Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт перспективной инженерии
Департамент цифровых, робототехнических систем и электроники

ОТЧЕТ

По лабораторной работе №2

Дисциплины «Анализ данных»

Выполнил:

Пустяков Андрей Сергеевич

3 курс, группа ИВТ-б-о-22-1,

09.03.01 «Информатика и вычислительная техника (профиль) «Программное обеспечение средств вычислительной техники и автоматизированных систем», очная форма обучения

____

(подпись)

Проверил:

Мельников С. В. к.т.н. доцент департамента цифровых робототехнических систем и электроники института перспективной инженерии

____

(подпись)

Отчет защищен с оценкой ___________ Дата защиты__________________

Ставрополь, 2024 г.

Тема: Построение кривой нормального распределения по опытным данным. Проверка гипотезы о нормальном распределении выборки.

Цель: овладеть способами построения эмпирической и теоретической (нормальной) кривой распределения, выработать умения и навыки применения критериев согласия для проверки выдвинутой статистической гипотезы.

Ход работы:

Вариант 25

Имеются данные замеров температуры масла двигателя автомобиля ГАЗ-53 А (табл. 1).

Таблица 1 – Данные замеров температуры масла.

19	29	21	39	25	26	32	25	28	26	36	30	31
29	35	23	32	27	27	26	26	30	27	25	28	28
36	29	35	26	32	29	38	28	25	29	34	28	29
32	34	28	28	29	33	27	34	25	28	26	30	38
39	32	29	29	34	35	32	27	26	25	26	35	36
30	28	33	26	28	26	28	27	33	33	29	32	25
38	26	36	23	24	27	26	30	34	25	24	33	

Задание 1.

На основе данных замеров температуры масла был получен непрерывный вариационный ряд, найден размах варьирования признака, найдено число интервалов вариационного ряда, длина частичных интервалов, были получены середины частичных интервалов для построения дискретного вариационного ряда (рис. 1).

Варианты, xi	19.0	20.3	21.6	22.9	24.2	25.5	26.8	28.1	29.4	30.7	32.0	33.3	34.6	35.9	38.1
Частоты, ni	1	0	1	2	2	20	7	11	15	1	7	5	9	4	5

Рисунок 1 – Дискретный вариационный ряд

Также было найдено среднее значение и среднеквадратическое отклонение (рис. 2).

```
Среднее квадратичное отклонение S =  4.254226076053415
Среднеарифметическое дискретного вариационного ряда: Xcp =  28.160000000000004
```

Рисунок 2 – Среднеквадратическое и среднее значения

Эмпирическая кривая распределения представляет собой полигон частот. Для построения теоретической (нормальной) кривой найдем координаты точек (x_i, n_i\`), для чего рассчитаем теоретические частоты $n_i$\`. Была создана расчётная таблица для нахождения теоретических частот (рис. 3).

```
+------+----+---------------------+--------------+--------+--------------------+-----+
|  xi  | ni |        xi-xcp       | ui=(xi-xcp)/S | ф(ui)  |         yi         | ni` |
+------+----+---------------------+--------------+--------+--------------------+-----+
| 19.0 | 1  | -10.435555555555553 |     -2.5     | 0.0175 | 0.4812861289918652 |  0  |
| 20.3 | 0  |  -9.135555555555552 |     -2.1     | 0.044  | 1.210090838608118  |  1  |
| 21.6 | 1  |  -7.835555555555551 |     -1.8     | 0.079  | 2.1726630965918483 |  2  |
| 22.9 | 2  |  -6.535555555555554 |     -1.5     | 0.1295 | 3.561517354539802  |  4  |
| 24.2 | 2  |  -5.235555555555553 |     -1.2     | 0.1942 | 5.340900928584012  |  5  |
| 25.5 | 20 | -3.9355555555555526 |     -0.9     | 0.2661 | 7.318299367127731  |  7  |
| 26.8 | 7  |  -2.635555555555552 |     -0.6     | 0.3332 | 9.163687896005111  |  9  |
| 28.1 | 11 | -1.3355555555555512 |     -0.3     | 0.3814 | 10.489287405571277 |  10 |
| 29.4 | 15 | -0.03555555555555401 |     -0.0     | 0.3989 | 10.97057353456314  |  11 |
| 30.7 | 1  |  1.2644444444444467 |     0.3      | 0.3814 | 10.489287405571277 |  10 |
| 32.0 | 7  |  2.5644444444444474 |     0.6      | 0.3332 | 9.163687896005111  |  9  |
| 33.3 | 5  |  3.8644444444444446 |     0.9      | 0.2661 | 7.318299367127731  |  7  |
| 34.6 | 9  |  5.164444444444449  |     1.2      | 0.1942 | 5.340900928584012  |  5  |
| 35.9 | 4  |  6.464444444444446  |     1.5      | 0.1295 | 3.561517354539802  |  4  |
| 38.1 | 5  |  8.664444444444449  |     2.0      | 0.054  | 1.4851114837463264 |  1  |
+------+----+---------------------+--------------+--------+--------------------+-----+
```

Рисунок 3 – Расчетная таблица для теоретических частот

Код функции нахождения значений и построения расчетной таблицы:

```
def calculation_theoretical_table(any_list, frequency, x_mid, rms, num_h):
    """
    Функция нахождения значений и построения расчетной таблицы.

    :param any_list: Дискретный вариационный ряд;
    :param frequency: частоты дискретного вариационного ряда;
    :param x_mid: среднее значение исходя из дискретного вариационного ряда;
    :param rms: среднеквадратическое отклонение;
    :param num_h: длина частичного интервала.
    :return: расчетная таблица и список теоретических частот.
    """

    # Создаем двумерный пустой массив.
    compute_table = [[0 for _ in range(7)] for _ in range(len(any_list))]

    # Заполним созданный список расчетными значениями
    for i in range(0, len(any_list)):
```

```
        compute_table[i][0] = any_list[i]

    for i in range(0, len(any_list)):
        compute_table[i][1] = frequency[i]

    for i in range(0, len(any_list)):
        compute_table[i][2] = any_list[i] - x_mid

    for i in range(0, len(any_list)):
        compute_table[i][3] = round((compute_table[i][2] / rms), 1)

    f_ui = [
        0.0175, 0.044, 0.079, 0.1295, 0.1942, 0.2661, 0.3332, 0.3814,
        0.3989, 0.3814, 0.3332, 0.2661, 0.1942, 0.1295, 0.054
    ]

    for i in range(0, len(any_list)):
        compute_table[i][4] = f_ui[i]

    for i in range(0, len(any_list)):
        compute_table[i][5] = (90 * num_h * compute_table[i][4]) / rms

    for i in range(0, len(any_list)):
        compute_table[i][6] = round(compute_table[i][5])

    # Создание таблицы
    table = PrettyTable()

    # Установка заголовков столбцов (например, Col1, Col2, ..., Col8)
    columns = ['xi', 'ni', 'xi-xcp', 'ui=(xi-xcp)/S', 'φ(ui)', 'yi', 'ni`']
    table.field_names = columns

    # Добавляем строки из списка в таблицу
    for row in compute_table:
        table.add_row(row)

    # Вывод таблицы
    print('Расчетная таблица:')
    print(table)

    theoretical_frequency = []
    for i in range(0, len(any_list)):
        theoretical_frequency.append(compute_table[i][6])

    # Вернем табличные значения для дальнейших вычислений
    return compute_table, theoretical_frequency
```

По данным расчетной таблицы, а также по данным дискретного вариационного ряда и его частотам на одном графике были построены кривые нормального распределения и кривая эмпирического распределения (рис. 4). Исходя из данного графика можно сделать вывод о том, что распределение слабо похоже на нормальное.

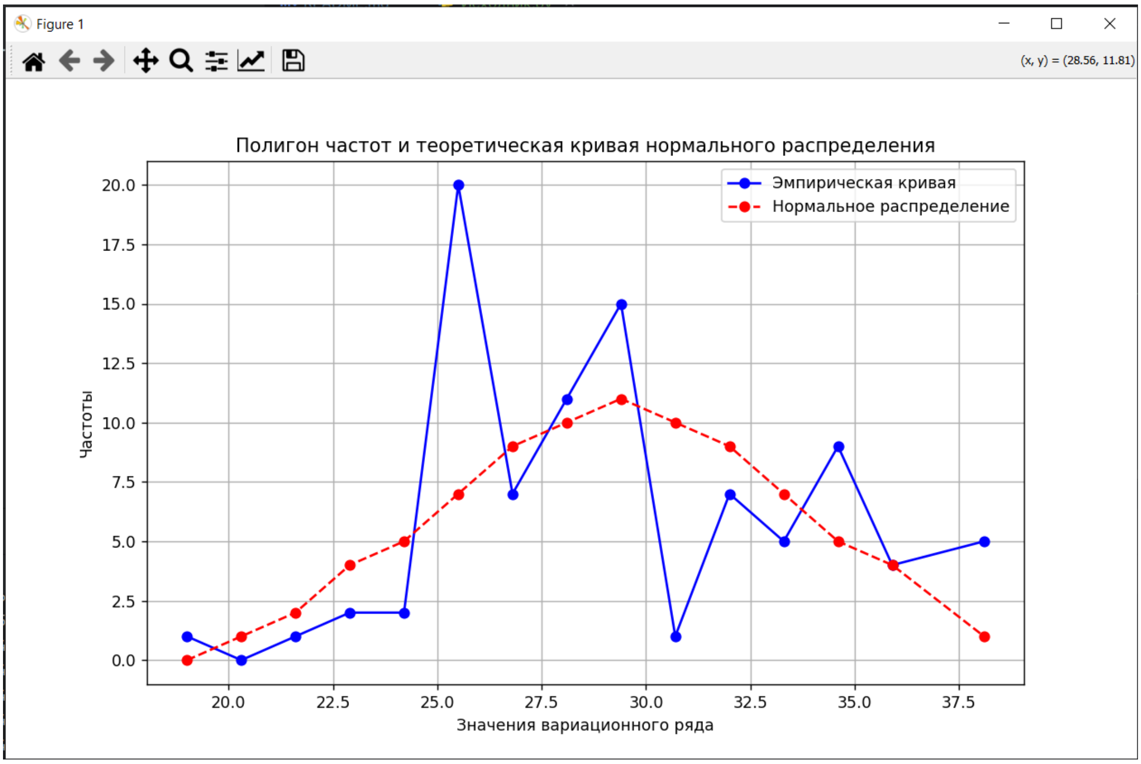

Рисунок 4 – Полигон частот и кривая нормального распределения

Задание 2.

Была проверена согласованность эмпирического распределения температуры масла двигателя автомобиля ГАЗ-53 А с теоретическим нормальным согласно критерию Пирсона. Было вычислено значение величины χ^2 по формуле:

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - n_i')^2}{n_i'}.$$

Код функции, выполняющей вычисление данного значения:

```
def pirson(compute_table, any_list):
    """
    Функция проверки согласованности распределения по критерию Пирсона.

    :param compute_table: Расчетная таблица
    :param any_list: дискретный вариационный ряд.
    :return: хи квадрат, коэффициент.
    """

    # Вычислим значение величины Хи квадрат
    xi_2 = 0
    for i in range(len(any_list)):
        xi_2 += ((compute_table[i][1] - compute_table[i][6]) *
                 (compute_table[i][1] - compute_table[i][6])) /
compute_table[i][5]
    print("Величина Хи^2 = ", xi_2)
```

```
    # Пусть уровень значимости a = 0,95, тогда
    num_k = len(any_list) - 3
    a = 0.95
    print("Число степеней свободы k = ", num_k)
    print("Уровень значисмости a = ", a)

    # Исходя из приложения 1:
    xi_kr_2 = 5.23
    print("Критическое значение Хи^2", xi_kr_2)

    if xi_2 < xi_kr_2:
        print("Хи^2 < Хкр^2, распределение соотвествует нормальному согласно
критерию Пирсона.")
    else:
        print("Хи^2 > Хкр^2, распределение не соотвествует нормальному
согласно критерию Пирсона.")

    return xi_2
```

Величина χ^2 оказалась равной 53.73032135795478, причем это значение оказалось большим чем критическое значение $\chi^2_{кр}$ (рис. 5). Исходя из этого можно сделать вывод о том, что распределение не соответствует нормальному, но при этом критерий Пирсона не совсем подходит для данной оценки, поскольку объем выборки мал для данного критерия. Объем выборки меньше 100 (90 измерений температуры масла). Данное соответствие было проверено с помощью другого, более мощного критерия Колмагорова, который больше подходит для данной выборки.

```
Величина Хи^2 =  53.73032135795478
Число степеней свободы k =  12
Уровень значисмости a =  0.95
Критическое значение Хи^2 5.23
Хи^2 > Хкр^2, распределение не соотвествует нормальному согласно критерию Пирсона.
```

Рисунок 5 – Согласованность по критерию Пирсона

Была проведена проверка близости эмпирического распределения к нормальному согласно критерию Колмагорова. По ранее полученным данным была вычислена статистика Колмагорова по формуле:

$$\lambda = D / \sqrt{n},$$

где D – максимум абсолютного значения разности между накопленными эмпирическими частотами и накопленными теоретическими частотами, n –

объем выборки. Значение λ составило 0.008680762204383779. По данному значению вручную по формуле:

$$K(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$$

Значение функции получилось равным 0.9835571947084456, что больше, чем 0.05, а значит распределение не соответствует нормальному согласно критерию Колмагорова (рис. 6). Предыдущий критерий подтвердился, мы не можем принять гипотезу о том, что распределение нормальное.

```
Статистика Колмагорова =  0.008680762204383779
Значение функции K(lambda_k) =  0.9835571947084456
Распределение не соответствует нормальному согласно критерию Колмагорова.
```

Рисунок 6 – Согласованность по критерию Колмагорова

Код функции, осуществляющей проверку согласованности нормальному распределению согласно критерию Колмагорова:

```
def kolmagorov(any_list, frequency, theoretical_frequency):
    # Вычислим объем выборки
    sample_size = len(any_list)

    # Находим общие суммы для нормализации частот
    sum_empirical = sum(frequency)
    sum_theoretical = sum(theoretical_frequency)

    # Вычисляем накопленные частоты вручную
    cumulative_empirical = []
    cumulative_theoretical = []

    cumulative_sum_empirical = 0
    cumulative_sum_theoretical = 0

    for i in frequency:
        cumulative_sum_empirical += i
        # Нормализуем накопленную сумму и добавляем в список накопленной
частоты
        cumulative_empirical.append(cumulative_sum_empirical / sum_empirical)

    for i in theoretical_frequency:
        cumulative_sum_theoretical += i
        # Нормализуем накопленную сумму и добавляем в список накопленной
частоты
        cumulative_theoretical.append(cumulative_sum_theoretical /
sum_theoretical)

    # Вычисляем абсолютные разности и находим максимальную разность
    max_difference = max(abs(emp - theo) for emp, theo in
zip(cumulative_empirical, cumulative_theoretical))

    # Делим на корень из объема выборки для получения статистики Колмогорова
```

```
    lambda_k = max_difference / math.sqrt(sample_size)

    print("Статистика Колмагорова = ", lambda_k)

    # Исходя из статистики Колмагорова найдем значение функции K(lambda_k)
    summation = 0
    for num_k in range(1, 11):
        summation += (-1) ** (num_k - 1) * math.exp(-2 * (num_k ** 2) *
(lambda_k ** 2))

    k_lambda_k = 1 - 2 * summation

    print("Значение функции K(lambda_k) = ", k_lambda_k)

    if k_lambda_k < 0.05:
        print("Распределение соответствует нормальному согласно критерию
Колмагорова.")
    else:
        print("Распределение не соответствует нормальному согласно критерию
Колмагорова.")

    return k_lambda_k
```

Была проведена близость рассматриваемой выборки к нормальному распределению по приближенному критерию. Были вычислены выборочные статистики: асимметрия и эксцесс. Коэффициент асимметрии составил A_s = 0,36190542870479026, а коэффициент эксцесса E_x = 0,4342819355657799. Были найдены среднеквадратические отклонения асимметрии и эксцесса по формулам:

$$S_{A_s} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}},$$

$$S_{E_x} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}.$$

Среднеквадратическое отклонение асимметрии составило S_{As} = 0,2511935352083134, среднеквадратическое отклонение эксцесса составило S_{Ex} = 0,4861099990472253 (рис. 7). Так как $|A_s| > S_{As}$ и $|E_x| < S_{Ex}$, то можно считать, что рассматриваемая выборка не подчиняется нормальному распределению.

```
Коэффициент асимметрии: As =  0.36190542870479026
Коэффициент эксцесса: Ex =  0.434281935565779 9
Среднеквадратическое отклонение асимметрии S_As =  0.2511935352083134
Среднеквадратическое отклонение эксцесса S_Ex =  0.4861099990472253
Выборочная совокупность не подчиняется нормальному закону распределения согласно приближенному критерию.
```

Рисунок 6 – Согласованность по приближенному критерию

Код функций, осуществляющий проверку согласованности по приближенному критерию и вычисляющей среднеквадратические отклонения асимметрии и эксцесса:

```
def rsm_asymmetry_excess(any_list):
    """
    Функция нахождения среднеквадратического отклонения асимметрии и
эксцесса.

    :param any_list: выборка.
    :return: среднеквадратическое отклонение асимметрии и
среднеквадратическое отклонение эксцесса.
    """

    # Найдем объем выборки
    num_n = len(any_list)

    # Вычислим среднеквадратическое отклонение асимметрии
    s_as = math.sqrt((6 * (num_n - 1)) / ((num_n + 1) * (num_n + 3)))

    # Вычислим среднеквадратическое отклонение эксцесса
    s_ex = math.sqrt(((24 * num_n) * (num_n - 2) * (num_n - 3) /
                      ((num_n - 1) * (num_n - 1) * (num_n + 3) * (num_n +
5))))

    print("Среднеквадратическое отклонение асимметрии S_As = ", s_as)
    print("Среднеквадратическое отклонение эксцесса S_Ex = ", s_ex)

    return s_as, s_ex


def approximate_criterion(a_s, e_x, s_as, s_ex):
    """
    Функция проверки согласованности нормальному закону распределения
согласно приближенному критерию.

    :param a_s: коэффициент асимметрии;
    :param e_x: коэффициент эксцесса;
    :param s_as: среднеквадратическое отклонение асимметрии;
    :param s_ex: среднеквадратическое отклонение эксцесса.
    :return: ничего.
    """

    if abs(a_s) <= s_as and abs(e_x) <= s_ex:
        print("Выборочная совокупность подчиняется нормальному закону
распределения согласно приближенному критерию.")
    else:
        print("Выборочная совокупность не подчиняется нормальному "
              "закону распределения согласно приближенному критерию.")
```

Ответы на контрольные вопросы:

1. Рассказать о возможных вариантах построения кривой нормального распределения по опытным данным.

Возможные варианты построения кривой нормального распределения по опытным данным:

- Графический метод: опытные данные наносятся на вероятностную бумагу и сравниваются с графиком теоретической функции, которая на вероятностной сетке выглядит как прямая линия. Отклонения от прямой линии показывают расхождение с нормальным распределением.

- Эмпирический метод: на основе дискретного вариационного ряда строится эмпирическая кривая (полигон частот), затем в тех же координатах строится теоретическая кривая нормального распределения на основе выборочных средних и стандартного отклонения.

2. Дать определение статистической гипотезы.

Статистическая гипотеза — это предположение относительно свойств генеральной совокупности, которое проверяется на основе выборочных данных. Нулевая гипотеза H_0 предполагает отсутствие различий между эмпирическими и теоретическими распределениями.

3. Что называется статистическим критерием?

Статистический критерий — это правило или алгоритм, который определяет, принимается или отвергается нулевая гипотеза на основе выборочных данных и заданного уровня значимости.

4. Описать алгоритм применения любого статистического критерия для обработки экспериментальных данных.

Алгоритм применения статистического критерия:

- Вычисление статистики критерия на основе выборочных данных.

- Определение критического значения статистики для заданного уровня значимости и числа степеней свободы.

- Сравнение вычисленной статистики с критическим значением и принятие решения о принятии или отклонении гипотезы.

5. Сформулировать правило применения критерия согласия Пирсона для проверки гипотезы согласованности эмпирического распределения с теоретическим нормальным.

– Определяются теоретические частоты на основе нормального распределения.

– Вычисляется статистика X^2.

– Находится число степеней свободы и критическое значение.

– Если вычисленное значение X^2 меньше критического, гипотеза о нормальности не отвергается.

6. Рассказать о применении критерия согласия Романовского для оценки близости эмпирического распределения к теоретическому нормальному.

Критерий Романовского применяется для оценки близости эмпирического распределения к нормальному. Вычисляется отношение X^2/k, где X^2 — статистика Пирсона, k — число степеней свободы. Если это отношение меньше 3, расхождение незначительно, и распределение можно считать нормальным.

7. Описать алгоритм применения критерия Колмогорова для проверки соответствия эмпирического распределения нормальному теоретическому распределению.

– Вычисляется статистика D — максимум разности между эмпирической и теоретической функциями распределения.

– Вычисляется значение $\lambda = \sqrt{nD}$, где n— объем выборки.

– Определяется вероятность по таблице функции Колмогорова. Если вероятность менее 0.05, гипотеза о нормальном распределении отвергается.

8. Рассказать о применении критерия Ястремского для проверки соответствия данной выборочной совокупности нормальному распределению.

Критерий Ястремского проверяет соответствие выборки нормальному распределению через неравенство $J <= 3l + \theta$, где J — вычисленное значение критерия, l — число интервалов вариационного ряда. Если условие выполняется, выборка считается нормально распределенной.

9. Рассказать о приближенных критериях, применяемых для проверки гипотезы о нормальном распределении выборочной совокупности.

Используются статистики асимметрии и эксцесса. Вычисляются их средние квадратические отклонения, и, если отклонения незначительны, выборка считается нормально распределенной. Также можно использовать модификации критерия X^2.

Вывод: в ходе выполнения лабораторной работы были изучены и закреплены способы построения эмпирической и теоретической (нормальной) кривой распределения на одном графике. Была построена эмпирическая и теоретическая кривая нормального распределения на одном графике согласно варианту (температура масла двигателя) и по ней был сделал вывод о возможном несоответствии распределения нормальному распределению. Также были выработаны умения и навыки применения критериев согласия для проверки выдвинутой статистической гипотезы. Был применен критерий согласия Пирсона, согласно которому был сделан вывод о несогласии распределения нормальному несмотря на то, что критерий Пирсона не подходит данной выборке так как ее объем довольно мал (меньше 100). Также был применен более мощный и подходящий для выборки критерий согласия Колмагорова в ходе которого был подтвержден вывод о несогласии распределения нормальному. Был изучен и применен приближенный критерий согласия, который также подтвердил несоответствие распределения нормальному.