Гипотезы, переменные, валидность, данные

Введение в статистику, лекция 1.

Как начинается исследование?

- Сначала вы наблюдаете то, что вы хотите понять.
- Потом вы придумываете некоторые объяснения того, что вы хотите понять. Эти объяснения в статистике называются теорией.
- Теория позволяет вам сделать некоторые предположения о зависимостях между вашими наблюдениями. Такие предположения называются гипотезами.
- Чтобы проверить гипотезы, вам нужны данные. Вы их собираете.
- После того, как вы их собрали, вы их анализируете.
- Анализ данных либо подтверждает теорию, либо ее

Что такое валидное исследование?

- **Валидное исследование спланировано так, чтобы исключить альтернативные объяснения наблюдаемого явления**.
- **Условия валидности (**условия для установления причинно-следственной зависимости от явления А к явлению В)**:**
- Во-первых, А должно предшествовать по времени В; это **хронологическая валидность**.
- Во-вторых, должна существовать статистическая зависимость между А и В; т.е. должно быть установлено, что А сопутствует В. Это – **валидность статистического вывода**.
- В третьих, не должно быть альтернатив причине появления В помимо А. Это условие называется **внутренней валидностью**.
- Существует и **конструктная валидность**, которая связана с верным выбором теории.
- Наконец, в-пятых, существует **внешняя валидность** – возможность обобщения результатов для различных периодов времени, условий и групп.

Зависимые и независимые переменные

- Для того, чтобы проверить гипотезы, мы должны измерить переменные.
- Переменные меняются (варьируются) между людьми (IQ, рост, вес), от условий (работающие или безработные), во времени (настроение, доход, количество детей).
- Большинство гипотез может быть выражено в терминах двух переменных: причина и следствие.
- Те переменные о которых мы думаем, что они причина, называются независимыми.
- Те переменные, которые мы воспринимаем как следствие – называются зависимыми.

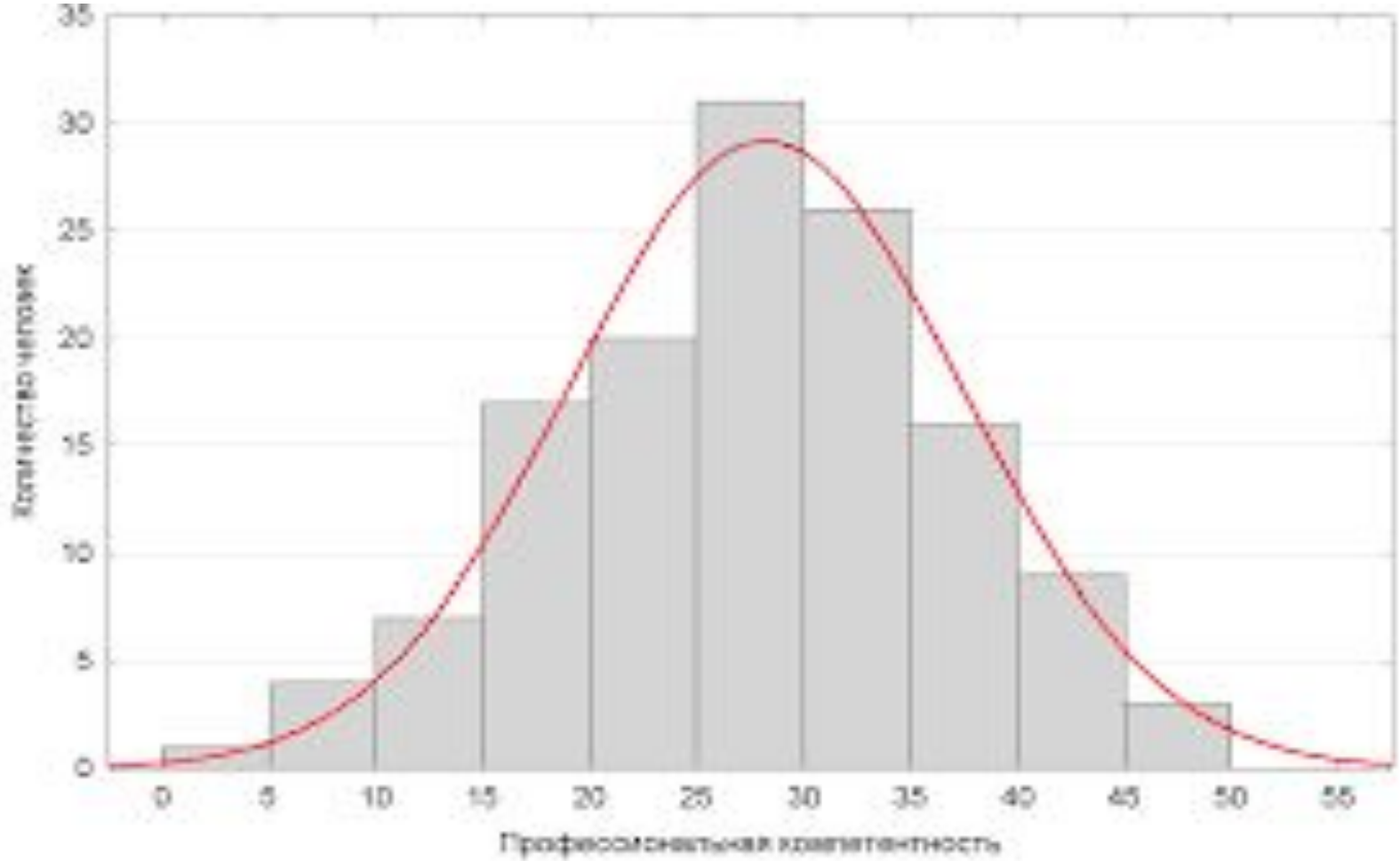
Уровни измерения

- Переменные бывают дискретные и непрерывные. В статистике дискретные переменные называют категориальными.
- Категориальные переменные бывают:
  - Биноминальными
  - Номинальными
  - Порядковыми
- Непрерывные переменные бывают:
  - Метрические
  - Интервальные

Частотное распределение

- После того, как вы собрали данные, полезно для каждой переменной посчитать, сколько раз встречается каждое ее значение и построить график.
- Такие расчеты называются частотным распределением, а график – гистограммой.
- В идеальном мире наше распределение должно быть нормальным.
- Потому что все случайные переменные распределены нормально.

Гистограмма и нормальное распределение



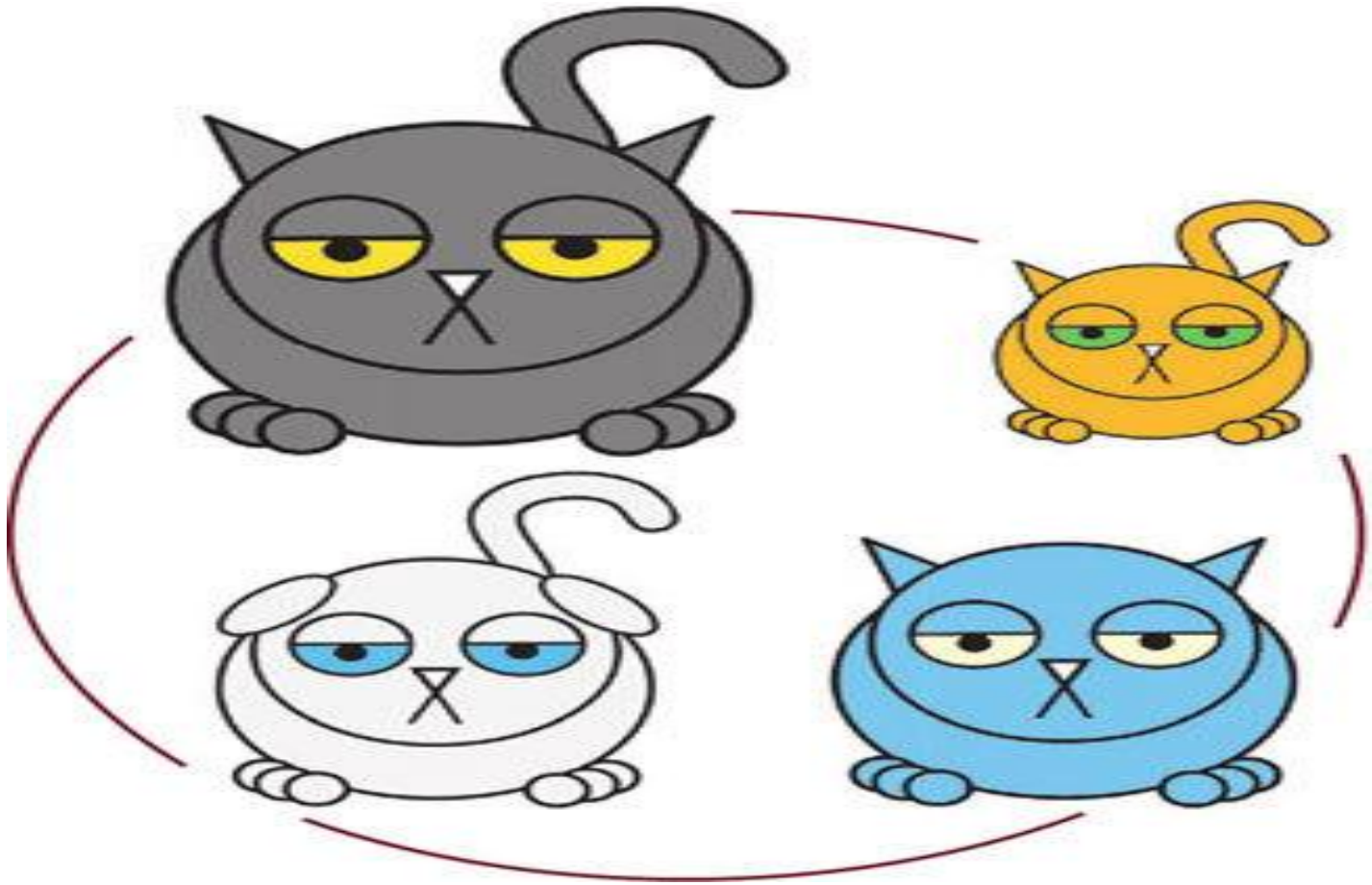
Центральная тенденция

- После того, как мы сделали частотное распределение, нам нужно найти его центр, который называют центральной тенденцией.
- Есть три основных измерения центральной тенденции: среднее, мода и медиана.

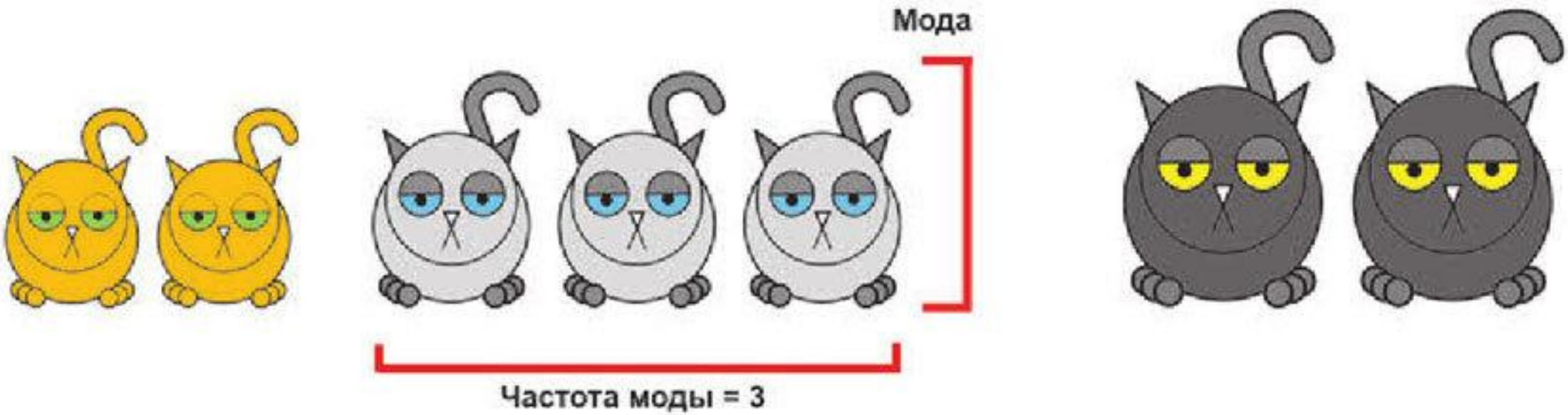
Мода

- Мода – это значение, которое встречается чаще всего.
- Ее легко увидеть на графике.
- Ее легко вычислить: надо посчитать сколько раз встречается то или иное значение переменной и выбрать то, которое встречается чаще.

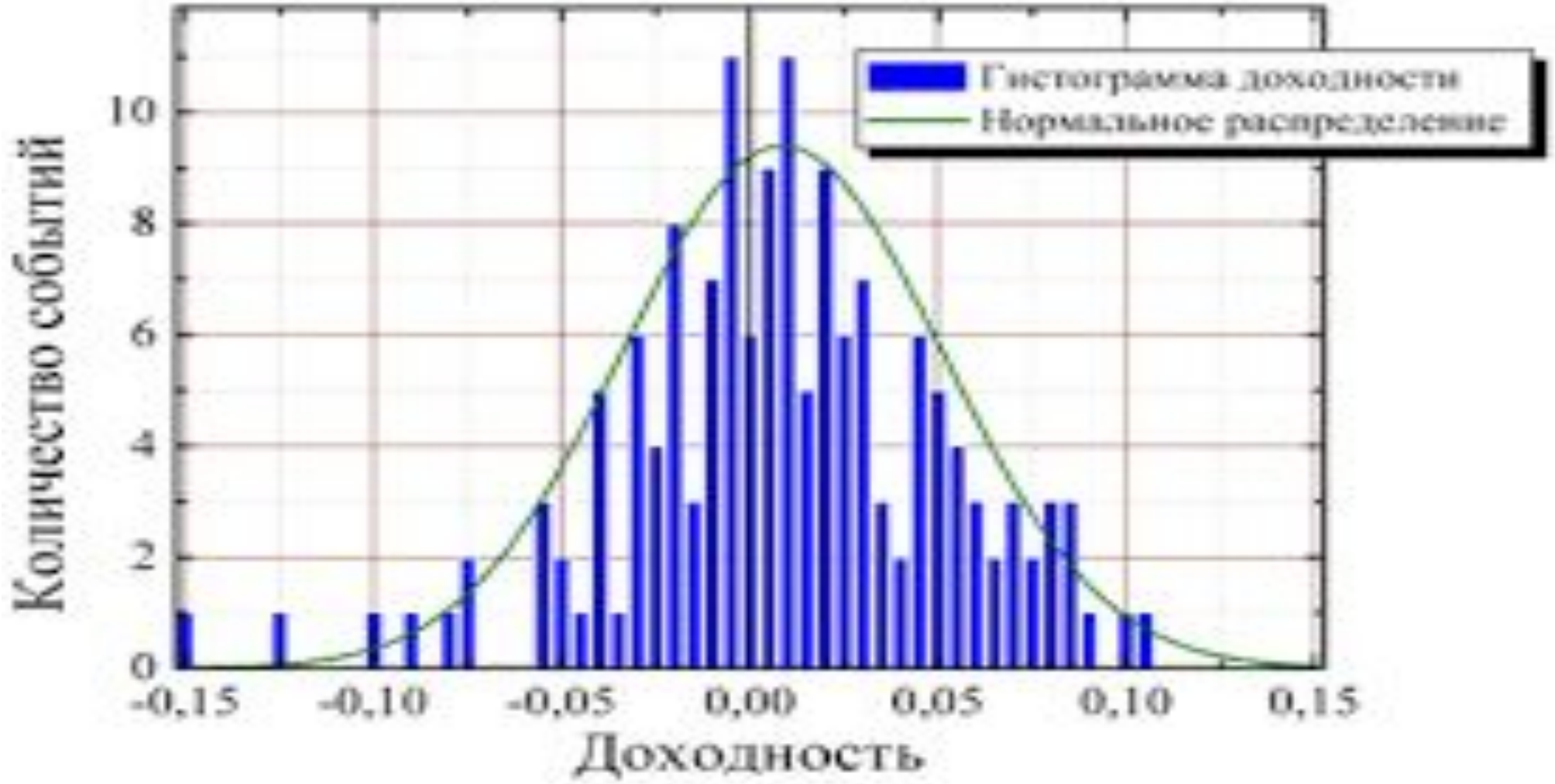
Котики бывают разные…



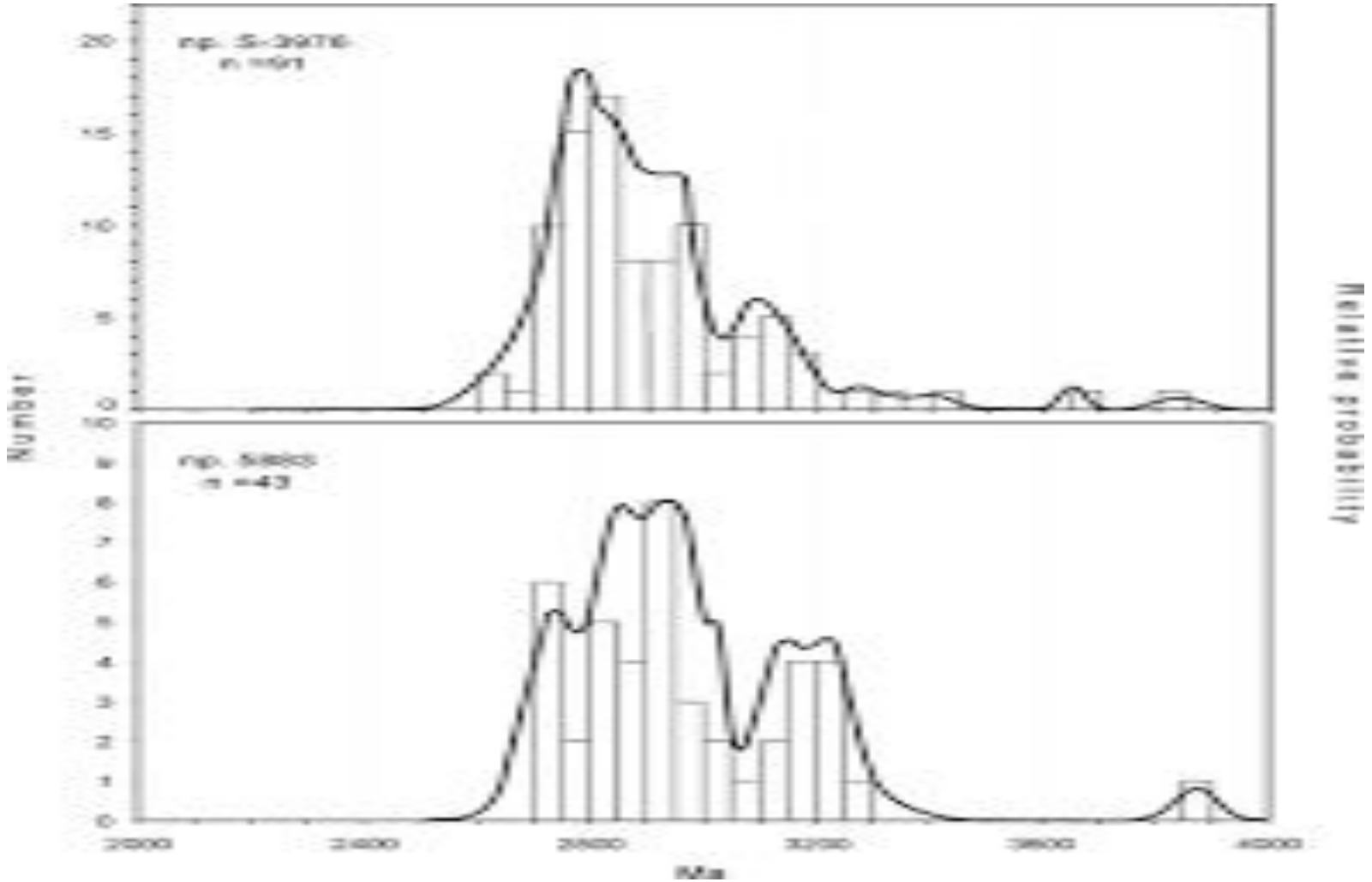
Как найти моду?



Бимодальное распределение



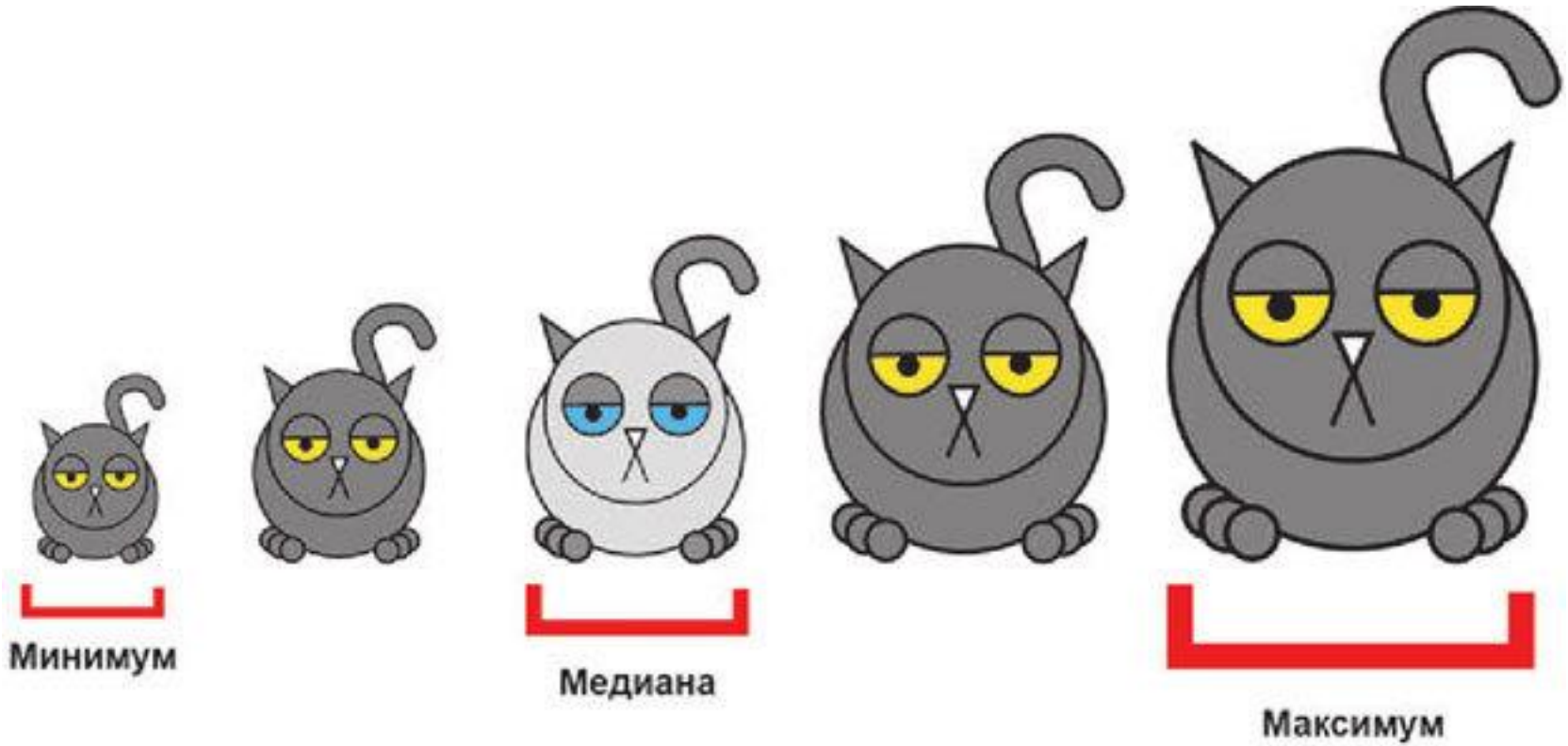
Мультимодальное распределение



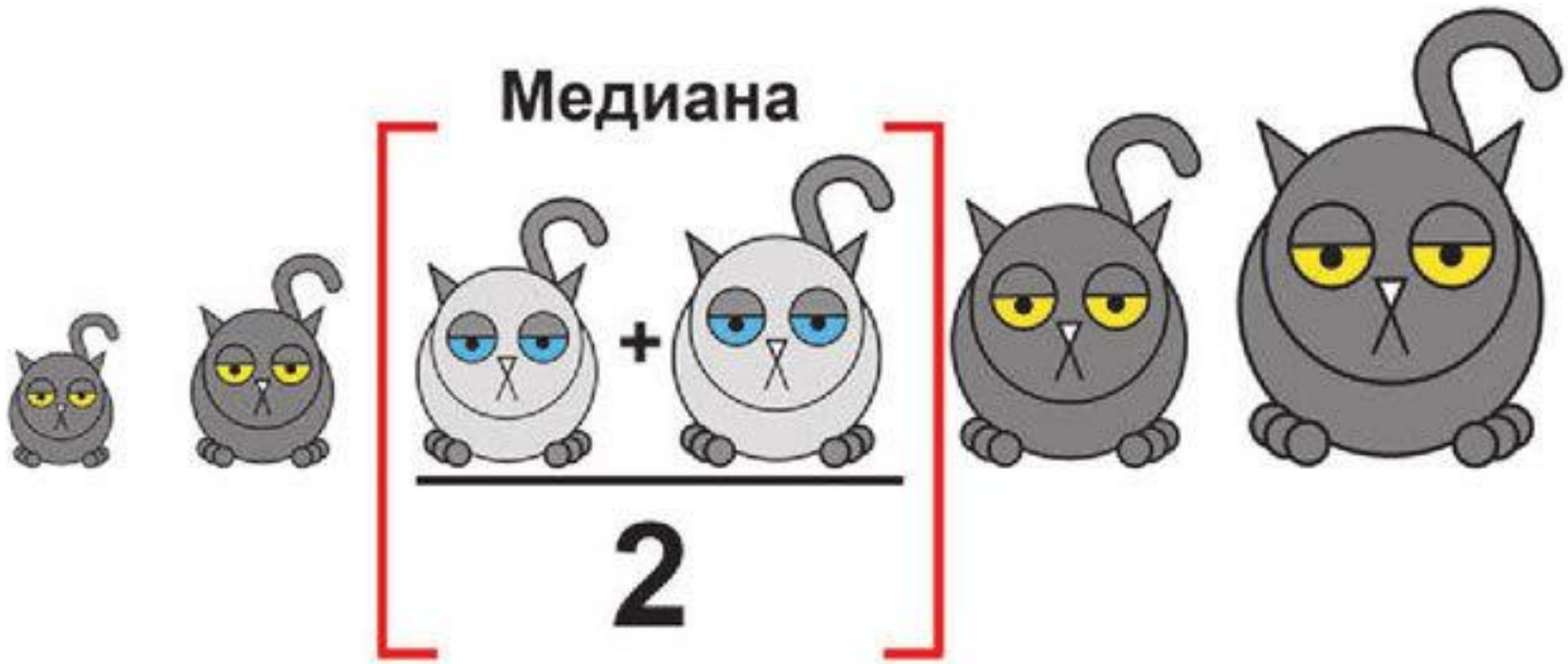
Медиана

- Еще один способ определить центр распределения – это посчитать медиану.
- Медиана – это значение, которое делит нашу выборку пополам, т.е половина выборки имеет значение этого параметра ниже, чем медиана, а вторая половина выборки – выше, чем медиана.
- Пример: количество друзей в Facebook: 108, 103, 252, 121, 93, 57, 40, 53, 22, 116, 98
- Для того, чтобы посчитать медиану, надо расположить значения в порядке возрастания: 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252
- Затем найдем элемент выборки, который находится посередине: n=11, (n+1)/2=6
- Значение 6-го элемента равно 98.
- Медиана=98.
- **У номинальных переменных медианы нет!!! Они не числовые!!!!**

Медиана размера котиков



А если у нас четное число котиков?



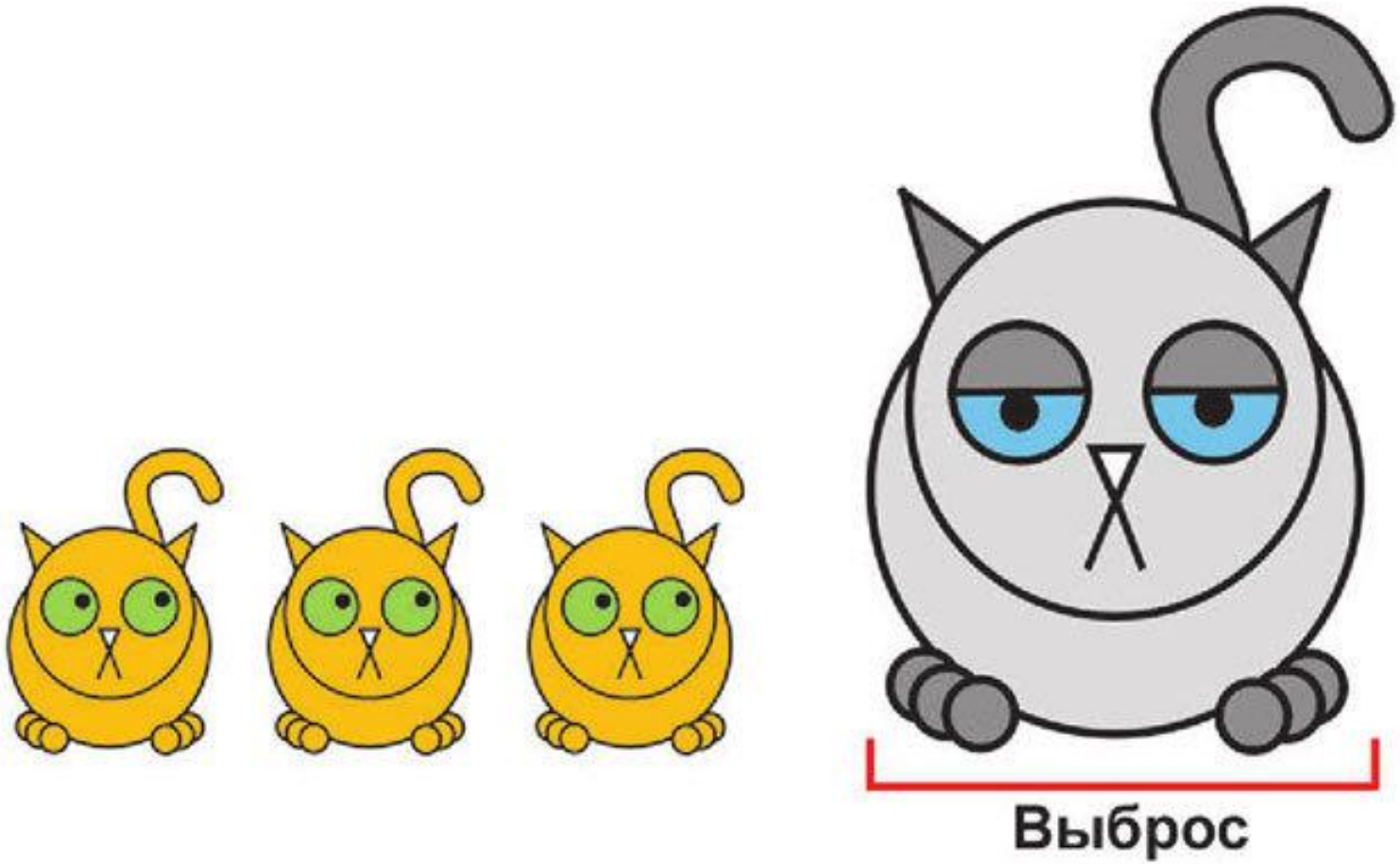
Среднее (Mean)

- Среднее – это среднестатистическое значение нашего распределение (average)
- Для того, чтобы его вычислить надо сложить все значения нашего распределения и поделить на размер выборки:
- $\Sigma(xi) = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 252$

$= 1063$

- $\overline{X}=\Sigma(xi)/n=1063/11=96.64$

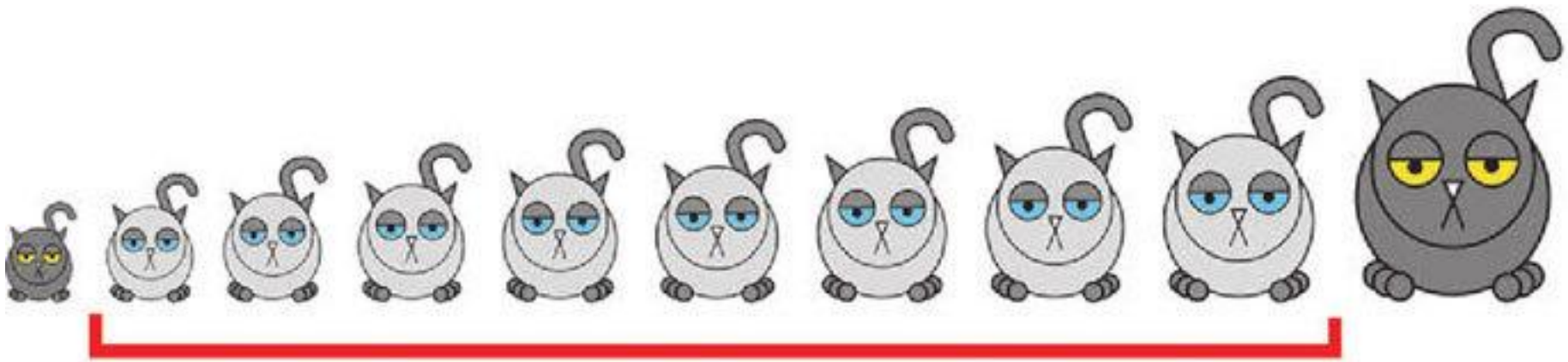
Почему среднее не всегда является лучшим показателем типичности?



Как корректировать данные при выбросах?

- Надо убрать 5-10% самых больших и самых маленьких значений, и посчитать среднее для оставшихся величин.
- Такой показатель называется усредненное среднее.

Усредненное среднее



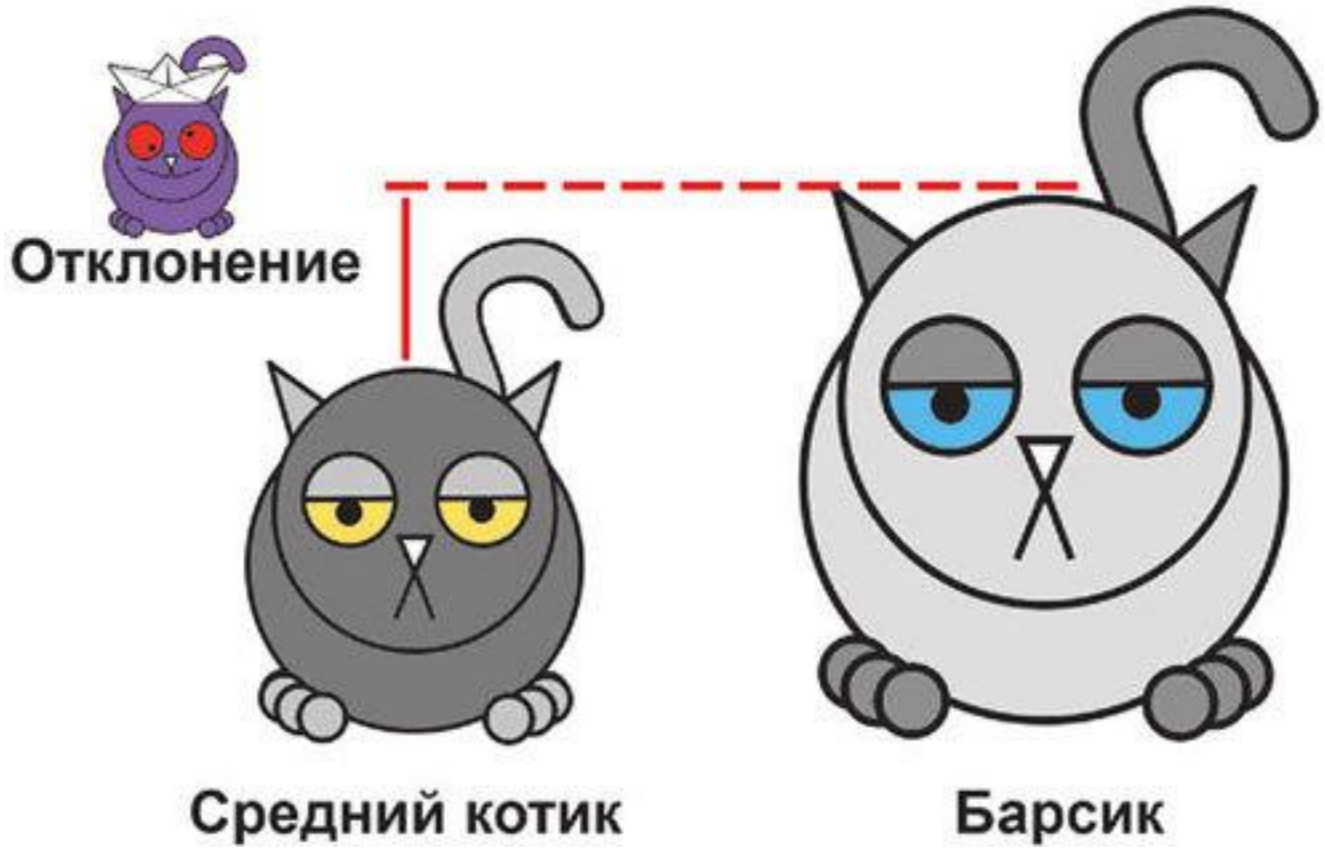
Котики для усеченного среднего

Меры разнообразия

- Межквартильный размах.
- Размах – различие между самой большой и самой маленькой величиной.
- Если мы уберем 25% самых больших значений и самых маленьких значений, то получим межквартильный размах.

С котиками все то же самое…

Дисперсия и стандартное отклонение



Как посчитать дисперсию и стандартное отклонение?

- Если взять и сложить все отклонения от среднего, то получится 0, так как отклонения бывают в разную сторону.
- Поэтому отклонения от среднего надо возвести в квадрат, а потом уже сложить.
- Полученную сумму надо разделить на общее количество наблюдений.
- $\sigma_x^2=\Sigma(x_i-\mu_x)^2/N$
- **σ** (корень из σ_x^2)**- стандартное отклонение**

Меры разнообразия

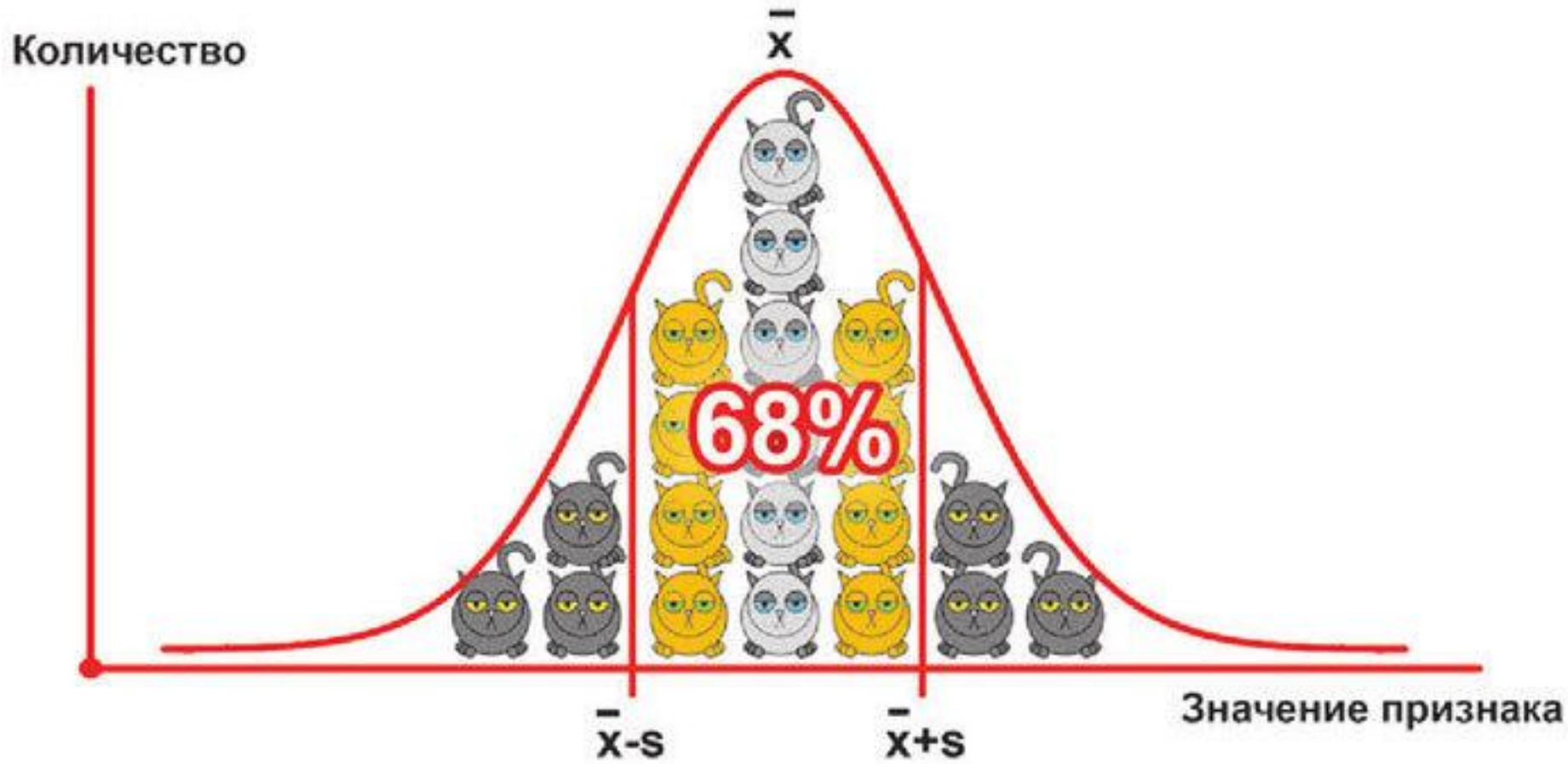
Дисперсия

Среднее отклонение

Важно помнить!

Важно помнить!

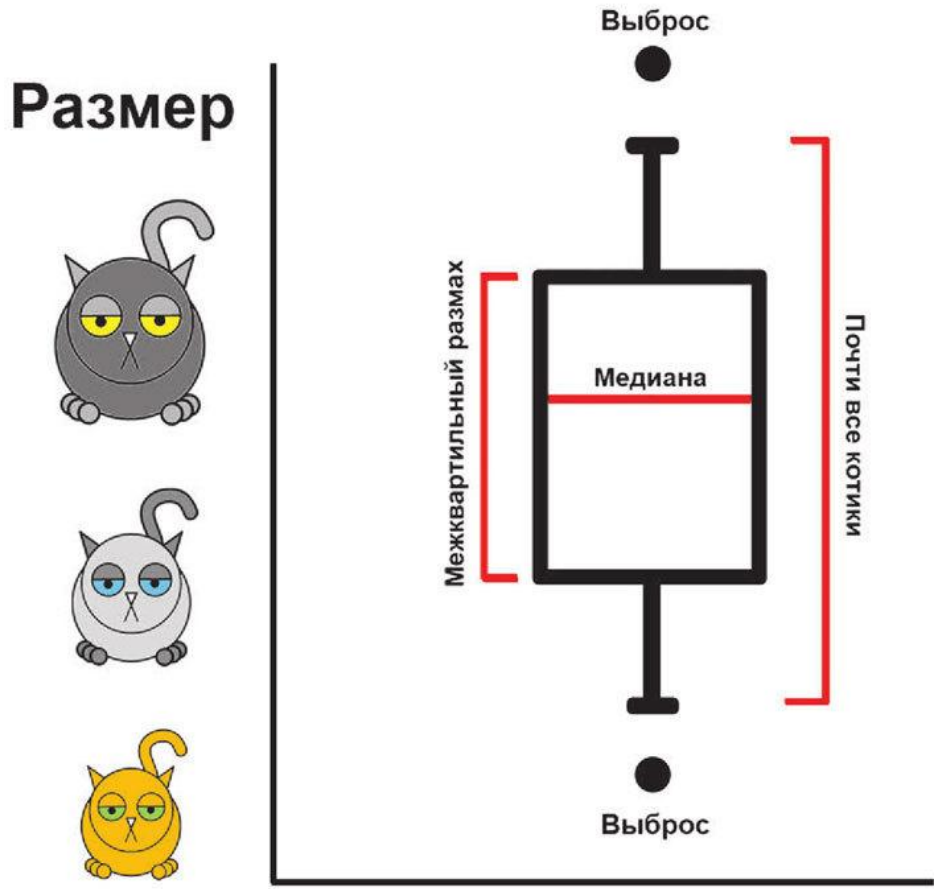
Свойства нормального распределения



Особенность нормального распределения

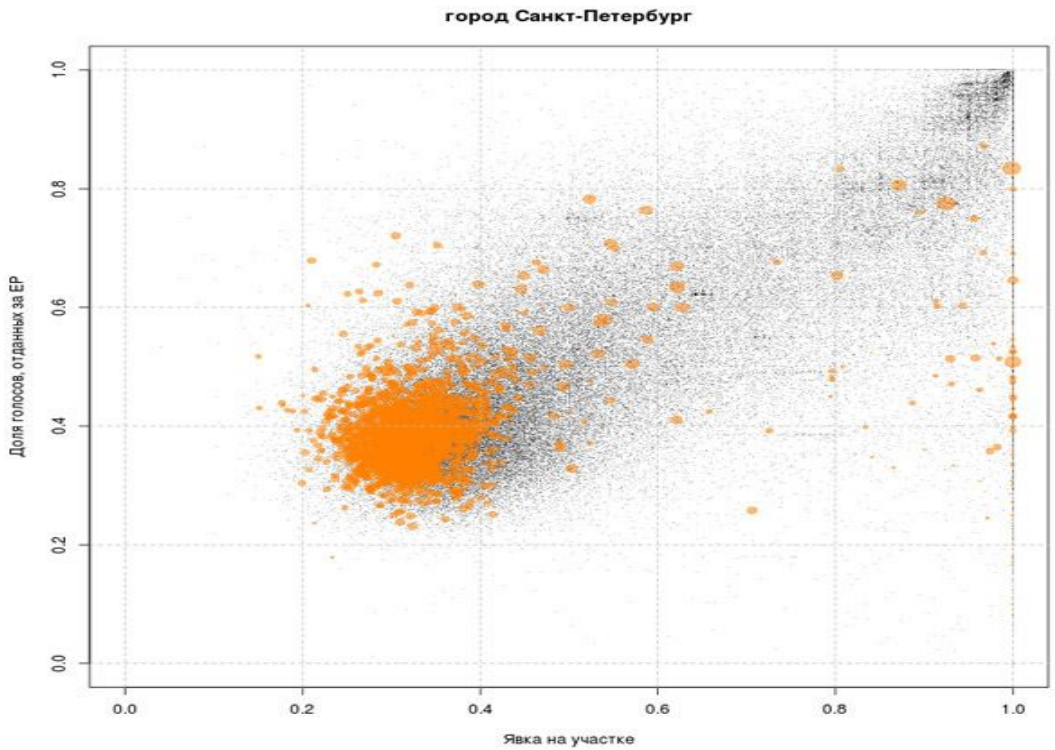
- Особенностью нормального распределения является то, что 99,73% всех случаев находятся в пределах трех стандартных отклонений от среднего значения.
- В пределах двух стандартных отклонения находится 96% всех случаев.
- 95% всех случаев будут находиться в пределах ±1,96 стандартных отклонений от средней.

Визуализация мер типичности и разнообразия - «Ящик с усами»

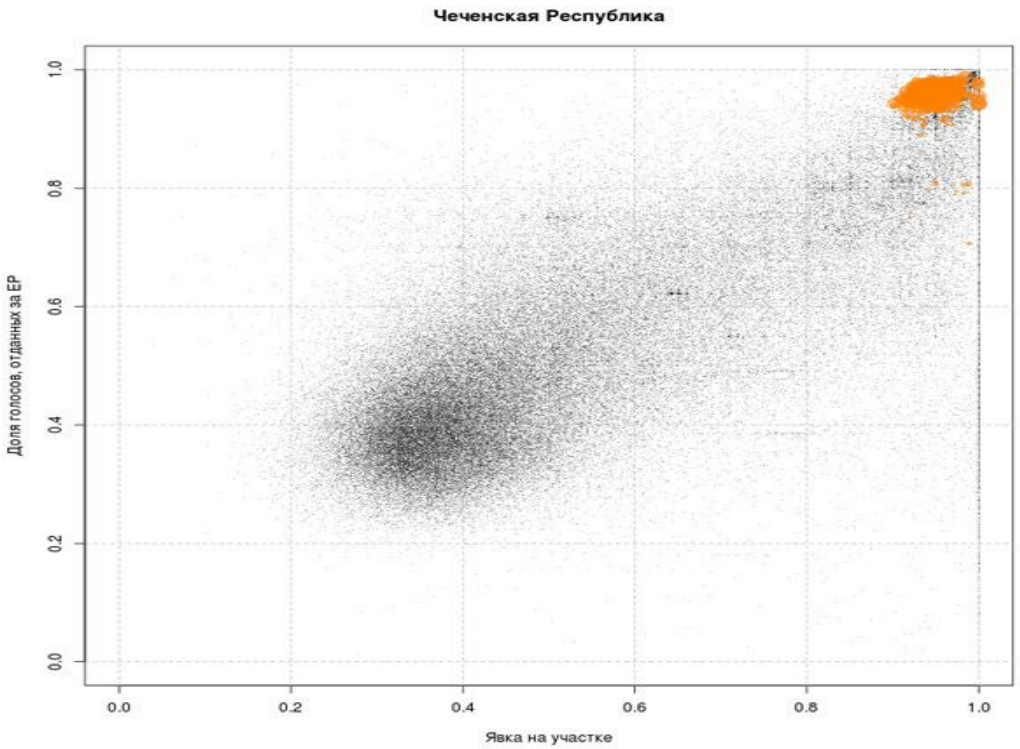


Явка и голосование за партию власти

Санкт-Петербург

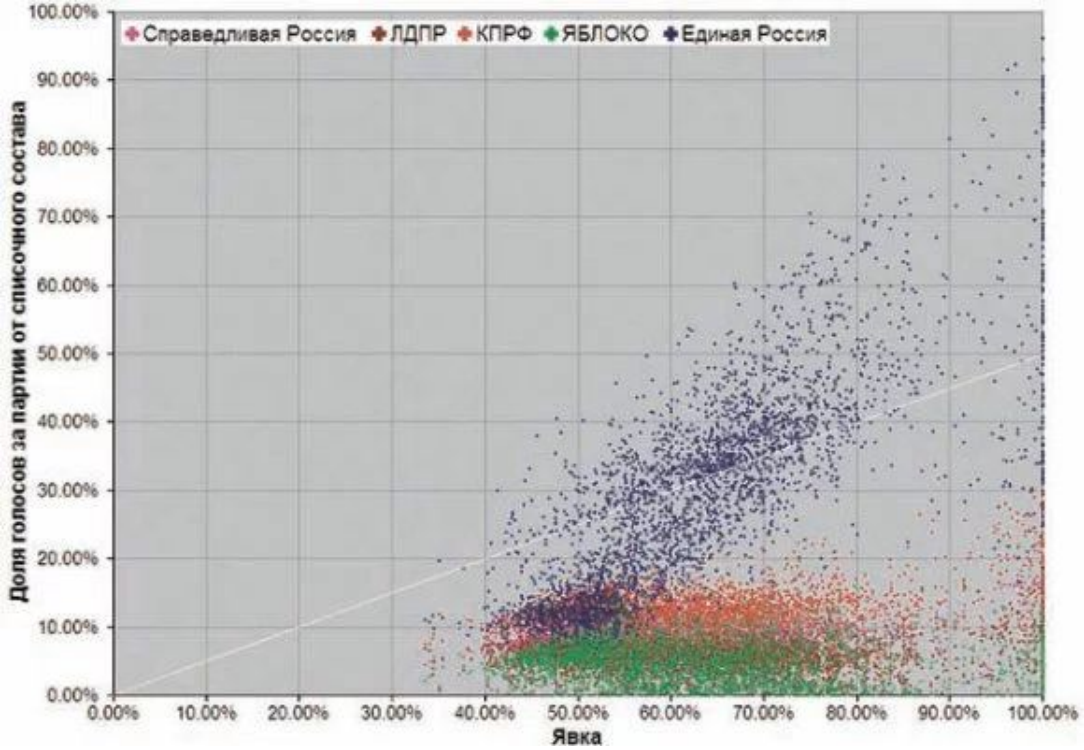


Чечня

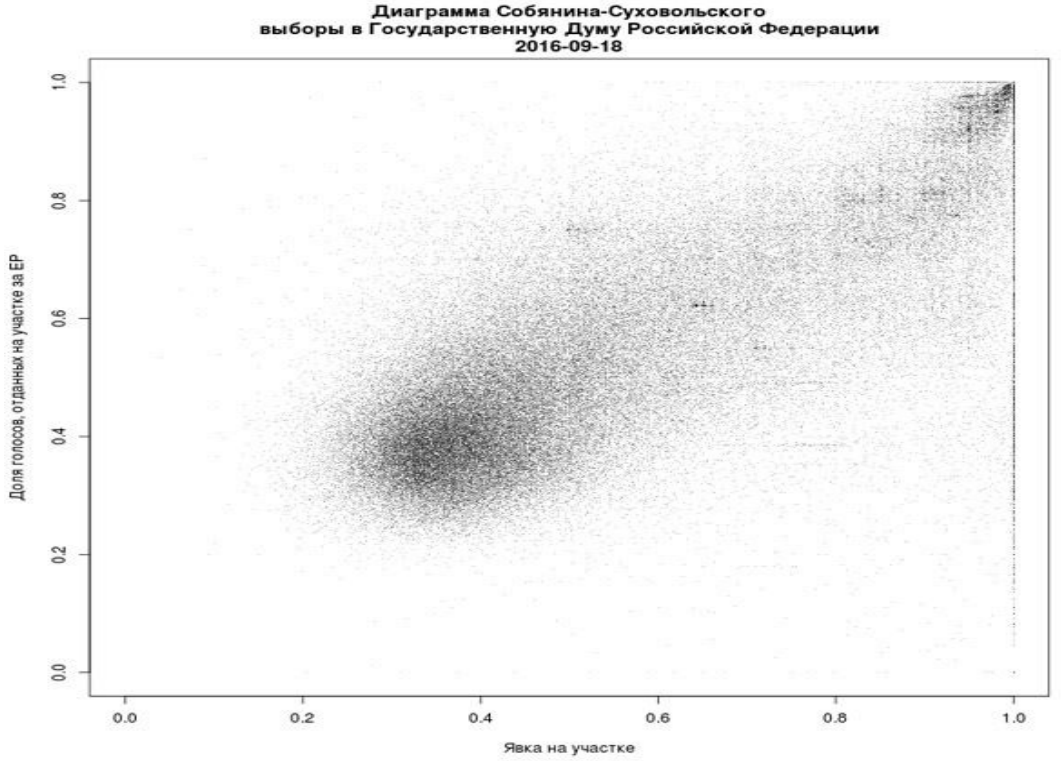


Явка и голосование за партию власти

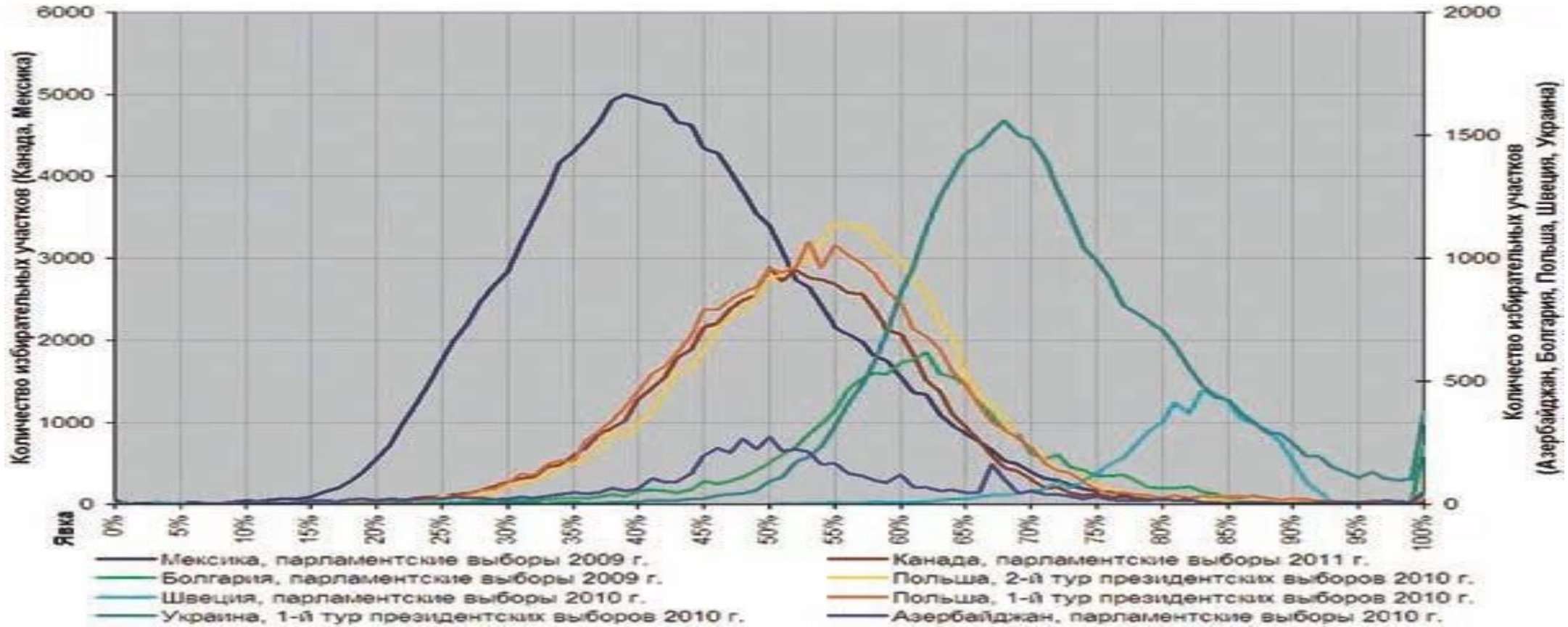
2011 год



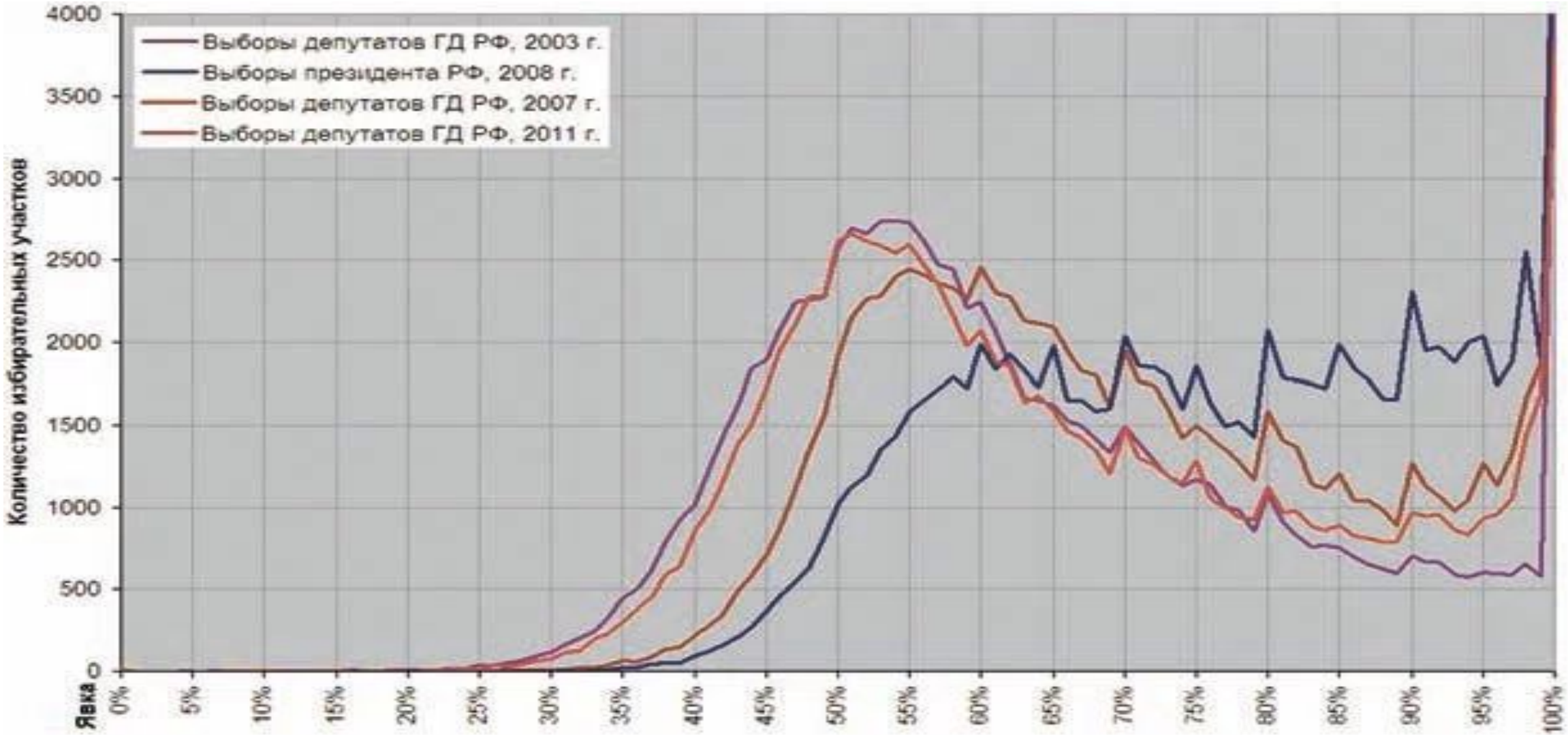
2016 год



Явка на избирательных участках

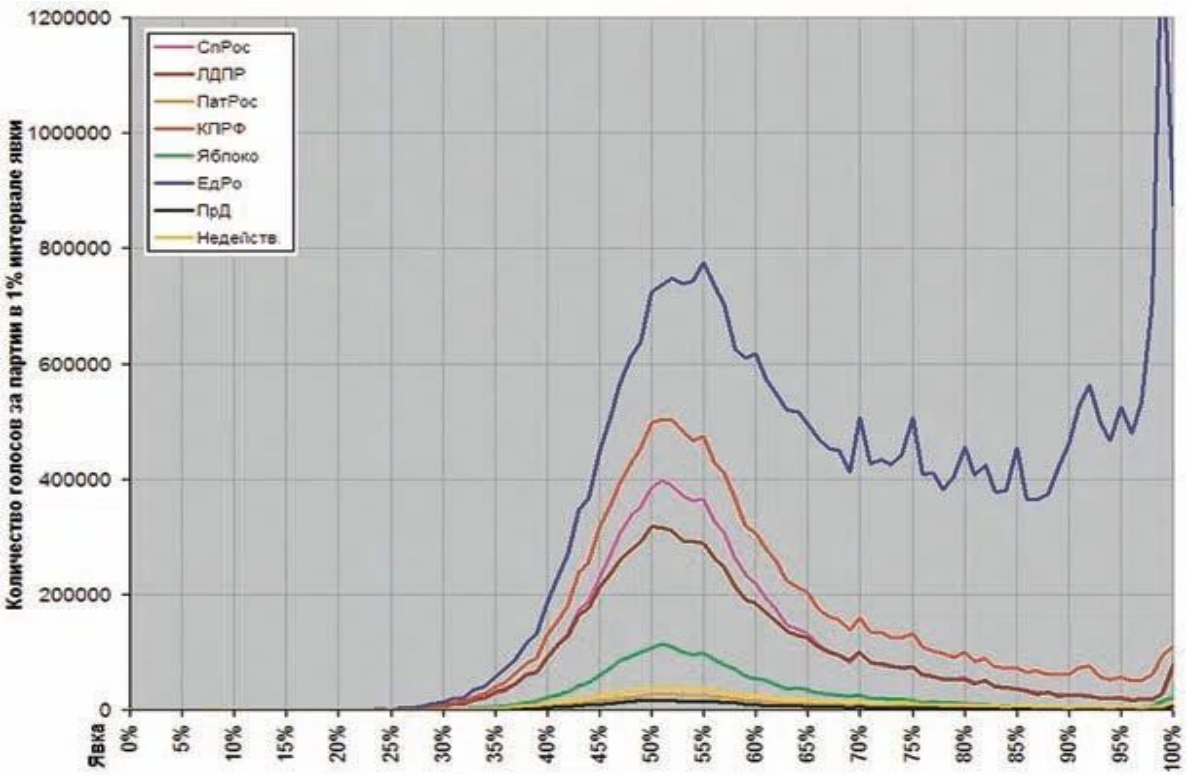


Явка на участки в России

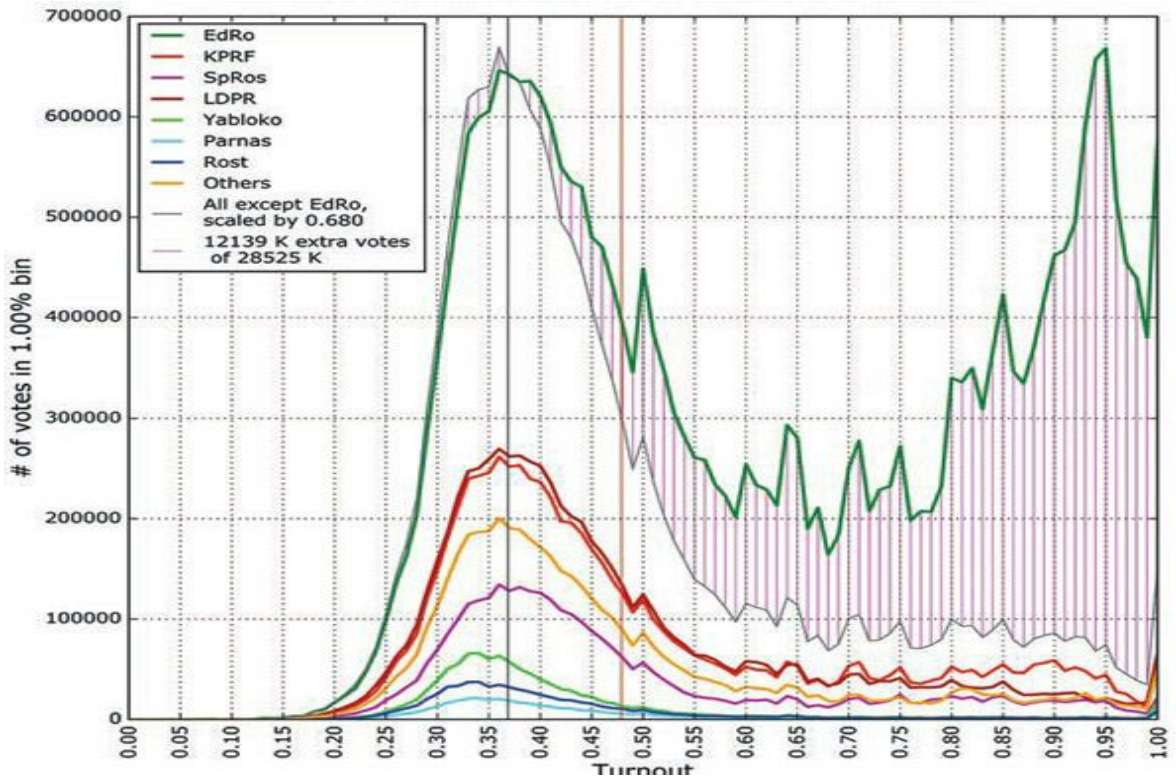


Распределение голосов от явки

2011

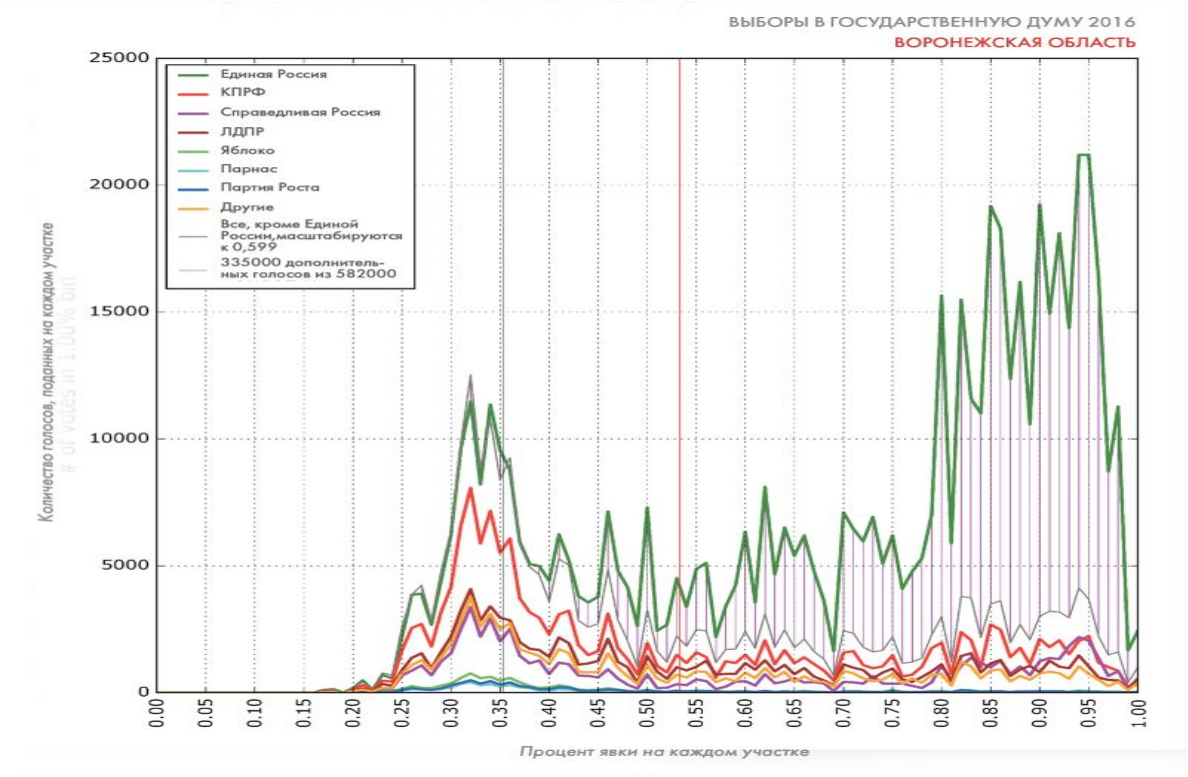


2016

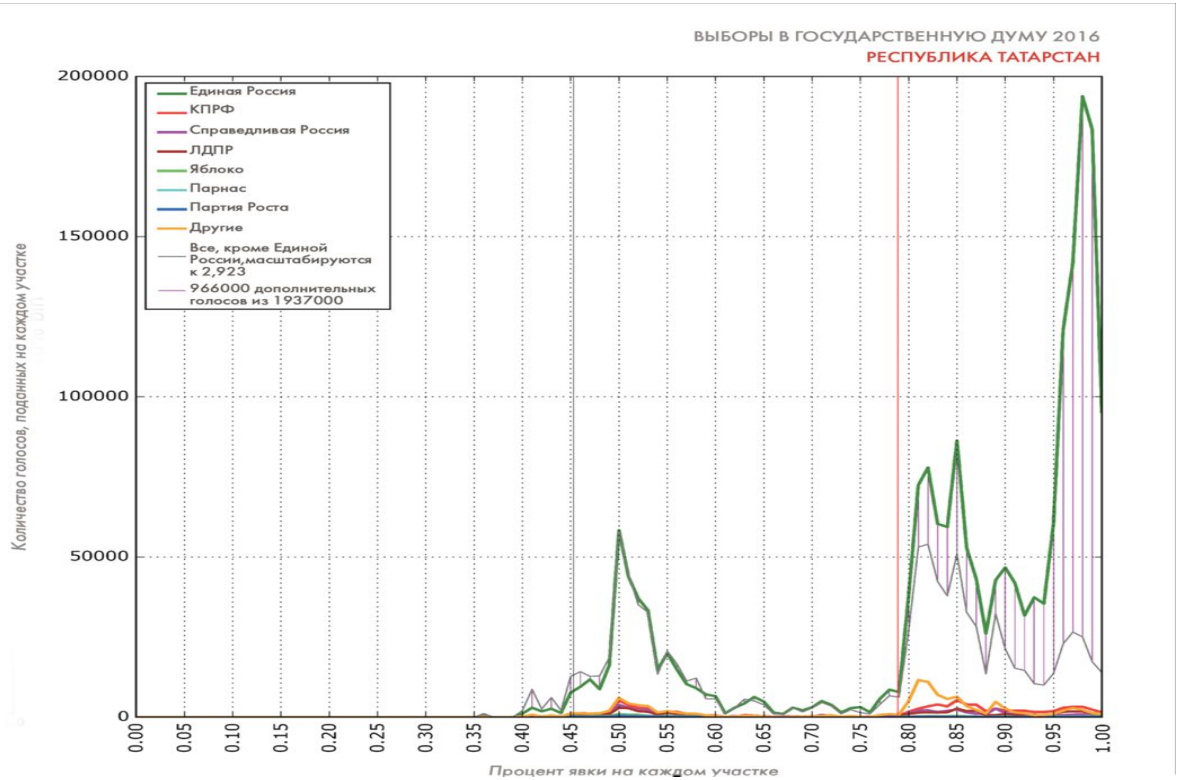


Аномалии в регионах

Воронежская область



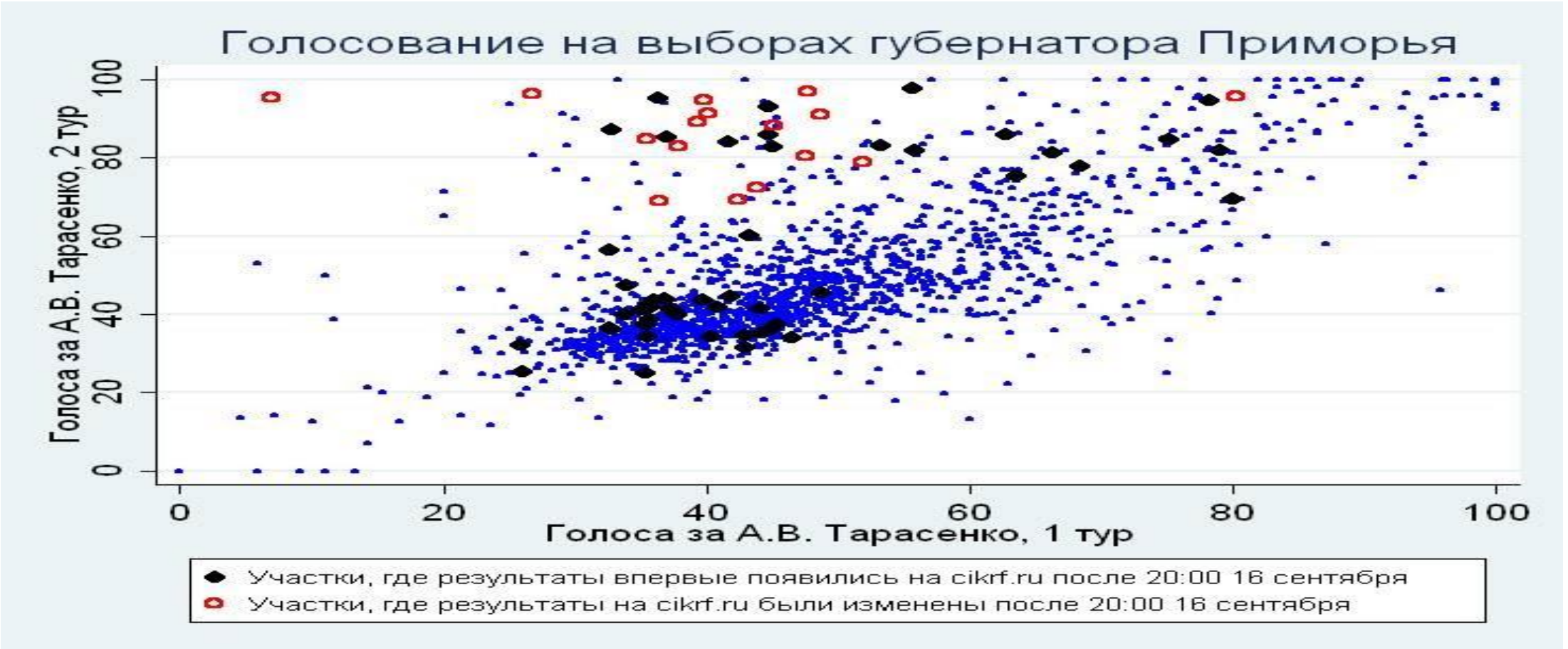
Татарстан



Реальная явка

Распределение голосов за партии

Выборы в Приморье



Выборы в Приморье