# CM 3120 Computational Statistics

## Chapter 2 The Bootstrap and Jackknife

Dr. Priyanga Talagala

12/08/2022

## Simulation (Chapter 1) vs Resampling (Chapter 2)

- In Chapter 1 we learnt how to use simulation techniques to sample and compute quantities from known distributions (Simulation - sampling from a known distribution)
- Naturally, we cannot simulate draws from an unknown distribution but we can draw from a sample of observations. If the sample is a good representation from the population, then our simulated draws from the sample should well approximate the simulated draws from a population.
- The process of sampling from a sample is called resampling.
- Resampling methods are a natural extension of simulation.



# Resampling

### What is resampling

- Resampling is a statistical technique to reuse data to generate new, hypothetical samples (called resamples) that are representative of an underlying population.
- Statistics of interest (eg: sample mean , median) are calculated for each new sample.
- The distribution of new statistics can be analysed to investigate different properties (eg: confidence intervals, the error, the bias) of statistics.

## Recap : CM 2110/ CM2130

- Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard deviation $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Standard error $SE_{\bar{x}} = \sqrt{\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n(n-1)}}$
- Bias of an estimator is the difference between the estimators expected value and the true value of the parameter being estimated.
- The 95% confidence interval is a range of values that you can be 95% confident, contain the parameters of the population.



### When to use resampling ?

- You don't know the underlying distribution for the population,
- Traditional formulas are difficult or impossible to apply
- As a substitute for traditional methods.
- The jackknife and the bootstrap are two nonparametric methods for estimating or approximating the sampling distribution of a statistic and its characteristics.
- They provide several advantages over the traditional parametric approach:
  - methods are easy to describe
  - can be applied to arbitrarily complicated situations; distribution assumptions, such as normality, are never made.



## Motivation

- A decision tree classifier is a systematic approach for multiclass classification.
- ML techniques: Decision Tree Ensembles- Bagging and Boosting
- Bootstrap aggregation, or bagging, is a popular ensemble method in machine learning that fits a decision tree on different bootstrap samples of the training dataset.
- Resampling methods are also useful to fit more accurate models, model selection and parameter tuning. (Cross-Validation for model selection)



# Jackknifing

## Jackknifing

- A resamling technique especially useful for finding standard error, variance and bias of estimators.
- The jackkife is a small, handy, rough-and-ready tool (a compact folding knife) that can improvise a solution for a variety of problems
- The jackknife technique was developed by Maurice Quenouille (1924–1973) from 1949 and refined in 1956.
- John Tukey expanded on the technique in 1958 proposed the name "jackknife" because, like a physical jack-knife , it is a rough-and-ready tool that can improvise a solution for a variety of problems even though specific problems may be more efficiently solved with a purpose-designed tool.



- The jackknife is also known as leave-one-out (LOO)
- The jackknife is a linear approximation of the bootstrap.
- This approach tests that some outlier datapoint is not having a disproportionate influence on the outcome.
- The jackknife deletes each observation and calculates an estimate based on the remaining $n-1$ values.
- It uses this collection of estimates to do things like estimate the bias and the standard error.



## Jackknifing: definition

- Let $x_1, \ldots, x_n$ be a dataset
- $\theta$ is a parameter you want to estimate from the data (eg: Population mean, median, standard deviation)
- Let $\hat{\theta}$ be the estimate based upon the entire dataset
- Let $\hat{\theta}_i$ be the estimate of $\theta$ obtained by deleting observation $x_i$
- Let $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$
  - Sometimes $\bar{\theta}$ is written as $\bar{\theta}_{(.)}$



- This provides an estimated correction bias due to the estimation method. the jackknife does not correct for a biased sample
- The jackknife estimate of bias is $B = (n-1)(\bar{\theta} - \hat{\theta})$.
  - In other words, it is the difference between the actual and the average of the delete-one estimates.
- We can then correct $\hat{\theta}$ (the estimator on the entire dataset), using

$$\hat{\theta}_{corrected} = \hat{\theta} - B$$

$$\hat{\theta}_{corrected} = n\hat{\theta} - (n-1)\bar{\theta}$$

- The $\hat{\theta}_{corrected}$ is the bias-corrected jackknife estimate of $\theta$ of the population.

- The jackknife method is more conservative than the bootstrap method.
- When the estimator is not normally distributed jackknifing may fail.
- Bootstrapping performs better for skewed distributions.
- The Jackknife gives the same results every time, because of the small differences between replications.
- The bootstrap gives different results each time that it's run.