Домашнее задание 1

Дедлайн: 2025-02-04, 23:59.

1. Случайные величины y_i независимы и одинаково распределены с $\mathbb{P}(y_i = 0) = a$, $\mathbb{P}(y_i = 1) = 2a$, $\mathbb{P}(y_i = 2) = 1 - 3a$. В выборке y_1, y_2, ..., y_n оказалось N_0 нулей, N_1 единиц и N_2 двоек.

   а) Найдите оценку $\hat{a}$ параметра a методом моментов используя $\mathbb{E}(y_i)$.

   б) Найдите оценку $\hat{a}$ параметра a методом моментов используя $\mathbb{E}(y_i^2)$.

   в) Найдите оценку $\hat{a}$ параметра a методом максимального правдоподобия.

2. Случайные величины y_i независимы и нормально распределены $\mathcal{N}(2a; a)$ с неизвестным параметром a.

   а) Найдите оценку $\hat{a}$ параметра a методом моментов используя $\mathbb{E}(y_i)$.

   б) Найдите оценку $\hat{a}$ параметра a методом моментов используя $\mathbb{E}(y_i^2)$.

   в) Найдите оценку $\hat{a}$ параметра a методом максимального правдоподобия.

3. В отделении банка 5 клиентских окошек. Время обслуживания каждого клиента имеет экспоненциальное распределение с неизвестной интенсивностью λ. Я был в очереди последним, и когда я встал к освободившемуся окошку номер 5, все остальные окошки ещё обслуживали клиентов. Через 3 минуты обслужили клиента в окошке 3, через 7 минут — клиента в окошке номер 4, а потом я освободился и ушёл.

   а) Найдите оценку $\hat{a}$ параметра a методом моментов, используя любое математическое ожидание.

   б) Найдите оценку $\hat{a}$ параметра a методом максимального правдоподобия.

   Примечание: если в данной задаче возникает нерешаемое в явном виде уравнение, то, конечно, можно и нужно воспользоваться подходящим численным методом.

Домашнее задание 2

Дедлайн: 2025-02-23, 23:59.
Оцениваемые задачи:

1. Величины y_1, y_2, ..., y_n независимы и равномерны отрезке на $[0; a]$ с неизвестным $a > 5$. Никола Тесла хочет оценить неизвестный параметр $b = \mathbb{P}(y_i > 5)$.

   Рассмотрим две оценки: $\hat{b}_n$ — доля наблюдений в выборке, оказавшихся больше 5 и $\hat{b}'_n = 1 - 2.5/\bar{y}$.

   а) Является ли оценка $\hat{b}_n$ несмещённой? состоятельной?

   б) Является ли оценка $\hat{b}'_n$ несмещённой? состоятельной?

2. Величины y_i независимы и имеют функцию плотности

$$f(y) = \begin{cases} 3y^2/\theta^3, \text{ если } y \in [0; \theta]; \\ 0, \text{ иначе.} \end{cases}$$

а) Найдите оценку $\hat{\theta}$ неизвестного параметра θ методом максимального правдоподобия.

б) Является ли оценка $\hat{\theta}$ несмещённой?

в) Является ли оценка $\hat{\theta}$ состоятельной?

г) Найдите функцию плотности оценки $\hat{\theta}$.

д) На какую величину нужно домножить оценку $\hat{\theta}$, чтобы она стала несмещённой?

Подсказка: ответ на пункт (б) можно получить без вычислений и интегралов :)

Неоцениваемые задачи в удовольствие:

3. Величина Y имеет биномиальное распределение $\text{Bin}(n, p)$.

   а) Является ли оценка $\hat{p} = Y/n$ для p несмещённой? Если является смещённой, то скорректируйте оценку так, чтобы она стала несмещённой.

   б) Чему равна теоретическая дисперсия σ^2 величины Y?

   в) Является ли оценка $\hat{\sigma}^2 = n\hat{p}(1 - \hat{p})$ для σ^2 несмещённой? Если является смещённой, то скорректируйте оценку так, чтобы она стала несмещённой.

4. Величины X_i независимы и одинаково распределены с неизвестными $\mathbb{E}(X_i) = \mu$ и $\mathbb{V}\text{ar}(X_i) = \sigma^2$.

   Рассмотрим четыре оценки:

$$\hat{\mu}_A = (X_1 + X_2)/2, \quad \hat{\mu}_B = (X_1 + X_2 + X_3)/3, \hat{\mu}_C = 2X_1 - X2, \quad \hat{\mu}_D = (X_1 + X_2 + \ldots + X_{20})/21.$$

   а) Какая из приведенных оценок для μ является несмещённой?

   б) У какой несмещённой оценки самая маленькая дисперсия?

   в) Выберите наиболее эффективную оценку в этом множестве по критерию MSE, если $\sigma = 0.5\mu$.

5. Величины X_1 и X_2 независимы и равномерны на отрезке $[0; a]$ с неизвестным a и $Y = \min\{X_1, X_2\}$.

   а) При каком β оценка $\hat{a} = \beta Y$ для параметра a будет несмещённой?

   б) При каком β оценка $\hat{a} = \beta Y$ для параметра a будет наиболее эффективной по критерию MSE?

6. Величины X_i независимы и имеют закон распределения

x	0	1	a
$\mathbb{P}(X = x)$	$1/4$	$1/4$	$2/4$

   а) Постройте состоятельную оценку для неизвестного a.

   б) Возможно ли в этой задаче построить несмещённую оценку для a?

Домашнее задание 3

Дедлайн: 2025-04-27, 23:59.
Оцениваемые задачи:

1. В одном тропическом лесу водятся удавы и питоны. Длина удавов имеет нормальное распределение $\mathcal{N}(\mu_X, \sigma^2_X)$. По выборке из 10 удавов оказалось, что $\sum X_i = 20$ метрам, а $\sum X_i^2 = 1000$. Длина питонов имеет нормальное распределение $\mathcal{N}(\mu_Y, \sigma^2_Y)$. По выборке из 20 питонов оказалось, что $\sum Y_i = 60$ метрам, а $\sum Y_i^2 = 4000$. Все наблюдения независимы между собой.

   а) Постройте точечные оценки для μ_X, σ^2_X, μ_Y, σ^2_Y.

   б) Постройте двусторонний 95%-й доверительный интервал для σ^2_X/σ^2_Y.

   в) Проверьте гипотезу H_0: $\sigma^2_X = \sigma^2_Y$ против альтернативной H_1: $\sigma^2_Y > \sigma^2_X$ на уровне значимости 5%. Укажите точное p-значение.

   г) Постройте примерный двусторонний 95%-й доверительный интервал для разницы $\mu_X - \mu_Y$ с помощью статистики Уэлча.

   д) Проверьте гипотезу H_0: $\mu_X = \mu_Y$ против альтернативной H_1: $\mu_Y > \mu_X$ на уровне значимости 5% с помощью теста Уэлча. Укажите точное p-значение.

2. Априорное распределение параметра θ является треугольным на отрезке $[0; 40]$ с модой в точке 30. Наблюдаемая величина X — это индикатор того, что $\theta > 20$. Оказалось, что $X = 1$.

   а) Найдите апостериорную плотность θ.

   б) Найдите апостериорное математическое ожидание θ.

   в) Найдите апостериорную медиану θ.

   г) Постройте 94% байесовский интервал наивысшей плотности для θ.

   д) Постройте 94% симметричный по вероятности байесовский интервал для θ.

   Определение треугольного распределения можно найти, например, на википедии :)

Неоцениваемые задачи в удовольствие:

3. Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} \theta x^{\theta - 1}, \text{ при } x \in [0; 1] \\ 0, \text{ иначе.} \end{cases}$$

   а) Оцените значение θ с помощью метода максимального правдоподобия.

   б) Оцените дисперсию оценки $\hat{\theta}_{ML}$ метода максимального правдоподобия.

   в) Как примерно распределена $\hat{\theta}_{ML}$?

   г) Оцените значение θ с помощью метода моментов.

   д) Оцените дисперсию оценки $\hat{\theta}_{MM}$ метода моментов.

   е) Как примерно распределена $\hat{\theta}_{MM}$?

4. Цыганка Роза ничего не понимает в статистике, но у неё всегда с собой колода из 36 карт.

   Помогите цыганке Розе построить точный 95%-й доверительный интервал для неизвестной вероятности p того, что клиента ждёт дальняя дорога и казённый дом.

5. Величины X_1, ..., X_n независимы и одинаково распределены с функцией плотности

$$f(x) = \begin{cases} \frac{\theta \exp(-\theta^2/2x)}{\sqrt{2\pi x^3}} \text{ при } x \in [0; +\infty), \\ 0, \text{ иначе.} \end{cases}$$

   а) Найдите оценку параметра θ методом максимального правдоподобия, если по выборке из 100 наблюдений оказалось $\sum 1/X_i = 12$.

   б) Найдите оценку параметра θ методом максимального правдоподобия для произвольной выборки.

   в) Найдите теоретическую информацию Фишера $I(\theta)$.

   г) Пользуясь данными по выборке постройте оценку $\hat{I}$ для информации Фишера.

   д) Постройте 90% доверительный интервал для θ. Подсказка: $\mathbb{E}(1/X_i) = 1/\theta^2$, интеграл берется, например, заменой $x = \theta^2 a^{-2}$.

6. Величины X_1 и X_2 независимы и распределены по Пуассону с интенсивностью a. Есть две гипотезы, H_0: $a = 1$ и H_a: $a = 2$. Мальвина отвергает H_0 в том случае, если $X_1 + X_2 \geq 2$. Найдите вероятность ошибок первого и второго рода.

7. Величины Y_1, ..., Y_n независимы и имеют распределение Бернулли с неизвестным p, $\hat{p} = \bar{Y}$.

   а) Постройте для неизвестного p доверительный интервал Вальда. Для этого вспомните про сходимость

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \xrightarrow{\text{dist}} \mathcal{N}(0; 1)$$

   и решите неравенство

$$-z_{\text{cr}} \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq z_{\text{cr}}.$$

   б) Постройте для неизвестного p доверительный интервал Вильсона. Для этого воспользуйтесь сходимостью

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{\text{dist}} \mathcal{N}(0; 1).$$

   На этот раз потребуется решить (о ужас!) квадратное неравенство.

   Обозначим центр интервала Вильсона с помощью $\hat{p}_w$.

   в) Докажите, что центр интервала Вильсона $\hat{p}_w$ можно представить как средневзвешенное классической оценки $\hat{p}$ и тривиальной оценки $1/2$,

$$\hat{p}_w = u\hat{p} + (1 - u)(1/2).$$

   Найдите веса u и $(1 - u)$.

г) Докажите, что центр интервала Вильсона $\hat{p}_w$ можно проинтерпретировать следующим образом: добавим f вымышленных единиц и f вымышленных нулей в выборку и посчитаем классическую оценку вероятности для выборки с вымышленными наблюдениями,

$$\hat{p}_w = \frac{\sum_{i=1}^n Y_i + f}{n + 2f}.$$

Какому целому числу примерно равно f для 95%-го доверительного интервала?

д) Докажите, что интервал Вильсона можно записать в виде

$$\hat{p}_w \pm z_{\text{cr}} \cdot \sqrt{\frac{u\hat{p}(1-\hat{p}) + (1-u)(1/2)^2}{n_w}}.$$

Найдите n_w, а также веса u и $(1-u)$.

Таким образом, интервал Вильсона слегка корректирует число наблюдений и использует в качестве оценки дисперсии Y_i средневзвешенное между классической оценкой $\hat{p}(1-\hat{p})$ и тривиальной оценкой $1/4$.

Доверительный интервал Агрести — Коулла для уровня доверия 95% строится следующим образом. В выборку мысленно добавляют два наблюдения равных единице и два наблюдений равных нулю, считают оценку доли

$$\hat{p}_{ac} = \frac{\sum_{i=1}^n Y_i + 2}{n + 4},$$

а затем строят классический интервал Вальда, используя $\hat{p}_{ac}$ вместо классической $\hat{p}$.

е) Правда ли, что при уровне доверия 95% центры интервала Агрести — Коулла и Вильсона совпадают?

ж) Какой 95%-й интервал шире, Агрести — Коулла или Вильсона?

з) С помощью симуляций на компьютере сравните фактическую вероятность накрытия неизвестного параметра p интервалами Вальда, Вильсона и Агрести — Коулла с номинальной 95%-й вероятностью. Для экспериментов возьмите $n = 50$ и различные p от 0 до 1 с шагом 0.1.

Домашнее задание 4

Дедлайн: 2025-05-18, 23:59.
Оцениваемые задачи:

1. Величины (y_i) независимы и одинаково непрерывно распределены. Всего есть 1000 наблюдений. Постройте 95%-й интервал для 90%-го квантиля с помощью выборочных квантилей.

   Если для вычисления необходимых выборочных квантилей использовался код, то приведите его.

2. Есть две выборки: $x = (2.7, 3.5, 4.2, 6.7)$ и $y = (1.6, 2.9, 3.9)$. Все наблюдения независимы. Величины (x_i) одинаково непрерывно распределены между собой, величины (y_i) одинаково непрерывно распределены между собой. Проверьте гипотезу H_0 об одинаковом законе распределения в двух выборках, против альтернативной $\mathbb{P}(x_i > y_j) > 0.5$ на уровне значимости 5%.

   а) Проведите тест Манна — Уитни, используя точное распределение статистики.

б) Проведите тест Манна — Уитни, используя нормальную аппроксимацию. Укажите p-значение.

Неоцениваемые задачи в удовольствие:

3. Рассмотрим тест знаковых рангов Уилкоксона и связанные пары наблюдений (x_i, y_i). При верной H_0 разницы $D_i = x_i - y_i$ одинаково непрерывно распределены и независимы.

   Рассмотрим сумму знаковых рангов $WSR = \sum_{i=1}^n \operatorname{sign}(D_i) \operatorname{rank}(|D_i|)$.

   Найдите ожидание $\mathbb{E}(WSR)$ и дисперсию $\mathbb{V}\mathrm{ar}(WSR)$ при верной H_0.

4. Величины (X_i) независимы и одинаково распределены с неизвестными $\mathbb{E}(X_i) = \mu$ и $\mathbb{V}\mathrm{ar}(X_i) = \sigma^2$. По выборке из 1000 наблюдений оказалось, что $\bar{X} = 30$, а несмещённая выборочная дисперсия равна 900.

   а) Постройте асимптотический 95%-й доверительный интервал для μ. Укажите p-значение для гипотезы H_0: $\mu = 35$ против альтернативной H_a: $\mu \neq 35$.

   б) Постройте асимптотический 95%-й предсказательный интервал для X_{1001}.

   в) Постройте асимптотический 95%-й предсказательный интервал для $(X_{1001} + X_{1002})/2$.

5. Бариста Борис заметил, что в последнее время посетители заказывают только капуччино и раф. Предположим, что посетители выбирают напиток независимо друг от друга, а вероятность выбора капуччино постоянна и равна неизвестному числу p.

   У Бориса есть только две гипотезы, $H_0 : p = 1/3$ и $H_a : p = 2/3$, в которые он до получения данных верит с вероятностями 0.6 и 0.4, соответственно.

   Из первых 100 утренних посетителей $S = 40$ выбрали капуччино. Борис хочет измерить разными способами, насколько этот наблюдаемый результат соотносится с гипотезами.

   а) Найдите $\mathbb{P}(H_0 \mid S = 40)$ и $\mathbb{P}(H_a \mid S = 40)$.

   Борис решил на следующий день повторить эксперимент и снова посчитать S_{new}, количество клиентов из первых ста, которые выберут капуччино.

   б) Найдите $\mathbb{P}(S_{\text{new}} \geq S \mid S = 40, H_0)$ и $\mathbb{P}(S_{\text{new}} \geq S \mid S = 40, H_a)$.

   в) Какие из вероятностей можно посчитать без мнения Бориса о $\mathbb{P}(H_0)$ и $\mathbb{P}(H_a)$?

   г) Какая из вероятностей называется p-значением для гипотезы H_0 и статистики S?

6. По таблице сопряжённости проверьте гипотезу о независимости двух признаков на уровне значимости 5% против альтернативной гипотезы о зависимости признаков. Укажите p-значение.

	$X = A$	$X = B$
$Y = C$	50	60
$Y = D$	20	30
$Y = E$	60	50

7. Рассмотрим таблицу сопряжённости

$X = A$	$X = B$	$X = C$	$X = D$
50	70	80	60

а) На уровне значимости 5% проверьте гипотезу об одинаковых вероятностях $p_a = p_b = p_c = p_d$ против альтернативной о том, что хотя бы одно из равенств нарушено.

б) На уровне значимости 5% проверьте гипотезу об одинаковых вероятностях $p_a = p_b = p_c = p_d$ против альтернативной о том, что $p_a \neq p_b = p_c$.

в) На уровне значимости 5% проверьте гипотезу об одинаковых вероятностях $p_a = p_b = p_c$ против альтернативной о том, что $p_a \neq p_b = p_c$.

В каждом случае укажите p-значение.

Домашнее задание 5

Дедлайн: 2025-06-01, 23:59.
Оцениваемые задачи:

1. Величины (y_i) независимы и экспоненциально распределены с интенсивностью λ.

   Количество наблюдений n велико. Тестируем гипотезу H_0: $\lambda = 2$ против альтернативы $\lambda \neq 2$.

   а) Выведите формулы для теста отношения правдоподобия LR, теста множителей Лагранжа LM и теста Вальда W.

   б) Проведите тесты для конкретной выборки с $n = 1000$, $\bar{y} = 2.2$ и уровня значимости 1%.

2. Величины (y_i) независимы и нормально распределены $\mathcal{N}(\mu, 1)$.

   Количество наблюдений n велико. Тестируем гипотезу H_0: $\mu = 0$ против альтернативы $\mu \neq 0$.

   а) Выведите формулы для теста отношения правдоподобия LR, теста множителей Лагранжа LM и теста Вальда W.

   б) Проведите тесты для конкретной выборки с $n = 1000$, $\sum y_i = 1000$, $\sum y_i^2 = 4000$ и уровня значимости 1%.

Неоцениваемые задачи в удовольствие:

3. Гипотеза H_0 описывается 5-ю независимыми уравнениями, неограниченный максимум лог-правдоподоб равен $\ell_{UR} = -200$, а ограниченный — $\ell_R = -209$. Число наблюдений n велико. Альтернативная гипотеза состоит в том, что хотя бы одно уравнение не выполнено.

   а) Отвергается ли H_0 на уровне значимости 1%?

   б) Найдите p-значение.

4. Оценка неизвестного вектора параметров $a = (a_1, a_2, a_3)$ равна $\hat{a} = (1, 2, 3)$ с оценкой ковариационной матрицы

$$\widehat{\mathbb{V}\mathrm{ar}}(\hat{a}) = \begin{pmatrix} 9 & -1 & 2 \\ & 16 & -1 \\ & & 10 \end{pmatrix}.$$

   Число наблюдений велико. Рассмотрим гипотезу H_0: $a_1 = a_2 = a_3$ против альтернативы о том, что хотя бы одно уравнение не выполнено.

а) Предложите естественную оценку $\hat{b}$ для вектора $b = (a_1 - a_2, a_2 - a_3)$.

б) Оцените ковариационную матрицу $\widehat{\mathbb{V}\mathrm{ar}}(\hat{b})$.

в) Переформулируйте H_0 в терминах вектора b.

г) Проведите тест Вальда гипотезы H_0 на уровне значимости 5%.

5. Мы оцениваем три неизвестных параметра, $(\theta_1, \theta_2, \theta_3)$. При максимизации с учётом ограничений гипотезы H_0 оказывается, что градиент лог-правдоподобия равен $\operatorname{grad} \ell = (-0.1, 0.2, 0)$, а матрица Гессе в точке ограниченного экстремума равна

$$H = \begin{pmatrix} -5 & -2 & 0 \\ & -6 & 0 \\ & & -10 \end{pmatrix}$$

Число наблюдений велико.

а) Чему равен градиент лог-правдоподобия в точке неограниченного экстремума?

б) Протестируйте H_0 на уровне значимости 1% с помощью теста множителей Лагранжа.

6. Вспомним классический хи-квадрат тест Пирсона на соответствие выборки заданному дискретному закону распределения со статистикой

$$S = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i},$$

где k — число клеток таблицы, f_i — количество наблюдений, попавших в i-ую клетку таблицы, а p_1, p_2, ..., p_k — вероятности, предполагаемые в H_0.

С каким тестом $(LR/LM/W)$ совпадает данная статистика?

Домашнее задание 6

Дедлайн: 2025-06-18, 23:59.
Оцениваемые задачи:

1. Случайные величины y_1, ..., y_n независимы и одинаково распределены с $\mathbb{P}(y_i = 1) = p$ и $\mathbb{P}(y_i = 0) = 1 - p$. Рассмотрим бутстрэп-выборку y_1^*, ..., y_n^*.

а) Найдите $\mathbb{P}(y_1^* = y_1)$ и $\mathbb{P}(y_1^* = y_2^*)$.

б) Найдите $\mathbb{E}(y_i^*)$ и $\mathbb{V}\mathrm{ar}(y_i^*)$.

в) Найдите $\mathbb{C}\mathrm{ov}(y_1^*, y_2^*)$.

2. Винни-Пух хочет проверить 5 нулевых гипотез. Он посчитал p-значения для каждой из них, $p = (0.03, 0.04, 0.08, 0.15, 0.30)$.

а) Какие гипотезы отвергнет алгоритм Холма — Бонферонни, гарантирующий $FWER = 0.2$?

б) Какие гипотезы отвергнет алгоритм Беньямини — Хохберга, гарантирующий $FDR = 0.2$?

Неоцениваемые задачи в удовольствие:

3. Величины y_1, ..., y_n независимы и одинаково распределены с функцией плотности

$$f(y) = \begin{cases} \exp(a - y) \text{ если } y \geq a, \\ 0, \text{ иначе.} \end{cases}$$

   а) Найдите оценку неизвестного параметра a методом максимального правдоподобия.

   б) Какова фактическая вероятность накрытия параметра a при построении наивного бутстрэп доверительного интервала с номинальной вероятностью накрытия 95%?

4. Опишите алгоритм наивного бутстрэпа для построения 95% доверительного интервала для истинной медианы распределения.

5. У исследователя всего две нулевых гипотезы. Каждая из них априорно верно с вероятностью 0.2 независимо от других. При верной отдельной нулевой гипотезы H_j распределение соответствующей ей тестовой статистики непрерывно. Для упрощения будем считать, что если отдельная нулевая гипотеза H_j не верна, то её p-значение в точности равно 0.

   а) Вспомните, как распределено p-значение при верной нулевой гипотезе.

   б) Рассмотри алгоритм Холма — Бонферонни, гарантирующий $FWER = 0.2$. Какова условная вероятность того, что он отвергнет конкретную нулевую гипотезу с известным p-значением равным u?

   в) Рассмотри алгоритм Беньямини — Хохберга, гарантирующий $FDR = 0.2$. Какова условная вероятность того, что он отвергнет конкретную нулевую гипотезу с известным p-значением равным u?

6. В одном из вариантов бутстрэпа (subsampling bootstrap) в бутстрэп выборку попадает $m < n$ наблюдений без повторов из исходной выборки в n наблюдений. Предположим, что исходные n наблюдений y_1, y_2, ..., y_n независимы и равномерны Unif$[0, 1]$. Рассмотрим бутстрэп выборку y_1^*, ..., y_m^*.

   а) Как распределено y_i^*?

   б) Найдите $\mathbb{C}\mathrm{orr}(y_1^*, y_2^* \mid y_1, y_2, \ldots, y_n)$.

   в) Найдите $\mathbb{C}\mathrm{orr}(y_1^*, y_2^*)$.