

Biostatistics 210: Building Regression Models

david.glidden@ucsf.edu

Dave Glidden
Division of Biostatistics
University of California at San Francisco

Contact

Dave Glidden

david.glidden@ucsf.edu

Zoom Sessions

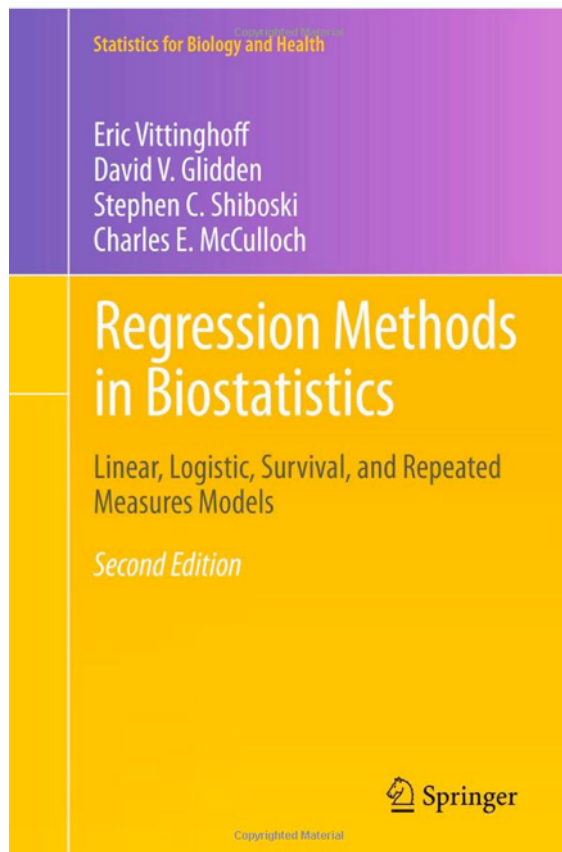
- Tues 10:30-12
- Afternoon HW review every other week
- Camera on, muted (within reason)
- Try hand raise/Q and A
- Feel free to speak up
- Remote instruction assessed in 4 weeks

Grading and HW

- Grade is 20% participation, 80% HW
- 8 HW, due Tuesday at noon
submit through the CLE
- HW issues/clarification use CLE forum
- HW review Tu 3:15-4:15, bi weekly
- TA's Kristen.Azar@ucsf.edu,
Safyer.McKenzie-Sampson@ucsf.edu

You will need

- Stata, version 14 or higher
- Text: Vittinghoff et al, *Regression Models in Biostatistics*, 2nd edition, ISBN #978-1461413523
- Data is always good



Course Textbook

Second edition,
Sept 2011

Regression

Diverse models with common themes

- Predictor selection
- Causal Inference
- Missing values
- Model checking
- Modeling continuous predictors
- Model comparison

Regression Models

Diverse Models

- Linear, Logistic, Cox etc
covered extensively in Biostat 208/209
- Model depends of outcome variable
measures of association vary by outcome
- Also dictated by correlation structure
e.g. generalized estimating equations

Regression Techniques

Outcome	Model	Measure of Association
Continuous	Linear Regression	Difference in Mean
Binary	Logistic Regression	Odds Ratio
Censored Survival	Cox Regression	Hazard Ratio
Count of Event	Poisson Regression	Rate Ratio

Regression

Diverse models with common themes

- Predictor selection
- Causal Inference
- Missing values
- Model checking
- Modeling continuous predictors
- Model comparison

Four General Problems

1. Developing a Model for Prediction/
Stratification
2. Adjusted Effect of Single Variable
3. Identifying Multiple Predictors of Outcome
4. Adjustment to Improve Efficiency in RCT

Example: Prediction

- Cohort of community-dwelling elders
- Predictors: age, co-morbidities, functional status
- Follow-up: mortality over 4 years
- Who is at the highest risk of death?

Lee, SJ et al. JAMA. 2006;295:801-808.

Prediction Issues

- How big should model be?
include every predictor with $p < 0.05$?
- Can continuous predictors be categorized?
- How regression leads to “score”
- How to assess predictive ability?
do we need a validation set?

Pregnancy in ARV initiators

- Cohort of HIV+ women in Uganda
- All given anti-retroviral drugs
- Followed for health outcomes: including pregnancy
- Predictors: CD4+, reproductive hx, clinical status, age, sexual behavior
- What are the predictors of pregnancy?

Pregnancy Study Issues

- Multiple hypotheses:
death of child influence pregnancy?
length of partnership? woman's health?
- Some desire for adjustment
- Missing data on sexual behavior
- Covariates that vary with time

Homsy J et al. PLOS One. 2009;4(1): e4149.

Post-Transplant Diabetes

- Case-control study of post-transplant diabetes following kidney transplant
- 16 cases, 32 controls
- Matched 2:1 on age
- Is FK506 associated with risk of post-transplant diabetes?

LC Greenspan, et al. Pediatric Nephrology 2002, 17:1-5.

Diabetes Study Issues

- Residual confounding
- Small sample size
- Possible interactions:
race/family history
- Dealing with matched structure
effect of pairs is a kind of clustering

SCUT Study

- Patients with bacterial ulcers
- Treated with moxifloxacin
- placebo v. steroid drops
- Outcome: vision at 0, 3, 12, 52 weeks
- Does steroid rx help vision?

SCUT Issues

- Randomized clinical trial
causal effects identified
- Repeated measures data
- Baseline strongly correlated with FU values
- Missing outcome data

Examples

Study	Regression Type
Predictors of Mortality in the Elderly	Prediction
FK506 in Post-Transplant Diabetes	Single Variable
Pregnancy in ARV initiators	Multiple Variable
Steroids in Corneal Ulcers	Efficiency in RCT

Regression Problem

Important difference in analysis advice

- e.g., how many variables in model?
- Single variable -- include many variables
- Prediction -- be more selective
- Multiple variable -- in-between
- RCT -- should be highly selective

Much more on this later

Ideology of Course

- Informed by real data analysis problems
- Emphasis on practical methods
- But theory gives generalizable knowledge
- Advanced but applicable

Choosing Between Methods

Unhelpful

- Quasi-religious treatment of statistics
 - “can’t do that”
 - “not fair”
 - “not kosher”
 - “must do ___ before doing ___”
- Emphasizes rigid rules over properties

Basic Statistical Tasks

- Estimating quantities (point estimation)
e.g., sample means, regression coefficient
- Testing hypotheses
two means different, coefficient is zero
- Constructing confidence interval
for a quantity that's been estimated

Framework for Methods

- Frequentist statistics
- Judge how methods perform in long run
- Long run: hypothetical repetitions of the same experiment
- Can be figured out mathematically
- Or by simulating known data and examining performance

Desirable Properties

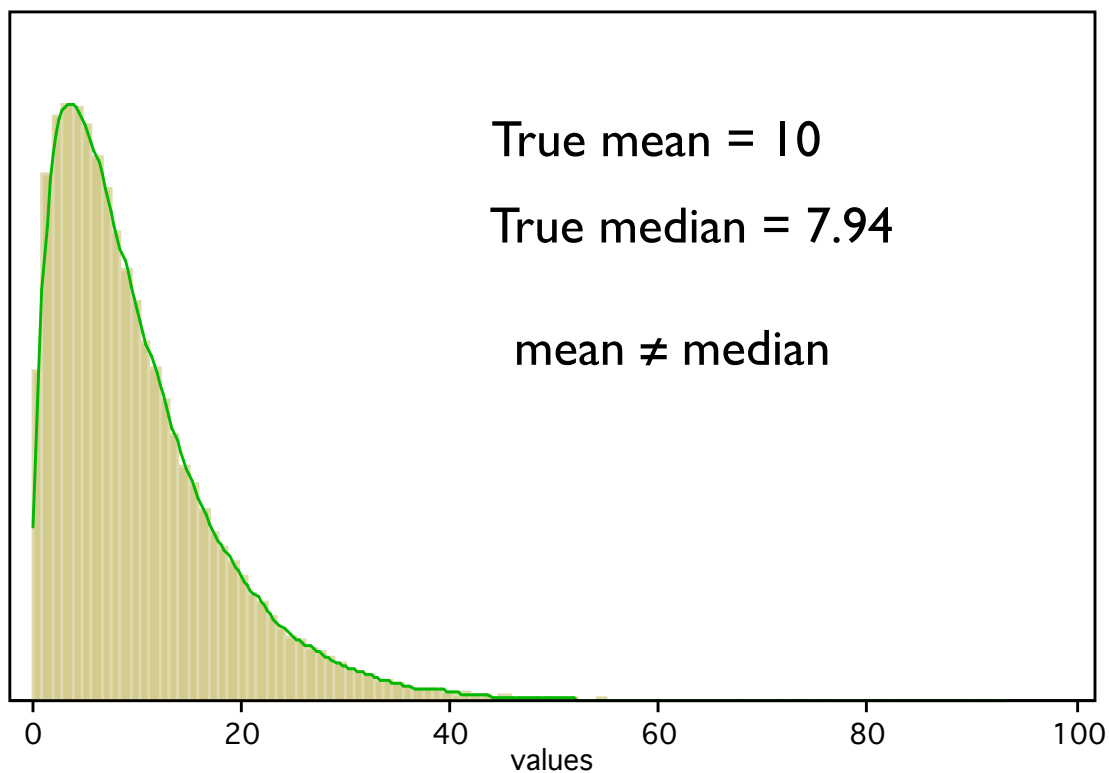
- Validity
- Efficiency
- Robustness
- All desirable -- sometimes in conflict

Validity

- It's name implies it's importance
- Point estimation: unbiasedness
- Testing hypotheses: rejecting 5% of time under null hypothesis
- Confidence interval: 95% chance of covering the true value

Simple Example

- A small ($n=25$) dataset
- Want to summarize the values
- The data is highly skewed
- Going to set the distribution:
know the true mean
know the true median



Issues

- Data is not normally distributed
- Should I calculate the mean? median?
- Are they valid? Are they efficient?
- Let's examine their validity

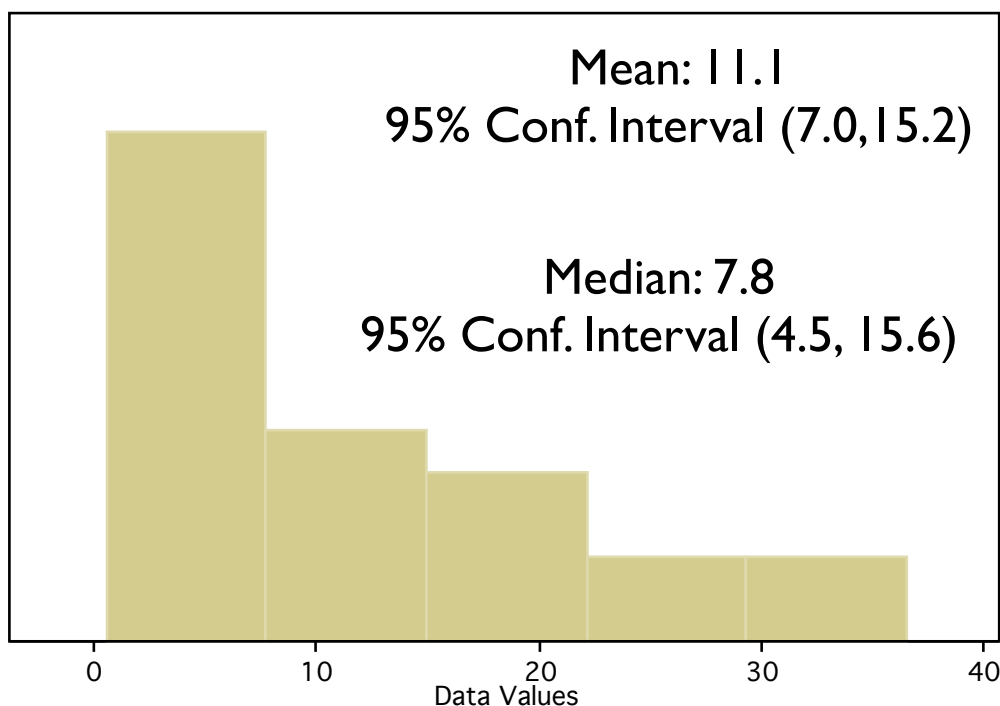
Simulation

- Repeated simulate samples of 25 obs
- From the distribution I showed
- Calculate mean (and 95% CI)
- Calculate median (and 95% CI)
- Examine their properties

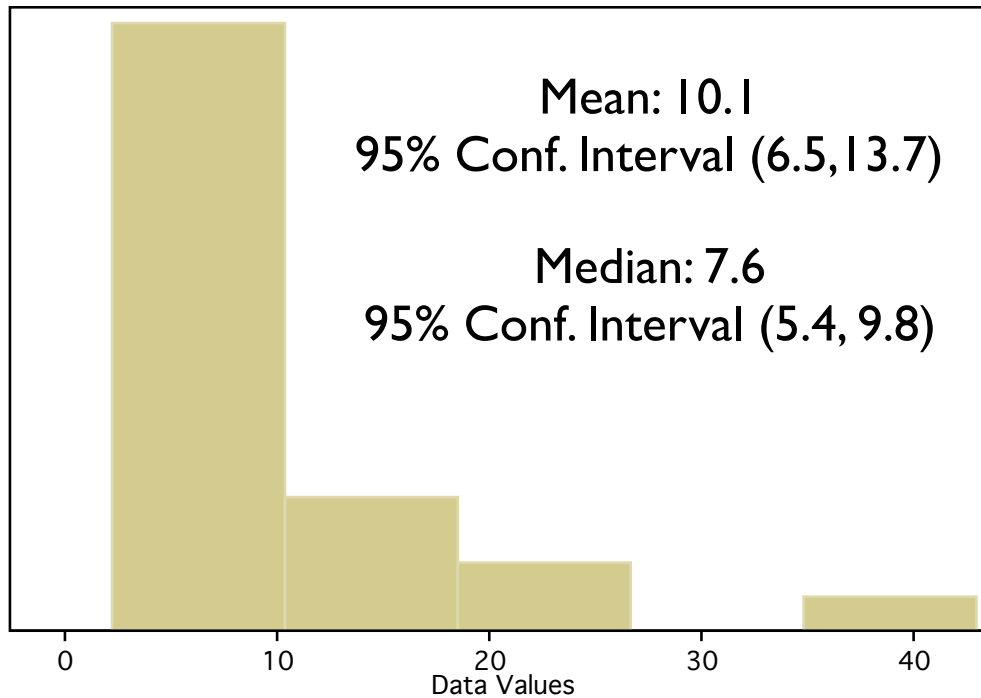
Central Limit Theorem

- Data might not follow Normal distribution
- But it's mean will (approximately)
- Approximation better if
 - sample size is larger
 - data is less skewed
- $n=25$: not large, data is very skewed

Dataset #1



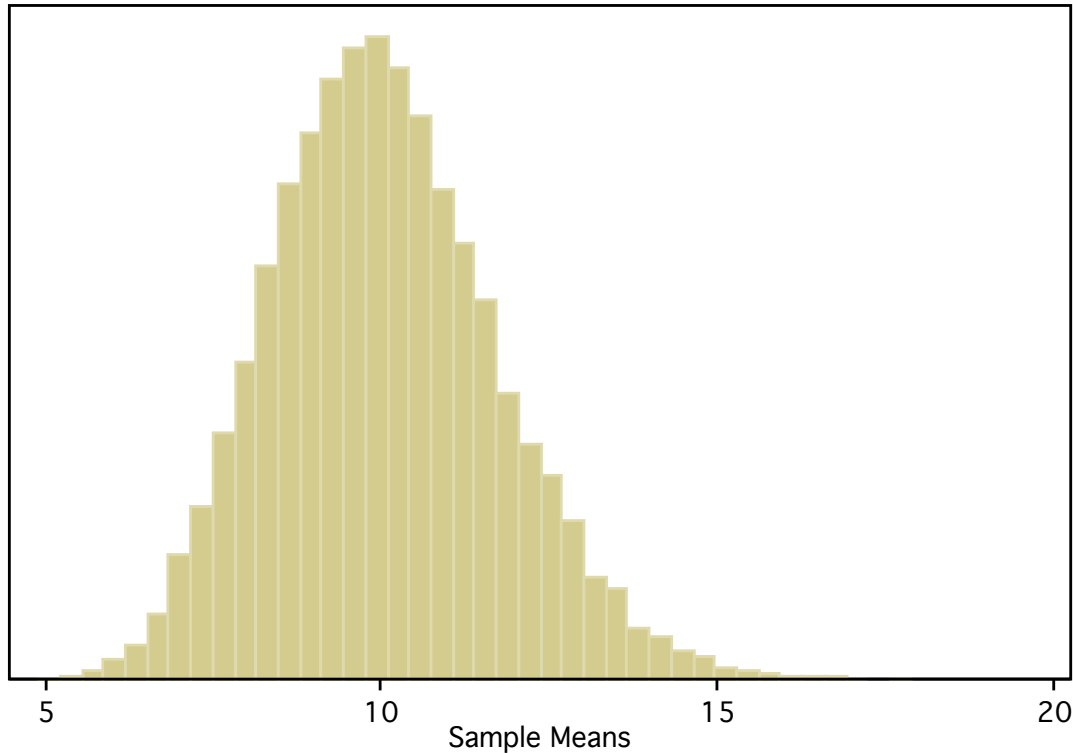
Dataset #2



Repeat This 10,000 Times

	xmean	xmedian
1	11.11253	7.750783
2	10.09266	7.608769
3	13.01521	9.782513
4	12.32289	8.71201
5	10.77202	8.075078
6	8.095377	7.596135
7	8.81252	7.16453
8	9.535892	5.978937
9	8.576806	6.26033
10	9.012348	7.342016
11	8.933349	8.224752
12	9.555133	9.188276
13	10.49054	7.187543
14	8.821921	6.818303
15	9.191763	6.643261

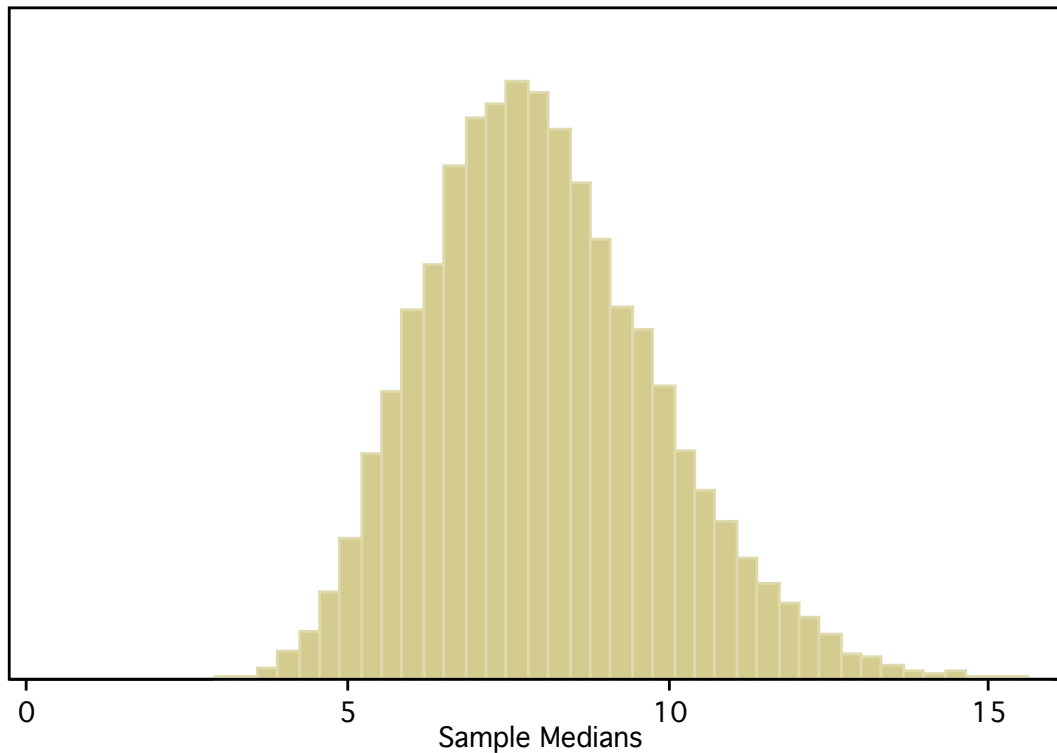
How do the means look?



Summary of “means”

- The long-run performance of sample mean
- In hypothetical replications of the data
- Average of 10,000 values: 10.0
- True Value 10
- The estimate is unbiased
- No systematic departure

How do the medians look?



Unbiasedness

- One aspect of validity
- Estimating the right quantity on average
don't always get it right
- Consistency: with ∞ data, would estimate the quantity exactly
- Inconsistent: with ∞ data, wouldn't estimate the quantity exactly
- Inconsistent = invalid

Unbiased Estimator

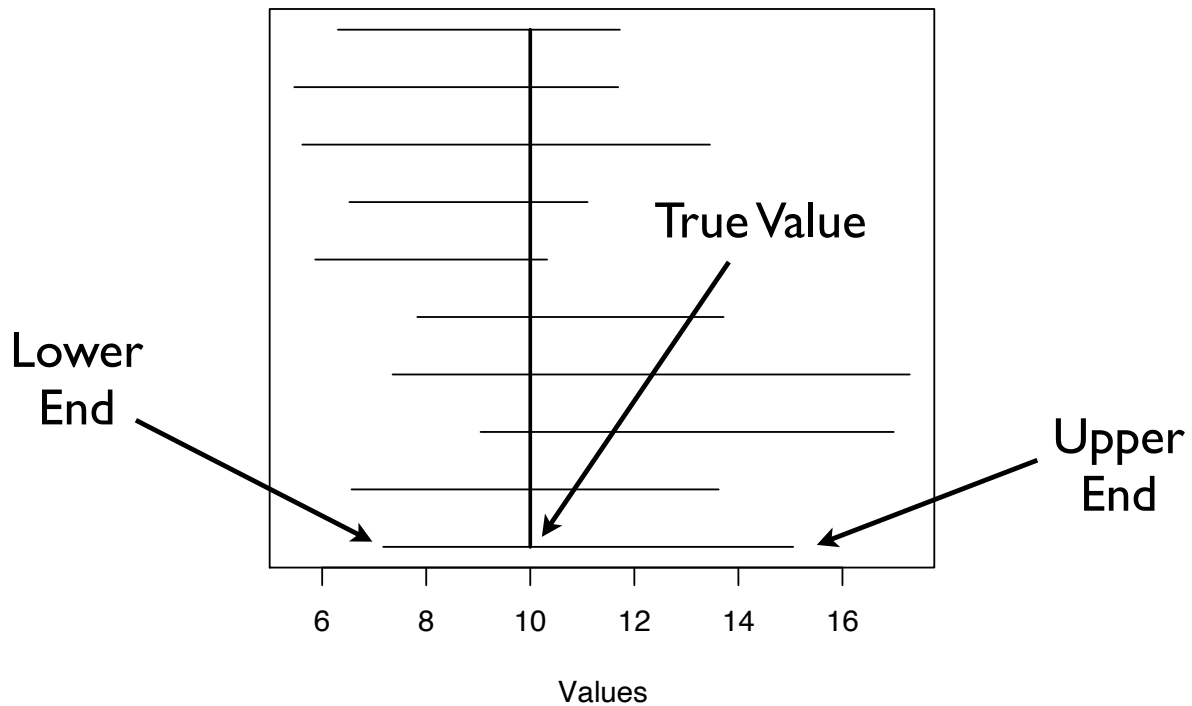


Correct “on average”

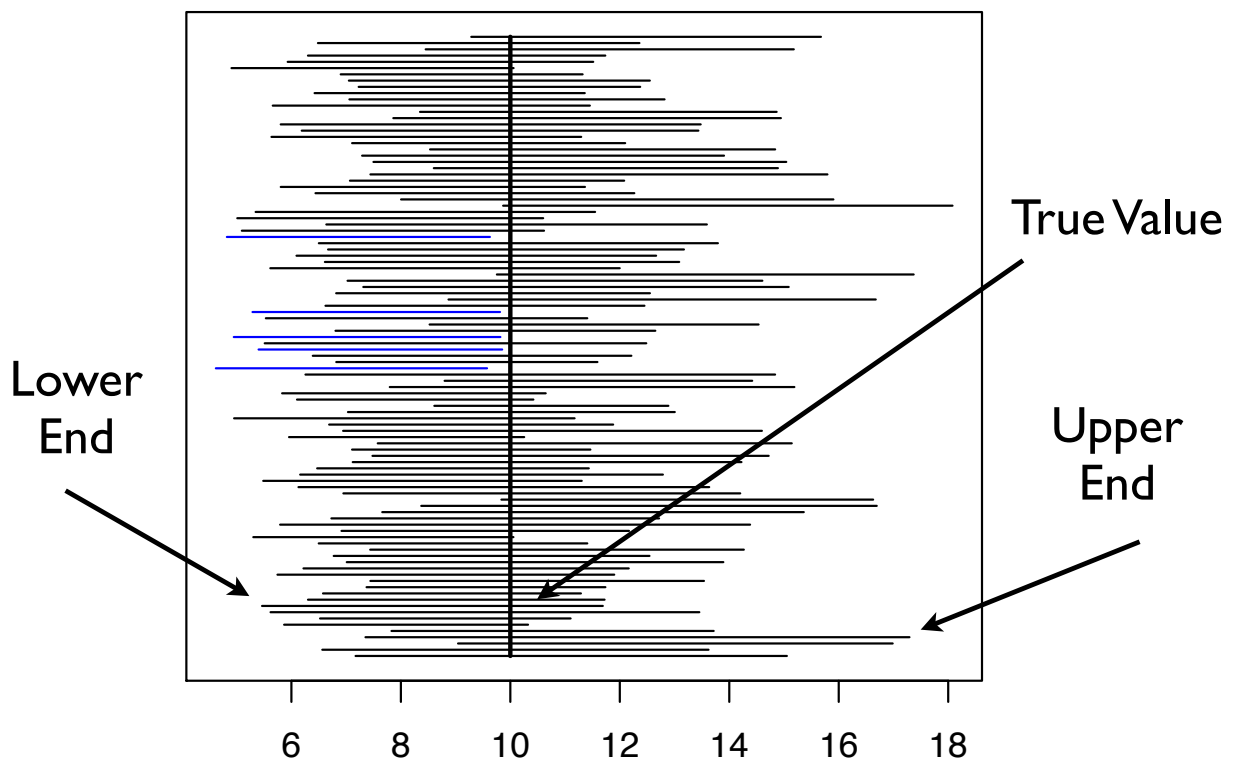
Confidence Interval

- Data #1 95% CI (7.0,15.2)
- A CI is a random function which tries to cover the true value
- Here $7.0 < 10.0 < 15.2$ so it covers it
- A valid 95% CI should cover the true value *with probability 0.95*
- Can check with simulation

Ten Confidence Intervals



100 Confidence Intervals



Confidence Interval

Mean

- 10,000 simulated datasets
- 9,221 CI include the value of 10.0
- 0.92 coverage (less than 0.95)
- 779 CI intervals don't include 10.0
 - 660 have upper limit < 10
 - 119 have lower limit > 10

Confidence Interval

Median

- 10,000 simulated datasets
- 9,486 CI include the value of 7.9
- 0.95 coverage
- 514 CI intervals don't include 7.9
 - 266 have upper limit < 7.9
 - 248 have lower limit > 7.9

Comparison

- mean & median estimate different quantities
- Both unbiased
valid as point estimators
- However, CI for mean “undercovers”
*0.95 CI covers less than 95% of time
bootstrap can't fix it*
- CI for median is almost exact
I'd tend to prefer median in this setting

Significance

- Should I use mean or median?
- Provides a meaningful comparison
- In this particular setting
n=25, highly skewed data
- With less skew, larger n
mean will perform much better

Summary

- Class focus: regression topics
across linear/logistic/Cox etc.
- Four types of model building
will inform many considerations in class
- Need to compare statistics
validity/efficiency/robustness
- Simulation is a valuable tool for this