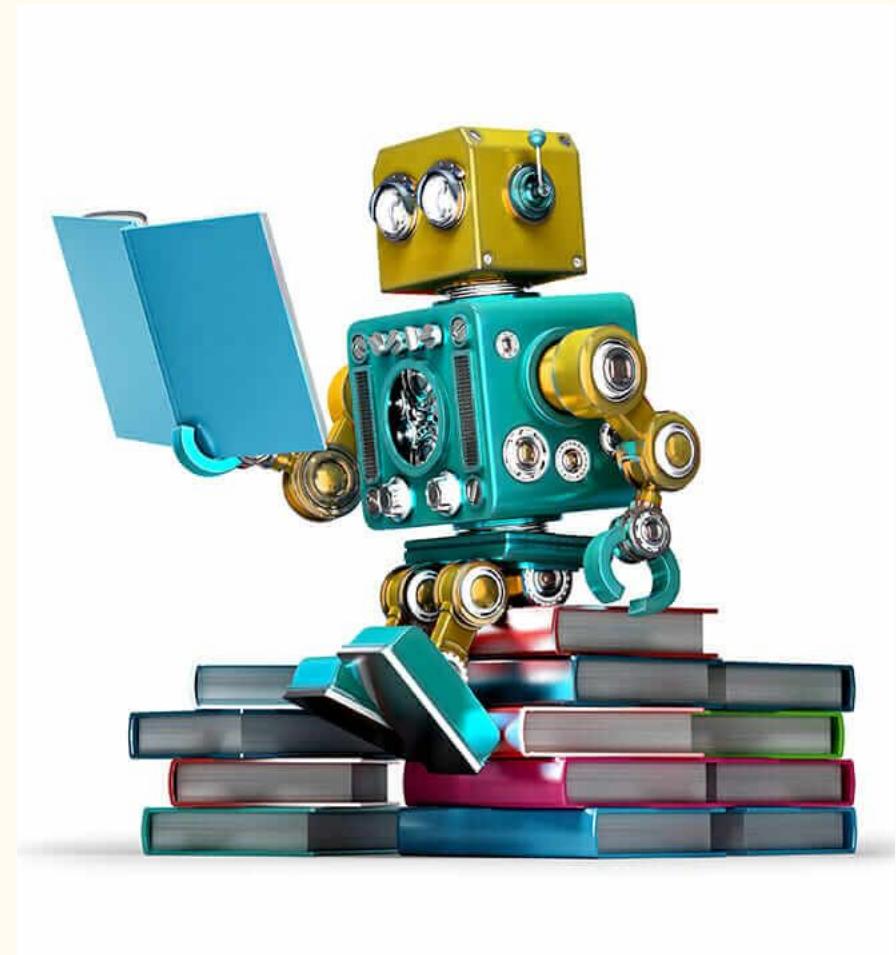


Lecture 1

Nadya Zueva, MIPT
Data Analysis,
Yandex

Plan

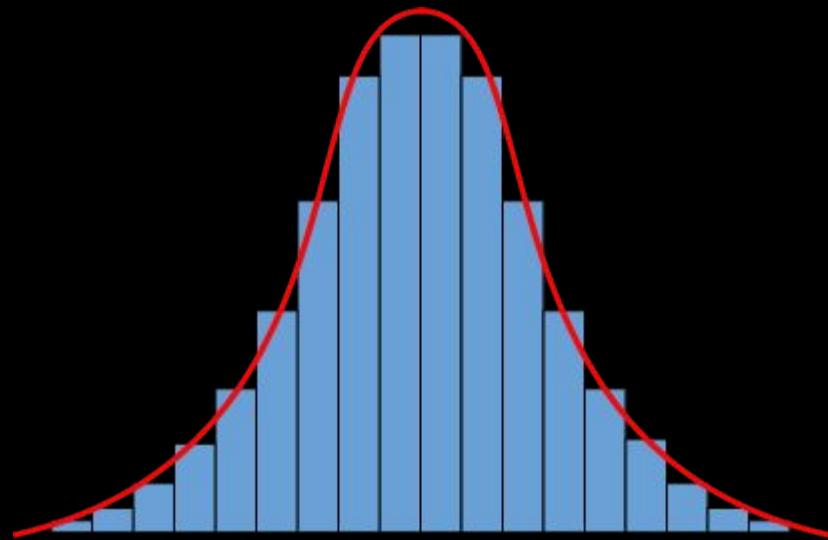
1. Структура курса
2. Recap (статистика)
3. Основные определения
4. Подходы к прогнозированию
 - a. Регрессия
 - b. STL - разложение
5. Автокорреляция и короллограммы
6. Стационарность
7. Дифференцирование ряда
8. MA, AR, ARMA,...
9. *Подбор параметров модели



Структура курса

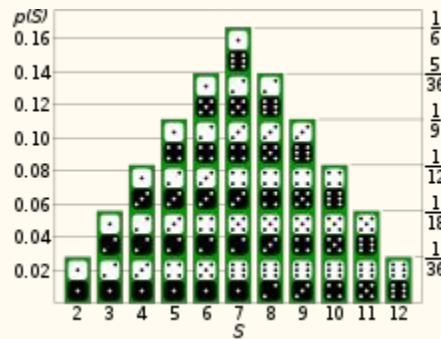
1. Обзор основных методов и повторение статистики
2. Предобработка данных
3. Методы прогнозирования
4. Проект и защита

Recap

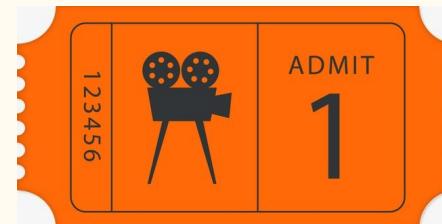


Случайная величина

переменная, значения которой представляют собой исходы какого-нибудь **случайного эксперимента** ξ



ξ, X



Выборка

Выборка — конечное множество объектов, отобранных из генеральной совокупности для проведения измерений.

$$X^n = (X_1, \dots, X_n).$$

n — объём выборки.

X^n — **простая выборка**, если X_1, \dots, X_n — независимые одинаково распределённые случайные величины (i.i.d.).

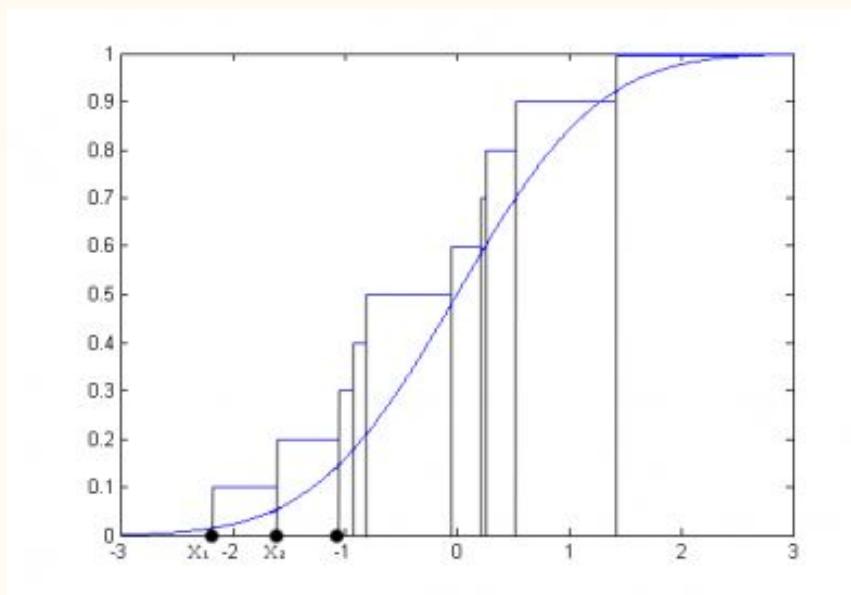
Пусть $F(x)$ — функция распределения элемента простой выборки:

$$F(x) = \mathbf{P}(X \leqslant x).$$

Основная задача статистики — описание $F(x)$ по реализации выборки.

Эмпирическая функция распределения

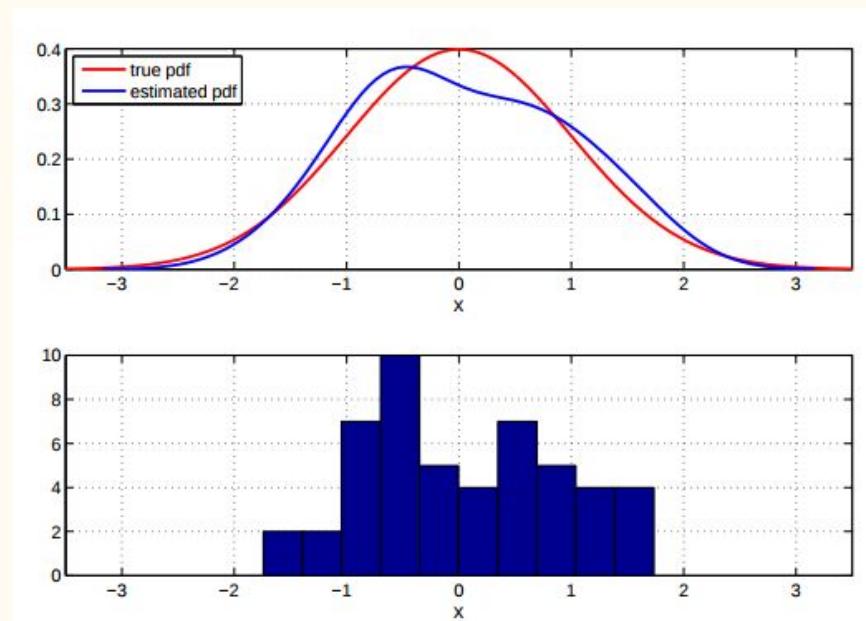
$$\widehat{F}_m(x) = \frac{1}{m} \sum_{i=1}^m I_{\{x_i \leq x\}}.$$



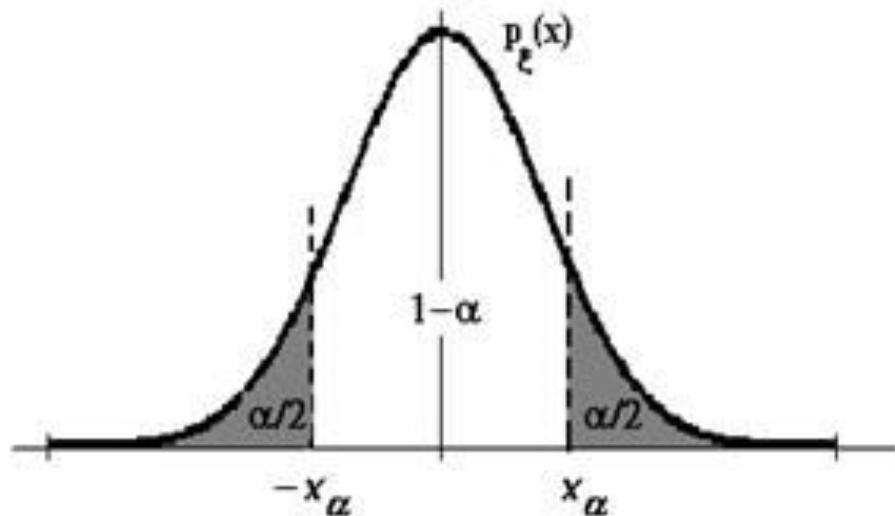
Плотность распределения

отражает изменение вероятности на конкретном элементе множества

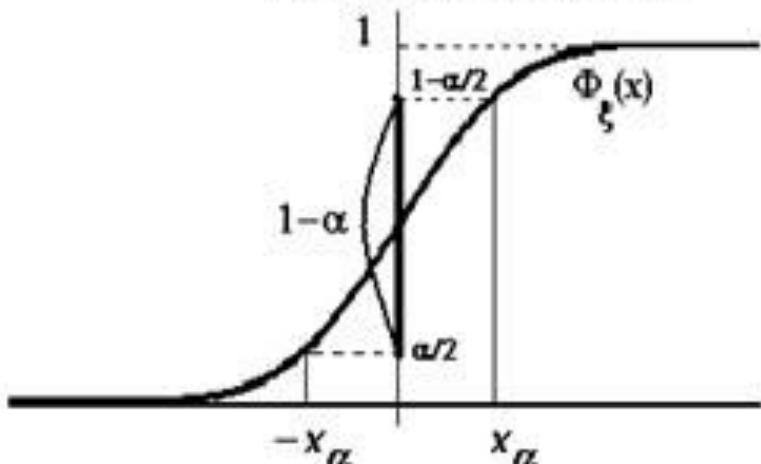
$$\mathbb{P}(B) = \int_B f(x) dx$$



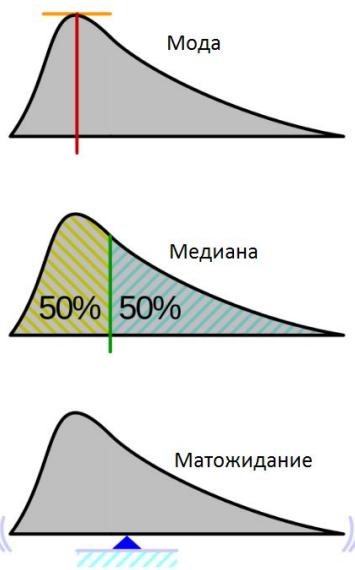
Плотность вероятностей



Функция распределения



Часто интерес представляют отдельные характеристики распределения $F(x)$:



- **матожидание** — среднее значение X :

$$\mathbb{E}X = \int x dF(x);$$

- **дисперсия** — мера разброса X :

$$\mathbb{D}X = \mathbb{E}((X - \mathbb{E}X)^2);$$

- **коэффициент асимметрии** (skewness):

$$\gamma_1 = \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sqrt{\mathbb{D}X}}\right)^3;$$

- **квантиль порядка** $\alpha \in (0, 1)$:

$$X_\alpha: \quad \mathbf{P}(X \leq X_\alpha) \geq \alpha, \quad \mathbf{P}(X \geq X_\alpha) \geq 1 - \alpha.$$

- **медиана** — квантиль порядка 0.5, «центральное» значение распределения:

$$\text{med } X: \quad \mathbf{P}(X \leq \text{med } X) \geq \frac{1}{2}, \quad \mathbf{P}(X \geq \text{med } X) \geq \frac{1}{2}.$$

Статистика $T(X^n)$ — любая измеримая функция выборки.

Примеры:

- выборочное среднее:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i;$$

- выборочная дисперсия:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2;$$

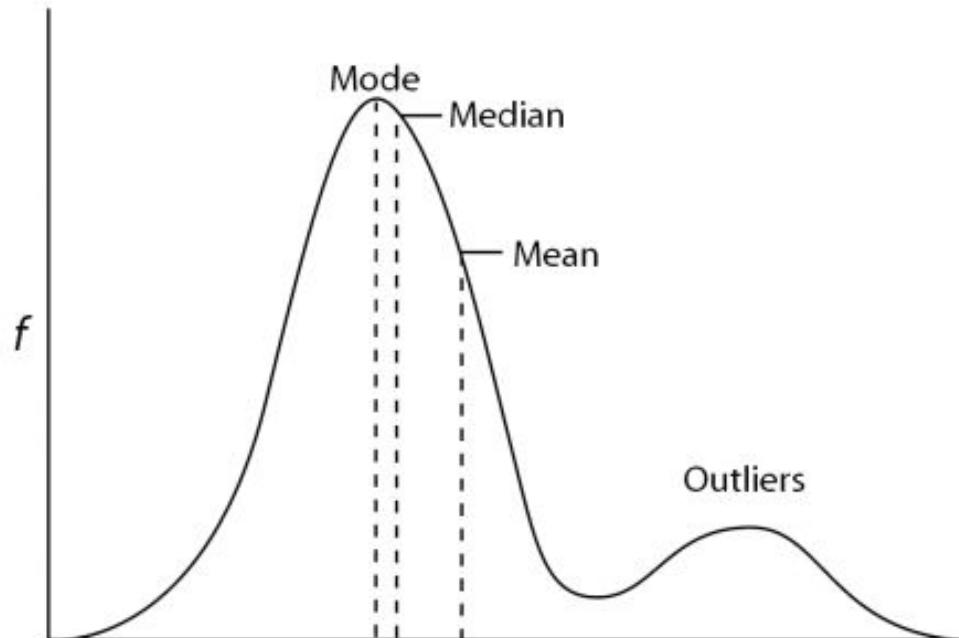
- выборочный коэффициент асимметрии:

$$g_1 = \frac{\sqrt{n} \sum_{i=1}^n (X_i - \bar{X})^3}{\left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)^{3/2}};$$

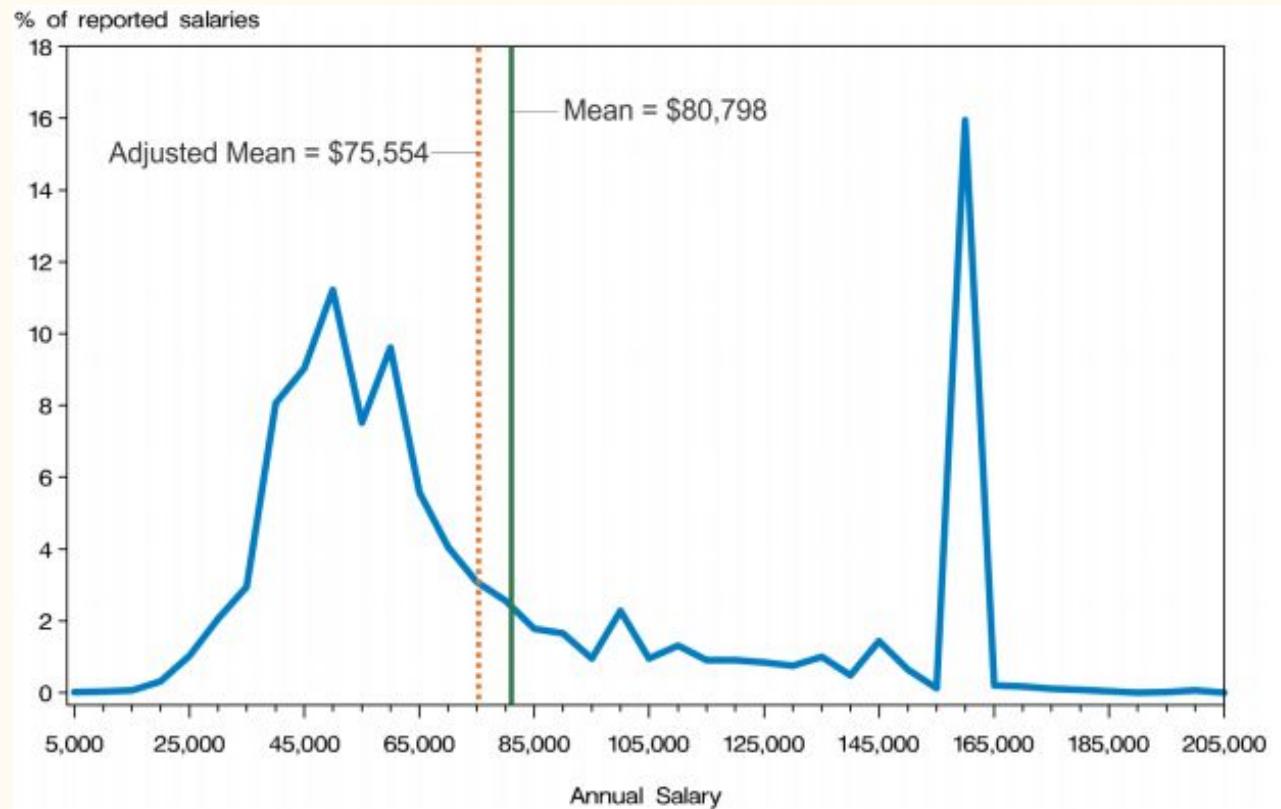
Выборочное среднее — среднее арифметическое по выборке.

Медиана — центральный элемент вариационного ряда.

Мода — самое распространённое значение в выборке.

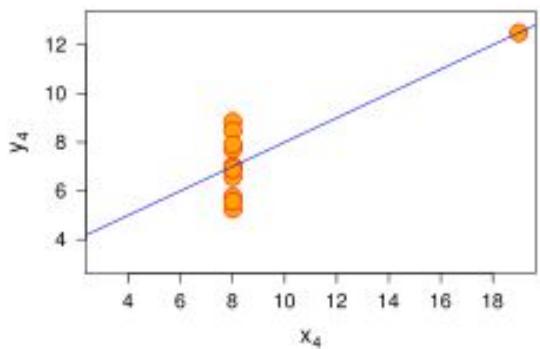
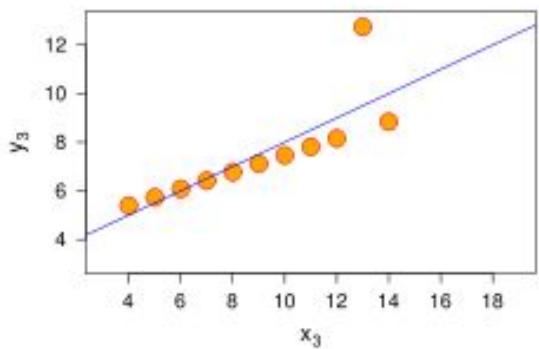
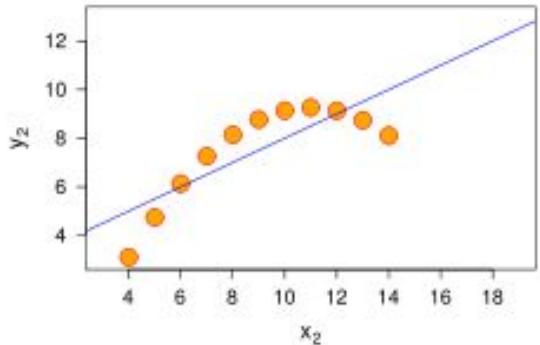
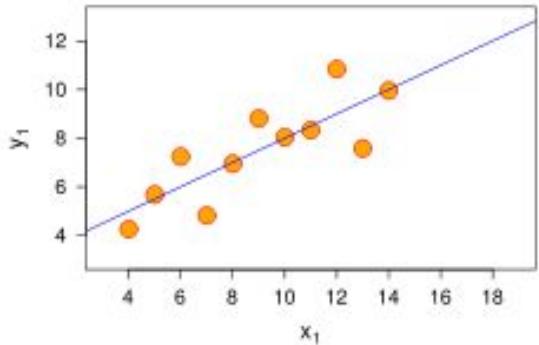


Визуализация



Квартет Энскомба

Характеристика	Значение
<u>Среднее значение</u> переменной $\{x\}$	9,0
<u>Дисперсия</u> переменной $\{x\}$	10,0
Среднее значение переменной $\{y\}$	7,5
Дисперсия переменной $\{y\}$	3,75
<u>Корреляция</u> между переменными $\{x\}$ и $\{y\}$	0,816
Прямая <u>линейной регрессии</u>	$y = 3 + 0,5x$



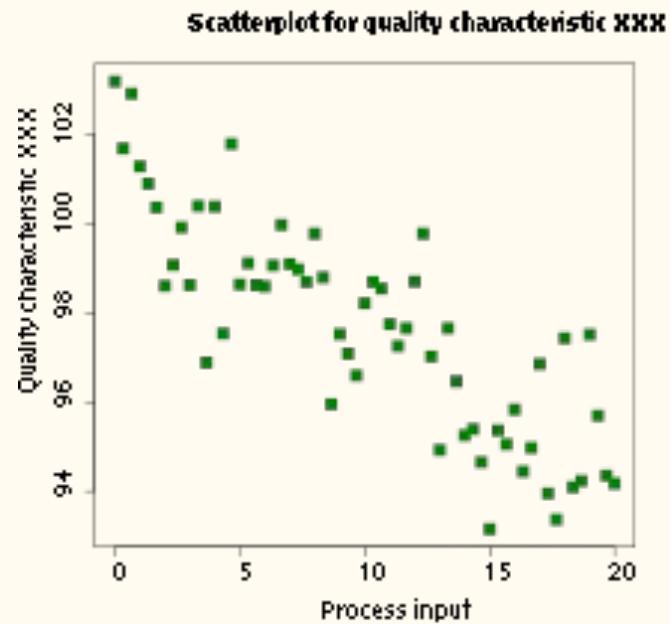
Корреляция

статистическая взаимосвязь двух или более случайных величин

$$\text{cov}_{XY} = \mathbf{M}[(X - \mathbf{M}(X))(Y - \mathbf{M}(Y))]$$

$$\text{cov}_{XY} = \mathbf{M}[(X - \mathbf{M}(X))(Y - \mathbf{M}(Y))] = \mathbf{M}(XY) - \mathbf{M}(X)\mathbf{M}(Y)$$

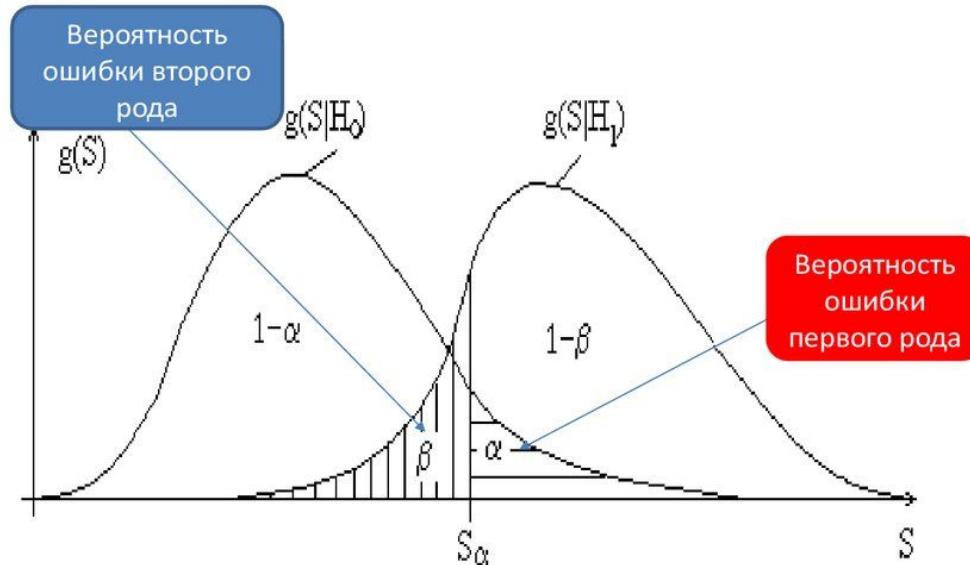
$$r_{XY} = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}.$$



ошибки 1 и 2 рода

H_0	верная	ложная
отклоняется	ошибка первого рода	решение верное
не отклоняется	решение верное	ошибка второго рода

Ошибки первого и второго рода



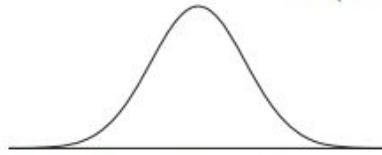
Проверка гипотез

выборка: $X^n = (X_1, \dots, X_n)$, $X \sim P \in \Omega$;

нулевая гипотеза: $H_0: P \in \omega$, $\omega \in \Omega$;

альтернатива: $H_1: P \notin \omega$;

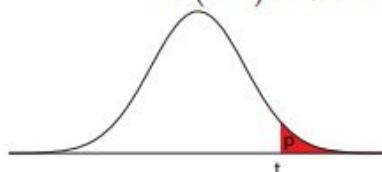
статистика: $T(X^n)$, $T(X^n) \sim F(x)$ при $P \in \omega$;
 $T(X^n) \not\sim F(x)$ при $P \notin \omega$;



реализация выборки: $x^n = (x_1, \dots, x_n)$;

реализация статистики: $t = T(x^n)$;

достигаемый уровень значимости: $p(x^n)$ — вероятность при H_0 получить
 $T(X^n) = t$ или ещё более экстремальное;



$$p(x^n) = P(T \geq t | H_0)$$

Гипотеза отвергается при $p(x^n) \leq \alpha$, α — уровень значимости.

пример проверки гипотезы

Продавец утверждает, что средний вес пачки чая составляет 100 г. Из партии извлечена выборка и взвешена. Вес каждой пачки - см. таблицу вариантов. Не противоречит ли это утверждению продавца? Доверительная вероятность 99%. Вес пачек чая распределен нормально.



	Выборка
5	98, 104, 97, 97, 101, 100, 99, 101, 99, 98

РЕШЕНИЕ.

Вычислим показатели выборки.

$$\bar{x} = \frac{1}{n} \sum x_i n_i = \frac{1}{10} 994 = 99,4.$$

Исправленная выборочная дисперсия

$$S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 n_i = \frac{1}{9} 42,4 \approx 4,71.$$

Выборочное исправленное среднеквадратичное отклонение:
 $S \approx 2,171.$

Расчеты в таблице:

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})^2 n_i$
97	2	194	11,52
98	2	196	3,92
99	2	198	0,32
100	1	100	0,36
101	2	202	5,12
104	1	104	21,16
Сумма	10	994	42,4

Введем нулевую гипотезу $H_0 : a = 100$ при конкурирующей гипотезе $H_1 : a \neq 100$.

Вычисляем наблюдаемое значение критерия

$$T_{\text{набл}} = \frac{(\bar{x} - a)}{S} \sqrt{n} = \frac{(99,4 - 100)}{2,171} \sqrt{10} \approx -0,87.$$

По таблице критических точек распределения Стьюдента найдем критическую точку по уровню значимости $\alpha = 0,01$ и числу степеней свободы $k = 9$, откуда $t_{kp} \approx 3,25$.

Так как $|T_{\text{набл}}| = 0,87 < 3,25 = t_{kp}$, то нулевую гипотезу о равенстве среднего веса 100 г можно принять.

Введем нулевую гипотезу $H_0 : a = 100$ при конкурирующей гипотезе $H_1 : a \neq 100$.

Вычисляем наблюдаемое значение критерия

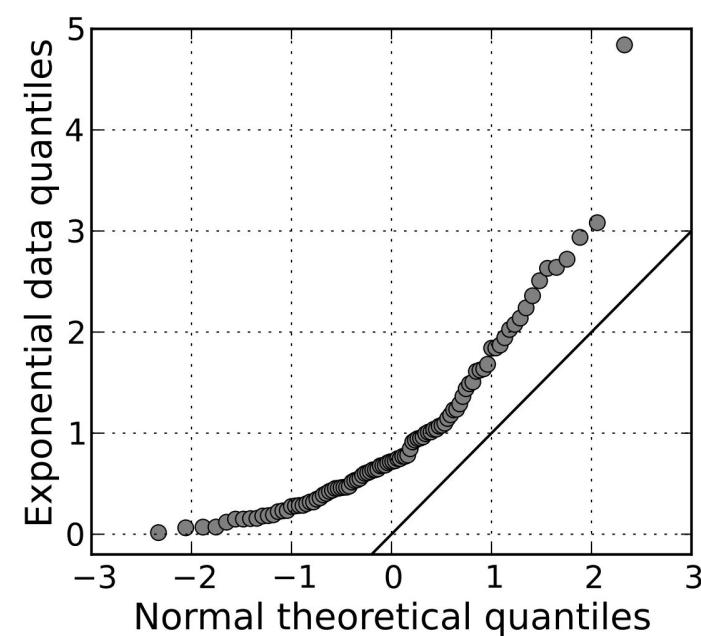
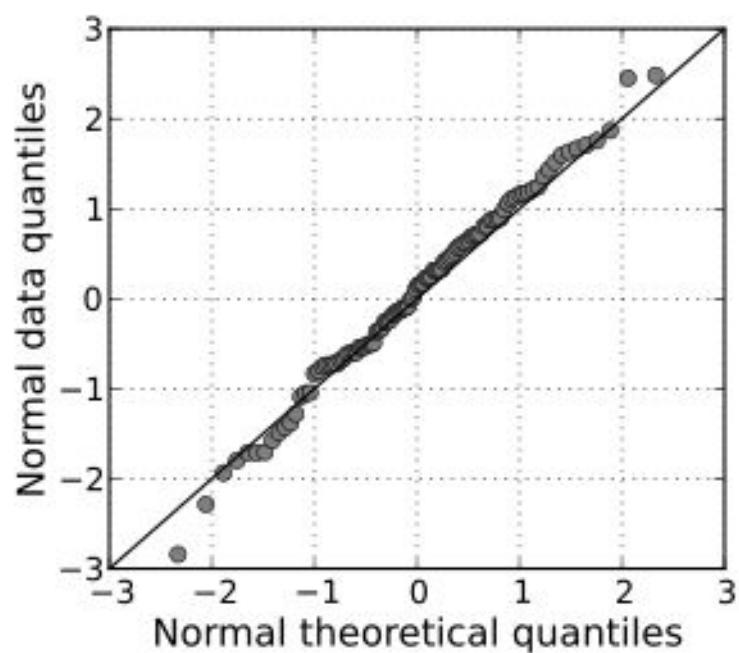
$$T_{\text{набл}} = \frac{\bar{x} - a}{S} \sqrt{n} = \frac{(99,4 - 100)}{2,171} \sqrt{10} \approx -0,87.$$

По таблице критических точек распределения Стьюдента найдем критическую точку по уровню значимости $\alpha = 0,01$ и числу степеней свободы $k = 9$, откуда $t_{kp} \approx 3,25$.

Так как $|T_{\text{набл}}| = 0,87 < 3,25 = t_{kp}$, то нулевую гипотезу о равенстве среднего веса 100 г можно ~~принять~~

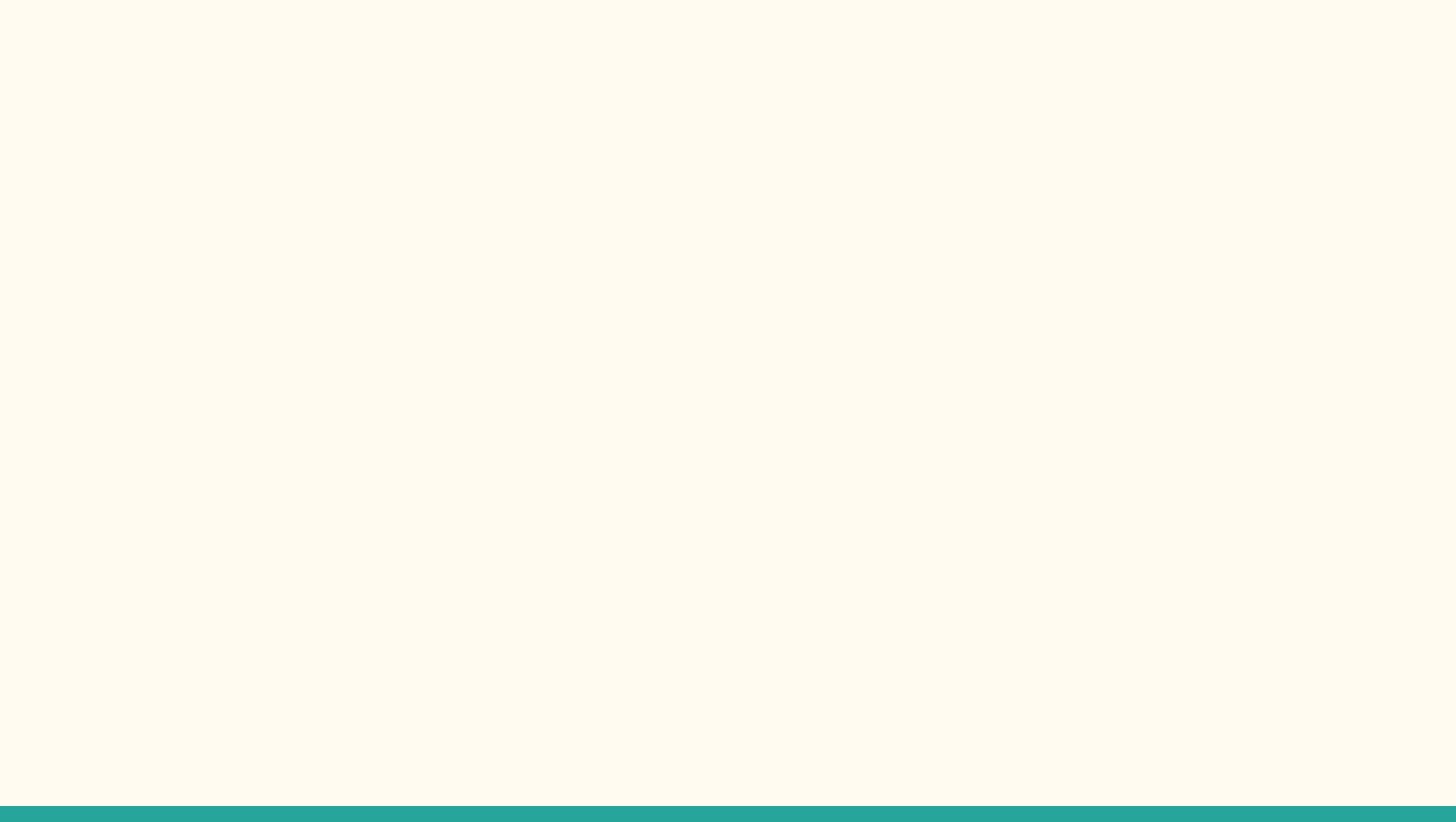
не отвергнуть

QQ-plot

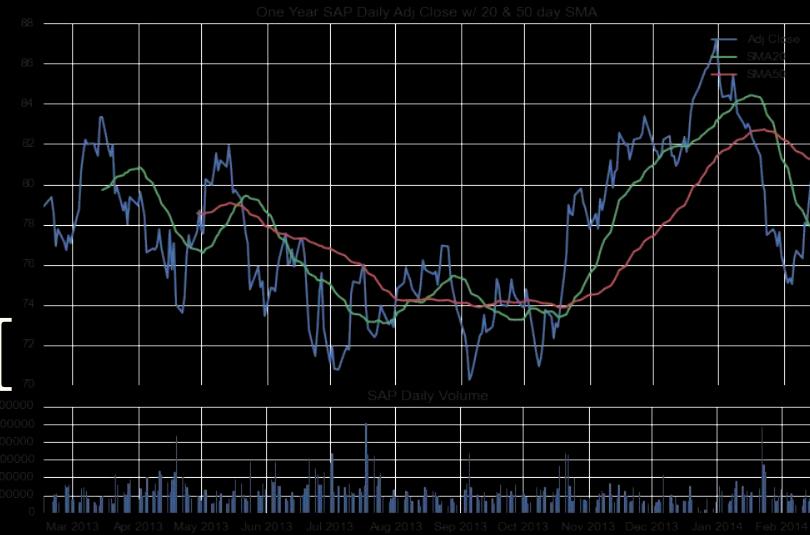




SciPy



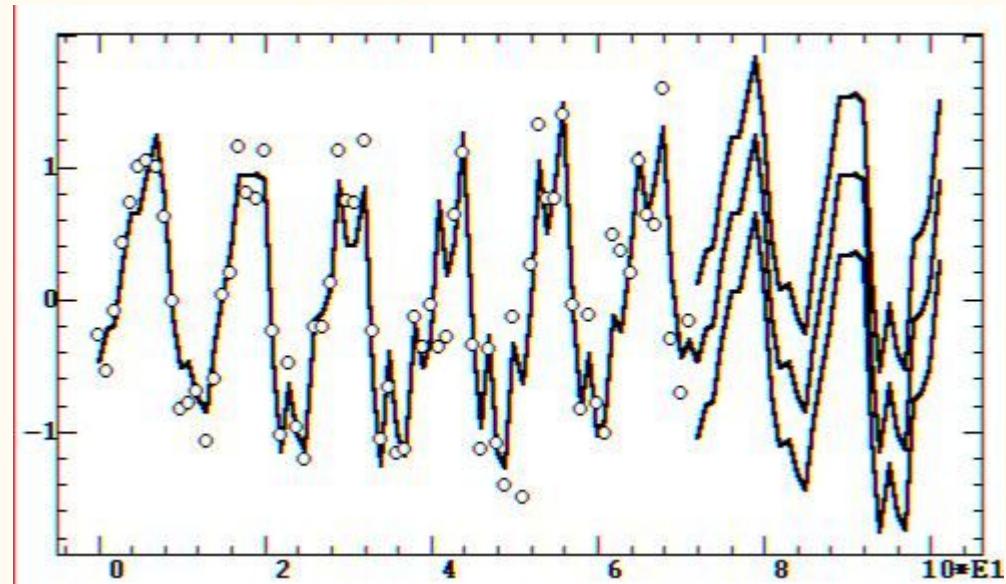
Основные определения



Временной ряд

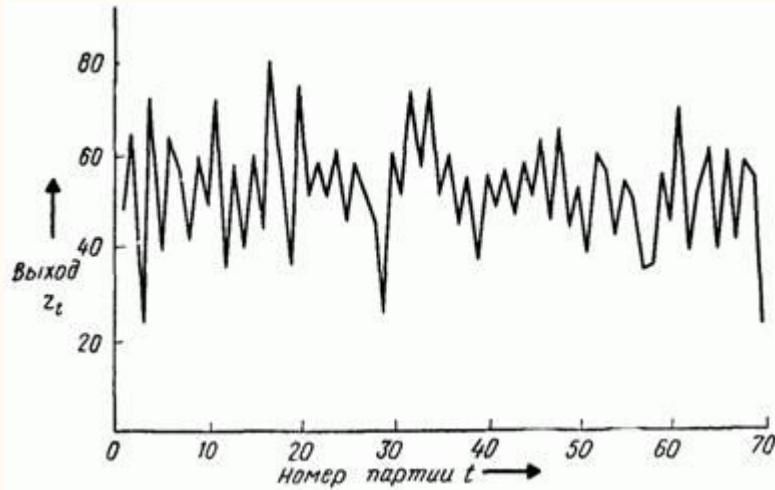
$y_1, \dots, y_T, \dots, y_t \in R$, —
значения признака, измеренные
через постоянные временные
интервалы.

(y_n, T_n)



Примеры временных рядов

1. Погодные данные
2. Акции компании
3. Количество продаж
4. Биржевой курс
5. Температура
6. ...



Задача прогнозирования: найти функцию f_T :

$$y_{T+h} \approx f_T(y_T, \dots, y_1, h) \equiv \hat{y}_{T+h|T},$$

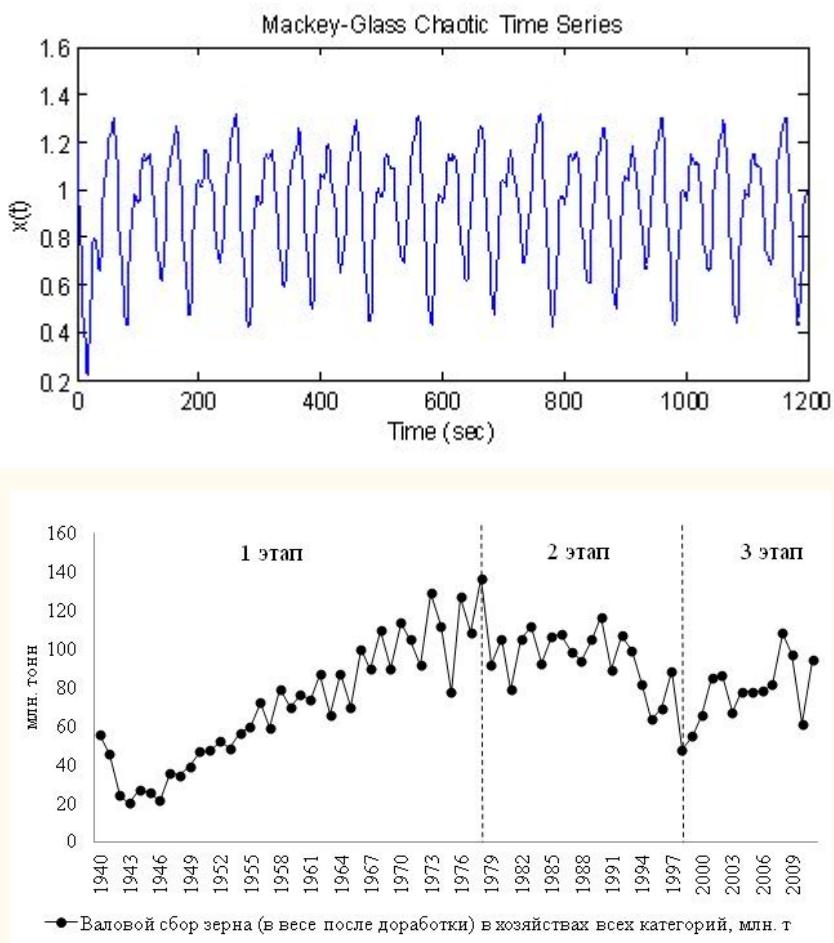
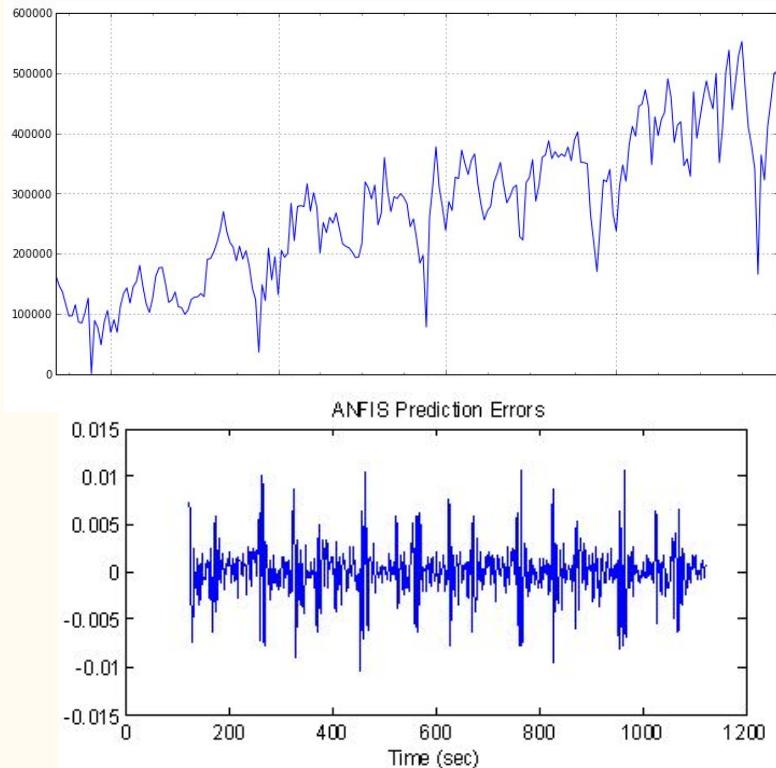
$h \in \{1, 2, \dots, H\}$, H — горизонт прогнозирования.

Особенности прогнозирования временных рядов

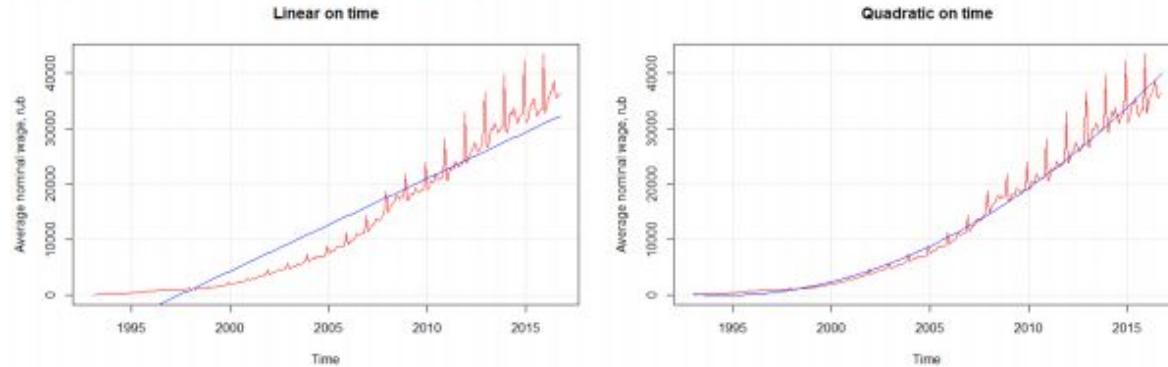
В классических задачах анализа данных предполагается независимость наблюдений

При прогнозировании временных рядов, наоборот, мы надеемся, что значения ряда в прошлом содержат информацию о его поведении в будущем

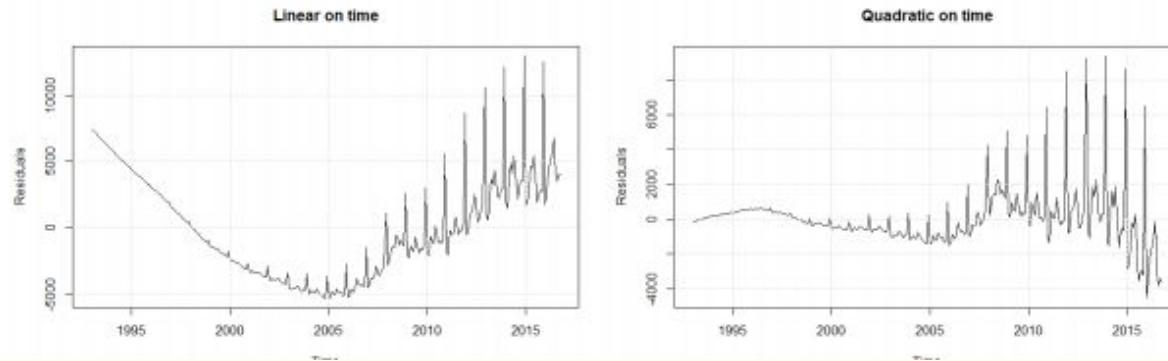
i.i.d?

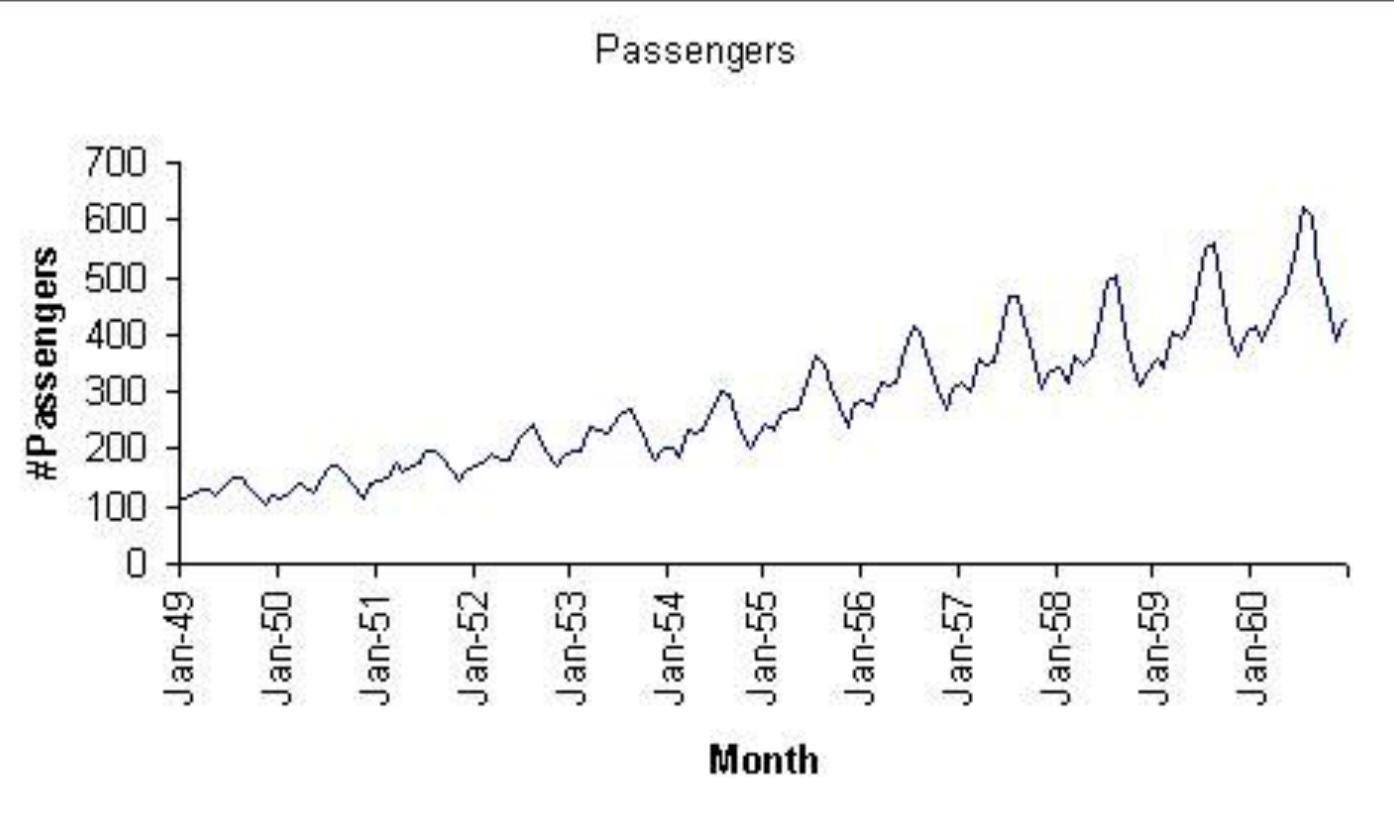


Идея: регрессия на время



Остатки не выглядят как шум:





STL-компоненты

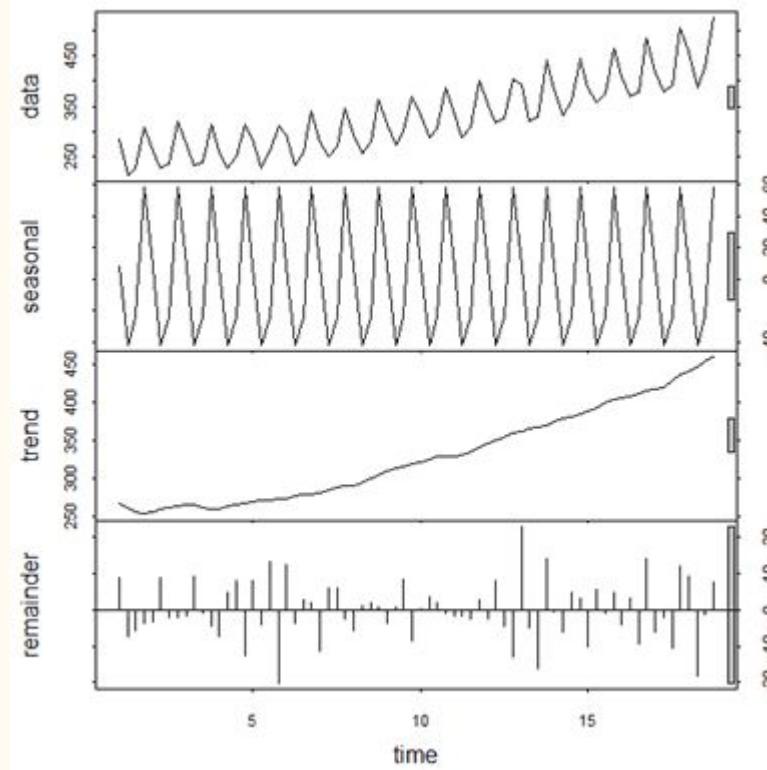
Тренд — плавное долгосрочное изменение уровня ряда.

Сезонность — циклические изменения уровня ряда с постоянным периодом.

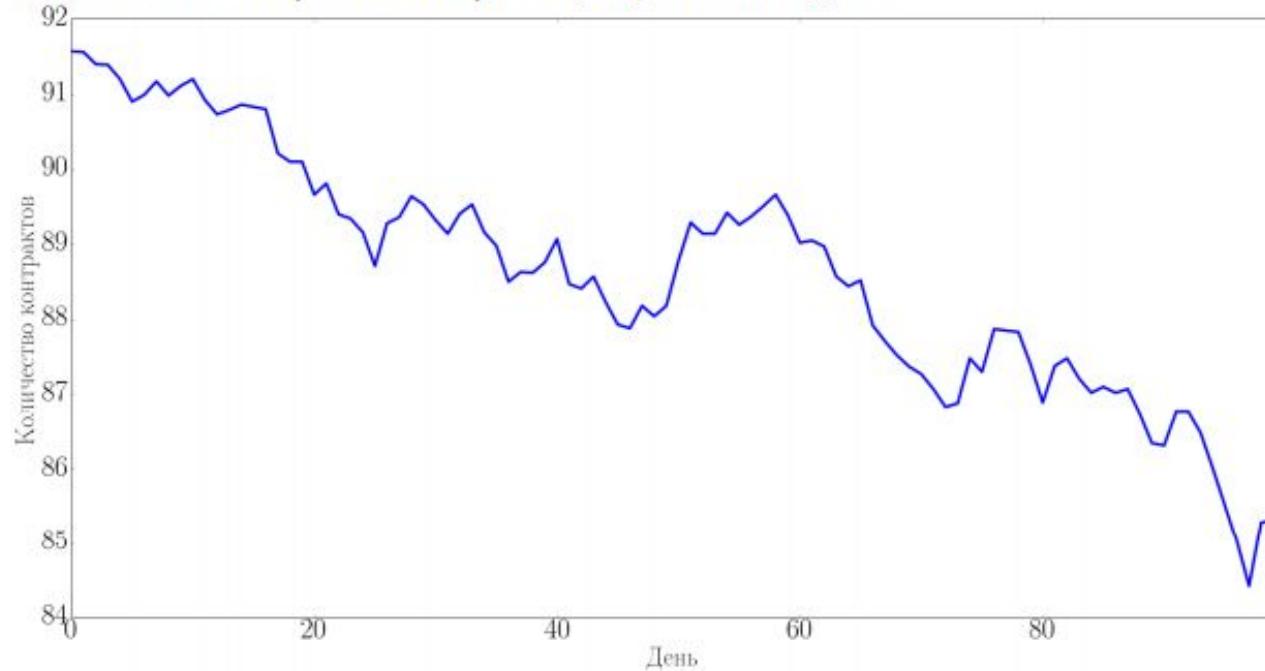
Цикл — изменения уровня ряда с переменным периодом (экономические циклы, периоды солнечной активности).

Ошибка — непрогнозируемая случайная компонента ряда

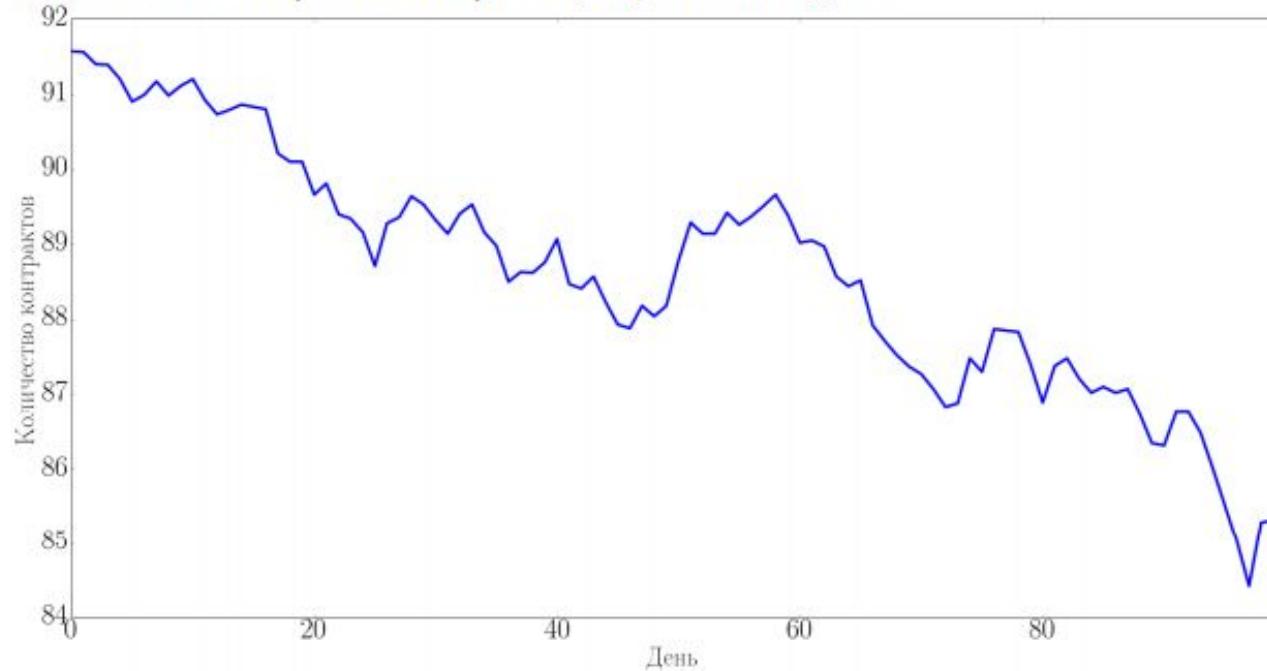
STL-декомпозиция



Количество контрактов сокровищницы США в день:

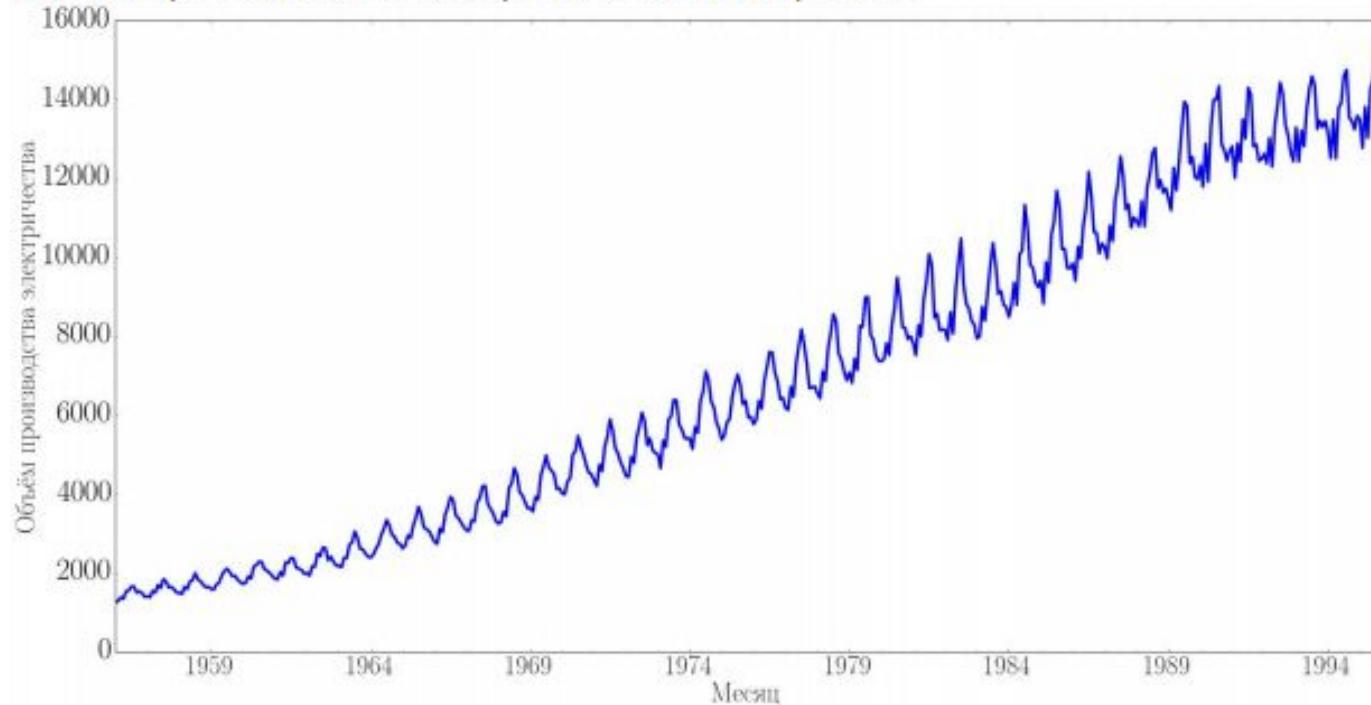


Количество контрактов сокровищницы США в день:

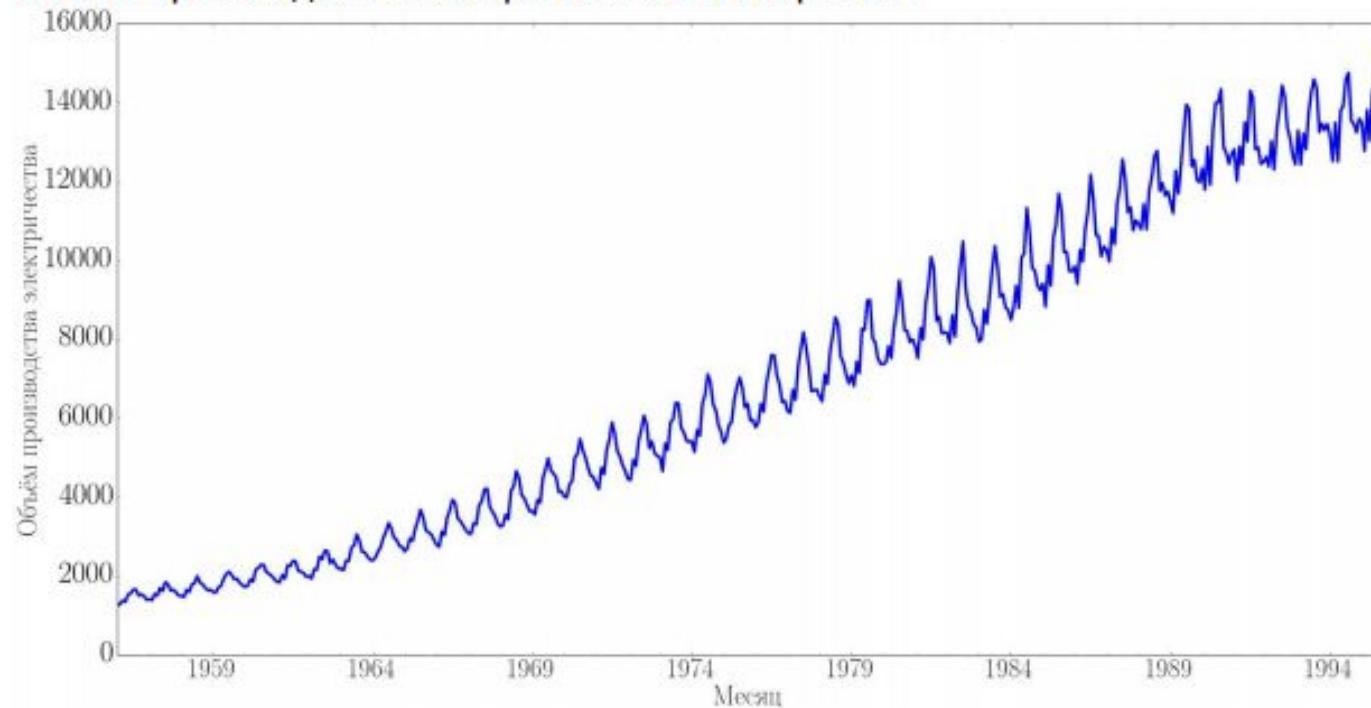


Тренд

Объём производства электричества в Австралии:

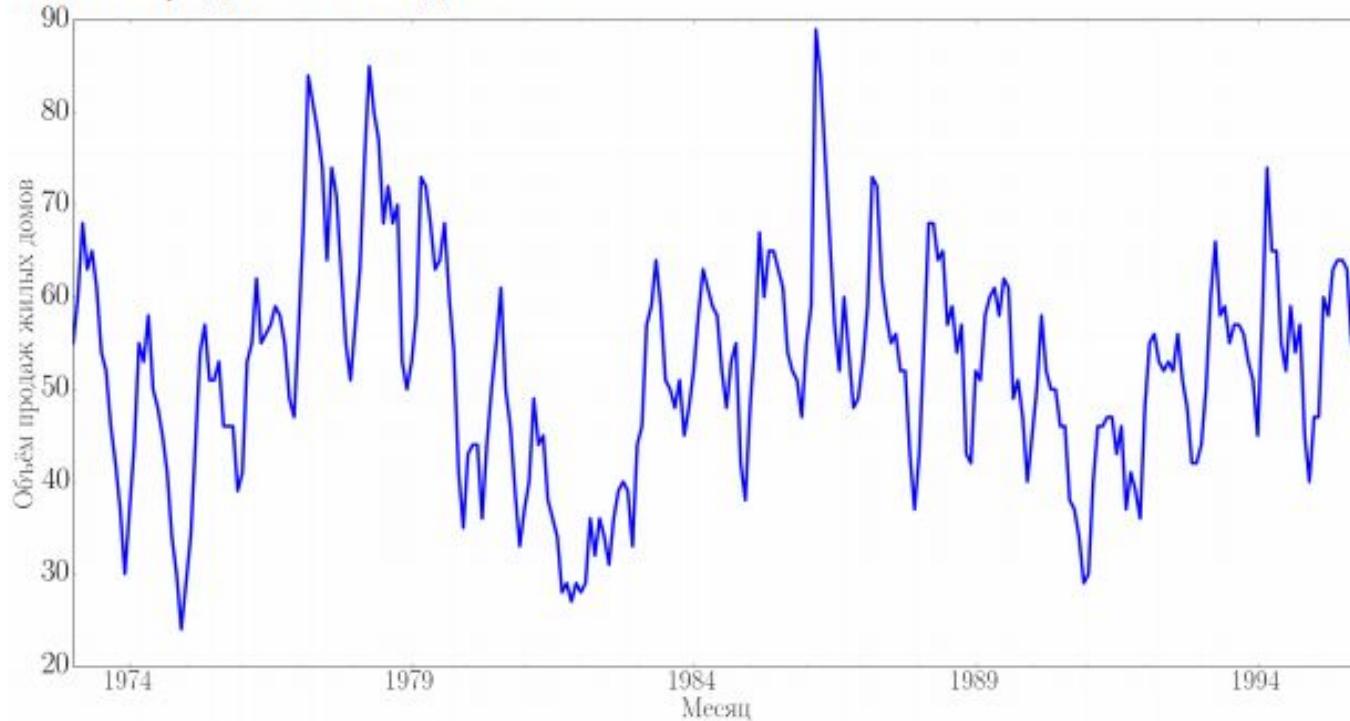


Объём производства электричества в Австралии:

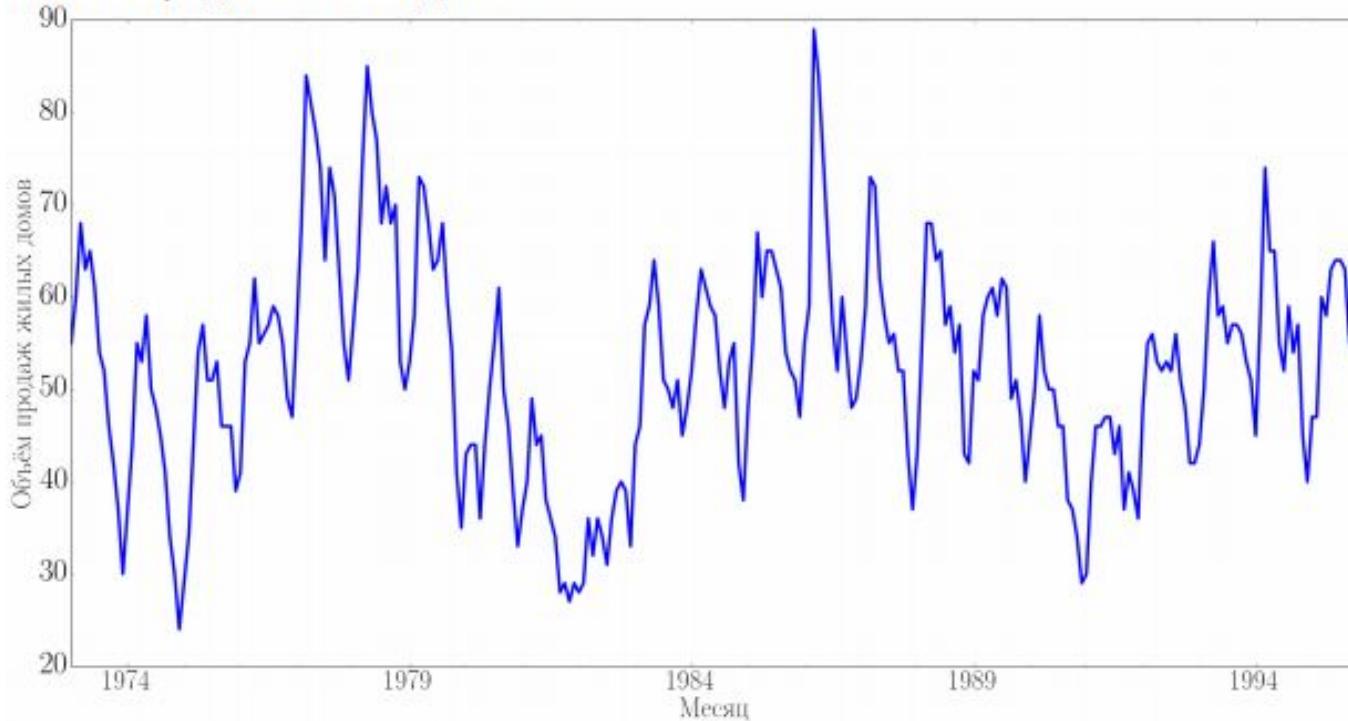


Тренд, годовая сезонность

Объём продаж жилых домов:

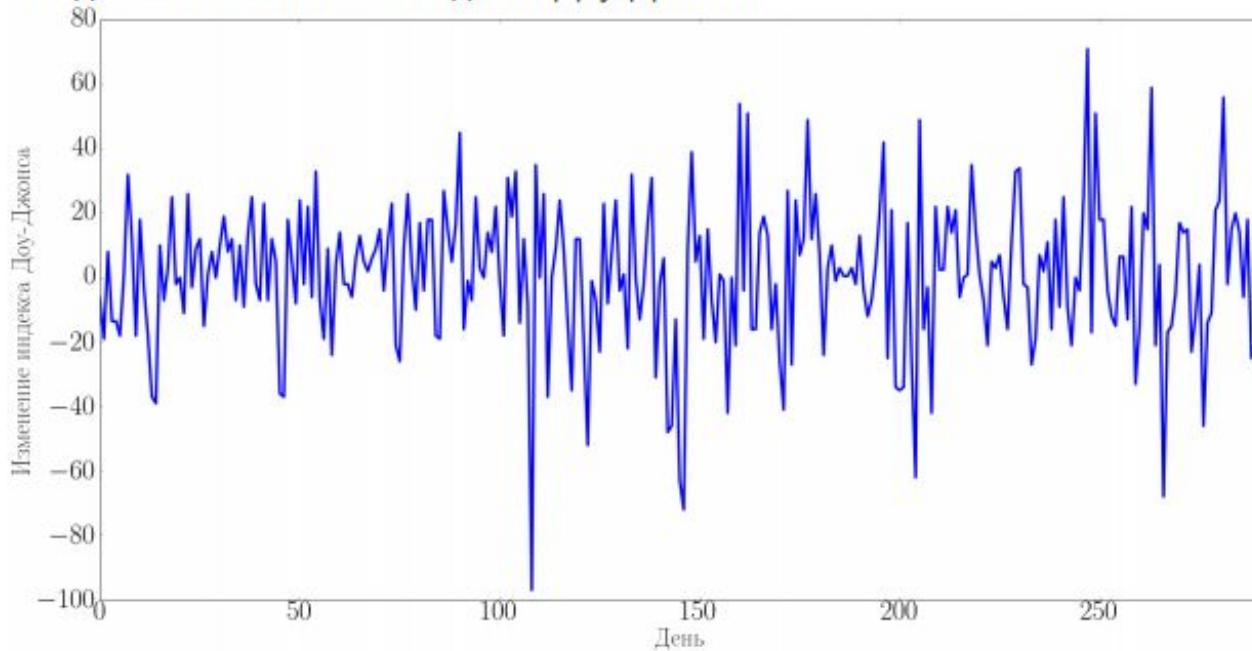


Объём продаж жилых домов:

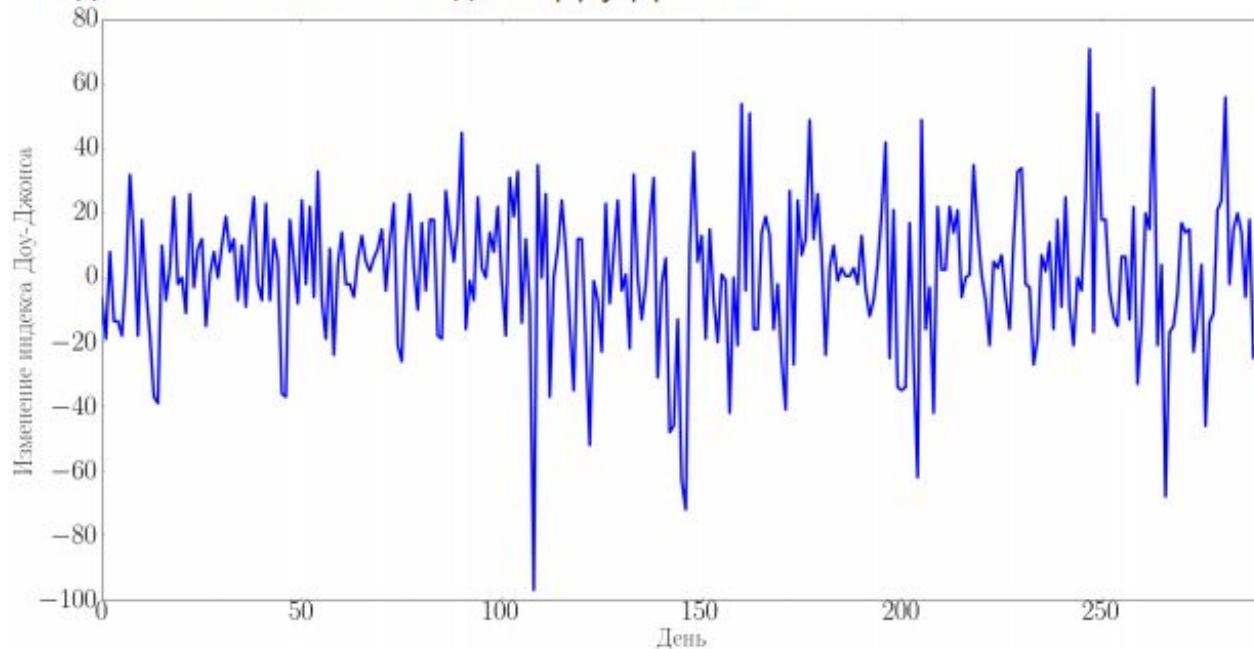


Циклы, сезонность

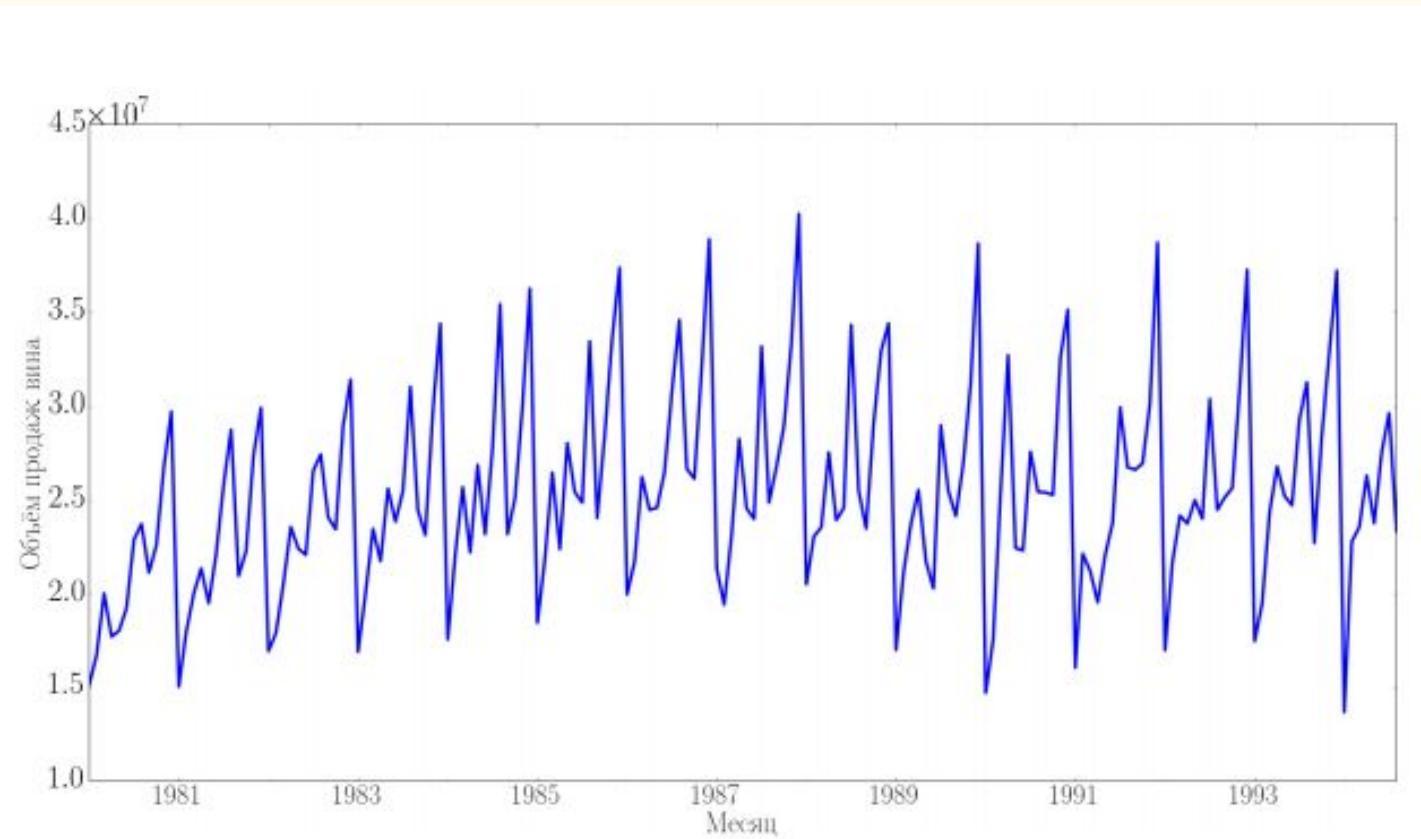
Ежедневные изменения индекса Доу-Джонса:



Ежедневные изменения индекса Доу-Джонса:

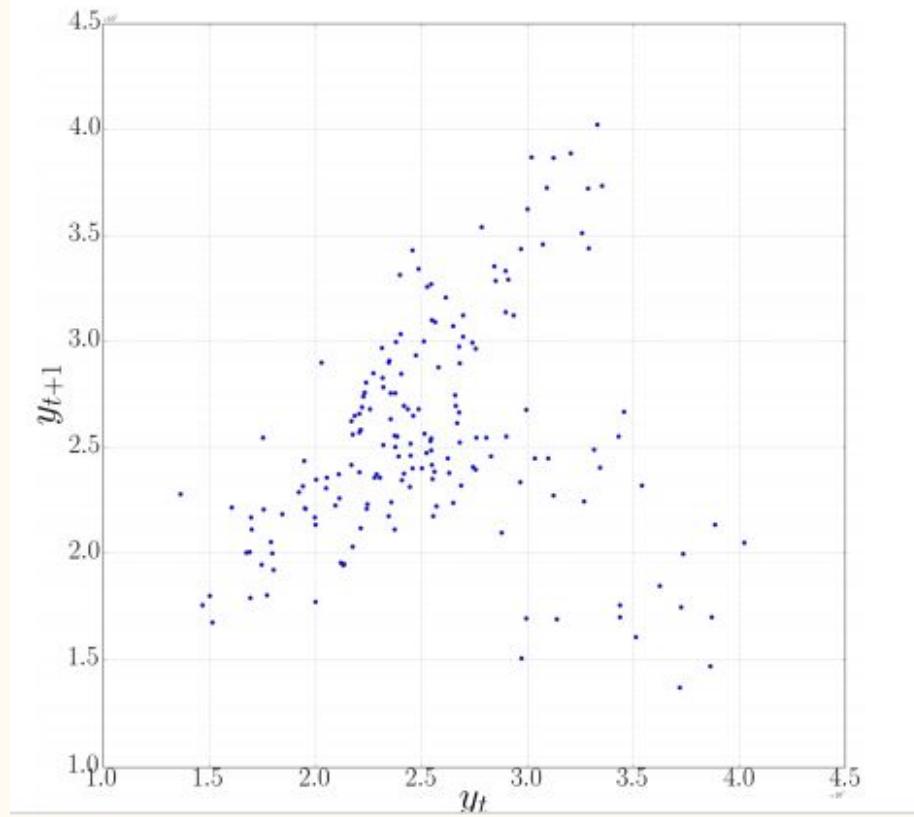


Продажи вина в австралии

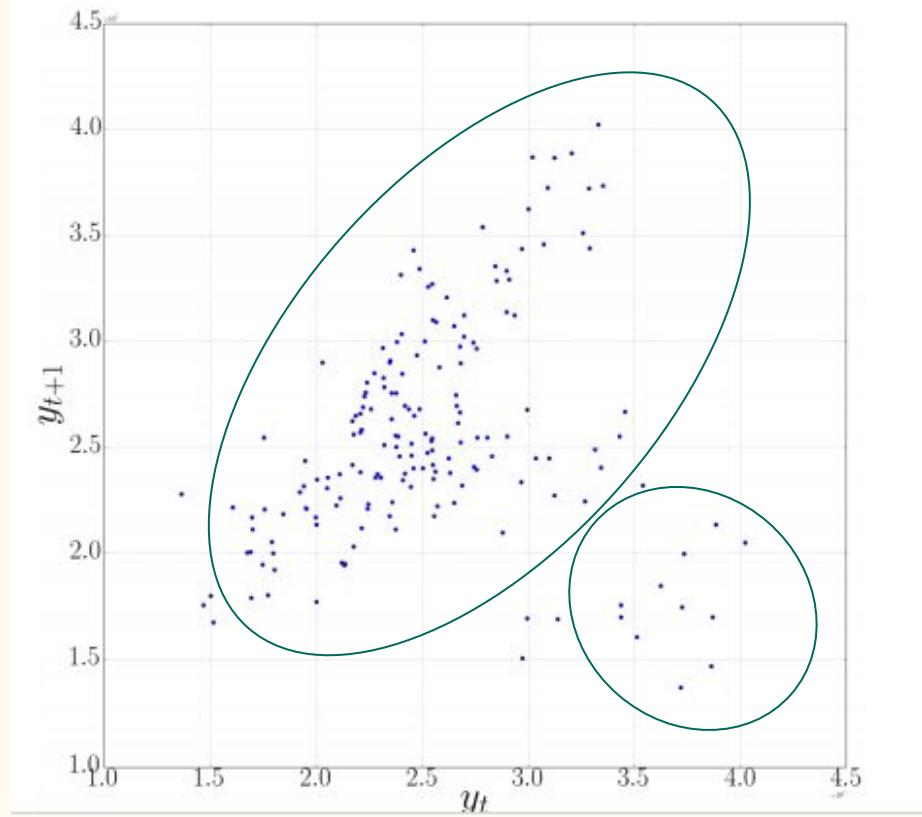


1. Каждый декабрь продажи большие
2. Каждый январь продажи падают

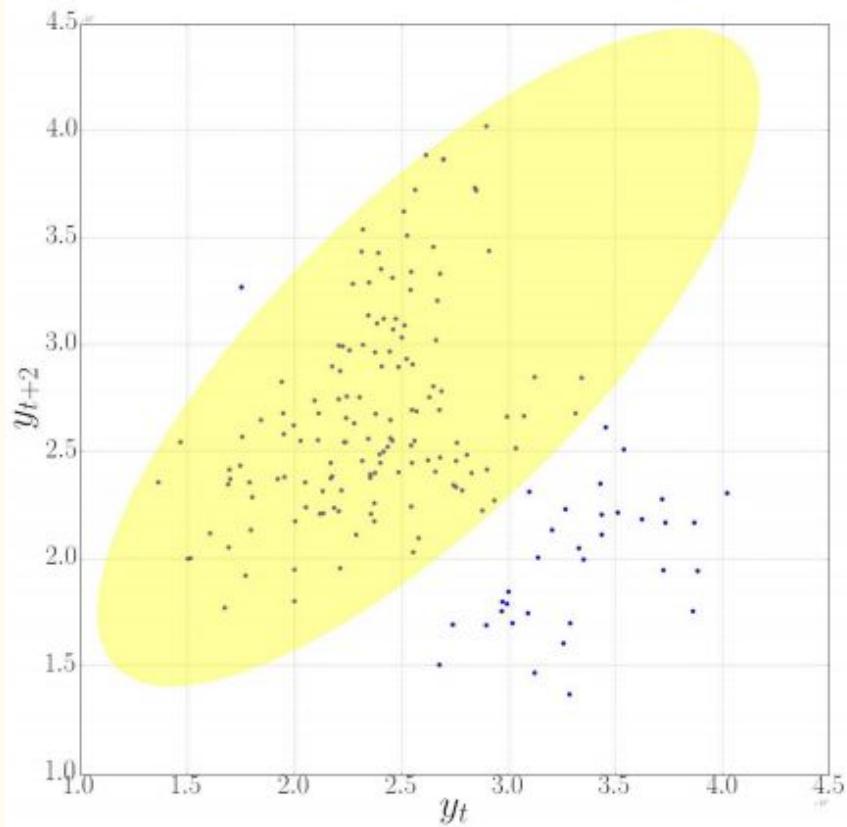
Объем продаж в соседние месяцы



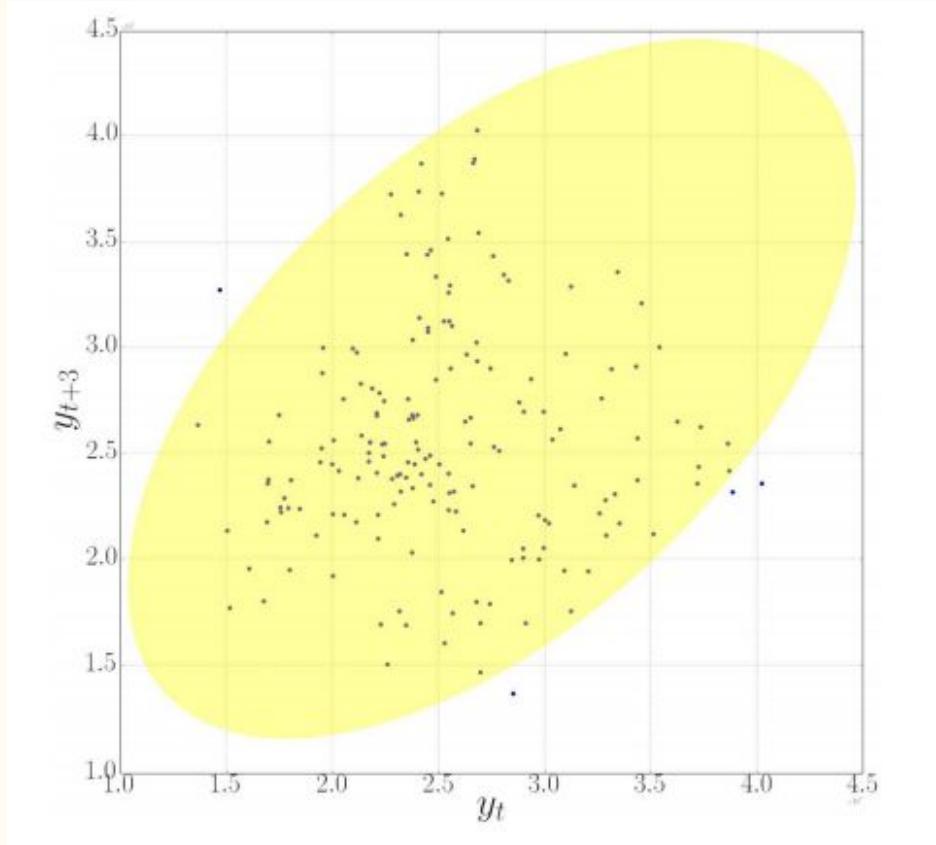
Объем продаж в соседние месяцы



Объем продаж через 1 месяц



Объем продаж через 2 месяца



Автокорреляция

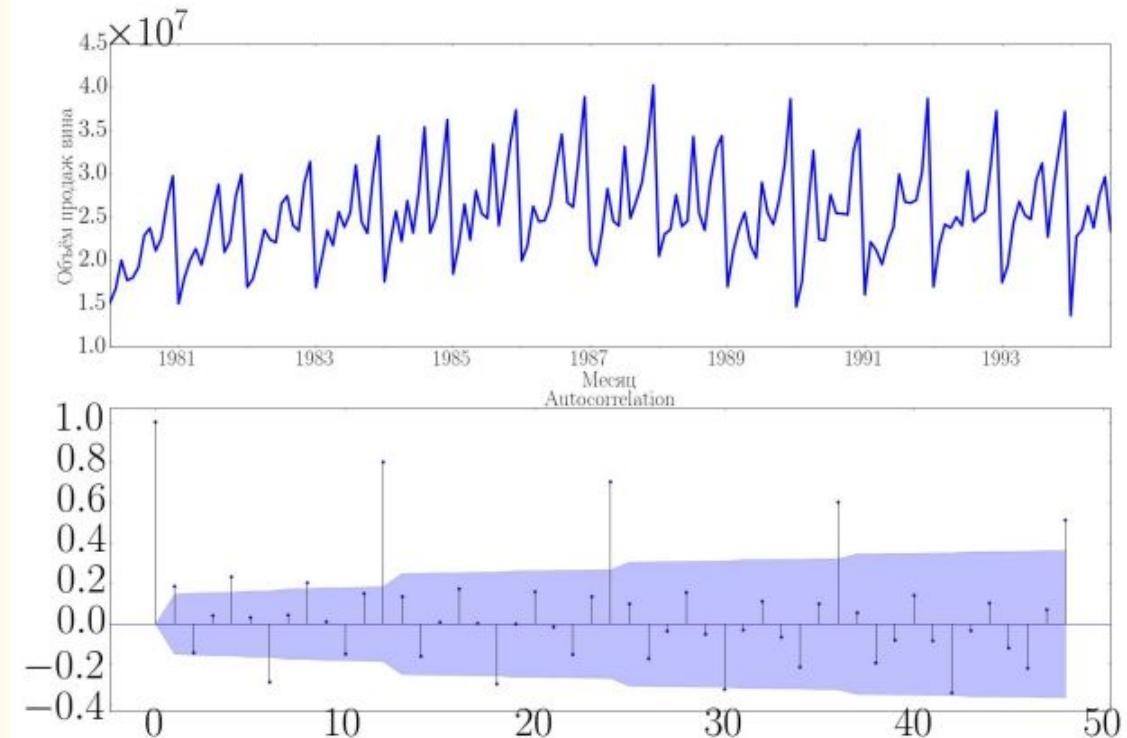
$$r_\tau = \frac{\mathbb{E}((y_t - \mathbb{E}y)(y_{t+\tau} - \mathbb{E}y))}{\mathbb{D}y}.$$

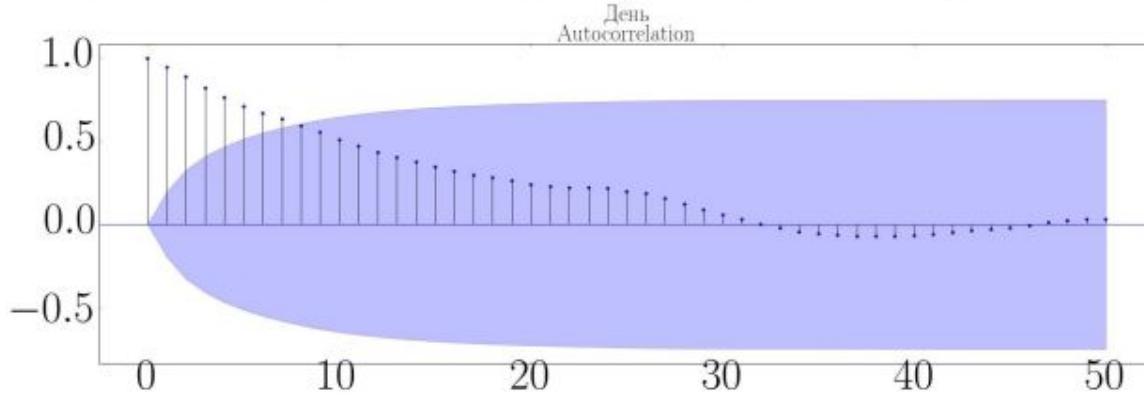
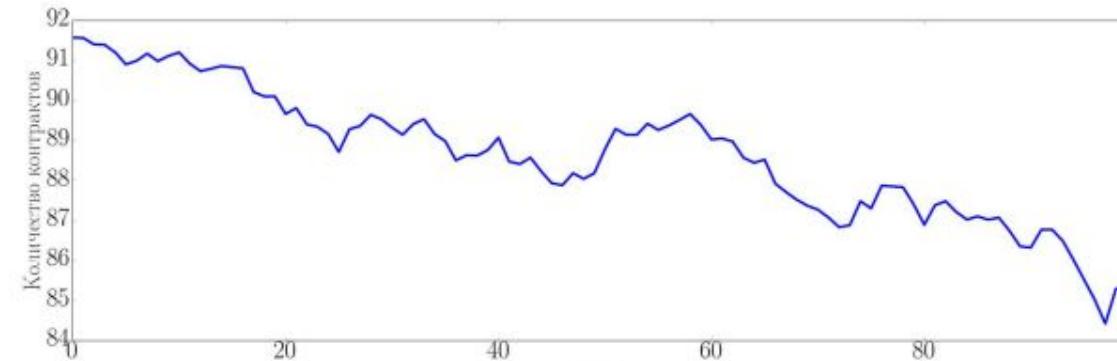
$r_\tau \in [-1, 1]$, τ — лаг автокорреляции.

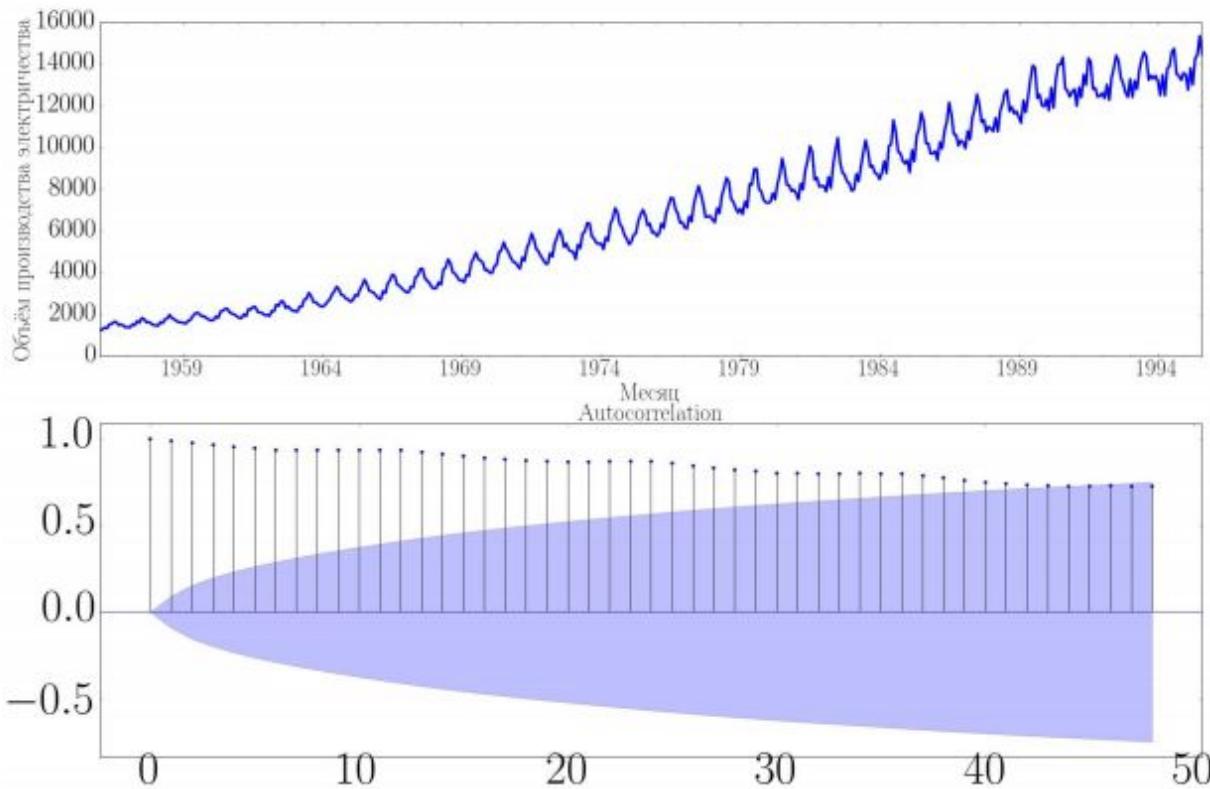
Выборочная автокорреляция:

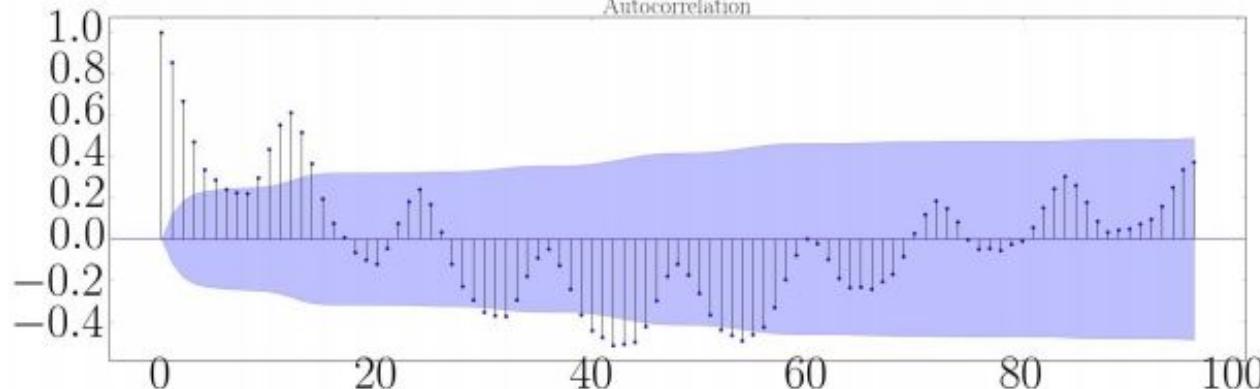
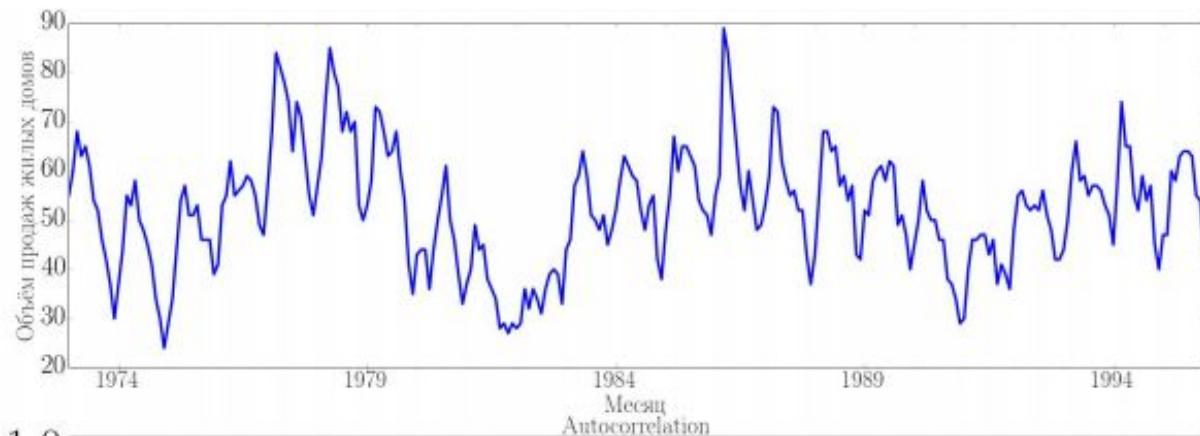
$$r_\tau = \frac{\sum_{t=1}^{T-\tau} (y_t - \bar{y})(y_{t+\tau} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

Коррелограмма

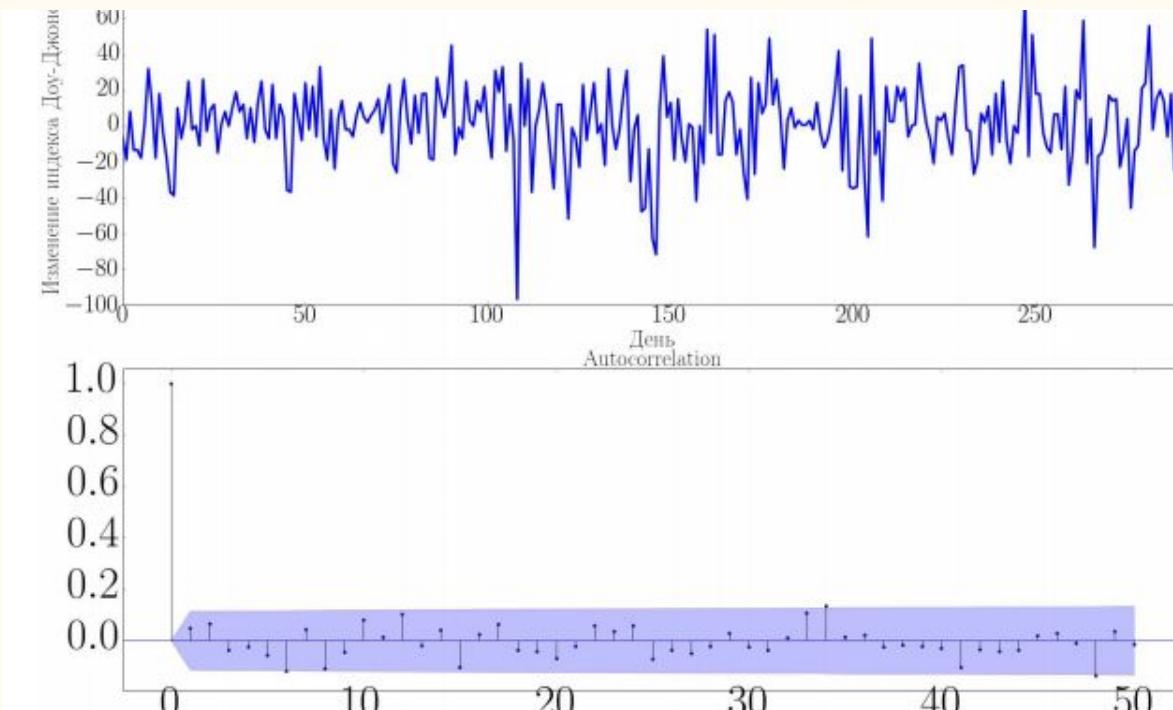




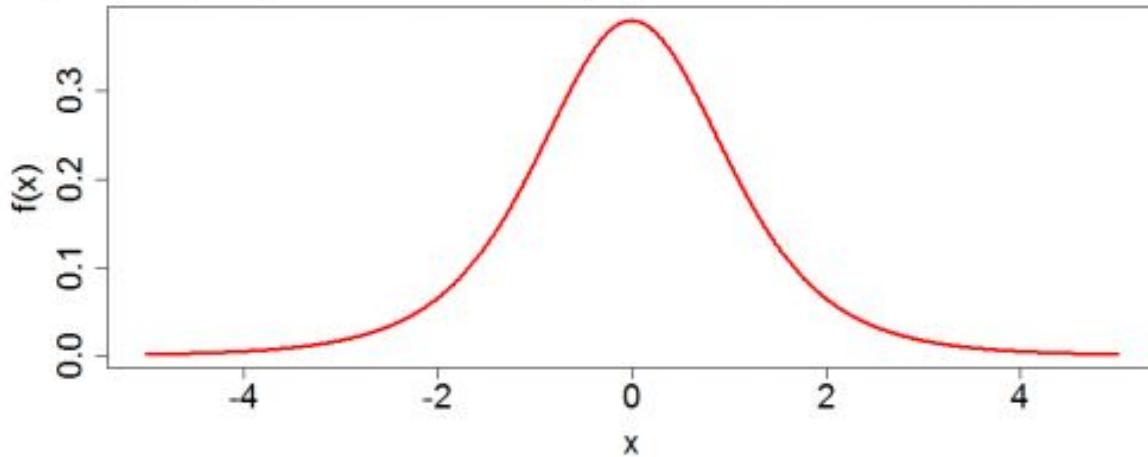




Значим ли результат?



- временной ряд: $y^T = y_1, \dots, y_T$;
- нулевая гипотеза: $H_0: r_\tau = 0$;
- альтернатива: $H_1: r_\tau \neq 0$;
- статистика: $T(y^T) = \frac{r_\tau \sqrt{T-\tau-2}}{\sqrt{1-r_\tau^2}}$;
- нулевое распределение: $T(y^T) \sim St(T - \tau - 2)$ при H_0 .



Q-критерий Льюнга-Бокса

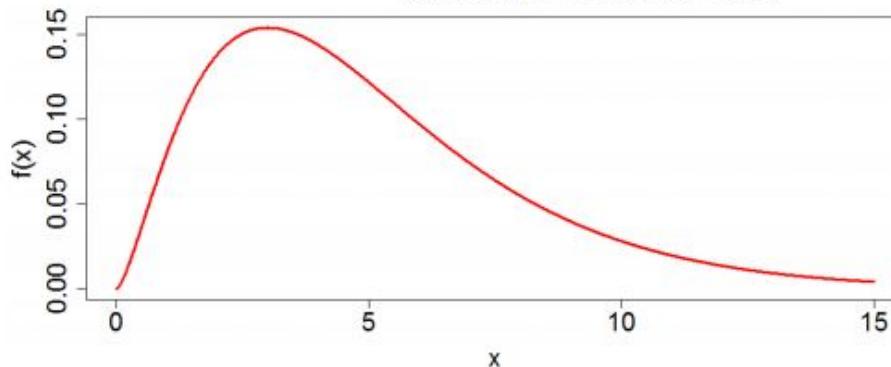
ряд ошибок прогноза: $\varepsilon^T = \varepsilon_1, \dots, \varepsilon_T$;

нулевая гипотеза: $H_0: r_1 = \dots = r_L = 0$;

альтернатива: $H_1: H_0$ неверна;

статистика: $Q(\varepsilon^T) = T(T+2) \sum_{\tau=1}^L \frac{r_\tau^2}{T-\tau}$;

нулевое распределение: χ_{L-K}^2 , K — число настраиваемых параметров модели ряда.



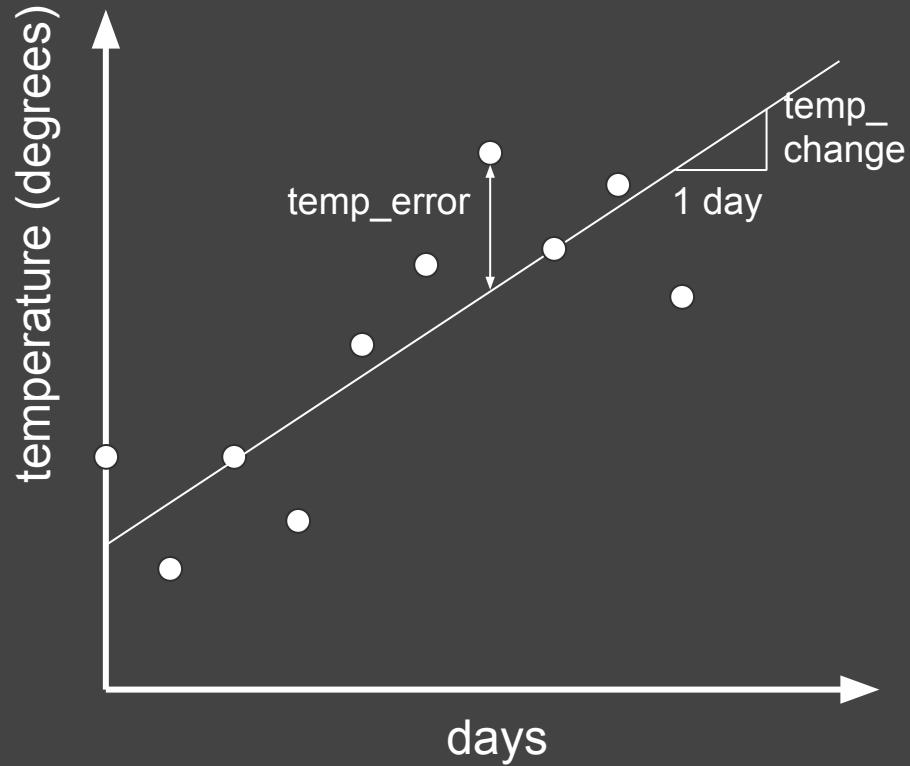
- Действительно ли данные случайны?
- Наблюдение связано со смежным наблюдением? Наблюдение связано с вдвое удаленным наблюдением? (и т.д.).
- Имеет ли наблюдаемый временной ряд белый шум?
- Действительно ли наблюдаемый временной ряд является синусоидальным?
- Действительно ли наблюдаемый временной ряд авторегрессивен?
- Какова соответствующая модель для наблюдаемого временного ряда?

Линейная регрессия

```
temps = np.array([68.2, 65.7, ...])
n_days = temps.size
days = np.arange(n_days)

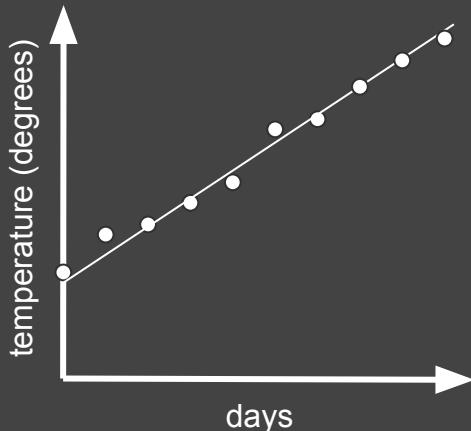
def find_temp_error(
    days, temps, intercept, slope):
    estimates = intercept + slope * days
    temp_error = temps - estimates
    return temp_error

slope, intercept = np.polyfit(
    days, temps, 1)
temp_error = find_temp_error(
    days, temps, intercept, slope)
```

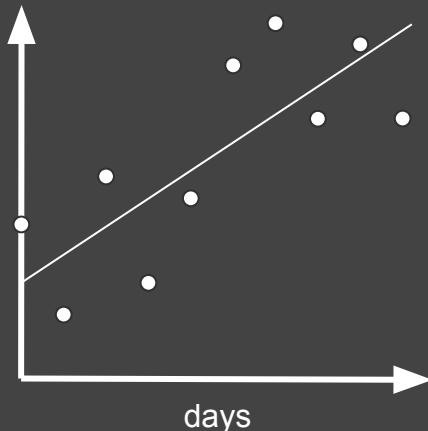


Корреляция

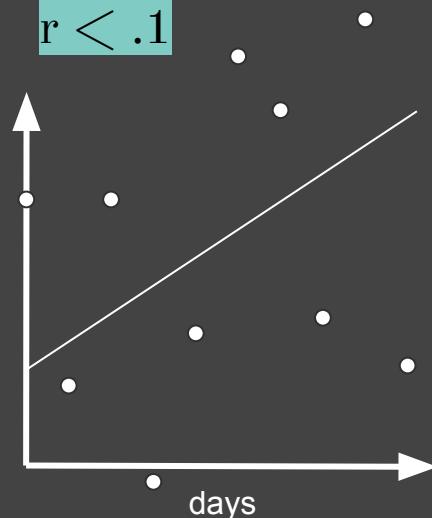
High correlation
 $r > .9$



Moderate correlation
 $.4 > r > .7$



Low correlation
 $r < .1$



Корреляция

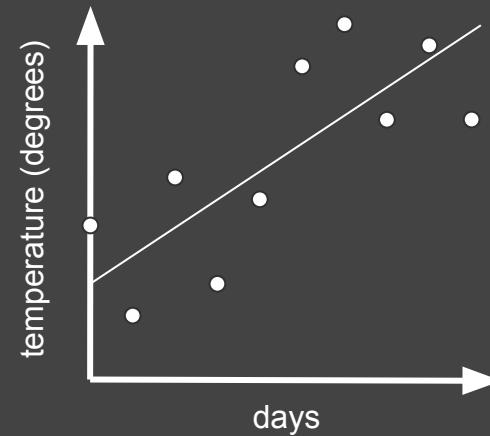
```
temps = np.array([68.2, 65.7, ...])
n_days = temps.size
days = np.arange(n_days)

np.corrcoef(days, temps)

> [[1.00, 0.79],
   [0.79, 1.00]]

np.corrcoef(days, temps)[0, 1]

> 0.79
```



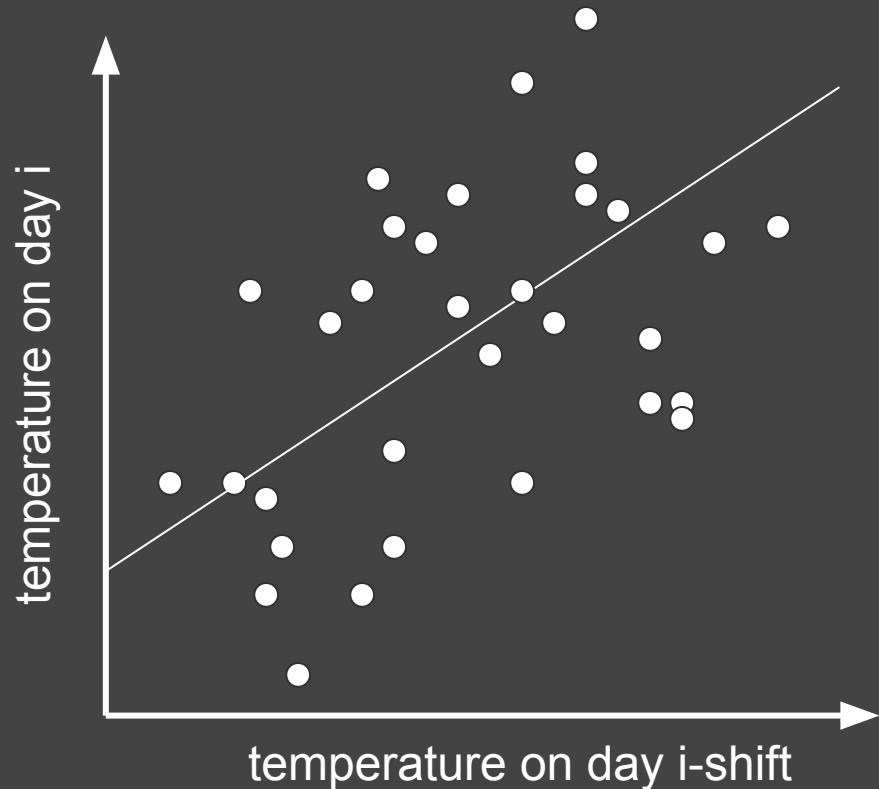
correlations

	days	temps
days	1.00	0.79
temps	0.79	1.00

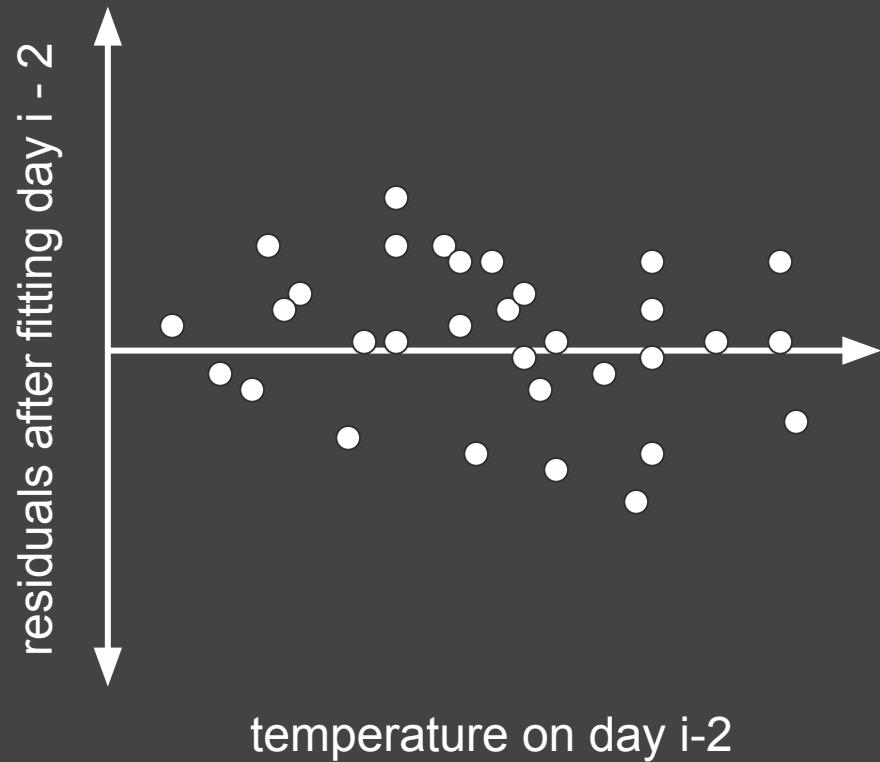
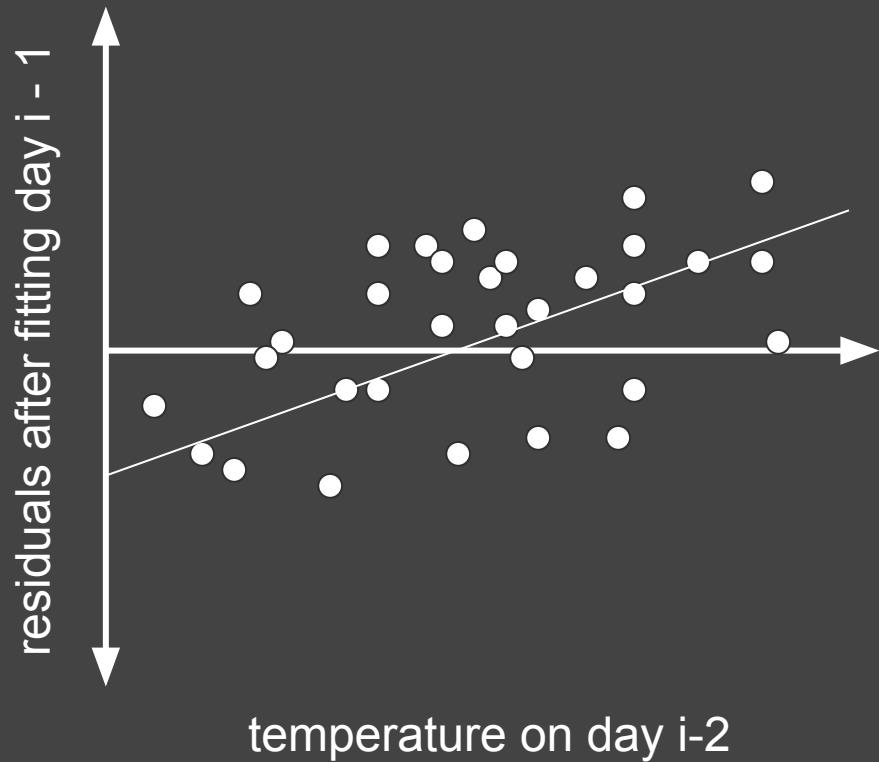
Автокорреляция

```
temp = np.array([68.2, 65.7, ...])
shift = 4

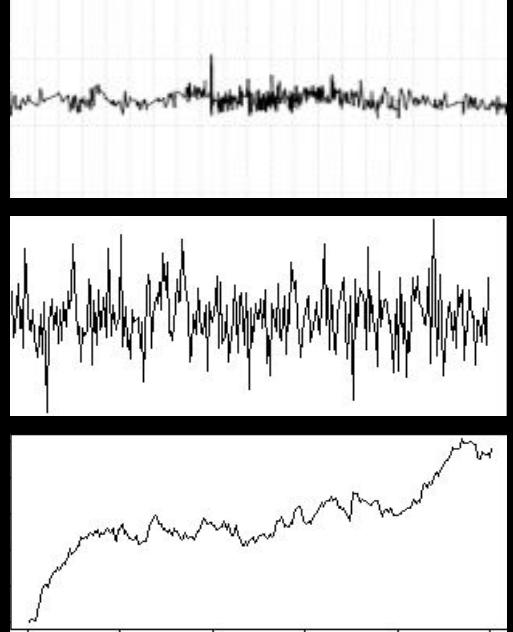
# autocorrelation
# Answers the question "How strongly
# related are today's temperatures with
# those of <shift> days ago?"
np.corrcoef(temp[:-shift],
            temp[shift:])[0, 1]
```



Частичная автокорреляция



Стационарность

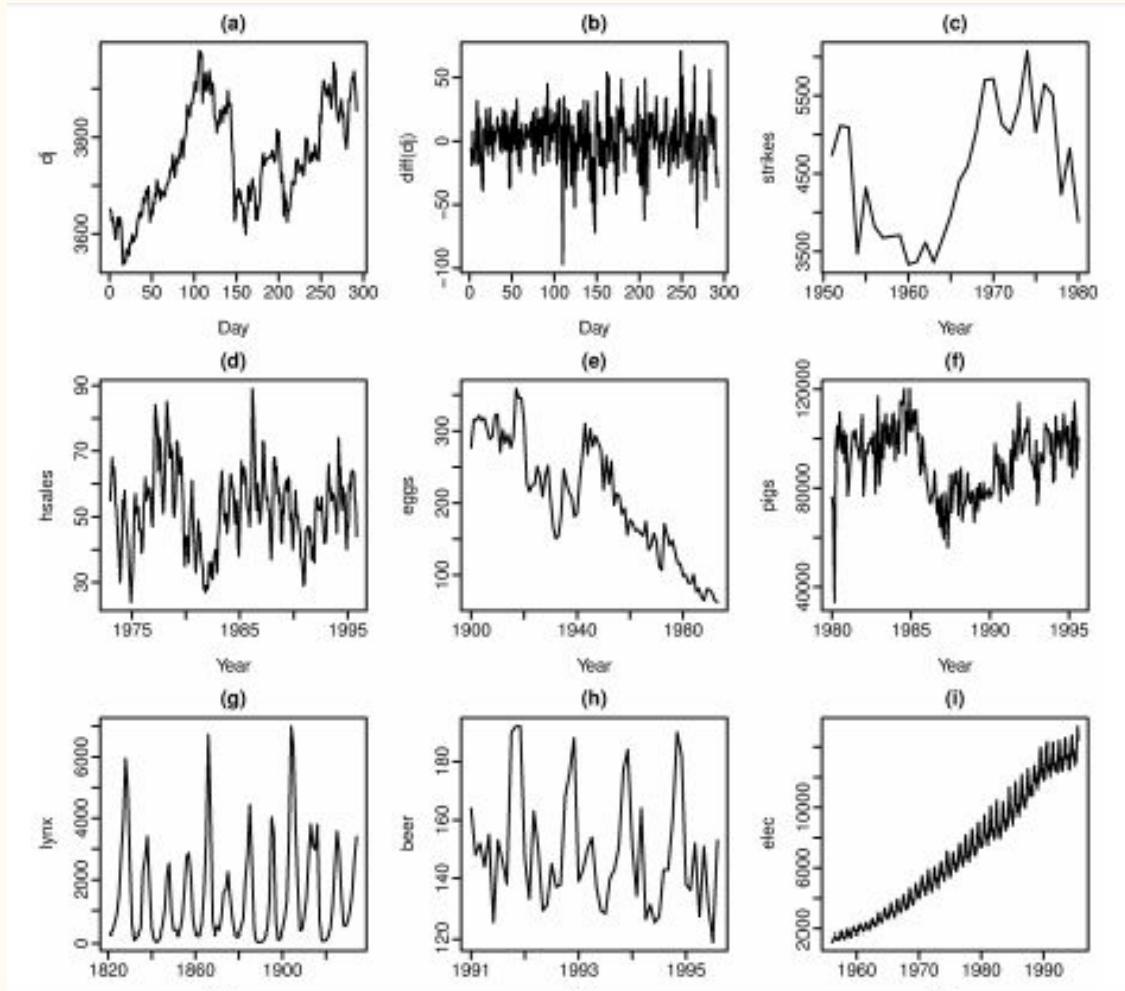


Стационарность

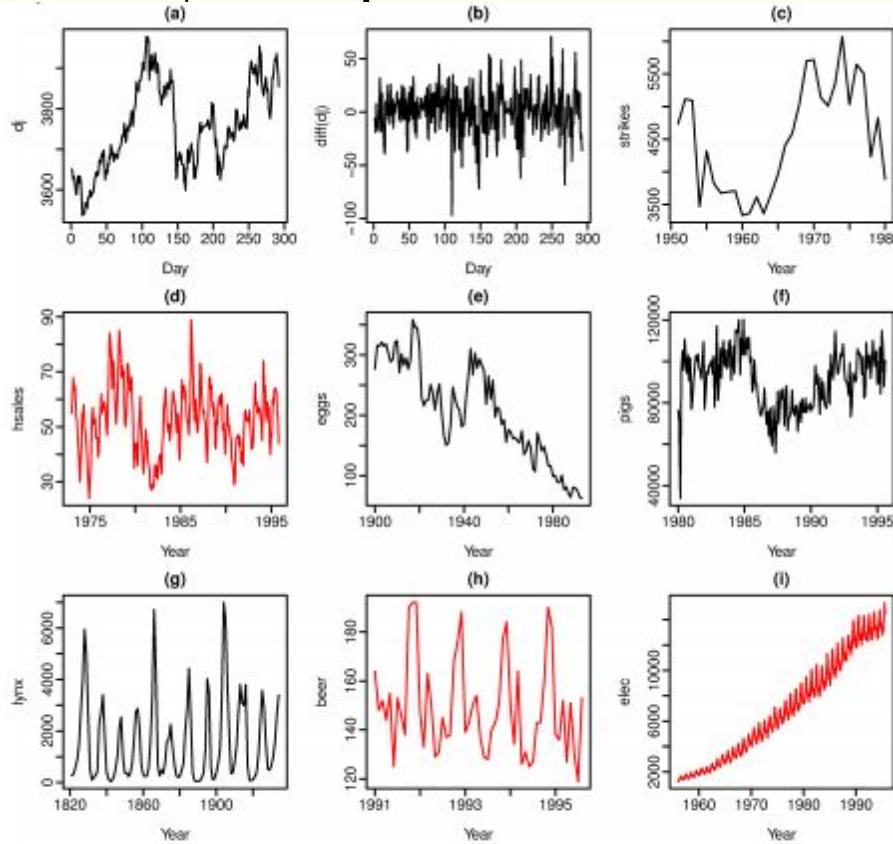
Ряд y_1, \dots, y_T -- стационарен, если $\forall s$ распределение y_t, \dots, y_{t+s} не зависит от t , т. е. его свойства не зависят от времени.

1. тренд \Rightarrow нестационарность
2. сезонность $\not\Rightarrow$ нестационарность цикл

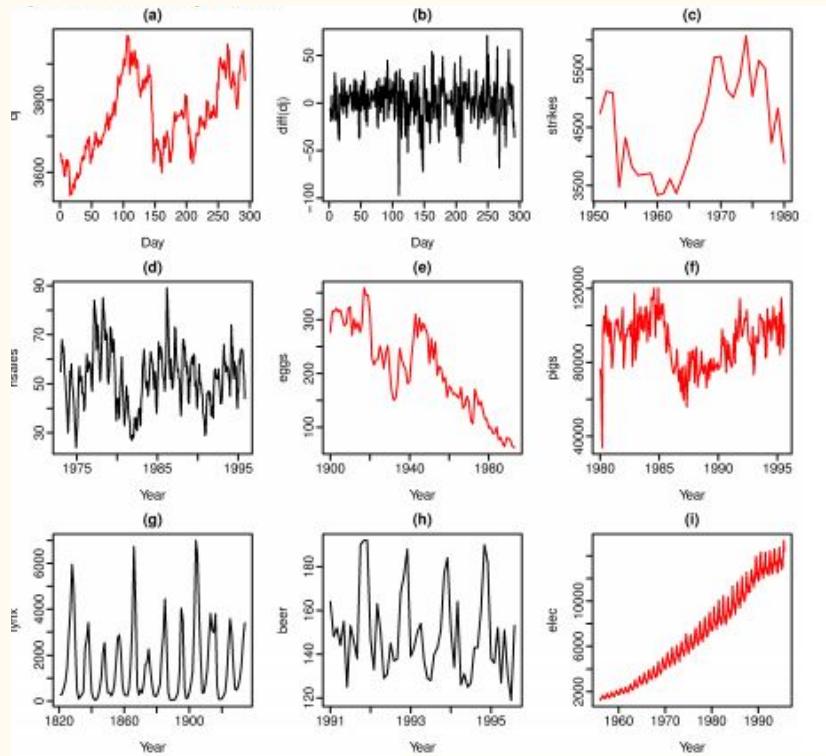
нестационарность (нельзя предсказать заранее, где будут находиться максимумы и минимумы)



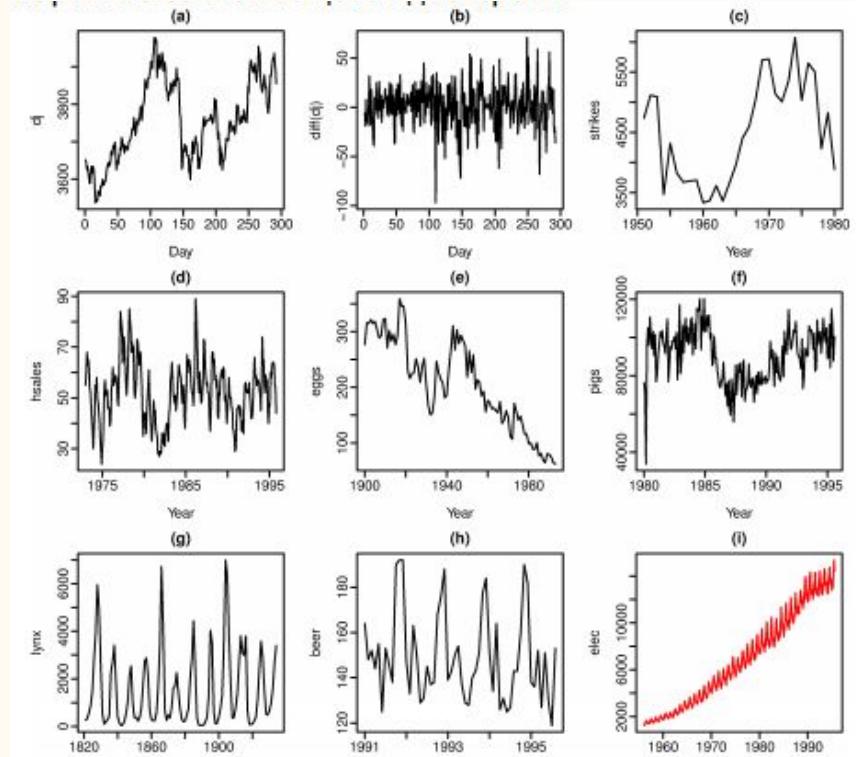
Сезонная нестационарность

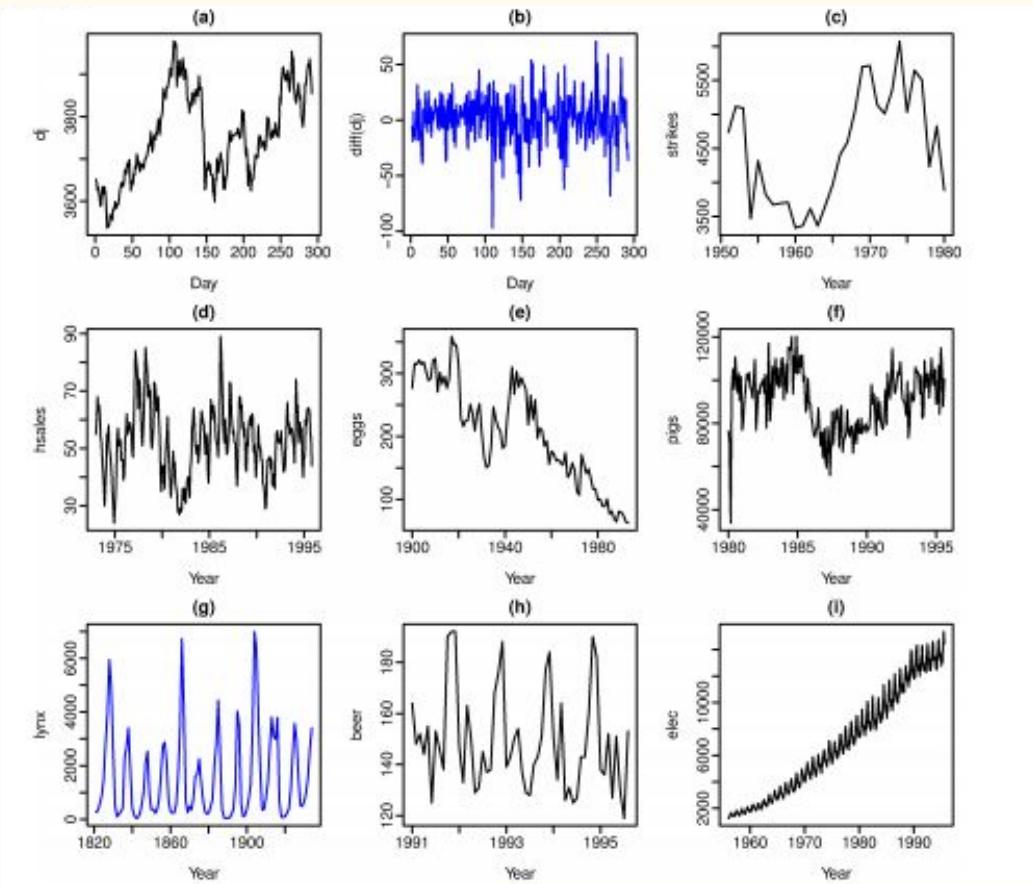


Нестационарность из-за тренда



Изменяющаяся дисперсия





Критерий KPSS (Kwiatkowski-Philips-Schmidt-Shin)

ряд: $y = y_1, \dots, y_T$;

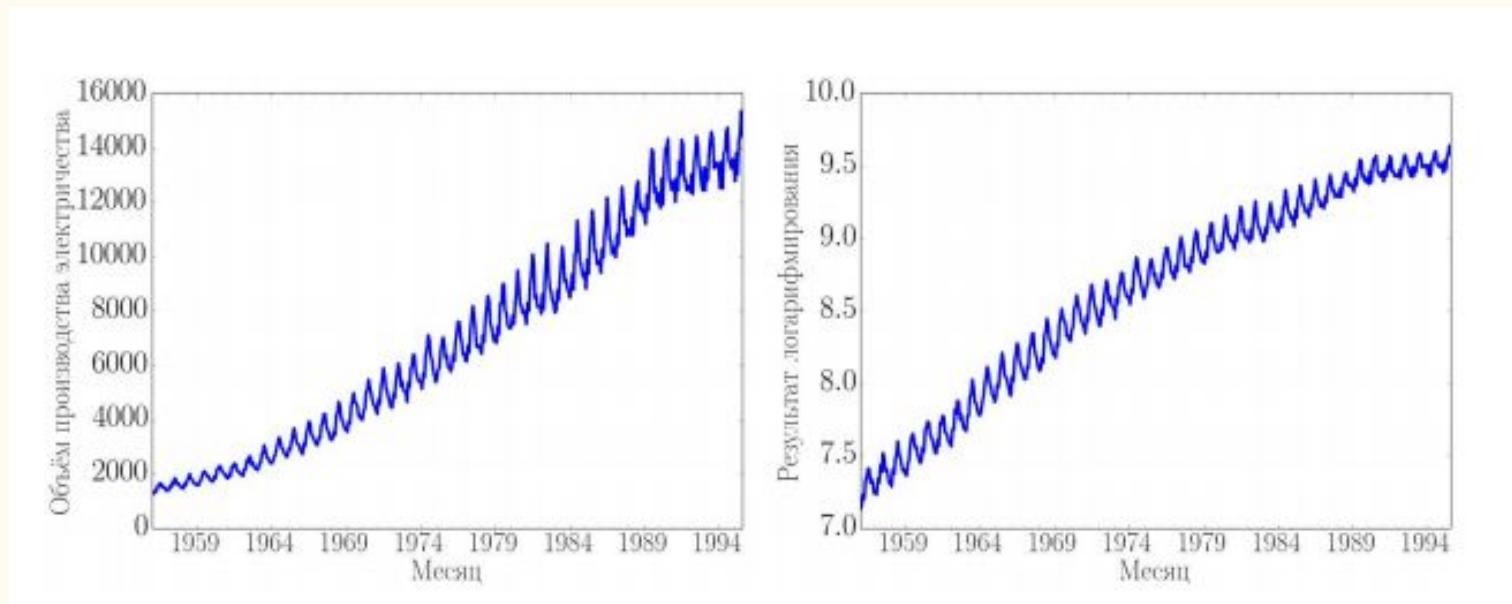
нулевая гипотеза: H_0 : ряд y стационарен;

альтернатива: H_1 : ряд y описывается моделью вида $y_t = \alpha y_{t-1}$;

статистика: $KPSS(y) = \frac{1}{T^2} \sum_{i=1}^T \left(\sum_{t=1}^i y_t \right)^2 / \lambda^2$;

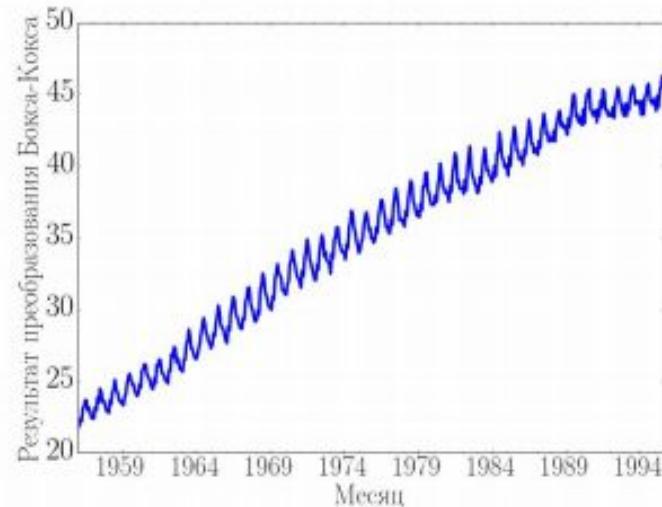
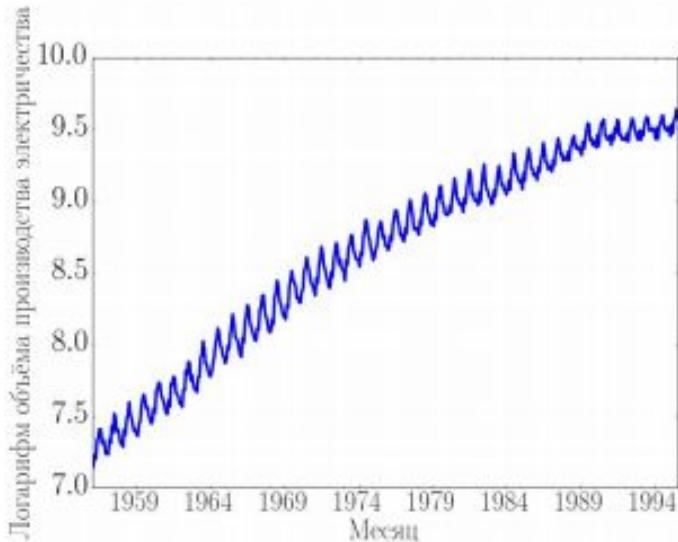
$KPSS(y)$ при H_0 имеет табличное распределение.

Стабилизация дисперсии (логарифмирование)

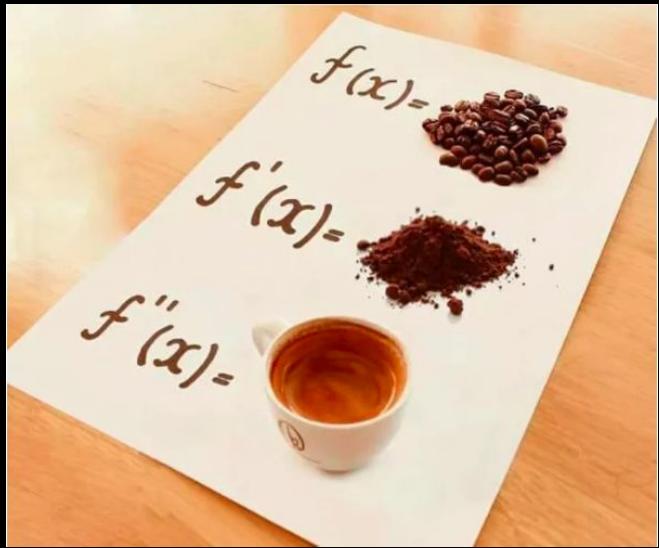


Преобразования Бокса-Кокса:

$$y'_t = \begin{cases} \ln y_t, & \lambda = 0, \\ (y_t^\lambda - 1) / \lambda, & \lambda \neq 0. \end{cases}$$



Дифференцирование ряда



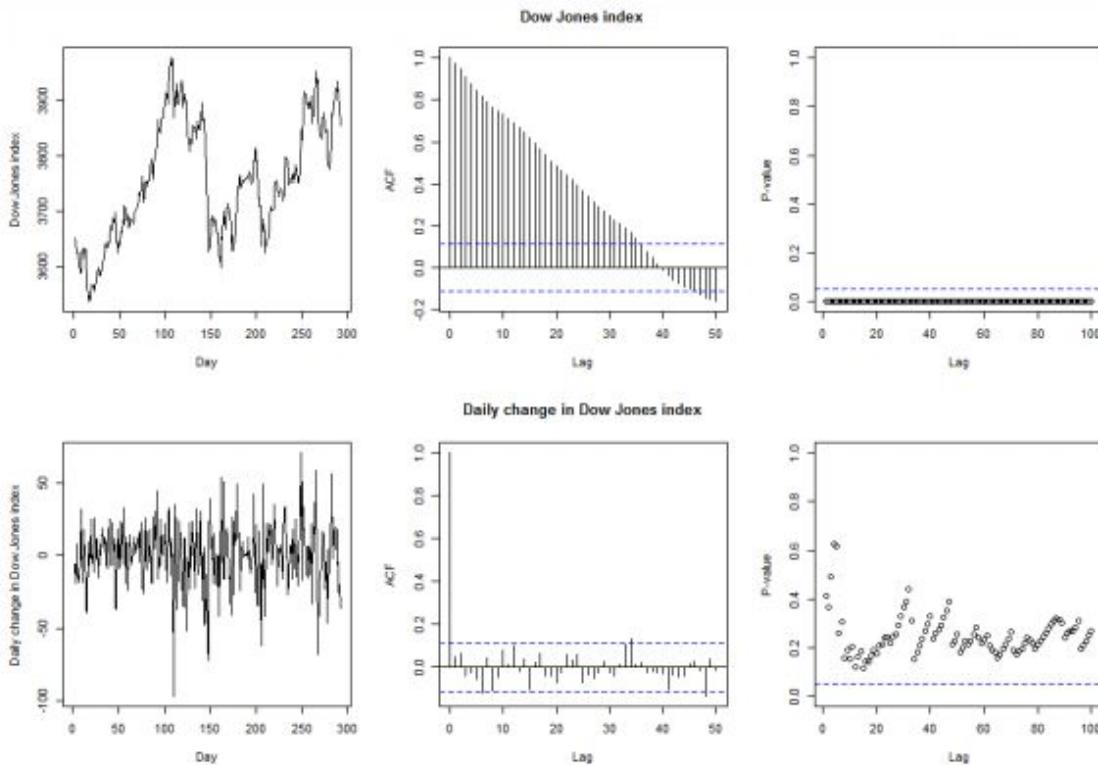
Дифференцирование ряда — переход к попарным разностям его соседних значений:

$$y_1, \dots, y_T \longrightarrow y'_2, \dots, y'_T,$$
$$y'_t = y_t - y_{t-1}.$$

Дифференцированием можно стабилизировать среднее значение ряда и избавиться от тренда и сезонности.

Может применяться неоднократное дифференцирование; например, для второго порядка:

$$y_1, \dots, y_T \longrightarrow y'_2, \dots, y'_T \longrightarrow y''_3, \dots, y''_T,$$
$$y''_t = y'_t - y'_{t-1} = y_t - 2y_{t-1} + y_{t-2}.$$



Критерий KPSS: для исходного ряда $p < 0.01$, для ряда первых разностей — $p > 0.1$.

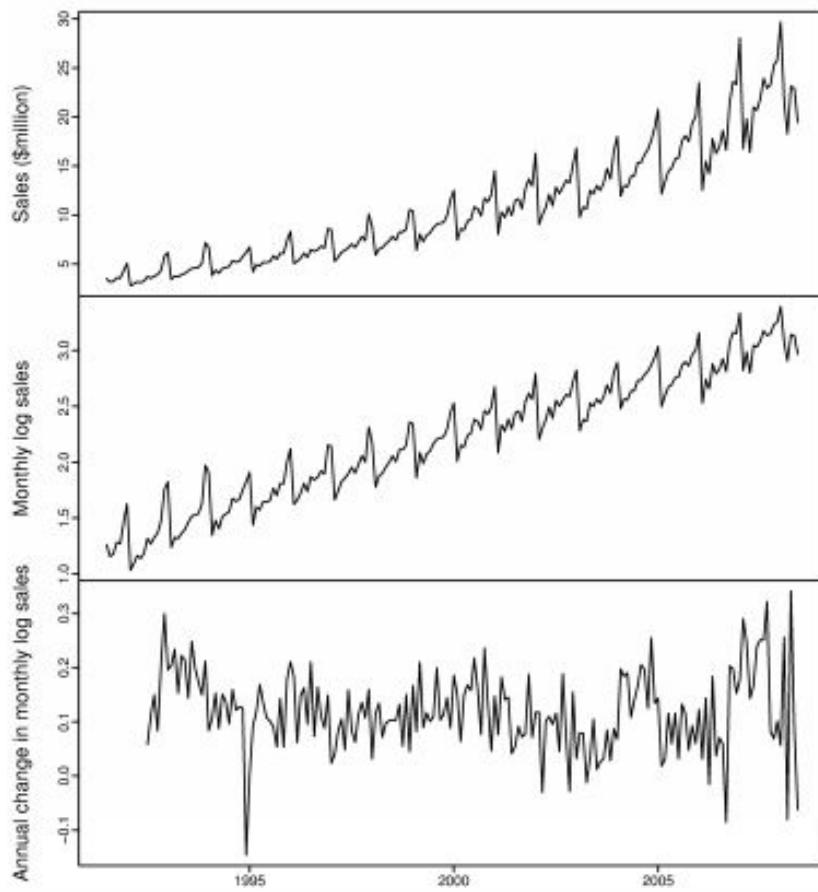
Сезонное дифференцирование

Сезонное дифференцирование ряда — переход к попарным разностям его значений в соседних сезонах:

$$y_1, \dots, y_T \longrightarrow y'_{s+1}, \dots, y'_T,$$

$$y'_t = y_t - y_{t-s}.$$

Antidiabetic drug sales



Критерий

KPSS: для исходного ряда $p < 0.01$, для логарифмированного — $p < 0.01$, после сезонного дифференцирования — $p > 0.1$.

Авторегрессия

Что если делать регрессию ряда на собственные значения в прошлом?

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

где y_t — стационарный ряд, ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2 .

Другой способ записи:

$$\phi(B) y_t = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) y_t = \varepsilon_t,$$

где B — разностный оператор ($By_t = y_{t-1}$).

Модель авторегрессии порядка p ($AR(p)$):

y_t — линейная комбинация p предыдущих значений ряда и шумовой компоненты.

Чтобы ряд AR(p) был стационарным, должны выполняться ограничения на коэффициенты. Например,

- в AR(1) необходимо $-1 < \phi_1 < 1$;
- в AR(2) необходимо $-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$, $\phi_2 - \phi_1 < 1$.

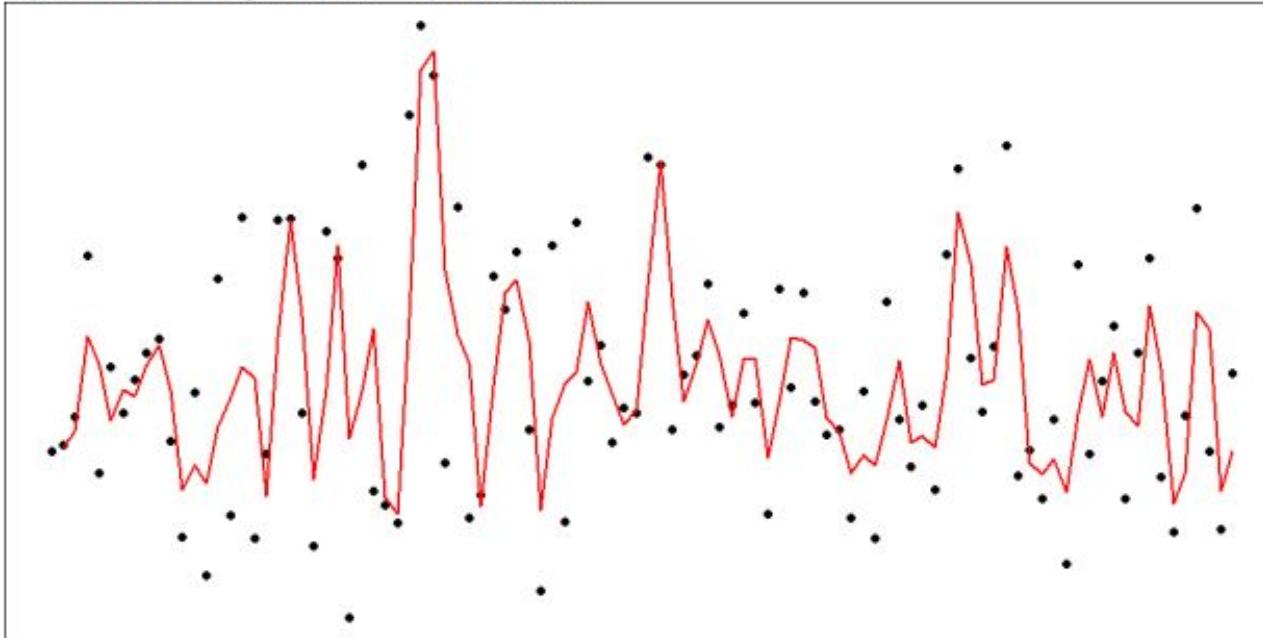
С ростом p вид ограничений усложняется.

Скользящее среднее

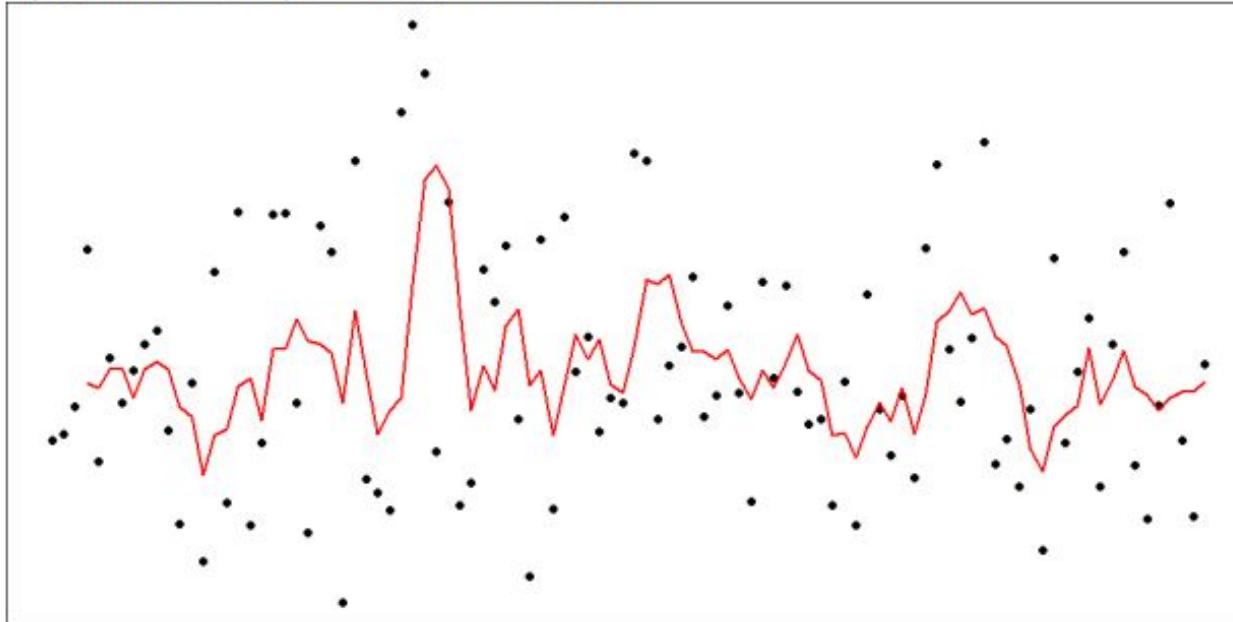
Пусть у нас есть независимый одинаково распределённый шум ε_t :



Среднее по двум соседним точкам:



Среднее по четырём соседним точкам:



Обобщим и добавим веса:

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

где y_t — стационарный ряд, ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2 .

Другой способ записи:

$$y_t = \theta(B) \varepsilon_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) \varepsilon_t,$$

где B — разностный оператор.

Модель скользящего среднего порядка q ($MA(q)$):

y_t — линейная комбинация q последних значений шумовой компоненты.

Чтобы ряд модель MA(q) была обратимой, должны выполняться ограничения на коэффициенты. Например,

- в MA(1) необходимо $-1 < \theta_1 < 1$;
- в MA(2) необходимо $-1 < \theta_2 < 1$, $\theta_1 + \theta_2 > -1$, $\theta_1 - \theta_2 < 1$.

С ростом q вид ограничений усложняется.

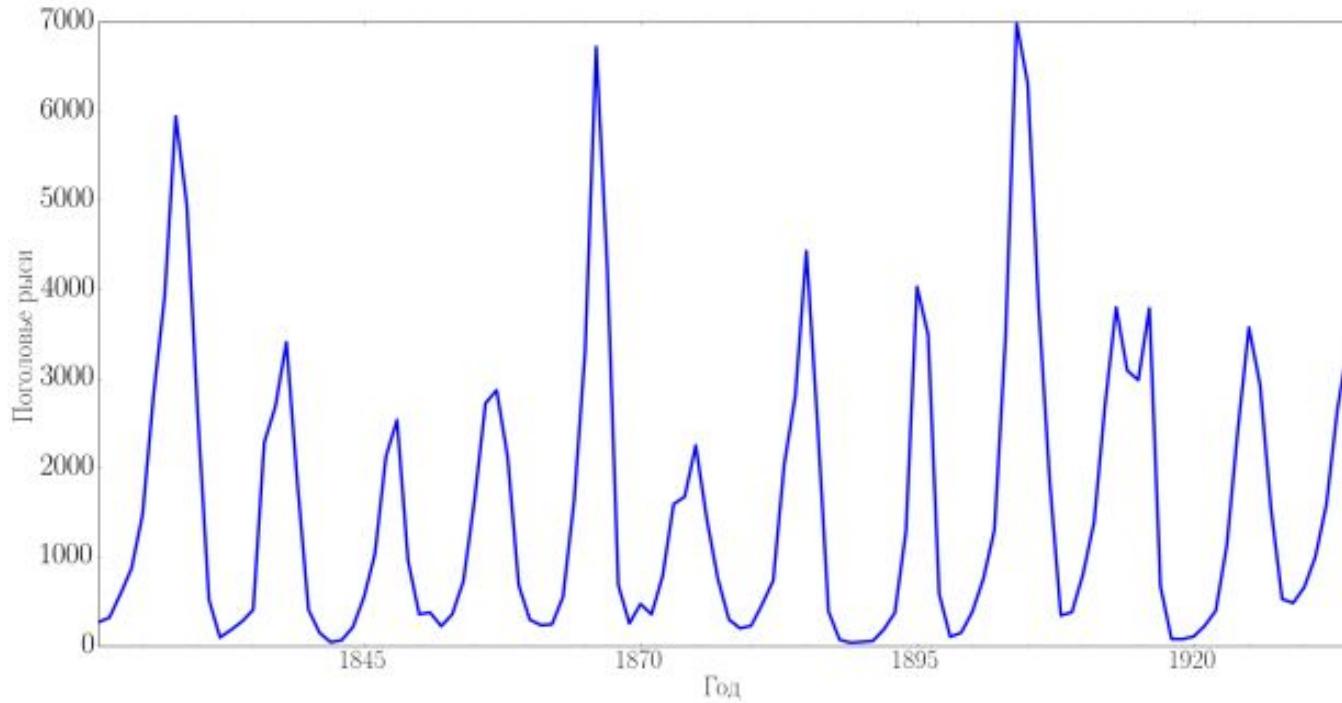
$$ARMA(p, q): \quad y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

где y_t — стационарный ряд, ε_t — гауссов белый шум с нулевым средним и постоянной дисперсией σ_ε^2 .

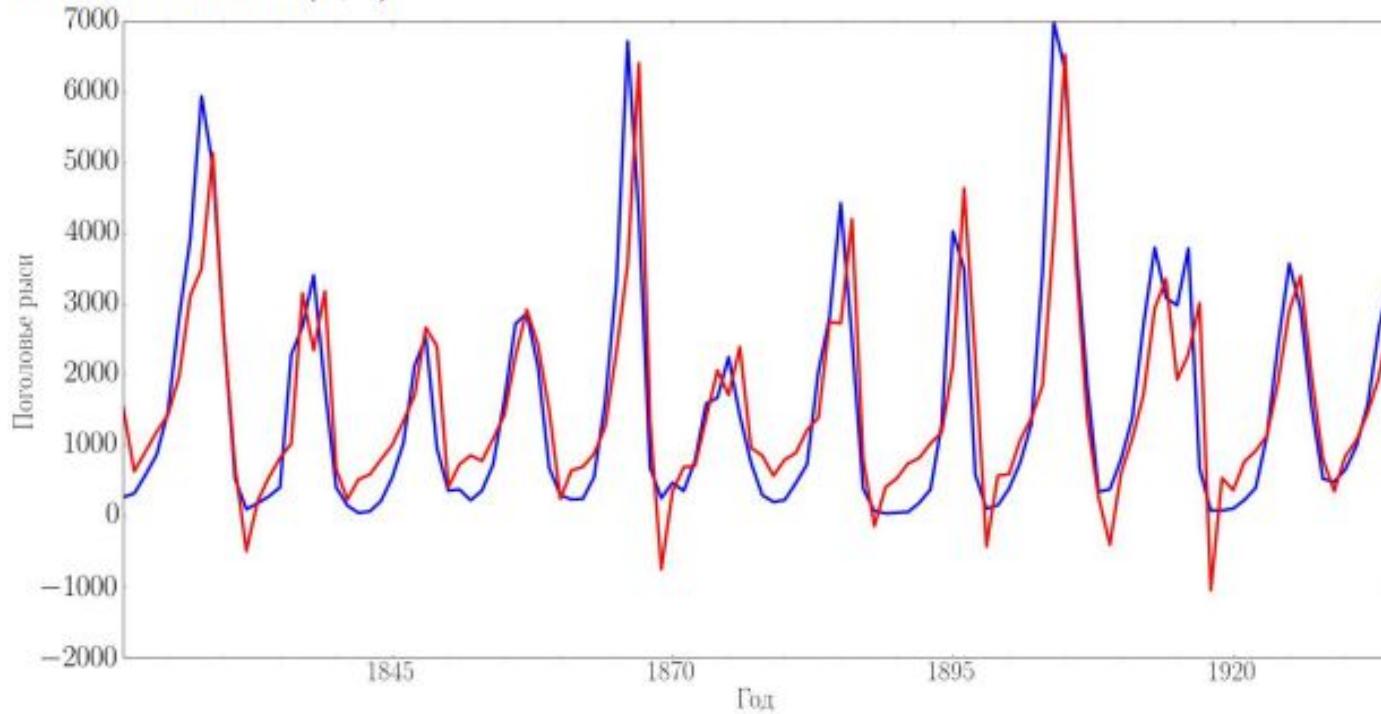
Другой способ записи:

$$\phi(B) y_t = \theta(B) \varepsilon_t.$$

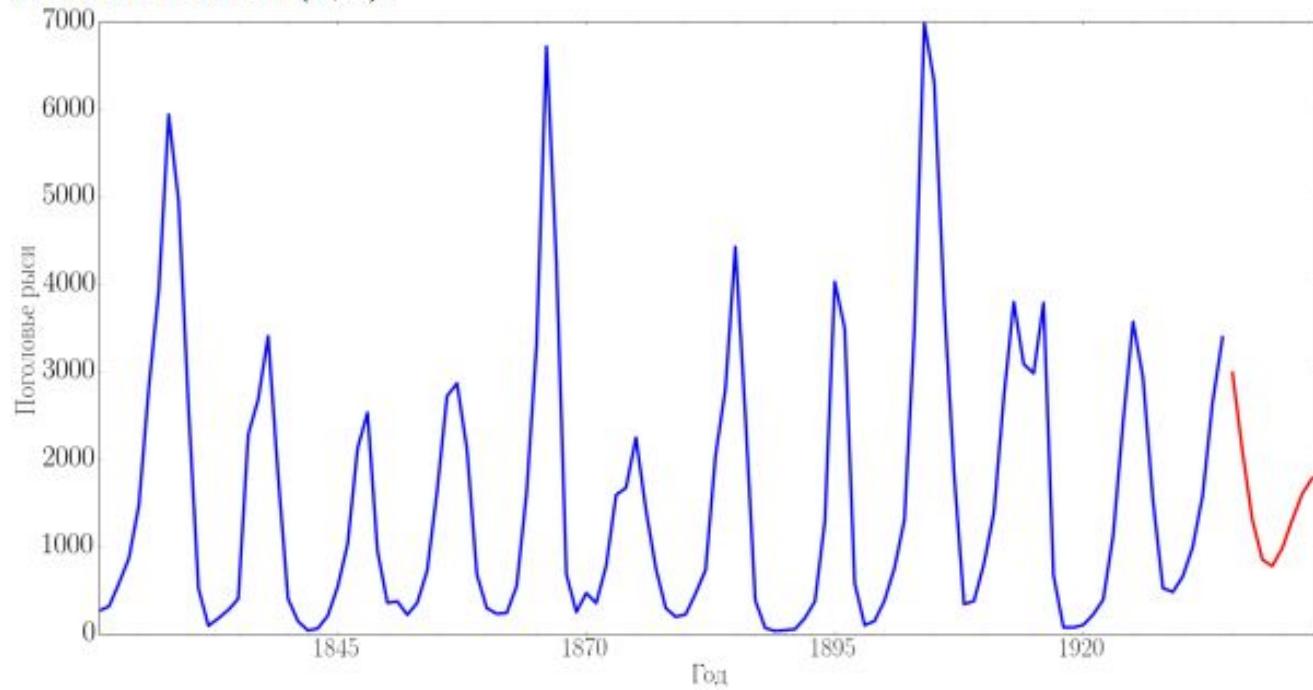
Теорема Вольда: любой стационарный ряд может быть описан моделью $ARMA(p, q)$ с любой точностью.

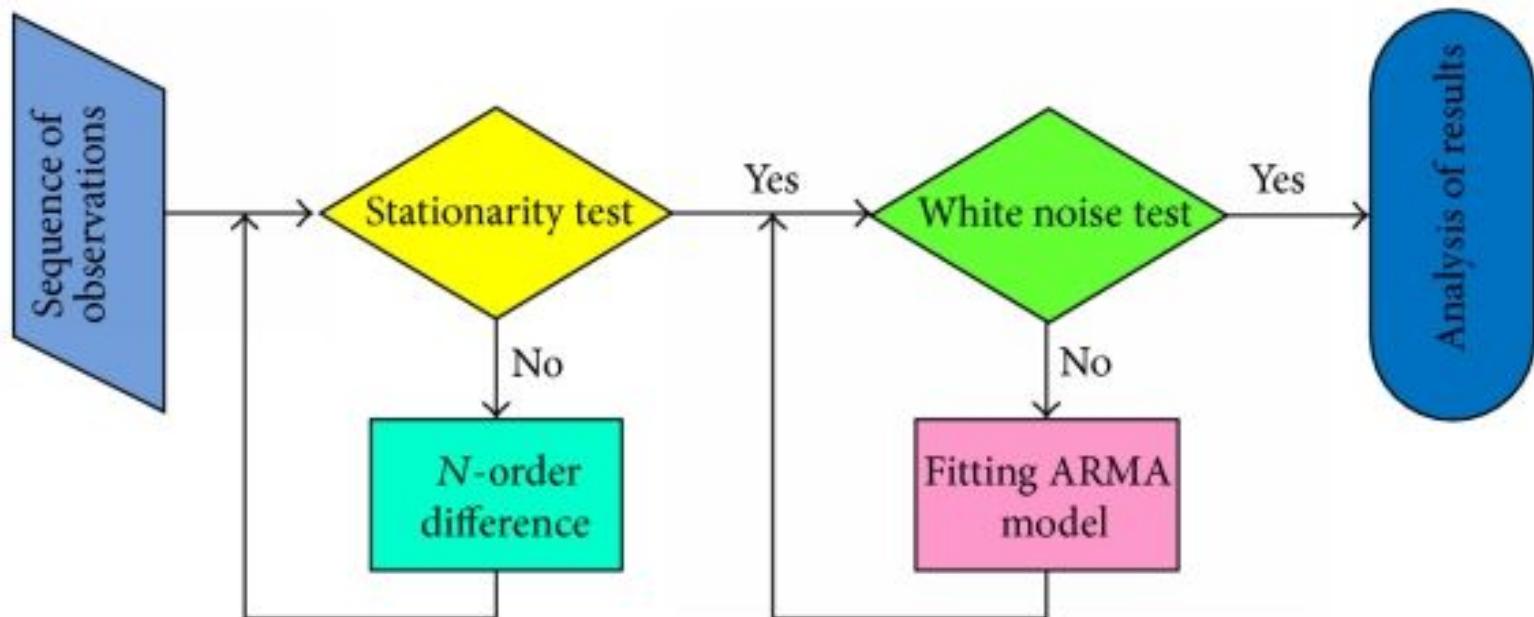


Модель $ARMA(2, 2)$:

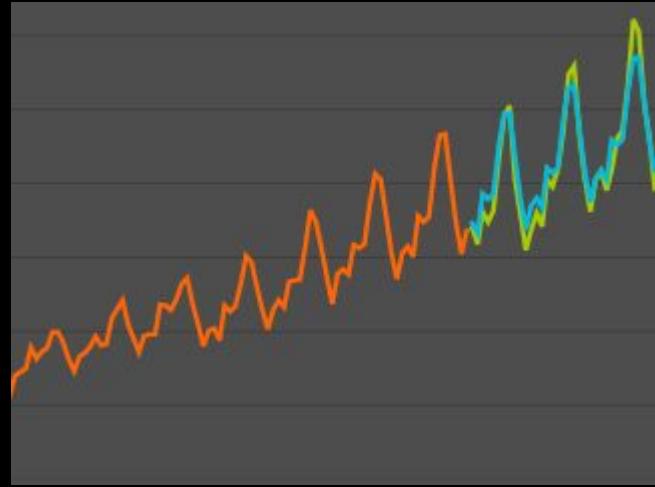


Модель $ARMA(2, 2)$:





ARIMA



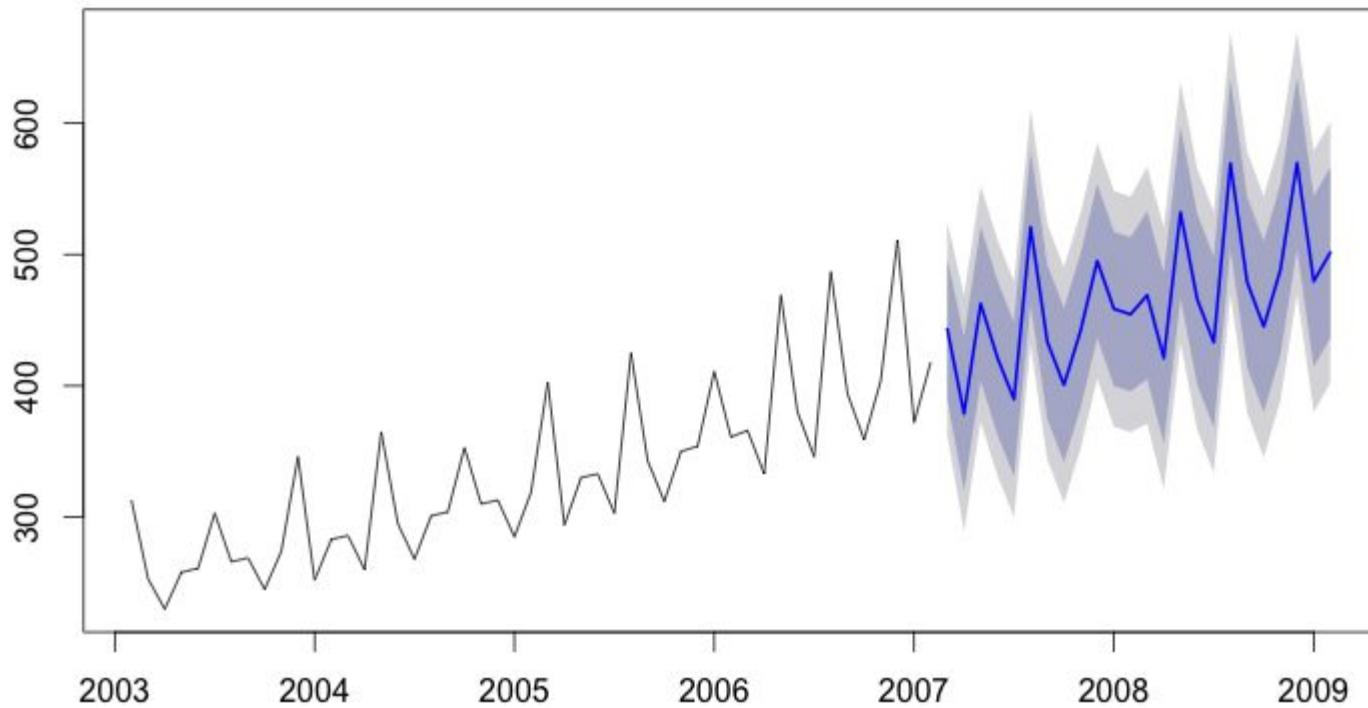
Ряд описывается моделью $ARIMA(p, d, q)$, если ряд его разностей

$$\nabla^d y_t = (1 - B)^d y_t$$

описывается моделью $ARMA(p, q)$.

$$\phi(B) \nabla^d y_t = \theta(B) \varepsilon_t.$$

Forecasts from ARIMA(0,0,1)(1,1,0)[12] with drift



Пусть ряд имеет сезонный период длины S .

Возьмём модель $ARMA(p, q)$:

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

и добавим P авторегрессионных компонент:

$$+\phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \cdots + \phi_{PS} y_{t-PS}$$

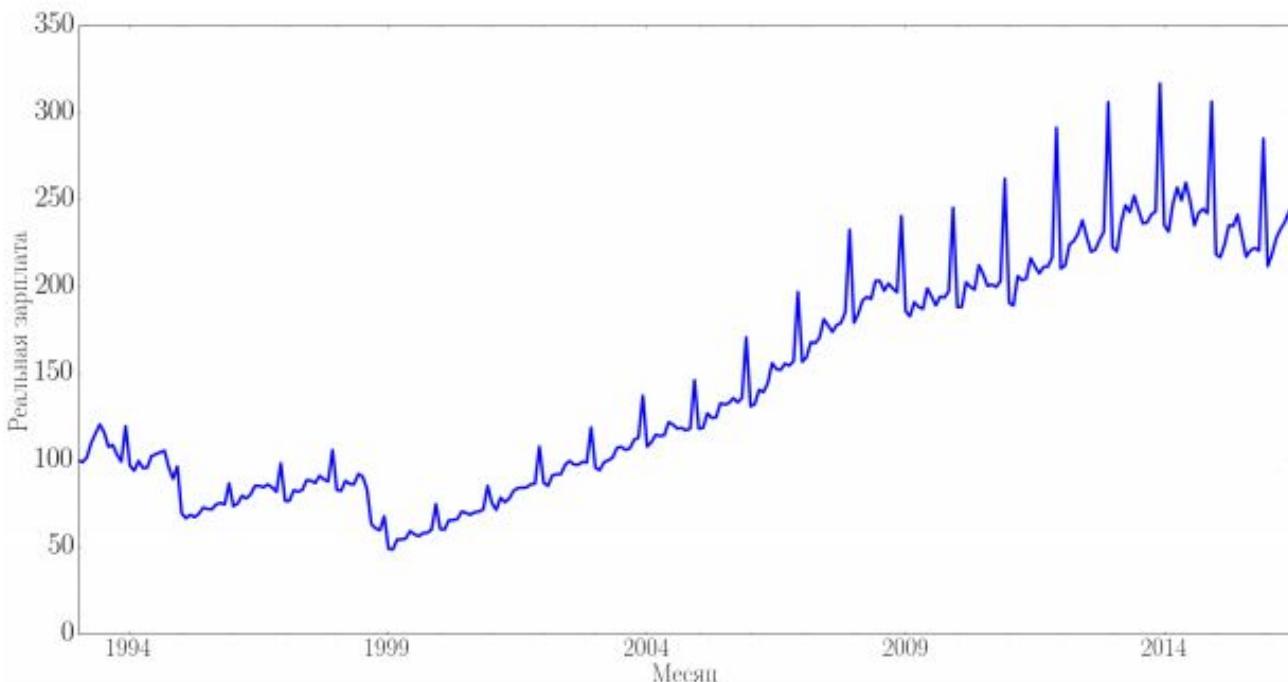
и Q компонент скользящего среднего:

$$+\theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \cdots + \theta_{PS} \varepsilon_{t-QS}.$$

Это модель $SARMA(p, q) \times (P, Q)$

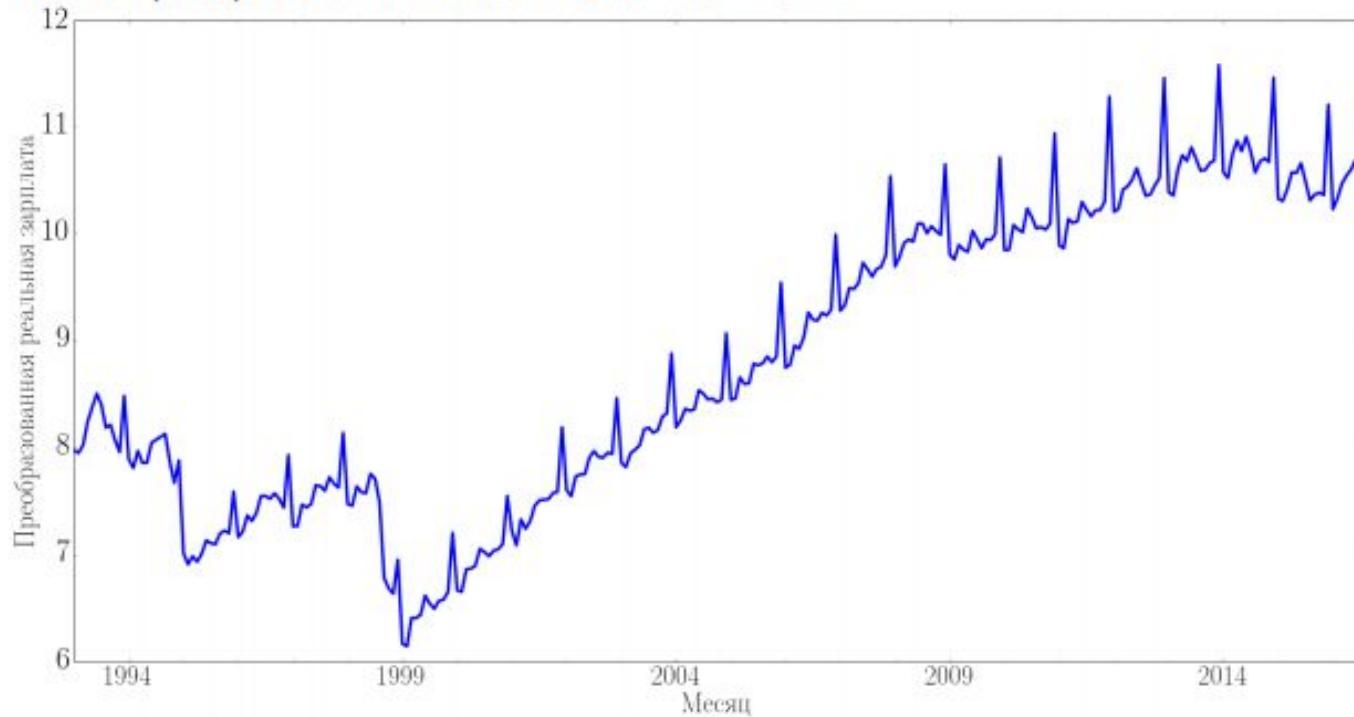
Модель $SARIMA(p, d, q) \times (P, D, Q)$ — модель $SARMA(p, q) \times (P, Q)$ для ряда, к которому d раз было применено обычное дифференцирование и D раз — сезонное.

Часто называют просто ARIMA.



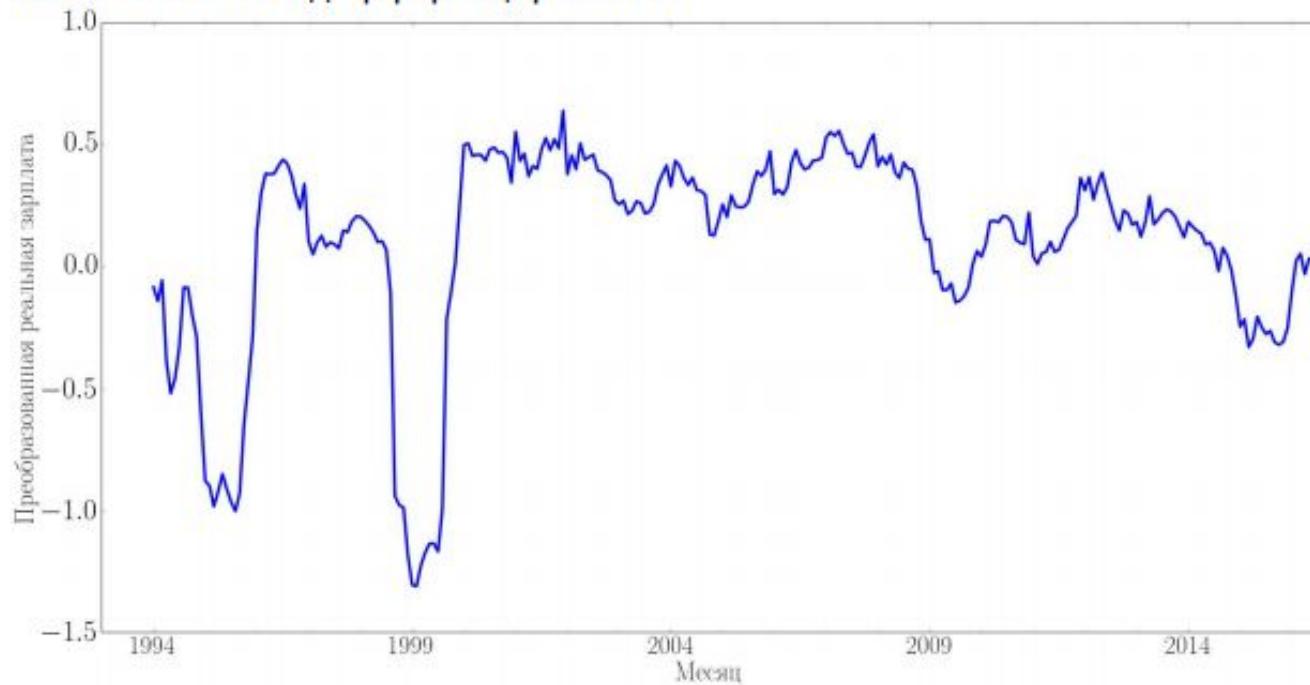
Критерий Дики-Фуллера: $p = 0.2265$.

После преобразования Бокса-Кокса с $\lambda = 0.22$:



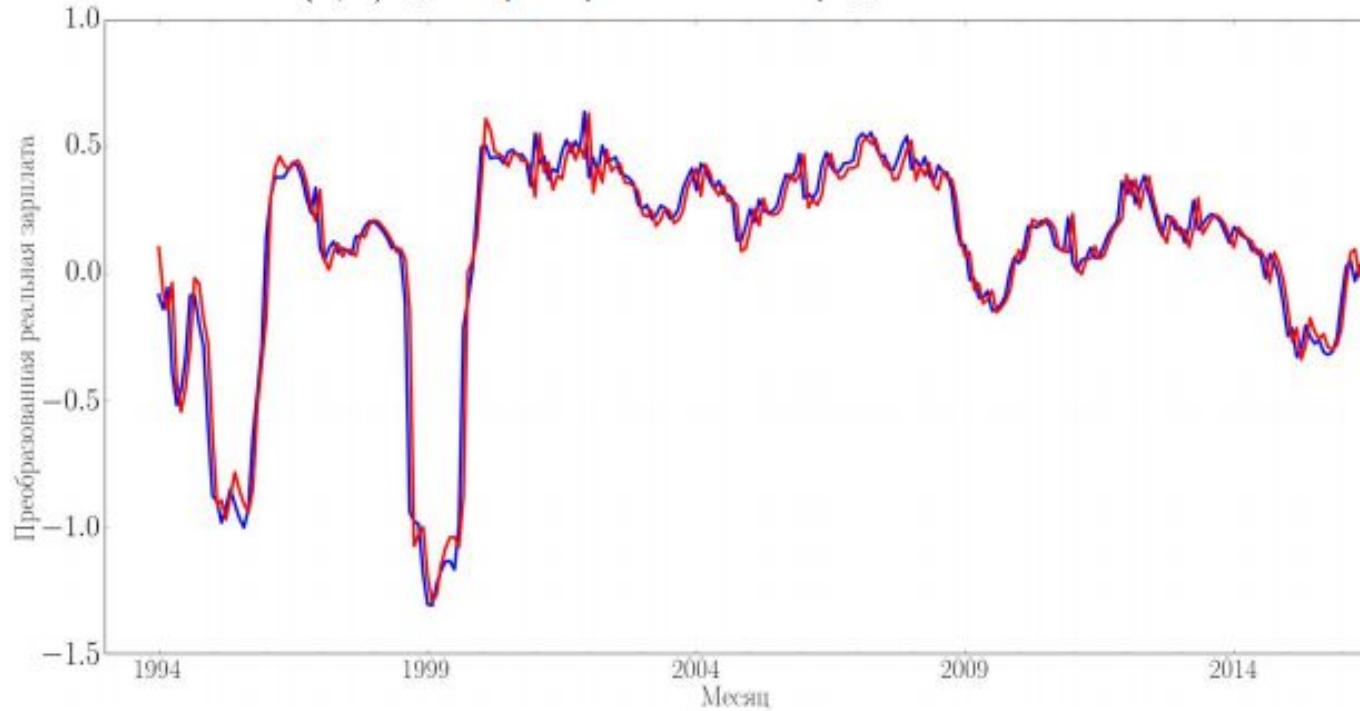
Критерий Дики-Фуллера: $p = 0.1661$.

После сезонного дифференцирования:

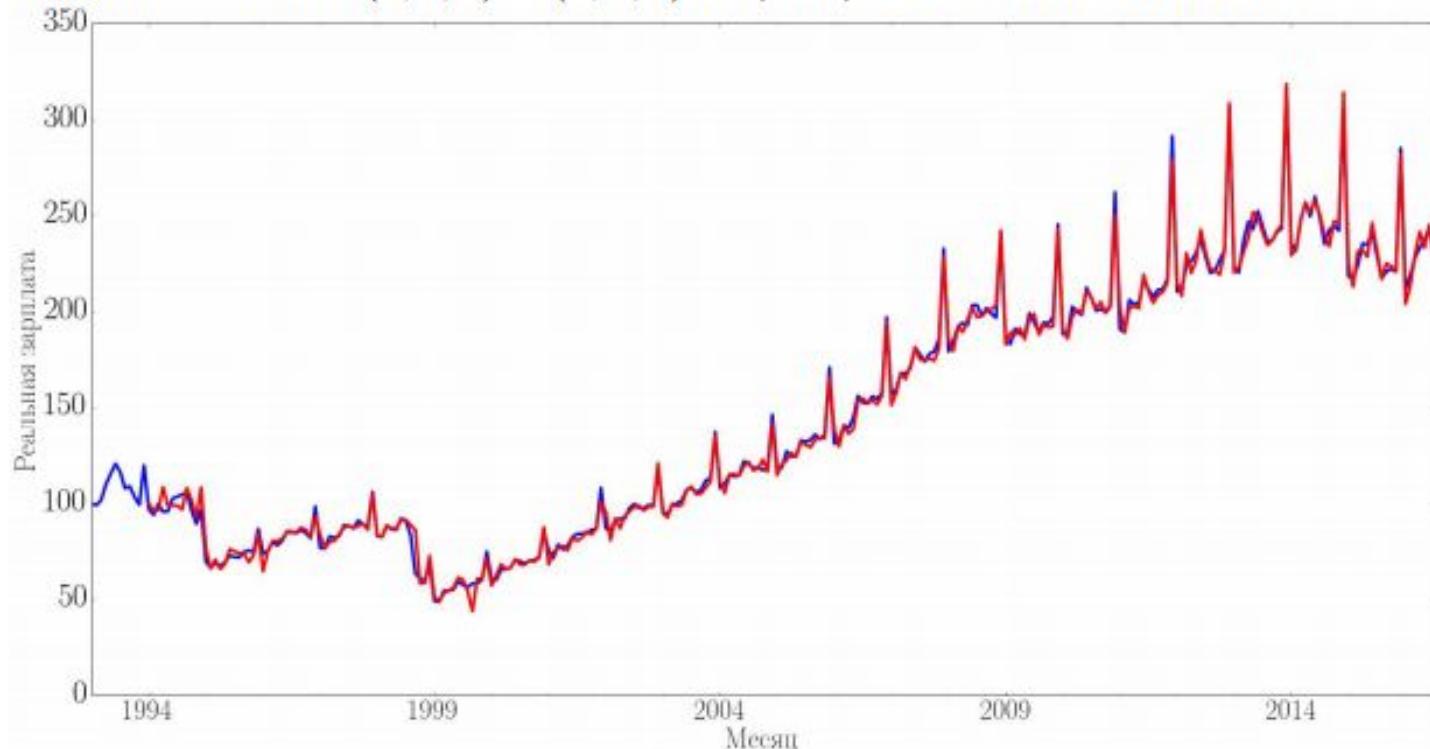


Критерий Дики-Фуллера: $p = 0.01$.

Модель $ARMA(2, 2)$ для преобразованного ряда:



Модель $SARIMA(2, 0, 2) \times (0, 1, 0)$ с преобразованием Бокса-Кокса:



Подбор параметров

- α, ϕ, θ
- d, D
- q, Q
- p, P

Подбор параметров

- α, ϕ, θ
- d, D
- q, Q
- p, P

1. Если все остальные параметры фиксированы, коэффициенты регрессии подбираются методом наименьших квадратов.
2. Чтобы найти коэффициенты θ , шумовая компонента предварительно оценивается с помощью остатков авторегressии.
3. Если шум белый (независимый одинаково распределённый гауссовский), то МНК даёт оценки максимального правдоподобия.

Подбор параметров

- α, ϕ, θ
- d, D
- q, Q
- p, P

1. Порядки дифференцирования подбираются так, чтобы ряд стал стационарным.
2. Ещё раз: если ряд сезонный, рекомендуется начинать с сезонного дифференцирования.
3. Чем меньше раз мы продифференцируем, тем меньше будет дисперсия итогового прогноза.

Подбор параметров

- α, ϕ, θ
- d, D
- q, Q
- p, P

1. Гиперпараметры нельзя выбирать из принципа максимума правдоподобия: L всегда увеличивается с их ростом.
2. Для сравнения моделей с разными q, Q, p, P можно использовать критерий Акаике: $AIC = -2 \log L + 2k$, $k = P + Q + p + q + 1$ — число параметров в модели.
3. Начальные приближения можно выбрать с помощью автокорреляций.

Частичная автокорреляция (PACF)

Частичная автокорреляция — автокорреляция после снятия авторегрессии на промежуточные значения:

$$\phi_{hh} = \begin{cases} r(y_{t+1}, y_t), & h = 1, \\ r(y_{t+h} - y_{t+h}^{h-1}, y_t - y_t^{h-1}), & h \geq 2, \end{cases}$$

где y_t^{h-1} — регрессия y_t на $y_{t+1}, y_{t+2}, \dots, y_{t+h-1}$:

$$y_t^{h-1} = \beta_1 y_{t+1} + \beta_2 y_{t+2} + \cdots + \beta_{h-1} y_{t+h-1},$$

$$y_{t+h}^{h-1} = \beta_1 y_{t+h-1} + \beta_2 y_{t+h-2} + \cdots + \beta_{h-1} y_{t+1}.$$

Выбор модели ARIMA

1. Смотрим на ряд.
2. При необходимости стабилизуем дисперсию.
3. Если ряд нестационарен, подбираем порядок дифференцирования.
4. Анализируем ACF/PACF, определяем примерные p , q , P , Q
5. Обучаем модели-кандидаты, сравниваем их по AIC, выбираем победителя.
6. Смотрим на остатки полученной модели, если они плохие, пробуем что-то поменять.

R

В пакете forecast в R есть функция, автоматически строящая прогноз ряда:

```
auto.arima(x, d=NA, D=NA, max.p=5, max.q=5,  
           max.P=2, max.Q=2, max.order=5, max.d=2, max.D=1,  
           start.p=2, start.q=2, start.P=1, start.Q=1,  
           stationary=FALSE, seasonal=TRUE,  
           ic=c("aicc","aic", "bic"), stepwise=TRUE, trace=FALSE,  
           approximation=(length(x)>100 | frequency(x)>12),  
           truncate=NULL, xreg=NULL, test=c("kpss","adf","pp"),  
           seasonal.test=c("ocsb","ch"), allowdrift=TRUE,  
           allowmean=TRUE, lambda=NULL, parallel=FALSE,  
           num.cores=2, ...)
```

Построить прогноз можно с помощью функции forecast:

```
forecast(object, h=ifelse(frequency(object)>1,2*frequency(object),10),  
         level=c(80,95), fan=FALSE, robust=FALSE, lambda=NULL,  
         find.frequency=FALSE, allow.multiplicative.trend=FALSE, ...)
```