Модульная домашняя работа №1

Мягкий дедлайн: 23:59 02/03/2020.

Жесткий дедлайн: 23:59 11/03/2020.

Форма, куда отправлять работу: https://forms.gle/tynNyExEZiEsqoLD6

Данные: https://drive.google.com/drive/folders/1RpHR-YH7CnxxbNDPLFqnnk6AthFcnSOo

Общие требования

- Для данного домашнего задания можно использовать языки программирования R, Python.
- Задания необходимо выполнять в RMarkdown либо jupyter notebook с комментариями и пояснениями, иначе работа проверяться не будет.
- Если работа будет прислана до мягкого дедлайна, то она будет проверена и вы будете иметь возможность исправить ее. Иначе ваша работа будет проверяться по мере возможности.
- После жесткого дедлайна никакие исправления нельзя будет вносить.
- Если у вас возникнут вопросы, то можно их задавать в телеграм-чате или лично в день занятий.

Требования к заданиям

1. В первом задании требуется сделать базовый EDA (exploratory data analysis). Найдите зависимости в показателях, посмотрите на распределения этих показателей и сделайте соответствующие выводы.

   Рекомендации подскажут вам с чего можно начать.

   - Постройте графики распределений показателей.
   - Посчитайте статистики, которые вы знаете.
   - Поищите зависимости между данными и подумайте о способах, которыми можно поискать эти зависимости.

2. Для третьего задания (про доверительные интервалы) в каждом варианте необходимо реализовать функцию, выполняющую построение доверительного интервала. Использовать встроенные функции (например, confint в R) нельзя. Также необходимо в задании про доверительный интервал продемонстрировать, что он соответствует заявленной вероятности (даже если этого не сказано явно в задании, это общее требование).

3. Если в задании сказано «исследуйте поведение такого-то показателя», это означает, что нужно обратить внимание как на сходимость, так и на её скорость.

Описание данных

Если у вас возникнут вопросы по данным (какие переменные и что значат), можно обращаться к Лене.

1. В var1 представлены данные по рождению детей: возраст родителей, уровень образования матери, пол ребенка, его вес на момент рождения, недоношенность плода.

2. В var2 представлена статистика по террористической угрозе в аэропортах за разные годы: год, количество различных видов оружия, обнаруженных сотрудниками, количество взрывов и ложных тревог.

3. В var3 представлена статистика по книгам: различные физические параметры, цена, год публикации, автор и издание.

4. В var4 представлены данные турнира по стрельбе из лука среди женщин: очки в разных частях турнира, размер наконечника стрел, страна.

5. В var5 представлены данные по броскам мяча в бейсболе.

6. var6-var7: представлены различные физические измерения среди мужчин.

9. var8-var9: представлены различные показатели по алкоголикам (непонятные буковки в названии колонки — это какой-то медицинский показатель, не стоит пытаться как-то понять его).

11. В var10 представлены замеры различных показателей у людей с депрессией (США).

12. В var11 представлены данные турнира по стрельбе из лука среди женщин: очки в разных частях турнира, размер наконечника стрел, страна.

Вспомогательные определения

Смещение

Пусть θ — параметр распределения, а $\hat{\theta}_n = \bar{\theta}_n(x_1, \ldots, x_n)$ — оценка этого параметра, полученная по выборке. Тогда

- Если выполняется $\mathbb{E}\hat{\theta}_n = \theta$, то оценка $\hat{\theta}_n$ является *несмещенной (unbiased)*.
- Если выполняется $\mathbb{E}\hat{\theta}_n \xrightarrow[n \to \infty]{} \theta$, то оценка $\hat{\theta}_n$ является *асимптотически несмещенной*.
- *Смещением (bias)* называется величина $\mathbb{E}\hat{\theta}_n - \theta$.

Вариант 1

1. Данные var1.tsv.

2. Существуют две оценки дисперсии: выборочная $\bar{s}_n^2$ и исправленная выборочная $\tilde{s}_n^2$.

   - Продемонстрируйте с помощью моделирования, что у $\tilde{s}_n^2$ отсутствует смещение, а $\bar{s}_n^2$ смещена.
   - Покажите поведение смещения при $n \to \infty$.

3. Пусть $x_1, \ldots, x_n \in \text{Bern}(p)$. Исследуйте зависимость ширины доверительного интервала для параметра распределения p от объема выборки.

4. Приведите пример распределения, для которого центральная предельная теорема не выполняется. Продемонстрируйте это.

Вариант 2

1. Данные var2.tsv.
2. Пусть $x_1, \ldots, x_n$ — равномерно распределены на $[0; \theta]$ и $\hat{\theta}_n = 2\bar{x}$.
   - Продемонстрируйте, что оценка является несмещенной.
   - Исследуйте дисперсию $\hat{\theta}_n$. Выполняется ли неравенство Рао-Крамера?
3. Известно, что дисперсия геометрического распределения равна $(1-p)/p^2$, а математическое ожидание равно $1/p$. Для построения доверительного интервала можно использовать следующее неравенство:

   $$z_1 < \sqrt{n}\frac{\bar{x} - 1/p}{\sqrt{(1-p)/p^2}} < z_2,$$

   где z_1 и z_2 — соответствующие квантили. Выпишите явное решение неравенства относительно параметра p и с его помощью постройте доверительный интервал для параметра p.
4. Продемонстрируйте, что значение выборочной функции распределения $\bar{F}_n(x)$ в точке x является несмещенной и состоятельной оценкой для теоретической функции распределения $F(x)$ в той же точке. Исследуйте асимптотическое поведение дисперсии этой оценки.

Вариант 3

1. Данные var3.tsv.
2. Для выборочной медианы верно следующее утверждение: если элементы выборки имеют плотность $p(x)$, причем $p(x_{1/2}) > 0$, то

   $$\sqrt{n}(\bar{med} - x_{1/2}) \xrightarrow[n \to \infty]{} N\left(0, \frac{1}{4p^2(x_{1/2})}\right),$$

   где $x_{1/2}$ — истинная (теоретическая) медиана.
   - Продемонстрируйте состоятельность выборочной медианы как оценки медианы.
   - Исследуйте и продемонстрируйте поведение дисперсии и смещения выборочной медианы в условиях утверждения.
3. Постройте доверительные интервалы для дисперсии в нормальной модели с неизвестным математическим ожиданием для разных объемов выборки. Продемонстрируйте поведение доверительных интервалов как для малых объемов выборки, так и для больших.
4. Рассмотрим выборку из двумерного нормального распределения $(x_i, y_i)^{\mathrm{T}}$, где

   $$x_i \sim \mathrm{N}(a_1, \sigma_1^2), \quad y_i \sim \mathrm{N}(a_2, \sigma_2^2).$$

   Обозначим ρ — корреляцию между x_i и y_i.

   При помощи моделирования покажите, что оценка $\bar{x}/\bar{y}$ не является состоятельной. Что будет происходить в случае $\rho = 0$ и $\rho = 1$?

Вариант 4

1. Данные var4.tsv.
2. Промоделируйте выборку из $\mathrm{N}(a, \sigma^2)$ (параметры a и σ^2 могут быть любыми) объема $N = 150$. Посмотрите на выборочную медиану и выборочное среднее. Затем промоделируйте выборку из $\mathrm{N}(a + 5\sigma, \sigma^2)$ объема $N = 5$ и добавьте эти наблюдения в исходную выборку.
   - Повторите моделирование (150 наблюдений из одного распределения и 5 из другого) несколько раз и исследуйте смещение среднего и медианы относительно a.
   - Для любого ли распределения можно использовать выборочную медиану для оценки среднего?
3. Пусть $x_1, \ldots, x_n$ — выборка из $\mathrm{Bin}(m, p)$. Исследуйте зависимость ширины доверительного интервала для параметра распределения p от объема выборки.
4. Продемонстрируйте, что скорость сходимости к нормальному распределению в рамках центральной предельной теоремы может быть различной для различных распределений.

Вариант 5

1. Данные var5.tsv.
2. Пусть $x_1, \ldots, x_n$ — равномерно распределены на $[0; \theta]$, $\theta > 0$. Известно, что ОМП для параметра θ равна
$$\hat{\theta}_n = \max x_i.$$
   - Продемонстрируйте скорость сходимости оценки $\hat{\theta}_n$ к истинному значению параметра θ.
   - Выполняется ли неравенство Рао-Крамера? Покажите это с помощью моделирования.
3. Пусть $x_1, \ldots, x_n \in \Pi(\lambda)$. Исследуйте зависимость ширины доверительного интервала для параметра распределения λ от объема выборки.
4. Продемонстрируйте выполнение следующего утверждения о скорости сходимости в теореме Пуассона (Ю.В. Прохоров):
$$\sum_{k=0}^{\infty} |\mathbf{P}(\xi_n = k) - \pi_k| \leq \frac{2\lambda}{n} \min(2, \lambda),$$
   где $\pi_k = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \ldots$.

Вариант 6

1. Данные var6.tsv.
2. Промоделируйте выборку из дискретного распределения, которое принимает значение в 1 и -1 с одинаковой вероятностью, равной 0.5.
   - Выясните, состоятельна ли выборочная медиана как оценка настоящей медианы, и продемонстрируйте это.

- Является ли она смещенной? Продемонстрируйте это

3. Постройте два доверительных интервала для среднего (не для нормальной модели): один с известной дисперсией, другой с оцененной по выборке. Исследуйте поведение ширины доверительных интервалов при n, стремящимся к бесконечности. Покажите, что среднее попадает в интервал с нужной вероятностью.

4. Пусть $\xi_i \in \mathrm{N}(0,1)$. Тогда $\eta_n = \sum_{i=1}^{n} \xi_i^2 \sim \chi^2(n)$. Покажите, что выполняется ЦПТ для η_n.

Вариант 7

1. Данные var7.tsv.

2. Промоделируйте выборку из $\mathrm{N}(a, \sigma^2)$ (параметры a и σ^2 могут быть любыми) объема $N = 150$. Посмотрите на выборочную медиану и выборочное среднее. Затем промоделируйте выборку из $\mathrm{N}(a + 5\sigma, \sigma^2)$ объема $N = 5$ и добавьте эти наблюдения в исходную выборку.

   - Повторите моделирование (150 наблюдений из одного распределения и 5 из другого) несколько раз и исследуйте смещение среднего и медианы относительно a.
   - Исследуйте поведение дисперсии выборочного среднего и выборочной медианы и сравните их поведение.

3. Известно, дисперсия распределения Бернулли равна $p(1-p)$. Для построения доверительного интервала можно использовать следующее неравенство:

$$z_1 < \sqrt{n}\frac{\bar{x} - p}{\sqrt{p(1-p)}} < z_2,$$

   где z_1 и z_2 — соответствующие квантили. Выпишите явное решение неравенства относительно параметра p и с его помощью постройте доверительный интервал для параметра p.

4. Рассмотрим последовательность случайных величин $\xi_n \in \mathrm{Bin}(n, p_n)$. Пусть $np_n \to \lambda > 0$ при $n \to \infty$. Продемонстрируйте выполнение теоремы Пуассона:

$$\mathbf{P}(\xi_n = k) \xrightarrow[n \to \infty]{} \pi_k = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \ldots$$

   (Таким образом, случайная величина ξ_n имеет асимптотически распределение Пуассона с параметром λ).

Вариант 8

1. Данные var8.tsv.

2. Пусть $x_1, \ldots, x_n \sim \mathrm{N}(a, \sigma^2)$.

   - Продемонстрируйте состоятельность выборочной дисперсии.
   - Продемонстрируйте, что выборочная дисперсия нормального распределения имеет распределение хи-квадрат: $n\frac{s_n^2}{\sigma^2} \sim \chi^2(n-1)$.

3. Случайная величина x распределена равномерно на отрезке $(0, \theta)$. Верно, что интервал

$$\left(\max_i x_i, \frac{\max_i x_i}{(1-\gamma)^{1/n}} \right)$$

является доверительным для уровня γ в такой модели. Постройте этот доверительный интервал, продемонстрируйте, что он действительно доверительный с указанной вероятностью.

4. Пусть $x_1, \ldots, x_n$ — независимые одинаково распределенные случайные величины с функцией распределения $F(x)$. Обозначим $F_n(x)$ эмпирическую функцию распределения для $x_1, \ldots, x_n$:

$$F_n(x) = \frac{\#\{i : x_i < x\}}{n}.$$

Продемонстрируйте выполнение теоремы Гливенко-Кантелли:

$$\sup_{x \in \mathbb{R}} |F(x) - F_n(x)| \xrightarrow[n \to \infty]{} 0.$$

Указание: в силу «ступенчатости» функции $F_n(x)$ супремум достигается на конечном множестве точек.

Вариант 9

1. Данные var9.tsv.

2. Пусть $x_1, \ldots, x_n$ — равномерно распределены на $[-\theta; 0]$, $\theta > 0$. Известно, что ОМП для параметра θ равна

$$\hat{\theta}_n = -\min_i x_i.$$

   - Продемонстрируйте скорость сходимости оценки $\hat{\theta}_n$ к истинному значению параметра θ.
   - Выполняется ли неравенство Рао-Крамера? Покажите это с помощью моделирования.

3. Постройте доверительные интервалы для математического ожидания в нормальной модели с неизвестной дисперсией для разных объемов выборки. Продемонстрируйте, почему при малых объемах такой доверительный интервал не имеет смысла.

4. Пусть $x_1, \ldots, x_n$ — независимые одинаково распределенные случайные величины с функцией распределения $F(x)$. Обозначим $F_n(x)$ эмпирическую функцию распределения для $x_1, \ldots, x_n$:

$$F_n(x) = \frac{\#\{i : x_i < x\}}{n}.$$

Введем величину D_n (так называемую статистику критерия Колмогорова-Смирнова):

$$D_n = \sqrt{n} \sup_x |F(x) - \bar{F}_n(x)|.$$

Величина D_n имеет предельное распределение при $n \to \infty$.

Вариант 10

1. Данные var10.tsv.
2. Существуют две оценки дисперсии: выборочная $\bar{s}_n^2$ и исправленная выборочная $\tilde{s}_n^2$.
   - Продемонстрируйте с помощью моделирования, что у $\tilde{s}_n^2$ отсутствует смещение, а $\bar{s}_n^2$ смещена.
   - Исследуйте поведение дисперсий $\bar{s}_n^2$ и $\tilde{s}_n^2$.
3. Для выборочной медианы верно следующее утверждение: если элементы выборки имеют плотность $p(x)$, причем $p(x_{1/2}) > 0$, то

$$\sqrt{n}(\bar{med} - x_{1/2}) \xrightarrow[n \to \infty]{} N\left(0, \frac{1}{4p^2(x_{1/2})}\right).$$

   Постройте доверительный интервал для выборочной медианы, исследуйте зависимость его ширины от объема выборки. Здесь $x_{1/2}$ — истинная (теоретическая) медиана.
4. Пусть $(\xi_1, \xi_2)^{\mathrm{T}} \sim \mathrm{N}(0, \Sigma)$, где $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

   Найдите по методу моментов оценку для параметра ρ (теоретически) и покажите с помощью моделирования, что она не совпадает с оценкой, полученной с помощью метода максимального правдоподобия.

   Указание: для ОМП можно использовать специальные функции в R/Python. Например, в R это функция mle2 из пакета bbmle.

Вариант 11

1. Данные var11.tsv.
2. Пусть $x_1, \ldots, x_n$ — равномерно распределены на $[0; \theta]$ и $\hat{\theta}_n = 2\bar{x}$.
   - Продемонстрируйте, что оценка является несмещенной.
   - Исследуйте дисперсию $\hat{\theta}_n$. Выполняется ли неравенство Рао-Крамера?
3. Известно, что для геометрического распределения дисперсия равна $(1-p)/p^2$, а математическое ожидание равно $1/p$. Для построения доверительного интервала можно использовать следующее неравенство:

$$z_1 < \sqrt{n}\frac{\bar{x} - 1/p}{\sqrt{(1-p)/p^2}} < z_2,$$

   где z_1 и z_2 — соответствующие квантили. Выпишите явное решение неравенства относительно параметра p и с его помощью постройте доверительный интервал для параметра p.
4. Продемонстрируйте, что скорость сходимости к нормальному распределению в рамках центральной предельной теоремы может быть различной для различных распределений.