Санкт-Петербургский государственный университет
Факультет прикладной математики — процессов управления

Лабораторная работа №2

Работу выполнил
Пшеничников Матвей
Группа 22.Б08-пу

Задание 1. Анализ датасета “Babyboom”

Описание: набор данных содержит время рождения, пол и вес при рождении каждого из 44 младенцев, родившихся в течение 24 часов в больнице Брисбена, Австралия.
Анализируемые переменные:

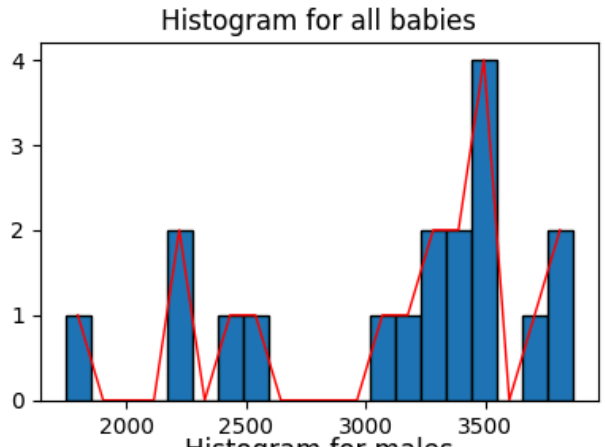
- ➢ Birth weight in grams (Вес при рождении в граммах)
- ➢ Number of minutes after midnight of each birth (Количество минут после полуночи каждого рождения)

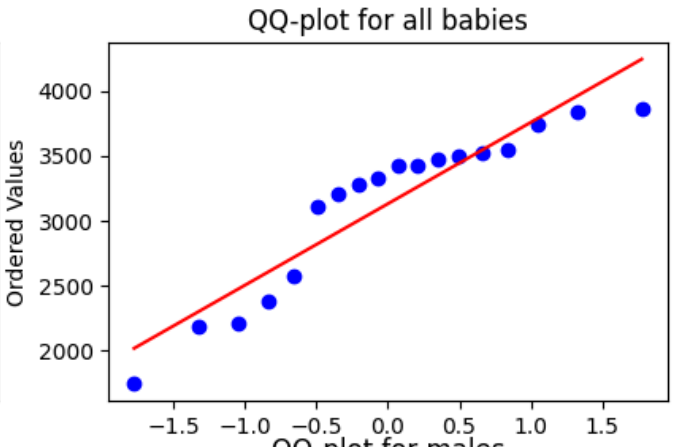
Задание 1.1 Проверьте вес младенцев на нормальность (сначала все данные, затем разделить по полу). При проверке гипотез использовать точечные оценки параметров. Построить доверительные интервалы для параметров нормального распределения.

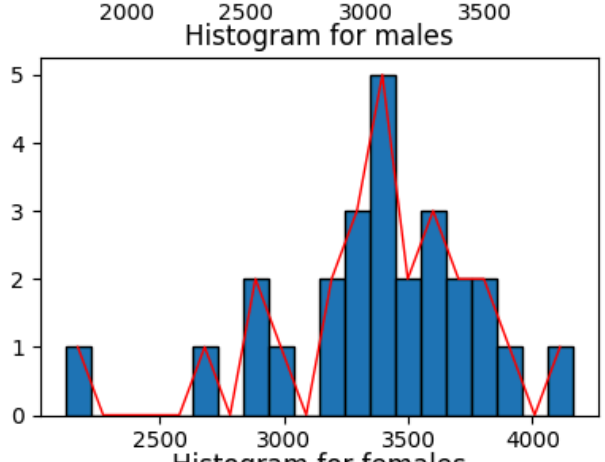
Для проверки на нормальность с использование точечных оценок был применён метод Шапиро-Уилкса с уровнем значимости 0.05. Результаты теста:

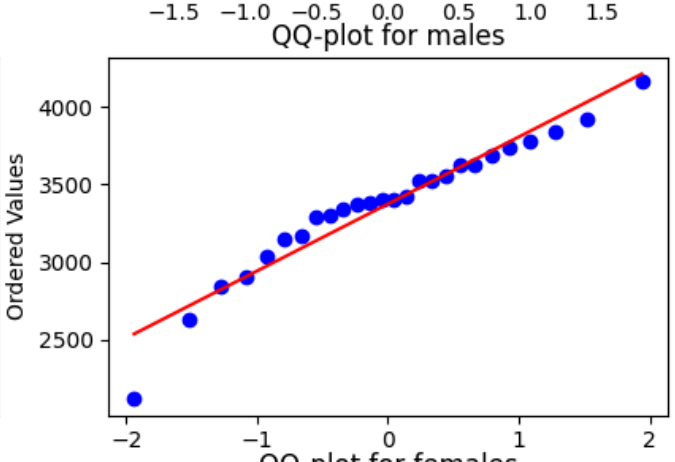
	Statistic	P-value	Is normally distributed
all babies	0.870283	0.0179848	No
males	0.947474	0.202248	Yes
females	0.898723	0.000994397	No

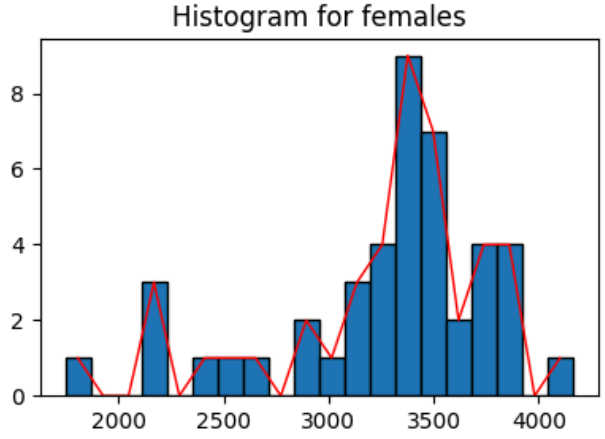
Также для наглядности были построены графики (гистограмма и QQ-plot), которые подтверждают нормальность данных только среди группы мальчиков:

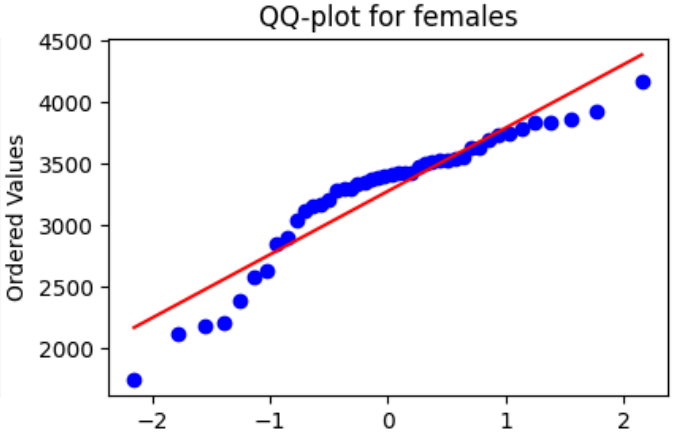

Histogram for all babies
0
1
2
3
4
2000
2500
3000
3500


QQ-plot for all babies
Ordered Values
2000
2500
3000
3500
4000
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Histogram for males
0
1
2
3
4
5
2500
3000
3500
4000


QQ-plot for males
Ordered Values
2500
3000
3500
4000
−2
−1
0
1
2


Histogram for females
0
2
4
6
8
2000
2500
3000
3500
4000


QQ-plot for females
Ordered Values
2000
2500
3000
3500
4000
4500
−2
−1
0
1
2

Были построены доверительные интервалы для параметров нормального распределения (мат. ожидание и дисперсия) с использованием распределения Стьюдента и “Хи-квадрат”.

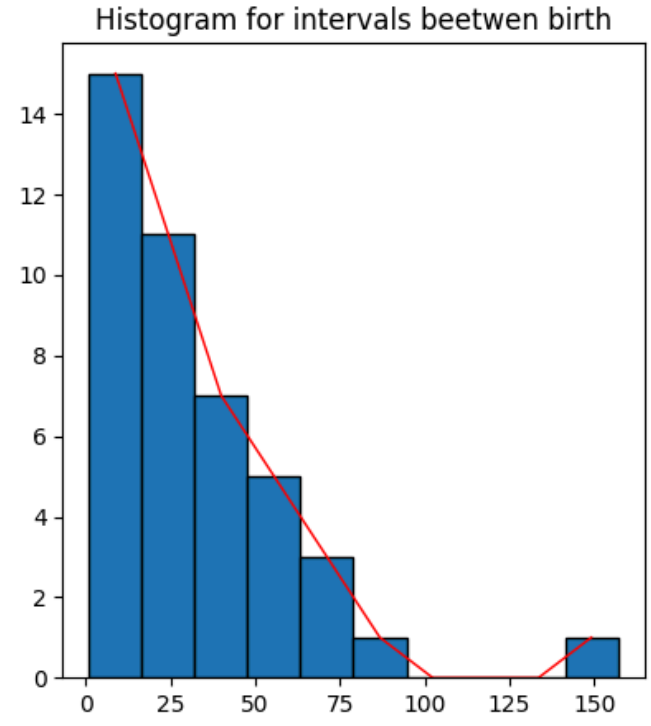
	Mean	Standard deviation
all babies	(2818.3658, 3446.523)	(473.9317, 946.8332)
males	(3202.4162, 3548.1992)	(335.6983, 590.8785)
females	(3115.418, 3436.4911)	(436.2725, 669.0306)

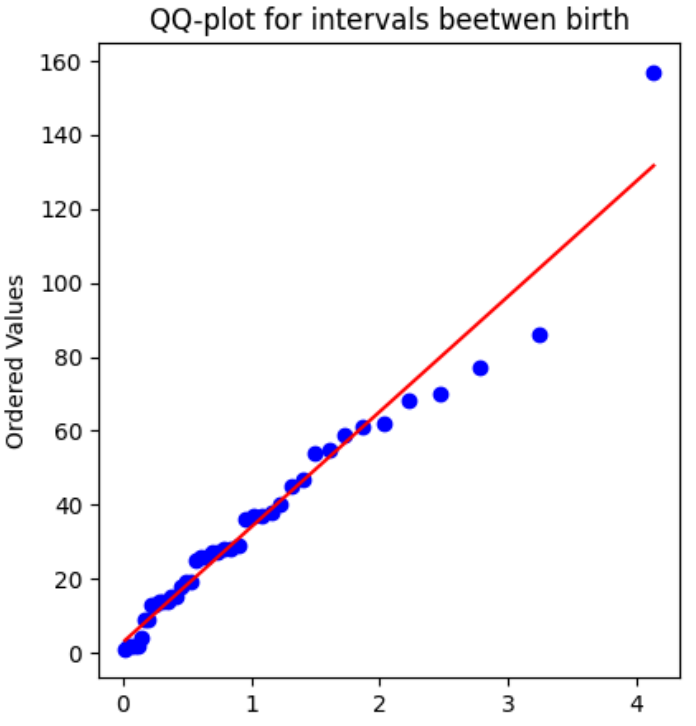
Задание 1.2 Проверить гипотезу о том, что время между рождением детей подчиняется экспоненциальному распределению (используя точечные оценки параметров).

Для проверки на экспоненциальное распределение был использован тест Колмогорова-Смирнова. Результаты теста:

Statistic	0.12461713801129565
P-value	0.47856775876549285
Is exponentially distributed	Yes

Также для наглядности была построена гистограмма и QQ-plot. Их вид подтверждает что время рождения распределено экспоненциально:

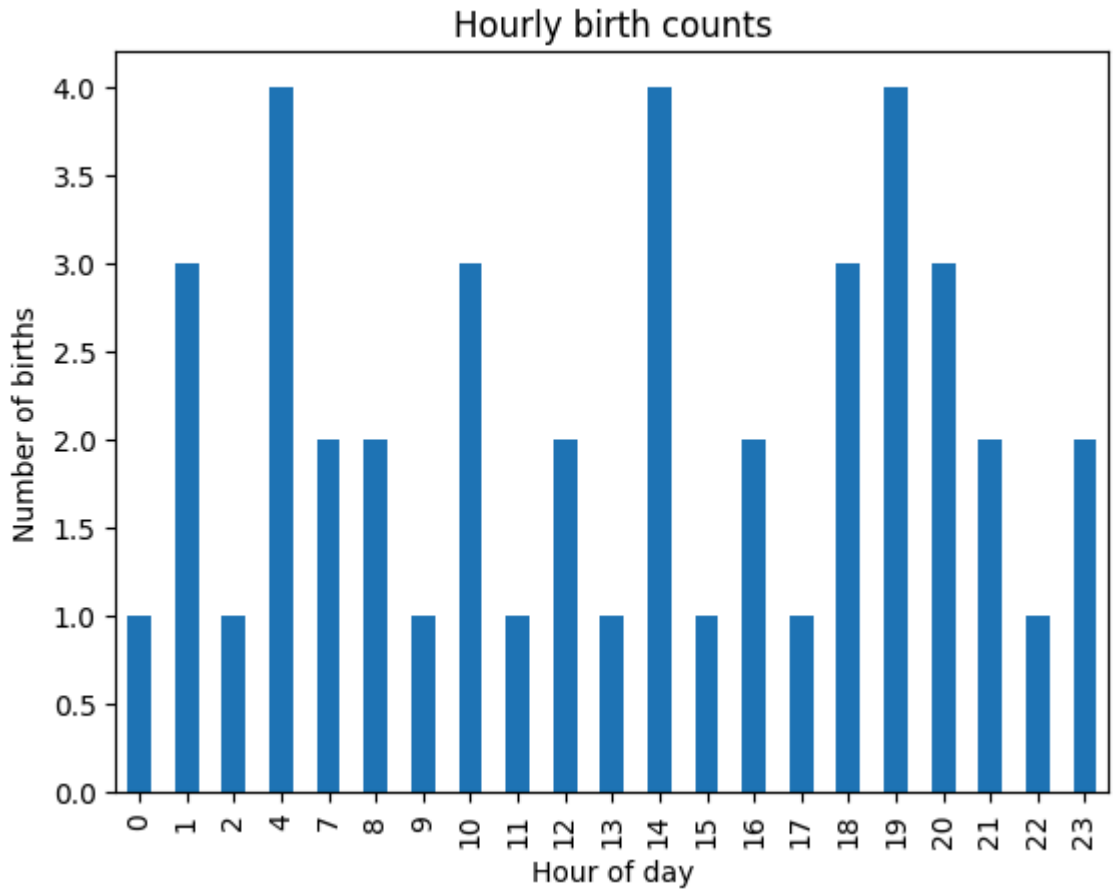




Задание 1.3 Проверить гипотезу, подчиняется ли количество рождений в час для каждого часа распределению Пуассона
гипотеза о распределении Пуассона проверена тестом “Хи-квадрат”. Результаты:

Statistic	3.9092527498131453
P-value	0.4184264420071915
Is it distributed according to Poisson	Yes

График кол-ва рождений в час:



Вывод: не отвергаем гипотезу о распределении интервалов между рождения по Пуассону

Задание 2. Анализ датасета “Euroweight”

Описание: Датасет содержит информацию о весе 2000 евро-монет, измеренном с точностью до миллиграмма в лабораторных условиях.
Анализируемые переменные:

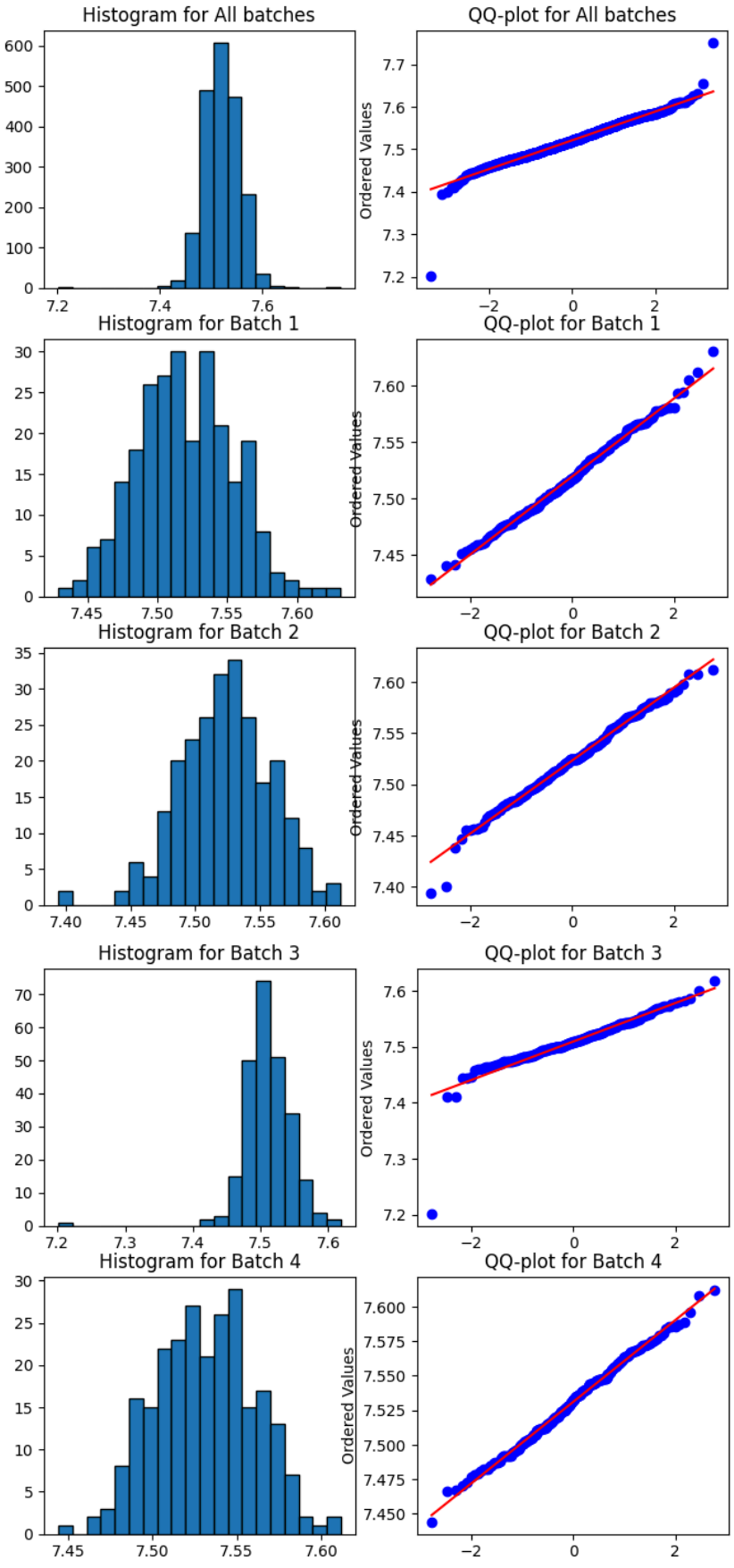
- ➢ Вес монеты (в граммах): измерен с точностью до миллиграмма. Используется для проверки предположения о нормальности распределения.
- ➢ Номер партии (batch): номер упаковки, к которой принадлежит монета. Может использоваться для оценки межпартийной вариативности.
- ➢ Идентификатор (ID): порядковый номер наблюдения, не несёт смысловой нагрузки, но нужен для навигации по данным.

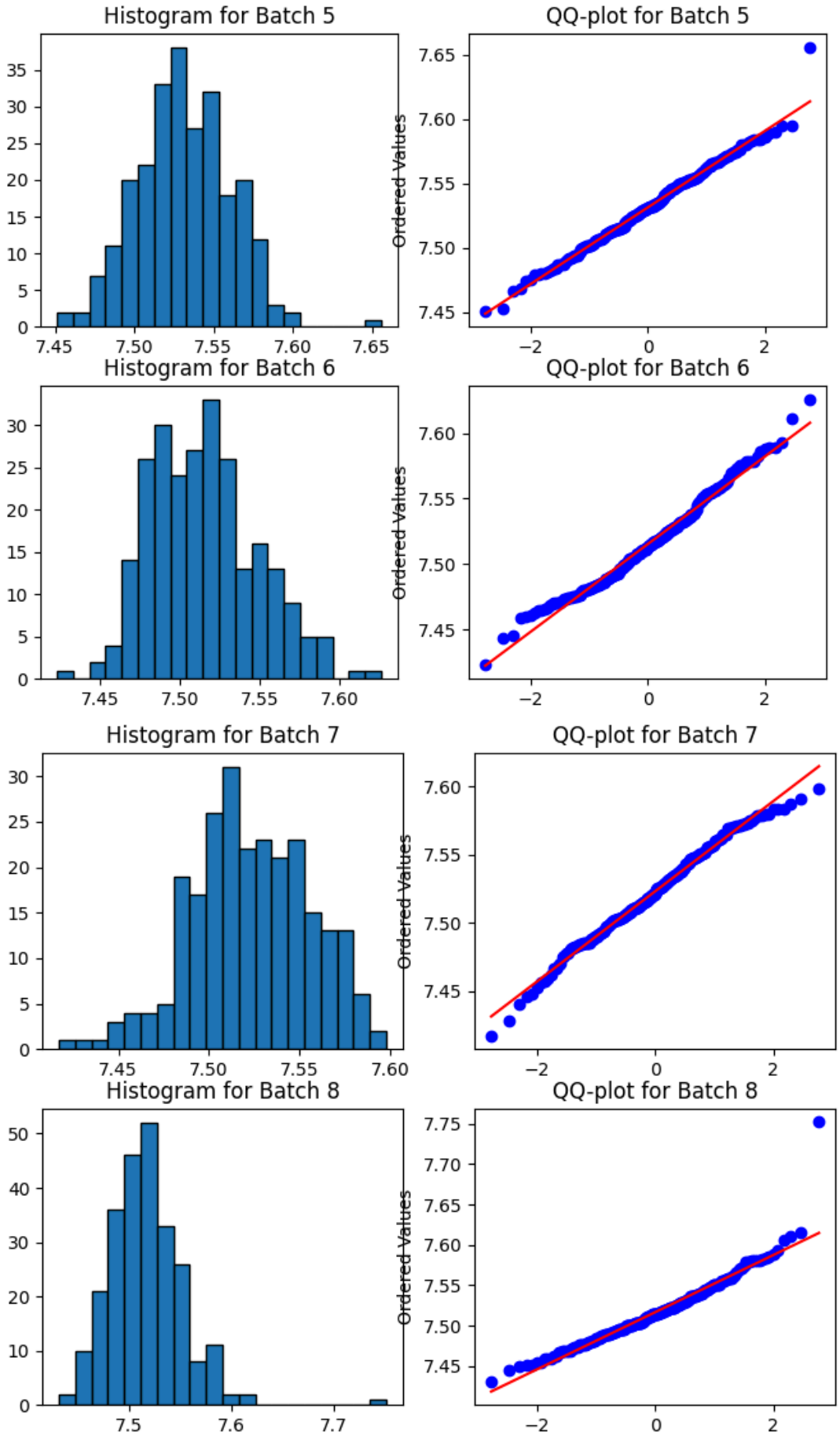
Задание 2.1 Проверить веса монет на нормальное распределение (сначала все, потом по партиям)

Для проверки на нормальное распределение был выполнен тест Шапиро для каждой исследуемой группы. Результаты:

	Statistic	P-value	Is normally distributed
All batches	0.975473	5.02328e-18	No
Batch 1	0.995507	0.683002	Yes
Batch 2	0.9909	0.121877	Yes
Batch 3	0.863432	4.08944e-14	No
Batch 4	0.995505	0.682659	Yes
Batch 5	0.991034	0.128993	Yes
Batch 6	0.984059	0.0067565	No
Batch 7	0.990701	0.111983	Yes
Batch 8	0.93672	6.8277e-09	No

Для подтверждения гипотезы были построены графики (диаграмма и QQ-plot) для каждой группы:


Histogram for All batches
QQ-plot for All batches
Ordered Values
Histogram for Batch 1
QQ-plot for Batch 1
Ordered Values
Histogram for Batch 2
QQ-plot for Batch 2
Ordered Values
Histogram for Batch 3
QQ-plot for Batch 3
Ordered Values
Histogram for Batch 4
QQ-plot for Batch 4
Ordered Values


Histogram for Batch 5
QQ-plot for Batch 5
Ordered Values
Histogram for Batch 6
QQ-plot for Batch 6
Ordered Values
Histogram for Batch 7
QQ-plot for Batch 7
Ordered Values
Histogram for Batch 8
QQ-plot for Batch 8
Ordered Values

Вывод: отвергаем гипотезу о нормальности для выборок, состоящих из всех монет и из монет партий 3, 6 и 8. Это подтверждается как точечными оценками (тест Шапиро), так и графиками (выбросы на QQ-plot и смещённость/вытянутость диаграмм).

Задание 2.2 Построить доверительные интервалы для параметров нормального распределения.

Были построены доверительные интервалы для параметров нормального распределения (мат. ожидание и дисперсия) с использованием распределения Стьюдента и “Хи-квадрат”. Результаты:

	Mean	Standard deviation
All batches	(7.5197, 7.5227)	(0.0333, 0.0355)
Batch 1	(7.5154, 7.5239)	(0.0316, 0.0377)
Batch 2	(7.5187, 7.5276)	(0.0326, 0.0389)
Batch 3	(7.5049, 7.5142)	(0.0341, 0.0406)
Batch 4	(7.5274, 7.5348)	(0.027, 0.0322)
Batch 5	(7.5277, 7.5351)	(0.0272, 0.0325)
Batch 6	(7.5111, 7.5194)	(0.0307, 0.0366)
Batch 7	(7.5189, 7.5271)	(0.0303, 0.0362)
Batch 8	(7.5122, 7.5213)	(0.0334, 0.0399)

Задание 3. Анализ датасета “Iris”

Описание: набор данных содержит информацию о характеристиках цветков трёх видов ирисов: *Iris Setosa*, *Iris Versicolour* и *Iris Virginica*.

Переменные:

- Sepal length (длина чашелистика)
- Sepal width (ширина чашелистика)
- Petal length (длина лепестка)
- Petal width (ширина лепестка)

Задание 3.1 Проверить гипотезу с помощью точечных оценок параметров о нормальном распределении длины цветков, сгруппировав их по типу ириса

Для проверки на нормальное распределение был выполнен тест Шапиро для каждой исследуемой группы. Результаты:

	Statistic	P-value	Is normally distributed
Iris-virginica	0.962186	0.109775	Yes
Iris-setosa	0.954946	0.0546505	Yes
Iris-versicolor	0.966004	0.158478	Yes

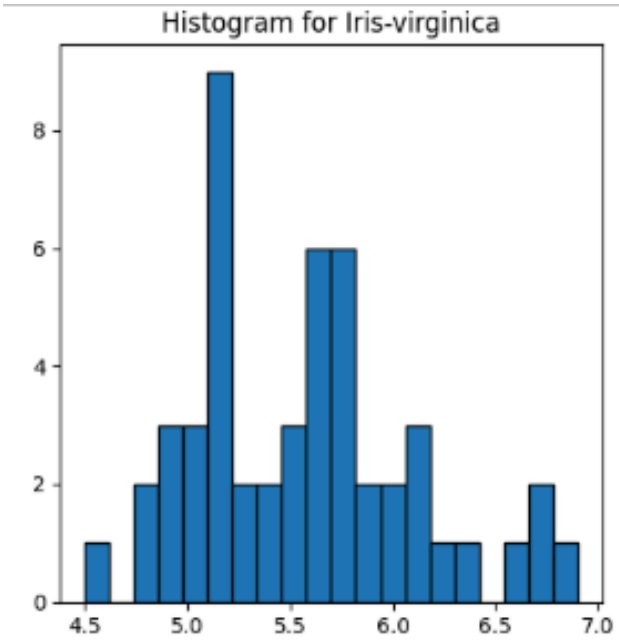
Также для подтверждения гипотезы для каждого типа ириса были построены графики (см. ниже).

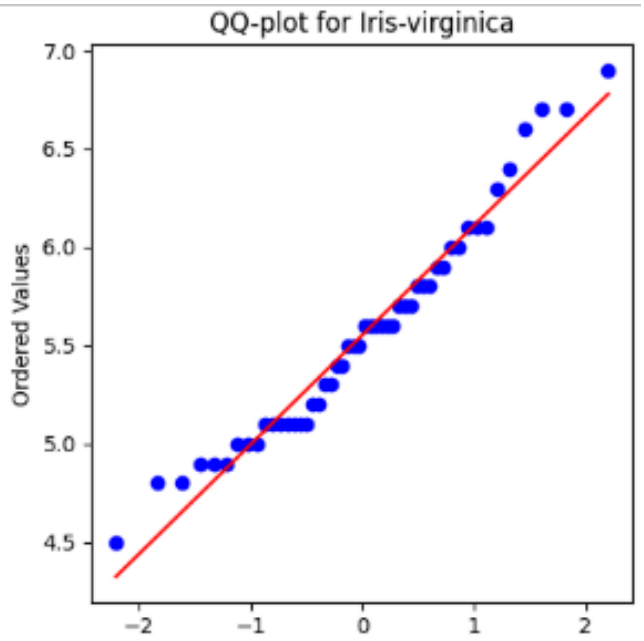
Вывод: длины цветков у ириса каждого типа подвержены нормальному распределению. Это следует как из точечных оценок параметров, так и из построенных графиков.

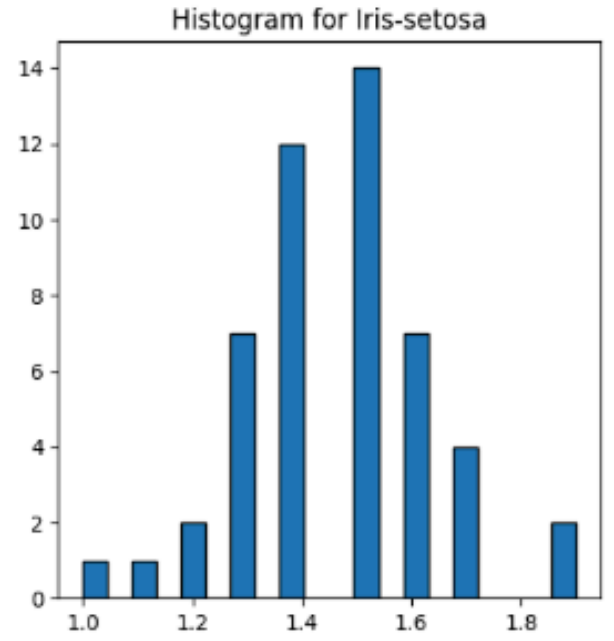
Задание 3.2 Построить доверительные интервалы для параметров нормального распределения

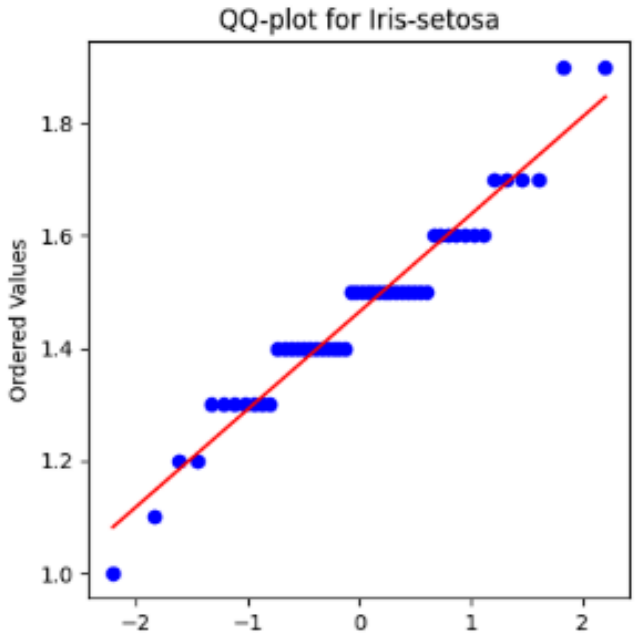
Были построены доверительные интервалы для среднего значения и стандартного отклонения длины цветка для каждого типа ириса. Результаты:

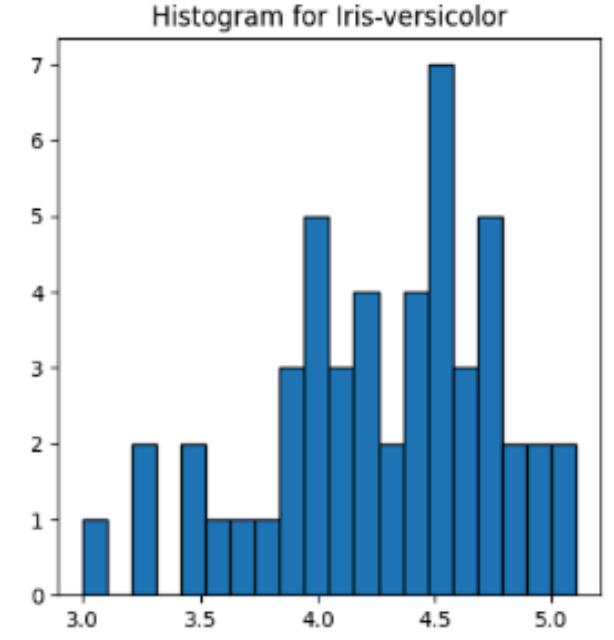
	Mean	Standard deviation
Iris-virginica	(5.3952, 5.7088)	(0.461, 0.6877)
Iris-setosa	(1.4147, 1.5133)	(0.1449, 0.2162)
Iris-versicolor	(4.1265, 4.3935)	(0.3925, 0.5856)

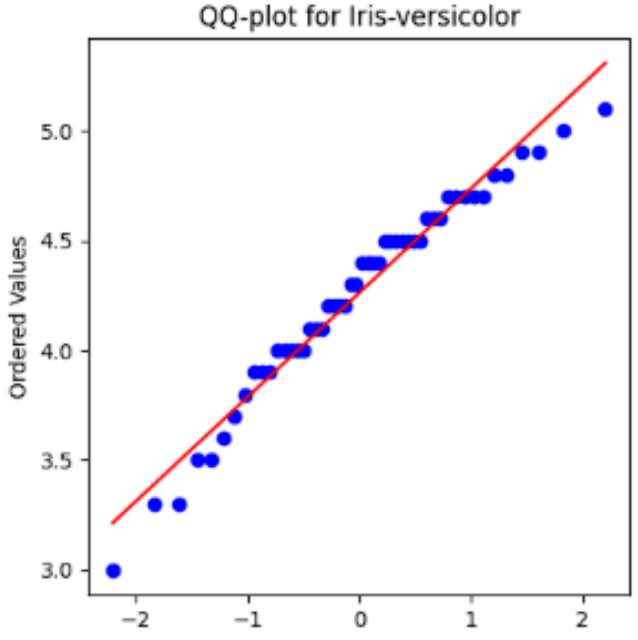

Histogram for Iris-virginica


QQ-plot for Iris-virginica
Ordered Values


Histogram for Iris-setosa


QQ-plot for Iris-setosa
Ordered Values


Histogram for Iris-versicolor


QQ-plot for Iris-versicolor
Ordered Values