Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

Лабораторная работа №4. Postgres

Выполнил:

Дощенников Никита

Группа: К3221

Проверила:

Александра Валерьевна Купера

Санкт-Петербург

2026

Содержание

Первичное описание выборки . . . 3
Вариационный ряд . . . 3
Эмпирическая функция распределения . . . 4
Гистограммы . . . 6
Числовые характеристики . . . 8
Описание формы распределений . . . 9
Гипотезы о виде закона распределения . . . 9
Оценивание параметров . . . 9
Метод моментов . . . 9
Нормальное распределение $N(a, \sigma^2)$ . . . 9
Равномерное распределение $U(a, b)$ . . . 9
Экспоненциальное распределение со сдвигом $\mathrm{Exp}_{\lambda,c}$ . . . 10
Результаты метода моментов . . . 10
Метод максимального правдоподобия . . . 10
Нормальное распределение . . . 10
Равномерное распределение . . . 10
Экспоненциальное распределение со сдвигом . . . 10
Результаты ММП . . . 11
Сравнение оценок двух методов . . . 11
Оценивание вероятности $P(X > x_0)$ . . . 11
Оценка моментов по сгруппированной выборке . . . 12
Доверительные интервалы . . . 12
Асимптотический доверительный интервал для EX . . . 12
Точные доверительные интервалы для нормального столбца . . . 13
Интерпретация доверительных интервалов . . . 13
Итоговый вывод . . . 14
Бонус: анализ столбца X_4 . . . 14
Первичное описание . . . 14
Кластеризация (k-средних, $k = 2$) . . . 14

Первичное описание выборки

В данном разделе выполняется первичный анализ столбцов X_1, X_2, X_3 из предоставленного CSV-файла объёмом $n = 200$ наблюдений.

Вариационный ряд

Вариационный ряд — упорядоченная по возрастанию последовательность наблюдений $x_1 \leq x_2 \leq ... \leq x_n$.

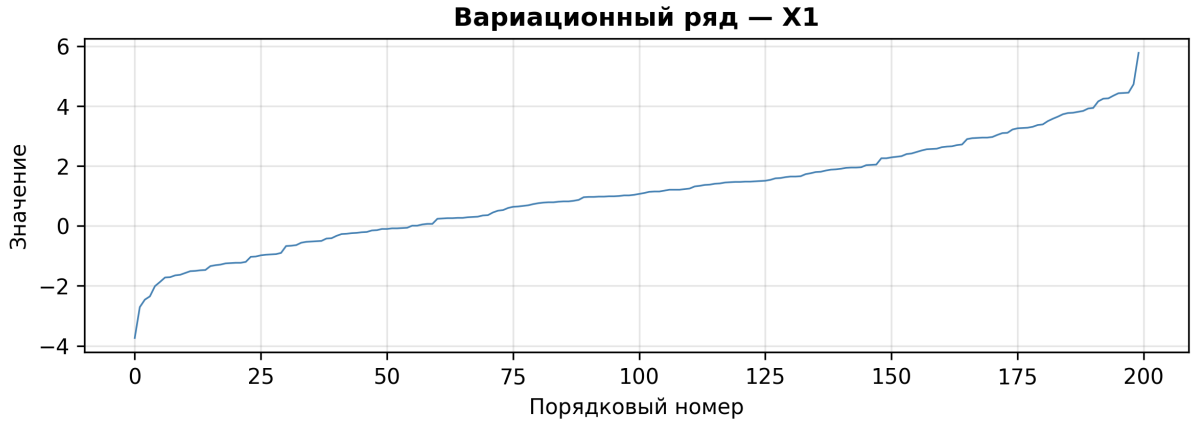

Рис. 1. Вариационный ряд столбца X_1

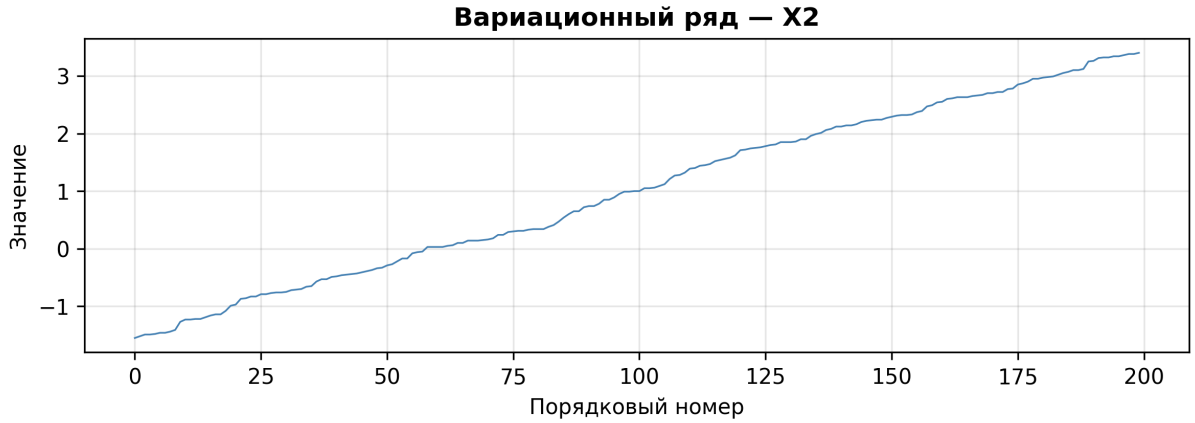

Рис. 2. Вариационный ряд столбца X_2

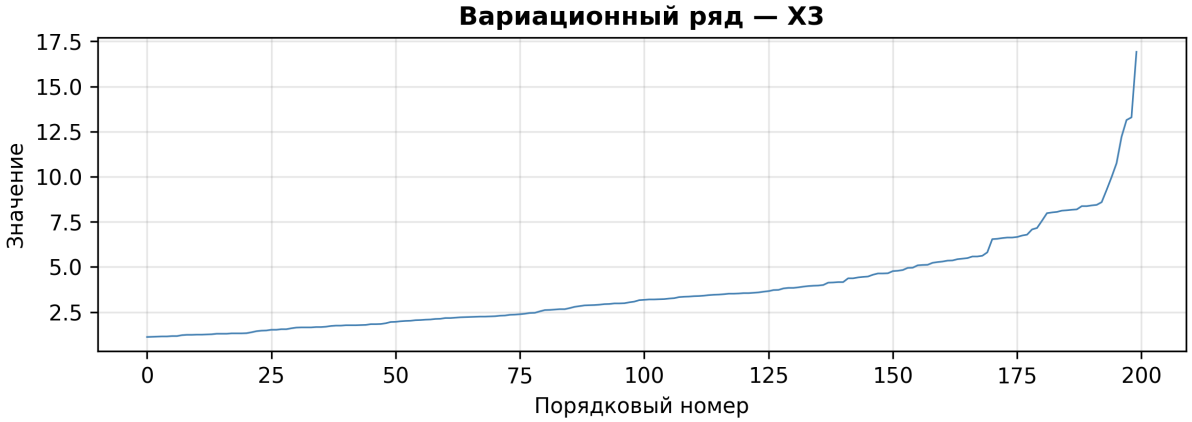

Рис. 3. Вариационный ряд столбца X_3

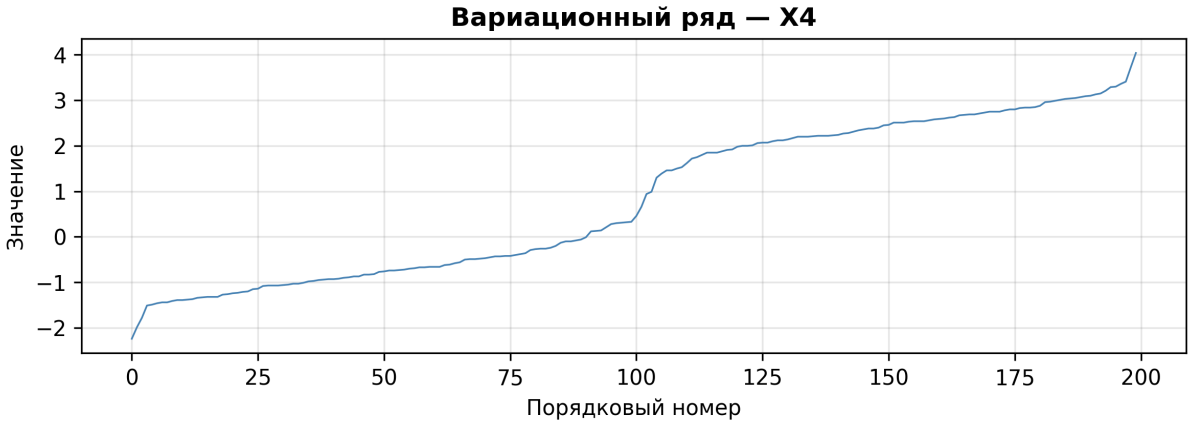

Рис. 4. Вариационный ряд столбца X_4

Эмпирическая функция распределения

Эмпирическая функция распределения определяется как

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

где $\nu_n(x)$ — число наблюдений, строго меньших x. График $F_n(x)$ является ступенчатой функцией: в каждой точке x_i происходит скачок на $\frac{1}{n}$.

Рис. 5. Эмпирическая функция распределения $F_n(x)$ для ряда X_1

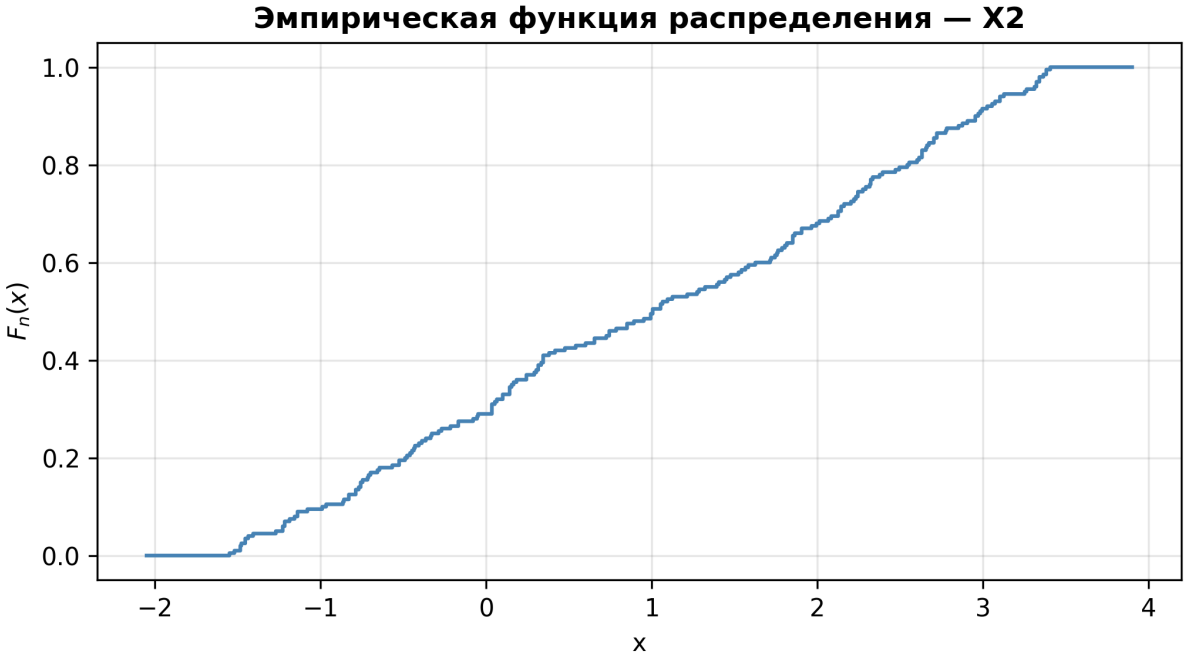

Рис. 6. Эмпирическая функция распределения $F_n(x)$ для ряда X_2

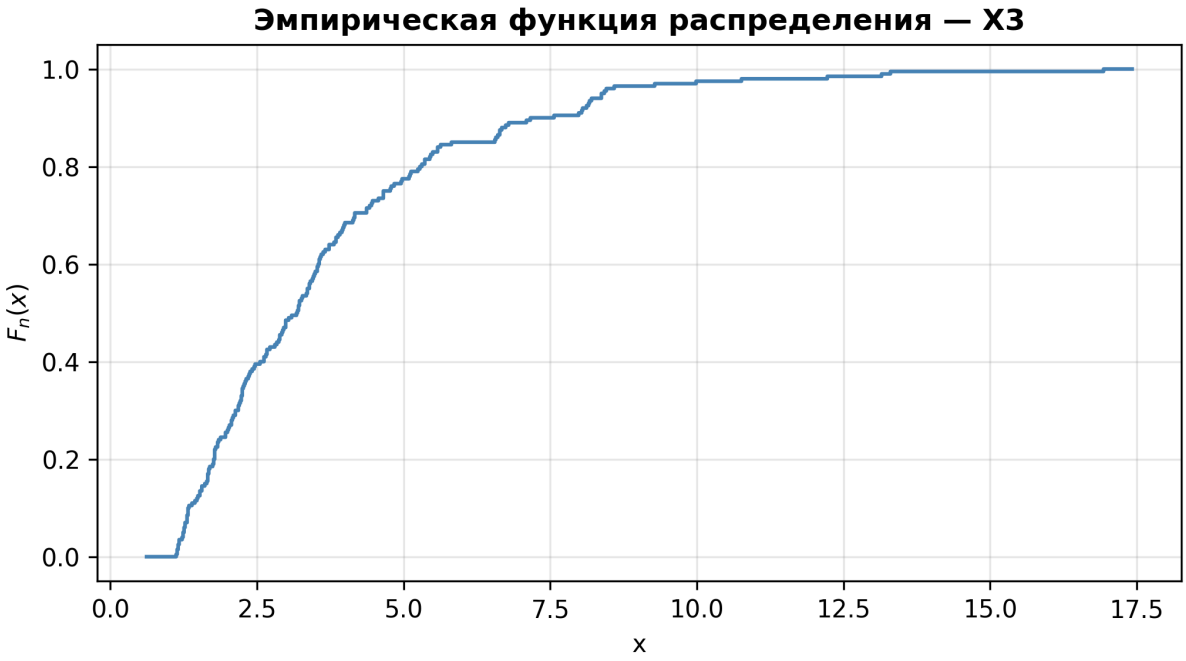

Рис. 7. Эмпирическая функция распределения $F_n(x)$ для ряда X_3

Рис. 8. Эмпирическая функция распределения $F_n(x)$ для ряда X_4

Гистограммы

Для выбора числа интервалов использовались три правила:

- **Правило Скотта:** $h = 3.5\hat{\sigma}n^{-\frac{1}{3}}$, откуда $k = \lceil \frac{x_{\max} - x_{\min}}{h} \rceil$.
- **Правило Фридмана–Диакониса:** $h = 2 \cdot \mathrm{IQR} \cdot n^{-\frac{1}{3}}$, $\mathrm{IQR} = q_{0.75} - q_{0.25}$.

- **Правило Стерджеса:** $k = 1 + \lfloor \log_2 n \rfloor$.

Результаты для $n = 200$:

Правило	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
Скотт	10	6	11
Фридман–Диакон.	12	6	17
Стерджес	8	8	8

Таблица 1. Число интервалов k по разным правилам

В дальнейшем используется правило Скотта как наиболее устойчивое к форме распределения.

[Вставить гистограммы X1 (три правила рядом)]

Рис. 9. Гистограммы X_1 по трём правилам

[Вставить гистограммы X2 (три правила рядом)]

Рис. 10. Гистограммы X_2 по трём правилам

[Вставить гистограммы X3 (три правила рядом)]

Рис. 11. Гистограммы X_3 по трём правилам

Числовые характеристики

Выборочное среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Смещённая и несмещённая дисперсии:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Характеристика	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
$\bar{x}$ — выборочное среднее	_____	_____	_____
S^2 — смещённая дисперсия	_____	_____	_____
$\hat{\sigma}^2$ — несмещённая дисперсия	_____	_____	_____
S — смещённое ст. откл.	_____	_____	_____
$\hat{\sigma}$ — несмещённое ст. откл.	_____	_____	_____
m_e — медиана	_____	_____	_____
$q_{\{0.25\}}$ — первый квартиль	_____	_____	_____
$q_{\{0.75\}}$ — третий квартиль	_____	_____	_____
$\mathrm{IQR} = q_{\{0.75\}} - q_{\{0.25\}}$	_____	_____	_____
$x_{\min}$	_____	_____	_____
$x_{\max}$	_____	_____	_____

Таблица 2. Числовые характеристики выборок

Описание формы распределений

Столбец X_1. *(Заполнить после анализа: симметрия/асимметрия, наличие выбросов, естественные границы значений, характер хвостов.)*

Столбец X_2. *(Заполнить после анализа.)*

Столбец X_3. *(Заполнить после анализа.)*

Гипотезы о виде закона распределения

На основании гистограмм, ЭФР и числовых характеристик для каждого столбца выдвигается гипотеза о законе распределения из списка: нормальное $N(a, \sigma^2)$, равномерное $U(a, b)$, экспоненциальное со сдвигом $\mathrm{Exp}_{\lambda,c}$.

Столбец X_1 — предполагаемый закон: _____.

(Обоснование: форма гистограммы, симметрия, соотношение среднего и медианы, характер ЭФР и т.д. — 2–6 предложений.)

Столбец X_2 — предполагаемый закон: _____.

(Обоснование.)

Столбец X_3 — предполагаемый закон: _____.

(Обоснование.)

Оценивание параметров

Метод моментов

Идея метода: приравнять теоретические моменты EX и DX к выборочным оценкам и решить систему уравнений.

Нормальное распределение $N(a, \sigma^2)$
Теория: $EX = a$, $DX = \sigma^2$. Оценки:

$$\hat{a} = \bar{x}, \quad \hat{\sigma}^2 = S^2.$$

Равномерное распределение $U(a, b)$
Теория: $EX = \frac{a+b}{2}$, $DX = \frac{(b-a)^2}{12}$. Оценки:

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

Экспоненциальное распределение со сдвигом $\mathrm{Exp}_{\lambda,c}$
Плотность: $f(x) = \lambda e^{\{-\lambda(x-c)\}}$, $x \geq c$, $\lambda > 0$.

Теория: $EX = c + \frac{1}{\lambda}$, $DX = \frac{1}{\lambda^2}$. Оценки:

$$\hat{\lambda} = \frac{1}{S}, \quad \hat{c} = \bar{x} - \frac{1}{\hat{\lambda}}.$$

Результаты метода моментов

Параметр	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
Параметр 1 (_____)	_____	_____	_____
Параметр 2 (_____)	_____	_____	_____

Таблица 3. Оценки параметров методом моментов

Метод максимального правдоподобия

Функция правдоподобия и логарифм правдоподобия:

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

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

Нормальное распределение

$$\ell(a, \sigma) = -n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - a)^2 + \text{const}.$$

Из условий первого порядка:

$$\hat{a}_{\text{МП}} = \bar{x}, \quad \hat{\sigma}^2_{\text{МП}} = S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Равномерное распределение

Плотность $f(x) = \frac{1}{b-a}$ при $x \in [a, b]$. Правдоподобие максимизируется при наименьшем возможном $(b - a)$, откуда:

$$\hat{a}_{\text{МП}} = \min\{x_1, ..., x_n\}, \quad \hat{b}_{\text{МП}} = \max\{x_1, ..., x_n\}.$$

Экспоненциальное распределение со сдвигом

$$\ell(\lambda, c) = n \ln \lambda - \lambda \sum_{i=1}^{n} (x_i - c).$$

Так как ℓ возрастает по c (при $c \le x_{\min}$):

$$\hat{c}_{\text{МП}} = \min\{x_1, ..., x_n\}, \quad \hat{\lambda}_{\text{МП}} = \frac{1}{\bar{x} - \hat{c}}.$$

Результаты ММП

Параметр	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
Параметр 1 (_____)	_____	_____	_____
Параметр 2 (_____)	_____	_____	_____

Таблица 4. Оценки параметров методом максимального правдоподобия

Сравнение оценок двух методов

	X_1		X_2		X_3	
Параметр	**ММ**	**МП**	**ММ**	**МП**	**ММ**	**МП**
Параметр 1	___	___	___	___	___	___
Параметр 2	___	___	___	___	___	___

Таблица 5. Сравнение оценок методом моментов и ММП

(Комментарий: для нормального распределения оценки совпадают. Для равномерного и экспоненциального оценки различаются — пояснить почему, 2–4 предложения.)

Оценивание вероятности $P(X > x_0)$

В качестве порога выбирается $x_0 = \bar{x} + \hat{\sigma}$.

Эмпирическая оценка:

$$\hat{p}_{\text{эмп}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[x_i > x_0].$$

Параметрическая оценка — подстановка найденных оценок в теоретическую формулу:

- Для $N(a, \sigma^2)$: $P(X > x_0) = 1 - \Phi(\frac{x_0 - \hat{a}}{\hat{\sigma}})$.
- Для $U(a, b)$: $P(X > x_0) = \frac{\hat{b} - x_0}{\hat{b} - \hat{a}}$ при $x_0 \in \left[\hat{a}, \hat{b}\right]$.
- Для $\text{Exp}_{\lambda,c}$: $P(X > x_0) = e^{\left\{-\hat{\lambda}(x_0 - \hat{c})\right\}}$.

Оценка	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
$x_0 = \bar{x} + \hat{\sigma}$	_____	_____	_____
$\hat{p}_{\text{эмп}}$	_____	_____	_____
$\hat{p}_{\text{пар}}$	_____	_____	_____
Расхождение $\lvert\hat{p}_{\text{эмп}} - \hat{p}_{\text{пар}}\rvert$	_____	_____	_____

Таблица 6. Сравнение эмпирической и параметрической оценок вероятности

(Комментарий к расхождению — 2–3 предложения.)

Оценка моментов по сгруппированной выборке

Пусть гистограмма содержит m интервалов, в k-й интервал попало n_k наблюдений, $\hat{X}_k$ — середина k-го интервала. Тогда:

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

Характеристика	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
$\hat{X}_g$ (по сгруппированной)	_____	_____	_____
$\bar{x}$ (по исходным)	_____	_____	_____
$\hat{\sigma}_g^2$ (по сгруппированной)	_____	_____	_____
$\hat{\sigma}^2$ (по исходным)	_____	_____	_____

Таблица 7. Сравнение оценок по сгруппированной и исходной выборкам

(Комментарий: потеря точности при группировке обусловлена заменой каждого наблюдения серединой интервала — 2–3 предложения.)

Доверительные интервалы

Уровень доверия $1 - \alpha = 0.95$, то есть $\alpha = 0.05$.

Асимптотический доверительный интервал для EX

По центральной предельной теореме для всех трёх столбцов:

$$EX \in \left(\bar{x} - z_{\{1-\frac{\alpha}{2}\}}\frac{\hat{\sigma}}{\sqrt{n}},\ \bar{x} + z_{\{1-\frac{\alpha}{2}\}}\frac{\hat{\sigma}}{\sqrt{n}}\right),$$

где $z_{\{0.975\}} \approx 1.960$.

Граница	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\boldsymbol{X_3}$
Нижняя граница	_____	_____	_____
Верхняя граница	_____	_____	_____
Ширина интервала	_____	_____	_____

Таблица 8. Асимптотические ДИ для EX на уровне 0.95

Точные доверительные интервалы для нормального столбца

Для столбца $X\_\_$ (отнесённого к $N(\mu, \sigma^2)$):

ДИ для μ при неизвестной σ^2 (распределение Стьюдента):

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

где $t_{\{0.975,199\}} \approx$ ____.

ДИ для σ^2 (распределение χ^2):

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

Интервал	**Значение**
ДИ для μ (асимптотический)	(_____, _____)
ДИ для μ (точный, t-распределение)	(_____, _____)
ДИ для σ^2	(_____, _____)

Таблица 9. Точные ДИ для параметров нормального распределения ($X\_\_$)

Интерпретация доверительных интервалов

Доверительный интервал с уровнем $1 - \alpha = 0.95$ означает следующее: если бы мы многократно повторяли эксперимент и строили интервал по каждой новой выборке, то примерно 95% таких интервалов покрывали бы истинное значение параметра. Это **не** означает, что истинный параметр попадает в данный конкретный интервал с вероятностью 95% — параметр либо лежит в интервале, либо нет. Вероятностный смысл относится к процедуре построения, а не к конкретному результату.

Чем уже интервал, тем точнее оценка. Ширина убывает как $O\left(\frac{1}{\sqrt{n}}\right)$, поэтому для вдвое более узкого интервала требуется вчетверо большая выборка.

Итоговый вывод

(Заполнить после получения числовых результатов — 5–12 строк.)

По результатам анализа установлено следующее соответствие столбцов и законов распределения:

- X_1 — _____ с параметрами _____;
- X_2 — _____ с параметрами _____;
- X_3 — _____ с параметрами _____.

Оценки методом моментов и методом максимального правдоподобия *(совпали / незначительно расходятся)*. Доверительные интервалы для среднего имеют ширину порядка _____, что свидетельствует о *(высокой / умеренной)* точности оценок при $n = 200$. Асимптотический и точный интервалы для нормального столбца практически совпадают, что подтверждает применимость ЦПТ при данном объёме выборки.

Бонус: анализ столбца X_4

Первичное описание

[Вставить гистограмму и ЭФР для X4]

Рис. 12. Гистограмма и ЭФР столбца X_4

(Описание формы: признаки бимодальности, неоднородности, наличие двух мод.)

Кластеризация (k-средних, $k = 2$)

(Описание метода и результат разбиения на два кластера.)

Характеристика	**Кластер 1**	**Кластер 2**
Объём n_j	_____	_____
Среднее $\bar{x}_j$	_____	_____
Ст. откл. $\hat{\sigma}_j$	_____	_____
Мин / Макс	_____ / _____	_____ / _____

Таблица 10. Характеристики кластеров столбца X_4

«Общее среднее» $\bar{x}$ плохо описывает смесь двух режимов, поскольку является взвешенным средним двух различных распределений и может не соответствовать ни одному из них. Например, если моды расположены на расстоянии d друг от друга, общее среднее окажется между ними в области низкой плотности.