**Введение в статистику**

**Статистика**— это математическая наука о сборе, анализе, интерпретации и представлении данных.

Знания в этой области позволяют использовать подходящие методы сбора и анализа данных, а также эффективно представлять результаты такого анализа. Статистика играет ключевую роль в научных открытиях, принятии решений и составлении прогнозов, основанных на данных. Она позволяет гораздо глубже разобраться в объекте исследования.

->)Один из основных принципов науки о данных — получение выводов из их анализа. Статистика отлично для этого подходит. Она является разновидностью математики и использует формулы.

Машинное обучение зародилось из статистики. Основой используемых в нём алгоритмов и моделей является так называемое статистическое обучение. Знание основ статистики крайне полезно вне зависимости от того, изучаете вы глубоко алгоритмы МО или просто хотите быть в курсе новейших исследований в этой сфере.

->)

* генеральная совокупность
* выборка
* матожидание
* квантиль
* медиана
* мода
* среднее арифметическое значение
* дисперсия
* интерквантильный размах

Основные понятия статистики: (сначала 2 потом 1)

1. генеральная совокупность

(Генеральная совокупность — это совокупность всех объектов или наблюдений, относительно которых исследователь намерен делать выводы при решении конкретной задачи. В ее состав включаются все объекты, которые подлежат изучению.)

Объем генеральной совокупности может быть очень велик, и на практике рассмотреть все ее элементы не представляется возможным. Поэтому обычно из генеральной совокупности извлекаются [выборки](https://wiki.loginom.ru/articles/sample.html), на основе анализа которых аналитик пытается сделать вывод о свойствах всей совокупности, скрытых в ней закономерностях, действующих правилах и т.д.

1. выборка — это подмножество наблюдений [генеральной совокупности](https://wiki.loginom.ru/articles/general-population.html), отобранных с целью изучения и анализа с помощью специальной процедуры (которая также называется выборкой), чтобы впоследствии обобщить полученные знания на всю совокупность. Выборки должны обладать свойством [репрезентативности](https://wiki.loginom.ru/articles/representativity.html).

(Под репрезентативностью в [статистике](https://wiki.loginom.ru/articles/mathematical-statistics.html) и [машинном обучении](https://wiki.loginom.ru/articles/machine-learning.html) понимается соответствие структурных характеристик [выборки](https://wiki.loginom.ru/articles/sample.html) характеристикам [генеральной совокупности](https://wiki.loginom.ru/articles/general-population.html), из которой она извлечена. Репрезентативность определяет, насколько возможно обобщать результаты исследования, полученные на основе выборочных данных на всю исходную совокупность.)

Причины, по которым анализируют выборки, а не всю совокупность, могут быть следующими:

* объем генеральной совокупности может быть очень велик, а её анализ сложен в вычислительном плане (особенно, если нельзя использовать [масштабируемые алгоритмы](https://wiki.loginom.ru/articles/scalable-algorithm.html));
* получить доступ ко всем элементам совокупности очень сложно, или вообще невозможно (например, опросить население всего города — кто-то уехал, кто-то просто отвечать не хочет, поэтому проводят выборочный опрос);
* при использовании методов [машинного обучения](https://wiki.loginom.ru/articles/machine-learning.html) требуется использовать несколько множеств: [обучающее](https://wiki.loginom.ru/articles/training-set.html), [тестовое](https://wiki.loginom.ru/articles/test-set.html) и [валидационное](https://wiki.loginom.ru/articles/validation-set.html), которые тоже являются выборками из исходного набора данных.

1. матожидание – это сумма произведений всех возможных значений случайной величины на вероятность этих значений. Означает среднее (взвешенное по вероятностям возможный значений) значение случайной величины.
2. квантиль – значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

на пример если квантиль у нас равен 0,25 то мы смотрим опять же на наше распределение и располагаем значение. это обязательно нужно запомнить в порядке возрастания и когда мы расположили наше значение в порядке возрастания мы смотрим где у нас находится примерно 25 процентов от выборки до 0.25 можно представить как 25 процентов не более двадцати пяти процентов от выборки и смотрим именно на этот значение вот нашем примере с возрастом квантиль 0.25 будет равен 24 года

(смотрим на значение до 25% от всей выборки)

1. медиана – это значение делящее распределение пополам( это такое число, что половина из элементов выборки больше него, а другая половина меньше)

Медиана – квантиль 0.5

1. мода – самое вероятное значение случайной величины( в нестрогом смысле)
2. среднее арифметическое значение – сумма деленная на количество значений.
3. дисперсия характеризует разброс случайной величины вокруг ее математического ожидания

возьмем какого-то среднего котика до который вы посчитали и сравним с каким-то значением значением другого котика например барсиком вы можете заметить что барсик отклоняется на какое-то число от среднего котиков и чем больше у нас будет таких котиков да тем больше у нас в конечном итоге будет разброс наших значений и как раз эти сам самые разнообразные значения/

**средне отклонений и будет являться дисперсий**

наши отклонения наше типичное отклонение в целом наше

1. интерквантильный размах – разность между первым и третьим квантилем

то есть 50% наших данных, выборка который в целом характеризует наши данные.