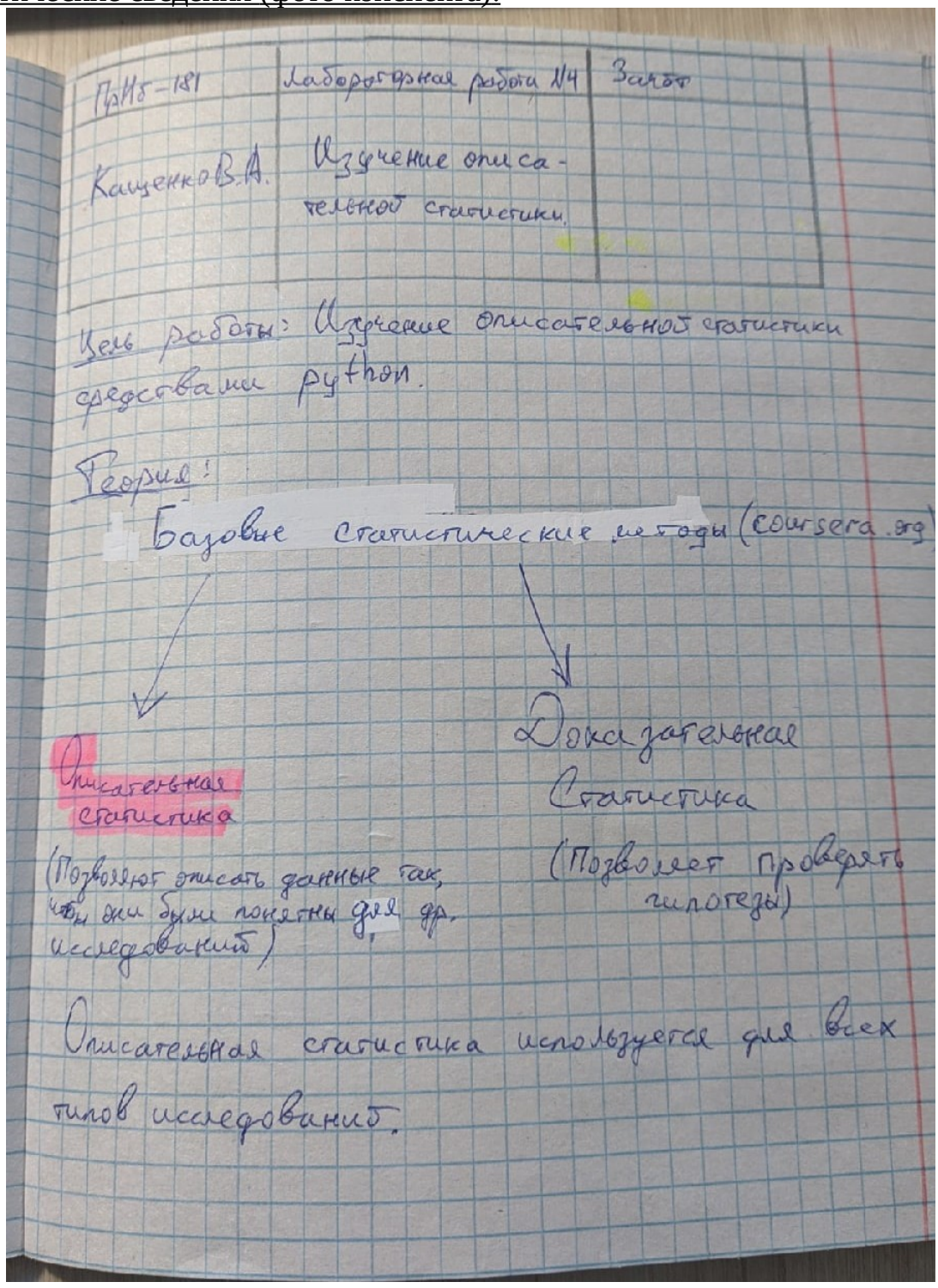
ПрИб-181	Лабораторная работа №4	Зачёт
Кащенко В. А.	Изучение описательной статистики	

Цель работы: Изучение описательной статистики средствами ЯП Python.

Теоретические сведения (фото конспекта):



Методы описательной статистики

Меры центральной тенденции

(для нахождения наиболее типичных значений переменной)

Меры:
- Мода (наиболее часто встреч. значение)
- Медиана (Середина упорядоч. ряда значений)
- Средн. арифметич (Сумма значен / кол-во)
  (не самая надежная мера)

(Если есть слишком большое значение, выброс, то он может смещать среднее)
(тогда либо выбрасывают значение, либо пользуются медианой)

Меры изменчивости

(на сколько сильно разбросаны значения от типичного)
(Чем больше мера изменчивости, тем более разнообразно будет значение по разным ху.)

Меры:
- размах (Макс. значение - мин. значен)
  (выбросы мешают)
- (убирают 25% самых больших знач. и 25% самых мелких)
  (Это - межквартильный размах)

- стандартное отклонение (диапазон от среднего нормальных разбросов) (~68%) (σ)
(если вычесть - нижн. граница нормы)
(если прибавить - верх. граница)

Меры позволяют кратко описать собранные данные

Описательная статистика (методика)

Раздел стат. науки, методы описания и представл. осн. свойств данных. Позволяет обобщить.

Этапы:

- сбор данных
- категоризация
- обобщение
- представление

Пусть $X_1, X_2 \dots X_n$ - выборка независ. случайных величин. Упорядочим по возрастанию (вариационный ряд)

$$X(1) < X(2) < \dots < X(n),$$

где $X(1)$ - min , $X(n)$ - max

Элементы вариац. ряда - порядков. статистики

Величины $d(i) = X(i+1) - X(i)$ называются спейсингами или расстояниями между порядковыми статистиками.

Размах выборки

$$R = X(n) - X(1) \quad (\max - \min)$$

Выборочное среднее:

$$\bar{X} = (X_1 + X_2 + \dots + X_n) / n$$

Среднее арифметическое:

Информативная мера "центрального положения" наблюдаемой переменной, если сообщается доверительный интервал (значения вокруг оценки, где с уровнем доверия, находится "истинное" (неизвестное) среднее.

Увеличение размера выборки делает оценку среднего более надёжной. (Увеличение разброса наблюд. значений уменьшает надёжность оценки.

Вычисление довер. интервалов основывается на предположении "нормальности" наблюдаемых величин. (Без него оценка может быть плохой, осо. для малых подборок.)

Нужно содержательно обобщать данные.

Диаграмма - отправн. точка. Можно стать инф-ции используя нужные характеристики данных.

Образ формул можно сформировать:

- зная из чего состоит величина
- зная рассеянность наблюдения.

формула:

$$\bar{X} = (X_1 + X_2 + \dots + X_n) / n$$

или

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Медиана:

При упорядочивании (min, max)

Делит ряд упоряд. значений пополам, с равным числом как выше, так и ниже.

При n - нечёт:

$$\frac{(n+1)}{2} \quad \left(\text{Для } n=11 \Rightarrow \frac{11+1}{2} = 6\right)$$

При n - чётн:

строго медианы нет. (средн. арифметич)

$n=20$ $\quad \frac{20}{2} = 10 \quad \frac{20}{2}+1 = 11$ (для упорядочен

Мода:

Мода M – наиболее частое значение

(если данные непрерывны – группируем и вычисляем модальную группу.

Не. наборы не имеют моды (∀ × только одна)

Иногда > 1:

Когда ≥ 2 значения встречаются одинаков. число раз и встречаются больше др. значений

Редко используют как общ. характеристику

Среднее арифметическое:

При несимметрич. распредел. не обобщающ. показатель

Взвешенное:

$$\frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n} = \frac{\sum w_i x_i}{\sum w_i}$$

Размах (интервал изменения):

max - min

При выбросах вводит заблуждение.

Квантили:

Предположим, что $X_0 \Rightarrow X_n$

Величина $X_{0,01}$ квантиль (до нее 1% наблюдений и выше 99%) (первый квантиль)

$X_{0,02}$ – второй квантиль и т. д.

Смысл: левее X_p лежит ~100p% наблюдений

$X_{0,1}, X_{0,2} \dots X_{0,9}$ – делят набор на 10 групп, децили (10-й, 20-й, ... 90-й квантили)

величины $X_{0,25}, X_{0,5}, \dots, X_{0,75}$ (деятка 4 группы) – квартили (25-й, 50-й, 75-й)

50-й квартиль – медиана.

Дисперсия – степень отклонения каждого наблюдения от средн. арифм.

$$D = \sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Если не генеральная совокупность, а выборка, то выборочная дисперсия

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Вариация в пределах субъектов и между ними. Объект не всегда дает точные и те же самые данные

Выброс

Отклоняющееся значение

Асимметрия

Мера асимметрии распределения вероятностей вещественной случайной величины.

Практика:

1) текст + конспект — ✓

2) ~~Ж~~ Я дисциплин — ✓ анализ — ✓

3) Построение гистограмм - ✓
График

Вывод: Изучена описательная статистика средствами python.

Постановка задачи:

Выберите не менее 9 дисциплин, которые читаются всему потоку (трем группам на одном курса разом). Каждую дисциплину для данного курса рассматривайте отдельно. Проведите анализ выбранных вами данных с использованием всех средств описательной статистики. Выполните построение соответствующих столбчатых диаграмм, поверх диаграмм должен располагаться график распределения, наиболее близко характеризующий вашу выборку.

Практическая часть (код программы):

```
# библиотеки
import seaborn as sns
import pandas as pd
import numpy as np
from scipy import stats

# функция создания подписей гистограмм описательной статистики
def set_up_statistic(marks, ax):
        # Среднее арифметическое
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.90,
        'Среднее арифметическое: {0}'.format(np.mean(marks)), fontsize=9)

        # Медиана
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.85,
        'Медиана: {0}'.format(np.median(marks)), fontsize=9)

        # Мода
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.80,
        'Мода: {0}'.format(stats.mode(marks)[0]), fontsize=9)

        # Среднее геометрическое
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.75,
        'Среднее геометрическое: {0}'.format(stats.hmean(marks)), fontsize=9)

        # Размах
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.70,
        'Размах: {0}'.format(np.ptp(marks)), fontsize=9)

        # Межквартальный размах
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.65,
        'Межквартальный размах: {0}'.format(stats.iqr(marks)), fontsize=9)

        # Межквантильный диапазон (Интердециальный размах)
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.60,
        'Интердециальный размах: {0}'.format(stats.iqr(marks, rng=(10, 90))),
        fontsize=9)

        # Дисперсия
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.55,
```

```
        'Дисперсия: {0}'.format(stats.variation(marks)), fontsize=9)

        # Среднеквадратичное отклонение
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.50,
        'Среднеквадратичное отклонение: {0}'.format(np.std(marks)), fontsize=9)

        # Коэффициент ассиметрии
        ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.45,
        'Коэффициент ассиметрии : {0}'.format(stats.skew(marks)), fontsize=9)

# Файл csv
data = pd.read_csv('marks_groups.csv')
# Ненужный столбец
data = data.drop('Группа', axis=1)
# Колонки
subjects = [column for column in data.columns][:]

# Создвние диаграмм для всех дисциплин
for subject in subjects:
        subject_data = data[[subject]]

        # Построение распределения
        g = sns.displot(subject_data, x=subject, binwidth=4, height=4,
        facet_kws=dict(margin_titles=True), kde=True, color = 'green')
        subject_marks = np.array([mark for mark in subject_data[subject]])

        # Настройка удобного отображения
        set_up_statistic(subject_marks, g.ax)

        # Вывод
        g.savefig('./diagrams/{0}.svg'.format(subject))
```

Этапы выполнения работы:

1. Подключаем необходимые библиотеки:

```
# библиотеки
import seaborn as sns
import pandas as pd
import numpy as np
from scipy import stats
```

2. Определим функцию для подписей графиков:

Это основная функция. Передаются значения интервалов для каждого отдельного вычисления, её подпись, размер шрифта и тип вычисления.

```
# функция создания подписей гистограмм описательной статистики
def set_up_statistic(marks, ax):
    # Среднее арифметическое
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.90,
    'Среднее арифметическое: {0}'.format(np.mean(marks)), fontsize=9)

    # Медиана
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.85,
    'Медиана: {0}'.format(np.median(marks)), fontsize=9)

    # Мода
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.80,
    'Мода: {0}'.format(stats.mode(marks)[0]), fontsize=9)

    # Среднее геометрическое
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.75,
    'Среднее геометрическое: {0}'.format(stats.hmean(marks)), fontsize=9)

    # Размах
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.70,
    'Размах: {0}'.format(np.ptp(marks)), fontsize=9)

    # Межквартальный размах
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.65,
    'Межквартальный размах: {0}'.format(stats.iqr(marks)), fontsize=9)

    # Межквантильный диапазон (Интердециальный размах)
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.60,
    'Интердециальный размах: {0}'.format(stats.iqr(marks, rng=(10, 90))),
    fontsize=9)

    # Дисперсия
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.55,
    'Дисперсия: {0}'.format(stats.variation(marks)), fontsize=9)

    # Среднеквадратичное отклонение
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.50,
    'Среднеквадратичное отклонение: {0}'.format(np.std(marks)), fontsize=9)
```

```
    # Коэффициент ассиметрии
    ax.text(ax.viewLim.intervalx.max(), ax.viewLim.intervaly.max() * 0.45,
    'Коэффициент ассиметрии : {0}'.format(stats.skew(marks)), fontsize=9)
```

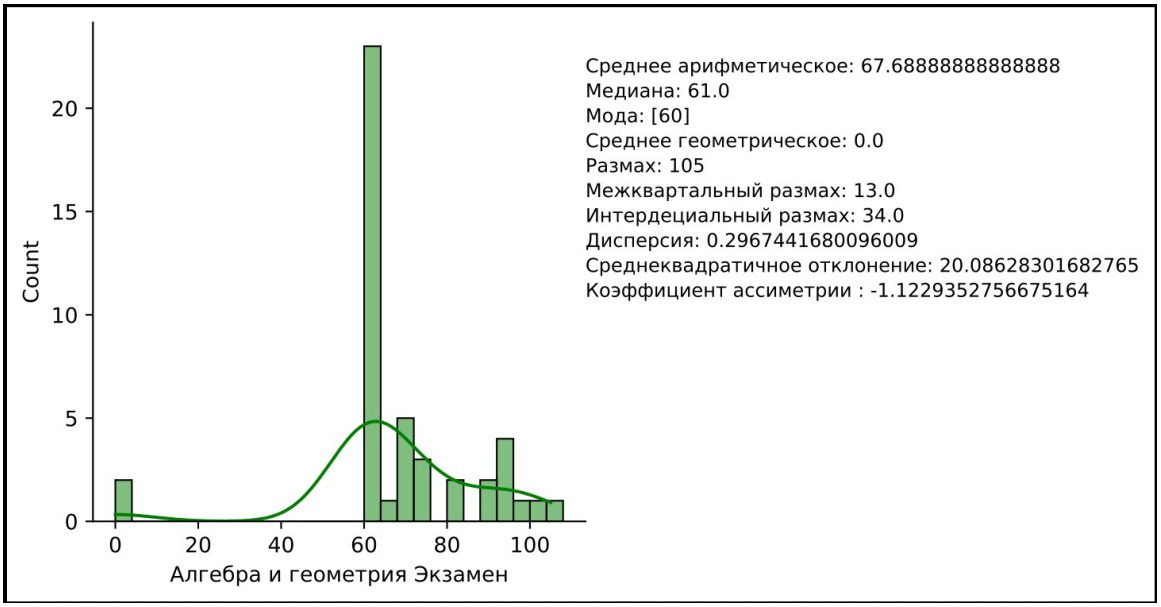
3. Загружаем и подготавливаем графики:

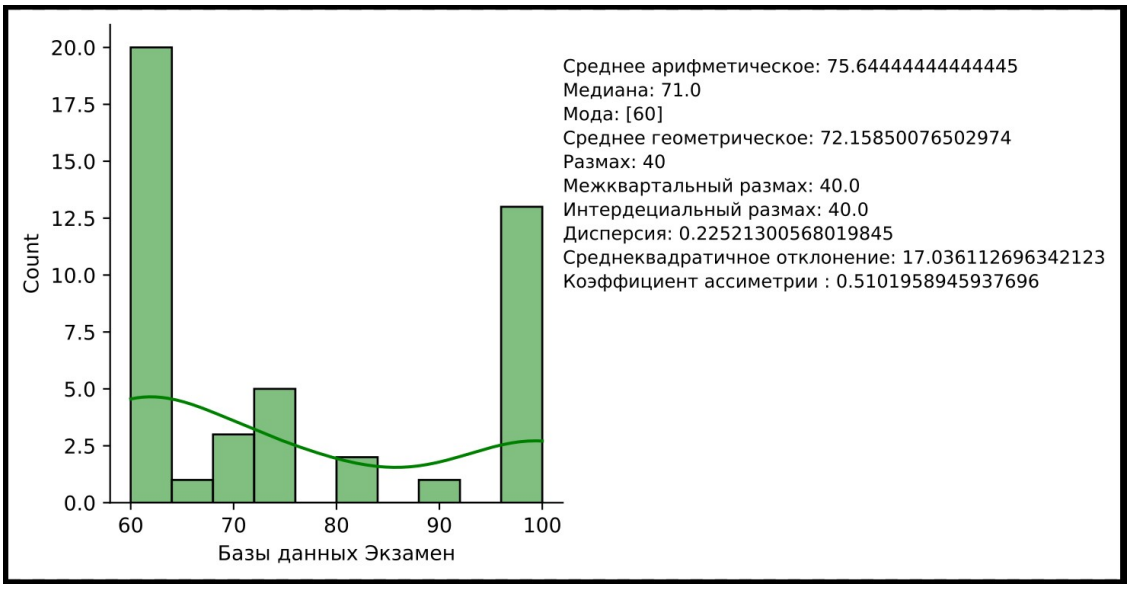
```
# Файл csv
data = pd.read_csv('marks_groups.csv')
# Ненужный столбец
data = data.drop('Группа', axis=1)
# Колонки
subjects = [column for column in data.columns][:]
```

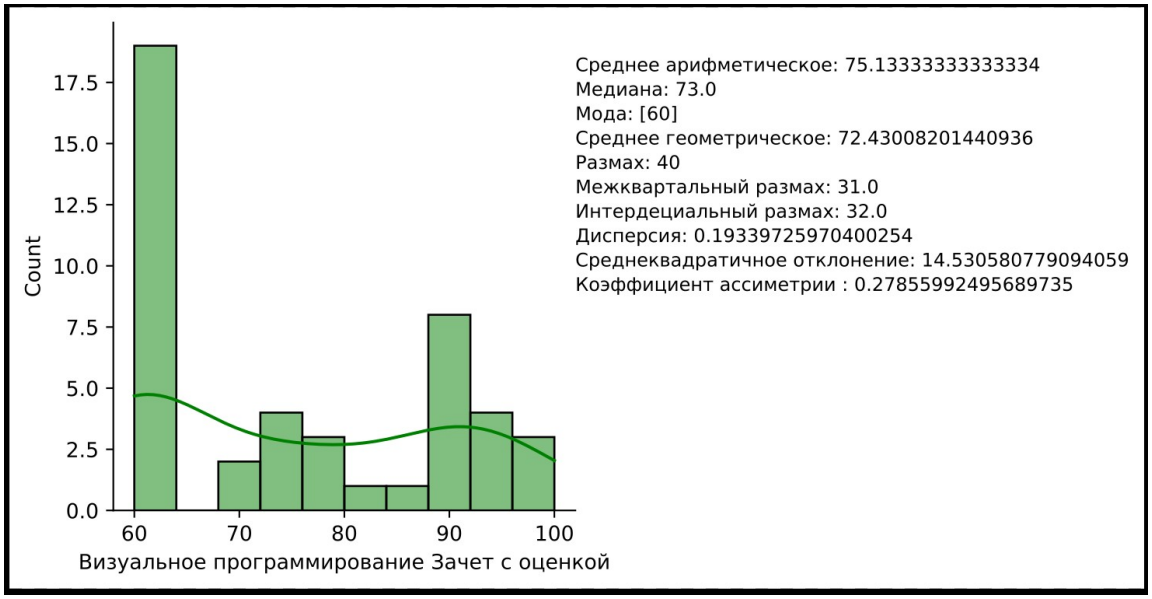
4. Проходим по всем колонкам и строим графики статистики, сохраняем их на диске.

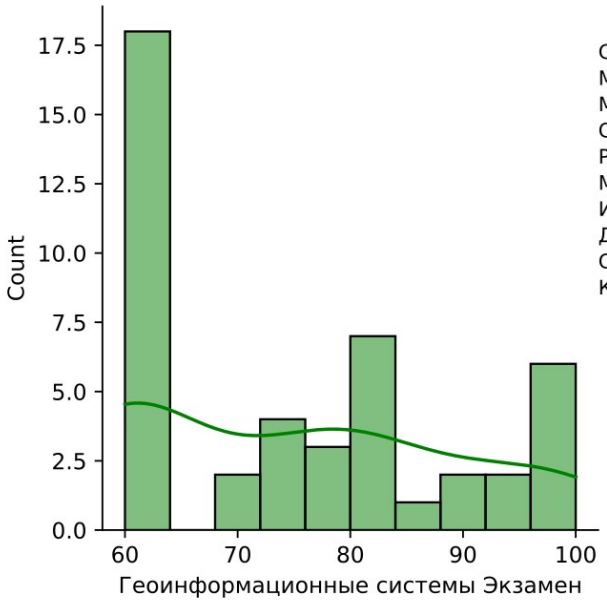
```
# Создание диаграмм для всех дисциплин
for subject in subjects:
    subject_data = data[[subject]]

    # Построение распределения
    g = sns.displot(subject_data, x=subject, binwidth=4, height=4,
    facet_kws=dict(margin_titles=True), kde=True, color = 'green')
    subject_marks = np.array([mark for mark in subject_data[subject]])

    # Настройка удобного отображения
    set_up_statistic(subject_marks, g.ax)

    # Вывод
    g.savefig('./diagrams/{0}.svg'.format(subject))
```

Практическая часть (результат подведения статистики):

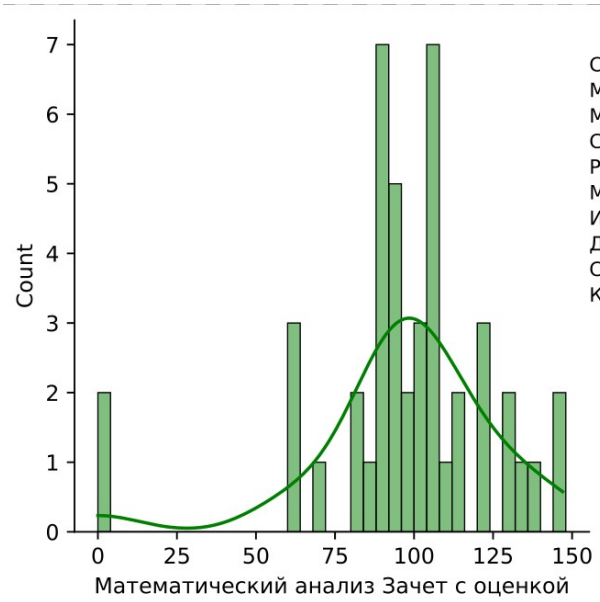




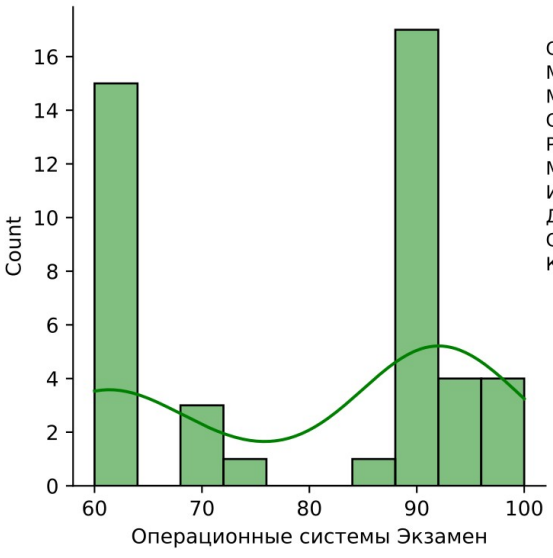




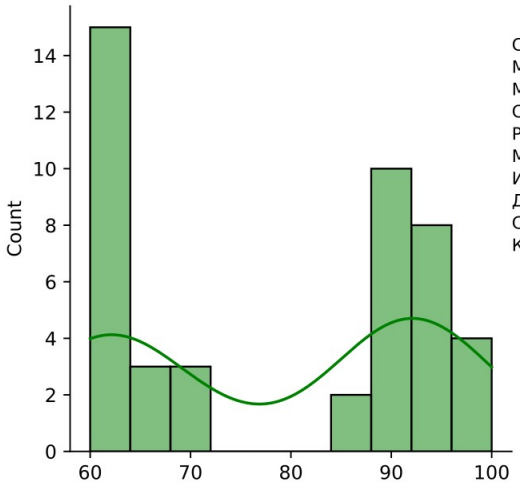
Среднее арифметическое: 74.91111111111111
Медиана: 75.0
Мода: [60]
Среднее геометрическое: 72.35377348814484
Размах: 40
Межквартальный размах: 23.0
Интердециальный размах: 38.0
Дисперсия: 0.18930551333407178
Среднеквадратичное отклонение: 14.181086343314577
Коэффициент ассиметрии : 0.37587349682423193



Среднее арифметическое: 96.86666666666666
Медиана: 97.0
Мода: [91]
Среднее геометрическое: 0.0
Размах: 147
Межквартальный размах: 17.0
Интердециальный размах: 64.19999999999999
Дисперсия: 0.2982003616817605
Среднеквадратичное отклонение: 28.88567503490653
Коэффициент ассиметрии : -1.3915638096645544

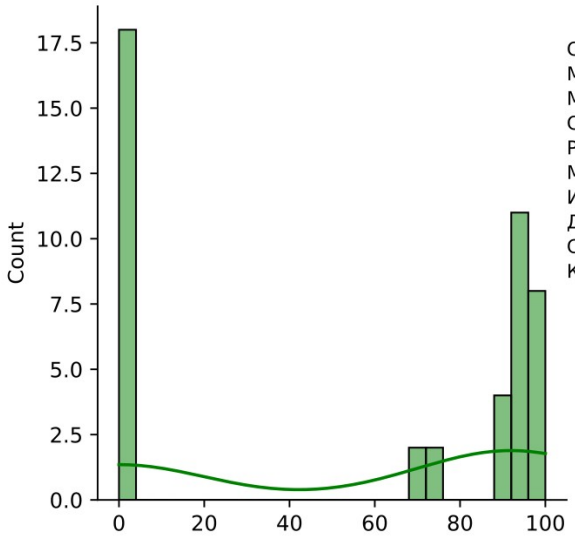


Среднее арифметическое: 80.02222222222223
Медиана: 91.0
Мода: [91]
Среднее геометрическое: 76.89333048197177
Размах: 40
Межквартальный размах: 30.0
Интердециальный размах: 35.0
Дисперсия: 0.19075103043539066
Среднеквадратичное отклонение: 15.264321346618706
Коэффициент ассиметрии : -0.31912060681374094



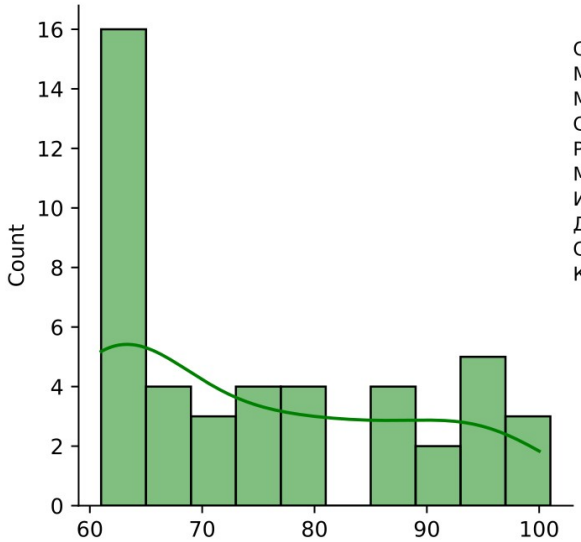
Среднее арифметическое: 78.62222222222222
Медиана: 85.0
Мода: [60]
Среднее геометрическое: 75.54729836263625
Размах: 40
Межквартальный размах: 30.0
Интердециальный размах: 33.0
Дисперсия: 0.19493857693235464
Среднеквадратичное отклонение: 15.326504115259349
Коэффициент ассиметрии : -0.1032811881076914

Производственная практика, научно-исследовательская работа Зачет с оценкой_x



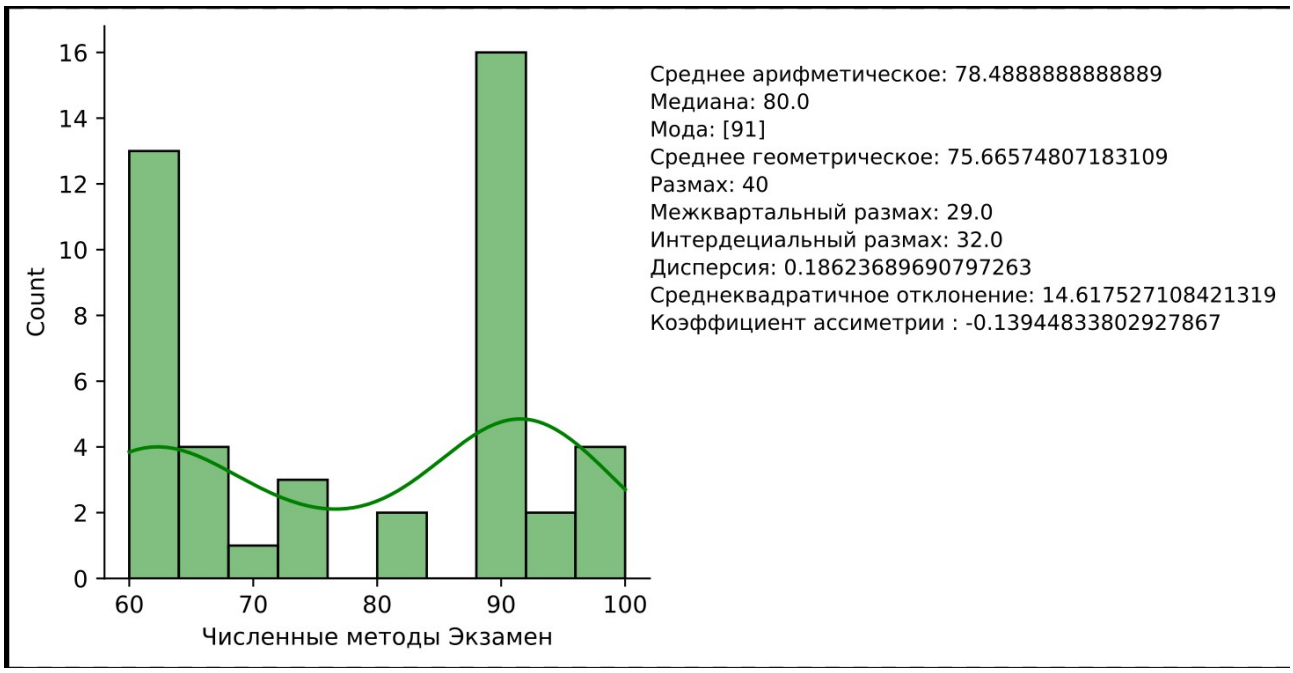
Среднее арифметическое: 54.577777777777776
Медиана: 91.0
Мода: [0]
Среднее геометрическое: 0.0
Размах: 100
Межквартальный размах: 93.0
Интердециальный размах: 97.2
Дисперсия: 0.8252709936546764
Среднеквадратичное отклонение: 45.04145689813078
Коэффициент ассиметрии : -0.34987421784185907

Производственная практика, научно-исследовательская работа Зачет с оценкой_y



Среднее арифметическое: 74.66666666666667
Медиана: 71.0
Мода: [61]
Среднее геометрическое: 72.43030268554797
Размах: 39
Межквартальный размах: 24.0
Интердециальный размах: 34.0
Дисперсия: 0.1790173004978877
Среднеквадратичное отклонение: 13.366625103842281
Коэффициент ассиметрии : 0.4779546133027651

Теория вероятностей и математическая статистика Зачет с оценкой



Практическая часть (подведение итогов):

Были проанализированы результаты сессий по 10 предметам. Данные визуализированы с помощью гистограмм и библиотек python для дальнейшей работы с ними. (описательная статистика)

Вывод: Изучена описательная статистика средствами Python.