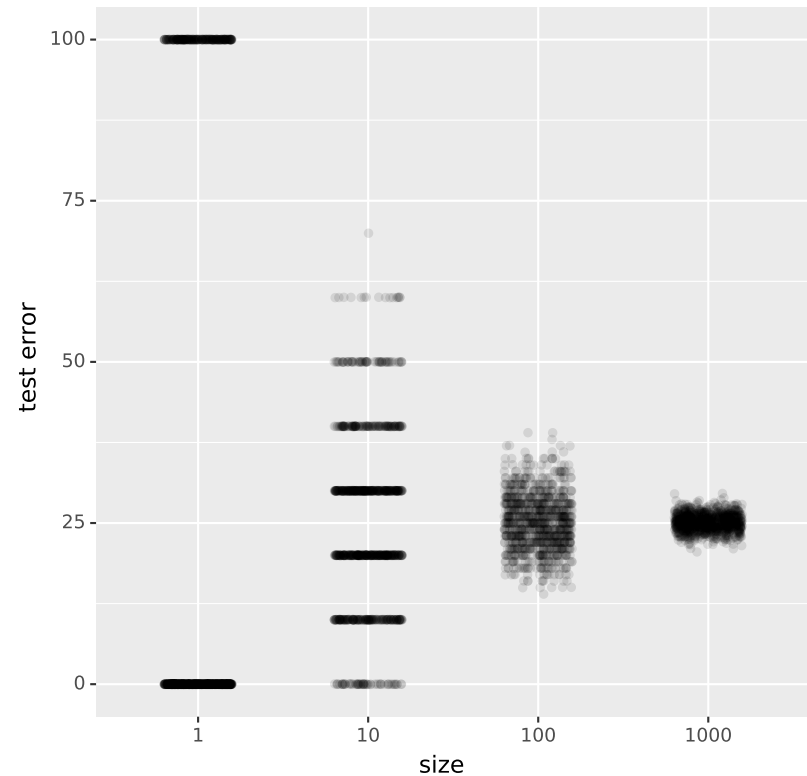
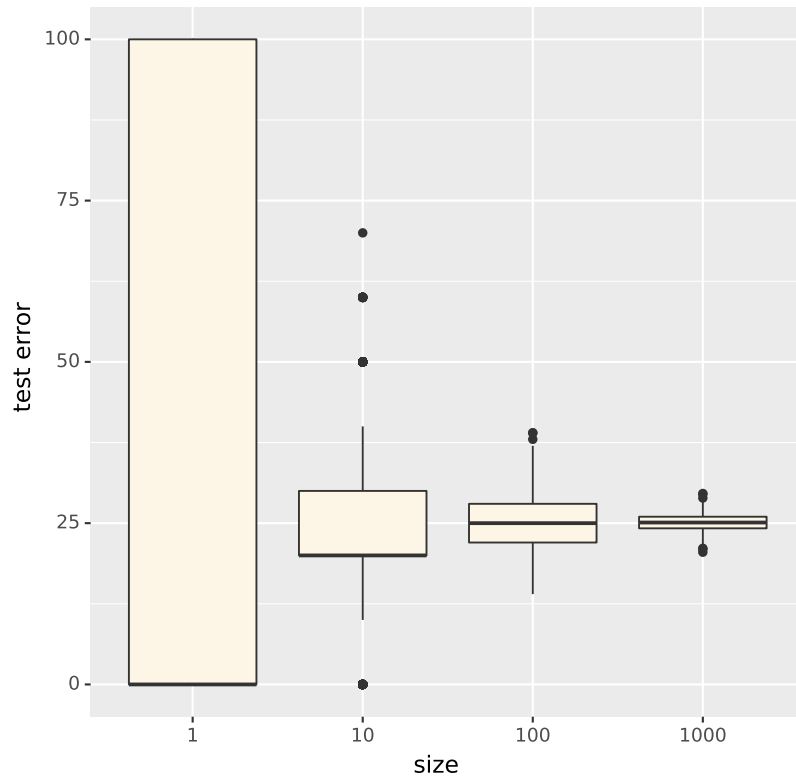


LTAT.02.004 MACHINE LEARNING II

## **Basics of probabilistic modelling**

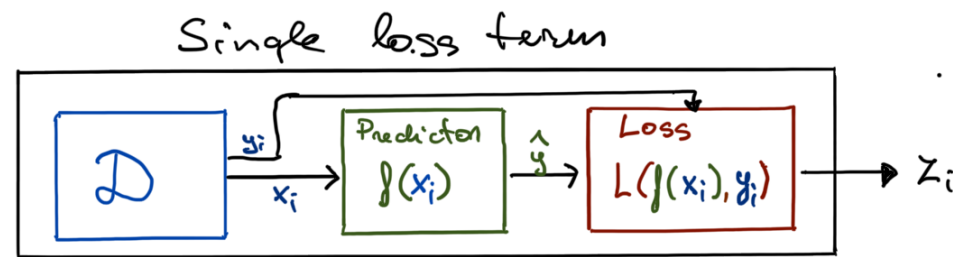
Sven Laur  
University of Tartu

# Why does empirical risk converge at all?



- ▷ Depends on the test set.
- ▷ Statistical fluctuations decrease with size.

## Empirical risk as mean of random variables



Recall that empirical risk is computed through the following formula

$$R_N(f) = \frac{1}{N} \cdot \sum_{i=1}^N L(f(\mathbf{x}_i), y_i) = \frac{1}{N} \cdot \sum_{i=1}^N z_i$$

where all samples  $(\mathbf{x}_i, y_i)$  are assumed to be

- ▷ independent from each other,
- ▷ coming from the same distribution.

## Law of large numbers

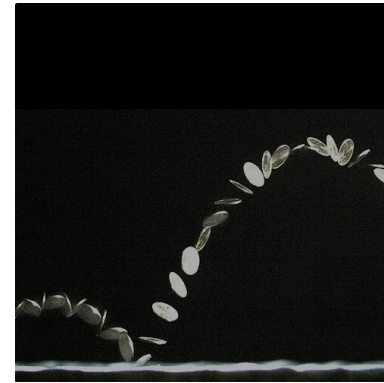
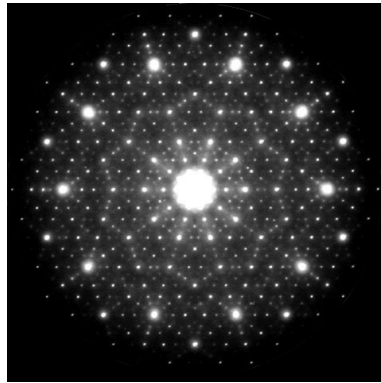
**Central limit theorem.** Let  $z_1, \dots, z_N$  be independent and identically distributed samples from a *real-valued distribution* with a *finite standard deviation*  $\sigma$  and *mean*  $\mu$ . Then the random variable

$$S = \sqrt{N} \left( \frac{1}{N} \cdot \sum_{i=1}^N z_i - \mu \right)$$

converges *in distribution* to normal distribution  $\mathcal{N}(\text{mean} = 0, \text{sd} = \sigma)$ .

- ▷ Under mild assumptions the empirical risk  $R_N(f)$  converges to risk  $R(f)$ .
- ▷ The result is not precise enough to quantify approximation error.

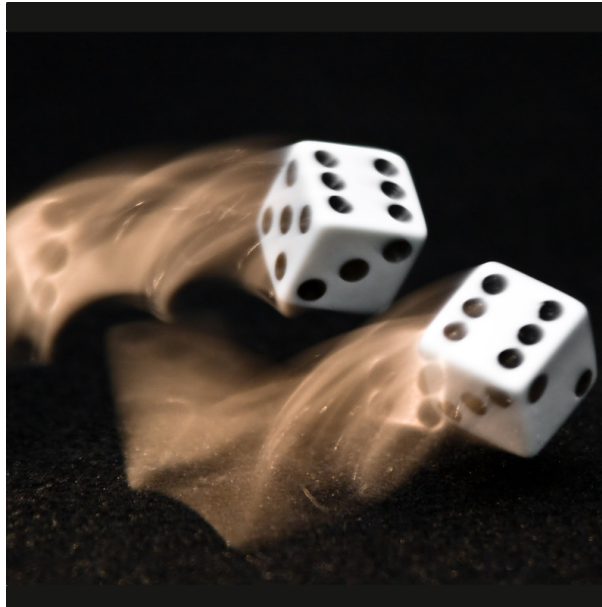
# What is probability?



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

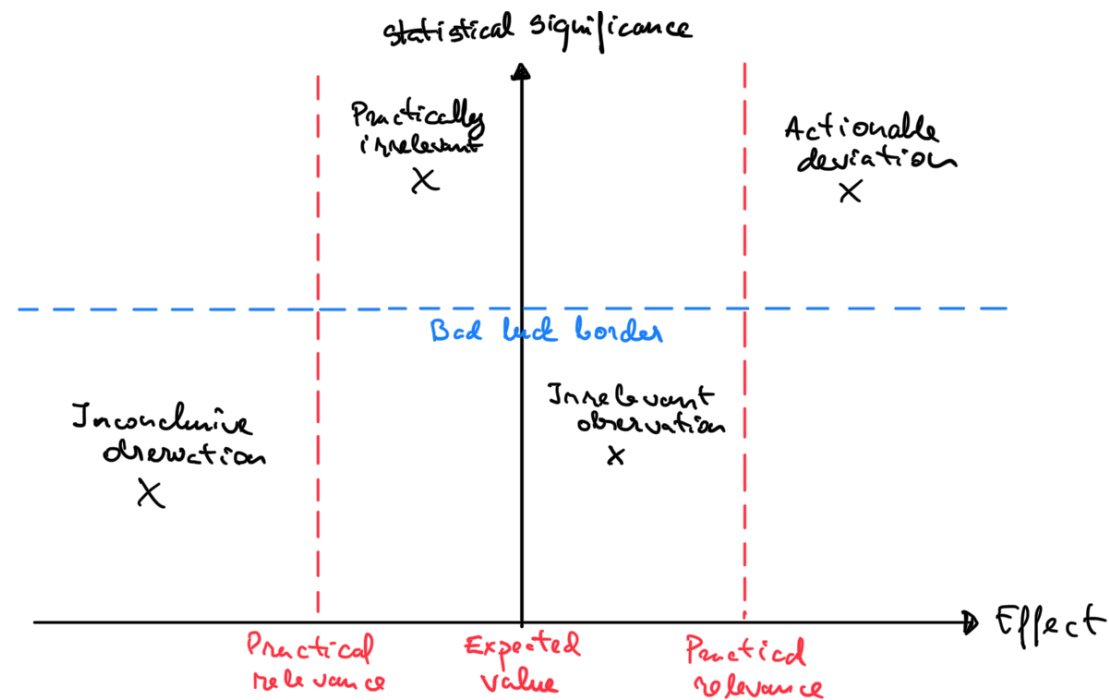
- ▷ The aim of statistics is to design algorithms that work well on average.
- ▷ For that one needs to specify probabilistic model for data sources.
- ▷ Confidence is the fraction of cases the algorithm works as specified.

## 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

# Central question in statistical testing



The question is my observation relevant has two aspects

- ▷ Can we explain the difference by sheer luck?
- ▷ Is the difference between expected and observed big enough?

## Causation between zero-one events

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

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

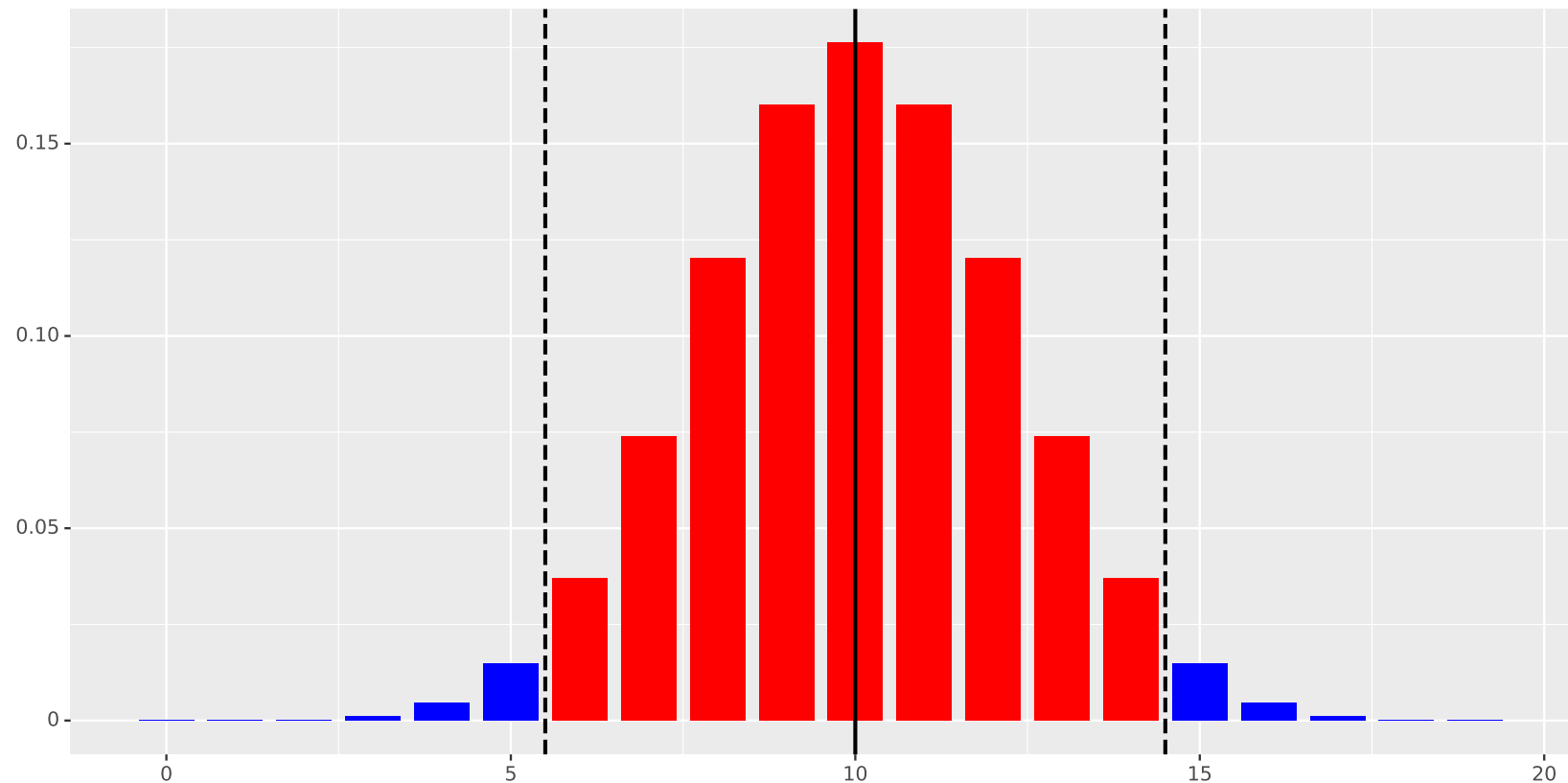
Then the probability is to get  $k$  ones in  $n$  independent trials is

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

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

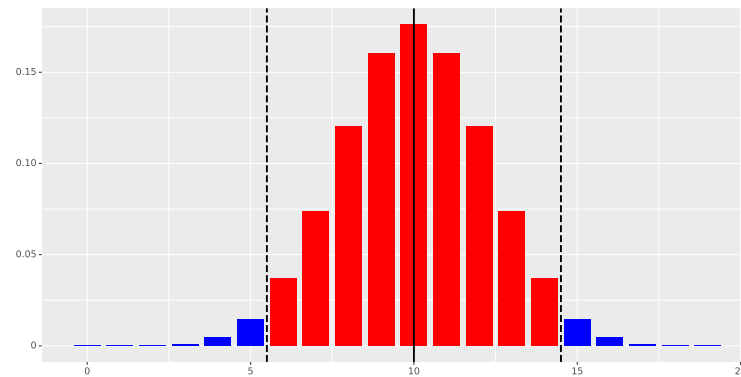
$$B_1 + \dots + B_n \sim \text{Bin}(n, p)$$

## 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.

## Does a classifier beat a random guess?



Consider three algorithms on twenty element test set:

- ◇ Algorithm A gets 9 correct answers;
  - ◇ Algorithm B gets 13 correct answers;
  - ◇ Algorithm C gets 17 correct answers.
- 
- ▷ Which of them are better than random classifiers?
  - ▷ Which of them are classifiers good enough for practical applications?

# 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 *fixed* classification *rule* 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:

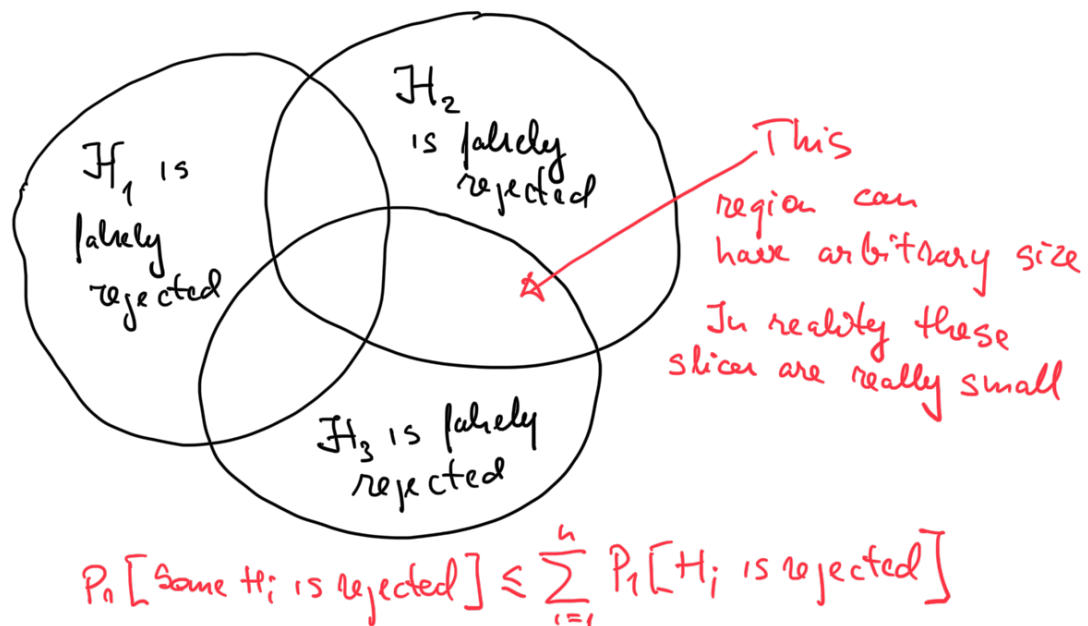
- ▷ Bonferroni correction is almost optimal
- ▷ FDR correction controls the expected number false positives



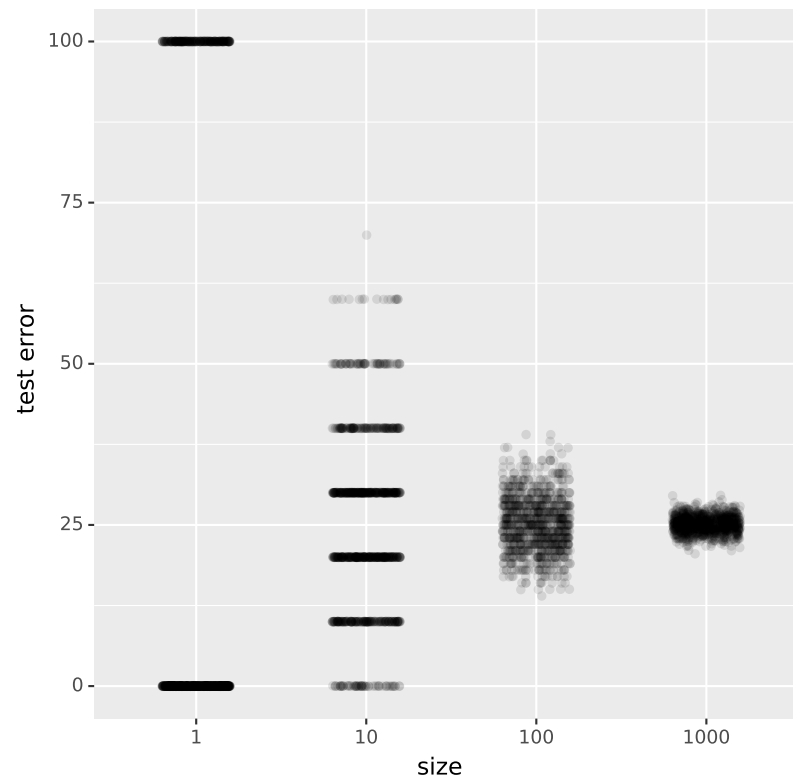
## Bonferroni correction for tests

Assume that data is generated so that null hypotheses  $\mathcal{H}_1, \dots, \mathcal{H}_n$  hold.

- ▷ Then we can still reject some the tests due to bad luck.
- ▷ We can use really naive enough bound visualised below.



## How far is the true risk?



- ▷ How wide error bars cover true risk for *all* observations?
- ▷ How wide error bars cover true risk for *most* observations?

# 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.

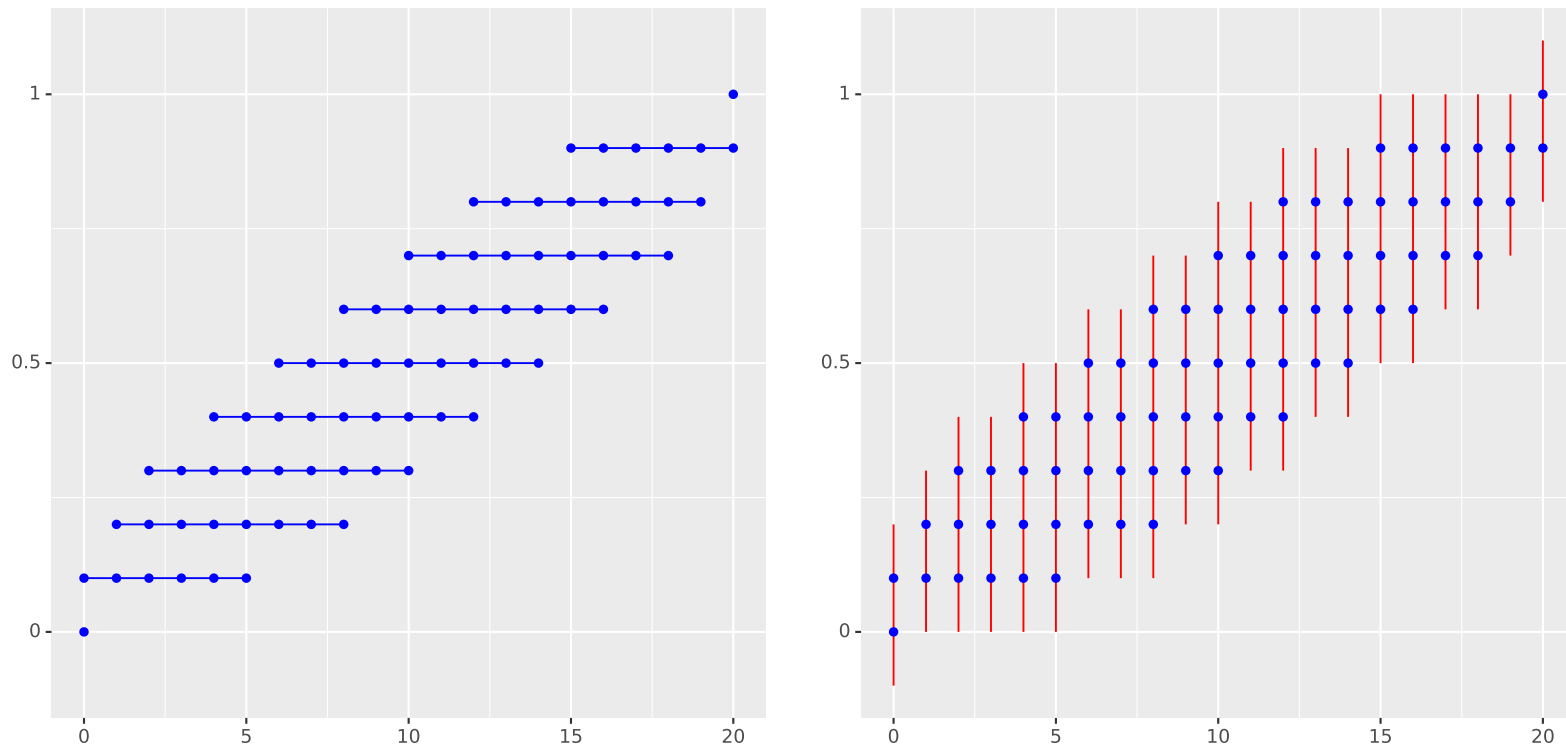
## II. Perform multiple hypothesis testing for all parameter values:

- ▷ Accept all parameters values for which p-value is greater than  $1 - \alpha$ .
- ▷ Output a minimal interval that covers all accepted parameter values.

## Rationale

- ▷ The true parameter value is rejected on  $\alpha$ -fraction of possible observations.
- ▷ For the remaining cases 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 in the range on  $\alpha$ -fraction of cases.

### Paradoxical inapplicability

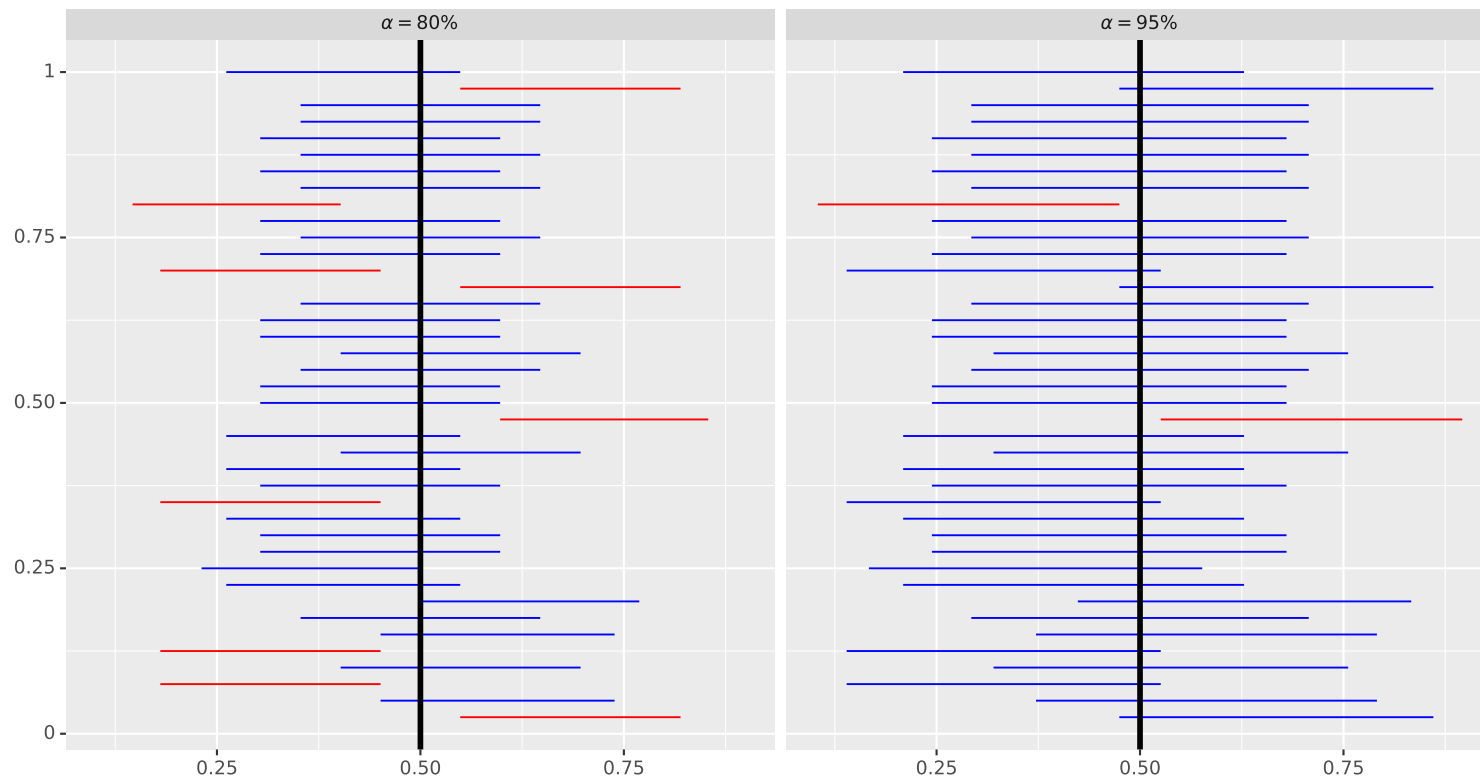
The definition does not state that the probability  $p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]$  is  $\alpha$ !

- ▷ 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  $1 - \alpha$  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.

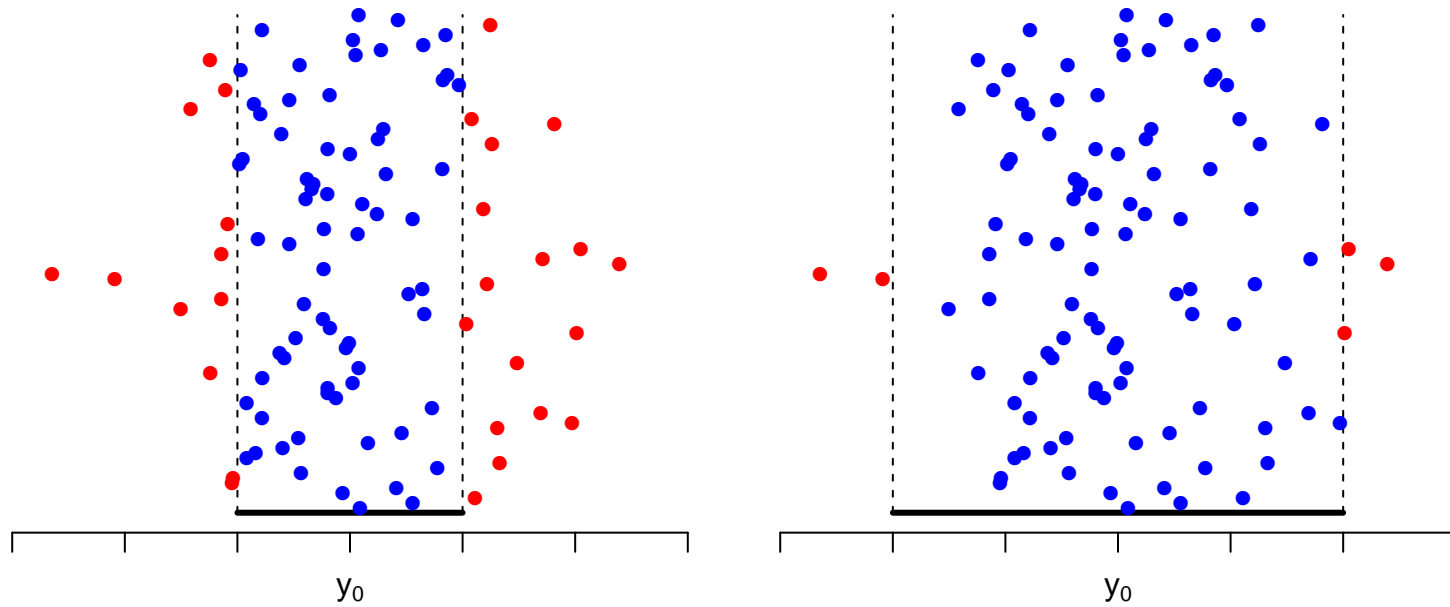
- ▷ Usually, the analysis is similar to confidence interval derivation.

Interpretation of prediction intervals is different from confidence intervals.

- ▷ The probability estimate holds for the particular interval.

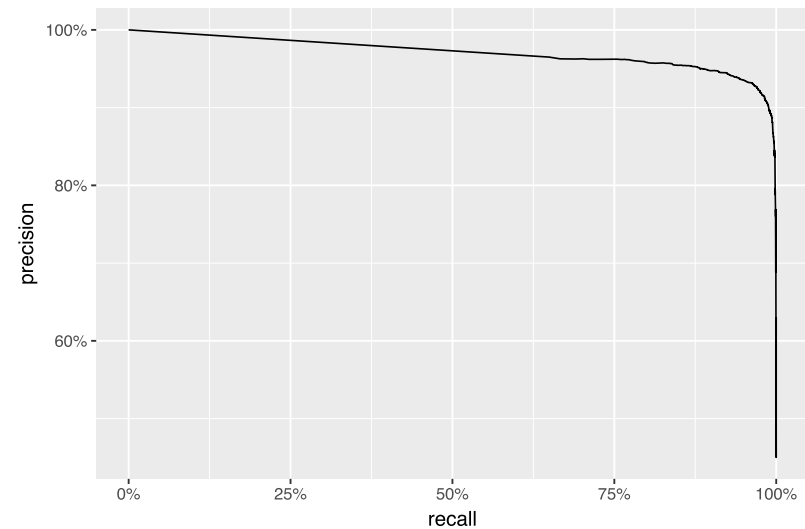
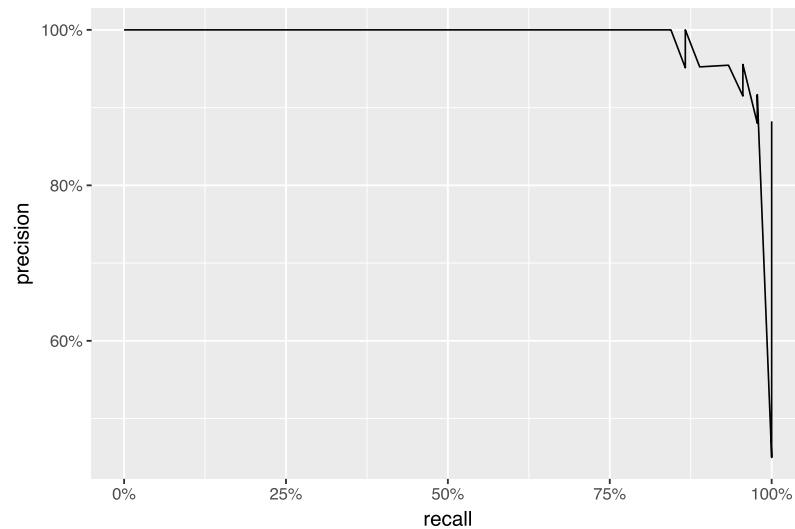


## Illustrative example



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

# Fluctuations in performance profiles



- ▷ Precision-recall graph is not smooth if the test set is small.
- ▷ How much the true graph can be different from observed?
- ▷ How many of samples are needed to get a decent resolution?

# Confidence envelopes

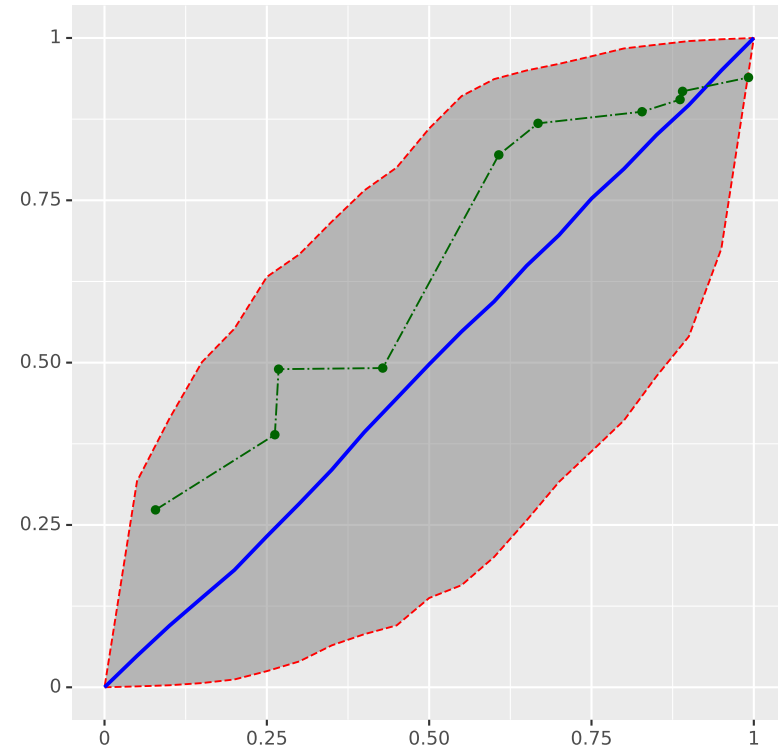
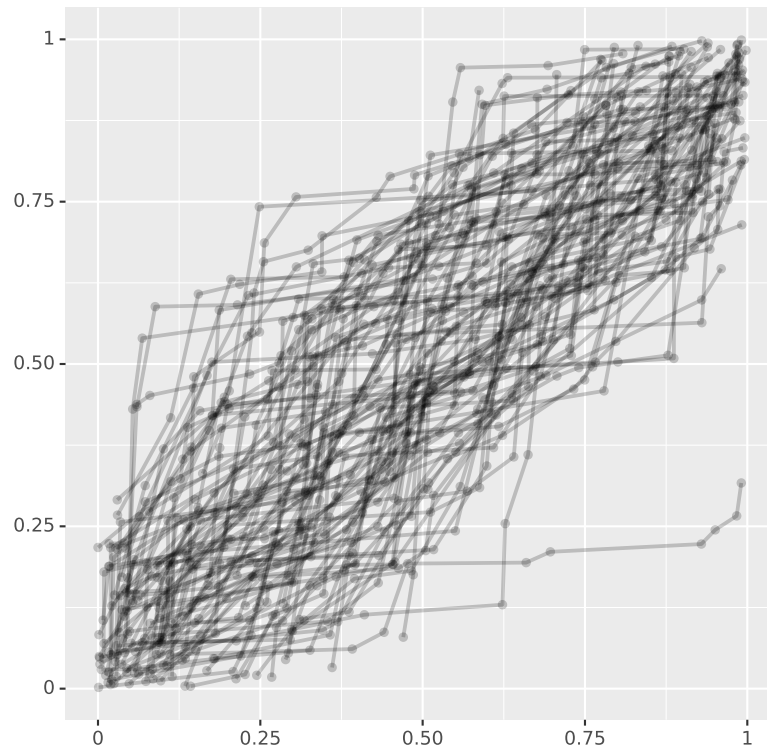
Confidence intervals is a good way to visualise uncertainty of a particular parameter. However, we are sometimes interested in the uncertainty many parameters or in the uncertainty of a function:

- ▷ 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  $(\mathbf{x}_i, \mathbf{y}_{\pi(i)})$  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!