Selected topics in Data Science

Grandjean, Martin (2014).
"La connaissance est un réseau".
Les Cahiers du Numérique

Wojciech Krzemień

23.10 2020, NCBJ

Last lecture recap

- Machine learning (ML) → learning from data without saying explicitly how.
- Main types of ML:
  - **Supervised Learning**– training set contains pairs {X,Y}, where Y is the expected output value/expected label
  - **Unsupervised Learning** – only feature vectors {X} given, the computer must classify/cluster by itself.
  - **Reinforcement Learning** – every action with environment provides „reward" looped-back to your model

- Dependent on outcome type:
  - Continuous output Y – **regression** model
  - Discrete output (labels) – **classification** model

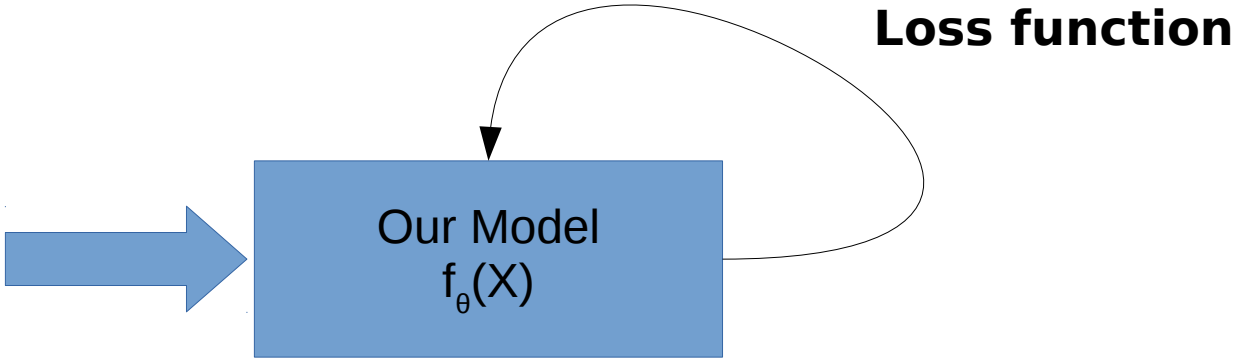
- Introduction to **K-Nearest Neighbors (KNN)** algorithm

Todays plan

- Elements of **Statistical Learning Theory:**
  - Loss function
  - Notions of generalization and stability
  - Empirical risk minimization
  - overfitting and model complexity
  - Prediction vs Inference
- Crash course in **Statistics**

Supervised learning

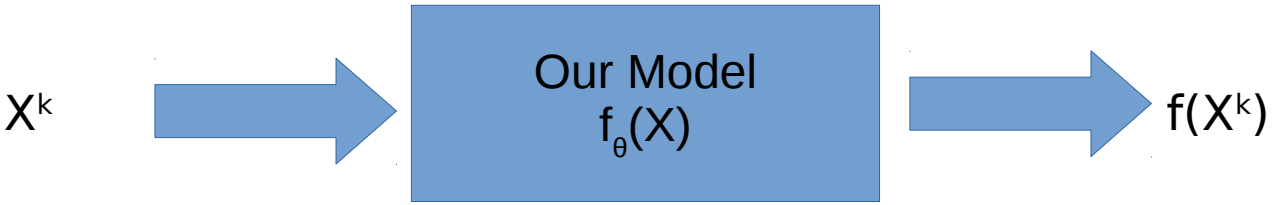
1.Training



$\{X^1,Y^1\}$, $\{X^2,Y^2\}$, ...

training data

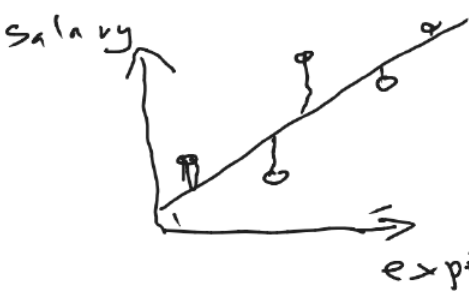
2.Prediction:



X – input, set of features (attributes)
Y – output
Training samples – pairs of $\{X^i, Y^i\}$ i=1,...N
θ – some parameters of the model

Loss function L to penalize the mistakes the model makes

regression



$\{x^1, y^1\}, \{x^2, y^2\} \ldots$

$L : X \times Y \to [0, \infty)$

quadratic loss → $L(y^i, f(x^i)) \to (y^i - f(x^i))^2$

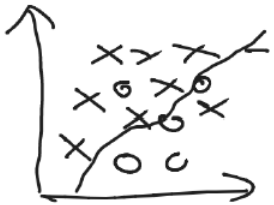
Mean squared error

training error → $E_T = \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i)) \to MSE = \frac{1}{N} \sum (y^i - f(x^i))^2$

$= \frac{1}{N} \sum |y^i - f(x^i)|$

L1, L2 norms

classification



$E_T = \frac{1}{N} \sum L(f(x^i), y^i)$

Indicator function → $\begin{cases} 0 \\ 1 \end{cases} \Sigma[f(x^i) \neq y^i]$

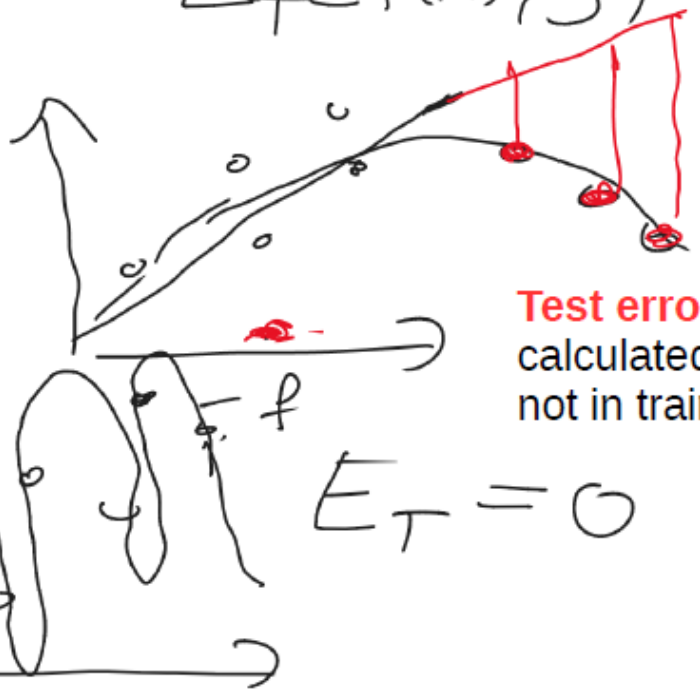
Training expressed as **minimization** problem

$f_\theta = \theta_0 + \theta_1 x_1$

Find parameters θ^* for which E_T has minimum

$E_T[f(x), y)$

f_{θ^*}



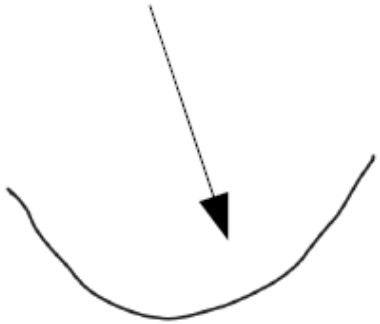
Test errors
calculated for points
not in training set

f

$E_T = 0$

OVERFITTING

Convex function



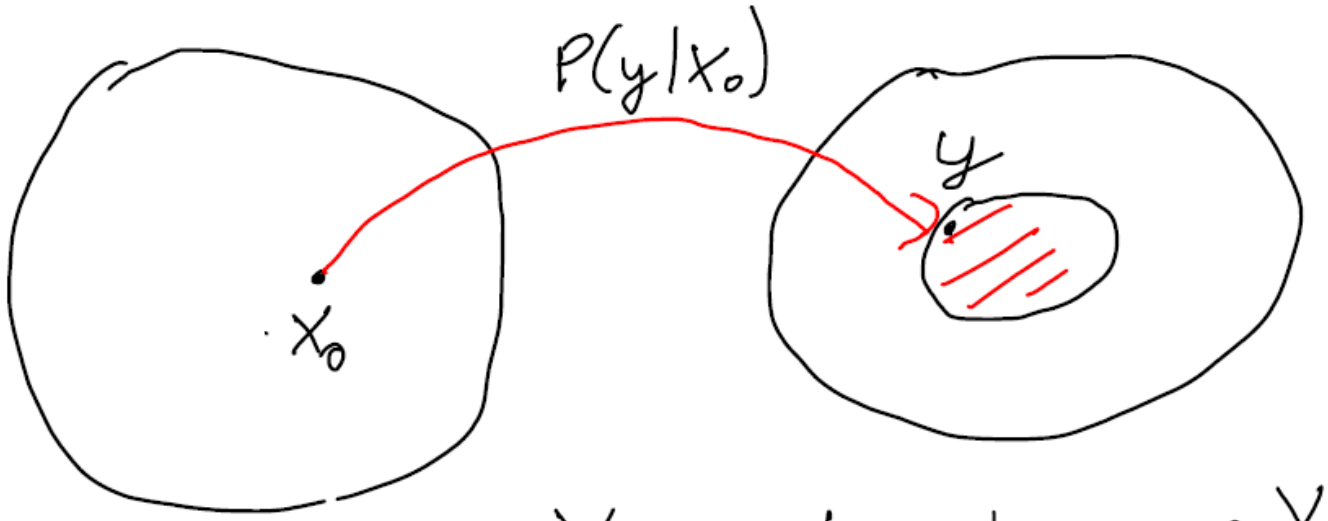
No more than one minimum

Generalization
Model should describe all the data

Statistical Learning Theory

$L : Y \times Y \to [0, \infty)$

Loss function



feature space X

e.g. $\mathbb{R}^p$

target space Y

e.g. $y \in \mathbb{R}$ - regression

$y \in \{0, 1\}$ - binary classification

$Z = X \times Y$, Z is a random variable with a joint probability density distribution

$$S_Z = S_{XY}$$

Distribution $p(z)$ (p_{XY}) is unknown.

We have only a training sample:

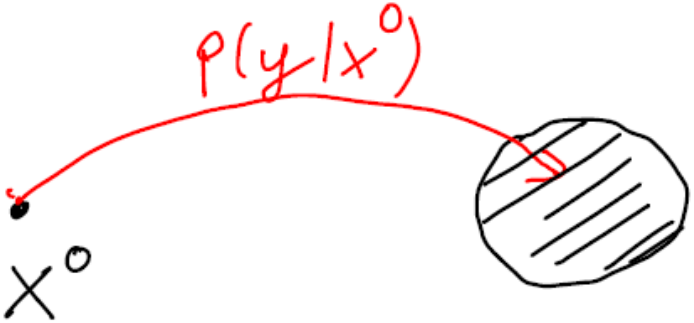
$$T = \{\{x^1, y^1\}, \ldots \{x^N, y^N\}\} \quad N - \text{number of samples}$$

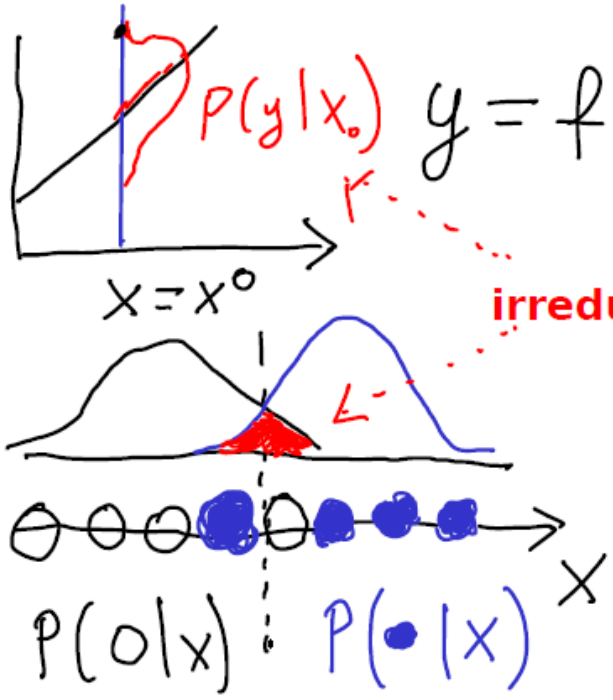
Assumption: Training pairs are independent and identically distributed (iid) from Z

$$P(z) = P_X(x) \cdot P_{Y|X}(y|x)$$

gaussian noise

$$y = f^*(x) + \varepsilon$$





A set of features x0 can correspond to different anwsers

$P_X(x)$ this probability incorporates e.g. uncertainty due to the sampling procedure.

Key question: Starting from given training set T, and without the knowledge about the underlaying distribution $p(z) = p_X(x) \cdot p_{y|X}(y|x)$ how to find a model f that makes 'good' predictions.

GENERALIZATION

Function (hypothesis) space H: $f : X \to Y$, contains all possible models we allow to be checked.
What conditions / constraints one should apply on function space H to ensure the generalization?

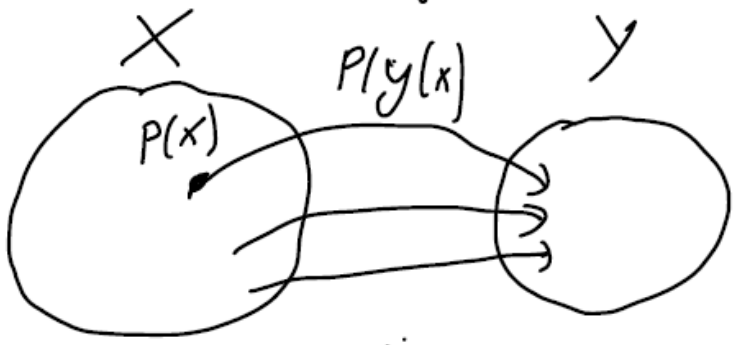
Vapnik's theory gave the answer.

Expected prediction error (true error, risk function).

Joint probability density function

$$E_{PE}[f] = \int\int L(f(x), y)\, \rho_{X,Y}\, dx\, dy$$

The error that given model f would commit calculated over the whole $Z = X \times Y$ space.



Since we don't know what is $\rho_{X,Y}$ we cannot compute it

$$E_T[f] = \frac{1}{N}\sum_i^N L(f(x^i, y^i)) \quad \text{- training error, empirical risk}$$

For the 'good' model we expect: $E_T \approx E_{PE}$

GENERALIZATION:

$$\forall_{T,S_{XY}} \lim_{N\to\infty} \left| E_{PE}(f_T) - E_T[f_T] \right| = 0$$

convergence in probability

Convergence in probability - simple example:

$$mean_T = \frac{1}{N}\sum X_i \longrightarrow mean_T \to \text{population expectation}$$

going with $N \to \infty$

sample $T \in \{X_1 \dots X_n\}$

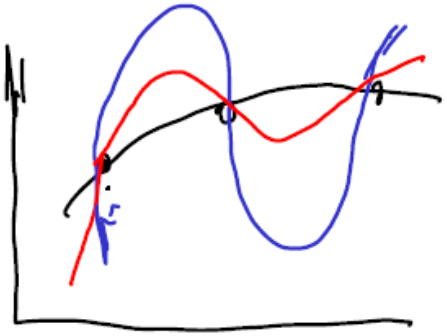
N numbers

Formally: $\forall_{\varepsilon>0} \lim_{N\to\infty} Pr\left(\left| mean_N - E \right| > \varepsilon \right) = 0$

Empirical risk minimization (ERM) by V. Vapnik et al.

He is the creator of Support Vector Machines algos

If we use training set to calculate E_T and find the optimal model f^* how we must restrict H space to fulfil generalization condition?



H space must be uniform Glivienko-Cantelli class.

For details see:

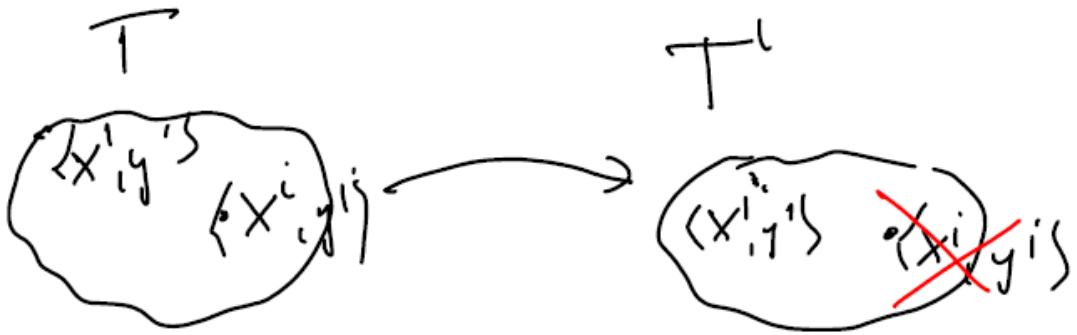
Poggio et al. proved that the ERM approach is equivalent to the requirements for a learning algorithm to be stable

Approach can be applied to analyze broader class of ML algos.

CV_{loo} - Cross-Validation leave-one-out stability

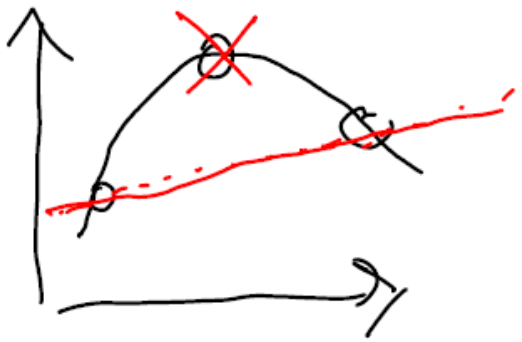
$$\forall_{S_{XY}} \lim_{N \to \infty} \sup_{i \in \{1, \dots N\}} \left| L[f_T(x^i), y^i] - L[f_{T^i}(x^i), y^i] \right| = 0$$

in probability

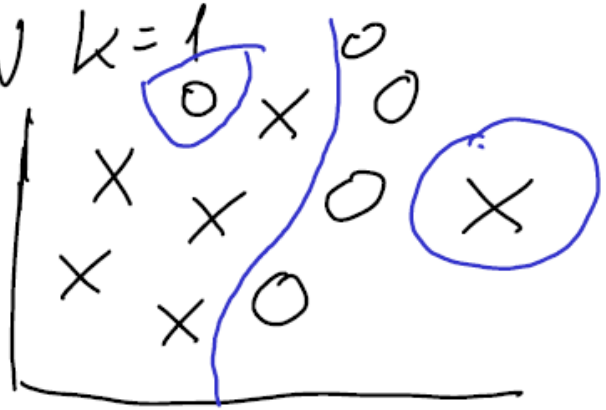


i-th point removed from the training sample

If we remove one point from our training sample how much does it perturbe the loss



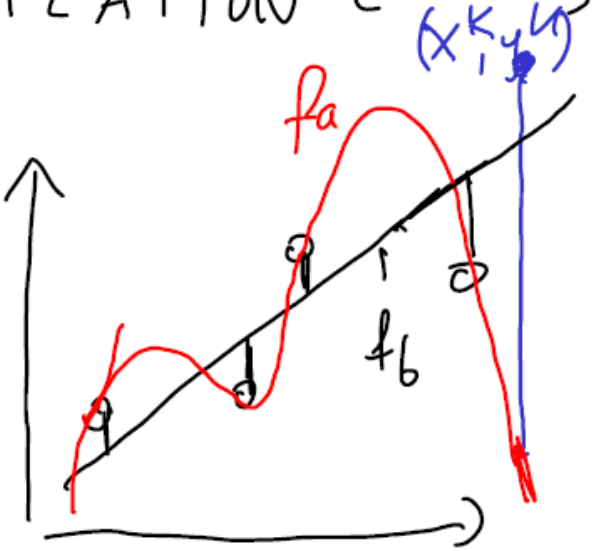
K-NN $k=1$



GENERALIZATION ⟷ STABILITY

$E_T[f_a] = 0$

$E_T[f_b] > 0$



model complexity:

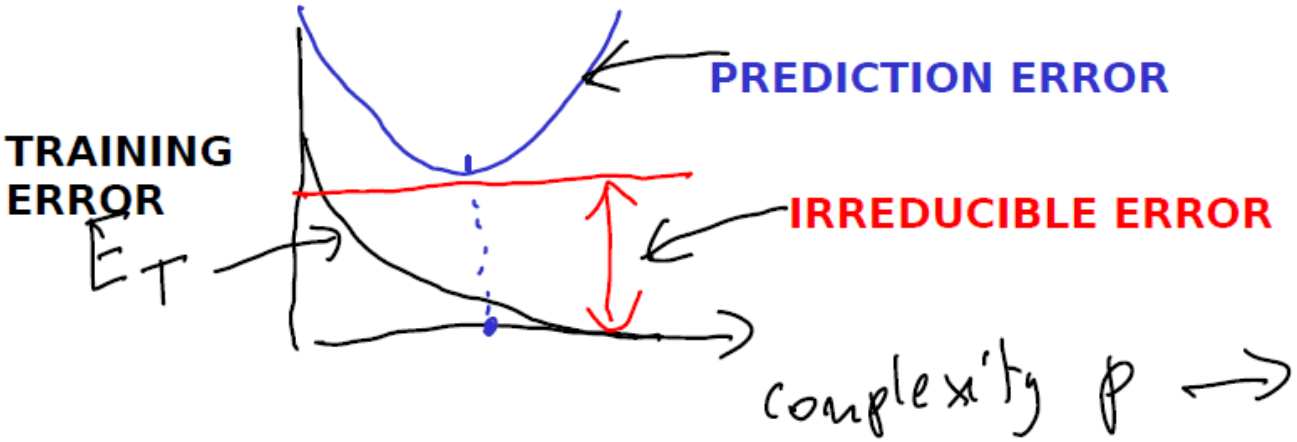
$f_1(x) = \theta_0 + \theta_1 x$

$f_2(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

⋮

$E[f_a(x^k, y^k)] \gg L[f_b(x^k, y^k)]$ violation of polynomial

↓ complexity



Statistical Learning Theory

Generalization ↔ Stability

- V. Vapnik „Principles of Risk Minimization for Learning Theory ”(link)

- P. Poggio et al. „General conditions for predictivity in learning theory”, Nature vol. 428 (2004) – (link)

Prediction vs inference

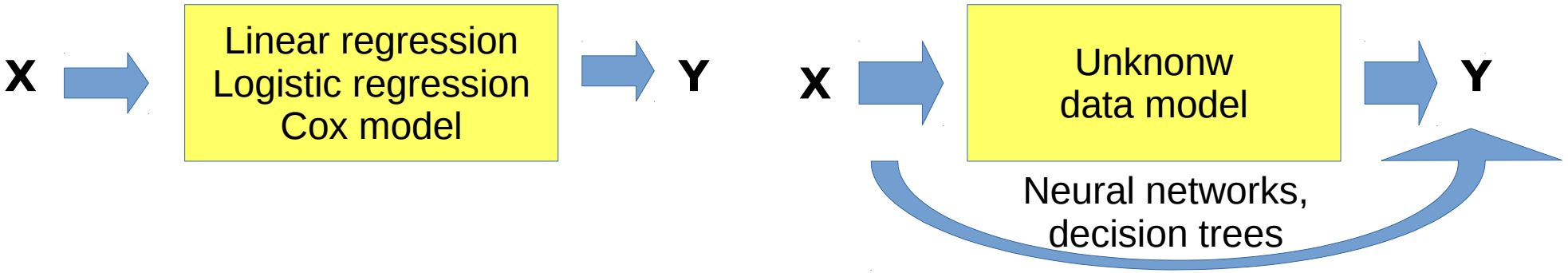
- **Prediction**: for a model $\mathbf{f_\theta(X)}$ we want to **predict**, with the best possible accuracy output $\mathbf{y^i}$ for the unknown vector $\mathbf{X^i}$
- **Inference** – for a model $\mathbf{f_\theta(X)}$, where $X = [x_1, x_2, x_3 \ldots, x_p]$ we would like to **understand** how the data is generated, what are the relations between the input and output variables e.g. how the change of given input variable $\mathbf{x_i}$ (element of the vector X) affects the change of **y**

Prediction vs inference

- **Statisticians' approach:** First we construct the statistical model that describes the **data generation**. Then we fit the parameters of the model and we can perform predictions and inference.

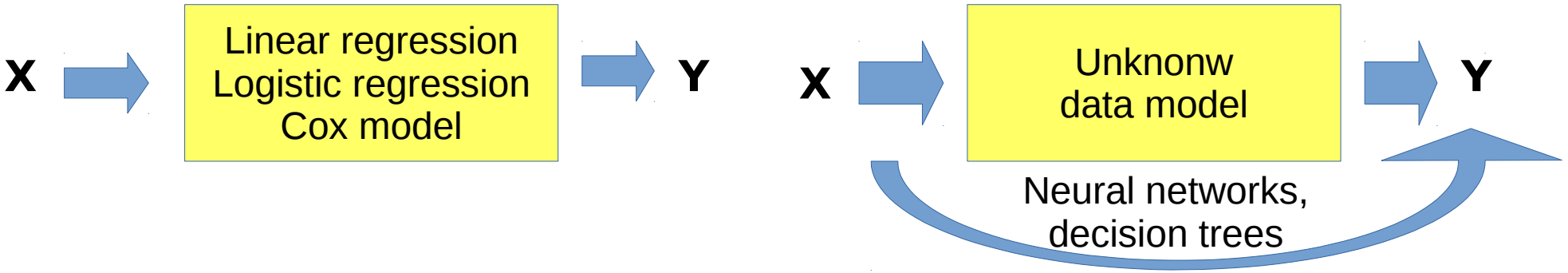
- **ML ('algorithmic') approach:** Who cares about how the data is generated ? The final goal is to have a an **algorithm** that accurately predicts the output.

Prediction vs inference



Leo Breiman „*Statistcal Modelling:The two cultures*” (link)

Prediction vs inference



Accuracy vs Interpretability

Leo Breiman „*Statistcal Modelling:The two cultures*” (link)

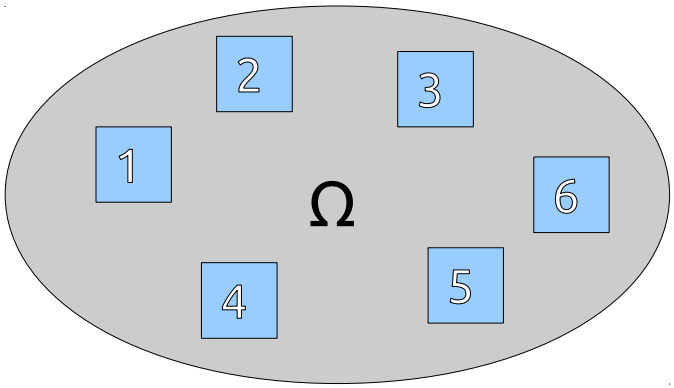
Quiz

Crash course in probability (reminder)

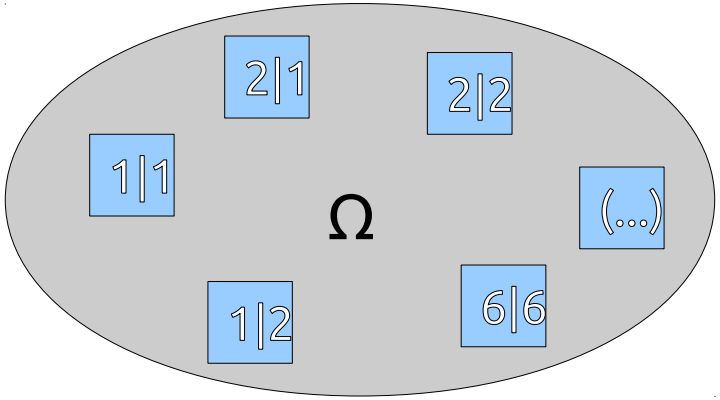
Probability reminder

Sample space Ω – set of all possible outcomes from a random experiment

Ex1. Rolling one die: $\Omega = \{1, 2, 3, 4, 5, 6\}$



Ex2. Rolling two dice: $\Omega = \{1|1, 1|2, 2|1, 2|2, .., 6|6\}$



Probability reminder

Event space F – set of events (outcome of experiments) $\mathbf{E_i}$ being a subsets of Ω.

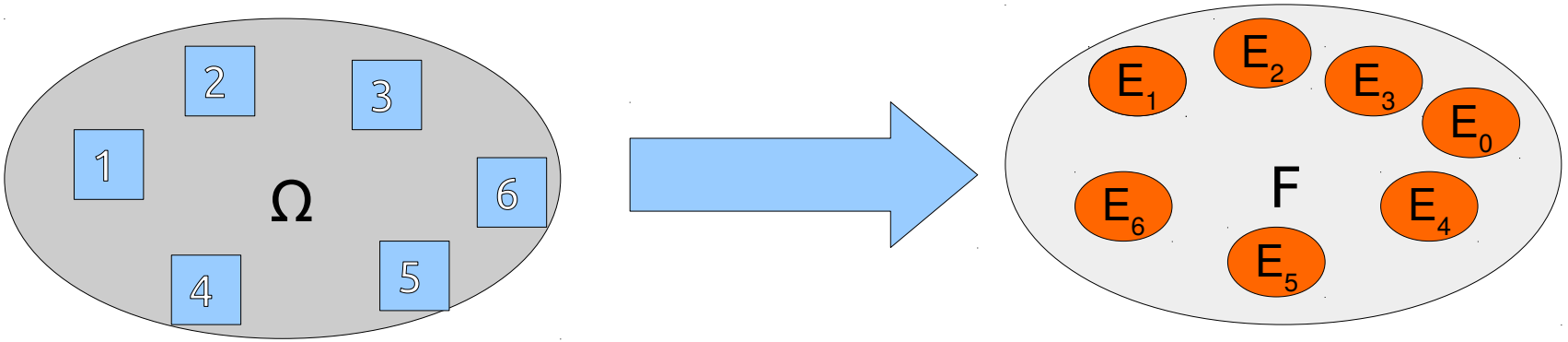
- We add an „impossible” event $E_0 = \{\emptyset\}$
- The way you define events is not unique

Ex3. Rolling one die: $\Omega = \{1, 2, 3, 4, 5, 6\}$

We are interested in probability of selecting given number:

$E_1=\{1\}$, $E_2 = \{2\}$ $E_3 =\{3\}$ $E_4 = \{4\}$, $E_5=\{5\}$, $E_6= \{6\}$

$F = \{E_1, E_2, E_3, E_4, E_5, E_6, E_0\}$



Probability reminder

Event space F – set of events (outcome of experiments) $\mathbf{E_i}$ being a subsets of Ω.

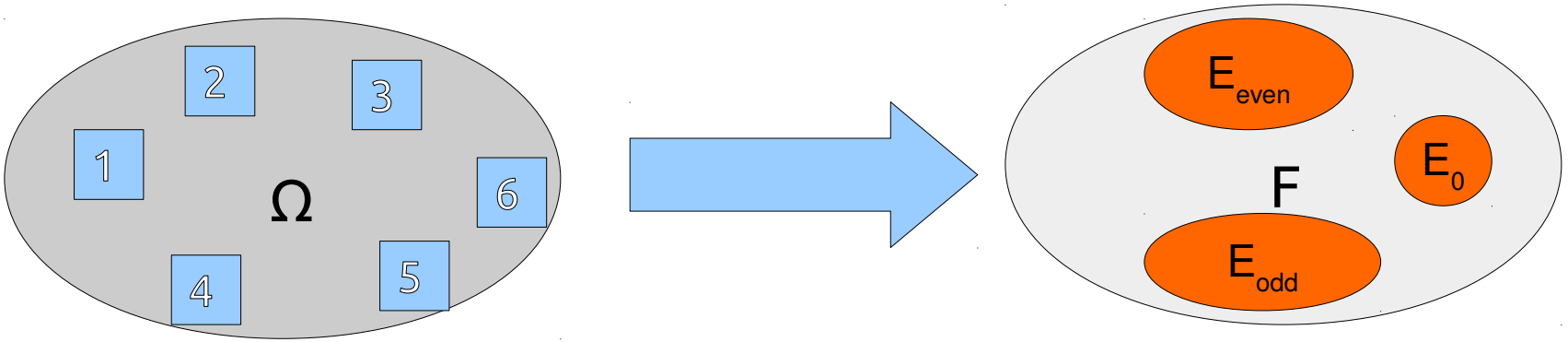
- We add an „impossible" event $E_0 = \{\emptyset\}$
- The way you define events is not unique

Ex4. Rolling one die: $\Omega = \{1, 2, 3, 4, 5, 6\}$

We are interested in probability of selecting even/odd number:

$E_{odd} = \{1, 3, 5\}$, $E_{even} = \{2, 4, 6\}$

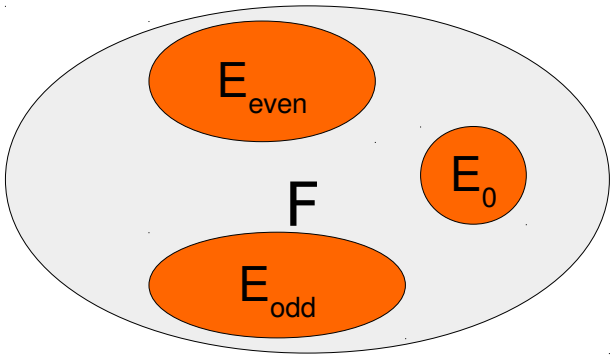
$F = \{E_{odd}, E_{even}, E_0\}$

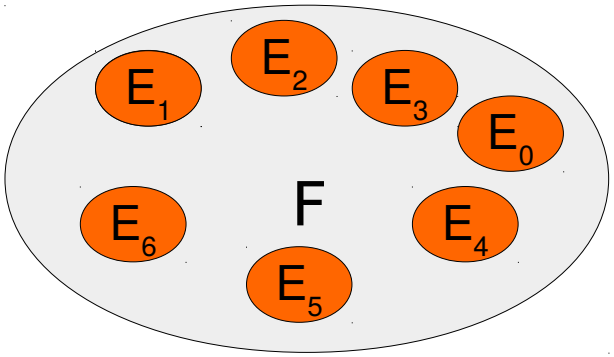


Probability reminder

Event space F – set of events (outcome of experiments) $\mathbf{E_i}$ being a subsets of Ω.

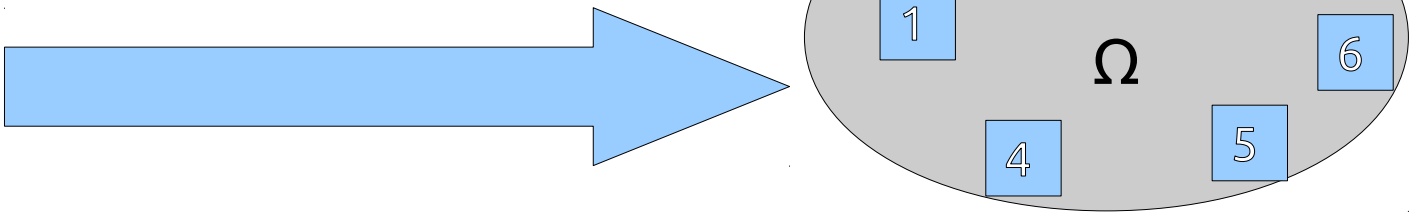
- We add an „impossible" event $E_0 = \{\emptyset\}$
- The way you define events is not unique





$\Omega = E_{even} \cup E_{odd}$

$\Omega = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5 \cup E_6$



***Remark:** Formally we should define σ-algebra ~ F must be closed under complement and under countable unions and intersections

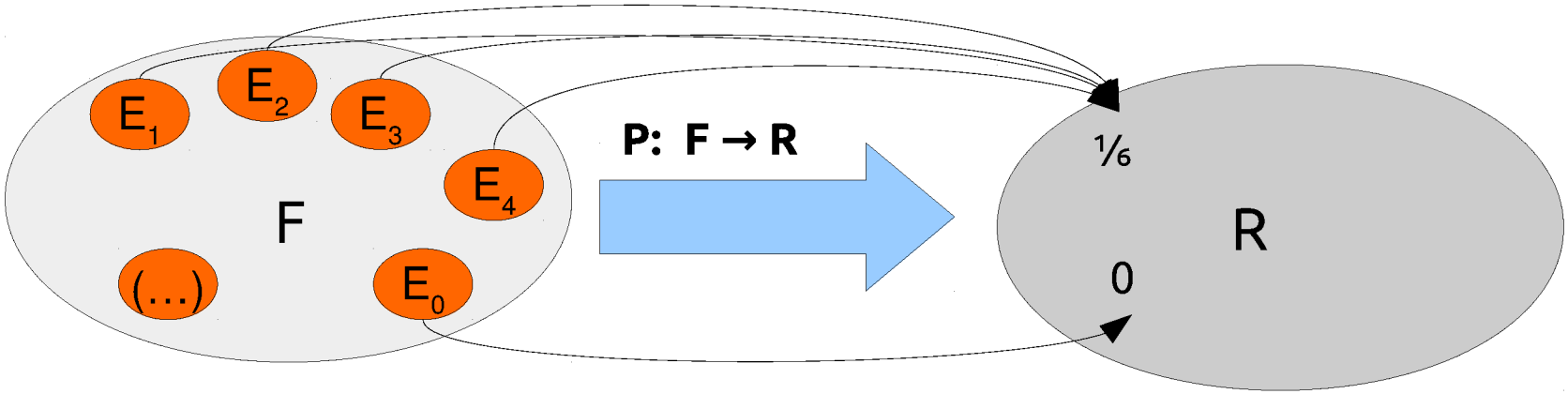
Problem Tossing a coin three times:

H – head,

T – tail

- What is the sample space Ω ?
- What is the event space F, for selection of at least two heads?

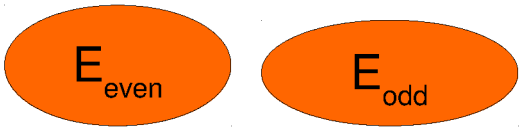
Probability measure P– Function that maps events from space **F** to real numbers



Kolmogorov's Axioms:

1) Probability is non-negative: $P(E) \geq 0$
2) Probability of the entire sample set : $P(\Omega) = 1$
3) If $E_1, \ldots E_n$ are disjoint events: $E_i \cap E_j = \emptyset$ for $i \neq j$

$$P(\cup E_i) = \sum P(E_i)$$



Some properties of probability measure

Properties: $E_i, Ej \in F$

- $P(\emptyset) = 0$
- $E_i \subseteq E_j \rightarrow P(Ei) <= P(E_j)$
- $P(E_i \cap E_j) = \min (P(E_i), P(E_j))$
- $P(E_i \cup E_j) <= P(E_i) + P(E_j)$
- $P(\Omega \backslash E) = 1 - P(E)$
- If $E_1, \ldots E_n$ are disjoint events: $E_i \cap E_j = \emptyset$ for $i \neq j$ and $\cup E_i = \Omega \rightarrow \sum P(E_i) = 1$

Conditional probability

A, B ∈ F such that P(B) ≠ 0

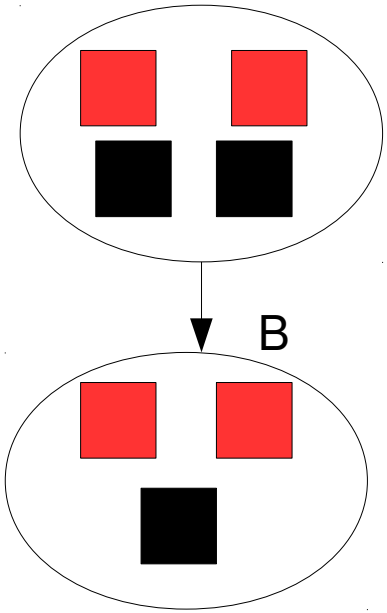
Probability that event A happens “after” event B occurred:

P(A|B) = P(A∩B)/P(B)

Ex5. Selecting a square from a bag with red (R) and black (B) squares

- A – selection of a red square
- B – selection of a black square

P(A) = ½ , P(B) = ½



A|B selection of a red square „after” selecting the black one:

P(A|B) = 2/3

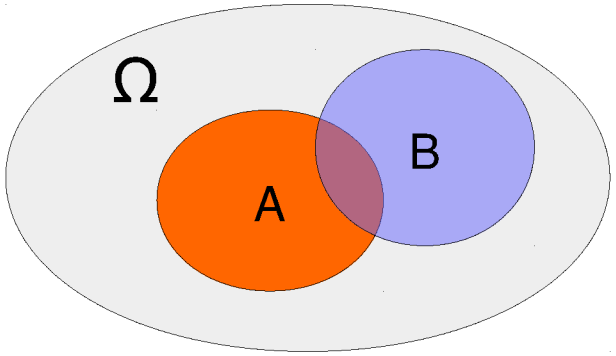
Conditional probability

$A, B \in F$ such that $P(B) \neq 0$

Probability that event A happens “after” event B occurred:

$$\mathbf{P(A|B) = P(A \cap B)/P(B)}$$

- P(A|B) fulfills Kolmogorov’s axiom → its a proper probability :-)



$$\mathbf{P(A|B) = P(A \cap B)/P(B)}$$

- P(A) can be treated as conditional $P(A) = P(A|\Omega)$

Independent events

A, B ∈ F are independent if:

$P(A \cap B) = P(A) * P(B) \leftrightarrow P(A|B) = P(A), P(B|A) = P(B)$

Ex6. Selecting (with returning) a square from a bag with red (R) and black (B) squares

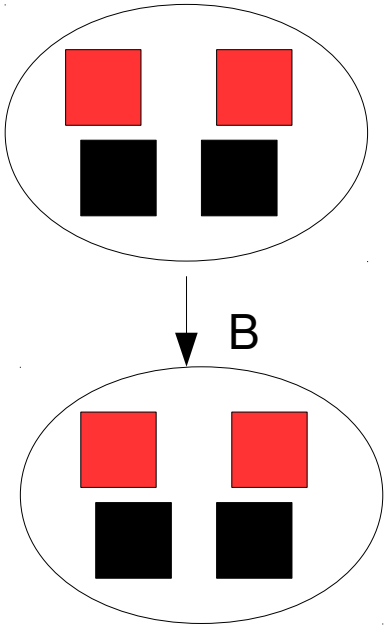
- A – selection of a red square
- B – selection of a black square

$P(A) = ½$, $P(B) = ½$

A|B selection of a red square „after" selecting the black one:

$P(A|B) = 2/4 = ½ = P(A)$

A and B are independent



Bayes's theorem

- A,B - events from Event space

$P(A|B) = P(A \cap B)/P(B) \rightarrow P(A \cap B) = P(A|B) * P(B)$

$P(B|A) = P(B \cap A)/P(B) \rightarrow P(B \cap A) = P(B|A) * P(A)$

but, $P(A \cap B) = P(B \cap A) \rightarrow P(A|B) * P(B) = P(B|A) * P(A)$

$$P(A|B) = [\, P(B|A) * P(A) \,] / P(B) \quad \text{for } P(B) \neq 0$$

Bayes's theorem

Law of total probabilities:

$P(B) = \Sigma_i P(B|A_i) * P(A_i)$ for A_i being disjoint subsets such that $\Omega = \cup A_i$

$$P(A|B) = [P(B|A) * P(A)] / P(B) \text{ for } P(B) \neq 0$$

$$P(A|B) = [P(B|A) * P(A)] / [\Sigma_i P(B|A_i) * P(A_i)] \text{ for } P(B) \neq 0$$

Problem :

Show that $P(B) = \Sigma_i P(B|A_i) * P(A_i)$

for A_i being disjoint sets, partitioning the whole sample space ($\Omega = U A_i$) and

$P(A_i) > 0$ for all i

Problem :

- A pharmaceutical company developed a test for detecting the rare disease, which is carried by 0.5 % cases of the whole population . Let's assume that the test gives the positive results for 96% of the cases if the patient is ill, but it also gives the positive results in 5% of the healthy patient. What is the probability that a patient is ill if his test gave a positive result?

Bayes's theorem

$$P(A|B) = [\, P(B|A) * P(A) \,] / P(B) \quad \text{for } P(B) \neq 0$$

posterior probability = [observation * prior probability] / evidence

How to interpret the posterior probability P(A|B) ?

Experiment measured a mass of some particle B what can one say about the "true" mass A?

Frequentist vs Bayesian interpretations of probability

- **Frequentist approach**: probability is a **relative frequency** of occurrences

$P(A|B) = P(A \cap B)/P(B) =$ 95 % → If we repeat the same experiment N times (N→ infinity), the ratio between occurrences of A together with B divided by the occurrences of B independently of A will be equal to 95%.

Frequentist vs Bayesian interpretations of probability

- **Frequentist approach**: probability is a **relative frequency** of occurrences

P(A|B) = P(A∩B)/P(B) = 95 % → If we repeat the same experiment N times (N→ infinity), the ratio between occurrences of A together with B divided by the occurrences of B independently of A will be equal to 95%.

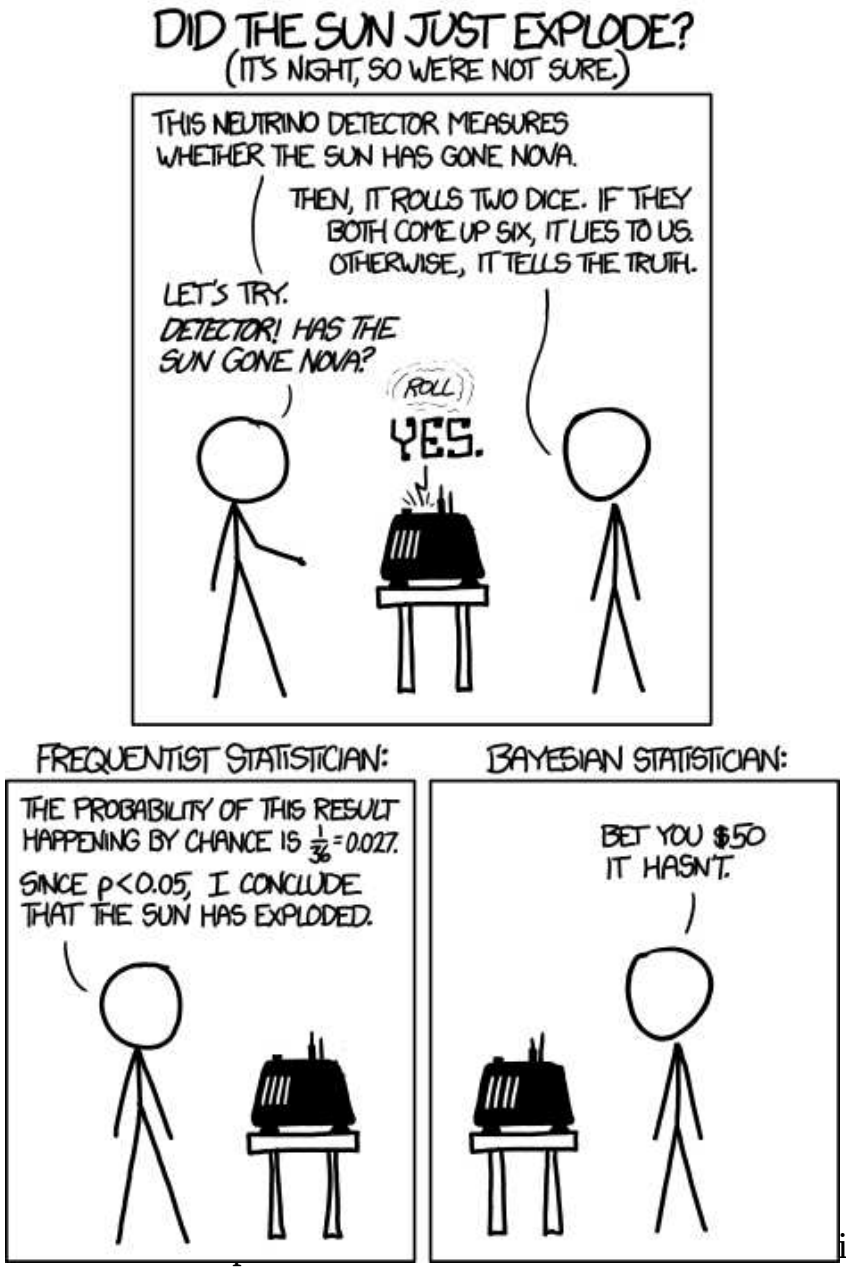
- **Bayesian approach**: probability is a measure of **degree of belief** of given hypothesis:

P(A| B) = 95 % → Probability that hypothesis is true if we measured(obtained) B is 95%.

P(theory |data) = P(data | model) * P(model)

In most of the cases equivalent results independent of the approach.

Interpretation problems leads to war ...



Physicists(at least particle) should be aware of this difference

- How to interpret unique phenomena in the frequentist framework e.g. Big Bang ?
- Different results can be obtained for upper limits calculations (depending on approach) → see Particle Data Group website

Further reading:

- B. Efron *Bayesians, Frequentists, and Scientists*

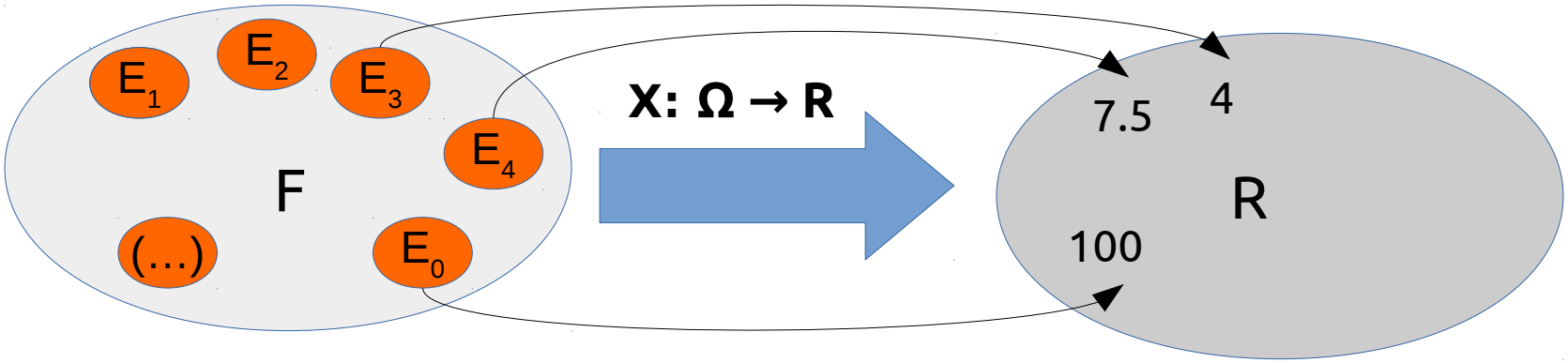
http://statweb.stanford.edu/~ckirby/brad/papers/2005BayesFreqSci.pdf

- L. Lyons, *Bayes and Frequentism: A particle physicist perspective*:

https://arxiv.org/pdf/1301.1273.pdf

Random variables

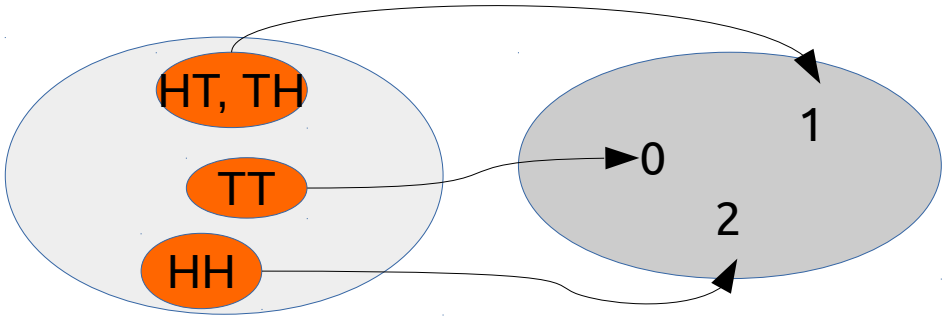
Random variable – function that maps events to some real numbers(*)



- Values are not limited to [0,1] range as probability
- $X(E_i) = x$, where: X- random variable, x – value of the variable

Ex7. Tossing two coins: Ω = { HT , TH , HH , TT }

X is a random variable - number of heads



X(E) = 1 if E = {HT, TH}
0 if E= {TT}
2 if E = {HH}

***Remark:** Formally we should allow only Borel functions

Random variables II

Random variables:

- **Discrete:** P(X = k) = P({E: X(E) = k})
- **Continuous:** P(a <= X <= b) = P({E: a <= X <b=})

Ex8. X numeric output of a die :

Ω = { 1 , 2 , 3 , 4 , 5, 6 } X = {1,2,3,4,5,6}

Ex9. Y - height measurement from a group of people:

Ω = { P1 , P2 , P3 ... }, Y = {105 cm, 183 cm, 160,5 cm, ...}

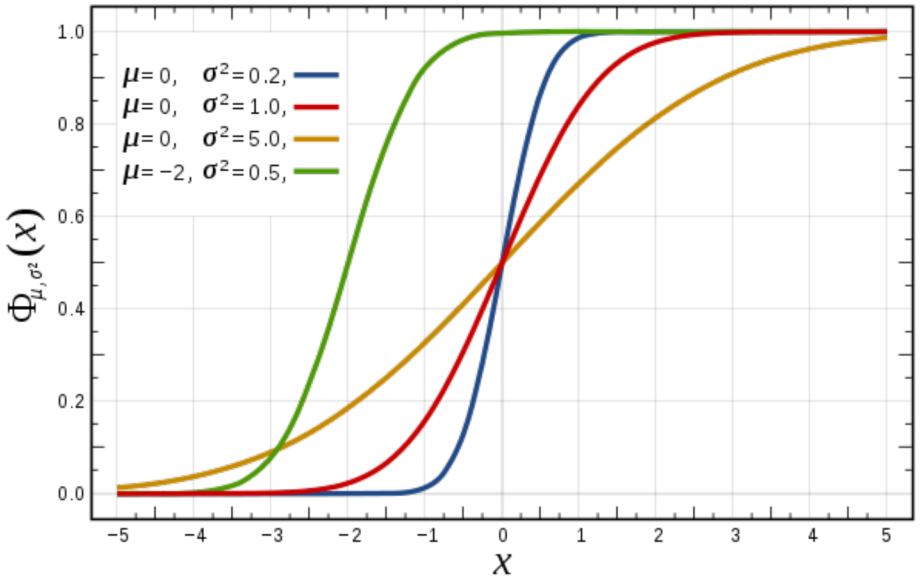
CDF

Cumulative Distribution Function (CDF) – describes probability measure for given random variable

- F_X:R → [0,1]
- $F_X(x) = P(X <= x)$

Properties:

- $0 <= F_X(x) <= 1$
- $\lim x \to -\infty F_X(x) = 0$
- $\lim x \to +\infty F_X(x) = 1$
- $p <= q \to F_X(p) <= F_X(q)$

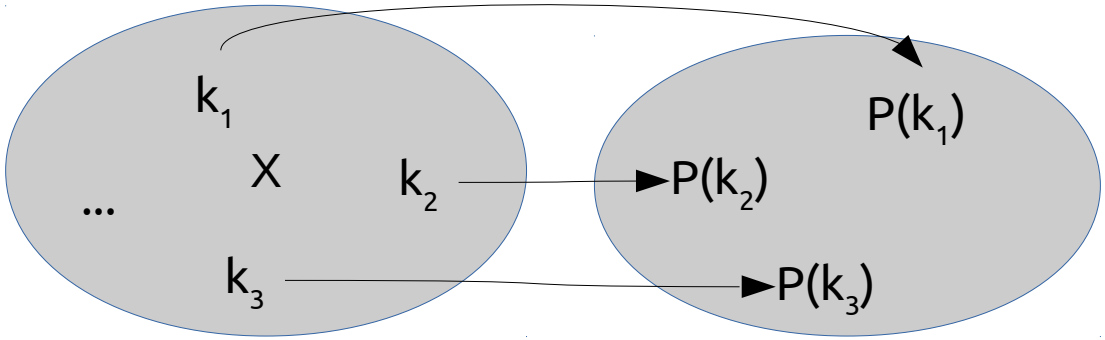


https://en.wikipedia.org/wiki/Cumulative_distribution_function

PMF

Probability Mass Function (PMF) – describes probability measure for a given discrete random variable:

- $p_X: C \rightarrow R$
- $p_X(k) = P(X=k)$



Properties:

- $0 <= p_X(x) <= 1$
- $\Sigma\, p_X(k) = 1$ for all possible k
- $\Sigma\, p_X(k) = P(X \in A)$ for $k \in A$

PDF

Probability Density Function (PDF) – describes probability measure for a given continuous random variable (F_X must be a differentiable function):

- $f_X: R \rightarrow R$
- $f_X(x) = dF_X(x)/dx$

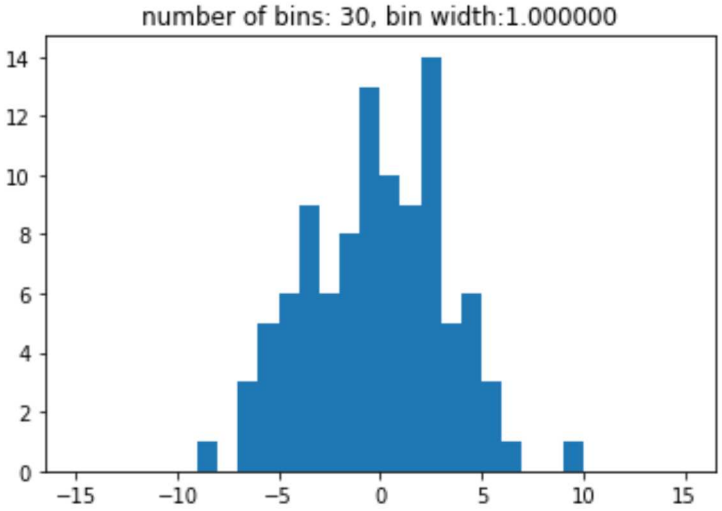
$$P(x < X < x+ \Delta x) = f_X(x)\, \Delta x$$

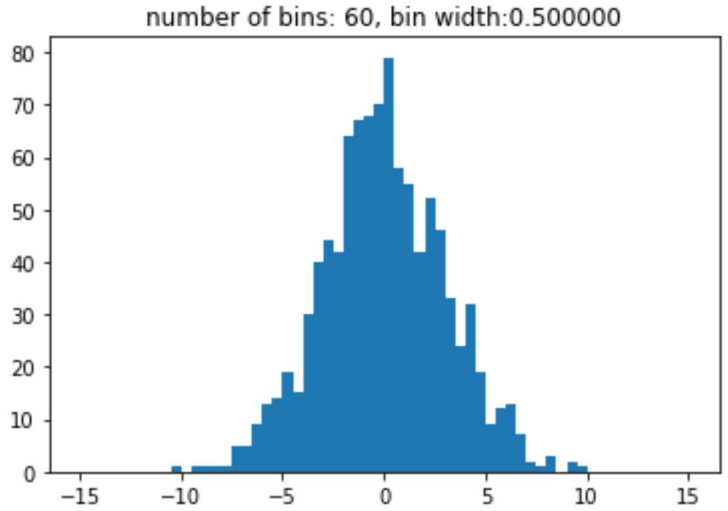
Properties:

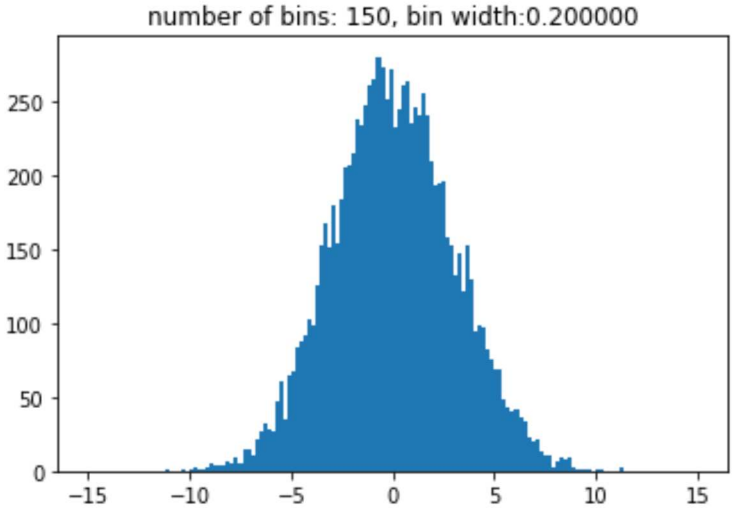
- $f_X(x) >= 0$
- $\int f_X(x)dx = 1$ for x $-\infty$ to ∞
- $\int f_X(x)dx = P(X \in A)$ for $x \in A$

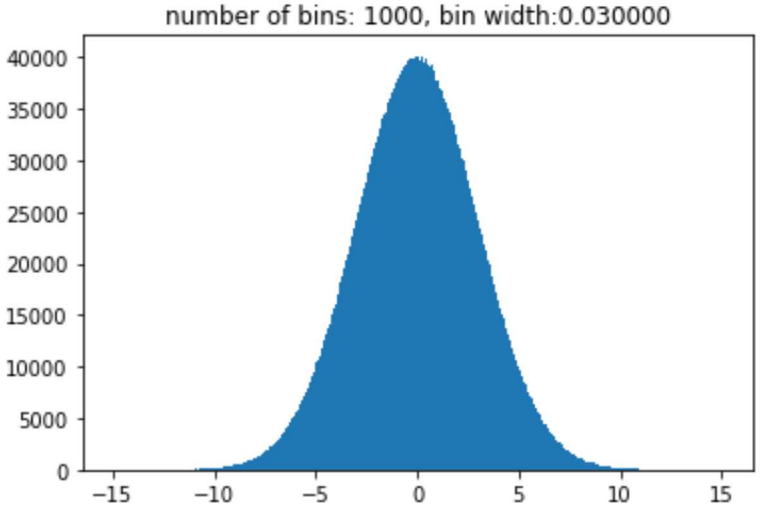
- **Remark:** f_x cannot express probability since can have values > 1
- **Remark2:** Especially, we know that for any continuous random variable P(X=x) =0

Histograms vs PDFs



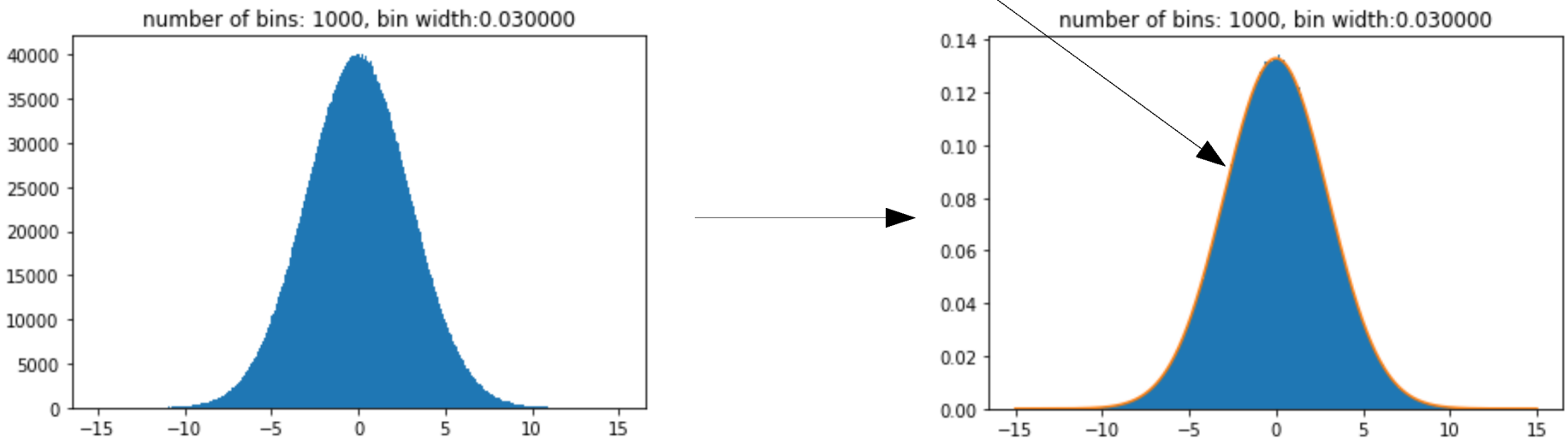






Histograms vs PDFs

red line – gaussian pdf



Dividing every bin by the total number of events and by the bin width

Notebook:
https://github.com/wkrzemien/dataScienceAndML2020/blob/master/notebooks/statistics/pdfs_histograms.ipynb

Problem Tossing two coins: Ω = { HT , TH , HH , TT }

X is a random variable - number of heads
X(E) = 1 if E = {HT, TH}
0 if E= {TT}
2 if E = {HH}

- What are CDF and PMF functions?
- Draw it

Problem What are the PDF and CDF functions for the uniform distribution defined for X =[0, a]?

Problem *
Write a program that simulates the tossing of two coins, and estimate the CDF and PMF functions for the first problem above

Thank you