МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет прикладной математики и информатики

Кафедра математического моделирования и анализа данных

Предварительный статистический анализ данных Фишера

Отчет по лабораторной работе №1
студента 3 курса 7 группы
Бовта Тимофея

Преподаватель
Сафиуллин Т. Т.

Минск, 2023

Оглавление

Предварительный статистический анализ данных Фишера

Цели

В ходе предварительного анализа данных с помощью графических и статистических результатов определить, является ли модель моделью «случайная выборка»; исследовать закон ее распределения и проверить присутствие аномальных наблюдений.

Постановка задачи

Для данных Фишера:

1) проверить предположение о том, что математической моделью данных является модель «случайная выборка»;
2) исследовать закон распределения случайных выборок, соответствующих переменным sepal length, sepal width, petal length, petal width;
3) проверить предположение о нормальном распределении выборок;
4) исследовать однородность данных по среднему значению и дисперсии значений переменных sepal length, sepal width, petal length, petal width, относящихся к различным типам ирисов, видам setosa, versicolor, virginica;
5) с помощью графического анализа и тестовых статистик проверить предположение о присутствии в данных аномальных наблюдений;
6) проверить гипотезу о том, что соответствующие выборки наблюдений) являются случайными выборками из многомерного нормального распределения; дать содержательную интерпретацию этому предположению.

Указания:

- исследования проводить по индивидуальным наборам данных для указанных переменных и выборок;

№	Фамилия, Имя, Отчество	Iris Type	Variable for analysis	Variable with Outlier
2	Бовт Тимофей	1	1	3

- описание данных см. в прилагаемом файле.

Результаты проведенного исследования

Предварительный анализ однородных данных

Сформируем исходную генеральную совокупность данных, с которыми мы будем работать. В качестве генеральной совокупности будут выступать данные Фишера. Генеральная совокупность имеет следующий вид (Таблица 1.1).

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target label	target name
0	5.1	3.5	1.4	0.2	0	setosa
1	4.9	3.0	1.4	0.2	0	setosa
2	4.7	3.2	1.3	0.2	0	setosa
3	4.6	3.1	1.5	0.2	0	setosa
4	5.0	3.6	1.4	0.2	0	setosa
...	...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2	virginica
146	6.3	2.5	5.0	1.9	2	virginica
147	6.5	3.0	5.2	2.0	2	virginica
148	6.2	3.4	5.4	2.3	2	virginica
149	5.9	3.0	5.1	1.8	2	virginica

Таблица 1.1 – Генеральная совокупность ирисов Фишера.

Проведем исследование закона распределения случайной величины SEPAL LENGTH для ирисов вида SETOSA.

Для проверки гипотезы о нормальности распределения была построена гистограмма (Рисунок 1.1). Проанализируем построенную гистограмму:

- гистограмма унимодальная;
- на гистограмме отсутствует ярко выраженная асимметрия;
- “хвосты” гистограммы не выходят существенно за границы трёх сигм;
- за пределами границ трёх сигм отсутствуют изолированные наблюдения.

Таким образом, на данном этапе у нас нет причин отклонить гипотезу о нормальности распределения исследуемой случайной величины.

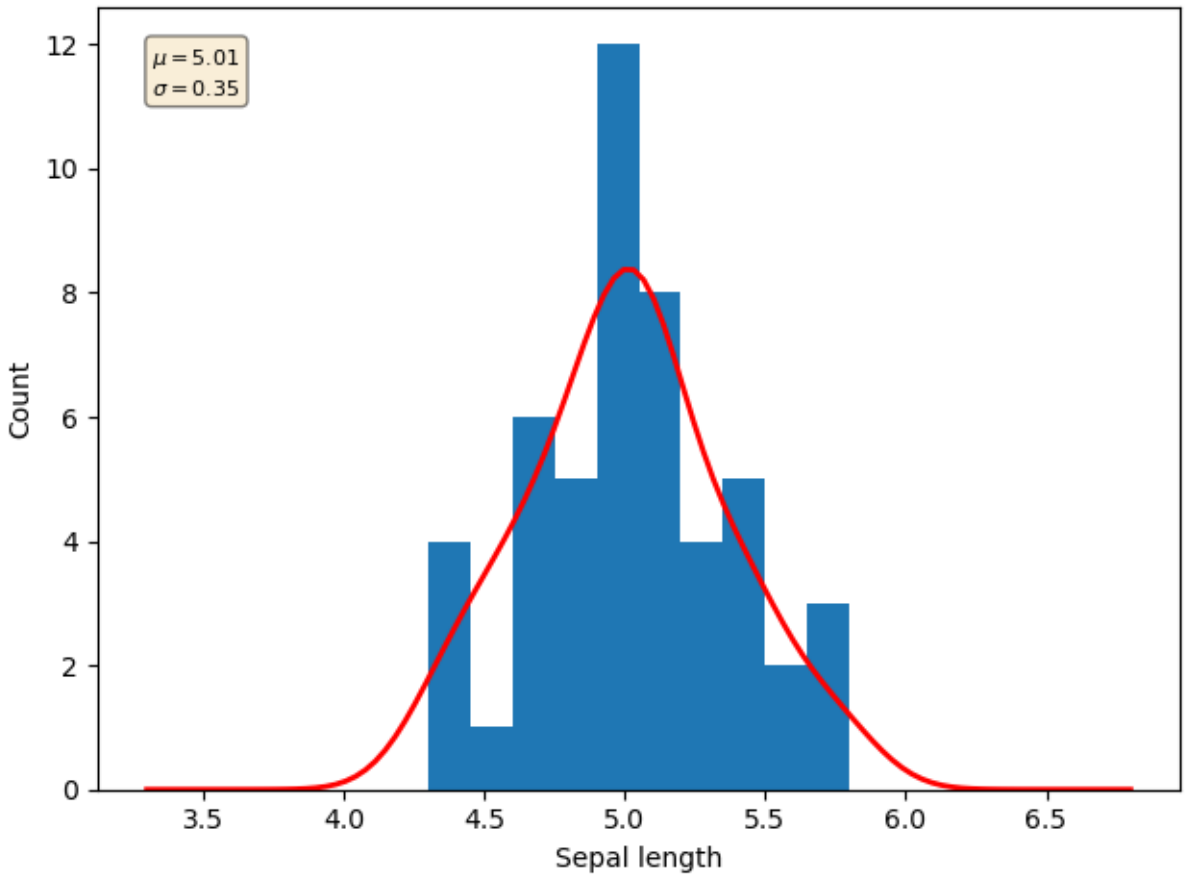

Рисунок 1.1 - Гистограмма распределения переменной SEPALLEN

Проведем анализ дескриптивных статистик (Таблица 1.2).

	count	mean	std	min	25%	50%	75%	max	median	mode	skewness	kurtosis
sepallen	50.0	5.006	0.35249	4.3	4.8	5.0	5.2	5.8	5.0	5.1	0.120087	-0.252689

Таблица 1.2 - Основные характеристики переменной SEPALLEN

Анализ показывает, что

- среднее, медиана распределения и мода выборки имеют близкие значения;
- близко к нулю значения коэффициента асимметрии (skewness), причем значение положительно, поэтому выборка смещена влево;
- близко к нулю значение коэффициента эксцесса (kurtosis), причем оно отрицательное, поэтому гистограмма имеет более пологую вершину;
- выполняется правило трёх сигмовых границ, то есть большинство значений лежит в пределах трёх сигм влево и вправо;

Исследуем наличие аномальных наблюдений с помощью графика «ящик с усами». Он представлен на рисунке 1.2.

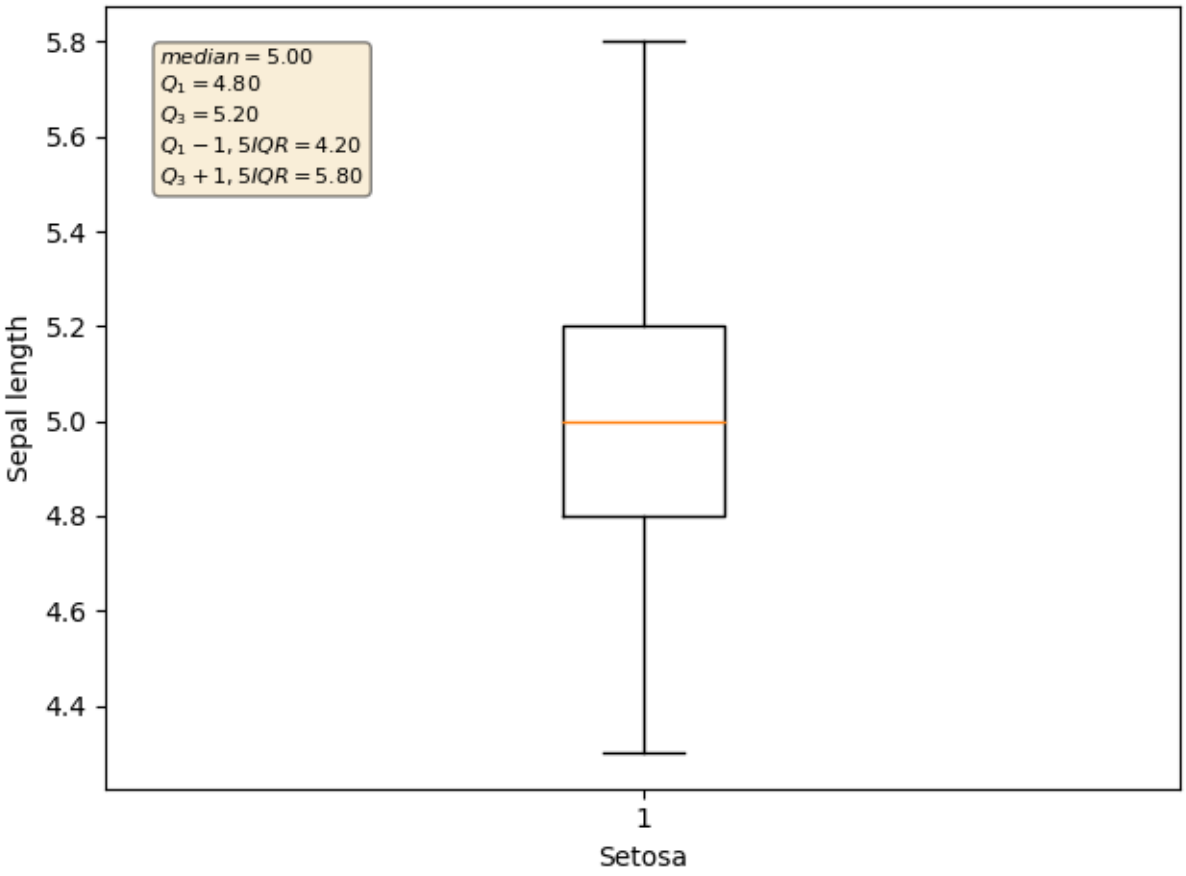

Рисунок 1.2 - График «ящик с усами» для переменной SEPALLEN

График дает удобное графическое представление о медиане, нижней и средней квартилях, минимальном и максимально значениях выборки. Расстояния между граничными частями графика дают представление о дисперсии и асимметрии выборки. Верхняя и нижняя границы “ящика” отражают значения в первом и третьем квартилях. На графике можно заметить положительную асимметрия в чем мы убедились ранее. На этом графике отсутствуют аномальные наблюдения, о чем говорит отсутствие значений за пределами “усов”. Параметры “ящика с усами” соответствуют нормальному распределению. Следовательно, мы всё так же не можем отклонить гипотезу о нормальности распределения.

Проведем проверку предположения о нормальности распределения, с помощью графика “квантиль-квантиль”.

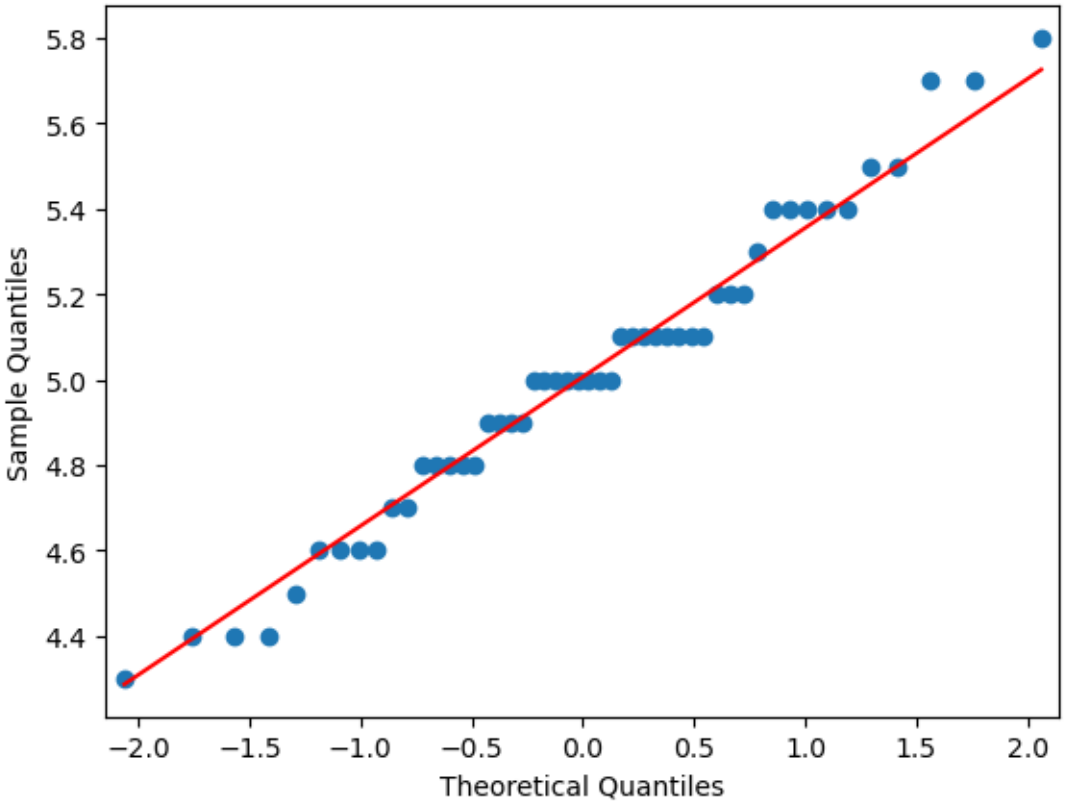

Рисунок 1.3 - График «квантиль-квантиль» для переменной SEPALLEN

График (Рисунок 1.3) показывает, что большая часть наблюдений в выборке близки к линии нормального распределения. На концах заметны несущественные отклонения. Однако они не позволяют нам отклонить гипотезу о нормальности распределения.

В заключение для проверки выборки на нормальность используем критерии согласия Шапиро-Уилка, Колмогорова-Смирнова и x^2-Пирсона. Если P-значения этих критериев будут больше, чем 0.05, то мы не отклоним гипотезу о нормальности распределения.

```
KstestResult(statistic=0.18, pvalue=0.3959398631708505, statistic_location=5.1, statistic_sign=1)
ShapiroResult(statistic=0.9776982069015503, pvalue=0.4595010578632355)
Power_divergenceResult(statistic=23.199999999999996, pvalue=0.0570921772726976)
```

Все три критерия свидетельствую в пользу нормальности распределения данной случайной величины, так как все они имеют P-значение больше 0.05.

Таким образом, по совокупности признаков мы можем принять гипотезу о нормальности исследуемого распределения, так как ни на одном из этапов мы не убедились в обратном. То есть выборка наблюдений случайной величины SEPAL LENGTH для ирисов вида SETOSA является случайной выборкой из нормального распределения.

Предварительный анализ данных с засорениями

Сформируем новую выборку из генеральной совокупности, взяв в качестве исходной случайной величины параметр PETAL WIDTH для ириса вида SETOSA. Добавим к данной выборке 5 случайных значений:

0	1.400000
1	1.400000
2	1.300000
3	1.500000
4	1.400000
	…
48	1.500000
49	1.400000
50	2.602068
51	8.598081
52	8.388875
53	3.545909
54	4.954904

Проведем исследование закона распределения полученной выборки по такому же алгоритму, как и прошлое исследование. Для проверки гипотезы о нормальности распределения была построим гистограмму распределения (Рисунок 2.1).

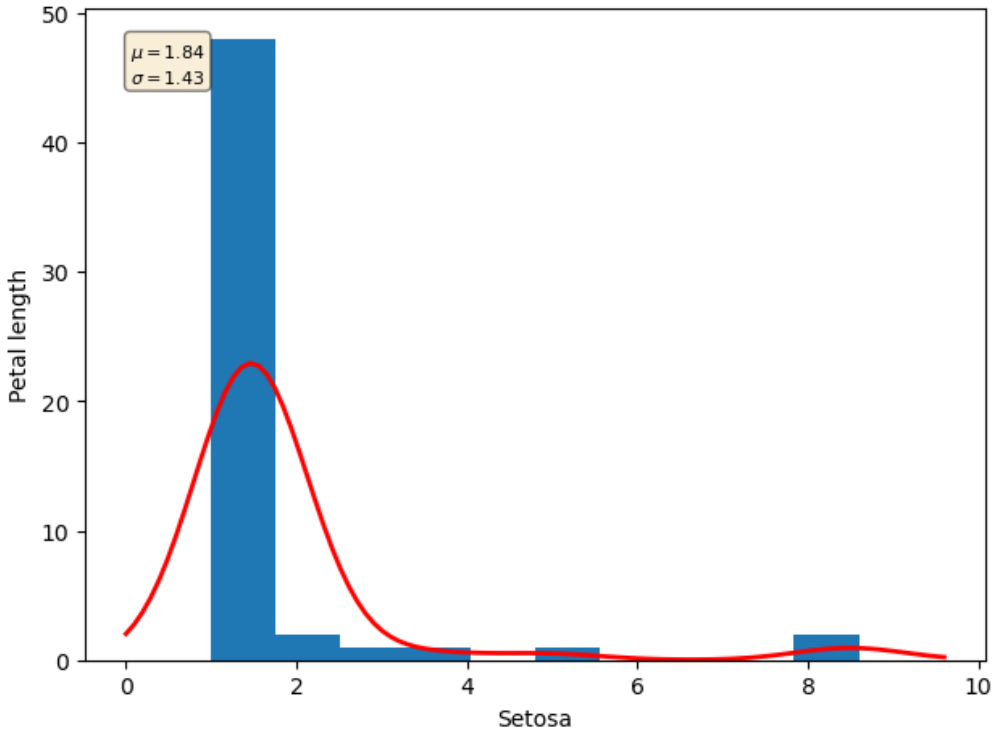

Рисунок 2.1 - Гистограмма распределения переменной PETALLEN

Проанализируем построенную гистограмму:

- гистограмма унимодальная;
- на гистограмме присутствует ярко выраженная асимметрия, график сильно смещён влево;
- график имеет очень острую вершину;
- правый “хвост” существенно выходит за границу трёх сигм;
- за пределами границ трёх сигм присутствуют изолированные наблюдения.

Все расхождения с нормальным распределением вызваны наличием аномальных наблюдений. Таким образом, на данном этапе мы не можем точно принять гипотезу о нормальности распределения исследуемой выборки.

Проведем анализ дескриптивных статистик (Таблица 2.2).

	count	mean	std	min	25%	50%	75%	max	median	mode	skewness	kurtosis
petallen	55.0	1.786904	1.490917	0.190563	1.4	1.5	1.6	9.194581	1.5	1.5	4.042253	16.163194

Таблица 2.2 - Основные характеристики переменной PETALLEN

Анализ показывает, что

- медиана распределения и мода выборки имеют близкие значения, а значение среднего отличается по причине неустойчивости среднего к аномальным значениям;
- значение коэффициента асимметрии положительно, поэтому выборка сильно смещена влево;
- велико и положительно значение коэффициента эксцесса, поэтому гистограмма имеет очень острую вершину;
- выполняется правило трёх сигмовых границ, то есть большинство значений лежит в пределах трёх сигм влево и вправо;

Следовательно, на данном этапе у нас имеются существенные отличия среднего от робастных аналогов, значения коэффициента ассиметрии от нуля, значение коэффициента эксцесса от нуля. Все эти проблемы вызваны наличие аномальных значений. Исследуем наличие аномальных наблюдений с помощью графика «ящик с усами». Он представлен на рисунке 2.2.

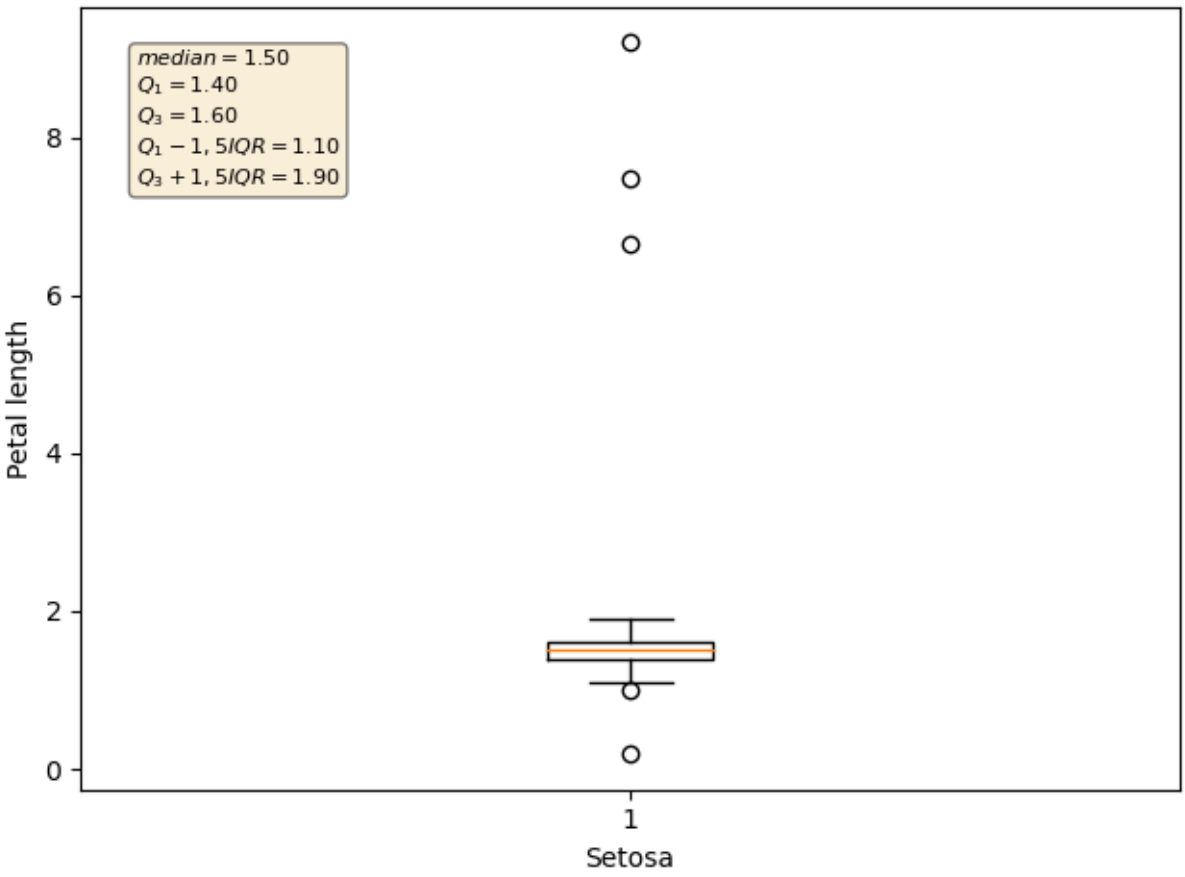

Рисунок 2.2 - График «ящик с усами» для переменной PETALLEN

На этом графике присутствуют аномальные наблюдения, о чем говорит наличие значений за пределами “усов”. Причем за пределами “усов” находятся как раз добавленные нами аномальные значения.

Проведем проверку предположения о нормальности распределения, с помощью графика “квантиль-квантиль”.

График (Рисунок 2.3) показывает, что большая часть наблюдений в выборке сильно отклоняются от линии нормального распределения. На концах заметны существенные отклонения.

Все проведенные исследования позволяют нам уже отклонить гипотезу о нормальности данного распределения. Причем это происходит из-за сильного влияния аномальных наблюдений на общую характеристику всей выборки.

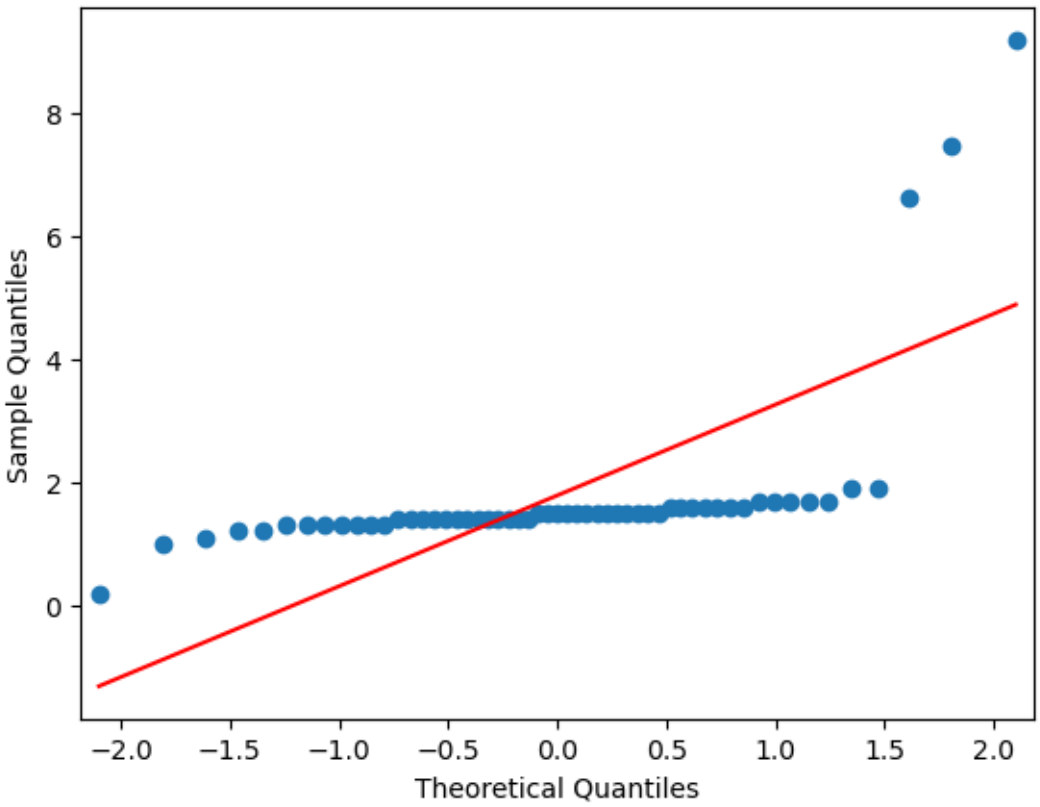

Рисунок 2.3 - График «квантиль-квантиль» для переменной PETALLEN

В заключение для проверки выборки на нормальность используем критерии согласия Шапиро-Уилка, Колмогорова-Смирнова и x^2-Пирсона. Если P-значения этих критериев будут больше, чем 0.05, то мы не отклоним гипотезу о нормальности распределения.

```
KstestResult(statistic=0.32727272727272727, pvalue=0.005240054478222634, statistic_location=1.7, statistic_sign=1)
ShapiroResult(statistic=0.3940703272819519, pvalue=1.0241968677240226e-13)
Power_divergenceResult(statistic=63.87272727272725, pvalue=1.0556358931874558e-08)
```

Все три критерия позволяют нам однозначно отклонить гипотезу о нормальности распределения исследуемой выборки.

Таким образом, по совокупности признаков мы можем отклонить гипотезу о нормальности исследуемого распределения. То есть выборка наблюдений случайной величины PETAL LENGTH с засорениями для ирисов вида SETOSA не является случайной выборкой из нормального распределения.