

Расчётно-графическая работа №1

Задание 1

Первое задание представлено в четырёх вариантах. Сами варианты.

1. В файле [iris.csv](#)¹ (здесь и далее ссылки кликабельны) представлены данные о параметрах различных экземплярах цветка ириса. Какой вид в датасете представлен больше всего, какой – меньше? Рассчитайте выборочное среднее, выборочную дисперсию, выборочную медиану и выборочную квантиль порядка 2/5 для суммарной площиади (более точно – оценки площиади) чашелистика и лепестка всей совокупности и отдельно для каждого вида. Построить график эмпирической функции распределения, гистограмму и box-plot суммарной площиади чашелистика и лепестка для всей совокупности и каждого вида.
2. В файле [sex_bmi_smokers.csv](#) приведены данные (пол, ИМТ, курит/не курит) о более 1000 испытуемых. Сравните количество курящих мужчин и некурящих женщин. Рассчитайте выборочное среднее, выборочную дисперсию, выборочную медиану и выборочную квантиль порядка 3/5 ИМТ всех наблюдателей и отдельно для каждой возможной комбинации пол-курение. Построить график эмпирической функции распределения, гистограмму и box-plot ИМТ для всех наблюдателей и отдельно для каждой возможной комбинации пол-курение.
3. В файле [cars93.csv](#) представлены данные об автомобилях, проданных в некотором автосалоне за 93 год. Какие типы автомобилей представлены в датасете? Какой тип наиболее распространен, какой – менее? Рассчитайте выборочное среднее, выборочную дисперсию, выборочную медиану и межквартильный размах мощности для всей совокупности автомобилей и отдельно для американских и не американских авто. Построить график эмпирической функции распределения, гистограмму и box-plot мощности для всей совокупности и отдельно для каждого типа авто.
4. В файле [mobile_phones.csv](#) приведены данные о мобильных телефонах. В сколько моделей можно вставить 2 сим-карты, сколько поддерживают 3-G, каково наибольшее число ядер у процессора? Рассчитайте выборочное среднее, выборочную дисперсию, выборочную медиану и выборочную квантиль порядка 2/5, построить график эмпирической функции распределения, гистограмму и box-plot для емкости аккумулятора для всей совокупности и в отдельности для поддерживающих/не поддерживающих Wi-Fi.

Задание 2

Предположите, какому вероятностному закону соответствует распределение показателя, рассмотренного (расчет выборочных характеристик и визуализация) в задании №1. Оцените параметры данного распределения методом максимального правдоподобия или методом моментов (**математическое обоснование оценки строго обязательно**). Какими статистическими свойствами обладает найденная оценка (**обосновать**)? Найти теоретические смещение, дисперсию, MSE, информацию Фишера (если определена для вашей модели).

Задание 3

Пусть P_θ – выбранное в предыдущем задании распределение, параметризующееся вектором θ (пример –), $\hat{\theta}$ – оценка параметра θ , полученная в предыдущем упражнении. Проведите численный эксперимент по следующей схеме:

- Зафиксируйте конкретное значение $\theta = \theta_0$
- Заведите массив $\{n_1, \dots, n_k\}$ объемов выборки

¹датасеты взяты с открытых источников, в частности с сайта для соревнований по Data Sciene и Machine Learning [Kaggle.com](#)

- Сгенерируйте из распределения P_{θ_0} достаточно большое количество M выборок объема n , где n принимает значения из массива $\{n_1, \dots, n_m\}$. Для каждой сгенерированной выборки вычислите оценку $\hat{\theta}$
- Эмпирически рассмотреть поведение оценки $\hat{\theta}$ в зависимости от объема выборки (можно для каждого объема выборки n_i вывести описательные статистики для оценок, изобразить гистограмму, box-plot, violin-plot).

Задание 4 (не обязательное, бонусное, со звёздочкой)

Небольшая теоретическая справка, не уверен что сей материал будет затронут на теории. Рассмотрим байесовскую постановку задачи точечного оценивания. В отличие от традиционной, или частотной, постановки, где параметр θ воспринимается как неизвестное, но фиксированное значение, в байесовской постановке θ есть случайная величина, имеющая некоторое априорное распределение.

Обозначим плотность данного распределения как $\pi(\theta)$ (в дискретном случае это функция вероятностей). Пусть $l(\theta, \hat{\theta})$ – функция потерь/ошибки (самые простые примеры: $l(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ – квадратичная, $l(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ – абсолютная). В традиционной постановке в качестве критерия эффективности/оптимальности можно использовать функцию $E_\theta l(\theta, \hat{\theta})$ (здесь математическое ожидание берется относительно распределения P_θ при **фиксированном** θ) – это фиксированная величина, но в байесовской постановке это уже случайная величина. Поэтому, чтобы получить конкретное число, нужно величину $E_\theta l(\theta, \hat{\theta})$ усреднить относительно априорного распределения:

$$\begin{aligned} r(\hat{\theta}) &= E_{\pi(\theta)}[E_\theta l(\theta, \hat{\theta})] \quad (\text{внешнее м.о. – усреднение относительно априорного распределения}) \\ &= El(\theta, \hat{\theta}) \quad (\text{здесь м. о. берется относительно совместного распределения } X_1, \dots, X_n, \theta). \end{aligned}$$

Данную функцию будем называть *байесовским риском*. Оценку, минимизирующую функцию $r(\cdot)$, будем называть *байесовской*. Можно показать, что байесовская оценка минимизирует функцию $E(l(\theta, \hat{\theta})|X)$ (здесь математическое ожидание берется относительно апостериорного распределения $\pi(\theta|X)$, которое находится по теореме Байеса)

$$\pi(\theta|X) = \frac{L(X|\theta)\pi(\theta)}{\int L(X|\theta)\pi(\theta)d\theta},$$

доказать в качестве дополнительного упражнения, что байесовская оценка действительно минимизирует среднюю относительно апостериорного распределения функцию потерь

Более того, если функция потерь квадратическая, то байесовская оценка – среднее относительно апостериорного распределения $E(\theta|X)$, **доказать в качестве дополнительного упражнения**.

Само задание. Найдите байесовскую оценку параметра θ (относительно среднеквадратической ошибки). Проведите эксперимент по схожей схеме, что и в задании №3 (здесь уже нужно учитывать, что θ – случайная величина).

Сами варианты (сначала указывается семейство распределений для выборки, затем – априорное распределение параметра, в конце – значения параметров для эксперимента, разделитель – точка с запятой):

1. $\mathcal{N}(\theta, b^2); \mathcal{N}(\mu, \sigma^2); \mu = 0, b = \sigma = 1$.
2. $\text{Pois}(\theta); \Gamma(k, \lambda), \lambda > 0, k \in \mathbb{N}$ (при решении явно указывайте используемую параметризацию); $\lambda = k = 1$.
3. $\text{Geom}(\theta); \text{Be}(a, b), a, b > 0$ (бета-распределение); $a = b = 1$.
4. $\text{Exp}(\theta); \Gamma(k, \lambda), \lambda > 0, k \in \mathbb{N}$ (при решении явно указывайте используемую параметризацию); $k = \lambda = 1$.