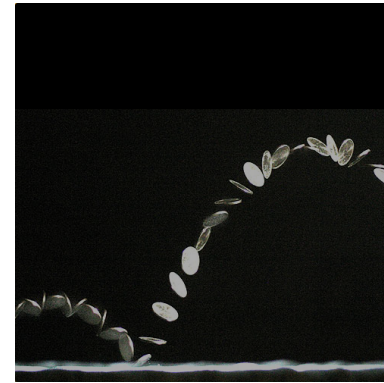
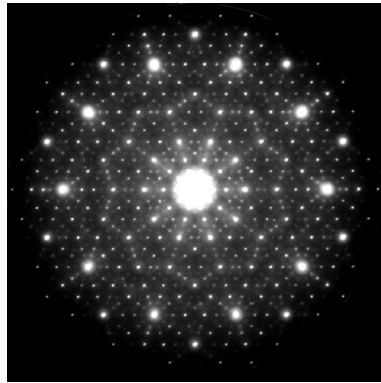


LTAT.02.004 MACHINE LEARNING II

## **Basics of probabilistic modelling**

Sven Laur  
University of Tartu

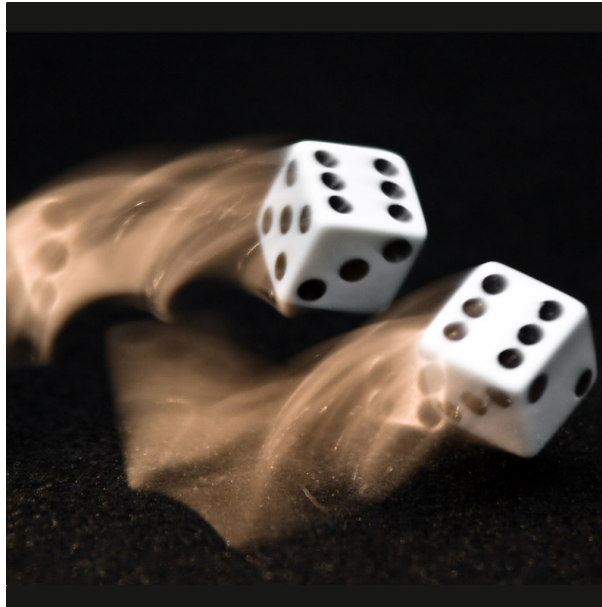
# What is probability?



Probability is a measure of uncertainty which can rise in several ways

- ▷ Intrinsic uncertainty in the system
- ▷ Uncertainty caused by inherent instability of the system
- ▷ Uncertainty caused by lack of knowledge or control over the system

# Frequentistic interpretation of probability



Probability is an average occurrence rate in long series of experiments.

- ▷ Law of large numbers
- ▷ Probability is a collective property
- ▷ Probabilities can be assigned only to future events

# Bayesian interpretation of probability



Probability reflects persons individual beliefs on future or unknown events.

- ▷ Belief updates through the Bayes rule
- ▷ Probability is an inherently subjective property
- ▷ Probabilities can be assigned to past, present and future events

# Ultra-frequentistic interpretation of probability



Events with small enough probability do not occur

- ▷ The main tool in classical statistics
- ▷ Errors in judgement does not matter if a gamma ray pulse kills us.
- ▷ One must avoid the lottery paradox in the reasoning

# The goal of statistical inference

## Frequentist goal

- ▷ The aim of statistics is to design algorithms that work well on average.
- ▷ For that one needs to specify probabilistic model for data sources.
- ▷ Confidence is the fraction of cases the algorithm works as specified.

## Bayesian goal

- ▷ The aim of statistics is to design algorithms that allow *rational individuals* to reliably update their beliefs through Bayes formula
- ▷ Besides the data source model one has to provide model for initial beliefs.
- ▷ Correctness of an algorithm does not make sense.

# Frequentistic methods

## Causation between zero-one events

Assume that condition  $A$  causes the event  $B = 1$  with probability  $p$ , i.e.,

$$\Pr[B = 1|A] = p$$

Then the probability is to get  $k$  ones in  $n$  independent trials is

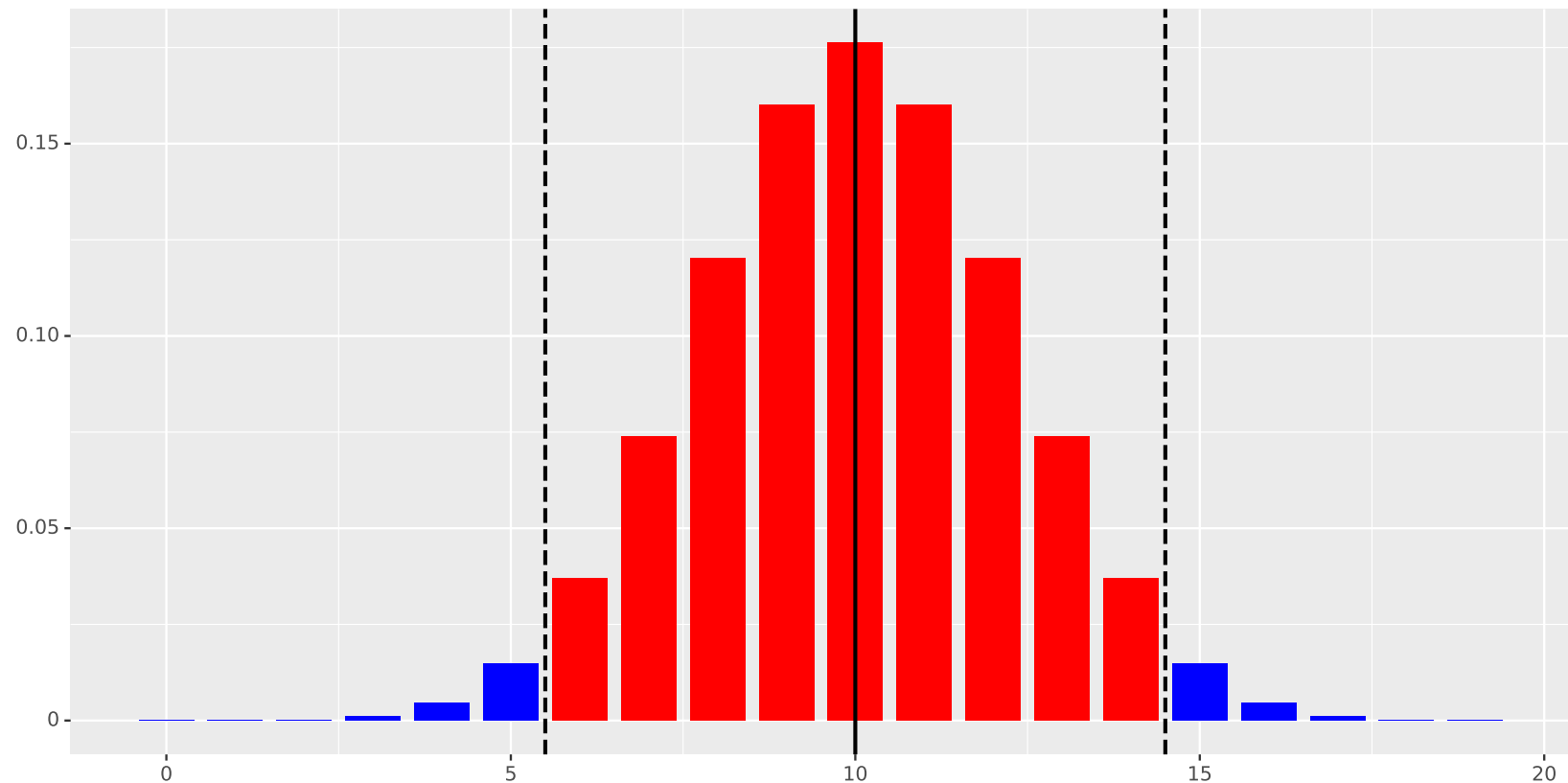
$$\Pr[B_1 + \dots + B_n = k|A] = \binom{n}{k} p^k (1 - p)^{n-k}$$

The number of ones is known to have a *binomial distribution*

$$B_1 + \dots + B_n \sim \text{Bin}(n, p)$$



## Illustration



The distribution of  $B_1 + \dots + B_n$  depends solely on the number of trials  $n$  and the probability  $p$ . Some values of  $B_1 + \dots + B_n$  are very unlikely.

# How to build a statistical test

## I. Null hypothesis:

- ▷ The probability of heads in a coinflip is  $\Pr[B_i = 1] = p$ .

## II. Choose value to compute aka test statistic:

- ▷ Our test statistic will be  $B_1 + \dots + B_n$ .

## III. Consequences on the observations:

- ▷ The observed sum  $B_1 + \dots + B_n \sim \text{Bin}(n = 20, p = 0.5)$ .
- ▷ Limit on the tail probability  $\Pr[|B_1 + \dots + B_n - 10| \geq 6] \leq 5\%$

## IV. Test procedure

- ▷ Reject null hypothesis at *significance level* 5% if  $|B_1 + \dots + B_n - 10| \geq 6$ .

## Properties of statistical tests

Statistical test is a classification algorithm designed to distinguish a fixed distribution of negative examples specified by a null hypothesis.

Any *static* classification algorithm can be converted to a statistical test by finding out the percentage of false positives aka *p-value*:

- ▷ There might exist a closed form solution.
- ▷ We can always estimate p-values using simulations.
- ▷ Observations must be compressed into a single decision value.

Testing several hypothesis in parallel increases the number of false positives. Several p-value adjustment methods are used to correct the issue:

- ▷ Bonferroni correction is almost optimal
- ▷ FDR correction controls the expected number false positives

# How to build confidence intervals

## I. Construct a family of statistical tests:

- ▷ Define a statistical test  $T_p$  for all possible parameter values  $p$ .
- ▷ All tests should share the same test statistic.

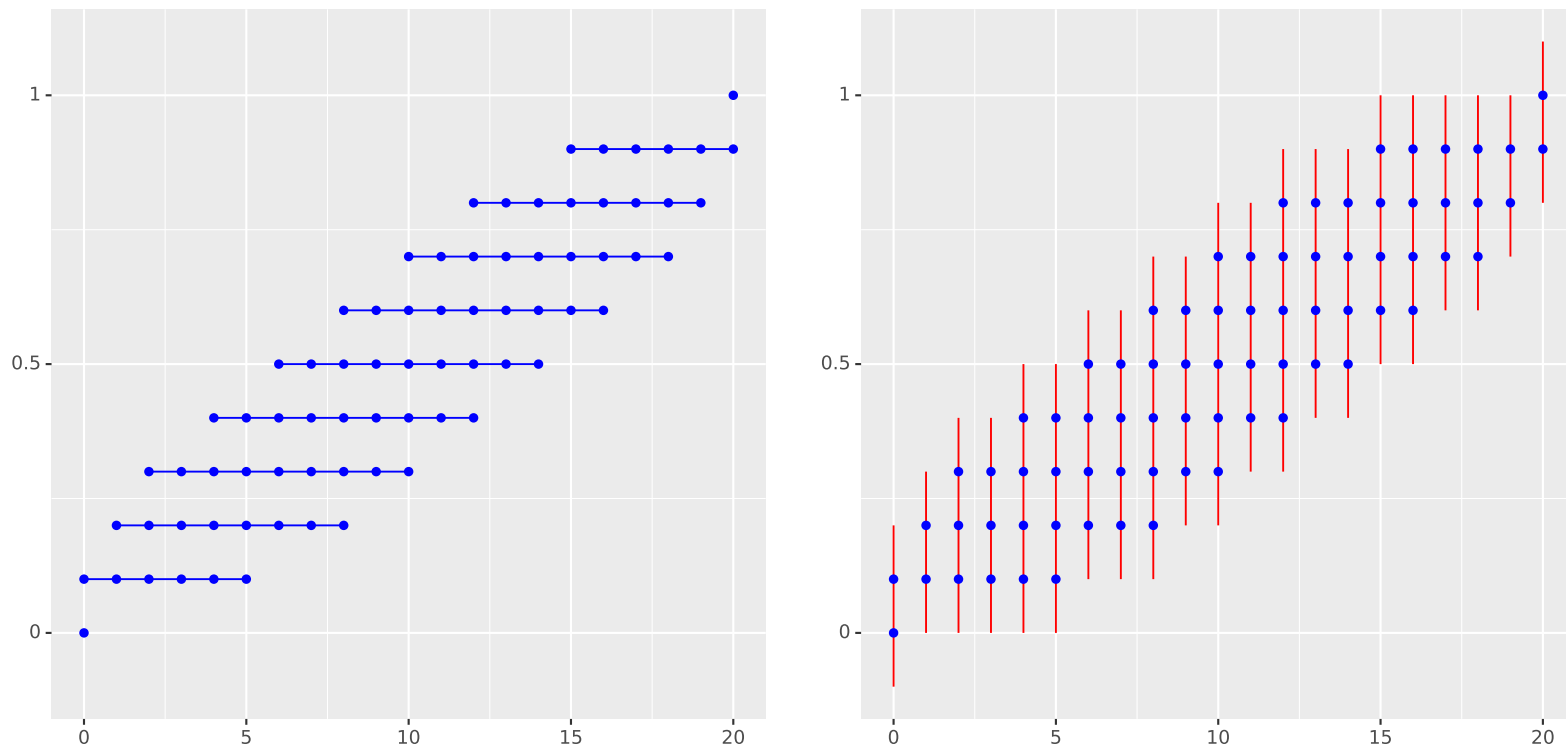
## II. Perform multiple hypothesis testing for all parameter values:

- ▷ Accept all parameters values for which pvalue is greather than  $\alpha$ .
- ▷ Output a minimal interval that covers all accepted parameter values.

## Rationale

- ▷ The true parameter value is rejected on  $\alpha$ -fraction of possible observations.
- ▷ Otherwise, the true value is inside the predicted interval.

# Illustration



- ▷ Acceptance ranges for different parameter values on the left.
- ▷ Extended parameter ranges covering all accepted parameters on the right.
- ▷ These ranges are the desired confidence intervals.

## Interpretation of confidence intervals

**Definition.** Confidence interval for a parameter  $p$  is an outcome of an approximation algorithm. The algorithm must output an interval  $[\hat{p} - \varepsilon, \hat{p} + \varepsilon]$  such that the true estimate is out of this range on  $\alpha$ -fraction of cases.

### Paradoxical inapplicability

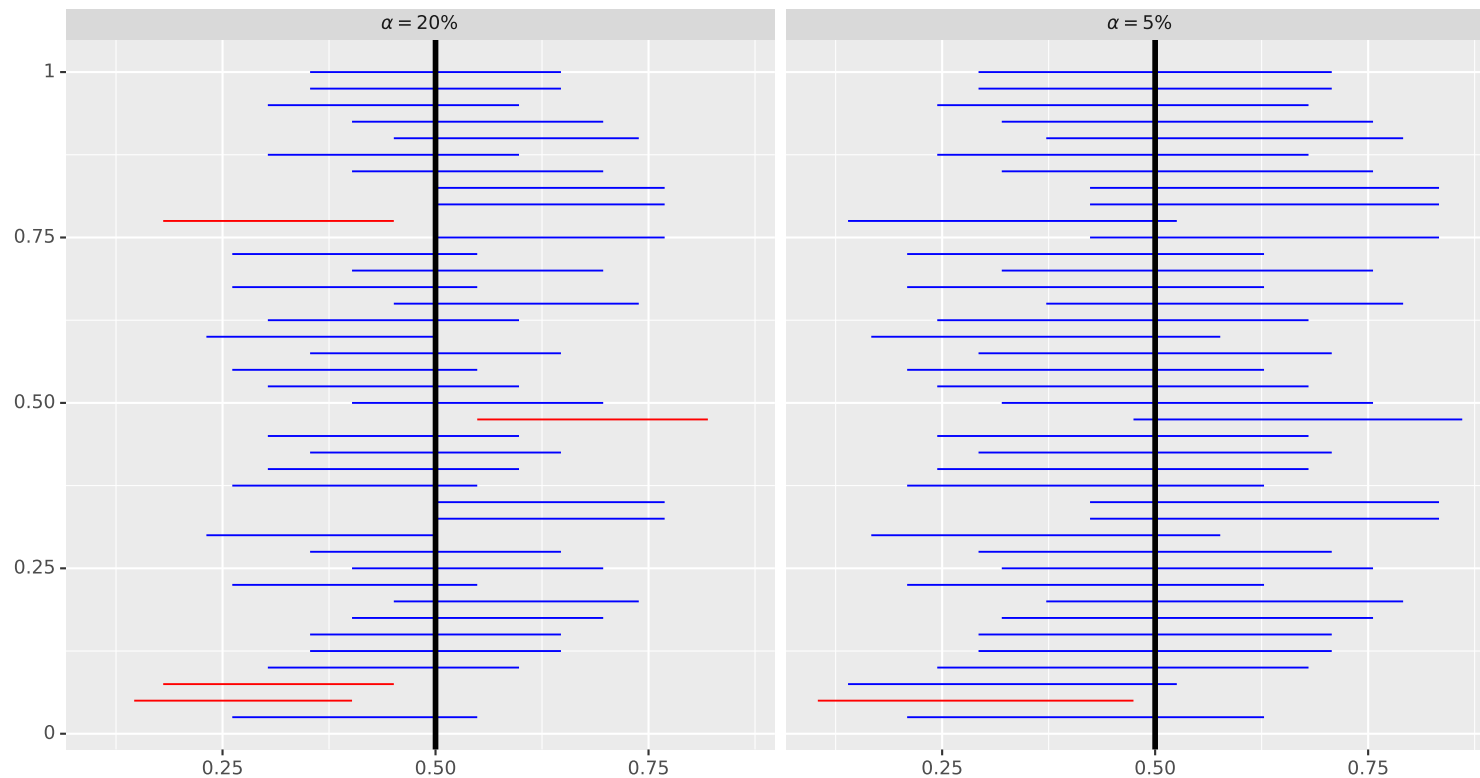
The definition does not state that the probability  $p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]$  is  $\alpha$ !

- ▷ The statement  $p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]$  is either true or false.
- ▷ There is no probability left. We just *do not know* the answer!

### Ultra-frequentistic resolution

- ▷ If  $\alpha$  is small enough say 5% then the algorithm is always correct.

## Illustrative example



By increasing the length of the interval we increase the fraction of runs for which the true value of  $p$  lies in the interval.

# Problems with confidence intervals

## Inability to capture background knowledge

- ▷ What if I know that  $p \in [0.1, 0.2]$  and observe  $B_1 = \dots = B_N = 1$ ?
- ▷ Then the estimate  $[\hat{p} - \varepsilon, \hat{p} + \varepsilon]$  is clearly wrong although on average this confidence interval is reasonable.

## Multiple hypothesis testing

- ▷ Using several confidence intervals in parallel increases the fraction of cases where some true estimate is out of the predicted range.
- ▷ We can use p-value adjustment methods are used to correct the issue.



## Prediction intervals

Even if we know the true relation  $y = f(x)$  we cannot predict the observation  $y_i = f(x_i) + \varepsilon_i$ , as the noise term  $\varepsilon_i$  is not known ahead.

- ▷ We cannot give upper and lower bounds for  $y_i$  which always hold.

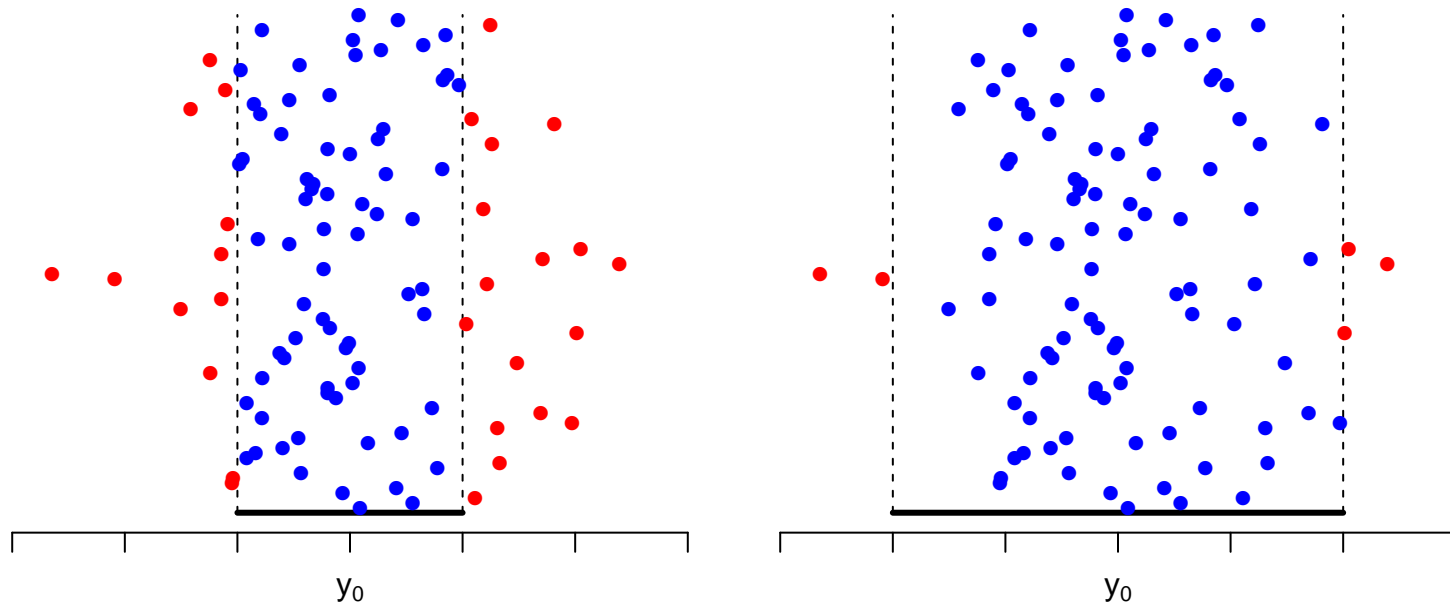
Instead, we can specify a prediction interval  $[y_* - \varepsilon, y_* + \varepsilon]$  so that with probability 95% the resulting measurement  $y_i$  is in the range.

- ▷ Usually, the analysis is similar to confidence interval derivation.

Interpretation of prediction intervals is different from confidence intervals.

- ▷ The probability estimate holds for the particular interval.

## Illustrative example



By increasing the length of the prediction interval we increase the fraction of future measurements which fall into interval.

# Confidence envelopes

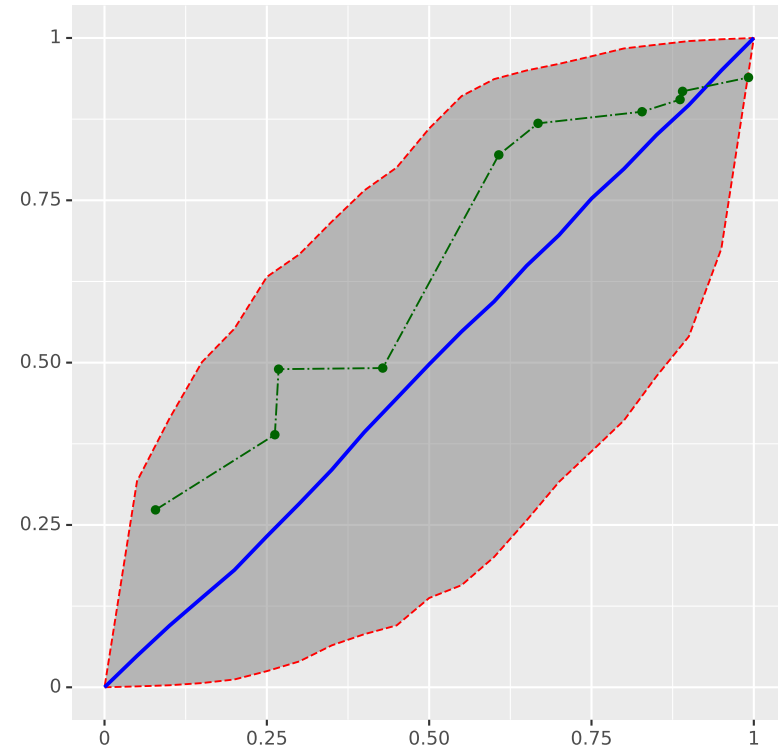
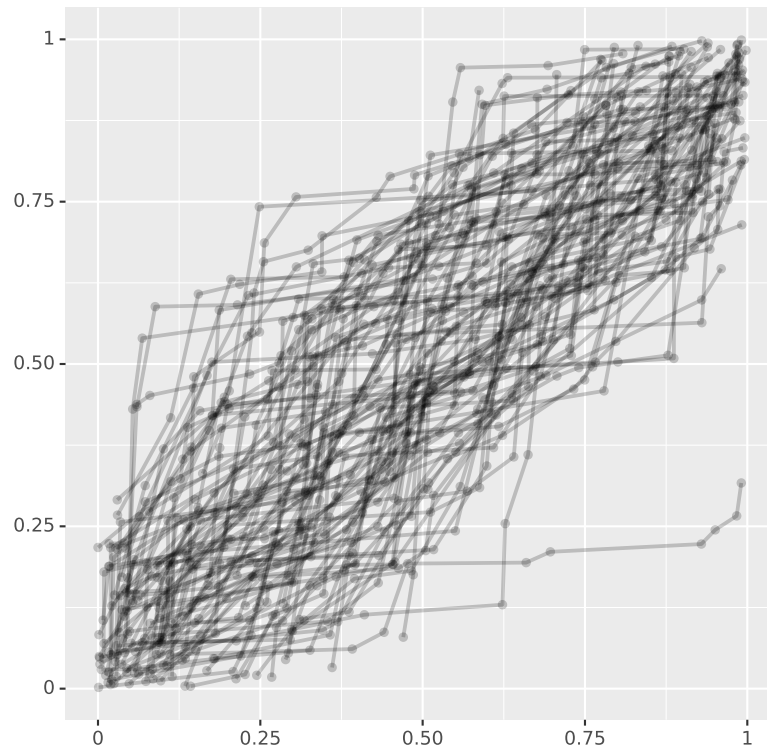
Confidence intervals is a good way to visualise uncertainty of a particular parameter. However, we are sometimes interested in the uncertainty many parameters or in the uncertainty of a function:

- ▷ How a predictor  $f : [0, 1] \rightarrow \mathbb{R}$  depends on the training set
- ▷ How a ROC curve  $\text{ROC} : [0, 1] \rightarrow [0, 1]$  depends on the test set
- ▷ How should a quantile-quantile plot be distributed.

Confidence bands are generalisations of confidence intervals

- ▷ Pointwise confidence band is a collection of confidence intervals
- ▷ Simultaneous confidence band must enclose  $\alpha$ -fraction of functions.
- ▷ Simultaneous confidence bands are much wider than pointwise bands.

## Illustrative example



- ▷ Distribution of qq-lines visualised through a sample on the left.
- ▷ A simulation based pointwise 95% confidence envelope on the right.
- ▷ The significance level that qq-line is inside the envelope is ca 50%.

# Permutation tests

## Baseline problem:

- ▷ Achievable accuracy depends on the data distribution.
- ▷ Artefacts in the dataset may bias performance measures.

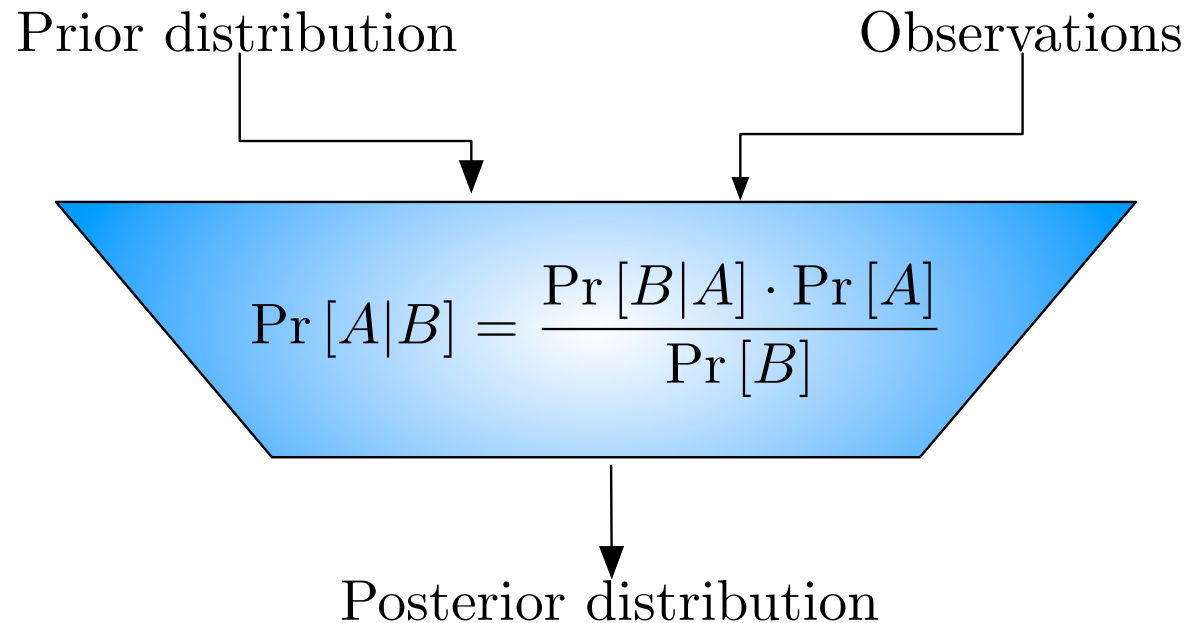
**Label permutation.** A random permutation  $\pi$  on outputs  $y_i$  destroys correlations between input-output pairs  $(x_i, y_{\pi(i)})$  but preserves marginal distribution of inputs and outputs.

**Permutation test.** Estimate how probable is to achieve equal or higher accuracy than was observed on the real data.

- ▷ If this probability is small then there must be signal in the data.
- ▷ The test completely neglect the effect size, i.e., how much results differ.
- ▷ Statistical significance does not imply utility!

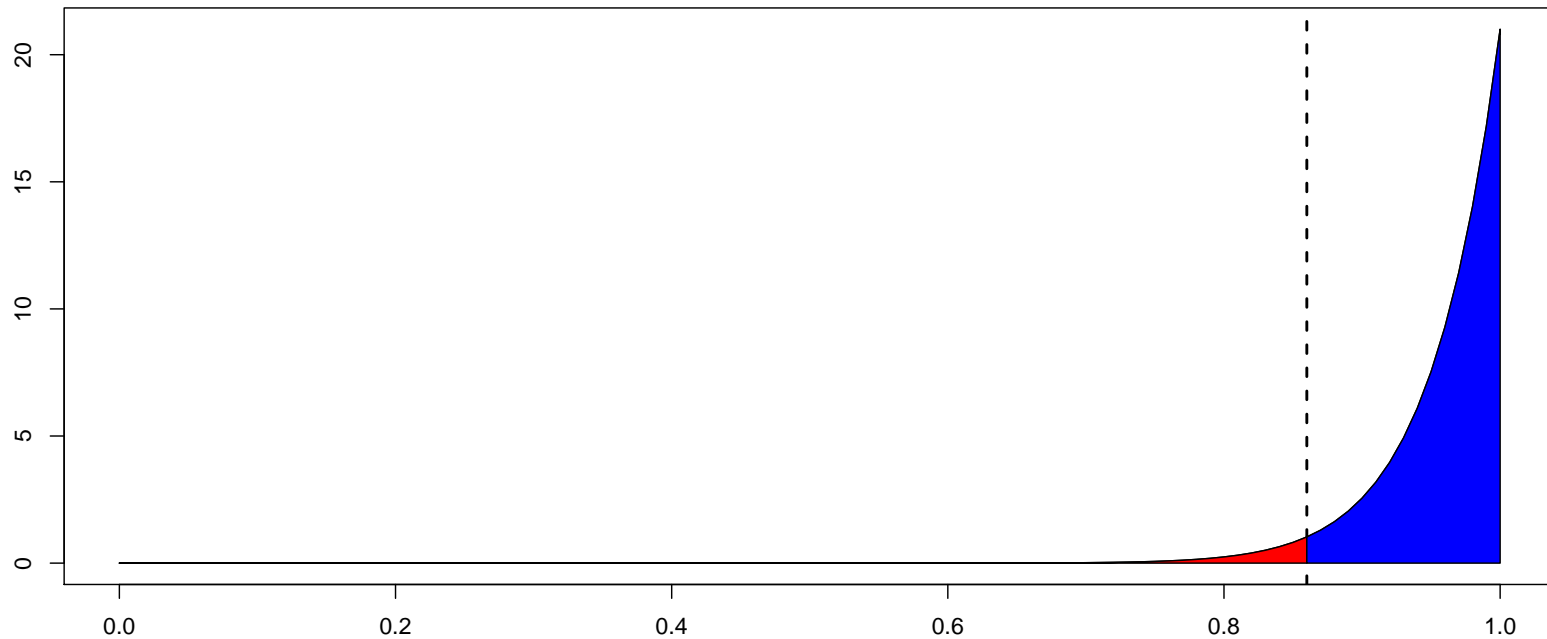
# Bayesian methods

# Bayesian inference procedure



- ▷ Prior distribution  $\Pr[A]$  encodes the background knowledge
- ▷ The model  $\Pr[B|A]$  determines how the posterior  $\Pr[A|B]$  is updated

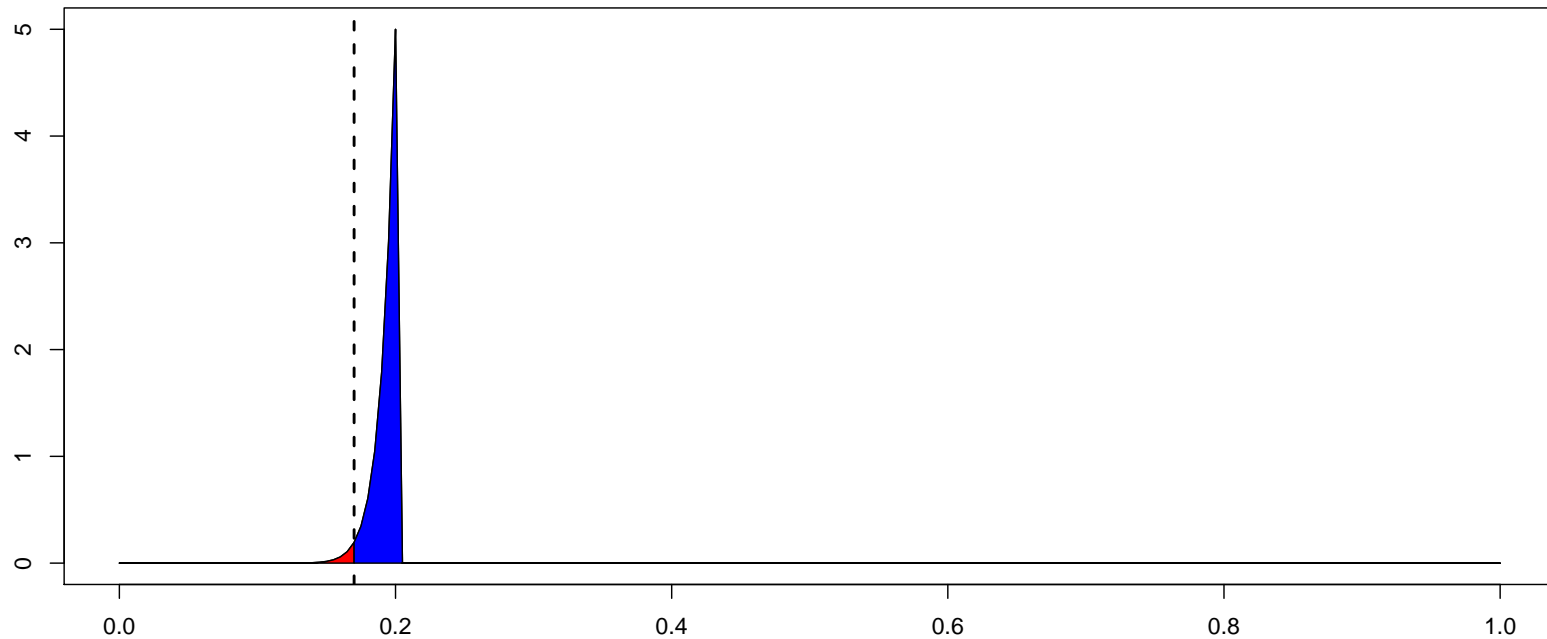
## Posterior of an uninformed person



With no preferences on the value of  $p$  the posterior is strongly skewed towards one and the range  $p \in [0.86, 1]$  contains 95% of posterior probability.

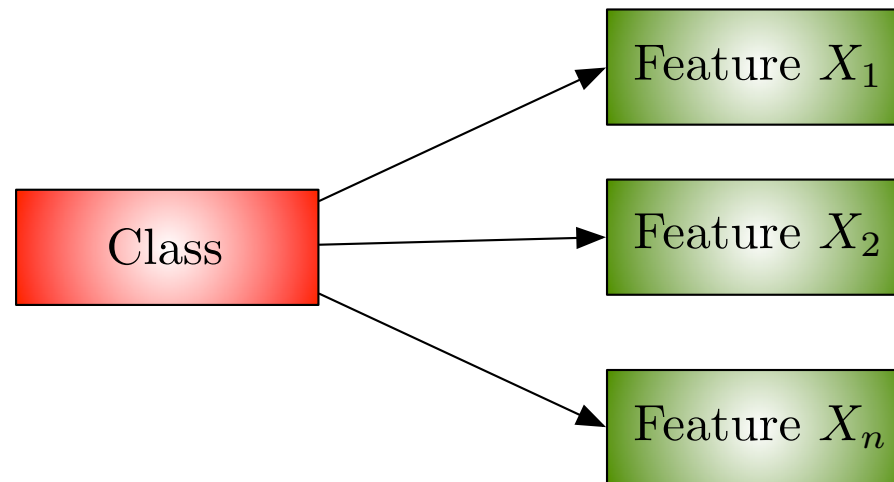


## Posterior of an informed person



With the knowledge  $p \in [0.1, 0.2]$  the posterior is strongly skewed towards 0.2 and the range  $p \in [0.17, 0.2]$  contains 95% of posterior probability.

## Model behind naive Bayes classifier



Underlying class value determines observed attributes

- ▷ Each attribute  $X_i$  is binary
- ▷ All variables are independent if class is fixed
- ▷ Sometimes we just ignore dependancies for easier modelling

## Likelihood of the data

Let us assume that we know the probabilities

$$p_i = \Pr [X_i = 1 | Class = 0]$$

$$q_i = \Pr [X_i = 1 | Class = 1]$$

Then using the independence assumption we get

$$\Pr [X_1 = a_1, \dots, X_n = a_n | Class = 0] = \prod_{i=1}^n p_i^{a_i} (1 - p_i)^{1-a_i}$$

$$\Pr [X_1 = a_1, \dots, X_n = a_n | Class = 1] = \prod_{i=1}^n q_i^{a_i} (1 - q_i)^{1-a_i}$$

## Prior and posterior for the class labels

Now it is straightforward to derive

$$\Pr [Class = 0 | \mathbf{X} = \mathbf{a}] = \frac{\prod_{i=1}^n p_i^{a_i} (1 - p_i)^{1-a_i} \cdot \Pr [Class = 0]}{\Pr [\mathbf{X} = \mathbf{a}]}$$

$$\Pr [Class = 1 | \mathbf{X} = \mathbf{a}] = \frac{\prod_{i=1}^n q_i^{a_i} (1 - q_i)^{1-a_i} \cdot \Pr [Class = 1]}{\Pr [\mathbf{X} = \mathbf{a}]}$$

which gives an *odd ratio*

$$\frac{\Pr [Class = 0 | \mathbf{X} = \mathbf{a}]}{\Pr [Class = 1 | \mathbf{X} = \mathbf{a}]} = \frac{\Pr [Class = 0]}{\Pr [Class = 1]} \cdot \frac{\prod_{i=1}^n p_i^{a_i} (1 - p_i)^{1-a_i}}{\prod_{i=1}^n q_i^{a_i} (1 - q_i)^{1-a_i}}$$

## The resulting classifier is a linear classifier

By taking logarithm form the odd ratio we get

$$\log \left( \frac{\Pr [Class = 0 | \mathbf{X} = \mathbf{a}]}{\Pr [Class = 1 | \mathbf{X} = \mathbf{a}]} \right) = w_0 + \sum_{i=1}^n w_i a_i$$

where

$$w_0 = \log \left( \frac{\Pr [Class = 0]}{\Pr [Class = 1]} \right) + \sum_{i=1}^n \log \left( \frac{1 - p_i}{1 - q_i} \right)$$

$$w_i = \log \left( \frac{p_i}{1 - p_i} \cdot \frac{1 - q_i}{q_i} \right)$$

## How to train the classifier?

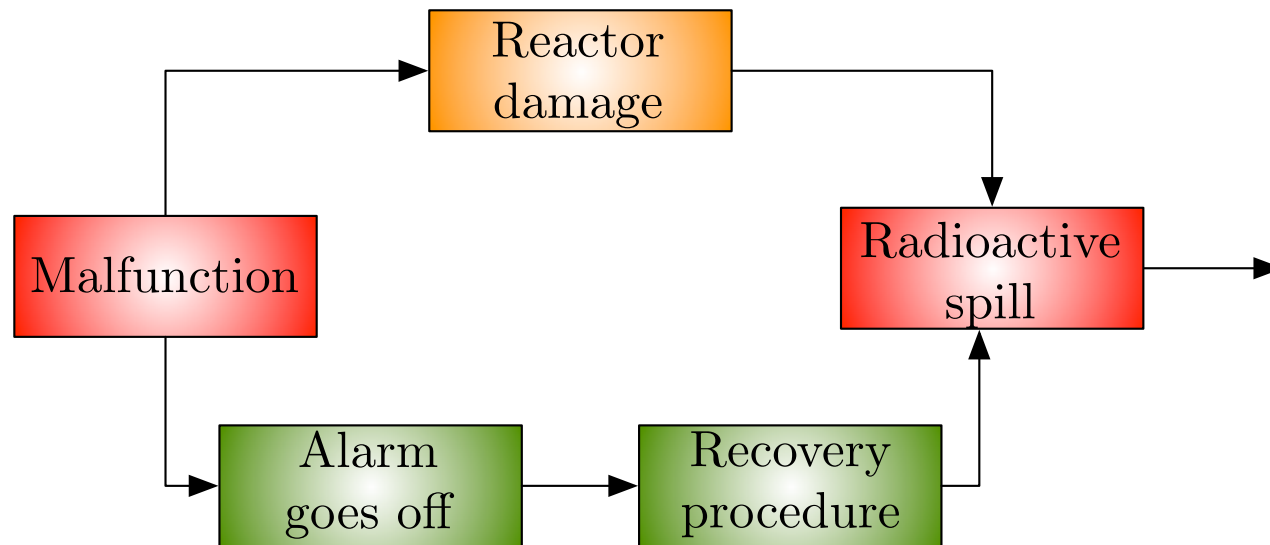
A frequentistic approach is to fix probabilities from the training sample

$$p_i = \frac{\# \{ \text{data points form class 0 with } X_i = 1 \}}{\# \{ \text{data points form class 0} \}}$$
$$q_i = \frac{\# \{ \text{data points form class 1 with } X_i = 1 \}}{\# \{ \text{data points form class 1} \}}$$

However if some value does not occur for  $X_i$  in the training sample we get overly confident results. Thus, Bayesian mean estimate is better alternative

$$p_i = \frac{\# \{ \text{data points form class 0 with } X_i = 1 \} + 1}{\# \{ \text{data points form class 0} \} + 2}$$
$$q_i = \frac{\# \{ \text{data points form class 1 with } X_i = 1 \} + 1}{\# \{ \text{data points form class 1} \} + 2}$$

## Going beyond naive Bayesian models



Complex causal models are often defined through Bayesian networks

- ▷ A complex processes is first split into sub-events
- ▷ Direct causal dependencies between sub-events are detected
- ▷ Causation mechanisms are characterised with probability tables

# Strength and weaknesses of Bayesian networks

## Strengths

- ▷ Bayesian networks are easy to interpret
- ▷ Bayesian networks are good for formalising fuzzy background knowledge
- ▷ Estimation of individual probability tables is tractable
- ▷ There are tools for doing inference with Bayesian networks

## Weaknesses

- ▷ You must know the causal structure of sub-events
- ▷ Identification of causal structure from data alone is very difficult
- ▷ It is notoriously difficult to model non-trivial causal dependencies
- ▷ Standard inference procedures often do not have closed solutions