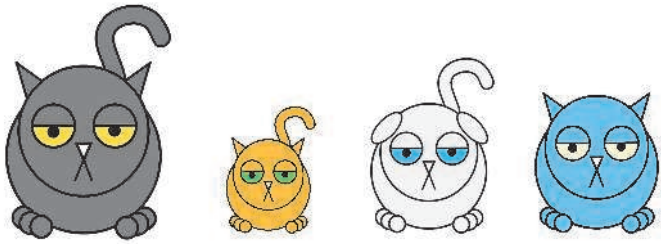


СТАТИСТИКА
И КОТИКИ

ИЗДАТЕЛЬСТВО АСТ
МОСКВА

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.

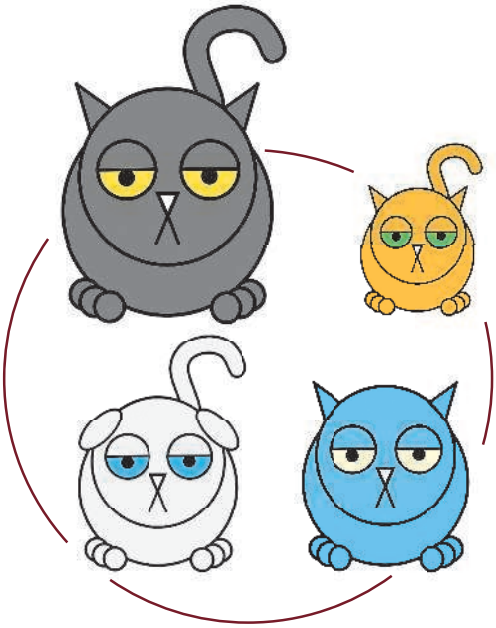
Когда мы изучаем данные, мы осознаем, что задача — найти соломинку в стоге иголок. И понять, сколько ещё стогов и соломы найдем дальше. Статистика в бизнесе помогает нам экономить деньги и открывать новые рынки. Экономия питает амбиции и потихоньку делает жизнь людей чуточку лучше.

Респект читателям. Респект автору.

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

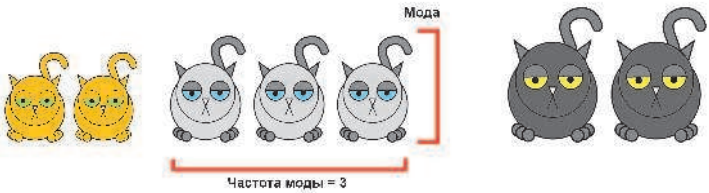
ГЛАВА 1.
КАК ВЫГЛЯДЯТ КОТИКИ
ИЛИ ОСНОВЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

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?

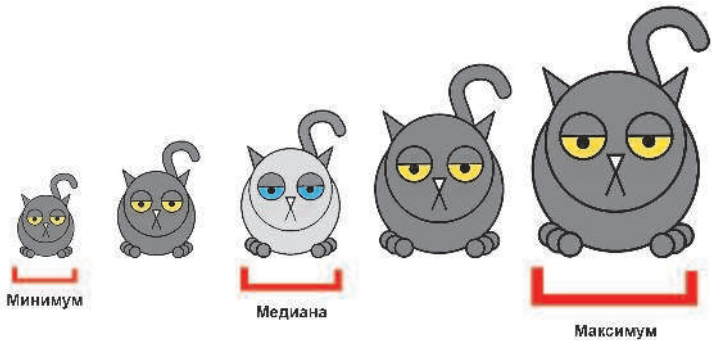


Для простоты мы возьмем такое котиковое свойство, как размер.

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

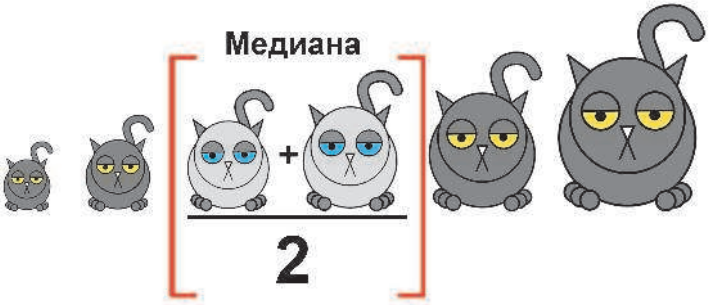


Второй способ: мы можем упорядочить всех котиков от самого маленького до самого крупного, а затем посмотреть на середину этого ряда. Как правило, там находится котик, который обладает самым типичным размером. И этот размер называется *медианой*.



Если же посередине находятся сразу два котика (что бывает, когда их четное количество), то,

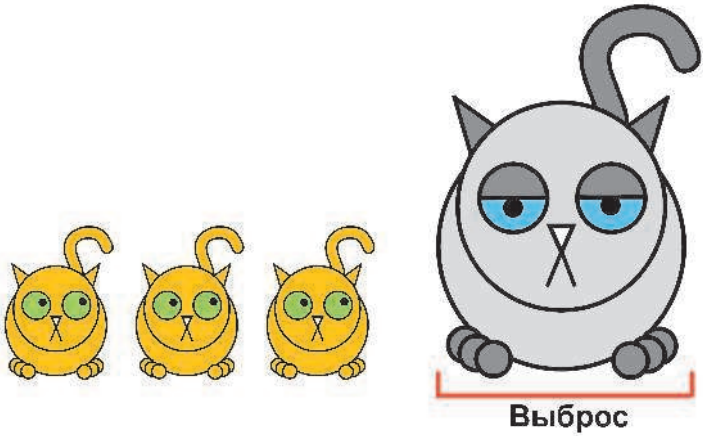
чтобы найти медиану, нужно сложить их размеры и поделить это число пополам.



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.

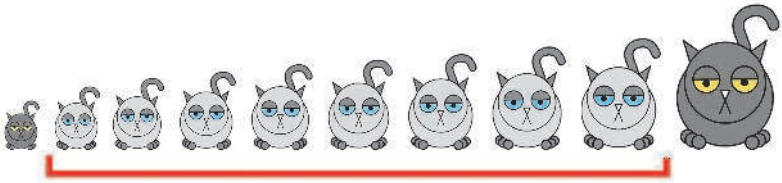
Предположим, что среди наших котиков есть один уникум размером со слона. Его присутствие может существенным образом сдвинуть среднее значение в большую сторону, и оно перестанет отражать типичный котиковый размер.



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*.



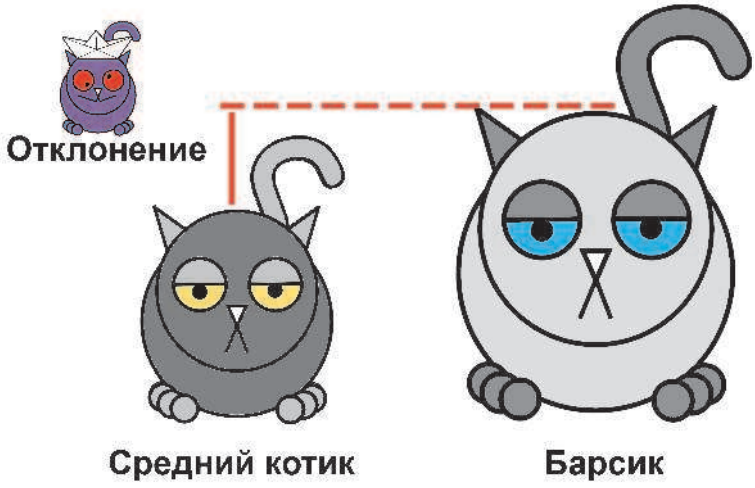
Котики для усеченного среднего

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*.



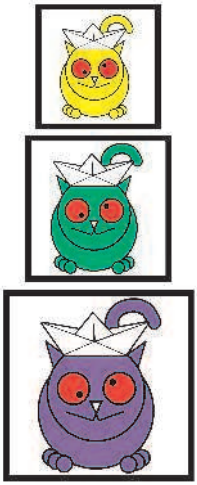
И совершенно очевидно, что чем сильнее Барсик будет отличаться от среднего котика, тем больше будет это самое отклонение.

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 средне-

го), 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.



$/\ 3$

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 санти-

метрах. 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 называе-