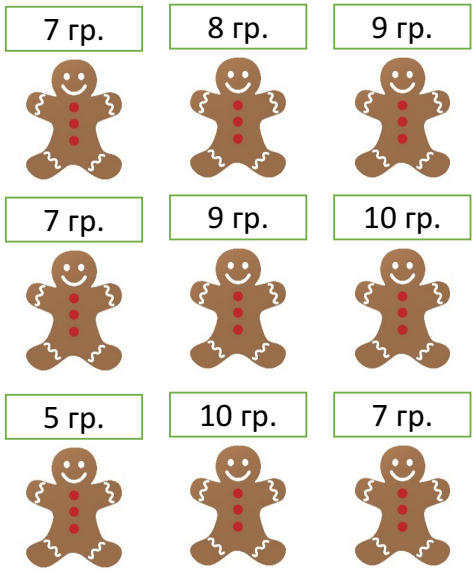
Качественные задачи
1
«Про данные»
лекция 2

Иван Станкевич,
НИУ ВШЭ

Показатели центра распределения

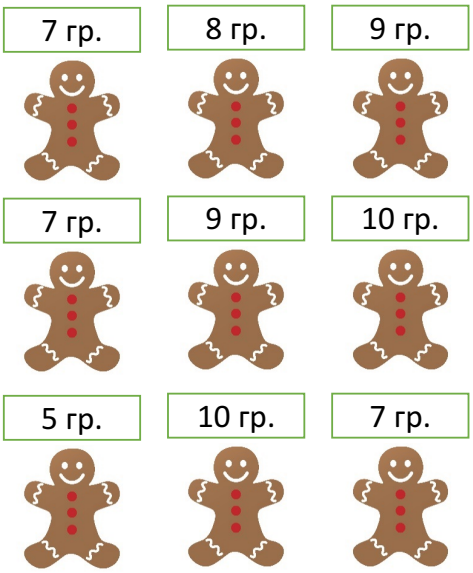
- Среднее – среднее арифметическое



$$\frac{7+7+5+8+9+10+9+10+7}{9}=\frac{72}{9}=8$$

Показатели центра распределения

- Мода – самое частое значение в выборке

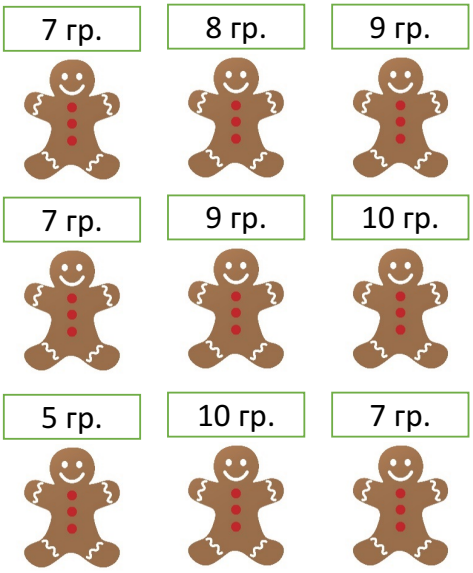


Значение	Частота
5	1
7	3
8	1
9	2
10	2

Мода!

Показатели центра распределения

- Медиана – такое число, что половина элементов выборки меньше него, а половина – больше



Упорядочим наблюдения по возрастанию

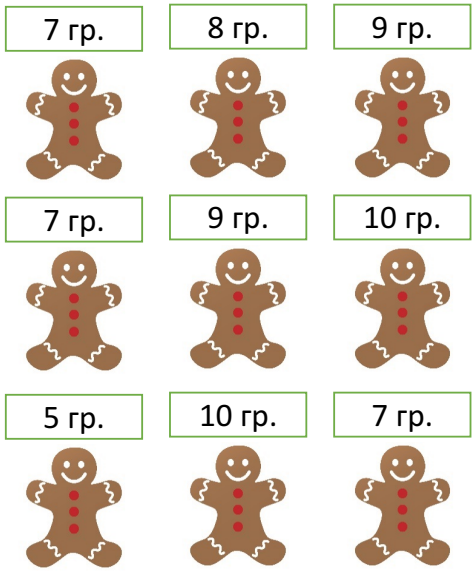
5 гр.
7 гр.
7 гр.
7 гр.
8 гр. — Медиана!
9 гр.
9 гр.
10 гр.
10 гр.

Показатели центра распределения

- При **большой неоднородности данных, медиана и среднее** могут дать очень сильно **различающиеся результаты**
- К примеру: **100 человек** с доходом около **10 т.р.** и **10 человек** с доходом около **100 т.р.**
- **Среднее** около **20** т.р.
- **Медиана** около **10** т.р.

Показатели неоднородности

- Размах – разность между максимальным и минимальным значением в выборке



10 гр. - максимальное

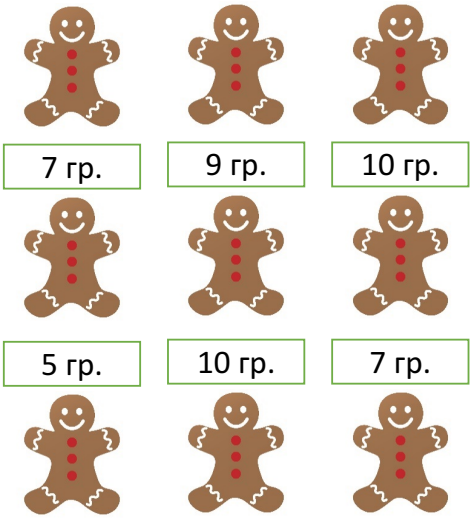
5 гр. - минимальное

Размах = 10 – 5 = 5

Показатели неоднородности

- Стандартное отклонение $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

7 гр.	8 гр.	9 гр.
7 гр.	9 гр.	10 гр.
5 гр.	10 гр.	7 гр.



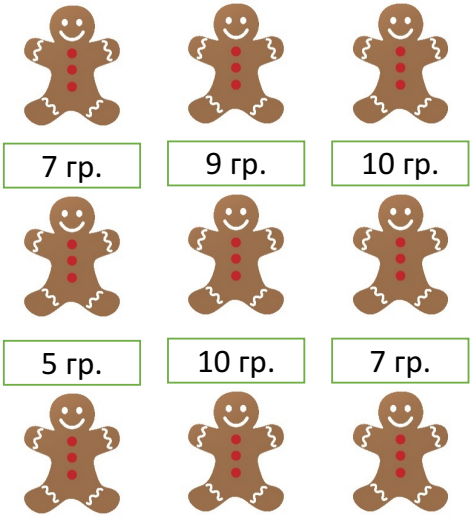
$$\sqrt{\frac{\left((7-8)^2+(7-8)^2+(6-8)^2+(8-8)^2+(9-8)^2+\right.}{\left.+(11-8)^2+(8-8)^2+(9-8)^2+(7-8)^2\right)}{9}}=$$

$$=\sqrt{\frac{(1+1+4+0+1+9+0+1+1)}{9}}=\sqrt{\frac{18}{9}}=\sqrt{2}\approx 1.4$$

Показатели неоднородности

- Коэффициент вариации – стандартное отклонение, деленное на среднее

7 гр.	8 гр.	9 гр.
7 гр.	9 гр.	10 гр.
5 гр.	10 гр.	7 гр.



1.4 – стандартное отклонение

8 среднее

Коэф. Вариации = 1.4/8 = 0.175

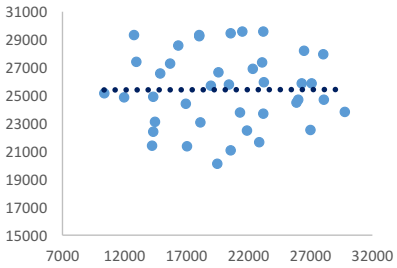
Корреляция

- Мера **линейной** связи между **двумя** показателями
- Рассчитывается по формуле:

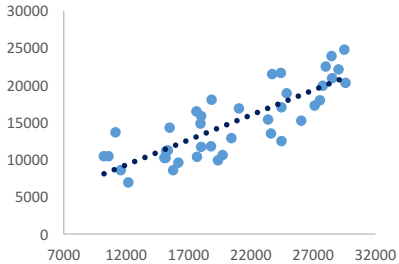
$$r_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

- Изменяется в пределах **от -1 до +1**
- Корреляция, равная **-1**, говорит об **отрицательной линейной связи** между переменными, **+1 – положительной**.
- **Нулевая** корреляция говорит об **отсутствии связи**.

Корреляция

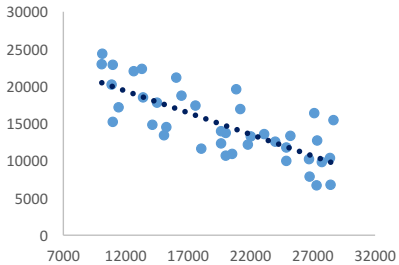


Корреляция около 0

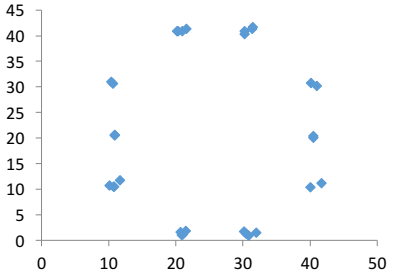


Корреляция +0.82

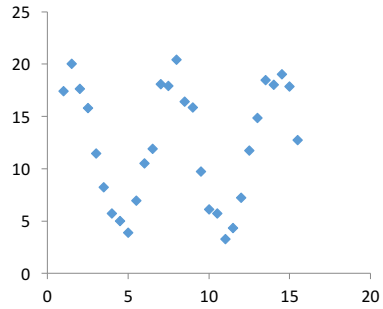
Корреляция -0.76

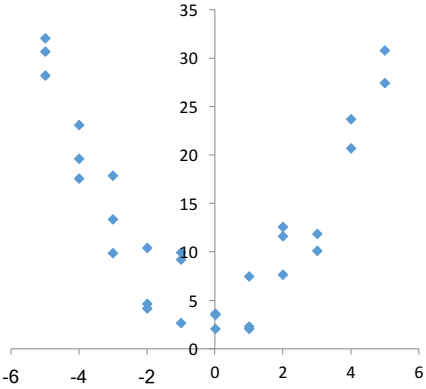


Корреляция - мера линейной связи



Корреляция везде около 0

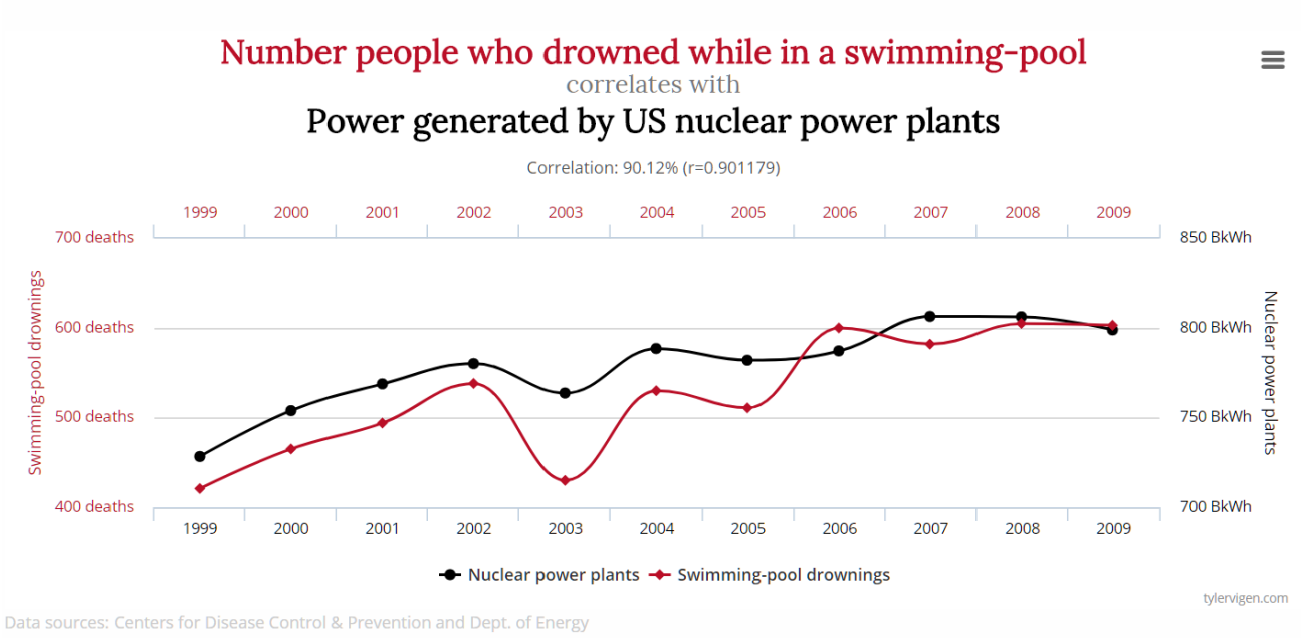




Ложная корреляция

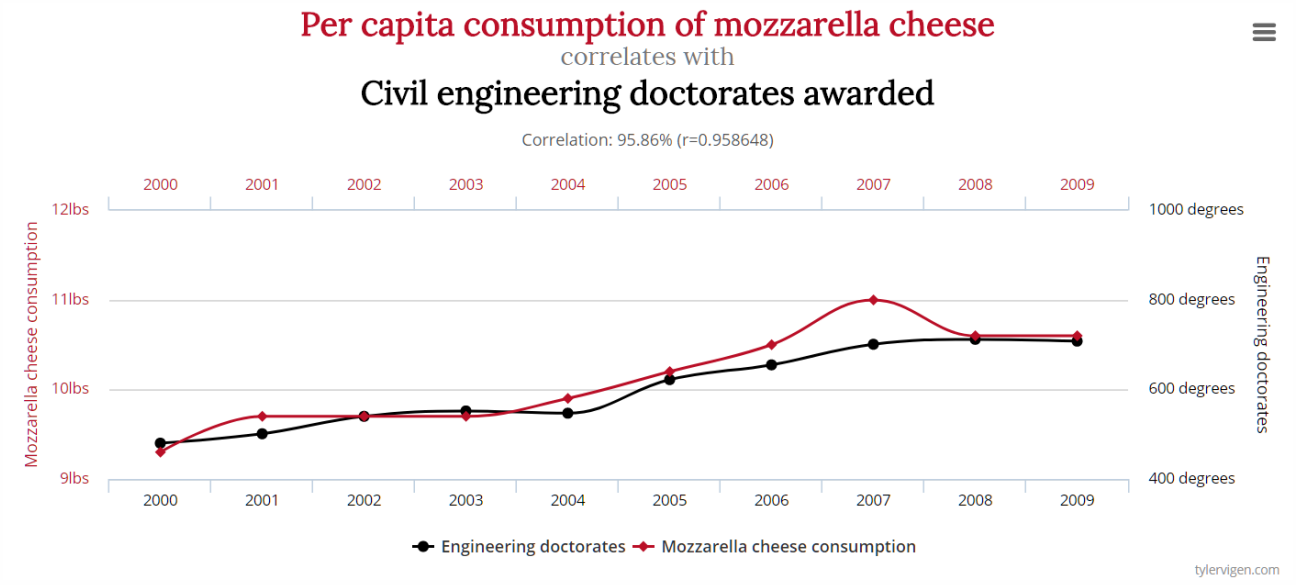
- Корреляция ещё **не означает**, что одна переменная **зависит** от другой! **Ложные корреляции** часто возникают, к примеру, когда мы работаем с **временными данными**
- По-хорошему, нужно **сначала предполагать наличие связи** между переменными из экономических предположений (теории, здравого смысла), и лишь **затем проверять эти предположения** на данных
- Иначе можно получить **бессмысленные результаты**

Ложная корреляция



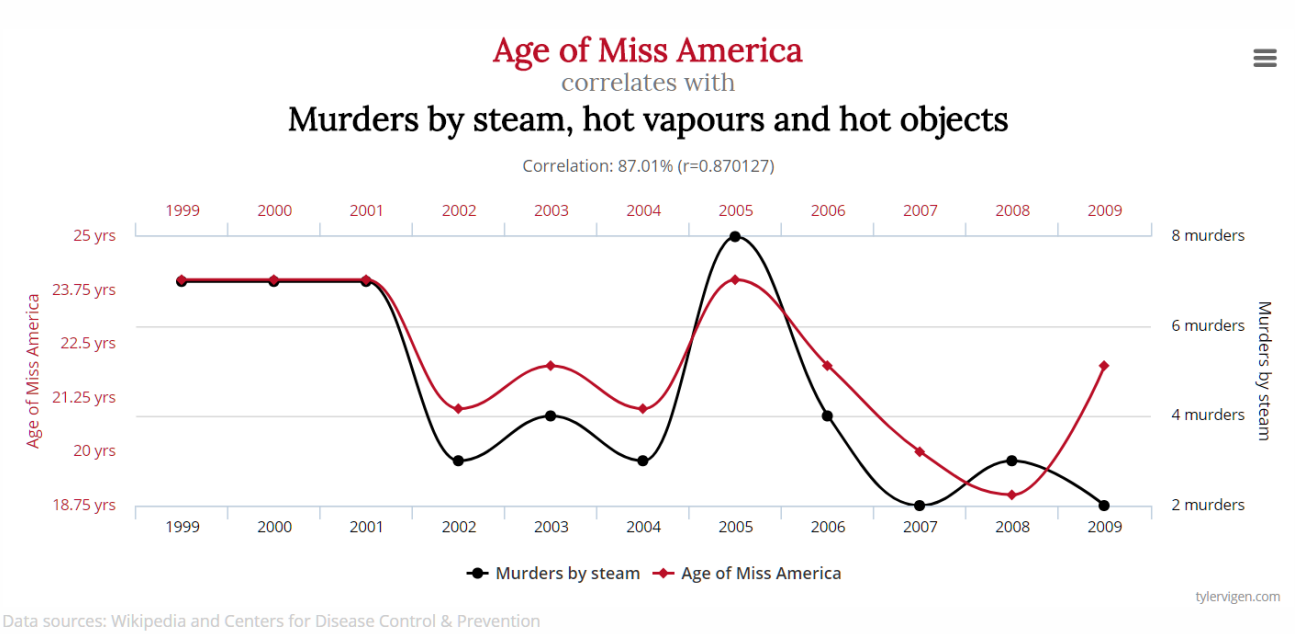
Data sources: Centers for Disease Control & Prevention and Dept. of Energy

Ложная корреляция



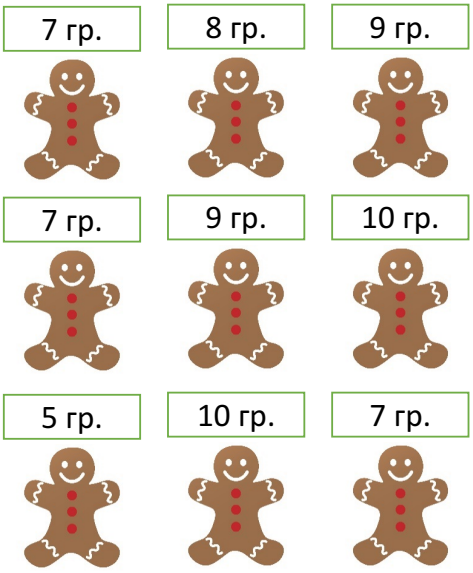
Data sources: U.S. Department of Agriculture and National Science Foundation

Ложная корреляция



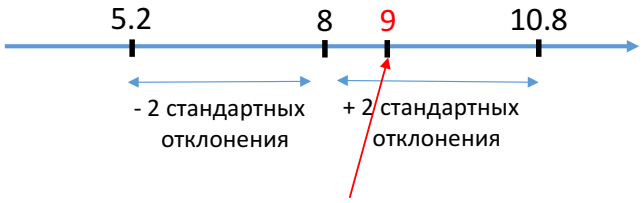
Гипотеза о среднем

- Вопрос: можно ли утверждать, что средний вес пряничного человечка значимо отличается от 9 грамм?



Среднее = 8

Стандартное отклонение = 1.4



9 внутри интервала +/- 2 стандартных отклонений от 8. Значит, среднее от 9 отличается незначимо!

Линейная регрессия

- Модель зависимости **одной** переменной **Y** от **одной или нескольких** независимых переменных **X**
- Зависимость **линейная**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Коэффициенты **β оцениваем по данным**

Линейная регрессия графически

Линейная регрессия в Excel

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0.257779322
R-квадрат	0.066450179
Нормированный R-квадрат	0.05692416
Стандартная ошибка	11.97786347
Наблюдения	100

Это показатель качества модели, чем ближе к 1 – тем лучше

Это свободный коэффициент

Это коэффициент при age

Если это число меньше 0.05, то age влияет на wage, иначе считается, что влияния нет

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	1000.791129	1000.791129	6.975650781	0.009617111
Остаток	98	14059.98289	143.4692132		
Итого	99	15060.77402			

	Коэффициенты	*Стандартная оши*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	30.19783836	4.36147503	6.923767339	4.61567E-10	21.54263311	38.85304361
age	-0.276632826	0.104739705	-2.641145733	0.009617111	-0.484485363	-0.068780289

Меньше 0.05 -> значит, есть связь!

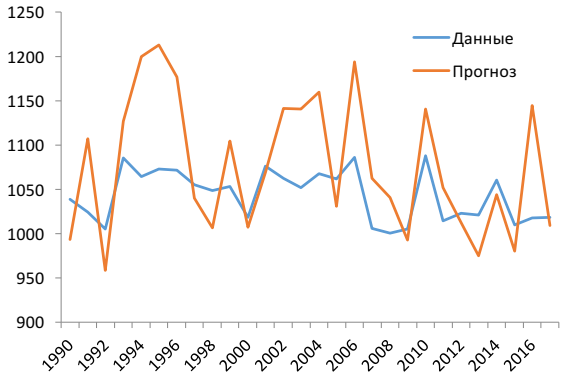
Тогда зависимость имеет вид: $wage = 30.2 - 0.277 * age$

Базы данных

- Россия
  - **РосСтат** – официальная статистика, широкий круг показателей и вопросов: http://gks.ru
  - **ЦБРФ** – статистика по банковской, денежно-кредитной сфере, немного макроэкономики: http://cbr.ru/statistics/
  - **РМЭЗ** – ежегодный опрос населения по очень(!) широкому кругу вопросов: https://www.hse.ru/rlms
  - А ещё: zakupki.gov.ru, sophist.hse.ru, fsin.su, stat.gibdd.ru, wciom.ru, cikrf.ru, …
- Мир
  - **World bank** – обычно макроэкономические данные по большому количеству стран: http://data.worldbank.org/
  - **ФРС США** – макроэкономические данные по США: https://www.stlouisfed.org/
  - **Бюро трудовой статистики США** – ещё макроэкономические данные по США: https://www.bls.gov/
  - **Quandl** – агрегатор, разные источники макроэкономической и финансовой статистики: https://www.quandl.com/
  - **Kaggle** – соревнования по машинному обучению, куча интересных и бесплатных данных: https://www.kaggle.com/
  - И практически бесконечность других источников…

Немного читинга с графиками

Меняем границы осей

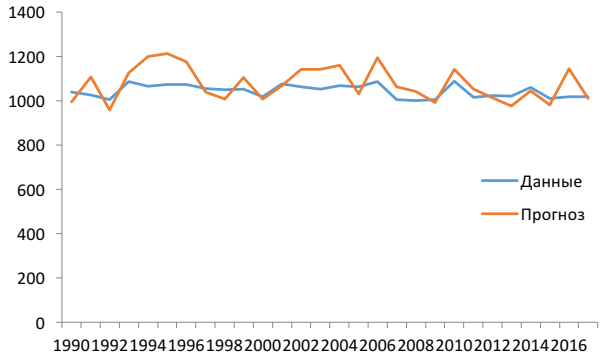


Лёгким движением руки плохой прогноз

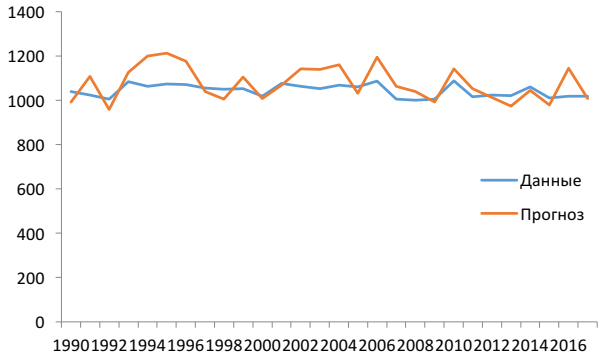
Меняем границы осей

Лёгким движением руки плохой прогноз

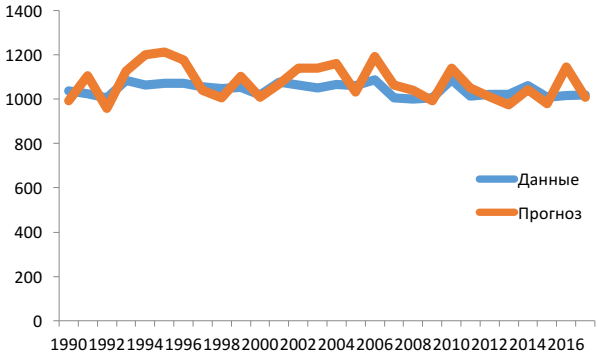
Становится очень неплохим прогнозом



И делаем линии пожирнее



И даже ещё более неплохим!



Создаём спады и восстановления

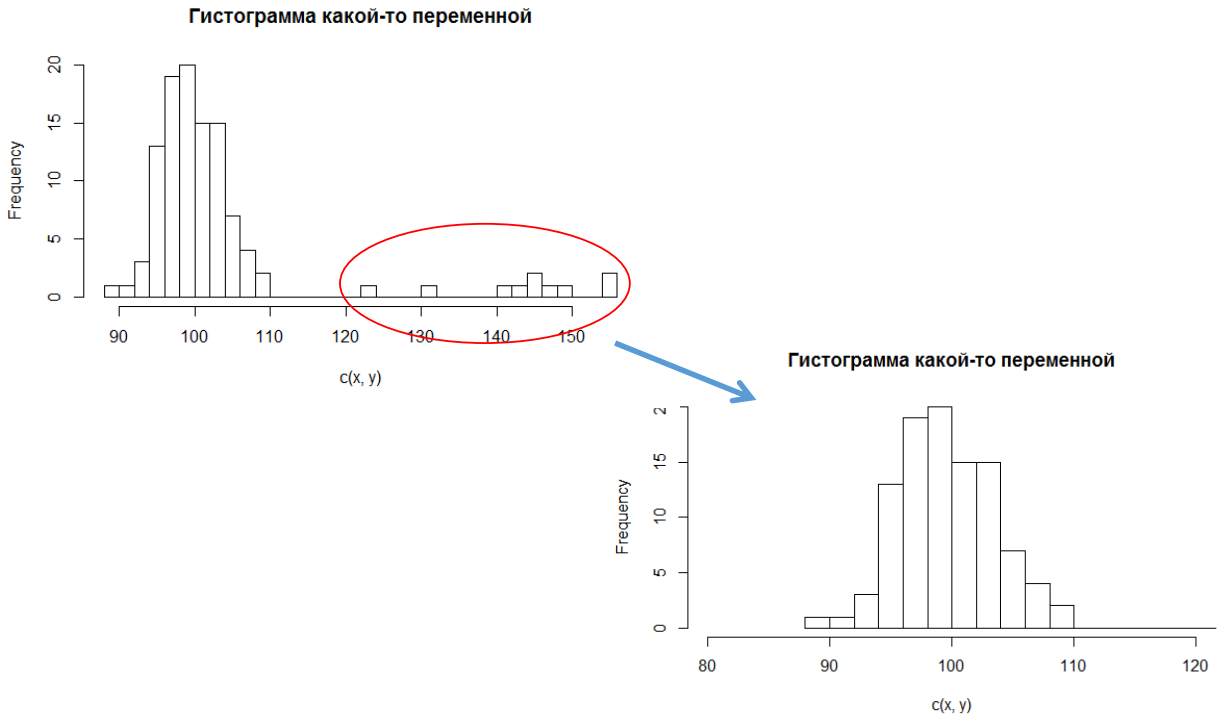
Обычный, стабильный во времени показатель

Создаём спады и восстановления

Обычный, стабильный во времени показатель

Но нет!

Прячем неугодные наблюдения



Или вот так