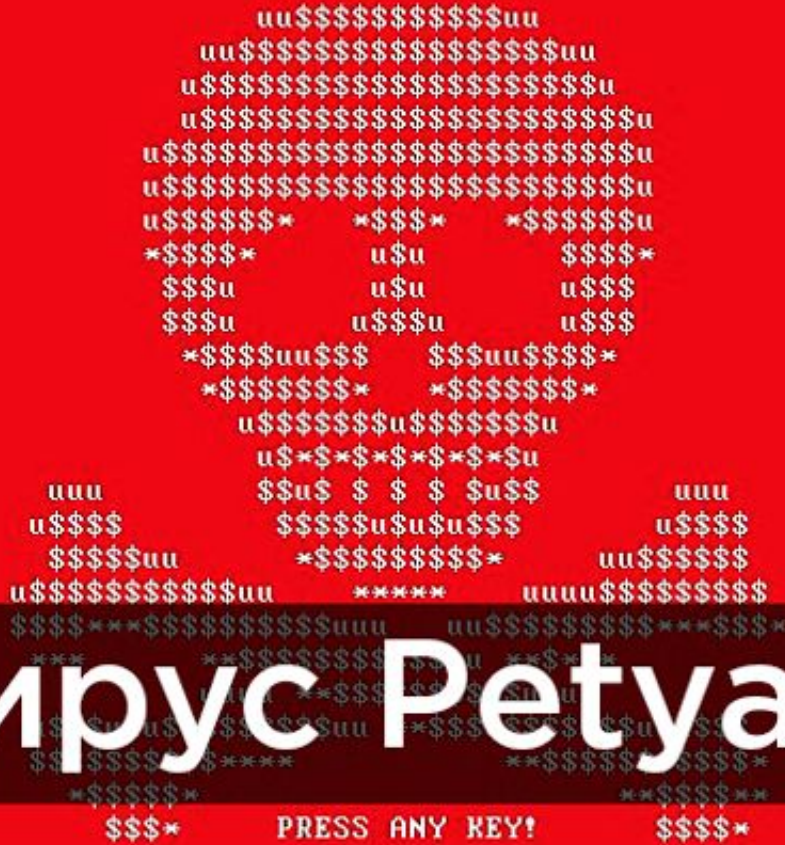




**Магия цифр:
когда стоит
сомневаться в
статистике?**

КЕЙС 1



Вирус Рetya.A

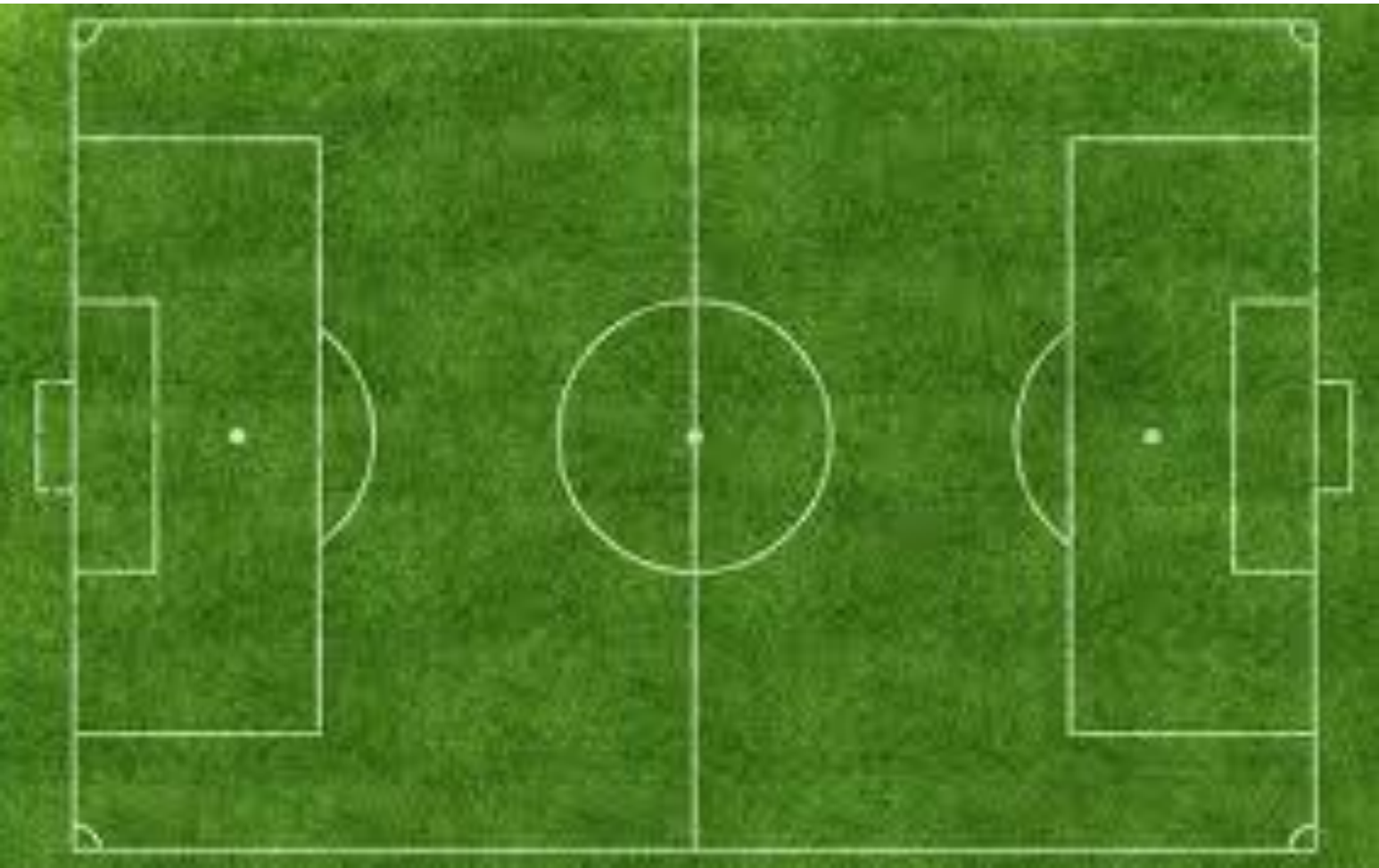
Статья о новом вирусе, появившемся в Интернете и замедляющем работу браузеров, а также отправку электронных сообщений во всем мире. Сколько компьютеров были им поражены? Эксперты, слова которых приводятся в статье, утверждают, что зараженными оказались 39 тыс. компьютеров, что повлияло еще на сотни тысяч других систем.



Но откуда они взяли это число? Разве его так просто определить? Неужели были проверены все имеющиеся компьютеры, чтобы установить, не заражены ли они? Тот факт, что статья была написана менее чем через сутки после атаки, наталкивает на мысль, то приводимая цифра — это предположение. Тогда почему же не сказать не 39, а 40 тыс.?



КЕЙС 2



Статья об уровне освоения земель и торговле земельными участками в стране. Учитывая количество зданий, которые, скорее всего, будут построены в данной местности, это очень важный вопрос. Приводятся статистические данные, касающиеся акров пахотной земли, которая ежегодно теряется из-за застройки, и все это превращается в квадратные мили. В качестве дополнительной иллюстрации того, как много земли теряется, эта площадь представлена также в соответствующем количестве футбольных полей. В этом конкретном случае эксперты отмечают, что в центре штата Огайо в год теряется 150 тыс. акров земли, что составляет 234 квадратные мили или 115 385 футбольных полей (включая зону защиты). Но как были получены эти цифры и насколько они точны? И неужели проще представить количество потерянной земли с помощью футбольных полей?



Но как были получены эти цифры и насколько они точны? И неужели проще представить количество потерянной земли с помощью футбольных полей?



КЕЙС

2



Статья о возрастающем количестве аварий на мотоциклах

с 1997 года это число увеличилось более чем на 50%, и никто не может объяснить причину.

Статистика приводит следующий интересный факт: в 1997 году погибло **2 116** мотоциклистов, в 2001 году погибших было **3 181**, как показывают данные Национальной администрации безопасности дорожного движения (НАБДД).

В статье рассматриваются многие возможные причины увеличения количества смертей, в том числе и тот факт, что **сегодня мотоциклисты стали старше** (средний возраст погибших мотоциклистов увеличился с 29,3 года в 1990 году до 36,3 лет в 2001 году). Еще одно из возможных объяснений — **увеличение размеров мотоциклов**. Размер двигателя среднего мотоцикла вырос почти на 25% — с 769 см³ в 1990 году до 959 см³ в 2001 году.

Дополнительный вариант — это тот факт, что некоторые штаты США делают **по послабления в законе относительно ношения шлема**. Специалисты, слова которых цитируются в статье, говорят, что необходимо более обширное изучение причин, но оно, вероятно, так и не будет проведено, потому что затраты на него составят от 2 до 3 млн. долл. При этом в статье **ничего не говорится о количестве людей, которые ездят на мотоциклах**, в 2001 и 1997 году. Естественно, что большее число людей на дорогах означает больше аварий, даже если все остальные факторы остаются прежними. Однако в статье приведен еще и график, отображающий количество смертей мотоциклистов на 100 млн. миль, которые были преодолены в США с 1997 по 2001 год. Касается ли это увеличения количества людей на дорогах? Здесь же приводится и столбиковая диаграмма, в которой сравниваются число смертей мотоциклистов с количеством людей, погибших в авариях на других видах транспорта. Из этой диаграммы видно, что **уровень смертности мотоциклистов составляет 33,4 смерти на 100 млн.** преодоленных миль по сравнению с показателем всего лишь в 1,7 на то же количество миль, преодоленных на машине.

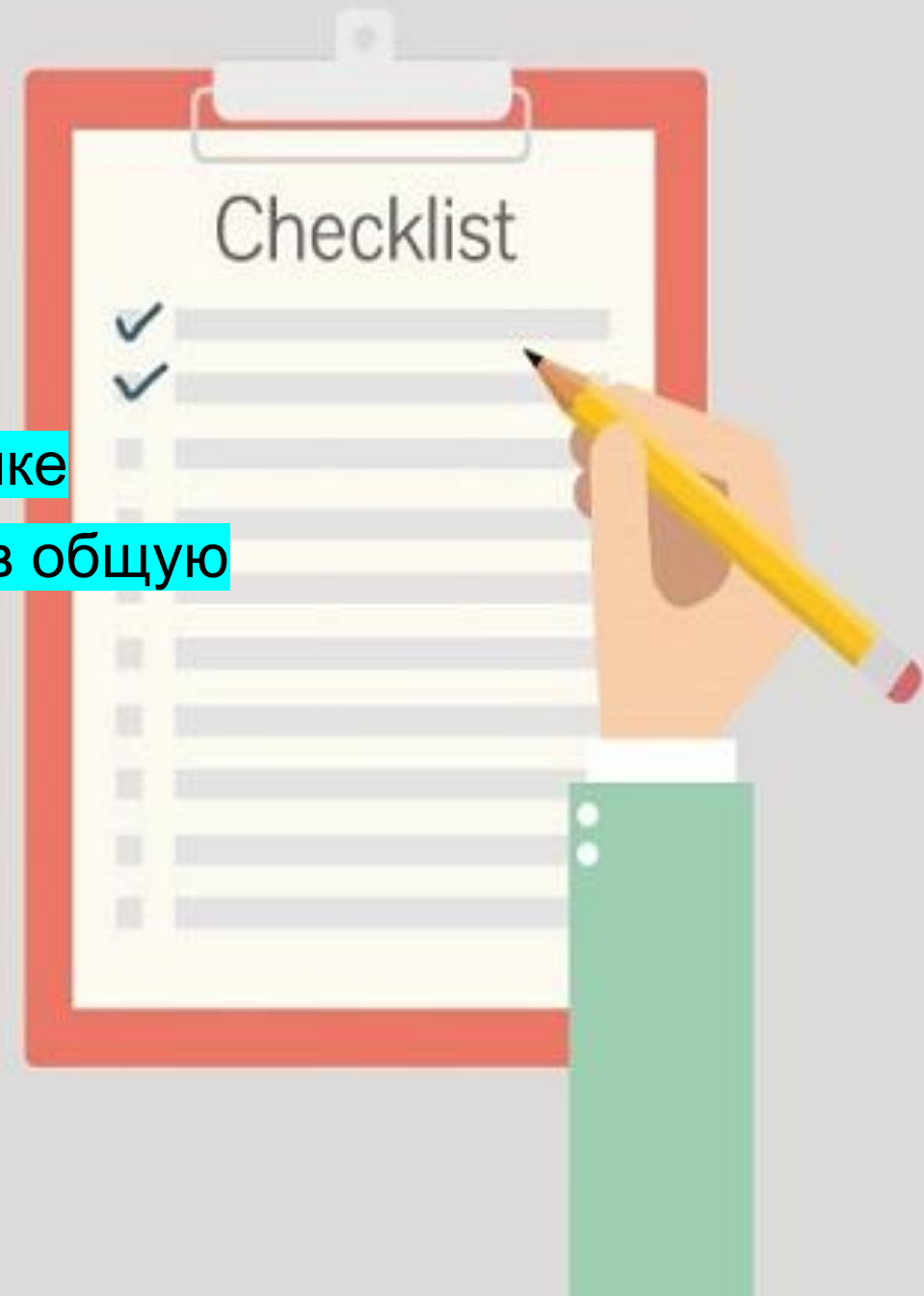


В этой статье множество цифр и самых разных статистических данных, но что все это значит? Объем и разнообразие статистических данных очень скоро может сбить с толку.



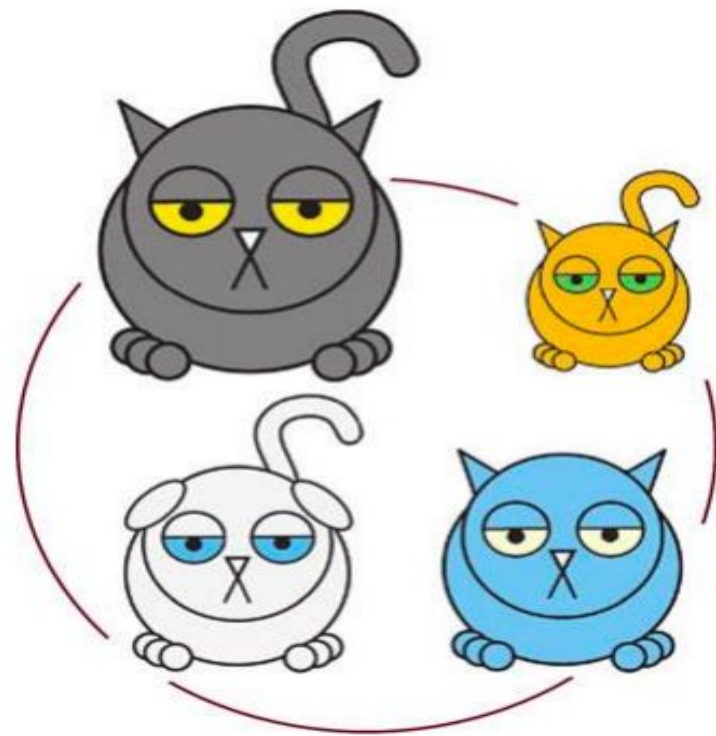
Статья о последних исследованиях в сфере страхования врачей на случай судебного преследования. Итак, насколько серьезна данная проблема? В статье сказано, что один из пяти врачей в штате Джорджия отказался от проведения опасных процедур (например, принятие родов) из-за постоянно растущих страховых ставок от судебного преследования в этом штате. Это описывается как “национальная эпидемия” и “кризис здравоохранения” в стране. Приводятся некоторые сведения об исследовании проблемы; в статье утверждается, что из 2200 врачей штата Джорджия, принявших участие в опросе, 2800 (которые, как говорится, составляют 18% от общего числа участников) скорее всего, откажутся от проведения рискованных процедур.

- Непонятна методология
- Используются метафоры
- Не соответствует математической логике
- Цифр много, но они не складываются в общую картинку



Статистика на котиках

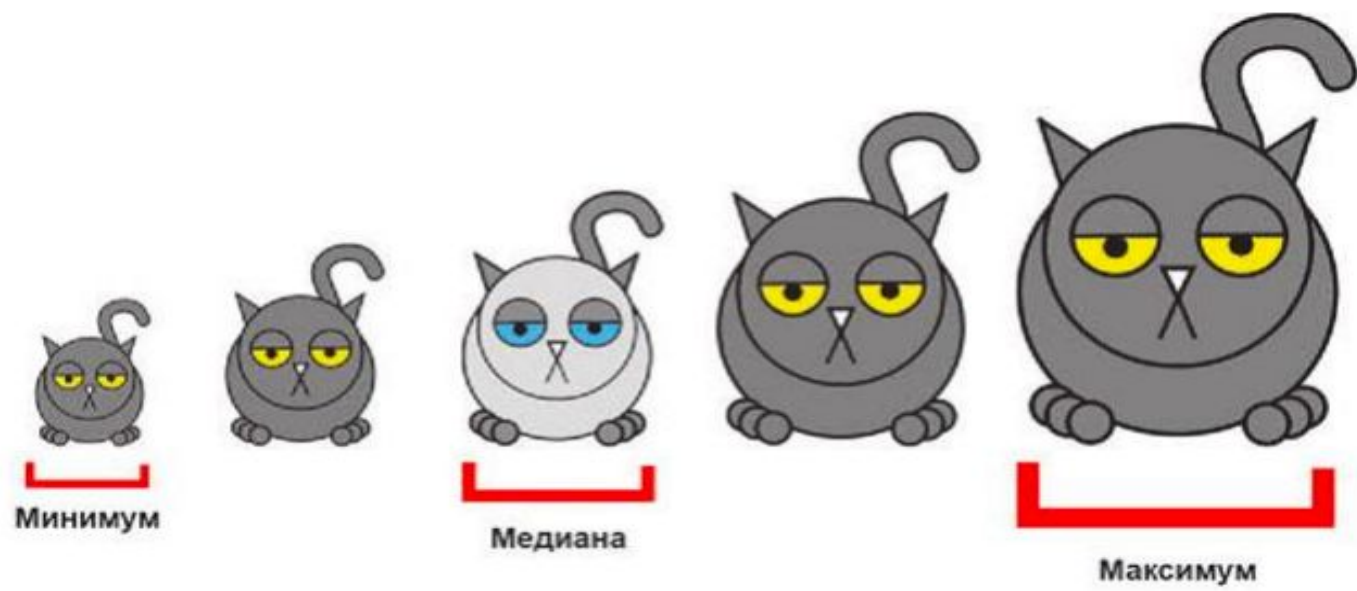
Котики бывают разные. Есть большие котики, а есть маленькие. Есть котики с длинными хвостами, а есть и вовсе без хвостов. Есть котики с висячими ушками, а есть котики с короткими лапками. Как же нам понять, как выглядит типичный котик?



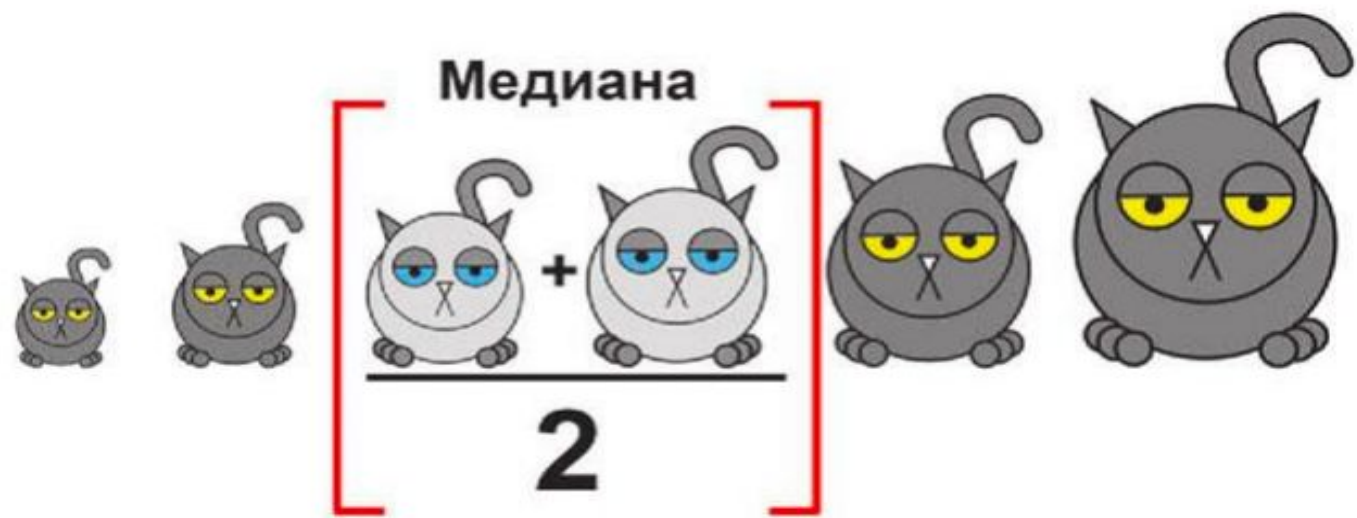
Для простоты мы возьмем такое котиковое свойство, как размер.

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

Мода — значение во множестве наблюдений,
которое встречается наиболее часто.
(**Мода** = типичность)



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.



Медиана — варианта, которая находится в середине вариационного ряда.

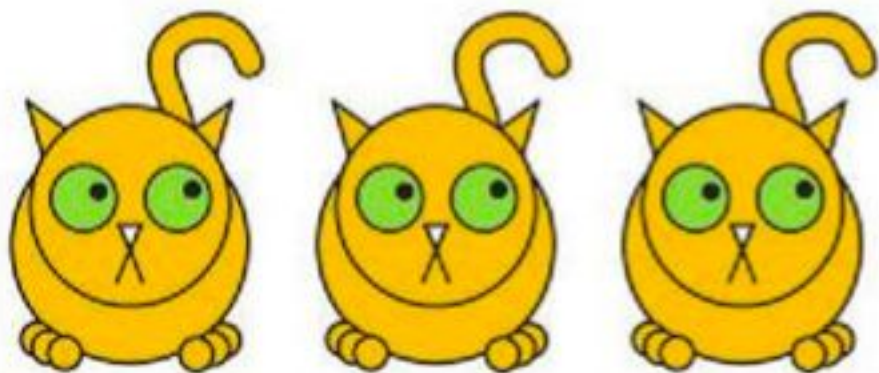
Медиана делит ряд пополам



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.

Среднее значение — некоторое число, заключённое между наименьшим и наибольшим из их значений



Выброс



Средний котик

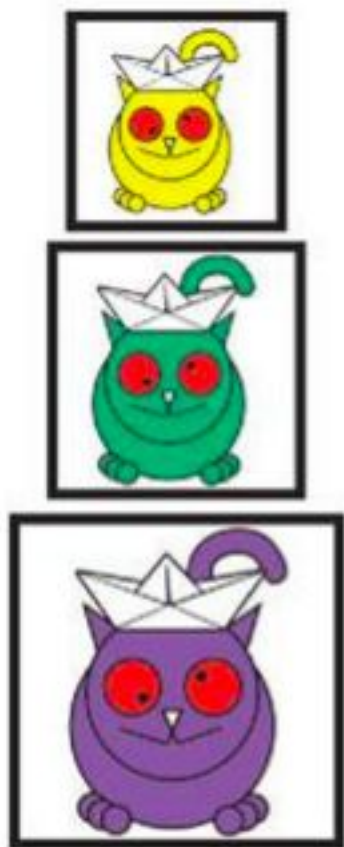


Барсик

**Отклонение – разность размера
конкретного котика и типичного Барсика**

**Как понять, какое из отклонение
нормально?**

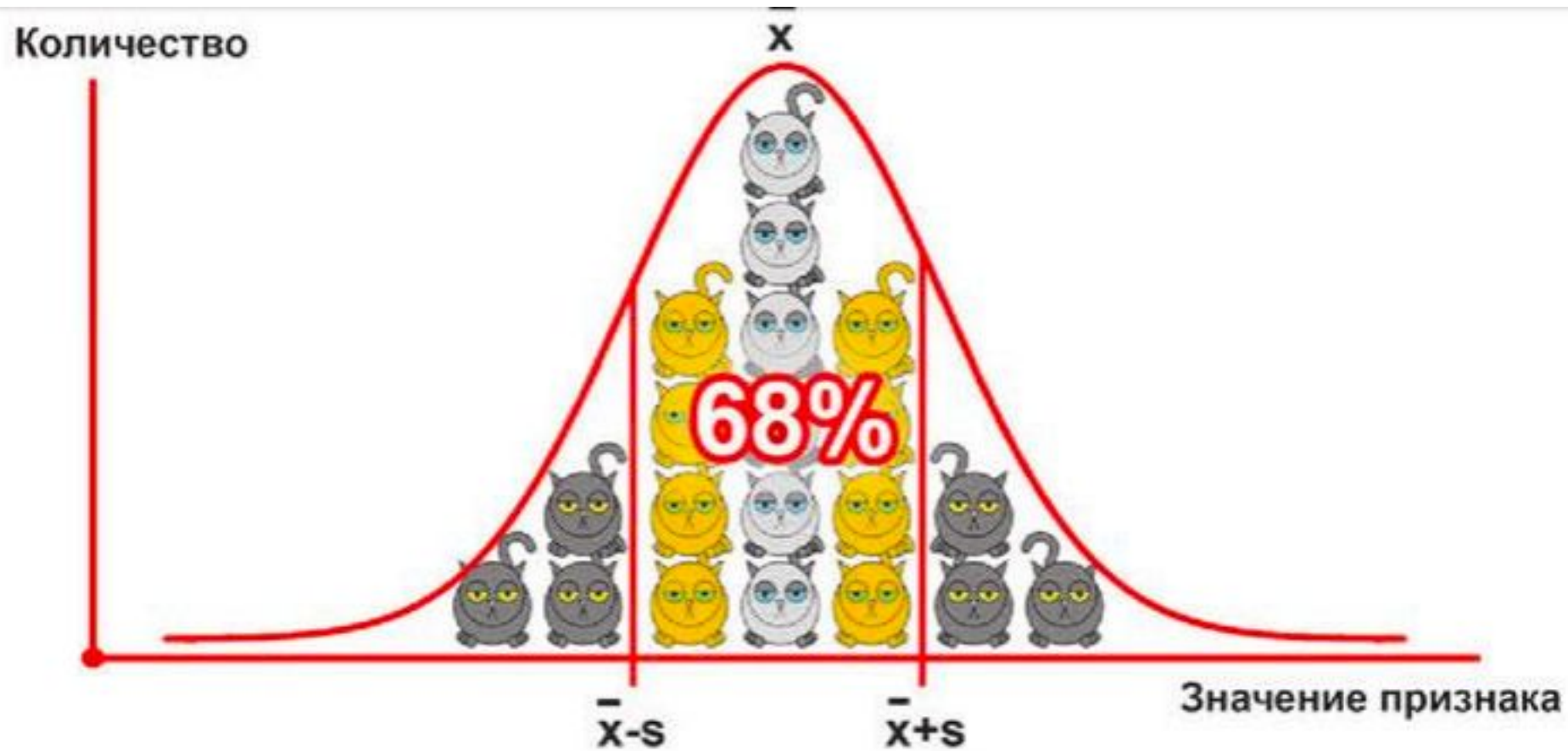
**Дисперсия – среднее от квадратов
отклонений**



/ 3



Дисперсия D



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.

Генеральная совокупность (от лат. generis — общий, родовой) — **совокупность** всех объектов (единиц), относительно которых предполагается делать выводы при изучении конкретной задачи.

Выборка — множество случаев (испытуемых, объектов, событий, образцов),

с помощью определённой процедуры выбранных из генеральной совокупности для участия в исследовании.

Вадим Палько

ВЦИОМ: почти 40 процентов россиян хотят жить в эпоху Брежнева

НОВОСТИ

Всероссийский центр изучения общественного мнения (ВЦИОМ) выяснил, что 37 процентов россиян хотели бы жить в годы правления Леонида Брежнева (с 1964 по 1982 годы), сообщает РИА «Новости»

18 АПРЕЛЬ 2019 22:39

Большее всего опрошенных хотят жить в современной России — таких 40 процентов, всего на три процента больше тех, кто выбрал эпоху Брежнева. На третьем месте по популярности оказался период правления Иосифа Сталина. В то время хотели бы жить всего пять процентов россиян. Четыре процента выбрали бы Россию на рубеже 19 и 20 века, а еще три процента — годы президентства Бориса Ельцина. Остальные опрошенные затруднились



НОВОСТИ



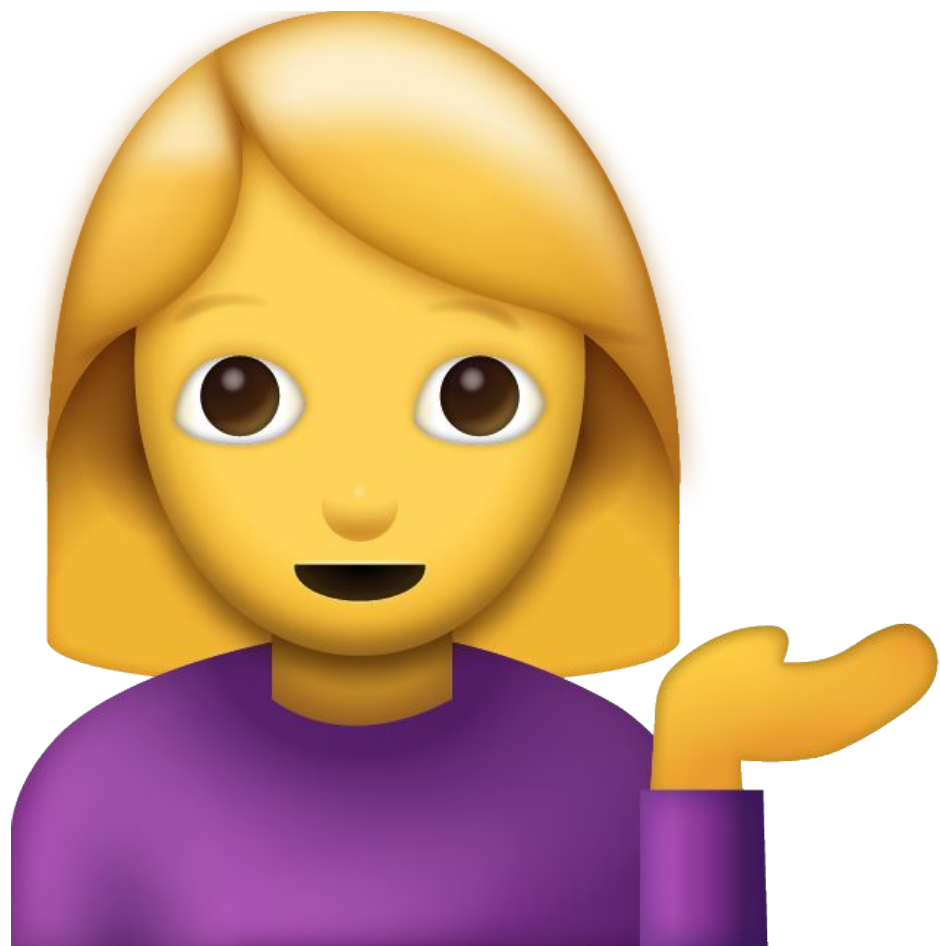
Об опросе:

Он состоялся 18 апреля методом телефонного интервью. В нем приняли участие 1,6 тысячи совершеннолетних россиян.

**А как это вы опросили 1,6 тысячи, а
сделали выводы за всю Россию?**

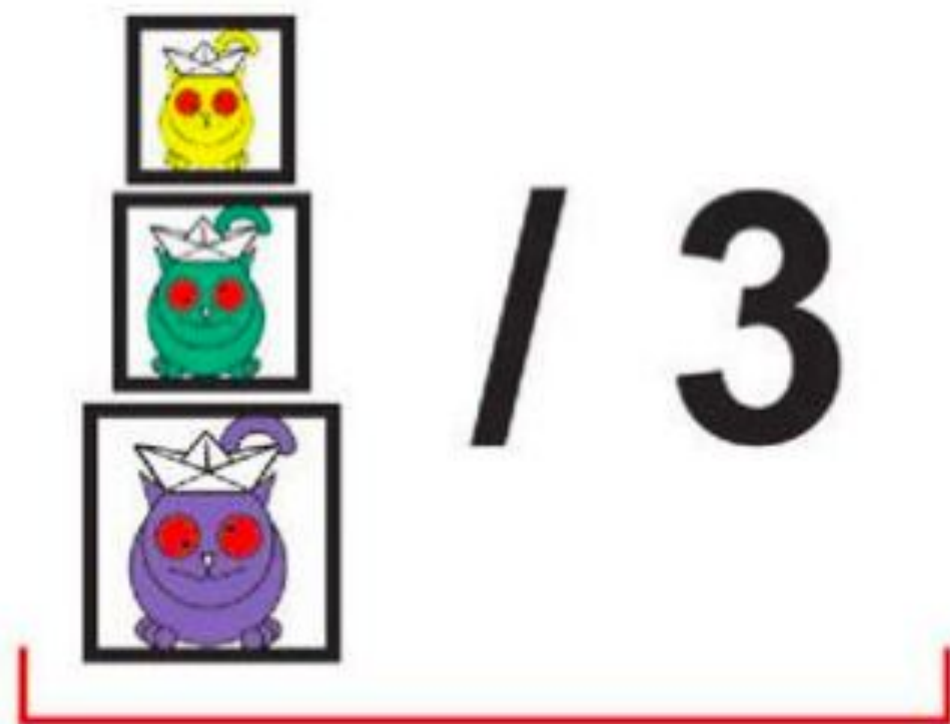


Выборка будет репрезентативной при обследовании большой группы людей, если внутри этой группы есть представители разных подгрупп, только так можно сделать верные выводы.

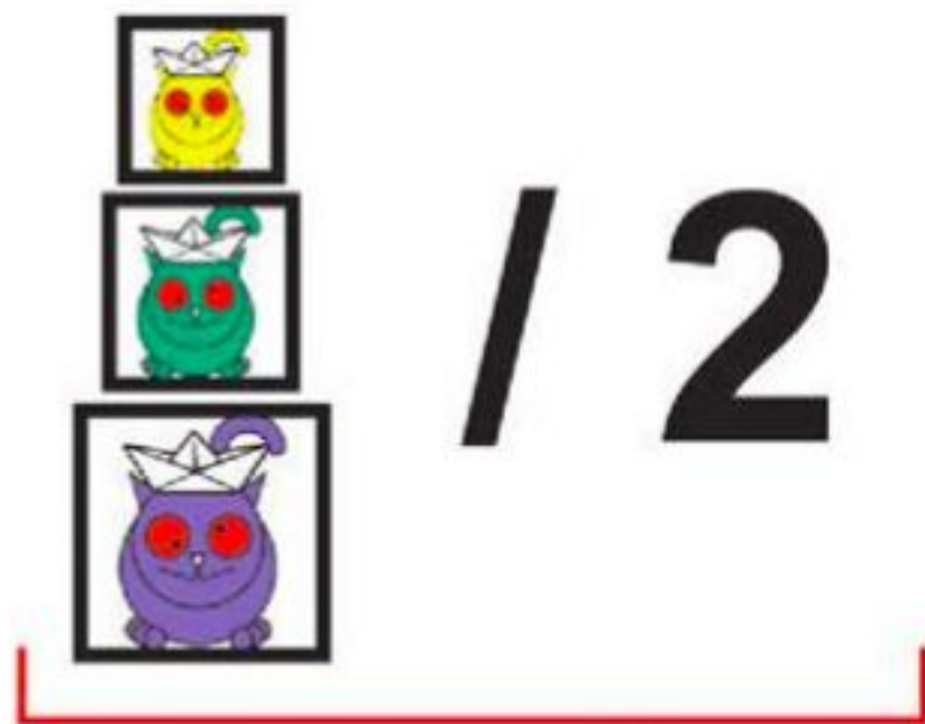




Очень важно, чтобы выборка была максимально похожа на генеральную совокупность. Степень такой похожести называется *репрезентативностью* .



Дисперсия генеральной
совокупности



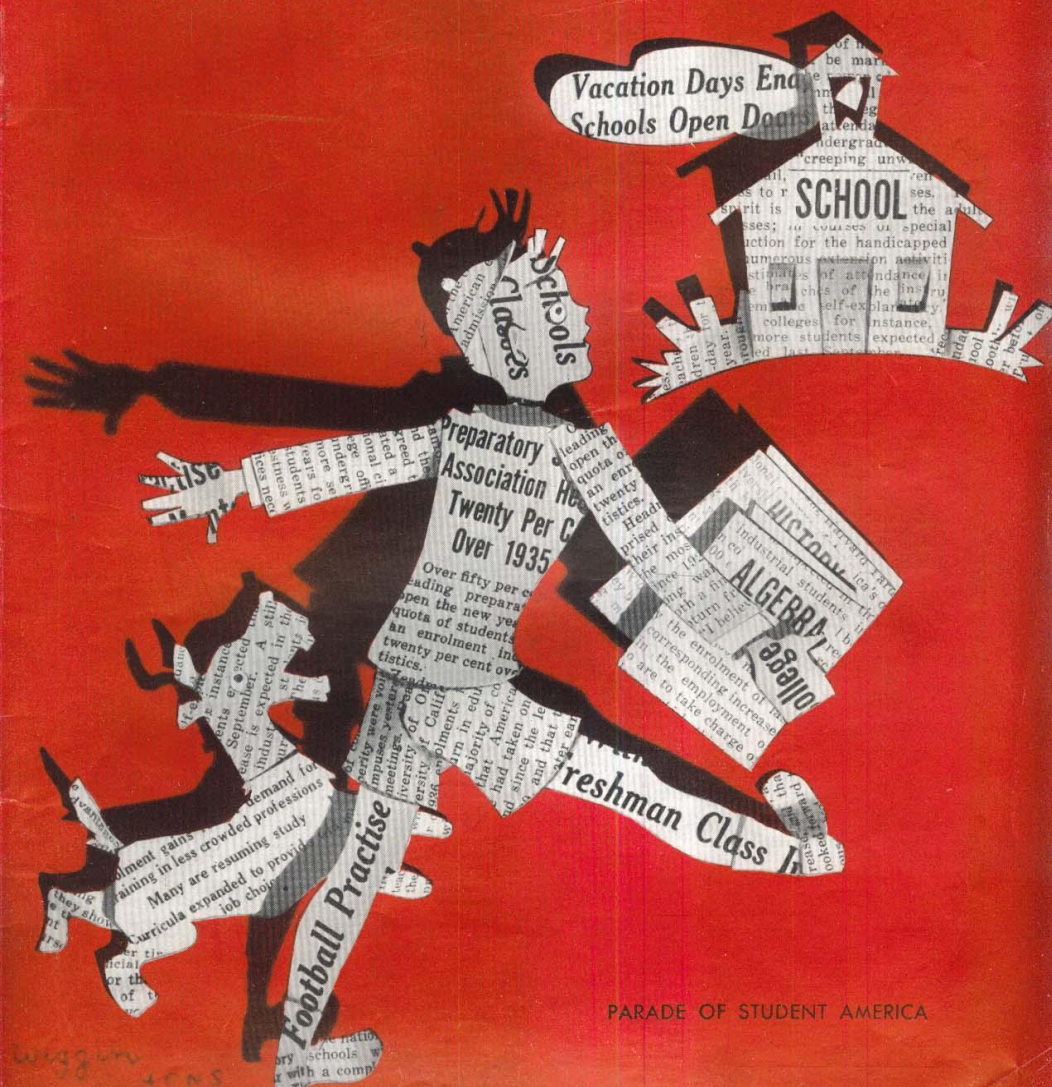
Дисперсия выборки

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

The Literary Digest

SEPTEMBER 19, 1936

TEN CENTS



США, выборы президента 1936 года

Журнал «Литрери Дайджест», успешно прогнозировавший события нескольких предшествующих выборов, ошибся в своих предсказаниях, разослав 10 млн пробных бюллетеней своим подписчикам, а также людям, выбранным по телефонным книгам всей страны и людям из регистрационных списков автомобилей. В 25 % вернувшихся бюллетеней (почти 2,5 миллиона) голоса были распределены следующим образом:

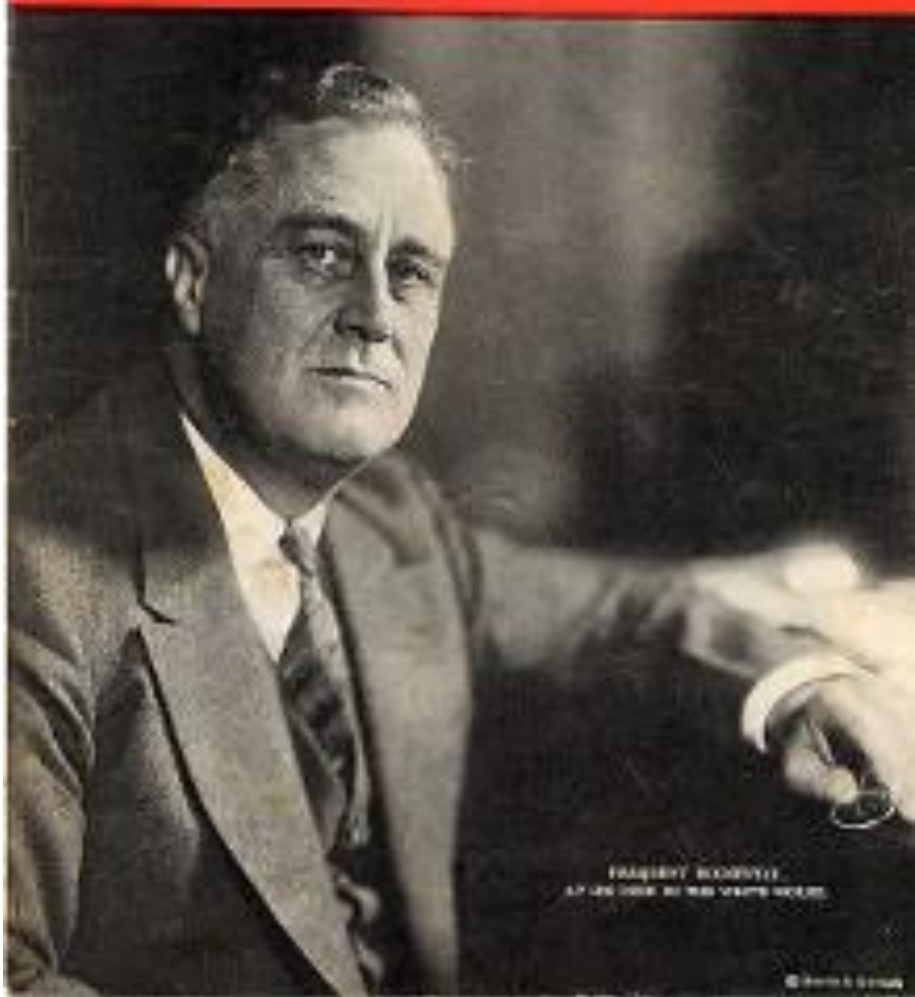
57 % голосов - кандидату-республиканцу Альфу Лэндону
 40 % выбрали действующего в то время президента-демократа Франклина Рузвельта

The Literary Digest

1933

July 1, 1933

Five Dollars



FRANKLIN ROOSEVELT
AT THE GROUND BY THE GREAT MOUNTAIN

© 1933 by The Literary Digest

победил Рузвельт, набрав более 60 % голосов.

Ошибка «Литрери Дайджест»: желая увеличить репрезентативность выборки, — так как им было известно, что большинство их подписчиков считают себя республиканцами, — они расширили выборку за счёт людей, выбранных из телефонных книг и регистрационных списков. Однако они не учли современных им реалий и в действительности набрали ещё больше республиканцев: во время Великой Депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и высшего класса (то есть большинство республиканцев, а не демократов).

Hours TO GET TO VACATIONLAND

Before you start, telephone ahead for reservations. After you arrive, telephone home—to tell friends about your trip—to take care of forgotten errands.

And later, call them often—to get the latest news, invite them for the week-end, tell them what you're doing.

It's quick, convenient, inexpensive. Rates for station-to-station and person-to-person calls to most points are lower after 7 every evening and all day Sunday. Keep in touch with family and friends . . . by telephone.

Minutes TO REACH HOME

во время Великой Депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и высшего класса (то есть большинство республиканцев, а не демократов).

В городе пять школ. В таблице приведен средний балл, полученный выпускниками каждой из этих школ за экзамен по математике. Найдите средний балл выпускного экзамена по математике по всему городу?

| | | | | | |
|-------------------------------|-----------|-----------|-----------|-----------|-----------|
| Номер школы | 1 | 2 | 3 | 4 | 5 |
| Количество выпускников | 60 | 70 | 30 | 50 | 70 |
| Средний балл | 60 | 54 | 68 | 72 | 54 |

Решение:

Чтобы найти средний балл выпускного экзамена по математике по всему городу, нужно сложить баллы всех выпускников и поделить на общее количество выпускников.

1. Общее количество выпускников равно

$$60+70+30+50+70=280$$

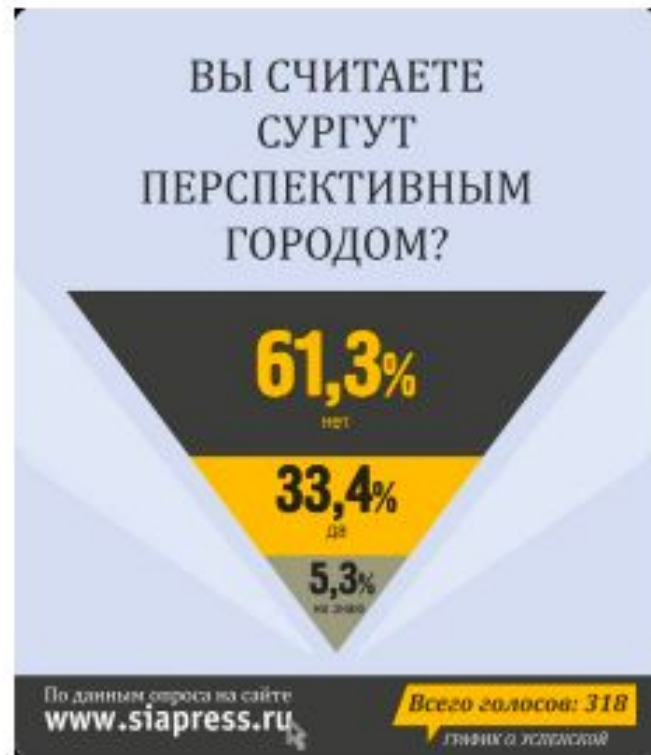
2. Если умножить количество учеников в школе на средний балл по школе, то получится сумма баллов в этой школе, а если сложить все такие произведения, то сумма всех баллов по городу равна

$$60 \cdot 60 + 70 \cdot 54 + 30 \cdot 68 + 50 \cdot 72 + 70 \cdot 54 = 3600 + 3780 + 2040 + 3600 + 3780 = 16800$$

3. Средний балл по городу равен $16800:280=60$

Ответ: 60.

Сургут без перспектив ? // ИНФОГРАФИКА



Большая часть аудитории siapress.ru считают, что Сургут неперспективный город. Такую позицию заняли 61,3 процента опрошенных. В свою очередь противоположную позицию заняли 33,4 процента респондентов и 5,3 процента не смогли определиться с ответом.

Отметим, что сегодня блогер **Илья Варламов** дал свою характеристику Сургуту, подробнее читайте по [ссылке](#).

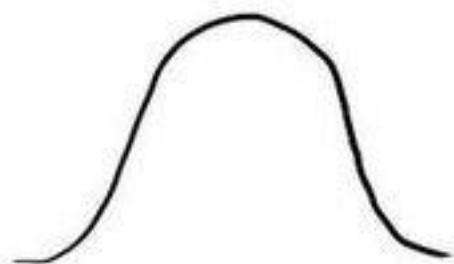
Почему большинство телефонных опросов, касающихся доходов населения, можно считать нерепрезентативными?

Ошибка выборки

Систематическая ошибка зависит от различных факторов, оказывающих постоянное воздействие на исследование и смещающих результаты исследования в определенную сторону.

Пример:

- Использование любых вероятностных выборок занижает долю людей с высоким доходом, ведущих активный образ жизни. Происходит это в силу того, что таких людей гораздо сложнее застать в каком-либо определенном месте (например, дома).



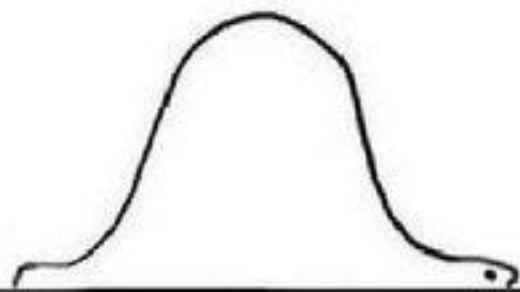
Нормальное распределение



He Нормальное распределение



Паранормальное распределение



Удав, который проглотил
нормальное распределение

ВЛАДИМИР САВЕЛЬЕВ



СТАТИСТИКА И КОТИКИ

ЛОНГ-ЛИСТ ПРЕМИИ
«ПРОСВЕТИТЕЛЬ»



БЕСТСЕЛЛЕР РУНЕТА



YAKABOOLJA

Легкий и увлекательный способ
познакомиться со статистикой



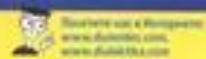
Статистика

ДЛЯ
"ЧАЙНИКОВ"™



Дебора Рамси
Директор Центра по изучению
математики и статистики при
Государственном университете Огайо

Объяснение человеческим языком



FAQ

Репрезентативность в опросах

Можно ли доверять результатам опросов общественного мнения, что такое выборка и генеральная совокупность и насколько массовые опросы политизированы

25 FEBRUARY 2016



Социология

409 публикаций

- РОССИЯ
- ОПРОС ОБЩЕСТВЕННОГО МНЕНИЯ
- ТЕЛЕФОННЫЕ ОПРОСЫ
- МАССОВЫЕ ОПРОСЫ
- ТЕОРЕМА БЕРНУЛЛИ
- ГРИГОРИЙ ЮДИН



FAQ

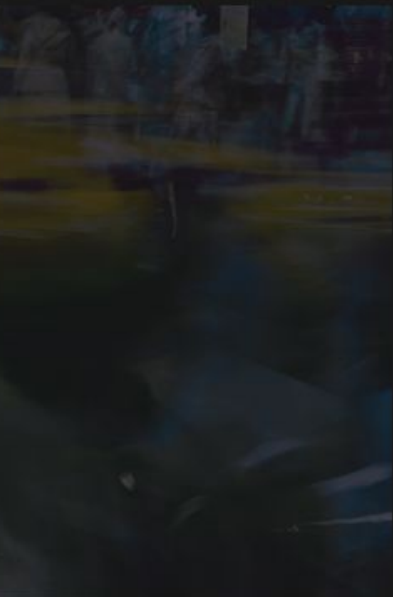
1157 публикаций

Читайте далее:

[Репрезентативность в опросах](#)



Лекции



11 минут

2/7 | Читает Григорий Юдин

11 минут

Как работают соцопросы?

Существует ли общественное мнение, не врут ли опросы и как Гэллапу удалось предсказать победителя выборов

Конспект

Скачать аудио

Поделиться:



11 минут

Мы — эгоисты

Зачем дарить подарки и срывать с нашей потребностью в обществ

Конспект

Снача

