Lecture 9: Resampling Methods

GENOME 560

Doug Fowler, GS (dfowler@uw.edu)

Review from last week

- Conceptually, what is a nonparametric test?

What is Resampling?

Two Broad Categories of Resampling

- In **randomization** we systematically shuffle observed data many times (no replacement)

Two Broad Categories of Resampling

- In **randomization** we systematically shuffle observed data many times (no replacement)

- In **bootstrapping** we draw samples with replacement from the observed data

Two Broad Categories of Resampling

- In **randomization** we systematically shuffle observed data many times (no replacement)
  - Unconcerned about population parameter estimates, used for hypothesis testing
- In **bootstrapping** we draw samples with replacement from the observed data
  - Focused primarily on estimating the accuracy of population parameter estimates

A Simple Example

- Let's say we measured transcript levels from two types of tissue

- A randomization approach would ask if it is likely that we would obtain a difference in levels as large as the one we observed assuming the tissue types had the same expression

- A bootstrap approach would give us variances of population parameters of interest (e.g. mean expression level)

Why Use Resampling Methods?

- Useful when we know little about the distribution from which a sample was drawn

- Useful when we know that assumptions required for other tests have been violated

Goals

- Randomization testing – exact
- Randomization testing – sampled
- Bootstrap

Randomization Testing

- Randomization testing has three steps:

Randomization Testing

- Randomization testing has three steps:
  - Consider an observed sample as one of many equally possible outcomes that could have arisen by chance

Randomization Testing

- Randomization testing has three steps:
  - Consider an observed sample as one of many equally possible outcomes that could have arisen by chance
  - Enumerate the possible outcomes that could be observed by randomly rearranging the data in the sample

Randomization Testing

- Randomization testing has three steps:
  - Consider an observed sample as one of many equally possible outcomes that could have arisen by chance
  - Enumerate the possible outcomes that could be observed by randomly rearranging the data in the sample
  - On the basis of the resulting distribution of outcomes, decide whether the observed outcome is improbable enough to warrant rejection of H_0

Exact Randomization Testing

- In exact cases, we can explicitly write down all possible outcomes

- Then, we can determine the proportion of outcomes as or more extreme than the observed one

Exact Randomization Testing

- For example, let's say that are measuring the growth rate of EMS mutagenized yeast with wild-type yeast.

WT	mutagenized
91	95
85	100
82	105

Exact Randomization Testing

- For example, let's say that are measuring the growth rate of EMS mutagenized yeast with wild-type yeast.
  - We could assume that these data are normally distributed, but with such a small N that might not be a good idea…

WT	mutagenized
91	95
85	100
82	105

Exact Randomization Testing

- For example, let's say that are measuring the growth rate of EMS mutagenized yeast with wild-type yeast.
  - So let's do a randomization test instead

WT	mutagenized
91	95
85	100
82	105

Exact Randomization Testing

- For example, let's say that are measuring the growth rate of EMS mutagenized yeast with wild-type yeast.

WT	mutagenized
91	95
85	100
82	105

- First, we have to choose a test statistic… thoughts?

Exact Randomization Testing

- For example, let's say that are measuring the growth rate of EMS mutagenized yeast with wild-type yeast.

	WT	mutagenized
	91	95
	85	100
	82	105
sum	258	300
mean	86	100
d	-14	

- Hypotheses?

Sampled Randomization Testing

- In some cases, we cannot feasibly write down all possible outcomes

Sampled Randomization Testing

- For example, let's say we had measured 100 different yeast cultures in each case. How many possibilities?

Sampled Randomization Testing

- In some cases, we cannot feasibly write down all possible outcomes

- How can we get around this problem?

Sampled Randomization Testing

- In some cases, we cannot feasibly write down all possible outcomes
- How can we get around this problem?
- We take random samples without replacement and compute our test statistic value

Sampled Randomization Testing

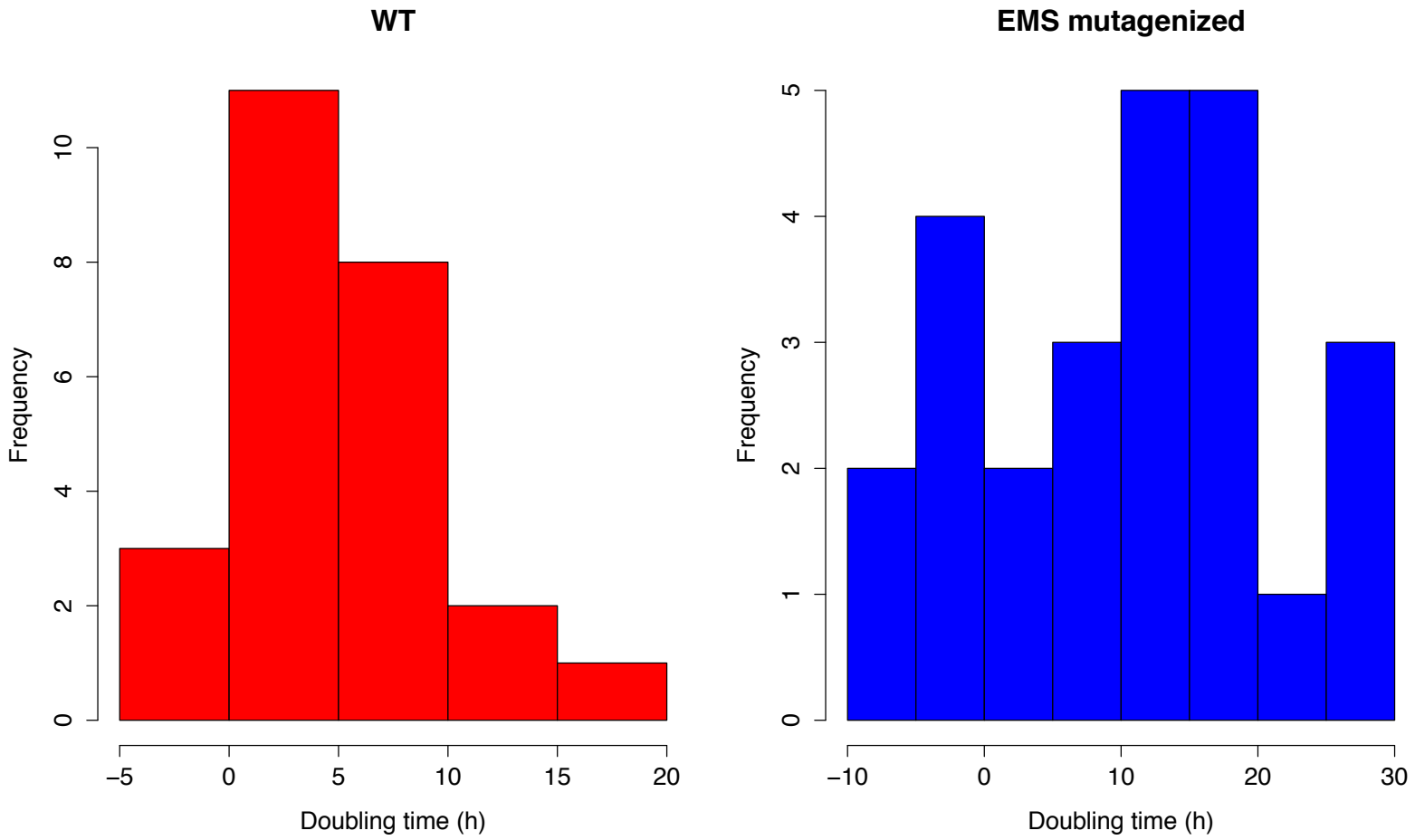
- In some cases, we cannot feasibly write down all possible outcomes
- How can we get around this problem?
- We take random samples without replacement and compute our test statistic value
- We do this enough times to generate distribution for the test statistic and then use that distribution to test the hypothesis

Sampled Randomization Testing

- For example, let's say that we are measuring the growth rate of EMS mutagenized yeast with wild-type yeast (50 total samples instead of 6).

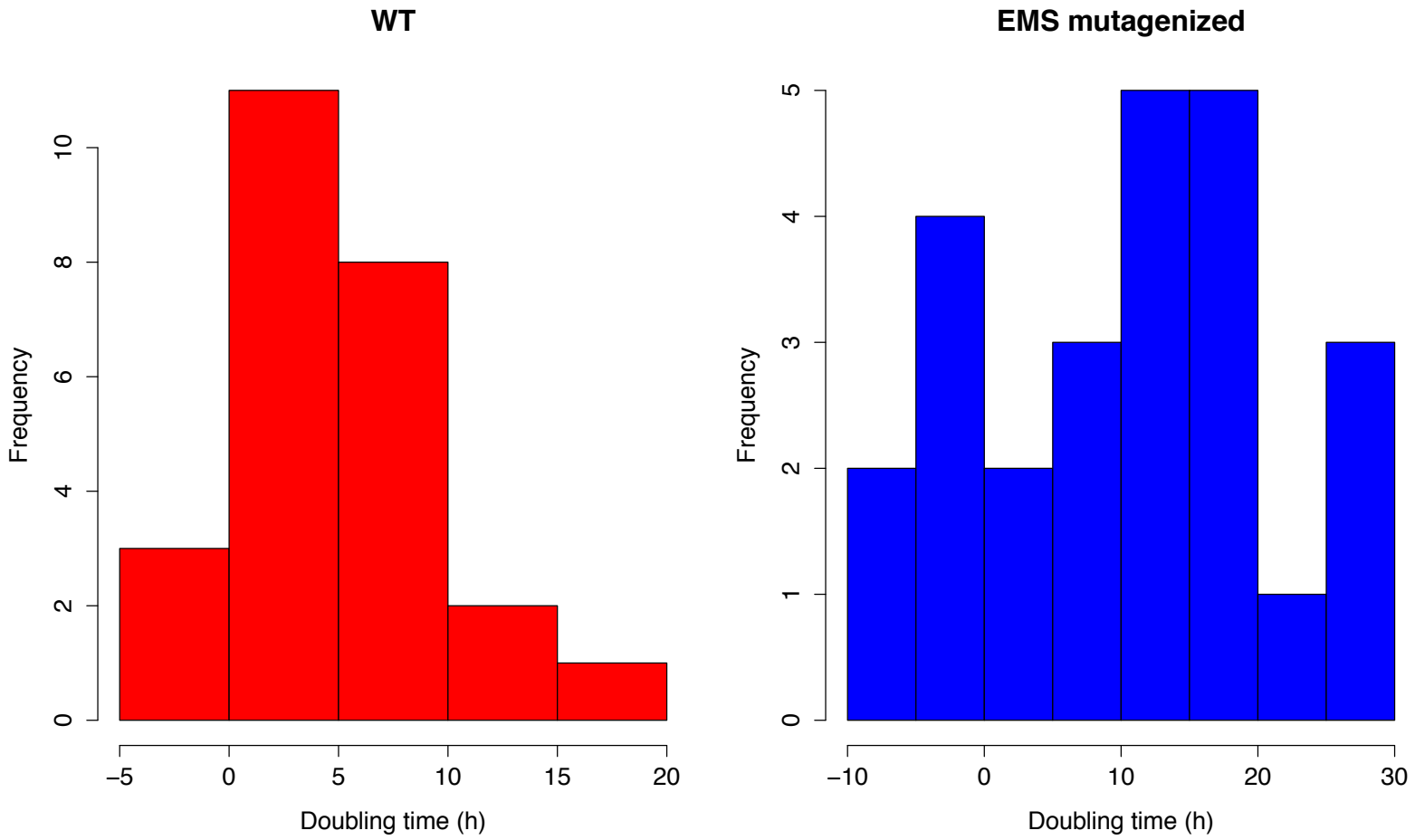
Sampled Randomization Testing

- For example, let's say that we are measuring the growth rate of EMS mutagenized yeast with wild-type yeast.



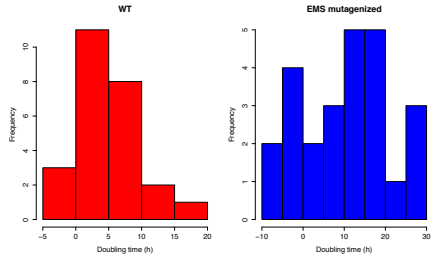
Sampled Randomization Testing

- What types of tests would/wouldn't you use here and why?



Sampled Randomization Testing

- For example, let's say that we are measuring the growth rate of EMS mutagenized yeast with wild-type yeast

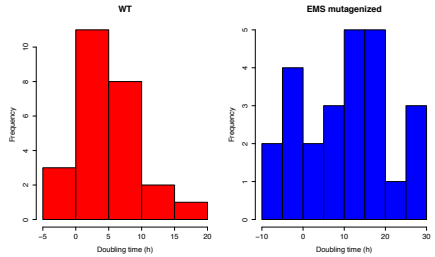


- First, we define a test statistic

$$d = \overline{WT} - \overline{EMS}$$

Sampled Randomization Testing

- For example, let's say that we are measuring the growth rate of EMS mutagenized yeast with wild-type yeast



$$H_0 : d = 0$$
$$H_1 : d \neq 0$$

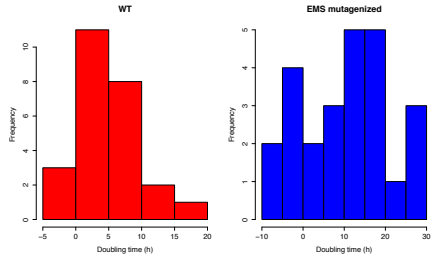
- First, we define a test statistic

$$d = \overline{WT} - \overline{EMS}$$

- Next, we define hypotheses

Sampled Randomization Testing

- For example, let's say that we are measuring the growth rate of EMS mutagenized yeast with wild-type yeast



$$H_0 : d = 0$$
$$H_1 : d \neq 0$$

- First, we define a test statistic

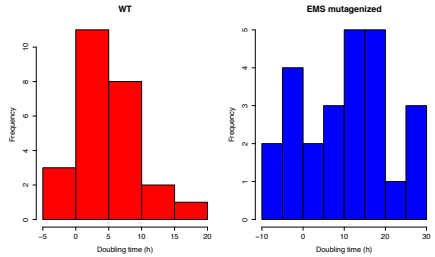
$$d = \overline{WT} - \overline{EMS}$$

- Next, we define hypotheses

- Finally we generate the test statistic distribution – how?

Sampled Randomization Testing

- For example, let's say that we are measuring the growth rate of EMS mutagenized yeast with wild-type yeast



$$H_0 : d = 0$$
$$H_1 : d \neq 0$$

- First, we define a test statistic

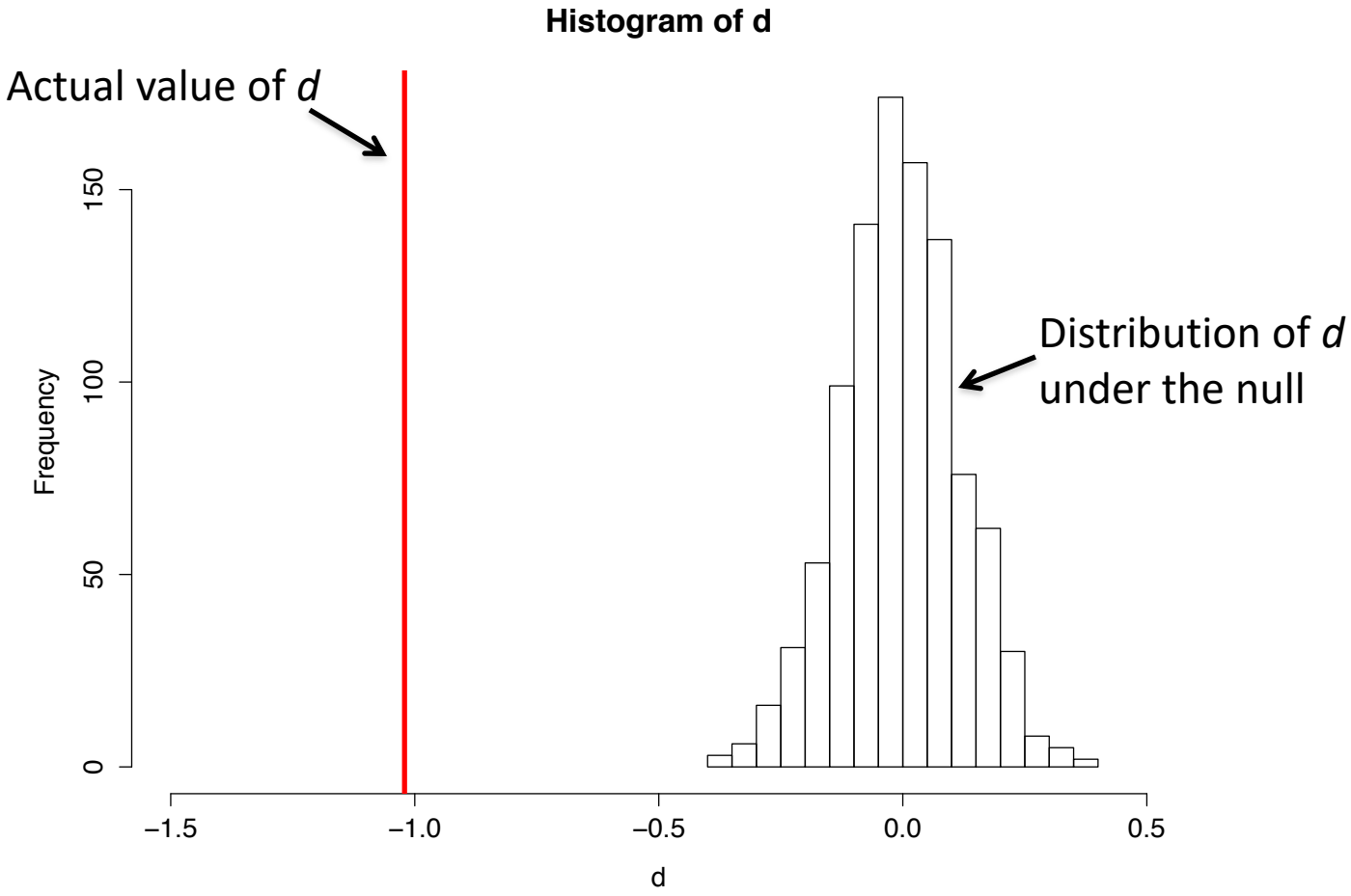
$$d = \overline{WT} - \overline{EMS}$$

- Next, we define hypotheses

- How many permutations is enough?

Compare value of *d* to distribution

- Here are our actual results… what do they mean?



Randomization Testing

- Randomization testing has three steps:
  - Consider an observed sample as one of many equally possible outcomes that could have arisen by chance
  - Enumerate (or estimate by random sampling) the possible outcomes that could be observed by randomly rearranging the data in the sample
  - On the basis of the resulting distribution of outcomes, decide whether the observed outcome is improbable enough to warrant rejection of H_0

Let's Bask in the Power

- We defined a test statistic *d* that suited our needs
- We easily could have picked others (e.g. a ratio or anything else)

Let's Bask in the Power

- We defined a test statistic *d* that suited our needs
- We easily could have picked others (e.g. a ratio or anything else)
- If we wanted to use parametric stats, we would have to derive the distribution of each stat (e.g. we'd have to be mathematicians)


- Resampling is an empirical way to do stats

Goals

- Randomization testing – exact
- Randomization testing – sampled
- Bootstrap

The Bootstrap

- Used to assign measures of accuracy (e.g. confidence intervals, variance, bias) to sample estimates of parameters

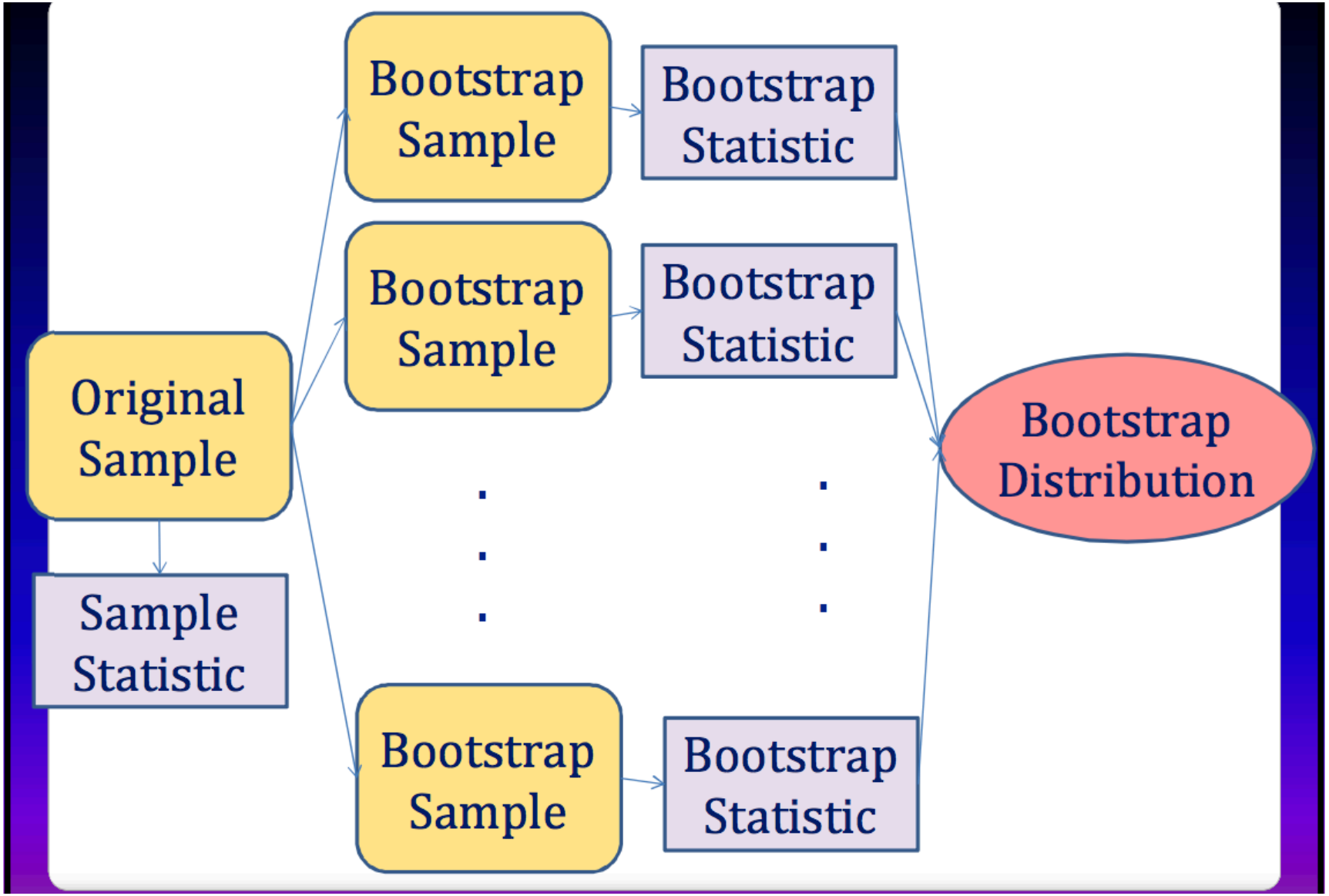
- Said another way, bootstrap is useful for estimating the sampling distribution of a statistic without using a parametric distribution (e.g. z-stat, t-stat)

- Assumes samples are independent and identically distributed

The Bootstrap - Procedure

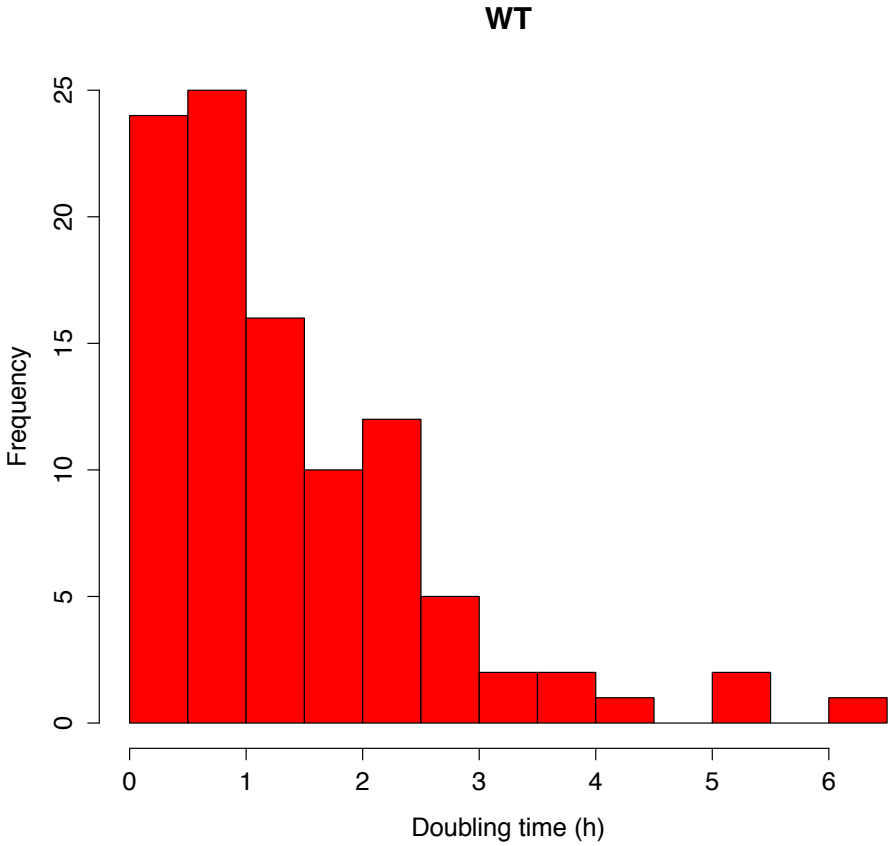
- Again, we treat the sample as the “population”
- Then we resample **with replacement** from the sample and recalculate our test/descriptive statistic
- Finally we use the resampled distribution to draw conclusions

The Bootstrap - Procedure



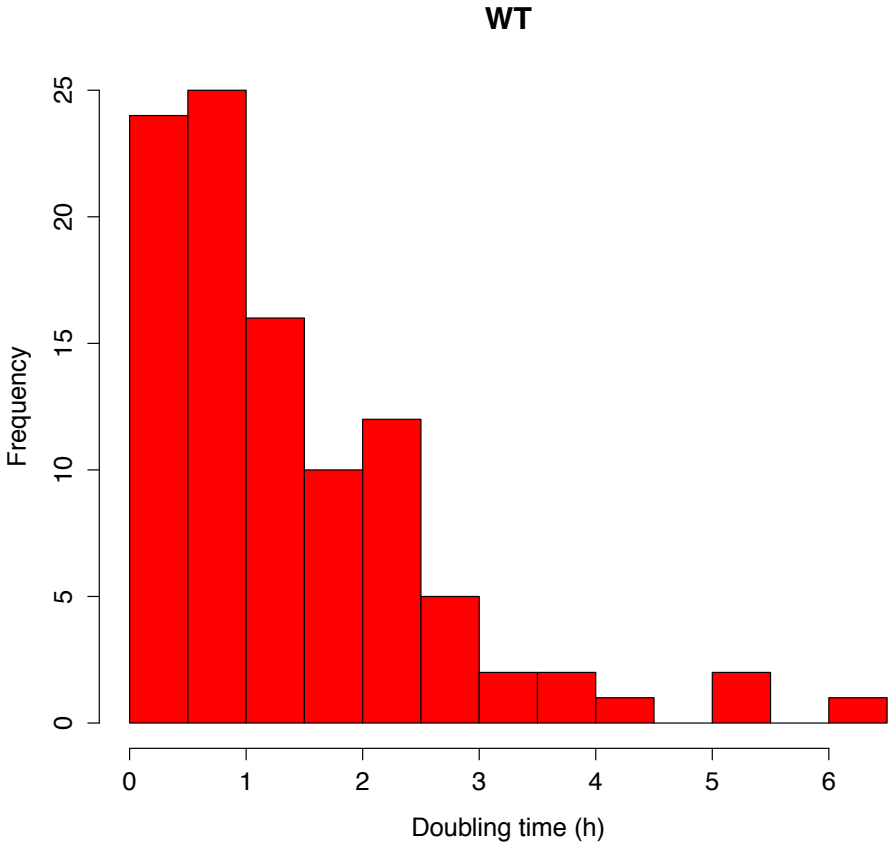
Bootstrap Example - CIs

- Let's say that we are measuring the growth rate yeast and would like to generate a 95% CI on the mean



Bootstrap Example - CIs

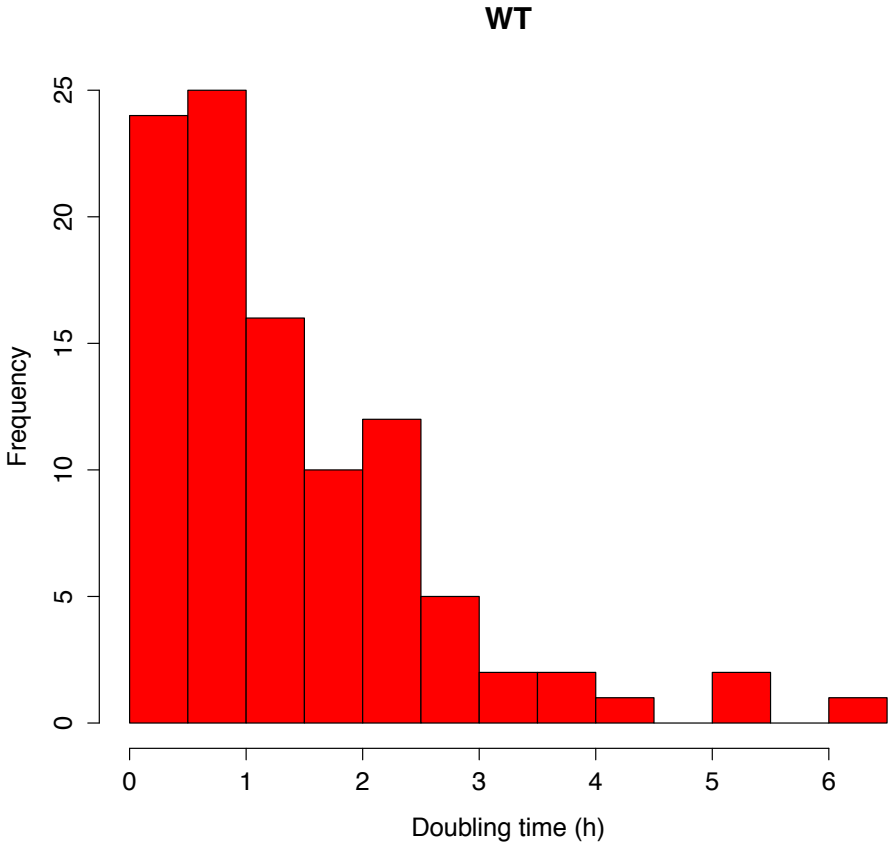
- Let's say that we are measuring the growth rate yeast and would like to generate a 95% CI on the mean



$$\bar{x} = 1.54$$

Bootstrap Example - CIs

- To calculate one bootstrapped mean we take a sample equal in size to the original one with replacement



$$\bar{x}_{boot} = \frac{x_1 + ... + x_n}{N}$$

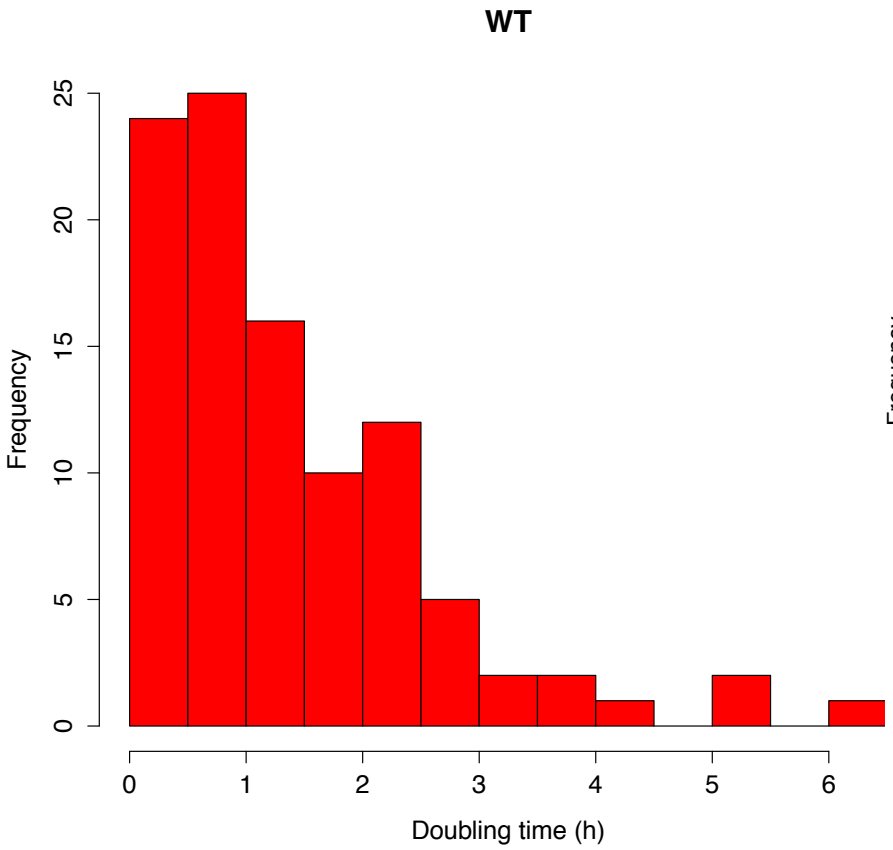
$$\bar{x}_{boot} = 1.75$$

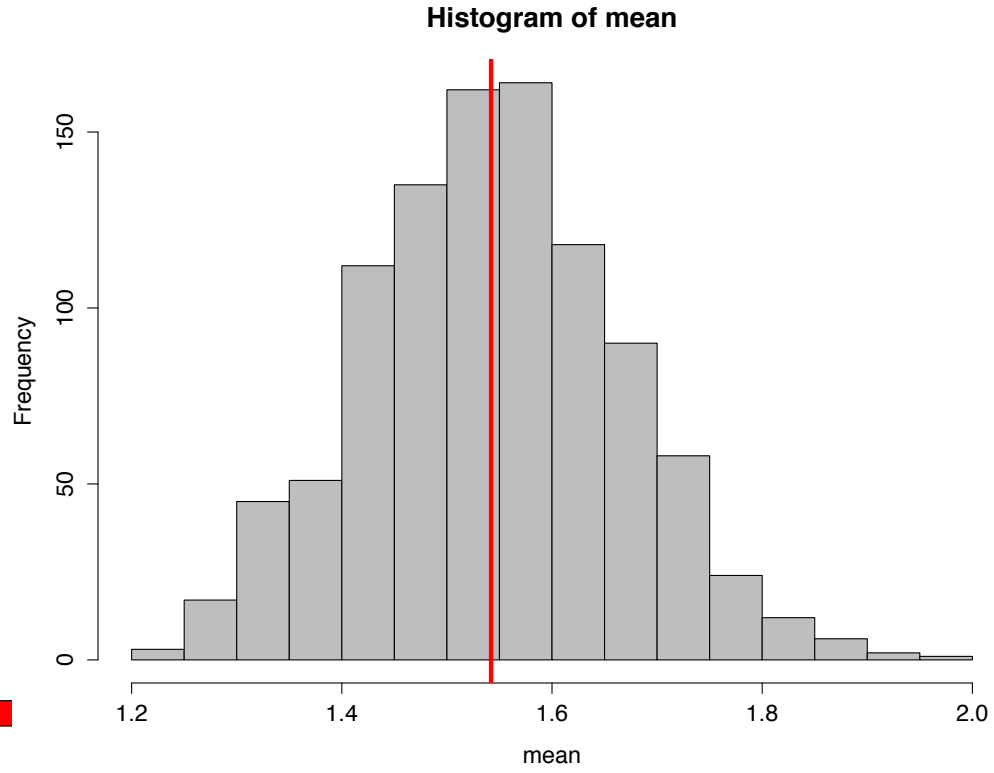
$$\dots$$

$$\bar{x} = 1.54$$

Bootstrap Example - CIs

- We sample with replacement 1,000 times to generate a bootstrapped sampling distribution of the sample mean

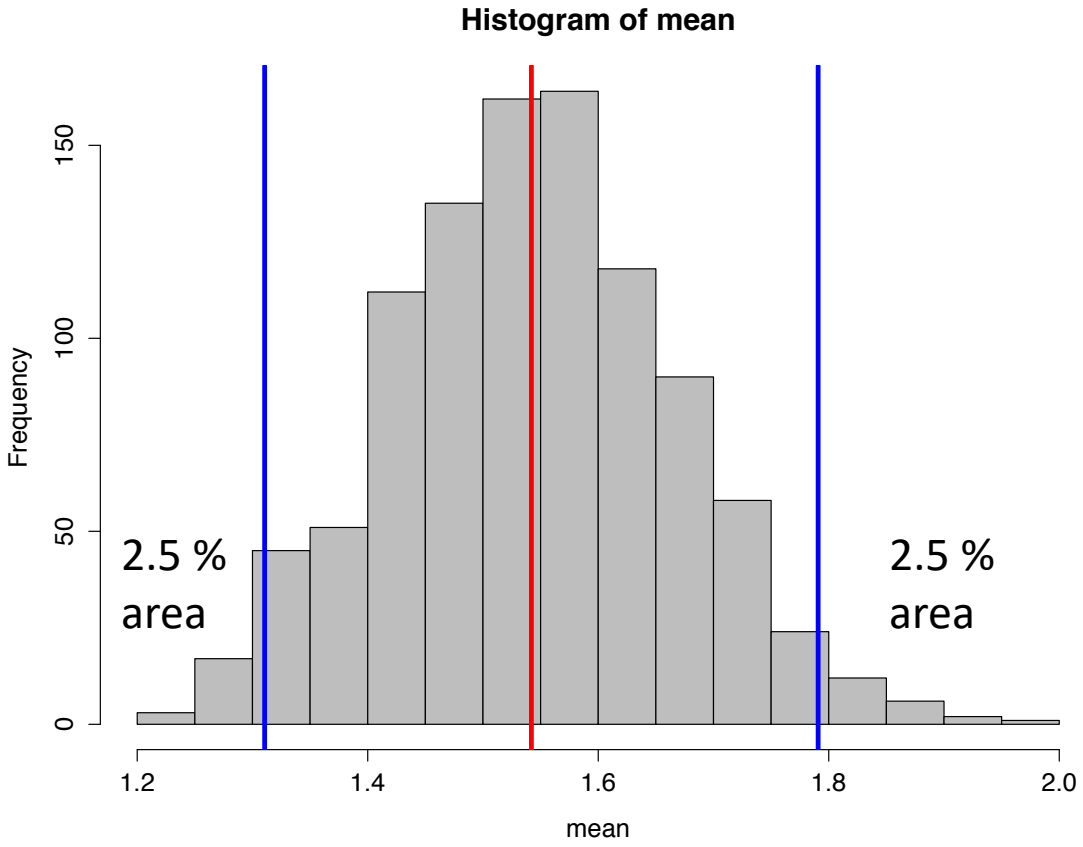




$$\bar{x} = 1.54$$

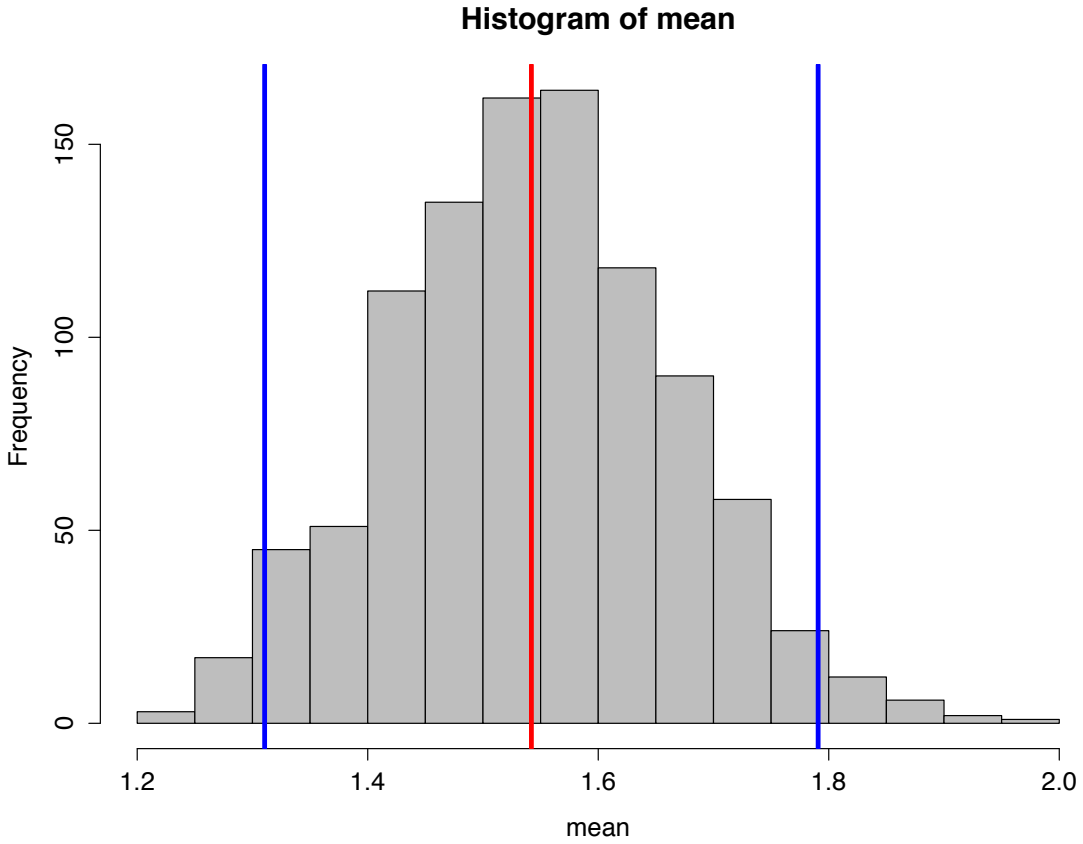
Bootstrap Example - CIs

- Now, we find the 2.5th and 97.5th percentile of the distribution



Bootstrap Example - CIs

- Now, we find the 2.5th and 97.5th percentile of the distribution

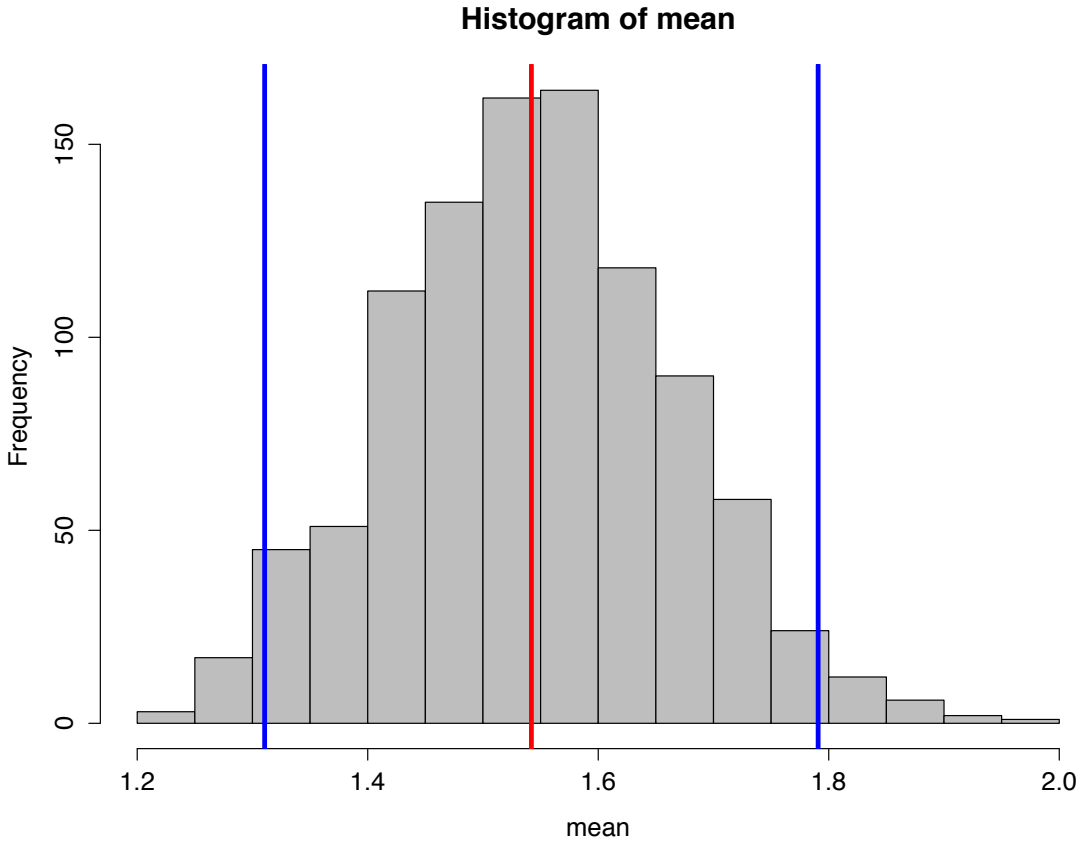


$$\bar{x} = 1.54$$

$$CI_{boot} = 1.33 - 1.80$$

Note, These are Not Symmetric!

- Does the CLT make you think it should be? Why is the distribution asymmetric?



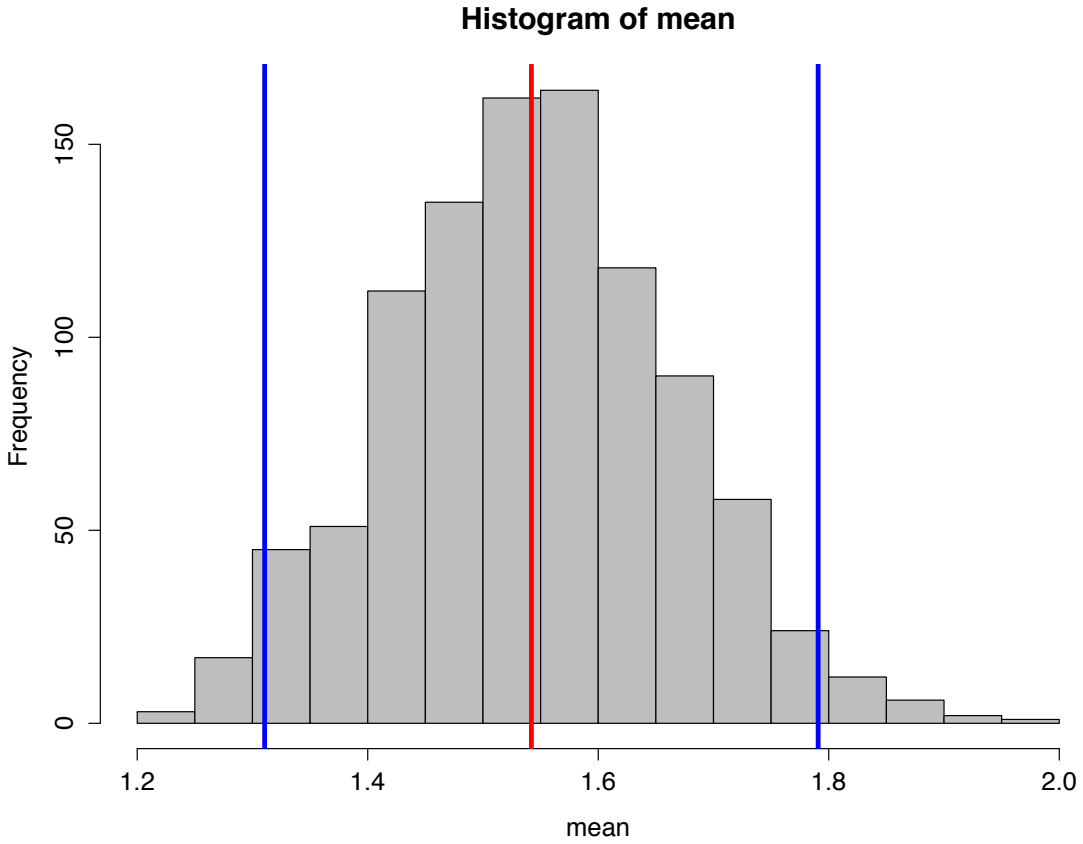
$$\bar{x} = 1.54$$

$$CI_{lower} = 1.54 - 0.214$$

$$CI_{upper} = 1.54 + 0.264$$

Note, These are Not Symmetric!

- We're sampling from a highly asymmetric population, and the sampling distribution only approaches normality



$$\bar{x} = 1.54$$

$$CI_{lower} = 1.54 - 0.214$$

$$CI_{upper} = 1.54 + 0.264$$

Other Applications of the Bootstrap

- Estimating variance
- Assessing bias

Two Broad Categories of Resampling

- In **randomization** we systematically shuffle observed data many times (no replacement)
  - Unconcerned about population parameters

- In **bootstrapping** we draw samples with replacement from the observed data
  - Focused primarily on estimating population parameters, CIs and variance

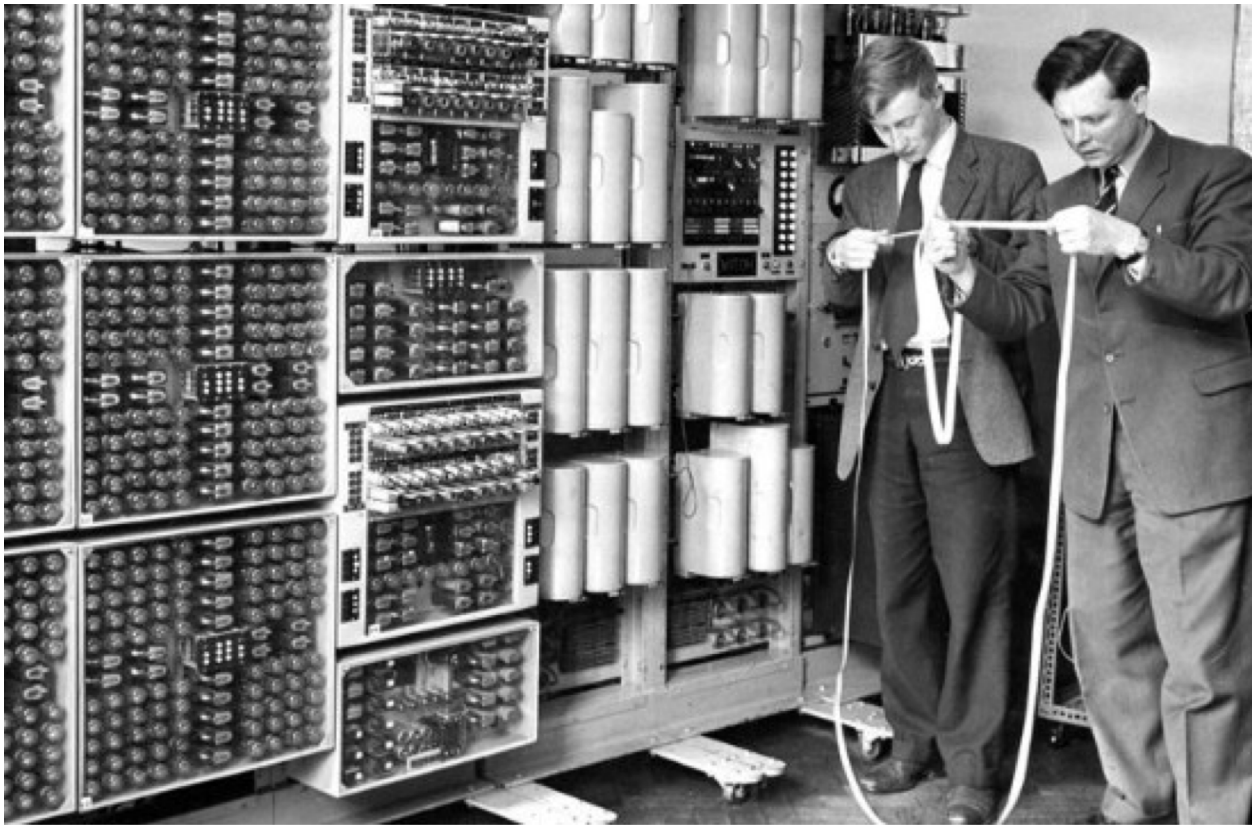
Why Use Resampling Methods?

- Useful when we know little about the distribution from which a sample was drawn
- Useful when we know that assumptions required for other tests have been violated
- Simple/straightforward

Why Haven't We Always Used Resampling?

- In fact, resampling methods have been around for a long time (<1900)
- They are powerful (about as powerful as equivalent parametric methods, generally speaking)
- So, why haven't they always been used?

NO FREE LUNCH (again)



Resampling methods are (extremely) computationally taxing

Parametric vs. Nonparametric vs. Resampling

	Basic Assumptions	Power	Conceptual complexity	Ease of computation
Parametric				
Nonparametric				
Resampling				

Parametric vs. Nonparametric vs. Resampling

	Basic Assumptions	Power	Conceptual complexity	Ease of computation
Parametric	Strong			
Nonparametric	Weak			
Resampling	Weaker			

Parametric vs. Nonparametric vs. Resampling

	Basic Assumptions	Power	Conceptual complexity	Ease of computation
Parametric	Strong	Best		
Nonparametric	Weak	Less		
Resampling	Weaker	Nearly as good as parametric		

Parametric vs. Nonparametric vs. Resampling

	Basic Assumptions	Power	Conceptual complexity	Ease of computation
Parametric	Strong	Best	High	
Nonparametric	Weak	Less	Intermediate	
Resampling	Weaker	Nearly as good as parametric	Low	

Parametric vs. Nonparametric vs. Resampling

	Basic Assumptions	Power	Conceptual complexity	Ease of computation
Parametric	Strong	Best	High	Easy
Nonparametric	Weak	Less	Intermediate	Intermediate
Resampling	Weaker	Nearly as good as parametric	Low	Hard

Reading/Resources

- http://www.statsoft.com/Textbook/Nonparametric-Statistics/button/2
- http://qed.econ.queensu.ca/working_papers/papers/qed_wp_1127.pdf
- http://anson.ucdavis.edu/~peterh/sta251/bootstrap-lectures-to-may-16.pdf