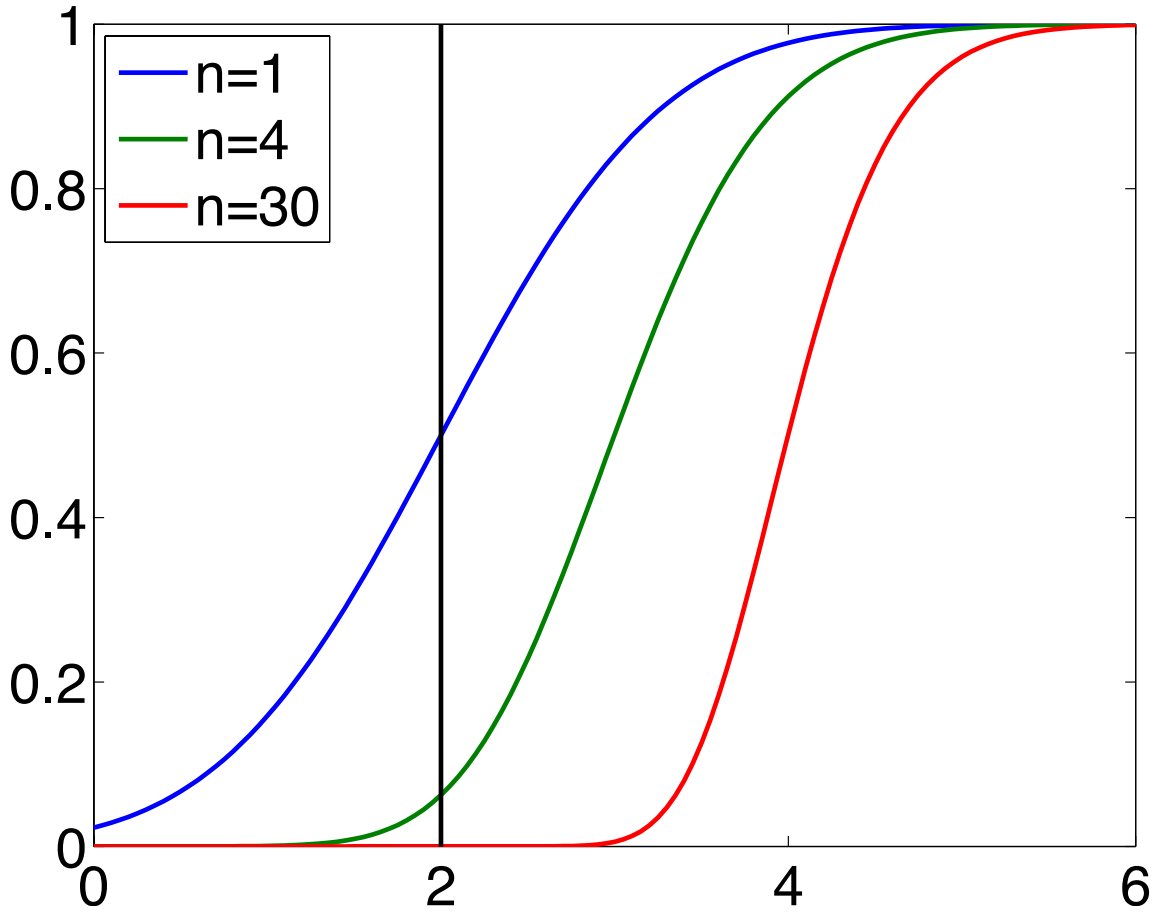
Accuracy & confidence

- Most of course so far: estimating stuff from data
- Today: how much do we trust our estimates?
- Last week: one answer to this question
  - prove ahead of time that training set estimate of prediction error will have accuracy ϵ w/ probability $1-\delta$
  - had to handle two issues:
    - limited data $\Rightarrow$ can't get exact error of single model
    - selection bias $\Rightarrow$ we pick "lucky" model r.t. right one

Selection bias

CDF of max of n samples of $N(\mu=2, \sigma^2=1)$

[representing error estimates for n models]



Overfitting

- Overfitting = selection bias when fitting complex models to little/noisy data
  - to limit overfitting: limit noise in data, get more data, simplify model class
- Today: not trying to limit overfitting
  - instead, try to *evaluate* accuracy of selected model (and recursively, accuracy of our accuracy estimate)
  - can lead to *detection* of overfitting

What is accuracy?

- Simple problem: estimate μ and σ^2 for a Gaussian from samples $x_1, x_2, \ldots x_N \sim \text{Normal}(\mu, \sigma^2)$

estimate → μ, σ^2 ← known

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad E(\bar{x}) = \frac{1}{N}\sum_i E(x_i) = \frac{1}{N} N\mu = \mu$$

bias = E(statistic) − parameter → unbiased

$$V(\bar{x}) = E\left[(\bar{x} - E(\bar{x}))^2\right] = E\left[(\bar{x} - \mu)^2\right] = E\left[\left(\frac{1}{N}\sum_i x_i - \mu\right)^2\right]$$

↳ $\mu = 0$

$$= \frac{1}{N^2} E\left(\sum_i \sum_j x_i x_j\right) = \frac{1}{N^2}\sum_i E(x_i^2) = \frac{1}{N^2} N\sigma^2 = \sigma^2/N$$

Bias vs. variance vs. residual

- Mean squared prediction error: predict x_{N+1}
  - $E\left[\left(\bar{x} - x_{N+1}\right)^2\right] = E\left[\left((\bar{x} - \mu) - (x_{N+1} - \mu)\right)^2\right]$

$= E\left[(\bar{x} - \mu)^2 - 2(\bar{x} - \mu)(x_{N+1} - \mu) + (x_{N+1} - \mu)^2\right]$

$(\bar{x} - \mu)^2$: σ^2/N → estimation error

$(x_{N+1} - \mu)^2$: σ^2 → residual

$= E\left[\left((\bar{x} - E(\bar{x})) - (\mu - E[\bar{x}])\right)^2\right] + \sigma^2$

$= E\left[(\bar{x} - E(\bar{x}))^2\right] - 2E\left[(\bar{x} - E(\bar{x}))(\mu - E(\bar{x}))\right] + E\left[(\mu - E(\bar{x}))^2\right] + \sigma^2$

$V(\bar{x})$ variance

$E\left[(\mu - E(\bar{x}))^2\right]$: bias^2

σ^2 → residual

Bias-variance tradeoff

- Can't do much about residual, so we're mostly concerned w/ estimation error = bias^2 + variance
- Can trade bias v. variance to some extent: e.g., always estimate 0 ⇒ variance=0, but bias big
- Cramér-Rao bound on estimation error:

$$E[\hat{\theta} - \theta] = b(\theta) \Rightarrow E[(\hat{\theta} - \theta)^2] \geq b(\theta)^2 + \frac{(1 + b'(\theta))^2}{I(\theta)}$$

$$b(\theta) = 0 \Rightarrow \geq 1/I(\theta)$$

$O(N)$

Fisher info

Prediction error v. estimation error

- Several ways to get at accuracy
  - prediction error (bias^2 + var + residual^2)
    - talks only about *predictions*
  - estimation error (bias^2 + var)
    - same; tries to concentrate on error due to estimation
  - parameter error $E((\mu - \hat{\mu})^2)$
    - talks about parameters r.t. predictions
    - in simple case, numerically equal to estimation error
    - but only makes sense if our model class is right

Evaluating accuracy

- In $N(\mu, \sigma^2)$ example, we were able to derive bias, variance, and residual from first principles
- In general, have to estimate prediction error, estimation error, or model error from data
- Holdout data, tail bounds, normal theory (use CLT & tables of normal dist'n), and today's topics: crossvalidation & bootstrap

Goal: estimate sampling variability

- We've computed something from our sample
  - classification error rate, a parameter vector, mean squared prediction error, …
  - for simplicity, a single number (e.g., i^{th} component of weight vector)
  - $t = f(x_1, x_2, \ldots, x_N)$
- How much would t vary if we had taken a different sample?
- For concreteness: f = sample mean (an estimate of population mean)

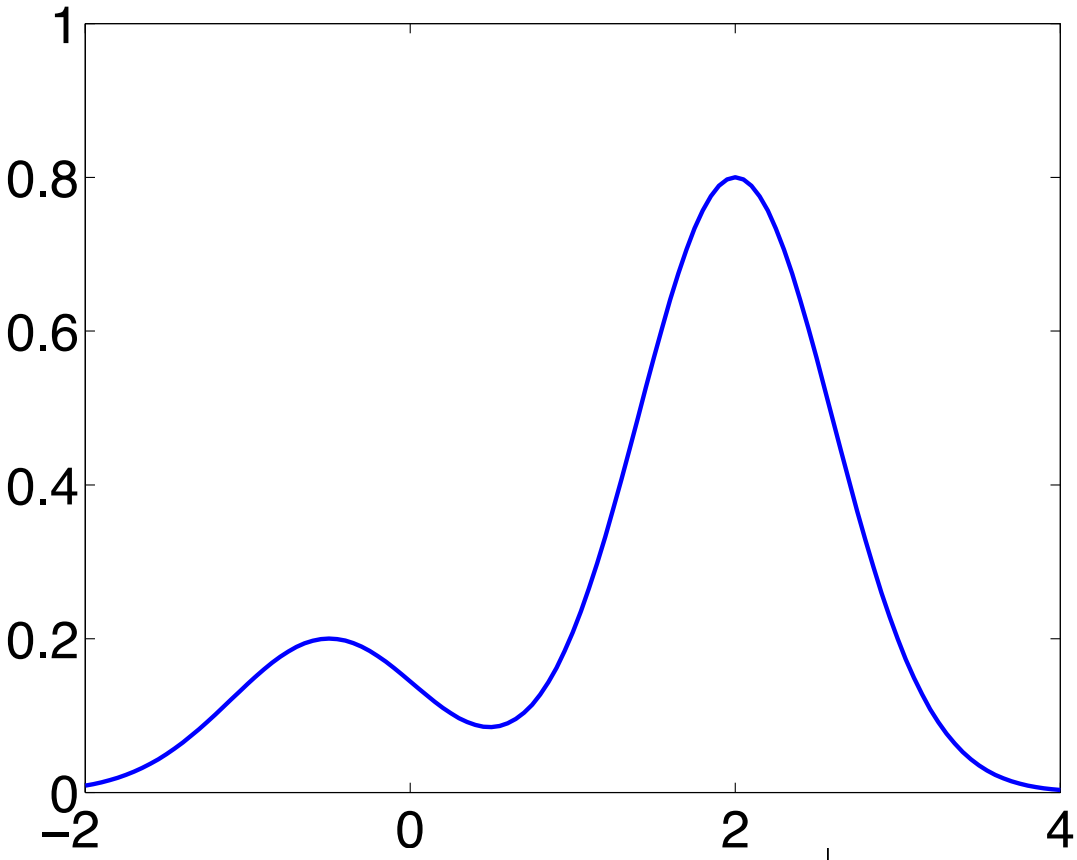
Gold standard: new samples

- Get M independent data sets $x_1^j \dots x_N^j \quad j = 1 \dots M$
- Run our computation M times: $t_1, t_2, \dots t_M$
  - $t_j = f(x_1^j \dots x_N^j)$
- Look at distribution of t_j
  - mean, variance, upper and lower 2.5% quantiles, …
- A tad wasteful of data…

Crossvalidation & bootstrap

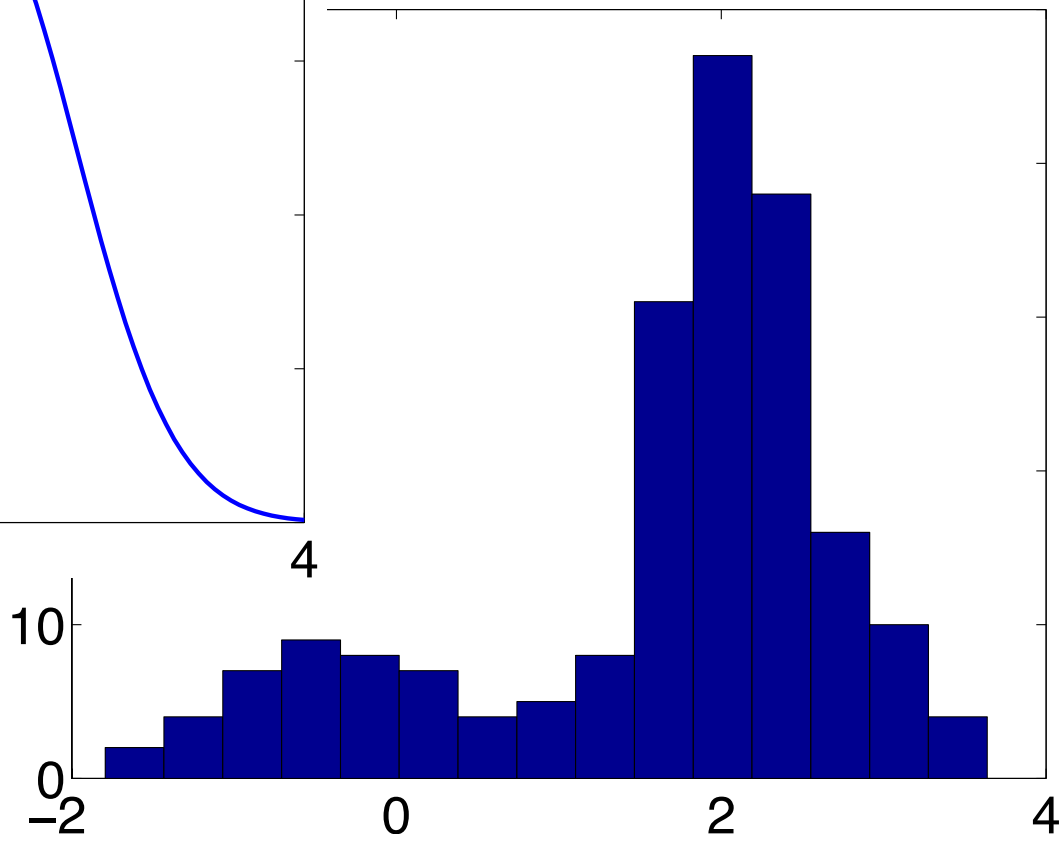
- CV and bootstrap: approximate the gold standard, but cheaper—spend computation instead of data
- Work for nearly arbitrarily complicated models
- Typically tighter than tail bounds, but involve difficult-to-verify approximations/assumptions
- Basic idea: surrogate samples
  - Rearrange/modify $x_1, \ldots, x_N$ to build each "new" sample
- Getting something from nothing? (hence name)

For example



μ=1.5

μ̂=1.6136



Basic bootstrap

- Treat $x_1 \ldots x_N$ as our estimate of true distribution
- To get a new sample, draw N times from this estimate (with replacement)
- Do this M times
  - each original x_i part of many samples (on average $1-1/e$ of them, about 63%)
  - each sample contains many repeated values (single x_i selected multiple times)

Basic bootstrap

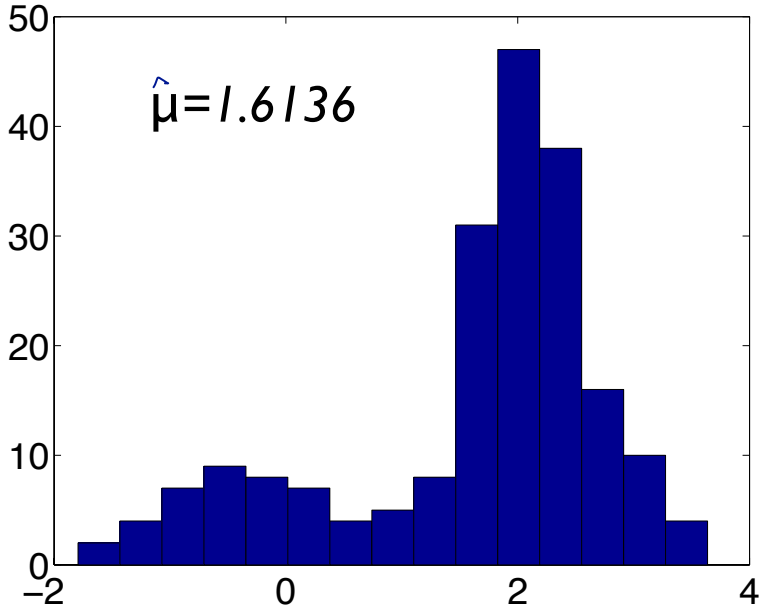
100k bootstrap resamples

$\sigma(\hat{\mu}) \approx .0818$

true $\sigma(\hat{\mu}) = .0825$

10 resamples

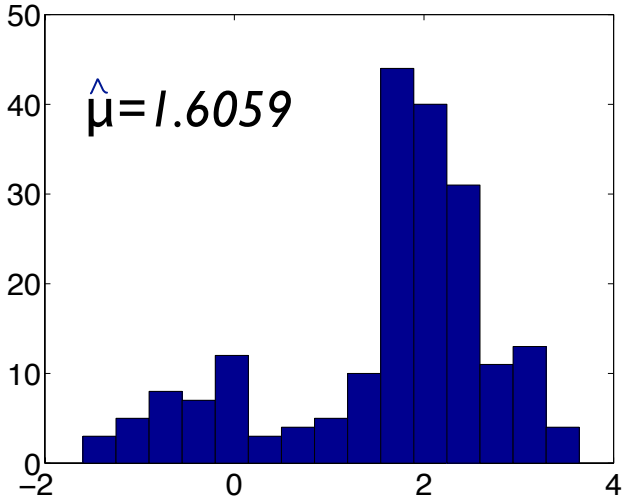
$\sigma(\hat{\mu}) \approx .07$ .09 .04

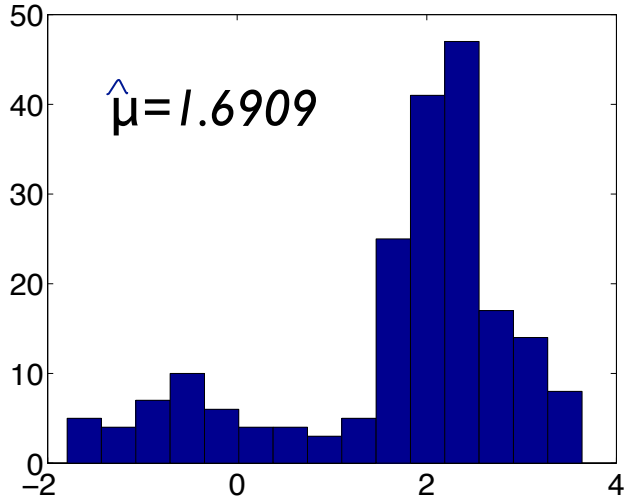


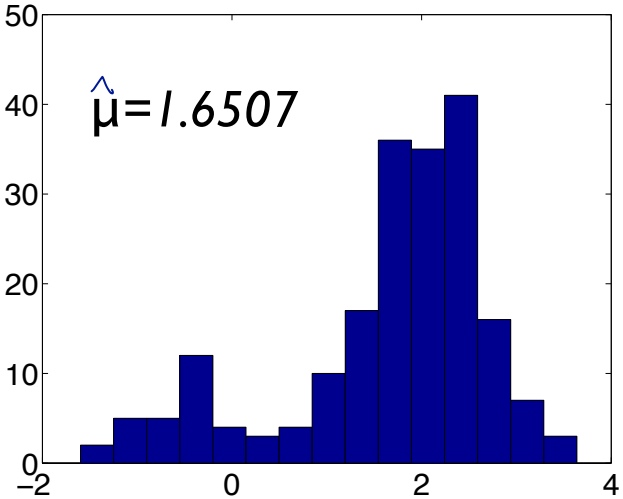
← original

resamples

↓







What can go wrong?

- Convergence is only asymptotic (large *original* sample)
  - here: what if original sample hits mostly the larger mode?

- Original sample might not be i.i.d.
  - unmeasured covariate

Types of errors

- "Conservative" estimate of uncertainty: tends to be high (too uncertain)
- "Optimistic" estimate of uncertainty: tends to be low (too certain)

Should we worry?

- New drug: mean outcome 1.327 [higher is better]
  - old one: outcome 1.242
- Bootstrap underestimates $\sigma = .04$
  - true $\sigma = .08$
- Tell investors: new drug better than old one
- Enter Phase III trials—cost $millions
- Whoops, it isn't better after all…

Blocked resampling

- Partial fix for one issue (original sample not i.i.d.)
- Divide sample into blocks that tend to share the unmeasured covariates, and resample blocks
  - e.g., time series: break up into blocks of adjacent times
    - assumes unmeasured covariates change slowly
  - e.g., matrix: break up by rows or columns
    - assumes unmeasured covariates are associated with rows or columns (e.g., user preferences in Netflix)

Further reading

- http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf
- Hesterberg et al. (2005). "Bootstrap methods and permutation tests." In Moore & McCabe, Introduction to the Practice of Statistics.