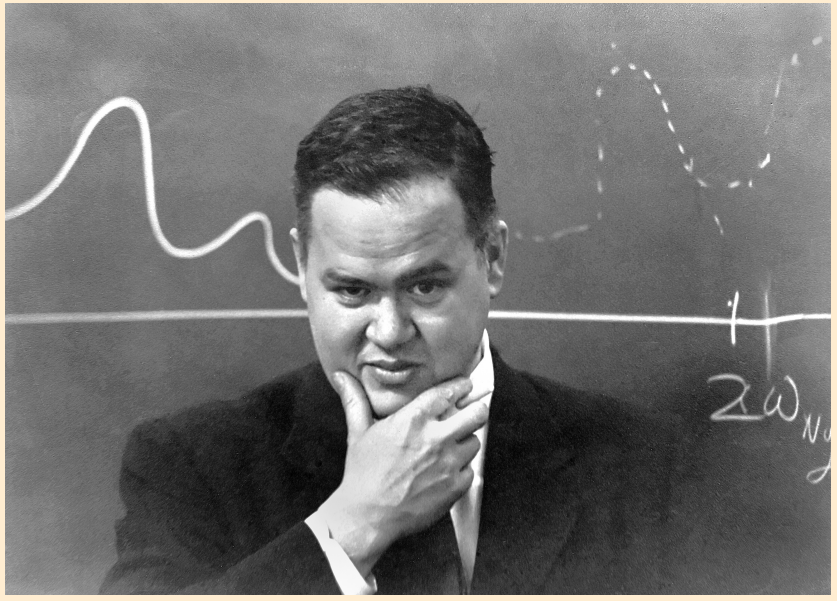
Introduction

Statistician John Tukey defined ***Data Analysis*** in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data".



John Wilder Tukey was an American mathematician best known for development of the Fast Fourier Transform (FFT) algorithm and the boxplot.

Introduction

Class eBook Section 1.6

- When the sample per se is not the target (i.e. in most of the cases), then one enters into the process of (statistical) inference.
- A fundamental aspect of statistical inference is the ability of constructing (manageable) measures of variability (probability) to any data treatment operated in order to produce information.
- Possible aims:
  - understand the phenomenon under investigation,
  - prediction,
  - evaluate research hypotheses, etc.
- In any cases, an **inferential risk** measure of risk associated to any inference made from the sample to the population, is needed.
- **Model selection** in statistics can be seen as a ***va-et-viens*** process between model setting and **model assessment**.

Introduction

- For statistical inference, useful techniques/methods exist.
- Mathematical statistics for deriving (exact) distributions (e.g. the Student t distribution).
- Asymptotic methods (expansions) for deriving asymptotic distributions.
- Plugin principle for estimating (asymptotic) variances or other statistics (e.g. prediction errors).
- Simulation methods (e.g. the **bootstrap**) for estimating variances or other statistics (e.g. prediction errors).
- These methods rely on the correct understanding of the models (probability distributions).

Introduction

- Model selection is a broad concept.
- Estimation (parametric, semi-parametric, nonparametric) is a form of model selection: for a family of models $F_{\boldsymbol{\theta}}$, there exists an (almost) infinite number of different ones according to the value of $\boldsymbol{\theta}$.
- Simultaneously managing the dimention p of $\boldsymbol{\theta}$: inferential process, with two sources of risk measures, estimation and model fit (see e.g. nonparametric estimation).
- A trade-off needs then to be made between the **model complexity** (i.e. p) and the **model adequacy** (i.e. its fit to the data).
- The choice of the measure associated to the trade-off is also important.

Introduction

- Reducing model complexity:
  - when we are mainly interested in hypothesis testing and the study of phenomena, then restricted models are much more interpretable,
  - when the objective is prediction only, more flexible models (larger p) provide less accurate predictions (!).
- Accuracy is understood here **out-of-sample**, i.e. in the population of interest.
- Optimizing a model within a sample, is rather straightforward... just make it complex enough...
- Optimizing a model out-of-sample, involves the computation (and development) of adequated probabilistic measures.

Introduction

- One can separate model selection procedures in three broad categories.
- **Subset Selection**: This approach involves identifying a subset of the p predictors (i.e. a non zero subset of $\boldsymbol{\theta}$).
- **Shrinkage**: This approach involves fitting a model involving all p predictors (or parameters in $\boldsymbol{\theta}$), under a constraint to minimize the Mean Squared Error (MSE). This shrinkage is also known as **regularization**.
- **Dimension Reduction**: This approach involves reducing the p predictors into a q-dimensional subspace, where $q < p$, using a trade-off between information loss due to the dimension reduction and model complexity. The q (orthogonal) axes of the subspace are then used as predictors to fit the model.

Introduction

- Subset selection and shrinkage methods allow for **model interpretability**.
- Dimension reduction is more appropriate for pure prediction.
- In terms of out-of-sample prediction error, one set of approaches it is not clear which method is the best.
- In this course we will mainly focus on subset selection while also presenting shrinkage methods.