Bootstrapping

PULLING YOURSELF UP THROUGH YOUR ASSUMPTIONS!

Psych Science (Journal)

http://www.psychologicalscience.org/index.php/news/releases/psychological-science-sets-new-standards-for-research-reporting.html

- In addition, the journal will encourage authors to use the "new statistics" of effect sizes, confidence intervals, and meta-analyses in an effort to avoid problems typically associated with null-hypothesis significance testing. To support this approach, the journal has published a statistics tutorial by Geoff Cumming of La Trobe University in Australia. "The New Statistics: Why and How" is freely available **online**.

What do they mean, "new statistics"?

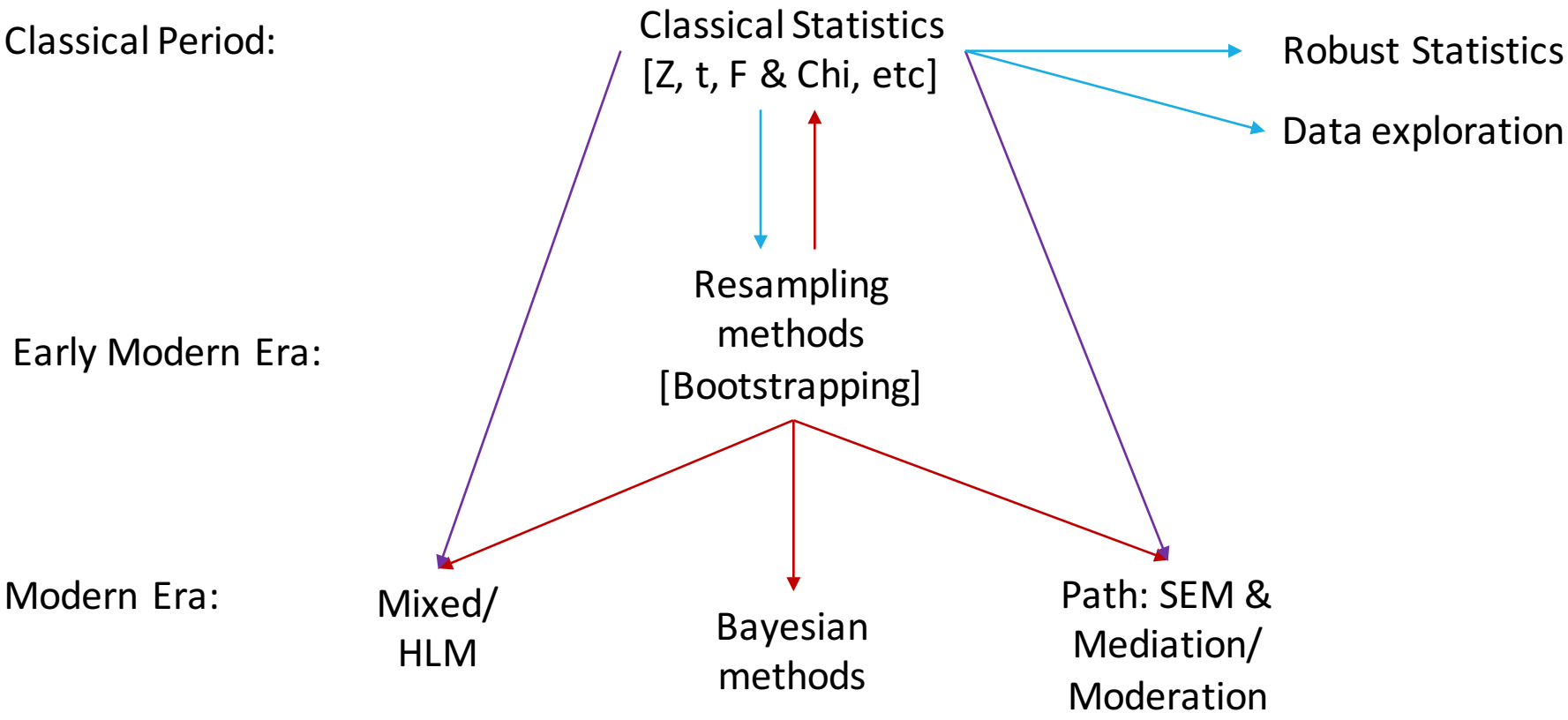
1) Data exploration: John Tukey's (1977) book Exploratory Data Analysis legitimated data exploration and also provides a wealth of practical guidance. There is great scope to bring Tukey's approach into the era of powerful interactive software for data mining and representation.

2) Bayesian methods: These are becoming commonly used in some disciplines, for example, ecology (McCarthy, 2007). Bayesian approaches to estimation based on credible intervals, to model assessment and selection, and to meta-analysis are highly valuable (Kruschke, 2010). I would be wary, however, of Bayesian hypothesis testing, if it does not escape the limitations of dichotomous thinking.

3) Robust methods: The common assumption of normally distributed populations is often unrealistic, and conventional methods are not as robust to typical departures from normality as is often assumed. Robust methods largely sidestep such problems and deserve to be more widely used (Erceg-Hurn & Mirosevich, 2008; Wilcox, 2011).

4) Resampling and bootstrapping methods: These are attractive in many situations. They often require few assumptions and can be used to estimate CIs (Kirby & Gerlanc, 2013).

“Modern” Statistics



Resampling Methods

Resample your observed data in ways to estimate the underlying distribution & error terms

Resampling methods have:

- Been a salvation to classical statistics
  - The are work around to classical assumptions for generating pvalues and confidence intervals
- Opened the door to questions we never could ask before!
  - Directly assess the difference between:
    - Medians
    - Variance
    - Constructed indexes

History of the assumptions regarding Pvalues

- Late 1800s – Probability values can only derived from ***population*** parameters
  - Independent sampling with replacement
  - Gaussian (normal) distributions (e.g., Z-tests)
- 1908 – Probability values can be approximated based on ***samples*** parameters
  - Independent sampling with replacement
  - Gaussian (normal) distributions,
    - Corrected for the sample size (e.g., *t*-distribution)
    - Reliance on the central limit theorem!
      - Mean + SD
- 1918/21/25 – Probability values can be corrected for the number of conditions
  - Independent sampling with replacement
  - Gaussian (normal) distributions, but corrected for number of conditions (e.g., *F*-distribution)

Rucci & Tweney (1980)

Expansions of ANOVA into psychology

- 1930-50 – Expansions of ANOVA into psychology
  - 1941: $t^2 = F$
  - 1946: RM ANOVA, ANOVA with unequal sample sizes
  - 1950: Homogeneity tests!
- Psychologists are pushing the ANOVA well past it design parameters and pushing its update
  - Humans are messy
    - Variance is not stable
    - Distributions are not always perfectly normal
    - Measurements are correlated
  - Result: Reliance on assumptions are problematic (but solutions are not readily apparent)
    - Error terms -> Pvalues

[1951 – UNIVAC is invented]

“Boy scout’s” Jackknife

- 1950s - Quenouille & Tukey: Invented a new ‘generalized’ method based on ‘resampling’ to better estimate specific statistic, i.e., theta - θ.
  - $where\ \theta =$ μ; σ^2; etc

$$\text{Estimate of } \mu; \overline{x_i} = \frac{1}{n-1}\sum_{j\neq i}^{n} x_j$$

Recalculate mean ($\overline{x_i}$) for a set of scores where we skip every j x-score.

5 7 9 7 9 5 0 7 6 0 7 4 6 8 7

j=2

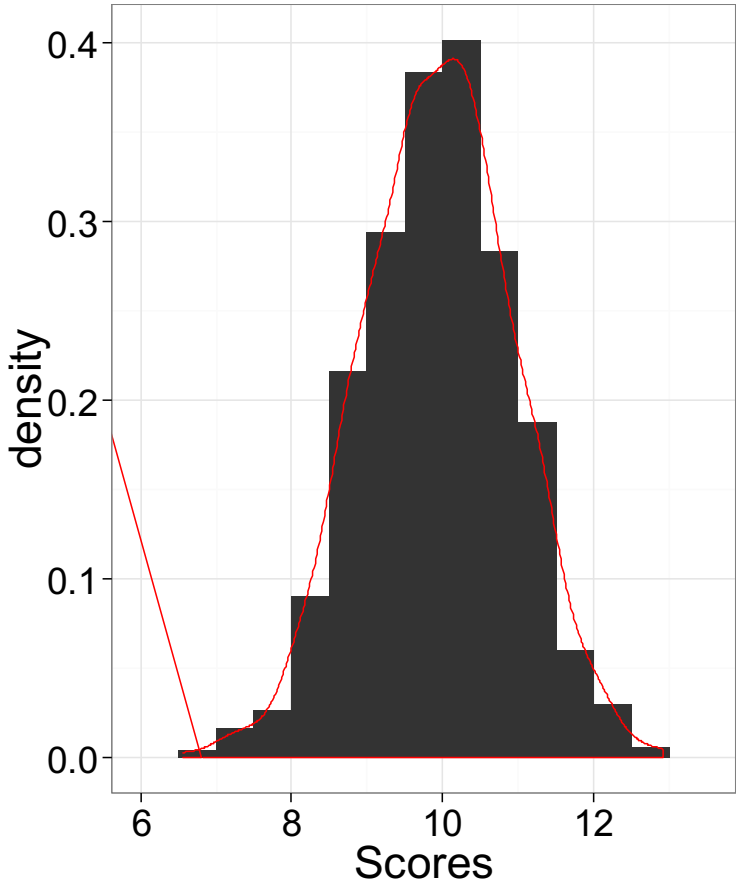
5 7 9 7 9 5 0 7 6 0 7 4 6 8 7

Sampling ***WITHOUT*** replacement!

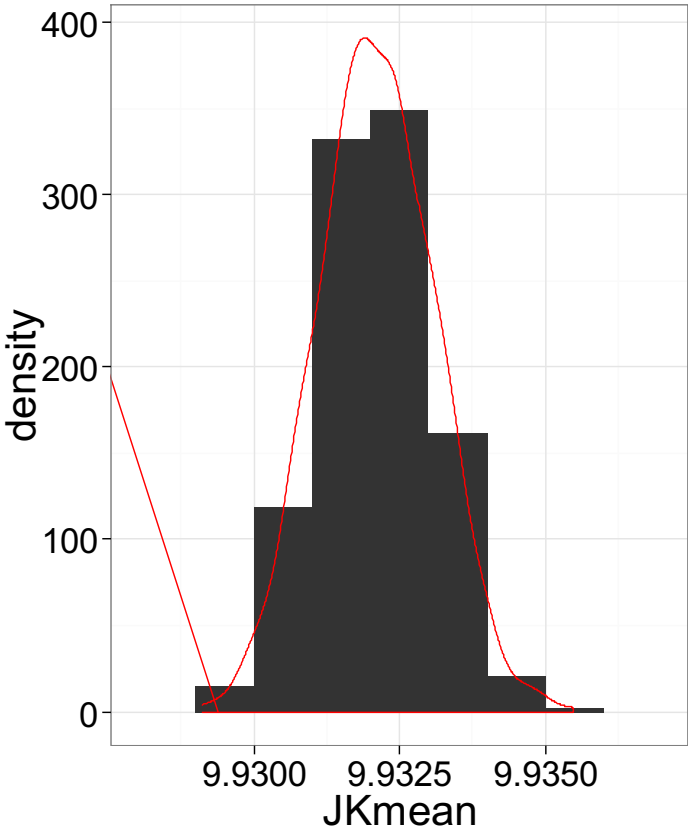
Jackknife Mean [Large sample]

Simulation Parameters
Gaussian Distribution
N= 1000
True mean = 10
True SD = 1

mean = 9.93
SD = .98



Jackknifed Distribution of Means



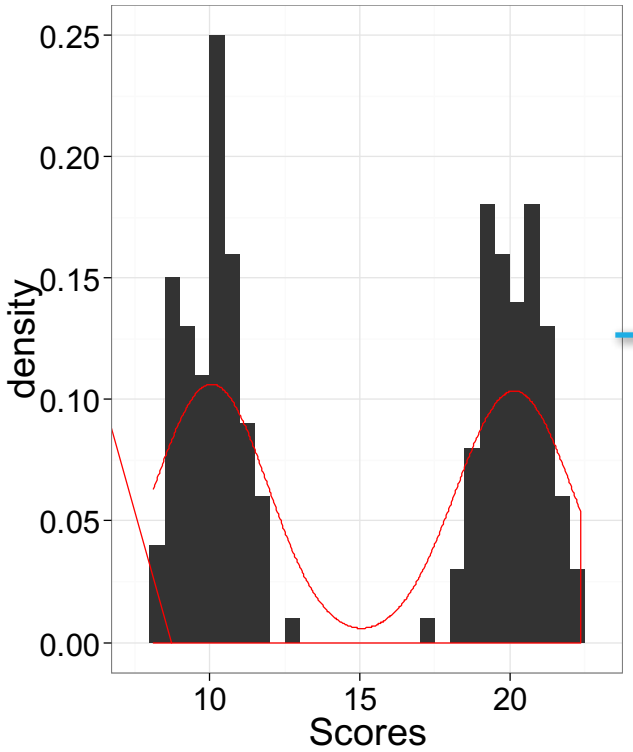
Jackknife Bias

- Bias = θ for all observed data – jackknifed θ
  - When distribution of a θ from set of observations is normal (i.e., the central limit theorem holds) than $bias \approx 0$

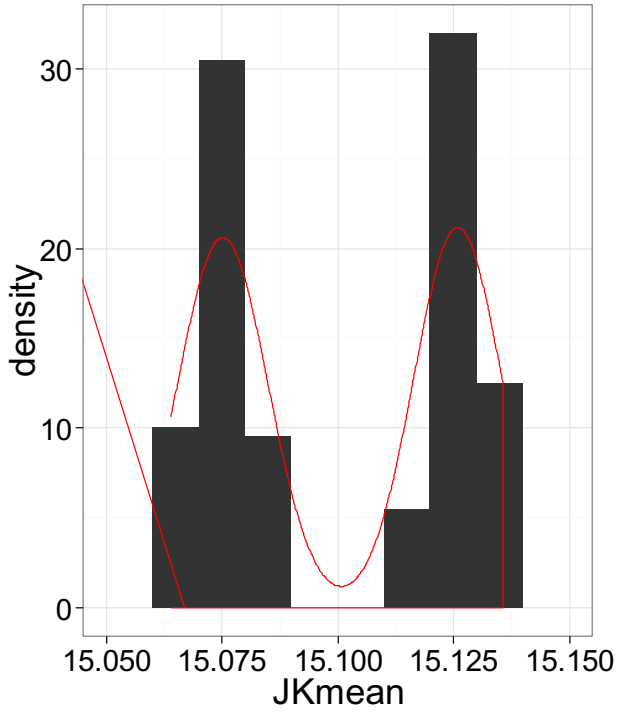
- When would bias be high and how might relate to how θ is defined?

Bimodal Distribution

Simulation Parameters
Gaussian Distribution
N= 30
True mean = 10
True SD = 1
+
N= 30
True mean = 20
True SD = 1

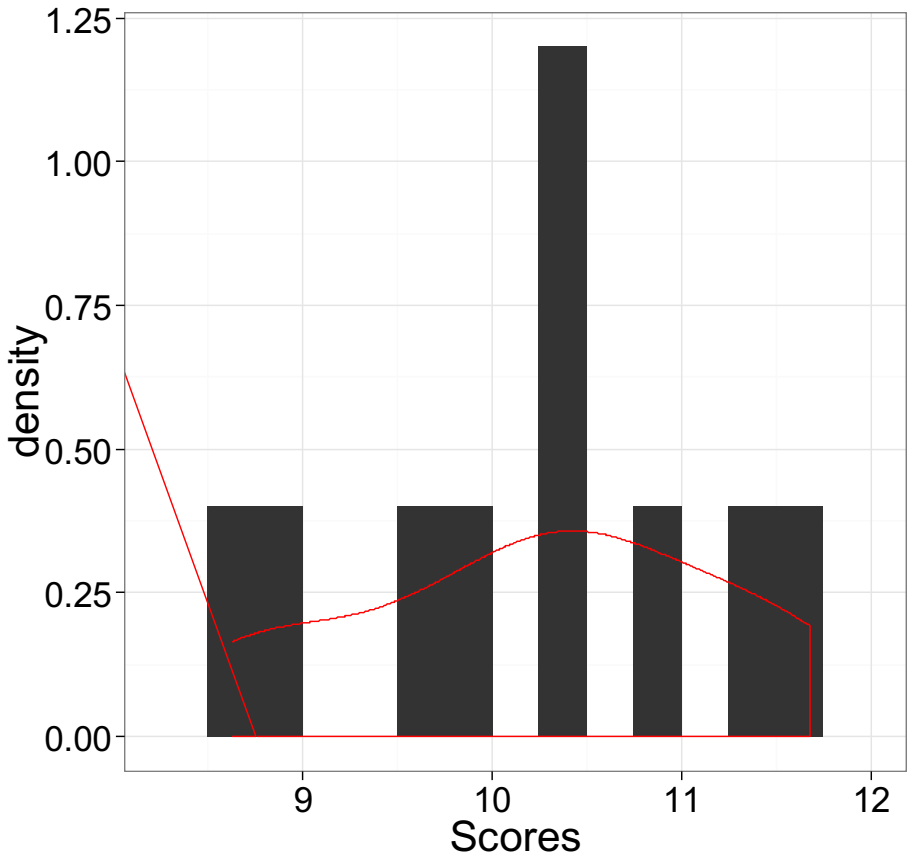


Jackknifed Distribution of Means



What would be bias here: High or Low?

Jackknife Mean [Small sample]



Simulation Parameters
Gaussian Distribution
N= 10
True mean = 10
True SD = 1

arithmetic mean = 10.239
arithmetic SD = 1.007

Jackknife mean = 10.239
Jackknife SD = 1.005

Jackknife Advantages & Limitations

Advantages

- Computationally cheap (could be done by hand)
- Calculate bias!
  - This concept becomes useful in correcting other resampling methods

Limitations

- Sampling without replacement
  - Results in non-normal distribution
  - The error term it calculates cannot always be used in significance testing
    - Cannot be used in calculating confidence intervals
- Works well with mean and SD, but can fail with other statistics such the median

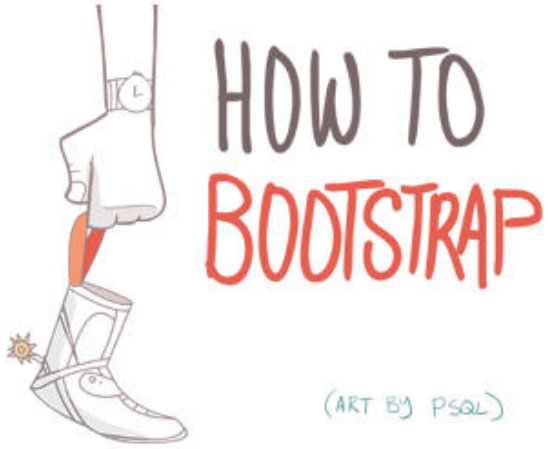
Birth of Modern Resampling Methods

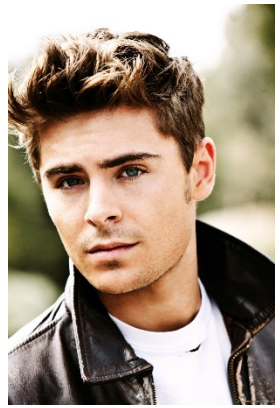
1970 era Main Frame

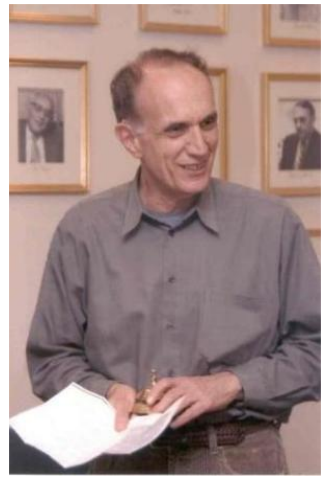
Bootstrapping

- Bootstrapping – “get (oneself or something) into or out of a situation using existing resources”
  - “Pull yourself up by your bootstraps”!

- 1979 Efron
  - Computationally intensive (expensive) method
    - “Conceptually” do Resampling ***WITH*** replacement
      - [Actually do a Taylor expansion]







The solution is in the computer?!

Bootstrap Logic

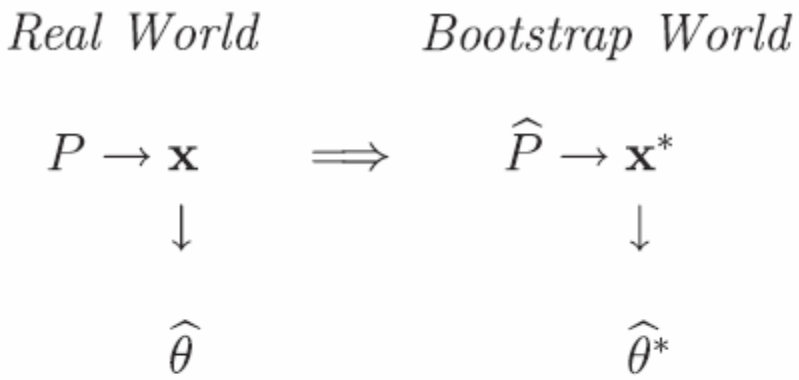


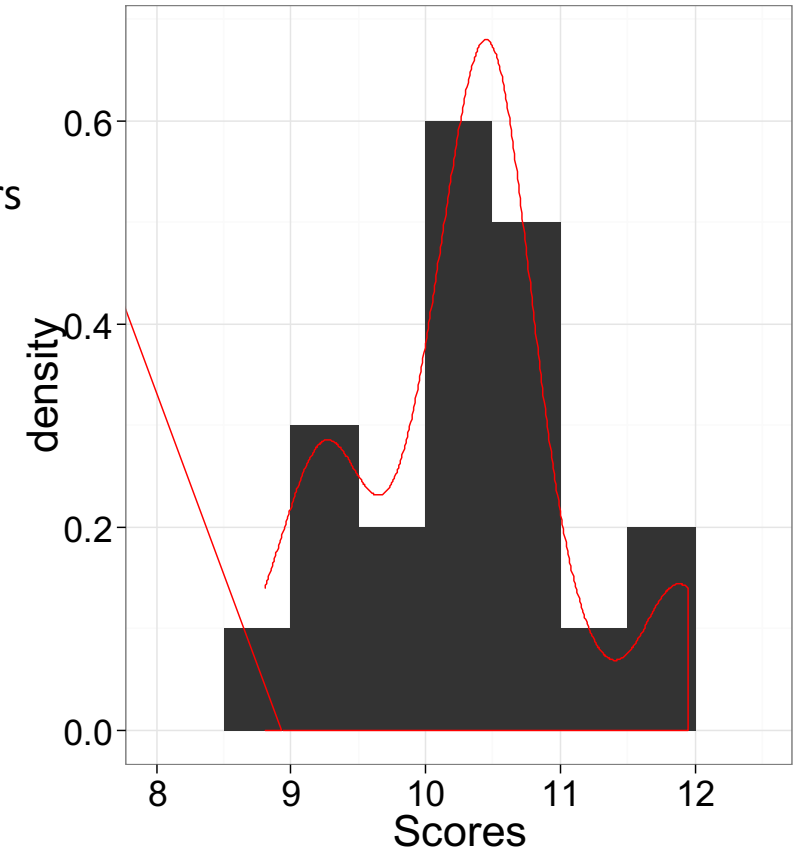
FIG. 1. *Typical bootstrap diagram. Unknown probability model P gives observed data $\mathbf{x}$ and we wish to know the accuracy of statistic $\widehat{\theta} = s(\mathbf{x})$ for estimating the parameter of interest $\theta = t(P)$. Point estimate $\widehat{P}$ for P yields bootstrap data sets $\mathbf{x}^*$. Accuracy is inferred from observed variability of bootstrap replications $\widehat{\theta}^* = s(\mathbf{x}^*)$.*

Bootstrap Basic set of θ

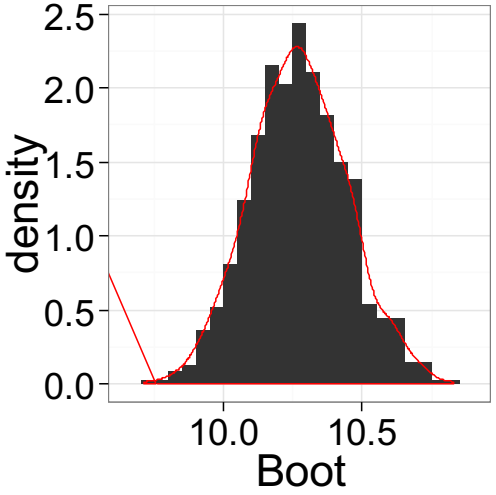
1000 bootstraps
[resampling's from original distribution]

Simulation Parameters
Gaussian Distribution
N= 20
True mean = 10
True SD = 1

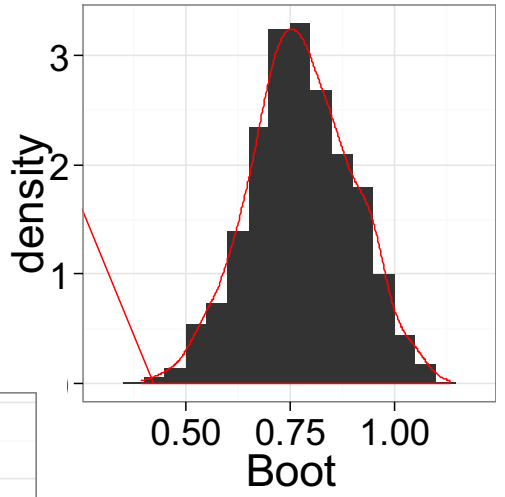
mean = 10.27
SD = .81
median = 10.38
CI95 = [9.89, 10.65]



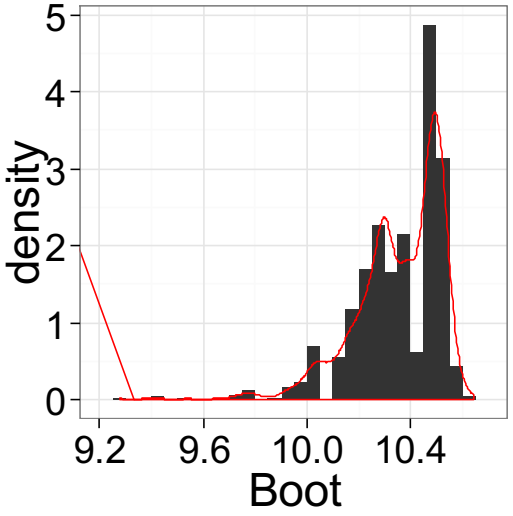
θ = Mean



θ = SD

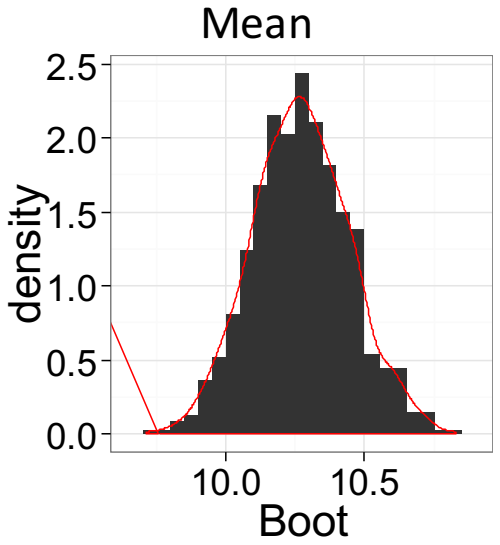


θ = Median



Error term on each θ [aka Early bootstrapping application]

- We can generate error terms for each statistic in which we are interested.
  - Classical methods we only had error on means!
  - Bootstrap we have an error on the Mean, SD value, median, kurtosis, anything
    - Based on the resampling of actual observed data and NOT our assumption of normal distribution of the means*



Classical M = 10.27, SD =.811 , N =20 , $SE = {}^{\sigma}/_{\sqrt{N}}$ =.181

Bootstrapped: M = 10.27, SE = SD(bootstrapped means) = .173*

*Standard approach to determining error applies in cases of normality

Construct custom bootstrapped "t" tests [aka Early bootstrapping application]

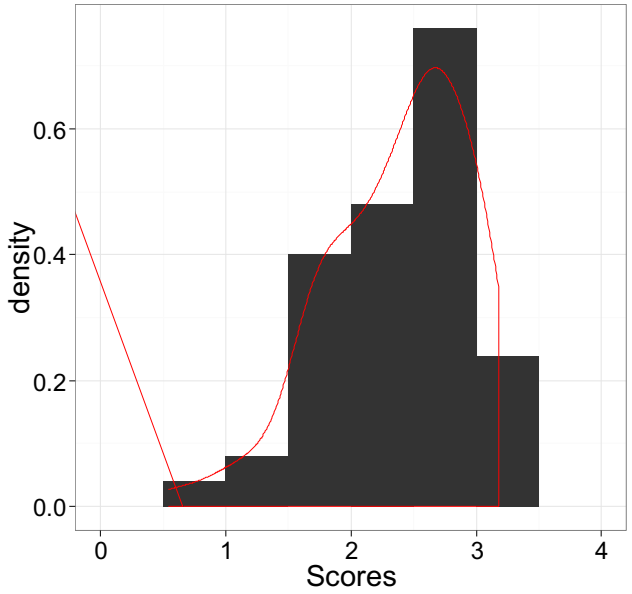
$$Classical\ t = \frac{\bar{x} - \mu}{SE}$$

$$\text{b}ootstrapped\ t = \frac{\bar{\theta}_B - \bar{\theta}}{SE_B}$$

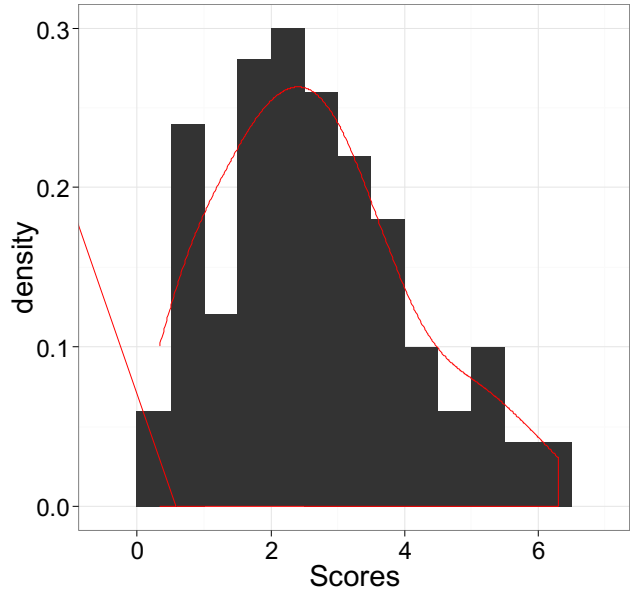
Reflects mean and CLT assumption

θ could be any point estimator (mean, SD, median, etc)

Example of a bootstrapped "t" tests [aka Early bootstrapping application]



Group 1,
M = 2.36

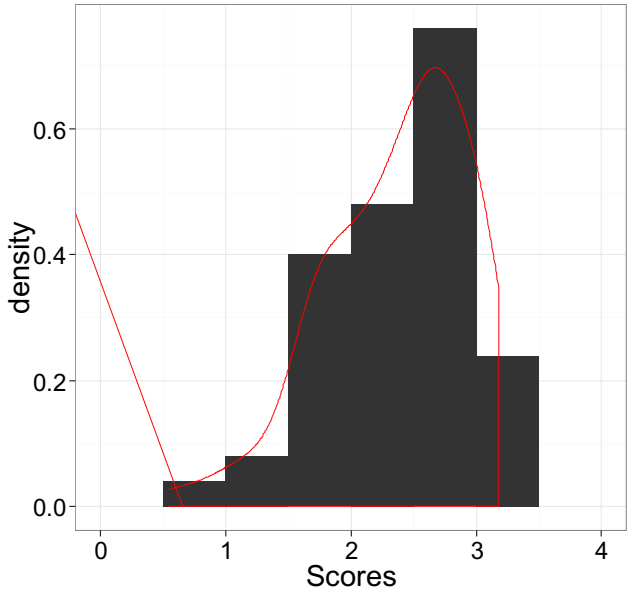


Group 2,
M = 2.66

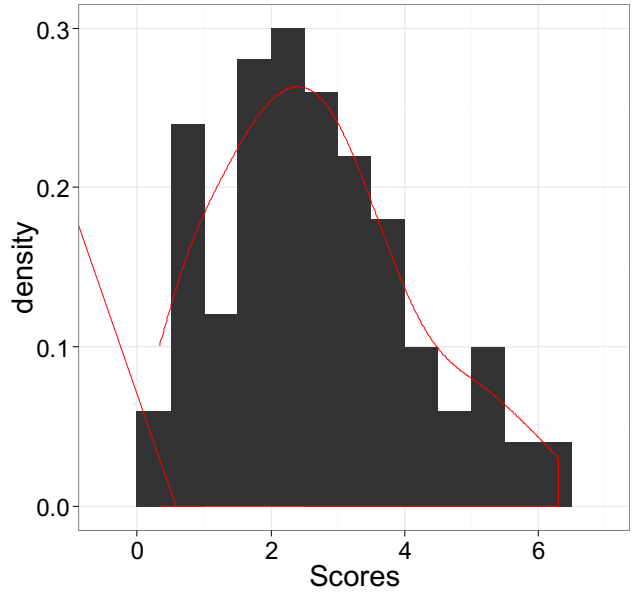
Is the mean of group 1 different from group 2?

Classical t-test on Means: t(198) = 1.921 , p = .057 (or .058 correcting for homogeneity of variance)

Example of a bootstrapped "t" tests [aka Early bootstrapping application]



Group 1,
M = 2.36

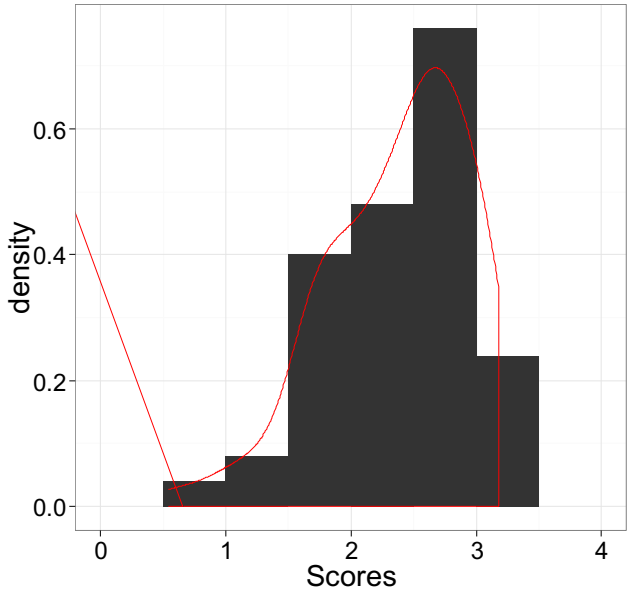


Group 2,
M = 2.66

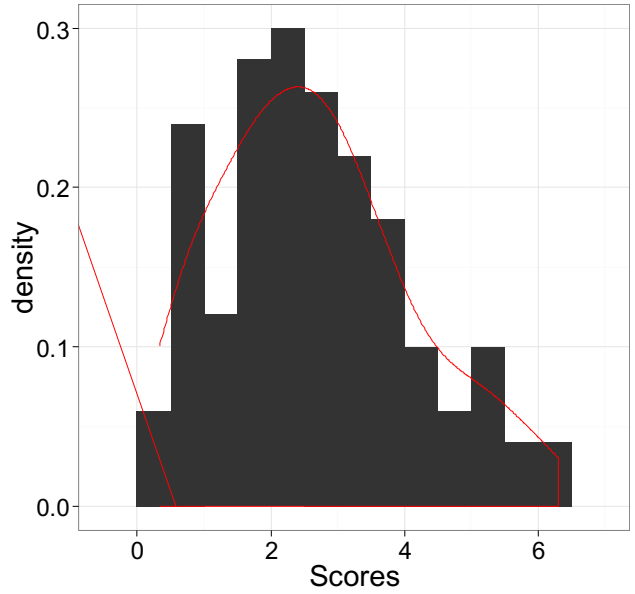
Is the mean of group 1 different from group 2?

Bootstrapped t-test on Means: p = .059

Example of a bootstrapped "t" tests [aka Early bootstrapping application]



Group 1,
Median = 2.48



Group 2,
Median = 2.51

Is the ***median*** of group 1 different from group 2?

Bootstrapped t-test on MEDIAN: p = .99. The answer is clearly NO!

Advantage of Bootstrapping

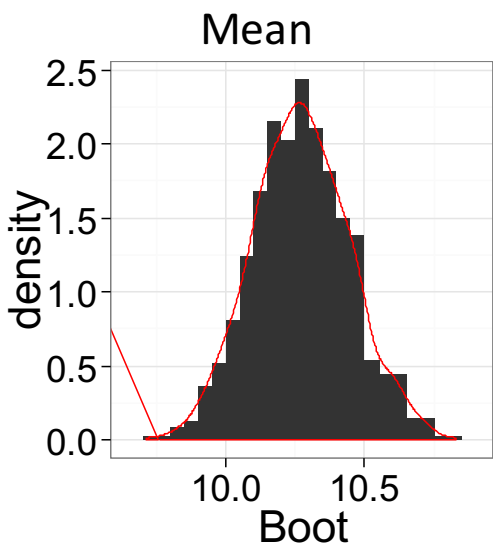
Since Bootstrapping does not rely on CLT, the point estimates and error estimates tend to better represent the population parameters.

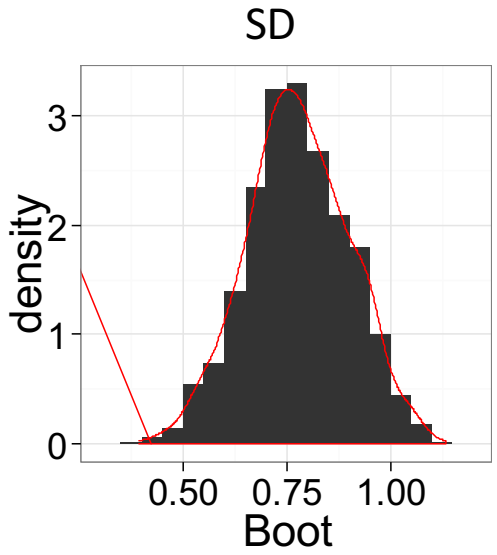
Most common modern usages: Get accurate confidence intervals on any θ!

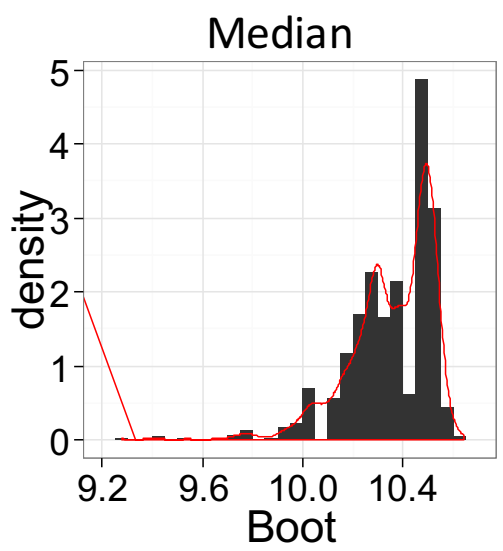
- Better estimate results in small samples
- You can test the difference between groups using any "statistic" you need
  - Means are no long KING!
  - You can test the difference between any custom "metric"
- Combine jackknife and bootstrap
  - use jackknife to estimate bias ('controls for outliners') and 'correct' your bootstrap to make it more accurate.

Confidence of each θ!

- We can generate confidence interval for each statistic in which we are interested.
  - Classical methods we only had CIs on means!
  - Bootstrap we have an CIs on the Mean, SD value, median, kurtosis, anything
    - Based on the resampling of actual observed data and NOT our assumption of normal distribution of the means*







Classical M = 10.27, CI95* = [9.89, 10.65]
Bootstrapped: M = 10.27, CI95 = [9.92, 10.61]

SD = .811
SD = .777, CI95 = [.60, 1.11]

Median = 10.38
Median = 10.35, CI95 = [9.93, 10.54]

Types of Bootstrapped Confidence Intervals

Type 1: Standard type AKA Normal type [not in SPSS]

- $\theta \pm Z\sigma_B$
- Produces symmetrical CI
- Worst of all type!

Type 2: Studentized AKA tbootstrap [not in SPSS]

- $\theta \pm t\sigma_B$
- Still produces symmetrical CI
- Not much better than type 1

Type 3: Percentile methods

- Take bootstrap distribution and find percentiles (usually, .025 and .975)
- Produces NON-symmetrical CI
- Better than type 1, 2

Types of Bootstrapped Confidence Intervals

Type 4: Bias corrected methods

- Do type 3
- Correct them for bias (jackknife) and for deviation from normality ('accelerated')
  - Bca method (in SPSS) [Bias corrected and accelerated]
- Produces NON-symmetrical CI (that are narrower then non-parametric)
  - ABC method [Accelerated, Bias corrected, and nonlinearity Coefficient
    - (alterative form of correction that is newer then Bca – not in SPSS but in R)
      - Note: You can do ABC and Bca non-parametric bootstrap as well (but not in SPSS)

Type 4 is best when you can do the ***correction***

- Correction is not always possible, so best to do type 3.

Types of Bootstrapped Confidence Intervals [Efron, 2003]

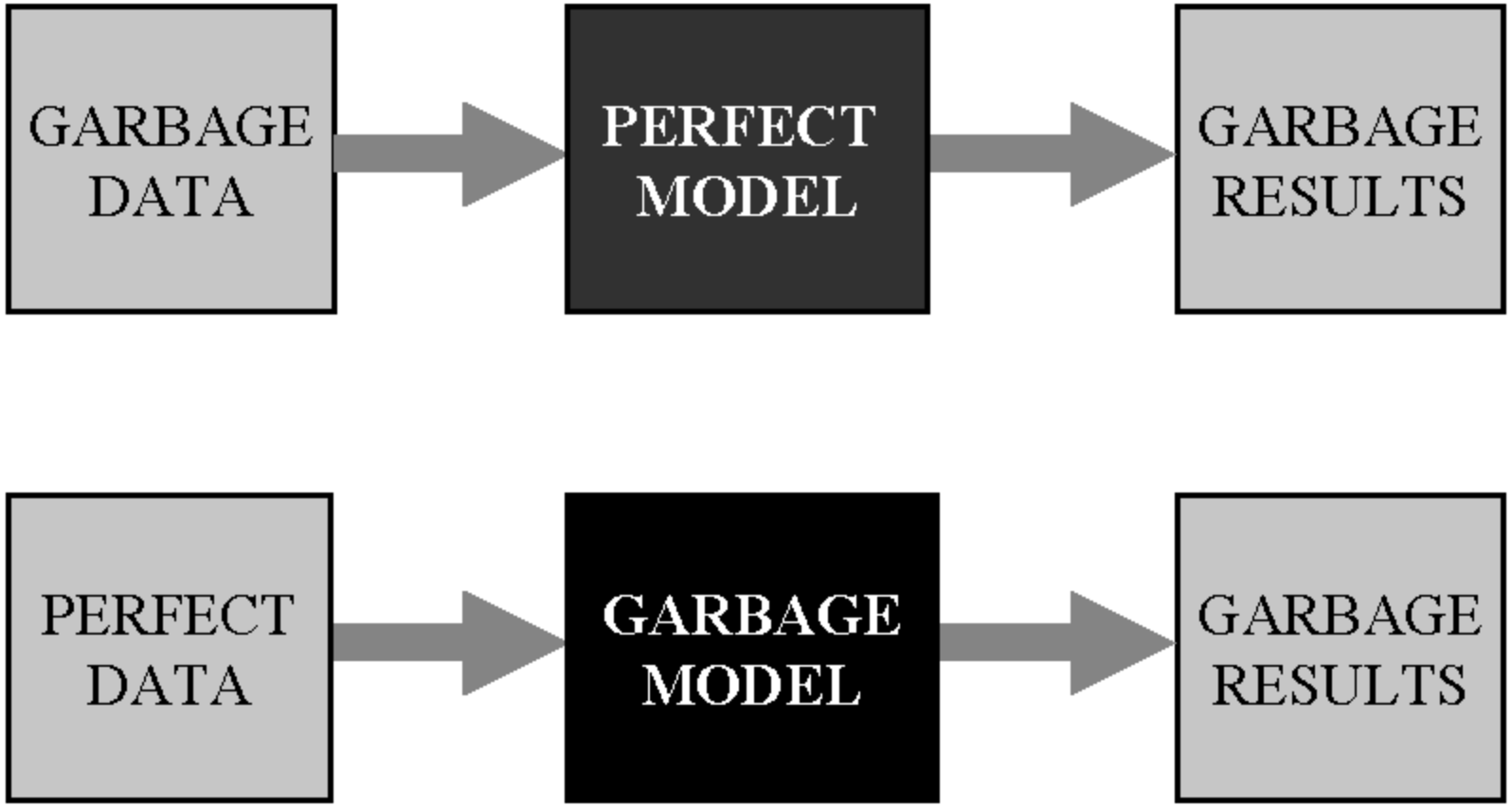
TABLE 1

Exact and approximate confidence intervals for the correlation coefficient of a bivariate normal sample, $n = 15$, with sample correlation coefficient 0.562. The tail area is the actual probability of exceeding 0.562 when the parameter value is the corresponding interval endpoint

	Limits		**Tail areas**	
	0.05	**0.95**	**0.05**	**0.95**
Exact	0.155	0.790	0.050	0.950
Parametric ABC	0.158	0.788	0.051	0.948
Nonparametric ABC	0.188	0.775	0.063	0.935
Standard	0.271	0.830	0.112	0.980

MODEL CALCULATIONS

”Garbage In-garbage Out” Paradigm



Other Resampling Methods

- Monte-Carlo simulations (very computationally expensive)
  - Broad class of methods becoming used in advanced modern statistics
  - Parametric Bootstrapping – Assume the data follow a normal distribution given the parameters extracted from the data
  - Permutations– Used for calculating pvalues for non-parametric tests (chi-square, binomial tests)

- Surrogate data analysis (very computationally expensive)
  - Common in time series analysis

- What can we do with all these modern methods?
  - Option 1: Salvage classical statistics hypothesis testing (journals are requesting bootstrapping already!)
  - Option 2: Resampling methods as gate way drug: abandon classical methods and switch to Bayesian methods
  - Option 3: Ask questions no one has ever been able to ask people!
    - Dual brain synchronization during joint action! [merging EEG, bootstrapping, and surrogate methods]

Modern uses of Bootstrapping

- Difference in Variances and Correlations:
  - With Bootstrapping you can check for differences between groups based on their variability!
    - CI around the point estimate of variance/correlation for a particular group and compare that to point estimate of variance/variance/correlation for another group
    - Devise an question that you could ***not*** ask with classical stats, but you ***can*** with bootstrapping.
- Regression-Based Approaches
  - Currently required for modern causal modeling (mediation and moderation, SEM)
  - Becoming required for getting pvalues in mixed models
- Signal processing/Time-series
  - Already used heavy in:
    - neuroscience (single cell recordings, modern types of fMRI/EEG, and computation neuro)
    - motor control
- Bayesian methods
- Machine Learning