

Лабораторная работа №1

Задача 1. Пусть выборка X_1, \dots, X_n соответствует классу распределений F_θ , $\theta \in E \subset \mathbb{R}$. При каком минимальном объеме выборки n равномерно для $\theta \in E$ выборочное среднее отличается от математического ожидания μ_θ не более чем на $\varepsilon > 0$ с вероятностью, не меньшей $1 - \delta$, $\delta \in (0, 1)$? Сгенерировать 500 выборок найденного объема при $\varepsilon = 0.01$ и $\delta = 0.05$ из указанного распределения F_θ при конкретном параметре θ и посчитать, сколько раз выборочное среднее отличается от математического ожидания μ_θ более чем на ε .

Задача представлена в семи вариантах. Для краткости указывается класс распределений, область ограничения параметра, значение параметра для эксперимента.

1. $\text{Bern}(p)$, $p \in (0, 1)$, $p = 2/3$,
2. $\text{Pois}(\lambda)$, $\lambda \in (0, 10]$, $\lambda = 2$,
3. $\text{Geom}(p)$ (указать вид используемой параметризации), $p \in (1/4, 1)$, $p = 4/5$,
4. $U[0, \theta]$, $\theta \in (0, 10)$, $\theta = 6$,
5. $U[-2\theta, 3\theta]$, $\theta \in (0, 5)$, $\theta = 3$,
6. $\text{Exp}(\lambda)$ (указать вид используемой параметризации), $\lambda \in (1, 5)$, $\lambda = 3$,
7. $\mathcal{N}(5, \sigma^2)$, $\sigma^2 \in (0, 4)$, $\sigma^2 = 2$.

Задача 2. Представлена в 4 вариантах.

1. В файле *iris.csv* 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 площади чашелистика и лепестка всей совокупности и отдельно для каждого вида. Построить график эмпирической функции распределения, гистограмму и box-plot суммарной площади 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.
3. В файле *cars93.csv* представлены данные об автомобилях. 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 каждого типа автомобиля. 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.

Ключевые понятия:

- Закон больших чисел (слабый, для независимых одинаково распределенных случайных величин)
- Центральная предельная теорема (для независимых одинаково распределенных случайных величин)
- Предположения на выборку
- Эмпирическая функция распределения, её состоятельность
- Выборочное среднее, его несмещенность, состоятельность и асимптотическая нормальность
- Смещенная и несмещенная выборочная дисперсия
- Теоретическая (в том числе непрерывный случай) и выборочная квантили
- Выборочная медиана