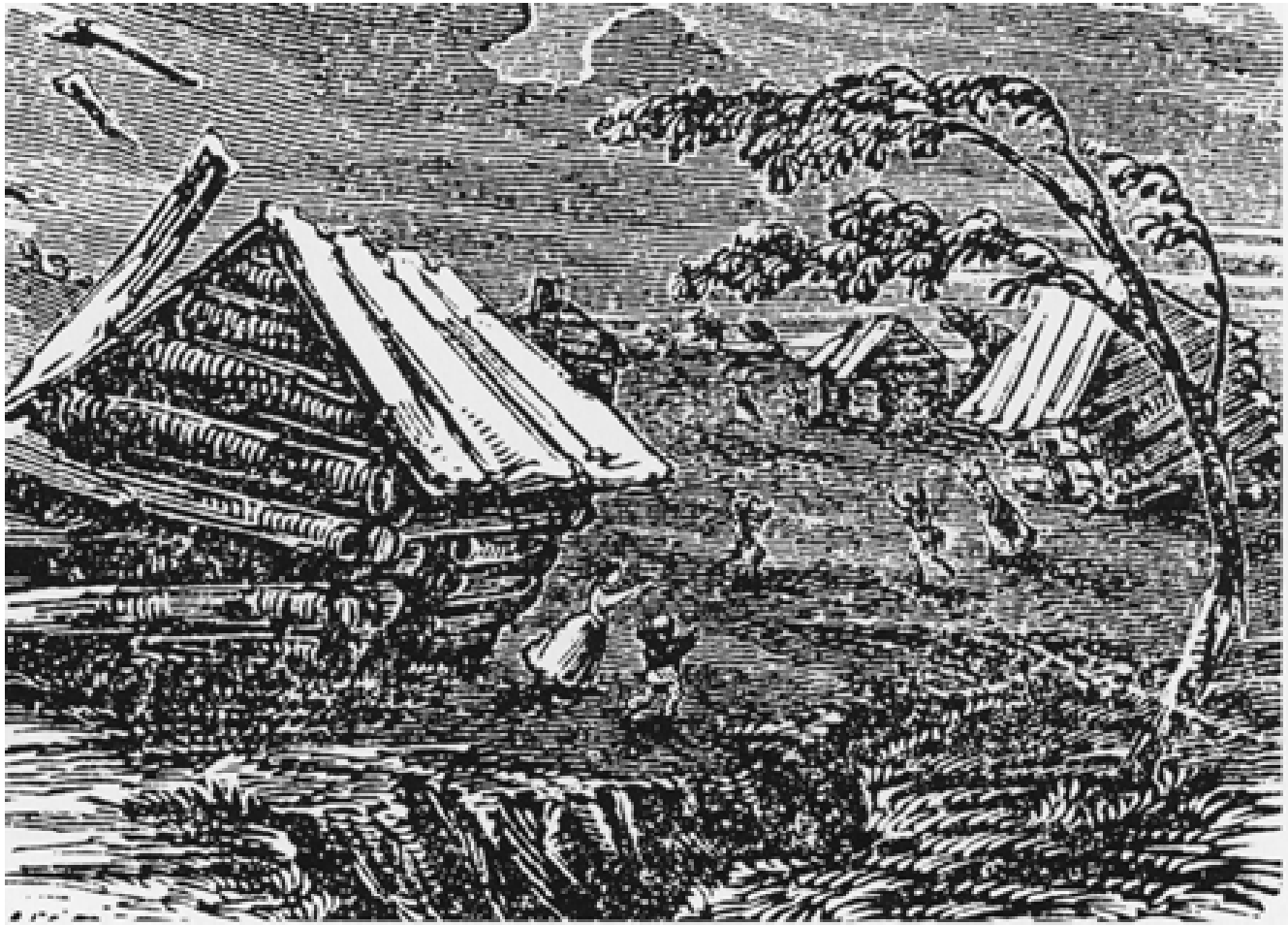
Oil, Earthquakes, and Survival:
some days in the life of a statistician

Tim Hesterberg

Google™

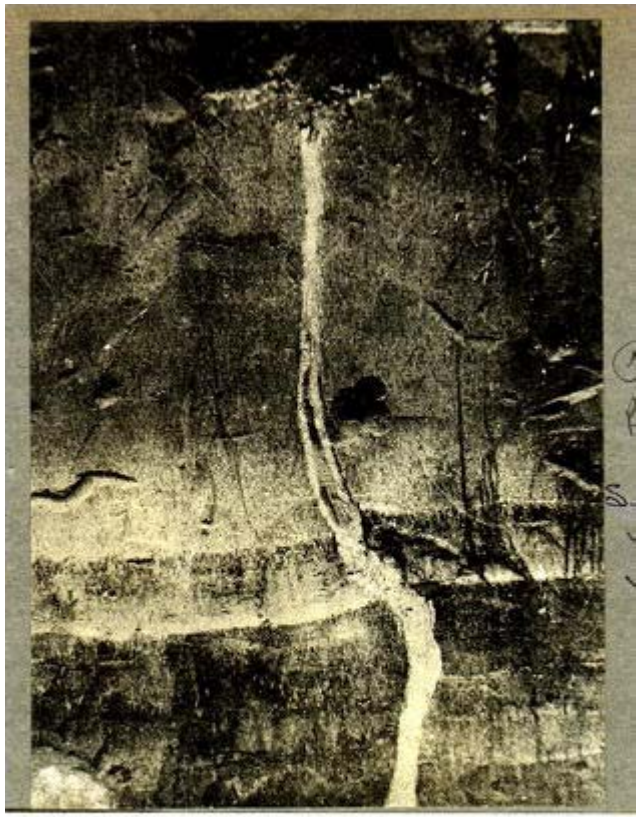
New Madrid Earthquakes



Sand Blows

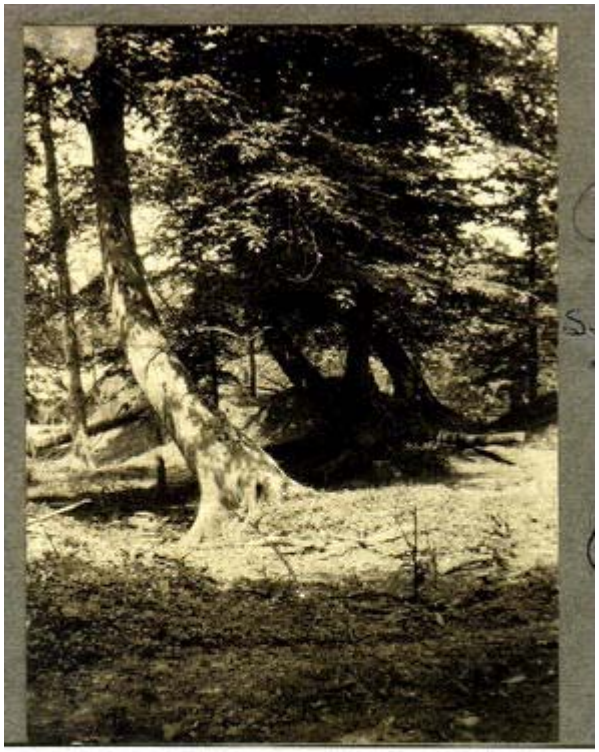
ID. FULLER, M. L. no. 137
Coalescent or linear blows obstructing drainage in the Arkansas district.
Sand blows of the New Madrid earthquake, Blytheville. Mississippi County, Missouri. 1904.

Fissures



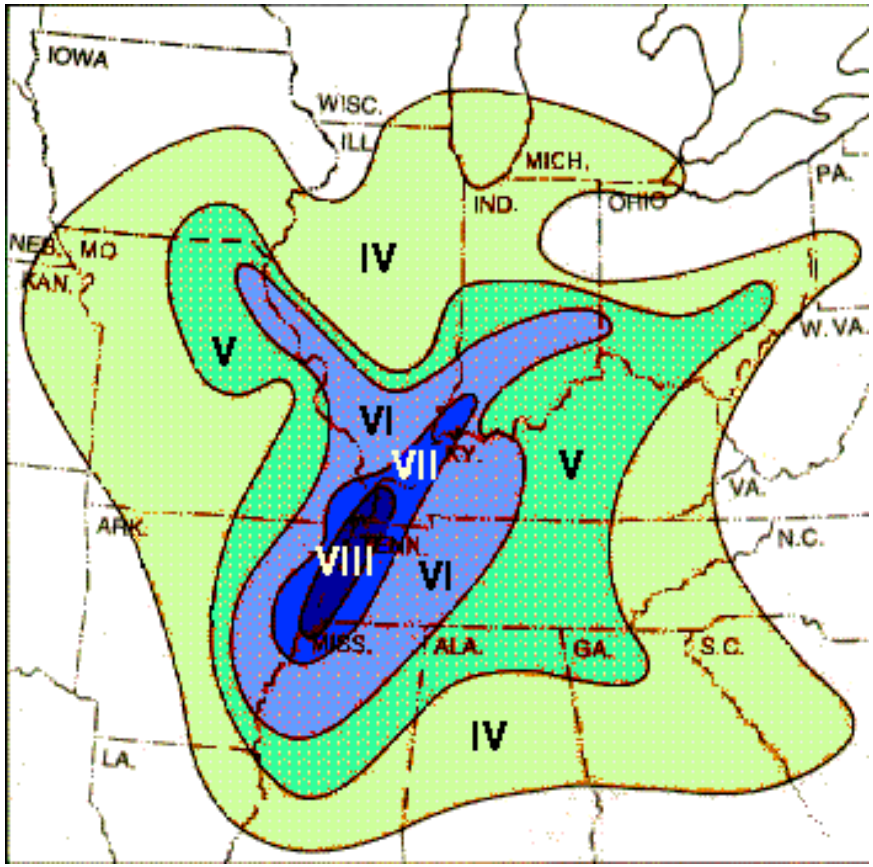
ID. FULLER, M. L. no. 336
Earthquake fissure filled with intruded sand, formed at the time of New Madrid earthquake. Mississippi County, Missouri. 1904.

Bent Trees



ID. FULLER, M. L. no. 356
Trees tilted by New Madrid earthquake, Chichasaw bluffs east side of Reelfoot Lake. Note twist of trees into upright position.

Damage



XI	Disastrous
X	Devastating
IX	Destructive
VIII	Major Damage
VII	Serious Damage
VI	Minor Damage

Comparison

- Northridge CA 1994
  - Magnitude 6.6
  - 60 dead, 1000 hospitalized
  - 20,000 homeless (after 10 days)
  - Sediment in some areas
  - Bedrock highly fractured, faulted, folded
  - Well-prepared Area
- Central USA
  - Magnitude 8.1-8.3 (New Madrid 1811-12)
  - Thousand feet of sediment
  - Bedrock old, stiff, stacked like layered cake
  - Affected area 5x larger for similar quake
  - Complacency
- Goal
  - Estimate deformation caused by seismic events
  - May help locate faults
- Seattle area: historic 9.0 quakes

Topography & Stream Gradients

- Topography has many causes, not all seismic
- Streams have natural gradients
  - Depend on rock & soil, vegetation cover
  - Depend on water flow
  - Streams adjust their channels when perturbed
  - Adjustment takes time
  - Deviations from natural profile may signal recent seismic deformation

Statistical Analysis of Stream Gradients

- Collect data from topographic maps
  - Cheap! Alternative to trenching
- Statistical Analysis
  - Estimate natural profile and deformation
  - Details omitted in this talk, but not beyond you: new ideas are being developed in statistics all the time, many of them not that complicated

Hesterberg History: Variety!

- 4 years college (junior year in Germany)
- 1 year Germany (fellowship)
- 4 years graduate school
- 3 years Pacific Gas & Electric
- 2 years St. Olaf – *teach & practicum*
- 6 years Franklin & Marshall – *teach & consult*
- 12 years Insightful
  - *Software, research, teach, consulting, …*
- Google

Pacific Gas & Electric

- Internal think tank/consulting
  - 20 people, ~10 countries
  - EE, CS, OR, Math, Stat
- Big project: fuel oil inventory

Fuel Oil Inventory

- PG&E has diverse electrical system
  - Hydro, nuclear, geothermal, coal, gas, oil
  - Oil is fuel of last resort (dirty & expensive)
  - How much to carry in inventory for winter?
  - Need depends on rain, snow, outages, …
  - Probabilistic modeling
    - Key new idea was pretty simple!

Monte Carlo

- Fuel = *h*(demand, hydro, …)
  - *h*(demand, supply)=max(0, demand-supply)
  - (take into account time periods, reservoirs, …)
  - Statistical modeling and historical data for distributions for the random quantities
  - E[Fuel] = $\int h(x_1, x_2, ...) f(x_1, x_2, ...) dx_1 dx_2 ...$
  - Generate random *x*'s, take average $(1/n)\Sigma h(x_i)$

Importance sampling

- Interesting cases are rare; so oversample them
- Reweight to correct for that bias
  - E[Fuel] = $\int h(x) f(x) dx = \int h(x) \frac{f(x)}{g(x)} g(x) dx$
- g is colder, drier, more outages than f
- Take weighted average, one of

$$(1/n) \sum h(x_i) f(x_i) / g(x_i)$$

$$\sum h(x_i) w(x_i) / \sum w(x_i) /$$

where $w(x_i) = f(x_i)/g(x_i)$

Teaching

- 2 years postdoc, 6 years small college
- Small colleges, small classes, lot of student involvement
- Summer research (self or with students)
- Consulting
  - 20 faculty (& their students), 12 departments
- Big pay cut (OK; more later)

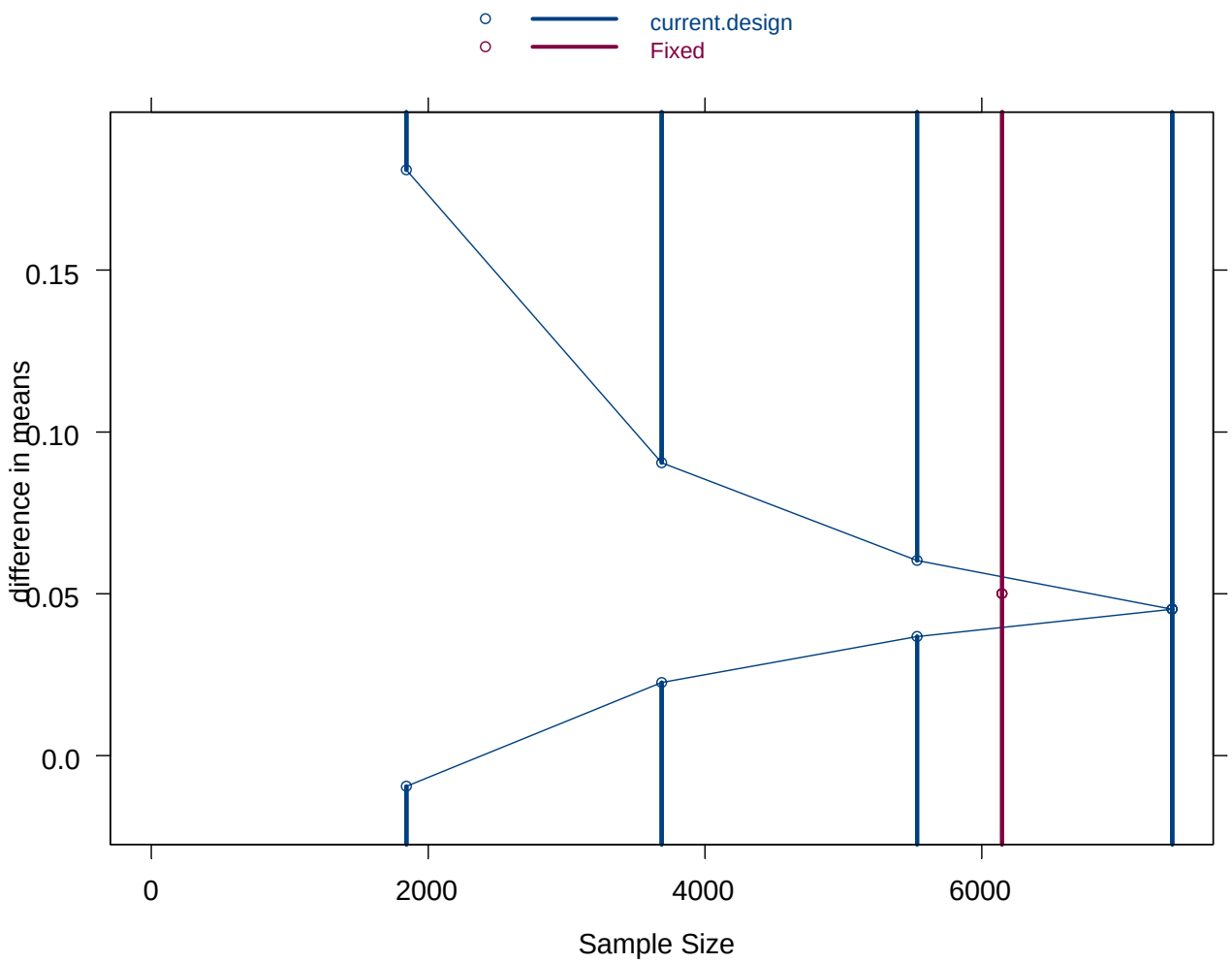
Insightful

- Statistical Software – S-PLUS
- Research
- Consulting
- Training
- Development, Tech Support, Documentation, Sales Support, …

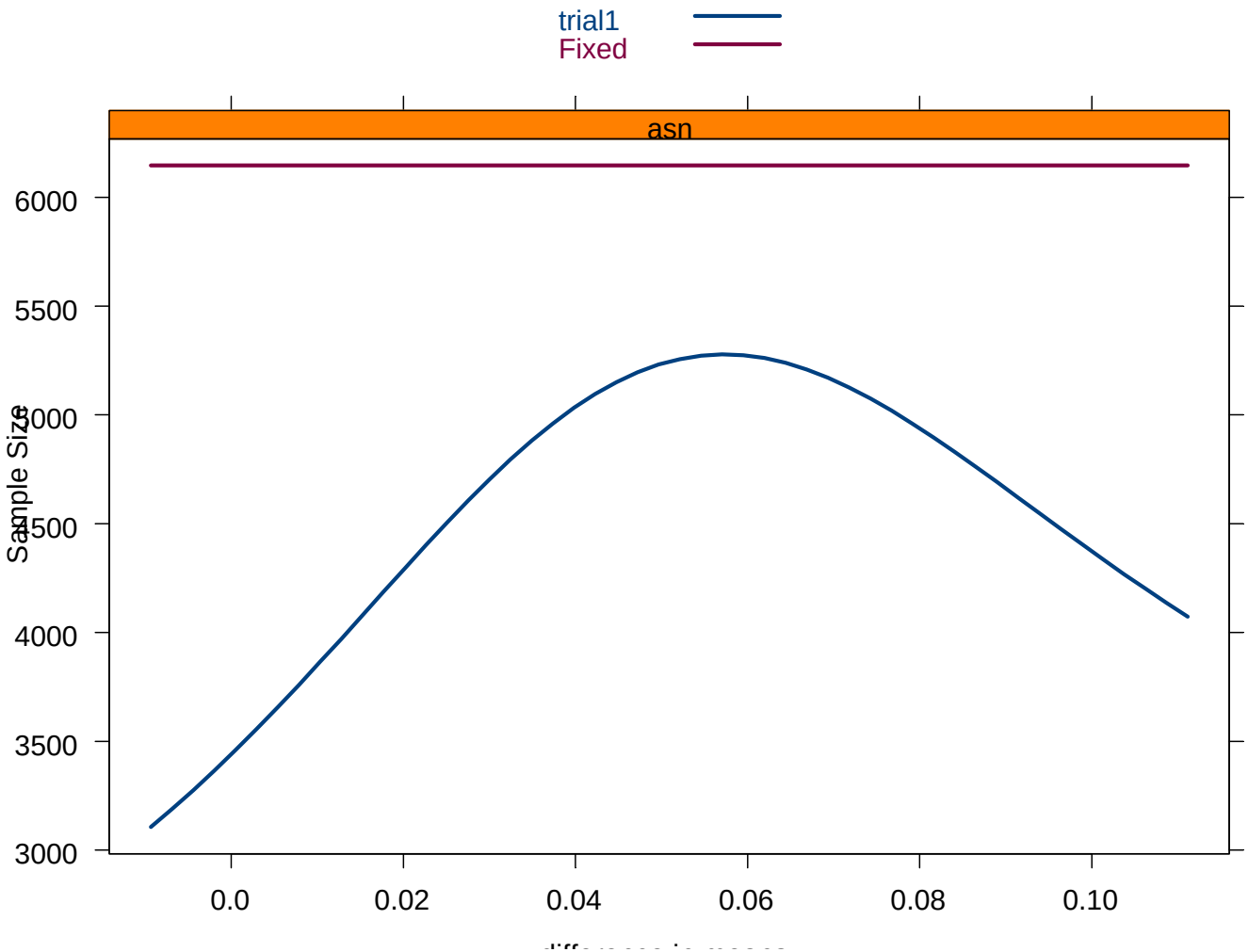
Research (at Insightful)

- Bootstrap
  - 3 interns
- Time series
- Sequential Designs for Clinical Trials
- Functional Data analysis
- Long-tailed distributions (sonar, finance)
- Econometrics (simulation-based)
  - 2 interns
- Least Angle Regression
- Survival Models

Sequential Stopping Rule



Average Sample Size



Consulting (at Insightful)

- Telecommunications
- Clinical Trials
- Risk Management in Finance
  - (June 05 in Zürich Switzerland)
- Pricing – logistic regression for large data
- Portfolio analysis
- Electric Power
- Survival Analysis

Training Courses

- Bootstrap, S-PLUS programming, Statistical Models
- Boston, San Francisco, New York, Pomona, Rochester, Cincinnati, Little Rock, Portland, Albuquerque, Philadelphia, San Antonio
- London, Basingstoke, Zürich, Basel, Manchester, Montpellier, Toronto, Manchester, Bedford,

Change Statistical Education

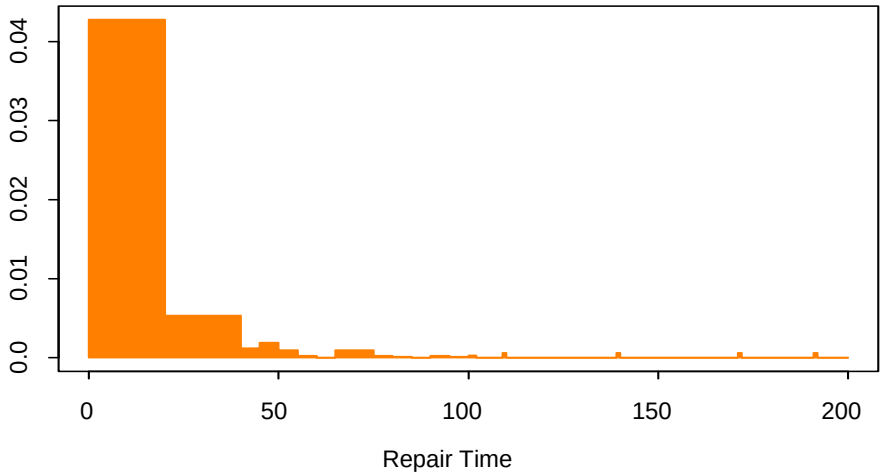
- Many reform movements – more data, less probability, …
- Bootstrap: replace formulas with computer simulation and visualization
  - Easier to understand
  - More accurate
  - Not limited to simple statistics like means

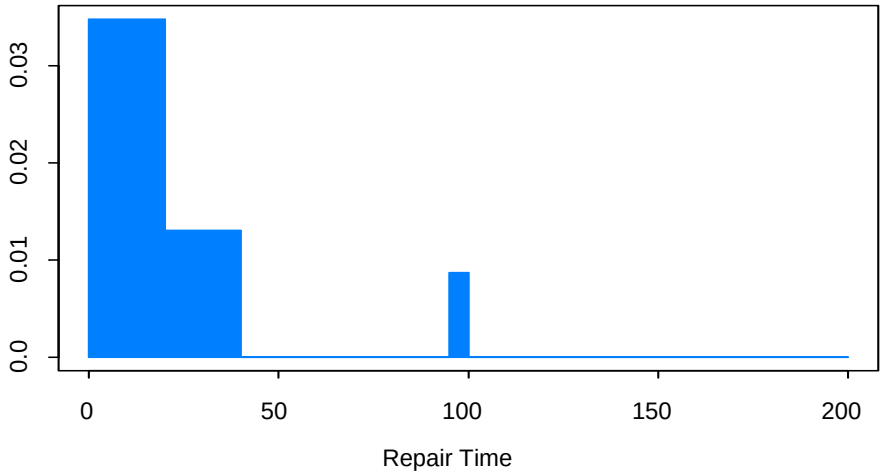
Example - Verizon

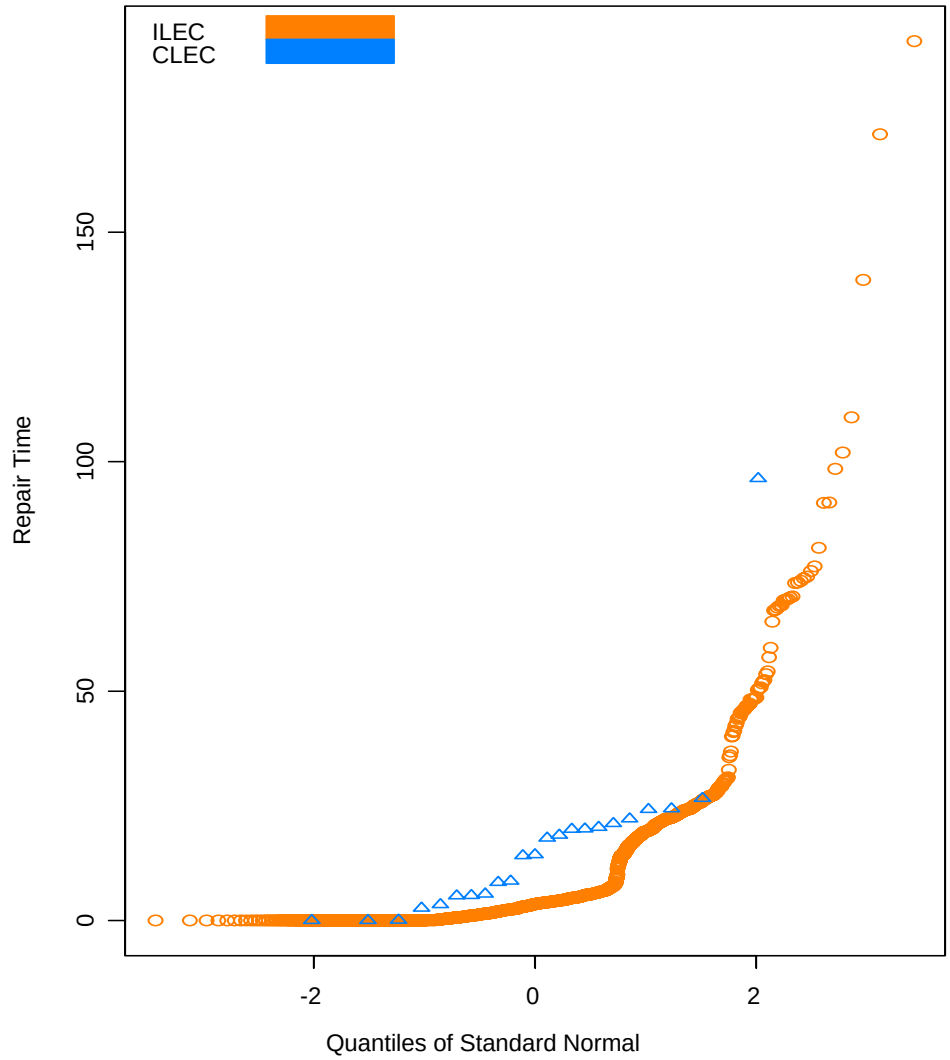
	Number of Observations	Average Repair Time
ILEC (Verizon)	1664	8.4
CLEC (other carrier)	23	16.5

Is the difference statistically significant?

Example Data



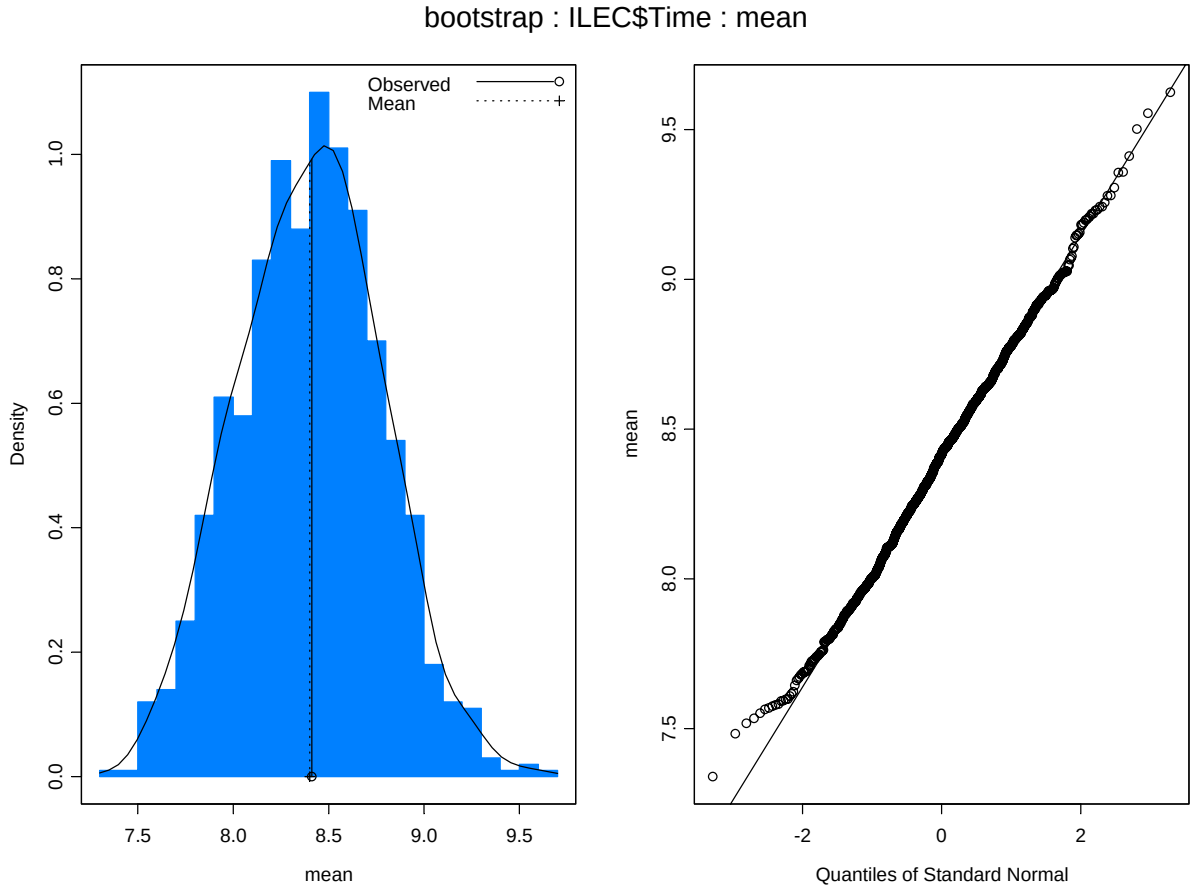




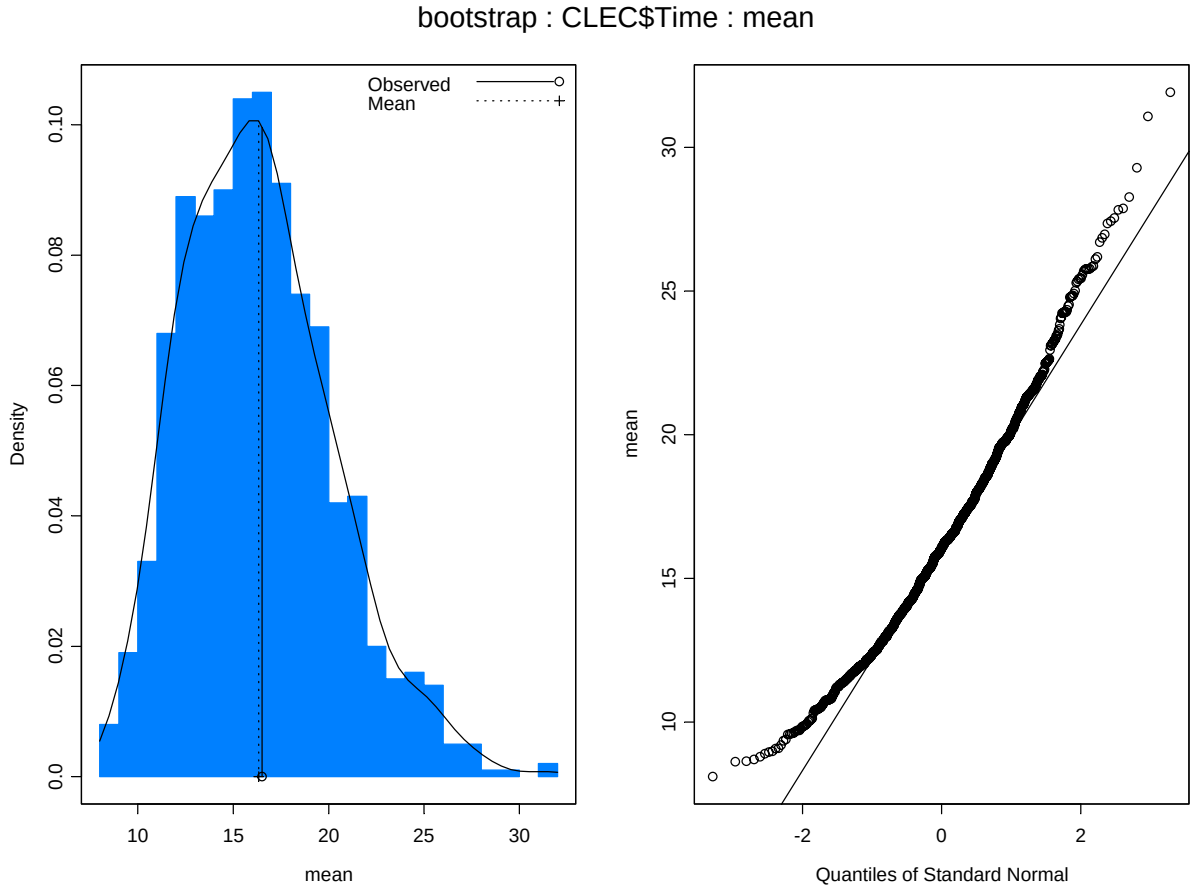
Bootstrap Procedure

- Repeat 1000 times
  - Draw a sample of size *n* with replacement from the original data (“bootstrap sample”, or “resample”)
  - Calculate the sample mean for the resample
- The 1000 bootstrap sample means comprise the bootstrap distribution.

Bootstrap Distn for ILEC mean



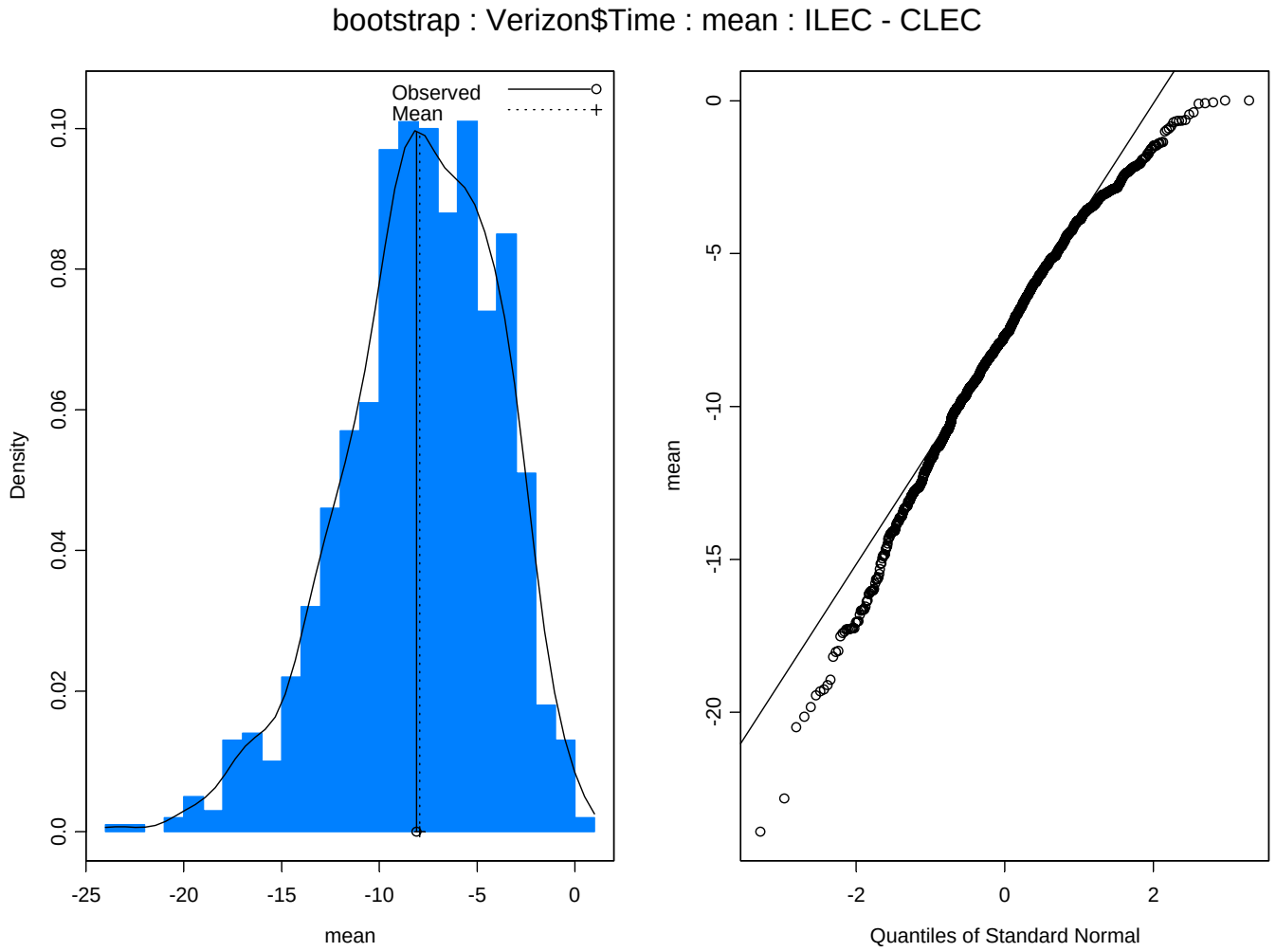
Bootstrap Distn for CLEC mean



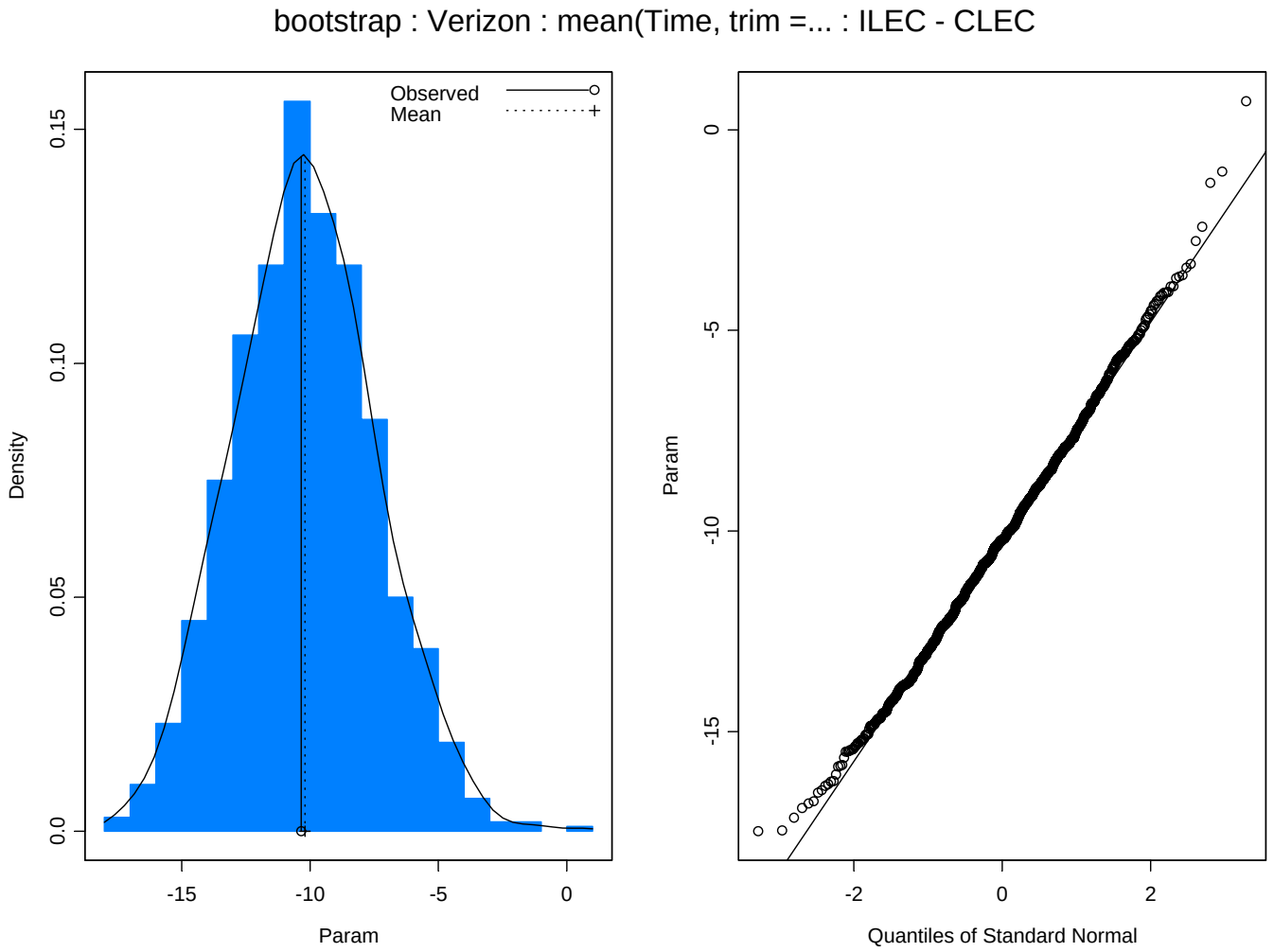
Take another look

- Take another look at the previous two figures.
- Is the amount of non-normality/asymmetry there a cause for concern?
- Note – we're looking at a sampling distribution, not the underlying distribution. This is *after* the CLT effect!

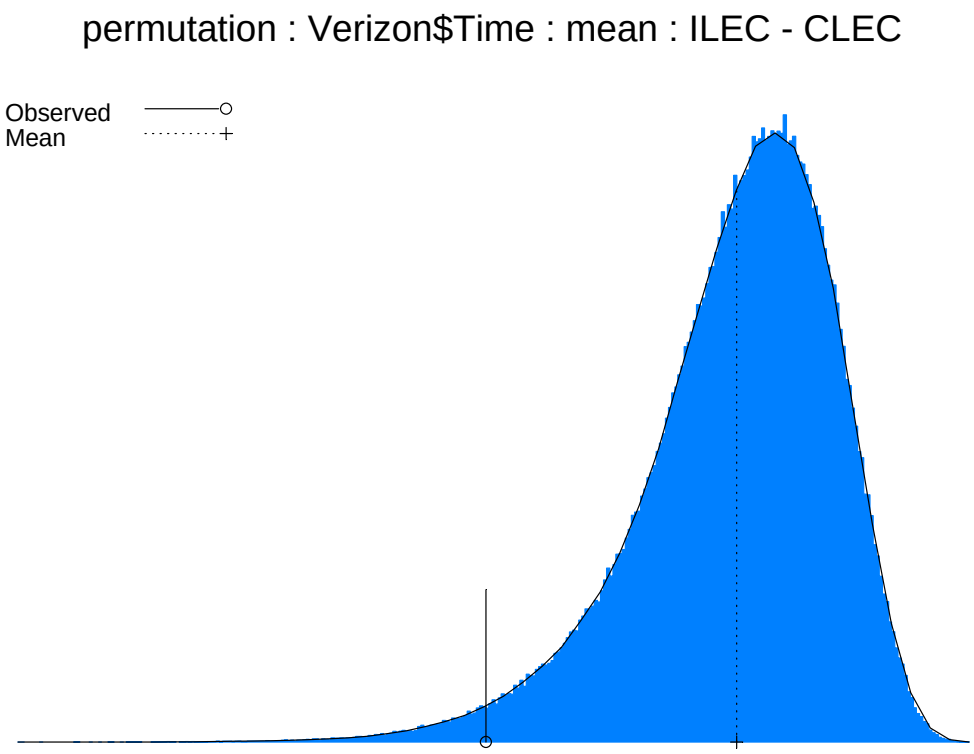
Verizon bootstrap diff in means



…difference in trimmed means



Verizon permutation test



Verizon test results

Pooled-variance t-test

t = -2.6125, df = 1685, **p-value = 0.0045**

Non-pooled-variance t-test

t = -1.9834, df = 22.3463548265907, p-value = 0.0299

permutationTestMeans(data = Verizon$Time,
treatment = Verizon$Group, B = 499999,
alternative = "less", seed = 99)

Number of Replications: 499999

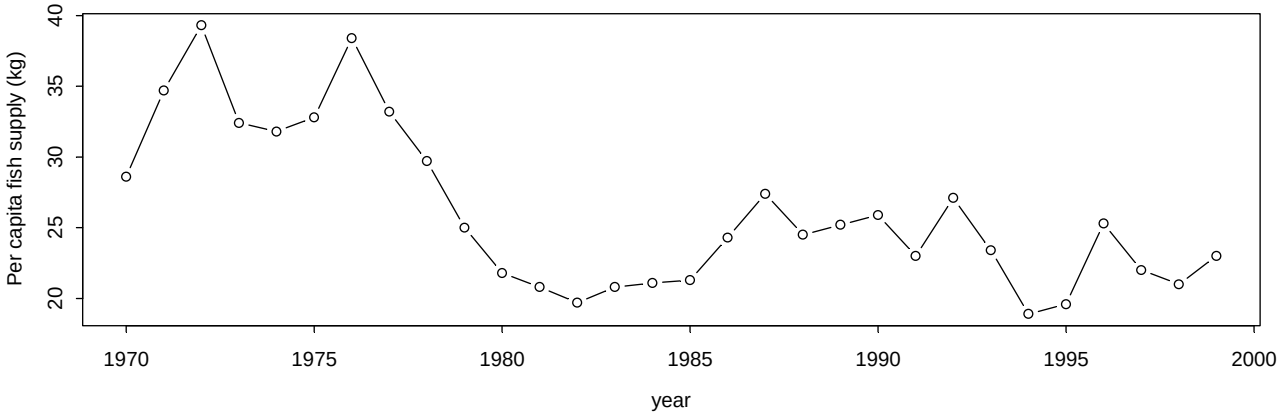
Summary Statistics:

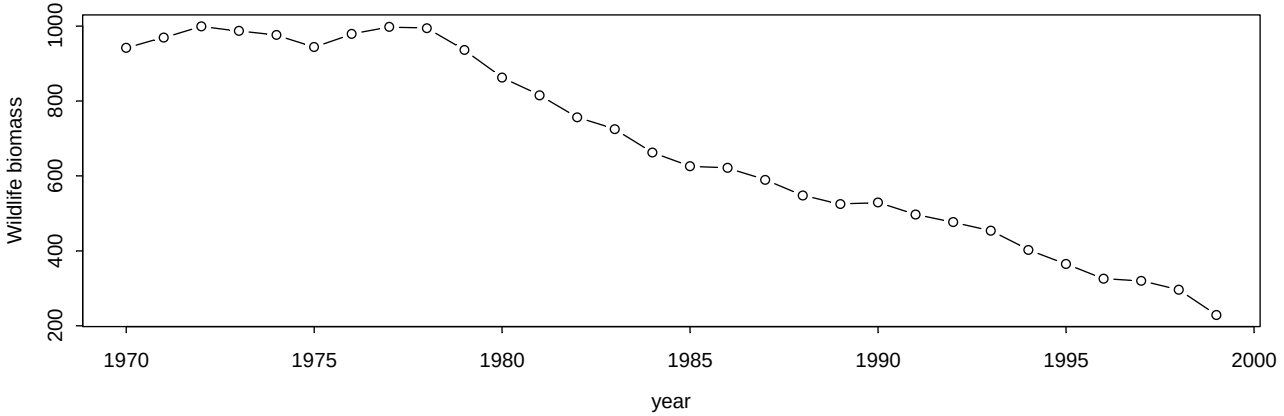
	Observed	Mean	SE	alternative	**p.value**
Var	-8.098	-0.001288	3.105	less	**0.01825**

Bushmeat

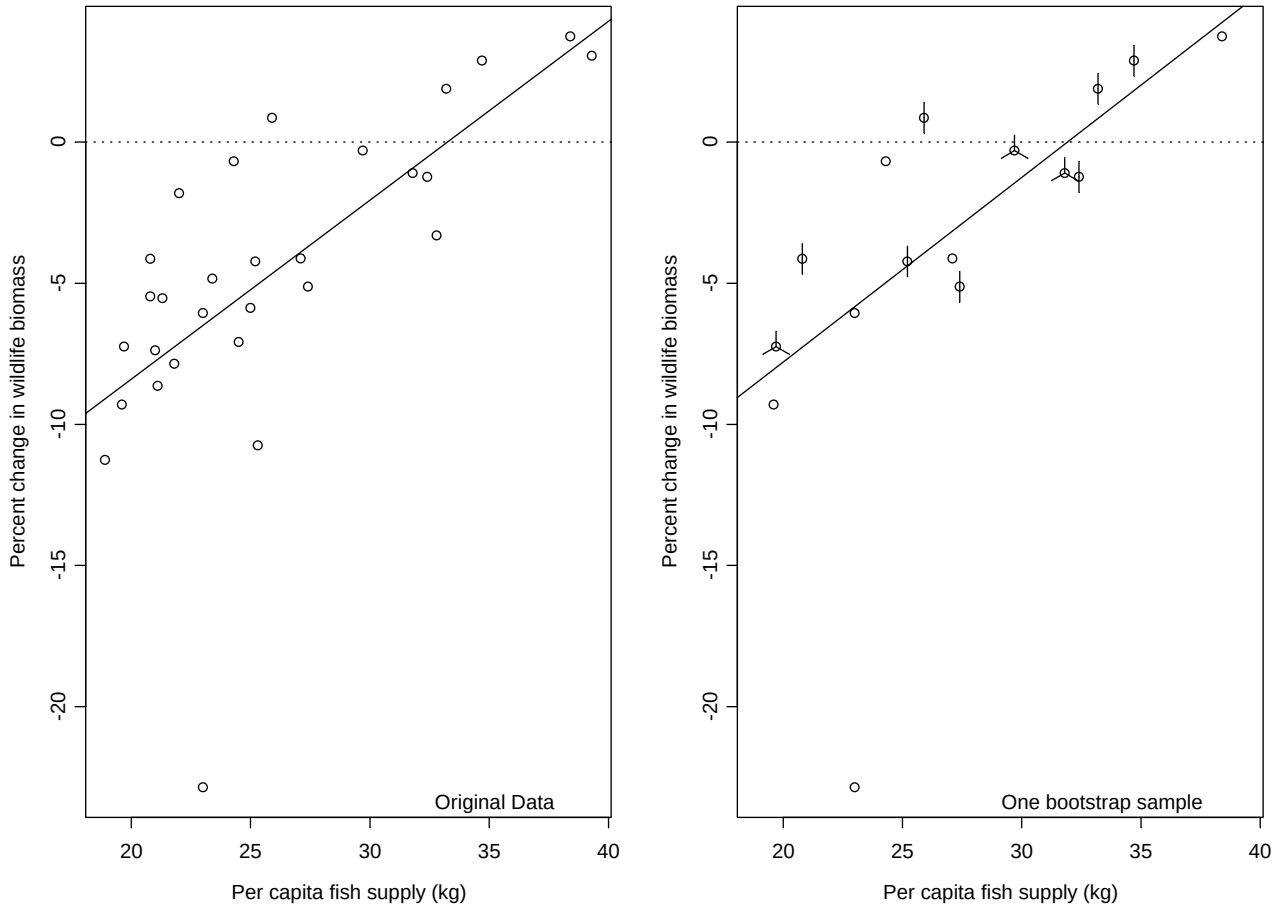
- Brashares et al., *Science* 2004
  - "Bushmeat hunting, wildlife declines, and fish supply in West Africa"
  - 1970-1999, biomass of 30 species in national parks, and fish supply (kg/person/year)
  - Relate change in biomass to fish supply

Bushmeat Data

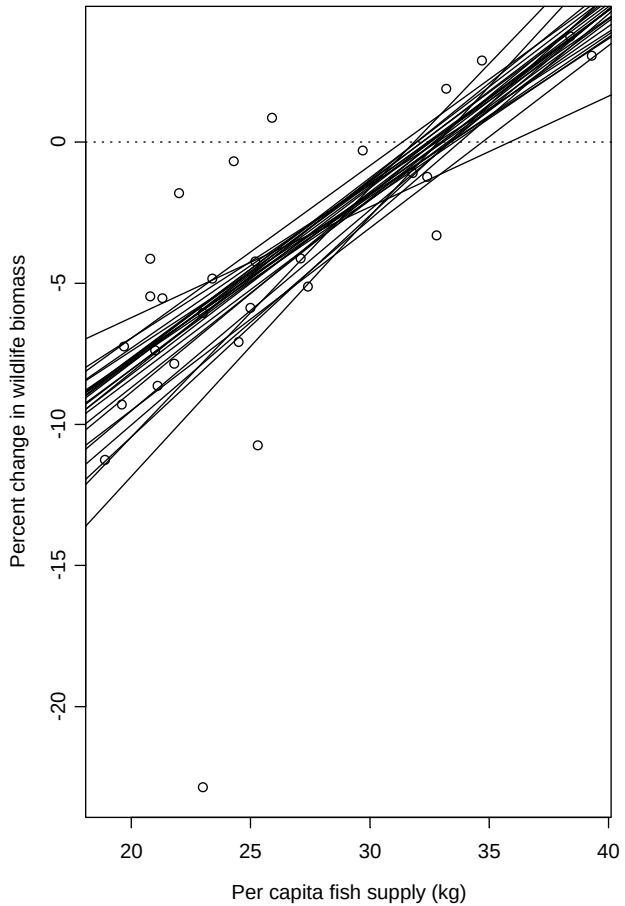


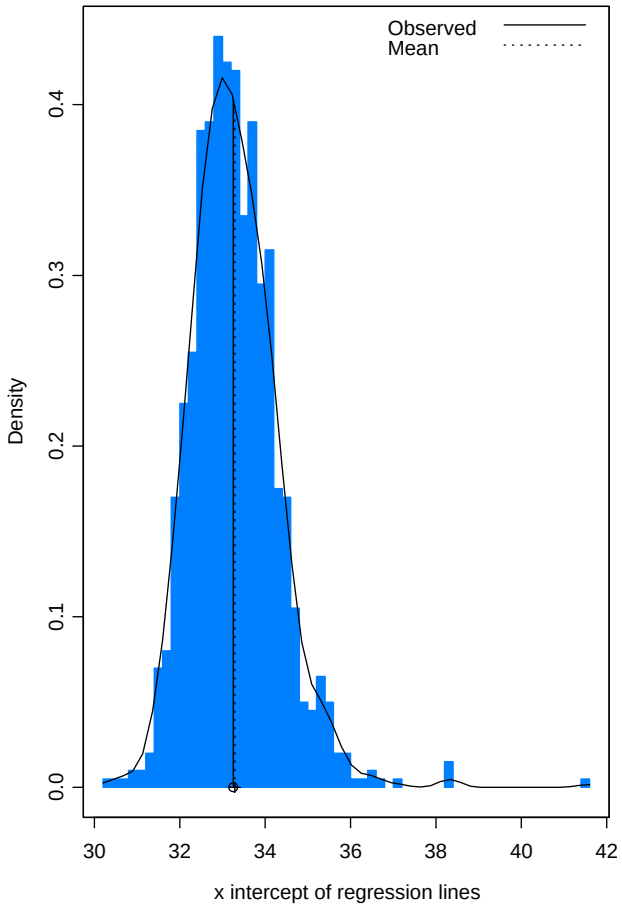


Bushmeat, ΔBiomass vs Fish



Bootstrap Bushmeat





- Google Web Optimizer
  - Optimize your web site
  - Create two versions, test which one works best
- Detecting Trends in Web Search
  - (related to the flu detector)
- …

Money and Job

- Choose interesting work!
- Money
  - Life frugally (house and car!)
  - Flexibility in job choice
  - Reduce stress
- Avoid driving
  - Time & money
  - Live close, bike, public transportation

Go abroad

- Empowering
- Interesting
- Different ways of doing things
  - Sprawl
  - Public transportation
  - Walk
- Resume value – you stand out
- Peace Corp? Housemate Reed Hastings

Keep fit

- Energy, health, longer career with high performance
- Exercise
  - Bike or walk to work
- Eat right
  - Avoid American-size portions

Get Involved

- Professional Groups
  - Raise your profile!
  - Give talks!
- Community Groups
  - Desperate for volunteers in modern America
- Do good
  - Own happiness
  - Time or money

Statistics as Career

- Interesting
- Variety
- High Demand
  - Including