Feature Engineering and Selection

Lecture Outline

Last time we studied the following topics:

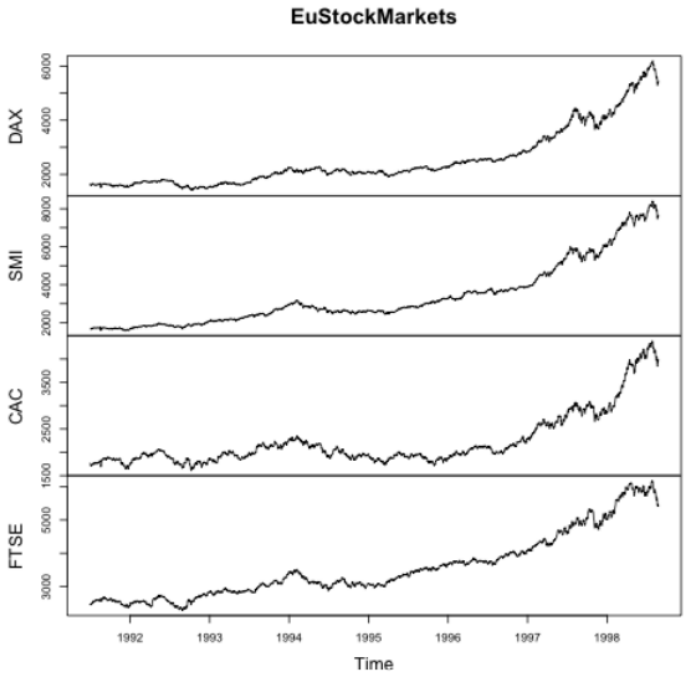
- the different components of a time series
- Why removing trends and stationarity
- how to handle missing values
- How to handle seasonality, trends and stationarity
- Useful transformations: box cox and differencing

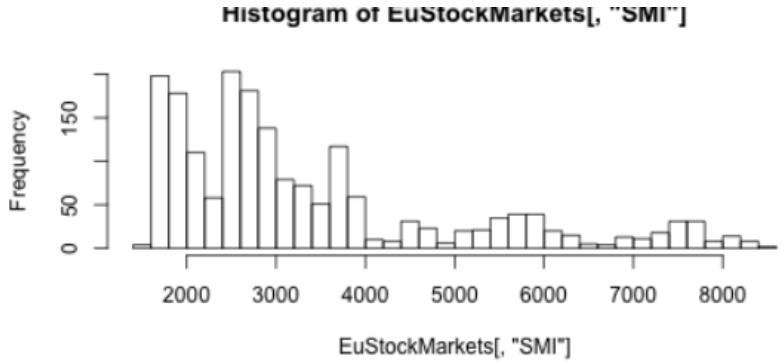
In this lecture we will study:

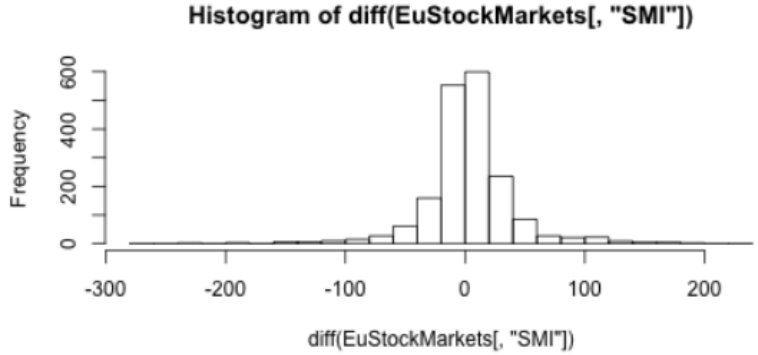
- Correlation and Autocorrelation
- Feature Engineering: how to create new features that can make the task easier for your prediction model
- Feature Selection: How to select which feature are useful or not for your model

Data Differencing

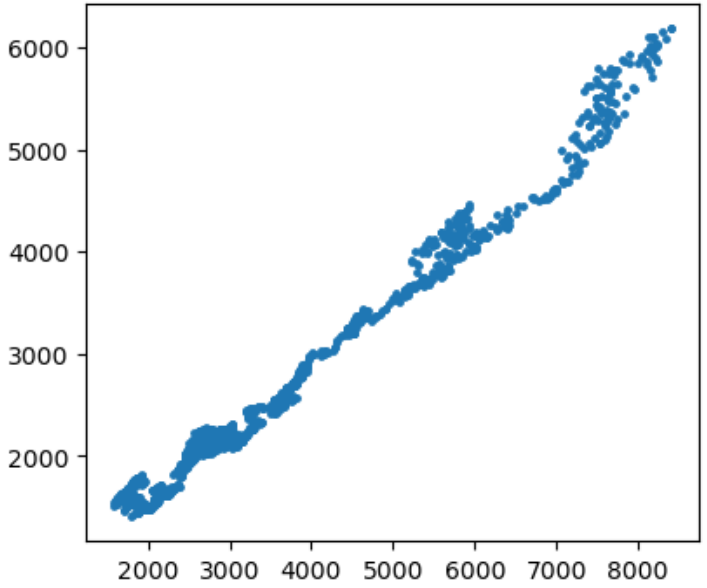
Variation is more interesting than actual values



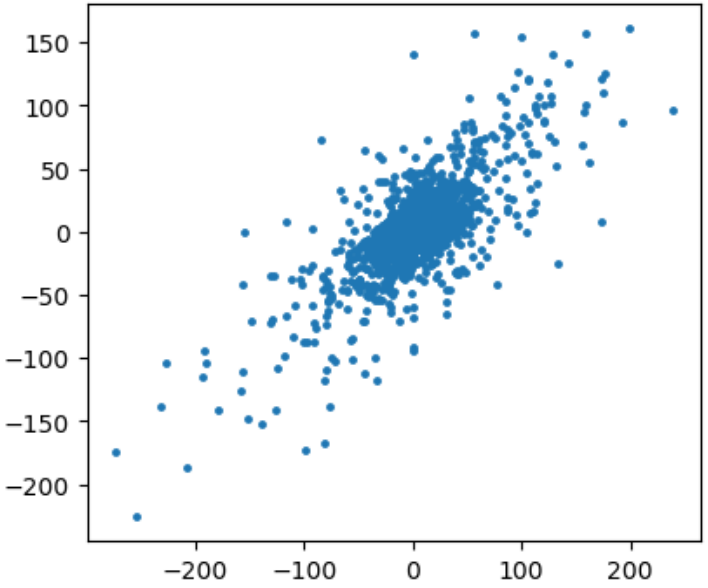




Importance of Correlation



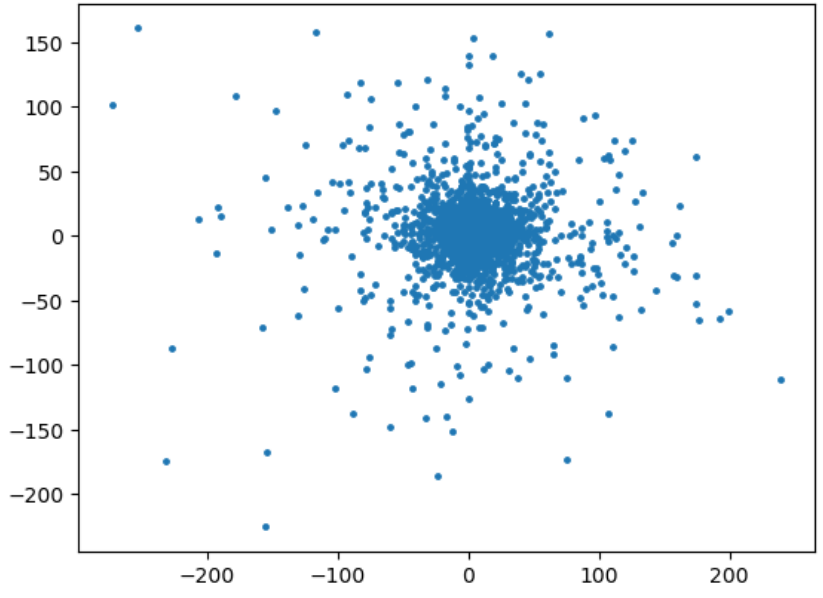
Stock variations when taken at the same time



What can you observe?

Correlation !

Is it useful?



Stock variations with one taken at time t and the other at t-1

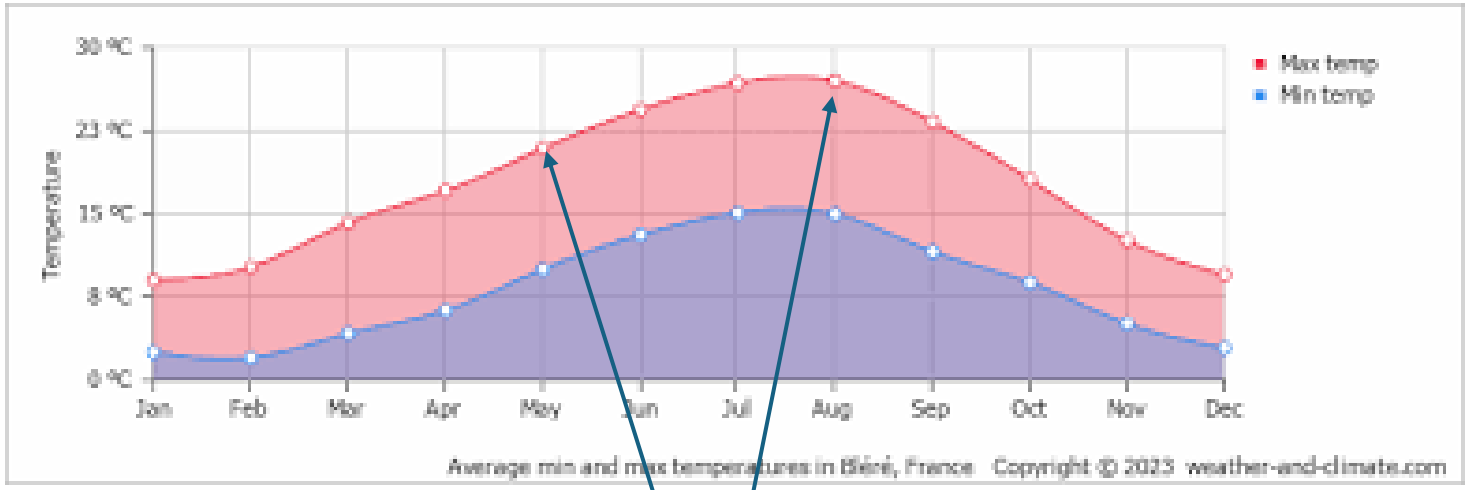
Correlation is what makes prediction possible, if there was no correlation then it means that the data are purely random and therefore not predictible.

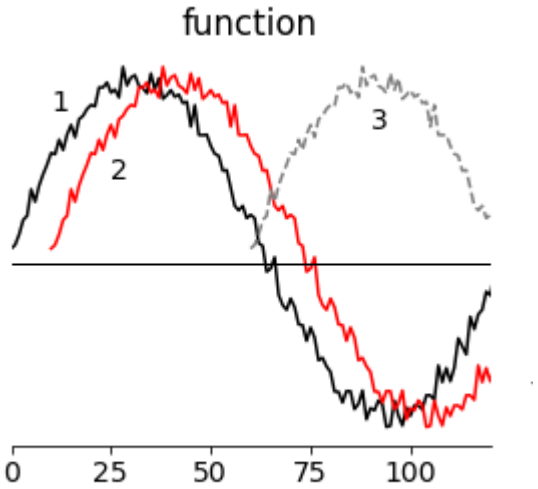
Auto-Correlation

Simo Alami

Self Correlation

Definition: Mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals

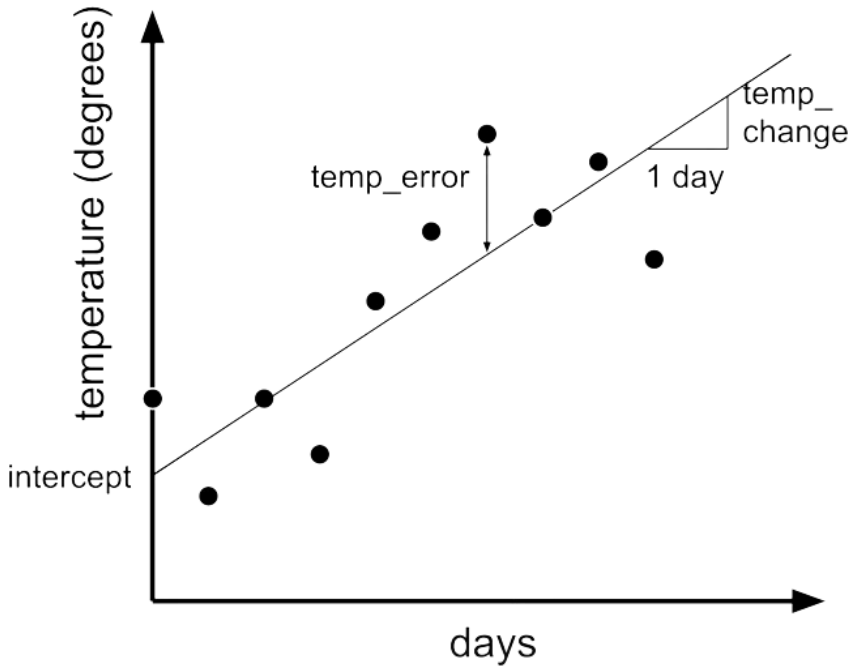




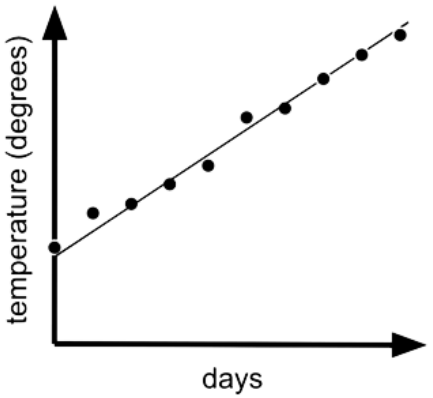
Compare these two values

- If you have **correlation**: hotter/cooler May temperature correlates with hotter/cooler August temps
  → There exists long term predictability
- If **no correlation**: knowing the temperature on May 15th does not alone give you any information about the likely range of temperatures on August 15th.
  → Information is still interesting

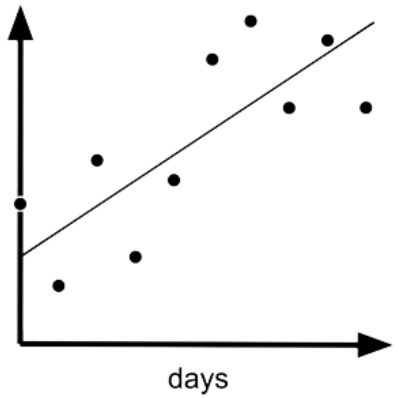
Correlation



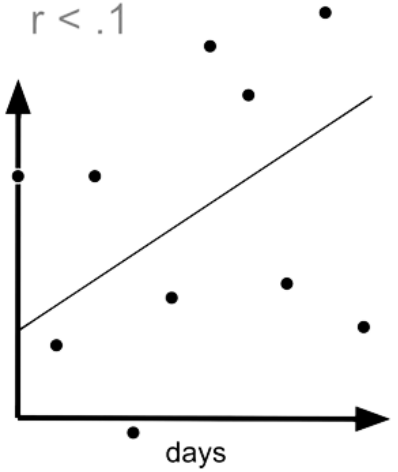
High correlation
r > .9



Moderate correlation
.4 > r > .7



Low correlation
r < .1



$$\rho(x,y) = \frac{\mathrm{cov}(x,y))}{\sigma_x \sigma_y}$$

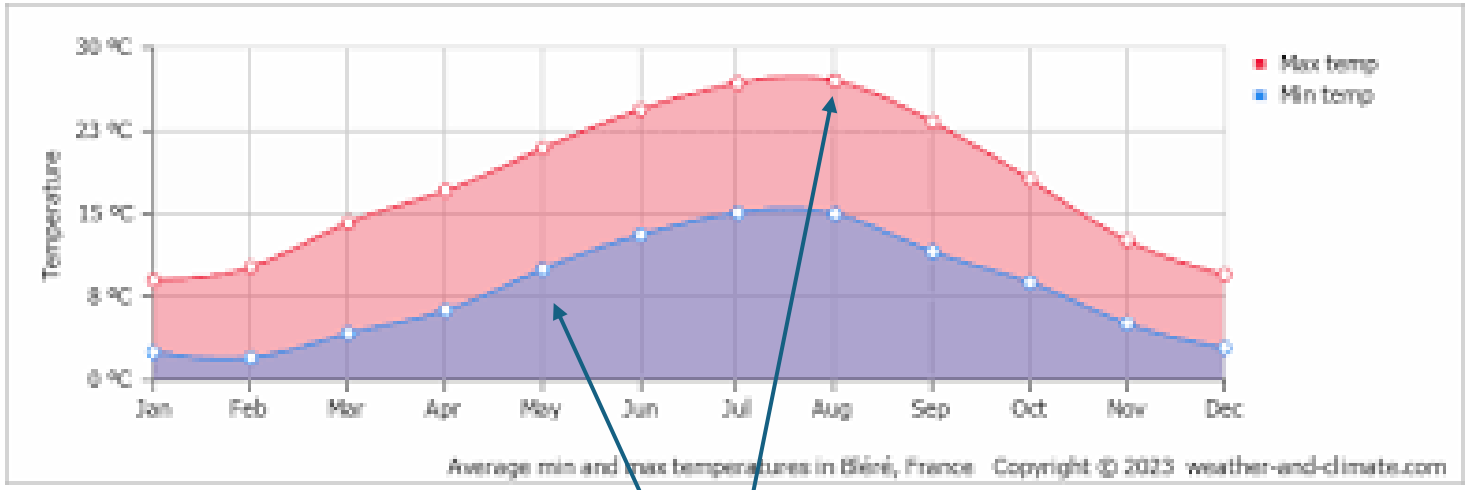
$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$$

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

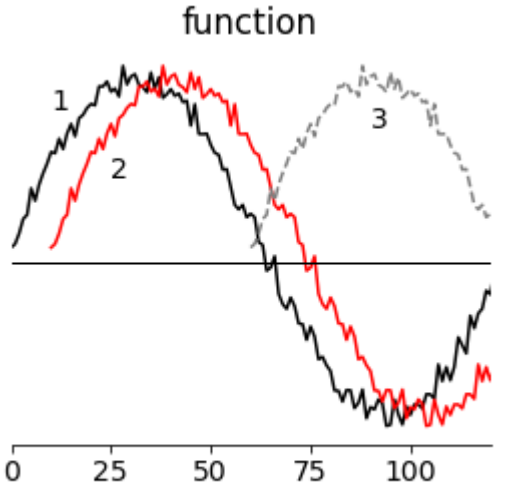
Autocorrelation

Autocorrelation generalizes self-correlation by not anchoring to a specific point in time

is there a correlation between any two points in a specific time series with a specific fixed distance between them?



Correlation compares these two values



Autocorrelation compares these two series

Definition: correlation of a signal with a delayed copy of itself as a function of the delay.

Intuition: how data points at different points in time are linearly related to one another as a function of their time difference.

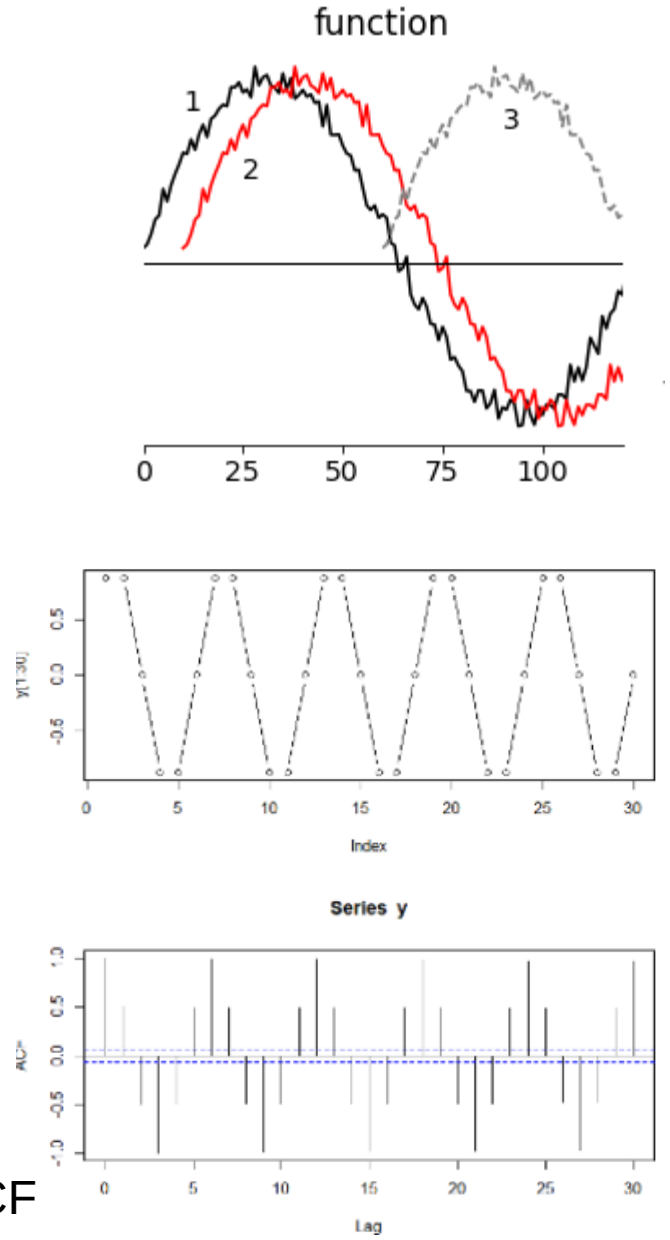
Autocorrelation

given a time series and a lag value k:

- Create two time series, one with the values x_i and another with the values x_{i-k}
- Calculate the correlation between these two time series

$$\rho_k(x, y) = \frac{\operatorname{cov}(x_i, x_{i-k})}{\sigma_{x_i}\sigma_{x_{i-k}}}$$

- Points that have a lag between them of 0 have a correlation of 1 (always true),
- Points separated by 1 lag have a correlation of 0.5.
- Points separated by 2 lags have a correlation of -0.5, and so on.



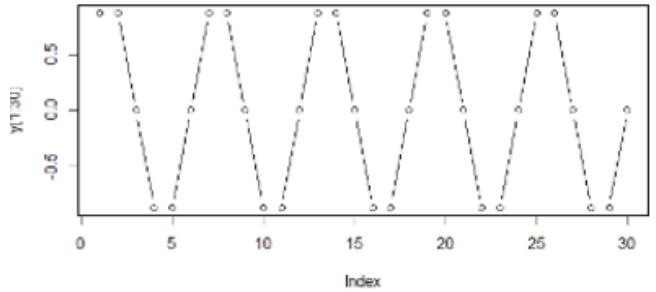
Plot of a sine function and its ACF

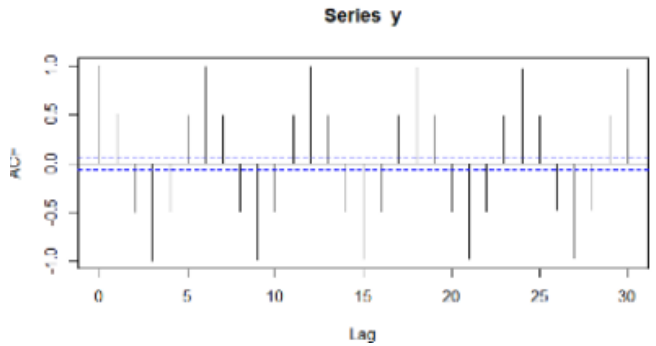
Autocorrelation in deterministic systems

This sine series is a simple function and a fully determined system given a known input sequence. Nonetheless, we do not have a correlation of 1.
→ Why is that?

- Given a few points in a row, one can predict which direction the process is going

- But when given only one point, most values do not have a unique subsequent value but rather more than one.
  → ACF is a 1:1 correlation measure
  → Correlation of less than 1 because most values do not have a unique subsequent value but rather more than one.

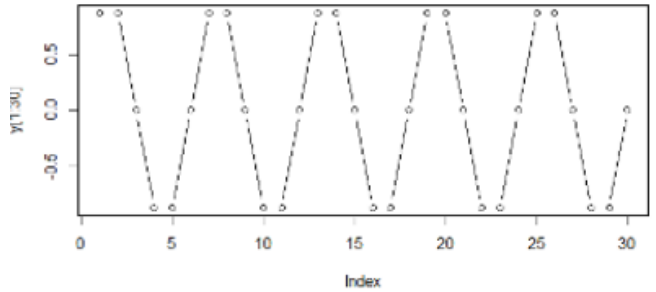


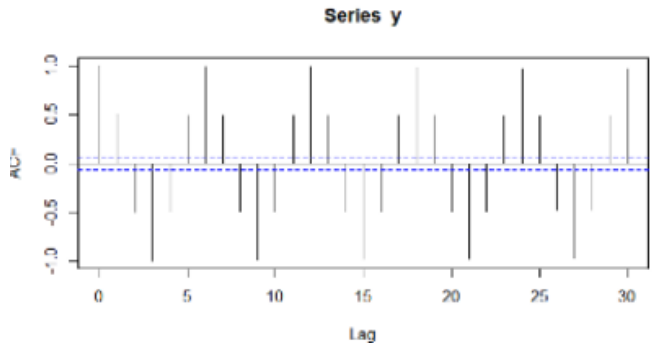


Nonunitary correlation does not mean you necessarily have a probabilistic or noisy time series.

Properties of the ACF

- The ACF of stationary data should drop to zero quickly.
- For nonstationary data the value at lag 1 is positive and large.
- The ACF of a periodic function has the same periodicity as the original process
- The autocorrelation of the sum of periodic functions is the sum of the autocorrelations of each function separately.
- All time series have an autocorrelation of 1 at lag 0
- The autocorrelation of a sample of white noise will have a value of approximately 0 at all lags other than 0
- The ACF is symmetric with respect to negative and positive lags, so only positive lags need to be considered explicitly.
- A statistical rule for determining a significant nonzero ACF estimate is given by a “critical region” with bounds at +/–1.96 × √n. This rule relies on a sufficiently large sample size and a finite variance for the process.





Time for Exercise 1

Partial Auto-Correlation

Simo Alami

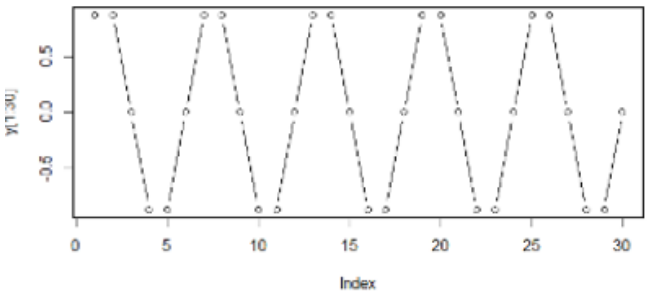
Partial Correlation

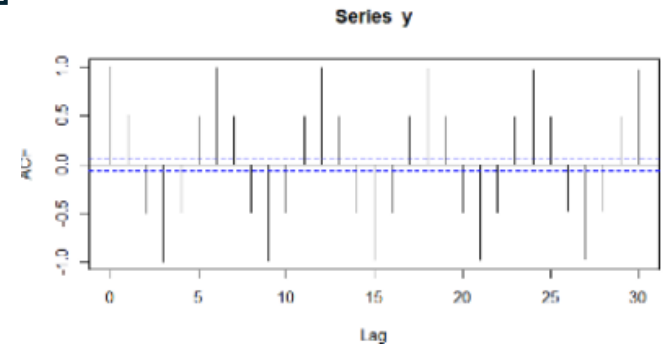
Example: Suppose you want to find the relationship between the number of hours studied X and exam score Y, but you suspect that both are influenced by intelligence Z.
→ Correlation between X and Y will include the effect of Z

Partial correlation: computes the correlation between hours studied and exam scores while removing the effect of intelligence.

Definitions:

- **Correlation:** Measures the direct relationship between two variables without accounting for any other influences.
- **Partial Correlation**: ensures the relationship between two variables while holding constant the effect of a third (or more) variable(s), removing any influence that third variable may have on both.





The partial correlation between X and Y, controlling for Z, is given by the formula:

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} \times r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Computing Partial Correlation (1/2)

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} \times r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

Necessary to compute the correlation between each pair of variables:

→ If you have a lot of variables and a lot of data, it is computationally costly

→ Even more serious in the case of multiple controlling variables

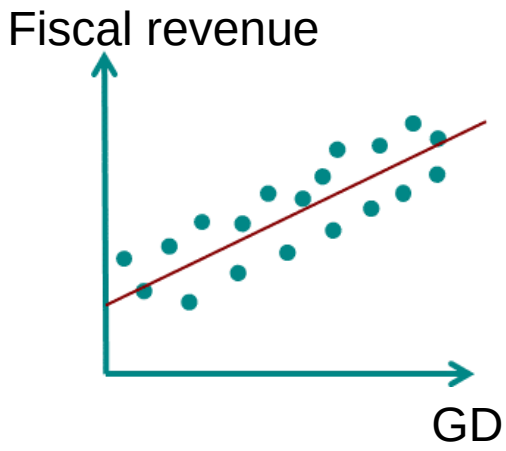
→ **Solution:** Use the residuals

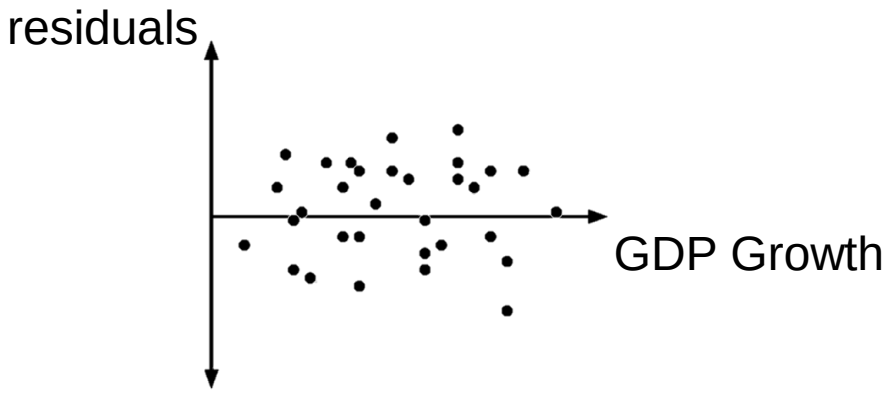
Computing Partial Correlation (2/2)

Recall: correlation captures the strength of a linear relationship between two variables.

If you want to get rid of the effect af all other variables, just fit a linear regression between each variable of interest and controling variables.

Example: Unemployment, Taxes, and Economic Conditions





Unemployment residuals: part of unemployment **not** explained by GDP
Taxes residuals: part of fiscal revenues not explained by GDP

→ Correlation between **Unemployment residuals** and **Taxes residuals** = Correlation between Unemployment and taxes without GDP effect

Partial Correlation Recipe

Step 1: predict X given Z and Y given Z. If you have multiple control variables, you predict X given all the control variables.

$$X = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \dots$$

Step 2: compute the residuals, $\mathrm{E_X} = \widehat{\boldsymbol{X}} - \boldsymbol{X}$ and $\mathrm{E_y} = \widehat{\boldsymbol{Y}} - \boldsymbol{Y}$

Step 3: Compute the correlation between E_X and E_Y

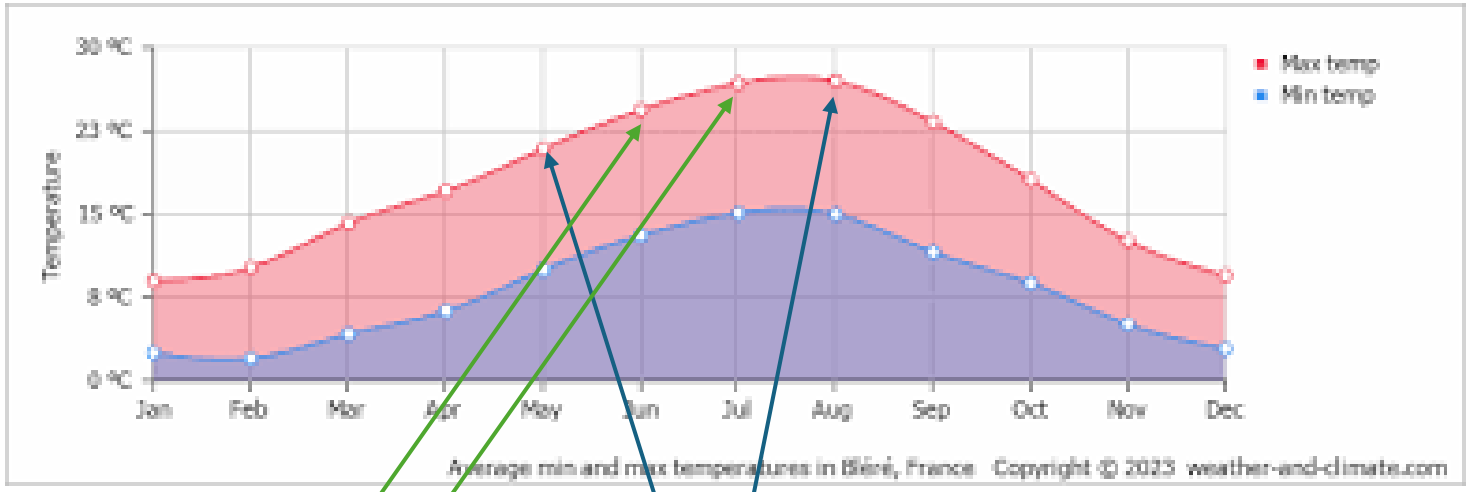
By calculating the correlation between the residuals, we are essentially:

- Measuring how much the parts of X and Y that are not related to Z are correlated with each other.
- This gives the partial correlation because the effect of Z has already been "factored out" in the residuals.

Time for Exercise 2

Partial Autocorrelation

Definition: The partial autocorrelation of a time series for a given lag is the partial correlation of the time series with itself at that lag given all the information between the two points in time removed.



Partial autoorrelation compares these two values without the effect of these

Unlike autocorrelation, which measures the correlation between a time series and its lagged version (including all prior lags), partial autocorrelation isolates the direct relationship between a time series and a specific lag, excluding the influence of shorter lags.

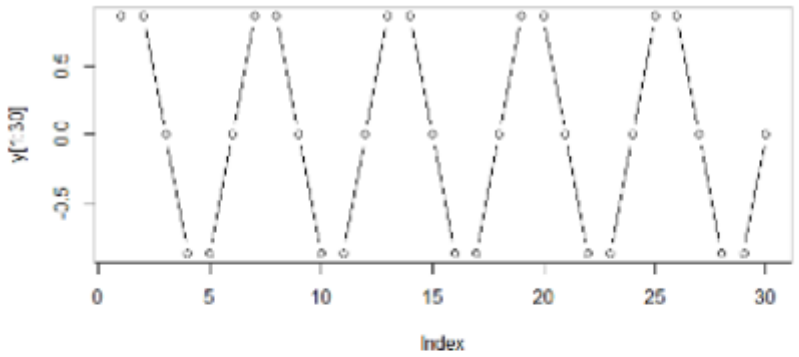
Visualizing PACF

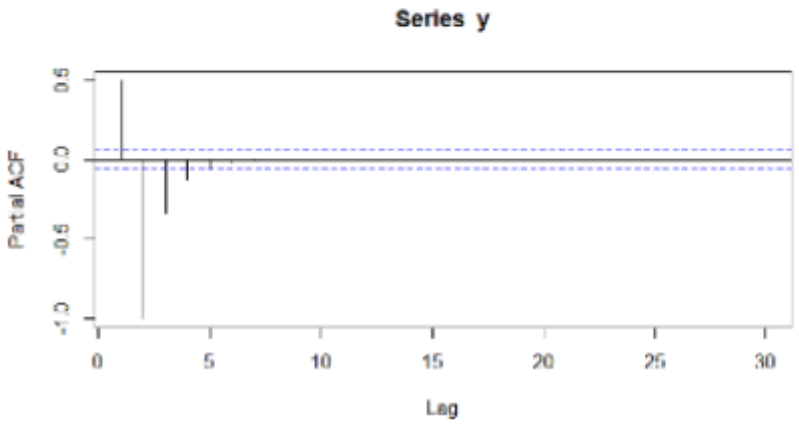
- The PACF shows which data points are informative and which are harmonics of shorter time periods.

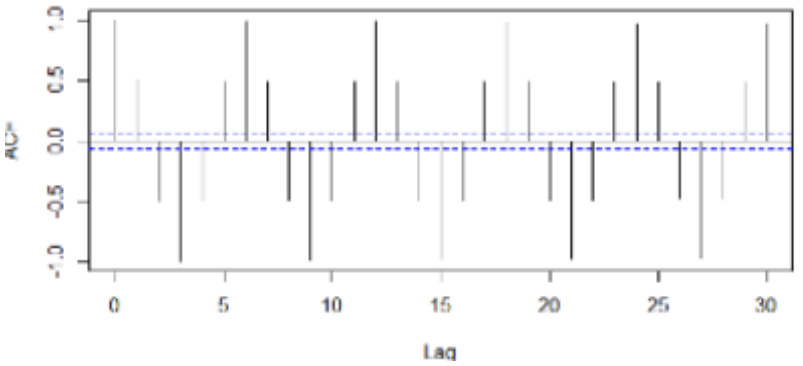
- For a seasonal and noiseless process, such as the sine function, with period T, the same ACF value will be seen at T, 2T, 3T, and so on up to infinity. An ACF fails toweed out redundant correlations.

- The PACF, on the other hand, reveals which correlations are “true” informative correlations for specific lags rather than redundancies.

→ Allows to know which temporal scale to use in a predictive model.
→ In the example, PACF shows that using more tan 5 lags is not useful for prediction







Computing PACF

Principle:

- The autocorrelation at lag k measures how strongly y_t is correlated with y_{t-k}.
- The partial autocorrelation at lag k removes the effect of the intermediate lags by using the residuals method between y_t , y_{t-k} while the controlling variables are all the intermediate lags

Recipe:

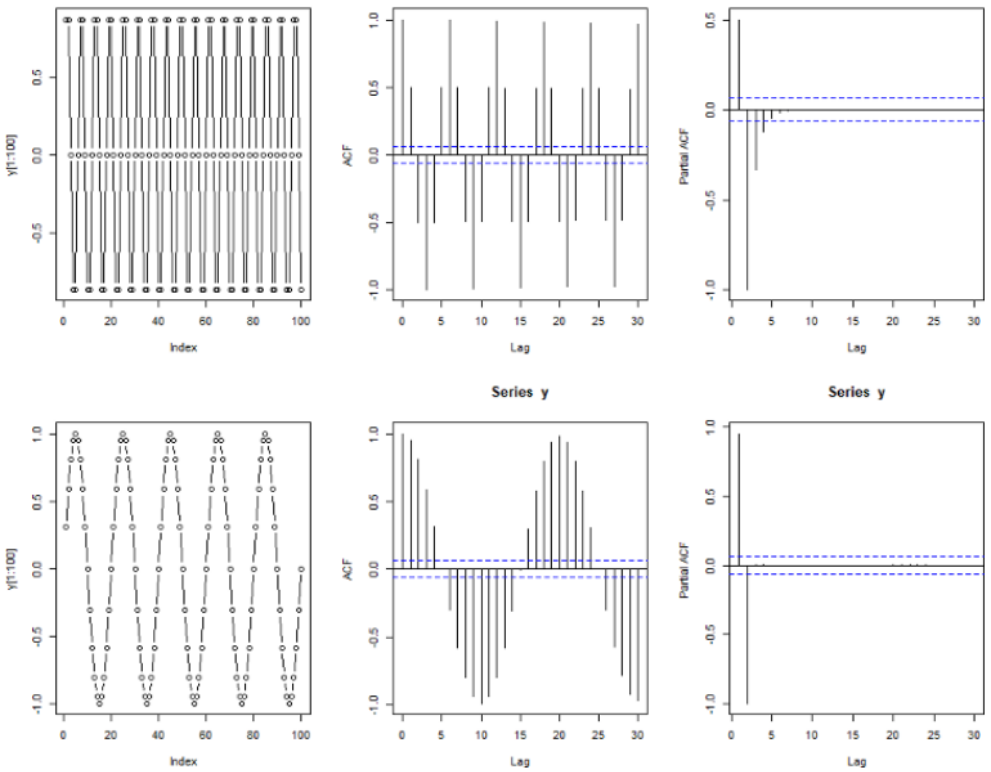
1. Regress y_t on all lagged values from 1 to k-1, i.e., $y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_{k-1} y_{t-k+1} + \epsilon_t$. The residual ϵ_t is the part of y_t that is not explained by the intermediate lags.

2. Regress y_{t-k} on the same intermediate lags from 1 to k-1, i.e., $y_{t-k} = \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_{k-1} y_{t-k+1} + \epsilon_{t-k}$.The residual ϵ_{t-k} represents the part of y_{t-k} that is not explained by the intermediate lags.

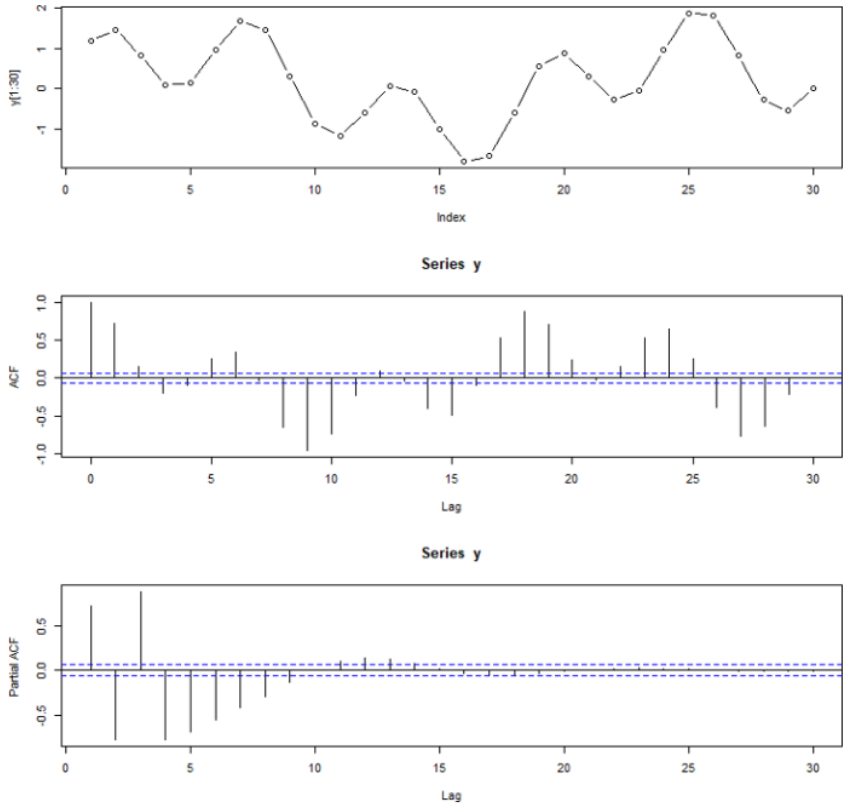
3. The PACF at lag k is then the correlation ϵ_t and ϵ_{t-k}, the residuals from both regressions. This gives the "pure" correlation between y_t and y_{t-k} after removing the influence of shorter lags.

The critical region for the PACF is the same as for the ACF. The critical region has bounds at ±1.96/√n. Any lags with calculated PACF values falling inside the critical region are effectively zero.

Sum of PACF



Sum



$y = x \times \pi/3$ and $y = x \times \pi/10$

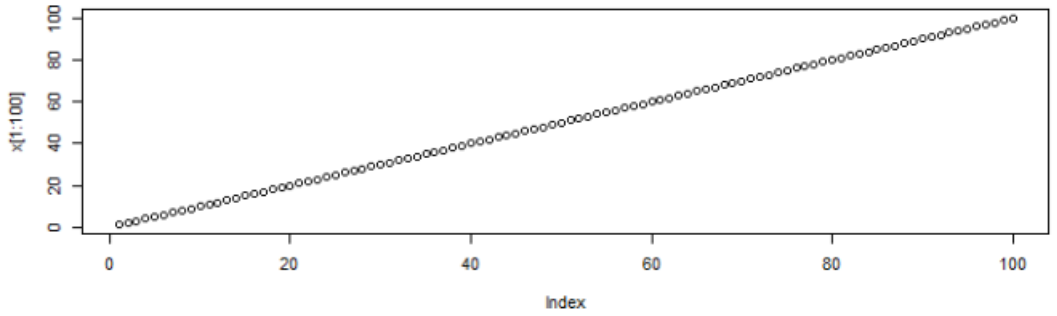
- ACF of the sum of two periodic series is the sum of the individual ACFs
- The PACF is not a straightforward sum of the PACF functions of the individual components.
- Partial autocorrelation is more substantial in the summed series than in either of the original series.

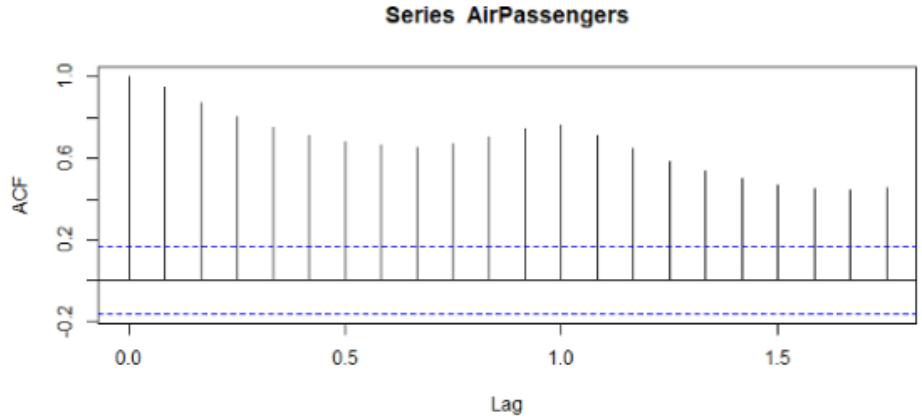
→ two different periods: any given point is less determined by the values of neighboring points since the location within the cycle of the two periods is less fixed now as the oscillations continue at different frequencies

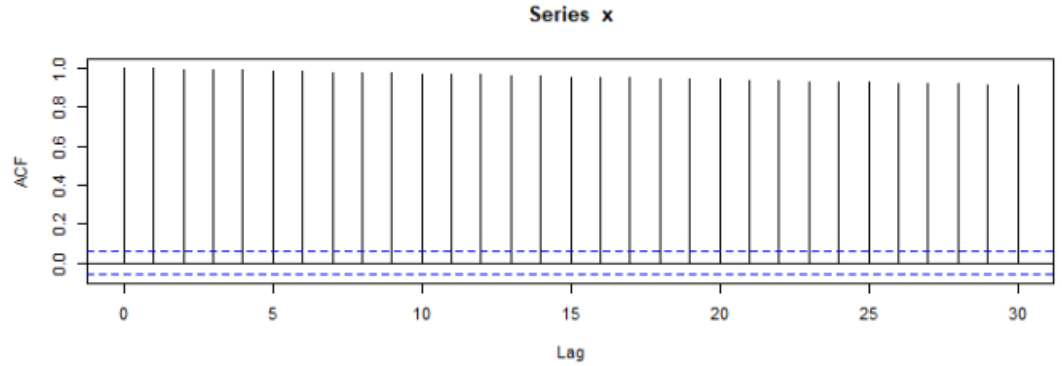
ACF and PACF for Non Stationary Data

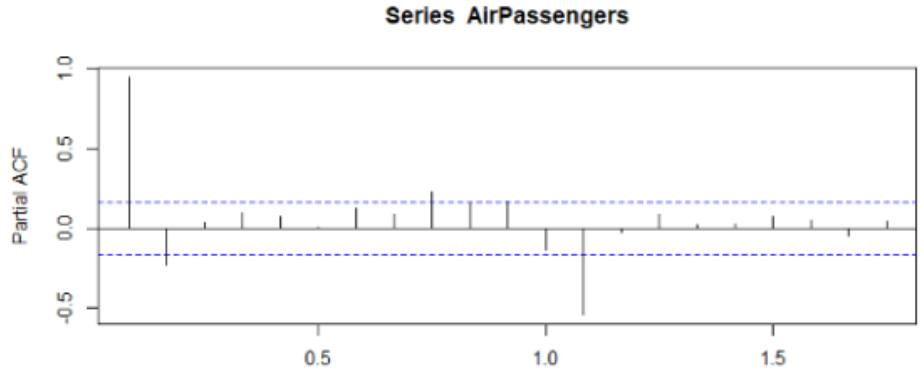
- The ACF is not informative.
- The only significant PACF correlation is at lag 1

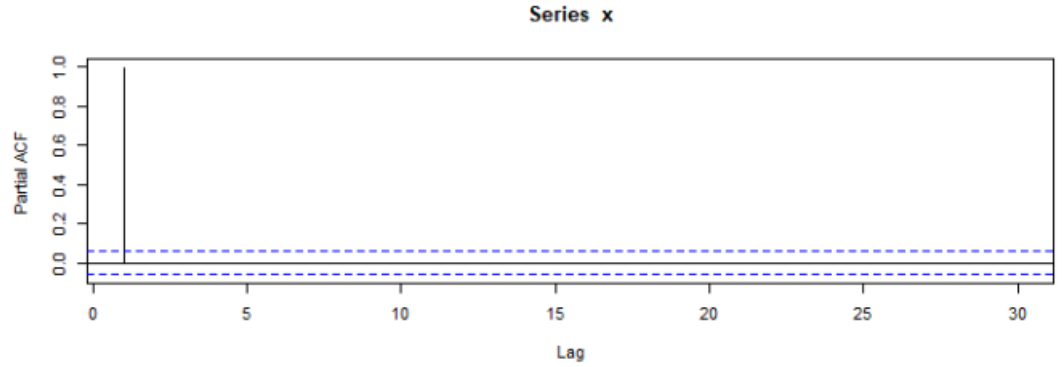
Why? For a given point in time, once you know the point just before it, you know all the necessary information











Spurious Correlations

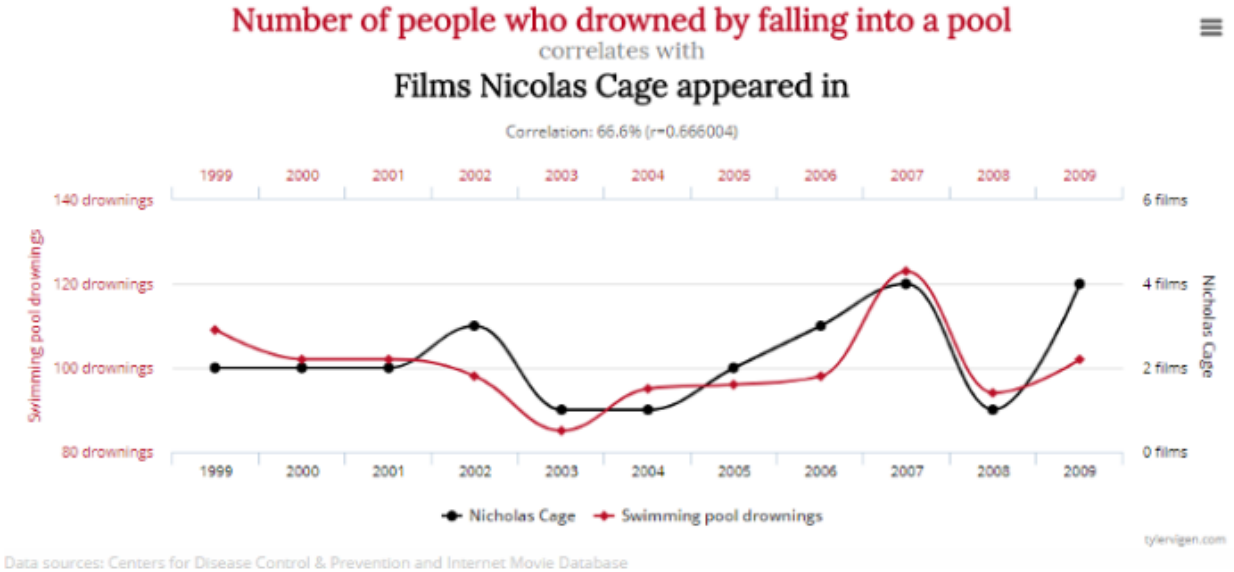
⚠ **Rookie mistake:** Computing correlations between any 2 variables and find correlations everywhere

Trends: Data with an underlying trend is likely to produce spurious correlations.

→ There is more information in a trending time series than in a stationary time series, so there are more opportunities for data points to move together.

Seasonality: think of a spurious correlation between hot dog consumption and death by drowning (summer)

Cumulatively summed quantities: this is a trick used in certain industries to make models or correlations look better than they are.



Make your data stationary !

Time for Exercise 3

Feature Engineering

Simo Alami

Feature Engineering

Feature engineering efforts mainly have two goals:

- Creating the correct input data set to feed the machine learning algorithm

- Increasing the performance of machine learning models: generating valid relationships between input features and the output feature or target variable to be predicted.

We will cover four different categories of time features:

- Date time features
- Lag features and window features
- Rolling window statistics
- Expanding window statistics

Date Time Features

Date Time Features are features created from the time stamp value of each observation:

- Weekend or not
- Minutes in a day
- Daylight savings or not
- Public holiday or not
- Quarter of the year
- Hour of day
- Before or after business hours
- Season of the year

Lag Features and Window Features

Lag features: values at prior timesteps that are considered useful because they are created on the assumption that what happened in the past can influence or contain a sort of intrinsic information about the future.

Example:

- Create features for sales yesterday 4:00 pm to help predict sales for today 4:00pm
- Use all last week's sales values to predict next week's sales

Question: The lag length is called the window → How large should it be?

→ To predict next week's sales, should we use last week's values? last month?...

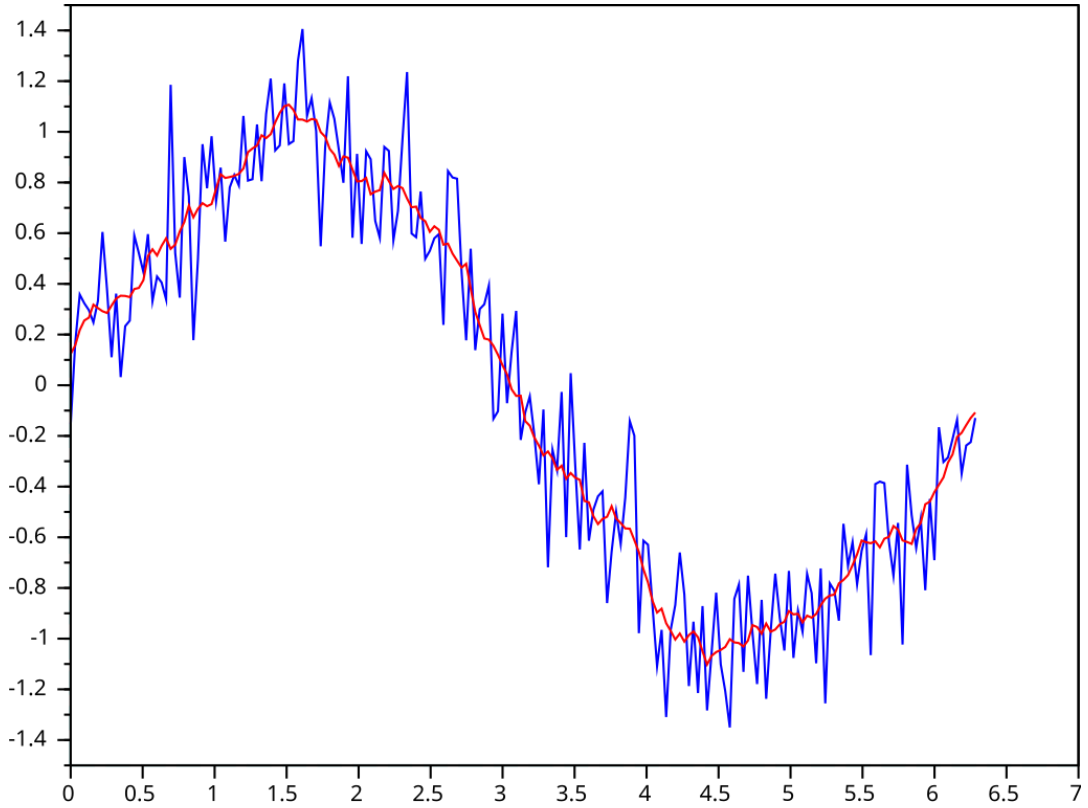
→ **To be studied in the feature selection part**

Rolling window statistics

Definition: compute statistics on the values from a given data sample by defining a range that includes the sample itself as well as some specified number of samples before and after the sample used.

Example: Moving average

Question: What is a good window size to use?



Caution: The lag and window features create NAN values for the first rows depending on your lag or window size. Please be sure to handle that properly

Expanding window statistics

Definition: Unlike a rolling window, which looks at a fixed number of observations, an expanding window increases its size with each time step, including all previous observations up to the current point.

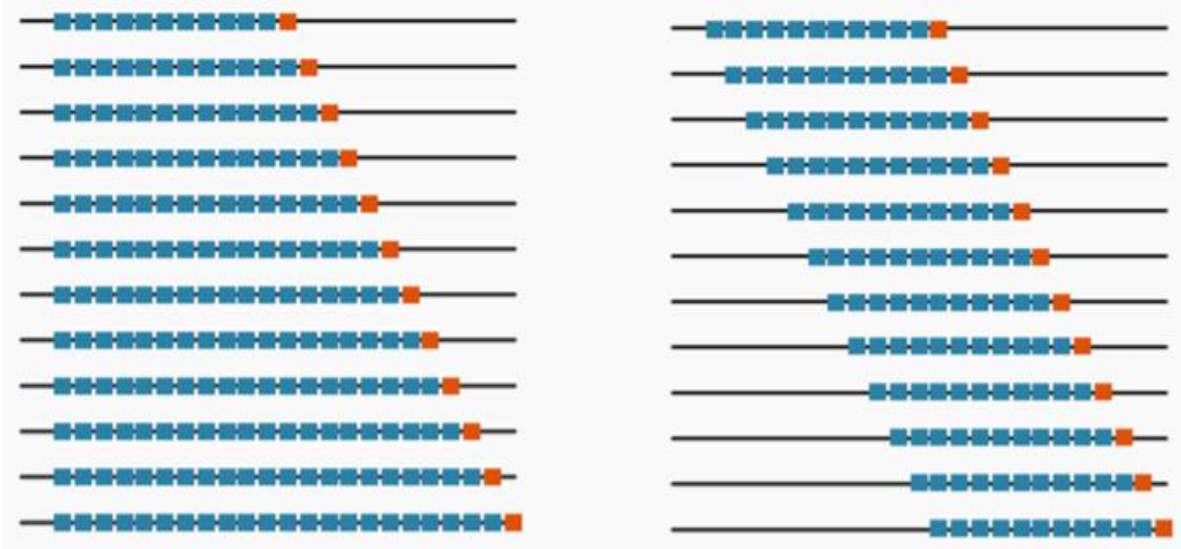
→ allows to analyze how a statistic changes as more data becomes available over time.

→ considers all previous data points, so the size of the window grows as you progress through the time series .

→ **Useful for Trends**: This method is particularly useful for understanding trends over time, as it provides insights into how metrics like mean, sum, or variance evolve as more data becomes available.

Example:

- Expanding sum
- Expanding mean
- Expanding variance



Feature selection

Too many features will harm your model, making it harder to find real correlations and causations
→ After feature engineering, feature selection is necessary

- **Univariate Feature Selection**: evaluate each feature independently to determine its relationship with the target variable.
  - **Correlation Analysis**: Computing the correlation coefficients between features and the target variable.
  - **Statistical Tests**: Measuring the significance of features against the target.

- **Recursive Feature Elimination (RFE)**: wrapper method that recursively removes the least important features based on the model's performance. It works by:
  - Training a model on the entire feature set.
  - Evaluating the importance of each feature
  - Removing the least significant features and repeating the process

- Feature Importance from Tree-based Models

- Lasso Regression (L1 Regularization)

- Principal Component Analysis (PCA)

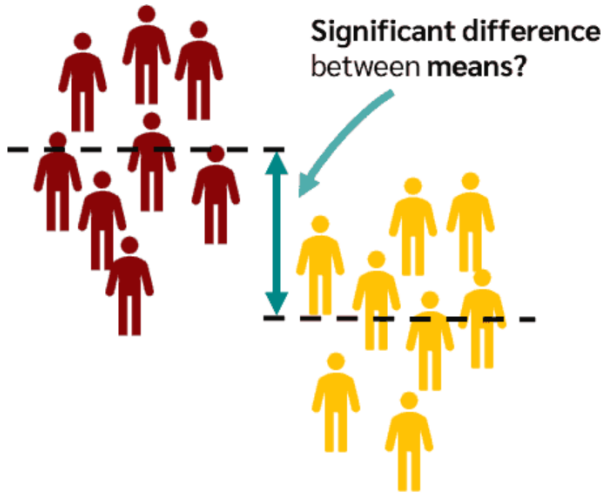
→ we will focus on Univariate Feature Selection

T-test and ANOVA

Context: Electric load forecasting; **Question:** Is there a different power consumption during week-ends?

Strategy: Compare the mean consumption during week-ends and during the rest of the week, compare the means and decide if the difference is statistically significant.

Solution: T-test



⚠ DO NOT DO THAT

T-test and ANOVA are not appropriate to for time series, they do not handle dependencies and autocorrelations usually present in time series:

- Autocorrelation: they assume independent observations
- Stationarity
- Non linear relationships: they assess linear relationships between variables, usually not the case in time series

Use Granger causality test and information gain instead

F-statistic

$$y = \alpha + \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

We want to test the regression coefficients i.e. we want to verify if there is a real impact on having a feature (i.e. $\beta_i \neq 0$) compared to a case where we did not have it.

Null Hypothesis h_0 :

$$h_0 : \beta_i = 0 \quad \forall i$$

Alternative hypthesis h_1:

At least one $\beta_i \neq 0$

F-statistic recipe (1/2)

1. First create an **unrestricted** regression that contains all the variables and fit it
2. Calculate the Sum of square residuals, i.e. the sum of errors:

$$SSR_{ur} = \sum_{i=1}^{N} u_i^2$$

3. Create a **restricted** regression where we let aside all independent variables that are not of interest. We can even have: $y = \alpha$
4. Calculate the SSR of the restricted regression: SSR_r
5. **Question:** is SSR_r significantly different from SSR_{ur} ? To do so we compute the F statistic:

$$F = \frac{(SSR_r - SSR_{ur})/P}{SSR_{ur}/(N - p - 1)}$$

P: the number of predictors; N the number of samples. They are called **degrees of freedom**

Intuition: if the numerator is particularly large relatively to the denominator, that means that the unrestricted module explains the error way more than the restricted one.
→If the F statistic is larger, we are more likely to reject the null hypothesis

F-statistic recipe (2/2)

5. Check the F-value in the F-table. An F statistic has two degrees of freedom. The first df is the number of predictors P, and the second one is N-P-1.

   → Considering a p-value of 0.05, then you will find a critical value $F_{0.05}$

6. If $F > F_{0.05}$ then you reject the null hypothesis

F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Granger causality test

Idea: Granger causality tests whether past values of one time series X contain information that can help predict future values of another time series Y, beyond what is already contained in the past values of Y itself.

Caution: Granger causality does not imply true causation; it only indicates predictive causation. A positive result suggests that X is useful for forecasting Y, but it doesn't necessarily mean that X causes Y in a causal sense.

Suppose we have two time series: X_t and Y_t. The test involves comparing two models:

- **Model 1** (Unrestricted Model): Regress Y_t on its own past values and the past values of X

$$Y_t = \alpha + \sum_{i=1}^{k} \beta_i Y_{t-i} + \sum_{j=1}^{p} \gamma_j X_{t-j} + \epsilon_t$$

- **Model 2** (restricted Model): Regress Y_t only on its own past values:

$$Y_t = \alpha + \sum_{i=1}^{k} \beta_i Y_{t-i} + \epsilon_t$$

Granger causality test

Model 1:

$$Y_t = \alpha + \sum_{i=1}^{k} \beta_i Y_{t-i} + \sum_{j=1}^{p} \gamma_j X_{t-j} + \epsilon_t$$

Model 2:

$$Y_t = \alpha + \sum_{i=1}^{k} \beta_i Y_{t-i} + \epsilon_t$$

H_0: the coefficients of past values of X in Model 1 are zero:

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0$$

→ If we reject H_0, we conclude that X "Granger-causes" Y, meaning that past values of X improve the forecast of Y

Calculate the F-statistic given these models, with degrees of fredom parameters P and N-P-k-1

$$F = \frac{(SSR_r - SSR_{ur})/P}{SSR_{ur}/(N-k-p-1)}$$

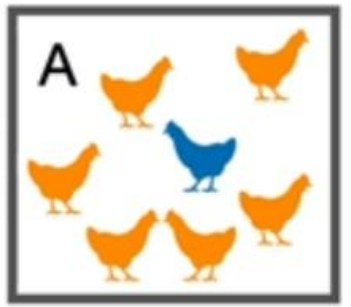
Limit: Only handles continuous variables. When it comes to categorical values, the assumptions of the traditional Granger causality test don't hold

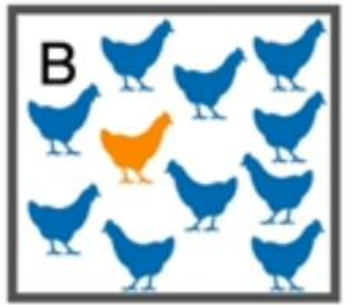
→ **Solution: Information gain**

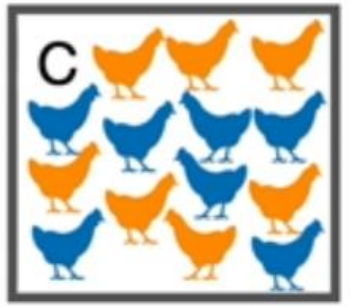
Entropy (surprise)

- If we randomly draw a chicken in zone A, we have a higher probability of drawing an orange one;
- conversely in zone B

→ If we draw an orange one in zone 1, it wouldn't be very surprising
→ if we had drawn a blue one, then it would be very surprising.







Surprise is related to the inverse of probability. When probability is low, surprise is high, and vice versa.

$$\text{surprise} = \log(\frac{1}{p})$$

Example: Imagine a coin that lands heads with p=0.9 and tails with p=0.1.The surprise for heads is 0.15, and for tails, it is 3.32.The surprise associated with the sequence HHT is 0.15+0.15+3.32=3.62.We can also estimate the expected surprise:

$$\mathbb{E}(\text{surprise}) = 0.9 \times 0.15 + 0.1 \times 3.32 = 0.47$$

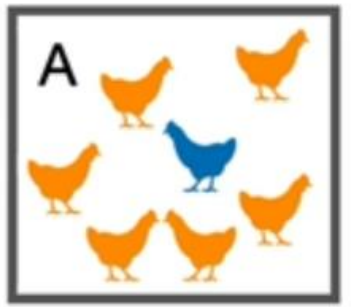
Entropy

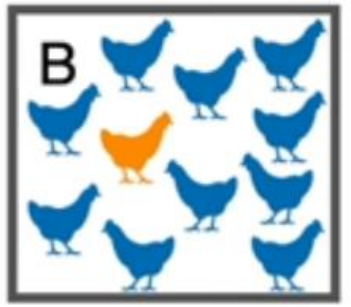
More generally, if x is the surprise:

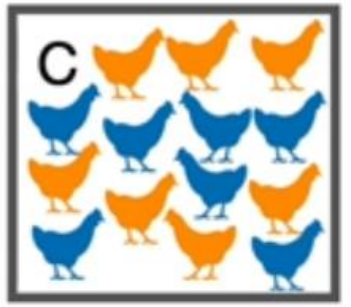
$$\begin{aligned}\text{Entropy} &= \sum_x xp(x) \\ &= \sum p(x)\log(\frac{1}{p(x)}) \\ &= \sum p(x)[\log(1) - \log(p(x))] \\ &= \sum -p(x)\log(p(x)) \\ &= -\sum p(x)\log(p(x))\end{aligned}$$

If we calculate the entropy in the case of the three chicken zones, we get:

- E(A)=0.59;
- E(B)=0.44,
- E(C)=1.







- We can use entropy to quantify the similarity or difference between zones.
- Entropy is maximal if the number of chickens of each type is the same.
- The more the imbalance between the two types increases, the lower the entropy becomes.

Information gain

Idea: The information gain measures how much it becomes easier to predict your target given a feature.

→ Metric derived from information theory used to quantify the amount of information obtained about one random variable through another random variable.
→ In the context of feature selection, it measures how much knowing the value of a feature reduces the uncertainty (entropy) about the target variable.

$$IG(Y;X) = H(Y) - H(Y|X)$$

- H(Y): entropy of the target variable Y, Measures the uncertainty inherent in the random variable Y.

$$H(Y) = -\sum_i p(y_i) \log p(y_i)$$

- H(Y|X): conditional entropy of Y given the feature X, measures the remaining uncertainty in Y after knowing X

$$H(Y|X) = -\sum_{i,j} p(x_i, y_j) \log p(y_j|x_i)$$

$$H(Y|X) = H(Y,X) - H(X)$$

Information gain recipe

1. The variables Y and X have to be categorical. If you have continuous values, like a stock value or load value, you have to discretize it.

2. Calculate the entropy: to do so, you need to compute P(X) for each X. This can be done easily by computing the number of elements in a bin divided by the total number of elements

3. Calculate the joint entropy. Same as for entropy, you just calculate the number of appearences of each possible couple (X,Y)

4. Use the formula to calculate the conditional entropy and Information Gain

Notes:

- IG should generally not be negative, as it represents the reduction in uncertainty
- Interpreting Information Gain Values:
  - If IG(Y; X) = 0 , then X provides no additional information about Y
  - If IG(Y; X) > 0, then X reduces uncertainty about Y, making X a potentially useful feature

Time for Exercise 4