### Центральная предельная теорема и статистический анализ данных в Python

#### **Рассматриваемые темы:**

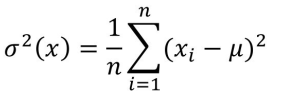
* Линейная регрессия;
* Центральная предельная теорема и статистический анализ данных в python;
* Виды распределений. Собственные вектора.

#### **Нахождение зависимости случайных величин**

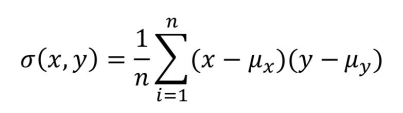
##### **Основные понятия:**

**Дисперсия** — квадратный корень среднеквадратичного отклонения от среднего

значения (насколько данные разбросаны)



**Ковариация** — наличие зависимости между величинами



**Ковариация** — это дисперсия, если две переменных — одна и та же x.

Ковариация не равна нулю — можно предположить зависимость.

#### **Корреляция Пирсона**

Коэффициент корреляции - это статистический показатель зависимости двух случайных величин. Коэффициент корреляции может принимать значения от -1 до +1. При этом, значение -1 будет говорить об отсутствии корреляции между величинами, 0 - о нулевой корреляции, а +1 - о полной корреляции величин. Т.е., чем ближе значение коэффициента корреляции к +1, тем сильнее связь между двумя случайными величинами.

Критерий корреляции Пирсона – это метод статистики, позволяющий определить наличие или отсутствие линейной связи между двумя количественными показателями, а также оценить ее тесноту и статистическую значимость.

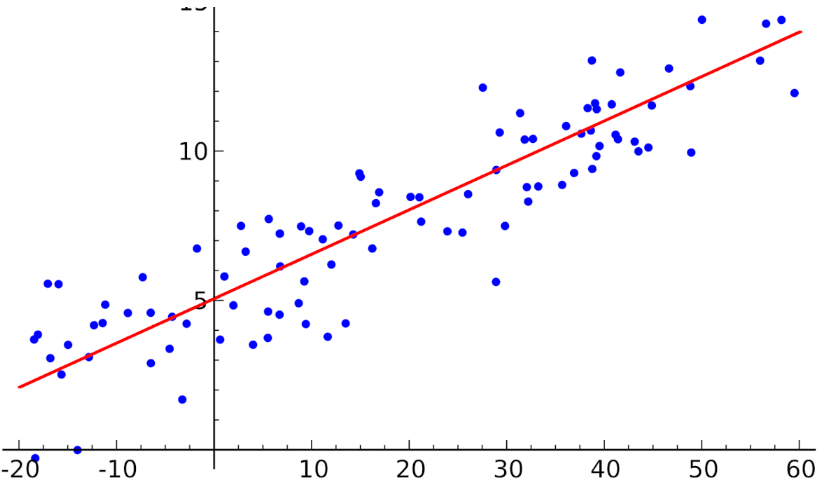
Другими словами, критерий корреляции Пирсона позволяет определить, есть ли линейная связь между изменениями значений двух переменных.

#### **Линейная регрессия**

**Линейная регрессия** — модель зависимости переменной x от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости.

Линейная регрессия относится к задаче определения «линии наилучшего соответствия» через набор точек данных и стала простым предшественником нелинейных методов, которые используют для обучения нейронных сетей. В этой статье покажем вам примеры линейной регрессии.

Цель линейной регрессии — поиск линии, которая наилучшим образом соответствует этим точкам.



#### **Функция потерь**

**Функция потерь** — мера количества ошибок, которые линейная регрессия делает на наборе данных

#### **Матрица корреляций**

Если две величины связаны между собой, то между ними есть корреляция.

При большом количестве данных, когда коэффициенты корреляции необходимо последовательно вычислять из нескольких рядов этих данных, для удобства получаемые коэффициенты сводят в таблицы, называемые корреляционными матрицами.

Корреляционная матрица — это таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

Матрица корреляций подсчитывается с помощью формул, которые

показывают как данные зависят друг от друга в пространстве n значений

#### **Транспонирование матрицы**

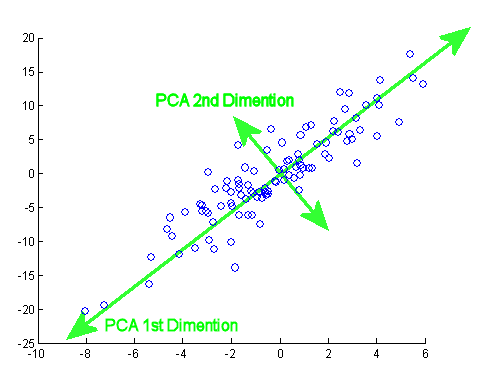
Для того, чтобы получить транспонированную матрицу, нужно в исходной матрице заменить столбцы соответствующими строками по такому принципу: была первая строка – станет первый столбец; была вторая строка – станет второй столбец; была третья строка – станет третий столбец и так далее.

#### **Геометрический смысл ковариационной матрицы**

Ковариационная матрица позволяет подсчитать собственные вектора и собственные значения.

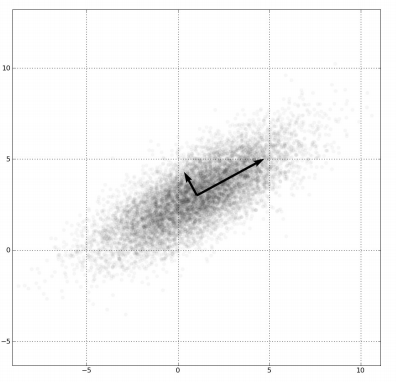
Позволяет найти такой вектор, при проецировании данных на который вариация

максимальна. Этот вектор называется собственный вектор.



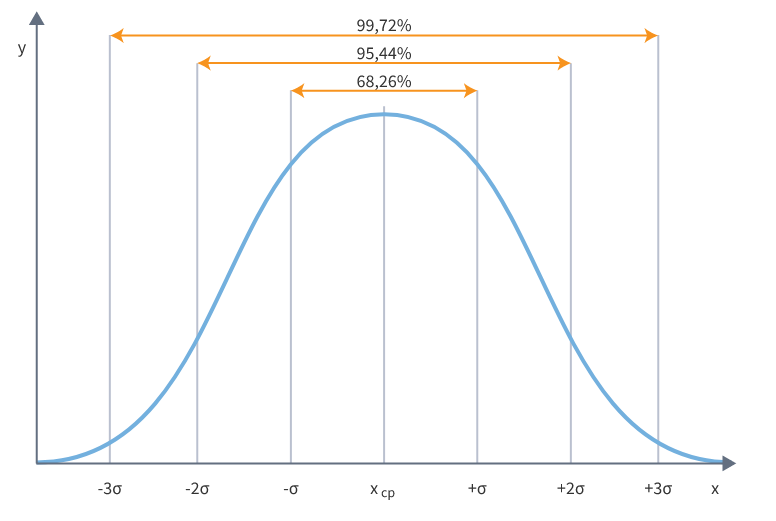
#### **Геометрический смысл собственных векторов**

Это собственные вектора, помноженные на корень квадратный из собственного значения.



#### **Правило трех сигм**

Правило, утверждающее, что вероятность того, что случайная величина отклонится от своего математического ожидания более чем на три среднеквадратических отклонения, практически равна нулю. Правило справедливо только для случайных величин, распределенных по нормальному закону.

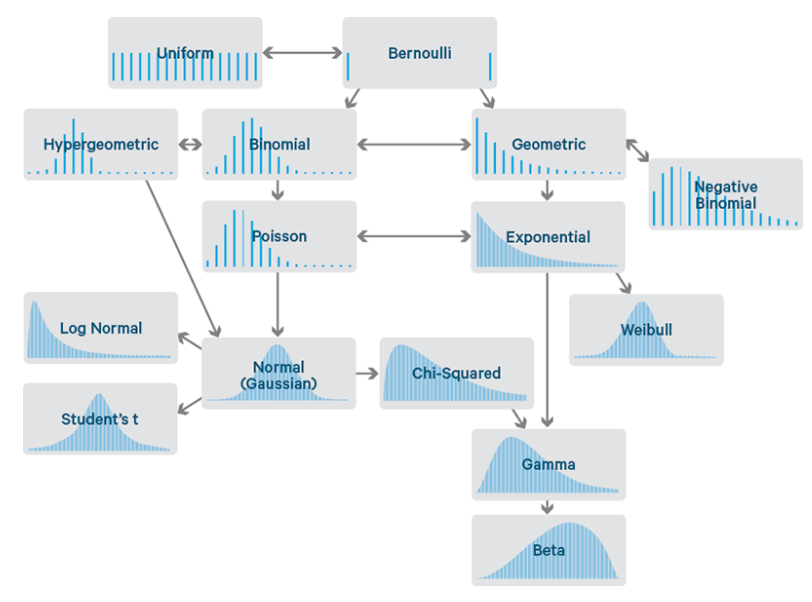


На рисунке видно, что в пределах одного среднеквадратического отклонения лежит 68,26% значений, принимаемых нормально распределенной случайной величиной (соответствует доли площади под кривой распределения). В пределах двух среднеквадратических отклонений — уже 95,44%, а в пределах трех — 99,72%. Это означает, что вероятность того, что случайная величина примет значение, отклоняющееся от математического ожидания больше чем на три среднеквадратических отклонения, не превышает 0,28%, т.е. пренебрежимо мала.

#### 

#### **Виды распределений**

Карта связей



* Бернулли и равномерное
* Биномиальное и гипергеометрическое
* Пуассон
* Геометрическое и отрицательное биномиальное
* Экспоненциальное и Вейбула
* Нормальное, логнормальное, Стьюдента и хи-квадрат
* Гамма и бета

#### **Центральная предельная теорема**

Класс теорем в [теории вероятностей](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9), утверждающих, что сумма достаточно большого количества [слабо зависимых](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B7%D0%B0%D0%B2%D0%B8%D1%81%D0%B8%D0%BC%D0%BE%D1%81%D1%82%D1%8C_(%D1%82%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9)) [случайных величин](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D0%B0%D1%8F_%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD%D0%B0), имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет [распределение](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9), близкое к [нормальному](https://ru.wikipedia.org/wiki/%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5).

Так как многие случайные величины в приложениях формируются под влиянием нескольких слабо зависимых случайных факторов, их распределение считают нормальным. При этом должно соблюдаться условие, что ни один из факторов не является доминирующим.

Центральные предельные теоремы в этих случаях обосновывают применение нормального распределения.

#### **Доверительные интервалы**

Доверительный интервал — термин, используемый в математической статистике при интервальной оценке статистических параметров, более предпочтительной при небольшом объеме выборки, чем точечная.

Доверительным называют интервал, который покрывает неизвестный параметр с заданной надежностью.

Доверительным называется интервал, в который попадают измеренные в эксперименте значения, соответствующие доверительной вероятности.

#### **Дискретные и непрерывные распределения**

**Дискретной случайной величиной** называется случайная величина, которая в результате испытания принимает отдельные значения с определенными вероятностями. Число возможных значений дискретной случайной величины может быть конечным и бесконечным. Примеры дискретной случайной величины: запись показаний спидометра или измеренной температуры в конкретные моменты времени.

**Непрерывной случайной величиной** называют случайную величину, которая в результате испытания принимает все значения из некоторого числового промежутка. Число возможных значений непрерывной случайной величины бесконечно. Пример непрерывной случайной величины: измерение скорости перемещения любого вида транспорта или температуры в течение конкретного интервала времени.

**Дискретное распределение** характеризуется тем, что оно сосредоточено

в конечном или счетном числе точек.

**Непрерывное распределение** "распределено" по некоторому вещественному интервалу.

#### **Математическое ожидание случайной величины**

**Математическое ожидание** — одно из важнейших понятий в теории вероятностей, означающее среднее (взвешенное по вероятностям возможных значений) значение случайной величины

На практике математическое ожидание обычно оценивается как среднее арифметическое наблюдаемых значений случайной величины (выборочное среднее, среднее по выборке).

#### **Распределение Стьюдента**

Распределение Стьюдента (t-распределение) в теории вероятностей — это однопараметрическое семейство абсолютно непрерывных распределений.

Распределение Стьюдента может быть использовано для оценки того, насколько вероятно, что истинное среднее находится в каком-либо заданном диапазоне.