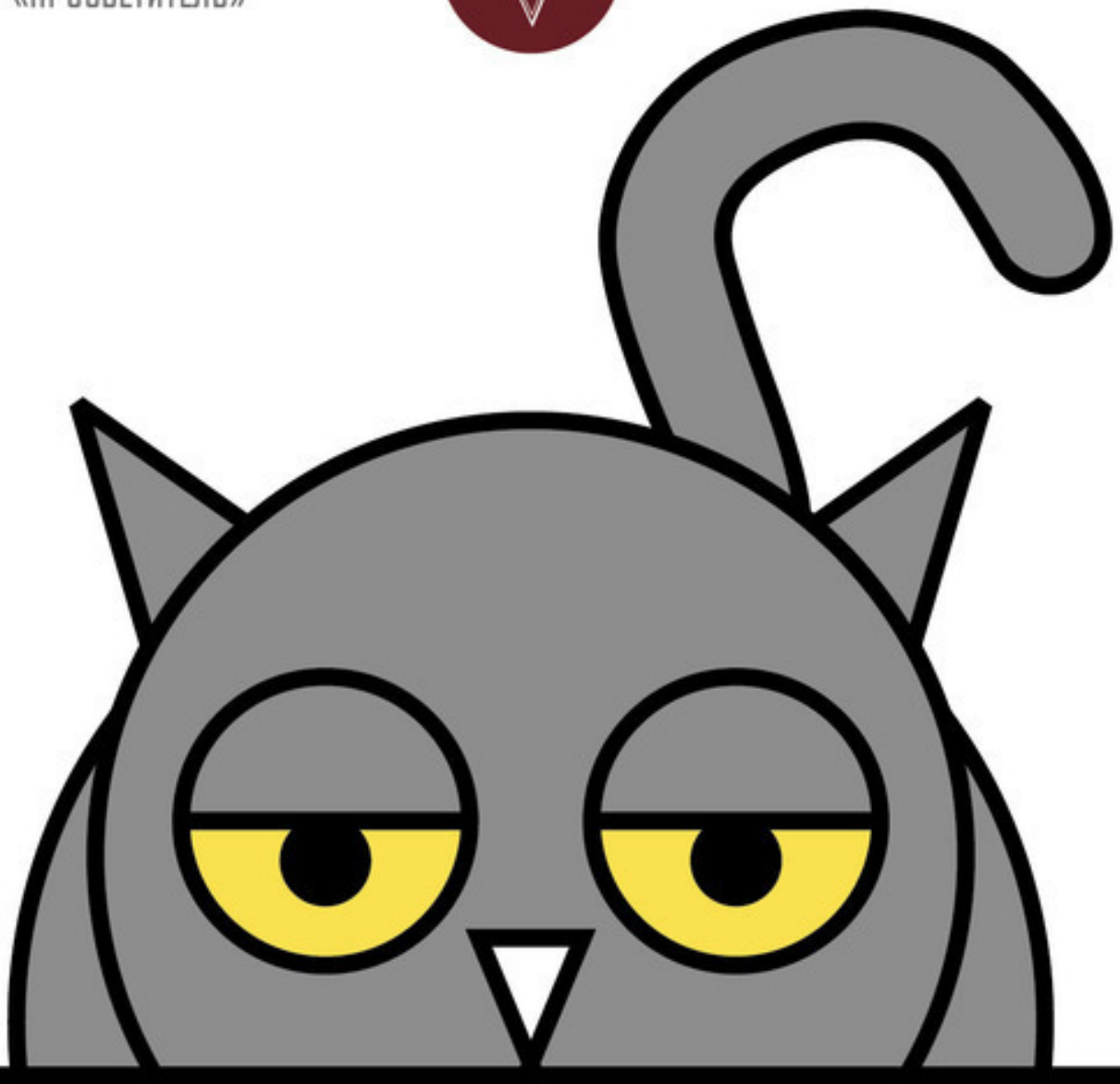
ВЛАДИМИР САВЕЛЬЕВ

# СТАТИСТИКА И КОТИКИ

ЛОНГ-ЛИСТ ПРЕМИИ «ПРОСВЕТИТЕЛЬ»

БЕСТСЕЛЛЕР РУНЕТА



Звезда Рунета. Бизнес

Владимир Савельев

# Статистика и котики

«АСТ»
2017

УДК 61
ББК 5

**Савельев В.**

Статистика и котики / В. Савельев — «АСТ», 2017 — (Звезда Рунета. Бизнес)

ISBN 978-5-17-106143-2

Из этой книги вы узнаете, что такое дисперсия и стандартное отклонение, как найти t-критерий Стьюдента и U-критерий Манна-Уитни, для чего используются регрессионный и факторный анализы, а также многое и многое другое. И все это – на простых и понятных примерах из жизни милых и пушистых котиков, которые дарят нам множество приятных эмоций.

УДК 61
ББК 5

ISBN 978-5-17-106143-2

# Содержание

# Савельев Владимир
# Статистика и котики



* * *

# Предисловие. От автора

Мало кто любит статистику.

Одни считают эту науку сухой и безжизненной. Другие боятся и избегают ее. Третьи полагают, что она бесполезна. Но у меня другое мнение на этот счет.

На мой взгляд, статистика обладает своей особой внутренней красотой. Ее можно увидеть, вглядываясь в корреляционную матрицу, рассматривая дендрограммы или интерпретируя результаты факторного анализа. За каждым статистическим коэффициентом стоит маленькое чудо, раскрывающее скрытые закономерности окружающего нас мира.

Но чтобы найти эту красоту, чтобы услышать поэзию, которая пронизывает статистику насквозь, необходимо преодолеть первоначальный страх и недоверие, вызванное внешней сложностью этого предмета.

Для того и написана эта книга. Чтобы показать, что статистика не такая страшная, как о ней думают. И что она вполне может быть такой же милой и пушистой, как котики, которые встретятся вам на страницах этой книги.

# От партнера издания

При слове «статистика» я вспоминаю британских ученых и выборы. Статистика – это многогранный инструмент. Иногда статистикой манипулируют, а можно открывать знания о реальном мире.

Автор написал книгу о базовой статистике в забавном формате. Старая система образования выдает порцию неинтересных и бесполезных знаний. А котики обучают, развлекая.

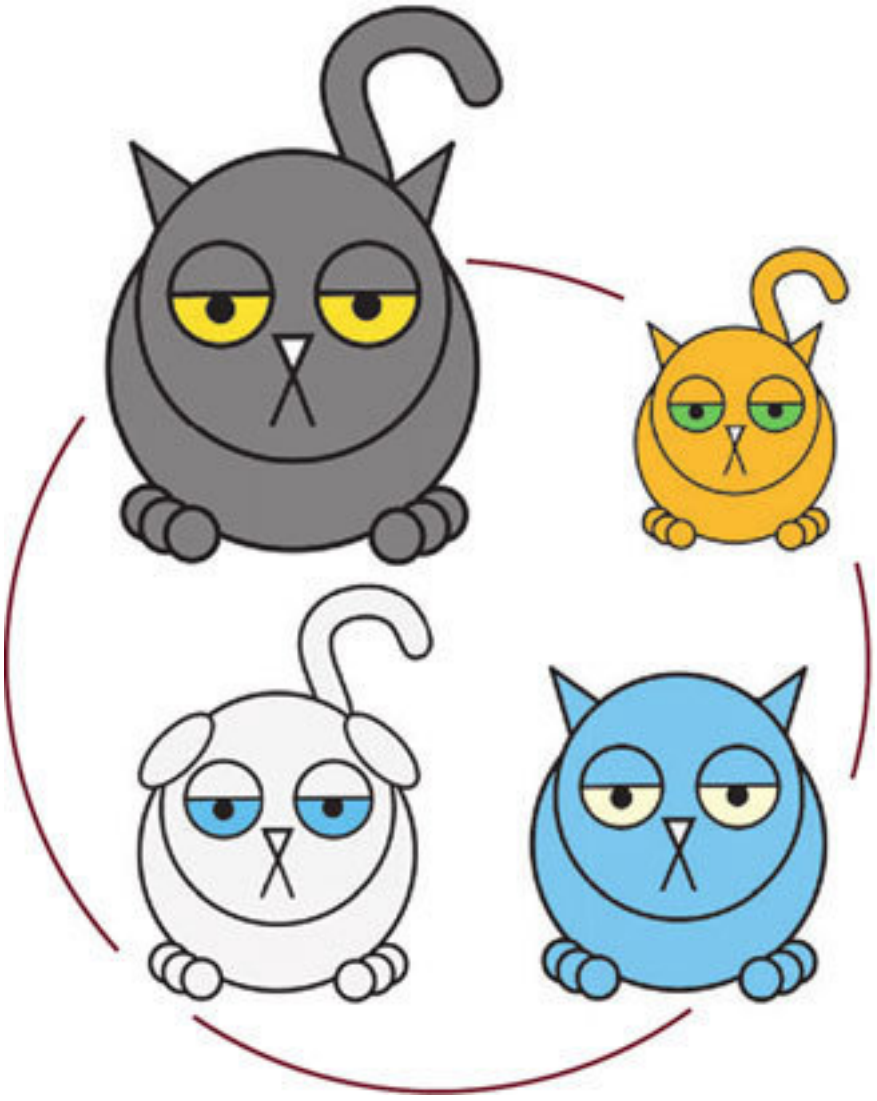
Когда мы изучаем данные, мы осознаем, что задача – найти соломинку в стоге иголок. И понять, сколько ещё стогов и соломы найдем дальше. Статистика в бизнесе помогает нам экономить деньги и открывать новые рынки. Экономия питает амбиции и потихоньку делает жизнь людей чуточку лучше.

*Респект читателям. Респект автору.*

*Юрий Корженевский,*
*Центр Исследований и Разработки.*
*www.rnd.center*

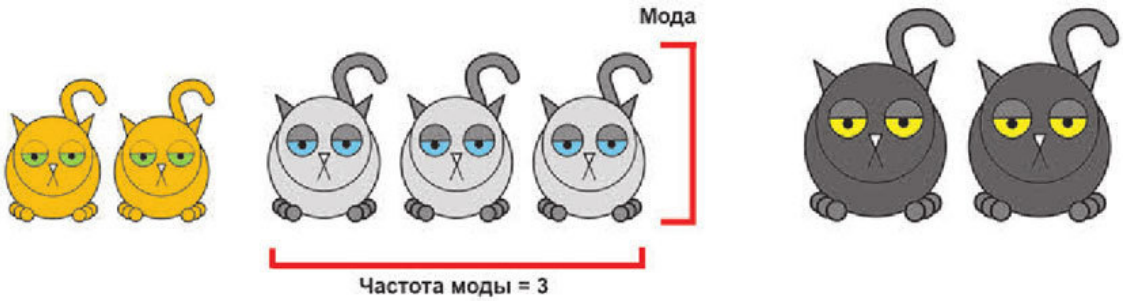
# Глава 1. Как выглядят котики или Основы описательной статистики

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?

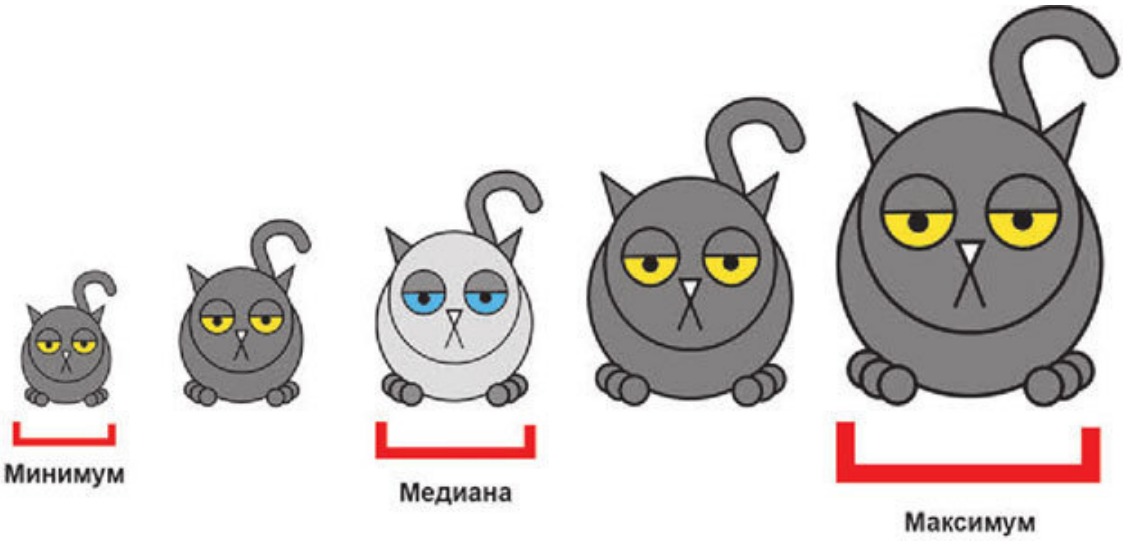


Для простоты мы возьмем такое котиковое свойство, как размер.

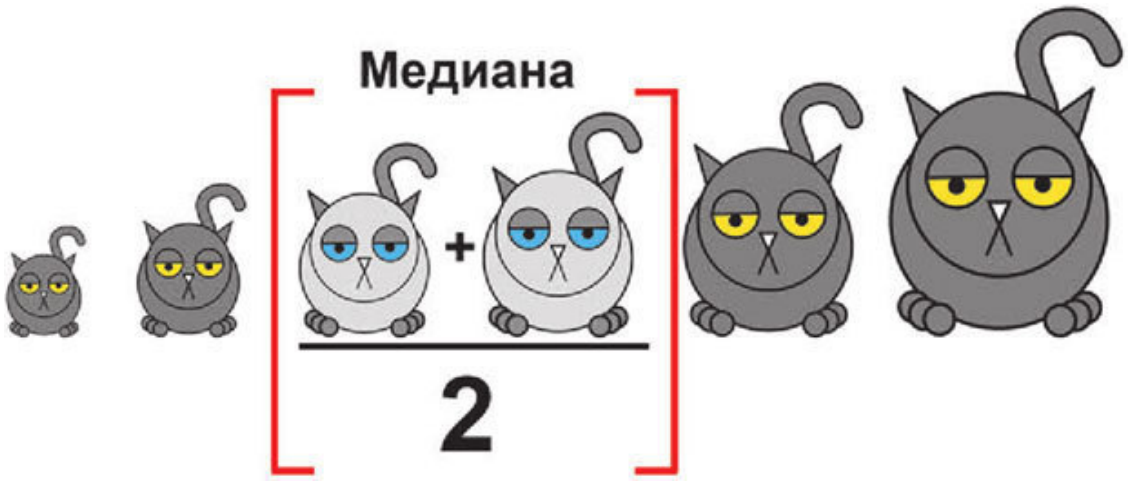
Первый и наиболее очевидный способ – посмотреть, какой размер котиков встречается чаще всего. Такой показатель называется *модой*.




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*.



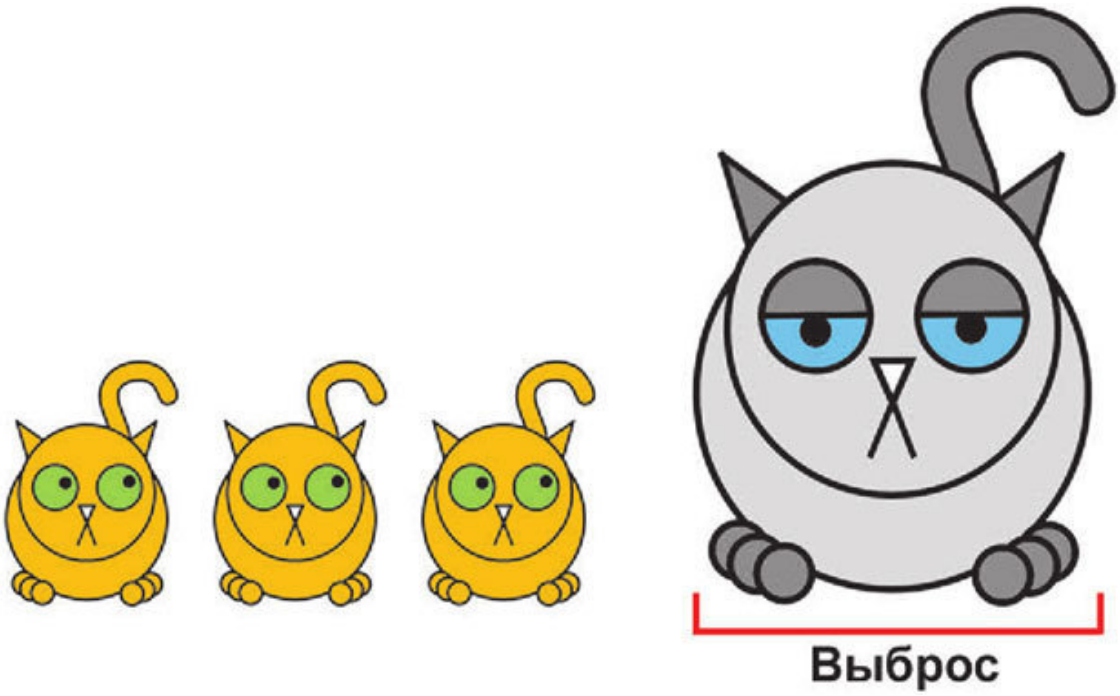

Если же посередине находятся сразу два котика (что бывает, когда их четное количество), то, чтобы найти медиану, нужно сложить их размеры и поделить это число пополам.



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.

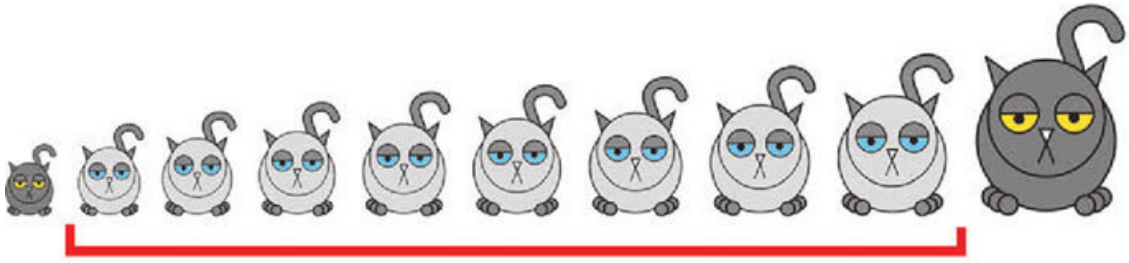

Однако, среднее арифметическое далеко не всегда является лучшим показателем типичности.

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*.



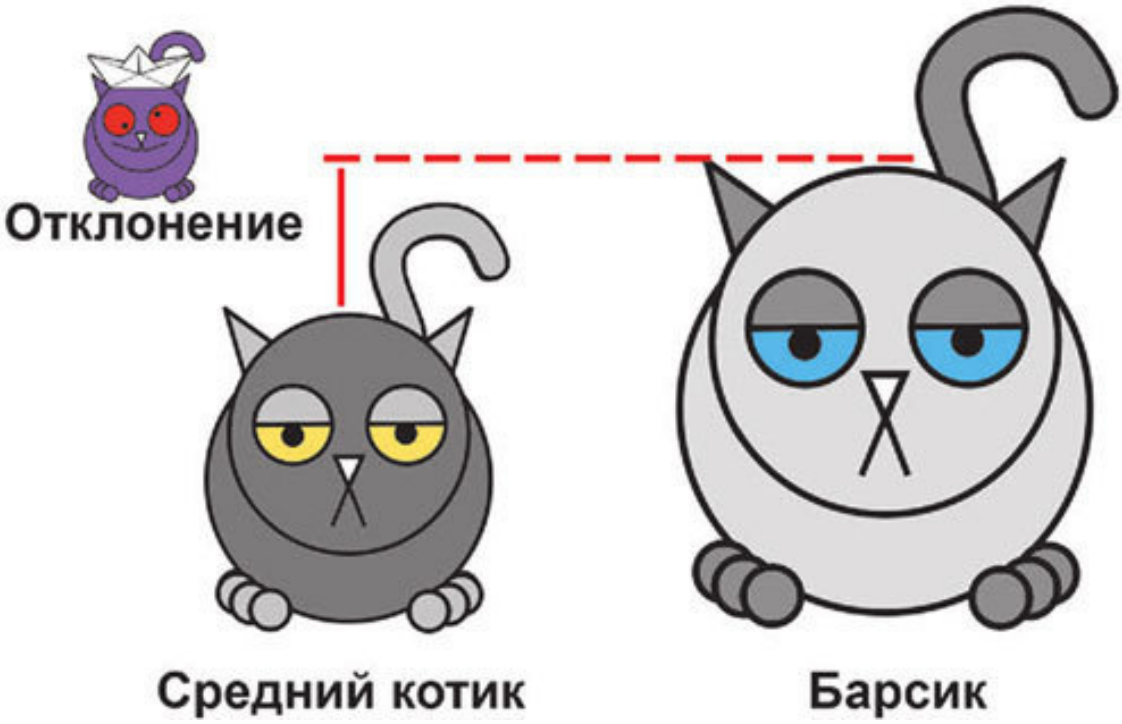
Котики для усеченного среднего

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).

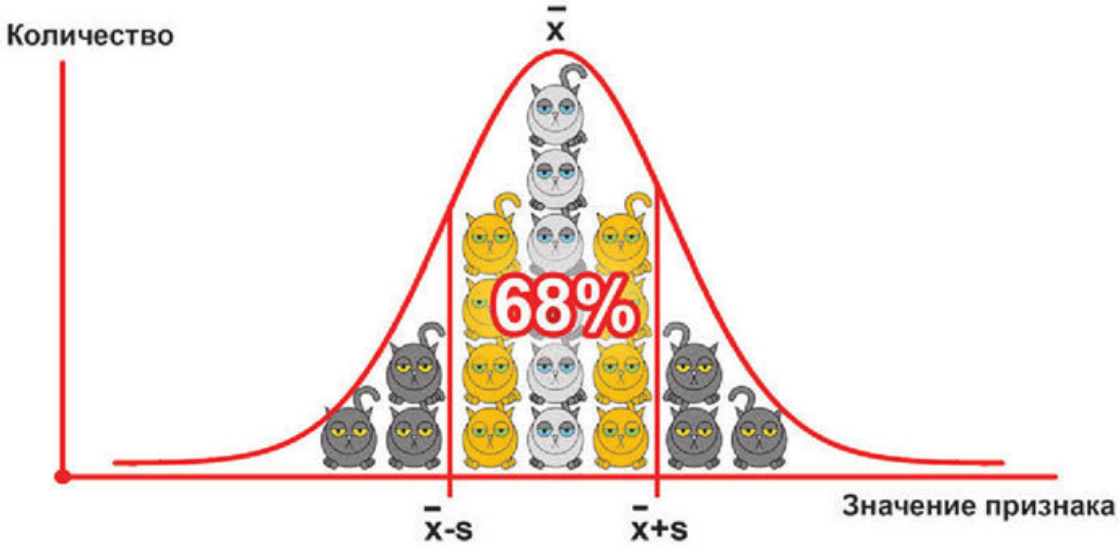
/ 3

Однако если мы это сделаем, то получим 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.


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*.


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.


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

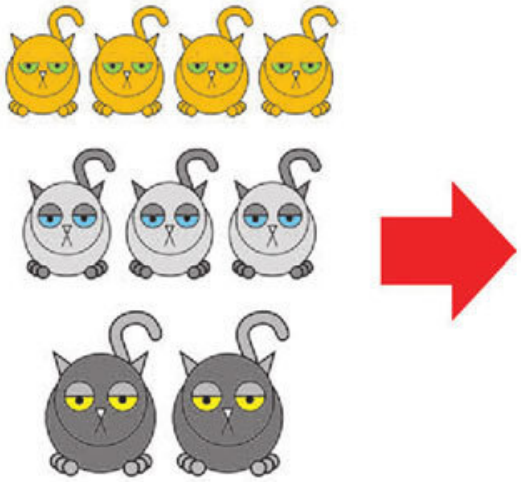
Однако не будет большой ошибкой, если вы будете пользоваться терминами *стандартное отклонение генеральной совокупности* и *стандартное отклонение выборки.* Чаще всего именно последнее и рассчитывается для реальных исследований.

# 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*.

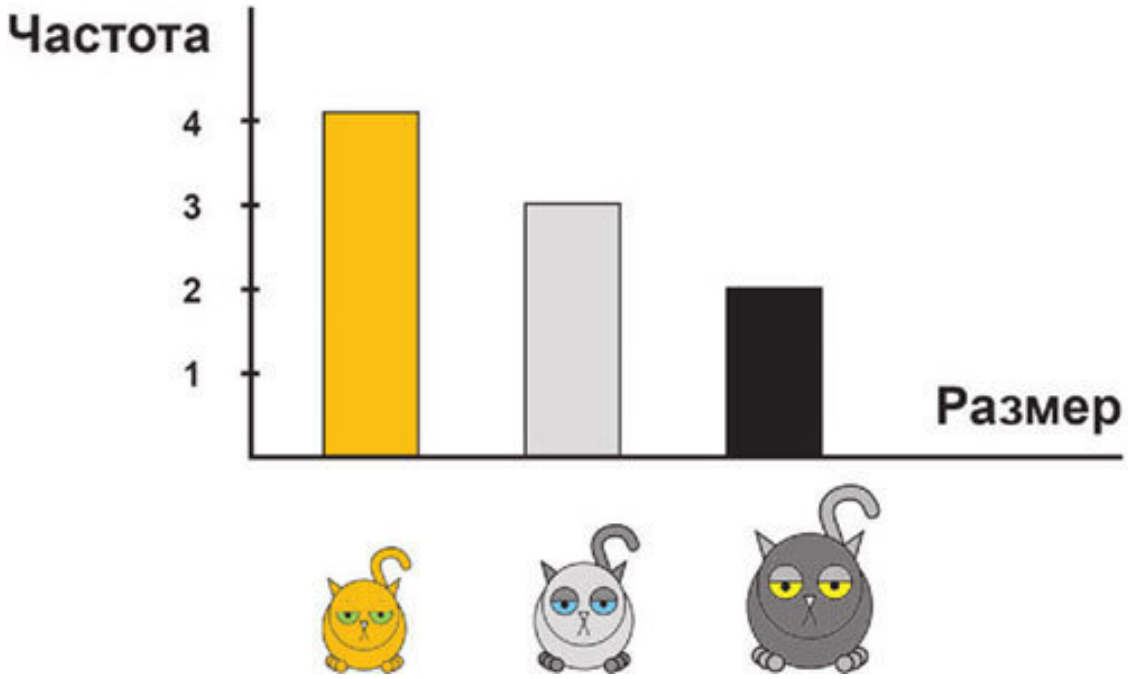
Первая группа средств показывает, сколько котиков обладает тем или иным размером. Для их использования необходимо предварительно построить так называемые *таблицы частот*. В этих таблицах есть два столбика: в первом указывается размер (или любое другое котиковое свойство), а во втором – количество котиков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

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

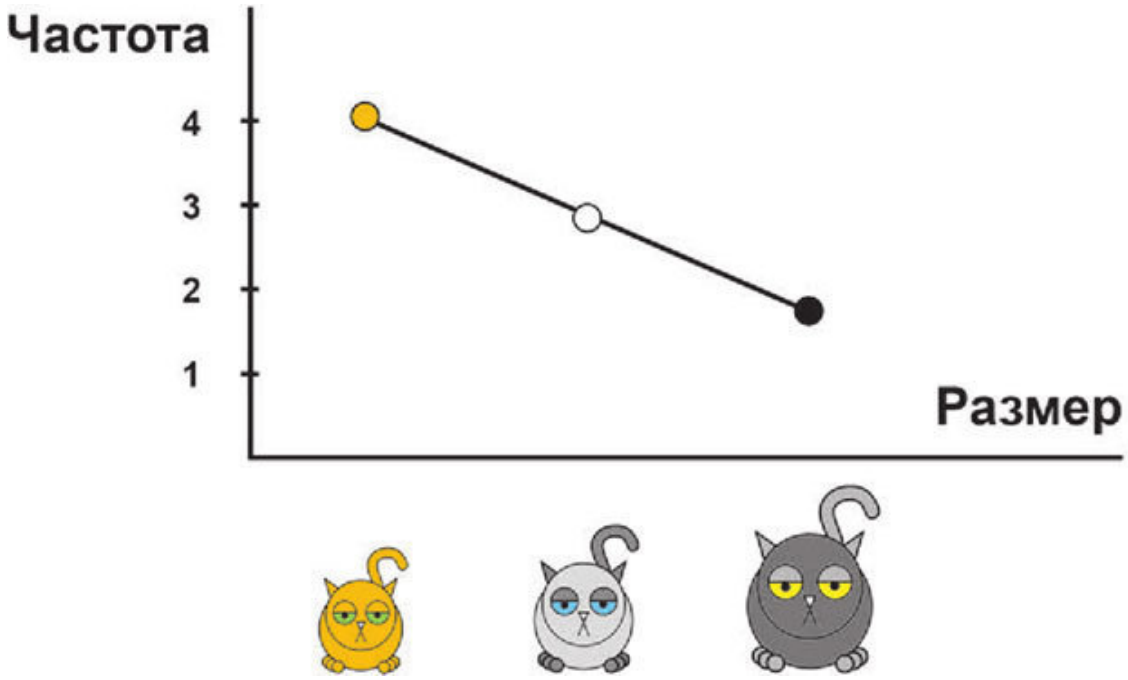


| Размер | Частота |
|---|---|
|  | 4 |
|  | 3 |
|  | 2 |

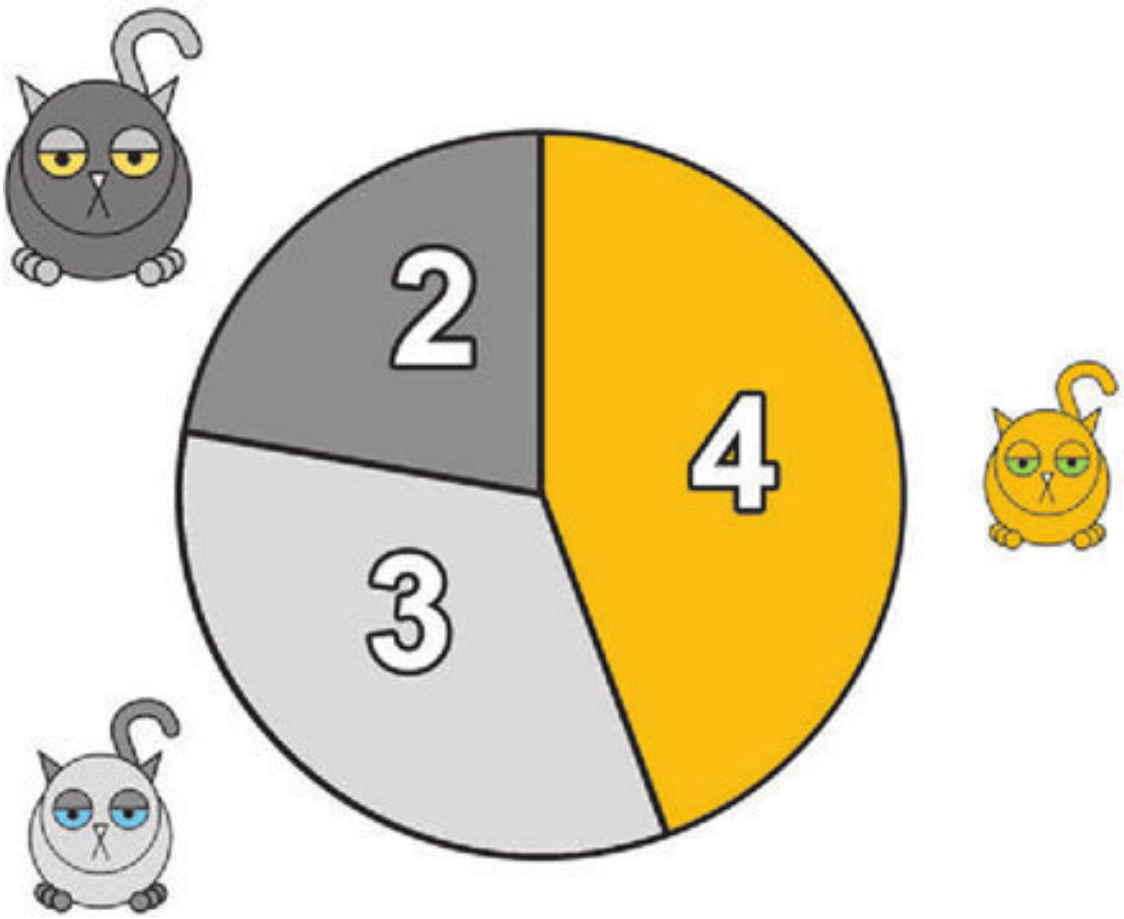
С таблицами частот можно делать много интересных вещей. Например, построить *столбиковую диаграмму*. Для этого мы откладываем две перпендикулярных линии: горизонтальная будет обозначать размер, а вертикальная – частоту. А затем – рисуем столбики, высота которых будет соответствовать количеству котиков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



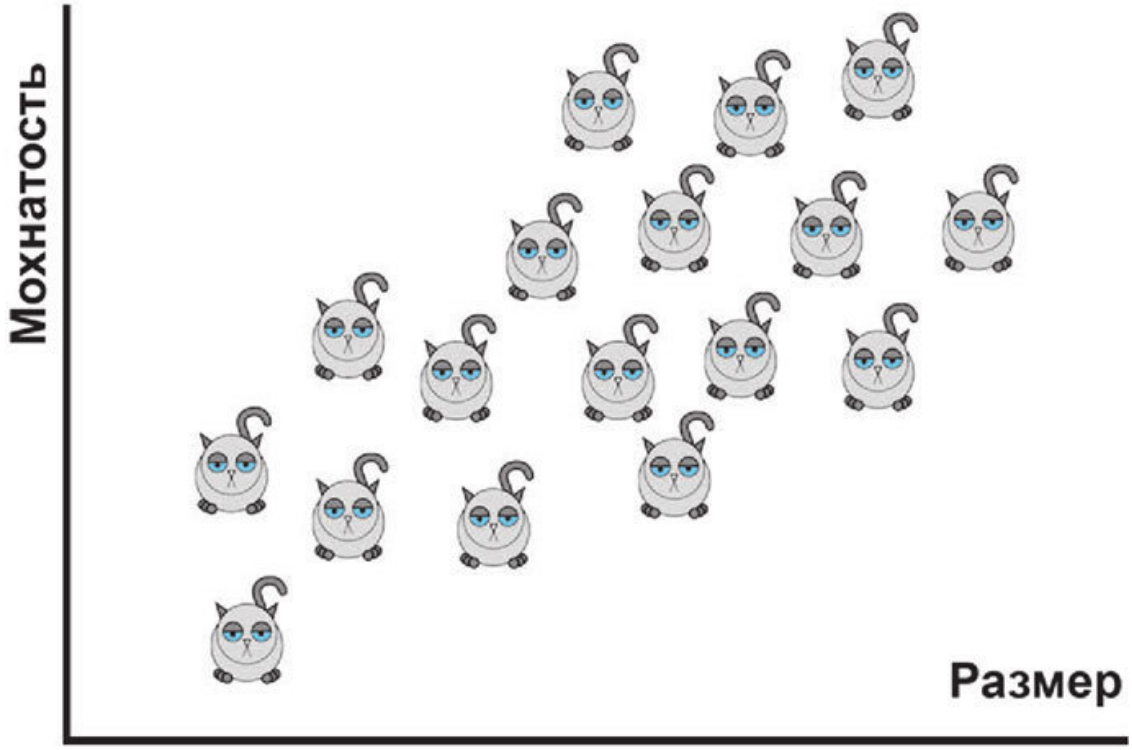

А еще мы можем вместо столбиков нарисовать точки и соединить их линиями. Результат называется *полигоном распределения*. Он довольно удобен, если котиковых размеров действительно много.



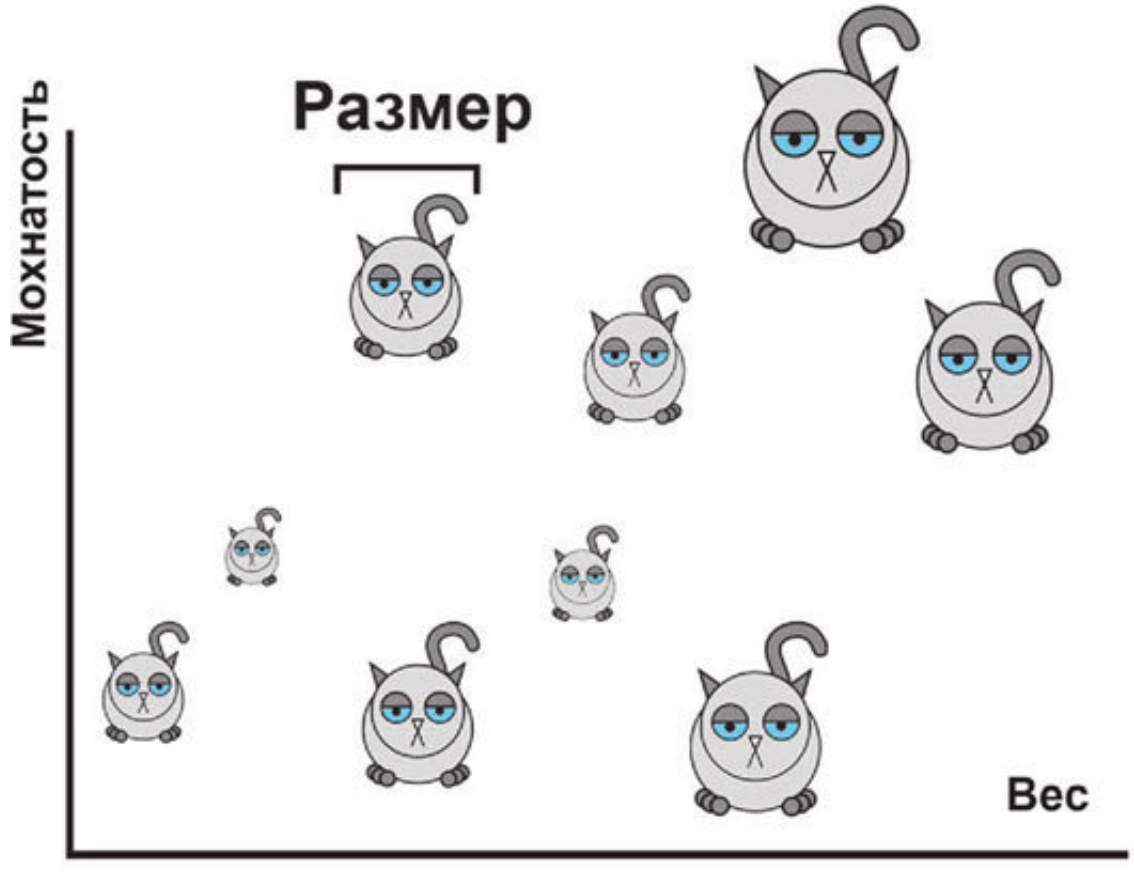

Наконец, мы можем построить *круговую диаграмму*. Величина каждого сектора такой диаграммы будет соответствовать проценту котиков определенного размера.



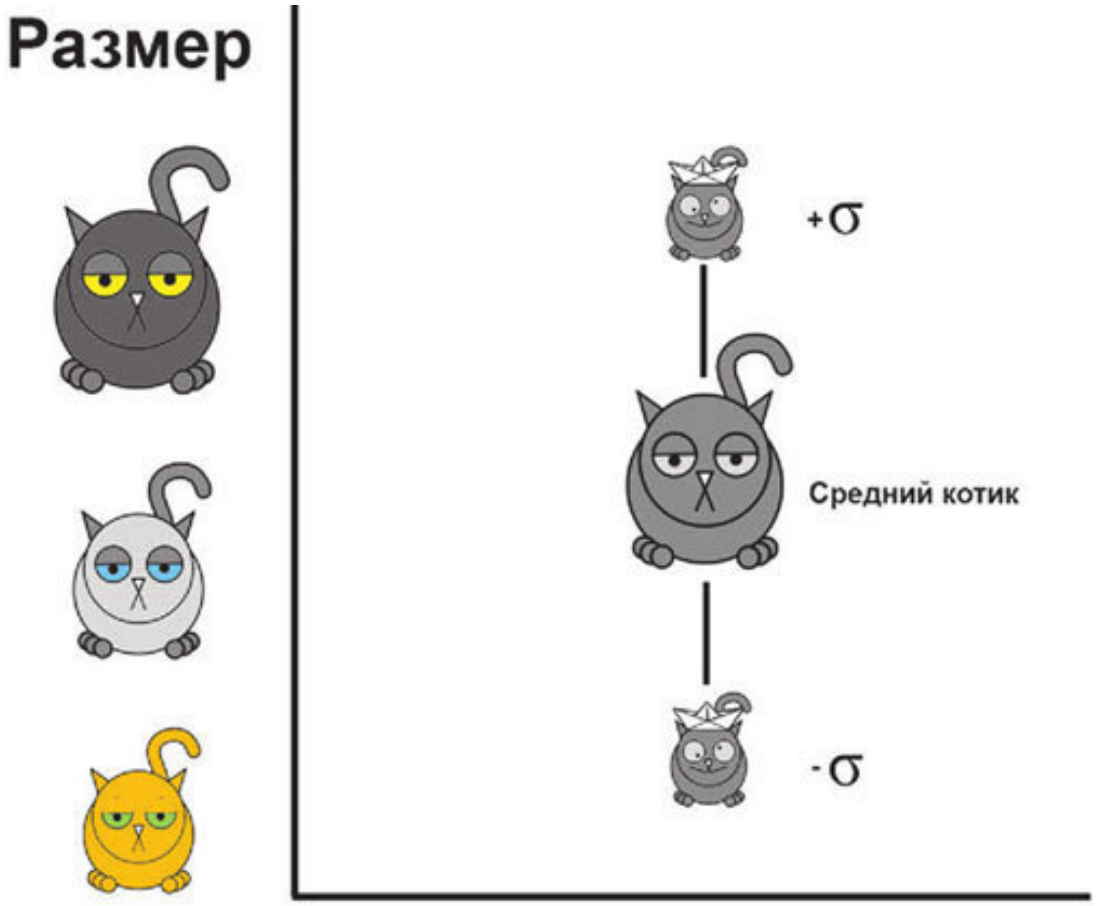

Следующая группа средств визуализации позволяет отобразить сразу два котиковых свойства. Например, размер и мохнатость. Как и в случае со столбиковыми диаграммами, первым шагом рисуются оси. Только теперь каждая из осей отображает отдельное свойство. А после этого каждый котик занимает на этом графике 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.

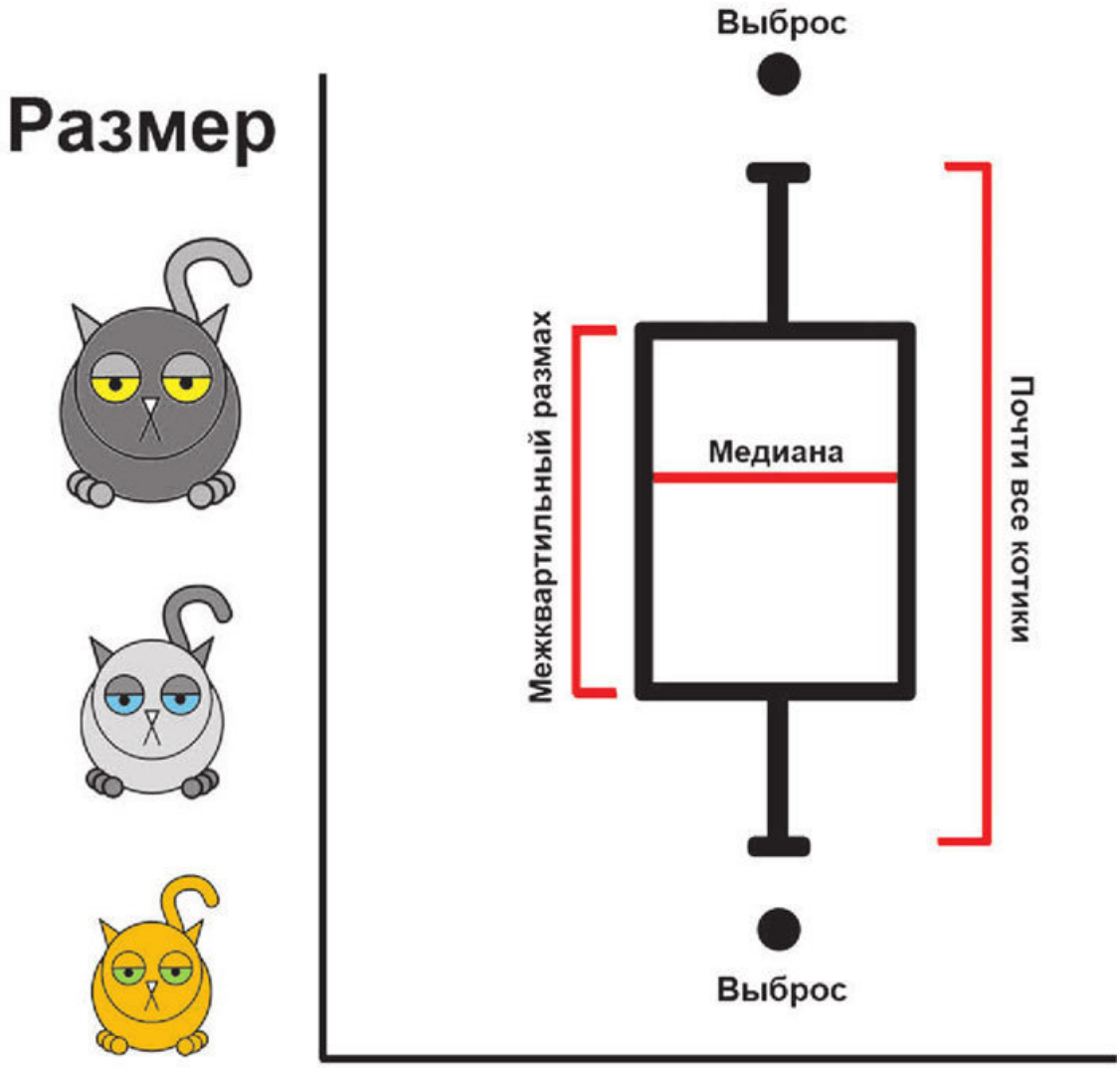



Последняя крупная группа средств визуализации позволяет графически изобразить меры центральной тенденции и меры изменчивости. В простейшем виде это точка на графике, обозначающая, где находится средний котик, и линии, длина которых указывает на величину стандартного отклонения.




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.



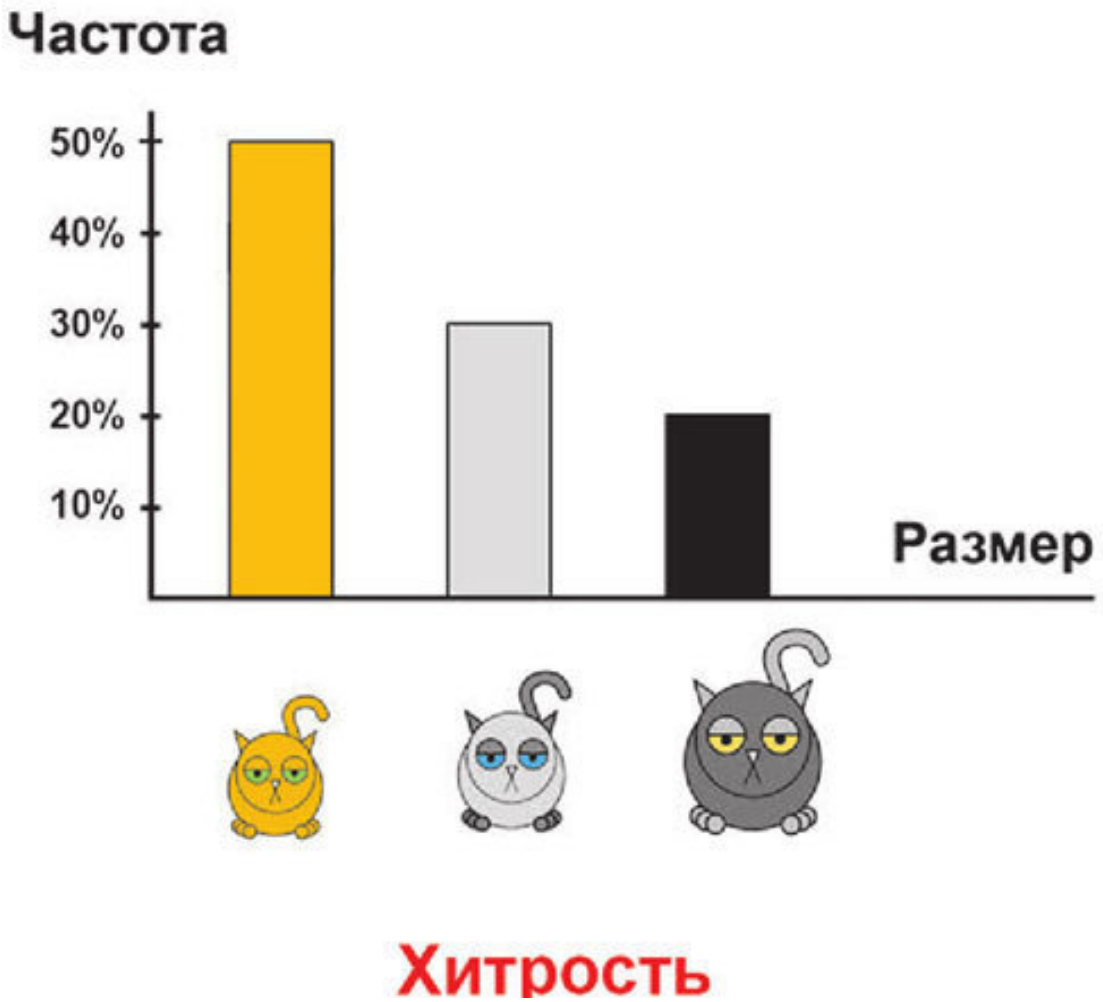

На диаграмме не должно быть ничего лишнего. Если на ней есть элемент, не несущий какой-либо смысловой нагрузки, его лучше убрать. Потому что чем больше лишних элементов, тем менее понятной будет диаграмма.

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.

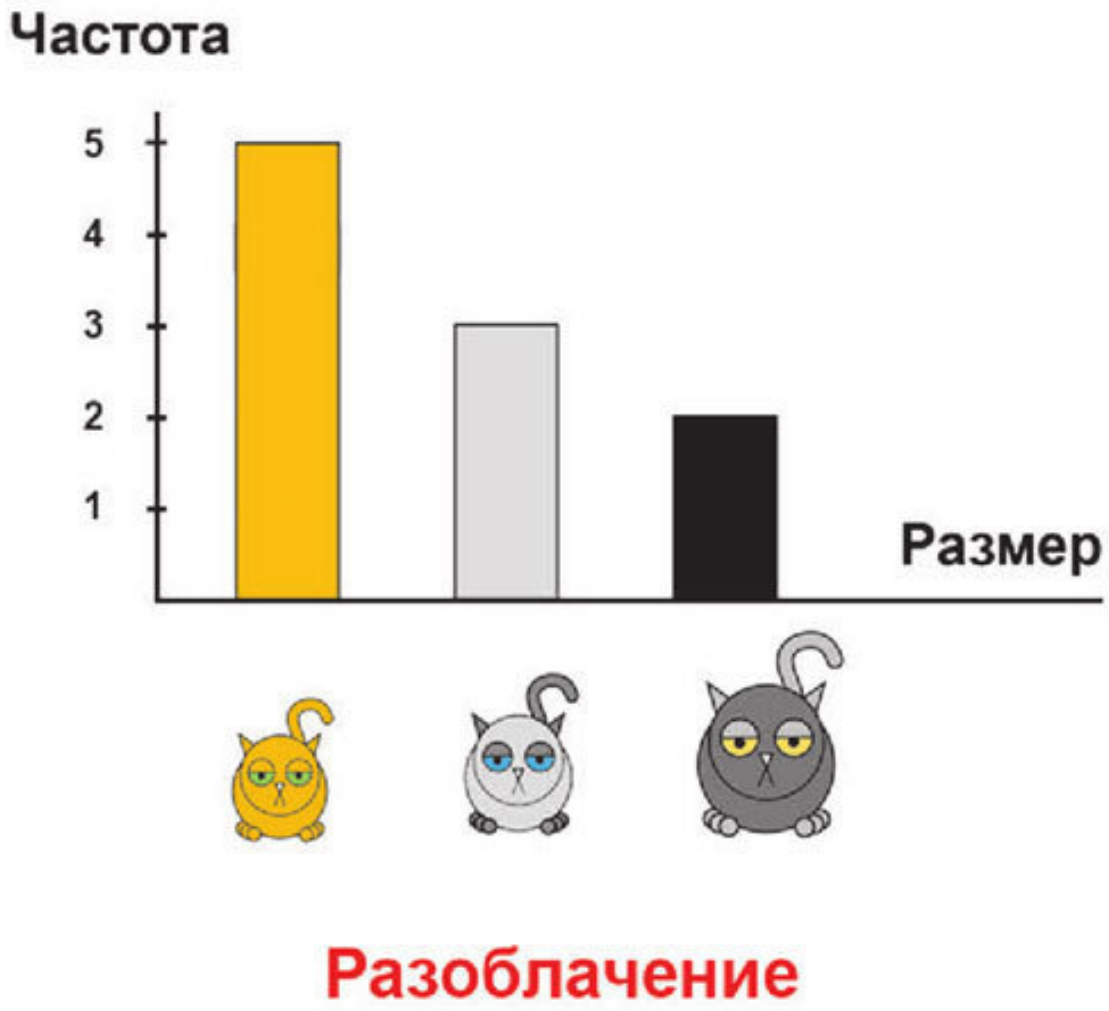
## НЕМАЛОВАЖНО ЗНАТЬ!

### Темная сторона визуализации

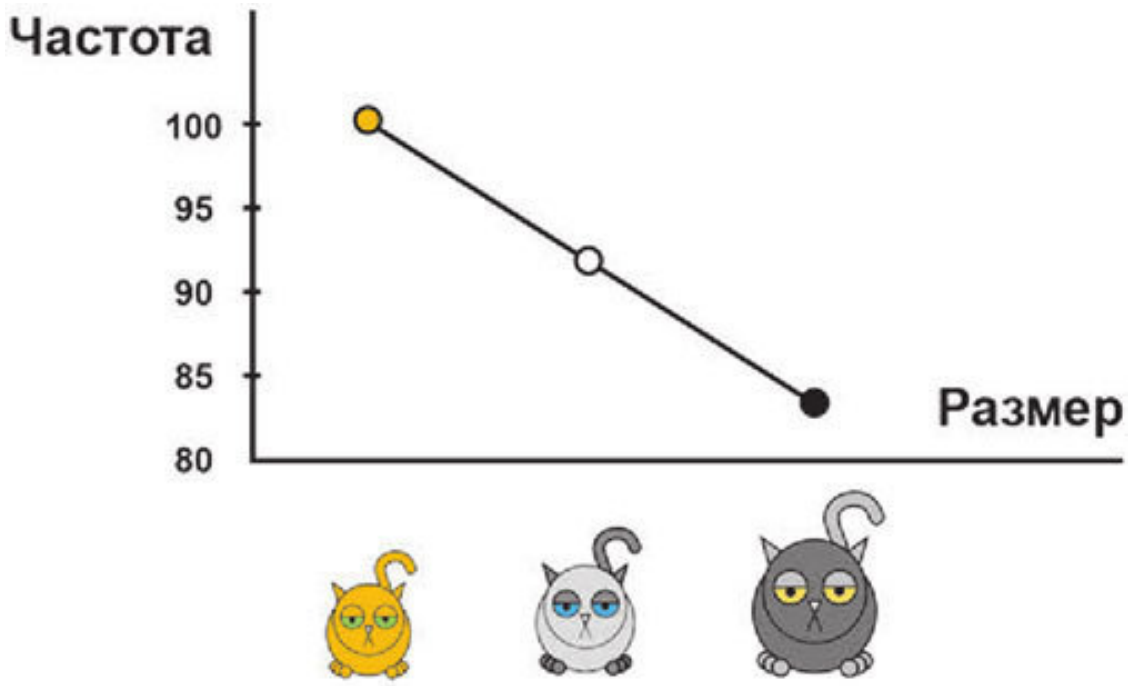
Несмотря на то, что средства визуализации помогают облегчить восприятие данных, они так же легко могут ввести в заблуждение, чем, к сожалению, часто пользуются разные хитрые люди. Ниже мы приведем самые распространенные способы обмана с помощью диаграмм и графиков.



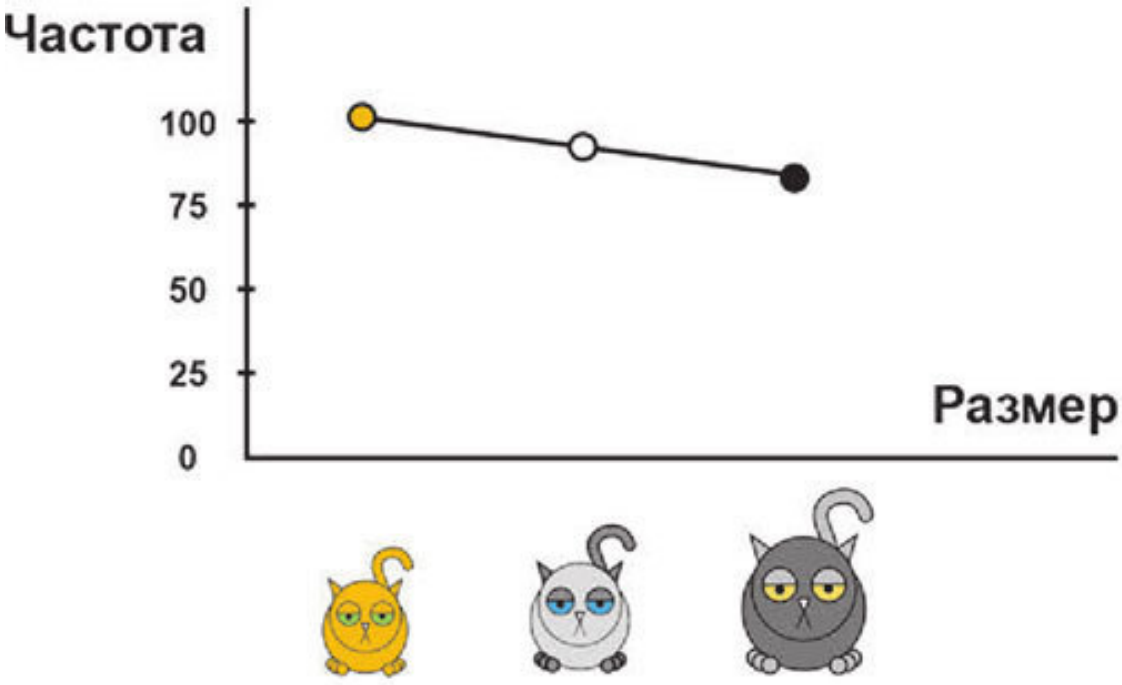

**Проценты вместо абсолютных величин.** Очень часто, чтобы придать своим данным значимости, хитрые люди переводят абсолютное количество котиков в проценты. Согласитесь, что результаты, полученные на 50 % котиков, выглядят куда солиднее, чем на пяти.




**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.



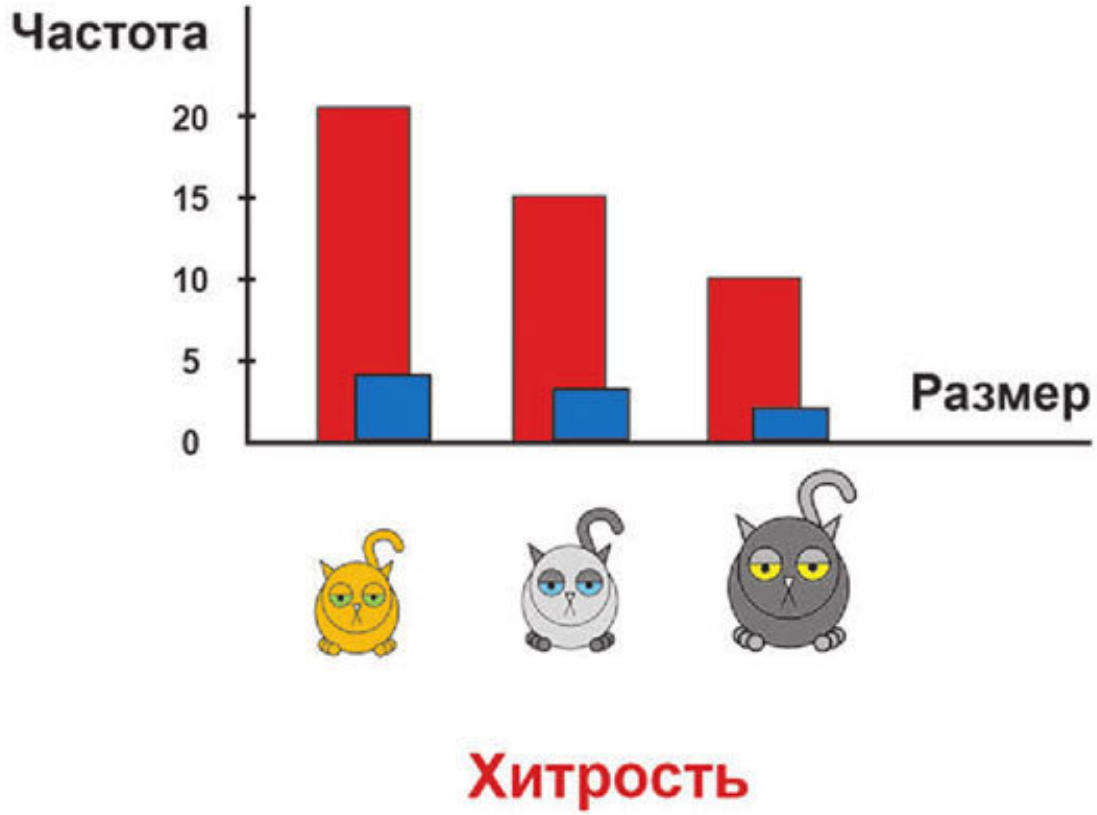

Хитрость

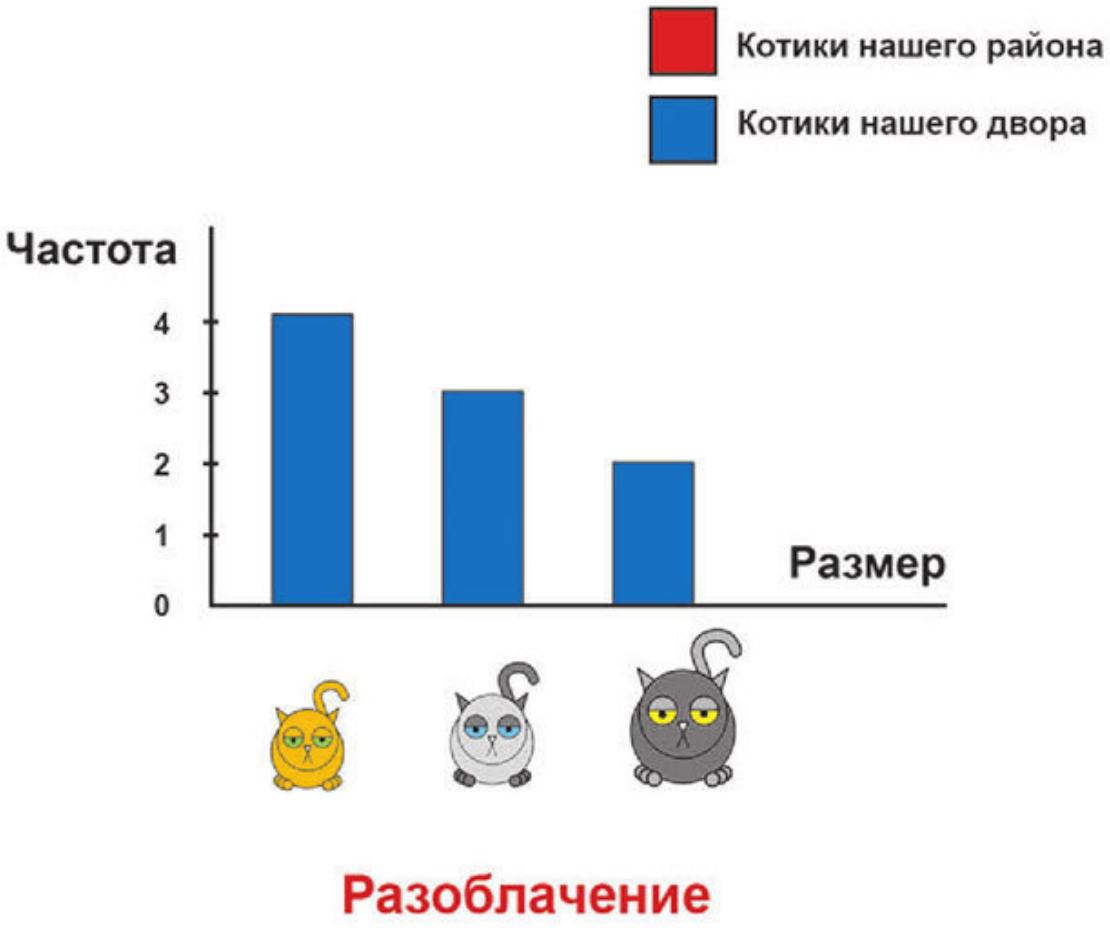



Разоблачение

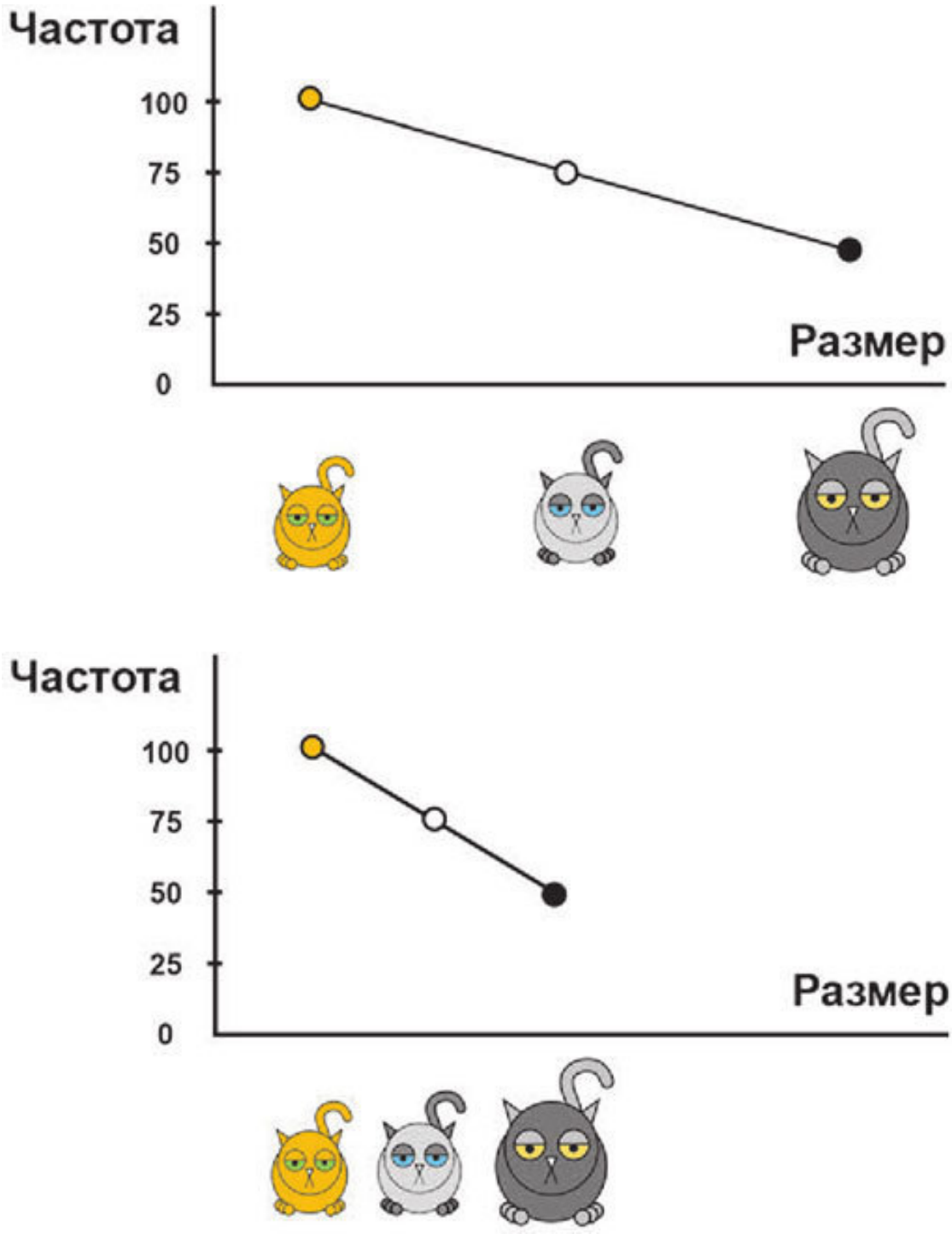
**Сокрытие данных.** Если же цель хитрого человека в том, чтобы скрыть значимые различия в данных, то их можно разместить на одной шкале с другими данными, которые на

порядок отличаются от первых. На их фоне любые различия или изменения будут выглядеть незначительно.





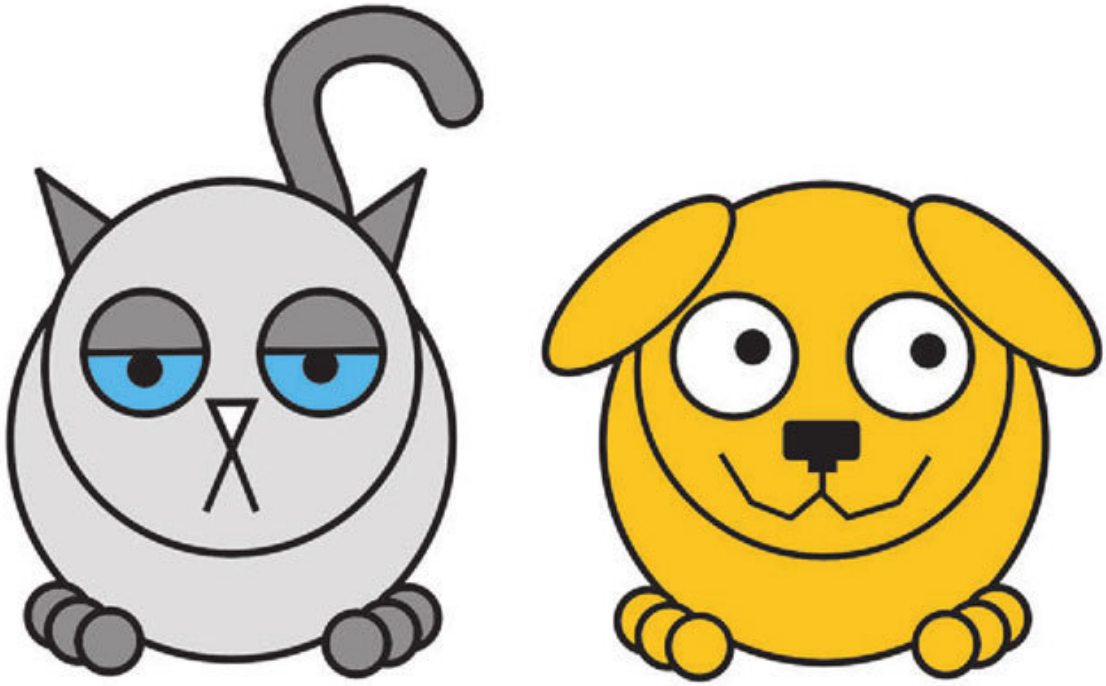

**Изменение масштабов.** Более мягкий вариант создания иллюзии значимости – это изменение масштабов шкал. В зависимости от масштаба одни и те же данные будут выглядеть по-разному.



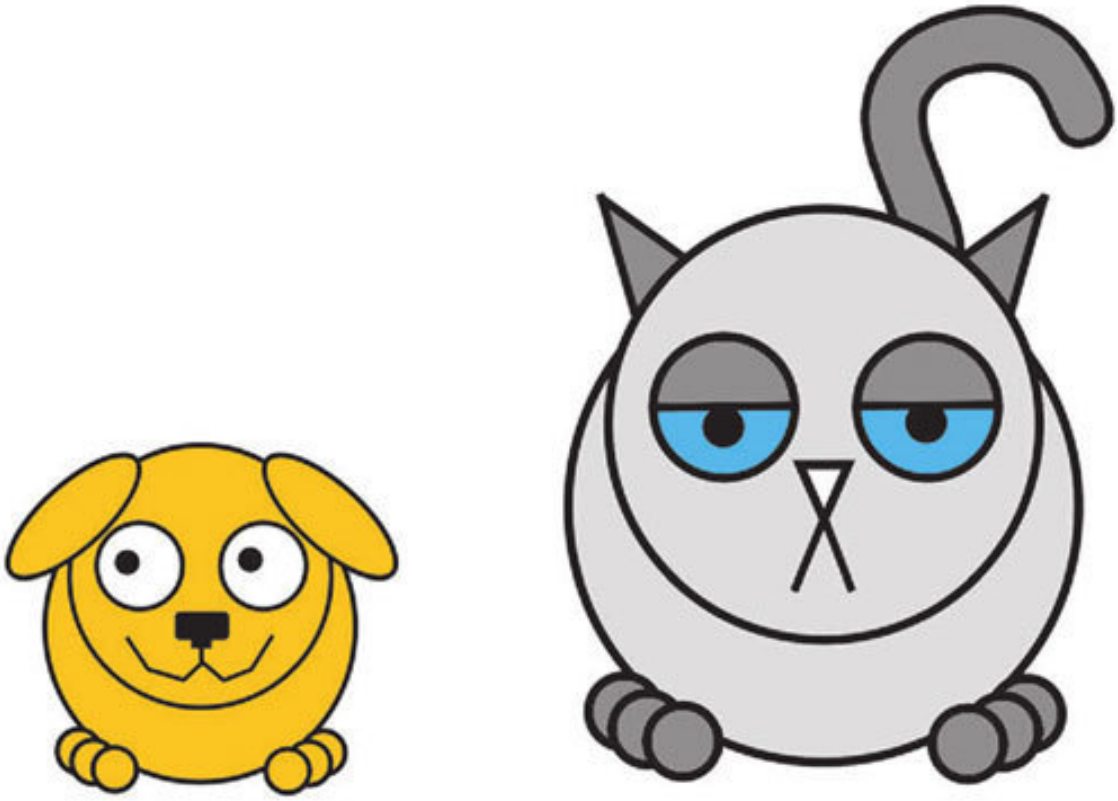

Таким образом, надо быть очень аккуратным, интерпретируя данные, представленные в виде графиков и диаграмм. Гораздо меньше подвержены манипуляции данные, представленные в табличной формуле. Однако и здесь можно использовать некоторые хитрости, которые могут ввести в заблуждение непосвященную публику.

# Глава 3. Чем отличаются котики от песиков или Меры различий для несвязанных выборок

Есть котики, а есть песики. Песики чем-то похожи на котиков: у них четыре лапы, хвост и уши. Однако они также во многом различаются – например, котики мяукают, а песики лают.



Но не все различия между ними настолько очевидны. Например, довольно трудно судить о том, различаются ли песики и котики по размеру – ведь есть как очень большие котики, так и очень маленькие песики.



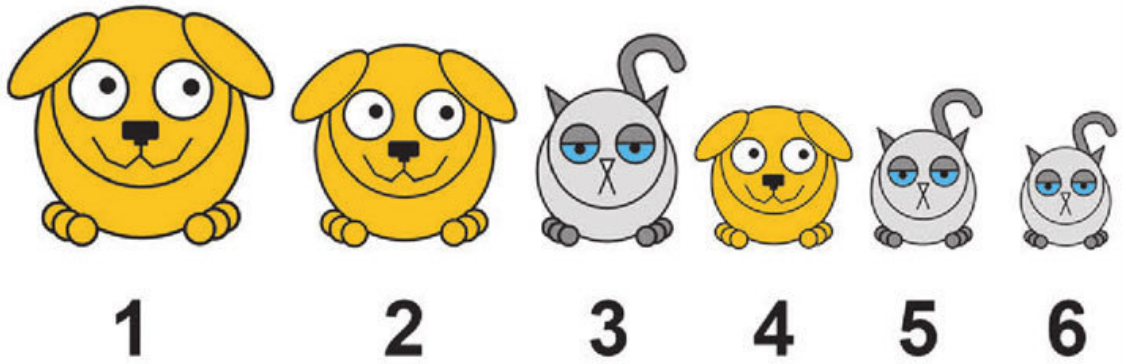
Чтобы понять, насколько они отличаются друг от друга, необходимы так называемые *меры различий для несвязанных выборок*. Большая часть таких мер показывает, насколько типичный песик отличается от типичного котика. Например, самая популярная из них – *t-критерий Стьюдента для несвязанных выборок* – оценивает, насколько различаются их средние размеры.

Чтобы рассчитать этот критерий, необходимо из среднего размера песиков вычесть средний размер котиков и поделить их на *стандартную ошибку* этой разности. Последняя вычисляется на основе стандартных отклонений котиковых и песиковых размеров и нужна для приведения t-критерия к нужной размерности.

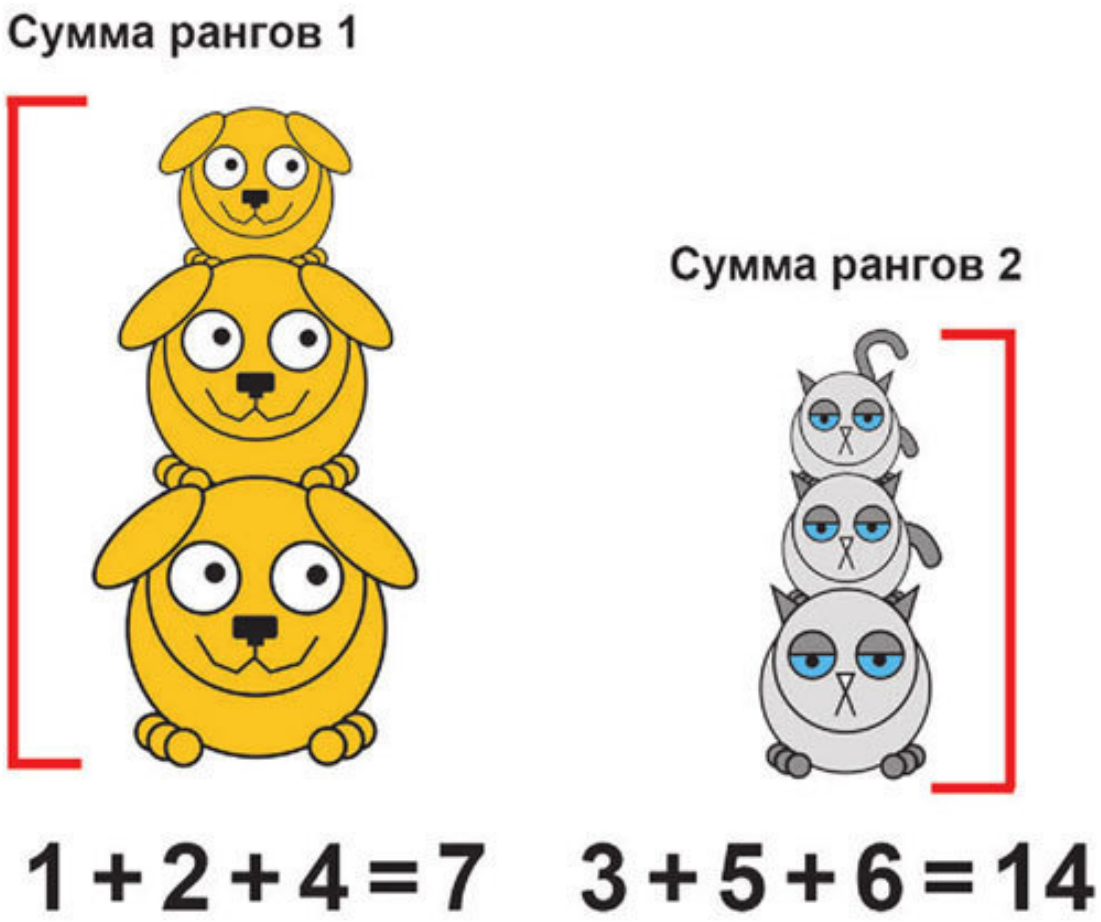

Если разность средних достаточно большая, а стандартная ошибка очень маленькая, то значение t-критерия будет весьма внушительным. А чем больше t-критерий, тем с большей уверенностью мы можем утверждать, что в среднем песики отличаются от котиков.

К большому сожалению, поскольку формула t-критерия включает в себя средние значения, то этот критерий будет давать неадекватные результаты при наличии котиков и песиков аномальных размеров (т. е. выбросов, о которых подробно рассказано в первой главе). Чтобы этого избежать, вы можете либо исключить этих котиков и песиков из анализа, либо воспользоваться непараметрическим *U-критерием Манна-Уитни*. Этот критерий, кстати, используется и в тех ситуациях, когда точные (сантиметровые) размеры животных нам неизвестны.

Чтобы рассчитать критерий Манна-Уитни, необходимо выстроить всех песиков и котиков в один ряд, от самого мелкого к самому крупному, и назначить им ранги. Самому большому зверьку достанется первый ранг, а самому маленькому – последний.




После этого мы снова делим их на две группы и считаем суммы рангов отдельно для песиков и для котиков. Общая логика такова: чем сильнее будут различаться эти суммы, тем больше различаются песики и котики.




Наконец, мы проводим некоторые преобразования (которые в основном сводятся к поправкам на количество котиков и песиков) и получаем критерий Манна-Уитни, по которому судим, в действительности ли котики и песики отличаются по размеру.

Помимо определения различий между типичными представителями котикового и песикового видов, в некоторых случаях нас могут интересовать различия по их разнообразию. Иными словами, мы можем посмотреть, являются ли песики более разнообразными по размеру, чем котики, или же нет. Для этого мы можем воспользоваться *F-критерием равенства дисперсий Фишера,* который укажет нам, насколько различаются между собой эти показатели.


Необходимо заметить, что в этой формуле сверху всегда должна стоять большая дисперсия, а снизу – меньшая.

Все вышеперечисленные критерии замечательно работают в случаях, когда нам известны точные или хотя бы приблизительные размеры котиков и песиков. Однако такие ситуации встречаются далеко не всегда. Иногда мы можем иметь только указание на то, является ли наш зверь большим или маленьким. В таких нелегких условиях определить различия между котиками и песиками нам поможет *критерий Хи-квадрат Пирсона.*

Чтобы вычислить этот критерий, нужно построить так называемые *таблицы сопряженности*. В простейшем случае это таблицы 2×2, в каждой ячейке которых – количество (или, по-научному, частота) песиков и котиков определенного размера. Впрочем, бывают таблицы сопряженности и с большим количеством столбцов и строчек.


Очевидно, что если котики и песики как биологические виды не отличаются по размеру, то больших котиков должно быть столько же, сколько и больших песиков (в процентном соотношении). И основная идея критерия Хи-квадрат состоит в том, чтобы сравнить такую таблицу, в которой песики не отличаются от котиков (иначе – *таблицу теоретических частот*), с той, что есть у нас (*таблицей эмпирических частот*).

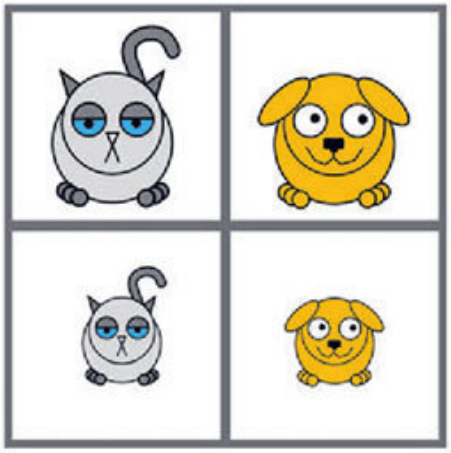


**Таблица эмпирических частот**

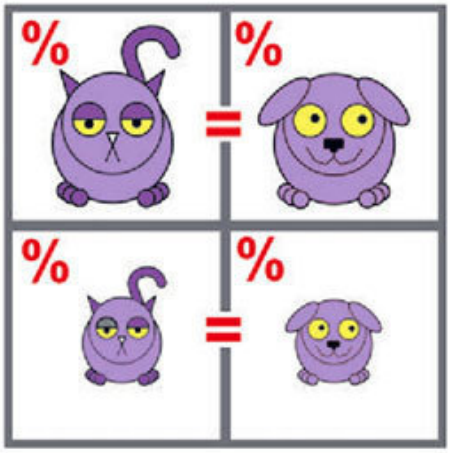
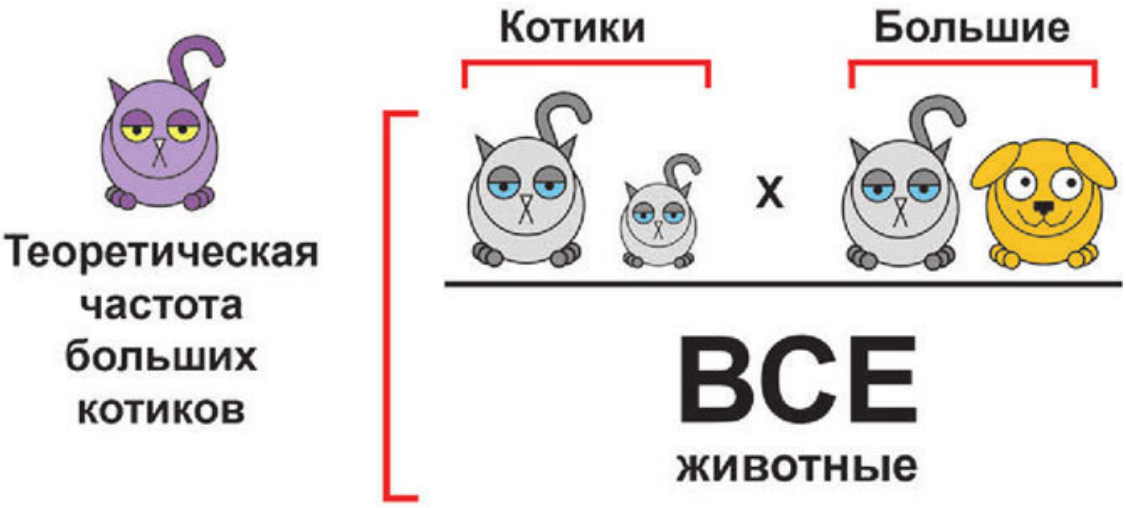
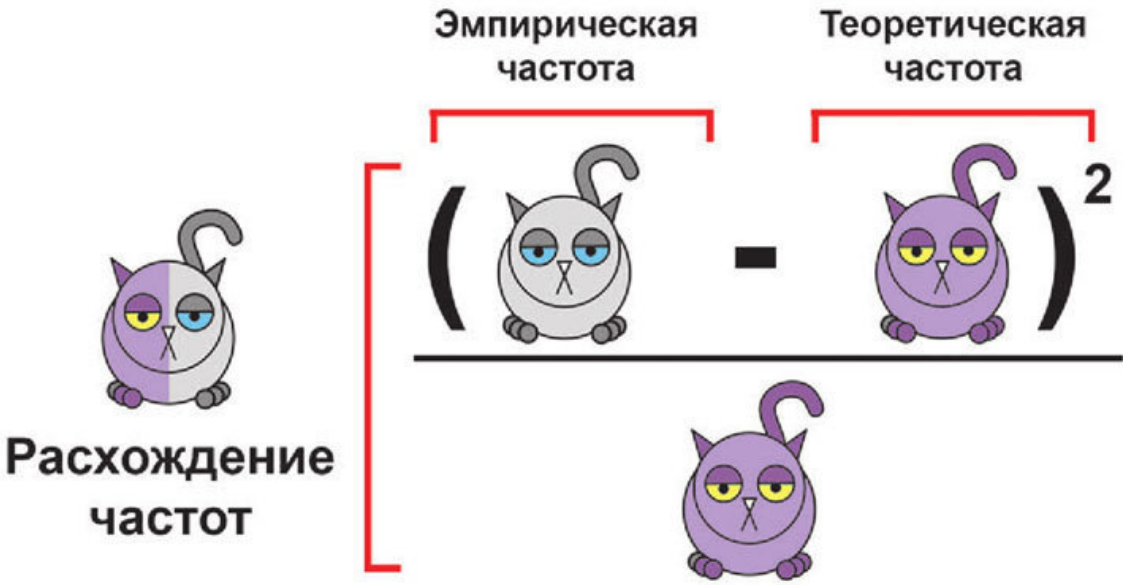



**Таблица теоретических частот**

Перво-наперво необходимо получить таблицу теоретических частот. Для этого для каждой ячейки подсчитывается *теоретическая частота* по такой формуле.




Следующим шагом мы смотрим, насколько сильно различаются между собой соответствующие ячейки в наших таблицах. Делается это вот так.




Квадрат в числителе этой формулы убирает знак, а знаменатель приводит Хи-квадрат в нужную размерность. Заметим, что если теоретическая частота равна эмпирической, то, применив эту формулу, мы получим 0.

Последним шагом мы складываем все получившиеся значения. Это и будет Хи-квадрат Пирсона. Чем он больше, тем сильнее отличаются песики от котиков.


Помимо всего вышеперечисленного существуют и другие статистические критерии, которые позволяют нам определить, чем песики отличаются от котиков. Они, как правило, имеют разные механизмы вычисления и требования к данным. Но вне зависимости от того, каким критерием вы воспользовались, мало просто его вычислить. Необходимо еще и уметь его интерпретировать. И этому вопросу будет посвящена следующая глава.

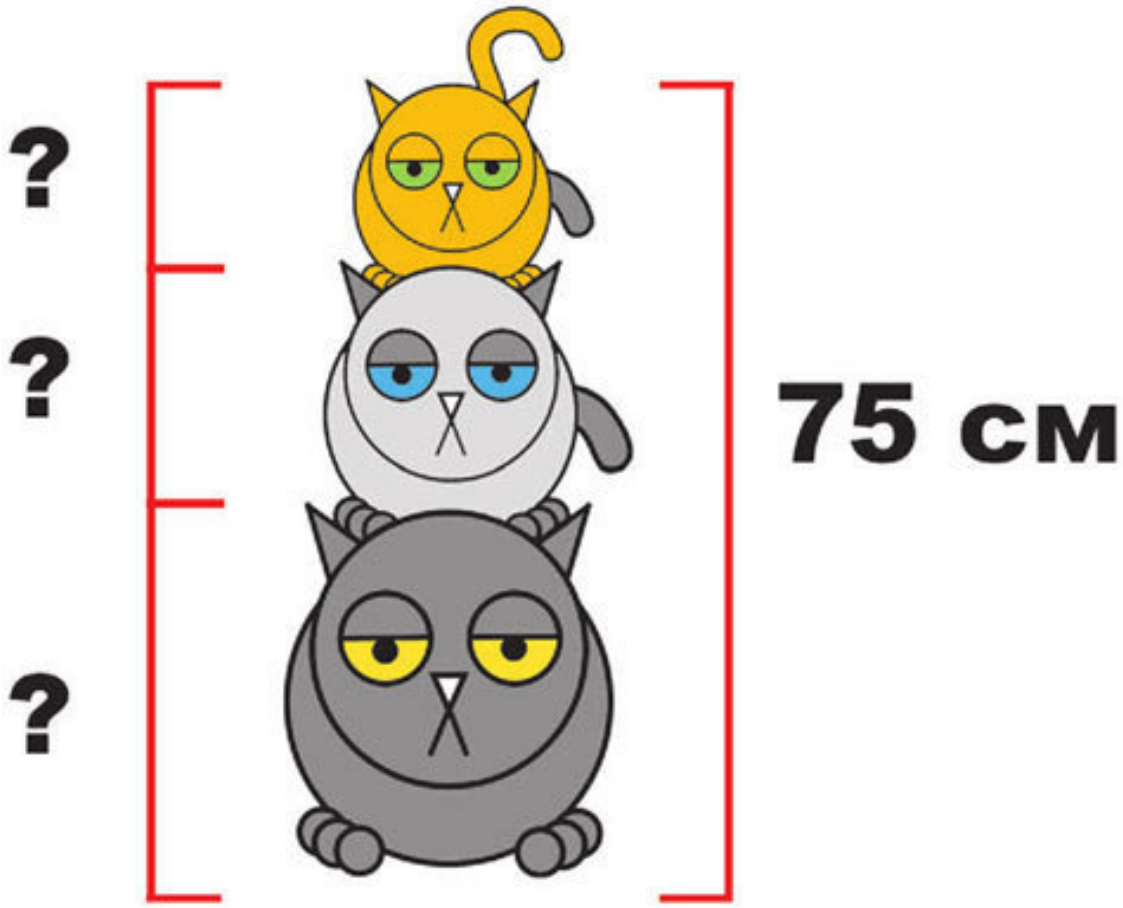
## НЕМАЛОВАЖНО ЗНАТЬ!

### Загадочные степени

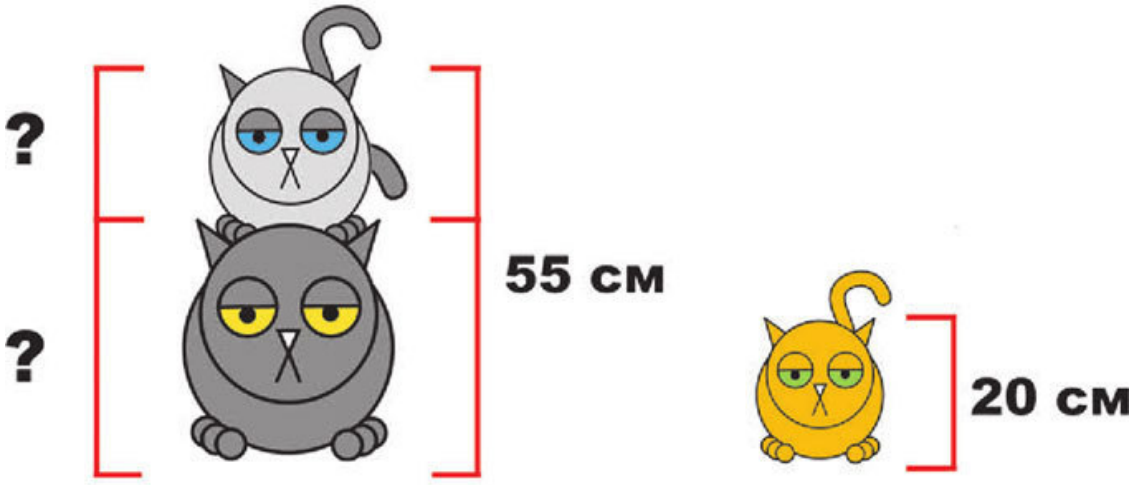
### свободы

Многих изучающих статистику ставит в тупик понятие «степень свободы», которое часто встречается в учебниках.

Предположим вы знаете, что сумма размеров всех ваших котиков равна 75 см, но не знаете величину каждого конкретного котика. Эти величины будут неизвестны ровно до тех пор, пока вы не начнете их измерять.

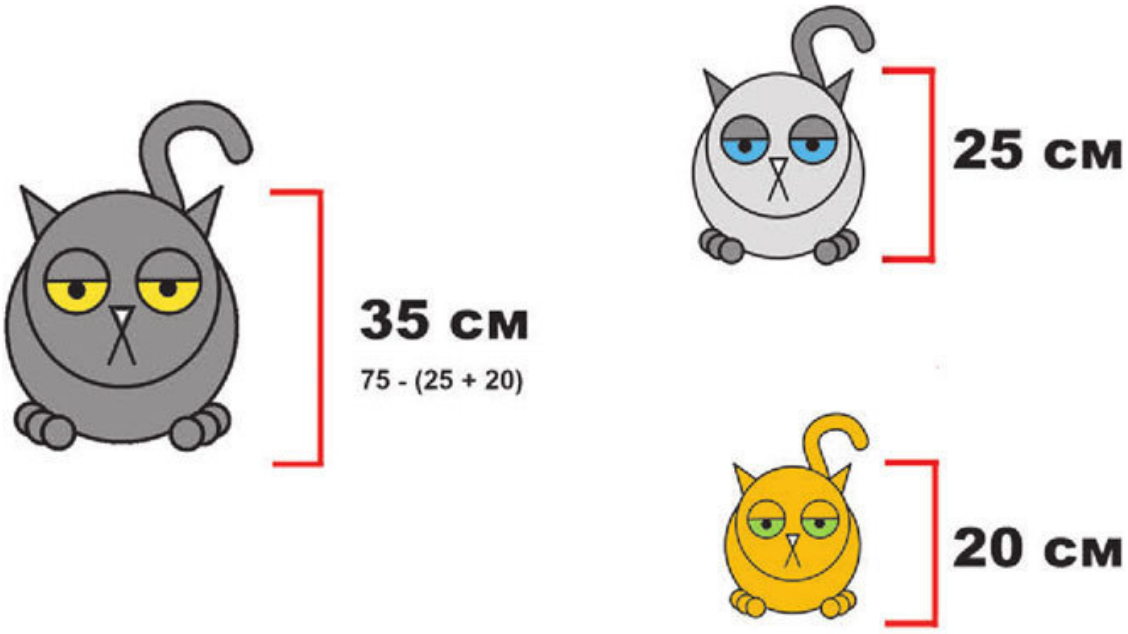



Представим, что вы узнали размер первого котика и он оказался равен 20 см. После несложных вычислений можно убедиться, что сумма размеров оставшихся котиков будет 55 см. При этом их конкретные размеры до сих пор неизвестны.

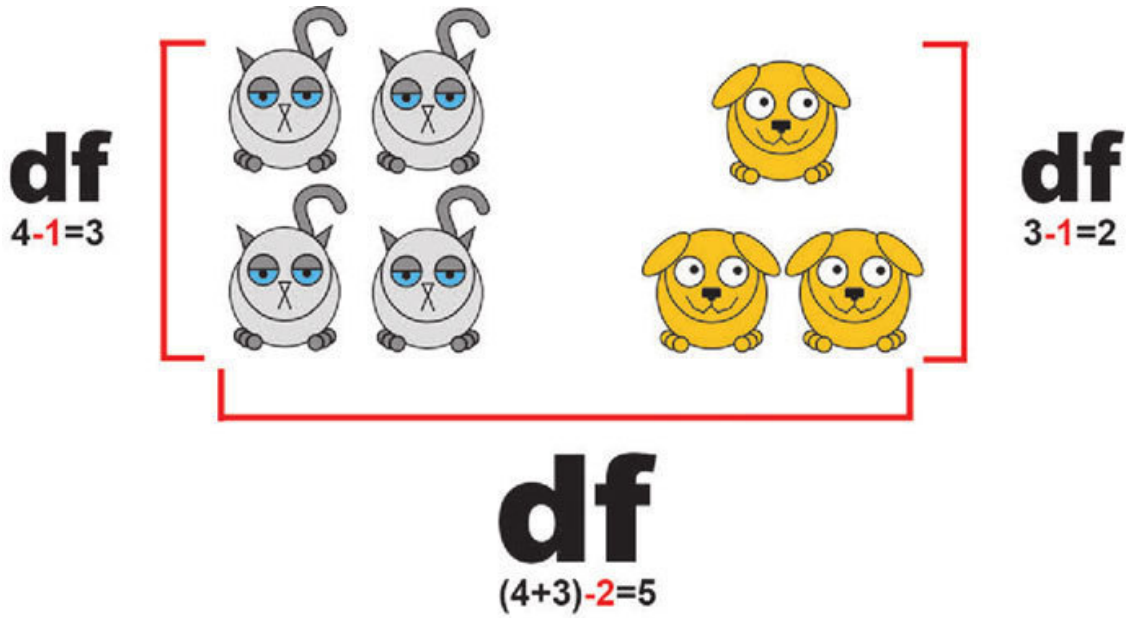



Измерим второго котика. Он оказался равен 25 см. Что мы можем сказать о размере третьего? А то, что он перестал быть неизвестным – теперь мы можем его вычислить. И действительно, вычтя из общей суммы размеры первого и второго котика мы получаем размер третьего.

Число степеней свободы – это то количество котиков, которое мы должны измерить, чтобы однозначно узнать размер всех котиков при известном среднем или дисперсии. Если у вас только одна котиковая выборка, то это количество котиков минус единица.




Если к ним добавляются еще и выборка пёсиков (например, при вычислении t-критерия Стьюдента), то общее количество степеней свободы – это просто сумма степеней свободы котиков и пёсиков. Или по-другому – общее количество животных вычесть двойку.

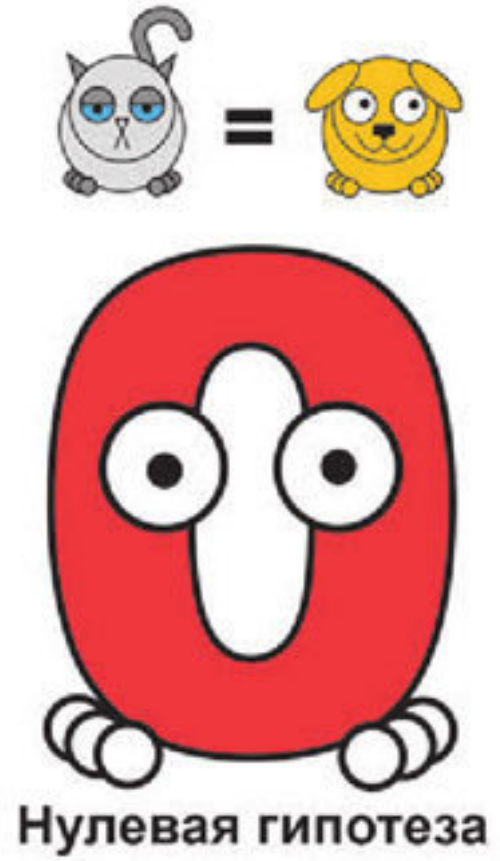



Истоки этого понятия – в самых основах теории вероятности и математической статистики, которые выходят за пределы нашей книги. С практической же точки зрения, знание о степенях свободы нужно при работе с таблицами критических значений и расчёте p-уровня значимости, о которых вы узнаете из следующей главы.

# Глава 4. Как понять, что песики отличаются от котиков или p-уровень значимости

Предположим, что вы вычислили t-критерий Стьюдента. Или U-критерий Манна-Уитни. Или какой-нибудь другой. Как же по нему понять, действительно ли песики и котики различаются по размеру? Чтобы это выяснить, статистики используют весьма нетривиальный подход.

Во-первых, они делают предположение, что котики и песики, как биологические, виды абсолютно не отличаются друг от друга. Это предположение называется *нулевой гипотезой.*



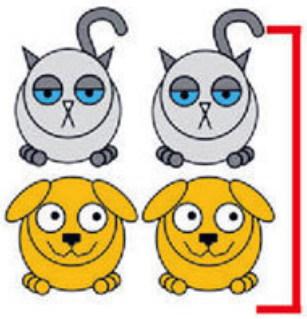

Следующим шагом они вычисляют вероятность того, что две случайно выбранные группы котиков и песиков дадут значение критерия большее или равное тому, которое мы получили (чаще всего без учета его знака). Эта вероятность называется *p-уровнем значимости.*

Если p-уровень значимости меньше 5 % (чаще записывается как 0,05), то нулевая гипотеза отвергается и принимается гипотеза о том, что котики и песики все-таки различаются. Такая гипотеза называется *альтернативной*.


Если же p-уровень значимости больше 0,05, то нулевая гипотеза не отвергается.

Однако то, что она не отвергается, еще не значит, что она верна. Это означает только то, что в данном опыте мы не обнаружили значимых различий.


В специальных статистических программах p-уровень значимости вычисляется автоматически, и нам достаточно просто найти его в соответствующей таблице. Однако, если у вас таких программ нет, то вам придется пользоваться *таблицами критических значений*.




| n-2 | Критическое значение для p=0,05 |
|---|---|
| 1 | 12,7 |
| 2 | 4,3 |
| 3 | 3,2 |

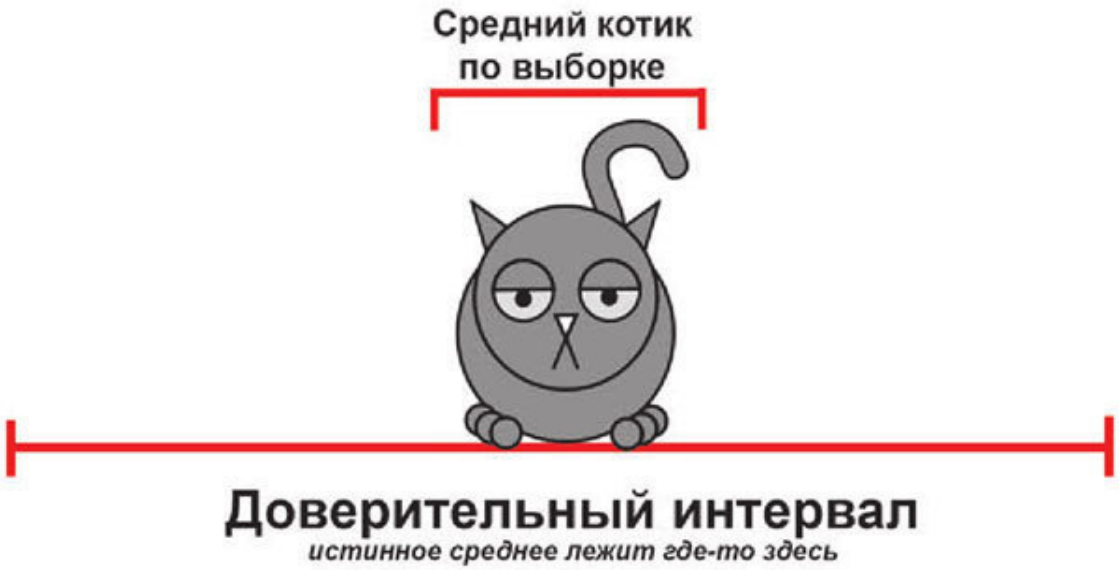
Работать с ними просто: найдите нужную строчку и посмотрите на значение критерия, которое там указано. Если то, что вы получили, превышает это значение, то котики и песики отличаются друг от друга. Правда, для этого правила есть исключения – это U Манна-Уитни и родственные ему критерии.

## НЕМАЛОВАЖНО

## ЗНАТЬ!

### Альтернативные подходы

Определение различий по p-уровню значимости в последнее время подвергается жесткой критике. Поэтому немаловажно знать о том, что существуют и альтернативные подходы, которые используются при определении значимости полученных результатов.




**Доверительные интервалы.** Как уже было сказано ранее, ученые чаще всего проводят свои исследования не на всех котиках, а на какой-то выборке. Соответственно, они не знают истинного среднего размера по всем котикам. Однако они могут прикинуть, в каком диапазоне он находится. Такой диапазон называется *доверительным интервалом*.

Рядом с доверительным интервалом всегда указывается вероятность. 95 %-ый доверительный интервал означает, что мы с точностью в 95 % можем утверждать, что истинный средний размер котиков находится в этом диапазоне.

Чем шире такой интервал, тем менее точной считается статистическая оценка. Что касается различий между песиками и котиками, то они имеют место быть, когда их доверительные интервалы не пересекаются.

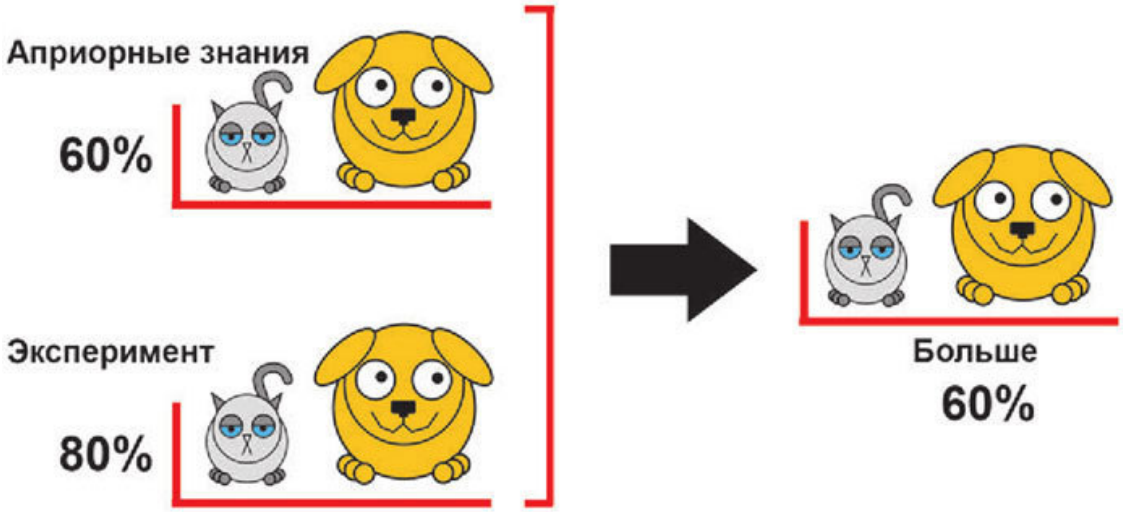
**Байесовская статистика.** Все вышеприведенные способы определения значимости не учитывают наши предыдущие *(априорные)* знания о том, каких размеров бывают котики и песики. Каждый раз, когда мы определяем p-уровень значимости или доверительный интервал, мы ведем себя так, как будто никогда не видели ни тех, ни других.

Но ведь это не так! Мы ведь достаточно четко представляем себе, как они выглядят! Нельзя просто так брать и отбрасывать предыдущий опыт!

Проблему сопоставления наших предыдущих знаний и новых данных пытается решить группа методов, основанных на теореме английского священника Томаса Байеса.

Не вдаваясь в математические подробности, опишем общую логику. Предположим, что из предыдущих опытов мы выяснили, что в 60 % случаев случайно выбранный песик больше случайно выбранного котика. Проведя собственный эксперимент, мы обнаружили, что это

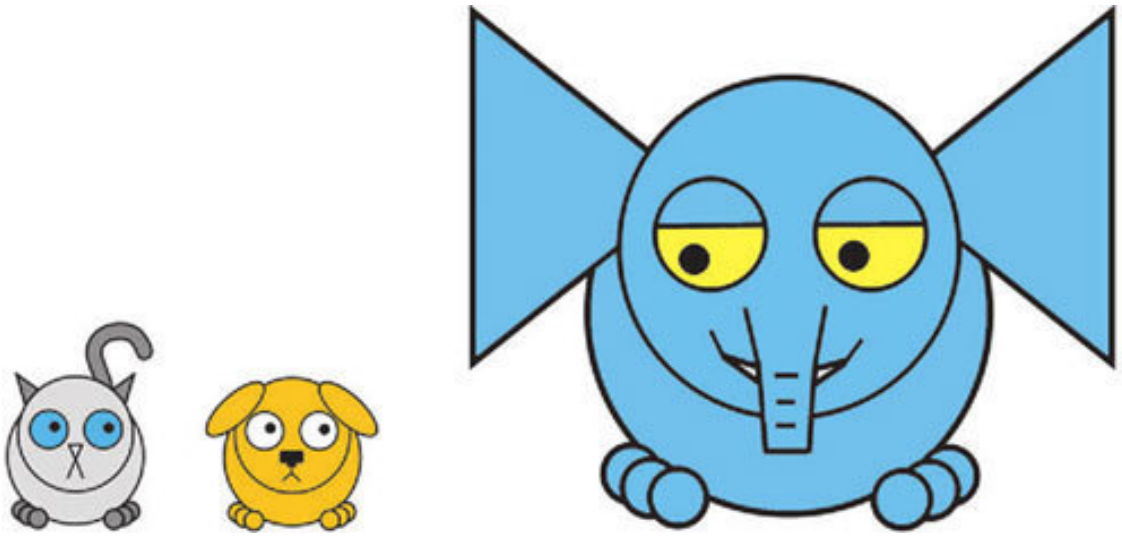
число гораздо выше – 80 %. Следует ли из этого, что нам нужно забыть наш предыдущий опыт и заменить старые данные новыми? Разумеется нет. Новый опыт только подправит предыдущую вероятность, и в следующий раз мы будем считать, что она несколько выше.



# Глава 5. Котики, песики, слоники или Основы дисперсионного анализа

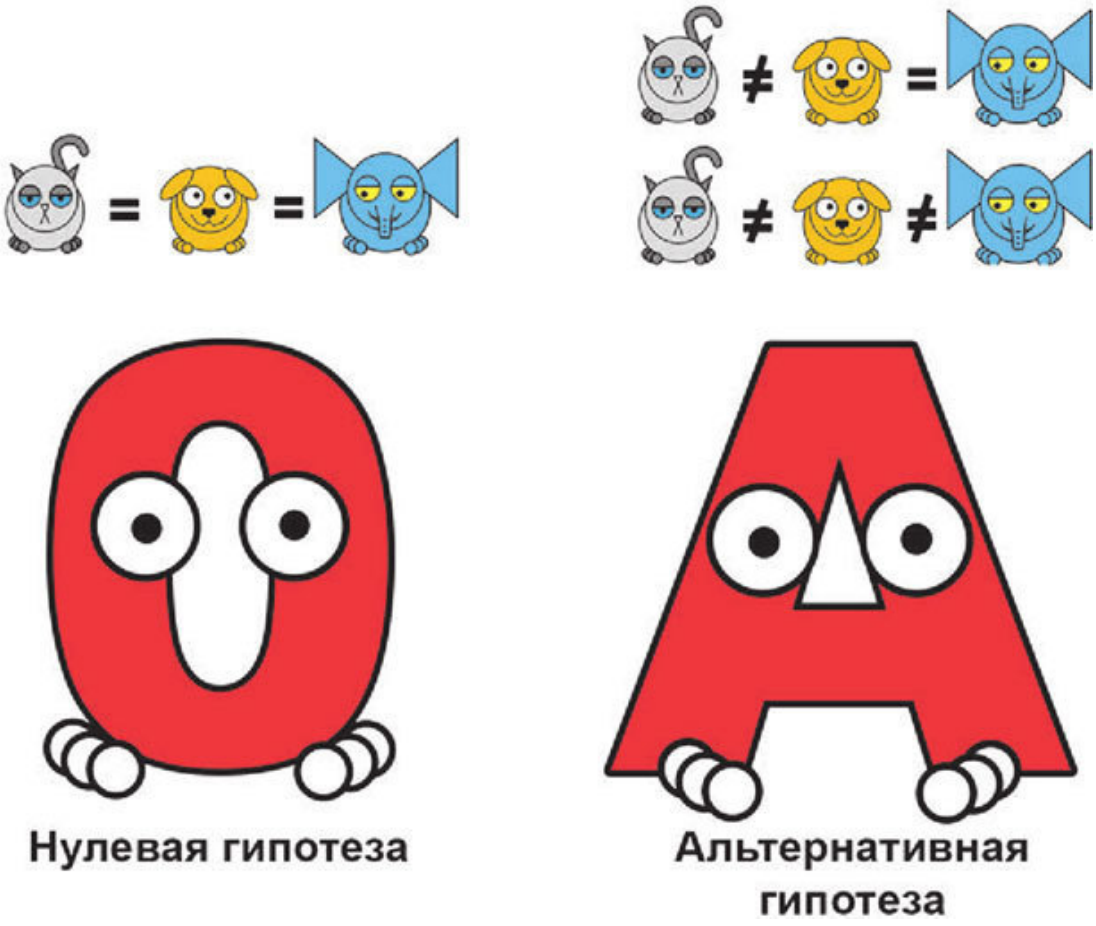
Из предыдущих разделов мы узнали, как определить, различаются ли между собой песики и котики по размеру. И если мы отвечаем на этот вопрос положительно, то мы, по сути, устанавливаем связь между двумя признаками: размером и биологическим видом, к которому принадлежат эти животные.

Однако, согласитесь, что мир не ограничивается только лишь котиками или песиками. Ведь существует еще и множество других животных. Например, слоники.



И, если мы добавим их к нашему небольшому зоопарку, мы не сможем применить обычное попарное сравнение (например, по t-критерию Стьюдента или U-критерию Манна-Уитни) для определения того, связан ли размер с биологическим видом. В этих случаях необходимо использовать другие методы. Например, *дисперсионный анализ*.

Дисперсионный анализ хорош тем, что позволяет сравнивать между собой любое количество групп (две, три, четыре и т. д.) Его нулевая гипотеза состоит в том, что животные абсолютно не различаются между собой по размеру. Альтернативная гипотеза – хотя бы один вид значимо отличается от остальных.

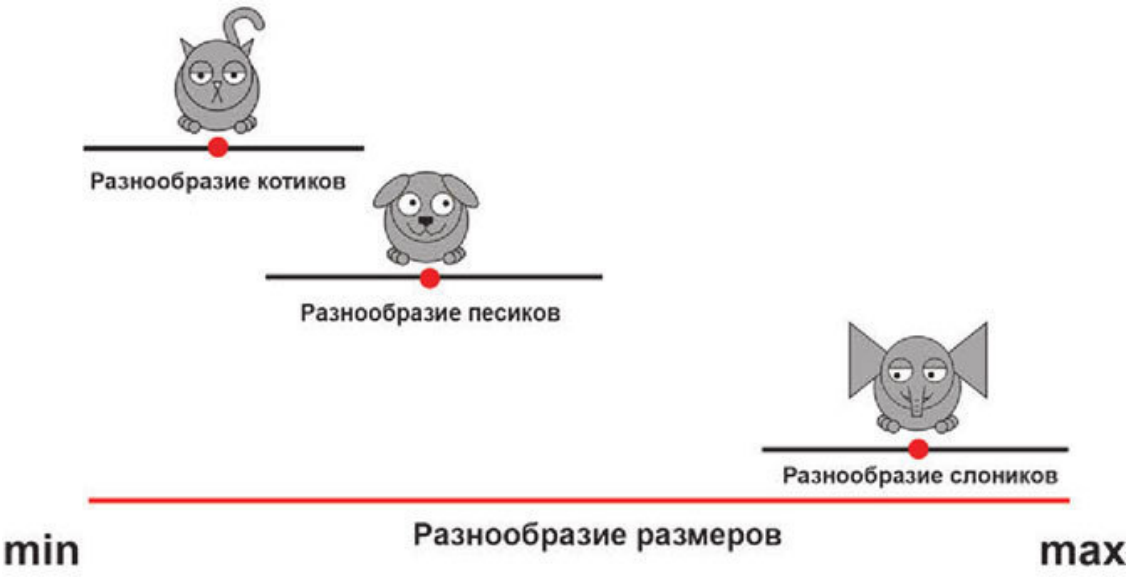



Теперь посмотрим, как это работает.

Во-первых, давайте объединим котиков, песиков и слоников вместе и отметим их общее разнообразие. Мы можем заметить, что размеры их типичных представителей могут существенно различаться. Например, средний слоник намного больше среднего котика.


Теперь предположим, что мы убрали отсюда всех слоников. Как вы можете заметить, разнообразие размеров сильно уменьшилось, поскольку слоники вносили в него существенный вклад. И чем сильнее типичные слоники отличались от остальных, тем больше был этот вклад.


Однако отметим, что котики, песики и слоники по отдельности также бывают весьма различными в зависимости от возраста, генов и режима питания. Теоретически мы можем встретить как очень большого котика, так и весьма маленького слоника.




Таким образом, разнообразие размеров складывается как из принадлежности животного к тому или иному виду, так и из абсолютно «левых» факторов. И наша задача – сравнить между собой их вклады.

Как мы помним, одной из основных мер, определяющих разнообразие, является дисперсия. И дисперсионный анализ работает именно с ней. Он выделяет ту часть дисперсии, которая обусловлена фактором вида (*межгрупповую дисперсию*), и ту, которая определяется прочими факторами (*внутригрупповую дисперсию*), а затем сравнивает их по F-критерию Фишера, с которым мы встречались раньше. И чем больше будет значение этого критерия, тем сильнее фактор вида влияет на размер животных.

# Конец ознакомительного фрагмента.