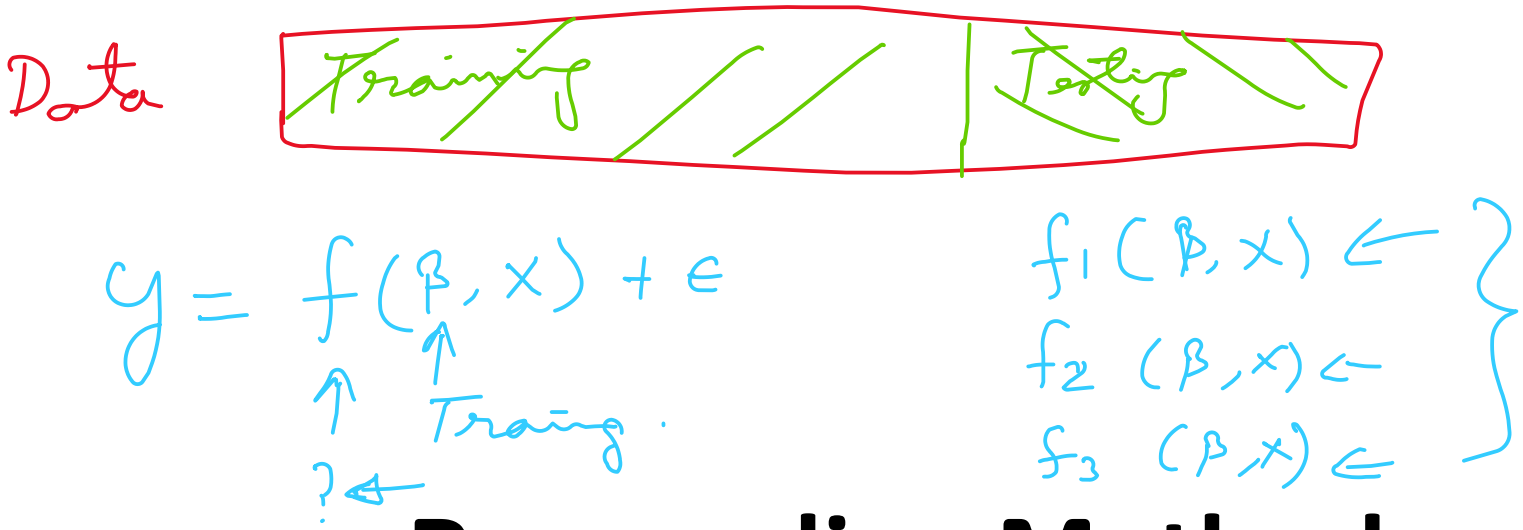


Resampling Methods

Prof. Asim Tewari
IIT Bombay

Resampling

- *Resampling* involve repeatedly drawing samples from a training set and refitting a model of interest on each sample

- Can be computationally expensive

- Resampling methods
  - *Cross-validation*
  - *Bootstrap*

Resampling

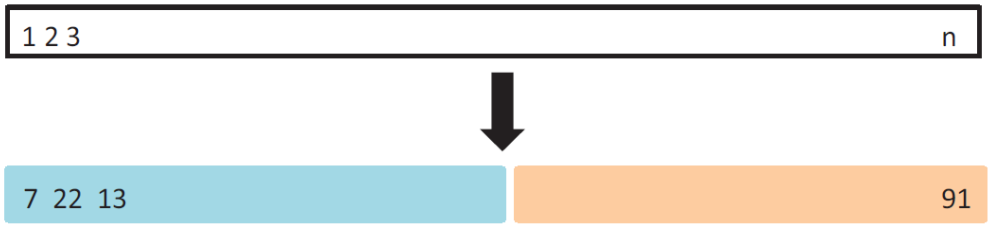
- *Model assessment:* The process of evaluating a model's performance

- *Model selection:* The process of selecting the proper level of flexibility.

Resampling

Cross-Validation

- *The Validation Set Approach*
  - It involves randomly dividing the available set of observations into two parts, a training set and a validation set or hold-out set.

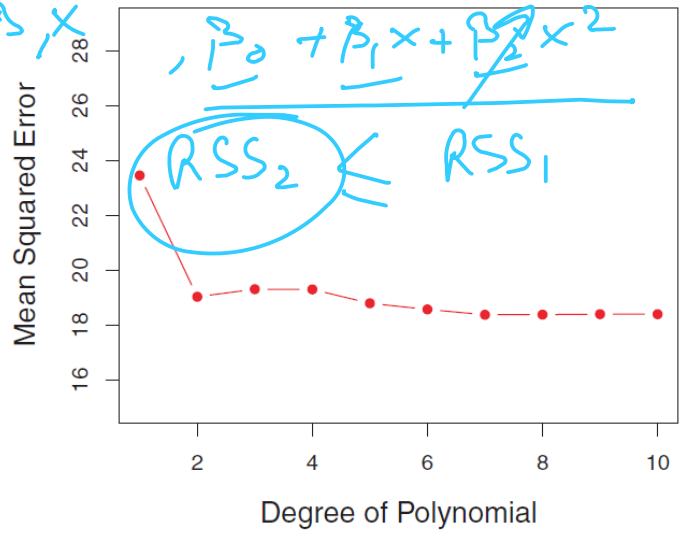


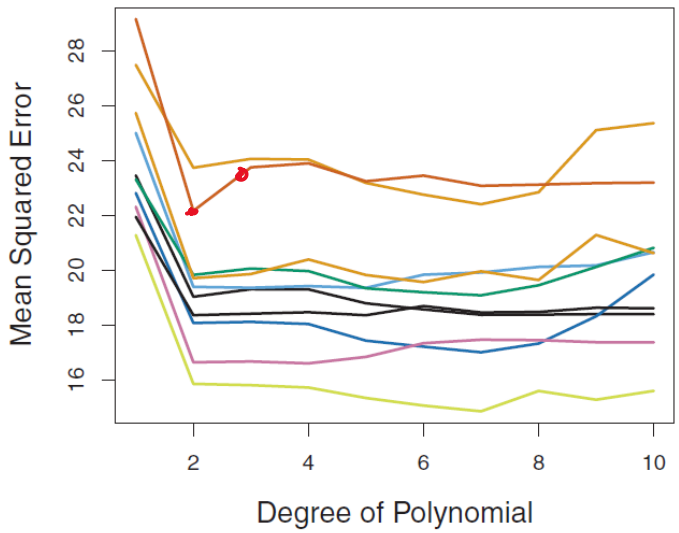
A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

Resampling

Cross-Validation

- *The Validation Set Approach*





Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Resampling

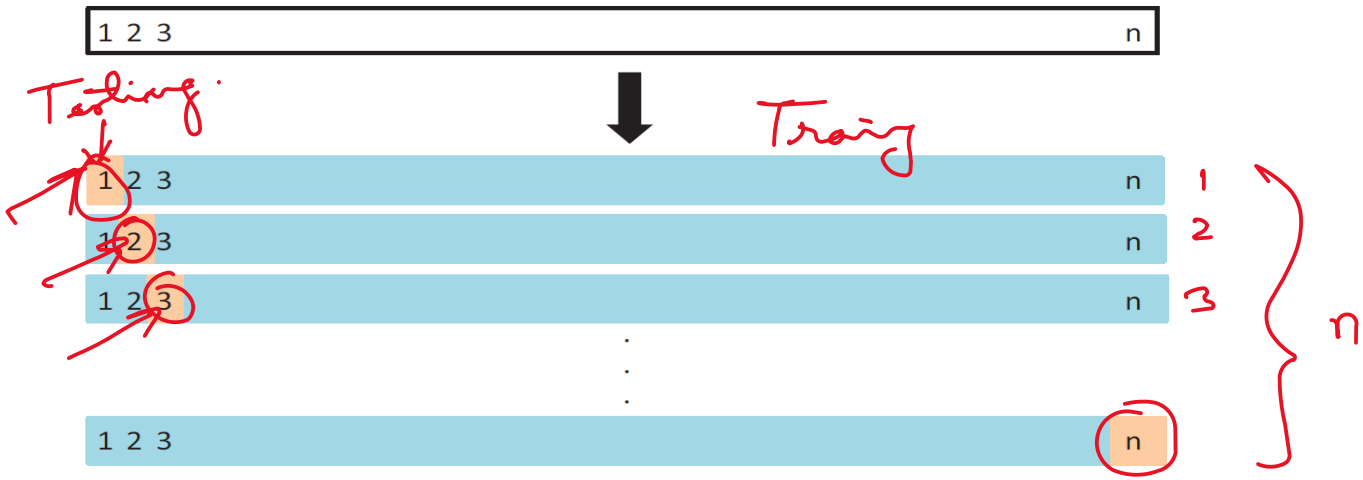
Cross-Validation

- *The Validation Set Approach*
  - Test error rate can be highly variable, depending on which observations are included in the training set and the validation set.
  - In the validation approach, only a subset of the observations are used to fit the model. This is a problem since statistical methods tend to perform worse when trained on fewer observations.

Resampling

Cross-Validation

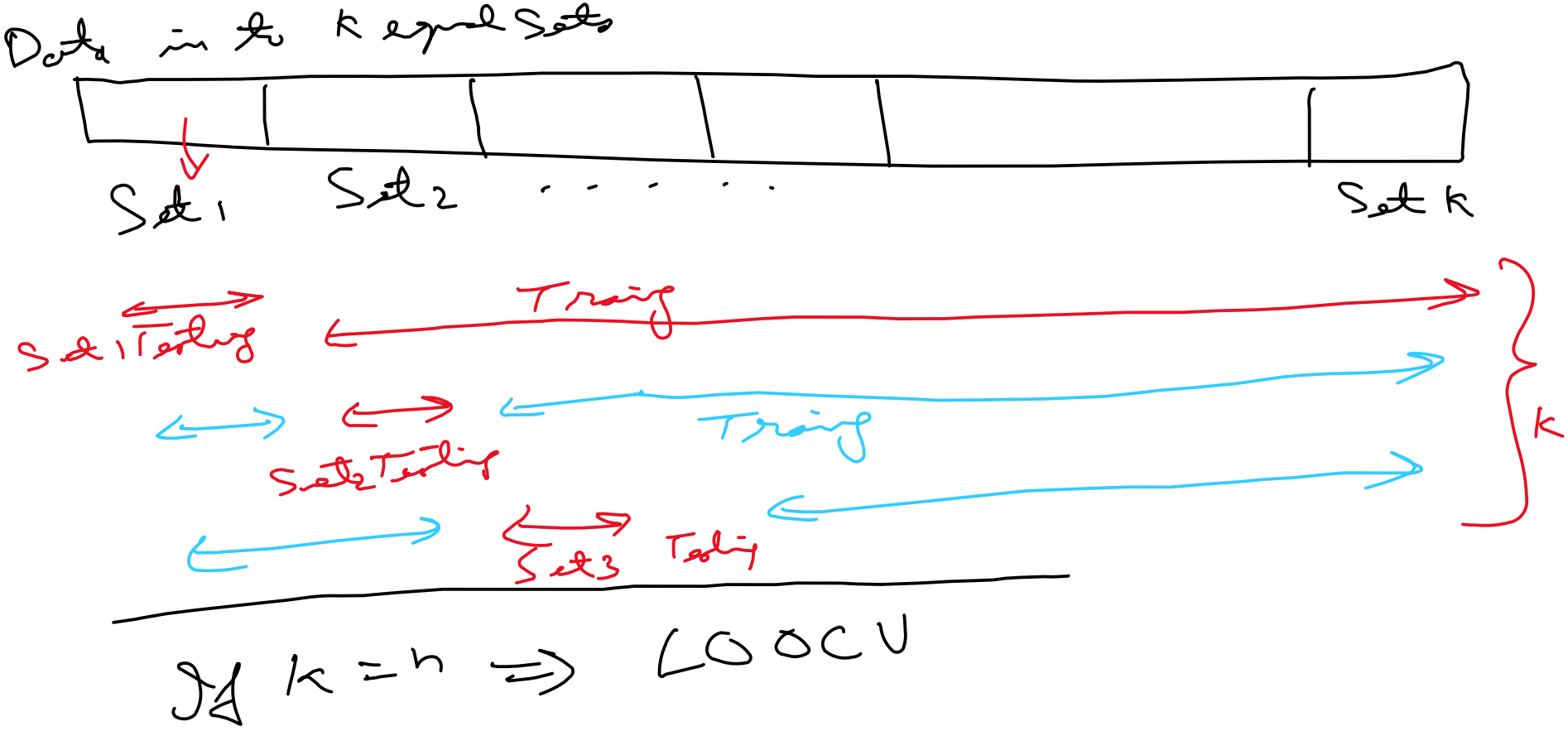
- *Leave-One-Out Cross-Validation*



A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.

Resampling

Cross-Validation



The Bootstrap Method

- Can be used to estimate the standard errors of the coefficients. But not very useful for linear models since the standard errors of the coefficients can be directly estimated.

The Bootstrap Method

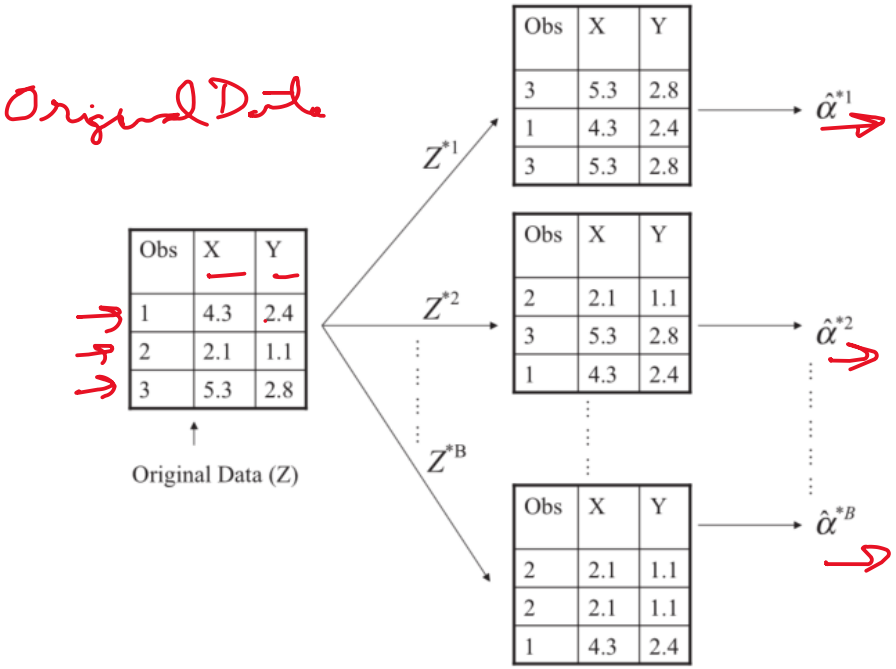
Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities. We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y. Since there is variability associated with the returns on these two assets, we wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $\mathrm{Var}(\alpha X + (1 - \alpha)Y)$. One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

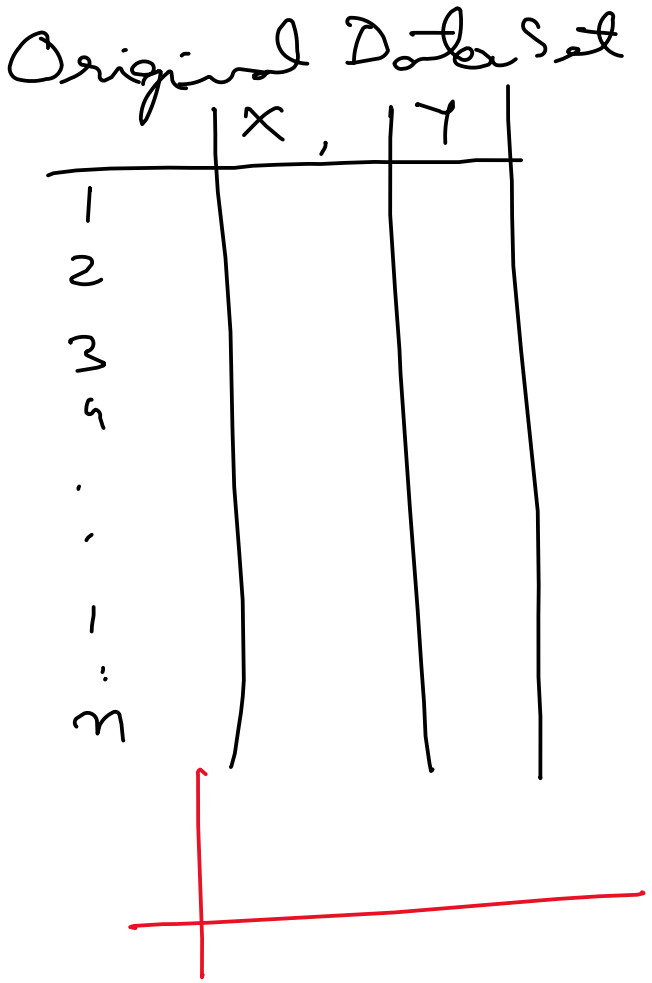
The Bootstrap Method

- The Bootstrap method can be used to estimate the standard errors of the coefficients. But not very useful for linear models since the standard errors of the coefficients can be directly estimated

A graphical illustration of the bootstrap approach on a small sample containing n = 3 observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α



The Bootstrap Method



Randomly choose n data points with replacement from original data

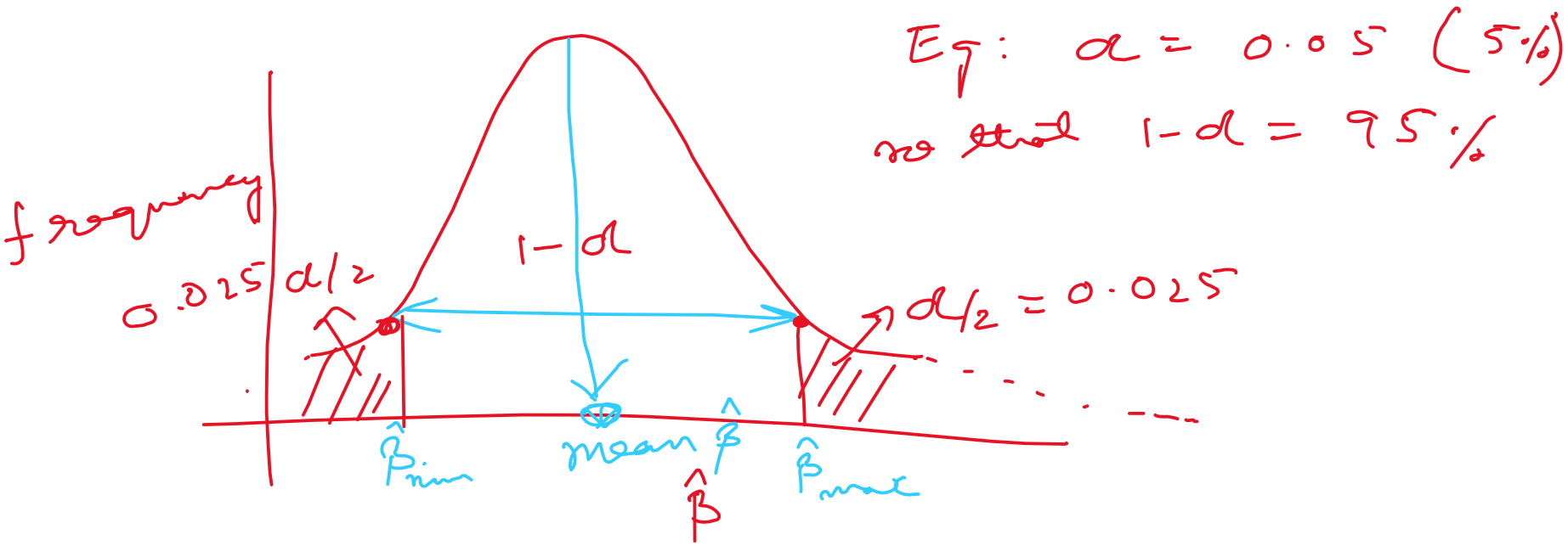
Bootstrap Data set 1 → $(\hat{\beta})_1$

BS Data Set 2 → $(\hat{\beta})_2$

BS Data Set 3 → $(\hat{\beta})_3$

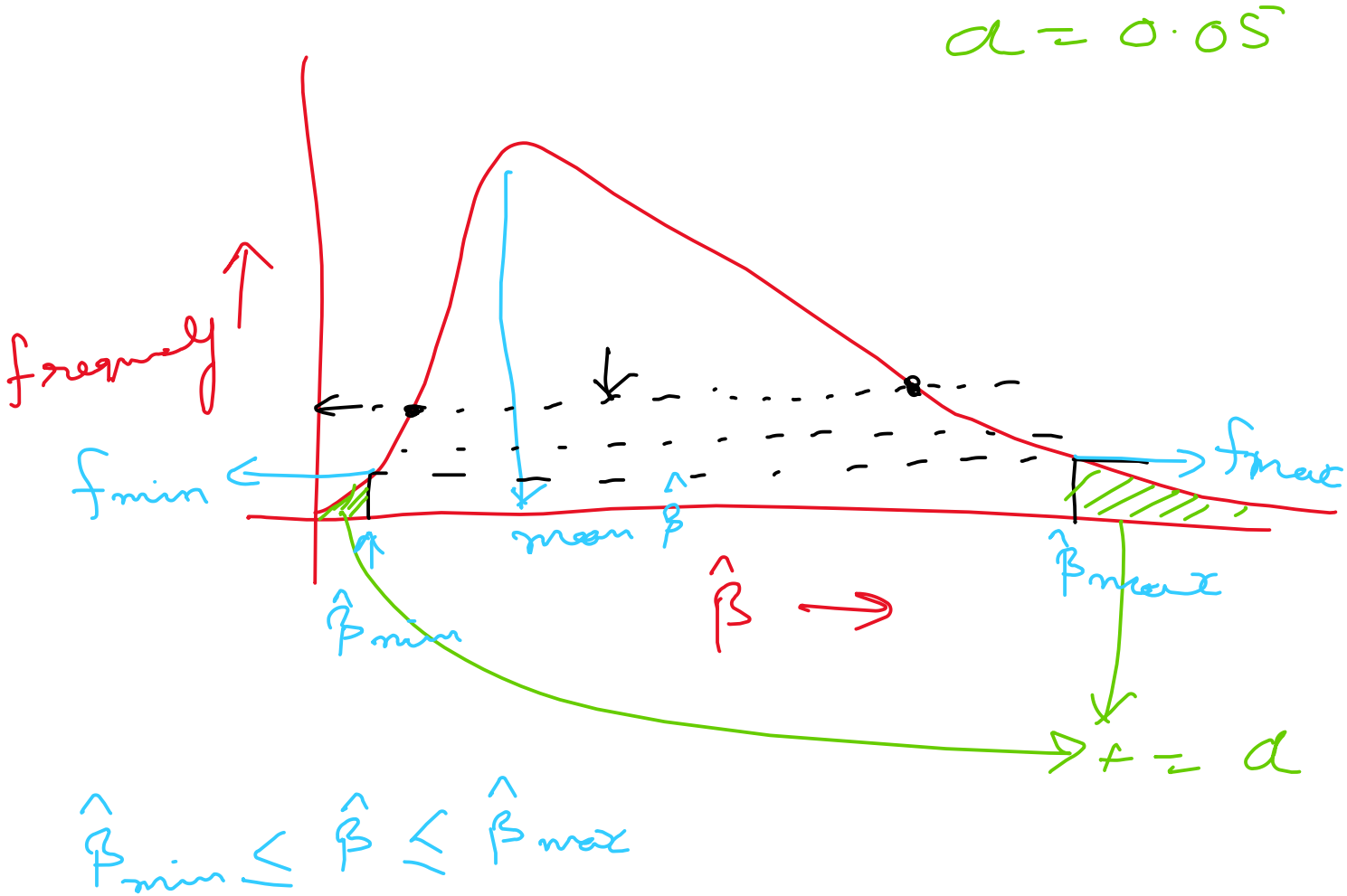
10^4 → $(\hat{\beta})_{10^4}$

The Bootstrap Method



$$\hat{\beta}_{min} \leq \hat{\beta} \leq \hat{\beta}_{max} \quad 95\% \text{ C.I.}$$

The Bootstrap Method



The Bootstrap Method