Практическая работа №1. Первичная обработка данных

Юрченков Иван Александрович, ассистент кафедры ПМ

2022-10-22

Введение

Для прикладных задач математической статистики базовым элементом анализа статистических данных является процедура первичного исследования имеющихся выборок. В процедуру первичного исследования входят такие техники, как построение вариационных таблиц для оценки эмпирического распределения данных, оценка выборочных статистик, оценка вариабельности данных и построение графиков гистограмм выборочных переменных для визуального анализа на принадлежность выборки к одному из теоретических распределений, на наличие выбросов и характерных паттернов. Данные задачи решаются не в отрыве от устоявшейся процедуры статистического анализа данных, а наборот являются базой для проведения других более сложных статистических процедур, позволяющих нам получать ответы на важные вопросы на основе данных.

Современные задачи статистики выполняются, в основном, с применением компьютерных вычислительных средств, языков программирования и статистических пакетов для выполнения сложных процедур обработки данных над всё большим их числом. Тенденция к росту объема накопленной статистической информации подогревает интерес к высокопроизводительным компьютерным вычислениям, так что без современных инструментов базовые задачи анализа выборки чаще всего уже не решаются.

Однако при большом засилии статистических программных продуктов и сред разработки конвейеров обработки больших данных, за которыми скрыты преднаписанные алгоритмы, концепции, приемы и техники проведения первичного анализа данных важны для изучения и полного освоения будущими специалистами данной области. Получение полного испчерпывающего знания о практических приложениях обработки статистических данных и способов расчета и проведения статистических процедур позволяет двигаться дальше в направлении изучения науки о данных с большей скоростью и пониманием за счет взаимосвязи большинства методов и концепций на более высоком уровне.

Цель работы

Данная практическая работа ставит перед собой задачу познакомить с процессом вычисления стандартных описательных статистик для выборок данных раазличных типов и научить использовать методы агрегации статистических данных с целью получения новых знаний об их эмпирическом распределении.

Постановка задачи

Для выполнения задачи 1 раздела необходимо разбиться на две подгруппы.

Студенты первой подгруппы должны собрать со всей учебной группы данные о росте в сантиметрах и данные о месяцах рождения.

Студенты второй подгруппы должны собрать со всей учебной группы данные о росте в сантиметрах и данные о загаданном случайном целом числе на интервале $[0; 8]$.

Для целочисленных данных необходимо:

1. Построить вариационный ряд распределения абсолютных и относительных частот появления событий по выборке дискретных данных.
2. Построить полигон относительных частот для событий вариационного ряда.
3. Вычислить эмпирическую функцию распределения и построить её график.
4. Рассчитать выборочные описательные статистики:

- среднее $\overline{x}$;
- математическое ожидание μ_x;
- дисперсию D_x;
- стандартное отклонение σ_x;
- среднеквадратическое отклонение $\widehat{\sigma}_x$;
- медиану зафиксированной выборки m_x;
- первый и третий квартиль $\tau_{x,\,0.25}$, $\tau_{x,\,0.75}$;
- межквартильный размах IQR_x;
- коэффициент вариации ν_x.

Для вещественных данных необходимо:

1. Рассчитать число групп g, необходимых для квантования исходных данных по правилу Стёрджесса.
2. Вычислить значения границ групп $Z_i, i = 0, 1, ..., g$ для значений выборки по правилу фиксированной величины интервала.
3. Построить вариационный ряд для выборки интервальных данных.
4. Построить гистограмму распределения относительных частот для рассчитанных интервалов выборки.
5. Вычислить эмпирическую функцию распределения и построить её график.
6. Рассчитать выборочные описательные статистики:

- среднее $\overline{x}$;
- математическое ожидание μ_x;
- дисперсию с использованием выборочного среднего D_x;
- стандартное отклонение σ_x;
- среднеквадратическое отклонение $\widehat{\sigma}_x$;
- медиану зафиксированной выборки m_x;
- первый и третий квартиль $\tau_{x,\,0.25}$, $\tau_{x,\,0.75}$;
- межквартильный размах IQR_x;
- коэффициент вариации ν_x.

Пример расчета

На рассмотрение выносится набор данных или выборка с двумя переменными, целочисленного X_1 и вещественного X_2 типа:

$$\begin{pmatrix} X_1 \\ x_{11} \\ x_{12} \\ x_{13} \\ \vdots \\ x_{1n} \end{pmatrix} \begin{pmatrix} X_2 \\ x_{21} \\ x_{22} \\ x_{23} \\ \vdots \\ x_{2n} \end{pmatrix}$$

Выборочное среднее целочисленных данных рассчитывается по соотношению:

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

Математическое ожидание целочисленных данных рассчитывается по определению математического ожидания дискретного ряда:

$$\mu_x = \sum_{i=1}^{g} x_i \cdot p_i.$$

Дисперсия целочисленных данных рассчитывается по определению дисперсии дискретного ряда с использованием выборочного среднего:

$$D_x = \sum_{i=1}^{g} (\zeta_i - \overline{x})^2 \cdot p_i.$$

Стандартное отклонение с использованием дисперсии:

$$\sigma_x = \sqrt{D_x} = \sqrt{\sum_{i=1}^{g} (\zeta_i - \overline{x})^2 \cdot p_i}.$$

Среднеквадратическое отклонение:

$$\widehat{\sigma}_x = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (\overline{x} - x_i)^2}.$$

Медиана выборки — серединное значение отсортированной выборки:

$$[x] = sort(x), \quad m_x = \begin{cases} [x]_{n/2}, & n \ - \text{ нечетное}, \\ \left([x]_{\lfloor n/2 \rfloor} + [x]_{\lceil n/2 \rceil}\right)/2, & n \ - \text{ четное}. \end{cases}$$

Вопросы на защиту практической работы