

R для тервера и матстата

Посиделка первая: что такое случайные величины и как их генерировать

Чем будем заниматься

- Поиск фальсификаций, генерация случайных величин
- Варка распределений, ЗБЧ, путешествие в Монте-Карло
- ЦПТ, сходимости и тяжёлые хвосты чёрных лебедей
- Свойства оценок, максимальное правдоподобие
- Как выжить на Титанике и доверительные интервалы
- Как придумывать и проверять гипотезы, АВ-тесты
- Как преподобный Байес защищает вас от спама
- И многое другое!
- На каждой паре игра с глубокой моралью

В чём будем заниматься

- Конечно же R!
- R очень лёгок в освоении
- Одна из самых красивых визуализаций данных
- Очень хорош в работе со статистикой!
- Много готовых пакетов и большое комьюнити
- Используется многими компаниями в работе. Вот устаревший [мини-список](#).

Будут ли активности

- Если самому не делать, никогда и не научишься => к каждой паре прилагается домашка
- Давайте поговорим о комитменте (commitment)



Будут ли отсылки

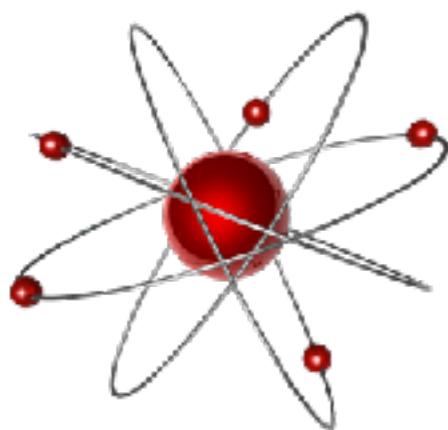
- Конечно! Поп-культура — это наше всё!



Теория вероятностей

- Мы изучали вероятности!
- Мы изучали события!
- Мы изучали случайные величины и их распределения!
- Зачем мы это делали?

Случайные ли величины?





Случайности неслучайны

Демон Лапласа



Байесовский взгляд на вероятность

- Лаплас: детерминизм, мы могли бы идеально прогнозировать вселенную, если бы измерили точное положение каждого атома. Издержки этого огромны.
- Между совершенством природы и несовершенством человеческого познания огромный разрыв.
- Неопределённость результат этого разрыва. Случайность это результат нашего невежества, а вероятность способ это невежество измерить.

Частотный взгляд на вероятность

- Наука не может рассматривать вероятность как нечто субъективное.
- Можно оценивать вероятности только тех событий, которые происходят более одного раза.
- Вопрос «Какова вероятность, что кандидат N победит на выборах?» не имеет ответа, так как событие уникально и не обладает «частотой».

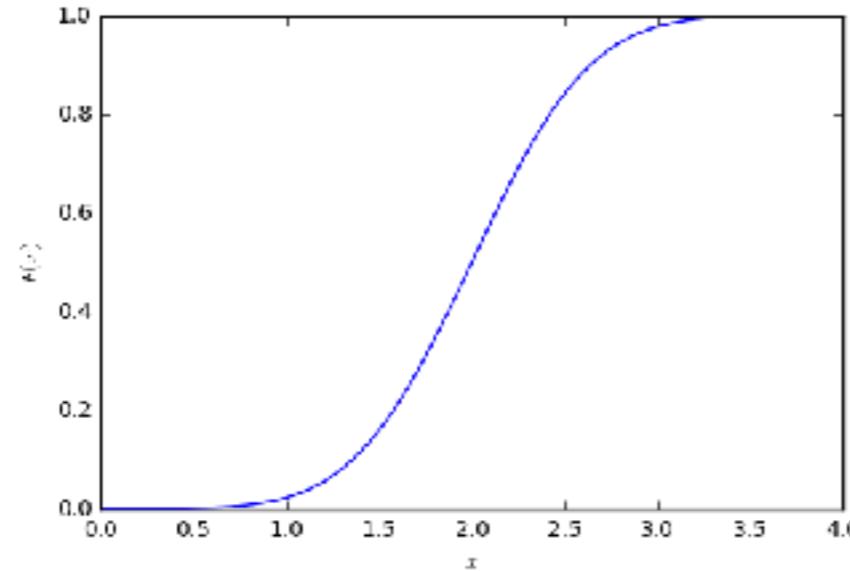
Что мы имеем



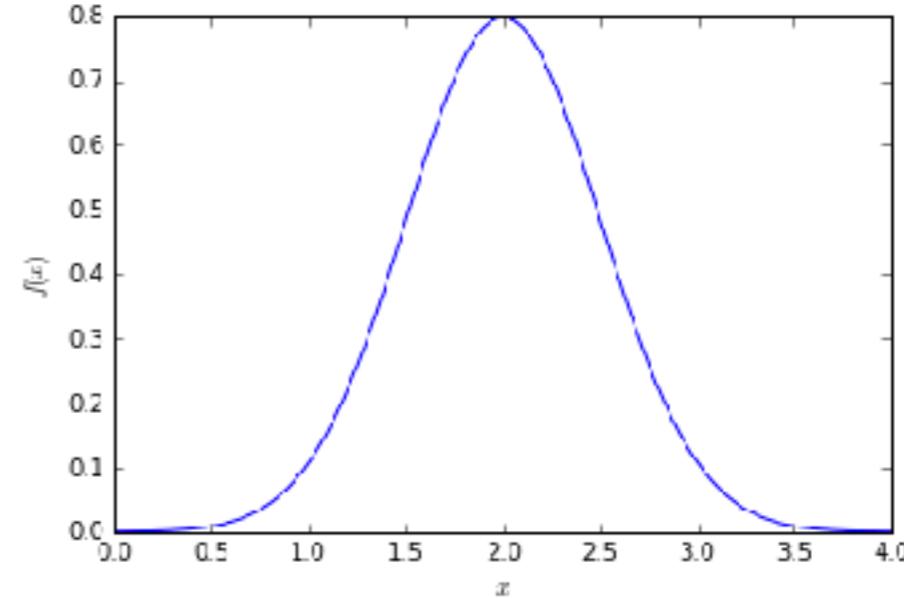
X

- Сундук — различные процессы порождения данных. Теория вероятностей изучает этот сундук. В реальности мы не видим сундука.
- Сундук порождает выборки, которые мы видим в реальном мире. Математическая статистика изучает, что сундук породил и по этому пытается восстановить его внутренности.

Барахло из сундука



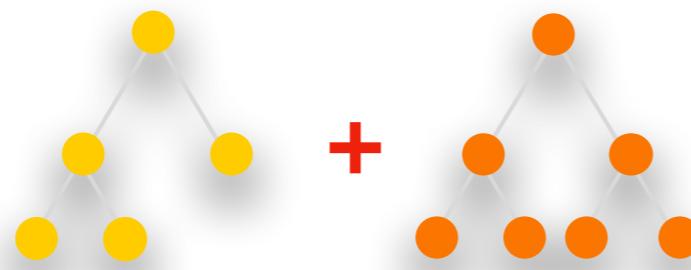
$$F_X(x) = P(X \leq x)$$



$$F_X(x) = \int f_X(x) dx$$

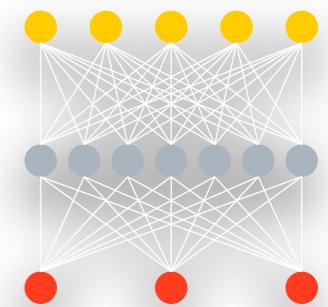
- На теории вероятностей мы приоткрыли сундук и на нас выпрыгнули различные распределения. В сундуке осталось много неизведанного. Например:

$$y_i = \sum_{j=0}^m w_j X_{ij} + \epsilon_i$$



Линейные модели

$$M = U V^T$$

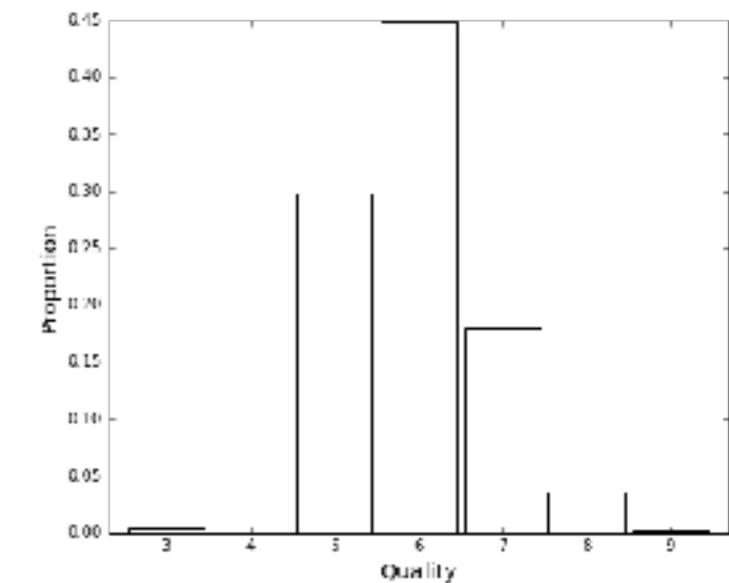
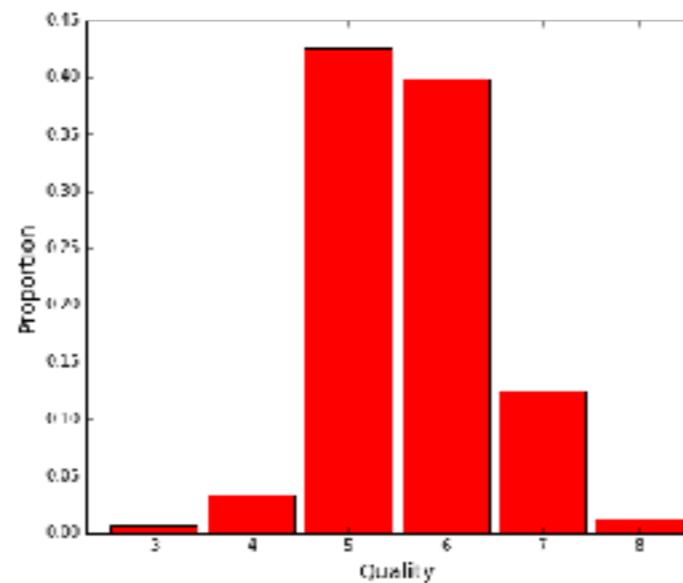
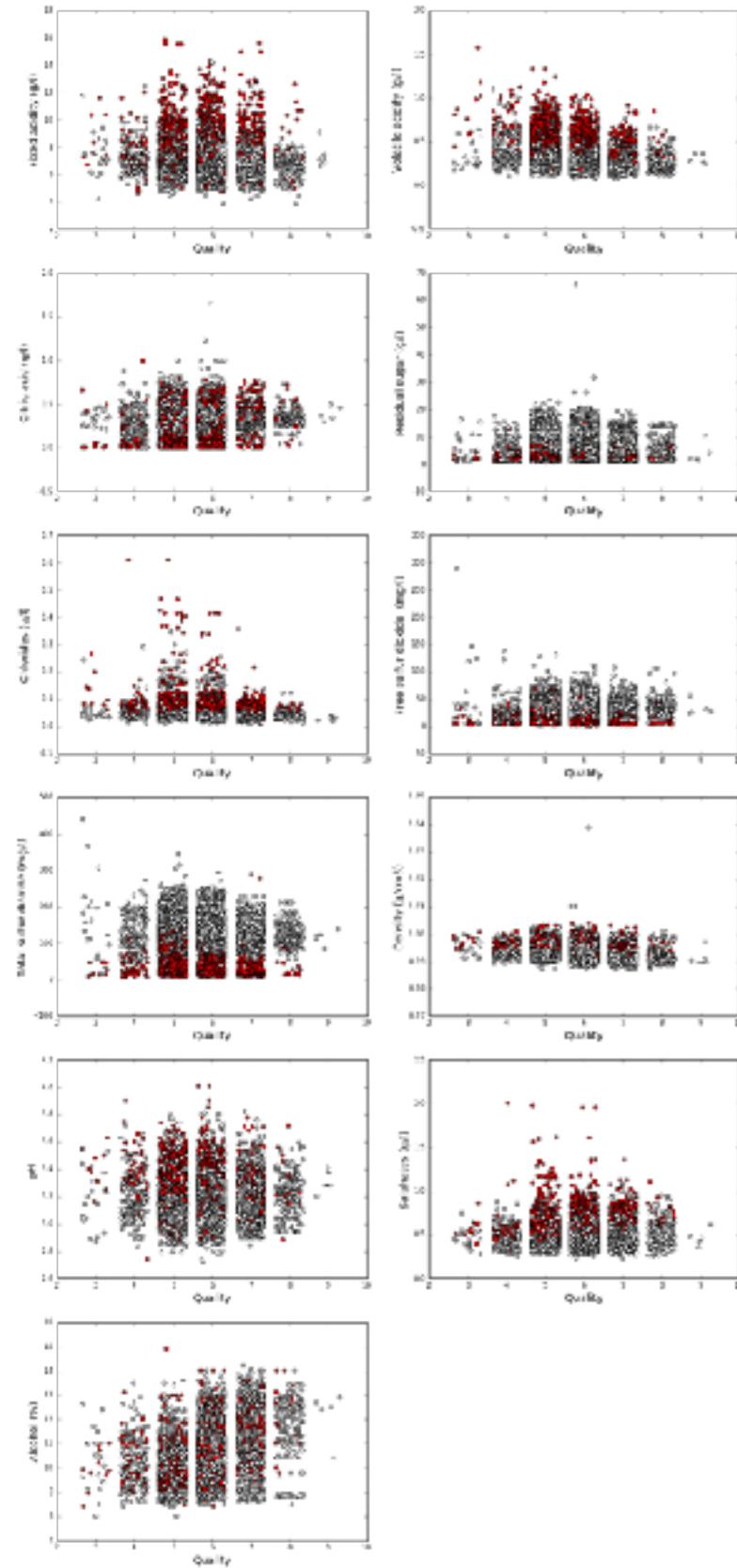


Матричные разложения

Деревья

Нейронные сети

Что извергает сундук



- Выборки, по которым можно попробовать восстановить внутренности сундука!

| Type | Fixed acidity (g/l) | Volatile acidity (g/l) | Citric acid (g/l) | Residual sugar (g/l) | Chlorides (g/l) | Free sulfur dioxide (mg/l) | Total sulfur dioxide (mg/l) | Density (g/cm³) | pH | Sulphates (g/l) | Alcohol (%) | quality |
|------|---------------------|------------------------|-------------------|----------------------|-----------------|----------------------------|-----------------------------|-----------------|------|-----------------|-------------|---------|
| 1098 | 8.7 | 0.41 | 0.41 | 6.2 | 0.078 | 25.0 | 42.0 | 0.96530 | 3.24 | 0.77 | 12.6 | 7 |
| 4292 | 6.7 | 0.23 | 0.26 | 1.6 | 0.035 | 25.0 | 143.0 | 0.96265 | 3.30 | 0.54 | 10.3 | 6 |
| 5960 | 6.5 | 0.20 | 0.33 | 1.6 | 0.039 | 25.0 | 110.0 | 0.96008 | 3.22 | 0.86 | 12.0 | 6 |
| 2216 | 7.4 | 0.19 | 0.30 | 1.4 | 0.087 | 33.0 | 136.0 | 0.96300 | 3.12 | 0.50 | 9.8 | 6 |
| 743 | 11.6 | 0.41 | 0.68 | 2.8 | 0.095 | 25.0 | 101.0 | 1.00024 | 3.13 | 0.53 | 10.0 | 5 |

Базовые теоремы

- Все манипуляции по восстановлению сундука позволяет делать ряд теорем
- Две из них вам уже знакомы
- Какие?

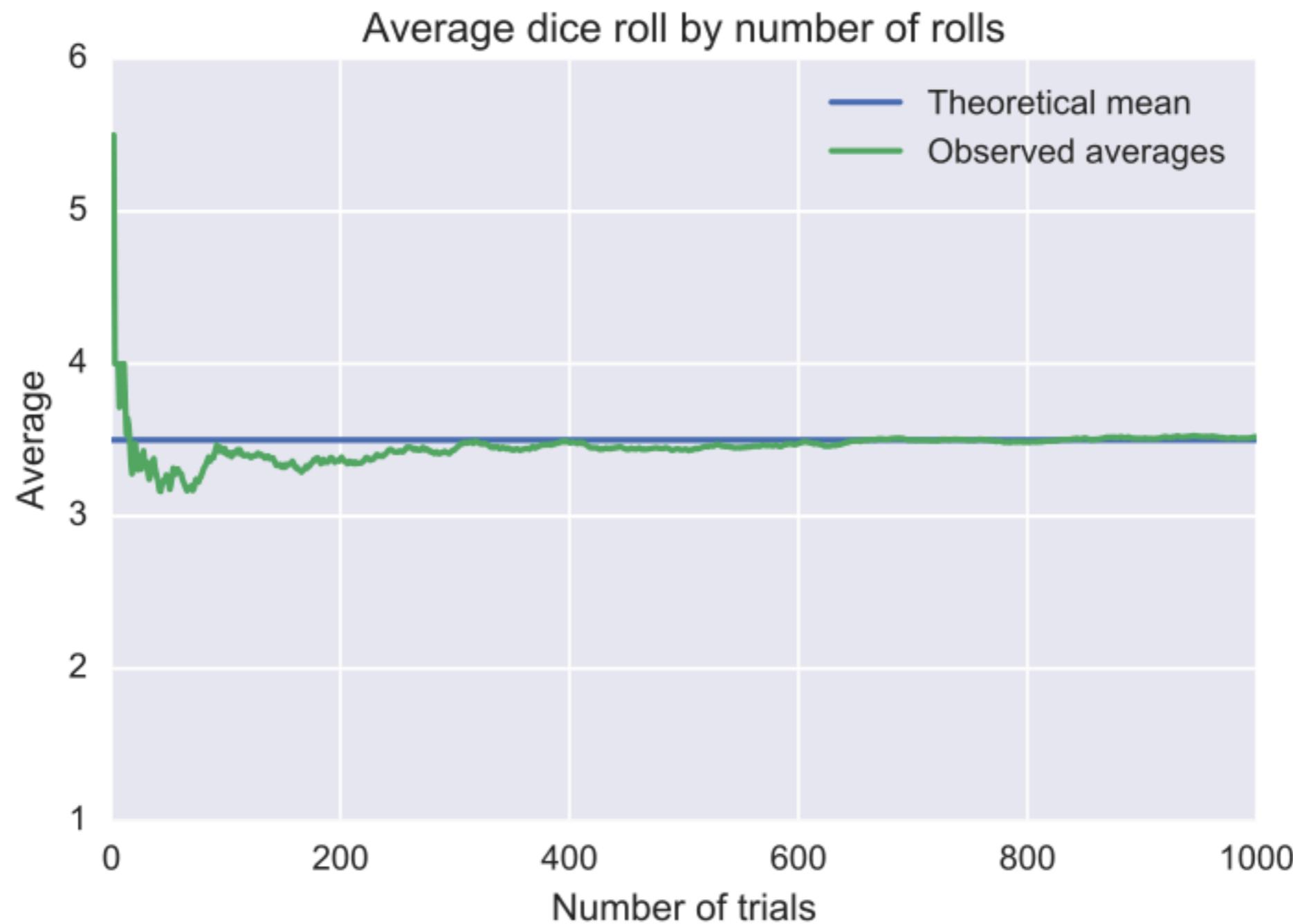
Закон больших чисел

- ЗБЧ утверждает, что среднее арифметическое большого числа похожих случайных величин «стабилизируется» с ростом их числа
- Как бы сильно случайные величины не отклонялись от своего среднего значения, эти отклонения взаимно косятся и среднее арифметическое приближается к постоянной величине
- В общем случае ЗБЧ называется любое утверждение, которое говорит, что:

$$\frac{X_1 + \dots + X_n}{n} - \frac{E(X_1) + \dots + E(X_n)}{n} \xrightarrow{P} 0$$

- ЗБЧ сформулировано довольно много: Чебышёва, Бернулли, Хинчина и тд
- Подробнее про ЗБЧ мы будем говорить на следующей паре!

Закон больших чисел

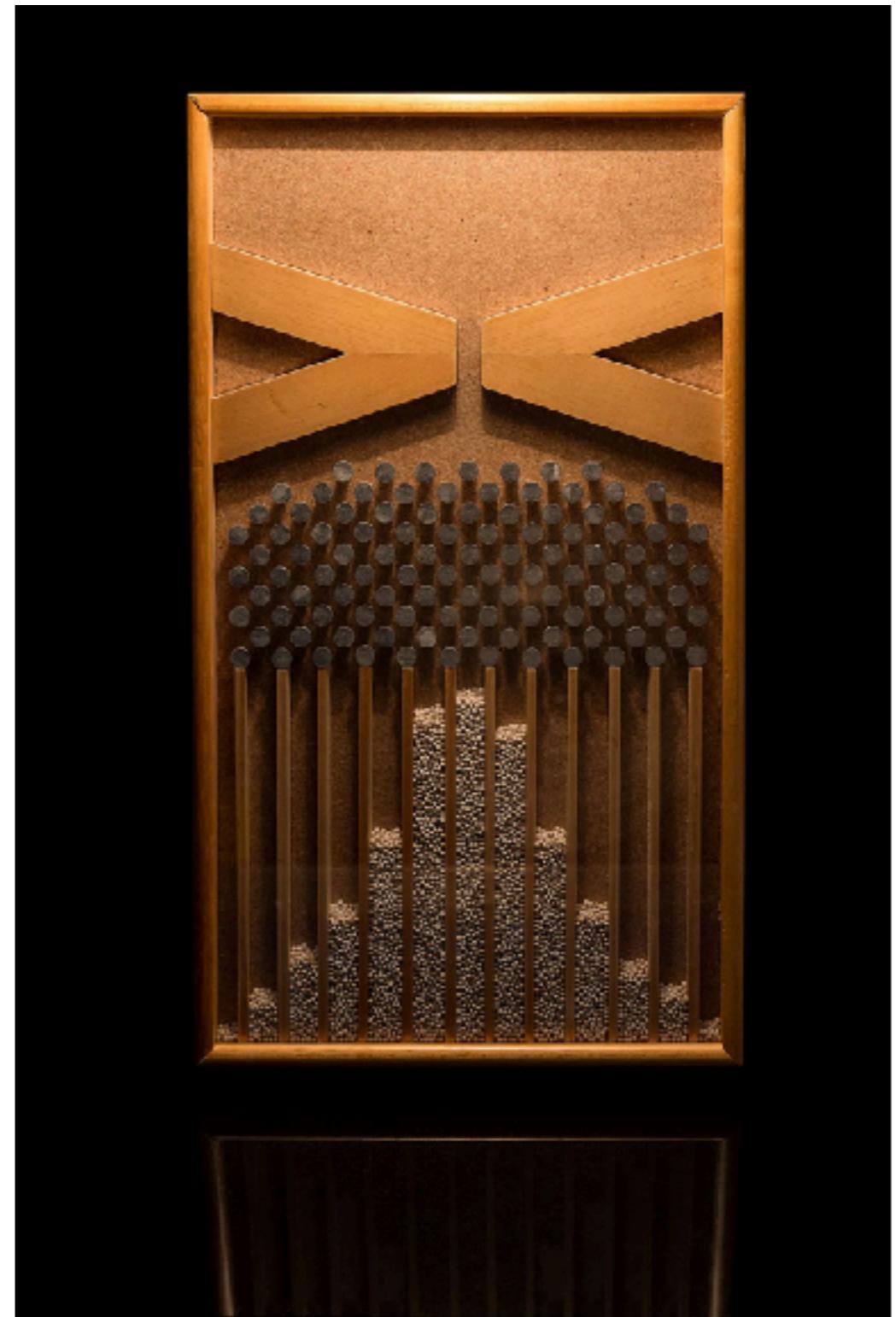
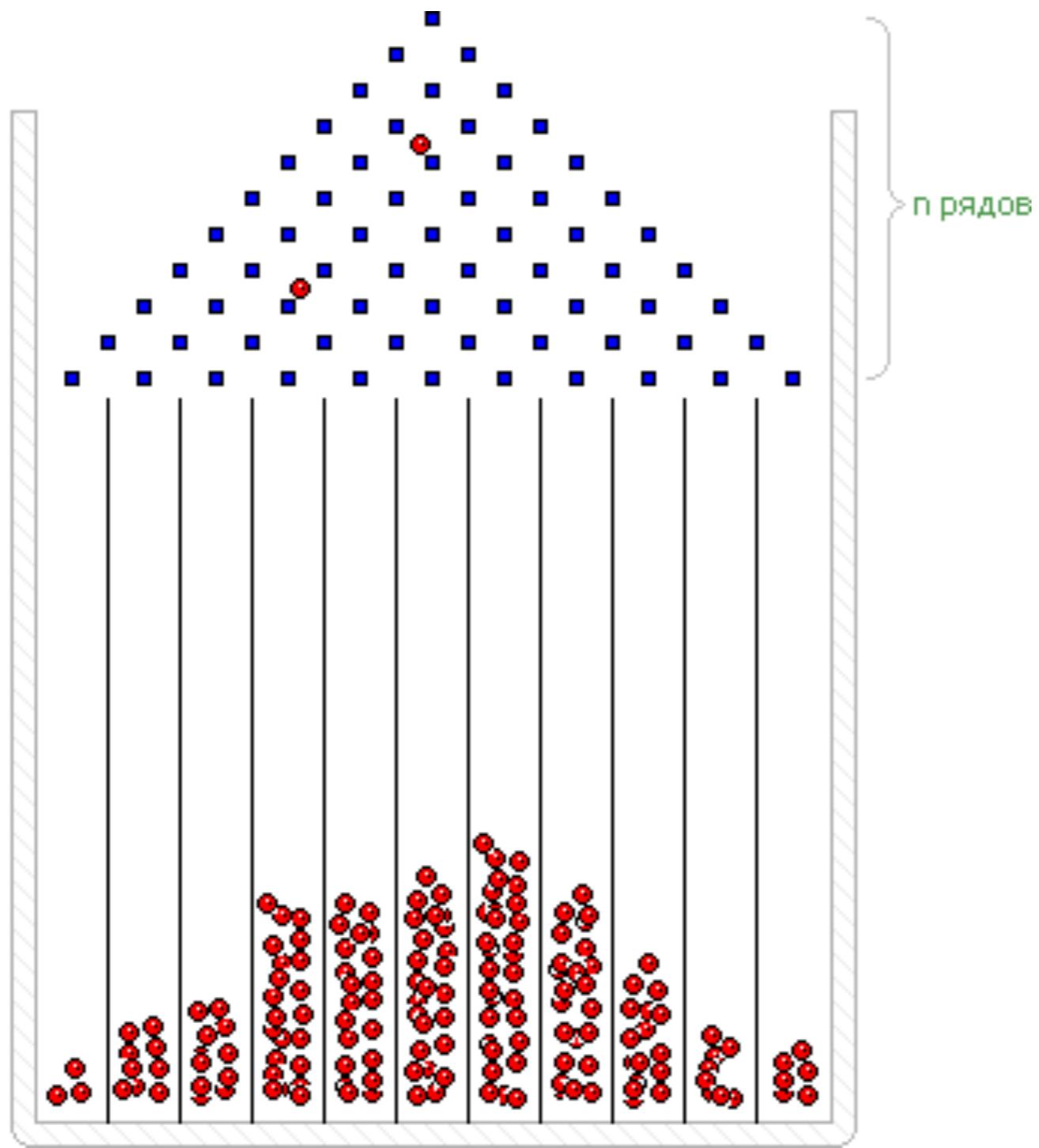


Центральная предельная теорема

- При определённых условиях сумма достаточно большого числа случайных величин имеет распределение близкое к нормальному.
- Главное: чтобы случайные величины были похожи и не было такого, что одна резко выделяется на фоне остальных
- В классической форме: пусть у нас есть бесконечная последовательность из одинаково распределённых случайных величин, имеющих конечное математическое ожидание и дисперсию, тогда:

$$\frac{(X_1 + \dots + X_n) - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$$

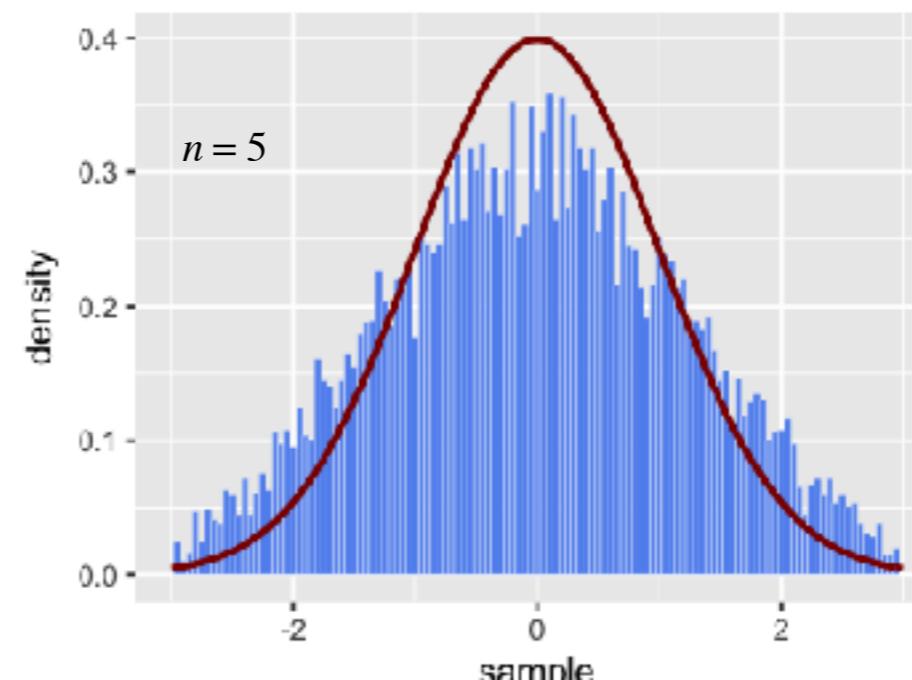
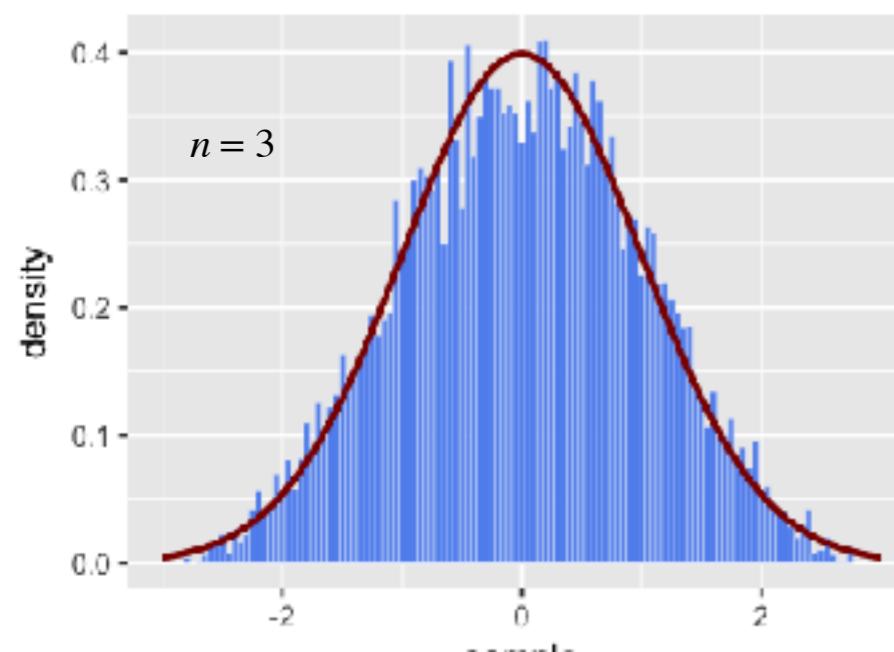
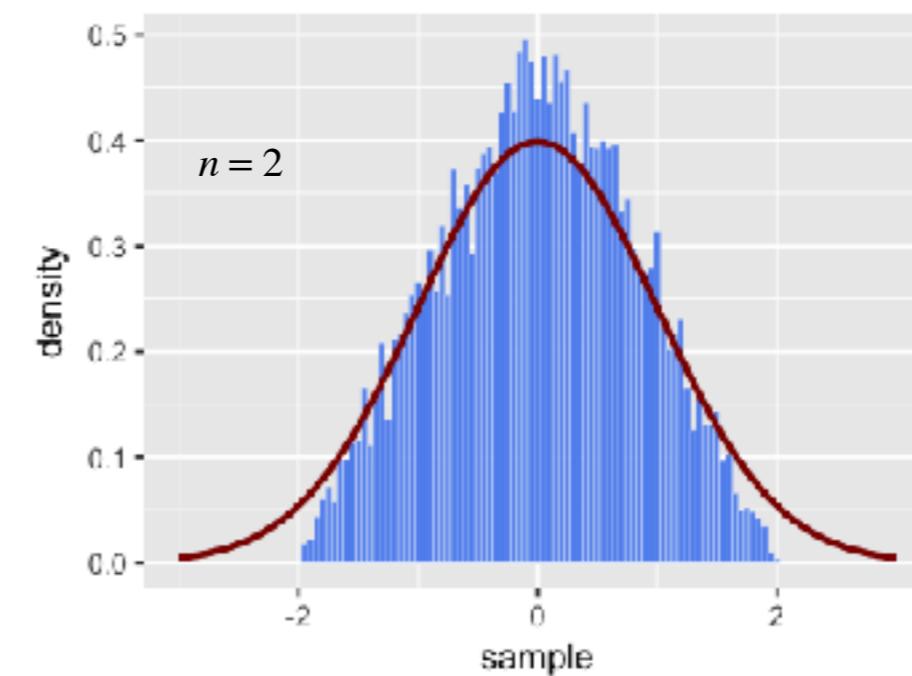
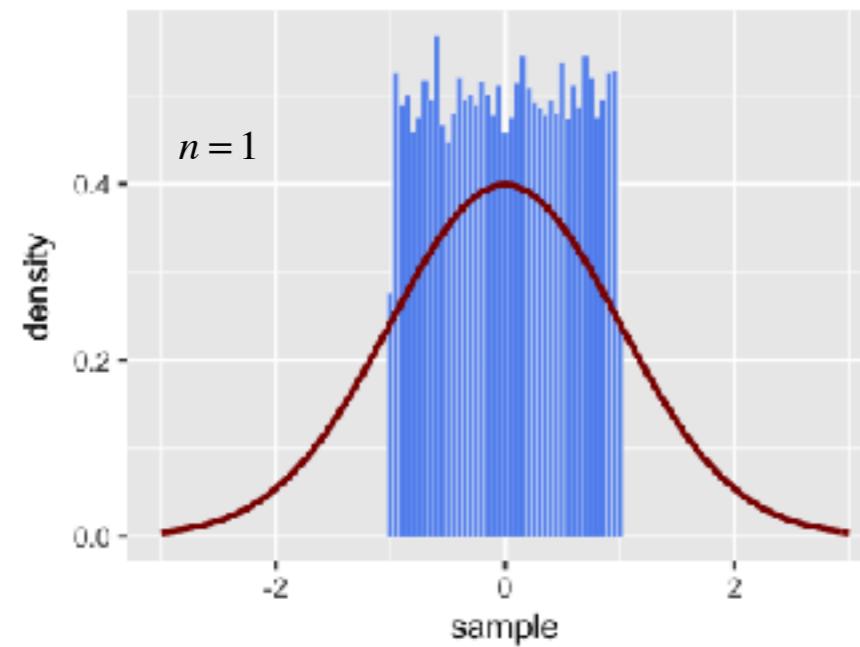
Доска Гальтона

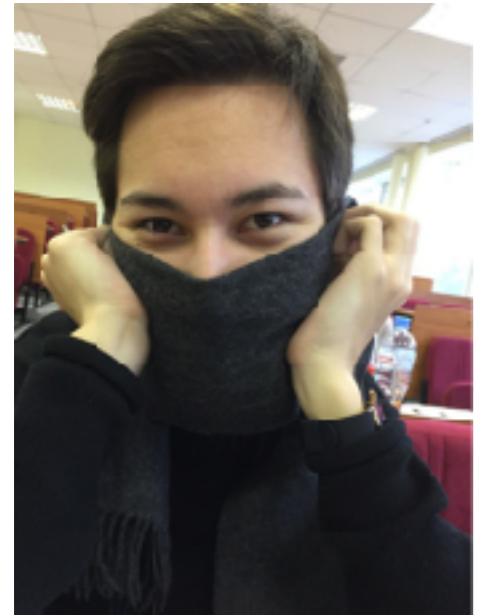


Центральная предельная теорема

$$X \sim U[-1;1]$$

$$Y = X_1 + \dots + X_n$$



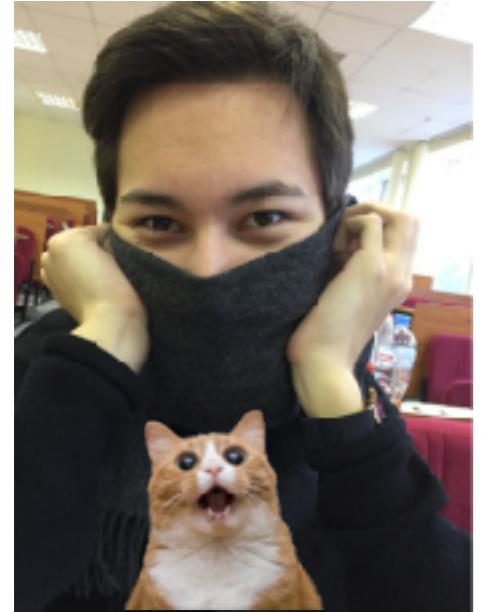


На пальцах

Саша

Время прихода
Саши на первую пару

Y



На пальцах

Саша

Время прихода
Саши на первую пару

Y

На Сашу прыгнул кот и он проснулся пораньше, ускорение на X_1

Пока готовил завтрак, убежало молоко, задержка на X_2

Быстро приехал автобус, ускорение на X_3

Встал в неожиданную пробку, задержка на X_4



На пальцах

Саша

Время прихода
Саши на первую пару

Y

На Сашу прыгнул кот и он проснулся пораньше, ускорение на X_1

Пока готовил завтрак, убежало молоко, задержка на X_2

Быстро приехал автобус, ускорение на X_3

Встал в неожиданную пробку, задержка на X_4

$$Y =$$



+



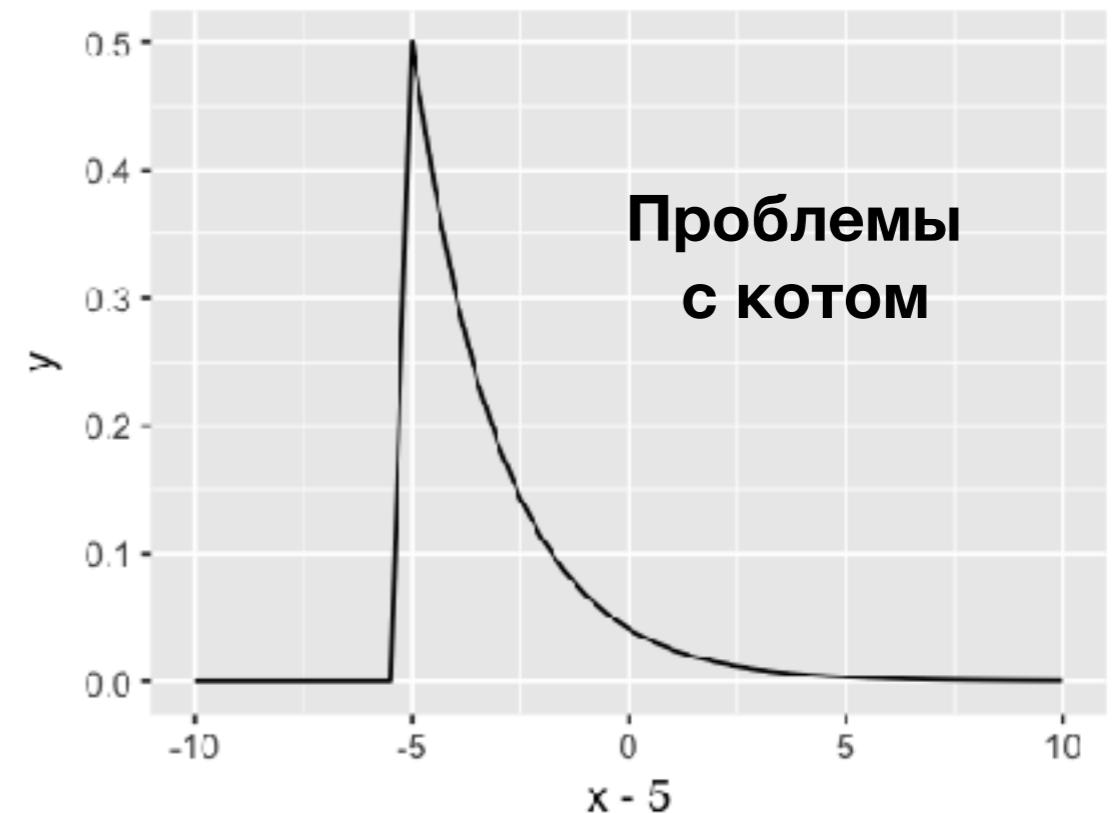
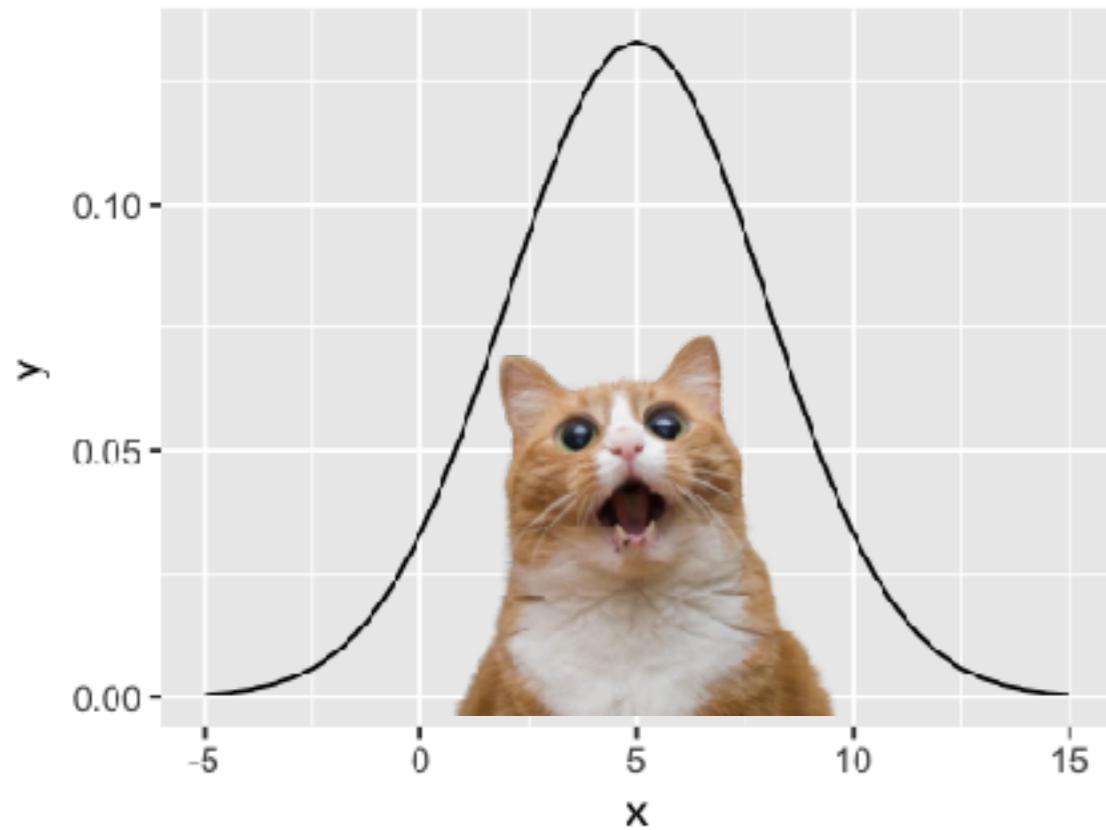
+



+



На пальцах



Если одна из величин выделяется, например кот требует внимание и сильно задерживает Сашу, то всё может сломаться и мы не получим нормального распределения

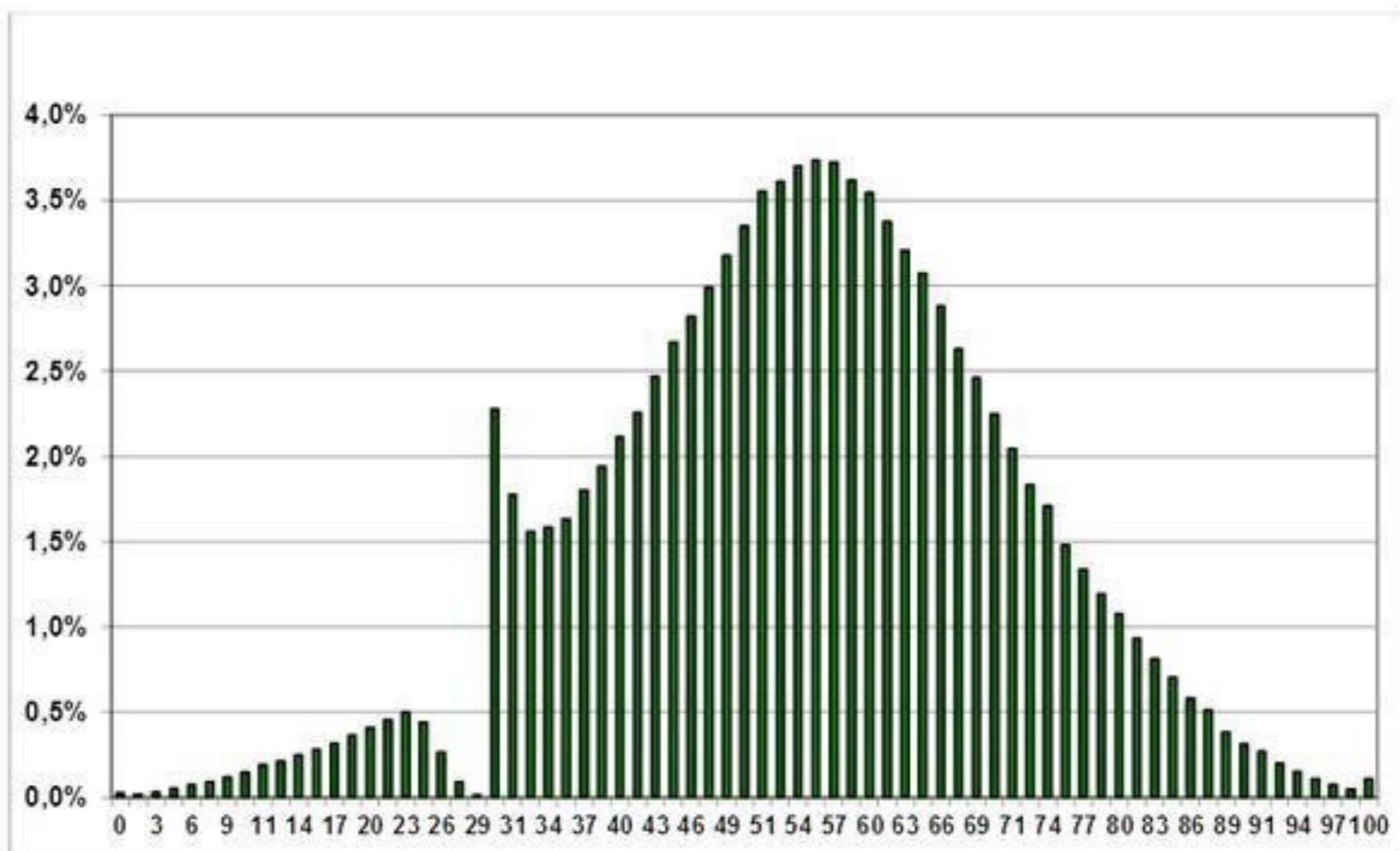
Крайнестан и среднестан

- А что если какая-то одна случайная величина выбивается?
- Тогда мы перемещаемся из среднестана в крайнестан и сталкиваемся с проблемой тяжёлых хвостов
- О тяжёлых хвостах, крайнестане и среднестане мы будем говорить через пару
- Пока что представим, что мы живём в среднестане

Фальсификации

Результаты выпускных экзаменов в Польше

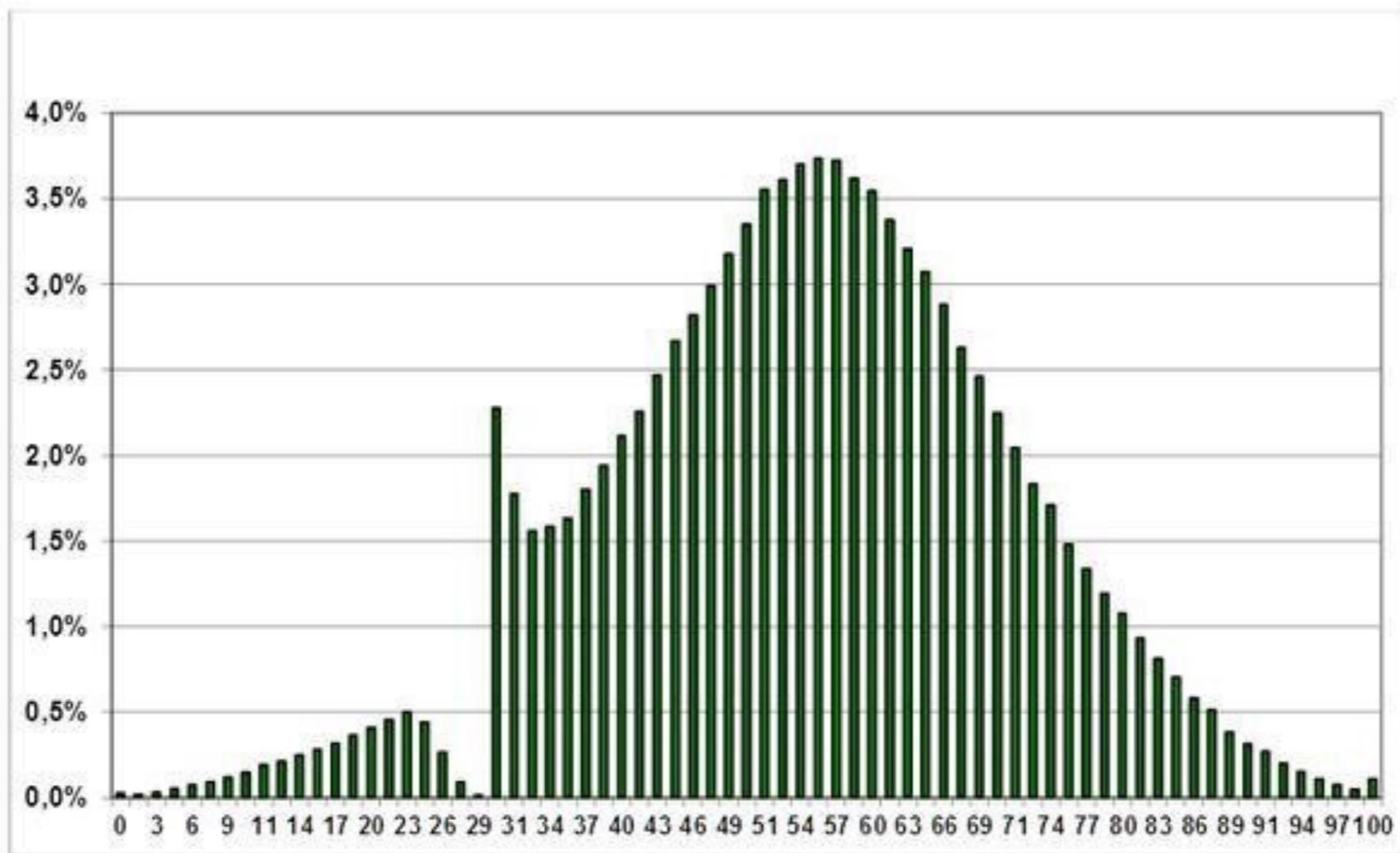
2.1. Poziom podstawowy



- Что здесь не так? Что здесь является случайной величиной?

Результаты выпускных экзаменов в Польше

2.1. Poziom podstawowy



- Подозрительный пик в районе проходного балла
- Подозрительный пик на 100 баллах
- Результаты теста вроде в среднестане

Идея!

- Если мы живём в среднестане, можно попробовать с помощью ЦПТ искать фальсификации!

Явка на выборы

- Находится ли явка на выборы в среднестане?
- Выборы – масштабное явление
- Очевидно, что на некоторых участках явка выше, на некоторых ниже
- Есть участки с явкой в 100%, но их мало
- Есть скрытые факторы, которые могут обломать весь наш анализ, но о них позже

Полезные ссылки

- Сайт центризберкома: <http://www.vybory.izbirkom.ru/region/izbirkom>
 - Код на питоне для сбора данных с сайта ЦИК: <http://nbviewer.jupyter.org/github/FUlyankin/Parsers/blob/master/Parsers\%20/Результаты\%20Выборов/Фальсификации\%2C\%20выборы.ipynb>
 - Статья на хабре, которая учит парсингу на python: <https://habrahabr.ru/company/ods/blog/346632/>
 - Уже готовые датасеты от движения за честные выборы «Голос»: http://els.golosinfo.org/ru/elections?utf8=%E1%&q=%5Beday_day_eq%5D=%E1%&q=%5Beday_month_eq%5D=%E1%&q=%5Beday_year_eq%5D=%E1%&q=%5Bregion_id_eq%5D=%E1%&q=%5Belevel_id_in_any%5D%5B%5D=1%E1%&q=%5Bname_ru_cont%5D=

Данные

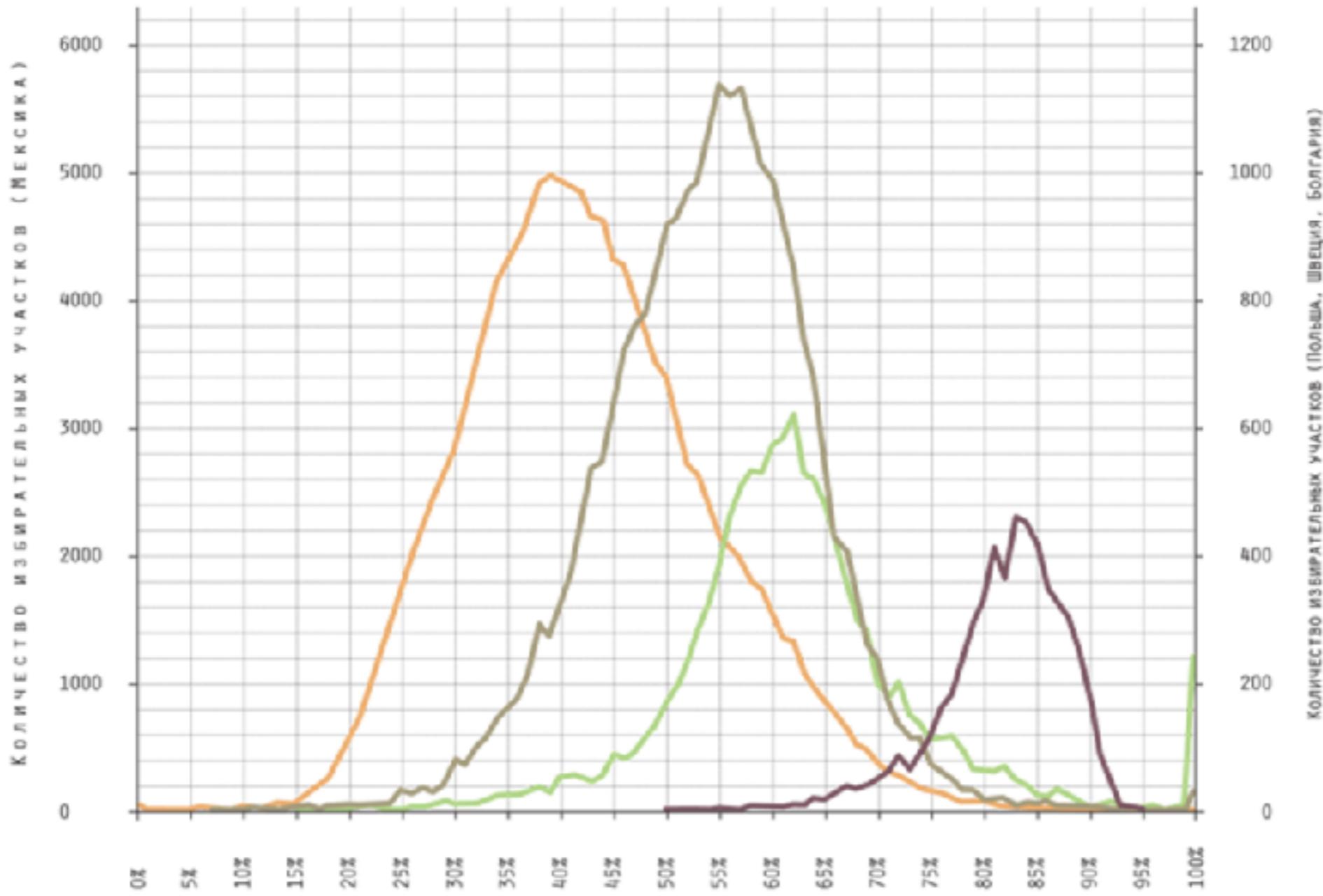
| Имя колонки | Регион | Район | Число избирателей, включенных в список избирателей | Число избирательных бюллетеней, полученных участковой избирательной комиссией | Жириновский Владимир Вольфович (абсолютно) | Зюганов Геннадий Андреевич (абсолютно) | Миронов Сергей Михайлович (абсолютно) | Прохоров Михаил Дмитриевич (абсолютно) | Путин Владимир Владимирович (абсолютно) |
|-------------|--------------------|---------------------------|--|---|--|--|---------------------------------------|--|---|
| УИК №1039 | Московская область | Котельниковская городская | 405 | 330 | 19 | 58 | 10 | 54 | 126 |
| УИК №1040 | Московская область | Котельниковская городская | 2684 | 2275 | 149 | 295 | 70 | 250 | 905 |
| УИК №1041 | Московская область | Котельниковская городская | 2375 | 2105 | 70 | 248 | 60 | 188 | 862 |
| УИК №1042 | Московская область | Котельниковская городская | 2049 | 1815 | 124 | 211 | 47 | 108 | 755 |
| УИК №1043 | Московская область | Котельниковская городская | 2102 | 1890 | 102 | 209 | 56 | 146 | 845 |

Выборы в Москве (дума, 2011)

город Москва



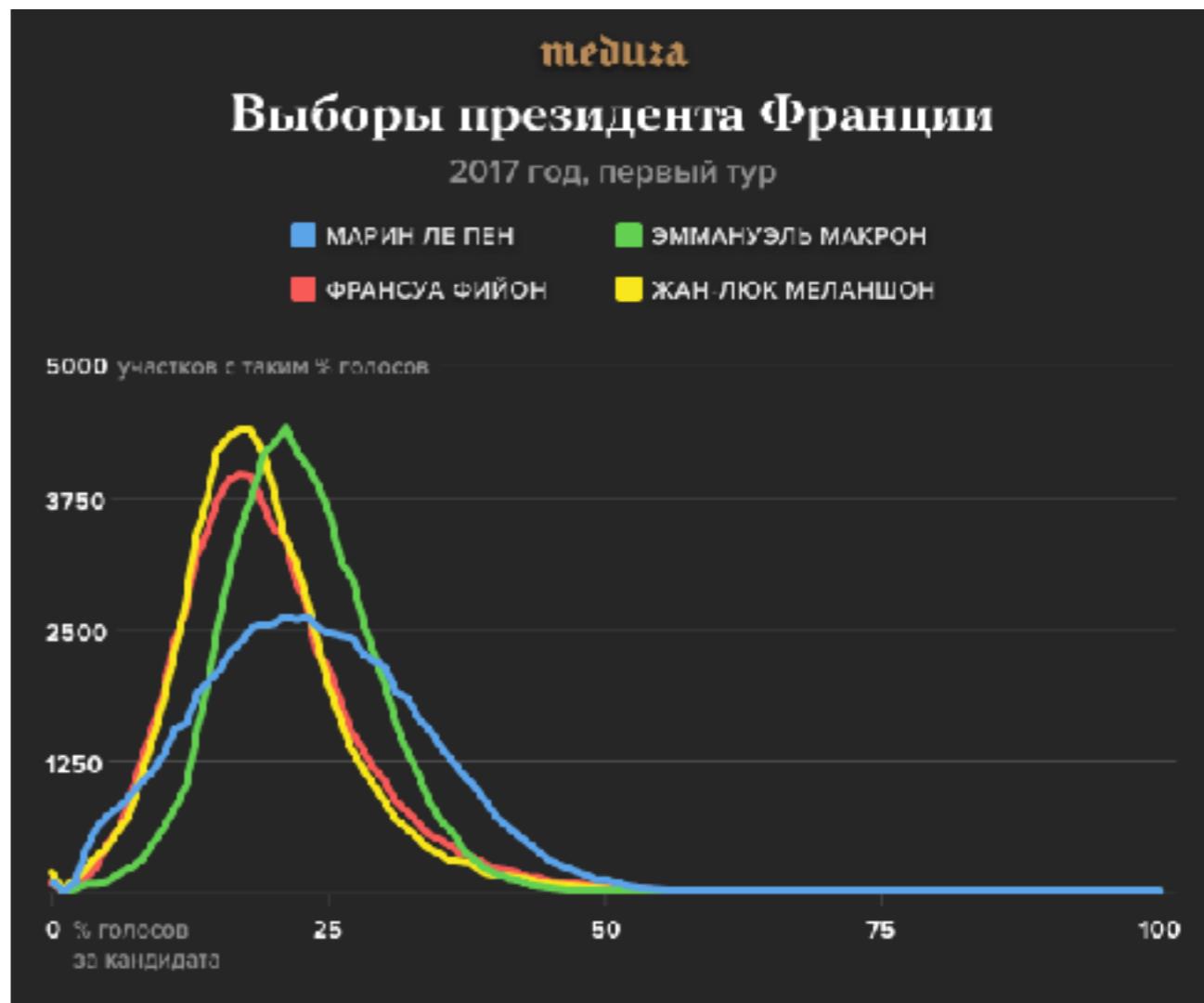
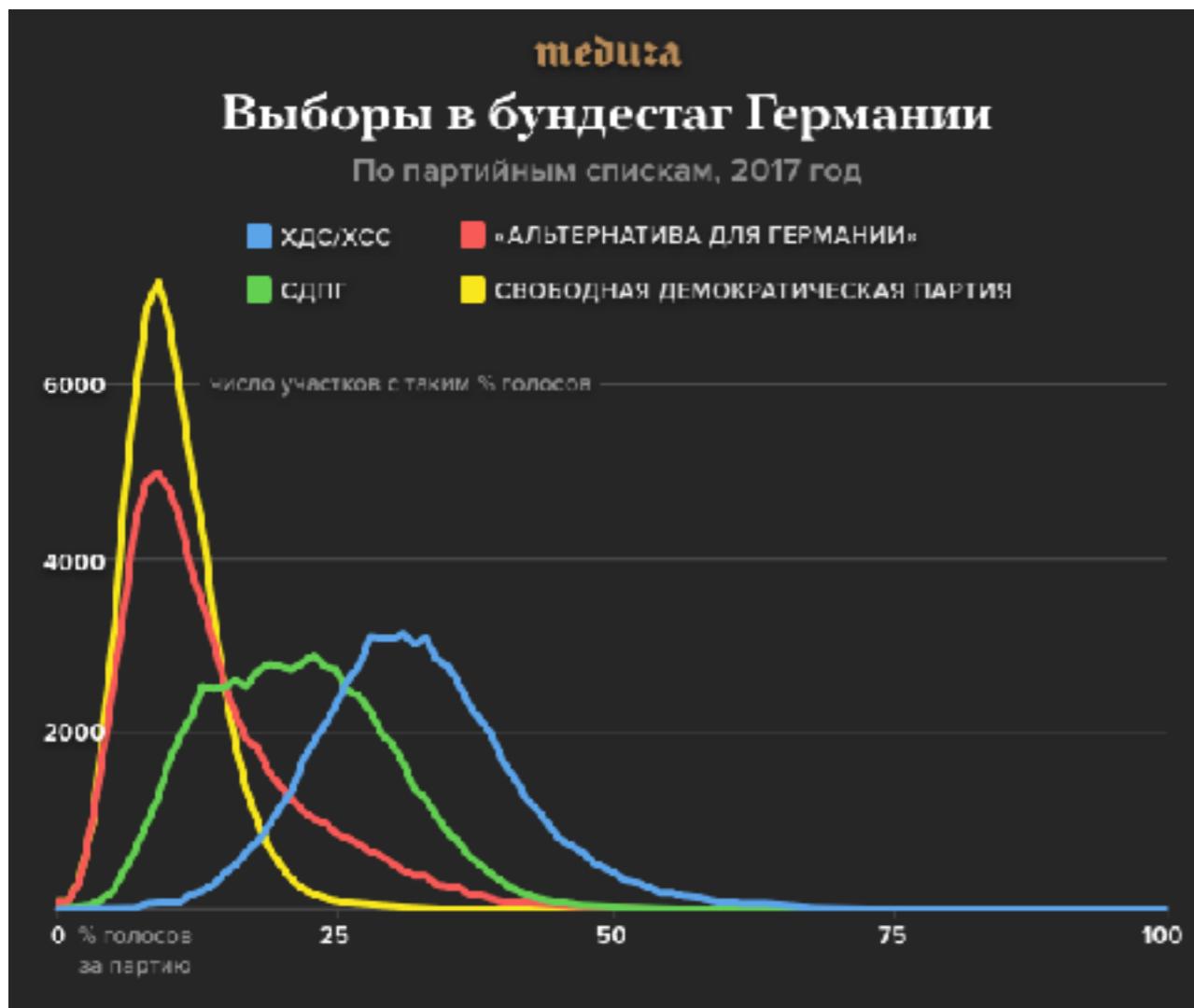
Другие страны



- Мексика, парламентские выборы, 2009 г.
- Польша, II тур выборов президента, 2007 г.
- Болгария, парламентские выборы, 2009 г.
- Швеция, парламентские выборы, 2010 г.

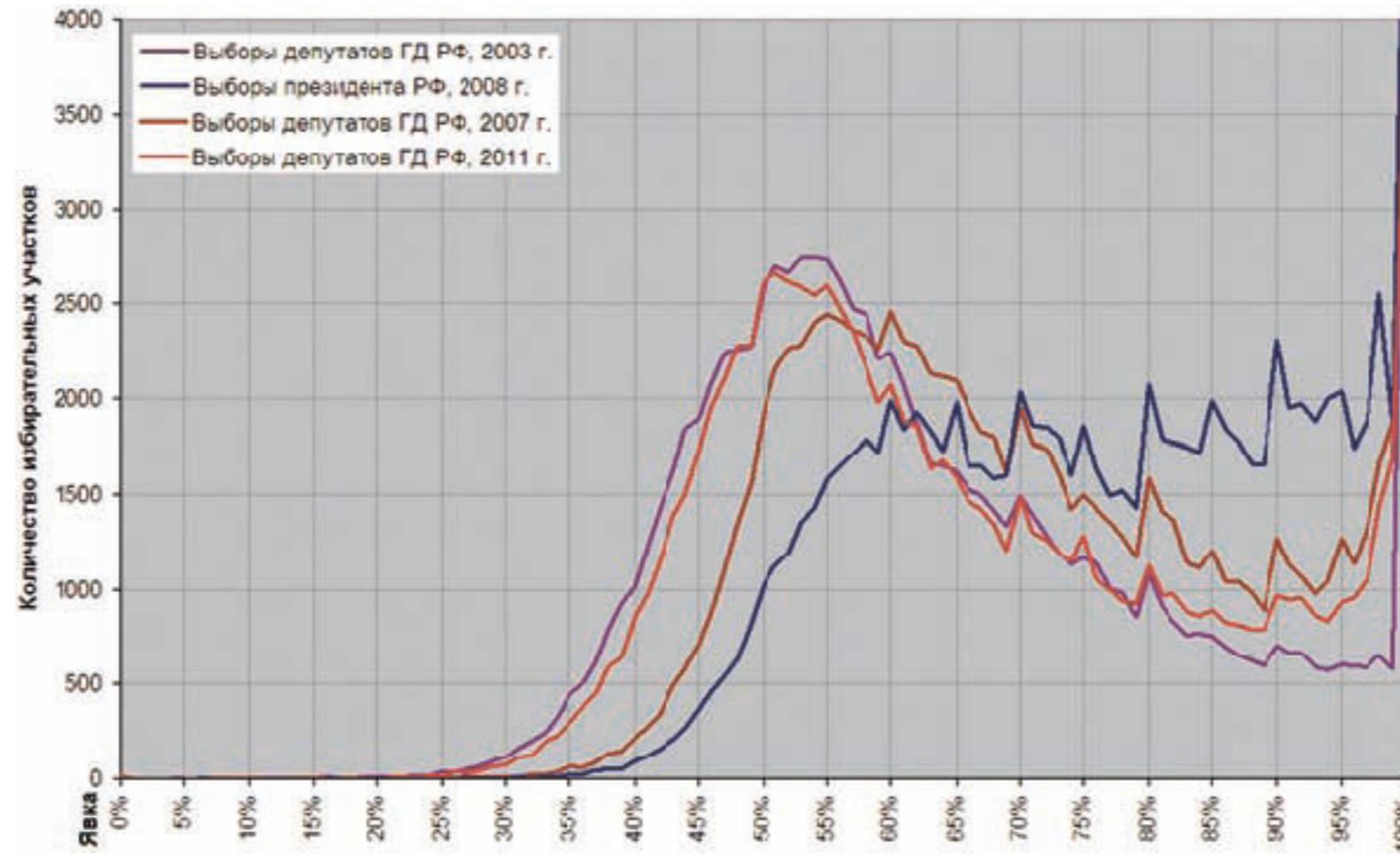
Источник: <https://esquire.ru/articles/2195-elections/>

Другие страны



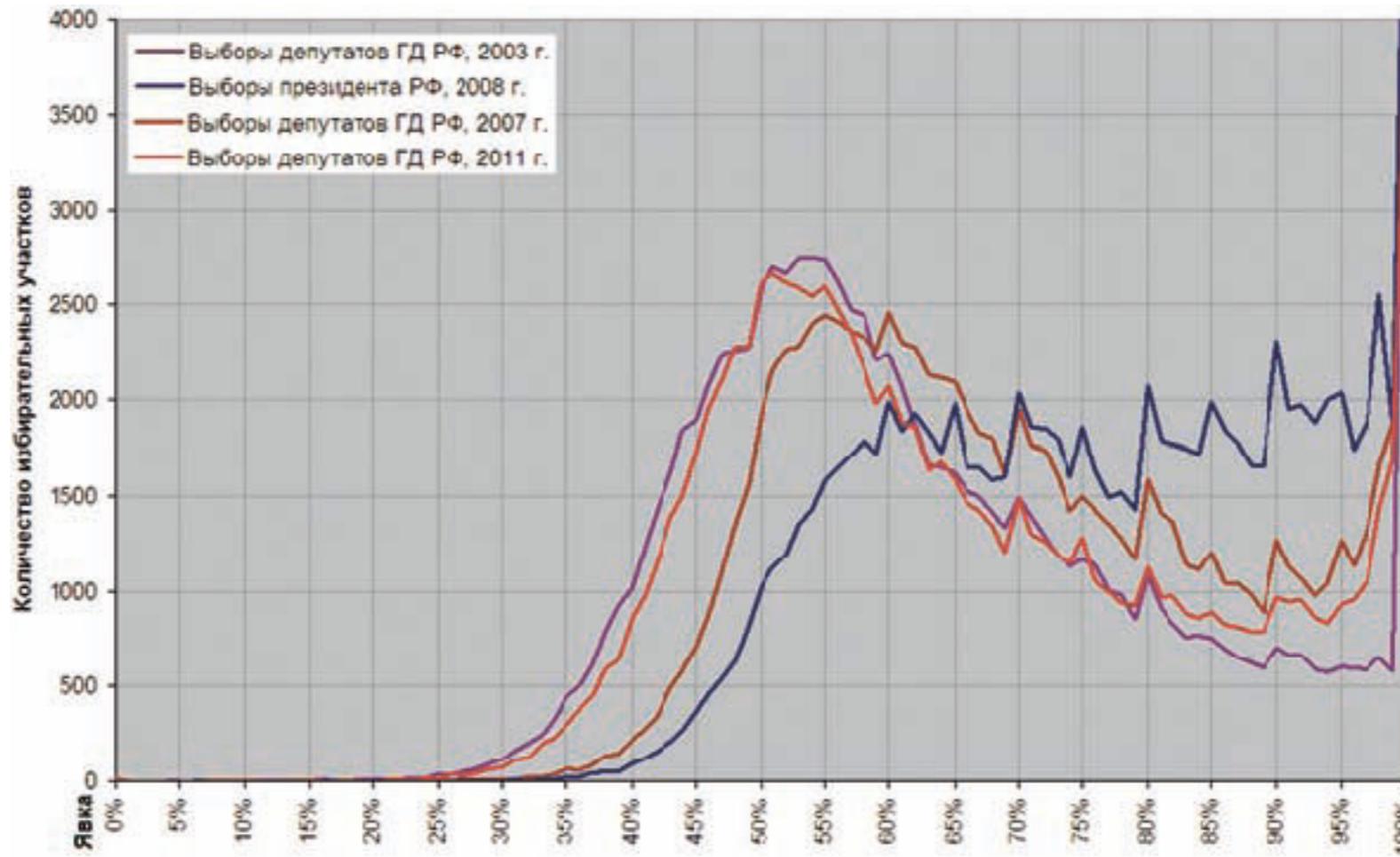
Источник: <https://meduza.io/feature/2018/03/13/tsentrizbirkom-prizyvaet-ne-otsenivat-vybory-po-gaussu-eto-i-pravda-ne-luchshiy-metod>

Выборы в России



- Что здесь не так?

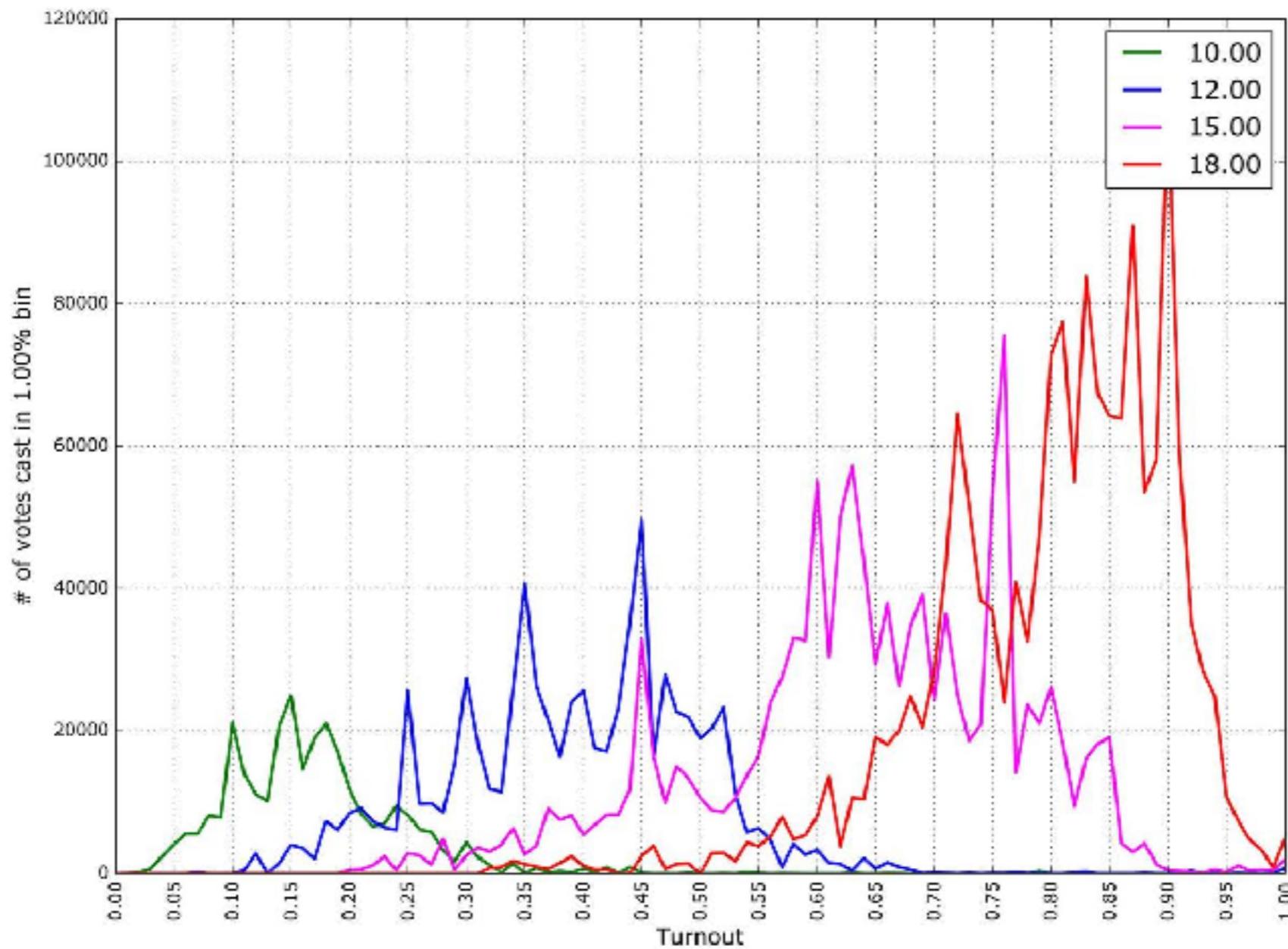
Выборы в России



- Странное распределение явки
- Пики в круглых числах: 55, 60, 65, 70 и тд

Выборы в Кемерово

Кемеровская область

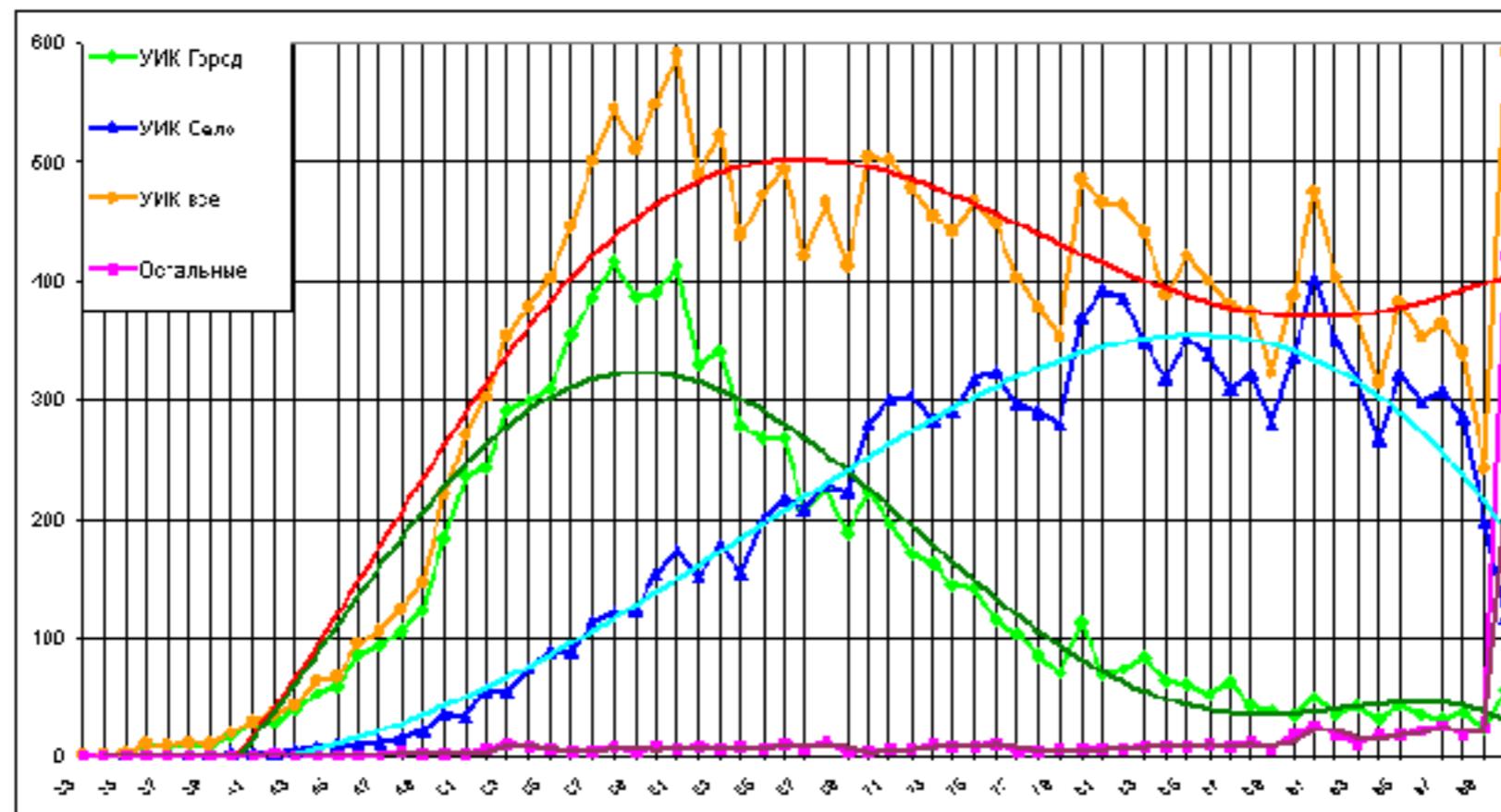


?????

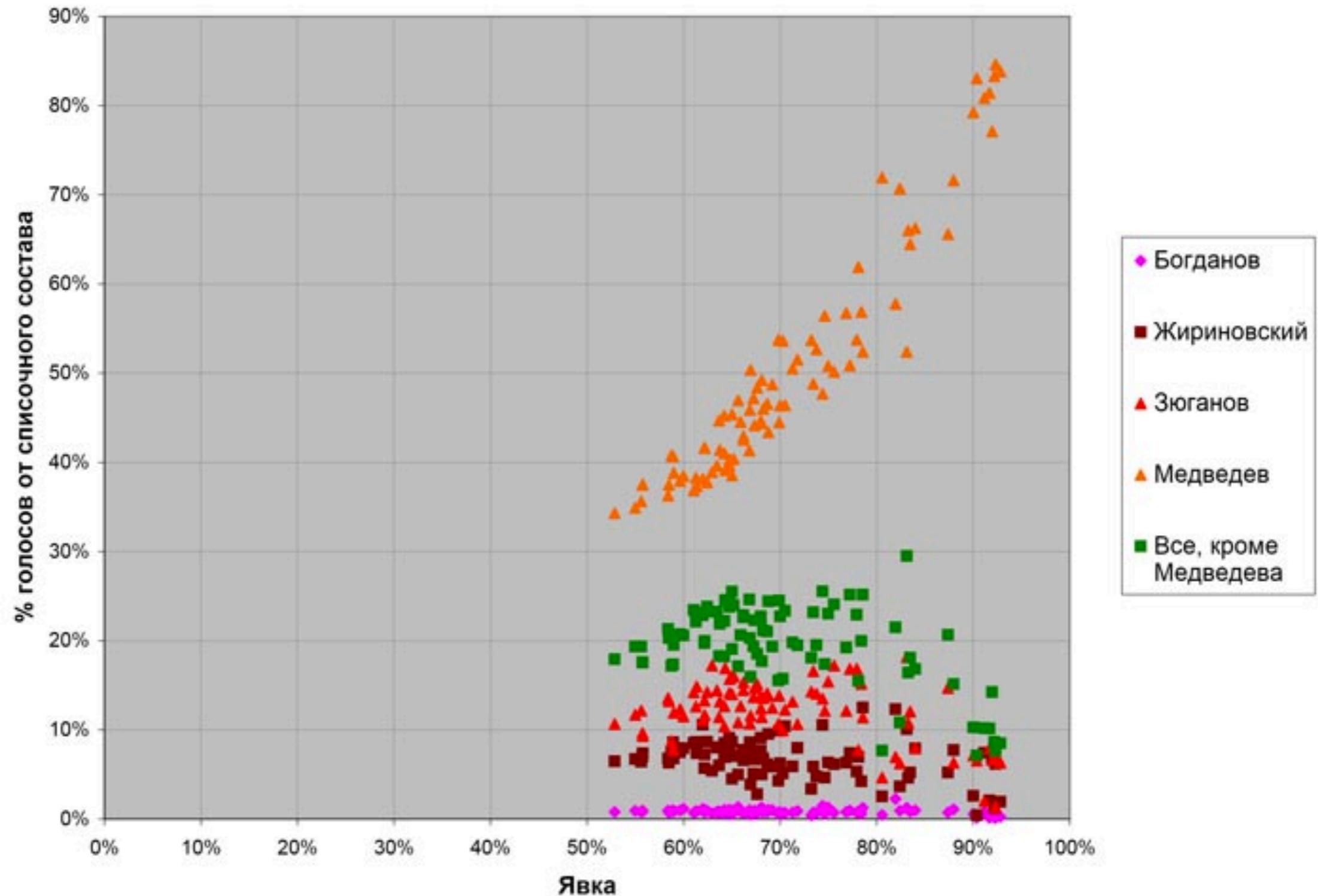
- Голосование противоречит ЦПТ и результатам электоральных исследований по другим странам
- На волеизъявление людей что-то повлияло

Ответ ЦИК

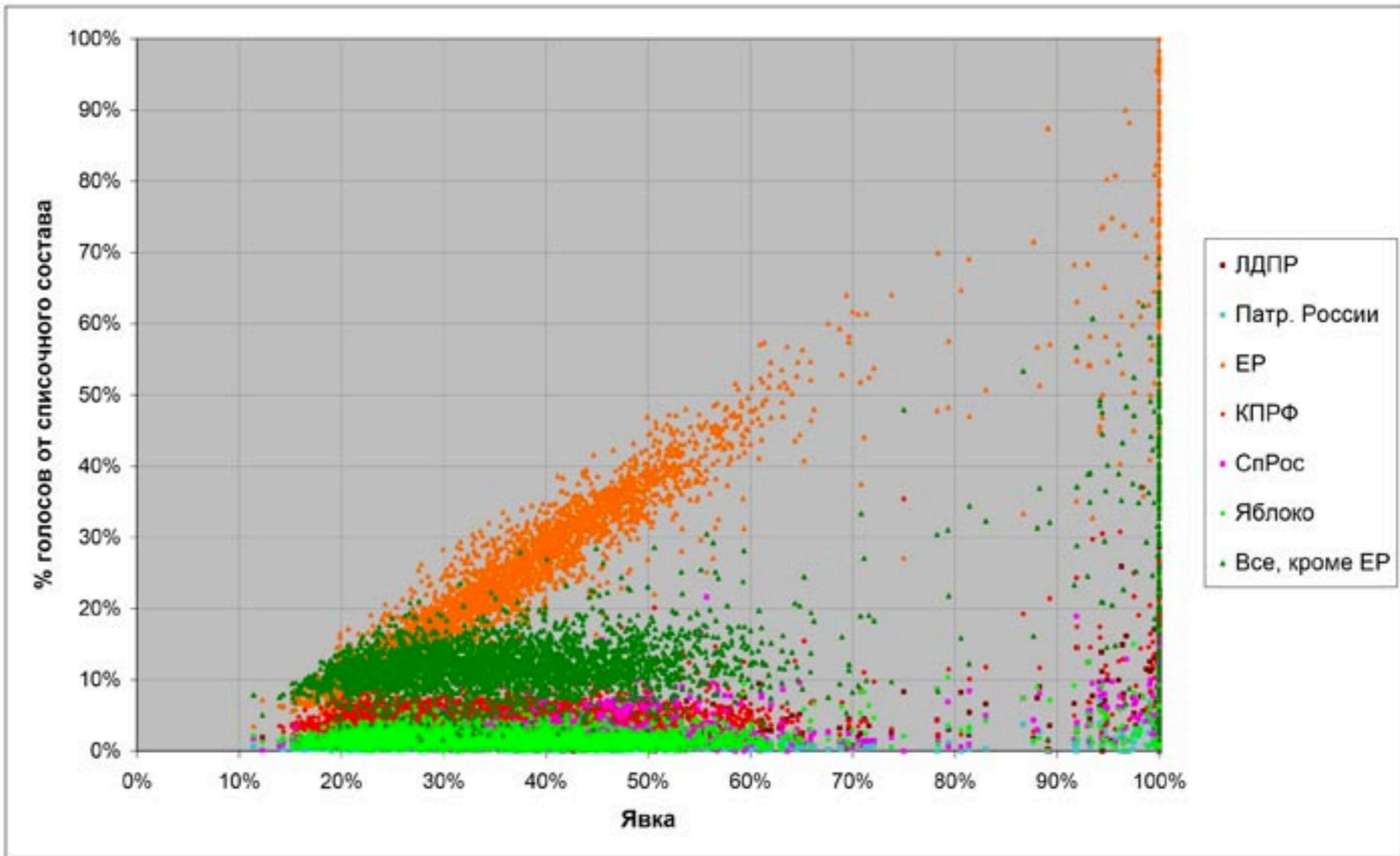
- Множество избирателей неоднородно
 - В сёлах сильна традиция и взгляды населения очень близки
 - Необходимо детализировать подмножества избирателей
 - Источник: <http://cikrf.ru/activity/relevant/detail/29380/>



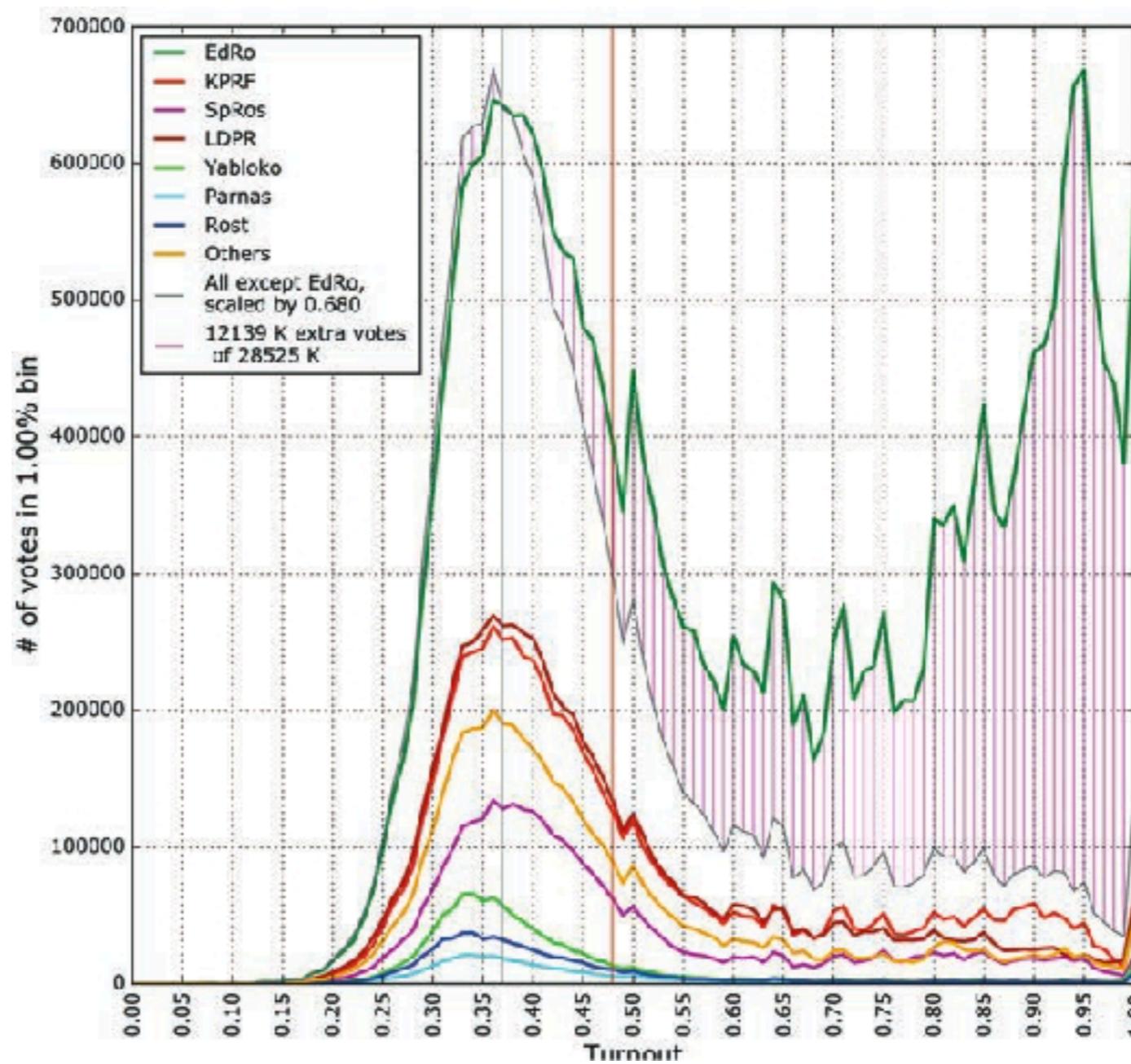
Зависимость результата от явки



Зависимость результата от явки



Попартийные гистограммы



- Природа сторонников власти отличается от природы сторонников других партий

Выводы

- Не факт, что явка на выборы должна быть куполообразной в неоднородных странах.
- Чисто теоретически доля за кандидата может зависеть от явки, но это неточно.
- Тем не менее, декомпозиция по кандидатам и участкам с автоподсчётом голосов могут являться косвенным доказательством наличия вбросов

Вперёд, в R!

Как бы вы замоделировали

- Подбрасывание монетки

Распределение Бернулли

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$$X \sim Bern(p)$$



Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину

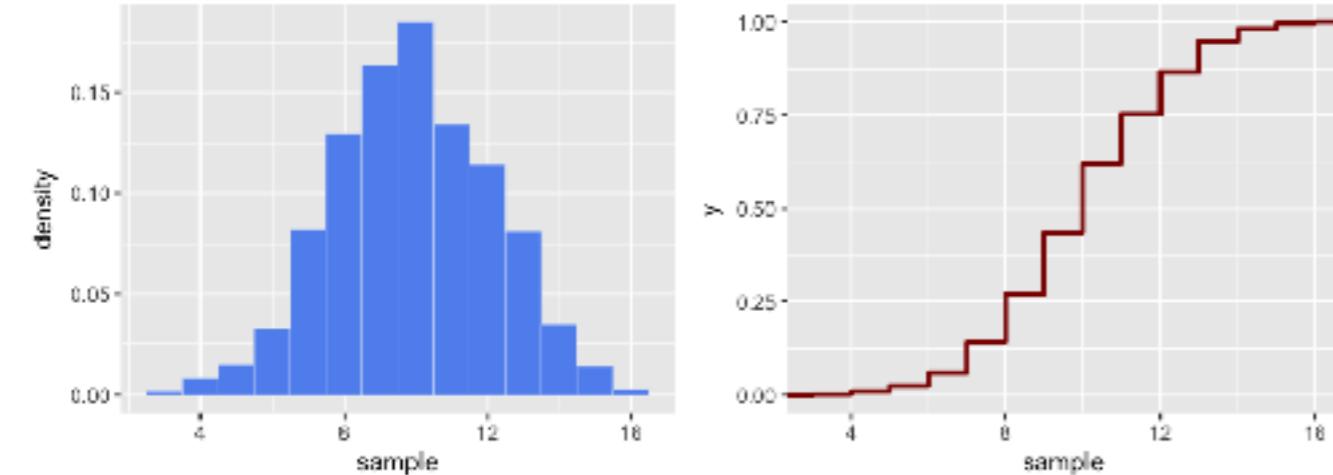
Биномиальное распределение

$$X_i \sim Bern(p)$$

$$Y = X_1 + \dots + X_n$$

$$Y \sim Bin(n, p)$$

$$P(Y = k) = C_n^k (1 - p)^{n-k} p^k$$



Как бы вы замоделировали

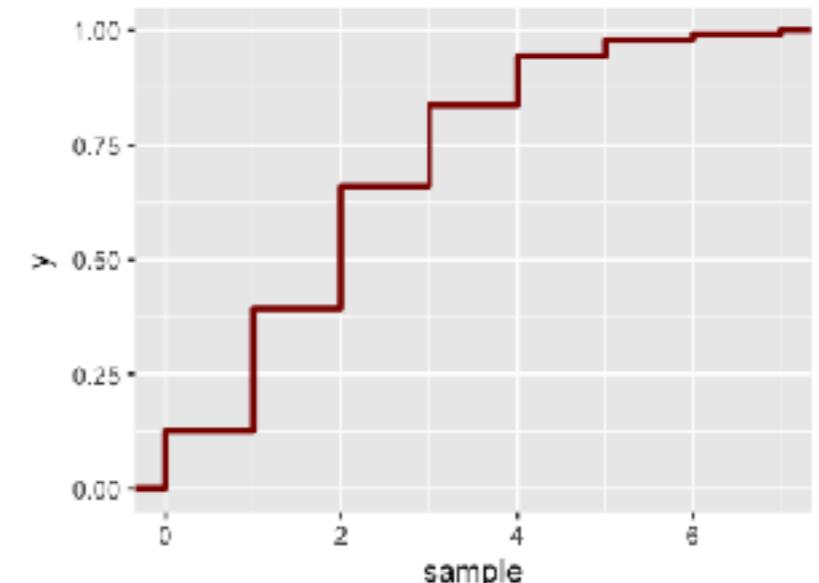
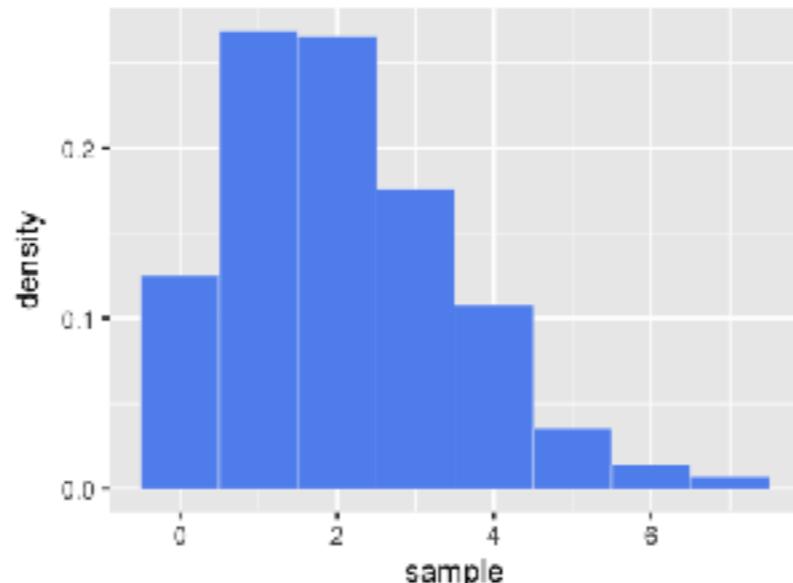
- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди

Распределение Пуассона

$X \sim Poiss(\lambda)$

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$\lambda > 0, k = 1, 2, \dots$



ЗИМА
БЛИЗКО



ВЫСОКИЙ
КАК ЧЕСТЬ



СЕМЬЯ
ДРОГ
ЧЕСТЬ



МЫ
НЕ
СЕEM



УСЛЫШЬ
МОЙ РЁВ!



НАМ
ЯРОСТЬ



ВЫРАСТАЯ
КРЕПНЕМ



НЕ ПРОЛОЖЕННЫЕ
НЕ СОПЛЮЩИЕСЯ
НЕ СЛОМАННЫЕ



ПЛАМЯ И
КРОВЬ



Нравится 15

Свойства распределения Пуассона

- Часто используется для моделирования случайных величин-счётчиков. Особенно редких счётчиков.
- Параметр λ – интенсивность потока событий

$$X \sim Pois(\lambda)$$

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

$$X_i \sim Pois(\lambda_i)$$

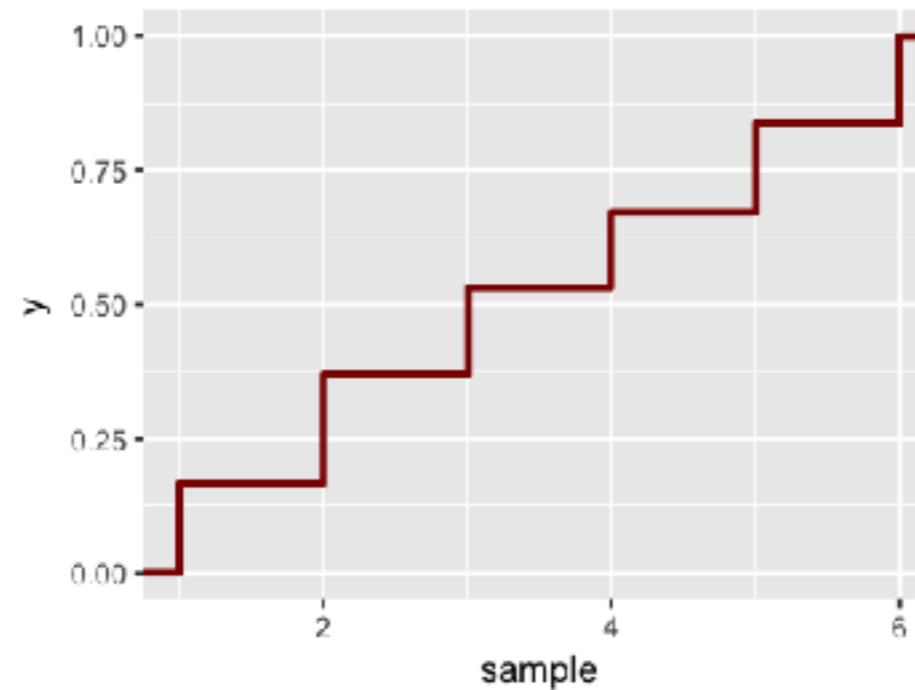
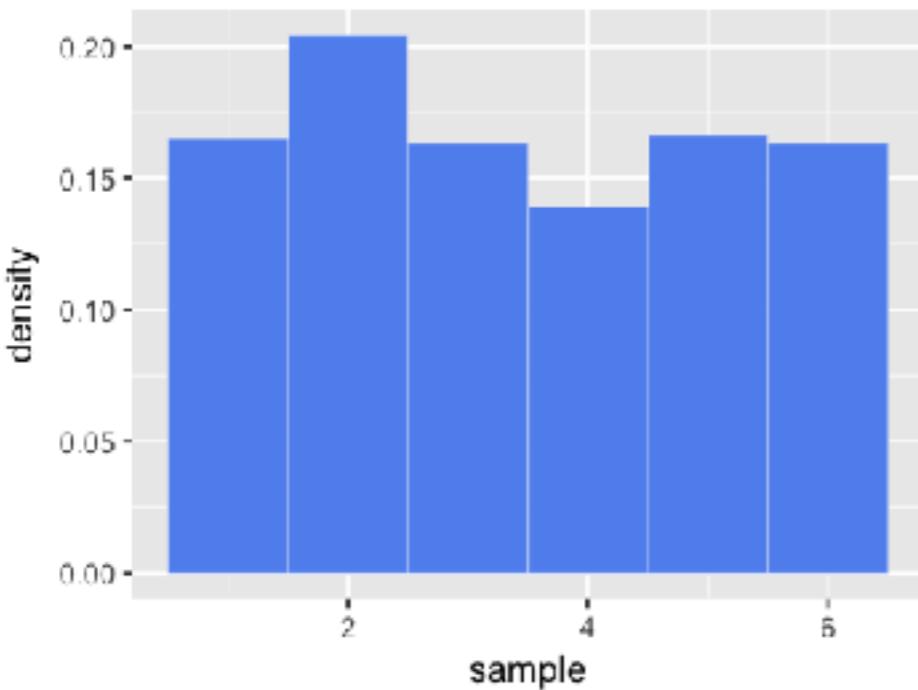
$$X_1 + \dots + X_n \sim Poiss(\lambda_1 + \dots + \lambda_n)$$

Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости

Произвольное дискретное распределение

| X | 1 | 2 | 3 | 4 | 5 | 6 |
|--------|-----|-----|-----|-----|-----|-----|
| P(X=k) | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |



Как бы вы замоделировали

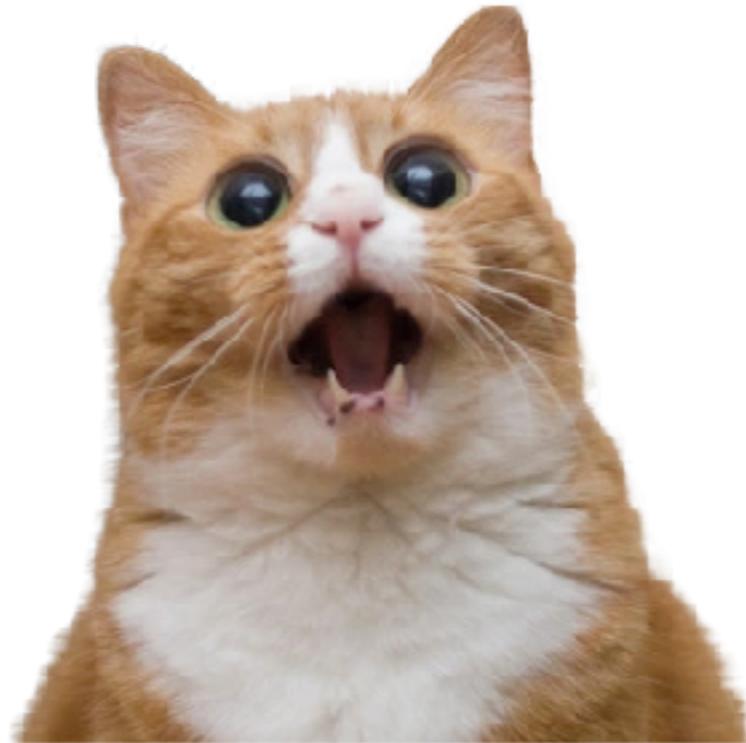
- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке

Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке
- Точное время прихода Саши Сидорова на первую пару

Вырожденное распределение

| | |
|----------|---------|
| X | Никогда |
| $P(X=k)$ | 1 |



Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке
- Точное время прихода Саши Сидорова на первую пару
- Погрешность весов

Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке
- Точное время прихода Саши Сидорова на первую пару
- Погрешность весов
- Время ожидания трамвая

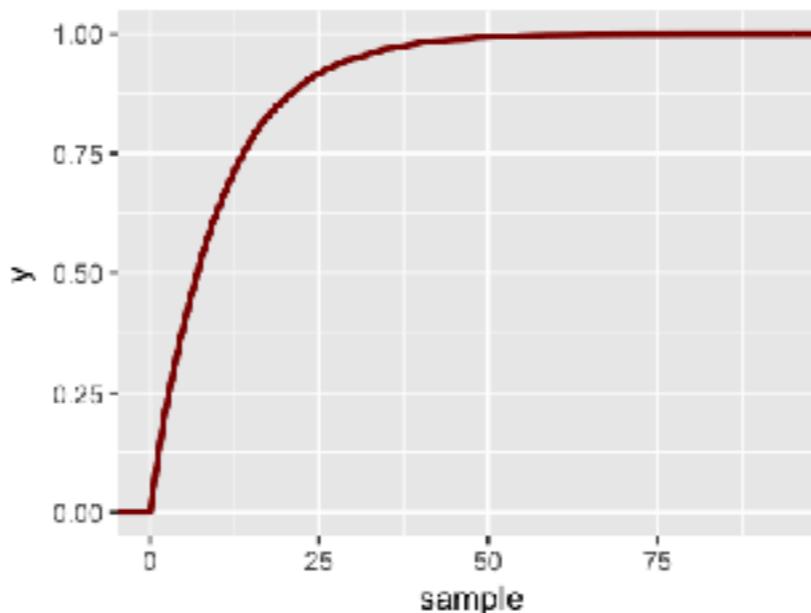
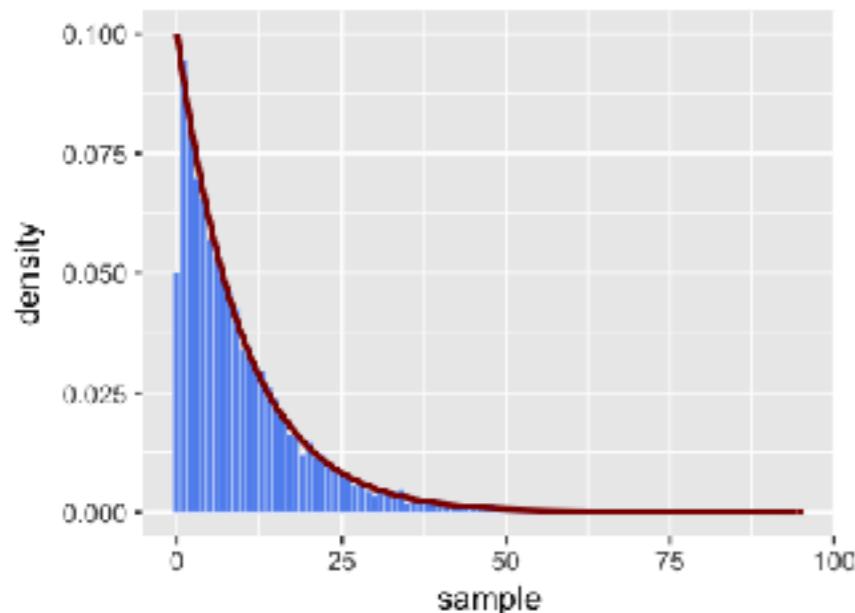
Экспоненциальное распределение

$$X \sim Exp(\alpha)$$

$$F_X(x) = 1 - e^{-\alpha x}, \quad x \geq 0$$

$$E(X) = \frac{1}{\alpha}$$

$$Var(X) = \frac{1}{\alpha^2}$$



Свойства экспоненциального распределения

- У экспоненциального распределения нет памяти

$$X \sim Exp(\lambda) \Rightarrow P(X > s + t \mid X \geq s) = P(X > t)$$

- Пример: пусть автобусы приходят на остановку случайно, но с некоторой фиксированной средней интенсивностью. Тогда количество времени, уже затраченное пассажиром на ожидание автобуса, не влияет на время, которое ему ещё придётся простоять.

$$X_i \sim Exp(\alpha_i)$$

$$\min X_i \sim Exp(\alpha_1 + \dots + \alpha_n)$$

$$Exp(\frac{1}{2}) = \chi^2_2$$

Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке
- Точное время прихода Саши Сидорова на первую пару
- Погрешность весов
- Время ожидания трамвая
- Время до прихода в очередь нового человека
- Время до поломки часов

Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке
- Точное время прихода Саши Сидорова на первую пару
- Погрешность весов
- Время ожидания трамвая
- Время до прихода в очередь нового человека
- Время до поломки часов
- Выборочное среднее выборки объёма 100

Как бы вы замоделировали

- Подбрасывание монетки
- Число попаданий в баскетбольную корзину
- Количество людей в очереди
- Результат подбрасывания игральной кости
- Количество изюма в булочке
- Точное время прихода Саши Сидорова на первую пару
- Погрешность весов
- Время ожидания трамвая
- Время до прихода в очередь нового человека
- Время до поломки часов
- Выборочное среднее выборки объёма 100
- Время рождения ребёнка

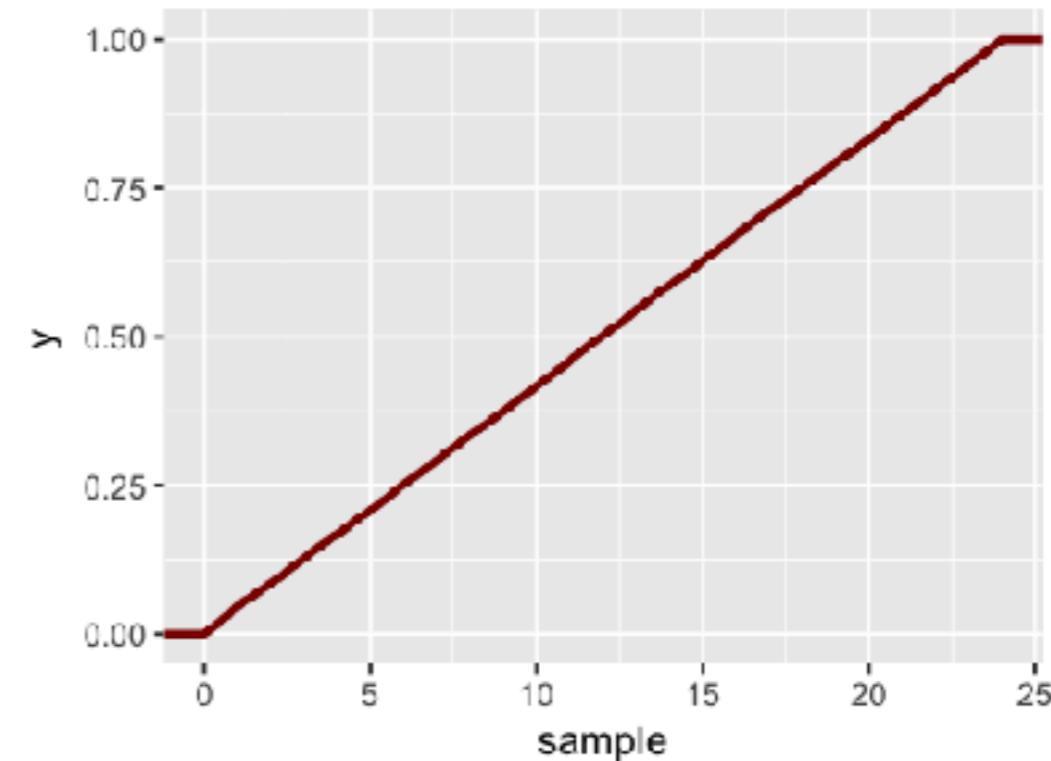
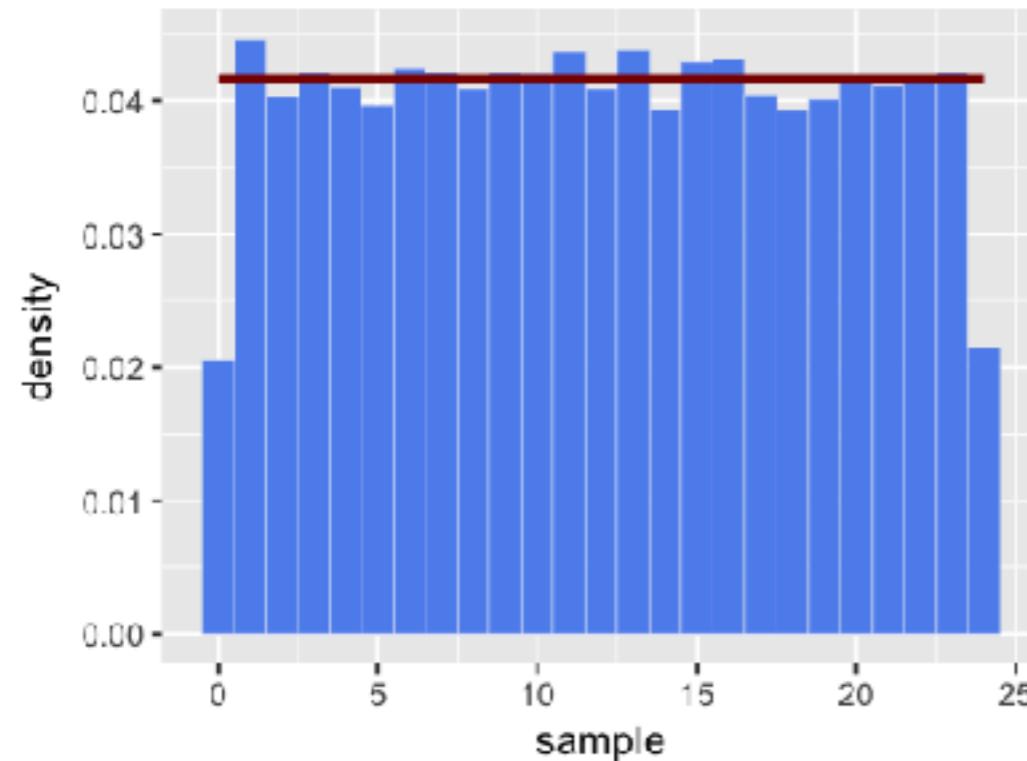
Равномерное распределение

$$X \sim U[a; b]$$

$$F_X(x) = \frac{x - a}{b - a}, \quad x \in [a; b]$$

$$E(X) = \frac{a+b}{2}$$

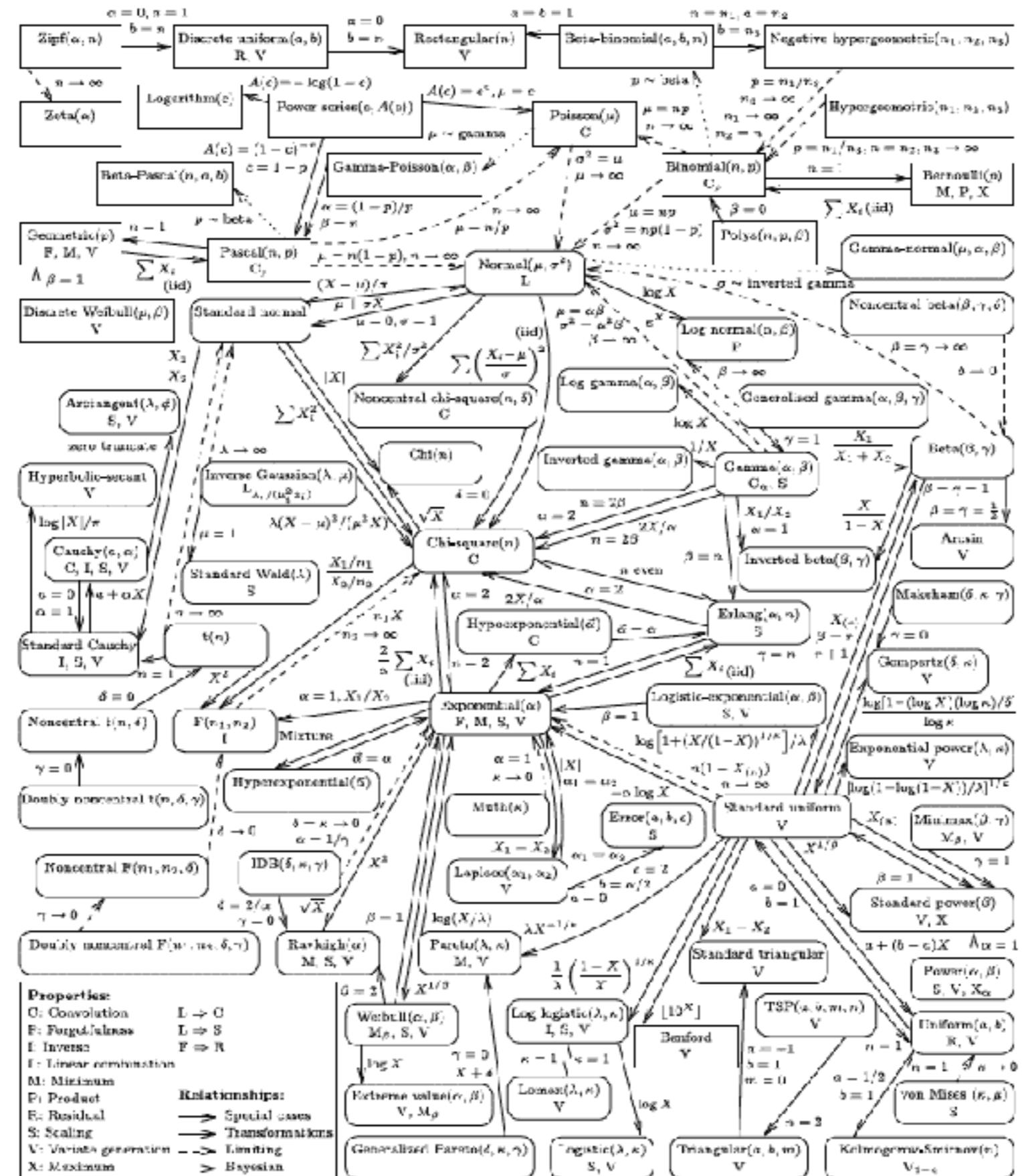
$$\text{Var}(X) = \frac{(b-a)^2}{12}$$



Наши примерные ответы, с которыми можно поспорить!

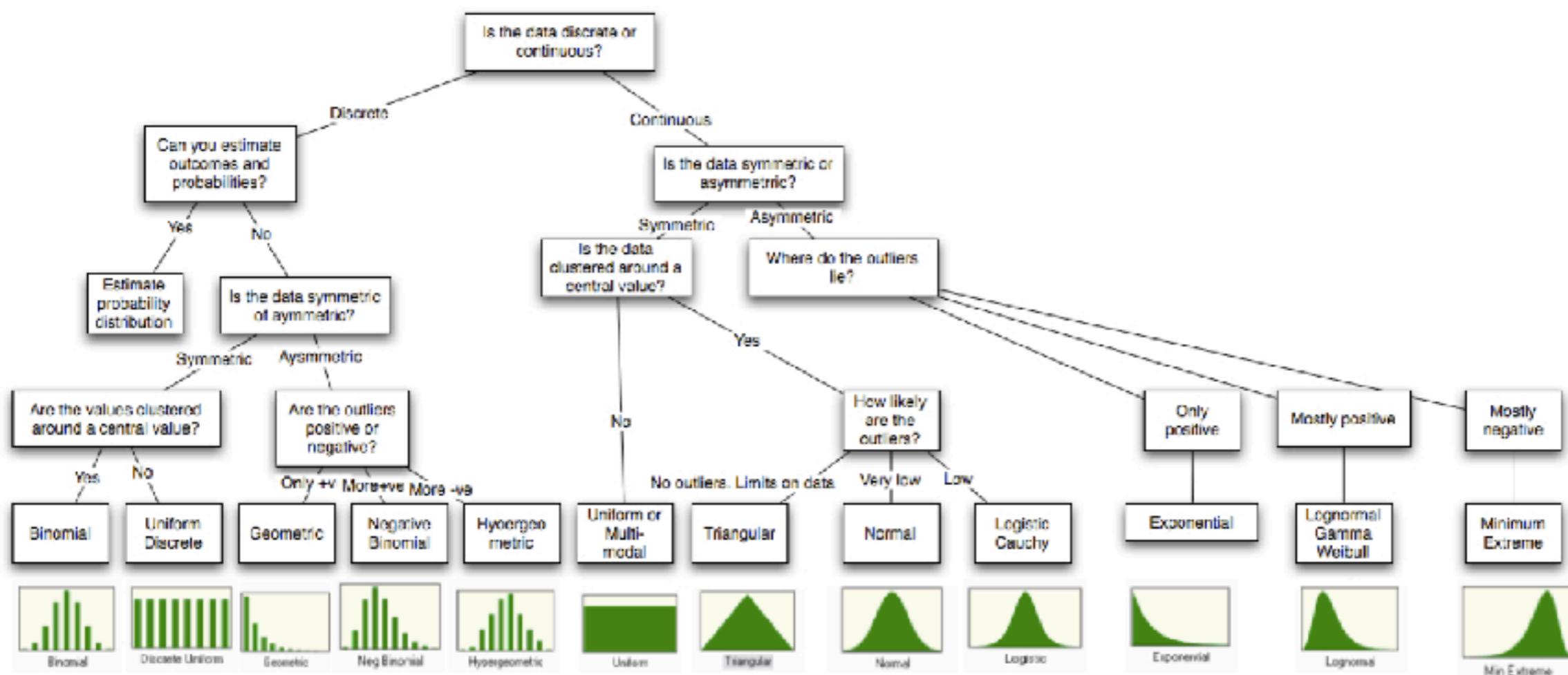
- Подбрасывание монетки $Bern(p)$
- Число попаданий в баскетбольную корзину $Binom(n, p)$
- Количество людей в очереди $Pois(\lambda)$
- Результат подбрасывания игральной кости $:)$
- Количество изюма в булочке $Binom(n, p)$
- Точное время прихода Саши Сидорова на первую пару $P(Never) = 1$
- Погрешность весов $N(0, \sigma^2)$
- Время ожидания трамвая $Exp(\alpha)$
- Время до прихода в очередь нового человека $Exp(\alpha)$
- Время до поломки часов $Exp(\alpha)$
- Выборочное среднее выборки объёма 100 $N(\mu, \sigma^2)$
- Время рождения ребёнка $U[0; 24]$

Связь распределений



Безумные рекомендации

Figure 6A.15: Distributional Choices



Игра в среднее

- Осмотритесь. Вокруг вас много лиц. Некоторые милые, некоторые бандитские.
- Ваша задача: угадать их мысли.
- Каждый участник загадывает число в диапазоне от 0 до 100 и пишет его на бумажке.
- После того как все сдадут свои бумажки, ищется среднее от этих чисел. Оно умножается на 2/3. Это итоговый ответ.
- Тот, кто написал наиболее близкое к итогу число получает приз.

We Can Do It!



Jeanne Miller

POST FEB. 15 TO FEB. 28



WAR PRODUCTION CO-ORDINATING COMMITTEE