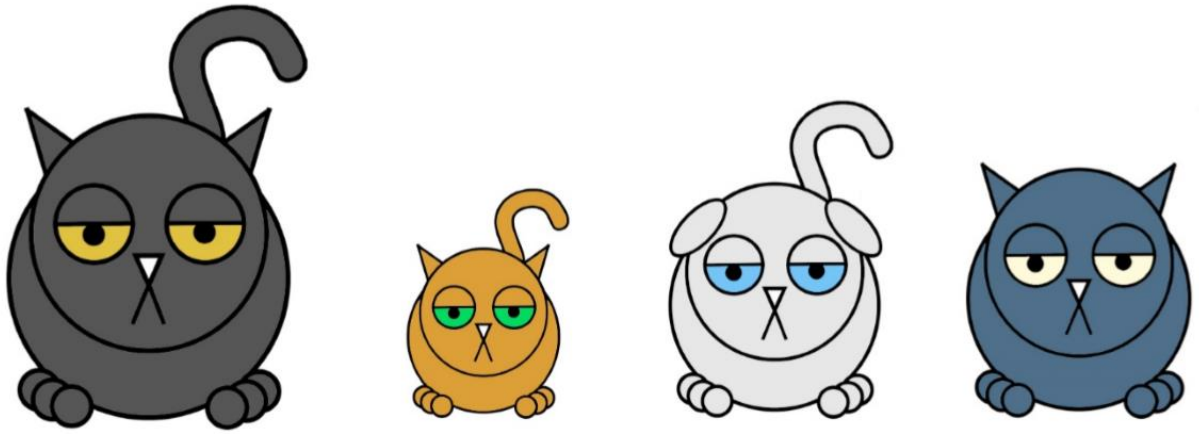
Лекция 2

ОСНОВЫ СТАТИСТИКИ

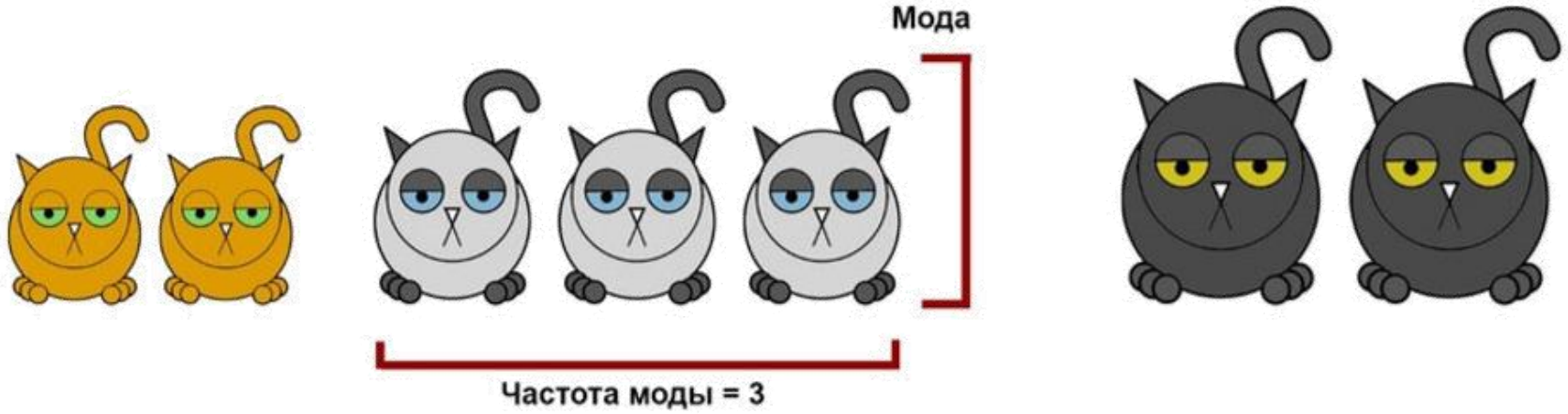
основы описательной статистики

Котики бывают разные. Есть большие котики, а есть маленькие. Есть котики с длинными хвостами, а есть и вовсе без хвостов. Есть котики с висячими ушками, а есть котики с короткими лапками. Как же нам понять, как выглядит типичный котик?



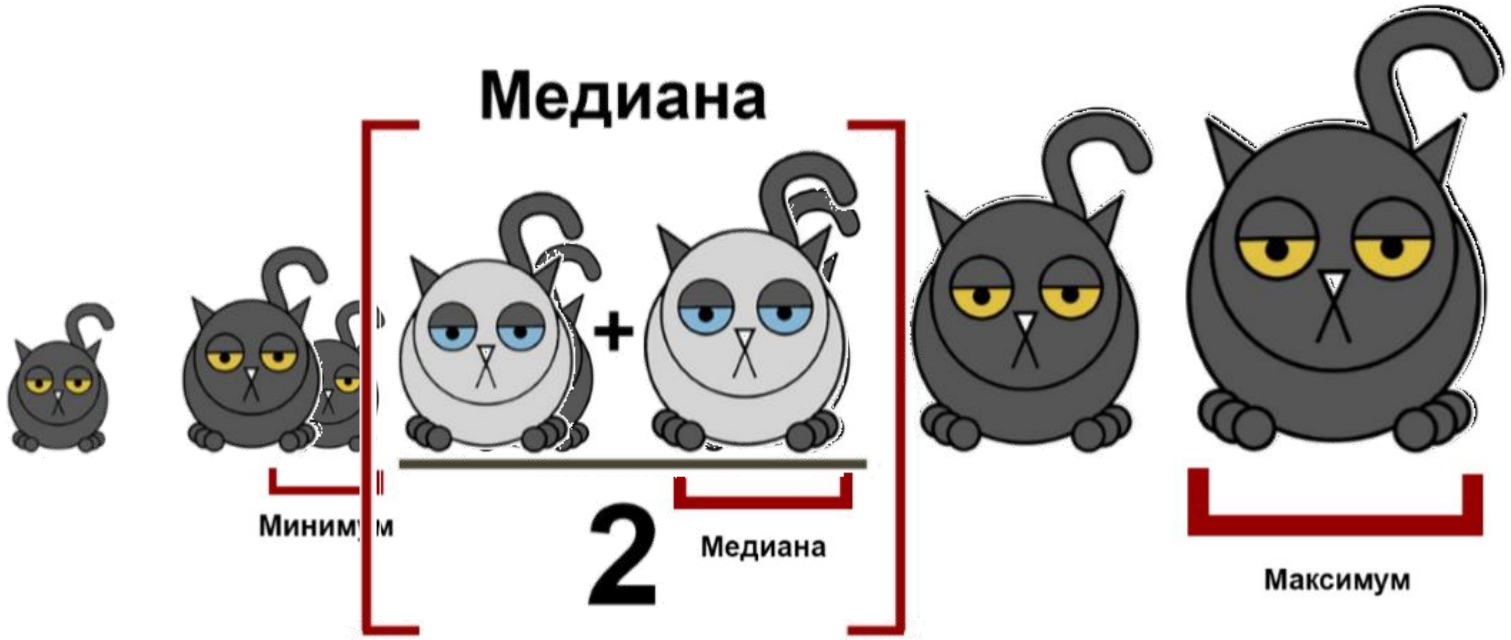
Типичный котик

Для простоты мы возьмем такое котиковое свойство, как размер. Первый и наиболее очевидный способ — посмотреть, какой размер котиков встречается чаще всего. Такой показатель называется МОДОЙ



Медиана

упорядочить всех котиков от самого маленького до самого крупного, а затем посмотреть на середину этого ряда. Как правило, там находится котик, который обладает самым типичным размером. И этот размер называется медианой

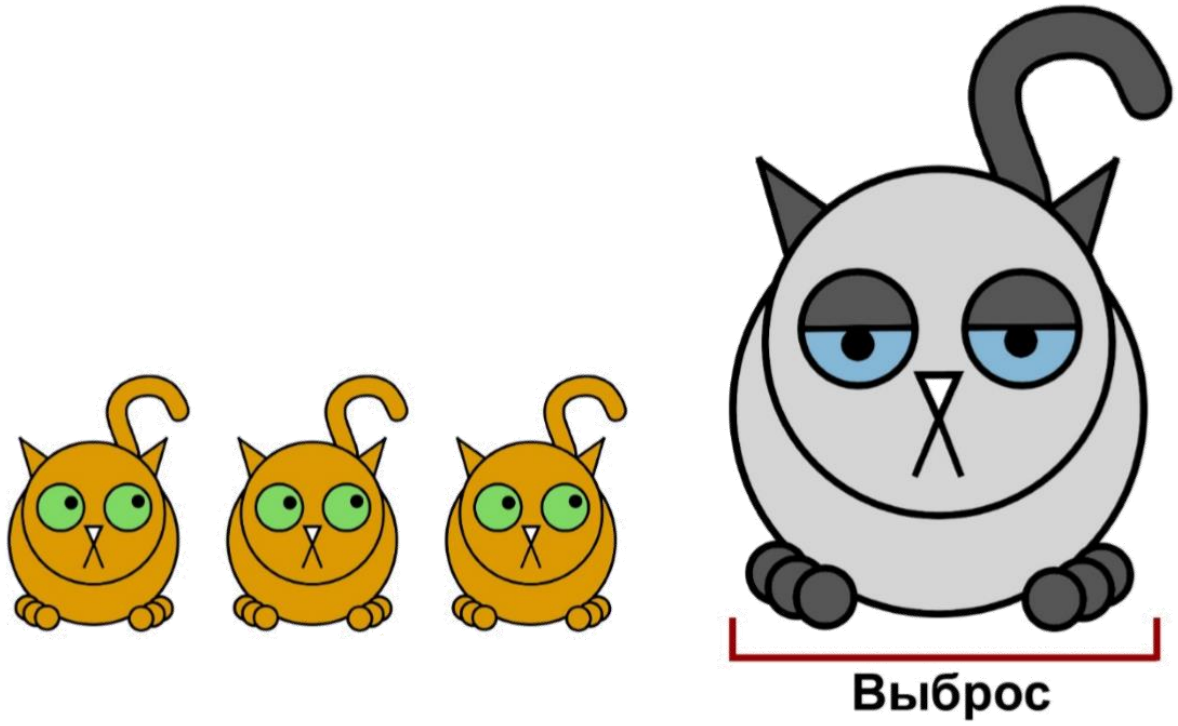


Среднее значение

сложить размер всех котиков и поделить на их количество. Полученное число называется средним значением, и оно является очень популярным в современной статистике

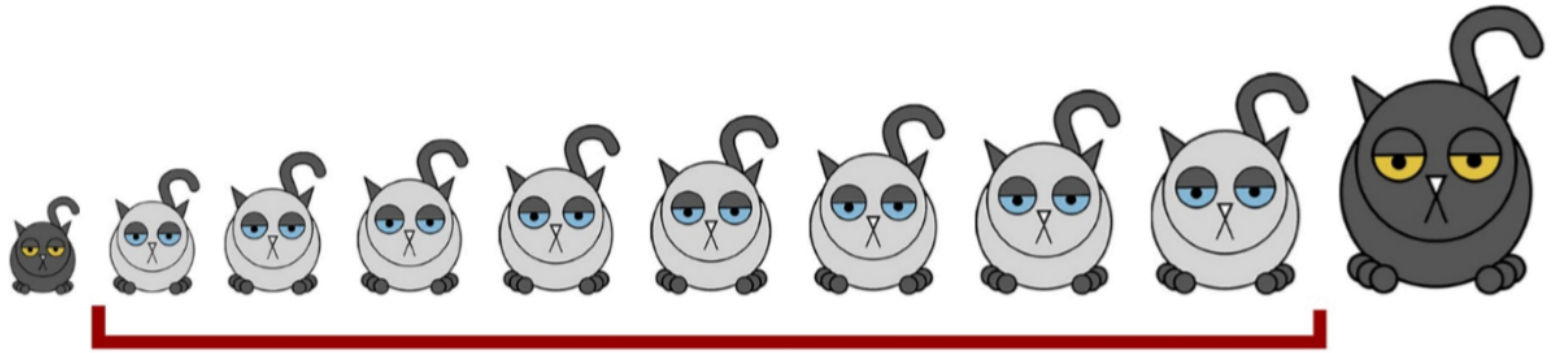
Выброс

Предположим, что среди наших котиков есть один уникум размером со слона. Его присутствие может существенным образом сдвинуть среднее значение в большую сторону, и оно перестанет отражать типичный котиковый размер



Усеченное (или урезанное) среднее

Чтобы избавиться от таких выбросов, иногда применяют следующий метод: убирают по 5—10% самых больших и самых маленьких котиков и уже от оставшихся считают среднее. Получившийся показатель называют усеченным (или урезанным) средним



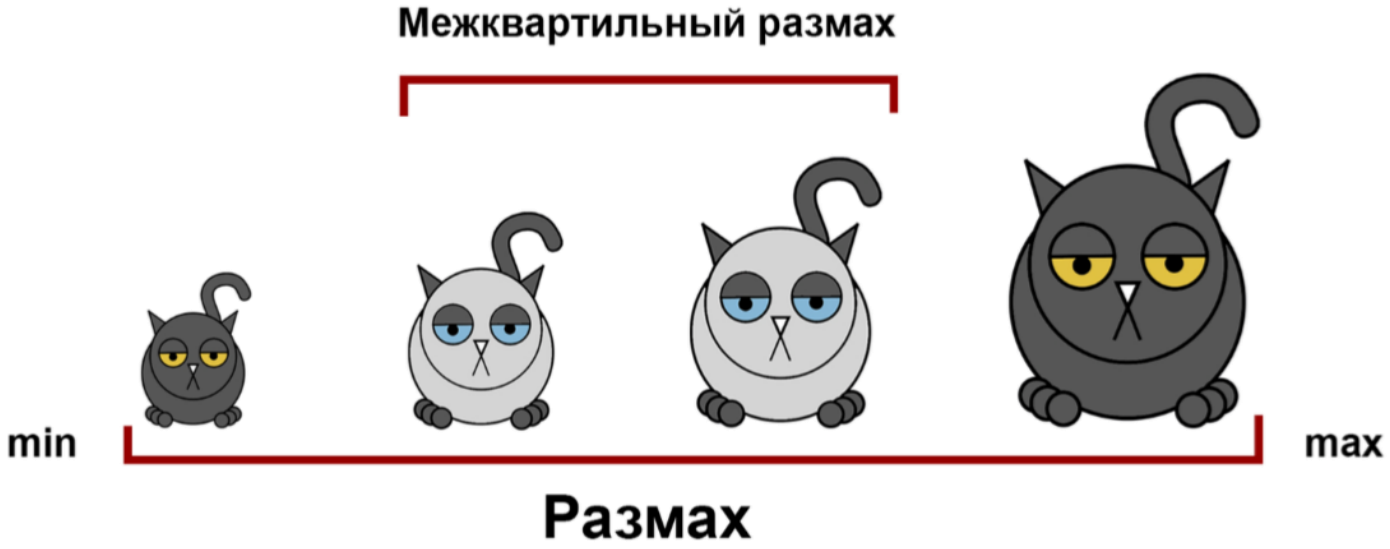
Котики для усеченного среднего

Меры изменчивости

Но, кроме типичности, нас довольно часто интересует, насколько разнообразными могут быть котики по размеру. И в этом нам помогают меры изменчивости

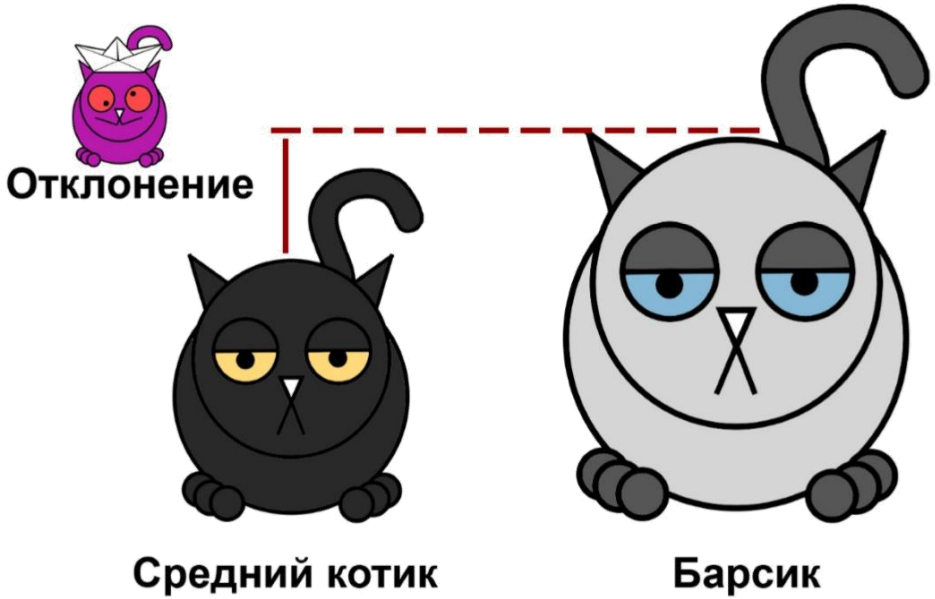
Размах

Является разностью между самым большим и самым маленьким котиком. Однако, как и среднее арифметическое, эта мера очень чувствительна к выбросам. И, чтобы избежать искажений, мы должны отсечь 25% самых больших и 25% самых маленьких котиков и найти размах для оставшихся. Эта мера называется межквартильным размахом



Отклонение

Предположим, что мы решили сравнить размер некоторого конкретного котика (назовем его Барсиком) со средним котиковым размером. Разница (а точнее разность) этих размеров называется отклонением. И совершенно очевидно, что чем сильнее Барсик будет отличаться от среднего котика, тем больше будет это самое отклонение.



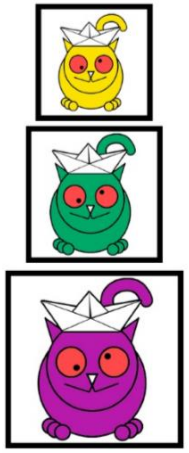
$$\sum_{i=1}^{n} \Delta x_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n \sum_{j=1}^{n} \frac{x_j}{n} = \sum_{i=1}^{n} x_i - \sum_{j=1}^{n} x_j = 0$$

Стандартное отклонение

И, чтобы понять, какое отклонение является для наших котиков наиболее типичным, мы можем просто найти среднее 8

значение по этим отклонениям (т. е. сложить все отклонения и поделить их на количество котиков).

/ 3



/ 3

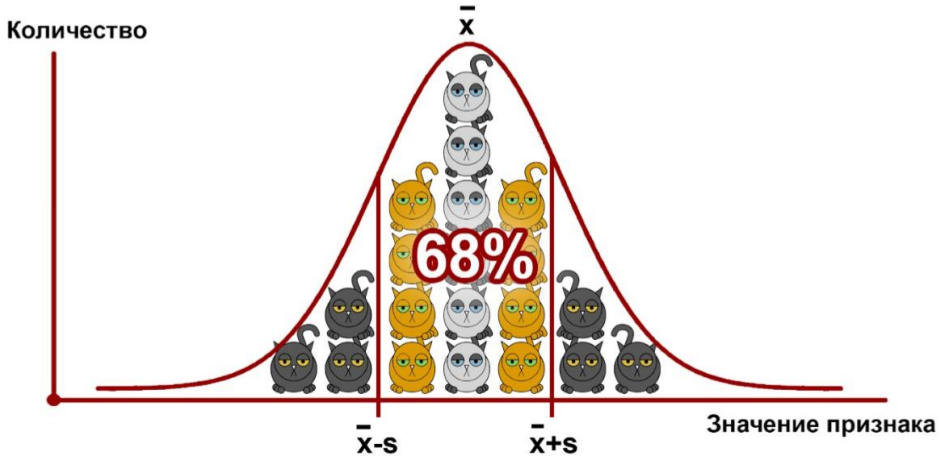
Дисперсия D

Среднеквадратическим отклонением

И, если мы найдем среднее от квадратов отклонений, мы получим то, что называется дисперсией. Однако, к большому сожалению, квадрат в этой формуле делает дисперсию очень неудобной для оценки разнообразия котиков: если мы измеряли размер в сантиметрах, то дисперсия имеет размерность в квадратных сантиметрах. Поэтому для удобства использования дисперсию берут под корень, получая по итогу показатель, называемый среднеквадратическим отклонением

Нормальное распределение

Среднее значение и среднеквадратическое отклонение очень часто совместно используются для описания той или иной группы котиков. Дело в том, что, как правило, большинство (а именно около 68%) котиков находится в пределе одного среднеквадратического отклонения от среднего. Эти котики обладают так называемым нормальным размером. Оставшиеся 32% либо очень большие, либо очень маленькие. В целом же для большинства котиковых признаков картина выглядит вот так



Выборка, генеральная совокупность

Чаще всего нас, как исследователей, интересуют все котики без исключения. Статистики называют этих котиков генеральной совокупностью. Однако на практике мы не можем замерить всю генеральную совокупность — как правило, мы работаем только с небольшим количеством котиков, называемым выборкой

Репрезентативность

Очень важно, чтобы выборка была максимально похожа на генеральную совокупность. Степень такой похожести называется репрезентативностью

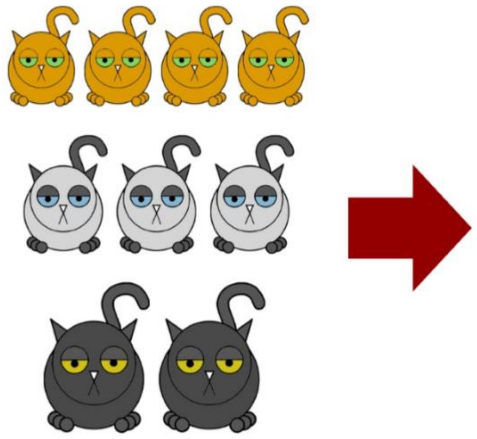
два вида дисперсии

Необходимо запомнить, что существует две формулы дисперсии: одна для генеральной совокупности, другая — для выборки. В знаменателе первой всегда стоит точное количество котиков, а у второй — ровно на одного котика меньше.

Средства визуализации данных

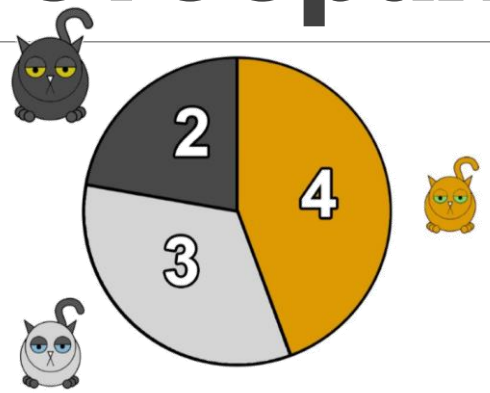
Таблицы частот

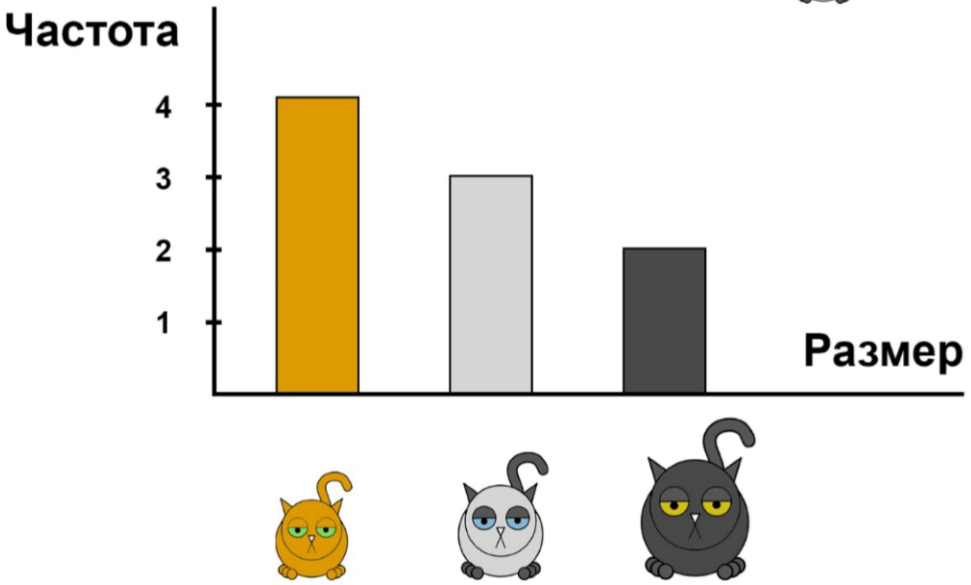
Этоколичество,кстати,иназываетсячастотой.Этичастотыбываютабсолютными(в котиках) и относительными(в процентах)

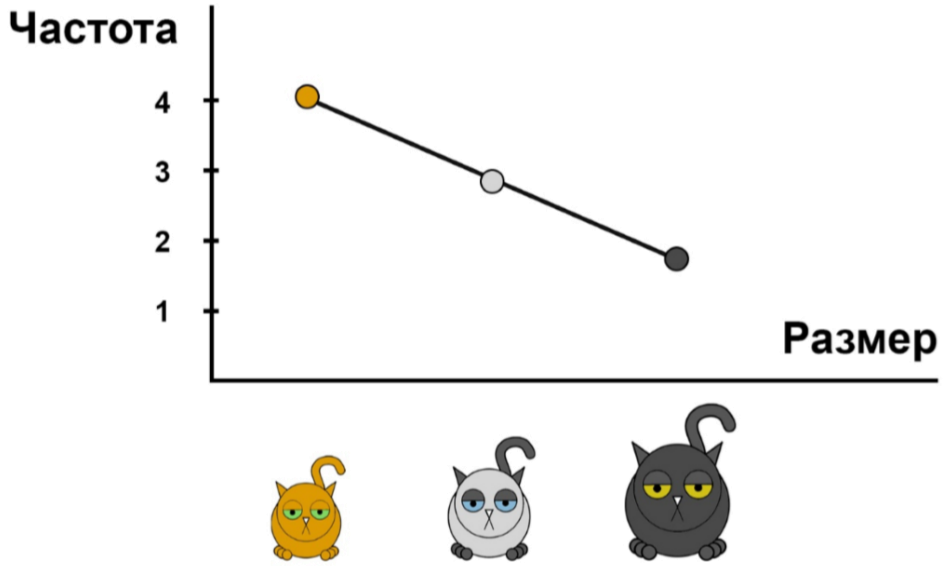


Размер	Частота
	4
	3
	2

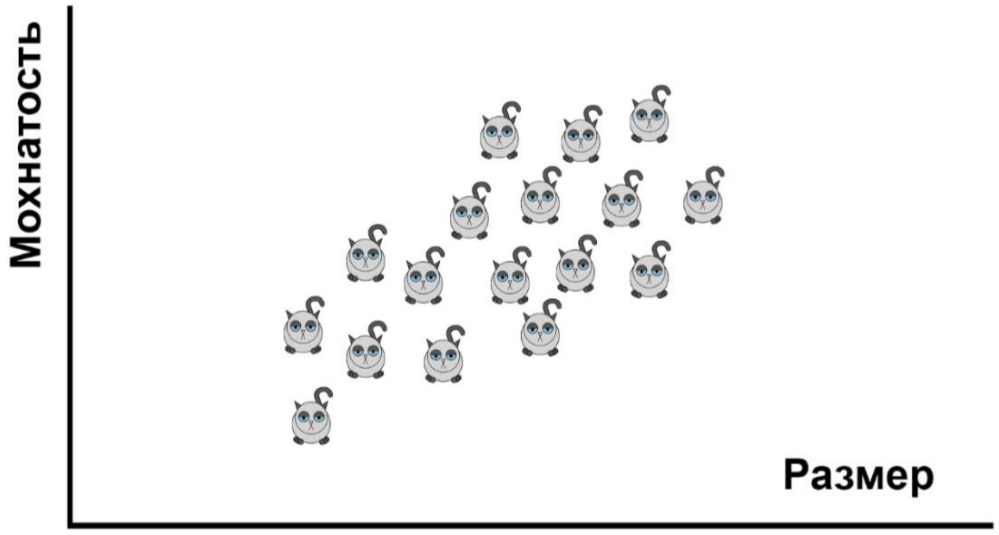
Варианты отображения



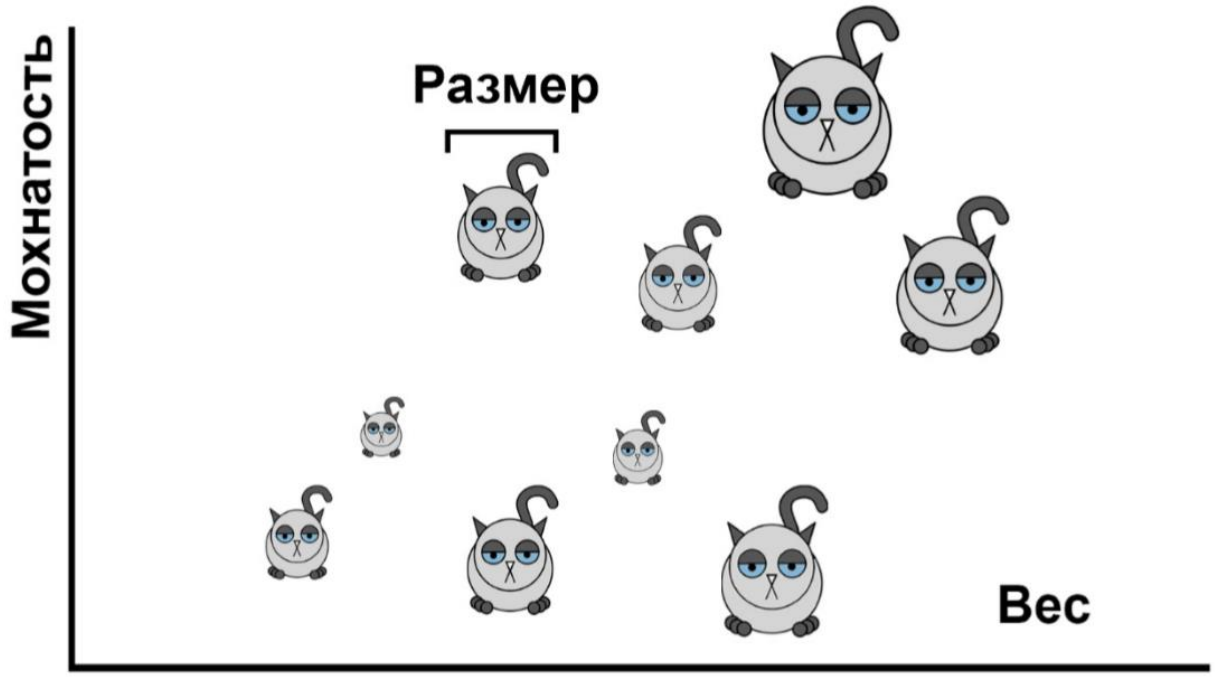




Точечная диаграмма (или диаграммой рассеяния)

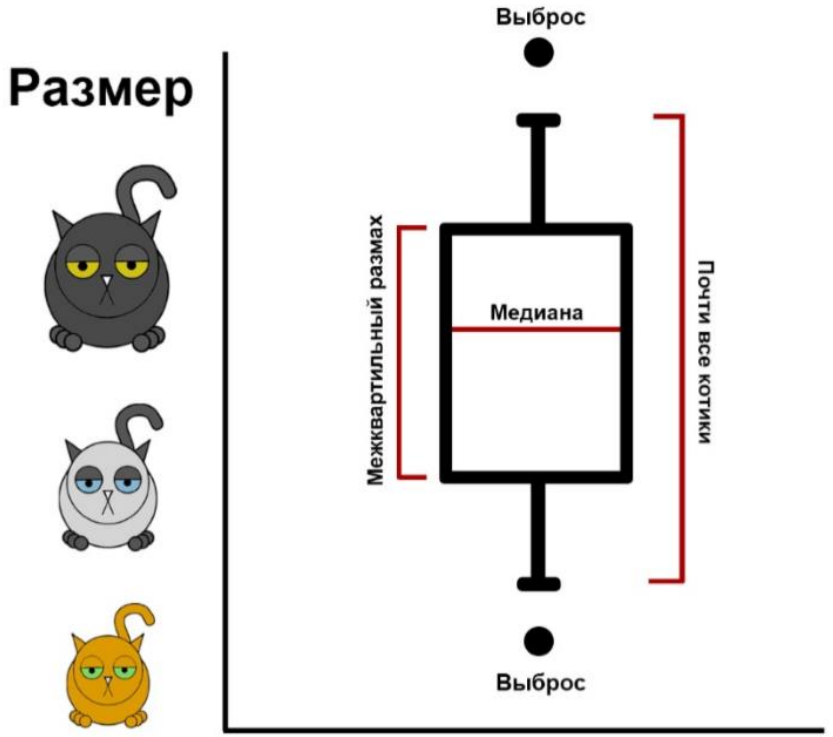


пузырьковая диаграмма

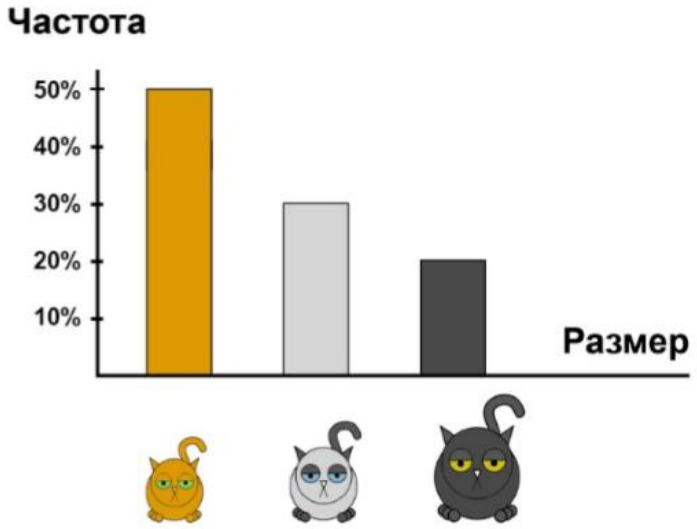


Боксплот

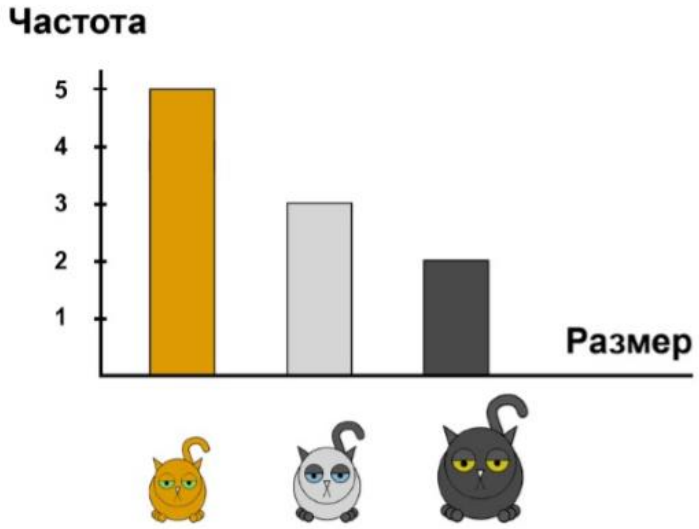
Более известным средством является так называемый боксплот (или «ящик с усами»).



Проценты вместо абсолютных величин

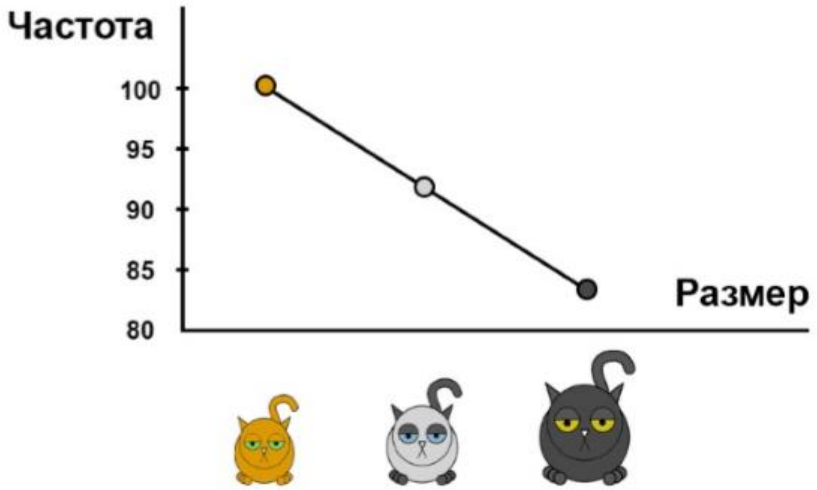


Хитрость

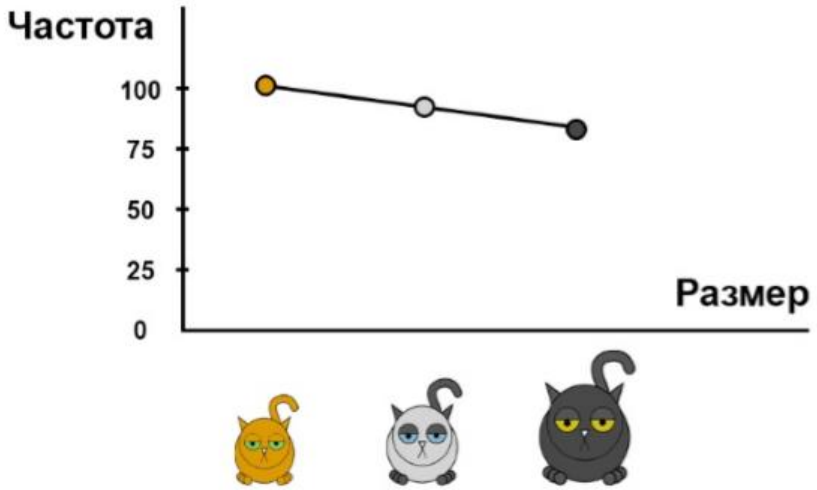


Разоблачение

Сдвиг шкалы

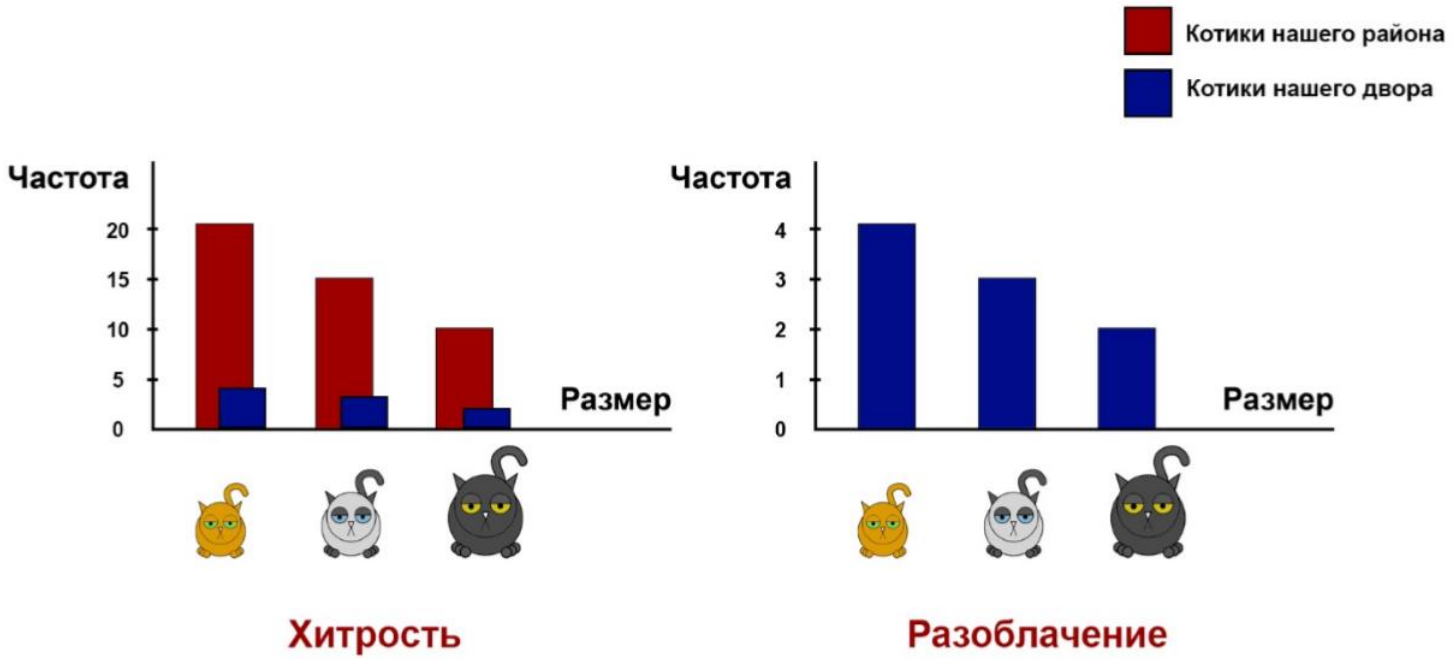


Хитрость



Разоблачение

Сокрытие данных



Изменение масштабов