Confidence Intervals for ‘the’ Generalization Error

Sebastian Fischer

Paper

- This presentation is based on the paper:
  “Constructing Confidence Intervals for “the” Generalization Error – a Comprehensive Benchmark Study”
- Authors: Fischer & Schulz-Kümpel & Hornung et. al
- Link: https://openreview.net/forum?id=x7kCj9OU2c

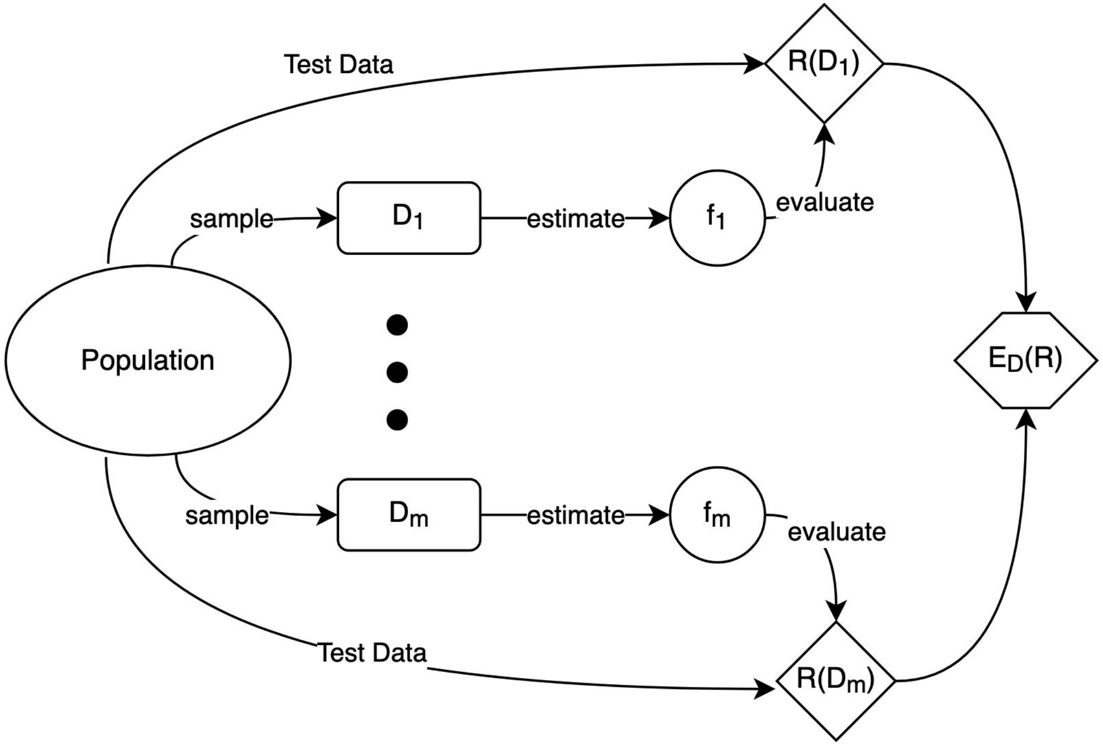
What is ‘the’ generalization error

(i) The risk, $\mathcal{R}_P(\hat{f}_{\mathcal{D}})$, *measures the error a specific model trained on specific data $\mathcal{D}$ will make on average when predicting for data from the same distribution.*

(ii) The expected risk, $\mathbb{E}\big[\mathcal{R}_P(\hat{f}_{\boldsymbol{\mathcal{D}}})\big]$, *measures the error of models that have been trained using inducer $\mathcal{I}$ on data of size n. Thus, it measures the quality of the general inducer on arbitrary data of size n from distribution P rather than the quality of a single model.*

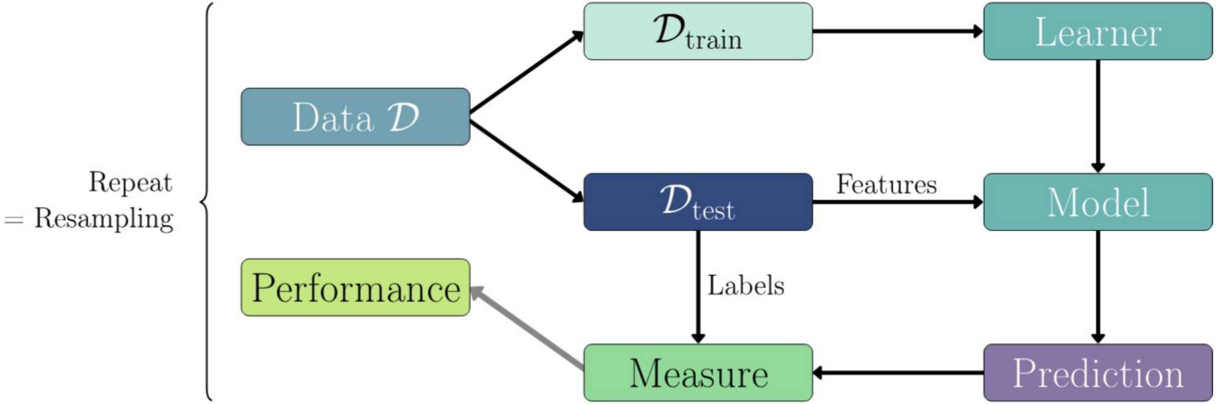
We will refer to both as Generalization Error

Risk vs Expected Risk



How to Estimate the (Expected) Risk?

In practice, we can't sample from the DGP, so we need to use our data D for both training and testing.



$$\hat{P}_n^{(H)} = \mathcal{R}_{\mathcal{D}_{\text{test}}}(\hat{f}_{\mathcal{D}_{\text{train}}}) = \frac{1}{n_{\text{test}}} \sum_{(x,y)\in\mathcal{D}_{\text{test}}} \mathcal{L}(y, \hat{f}_{\mathcal{D}_{\text{train}}}(x))$$

How to Construct a Confidence Interval

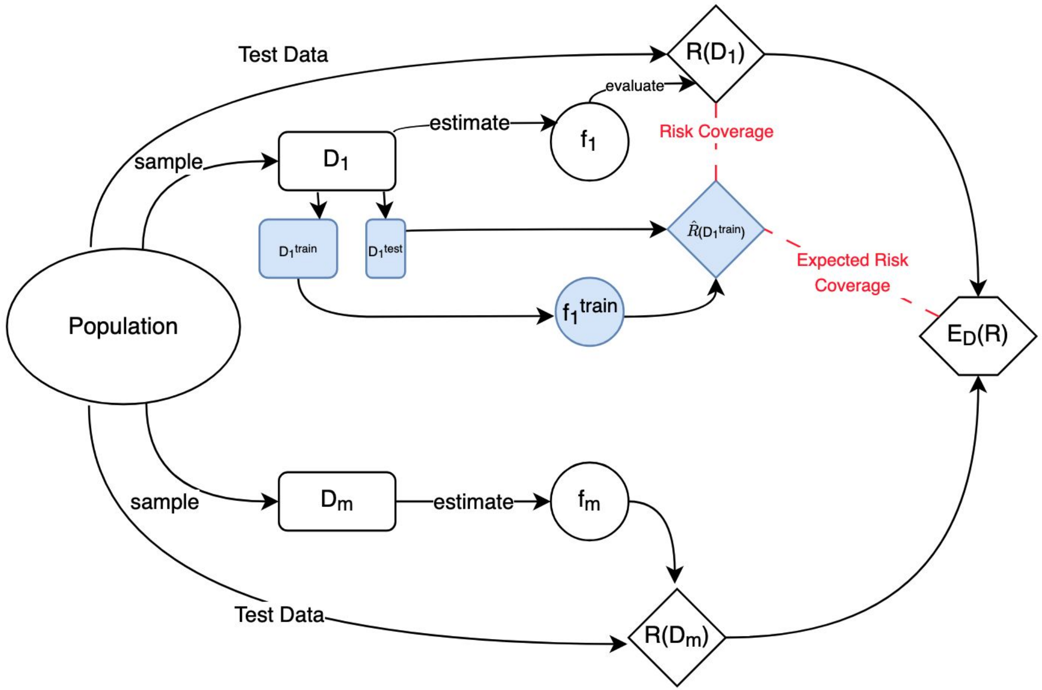
For the simple Holdout resampling, construction of the Confidence Interval is straightforward:

$$\hat{\sigma}_H^2 = \frac{1}{n_{\text{test}} - 1} \sum_{i \in J_{test}} \left(\mathcal{L}(y^{(i)}, \hat{f}_{\mathcal{D}_{\text{train}}}(x^{(i)})) - \mathcal{R}_{\mathcal{D}_{\text{test}}}(\hat{f}_{\mathcal{D}_{\text{train}}}) \right)^2$$

The corresponding CI is then given by

$$\left[\hat{P}_n^{(H)} \pm z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}_H}{\sqrt{n_{\text{test}}}} \right],$$

Estimator based on simple Train/Test (Holdout) Split



What does the estimator estimate?

- It kind of estimates both the Risk and Expected Risk, which is also intuitive as the formers is the expectation of the latter
- There is always a size bias, because we train the model on less data.
- In our experiments, we found little differences between the relative coverage frequencies of the risk and the expected risk

What makes a good Estimator & Confidence Interval?

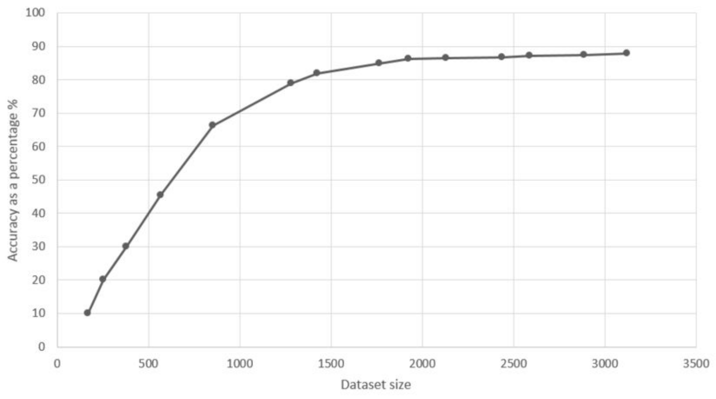
- It has the correct alpha level
- It is narrow (i.e. the estimator has low MSE)
- it is computationally feasible

This can be evaluated:

- mathematically, e.g. via asymptotic analysis
- empirically

What’s the problem of the simple holdout split?

- There is a tradeoff between the size bias and the estimation variance on the test set
- This is especially problematic for “small” datasets D



Other Resampling Methods to the Rescue

- Methods like CV “solve” this problem by simultaneously reducing bias and estimation variance

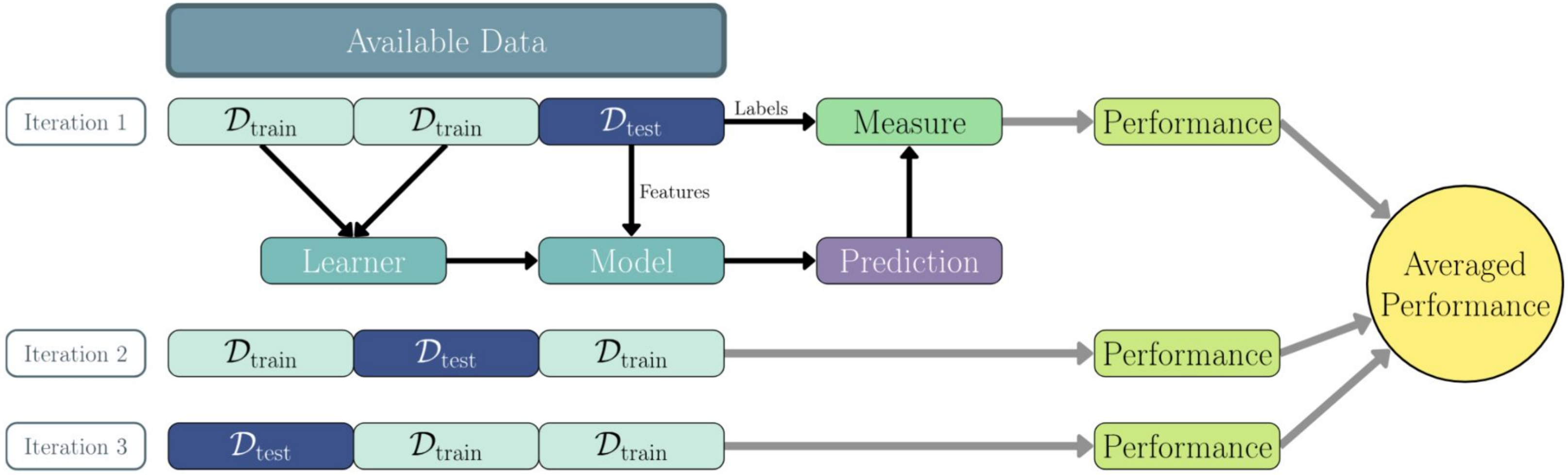


Figure from Chapter 3 of the mlr3 book

What does Cross-Validation estimate?

- Recommended Paper: “Cross-Validation: What Does It Estimate and How Well Does It Do It?” (Bates et. al, 2024)
- In a Nutshell: The CV point estimator is a better estimator for the Expected Risk than for the Risk

So, what’s the catch?

Difficulties of deriving CI methods for the CV point estimate

Question: Why can't we just use the "naive" estimator for the variance?

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in J_{\text{test},k}}(e_k[i]-\hat{P}_n)^2$$

Impossibility of Unbiased Estimator

Theoretical Result: It’s impossible to obtain an unbiased estimator of the variance of the CV point estimator using the results of a single Cross-Validation (Bengio & Grandvalet, 2004).

But what is the difficulty?

Covariance Structure of Cross-Validation Losses

Corollary 2 *The covariance matrix Σ of cross-validation errors $\mathbf{e} = (e_1, \ldots, e_n)'$ has the simple block structure depicted in Figure 2:*

1. *all diagonal elements are identical*

   $\forall i,\ \mathrm{Cov}(e_i, e_i) = \mathrm{Var}[e_i] = \sigma^2$;

2. *all the off-diagonal entries of the K $m \times m$ diagonal blocks are identical*

   $\forall (i, j) \in T_k^2 : j \neq i,\ \mathrm{Cov}(e_i, e_j) = \omega$;

3. *all the remaining entries are identical*

   $\forall i \in T_k,\ \forall j \in T_\ell : \ell \neq k,\ \mathrm{Cov}(e_i, e_j) = \gamma$.

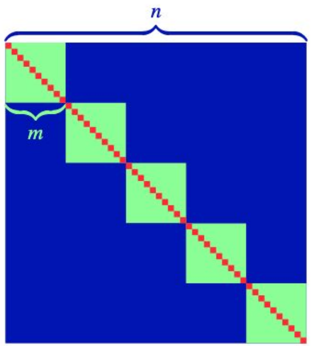


Figure 2: Structure of the covariance matrix.

[Grandvalet & Bengio (2004)](#)

Other methods

Method name	Resampling method**	Cost***	Theoretical guarantee	Reference
*Holdout (H)**	Holdout	1	yes	Nadeau and Bengio (2003)
*Replace-One CV (ROCV)**	(LOO)CV ($\hat{P}_n$), ROCV ($\hat{\sigma}$)	$(n/2+2)K$	yes	Austern and Zhou (2020)
*Repeated Replace-One CV (HRCV)**	Repeated CV ($\hat{P}_n$), RORCV ($\hat{\sigma}$)	$(n/2+2)RK$	no	
*CV Wald (CVW)**	(LOO)CV	K	yes	Bayle et al. (2020)
Corrected Resampled-T (CRT)	Subsampling	K	no	Nadeau and Bengio (2003)
Conservative-Z (CZ)	Subsampling ($\hat{P}_n$), Paired Subsampling ($\hat{\sigma}$)	$(2R+1)K$	no	Nadeau and Bengio (2003)
5×2 *CV* (5×2)	Repeated CV	10	no	Dietterich (1998)
Nested CV	Nested CV	RK^2	no	Bates et al. (2024)
Out-of-Bag (OOB)	Bootstrap	R	no	Efron and Tibshirani (1997)
632+ *Bootstrap* (632+)	Insample + Bootstrap	$R+1$	no	Efron and Tibshirani (1997)
BCCV Percentile (BCCVP)	BCCV ($\hat{q}$), LOOCV ($\hat{b}$)	$(0.632R+1)n$	no	Jiang et al. (2008)
Location-shifted Bootstrap (LSB)	Insample Bootstrap ($\hat{q}$), Insample + Bootstrap ($\hat{P}_n$)	$1+2K$	no	Noma et al. (2021)
Two-stage Bootstrap (TSB)	Two-stage Bootstrap ($\hat{q}$), Insample + Bootstrap ($\hat{P}_n$)	$(R+1)(K+1)$	no	Noma et al. (2021)

Our Benchmark Study

- We compared all methods we found in the literature on constructing CIs for the Generalization Error
- These inference methods differ w.r.t.:
  - The resampling scheme
  - How to construct the CI from the results of the resample experiment
- We evaluated them on:
  - Different Inducers: Linear/Logreg, Decision Tree, Random Forest, MLP (tuned), XGBoost (tuned)
  - 19 different DGPs
  - various loss functions
  - different hyperparameter configurations (of the inference methods)

Results of the Benchmark Study in a Nutshell

- Out of the 13 different CI methods, 5 performed decent and 3 rather well:
- Decent:
  - Holdout (good coverage, but too wide)
  - Wald CV (mostly reasonable coverage, but poor coverage with decision tree)
- Good:
  - Nested CV (good coverage, a bit conservative, a bit expensive)
  - Conservative Z (conservative, a bit expensive)
  - Corrected T (a bit too liberal)

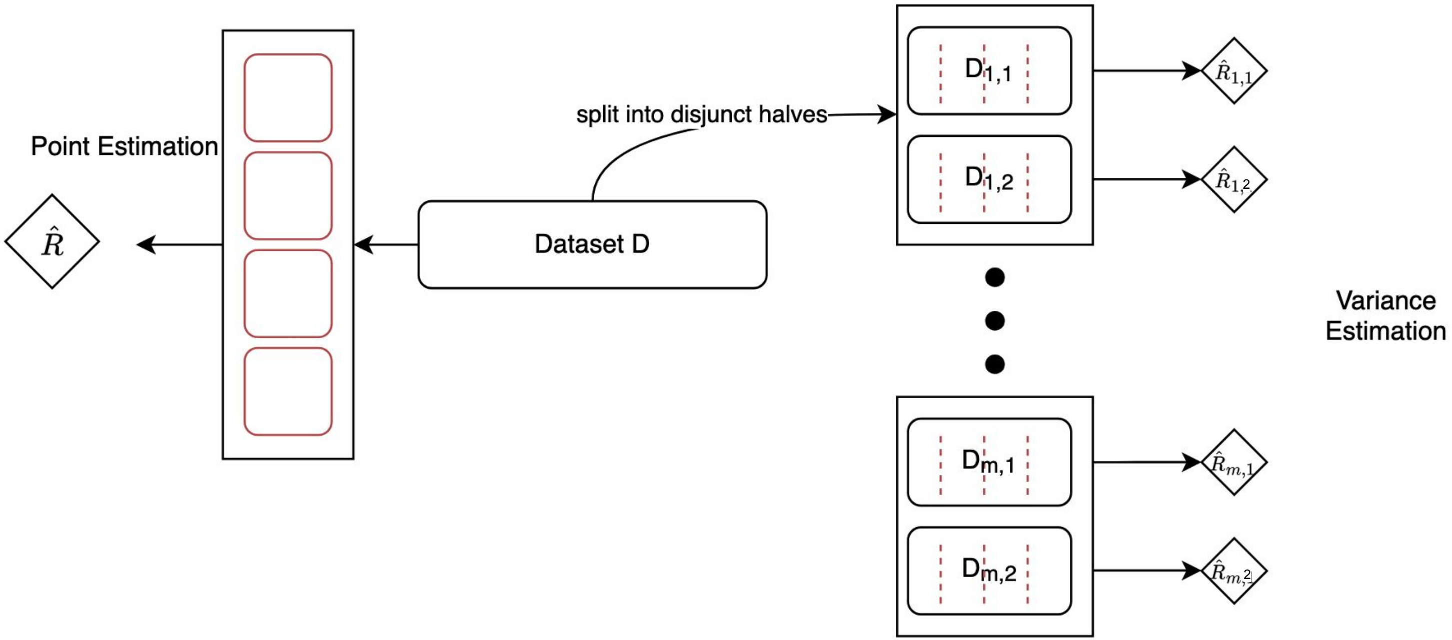
We will now present the main ideas of the methods

Corrected T

- The corrected t method is based on Subsampling, aka Repeated Holdout, aka Monte Carlo Cross-Validation
- The method corrects applies a (heuristic) correction factor to the estimator that assumes normality: K is number of Folds, and n2 is test set size

$$\widehat{\mathrm{SE}}^2_{CRT} = \left(\frac{1}{K} + \frac{n_2}{n - n_2}\right) \cdot \hat{\sigma}(\boldsymbol{\mathcal{D}}_n)^2 ,$$

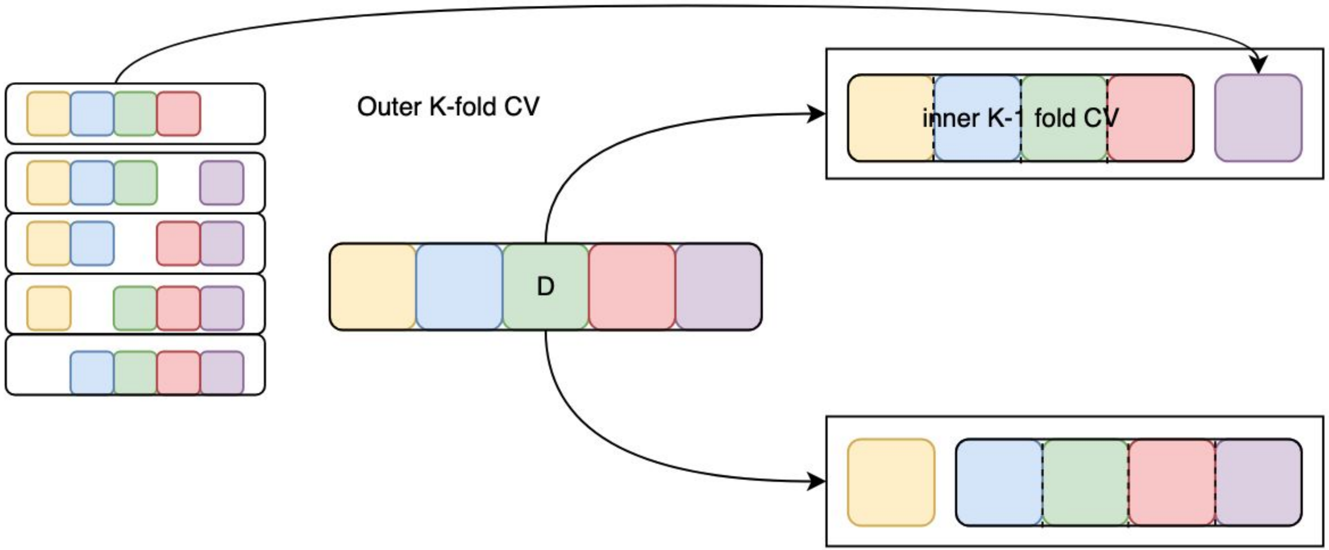
Conservative Z



The method is **conservative**, because the variance estimator uses datasets half the size of the actual data D

Nested CV

- Below, we see one iteration of Nested CV (which can be repeated).
- This is not the standard Nested CV for evaluating hyperparameter tuning, but a special method



Wald CV

- Earlier we said, we can't just use the "naive" CV estimator, because it ignores the covariance structure
- But, Bayle et. al, have shown asymptotical exactness for the estimator, albeit w.r.t. a different target quantity (a "proxy quantity")

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{(x,y)\in\boldsymbol{\mathcal{D}}_{\text{test},k}}\mathbb{E}[\mathcal{L}(y,\hat{f}_{\mathcal{I},\boldsymbol{\mathcal{D}}_{\text{test},k}}(x))|\boldsymbol{\mathcal{D}}_{\text{train},k}]$$

- It also works reasonably when evaluated w.r.t. the Risk/Expected Risk

Types of Hyperparameters

In general, the hyperparameters of the inference methods can be divided into three categories:

- There are hyperparameters that have a tradeoff, such as the ratio of training and test data (variance vs. bias)
- Other Hyperparameters reduce the variance or bias in the point estimate (Number of folds in CV, number of repetitions during Subsampling)
- Then there are hyperparameters that are primarily intended to reduce the estimation variance of the standard error estimate for more accurate CIs (Outer repetitions of Nested CV or Conservative Z)

What to do in practice?

Based on our empirical results, we recommend the following methods:

- For small data (up to $n = 100$):
  - Nested CV with at least 25 outer repetitions and $K = 5$, or
  - Conservative-Z with 25 outer repetitions and at least $K = 10$
- For larger data:
  - Corrected Resampled-T with a ratio of 0.9 and at least 25 repetitions, or
  - Conservative-Z with 10 outer repetitions and $K = 5$ for slightly wider CIs with very slightly more accurate coverage.

Some things to keep in mind

- We did not consider imbalanced data (“small data in disguise”)
- We did not consider grouping
- Inference methods can fail when the model fitting is very unstable
- Inference methods can completely fail when there are strong outliers in the data, but this can be somewhat mitigated by using robust loss functions for evaluation
- Some Inference methods (Holdout, Wald CV, Nested CV) only work with pointwise loss functions (not e.g. AUC)