Лабораторная работа №1

Задача 1. Пусть выборка $X_1, \ldots, X_n$ 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. Bern(p), $p \in (0, 1)$, $p = 2/3$,
2. Pois(λ), $\lambda \in (0, 10]$, $\lambda = 2$,
3. 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. Exp(λ) (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

Лабораторная работа №2

Задача представлена в 7 вариантах, каждому достанутся две задачи. Схема эксперимента везде одна и та же.

1. Методом моментов найти оценку параметра θ равномерного распределения на $[-\theta, \theta]$. Найти смещение оценки, дисперсию, среднеквадратическую ошибку. Эксперимент для $\theta = 10$.

2. Методом моментов найти оценку масштабирующего параметра θ распределения Лапласа (сдвиг считать нулевым). Найти смещение оценки, дисперсию, среднеквадратическую ошибку. Эксперимент для $\theta = 0.5$.

3. Методом максимального правдоподобия найти оценку параметра θ биномиального распределения $\mathrm{Bin}(n, \theta)$, считая n известным. Найти смещение оценки, дисперсию, среднеквадратическую ошибку. Является ли найденная оценка эффективной? Эксперимент при $n = 4$, $\theta = 1/5$.

4. Можно ли оценить параметр сдвига θ распределения Коши с известным масштабирующим параметром с помощью метода моментов? С помощью какой оценки можно оценить параметр θ? Показать её состоятельность (*подсказка*: см. теорему об асимптотическом поведении среднего члена вариационного ряда). Эксперимент для $\mathrm{Cauchy}(2, 1)$.

5. Найти оценку максимального правдоподобия параметра θ для распределения с плотностью

$$f_\theta(x) = \frac{2x}{\sqrt{2\pi}} \exp\left(\frac{(\theta - x^2)^2}{2}\right).$$

   Найти её смещение, дисперсию и среднеквадратическую ошибку. Какими свойствами обладает данная оценка? Эксперимент при $\theta = 5$.

6. С помощью метода моментов найти оценку параметра θ распределения с плотностью

$$f_\theta(x) = \frac{1}{(k-1)!\theta^k} x^{k-1} e^{-x/\theta} \mathbb{1}(x > 0),$$

   если $k \in \mathbb{N}$ – известный параметр. Какими свойствами обладает данная оценка? Эксперимент при $\theta = 2$, $k = 3$.

7. С помощью метода моментов найти оценку параметра θ геометрического распределения (указать вид используемой параметризации). Какими свойствами обладает оценка? Эксперимент при $\theta = 0.3$.

Сгенерируйте 500 выборок объема 50 с указанным значением параметра θ. Сколько раз оценка отклонится от истинного значения параметра более чем на 0.01? То же самое сделать для объемов выборки 100, 500, 1000, 2500. Визуализируйте результат. Как объяснить полученный результат?

Ключевые понятия:

- Постановка задачи точечного оценивания параметров
- Состоятельность, несмещенность, асимптотическая нормальность
- Эффективность оценки, информация Фишера, неравенство Рао-Крамера
- Метод моментов
- Метод максимального правдоподобия

Лабораторная работа №3

Задача 1. Предъявите доверительный интервал уровня $1-\alpha$ для указанного параметра при данных предположениях (с обоснованиями). Сгенерируйте 2 выборки объёма объёма 25 и посчитайте доверительный интервал. Повторить 1000 раз. Посчитайте, сколько раз 95-процентный доверительный интервал покрывает реальное значение параметра. То же самое сделайте для объема выборки 10000. Как изменился результат? Как объяснить?

Задача представлена в 3 вариантах. Везде даны две независимые выборки X, Y из нормальных распределений $\mathcal{N}(\mu_1, \sigma_1^2)$, $\mathcal{N}(\mu_2, \sigma_2^2)$ объема n, m соответственно. Сначала указывается оцениваемая функция, потом данные об остальных параметрах, затем параметры эксперимента и подсказки.

1. $\tau = \mu_1 - \mu_2$; σ_1^2, σ_2^2 известны; $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$; воспользуйтесь функцией
$$\frac{\overline{X} - \overline{Y} - \tau}{\sigma}, \quad \sigma^2 = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

2. $\tau = \mu_1 - \mu_2$; $\sigma_1^2 = \sigma_2^2$ неизвестна; $\mu_1 = 2$, $\mu_2 = 1$, $\sigma_1^2 = \sigma_2^2 = 1$; воспользуйтесь функцией
$$\sqrt{\frac{mn(m+n-2)}{m+n}} \frac{\overline{X} - \overline{Y} - \tau}{\sqrt{n\mathrm{Var}(X) + m\mathrm{Var}(Y)}},$$
где $\mathrm{Var}(.)$ – выборочная смещенная дисперсия. Смотрите в сторону распределения Стьюдента.

3. $\tau = \sigma_1^2/\sigma_2^2$; μ_1, μ_2 неизвестны; $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 2$, $\sigma_2^2 = 1$; воспользуйтесь функцией
$$\frac{n(m-1)\mathrm{Var}(X)}{m(n-1)\mathrm{Var}(Y)},$$
где $\mathrm{Var}(.)$ – выборочная смещенная дисперсия. Смотрите в сторону распределения Фишера.

Задача 2. Постройте асимптотический доверительный интервал уровня $1-\alpha$ для указанного параметра. Проведите эксперимент по схеме, аналогичной первой задаче.

Задача представлена в 7 вариантах. Сначала указывается класс распределений (однопараметрический) и оцениваемый параметр, затем параметры эксперимента и подсказки.

1. $\mathrm{Exp}(\lambda)$; медиана; $\lambda = 1$; воспользуйтесь предельной теоремой об асимптотическом поведении среднего члена вариационного ряда.

2. Распределение Лапласса с неизвестным параметром сдвига μ и единичным масштабирующим параметром; μ; $\mu = 2$; можно воспользоваться подсказкой для предыдущего варианта, хотя другие способы решения приветствуются.

3. $U[-\theta, \theta]$; θ; $\theta = 5$; воспользуйтесь предельной теоремой об асимптотическом поведении крайних членов вариационного ряда.

4. Geom(p); p; $p = 0.7$; тут рецепт стандартный).

5. Pois(λ); второй момент; $\lambda = 1$; воспользоваться асимптотической нормальностью второго момента.

6. $U[0; \theta]$; θ; $\theta = 2$, см. п. 3.

7. $U[-\theta; 0]$; θ; $\theta = 3$, см. п. 3.

Ключевые понятия:

- Доверительные интервалы. Доверительные интервалы для параметров нормального распределения. Теорема Фишера
- Доверительные интервалы. "Универсальный" рецепт.
- Асимптотические доверительные интервалы. "Обычный" рецепт.
- Теоремы об асимптотическом поведении среднего и крайних членов вариационного ряда.

Лабораторная работа №4

Задание представлено в 6 вариантах. Для каждого вопроса требуется формализовать задачу и проверить статистическую гипотезу двумя критериями, если не сказано иное.

Вариант 1. В файле *exams_dataset.csv* (исходник отсюда) представлены данные об экзаменуемых.

1. Часто результаты интеллектуальных тестов аппроксимируют нормальным распределением, в частности, итоги IQ-тест можно приблизить нормальным распределением со средним 100 и стандартным отклонением 15. Можно ли приблизить результаты по математике нормальным распределением?

2. Можно ли утверждать, что результаты по чтению и письменной части принципиально не отличаются?

3. Есть подозрение, что посещавшие подготовительные курсы более успешны на экзаменах. Проверите данное утверждение.

Вариант 2. В файле *mobile_phones.cvs* (исходник отсюда) представлены данные о мобильных телефонах.

1. Разумно ли считать, что емкость аккумулятора распределена равномерно?

2. Верно ли, что телефонов с поддержкой 3G больше моделей с Wi-Fi? А разнится ли количество телефонов с touch screen от моделей с двумя сим-картами? На каждый вопрос по тесту.

3. Есть подозрение, что цена зависит от объема оперативной памяти. Проверите данное утверждение.

Вариант 3. В файле *sex_bmi_smokers.csv* данные о пациентах.

1. Разумно ли индекс массы тела аппроксимировать нормальным законом?

2. Отличаются ли принципиально распределение индекса массы тела у мужчин и женщин?

3. Есть подозрение, что курящие склонны к ожирению. Кажется, что мужчины более склонны к ожирению. Проверьте данные утверждения, на каждую гипотезу по одному тесту.

Вариант 4. В файле *song_data.csv* (взято отсюда) приведены данные о музыкальных произведениях.

1. Разумно ли популярность песни аппроксимировать нормальным законом?

2. Отличается ли принципиально распределение рейтинга песни в зависимости от продолжительности (разбейте условно на "длинные" и "короткие", порог выбирайте сами)?

3. Зависит ли популярность песни от продолжительности?

Вариант 5. В файле *MEN_SHOES.csv* (исходник отсюда) приведены данные о продажах мужской обуви.

1. Разумно ли количество проданных экземпляров обуви аппроксимировать распределением Пуассона, а рейтинг – нормальным распределением (по 1 тесту на каждый вопрос)?

2. Верно ли что распределения количества проданных экземпляров существенно не отличаются в зависимости от бренда? Тот же вопрос для цены (по одному тесту на утверждение).

3. Есть подозрение, что рейтинг зависит от цены. Проверить данное предположение.

Вариант 6. В файле *cars93.csv* приведены данные об авто.

1. Разумно ли мощность считать равномерно распределенной, а цену – нормально (для каждого теста по вопросу)?

2. Верно ли, что распределения мощности для каждого типа авто принципиально не отличаются? Тот же вопрос про цену (для каждого вопроса по тесту).

3. Есть подозрение, что цена зависит от мощности авто. Проверьте данное предположение.

Ключевые понятия:

- Постановка задачи проверки статистических гипотез.
- Статистический критерий и его статистика. Области принятия и опровержения нулевой гипотезы, p-value.
- Ошибки I и II рода.
- Критерии согласия. Примеры критериев.
- Критерии однородности. Примеры критериев.
- Критерии независимости. Примеры критериев.

Лабораторная работа №5

Задание представлено в 4 вариантах. Для каждого варианта требуется построить линейную модель, вычислить оценки коэффициентов модели и остаточной дисперсии, построить для них доверительные интервалы, вычислить коэффициент детерминации, проверить указанные в условии гипотезы с помощью построенной линейной модели.

Указание: из встроенных функций разрешается пользоваться квантильными функциями и средствами для квадратичной оптимизации (иными словами, готовую обертку для построения линейной модели не использовать)

Вариант 1. В файле *cars93.csv* представлены данные о продажах различных авто.

1. Постройте линейную модель, где в качестве независимых переменных выступают расход в городе, расход на шоссе, мощность (вместе со свободным коэффициентом), зависимой – цена.

2. Проверьте следующие подозрения:

   - Чем больше мощность, тем больше цена
   - Цена изменяется в зависимости от расхода в городе
   - Цена зависит от расхода в городе и от расхода на шоссе

Вариант 2. В файле *mobile_phones.cvs* представлены данные о мобильных телефонах.

1. Постройте линейную модель, где в качестве независимых переменных выступают высота, ширина и емкость аккумулятора (вместе со свободным коэффициентом), зависимой – масса телефона.

2. Проверьте следующие подозрения:

   - Чем больше высота телефона, тем больше масса
   - Чем больше ширина, тем больше масса
   - Масса зависит и от ширины, и от высоты

Вариант 3. В файле *MEN_SHOES_.csv* приведены данные о мужской обуви.

1. Постройте линейную модель, где в качестве независимых переменных выступают количество проданных экземпляров и цена (вместе со свободным коэффициентом), зависимой – рейтинг.

2. Проверьте следующие подозрения:

   - Чем больше продажи, тем больше рейтинг
   - Рейтинг за зависит от цены
   - Рейтинг зависит и от цены, и от количества проданных экземпляров

Вариант 4. В файле *song_ data.csv* приведены данные о музыкальных произведениях.

1. Постройте линейную модель, где в качестве независимых переменных выступают продолжительность, "танцевальность" и энергичность (вместе со свободным коэффициентом), зависимой – популярность.

2. Проверьте следующие подозрения:

   - Чем больше энергичность, тем больше популярность
   - Популярность зависит от продолжительности
   - Популярность зависит от энергичности и "танцевальности"

Ключевые понятия:

- Линейная регрессия. Основные предположения
- Метод наименьших квадратов и его свойства
- Основная теорема о линейной регрессии. Следствия из нее
- Остаточная дисперсия. Коэффициент детерминации.