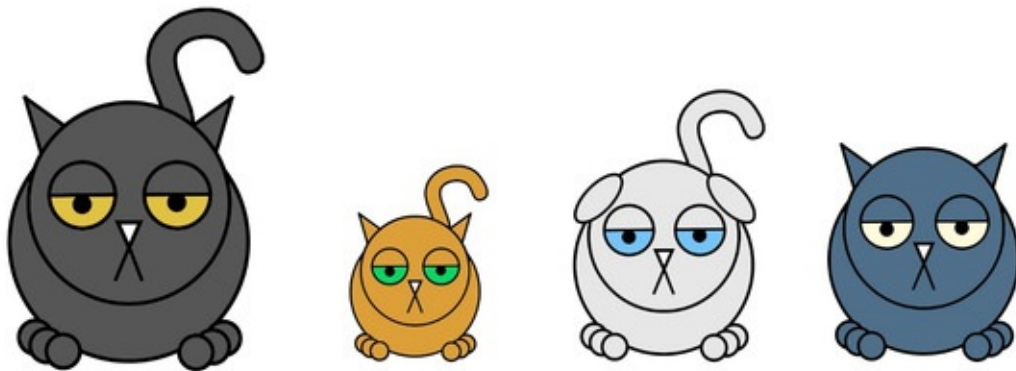


Глава 1.

Как выглядят котики

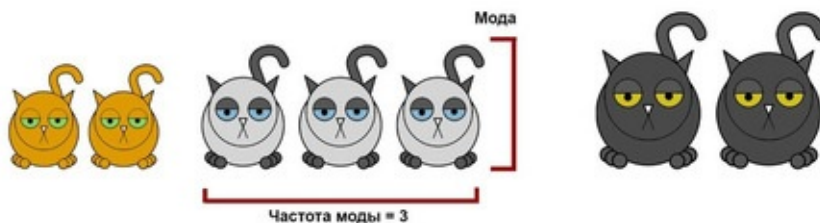
или основы описательной статистики

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?

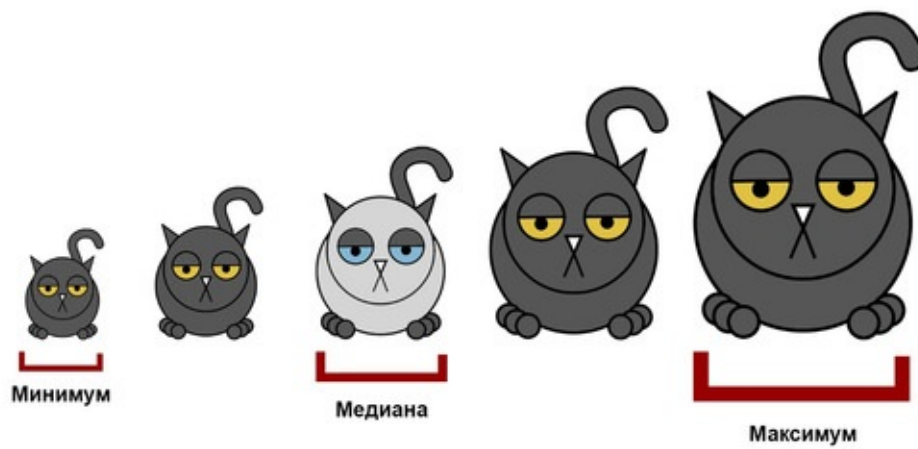


Для простоты мы возьмем такое котиковое свойство, как размер.

Первый и наиболее очевидный способ — посмотреть, какой размер котиков встречается чаще всего. Такой показатель называется *модой*.



Второй способ: мы можем 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*.



Если же посередине находятся сразу два котика (что бывает, когда их четное количество), то, чтобы найти медиану, нужно сложить их размеры и поделить это число пополам.

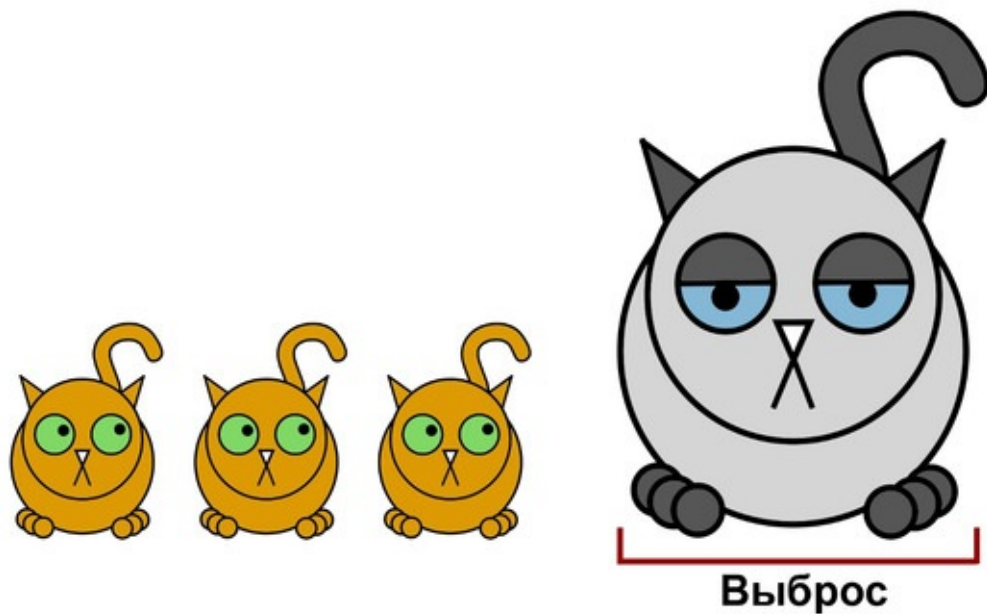


Последний способ нахождения наиболее типичного котика — это сложить размер всех котиков и поделить на их количество. Полученное число называется *средним значением*, и оно является очень популярным в современной статистике.



Однако, среднее арифметическое далеко не всегда является лучшим показателем типичности.

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.



Такой «слоновый» котик, так же как и котик размером с муравья, называется *выбросом*, и он может существенно исказить наши представления о котиках. И, к большому сожалению, многие статистические критерии, содержащие в своих формулах средние значения, также становятся неадекватными в присутствии «слоновых» котиков.

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*.



Альтернативный вариант — применять вместо среднего медиану.

Итак, мы рассмотрели основные методы нахождения типичного размера котиков: моду, медиану и средние значения. Все вместе они называются *мерами центральной тенденции*. Но, кроме типичности, нас довольно часто интересует, насколько разнообразными могут быть котики по размеру. И в этом нам помогают меры изменчивости.

Первая из них — *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*.



Вторая и третья меры изменчивости называются *дисперсией* и *стандартным отклонением*. Чтобы разобраться в том, как они устроены, 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.



Логично было бы предположить, что чем больше у нас будет котиков с сильным отклонением, тем более разнообразными будут наши котики по размеру. И, чтобы понять, какое отклонение является для наших котиков наиболее типичным, мы можем просто найти среднее значение по этим отклонениям (т. е. сложить все отклонения и поделить их на количество котиков).



Однако если мы это сделаем, то получим 0. Для недоверчивых привожу доказательство:

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

Это происходит, поскольку одни отклонения являются положительными (когда Барсик больше среднего), а другие — отрицательными (когда Барсик меньше среднего). Поэтому

необходимо избавиться от знака. Сделать это можно двумя способами: либо взять модуль от отклонений, либо возвести их в квадрат, который, как мы помним, всегда положителен. Последнее применяется чаще.


$$\frac{\text{Квадраты отклонений}}{3}$$

Дисперсия D

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*.

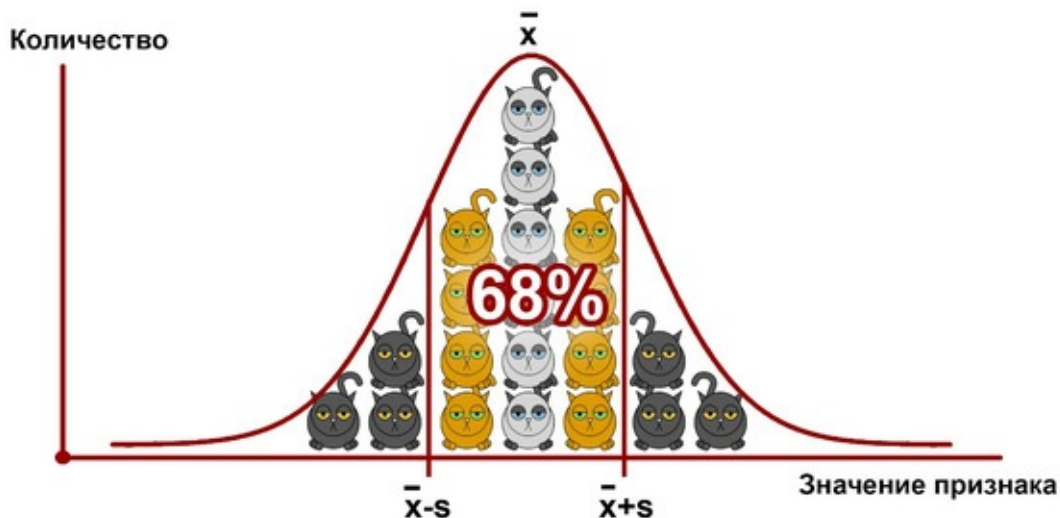

$$\sqrt{\frac{\text{Квадраты отклонений}}{3}}$$

Среднеквадратическое отклонение σ

К несчастью, дисперсия и среднеквадратическое отклонение так же неустойчивы к выбросам, как и среднее арифметическое.

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.



Такой график называется *нормальным распределением признака*.

Таким образом, зная всего два показателя, вы можете с достаточной долей уверенности сказать, как выглядит типичный котик, насколько разнообразными являются котики в целом и в каком диапазоне лежит норма по тому или иному признаку.

НЕМАЛОВАЖНО ЗНАТЬ!

Выборка, генеральная совокупность и два вида дисперсии

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*.



Очень важно, чтобы выборка была максимально похожа на генеральную совокупность.

Степень такой похожести называется *репрезентативностью*.

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.


$$\frac{\text{Дисперсия генеральной совокупности}}{3}$$


$$\frac{\text{Дисперсия выборки}}{2}$$

Корень из дисперсии генеральной совокупности, как уже было сказано, называется *среднеквадратическим отклонением*. А вот корень из дисперсии по выборке называется *стандартным отклонением*.

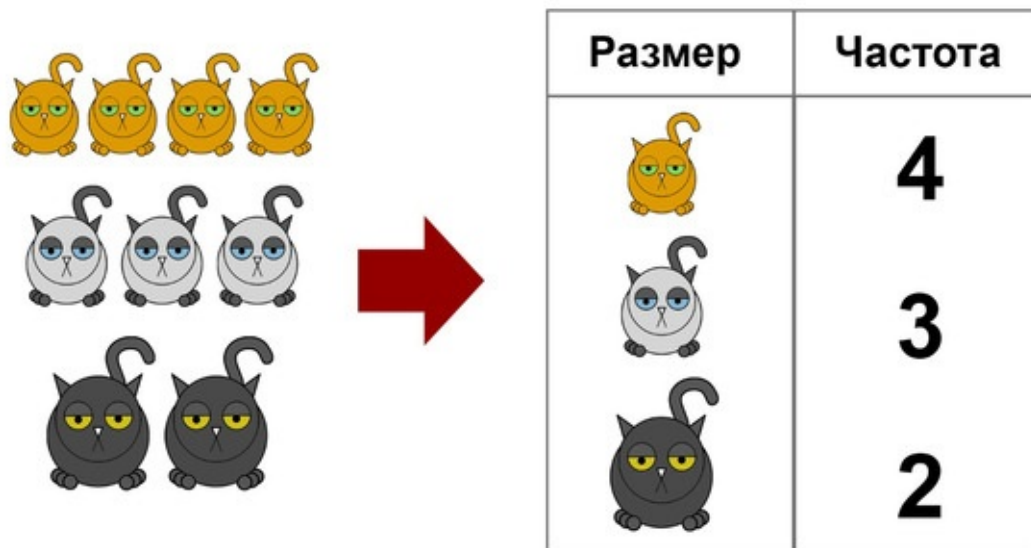
Однако не будет большой ошибкой, если вы будете пользоваться терминами *стандартное отклонение генеральной совокупности* и *стандартное отклонение выборки*. Чаще всего именно последнее и рассчитывается для реальных исследований.




Глава 2.

Картинки с котиками или средства визуализации данных

В предыдущей главе мы говорили про показатели, которые помогают определить, какой размер является для котиков типичным и насколько он бывает разнообразным. Но когда нам требуется получить более полные и зрительно осязаемые представления о котиках, мы можем прибегнуть к так называемым *средствам визуализации данных*.

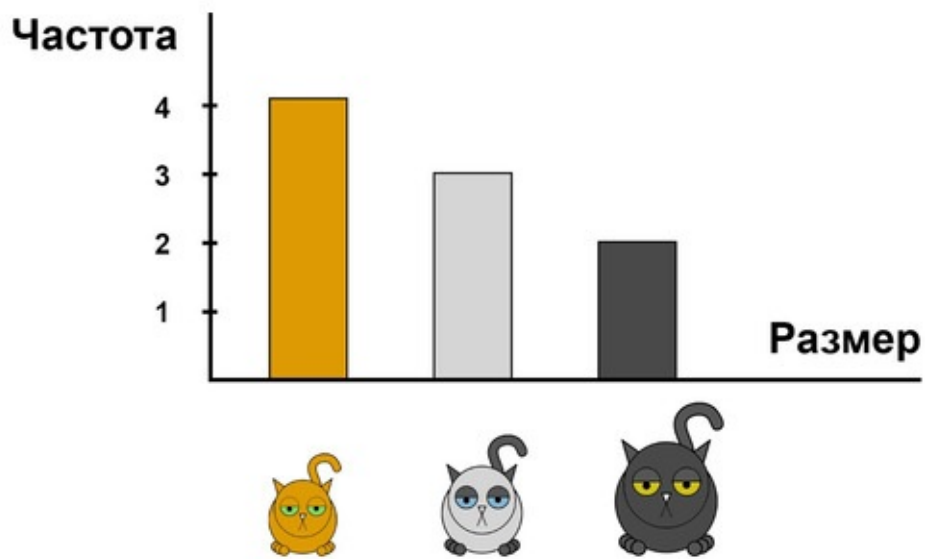
Первая группа средств показывает, сколько котиков обладает тем или иным размером. Для их использования необходимо предварительно построить так называемые *таблицы частот*. В этих таблицах есть два столбика: в первом указывается размер (или любое другое котовое свойство), а во втором — количество котиков при данном размере.



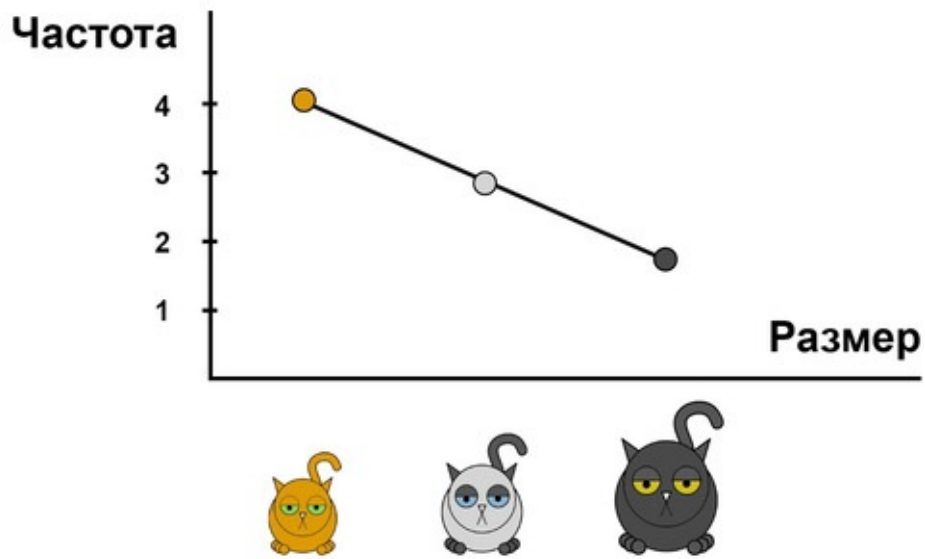
Размер	Частота
	4
	3
	2

Это количество, кстати, и называется *частотой*. Эти частоты бывают *абсолютными* (в котиках) и *относительными* (в процентах).

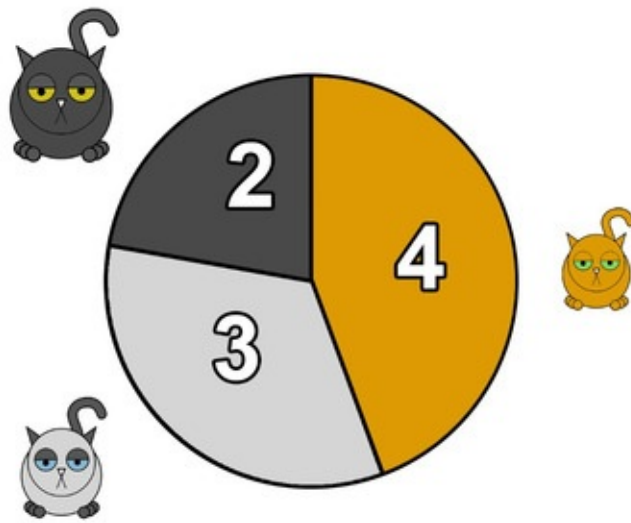
С таблицами частот можно делать много интересных вещей. Например, построить *столбиковую диаграмму*. Для этого мы откладываем две перпендикулярных линии: горизонтальная будет обозначать размер, а вертикальная — частоту. А затем — рисуем столбики, высота которых будет соответствовать количеству котиков того или иного размера.



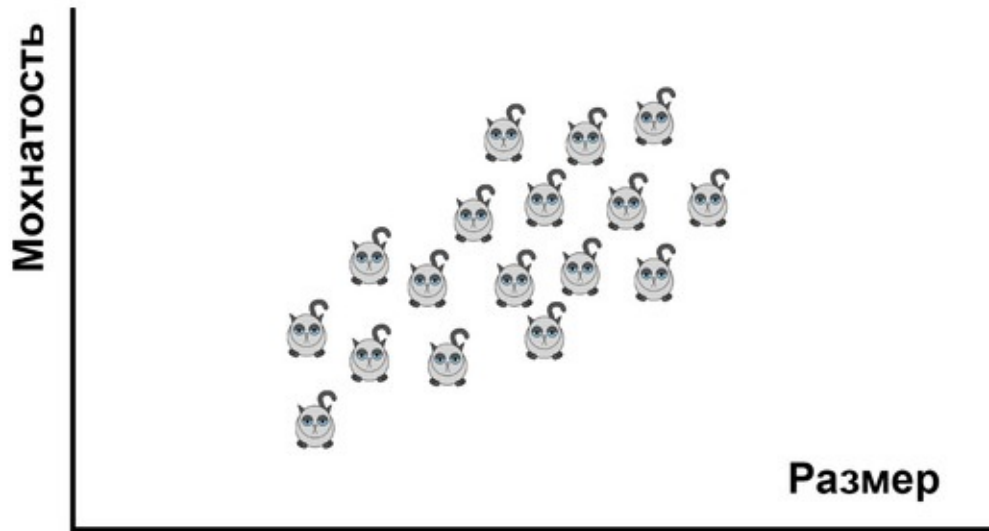
А еще мы можем вместо столбиков нарисовать точки и соединить их линиями. Результат называется *полигоном распределения*. Он довольно удобен, если котиковых размеров действительно много.



Наконец, мы можем построить *круговую диаграмму*. Величина каждого сектора такой диаграммы будет соответствовать проценту котиков определенного размера.



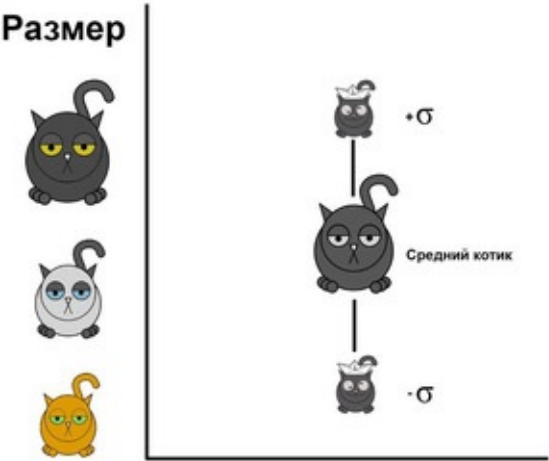
Следующая группа средств визуализации позволяет отобразить сразу два котиковых свойства. Например, размер и мохнатость. Как и в случае со столбиковыми диаграммами, первым шагом рисуются оси. Только теперь каждая из осей отображает отдельное свойство. А после этого каждый котик занимает на этом графике свое место в зависимости от степени выраженности этих свойств. Так, большие и мохнатые котики занимают место ближе к правому верхнему углу, а маленькие и лысые — в левом нижнем.



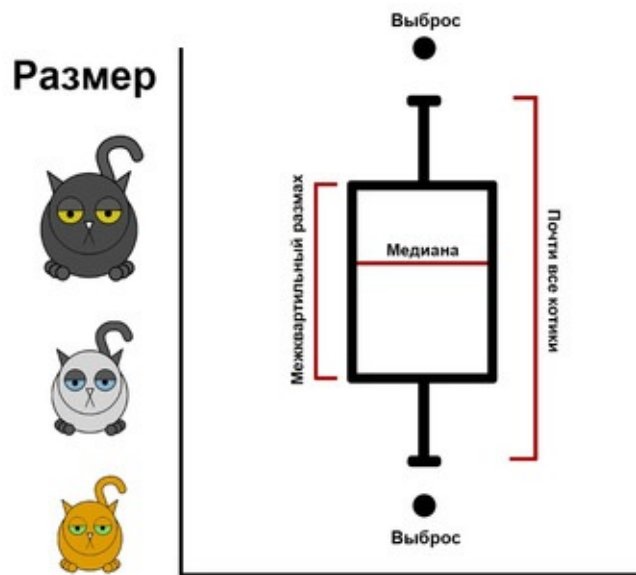
Поскольку обычно котики на данной диаграмме обозначаются точками, то она называется *точечной* (или *диаграммой рассеяния*). Более продвинутый вариант — *пузырьковая диаграмма* — позволяет отобразить сразу три котиковых свойства одновременно (размер, мохнатость и вес). Это достигается за счет того, что сами точки на ней имеют разную величину, которая и обозначает третье свойство.



Последняя крупная группа средств визуализации позволяет графически изобразить меры центральной тенденции и меры изменчивости. В простейшем виде это точка на графике, обозначающая, где находится средний котик, и линии, длина которых указывает на величину стандартного отклонения.



Более известным средством является так называемый *боксплот* (или «ящик с усами»). Он позволяет компактно отобразить медиану, общий и межквартильный размах, а также прикинуть, насколько распределение ваших данных близко к нормальному и есть ли у вас выбросы.



Помимо вышеперечисленных средств существует еще немало специфических, заточенных под определенные цели (например диаграммы, использующие географические карты). Однако, вне зависимости от того, какой тип диаграмм вы хотели бы использовать, существует ряд рекомендаций, которые желательно соблюдать.

На диаграмме не должно быть ничего лишнего. Если на ней есть элемент, не несущий какой-либо смысловой нагрузки, его лучше убрать. Потому что чем больше лишних элементов, тем менее понятной будет диаграмма.

То же самое касается цветов: лучше ограничить их количество до трех. А если вы готовите графики для публикации, то лучше их вообще делать черно-белыми.

НЕМАЛОВАЖНО ЗНАТЬ!

Темная сторона визуализации

Несмотря на то, что средства визуализации помогают облегчить восприятие данных, они так же легко могут ввести в заблуждение, чем, к сожалению, часто пользуются разные хитрые люди. Ниже мы приведем самые распространенные способы обмана с помощью диаграмм и графиков.

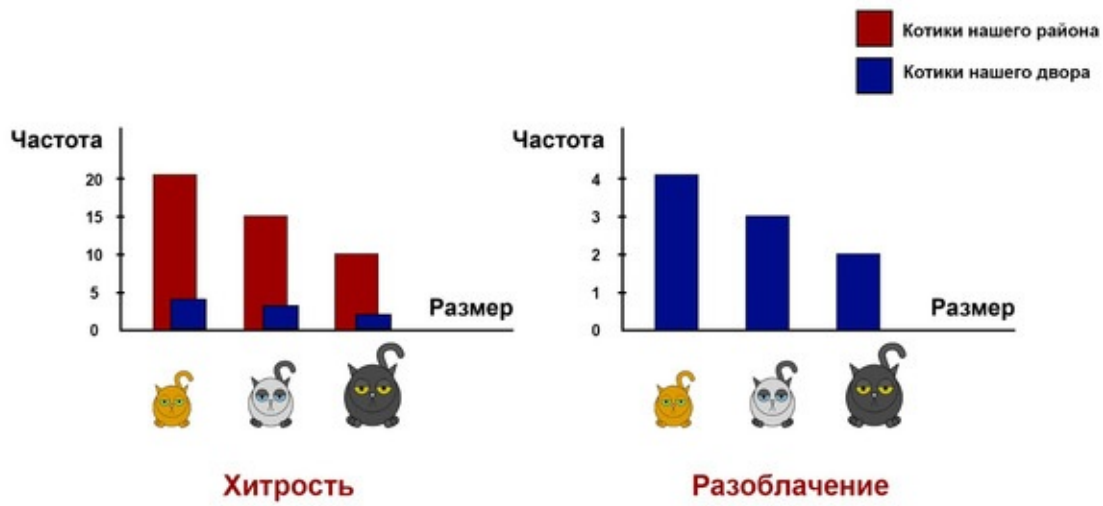
Проценты вместо абсолютных величин. Очень часто, чтобы придать своим данным значимости, хитрые люди переводят абсолютное количество котиков в проценты. Согласитесь, что результаты, полученные на 50% котиков, выглядят куда солиднее, чем на пяти.



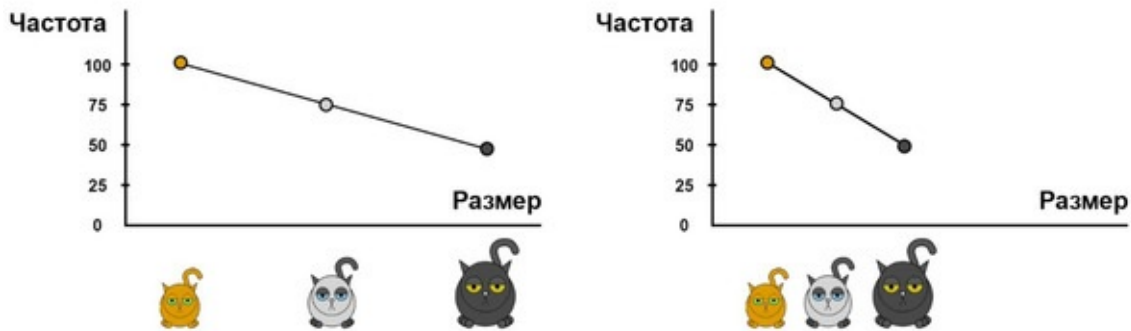
Сдвиг шкалы. Чтобы продемонстрировать значимые различия там, где их нет, хитрые люди как бы «сдвигают» шкалы, начиная отсчет не с нуля, а с более удобного для них числа.



Соккрытие данных. Если же цель хитрого человека в том, чтобы скрыть значимые различия в данных, то их можно разместить на одной шкале с другими данными, которые на порядок отличаются от первых. На их фоне любые различия или изменения будут выглядеть незначительно.



Изменение масштабов. Более мягкий вариант создания иллюзии значимости — это изменение масштабов шкал. В зависимости от масштаба одни и те же данные будут выглядеть по-разному.



Таким образом, надо быть очень аккуратным, интерпретируя данные, представленные в виде графиков и диаграмм. Гораздо меньше подвержены манипуляции данные, представленные в табличной формуле. Однако и здесь можно использовать некоторые хитрости, которые могут ввести в заблуждение непосвященную публику.