Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

Лабораторная работа №7

за 7 семестр

По дисциплине: «КМиАД»

Выполнила: Студентка 4 курса

Группы ПО-4(2)

Шиба А.А.

Проверил: Чичурин А. В.

2022

**Статистический анализ данных**

Wolfram Language объединяет многие аспекты статистического анализа данных, от получения и изучения данных до построения высококачественных моделей и вывода последствий.

Wolfram Language предоставляет несколько способов получения данных, начиная со встроенных кураторских источников данных, импорта из различных форматов файлов или подключения к базам данных. Базовая обработка данных, включая вычисление статистических величин, сглаживание, тестирование и визуализацию, дает первый уровень анализа.

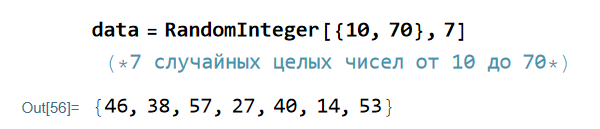
Добавляя в микс модели, такие как модели распределения или регрессии, можно ответить на более широкий круг вопросов анализа или даже предоставить возможности прогнозирования.

**Описательная статистика**

Описательные статистические функции Wolfram Language работают как на явных данных, так и на символьных представлениях статистических распределений.

При работе с явными данными функции регулярно обрабатывают огромные наборы данных, которые могут содержать не только числа, но и символьные элементы, представляющие, например, параметризованные или неизвестные данные.

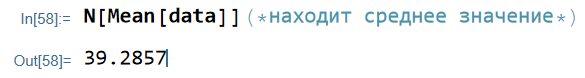
Наиболее часто статистические программы используются для генерации случайных чисел и массивов данных с ними. Для отдельных псевдослучайных реальных чисел предусмотрена функция RandomReal[]. При каждом обращении к этой функции генерируется случайное (точнее, псевдослучайное) число в интервале от 0 до 1 с равномерным распределением. Функция RandomReal[xmin, xmax] генерирует случайное число в заданном интервале изменения переменной x. Также есть функция для генерации целых чисел RandomInteger, работающая по тому же принципу.



1. Статистика местоположения (**Mean**, **Median**)

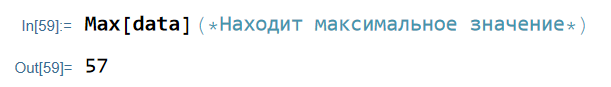
Статистика местоположения описывает, где находятся данные.

При наличии списка с n элементами  *среднее* значение Mean[list] определяется как .



*Медиана* [Median](https://reference.wolfram.com/language/ref/Median.html)[*list*] эффективно дает значение в середине отсортированной версии *списка*. Его часто считают более надежной мерой центра распределения, чем среднее, поскольку оно меньше зависит от отдаленных значений.

1. Статистика дисперсии(**Variance**, **StandartDeviation**)
2. Статистика заказов(**Max**, **Quantile**, **Quartiles**)



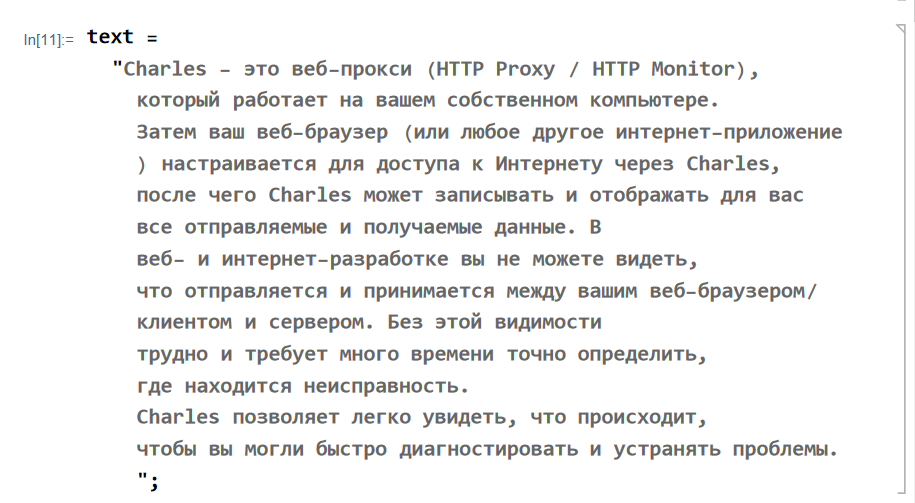
1. Символические особенности дистрибутивов(**PDF**, **CharacteristFunction**)
2. Статистическая визуализация(**Histogram&Plot**)

Массивы данных часто представляются гистограммами по столбцам или строками. Если массив одномерный — его представляют столбиковыми гистограммами. При этом данные по горизонтали разбиваются на N участков по числу столбцов (или строк) будущей гистограммы. В каждом из частичных участков подсчитывается сумма данных — число, и оно определяет высоту (или длину) столбца. Таким образом, проводится первичная статистическая обработка данных.

**Работа с текстом**

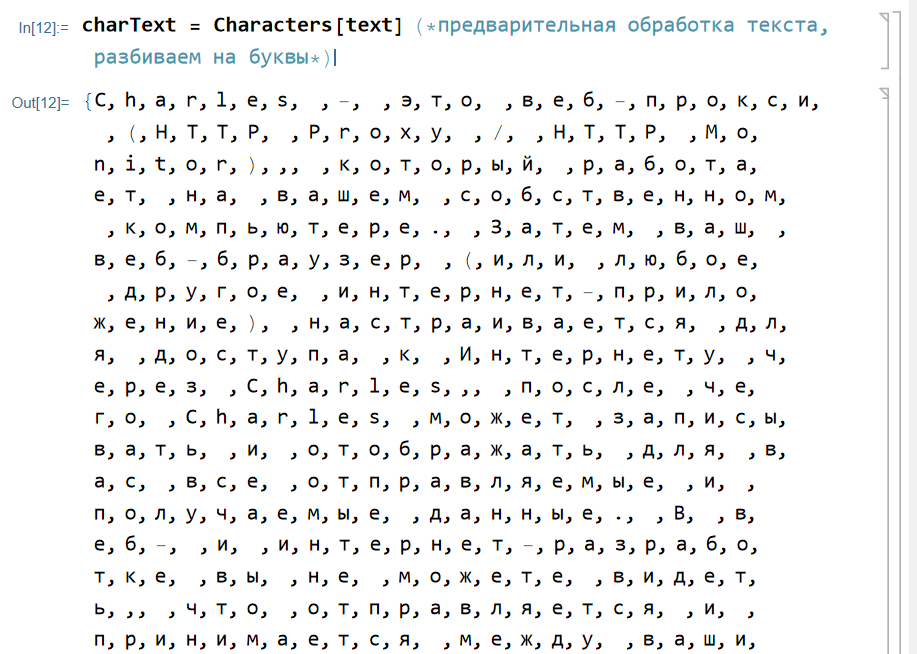
Для аналитиков нередко бывает полезно иметь возможность быстро обработать текст и выявить наиболее часто встречающиеся слова, определить тему статьи или проверить текст на содержание запрещенных выражений. Средства Wolfram Mathematica позволяют пользователям работать с текстом максимально простыми и понятными функциями.

Возьмем для примера любой текст:

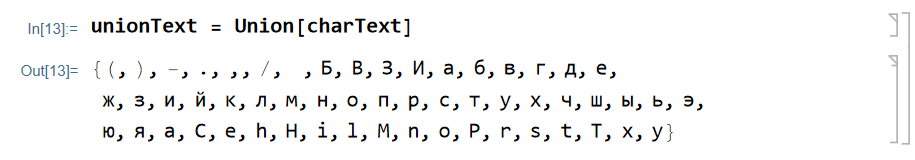


Попробуем поработать с частотой встречания букв и символов в тексте.

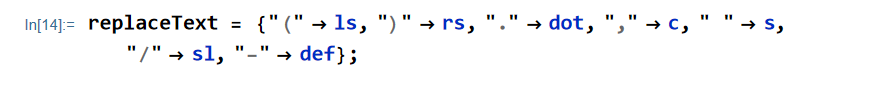
Для этого необходимо разбить текст на отдельные символы. Это делает команда Characters["string"].



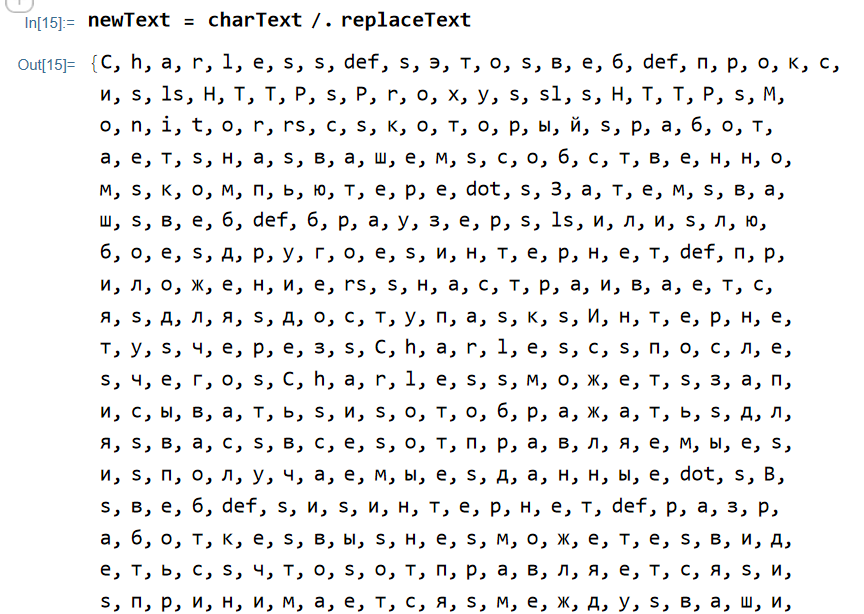
Далее определим множество букв и символов, встречаемых в тексте. Для этого воспользуемся командой Union[Subscript[list, 1],Subscript[list, 2],…] - дает отсортированный список всех отдельных элементов, которые появляются в любом из сабскриптов.



Чтобы было удобнее работать с символами, дадим им другие названия.



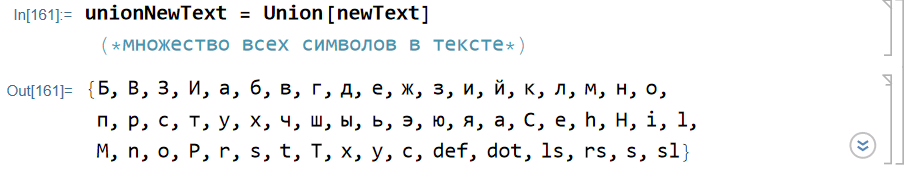
Теперь применим наше новое правило к результату, полученному после разбивания текста на отдельные символы.



Узнаем длину полученного списка.



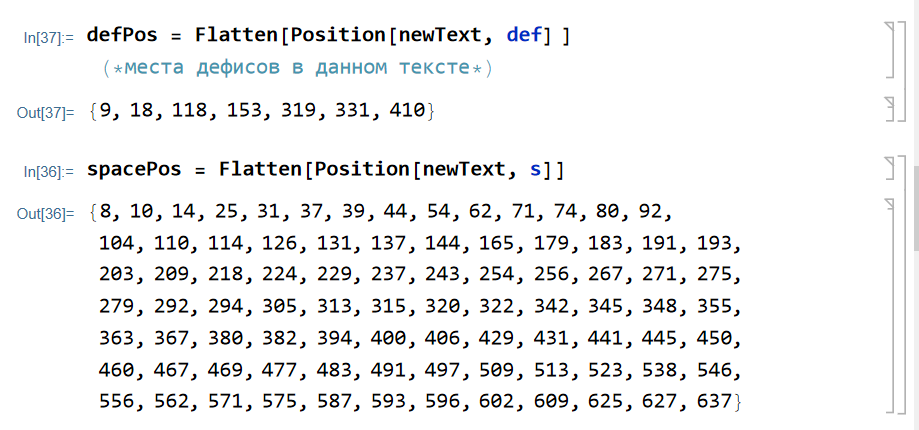
Построим множество символов еще раз.



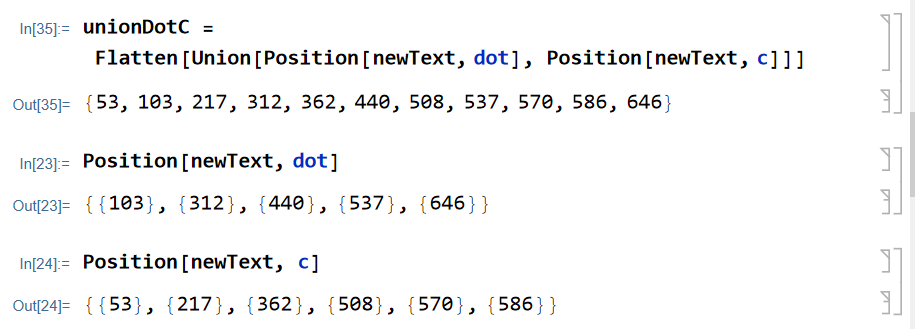
И узнаем длину полученного списка.



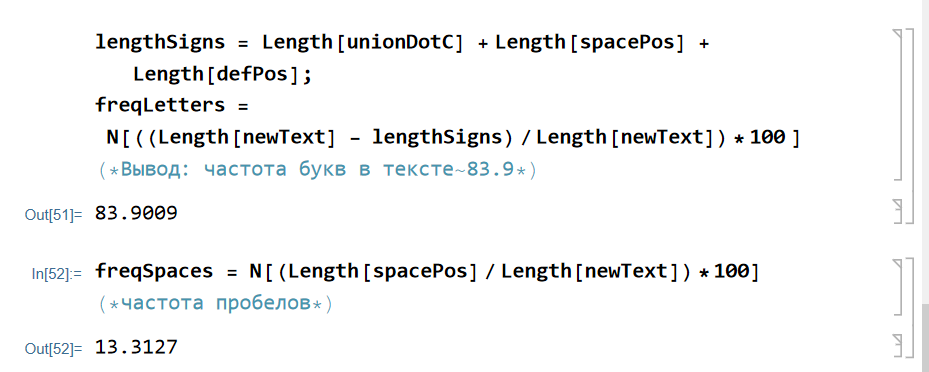
Далее найдем позиции в тексте, на которых находятся дефисы и пробелы.



Объединим позиции точек и запятых и получим список.



Теперь мы можем узнать частоту встречания букв и пробелов в тексте. Для этого вычтем из количества всех символов в тексте количество всех знаков препинания и пробелов и поделим полученный результат на длину всего текста. Чтобы получить значение в процентах, умножим на 100. Аналогично с пробелами.



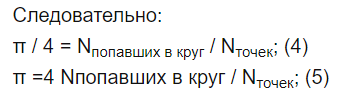
**Метод Монте-Карло. Расчет числа Пи**

Существует много способов вычисления числа Пи. Самым простым и понятным является численный метод Монте-Карло, суть которого сводится к простейшему перебору точек на площади. Суть расчета заключается в том, что мы берем квадрат со стороной a = 2 R, вписываем в него круг радиусом R. И начинаем наугад ставить точки внутри квадрата. Геометрически, вероятность P1 того, что точка попадет в круг, равна отношению площадей круга и квадрата:



Вероятность попадания точки в круг можно также посчитать после численного эксперимента ещё проще: посчитать количество точек, попавших в круг, и поделить их на общее количество поставленных точек:





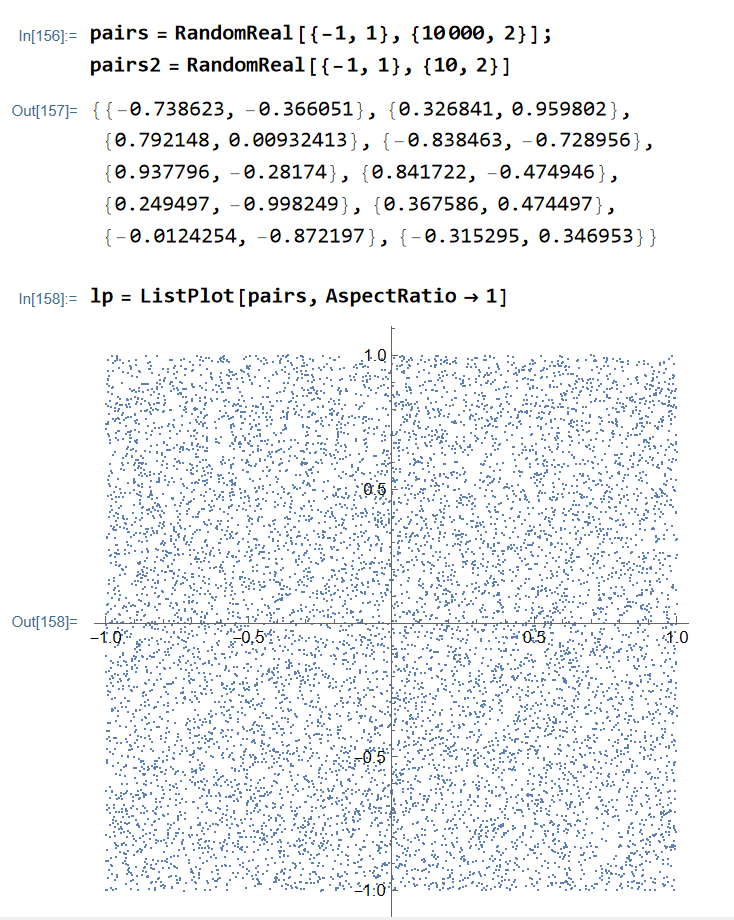
НО! При моделировании мы применяем псевдослучайные числа, которые не являются случайным процессом.

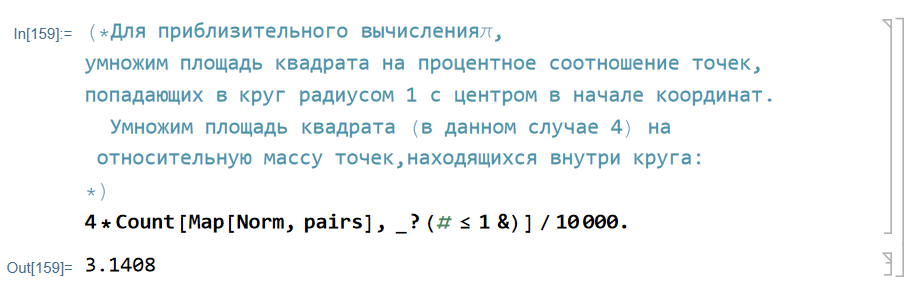
Поэтому, выражение (5), к сожалению, строго не выполняется.

Моделирование по методу Монте Карло позволяет вычислить множество значений. Используя эти значения, определяется искомый результат путем вычисления среднего арифметического или диапазон, в котором может находиться нужный результат.

К примеру, бегун способен пробежать дистанцию в 10 км за 50 мин. Означает ли это, что 20 км он пробежит за 1 час 40 минут? Конечно же, нет: человек – не машина. Если 10 км можно пробежать без остановки, то расстояние вдвое больше требует значительных затрат энергии. Так, необходимо замедлиться, чтобы попить воды, завязать шнурки. При сильном учащении пульса – перейти на шаг или легкий бег и т.д.

Таким образом, прогнозирование времени путем простых математических расчетов – способ, который даст весьма неточный результат. Правильнее будет отобрать в случайном порядке результаты забегов нескольких спортсменов (чем больше, тем лучше) такого же пола, примерно того же возраста и уровня подготовки, которые несколько раз бежали дистанции по 20 км, и вычислить среднее арифметическое результатов. Тогда мы получим значение, на которое можно ориентироваться.





**Прогноз погоды**

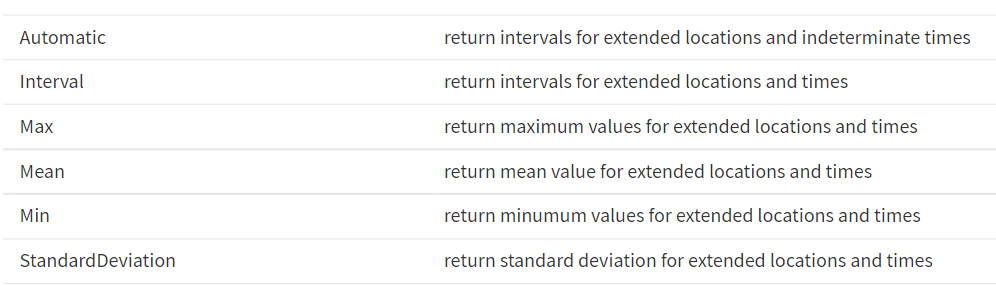
WeatherForecastData[loc] дает самый последний прогноз для всех свойств прогноза погоды для указанного местоположения.

Данные WeatherForecastData собираются из различных источников, включая Глобальную систему прогнозов и Североамериканскую систему мезомасштабных прогнозов.

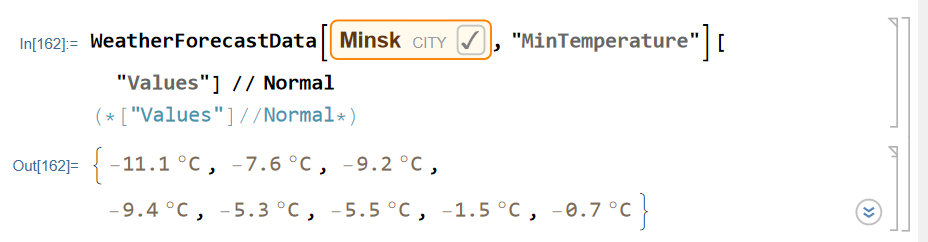
Дата может быть указана как DateObject или {start,end,step}, где start и end оба являются DateObject(например, tomorrow).

Параметры:









**Вывод:** Изучила теоретический материал и провёл статистический анализ данных в системе Wolfram Mathematica.