Frequentist Statistics

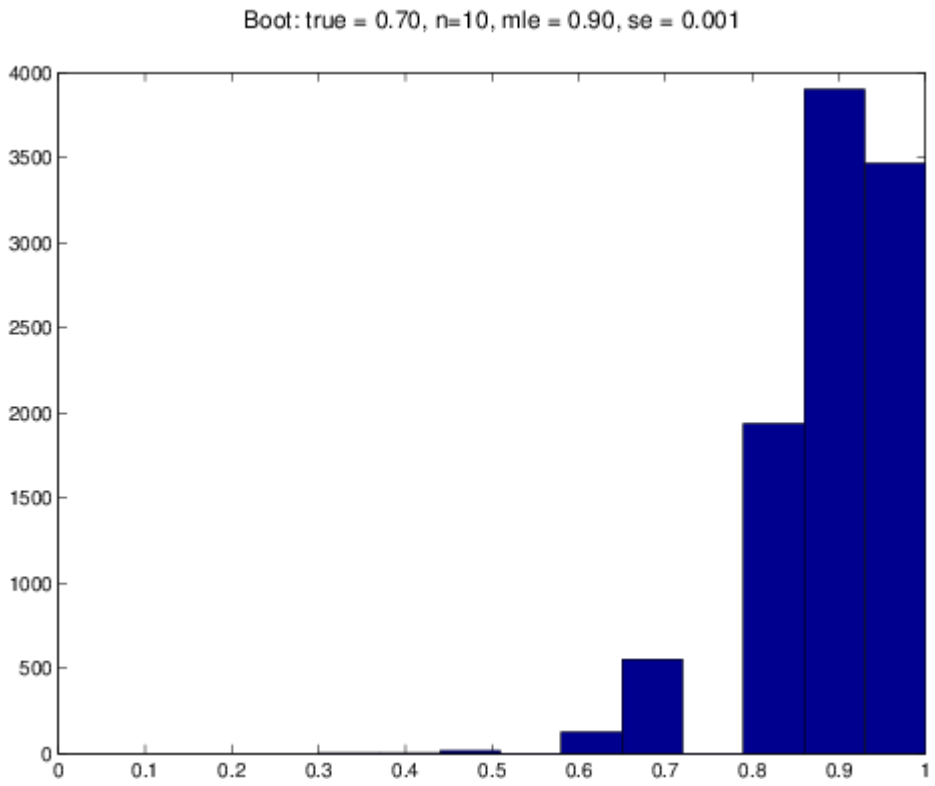
ddebarr@uw.edu

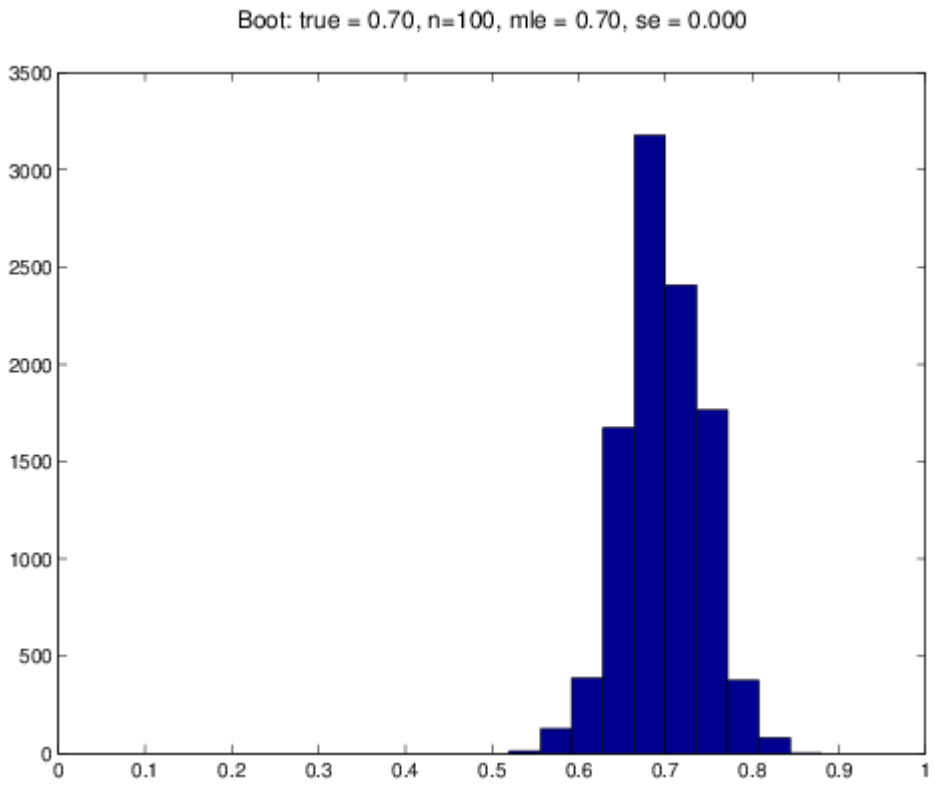
2016-05-12

Agenda

- Sampling Distribution of an Estimator
- Frequentist Decision Theory
- Desirable Properties of Estimators
- Empirical Risk Minimization
- Pathologies of Frequentist Statistics

Bootstrap: 10,000 samples





Bootstrap estimates don’t require significant theory to generate estimates; e.g. confidence intervals

Bootstrap Example

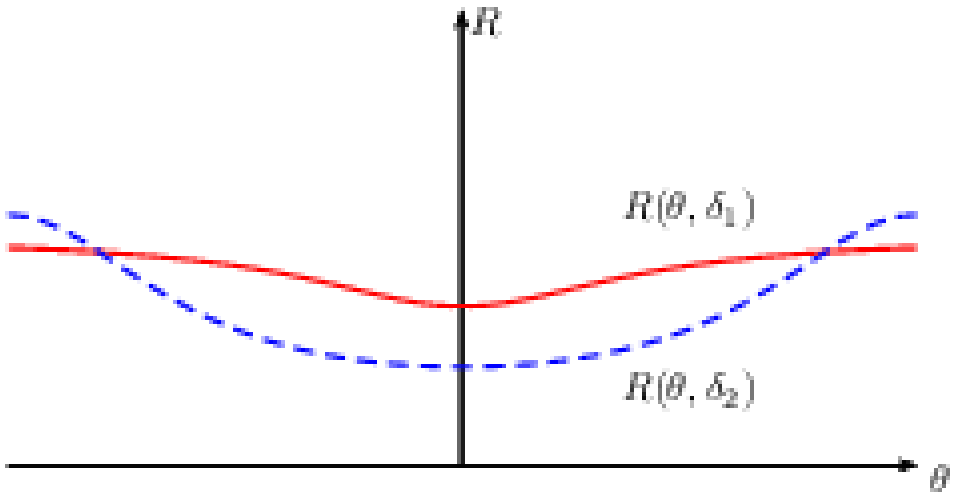
```
> set.seed(123)
> n <- 1000
> x <- rnorm(n)
> y <- 2 * x + 2 * rnorm(n)
>
> # parametric
> Pearson.Correlation.Confidence.Interval <- function(vector1, vector2, confidence = 0.95) {
+     z <- qnorm(1 - (1 - confidence)/2)
+     n <- length(vector1)
+     r <- cov(vector1, vector2) / (sd(vector1) * sd(vector2))
+     return(tanh(atanh(r) + z * c(-1, 0, 1) * sqrt(1 / (n - 3))))
+ }
> Pearson.Correlation.Confidence.Interval(x, y)
[1] 0.7012618 0.7314560 0.7590307
>
> # non-parametric
> Pearson.Correlation <- function(data, selected) {
+     return(cov(data[selected,1], data[selected,2]) / (sd(data[selected,1]) * sd(data[selected,2])))
+ }
> library(boot)
> boot.ci(boot(cbind(x, y), Pearson.Correlation, 1000), type="bca")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot(cbind(x, y), Pearson.Correlation, 1000),
    type = "bca")

Intervals :
Level       BCa
95%   ( 0.7015,  0.7593 )
Calculations and Intervals on Original Scale
```

The parametric estimate is based on the hyperbolic arctangent transform, while the non-parametric bootstrap estimate yields a similar result

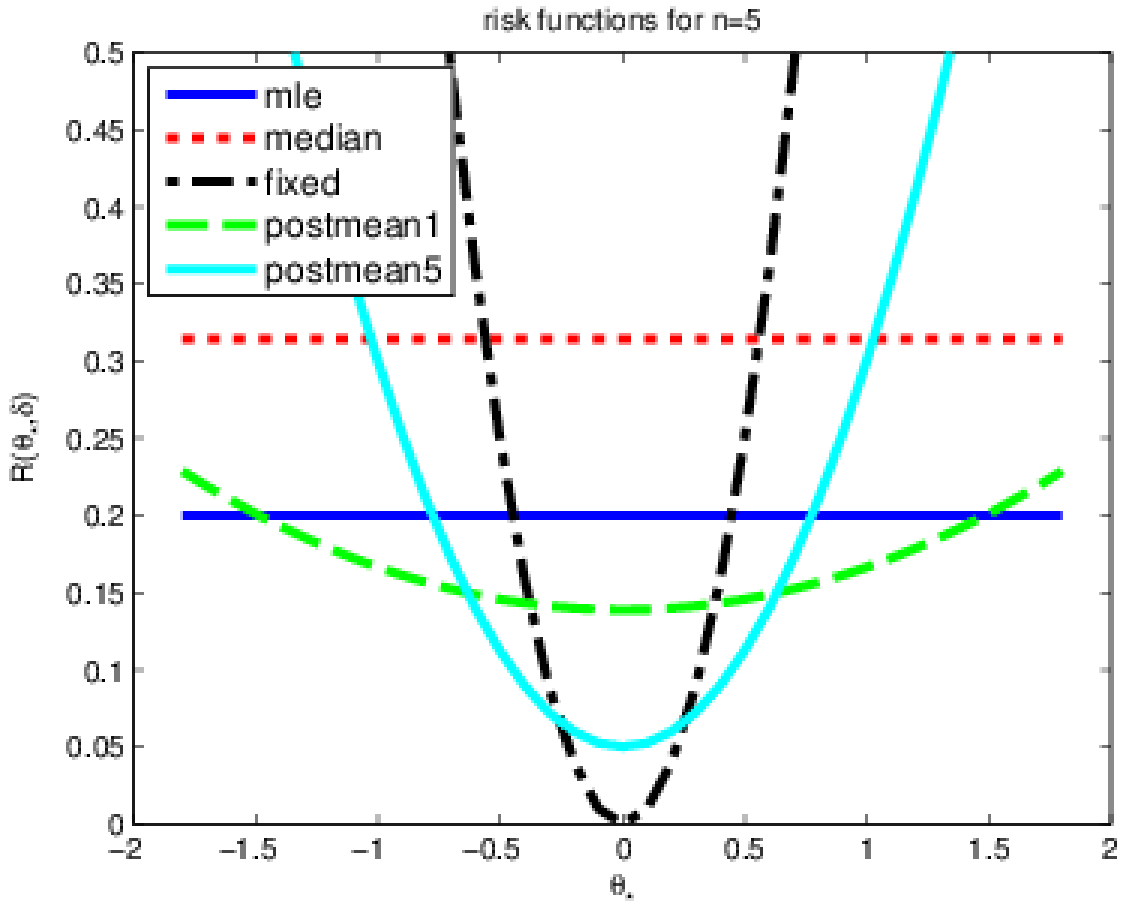
Risk Functions

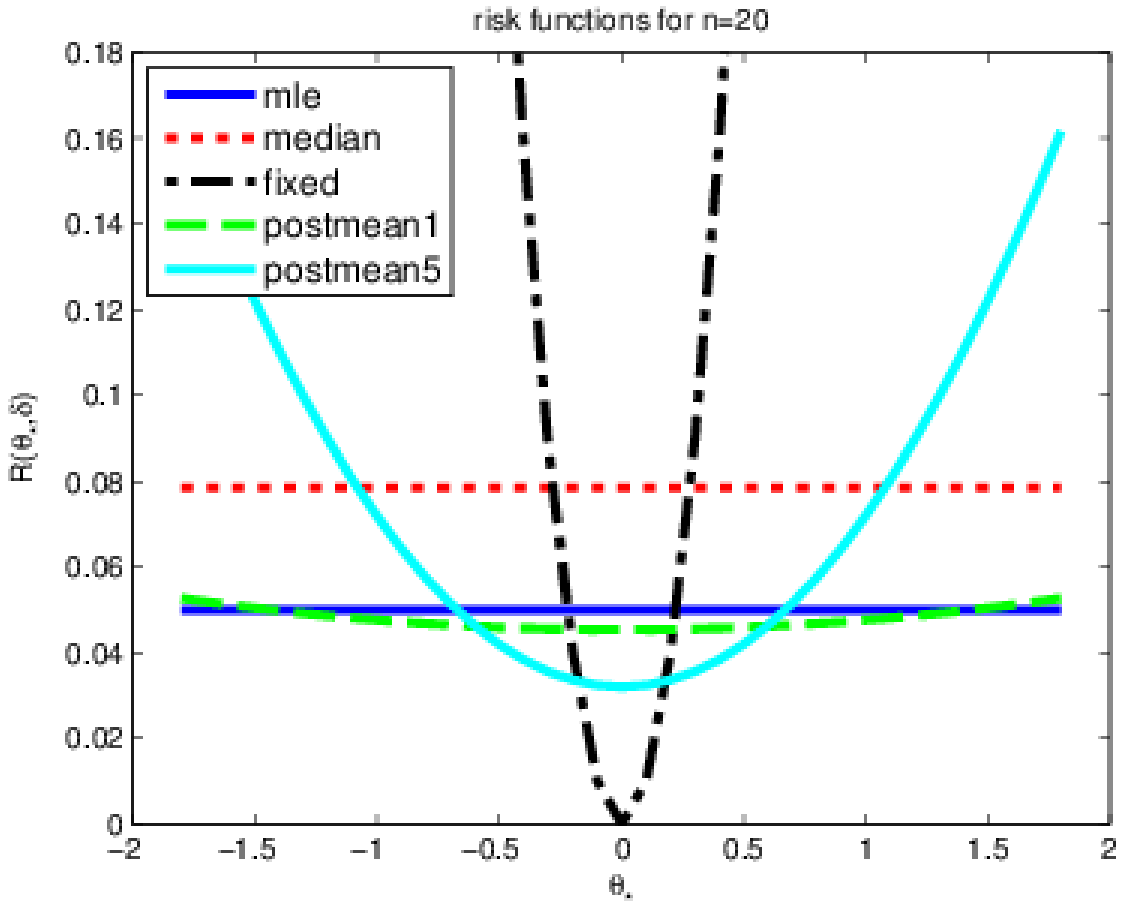


Minimax risk: choose the model that minimizes the maximum risk [the red line above]

An estimator is admissible if it is not strictly dominated by any other estimator

Risk Functions for Estimating a Mean

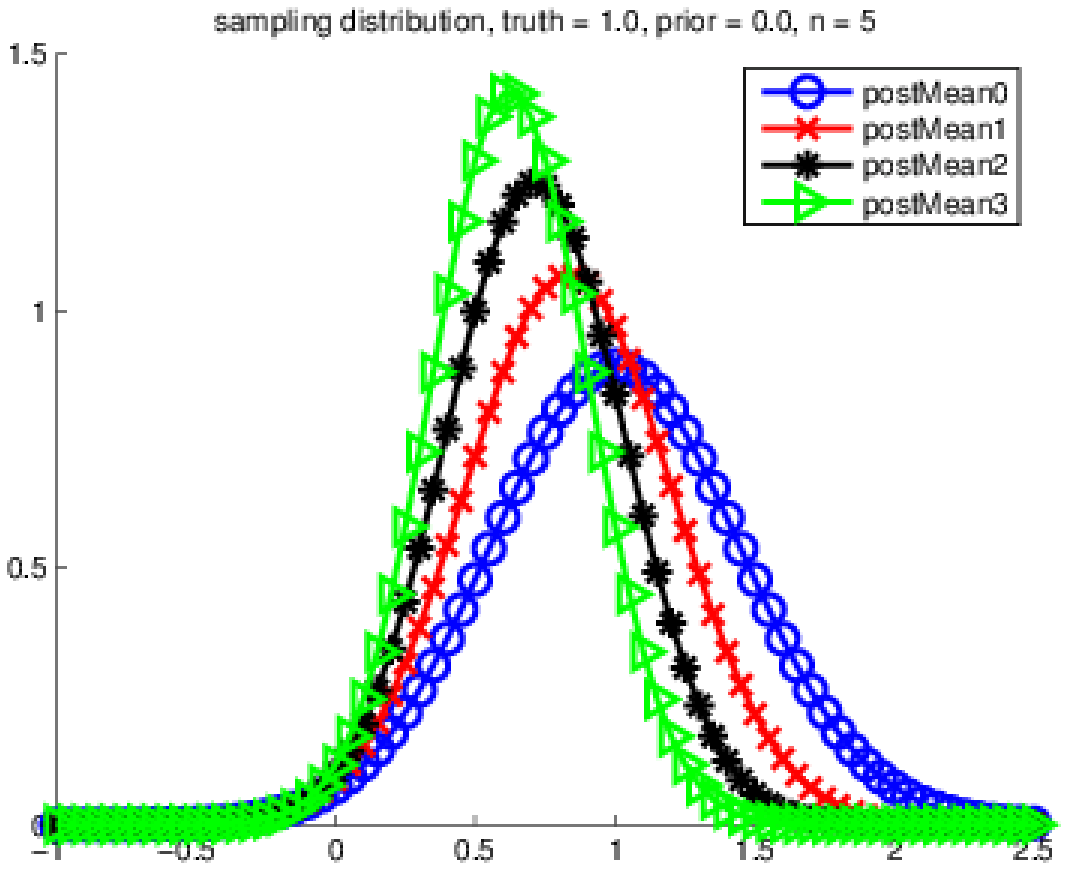


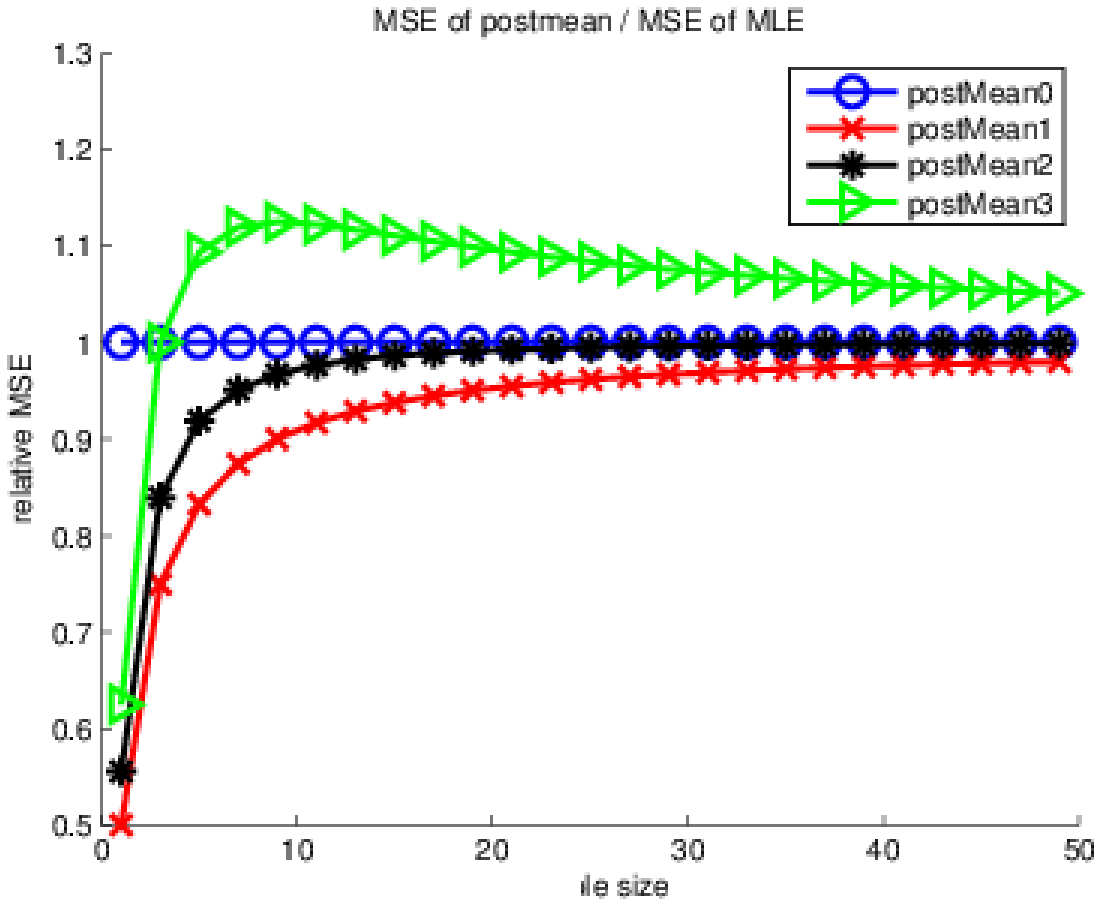


$$\delta_\kappa(\mathbf{x}) = \frac{N}{N+\kappa}\overline{x} + \frac{\kappa}{N+\kappa}\theta_0 = w\overline{x} + (1-w)\theta_0$$

Sampling Distribution of a MAP Estimate

An estimator is consistent if it approaches the true parameter as the sample size goes to infinity





$$\tilde{x} \triangleq \frac{N}{N+\kappa_0}\overline{x} + \frac{\kappa_0}{N+\kappa_0}\theta_0 = w\overline{x} + (1-w)\theta_0$$

Bias-Variance Trade-Off (for Regression)

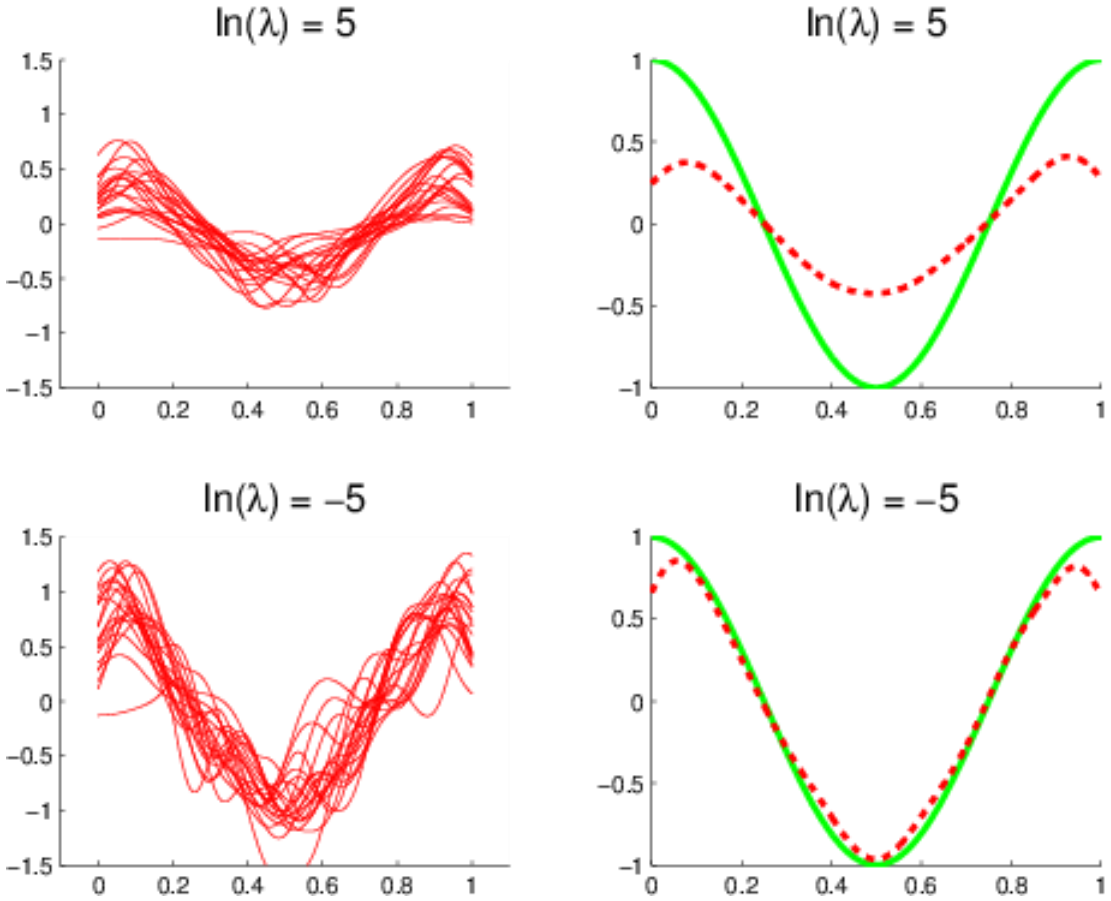
$$\text{MSE} = \text{variance} + \text{bias}^2$$

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta} - \theta^*)^2\right] &= \mathbb{E}\left[\left[(\hat{\theta} - \overline{\theta}) + (\overline{\theta} - \theta^*)\right]^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta} - \overline{\theta}\right)^2\right] + 2(\overline{\theta} - \theta^*)\mathbb{E}\left[\hat{\theta} - \overline{\theta}\right] + (\overline{\theta} - \theta^*)^2 \\
&= \mathbb{E}\left[\left(\hat{\theta} - \overline{\theta}\right)^2\right] + (\overline{\theta} - \theta^*)^2 \\
&= \text{var}\left[\hat{\theta}\right] + \text{bias}^2(\hat{\theta})
\end{aligned}$$

Bias-Variance Trade-Off for Ridge Regression

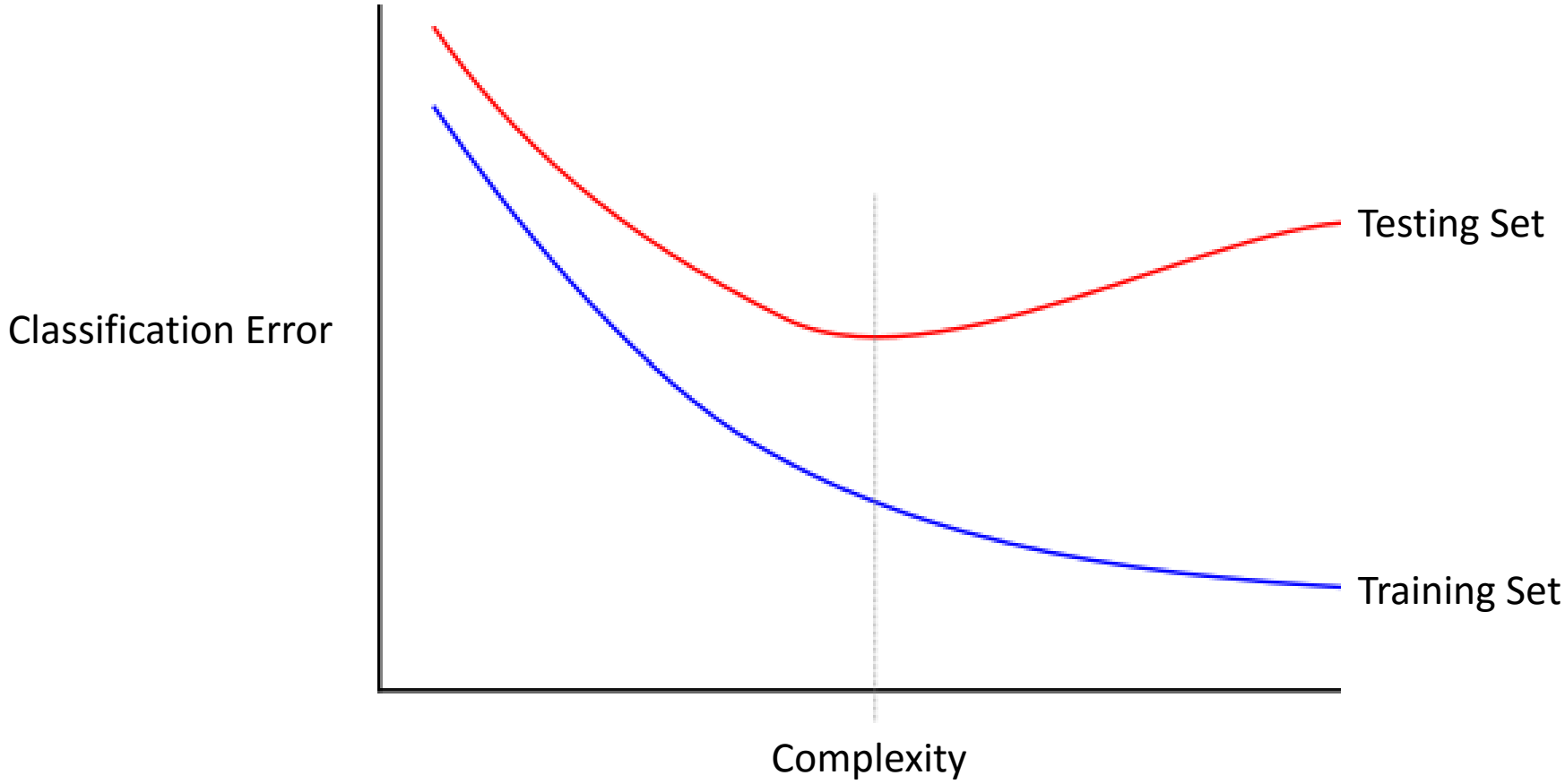
High Bias

High Variance



True Function: green
Averaged Estimates: red

Bias-Variance Trade-off (for Classification)



On the left side of the graph, we say the model exhibits high-bias (underfitting). This is often associated with low variance. For example a majority classifier assigns the same class to observations, regardless of how the features change.
On the right side of the graph, we say the model exhibits high-variance (overfitting). This is often associated with low bias. In the extreme, the model responds to the smallest of changes in the data because it has memorized the training data.

Risk Minimization

- Empirical Risk

$$R_{emp}(\mathcal{D}, \delta) \triangleq R(p_{\text{emp}}(\cdot|\mathcal{D}), \delta) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \delta(\mathbf{x}_i))$$

- Structural Risk

$$\hat{\delta}_\lambda = \underset{\delta}{\text{argmin}} \left[R_{emp}(\mathcal{D}, \delta) + \lambda C(\delta) \right]$$

- Upper Bound from Statistical Learning Theory Bound

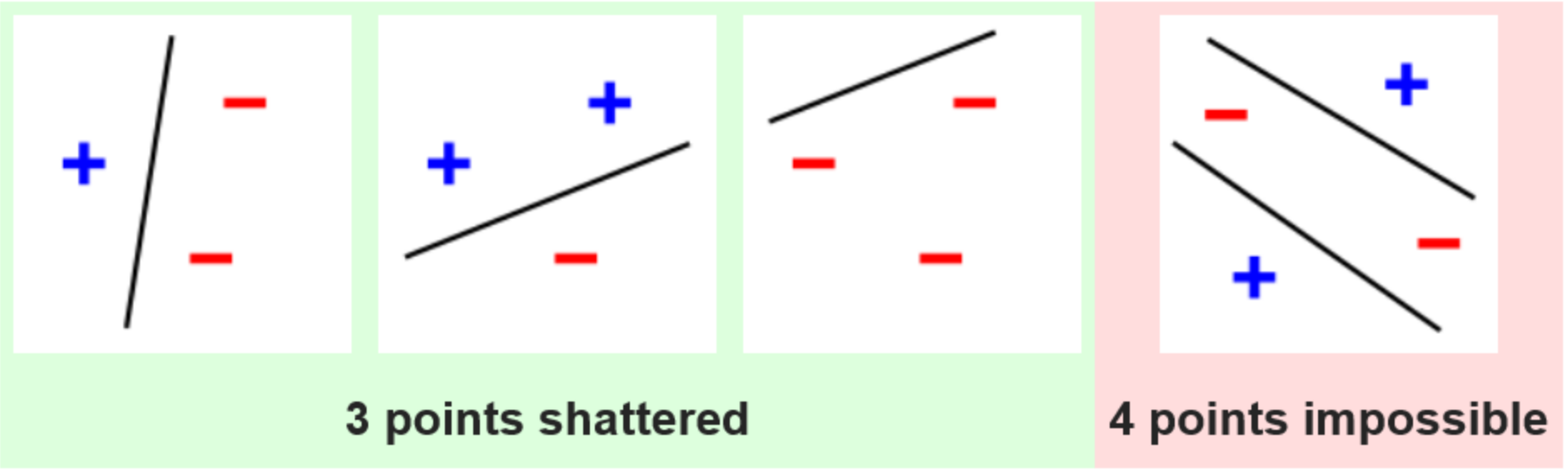
$$P\left(\max_{h \in \mathcal{H}} |R_{emp}(\mathcal{D}, h) - R(p_*, h)| > \epsilon \right) \leq 2 \dim(\mathcal{H}) e^{-2N\epsilon^2}$$

Never report the risk value associated with the training data.
Evaluation of the training data is for detecting problems, not estimating generalization performance.

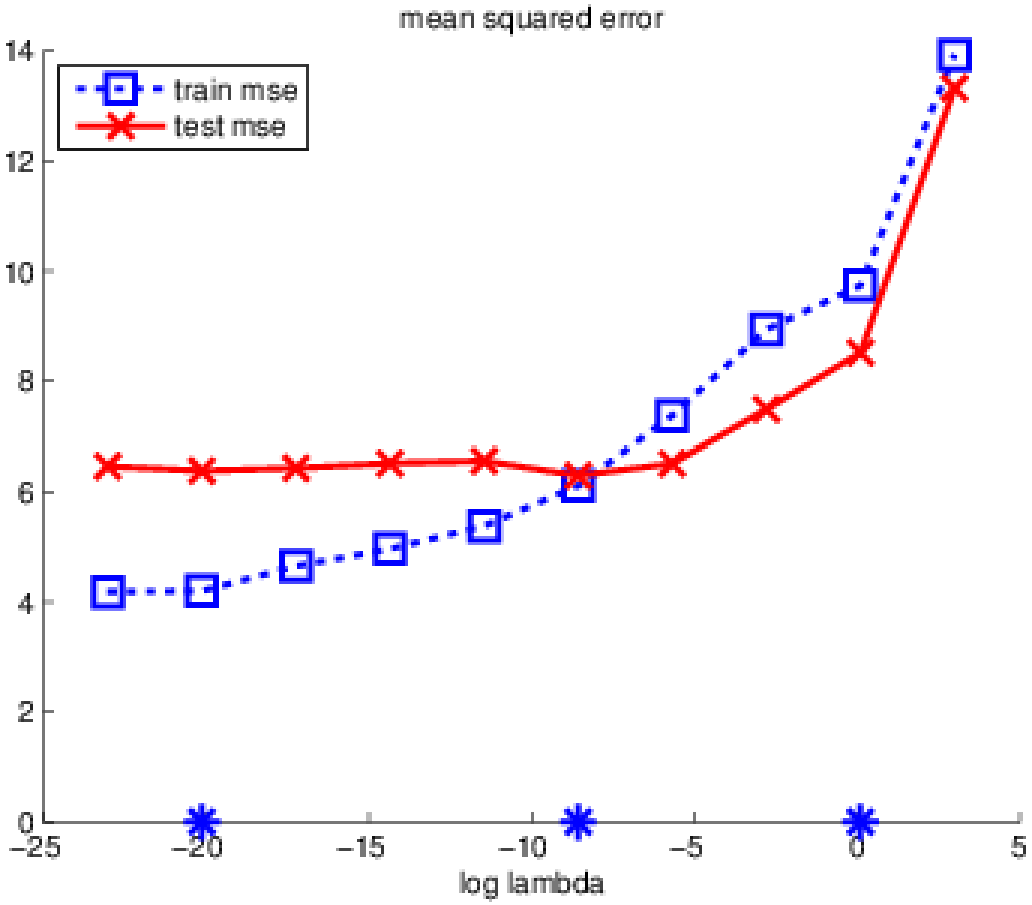
Strucutural Risk Minimization

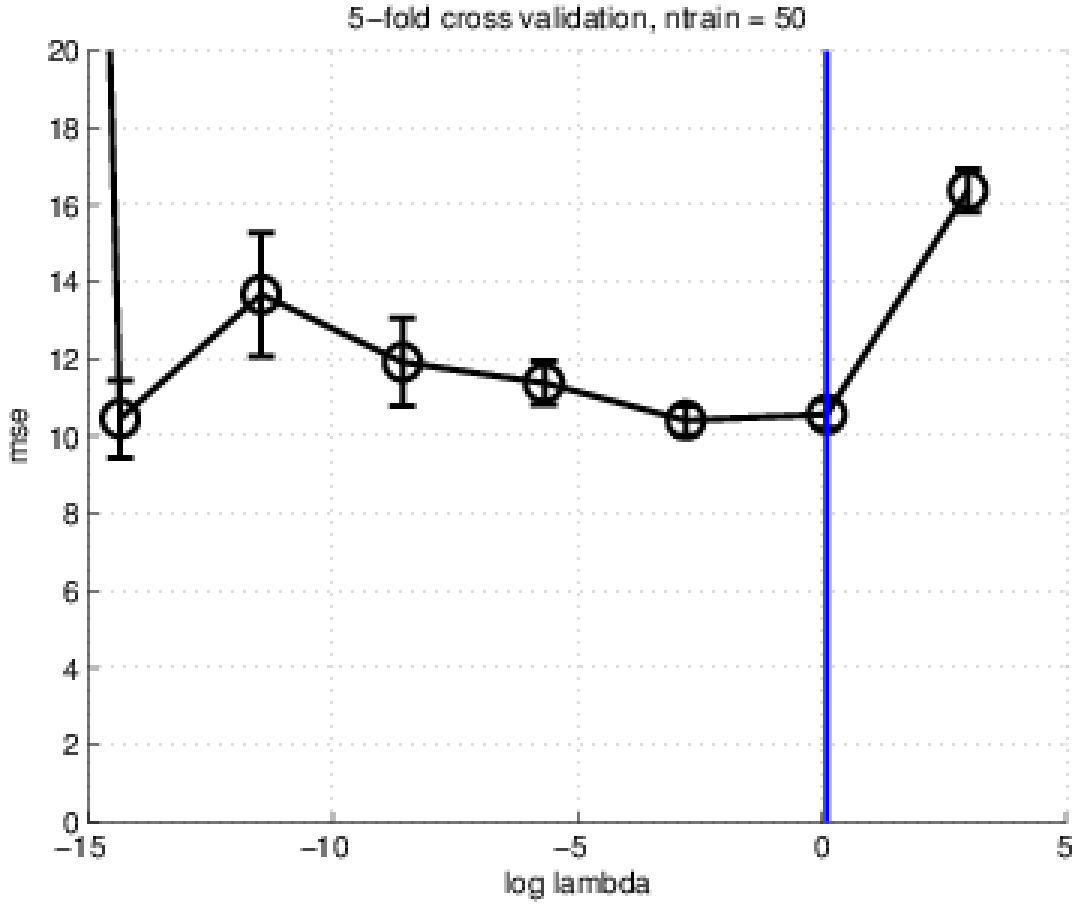
Vapnik Chervonenkis (VC) Dimensionality

- Measures the complexity of a classifier
- Number of non-collinear data points that can be “shattered” by a classifier
- A linear classifier in ‘n’ dimensions can shatter ‘n’ + 1 data points



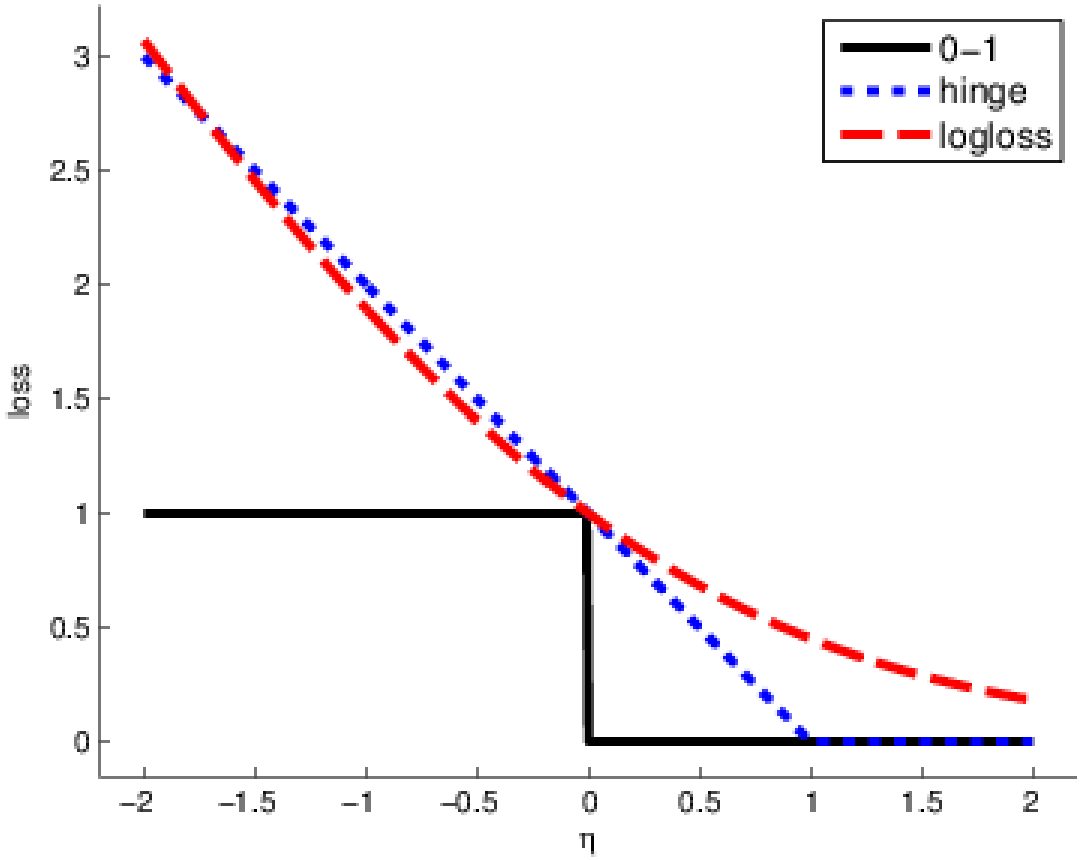
MSE for Regularized Polynomial Regression





If you have enough data, hold-out validation can be preferable.
Of course, cross validation can be viewed as repeated hold-out validation.

Loss Functions for Binary Classification



$$L_{01}(y, \eta) = \mathbb{I}(y \neq \hat{y}) = \mathbb{I}(y\eta < 0)$$

$$L_{\text{hinge}}(y, \eta) = \max(0, 1 - y\eta)$$

$$L_{\text{nll}}(y, \eta) = -\log p(y|\mathbf{x}, \mathbf{w}) = \log(1 + e^{-y\eta})$$

$$-log\big(probability(y_i|\,\boldsymbol{x_i}; \boldsymbol{w})\big)$$

$$= -log\left(\left(\frac{1}{1 + exp(-\boldsymbol{w}^t \boldsymbol{x_i})}\right)^{(y_i^*)} \left(1 - \frac{1}{1 + exp(-\boldsymbol{w}^t \boldsymbol{x_i})}\right)^{(1-y_i^*)}\right)$$

$$= -log\left(\frac{1}{1+exp(-y_i\, \boldsymbol{w}^t\, \boldsymbol{x_i})}\right) = log\big(1 + exp(-y_i\, \boldsymbol{w}^t\, \boldsymbol{x_i})\big).$$

Example: The Wald Confidence Interval

- The 95% Wald confidence interval:

$$\overline{x} \pm 1.96\sqrt{\overline{x}(1-\overline{x})/N}$$

- Suppose we are estimating classification error for a small test set, and we misclassify 1 out of 100 observations. The Wald confidence interval will include negative values!

```
> .01 + c(-1, 1) * qnorm(1-.05/2) * sqrt(.01 * .99 / 100)
[1] -0.009501  0.029501
```

- Always think about the models you are using and be on the lookout for pathological results [often times there are heuristics for avoiding behavior]

Rebuttal from a Frequentist

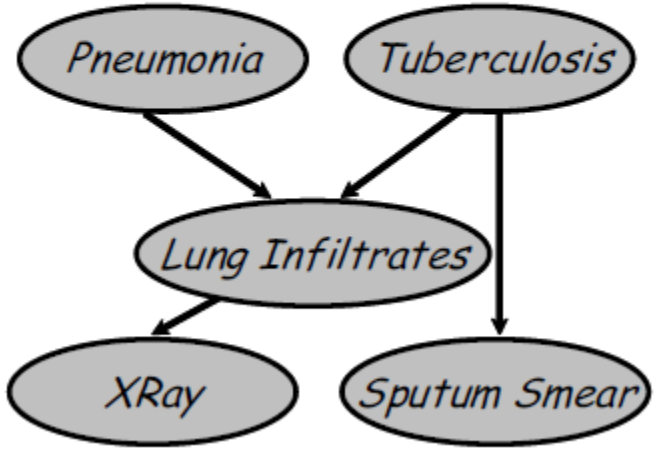
- From the paper by Bradley Efron [the guy who brought you bootstrap resampling] …

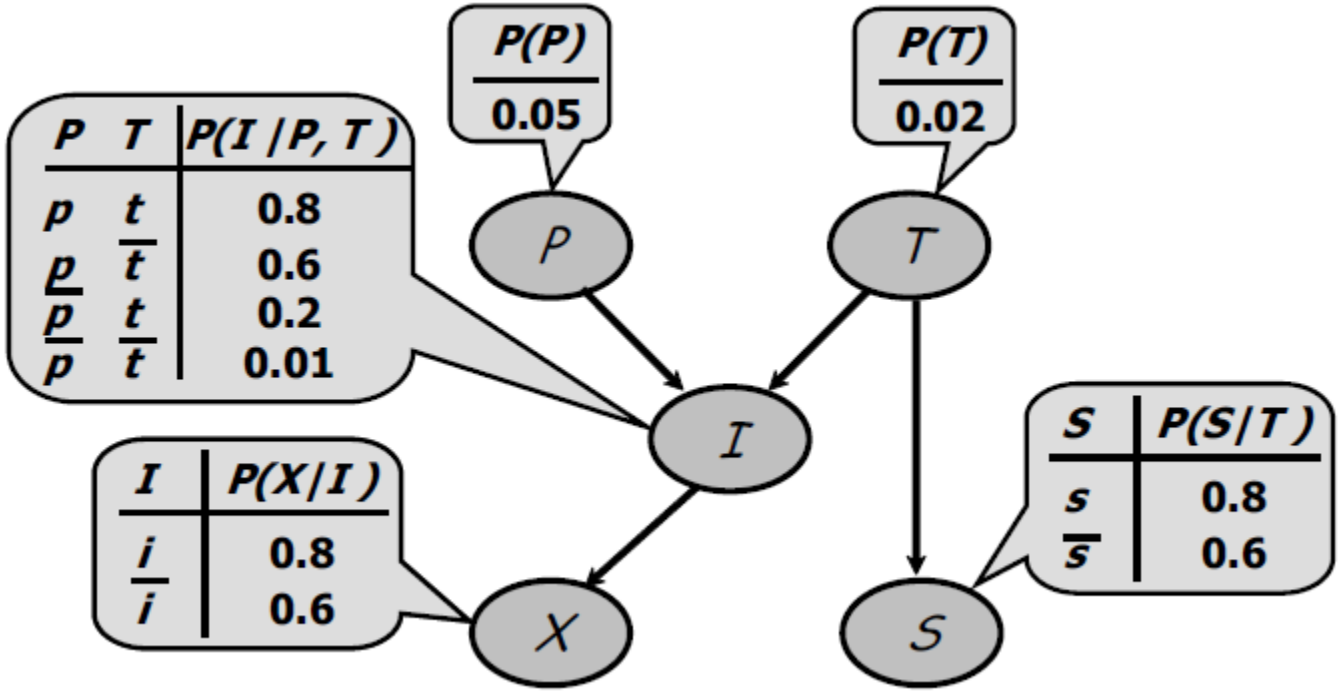
A summary of the major reasons why Fisherian and NPW ideas have shouldered Bayesian theory aside in statistical practice is as follows:

1. Ease of use: Fisher's theory in particular is well set up to yield answers on an easy and almost automatic basis.
2. Model building: Both Fisherian and NPW theory pay more attention to the preinferential aspects of statistics.
3. Division of labor: The NPW school in particular allows interesting parts of a complicated problem to be broken off and solved separately. These partial solutions often make use of aspects of the situation, for example, the sampling plan, which do not seem to help the Bayesian.
4. Objectivity: The high ground of scientific objectivity has been seized by the frequentists.

Graphical Models

Graphical Model Example





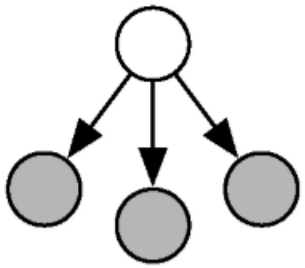
$$P(P,T,I,X,S) = P(P)P(T)P(I \mid P,T)P(X \mid I)P(S \mid T)$$

Example: P(P=0,T=0,I=0,X=1,S=1) = 0.95*0.98*0.99*0.6*0.6 = 0.3318

Sequence Prediction

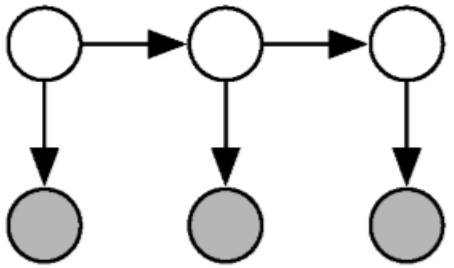
$$p(y, \mathbf{x}) = p(y) \prod_{k=1}^{K} p(x_k|y)$$

generative models
on top



Naive Bayes

SEQUENCE



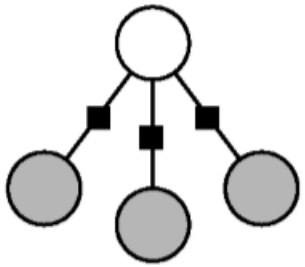
HMMs

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^{T} p(y_t|y_{t-1}) p(x_t|y_t)$$

CONDITIONAL

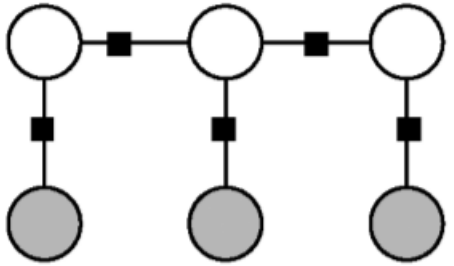
CONDITIONAL

discriminative models
on bottom



Logistic Regression

SEQUENCE



Linear-chain CRFs

References

- Graphical Models in a Nutshell
  - http://www.cs.umd.edu/srl-book/
  - http://ai.stanford.edu/~koller/Papers/Koller+al:SRL07.pdf
- Conditional Random Fields
  - http://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf
  - http://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers