \frac{32 \pm 21.94}{2}$$

Получаем два решения: $a_1 = \frac{32-21.94}{2} \approx 5.03$, $a_2 = \frac{32+21.94}{2} \approx 26.97$.

Так как $a < b$ для равномерного распределения, то $\hat{a}_{MM} = 5.03$, $\hat{b}_{MM} = 26.97$.

Ответ: $\hat{a}_{MM} \approx 5.03$, $\hat{b}_{MM} \approx 26.97$

7. Тренер и ученик стреляют в цель до первого попадания каждый. Известно, что тренер попадает в цель с вероятностью в два раза большей, чем ученик. Методом максимального правдоподобия оценить вероятность попадания учеником в цель при единичном выстреле, если известно, что тренер попал со второго раза, а ученик – с пятого. При решении задачи использовать логарифмическую функцию правдоподобия.

УКАЗАНИЕ:

$\ln ln\,(L(p)) = \ln ln\,(const) + 2\ln ln\,(p) + \ln ln\,(1-2p) + 4\ln ln\,(1-p) \Rightarrow \ln ln\,(L(p))' = \ldots = \frac{14p^2-12p+2}{2(p-0.5)(p-1)p} \approx \frac{7(p-0.631)(p-0.227)}{(p-0.5)(p-1)p}$

– сравниваем с нулем и получаем два максимума, один из них отбрасываем по смыслу переменной.

(корни числителя $\frac{3\pm\sqrt{2}}{7}$)

в качестве ответа берем $\hat{p} = 0.227$

8. В магазине работают три кассы. Время обслуживания покупателей каждым из кассиров распределено по показательному закону. При этом первый из кассиров самый опытный – среднее время обслуживания покупателя у него в два раза меньше чем у оставшихся двух. Первый кассир обслужил очередного покупателя за минуту, второй – за две минуты, третий – за полторы. Методом наибольшего правдоподобия оценить параметр λ – интенсивность для первого кассира.

В магазине работают три кассы. Время обслуживания покупателей каждым из кассиров распределено по показательному закону. При этом первый из кассиров самый опытный – среднее время обслуживания покупателя у него в два раза меньше чем у оставшихся двух. Первый кассир обслужил очередного покупателя за минуту, второй – за две минуты, третий – за полторы. Методом наибольшего правдоподобия оценить параметр λ – интенсивность для первого кассира.

УКАЗАНИЕ:

$\ln ln\,(L(\lambda, 1, 2, 1.5)) = \ln ln\left(f(\lambda,1) * f\left(\frac{\lambda}{2},2\right) * f\left(\frac{\lambda}{2},1.5\right)\right) = \ln ln\left(\lambda * e^{-\lambda*1} * \frac{\lambda}{2} * e^{-\frac{\lambda}{2}*2} * \frac{\lambda}{2} * e^{-\frac{\lambda}{2}*1.5}\right) = 3\ln ln\,(\lambda) - \frac{11}{4}\lambda + const$

Ответ: $\hat{\lambda} \approx 1.091$.

9. Андрей и Борис независимо друг от друга играют в покер в интернете. Выигрыш каждого из них за день – это случайная величина, распределенная по нормальному закону, причем известно, что у них одинаковое матожидание выигрыша m (тысяч рублей), но разные стандартные отклонения – у Андрея 1 а у Бориса 2 тысячи рублей. За последний день они выиграли по 2 и 3 тысячи рублей соответственно. Методом наибольшего правдоподобия оценить значение параметра m.

УКАЗАНИЕ:

$$\ln ln\,(L(m)) = ln\left(\frac{e^{-\frac{(2-m)^2}{2*1}}}{\sqrt{2\pi}*1}\,\frac{e^{-\frac{(3-m)^2}{2*2^2}}}{\sqrt{2\pi}*2}\right)$$

$$\ln ln\,(L(m)) = ln\left(\frac{e^{-\frac{(2-m)^2}{2*1}}}{\sqrt{2\pi}*1}\,\frac{e^{-\frac{(3-m)^2}{2*2^2}}}{\sqrt{2\pi}*2}\right) = \ln ln\,(const) - \frac{(2-m)^2}{2*1} - \frac{(3-m)^2}{2*2^2} \Rightarrow \ln ln\,(L(m))' = (2-m) + \frac{3-m}{4} = \frac{11}{4} - \frac{5}{4}m = 0 \Rightarrow m = \frac{11}{5},$$

проверяем, что это максимум.

Ответ: $\hat{m} = 2.2$

10. 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 $\sigma = 12$:

    (a) Постройте 98% доверительный интервал для среднего возраста μ всех клиентов.

    (b) Предположим, что требуется, чтобы 92% доверительный интервал был строго равен $[\bar{X} - 2, \bar{X} + 2]$. Какой размер выборки для этого потребуется?

    Решение:

    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:

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

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

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

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

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

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

11. Проведена случайная выборка из 200 студентов. 30 из них говорят, что им "очень нравится"статистика.

    (a) Вычислите долю студентов в этой выборке, которые говорят, что им "очень нравится"статистика, и затем постройте 95% доверительный интервал для истинной доли.

    (b) 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.

    Решение:

    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 $\hat{p}$.

    (a) $\tilde{p} = \frac{30}{200} = 0.15$, $z_{0.025} = 1.96$:

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

    (b) $\tilde{p} = \frac{16}{40} = 0.40$, $z_{0.025} = 1.96$:

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

    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.

12. 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.

Решение:

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.

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

$$79 \quad 73 \quad 68 \quad 77 \quad 86 \quad 71 \quad 69$$

(a) Найдите выборочное среднее и выборочную дисперсию.

(b) Указав все необходимые предположения, постройте 90% доверительный интервал для средней скорости всех автомобилей, проезжающих по этому участку шоссе.

Решение:

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

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

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

14. 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?

Решение:

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 нуль и полностью находится в положительной области, что означает статистически значимое снижение заболеваемости в группе лечения. Вакцина эффективна.

15. Пусть $\bar{x}$ и $\bar{y}$ — выборочные средние двух независимых случайных выборок $X_1, \ldots, X_8$ и $Y_1, \ldots, Y_{10}$ из распределений $\mathcal{N}(\mu_X, \sigma^2)$ и $\mathcal{N}(\mu_Y, \sigma^2)$ соответственно, где общая дисперсия неизвестна. Известны собранные данные: $\bar{x} = 5$, $\bar{y} = 3$, $\sum\limits_{i=1}^{8} x_i^2 = 215.75$, $\sum\limits_{i=1}^{10} y_i^2 = 107.64$.

(a) Вычислите 95% доверительные интервалы для μ_X и μ_Y.

(b) Вычислите 90% доверительный интервал для $\mu_X - \mu_Y$.

(c) Получите теоретическую формулу, а потом на имеющихся данных вычислите 95% доверительный интервал для $\theta = \frac{1}{3}\mu_X + \frac{2}{3}\mu_Y$.

Решение:

Исследуемые случайные величины: X имеет нормальное распределение $\mathcal{N}(\mu_X, \sigma^2)$, Y имеет нормальное распределение $\mathcal{N}(\mu_Y, \sigma^2)$, где $\mu_X \equiv E[X]$, $\mu_Y \equiv E[Y]$, а общая дисперсия σ^2 неизвестна.

Так как известно, что обе случайные величины имеют нормальное распределение с одинаковой дисперсией, то для разности выборочных средних $\bar{X} - \bar{Y}$ выполняется:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{(n+m-2)}$$

где S_p^2 — объединённая выборочная дисперсия.

Посмотрим, как поработать с формулой для выборочной дисперсии, чтобы использовать наши имеющиеся данные:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right) =$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(x_i^2\right) - 2\bar{x}\,\bar{x}n + n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(x_i^2\right) - n\bar{x}^2\right)$$

$s_X^2 = \frac{1}{7}\left(215.75 - 8 \cdot 5^2\right) = 2.25$, $s_X = 1.5$

$s_Y^2 = \frac{1}{9}\left(107.64 - 10 \cdot 3^2\right) = 1.96$, $s_Y = 1.4$

Объединённая дисперсия: $s_p^2 = \frac{7 \cdot 2.25 + 9 \cdot 1.96}{16} = 2.087$, $s_p \approx 1.445$

(a) $t_{7,0.025} = 2.365$, $t_{9,0.025} = 2.262$:

$$\mu_X \in \left(5 - 2.365 \cdot \frac{1.5}{\sqrt{8}},\ 5 + 2.365 \cdot \frac{1.5}{\sqrt{8}}\right) = (3.75, 6.25)$$

$$\mu_Y \in \left(3 - 2.262 \cdot \frac{1.4}{\sqrt{10}},\ 3 + 2.262 \cdot \frac{1.4}{\sqrt{10}}\right) = (2.00, 4.00)$$

(b) $t_{16,0.05} = 1.746$:

$$\mu_X - \mu_Y \in \left(5 - 3 - 1.746 \cdot 1.445\sqrt{\frac{1}{8} + \frac{1}{10}},\ 5 - 3 + 1.746 \cdot 1.445\sqrt{\frac{1}{8} + \frac{1}{10}}\right) = (0.80, 3.20)$$

(c) По аналогии с разностью выборочных средних, тут нам нужно поработать с точечной оценкой $\hat{\theta} = \frac{1}{3}\bar{X} + \frac{2}{3}\bar{Y}$ — несмещённая оценка для $\theta = \frac{1}{3}\mu_X + \frac{2}{3}\mu_Y$.

$$E[\hat{\theta}] = \frac{1}{3}E[\bar{X}] + \frac{2}{3}E[\bar{Y}] = \frac{1}{3}\mu_X + \frac{2}{3}\mu_Y = \theta$$
$$Var[\hat{\theta}] = \frac{1}{9}Var[\bar{X}] + \frac{4}{9}Var[\bar{Y}] = \frac{1}{9}\frac{\sigma^2}{n} + \frac{4}{9}\frac{\sigma^2}{m} = \sigma^2\left(\frac{1}{9n} + \frac{4}{9m}\right)$$

При неизвестной дисперсии σ^2 используем t-распределение с $n + m - 2$ степенями свободы. Стандартизированная случайная величина:

$$\frac{\hat{\theta} - \theta}{S_p\sqrt{\frac{1}{9n} + \frac{4}{9m}}} \sim t_{(n+m-2)}$$

Построение доверительного интервала:

$$1-\alpha = P(L<\theta<U) = P(-U<-\theta<-L) =$$

$$P\left(\frac{\hat\theta-U}{S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}} < \frac{\hat\theta-\theta}{S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}} < \frac{\hat\theta-L}{S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}}\right) = P\left(-t_{(n+m-2,\alpha/2)} < t_{(n+m-2)} < t_{(n+m-2,\alpha/2)}\right)$$

Выполняем обратное преобразование и находим теоретические границы:

$$t_{(n+m-2,\alpha/2)} = \frac{\hat\theta-L}{S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}} \Rightarrow L = \hat\theta - t_{(n+m-2,\alpha/2)}S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}$$

$$-t_{(n+m-2,\alpha/2)} = \frac{\hat\theta-U}{S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}} \Rightarrow U = \hat\theta + t_{(n+m-2,\alpha/2)}S_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}$$

Получаем, что реализация $(1-\alpha)100\%$ доверительного интервала для θ:

$$\theta \in \left(\frac{1}{3}\bar x + \frac{2}{3}\bar y - t_{(n+m-2,\alpha/2)}s_p\sqrt{\frac{1}{9n}+\frac{4}{9m}},\ \frac{1}{3}\bar x + \frac{2}{3}\bar y + t_{(n+m-2,\alpha/2)}s_p\sqrt{\frac{1}{9n}+\frac{4}{9m}}\right)$$

$t_{16,0.025} = 2.120$:

$$\theta \in \left(\frac{11}{3} - 2.120\cdot 1.445\sqrt{\frac{1}{72}+\frac{4}{90}},\ \frac{11}{3} + 2.120\cdot 1.445\sqrt{\frac{1}{72}+\frac{4}{90}}\right) = (2.927, 4.406)$$