Resampling: The New Statistics

Frank Schieber

Heimstra Labs Colloquium

4 February 2013

The “Sampling Distribution” is the foundation of Statistical Inference

The sampling distribution represents the relative frequency of all possible values of a statistic given a well-defined set of conditions.

It is this knowledge that allows us to discriminate “likely” vs. “unlikely” (significant) events.

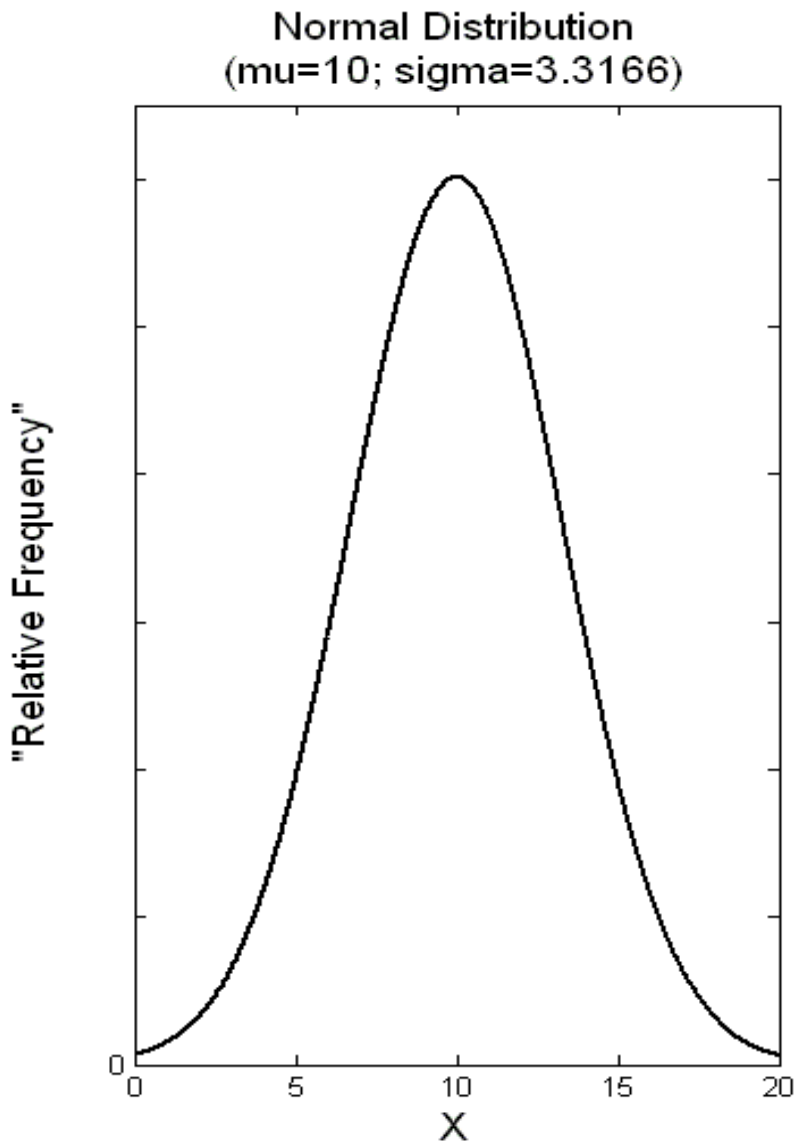
The most commonly used statistics have well-known, mathematically-defined sampling distributions: e.g., mean, binomial proportion, difference between sample means, etc.

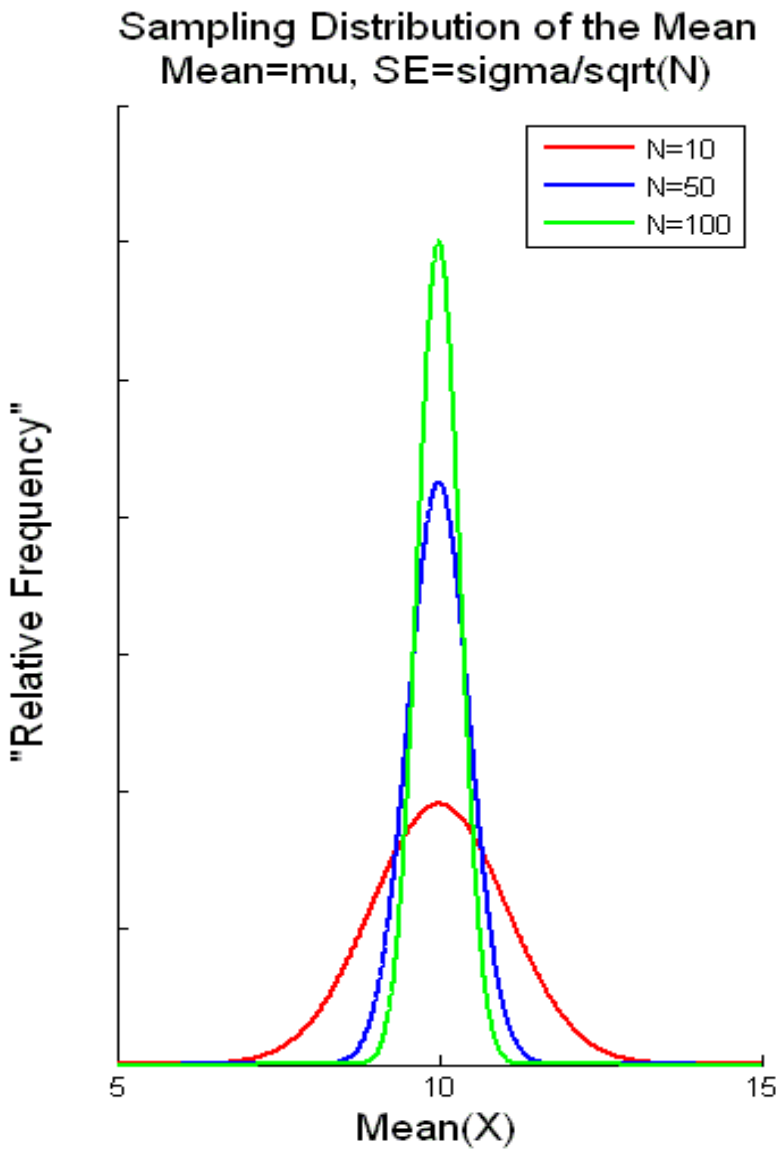
Population	Sample statistic	Sampling distribution
Infinite, $X \sim N(\mu, \sigma^2)$	Sample mean, $\bar{X}$	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
Finite (size N), $X \sim N(\mu, \sigma^2)$	Sample mean, $\bar{X}$	$\bar{X} \sim N\left(\mu, \frac{N-n}{N-1} \times \frac{\sigma^2}{n}\right)$
Infinite, $X \sim \mathrm{Binomial}(p)$	Sample proportion, $\bar{p}$	$\bar{p} \sim \mathrm{Binomial}(p)$
Infinite, $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$	Sample difference between means, $\bar{X}_1 - \bar{X}_2$	$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

The Sampling Distribution of the Mean

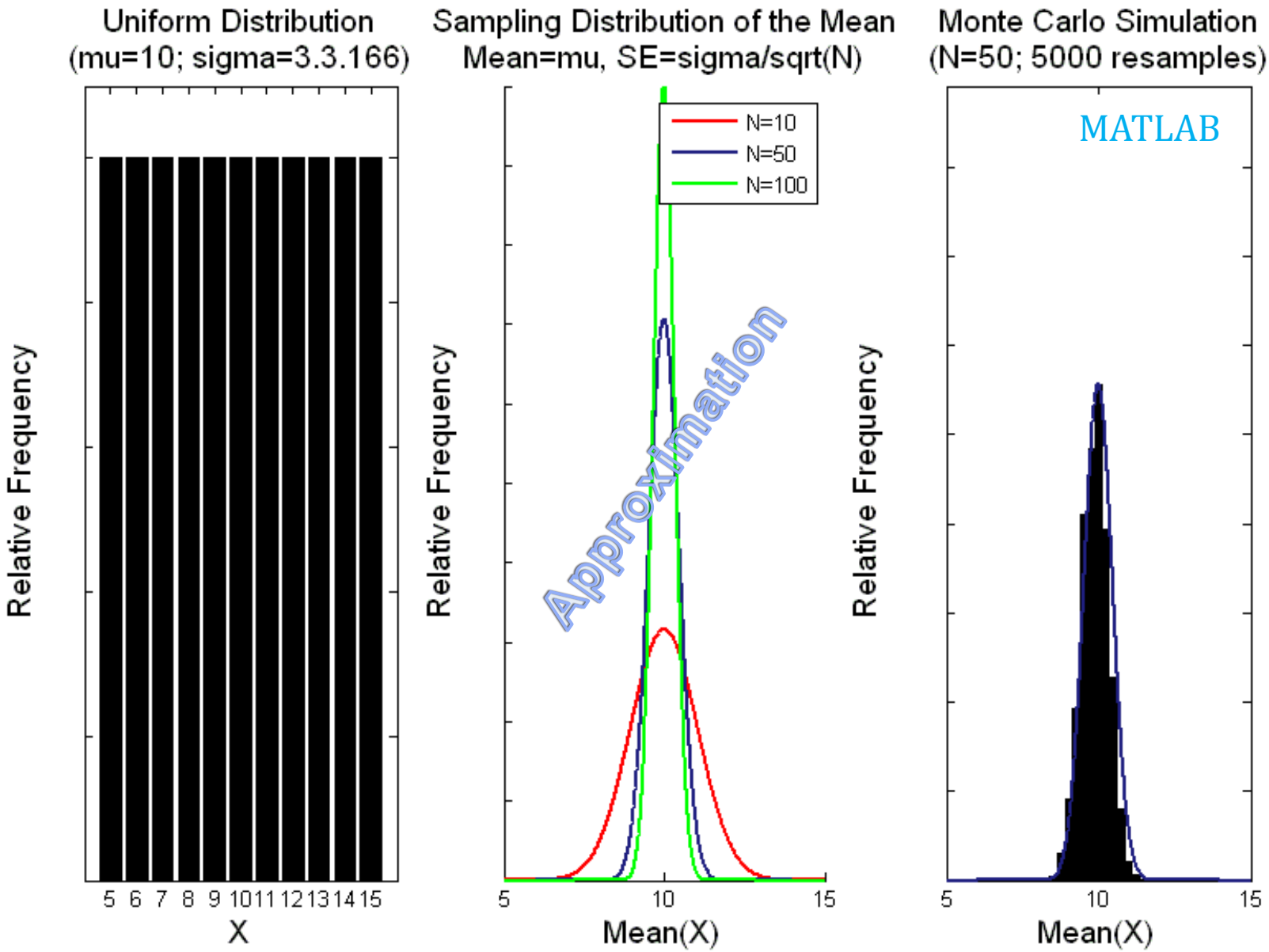
(Population is Normally Distributed)





The Sampling Distribution of the Mean

(Population is Uniformly Distributed)



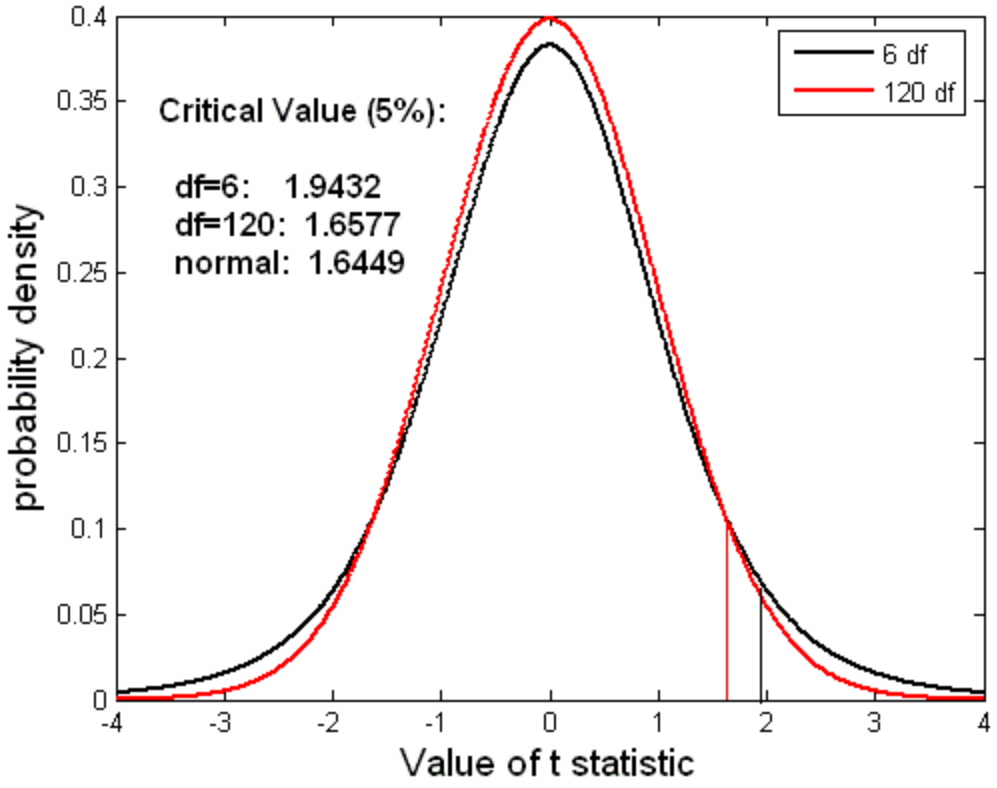
MATLAB Code Snippet

Monte Carlo Generation of Sampling Dist.

```
%Generate 5000 random samples of N=50 from a
%Uniform Distribution and save means in a list

universe=[5 6 7 8 9 10 11 12 13 14 15];
for i=1:5000
    rs=randsample(universe,50,true); %with replacement
    remeans(i)=mean(rs);  %plug-in statistic of interest
end
```

"Student's t"

Theoretical sampling distribution for the difference between means assumed to have been sampled from the same population (Assumes normality)



Let's Revisit the Sampling Distribution of a (any) Statistic

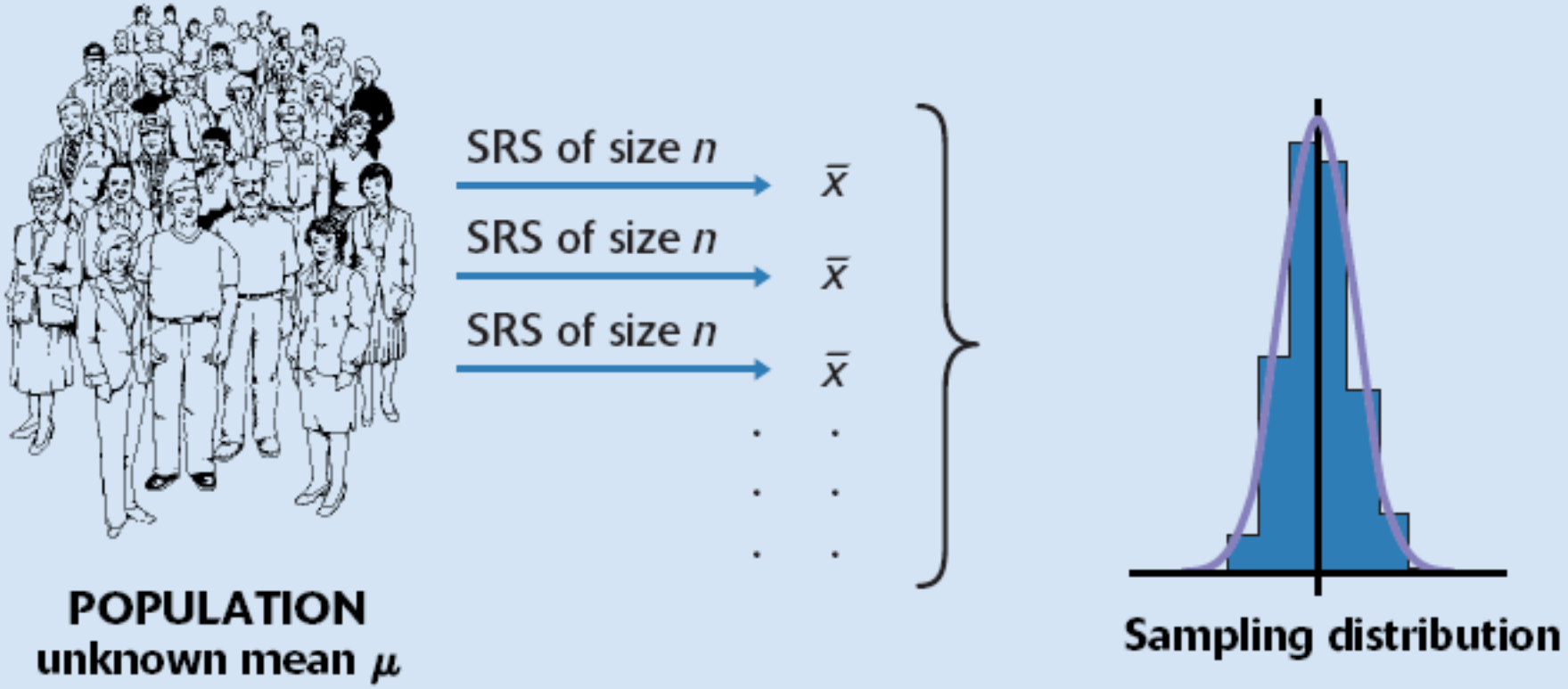
Three ways to generate a Sampling Distribution

Ideal Sampling Approach

Take a large number of samples of size N from your population of interest and generate a custom sampling distribution for your statistic

Best approach; Can be applied to ANY statistic;

Unrealizable (silly?) given normal constraints



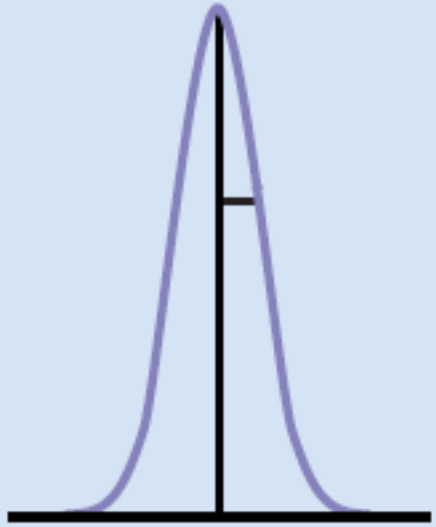
Theoretical Probability Approach

Apply mathematical theory to generate expected distributions of a given statistic. This is the traditional approach we all know and love(?)

Useful approach; Limited to "well" behaved statistics; opaque; Depends upon assumptions that are often difficult to evaluate

NORMAL POPULATION
unknown mean μ

Theory

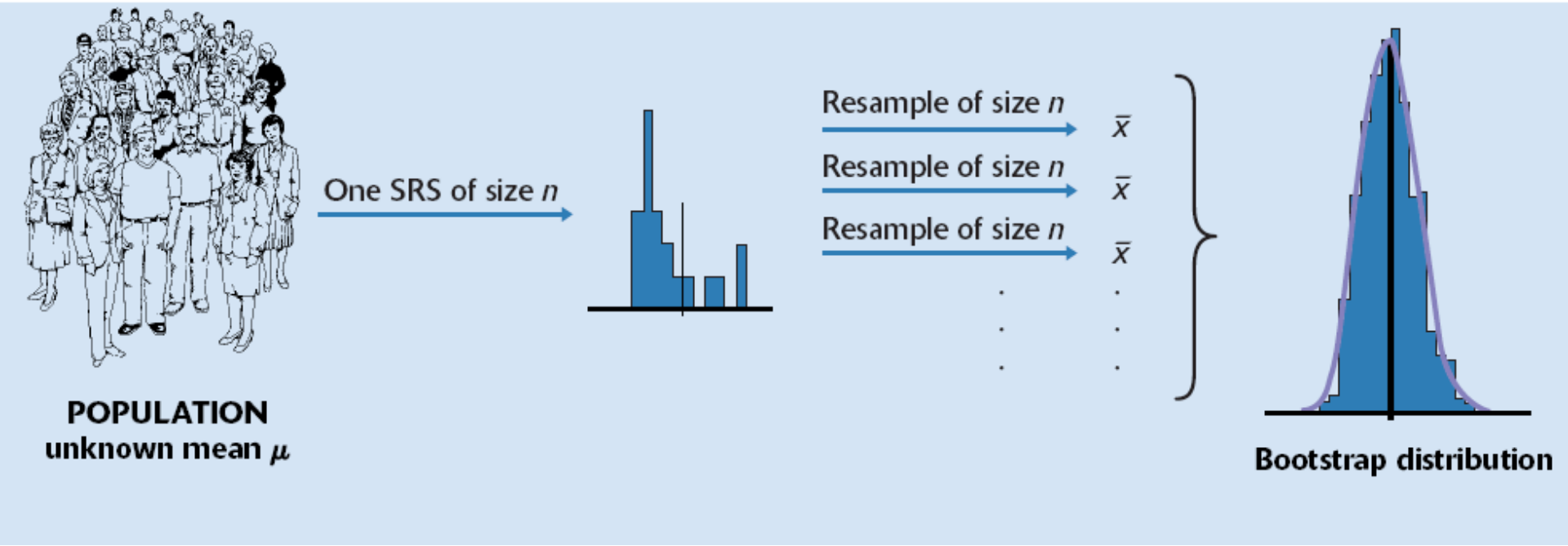


Sampling distribution
(t Distribution)

Resampling Approach

Draw a sample from your target population(s) and use Monte Carlo techniques to randomly resample in order to generate an empirically derived estimate of the sampling distribution of your statistic.

Computationally and conceptually simple; minimal reliance upon mathematical theory ("brute force"?); Can be applied to ANY statistic (median; ratio; mode); Makes bias, shape and spread of sampling distribution observable



Resampling Approach to Statistical Inference

What is Resampling?

- Resampling refers to a variety of statistical methods based on available data (samples) rather than a set of standard assumptions about underlying populations.
- Such methods include **bootstrap**, jackknife, and **permutation tests**.
- Resampling represents a “new” idea about statistical analysis which is distinct from that of traditional statistics.

Why resampling?

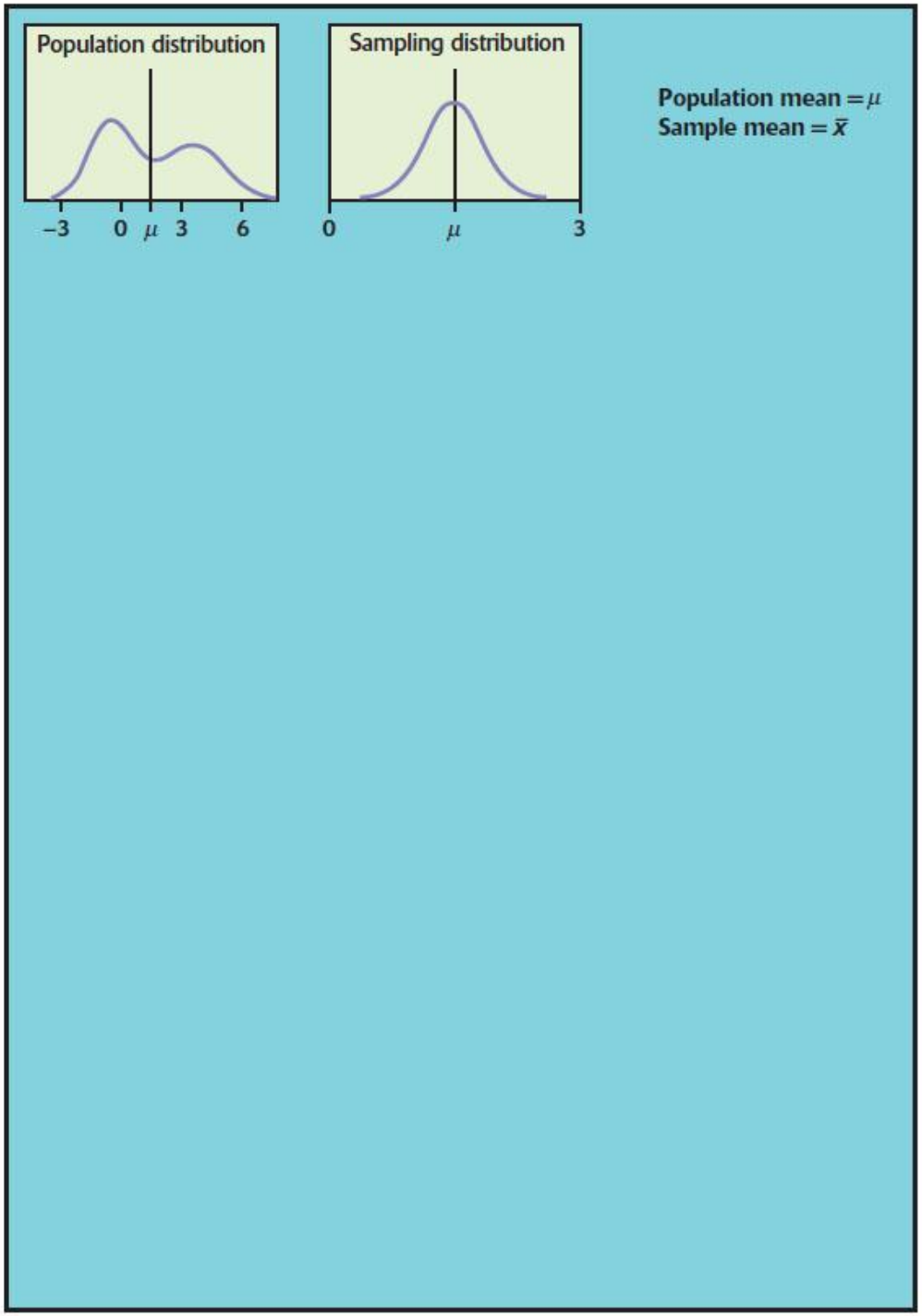
- In most cases we accept the assumptions for traditional statistics “as if” they are satisfied.
- Traditional statistics cannot be applied to some “awkward” but “interesting” statistics such as the median, mode, range, ratio, trimmed mean, etc.
- Generality of resampling saves us from onerous formulas for different problems.
- Permutation tests, and some bootstraping methods, are more accurate in practice than traditional methods.

What is bootstraping?

- Bootstraping is a statistical method for generating the sampling distribution of a statistic by sampling with replacement from the original data sample.
- Bootstraping can also be exploited to estimate confidence intervals and to conduct null hypothesis testing.

Procedure of bootstraping

- Get an original random sample from the population of interest.
- Create hundreds of new samples, called bootstrap samples (or resamples), by sampling with replacement. Each resample is the same size as the original random sample.
- Calculate the statistic for each resample. The frequency (probabiliy) distribution of these resample statistics is called the bootstrap distribution.
- Use the bootstrap distribution for establishing confidence intervals and/or null hypothesis testing.

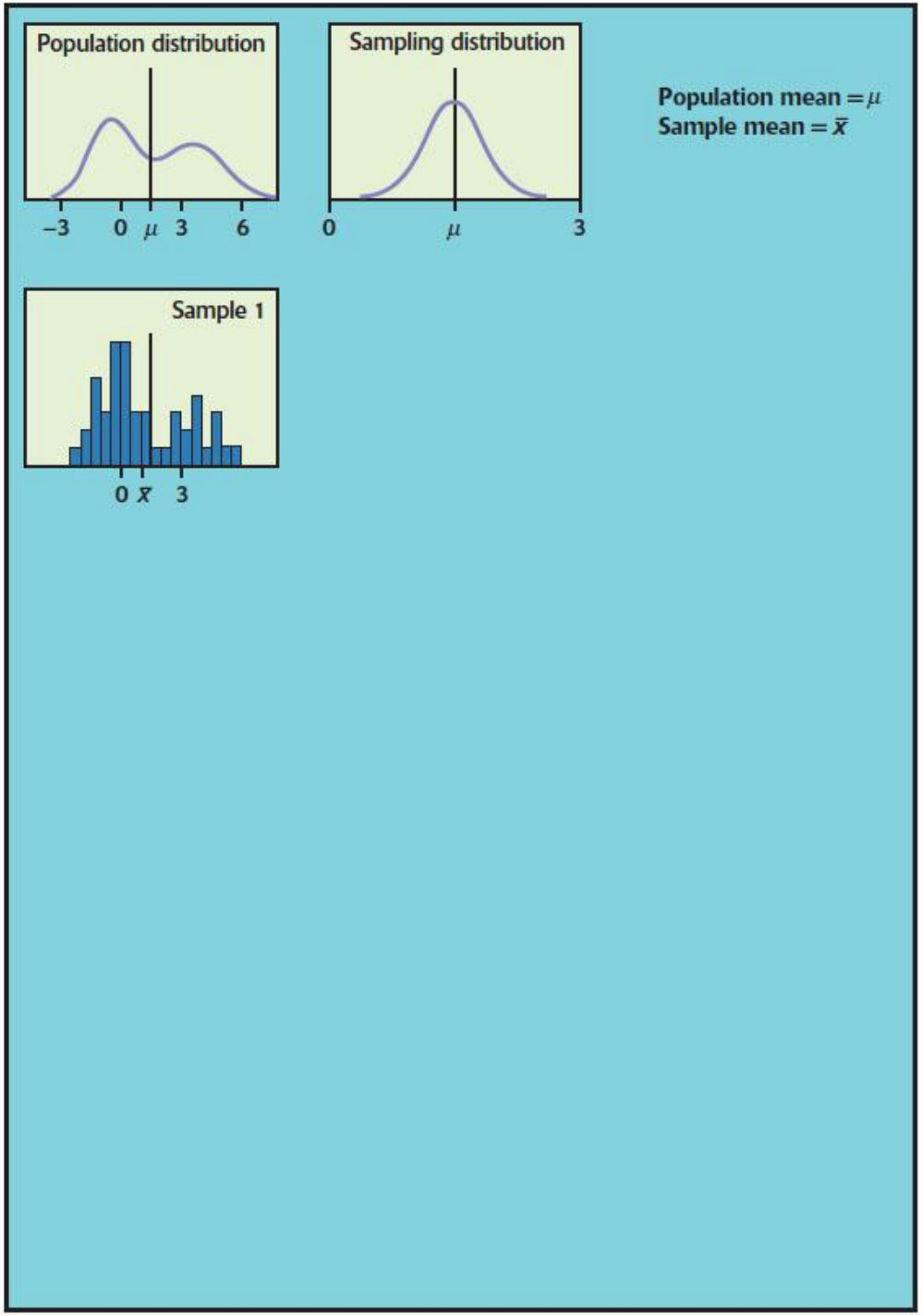


Let's start with the Ideal Approach for the Mean

Specify population
Take 1,000 random samples
Generate probability distribution
for all possible values of mean
for sample size (n) = 50

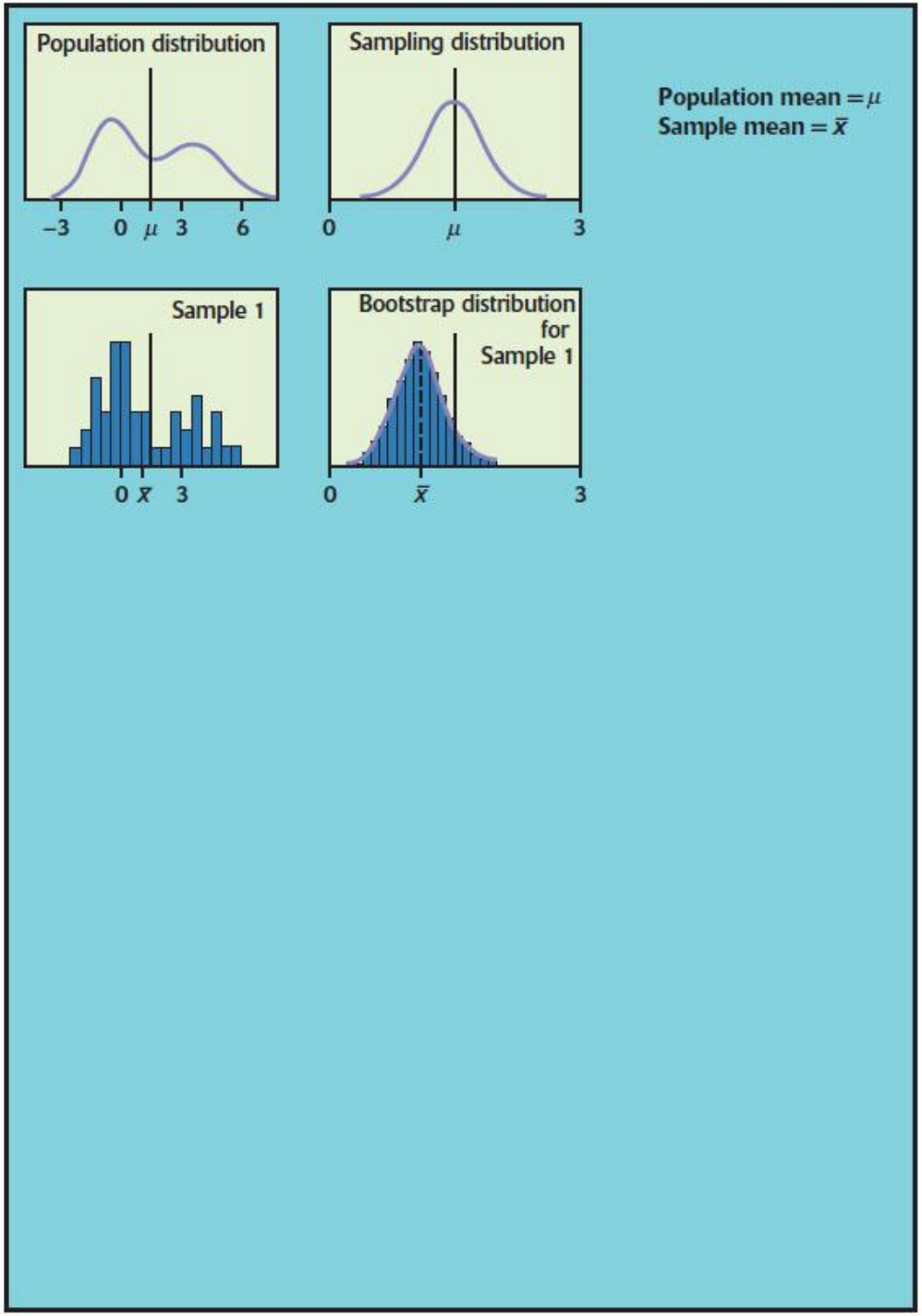
Theoretical approach

Assumes normality
E(M) of sampling dist = μ
S.D. of sampling dist = δ/√n
Normality assumption less important as N increases



Now let's explore the
Bootstrap Resampling Approach

Begin by taking a simple random sample from the target population (n = 50)



Now let's explore the
Bootstrap Resampling Approach

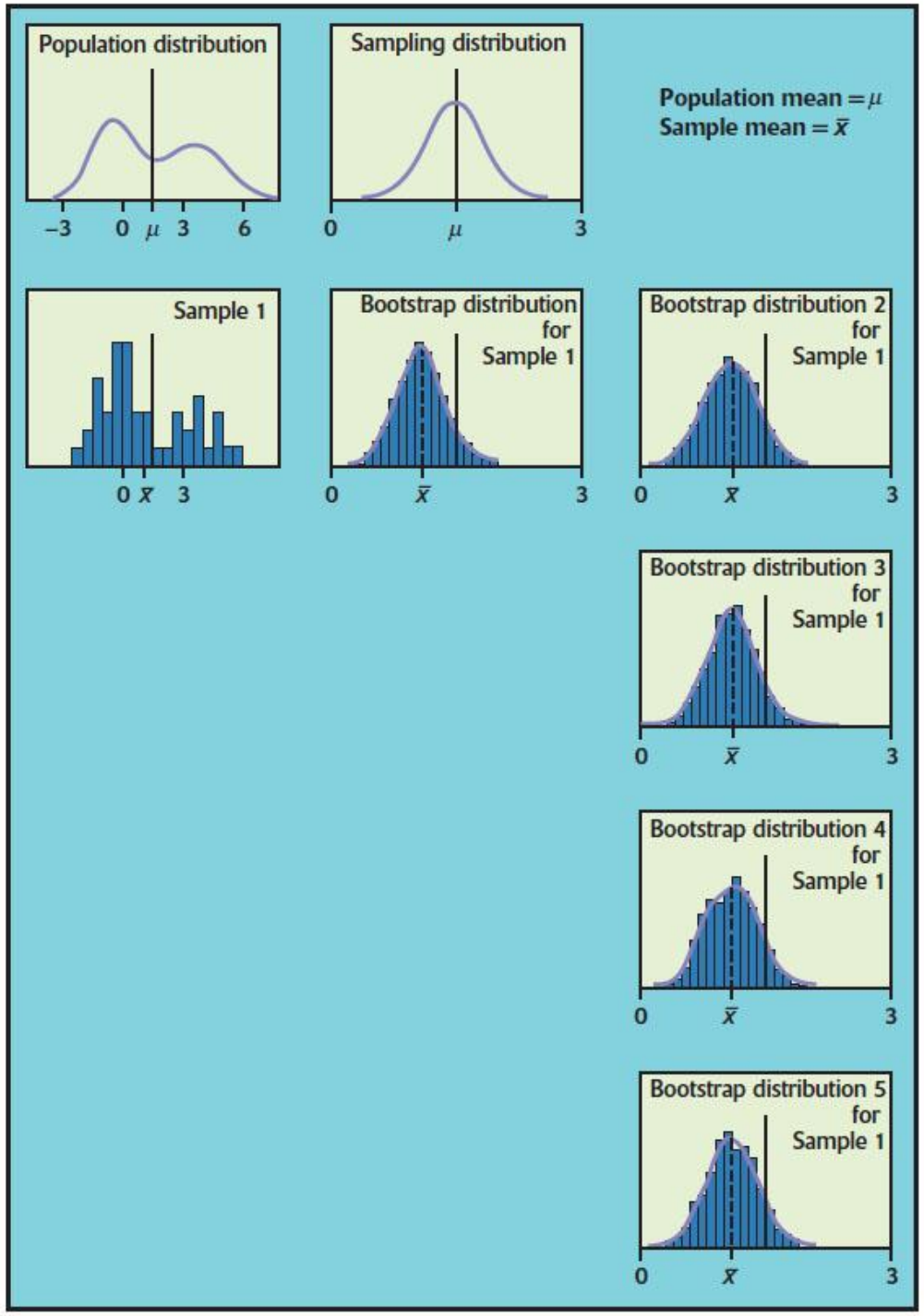
Begin by taking a simple random sample from the target population (n=50)

Next,

Draw 1000 (re)samples from the simple random sample (SRS);
Compute a mean for each resample;
Generate distribution of resampled means

All resamples of size n (=50)
Resampling is (obviously) sampling with replacement

Shape and spread of bootstrap distribution approximates the true (ideal) sampling distribution very well.

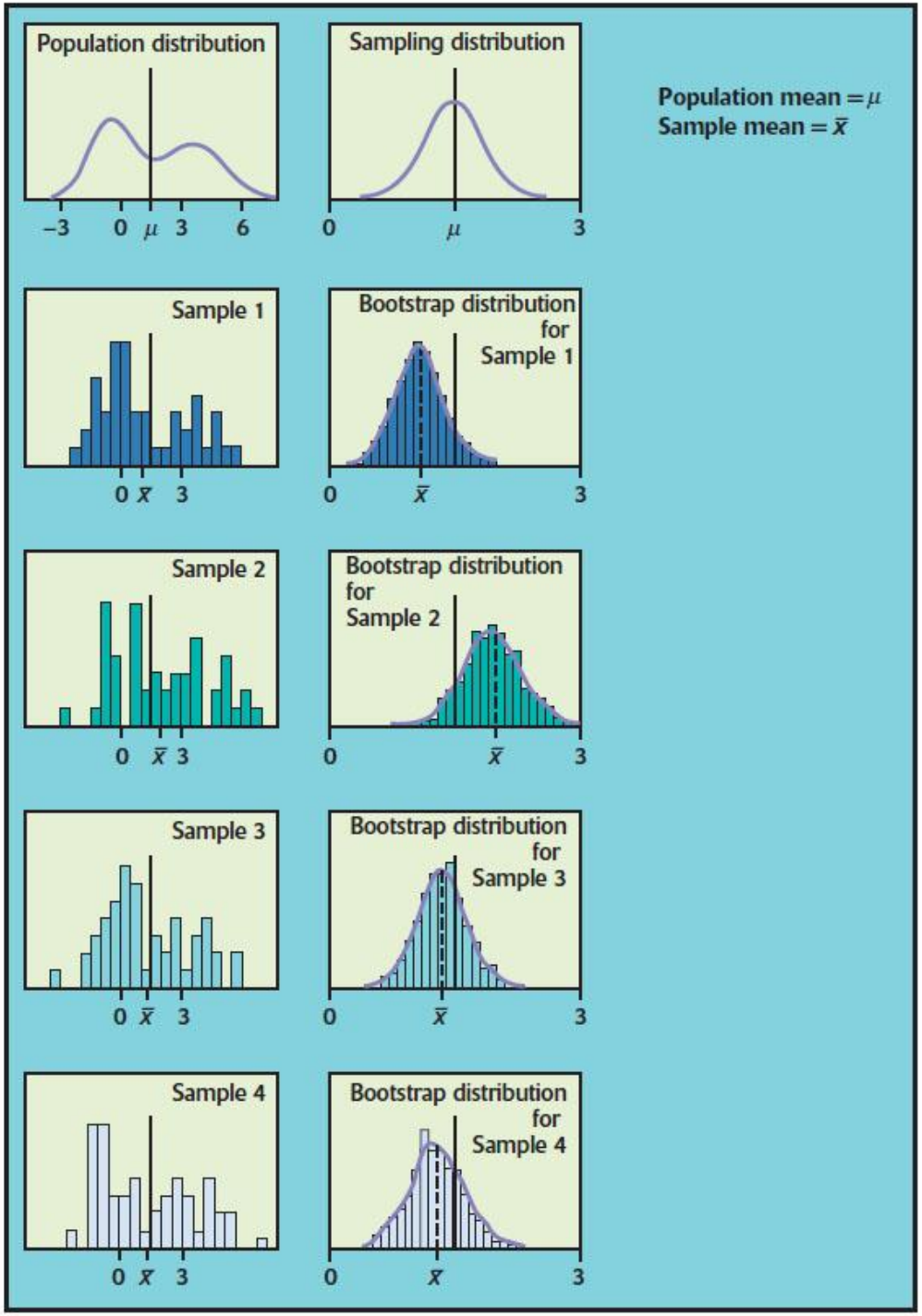


Now let's explore the
Bootstrap Resampling Approach

If we repeat the resampling procedure several times using the same original random sample from the population…

We can see that the amount of uncertainty introduced by the resampling procedure is minimal for 1000+ resamples

That is:
The shape, spread and bias is preserved across all five replications of the bootstrap distribution



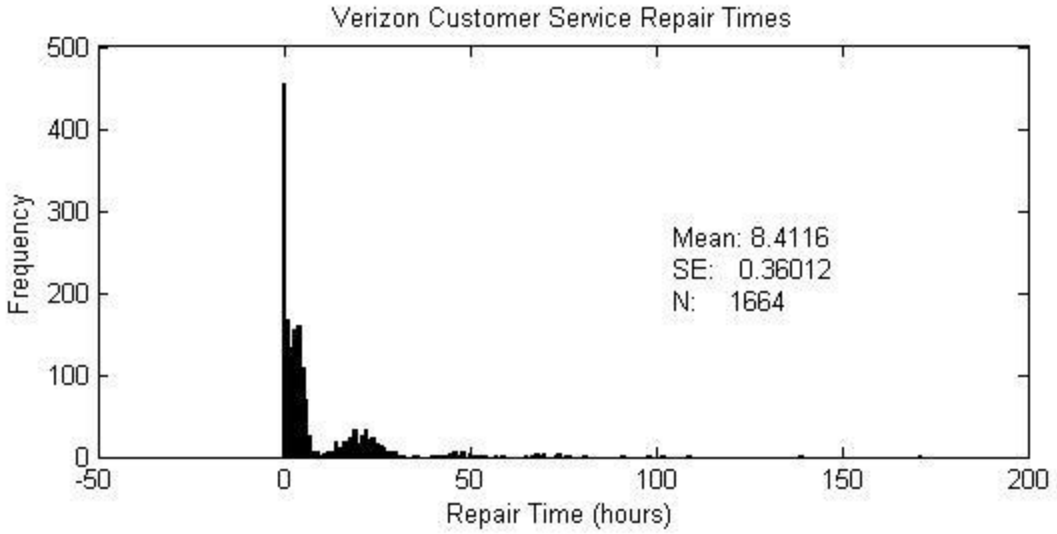
Now let's explore the
Bootstrap Resampling Approach

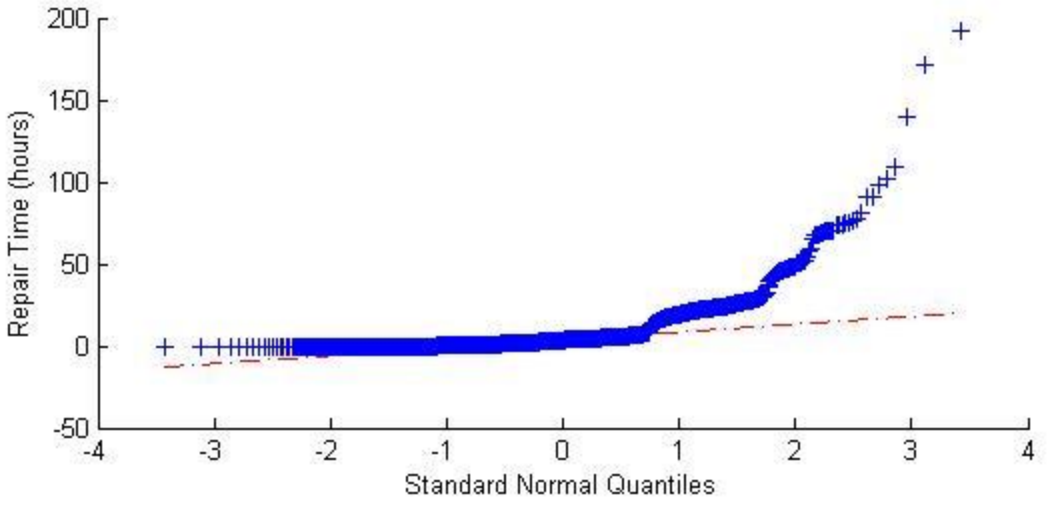
If we draw three additional simple random samples (n=50) from the target population…we can visualize the uncertainty introduced by the conventional sampling process

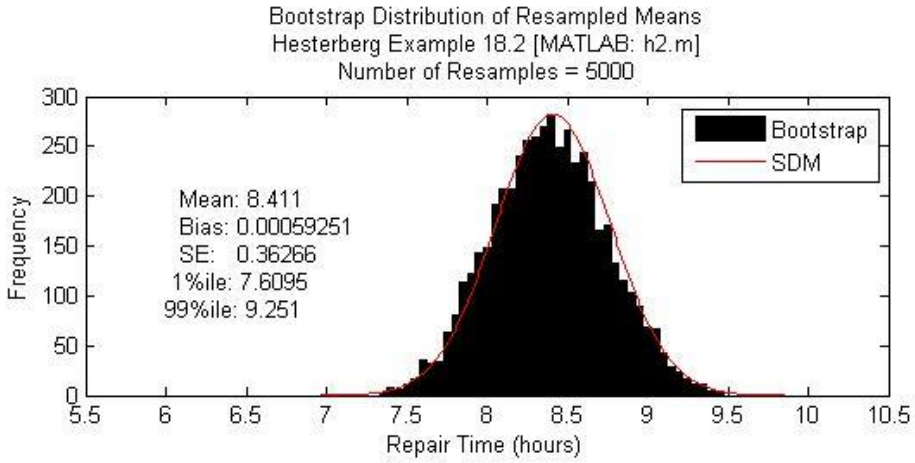
Variations in the shape and bias of the resulting bootstrap distributions reflect these uncertainties (which are exacerbated by violations of the normality assumption)

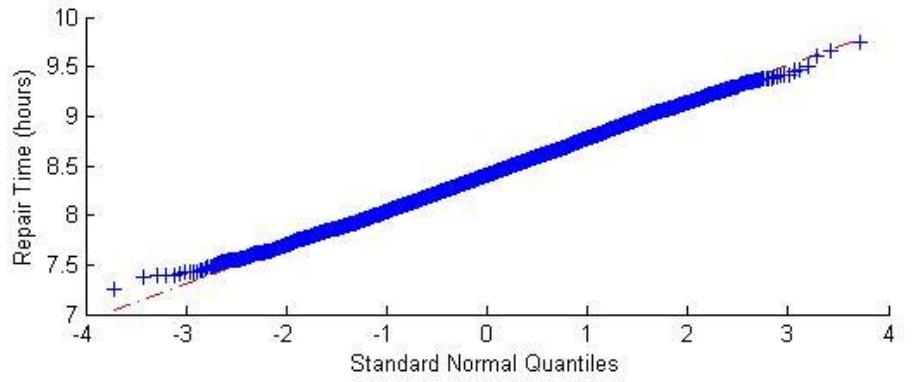
The bootstrap distribution makes these distortions directly observable. Many techniques for dealing with these distortions have emerged as understanding of resampling techniques accummulates

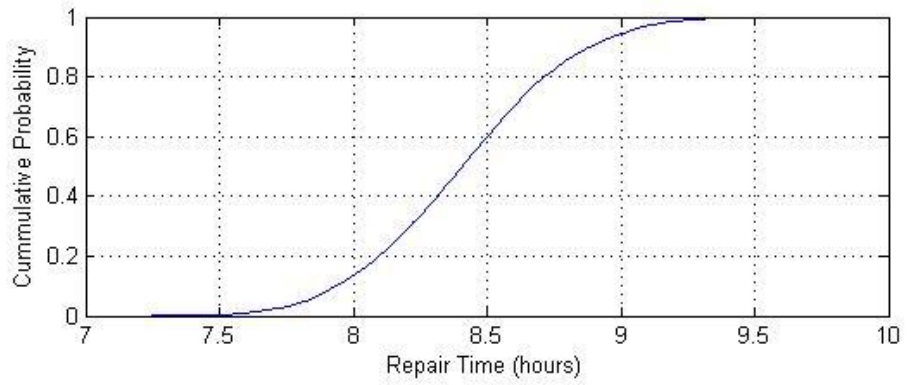
Bootstrap Example [h2.m]











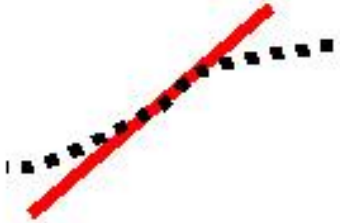
Bootstrap Example

[h2.m]

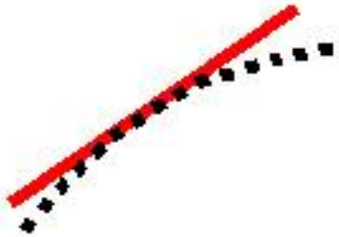
Quantile-Quantile Plot

Plot percentiles of distribution X against percentiles of the standard normal distribution
(Linear if from the same distribution)

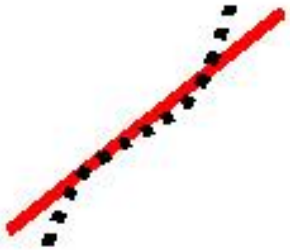
Right Skew



Truncated Tails



Left Skew



Extended Tails

2-Sample Applications of Bootstraping

1. Draw original SRS's of size n and m from each population
2. Generate resamples of size n and m from each SRS; compute statistics; compute and record differences between these stats
3. Repeat Step #2 1000 times
4. Construct bootstrap distribution and use it to evaluate the difference between the original two SRS's

The Permutation Test

(aka Randomization Test)

A resampling approach to

Null Hypothesis Testing

While the Bootstrap derives its power from the process of random sampling

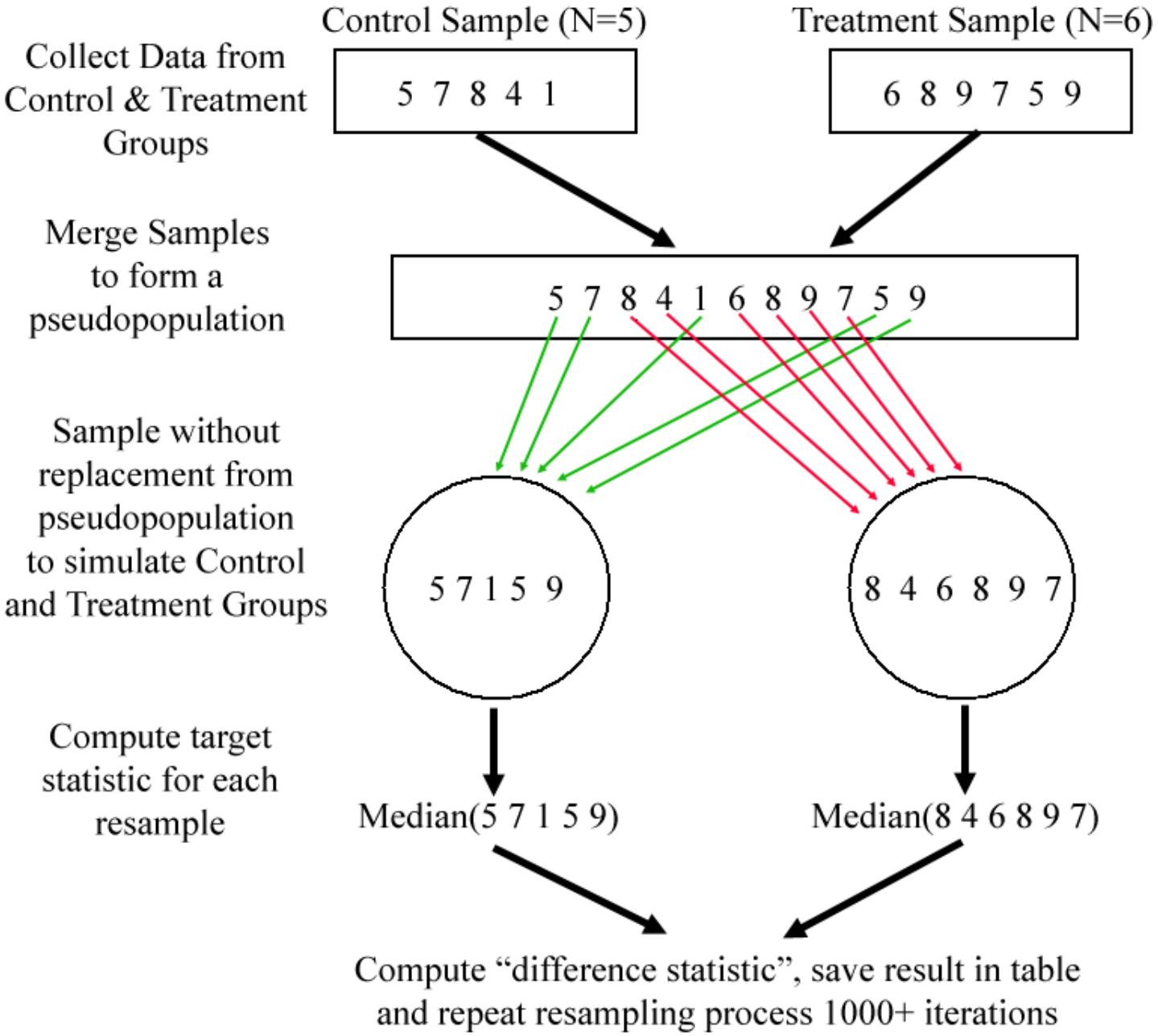
The Permutation Test derives its power from the process of random assignment

GENERAL PROCEDURE FOR PERMUTATION TESTS

To carry out a permutation test based on a statistic that measures the size of an effect of interest:

1. Compute the statistic for the original data.
2. Choose permutation resamples from the data without replacement in a way that is consistent with the null hypothesis of the test and with the study design. Construct the permutation distribution of the statistic from its values in a large number of resamples.
3. Find the *P*-value by locating the original statistic on the permutation distribution.

The Permutation Resampling Process



Example : Do structured reading exercises improve Degree of Reading Power (DRP) scores?

TABLE 18.4 DRP scores for third-graders

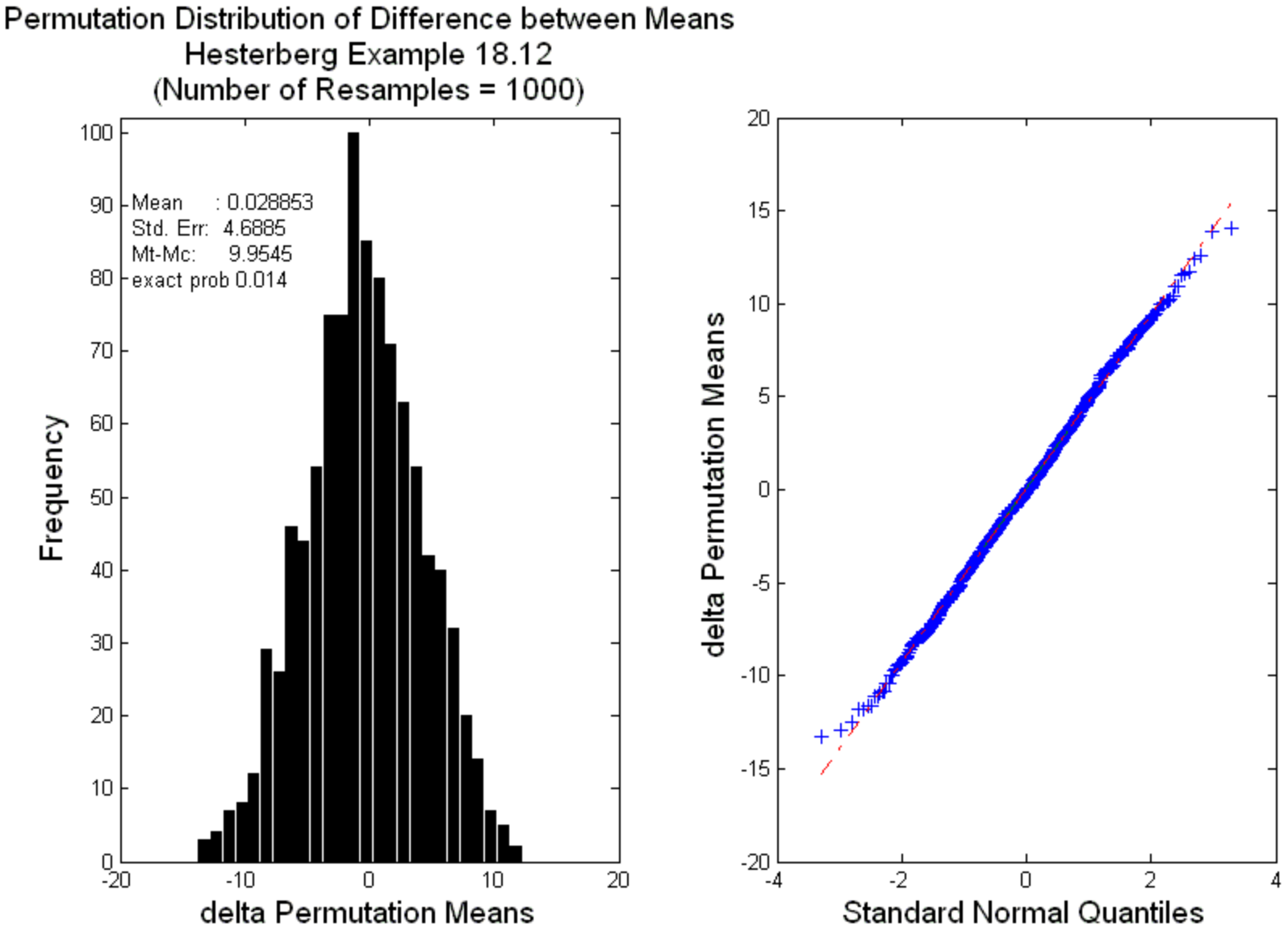
Treatment group				Control group			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Standard Parametric 1-tailed t-test:

H_0: $MEAN_{treatment} - MEAN_{control} = 0$
H_1: $MEAN_{treatment} - MEAN_{control} > 0$

Results: $t(42) = 2.26$, $p < 0.014$ (reject null hypothesis)

Resampling Approach to Same Reading Study: Monte Carlo Computation of Permutation Distribution



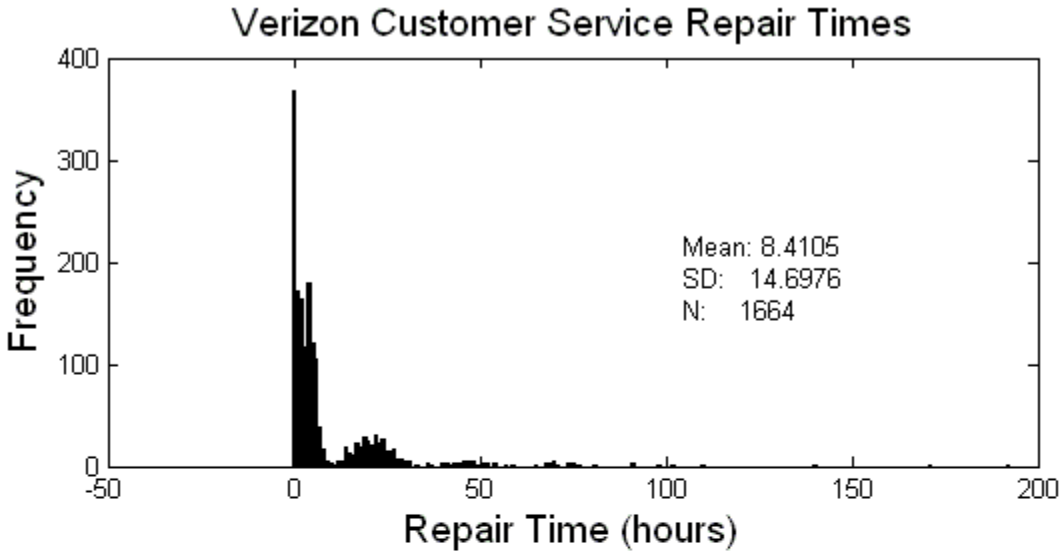
MATLAB Code Snippet
Permutation Resampling

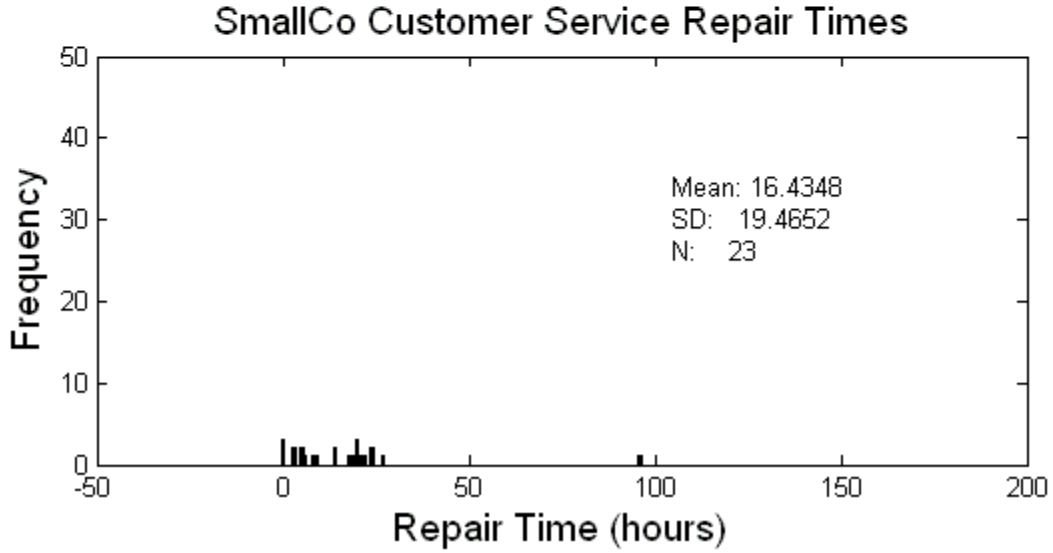
```
%Generate 1000 pairs of permutations representing
%random assignment to the treatment & control groups
%using sampling without replacement

universe=concat(treatment_group, control_group);
for i=1:1000
   [s1,s2]=randperm2(universe, 21);
   remeans(i)=mean(s1)-mean(s2);  %plug-in statistic
end
```

An example where the t-Test and Permutation Test yield Different Results

FCC regulations mandate that non-traditional telephone service providers be fined if they provide inferior service compared to the Legacy Provider in their area.

If the MEAN TIME TO ANSWER SERVICE CALLS is significantly longer than that obtained from a representative sample of the Legacy Co's logs then the competing Co. must pay a fine (alpha=1%)





<u>1-tailed t-Test</u>:

$t(1685) = -2.5877$

$p < 0.0048$

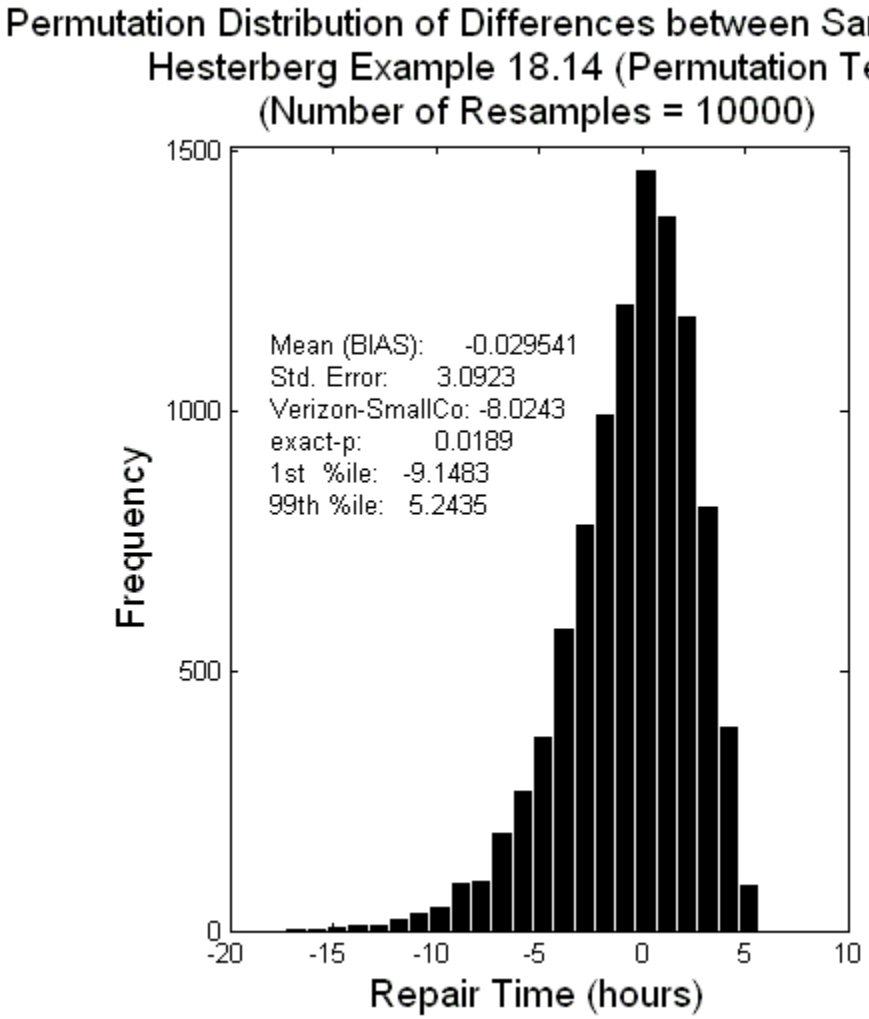
Statistically reliable increase in service call latency (alpha=1%)

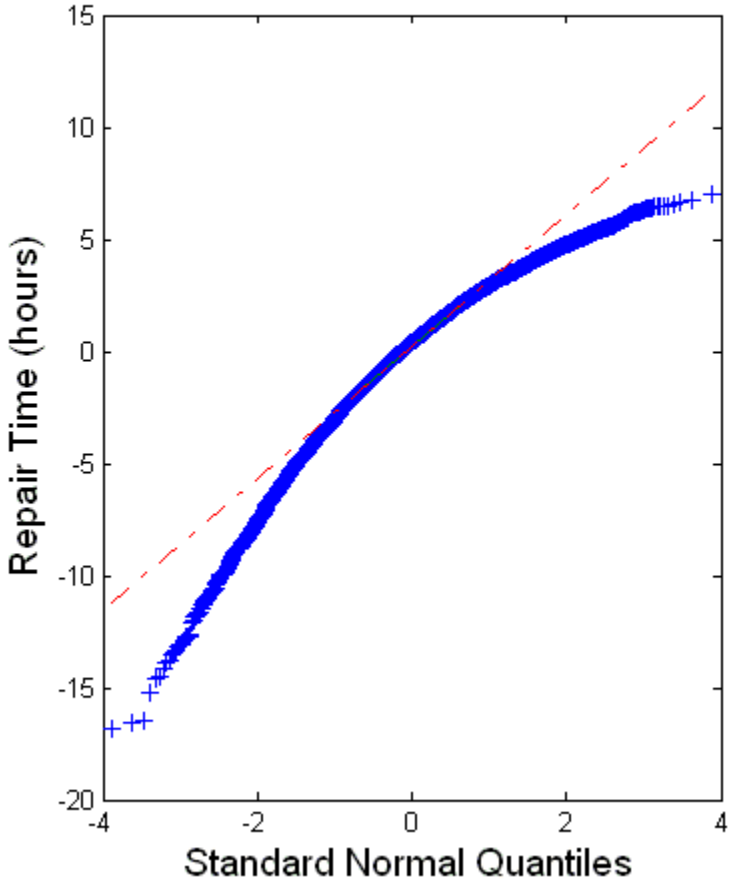
SmallCo must pay fine!

Permutation Sampling Distribution reveals:

Expected value of M1-M2 deviates from normal

M1-M2 not significant at alpha = 1%





Resampling for Correlated t-test [h15.m]

Scenario:

Twenty executives participated in a 2-week intensive foreign language course. Evaluate the effectiveness of the course based upon their "before" and "after" scores on a language assessment test.

Subject	1	2	3	4	18	19	20
Before	32	31	29	10	32	23	23
After	34	31	35	16	34	26	26

Bootstrap Example [h2.m]

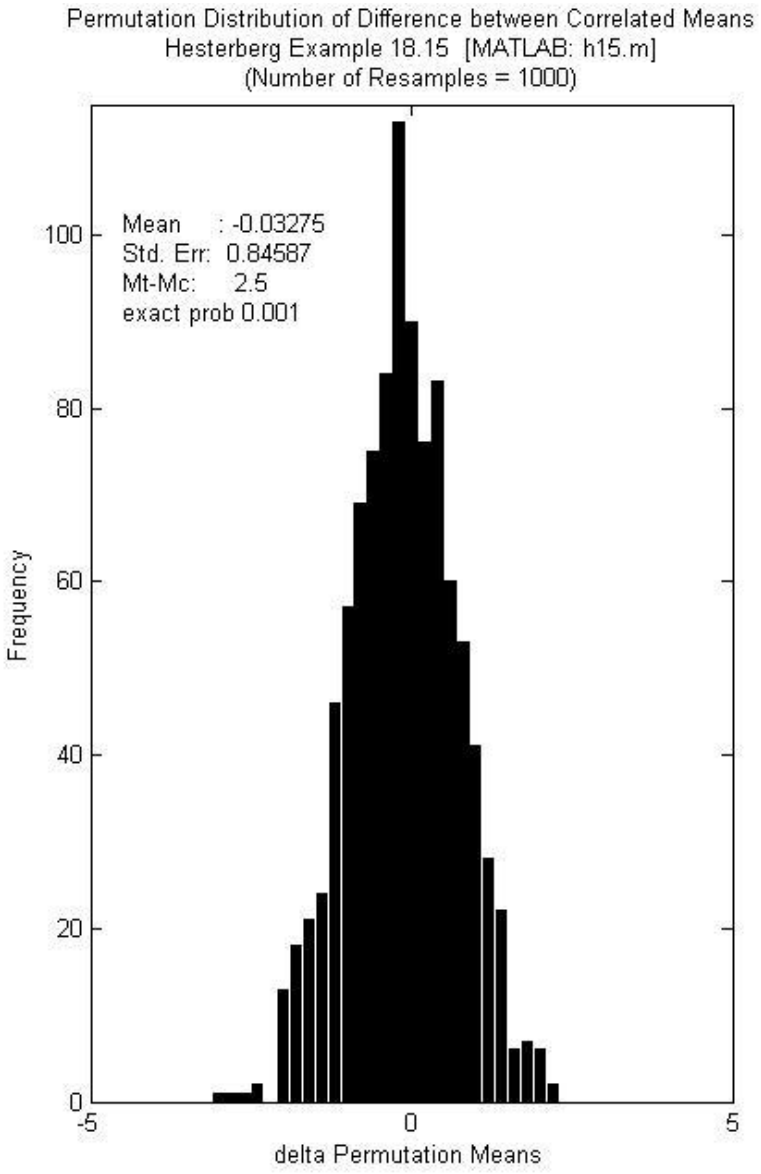
Use Resampling to generate Permutation Distribution
of Difference between Correlated Means

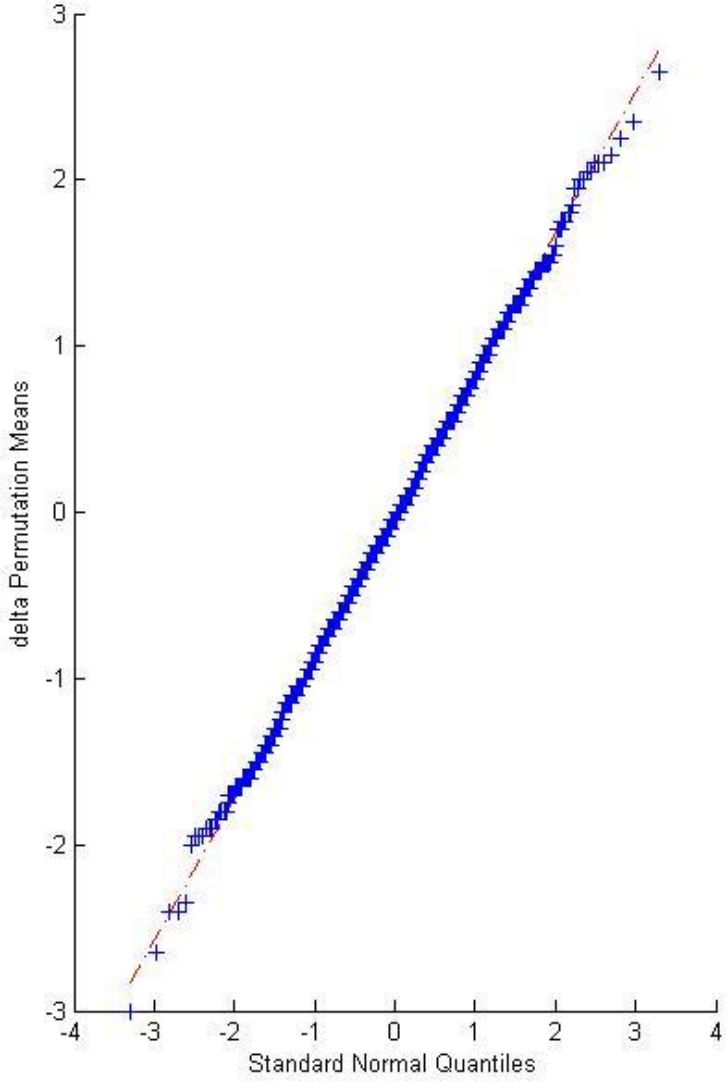
1. Generate list of subjects to resample (with replacement)
2. Randomly assign each resampled subject's bivariate data to the before vs. after treatment pool (null hypothesis)
3. Compute the means for the before vs. after pools
4. Compute difference between after-before resampled means
5. Repeat steps 1-4 for 1000 iterations
6. Generate and apply the permutation distribution

Bootstrap Example [h2.m]

```
subjects = [1:samplesize];
n_resamples=1000;
for i=1:n_resamples

    %resample N subjects (with replacement)
    s1 = randsample(samplesize, samplesize, true);

    %randomly assign subjects' bivariate sample pairs to before vs. after pool
    before_assignments = randsample(2, samplesize, true); %1 or 2
    after_assignments = 2-floor(before_assignments/2);       %complement
    for n=1:samplesize  %implement random assignment
        before_pool(n) = bivariate(s1(n),before_assignments(n));
        after_pool(n)  = bivariate(s1(n),after_assignments(n));
    end

    %compute difference between after-before means
    %and add it to the accumulating permutation distribution
    remdiff(i)=mean(after_pool)-mean(before_pool);
end
```

Resampling for Correlated t-test [h15.m]





Visit MATLAB Resampling Resource Page:

apps.usd.edu/coglab/psyc792/resampling/resampling.html