Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»

Математическая статистика

2025–2026 учебный год

Расчётно-графическая работа №1

Описание выборки. Оценивание параметров. Доверительные интервалы

Авторы: Смирнов И. С., Лукина М. В.

1. Общие сведения

- **Тема РГР №1:** разделы 1–2 дисциплины (описание выборки; точечное и интервальное оценивание).
- **Формат:** работа выполняется **в аудитории** на практических занятиях; допускается командная работа **2–3 человека**.
- **Инструменты:** Python/Colab, Excel, R и др. (на выбор команды). В отчёте должны присутствовать пояснения и обоснования.
- **Оценивание:** максимум **20 баллов** (шкала приведена в разделе 5). Возможны бонусные баллы (раздел 6).

2. Данные и вариант

Команде выдаётся вариант вида **A–1, B–7, . . .** и CSV-файл с данными.

В файле находятся четыре столбца X_1, X_2, X_3, X_4 объёма n ($n = 200$).

Важно. Столбцы X_1, X_2, X_3 имеют распределения из следующего списка:

- нормальное распределение $N_{a,\sigma}$;
- равномерное распределение $U_{a,b}$;
- экспоненциальное распределение со сдвигом $\mathrm{Exp}_{\lambda,c}$.

Соответствие $\{X_1, X_2, X_3\} \leftrightarrow \{\text{законы}\}$ **заранее не сообщается** и должно быть установлено по данным.

Столбец X_4 содержит данные более сложной структуры; требуется выполнить его первичное описание и сделать вывод о возможной неоднородности данных. Кластеризация X_4 предлагается как бонусное задание.

3. Что сдаём

Сдаётся короткий отчёт (допускается ноутбук), содержащий:

- идентификатор варианта и объём выборки n;
- графики и вычисления по заданиям;
- краткие выводы по каждому столбцу и общий итоговый вывод.

4. Задание (обязательная часть)

Все пункты ниже выполняются для столбцов X_1, X_2, X_3.

4.1. Первичное описание выборки

1. Постройте вариационный ряд.
2. Постройте эмпирическую функцию распределения $F_n(x)$.
3. Постройте гистограмму (обоснуйте выбор числа интервалов/ширины интервала; допускаются стандартные правила, см. Приложение А). Рекомендуется построить **несколько гистограмм** с разными правилами и сравнить результат.
4. Вычислите числовые характеристики: выборочное среднее $\overline{x}$, дисперсию (смещённую) S^2, дисперсию (несмещённую) $\hat{\sigma}^2$, стандартные отклонения S и $\hat{\sigma}$, медиану m_e и квантили (в том числе квартили).
5. Коротко опишите форму распределения: симметрия/асимметрия, наличие выбросов, наличие/отсутствие естественных границ значений.

4.2. Предположение о виде закона распределения

Используя гистограмму, ЭФР и числовые характеристики, для каждого столбца выберите наиболее подходящую модель из списка: $N_{a,\sigma}$, $U_{a,b}$, $\mathrm{Exp}_{\lambda,c}$. Дайте краткое обоснование (2–6 предложений).

4.3. Оценивание параметров: метод моментов и метод максимального правдоподобия

Для выбранной модели оцените параметры двумя способами:

1. **Метод моментов:** приравняйте теоретические моменты $\mathrm{E}X$ и $\mathrm{D}X$ к соответствующим выборочным характеристикам и найдите оценки параметров.
2. **Метод максимального правдоподобия:** выпишите функцию правдоподобия (или логарифм правдоподобия) и найдите оценки, максимизирующие её.
3. **Сравнение:** сравните численно оценки двух методов и прокомментируйте различия (если они есть).

4.4. Оценивание параметрической вероятности двумя способами

Для каждого столбца выберите порог x_0 (например, $x_0 = \overline{x} + \hat{\sigma}$) и оцените величину $P(X > x_0)$ двумя способами:

- **эмпирически** (по доле наблюдений, превышающих x_0);
- **параметрически** (подставив найденные оценки параметров в теоретическую формулу для выбранной модели).

Сравните две оценки и прокомментируйте расхождение.

4.5. Оценка моментов по сгруппированной выборке

Используя построенную гистограмму (частоты по интервалам и середины интервалов), оцените $\mathrm{E}X$ и $\mathrm{D}X$ по сгруппированной выборке и сравните с оценками по исходным данным. (Формулы см. Приложение А.)

4.6. Доверительные интервалы

Зафиксируйте уровень доверия $1 - \alpha$ (по умолчанию 0.95).

1. Для каждого столбца X_1, X_2, X_3 постройте **асимптотический** доверительный интервал для $\mathrm{E}X$ (по ЦПТ).
2. Для того столбца, который вы отнесли к $N(\mu, \sigma^2)$, постройте **точные** доверительные интервалы:
   - для μ при неизвестной σ^2;
   - для σ^2.
3. Дайте интерпретацию доверительных интервалов (2–5 предложений): что означает интервал и почему это **не** «вероятность параметра».

4.7. Итоговый вывод

Сформулируйте итог (5–12 строк): какая модель выбрана для каждого столбца, какие параметры получены, насколько «узкие» интервалы и какие практические выводы можно сделать.

5. Оценивание (20 баллов)

Рекомендуемая шкала (может уточняться преподавателем потока):

Элемент работы	**Баллы**
Описание выборки (графики + характеристики + выводы)	6
Гипотеза о виде закона + обоснование (для трёх столбцов)	2
Оценивание параметров: метод максимального правдоподобия	4
Оценивание параметров: метод моментов + сравнение с ММП	4
Доверительные интервалы + интерпретация	4
Итого	**20**

6. Бонус (по желанию): бимодальность и кластеризация (до +2 баллов)

Для столбца X_4:

1. Постройте гистограмму и эмпирическую функцию распределения; опишите форму распределения и возможные признаки неоднородности данных.
2. За дополнительные баллы: разделите выборку на два кластера (например, алгоритмом k-средних при $k = 2$ или разбиением порогом) и сравните характеристики подвыборок.
3. Объясните, почему «общее среднее» может плохо описывать смесь двух режимов.

Приложение А. Теоретическая справка

А.1. Эмпирическая функция распределения

Пусть $X = (X_1, \ldots, X_n)$ — выборка, а $x = (x_1, \ldots, x_n)$ — её реализация.

Эмпирическая частота $\nu_n(x)$ — число элементов выборки, меньших x

Эмпирическая функция распределения определяется как

$$F_n(x) = \frac{\nu_n(x)}{n}.$$

А.2. Выборочные характеристики

Выборочное среднее:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Выборочная дисперсия (смещённая):

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad S = \sqrt{S^2}.$$

Несмещённая выборочная дисперсия (исправленная):

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad \hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

(Для вычислений по данным подставляют x_i вместо X_i и $\overline{x}$ вместо $\overline{X}$.)

А.3. Выбор ширины интервала гистограммы (примеры)

- Правило Скотта: $h = 3.5\, S\, n^{-1/3}$.
- Правило Фридмана–Диакониса: $h = 2\,\mathrm{IQR}\, n^{-1/3}$, где $\mathrm{IQR} = q_{0.75} - q_{0.25}$.
- Правило Стерджеса: $k = 1 + \lfloor \log_2 n \rfloor$.

А.4. Метод моментов

Идея: приравнять теоретические значения $\mathrm{E}X$ и $\mathrm{D}X$ к соответствующим выборочным оценкам и решить систему уравнений относительно параметров.

A.5. Метод максимального правдоподобия

Пусть $f(x \mid \theta)$ — плотность (или функция вероятностей). Функция правдоподобия:

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \qquad \ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta).$$

Оценка ММП: $\hat{\theta} = \arg\max_\theta L(\theta)$ (эквивалентно $\arg\max_\theta \ell(\theta)$).

A.6. Формулы оценок для используемых моделей

A.6.1. Нормальное распределение $N_{a,\sigma}$

Теория: $\mathrm{E}X = a$, $\mathrm{D}X = \sigma^2$.

Метод моментов: $\hat{a} = \overline{x}$, $\hat{\sigma}^2 = S^2$ (или $\hat{\sigma} = S$).

Метод максимального правдоподобия: $\hat{a} = \overline{x}$, $\hat{\sigma}^2 = S^2$ (или $\hat{\sigma} = S$). (Часто дополнительно сравнивают с несмещённой оценкой $\hat{\sigma}^2$.)

A.6.2. Равномерное распределение $U(a, b)$

Теория: $\mathrm{E}X = \dfrac{a+b}{2}$, $\mathrm{D}X = \dfrac{(b-a)^2}{12}$.

Метод моментов:

$$\hat{a} = \overline{x} - \sqrt{3S^2}, \qquad \hat{b} = \overline{x} + \sqrt{3S^2}.$$

Метод максимального правдоподобия:

$$\hat{a} = \min\{x_1, \ldots, x_n\}, \qquad \hat{b} = \max\{x_1, \ldots, x_n\}.$$

A.6.3. Экспоненциальное распределение со сдвигом $\mathrm{Exp}_{\lambda,c}$

Под $\mathrm{Exp}_{\lambda,c}$ будем понимать распределение с плотностью

$$f(x) = \lambda e^{-\lambda(x-c)}, \quad x \geq c, \qquad \lambda > 0.$$

Теория: $\mathrm{E}X = c + \dfrac{1}{\lambda}$, $\mathrm{D}X = \dfrac{1}{\lambda^2}$.

Метод моментов:

$$\hat{\lambda} = \frac{1}{S}, \qquad \hat{c} = \overline{x} - \frac{1}{\hat{\lambda}}.$$

Метод максимального правдоподобия:

$$\hat{c} = \min\{x_1, \ldots, x_n\}, \qquad \hat{\lambda} = \frac{1}{\overline{x} - \hat{c}}.$$

A.7. Моменты по сгруппированной выборке

Пусть диапазон значений выборки разбит на m интервалов, в k-й интервал попало n_k наблюдений, а $\hat{X}_k$ — середина k-го интервала. Тогда оценки вычисляются по формулам

$$\hat{X}_g = \frac{1}{n}\sum_{k=1}^{m} n_k \hat{X}_k, \qquad \hat{\sigma}_g^2 = \frac{1}{n-1}\sum_{k=1}^{m} n_k (\hat{X}_k - \hat{X}_g)^2.$$

A.8. Доверительные интервалы

A.8.1. Асимптотический ДИ для $\mathrm{E}X$

$$\mathrm{E}X \in \left(\bar{x} - z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \right).$$

A.8.2. Точные ДИ для параметров нормального распределения

Если $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, то

$$\frac{\overline{X} - a}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}, \qquad \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1},$$

где $\hat{\sigma}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Отсюда:

$$a \in \left(\overline{x} - t_{1-\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}},\ \overline{x} + t_{1-\alpha/2,n-1}\frac{\hat{\sigma}}{\sqrt{n}} \right),$$

$$\sigma^2 \in \left(\frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}},\ \frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-1}} \right).$$