Resampling Methods.

* **Introduction:**

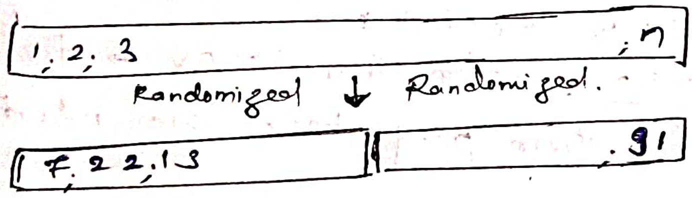
- Resampling method:
repeatedly drawing samples from a training set and fitting a model of interest on each sample in order to obtain additional information about the model.

- cross validation:
can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance or to select appropriate level of flexibility.

- model assesment: process of evaluating a model's performance
- model selection: process of selecting the proper level of flexibility.
- bootstrap: used in several contexts, most commonly to provide a measure of accuracy of a parameter estimate or of a given statistical learning method

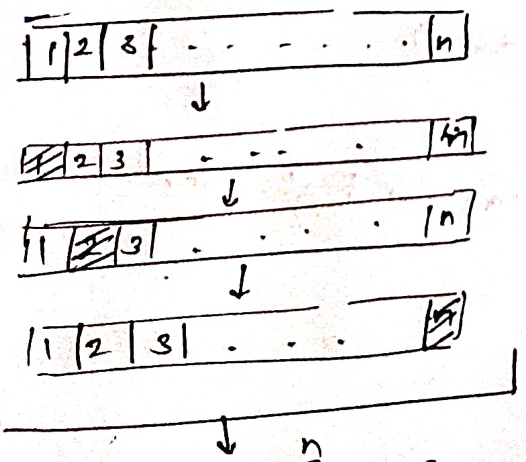
1. Cross validation.

5.1.1: The validation set approach.
- it involves randomly dividing the available set of observations into two parts: training set (used for model fitting) & a validation set/hold-out set (prediction & calculating estimate of test error)



Drawbacks
* validation estimate of the test error rate can be highly variable, depending on which observations included in train & test.
* if training set small - model may perform bad & validation set error may tend to overestimate the test error rate.

5.1.2 Leave One out cross validation. (LOOCV)
- attempts to address drawbacks of ~~cross~~ validation set approach.



$MSE_i = (y_i - \hat{y}_i)^2$ → not included in validation set the ith observatn.

→ consider all $\binom{n}{1}$ = n possible leave one out ~~cross~~ validation sets then get the average: that is the required estimation.

* Advantages
- less bias (on repeatitions).
- does not over estimate test error.
- no randomness involved in the train-validation split.

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Note: LOOCV has potential to be expensive for implementation.

For least squares linear / polynomial regression:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2 \longrightarrow$$ This formulae does not hold true in general!

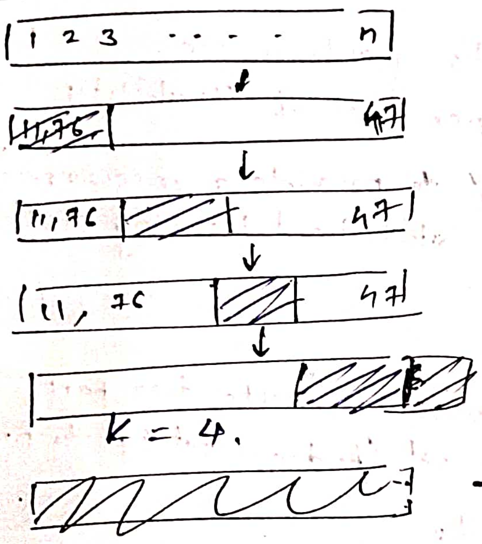
where: $\hat{y}_i$: ith fitted value from the original least squares fit.

h_i : leverage $= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

5.1.3: k-fold cross validation: (1) Randomly split into k many non overlapping groups of equal size.

(2) i^{th} group held out as validation set. remaing 'k-1' used for training model.

(3) repeat process for holding out each groups i : i = 1(1) k-1



Formula: $CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$

generally : k = 5 or 10 (computation perspective)

→ very less variability compared to validation set approach

5.1.4. Bias - Variance Trade off for k-Fold Cross-Validation.

- k fold cv gives more accurate estimate of the test error.
- no. of observations in k fold ≃ $\frac{(k-1)n}{k}$ → more than LOO CV but less than validation set approach ⟹ from perspective of bias reduction LOOCV is prefered to k - fold CV.
- LOOCV has higher variance than does k-fold CV with k<n.

→ LOOCV : - training sets almost identical - highly +vely correlated.
- avereging output of these n fitted outputs.

k fold : - less over lap in training sets ⟹ less correlated

var (mean of less corr. data) < var (mean of highly correlated data)

∴ var (k fold) < Var (LOOCV)

⟹ for k fold CV with k=5 or 10 : ~~not~~ no high bias nor high variance

5.1.5 Cross validation on classification problems.

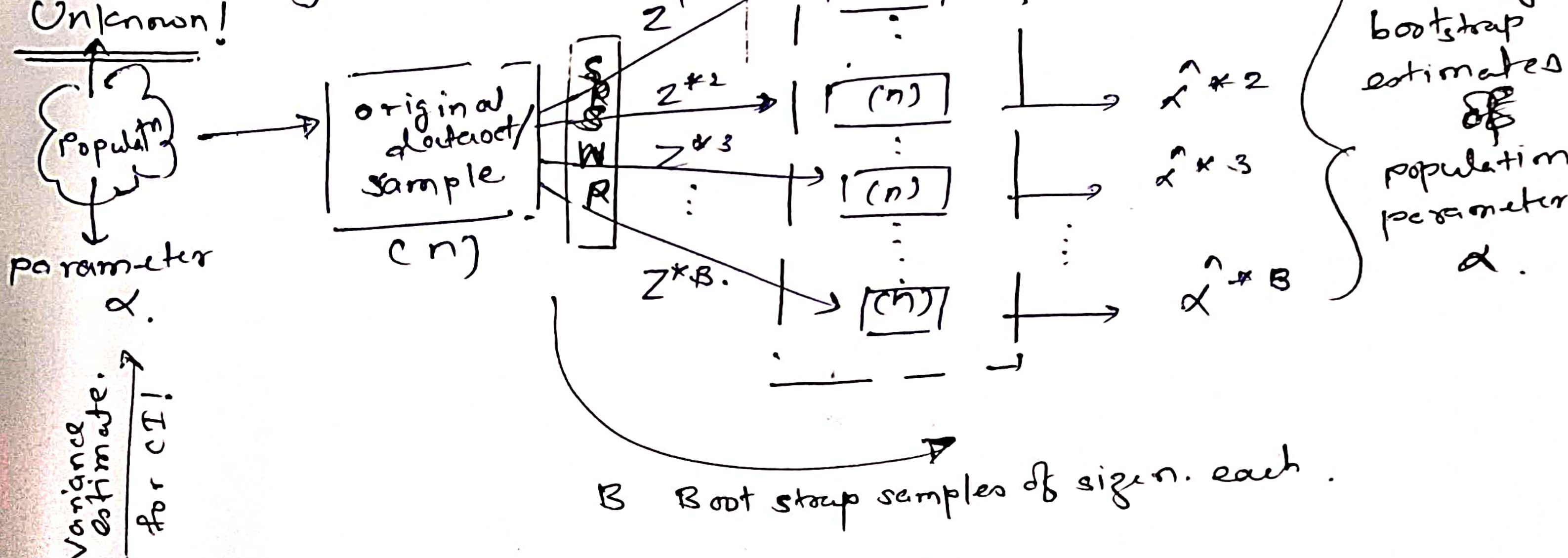
- Regression: MSE, classification: error rate/ ~~no. of~~ fraction of misclassified objects.

For LOOCV :
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \quad \text{where } Err_i = I(y_i \neq \hat{y}_i)$$

For K fold CV :
$$CV_{(K)} = \frac{1}{K}\sum_{i=1}^{K} Err_i$$

* 2. Bootstrap:

- widely applicable & extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- used when measure of variability is hard to obtain.
- rather than repeatedly obtaining independent data sets from the population we instead obtain distinct data sets by repeatedly sampling observations from the original data set.



$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\hat{\alpha}^{*r}\right)^2}$$

* Advantage : - Can be used in almost all situations.
  - No complicated mathematical ~~situat~~ formulae required.