Первичная обработка данных

Цель работы

Познакомиться с процессом вычисления стандартных описательных статистик для выборок данных различных типов и научить использовать методы агрегации статистических данных с целью получения новых знаний об их эмпирическом распределении.

Разделиться на две максимально близкие по численности подгруппы.

Каждому студенту необходимо будет решить две подзадачи:

1. Собрать данные.
2. Обработать собранные данные (рассчитать статистики и построить диаграммы).

Студенты обеих подгрупп должны собрать данные о росте учащихся всей учебной группы в сантиметрах. Собранные данные рассматривать в вещественных.

1. Студенты первой подгруппы должны собрать номера месяцев рождения со всей подгруппы. Собранные данные рассматривать, как целочисленные.
2. Студенты второй подгруппы должны собрать данные о загаданном случайном целом числе на интервале [0; 8]. Необходимо провести сбор данных так, чтобы опрашиваемые студенты не знали, какие числа загадали их одногруппники. Собранные данные рассматривать как целочисленные.

Для целочисленных данных необходимо:

- Построить вариационный ряд с абсолютными и относительными частотами по выборке дискретных данных

- Построить полигон относительных частот вариационного ряда

- Выписать выражение для эмпирической функции распределения и построить её график

- Рассчитать выборочные описательные статистики:
    - ❖ выборочное среднее
    - ❖ выборочную дисперсию
    - ❖ выборочное стандартное отклонение
    - ❖ выборочную медиану
    - ❖ коэффициент вариации

Для вещественных (непрерывных) данных необходимо:

- Рассчитать число групп (интервалов) m для квантования исходных данных по правилу Стёрджесса
- Вычислить значения $m+1$ границ групп для значений выборки по правилу фиксированной величины интервала
- Построить вариационный ряд для выборки интервальных данных
- Построить гистограмму распределения относительных частот для рассчитанных интервалов выборки
- Выписать выражение для эмпирической функции распределения, построить её график
- Рассчитать выборочные описательные статистики:
    - ❖выборочное среднее
    - ❖выборочную дисперсию
    - ❖выборочное стандартное отклонение
    - ❖выборочную медиану
    - ❖коэффициент вариации

- выборочное среднее

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

- выборочная дисперсия

$$D_x = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

- выборочное стандартное отклонение

$$\sigma_x = \sqrt{D_x}$$

- выборочная медиана

Для нечетного количества: центральный элемент в отсортированном массиве

Для четного количества элементов: среднее двух центральных элементов в отсортированном массиве

- коэффициент вариации

$$v_x = \frac{\sigma_x}{\overline{x}}$$

Чем больше его величина, тем больше разброс значений вокруг средней, тем менее однородна совокупность по своему составу
Если меньше 10%, то выборка компактна

[10, 4, 10, 10, 9, 9, 3, 9, 10, 2, 3, 2, 2, 1, 4]

- Вариационный ряд

[1, 2, 2, 2, 3, 3, 4, 4, 9, 9, 9, 10, 10, 10, 10]

Интервальный(непрерывные значения)

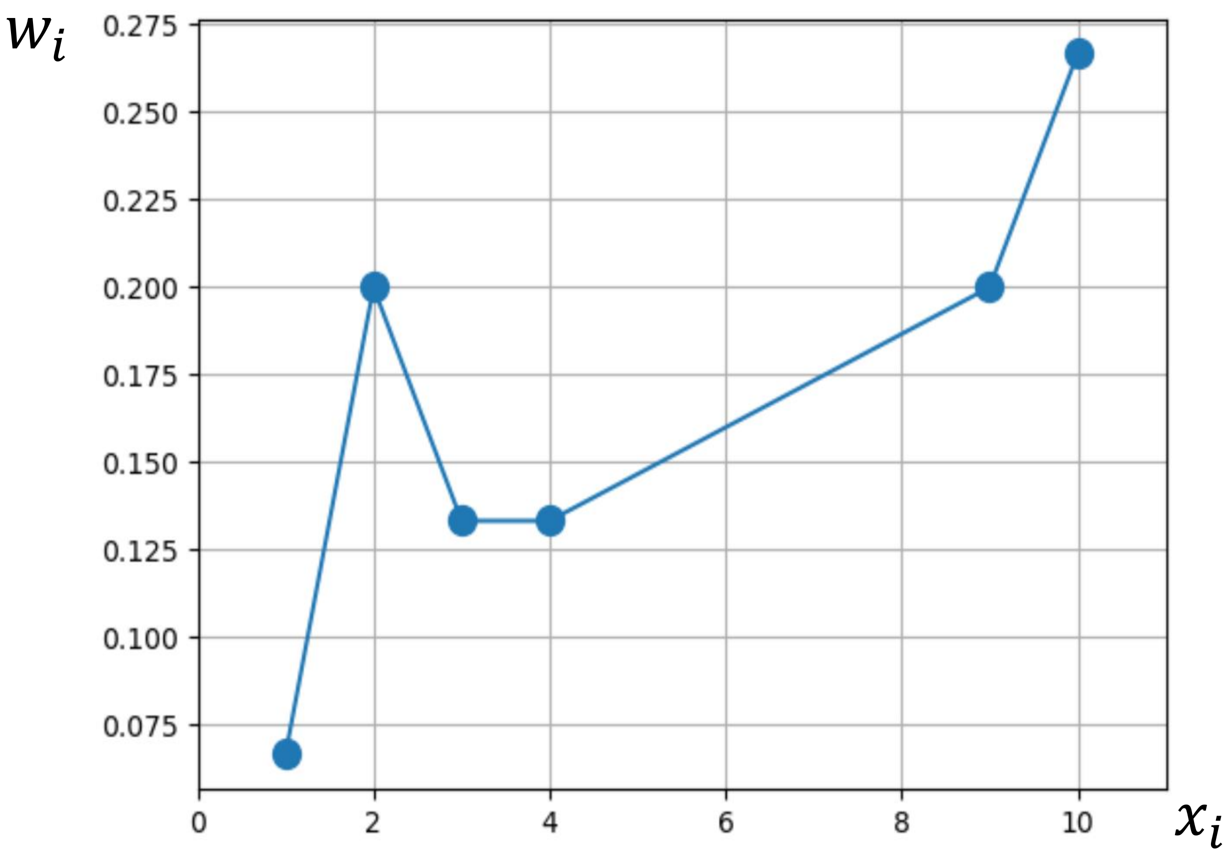
Дискретный(целые повторяющееся значения)

- Статистический ряд

	x_i	1	2	3	4	9	10
Абсолютная частота	n_i	1	3	2	2	3	4
Относительная частота	w_i	0.07	0.2	0.13	0.13	0.2	0.27

$$w_i = \frac{n_i}{n - \text{длина выборки}}$$

Для дискретного ряда строится полигон частот



x_i	1	2	3	4	9	10
n_i	1	3	2	2	3	4
w_i	0.07	0.2	0.13	0.13	0.2	0.27

Правило Стёрджеса

В случае непрерывного вариационного ряда составляют интервальный статистический ряд, под которым понимают упорядоченную совокупность интервалов значений случайной величины с соответствующими частотами или частостями (относительными частотами) попаданий в каждый из них значений случайной величины.

Формула Стерджеса используется для поиска оптимального количества интервалов

$$m = 1 + 3.332 \cdot \log_{10}(n), \mathrm{n} \text{ – длина выборки}$$

$$m = 1 + \log_2(n)$$

Если **m** окажется **дробным числом**, то за длину интервала принимается ближайшая целая величина.

$$\mathrm{h} = \frac{x_{max} - x_{min}}{m} \text{ – длина каждого интервала}$$

Длина интервалов должна быть такой, чтобы была возможность выявить характерные изменения случайной величины

x_{min}=156.0
x_{max} =188.0

Рекомендуется за начало первого интервала брать величину $x_{min} - \frac{h}{2}$

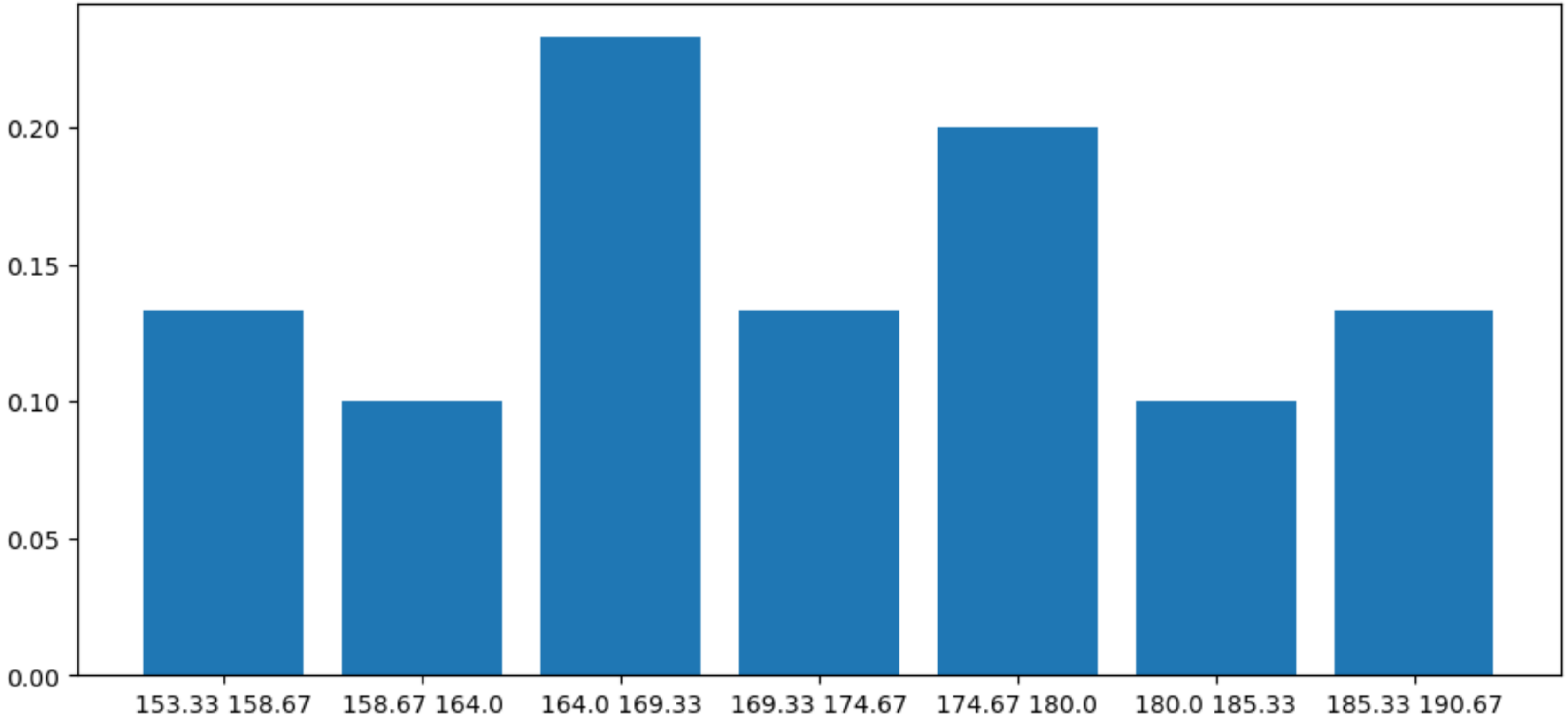
$$m = 1 + \log_2(31) = 5.95 \approx 6$$

$$\mathrm{h} = \frac{188.0 - 156.0}{6} = 5.33$$

	153.33	158.67	164.00	169.33	174.67	180.00	185.33
	158.67	164.00	169.33	174.67	180.00	185.33	190.67
ni	4	3	7	4	6	3	4
wi	0.13	0.10	0.23	0.13	0.20	0.10	0.13

Для интервального ряда строится гистограмма

	153.33	158.67	164.00	169.33	174.67	180.00	185.33
	158.67	164.00	169.33	174.67	180.00	185.33	190.67
ni	4	3	7	4	6	3	4
wi	0.13	0.10	0.23	0.13	0.20	0.10	0.13



Высоты в случае неравных интервалов равны плотности относительной частоты $\frac{w_i}{h_i}$.

Это необходимо сделать для устранения влияния величины интервала на распределение и иметь возможность сравнивать частоты.

Медиана и мода для интервального ряда

Найти интервал, в котором содержится медиана, путем подсчета накопленных частот или накопленных относительных частот. Медианным будет тот интервал, в котором накопленная частота впервые окажется больше $\frac{n}{2}$ или накопленная относительная частота – больше 0,5.

$$M = x_M + \frac{0.5n - n_{M-1}}{n_M} h_M$$

Мода – наиболее часто встречающееся значение в вариационном ряду

Найти интервал с наибольшей частотой (модальный интервал). Внутри модального интервала мода определяется по одной из следующих формул:

$$Mode = x_{Mode} + \frac{n_{Mode} - n_{Mode-1}}{2n_{Mode} - (n_{Mode-1} + n_{Mode+1})} h$$

Эмпирическая функция распределения (функция распределения выборки)

Статистический аналог интегральной функции распределения $F(x) = P(\mathrm{X} < x)$ случайной величины X в теории вероятностей – эмпирическая функция распределения.

Эмпирическая функция распределения $F^*(x)$ определяет для каждого значения x накопленную частость события $\mathrm{X} < x$. (случайная величина примет значение меньшее, чем x)

$$F^*(x) = \frac{n_x}{n}$$

n_x – число наблюдений, при которых наблюдается значение признака $\mathrm{X} < x$. Число n_x называется накопленной частотой, а отношение $\frac{n_x}{n}$ называется накопленной частостью.

x_i	1	5	8
n_i	10	15	25

$F^*(1) = 0$ – частость события, при котором случайная величина принимает значение меньше 1 равна 0, $\quad x < 1$

$F^*(5) = \frac{10}{50}$ – частость события, при котором случайная величина принимает значение меньше 5 равна 0.2, $\quad 1 \leq x < 5$

$F^*(8) = \frac{10 + 15}{50}$ – частость события, при котором случайная величина принимает значение меньше 8 равна 0.5, $\quad 5 \leq x < 8$

Эмпирическая функция распределения (функция распределения выборки)

$$F^*(x) = \frac{n_x}{n}$$

x_i	1	5	8
n_i	10	15	25

$F^*(1) = 0$ – частость события, при котором случайная величина принимает значение меньше 1 равна 0, $x < 1$

$F^*(5) = \frac{10}{50}$ – частость события, при котором случайная величина принимает значение меньше 5 равна 0.2, $1 \le x < 5$

$F^*(8) = \frac{10+15}{50}$ – частость события, при котором случайная величина принимает значение меньше 8 равна 0.5, $5 \le x < 8$

$$F^*(x) = \begin{cases} 0, & x < 1 \\ 0.2, & 1 \le x < 5 \\ 0.5, & 5 \le x < 8 \\ 1, & x \ge 8 \end{cases}$$