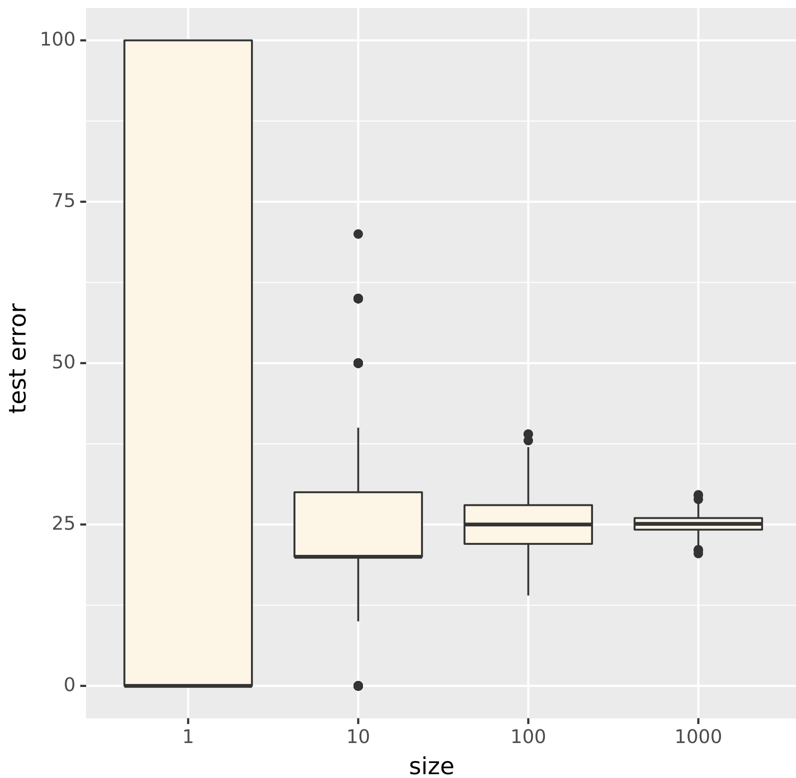
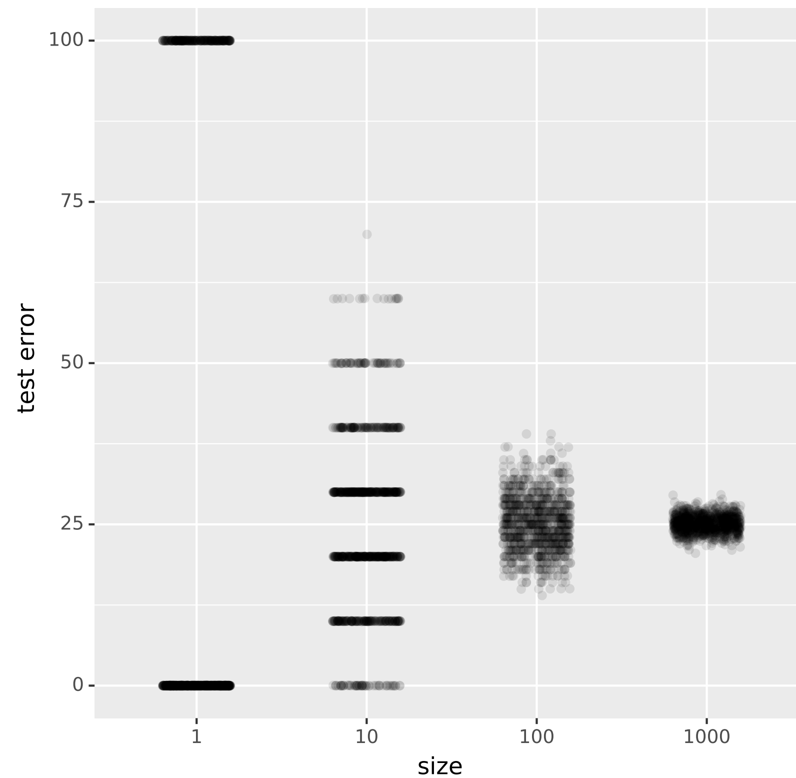
LTAT.02.004 Machine Learning II

Basics of probabilistic modelling

Sven Laur
University of Tartu

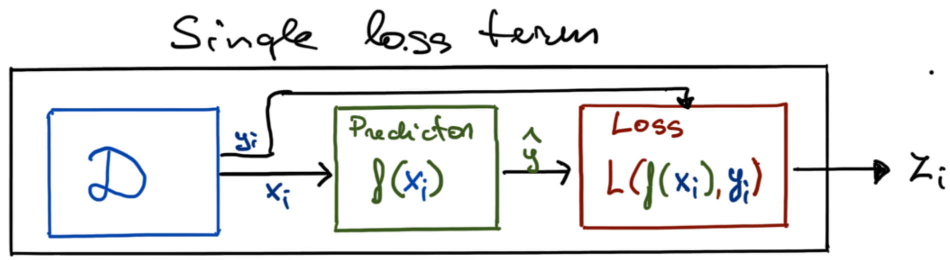
Why does empirical risk converge at all?





- ▷ 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

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

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

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

Law of large numbers

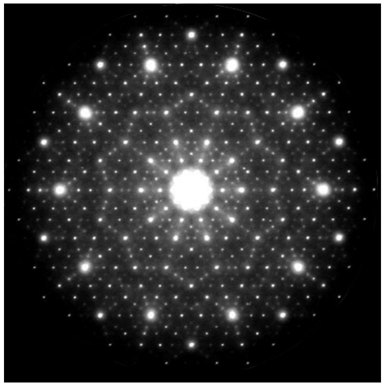
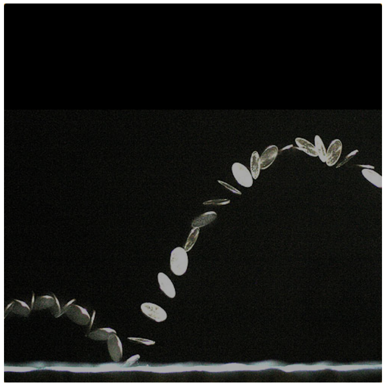
Central limit theorem. Let $z_1, \ldots, z_N$ be independent and identically distributed samples form a *real-valued distribution* with a *finite standard deviation* σ and *mean* μ. Then the random variable

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

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

- ▷ 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

Illustrative example

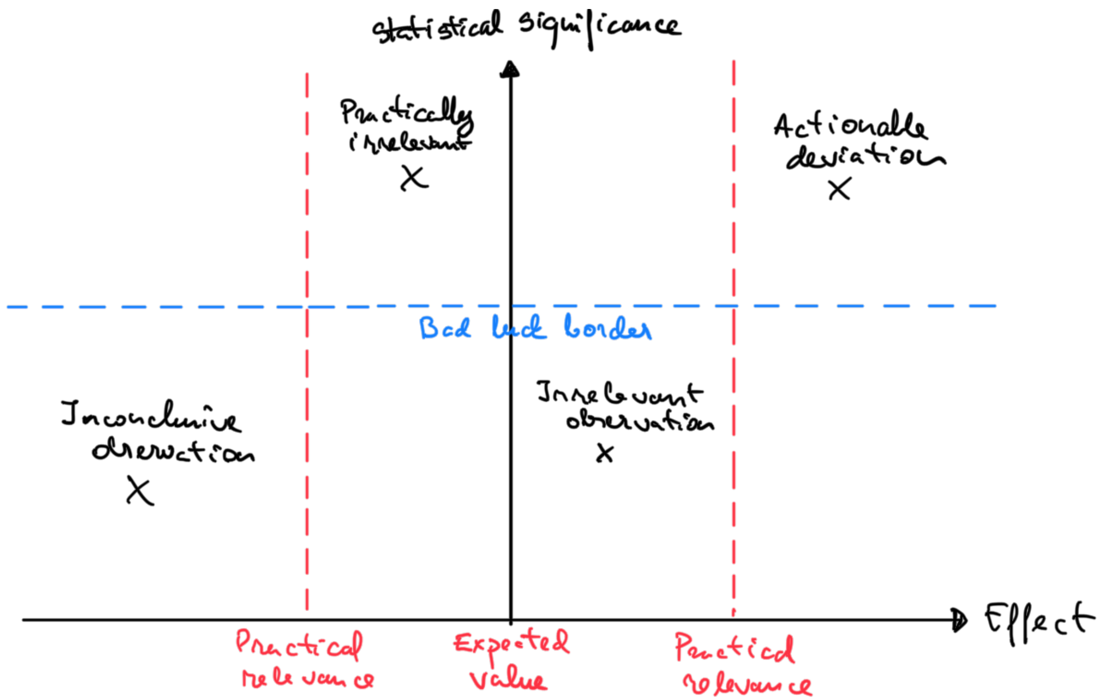
Consider an experiment that yields 2 heads and 8 tails.

- ▷ Frequency of heads is 20%.
- ▷ Can the coin be still fair?

Consider an experiment that yields 1,000,100 heads and 999,900 tails.

- ▷ Frequency of heads is 50.005%.
- ▷ Can the coin be still biased?

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.,

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

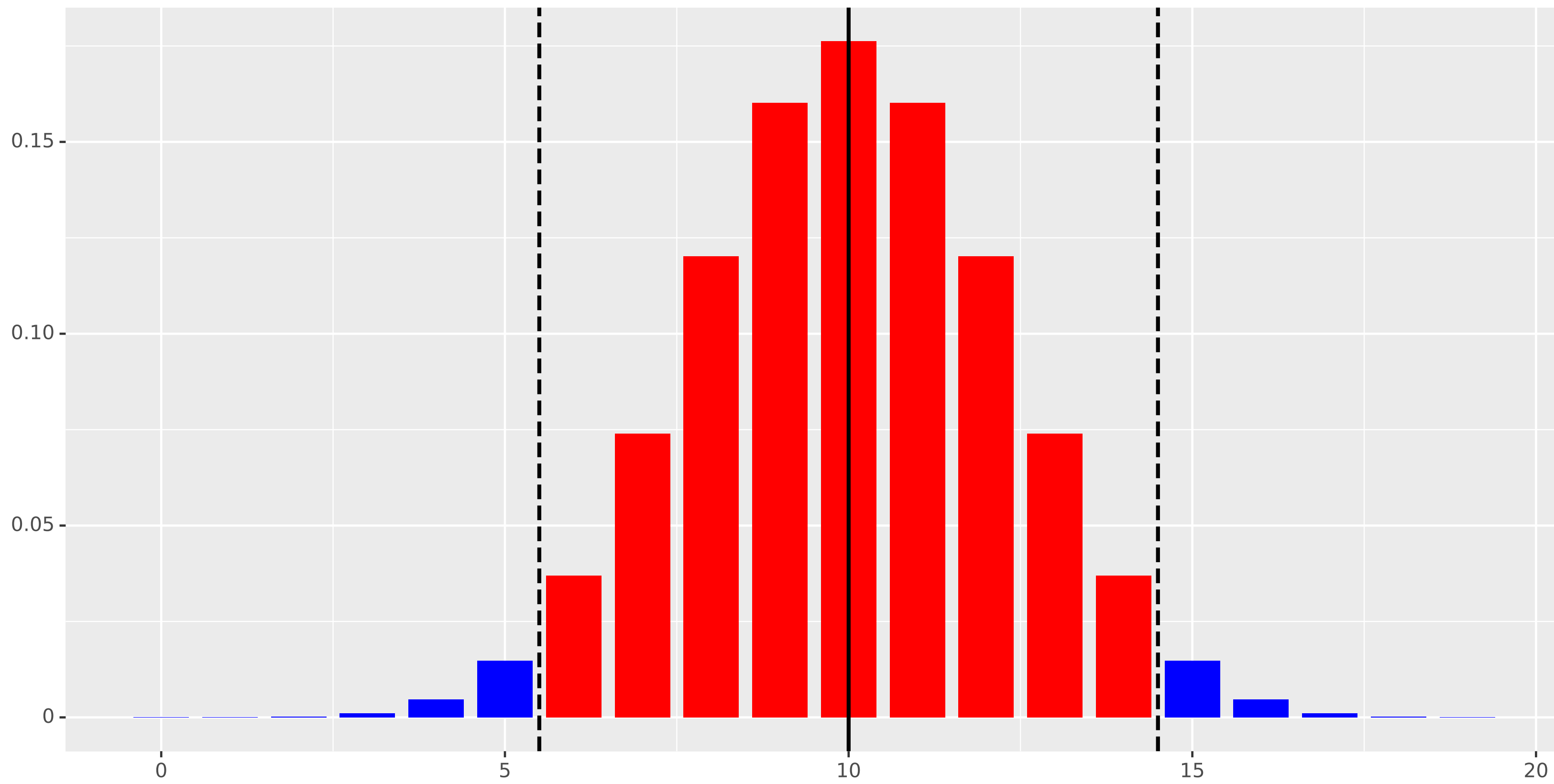
Then the probability is to get k ones in n independent trials is

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

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

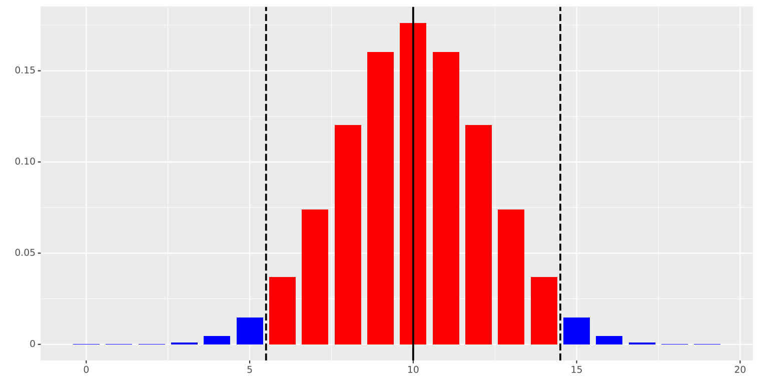
$$B_1 + \cdots + B_n \sim \mathsf{Bin}(n, p)$$

Illustration



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

III. Consequences on the observations:

▷ The observed sum $B_1 + \ldots + B_n \sim \mathrm{Bin}(n = 20, p = 0.5)$.

▷ Limit on the tail probability $\Pr[|B_1 + \ldots + B_n - 10| \geq 6] \leq 5\%$

IV. Test procedure

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

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

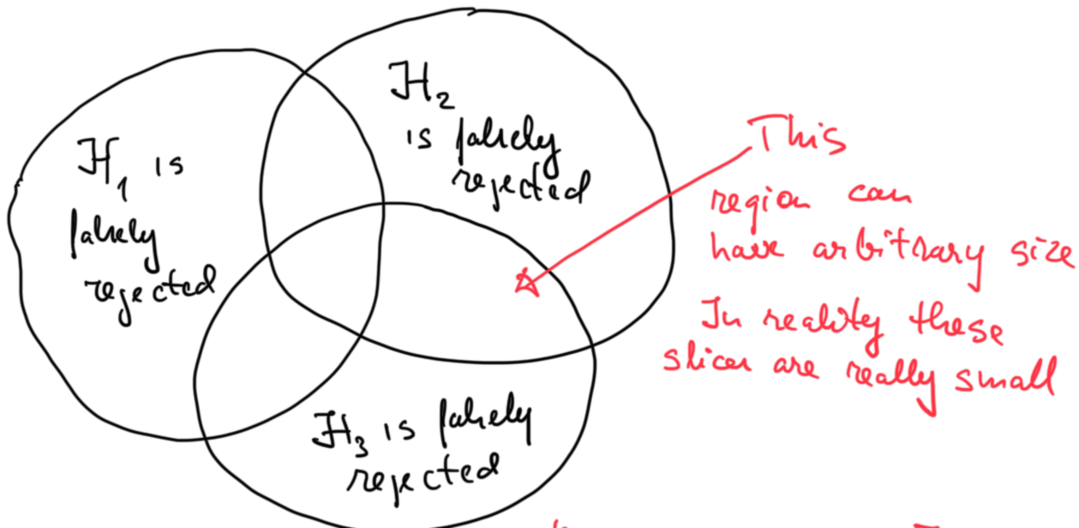
Testing several hypothesis in parallel increases the number of false positives. Several p-value adjustment methods are used to correct the issue:

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

Bonferroni correction for tests

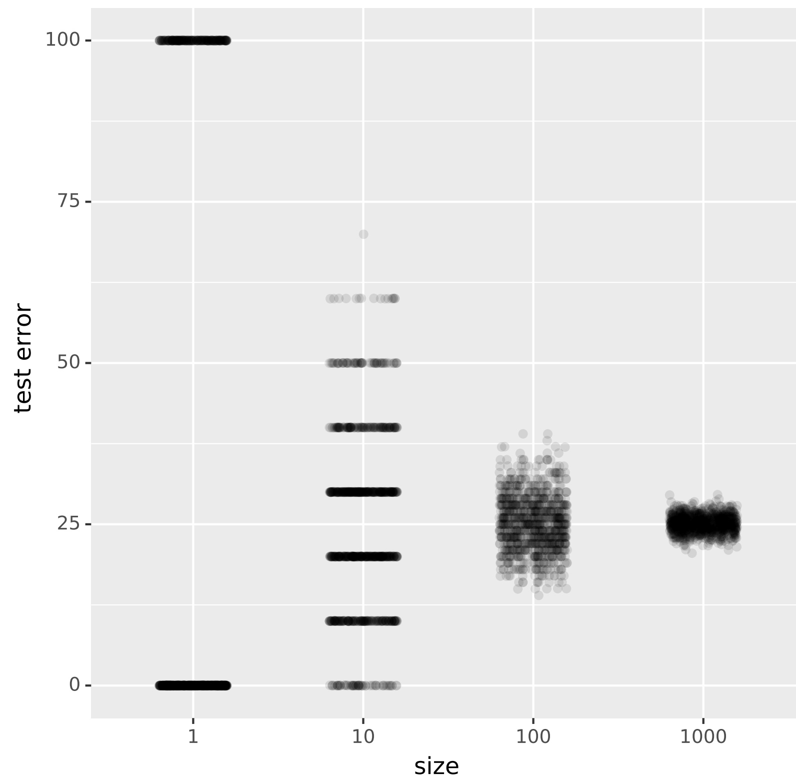
Assume that data is generated so that null hypotheses $\mathcal{H}_1, \ldots, \mathcal{H}_n$ hold.

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



$$P_1[\text{some } H_i \text{ is rejected}] \le \sum_{i=1}^{n} P_1[H_i \text{ is rejected}]$$

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.

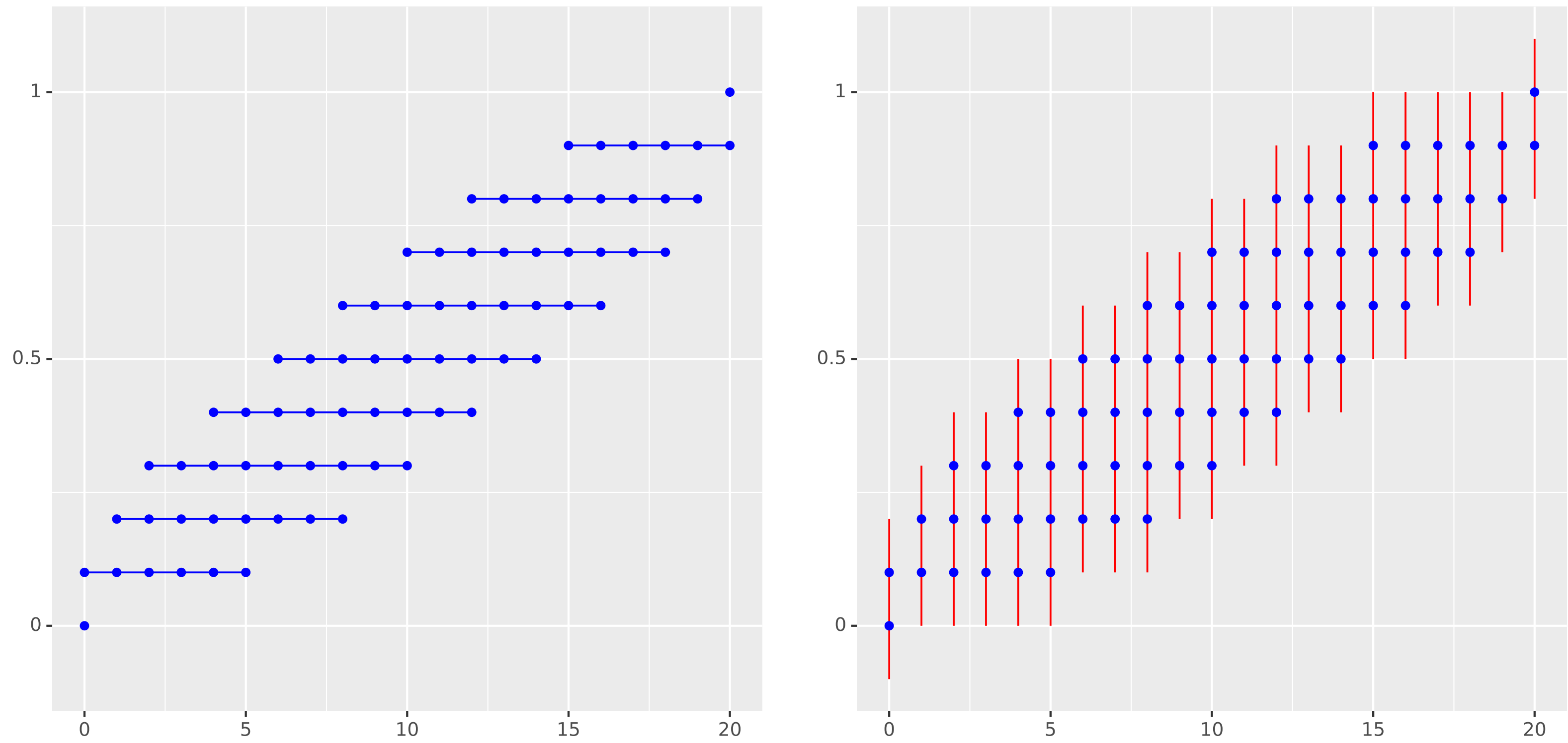
II. Perform multiple hypotesis testing for all parameter values:

- ▷ 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 α-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 α-fraction of cases.

Paradoxical inapplicability

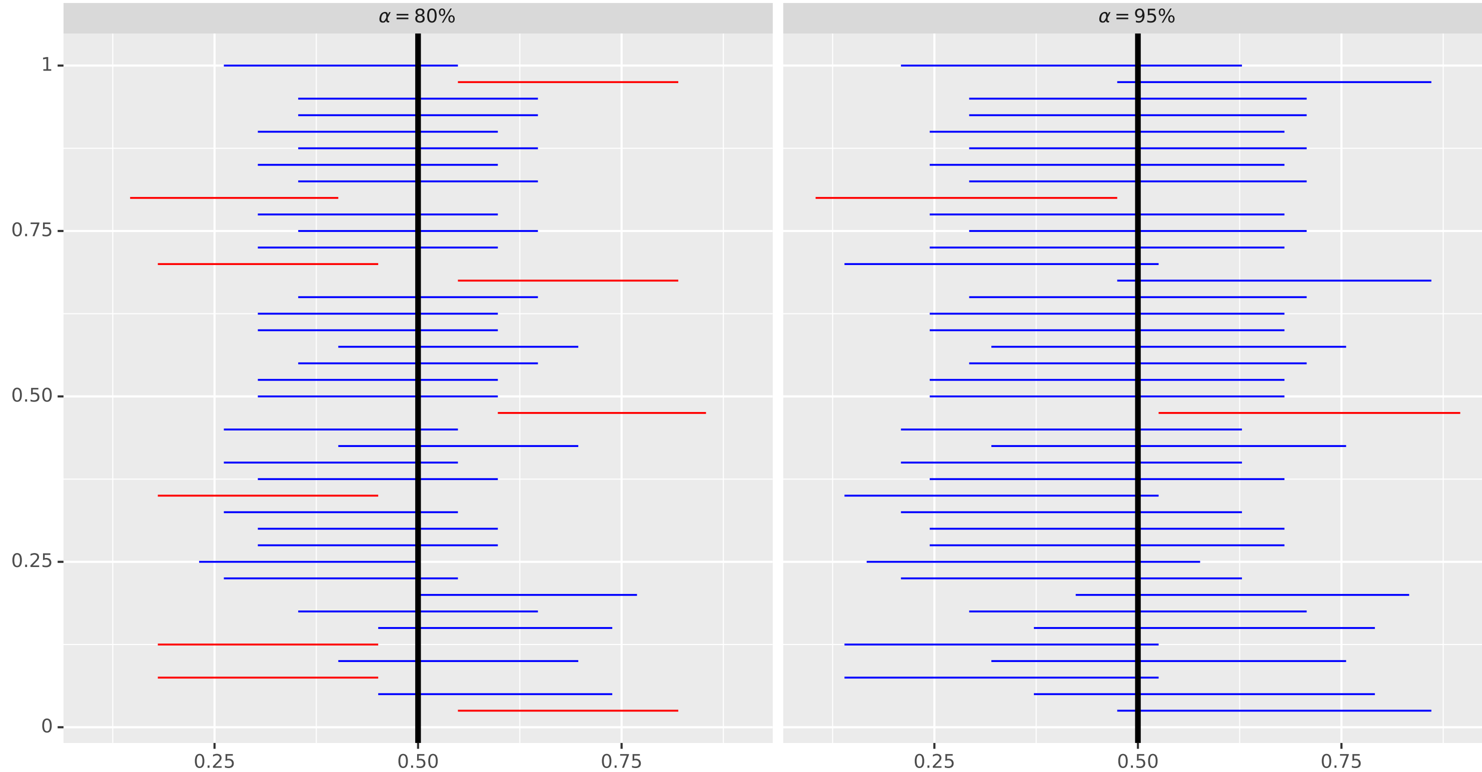
The definition does not state that the probability $p \in [\hat{p}-\varepsilon, \hat{p}+\varepsilon]$ is α!

- ▷ 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 $B_1 = \ldots = B_N = 1$?
- 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 $y = f(\boldsymbol{x})$ we cannot predict the observation $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, as the noise term ε_i is not known ahead.

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

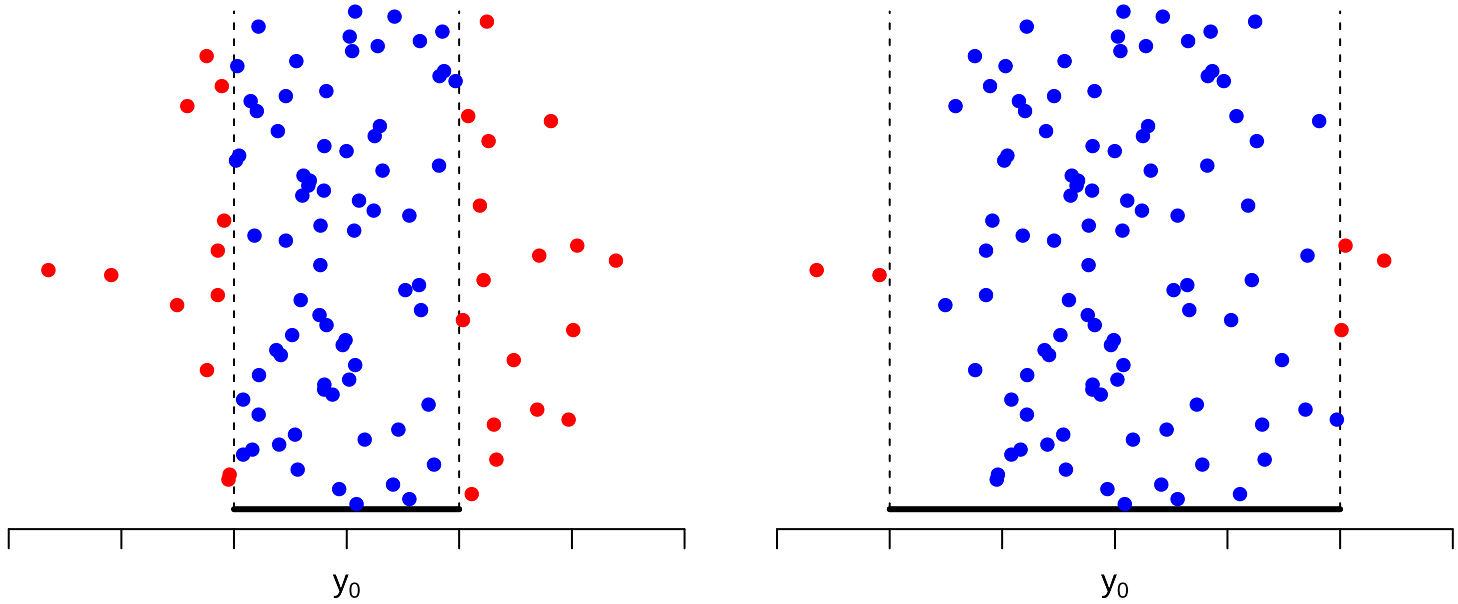
Instead, we can specify a prediction interval $[y_* - \varepsilon, y_* + \varepsilon]$ so that with probability 95% the resulting measurement y_i is in the range.

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

Interpretation of prediction intervals is different from confidence intervals.

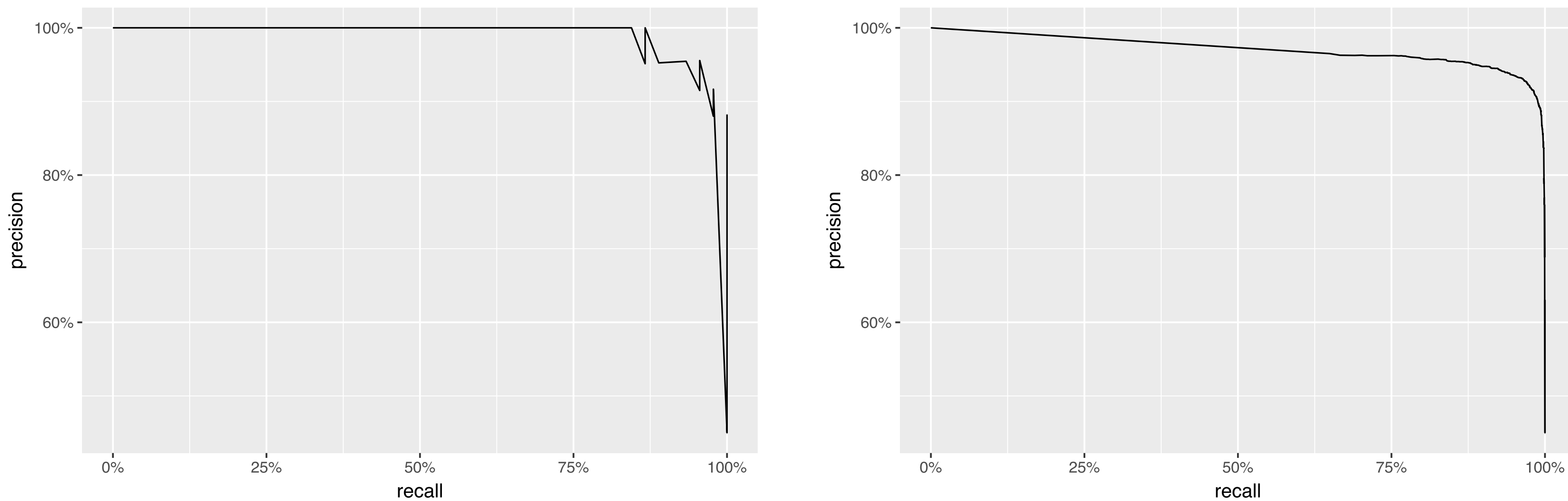
▷ The probability estimate holds for the particular interval.

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

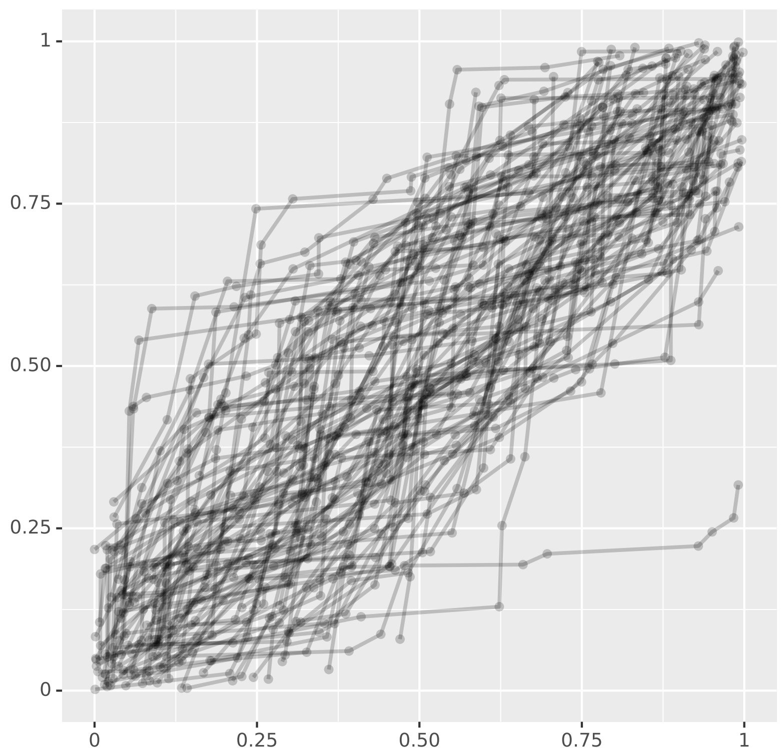
Confidence intervals is a good way to visualise uncertainty of a particular parameter. However, we are sometimes interested in the uncertainty many parameters or in the uncertainty of a function:

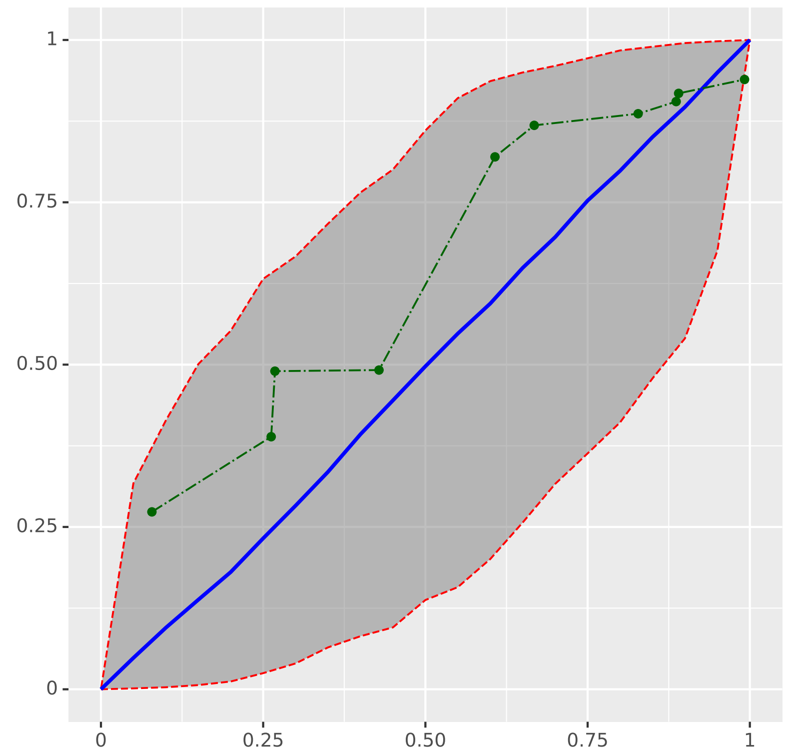
- ▷ How a predictor $f : [0, 1] \to \mathbb{R}$ depends on the training set
- ▷ How a ROC curve $\textsc{Roc} : [0, 1] \to [0, 1]$ depends on the test set
- ▷ How should a quantile-quantile plot be distributed.

Confidence bands are generalisations of confidence intervals

- ▷ Pointwise confidence band is a collection of confidence intervals
- ▷ Simultaneous confidence band must enclose α-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 π on outputs y_i destroys correlations between input-output pairs $(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})$ 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!

Crossvalidation

Empirical risk and law of large numbers

Under mild assumptions the empirical risk $R_N(f)$ converges to risk $R(f)$ and we can actually use normal distribution to estimate probabilities:

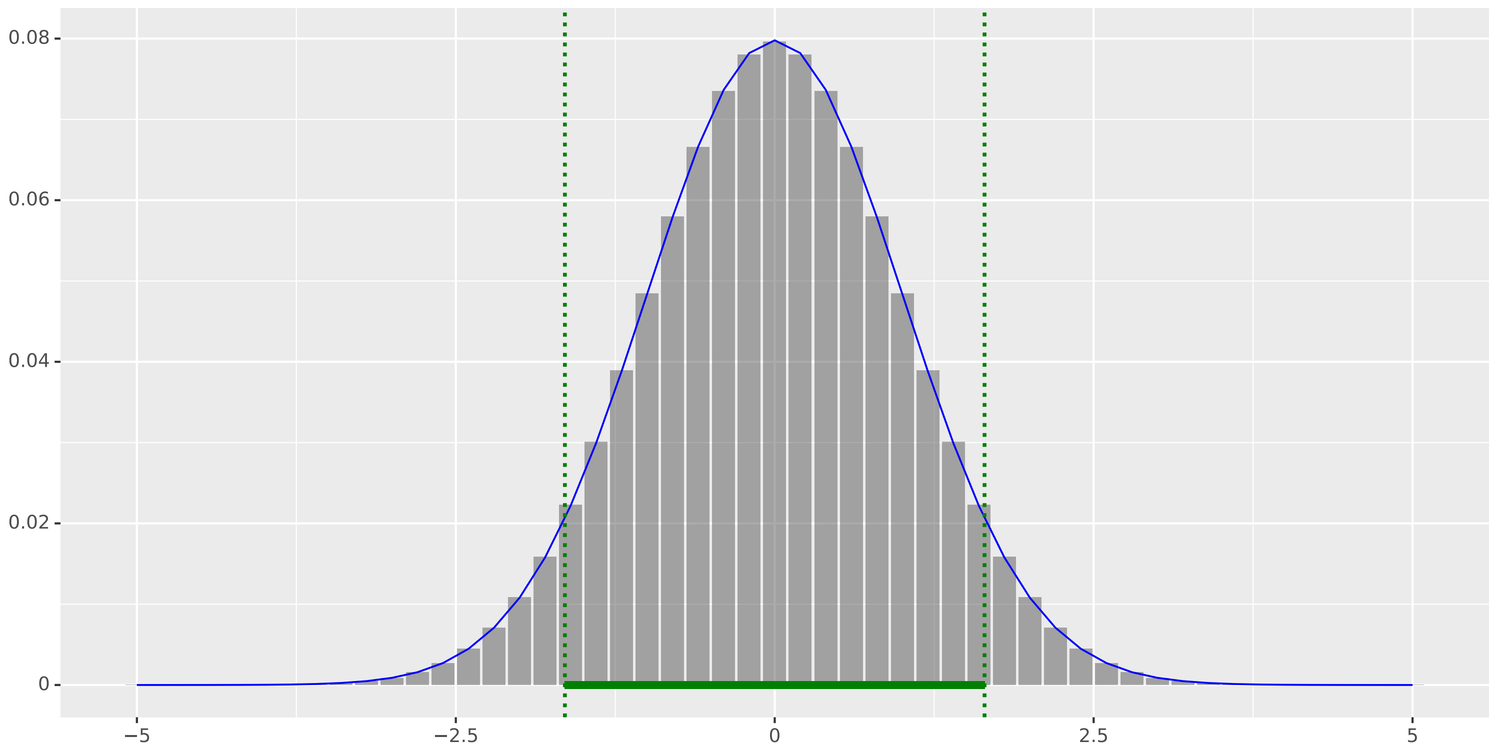
$$\Pr\left[|R_N(f) - R(f)| \geq \varepsilon\right] \lesssim 2 \cdot \int\limits_{-\infty}^{\varepsilon} \frac{\sqrt{N}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{Nt^2}{2\sigma^2}\right) dt$$

for a finite value σ where σ^2 is the variance of loss $\mathbf{D}(R(f))$.

What do we need to apply this result

- ▷ Test set elements must be independent and from the same distribution.
- ▷ CLT assumes that *risk μ is finite* and *standard deviation σ is finite*.
- ▷ Test set must be large enough that approximation is good enough.
- ▷ We *need to approximate* σ so that we can estimate the integral.

Visual representation



Convergence implies that the centre area of is well approximated

▷ 90% confidence intervals are roughly the same for both distributions

Moment matching

We know that the empirical risk $R_N(f)$ converges to normal distribution

- ▷ Normal distribution is fixed by a mean μ and variance σ^2
- ▷ We can estimate mean $\hat{\mu}$ and variance $\hat{\sigma}^2$ of a loss term $L(f(\boldsymbol{x}), y)$
- ▷ Then the estimates of mean and variance of the empirical risk are

$$\mathbf{E}(R_N(f)) \approx \hat{\mu}$$
$$\mathbf{D}(R_N(f)) \approx \frac{\hat{\sigma}^2}{N}$$

- ▷ This allows us to approximate $R_N(f)$ with normal distribution

Why do we need a test set at all

Machine learning algorithm

- ▷ Count number of zeroes n_0 and number of ones n_1 in training sample.
- ▷ If $n_0 > n_1$ output $f_0(x) \equiv 0$, otherwise output $f_1(x) \equiv 1$.

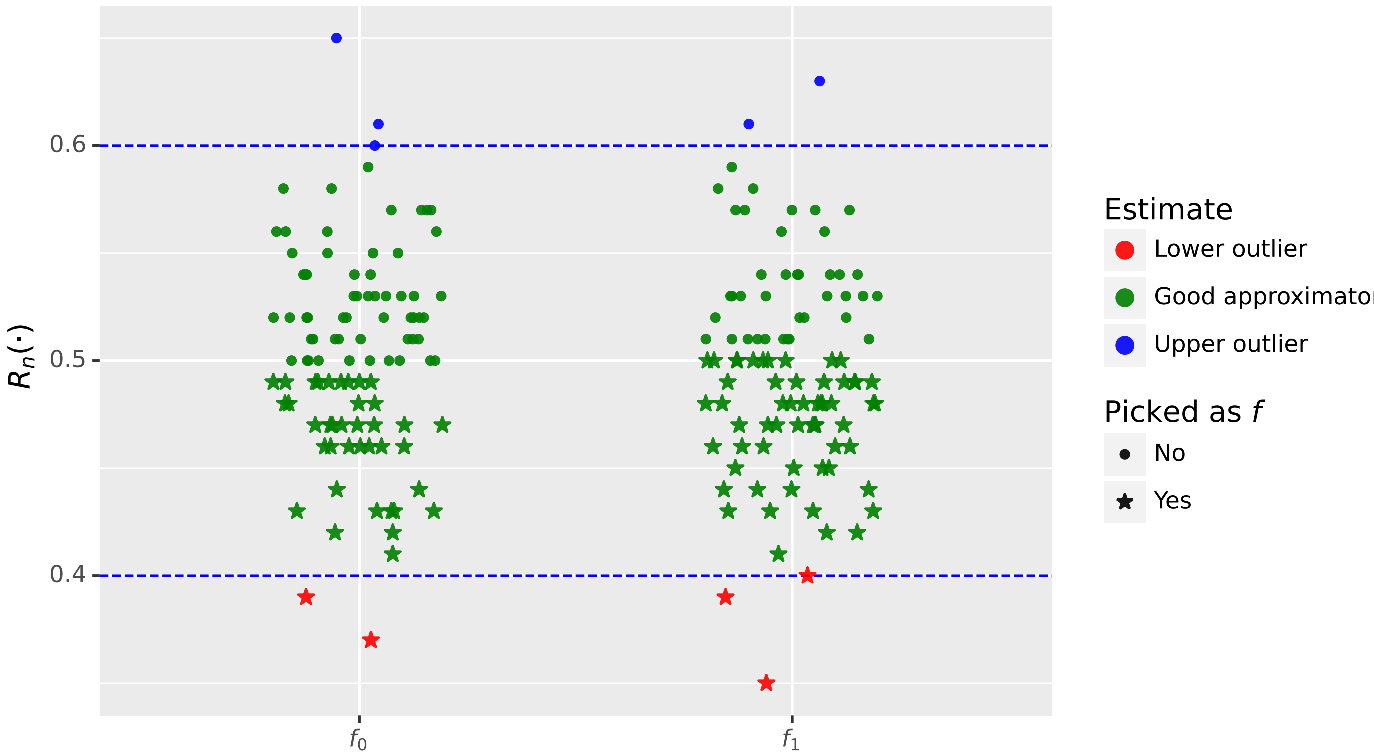
Data source

- ▷ Choose the input x randomly form the range $[0, 1]$
- ▷ Choose the label y randomly from the set $\{0, 1\}$.

True risk value

- ▷ Clearly the risk of both rules $R(f_0) = R(f_1) = 0.5$.
- ▷ The risk of of our learning algorithm $R(f)$ is also 0.5.

What happens during the training phase

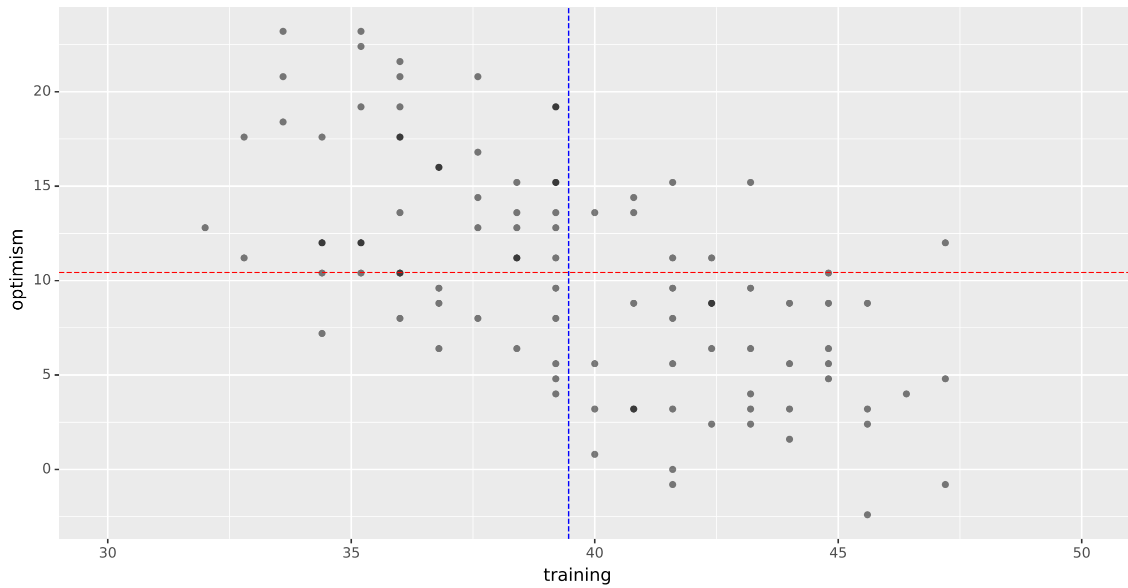


- ▷ We always choose the function f_i that underestimates the true risk!
- ▷ The probability that we go below the range effectively doubles.

What happens in real ML algorithms

- ▷ Not all function are achievable. More epochs more confidence intervals
- ▷ Not all measurements are independent. Many results are correlated.

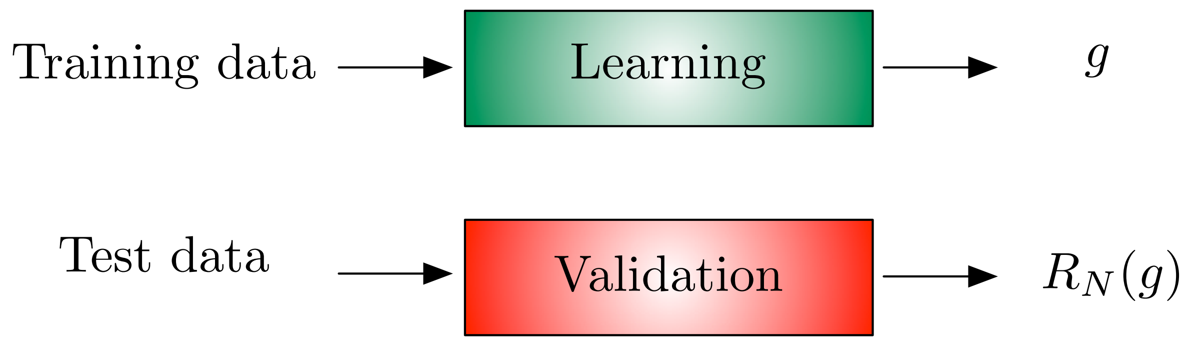
Generalisation gap aka optimism



By knowing the *optimism* $\Delta = R(f) - R_N(f_i)$ we can correct $R_N(f)$.

- ▷ Optimism is usually anti-correlated with empirical risk $R_N(f)$.
- ▷ Commonly mean value of Δ is used for the correction.
- ▷ Simple shifting does not resolve the systematical bias.

Why does the holdout testing work



By randomly splitting the data into training and test data we assure

- ▷ The training and test sets are independent under IID assumption.
- ▷ On a training set we compare many models and choose few winners.
- ▷ These functions are independent from the test set data.
- ▷ As there number of functions is small the law of large numbers holds.