Министерство образования Республики Беларусь

Учреждение образования

Брестский государственный технический университет

Кафедра ИИТ

Лабораторная работа №7

За 7 семестр

По дисциплине «Компьютерное моделирование и анализ данных»

Выполнил: студент 4 курса

Группы ПО-4(2)

Коташевич С.Н.

Проверил: Чичурин А.В.

Брест 2022

**Лабораторная работа №7**

**Элементарная описательная статистика**

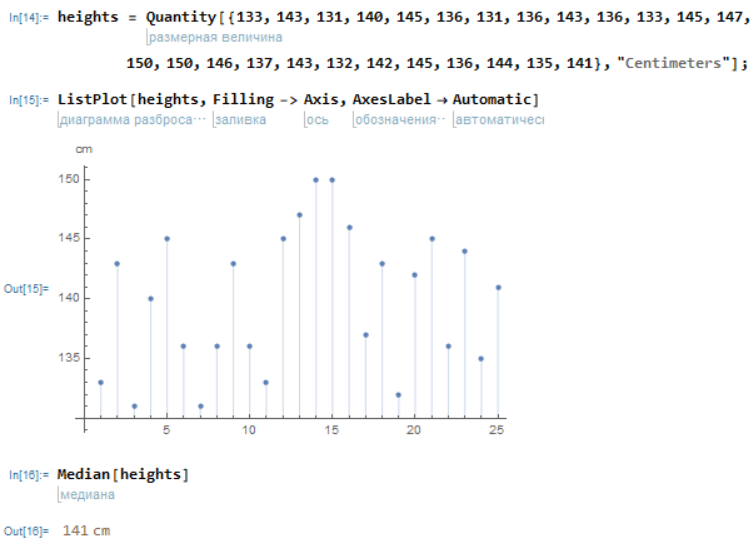
Функции описательной статистики Wolfram работают как с явными данными, так и с символическими представлениями статических распределений. При работе с явными данными функции обычно обрабатывают огромные наборы данных, которые могут содержать не только числа, но и символьные элементы, представляющие, например, параметризованные или неизвестные данные.

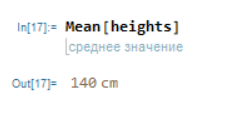
Описательная статистика относится к свойствам распределений, такими как местоположение, дисперсия и форма.

Статистика местоположения описывает, где находятся данные.

*Среднее значение.*

Функция **mean** используется для вычисления среднего арифметического значения элементов списка или распределения. Пример вычисления среднего значения роста учеников в кассе в сантиметрах.





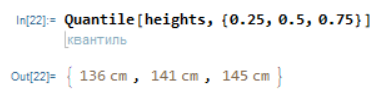
*Медиана.*

Функция **median** используется для нахождения медианы (срединное значение) элементов в списке, т.е. значение, находящееся в середине этого набора данных (половина элементов набора меньше него, половина набора элементов больше него) Пример нахождения медианы роста учеников в классе из примера выше.



*Квантиль.*

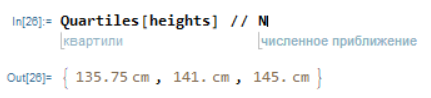
Функция **Quantile** даёт квантиль определенного уровня списка значений. Квантиль – значение, которое заданная случайная величина не превышает с фиксированной вероятностью. В данном примере находятся 0.25 квантиль (первый квартиль), 0.5 квантиль (второй квартиль) и 0.75 квантиль (третий квартиль).



Можно сказать, что 25% значений меньше 136 см, 50% меньше 141 см и 75% меньше 145 см.

*Квартиль.*

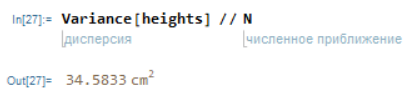
Функция **Quartiles** выдает все квартили списка значений. Квартиль – тип квантиля, который делит количество точек данных на четыре части или четверти более или менее одинакового размера. Пример все тот же. Первая квартиль – среднее между наименьшим числом и медианой. (25% данных находятся ниже этой точки) Второй квартиль – медиана данных (50% данных находятся ниже этой точки). Третий квартиль – среднее между медианой и наибольшим значением (75% данных находятся ниже этой точки).



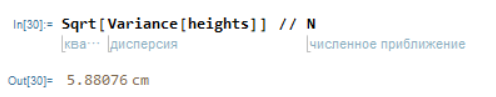
Статистика дисперсии суммирует разброс данных. Большинство таких функций описывают отклонение от определенного местоположения.

*Дисперсия.*

Функция **Variance** возвращает дисперсию элементов набора данных либо распределения. Дисперсия – мера того, насколько далеко набор чисел отличается от их среднего значения. Набор данных для примера тот же.

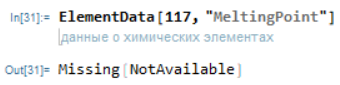


Квадратный корень из дисперсии равен среднеквадратическому отклонению.

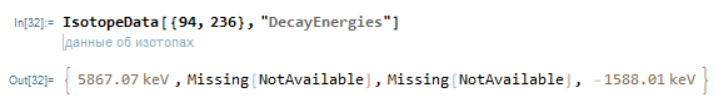


**Замена или удаление недостающих или недействительных данных**

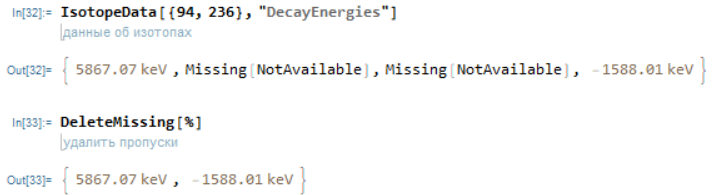
В Wolfram представления недостающих данных используется объект **Missing**. Как пример попытаемся через функцию *ElementData* получить свойство *MeltingPoint* 117 элемента.



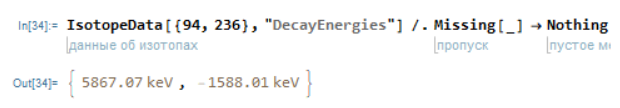
Для замены или удаления недостающих данных существует много функций. Для примера получим с помощью функции *IsotopeData* значения энергии распада для изотопов с атомным числом 94 и массовым числом 236.



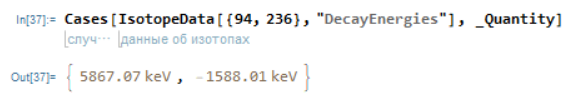
Как видим, получаем элементы, недостающие элементы. Удалим их с помощью функции **DeleteMissing**.



Можно заменить все сущности **Missing на Nothing** – элементы списка, которые автоматически удаляются.



Можно выбрать корректные данные из всего набора данных. Используем функцию **Сases**, которая выдает данные из нашего списка, которые удовлетворяют шаблону **Quantity**.



С другой стороны, вместо выборки только корректных данных, можно удалить данные, которые представляют недостающие данные, т.е. Missing. Делается через функцию **DeleteCases**, которая удаляет все элементы из исходного набора данных, которые удовлетворяю шаблону **Missing**.



Если в нашем наборе данных нет недостающих данных, но есть недействительные данные или так называемые аномалии, то от них можно избавиться с помощью функции **DeleteAnomalies**. Она изучает передаваемые данные и удаляет элементы, которые считает аномальными.

Пример с явными недействительные данными.

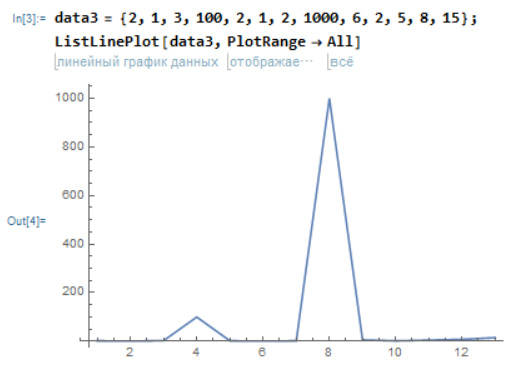
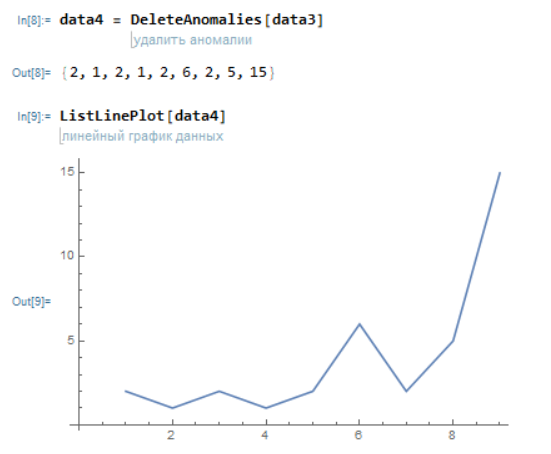


График и набор данных уже после применения функции **DeleteAnomalies**.



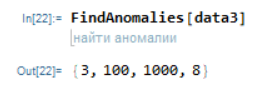
Как можно заметить, недействительные дынные были удалены и график уже стал похож на нормальный.

С другой стороны, если нам необходимо просто получить недействительные данные для дальнейшего анализа, а не удалять их, то можно использовать функцию **FindAnomalies**. Он возвращает набор данных, которые считает аномальными для исследуемого набора.

Для выше представленного набора данных

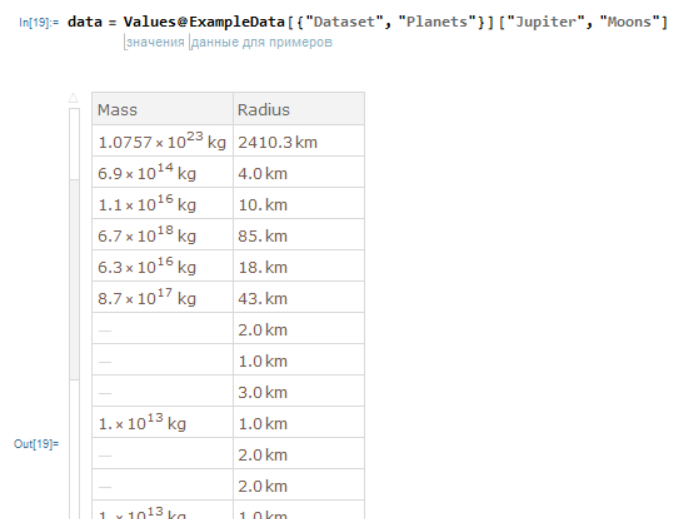


Аномальными будут

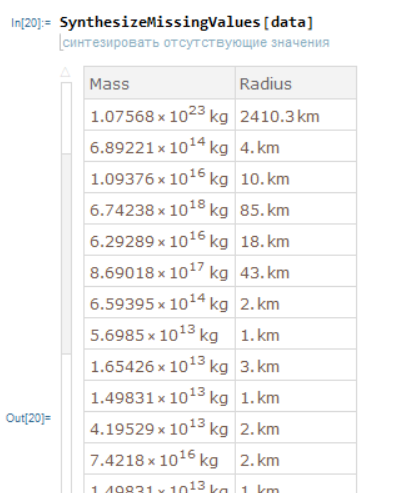


Из предыдущего примера можно убедиться, что в наборе данных после удаления аномалий как раз отсутствуют элементы, полученные недавно.

Вместо удаления недостающих данных, можно заменить их на какое-нибудь сгенерированное значение, которое будет похоже не значения из исходной выборки. Для этого используется функция **SynthesizeMissingValues**. Получим набор данных, связанный со свойствами спутников Юпитера.



Как можно заметить, присутствуют недостающие данные. Заменим их.

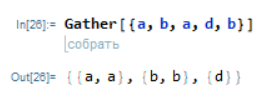


Все недостающие данные были заменены сгенерированными значениями.

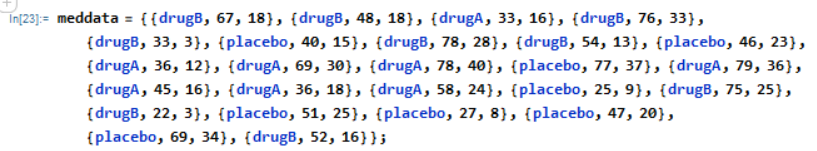
**Выполнение операций над подгруппами данных**

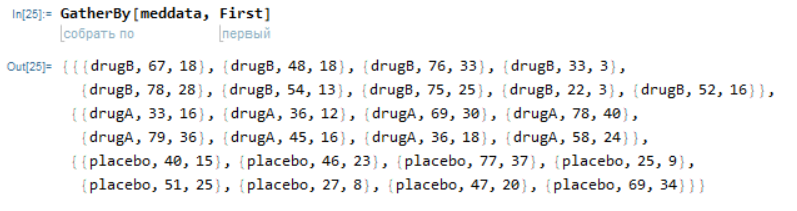
При обобщении данных часто полезно анализировать их по подгруппам. В Wolfram имеются функции для выполнения реорганизации списков данных любой структуры и любым количеством элементов.

Функция **Gather** объединяет элементы группы данных в подгруппы идентичных элементов. Простой пример объединения.

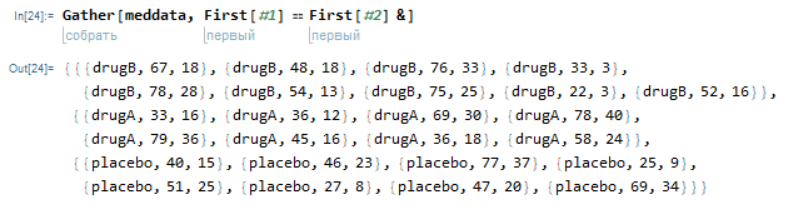


Функция **GatherBy** объединяет элементы группы данных в подгруппы, дающий одинаковый результат выполнения какой-либо функции. Пример объединения подгрупп, содержащих данные о типе препарата, возраст пациента и время восстановления, по типу препарата.

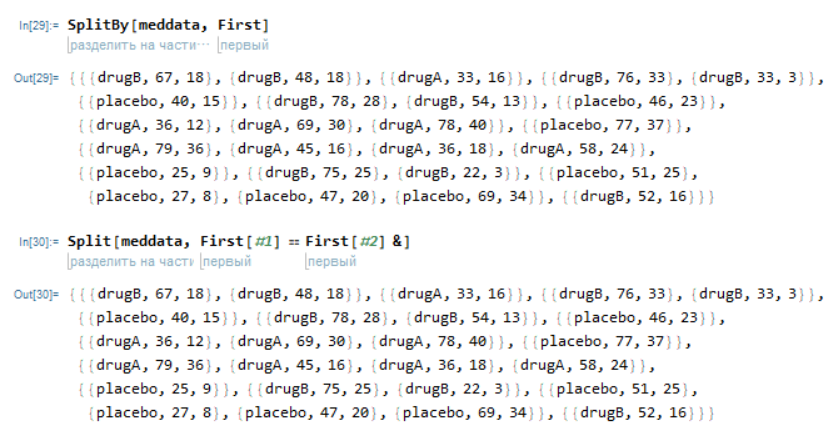




Идентично результата можно добиться и через **Gather**, только нужно указать тестовый параметр, по которому подгруппы будут считаться идентичными.



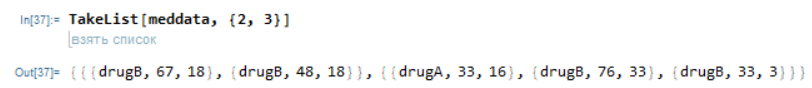
Функция **Split** используется для разделения серии идентичных элементов группы на подгруппы. Можно использовать тестовый параметр. Аналогично работает функция **SplitBy**.



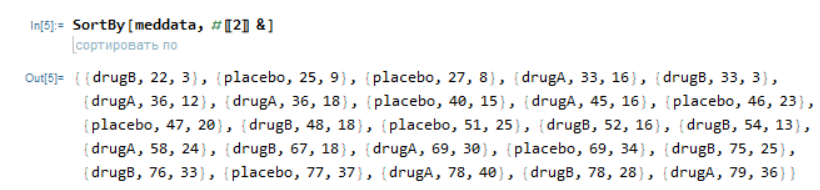
Функция **TakeList** позволяет взять конкретное количество подгрупп. Здесь берем первые две подгруппы.



Во втором случае функция возвращает первые две подгруппы, а затем следующие за ней три подгруппы.



Используя функцию **SortBy** можно отсортировать набор значений. Отсортируем наши подгруппы по возрасту пациента.



Можно удалить подгруппы, имеющие одинаковое значение определенных элементов. Для этого используются функции **DeleteDublicates** и **DeleteDublicatesBy**. Удалим подгруппы, имеющие одинаковое значение времени восстановления.

