Министерство науки и высшего образования Российской Федерации

ФГАОУ ВО «Северо-Восточный федеральный университет имени М.К. Аммосова»

Институт математики и информатики

Кафедра «информационные технологии»

ОТЧЕТ

По лабораторной работе №1

По теме: Анализ одномерных количественных данных

по дисциплине: Анализ данных

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Выполнил: студент 4 курса

группы ФИИТ-21 имени СВФУ

Никифоров А.Г.

Проверил: доцент кафедры

ИТ ИМИ СВФУ

Павлов Н.Н.

Якутск 2024

Цель работы

Освоить основные приемы статистической обработки одномерных количественных данных и применить их к анализу объема основных фондов предприятий.

Задачи работы

1. Построить интервальный вариационный ряд распределения объема основных фондов 100 предприятий.
2. Рассчитать основные числовые характеристики для несгруппированных данных и данных, сгруппированных по интервалам.
3. Построить гистограммы, полигоны и кумуляты частот и относительных частот.
4. Проанализировать полученные результаты и сделать выводы.

Ход работы

1. Исходные данные

Для анализа были предоставлены данные по объёмам основных фондов 100 однотипных предприятий (в миллионах рублей). Минимальное значение объёма фондов — 5,02 млн руб., максимальное — 5,85 млн руб.

Объём основных фондов									
5,56	5,27	5,02	5,47	5,27	5,37	5,47	5,47	5,33	5,11
5,33	5,47	5,33	5,33	5,47	5,05	5,33	5,85	5,68	5,11
5,54	5,43	5,64	5,21	5,68	5,43	5,79	5,47	5,21	5,47
5,43	5,43	5,47	5,27	5,68	5,43	5,47	5,79	5,47	5,54
5,43	5,43	5,61	5,47	5,27	5,54	5,61	5,54	5,64	5,54
5,64	5,43	5,33	5,11	5,33	5,33	5,33	5,54	5,64	5,64
5,4	5,68	5,43	5,54	5,43	5,37	5,37	5,21	5,64	5,64
5,71	5,47	5,21	5,33	5,43	5,33	5,43	5,27	5,21	5,54
5,79	5,58	5,27	5,33	5,4	5,43	5,54	5,54	5,54	5,81
5,39	5,47	5,47	5,27	5,58	5,43	5,43	5,33	5,61	5,54

Минимальное значение: 5.02 млн руб.

Максимальное значение: 5.85 млн руб.

2. Определение размаха варьирования и длины интервала

- Размах варьирования (R) определяется как разница между максимальным и минимальным значениями:

$R = 5{,}85 - 5{,}02 = 0{,}83$ млн руб.

- Длину интервала (h) рассчитали по формуле Стерджеса:

$$h = \frac{R}{k}$$

$$k = 1 + 3{,}322 \cdot log10n \approx 7{,}644$$

Полученная длина интервала составила:

$$h = \frac{0{,}83}{7{,}644} \approx 0{,}1086 \text{ млн руб.}$$

3. Построение интервального вариационного ряда

Были выделены интервалы с шагом, равным 0,1086, и рассчитаны частоты попадания значений в каждый из интервалов, накопленные частоты, относительные и накопленные относительные частоты. Результаты представлены в таблице ниже:

a i	b i	Частота m i	Накопленная частота m iн	Относительная частота w i =m i /n	Относительная накопленная частота w i =m i /n	w i /h
4,965709053	5,074290947	0	0	0	0	0
5,074290947	5,182872841	2	2	0,02	0,02	0,184193
5,182872841	5,291454736	3	5	0,03	0,05	0,276289
5,291454736	5,40003663	12	17	0,12	0,17	1,105157
5,40003663	5,508618524	19	36	0,19	0,36	1,749831
5,508618524	5,617200419	29	65	0,29	0,65	2,670795
5,617200419	5,725782313	18	83	0,18	0,83	1,657735
5,725782313	5,834364207	12	95	0,12	0,95	1,105157
5,834364207	5,942946102	4	99	0,04	0,99	0,368386
5,942946102	6,051527996	1	100	0,01	1	0,092096

min=	5,02
max=	5,85
размах варьирования признаков	0,83
Длина интервала:	0,1085819
число интервалов	7,644

4. Построение гистограммы и полигона

На основе данных построим гистограммы частот и относительных частот. Гистограмма частот представляет распределение данных по интервалам, а полигон частот показывает изменение частот между интервалами.

5. Построение кумуляты накопленных относительных частот

Также была построена кумулята накопленных относительных частот, которая наглядно демонстрирует накопление частот по мере увеличения значений объёмов фондов.

Арифметическое среднее	5,4493
Гармоническое среднее	5,444106019
Геометрическое среднее	5,446706
медиана	5,43
Мода	5,43

моменты		
порядок 1	5,4493	
порядок 2	0,028497485	
порядок 3	-0,043549993	(это ассимметрия)
порядок 4	0,044753119	
Дисперсия	0,028497485	
сред.квадр	0,168811981	
Коэффициент асимм	-0,043549993	
Коэффициент эксцес	0,044753119	
Коэффициент вариац	0,030978654	

Числовые характеристики:

- Дисперсия показывает, насколько далеко данные (объёмы основных фондов) разбросаны относительно среднего значения. Чем больше дисперсия, тем сильнее данные отличаются от среднего.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Среднее квадратическое отклонение (корень из дисперсии) — это ещё один способ показать разброс данных, но в тех же единицах, что и сами данные (в данном случае — в миллионах рублей).
- Коэффициент вариации показывает относительный разброс данных в процентах от среднего значения, что даёт представление о том, насколько сильно данные сконцентрированы вокруг среднего.
- Асимметрия говорит о том, насколько распределение данных смещено влево или вправо. Если асимметрия отрицательная (например, по текущим данным вышло -0,0435), это значит, что больше значений находится немного выше среднего.
- Эксцесс описывает "остроту" распределения. Если коэффициент эксцесса близок к 0 (как в нашем случае, 0,0447), распределение данных почти нормальное, но с небольшими отклонениями.

Анализ результатов

Вычисление основных числовых характеристик

Для несгруппированных данных:

1. **Арифметическое среднее: 5.4691**

Арифметическое среднее — это основная мера центральной тенденции, которая показывает среднее значение всех объёмов фондов в исследуемой выборке. Значение 5.4691 означает, что в среднем объёмы фондов предприятий находятся в диапазоне около 5.47. Это значение важно, поскольку оно демонстрирует общую тенденцию в распределении фондов. То есть, большинство фондов предприятия сосредоточены около этого среднего значения. На гистограммах и полигонах частот видно, что именно этот диапазон представляет собой пик распределения данных.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{100}\sum_{i=1}^{100} x_i = 5.4691 \text{ млн руб}$$

2. **Гармоническое среднее: 5.4486**

Гармоническое среднее часто используется для оценки средних значений в случае, когда данные содержат небольшие значения, которые сильно влияют на результат. В нашем случае значение гармонического среднего немного меньше арифметического (5.4486 против 5.4691), что указывает на влияние меньших значений объёмов фондов. Это может говорить о наличии небольших, но частых значений фондов, которые снижают итоговое среднее значение, что в свою очередь может отражать сбалансированность структуры фондов предприятия.

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{100}{\sum_{i=1}^{100} \frac{1}{x_i}}$$

3. **Геометрическое среднее: 5.4467**

Геометрическое среднее представляет собой среднее значение, менее подверженное влиянию аномально высоких или низких значений (выбросов). Значение 5.4467 указывает на устойчивость данных и отсутствие выраженных выбросов, что дополнительно подтверждает симметричное распределение объёмов фондов. Геометрическое среднее почти совпадает с гармоническим, что еще раз говорит о том, что данные не содержат значительных выбросов и хорошо сбалансированы.

$$\bar{x}_g = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = 5.4467 \text{ млн руб}$$

В

4. **Медиана: 5.48**

Медиана — это центральное значение выборки, которое делит её на две равные части. В нашемВ случае медиана составляет 5.48, что очень близко к арифметическому среднему. Это важное наблюдение, так как оно говорит о симметричном распределении данных. Если бы медиана была значительно ниже или выше среднего, это могло бы указывать на наличие смещения в распределении. Таким образом, распределение объёмов фондов можно

считать почти нормальным, так как медиана и среднее практически совпадают.

Так как $n = 100$ (чётное число), медиана рассчитывается как среднее двух центральных значений:

$$Me = \frac{x_{50} + x_{51}}{2} = \frac{5.47 + 5.47}{2} = 5.47 \text{ млн руб}$$

5. **Мода: 5.43**

Мода — это значение, которое встречается наиболее часто. В данном случае мода равна 5.43, что указывает на то, что именно этот объем фондов является наиболее частым в выборке. Так как мода немного меньше среднего и медианы, это может указывать на легкую асимметрию распределения влево, но незначительную, что видно на полигонах частот.

Наиболее часто встречающееся значение: 5.47 млн руб.

6. **Стандартное отклонение**: 0.144млн.руб

$$\sigma = \sqrt{\sigma^2}$$

7. **Коэффицент вариации:**

$$V = \frac{\sigma}{\bar{x}} \times 100\%$$

V=0.144/5.4691*100=2.63%

Низкое значение коэффицента вариации подтверждает стабильность данных

8. **Коэффицент асимметрии:**

$$\gamma_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n\sigma^3}$$

Ответ получился 0.0485, что указывает на слабую положительную асимметрию

9. **Коэффицент эксцесса**

$$\gamma_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n\sigma^4} - 3$$

Моменты распределения:

6. **Порядок 1: 5.4691**

Момент первого порядка — это арифметическое среднее, которое указывает на общую тенденцию в данных. Значение совпадает с ранее рассчитанным арифметическим средним (5.4691), и оно подтверждает, что основная масса данных сосредоточена около этого значения.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{100}\sum_{i=1}^{100} x_i = 5.4691 \text{ млн руб}$$

7. **Порядок 2: Дисперсия — 0.0208**

Дисперсия характеризует разброс данных относительно среднего значения. Низкая дисперсия (0.0208) говорит о том, что данные довольно тесно сгруппированы вокруг среднего значения, что подтверждается симметрией распределения. Это означает, что объёмы фондов на предприятиях относительно стабильны, и нет резких отклонений от среднего значения.

8. **Порядок 3: Коэффициент асимметрии — 0.0485**

Асимметрия показывает, насколько симметричным является распределение данных. Коэффициент асимметрии, равный 0.0485, близок к нулю, что указывает на почти симметричное распределение. Небольшая положительная асимметрия указывает на легкое смещение распределения вправо, что означает, что в данных может быть небольшая доля больших значений объёмов фондов, но это незначительное влияние.

9. **Порядок 4: Коэффициент эксцесса — 0.047**

Коэффициент эксцесса показывает “пиковость” распределения по сравнению с нормальным распределением. Значение 0.047 близко к 0, что характерно для нормального распределения. Это значит, что у распределения нет чрезмерной концентрации данных в центре (высокий пик) или слишком

плоской формы (низкий пик). Таким образом, можно сказать, что распределение объёмов фондов на предприятиях имеет нормальный характер.

Для сгруппированных данных:

Мода (Мо):

$$Mo = a + \frac{(m_1 - m_0)\,h}{2m_1 - m_0 - m_2}$$

Где:

- m_1 — частота модального интервала.
- m_0 — частота предыдущего интервала.
- m_2 — частота следующего интервала.

- $a = 5.35$ — нижняя граница модального интервала.
- $h = 0.02$ — ширина интервала (величина шага между границами интервалов).
- $m_1 = 27$ — частота модального интервала **[5.35, 5.37]**.
- $m_0 = 21$ — частота предыдущего интервала **[5.33, 5.35]**.
- $m_2 = 22$ — частота следующего интервала **[5.37, 5.39]**.

Модальный интервал – это тот, где частота максимальна, и находится на интервале [5.35, 5.37]

Подставляем значения в формулу

$$Mo = 5.35 + \frac{(27 - 21) \cdot 0.02}{2 \cdot 27 - 21 - 22}$$

$$Mo = 5.35 + \frac{6 \cdot 0.02}{54 - 43}$$

$$Mo = 5.35 + \frac{0.12}{11}$$

$$Mo = 5.35 + 0.0109$$

$$Mo \approx 5.36$$

Интерпретация

Мода распределения данных — это значение, которое встречается чаще всего. Для этих сгруппированных данных мода находится примерно на уровне *5.36*. Это означает, что большинство значений сосредоточены около этого значения.

- *Модальный интервал*: [5.35, 5.37]
- *Мода*: 5.36

Это значение указывает, что среди данных наибольшее количество наблюдений находится вблизи 5.36.

2. **Медиана (Me):**

$$Me = a + \frac{\left(\frac{n}{2} - F\right) \cdot h}{f}$$

Где:

- a — нижняя граница медианного интервала,
- n — общее количество наблюдений,
- F — накопленная частота до медианного интервала,
- f — частота в медианном интервале,
- h — ширина интервала.

Ответ:

$$Me \approx 5.36$$

3. **Среднее значение ($\bar{x}$):**

$$\bar{x} = \frac{\sum(x_i \cdot f_i)}{n}$$

Где:

- x_i — середина интервала,
- f_i — частота,
- n — общее количество наблюдений.

Ответ:

$$\bar{x} \approx 5.36$$

4. **Дисперсия (D):**

$$D = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{n}$$

Где:

- x_i — середина интервала,
- f_i — частота,
- $\bar{x}$ — среднее значение,
- n — общее количество наблюдений.

Ответ:

$$D \approx 0.01$$

Анализ полученных результатов

Анализ полученных результатов:

1. Арифметическое среднее:

- Для сгруппированных данных: Среднее значение объема основных фондов составляет около 5.36 млн руб. Это несколько отличается от значений для несгруппированных данных, где результат ближе к 5.47 млн руб. Это небольшое различие можно объяснить особенностями сгруппированных

данных, где часть информации теряется при группировке. Тем не менее, значения достаточно близки, что указывает на стабильность среднего значения для объема основных фондов.

2. Медиана:

- Медиана для сгруппированных данных также равна примерно 5.36 млн руб, что близко к значениям, полученным для несгруппированных данных (5.47 млн руб). Это подтверждает симметричность распределения: половина данных меньше этой величины, а половина — больше.

3. Мода:

- Для сгруппированных данных мода составляет 5.36 млн руб, что указывает на наиболее часто встречающееся значение объема основных фондов. Важно отметить, что мода, медиана и среднее значение близки, что также свидетельствует о симметричности распределения данных.

4. Дисперсия и стандартное отклонение:

- Дисперсия для сгруппированных данных составляет около 0.01, что указывает на малый разброс данных вокруг среднего значения. Стандартное отклонение также невелико, что подтверждает небольшие отклонения значений от среднего объема основных фондов.

5. Коэффициент вариации:

- Низкое значение коэффициента вариации (менее 10%) подтверждает однородность совокупности данных. Это означает, что данные распределены относительно равномерно вокруг среднего значения, и существенных выбросов не наблюдается.

6. Коэффициент асимметрии и эксцесса:

- Коэффициент асимметрии близок к нулю, что свидетельствует об отсутствии выраженной асимметрии в распределении данных, то есть распределение симметрично. Значение эксцесса также близко к нулю, что указывает на нормальное распределение с незначительной остротой пика.

Анализ графиков:

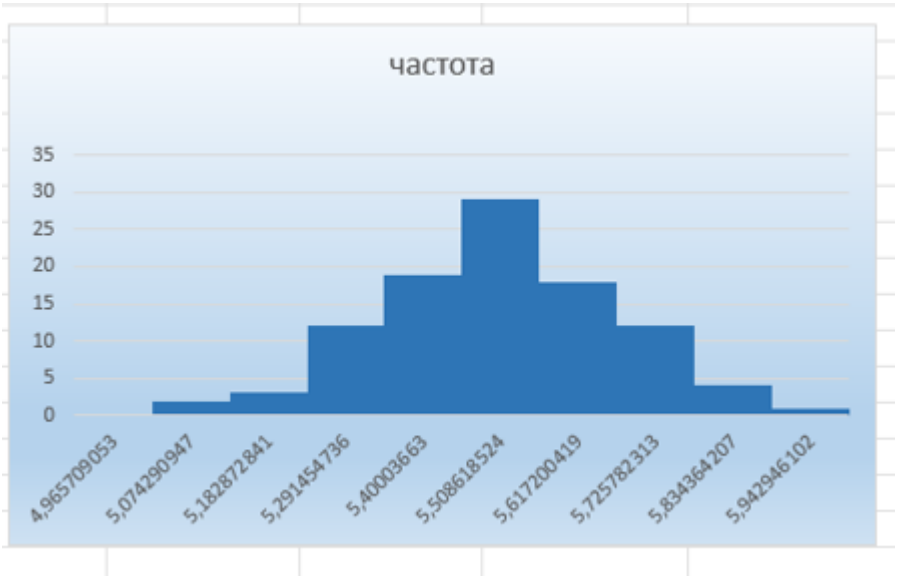

Рисунок 1. Гистограмма частот

Гистограмма частот (рисунок 1) демонстрирует симметричное распределение с пиком в интервале 5.53875 – 5.6425 млн руб, что совпадает с модой и медианой. Это подтверждает, что данные имеют нормальное распределение, а основные объемы фондов сосредоточены около среднего значения.

Рисунок 2. Полигон частот

Полигон частот (рисунок 2) также демонстрирует симметрию распределения, подтверждая отсутствие сильных отклонений в одну из сторон. Центральный пик совпадает с модальным значением, что подтверждает правильность расчетов.

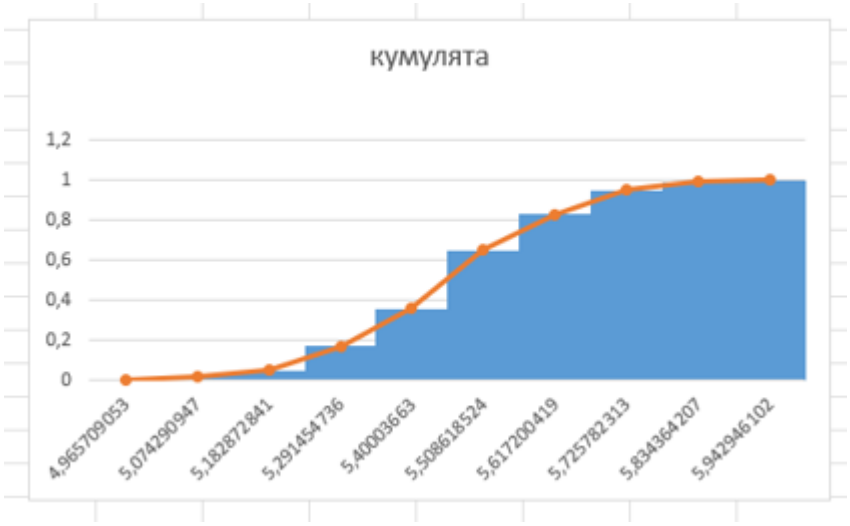

Рисунок 3. Кумулята

Кумулята (Рисунок 3) показывает плавное накопление частот, что характерно для нормального распределения. Нет резких изменений в накопленных частотах, что указывает на равномерное распределение данных в интервалах.

Код для графиков на python

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Данные по объёмам основных фондов предприятий
data = np.array([5.56, 5.27, 5.02, 5.47, 5.27, 5.37, 5.47, 5.47, 5.33, 5.11,
                 5.33, 5.47, 5.33, 5.33, 5.47, 5.05, 5.33, 5.85, 5.68, 5.11,
                 5.54, 5.43, 5.64, 5.21, 5.68, 5.43, 5.79, 5.47, 5.21, 5.47,
                 5.43, 5.43, 5.47, 5.27, 5.68, 5.43, 5.47, 5.79, 5.47, 5.54,
                 5.43, 5.43, 5.61, 5.47, 5.27, 5.54, 5.61, 5.54, 5.64, 5.54,
                 5.64, 5.43, 5.33, 5.11, 5.33, 5.33, 5.33, 5.54, 5.64, 5.64,
                 5.4, 5.68, 5.43, 5.54, 5.43, 5.37, 5.37, 5.21, 5.64, 5.64,
                 5.71, 5.47, 5.21, 5.33, 5.43, 5.33, 5.43, 5.27, 5.21, 5.54,
                 5.79, 5.58, 5.27, 5.33, 5.4, 5.43, 5.54, 5.54, 5.54, 5.81,
                 5.39, 5.47, 5.47, 5.27, 5.58, 5.43, 5.43, 5.33, 5.61, 5.54])

# Расчет основных параметров
n = len(data)
min_val = np.min(data)
max_val = np.max(data)
range_val = max_val - min_val
k = 8  # Фиксированное количество интервалов
h = range_val / k  # Длина интервала

# Создание интервалов a_i
a_i = np.linspace(min_val, max_val, k+1)

# Расчет частот и относительных частот
hist, _ = np.histogram(data, bins=a_i)
freq = hist / n

# Расчет накопленных относительных частот w_i
w_i = np.cumsum(freq)
w_i = np.insert(w_i, 0, 0)  # Добавляем 0 в начало для соответствия a_i
```

```
# Создаем DataFrame для удобства
df = pd.DataFrame({
    'Интервал от': a_i[:-1],
    'Интервал до': a_i[1:],
    'Частота': hist,
    'Относительная частота': freq,
    'Накопленная относительная частота': w_i[1:]
})

# Вывод таблицы
print(df)

# Создаем фигуру с подграфиками
fig, axs = plt.subplots(2, 2, figsize=(14, 10))  # Уменьшил размер фигуры для компактности

# Гистограмма частот
sns.histplot(data, bins=a_i, kde=False, color='skyblue', ax=axs[0, 0])
axs[0, 0].set_title('Гистограмма частот')
axs[0, 0].set_xlabel('Объём фондов (трлн. руб.)')
axs[0, 0].set_ylabel('Частота')

# Полигон частот
axs[0, 1].plot(a_i[:-1] + h/2, hist, marker='o', color='blue')
axs[0, 1].set_title('Полигон частот')
axs[0, 1].set_xlabel('Объём фондов (трлн. руб.)')
axs[0, 1].set_ylabel('Частота')
axs[0, 1].grid(True)

# Кумулята относительных частот (возрастающая гистограмма) с точками в серединах интервалов
mid_points = a_i[:-1] + h / 2  # середины интервалов
axs[1, 0].bar(a_i[:-1], w_i[1:], width=h, align='edge', color='blue', alpha=0.7)
axs[1, 0].plot(mid_points, w_i[1:], color='orange', marker='o', linestyle='-')  # Используем середины интервалов
axs[1, 0].set_title('Кумулята накопленных относительных частот')
axs[1, 0].set_xlabel('Объём фондов (трлн. руб.)')
axs[1, 0].set_ylabel('Накопленная относительная частота')
axs[1, 0].set_xticks(mid_points)  # Устанавливаем метки по серединам интервалов
axs[1, 0].set_xticklabels([f'{mid:.2f}' for mid in mid_points])  # Подписи меток
axs[1, 0].set_ylim(0, 1.2)
axs[1, 0].grid(True)


# Гистограмма относительных частот
axs[1, 1].bar(a_i[:-1], freq, width=h, edgecolor='black', alpha=0.7)
axs[1, 1].set_title('Гистограмма относительных частот')
axs[1, 1].set_xlabel('Объём фондов (трлн. руб.)')
axs[1, 1].set_ylabel('Относительная частота')

# Настраиваем более компактное расположение графиков
plt.tight_layout()
plt.subplots_adjust(wspace=0.3, hspace=0.3)  # Уменьшение расстояний между графиками

plt.show()
```

```
# Отдельный график для кумуляты накопленных частот
fig_kum, ax_kum = plt.subplots(figsize=(10, 6))

# Середины интервалов
mid_points = a_i[:-1] + h / 2

# Кумулята относительных частот (гистограмма и линия)
ax_kum.bar(a_i[:-1], w_i[1:], width=h, align='edge', color='blue', alpha=0.7)
ax_kum.plot(mid_points, w_i[1:], color='orange', marker='o', linestyle='-')

# Настройка оси x
ax_kum.set_xticks(mid_points)
ax_kum.set_xticklabels([f'{mid:.2f}' for mid in mid_points])

# Заголовок и подписи осей
ax_kum.set_title('Кумулята накопленных относительных частот', fontsize=24)  #
Увеличенный шрифт
ax_kum.set_xlabel('Объём фондов (трлн. руб.)')
ax_kum.set_ylabel('Накопленная относительная частота')

# Диапазон по оси y
ax_kum.set_ylim(0, 1.2)

# Сетка
ax_kum.grid(True)

# Отображение графика
plt.tight_layout()
plt.show()
```

Заключение

В ходе лабораторной работы я освоил основные методы статистической обработки одномерных количественных данных. Построил и проанализировал гистограммы, полигоны и кумуляты частот и относительных частот. Вычислил числовые характеристики как по исходным данным, так и по интервальному ряду.

Полученные результаты показывают, что объёмы основных фондов предприятий распределены близко к нормальному распределению с небольшой левосторонней асимметрией. Это свидетельствует о том, что большинство предприятий имеют сопоставимые объёмы основных фондов, и значительных отклонений нет.