# Bootstrap and resampling

## Outline for today

- Estimation
- The sampling distribution
- Standard error of an estimate
- The bootstrap standard error
- The bootstrap confidence interval
- Bootstrap used when estimating a phylogeny
- Comparing two groups
- Randomization test
- Summary

**Estimation**

The process of inferring a population parameter from sample data.

The value of a sample *estimate* is almost never the same as the *parameter* in the population because of random sampling error (chance).

The sampling distribution of an estimate gives all the values we might have obtained from our sample, and their probabilities of occurrence.

The standard error of an estimate is the standard deviation of its sampling distribution. No estimate is useful without it.

Your estimates are more useful than your *P*-values to future research.

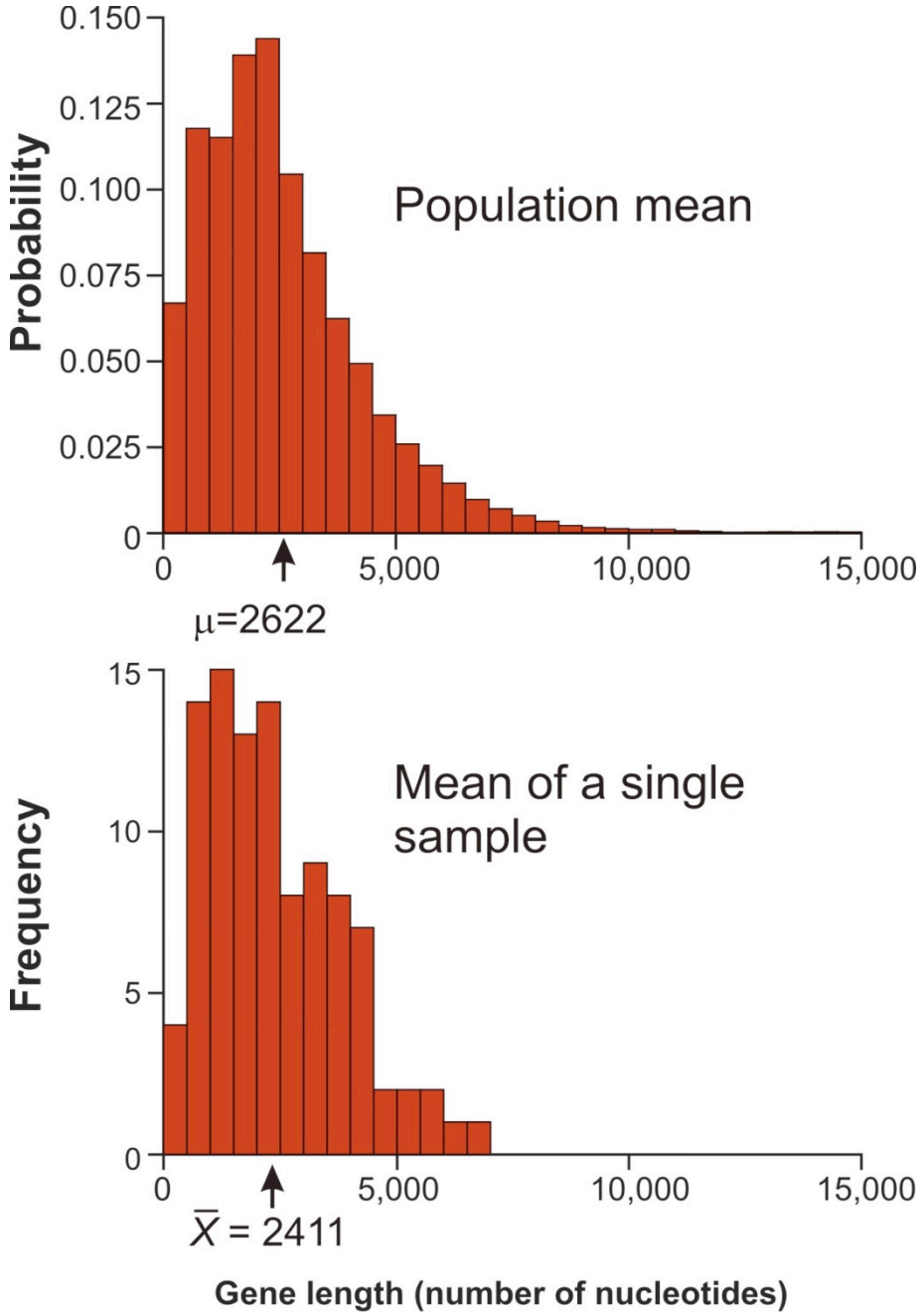
**Estimation of mean**

What we want:

The mean of a variable in the <u>population</u> (e.g., the lengths of all the genes in the human genome)

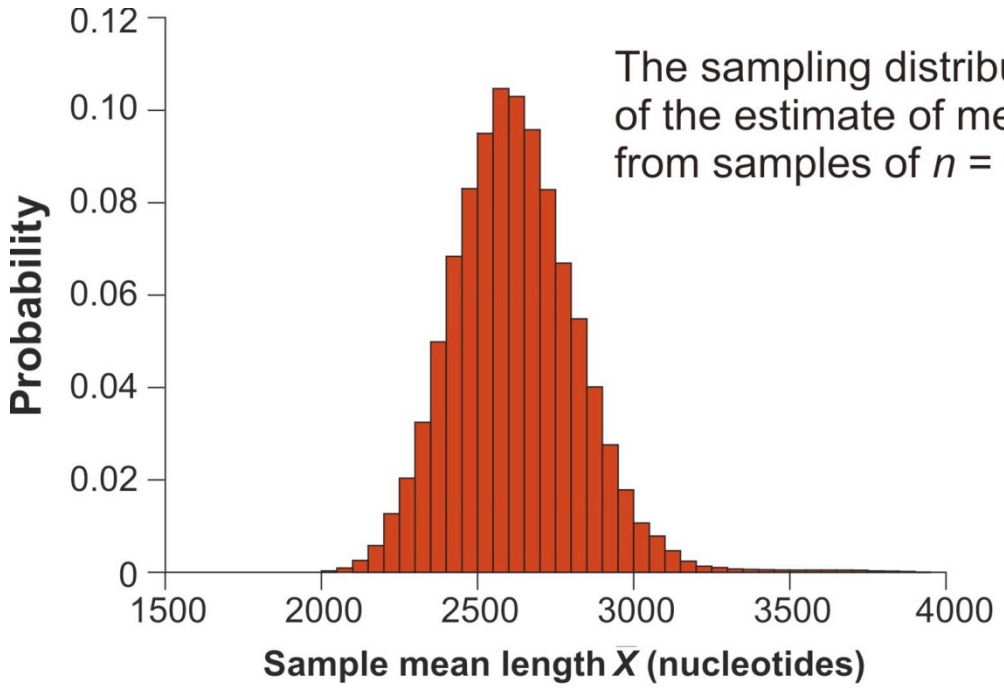
What we have instead:

The <u>sample</u> mean (e.g., a random sample of *n* = 100 genes)

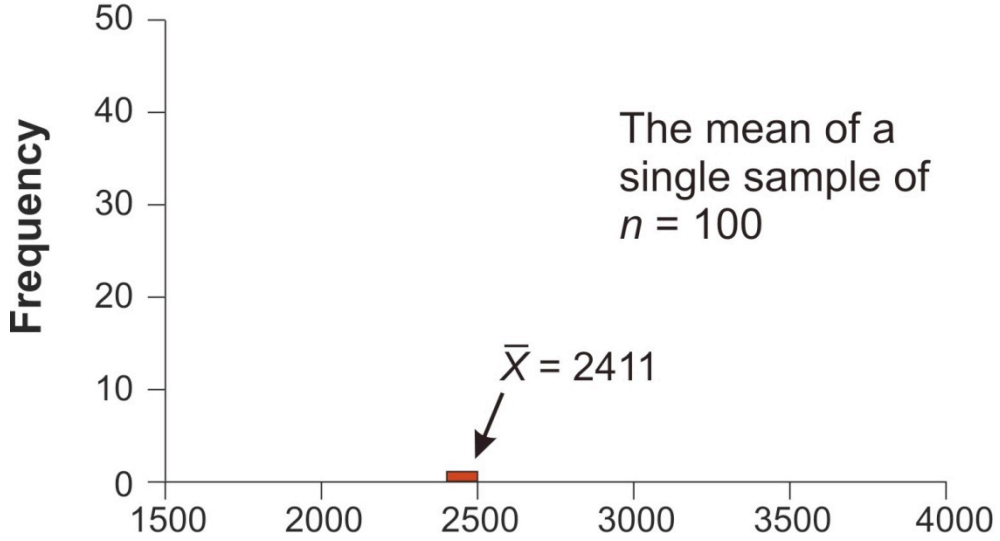


**The sampling distribution**

Since we don't have the true mean, we want the sampling distribution, giving all possible values of the estimate and their probabilities



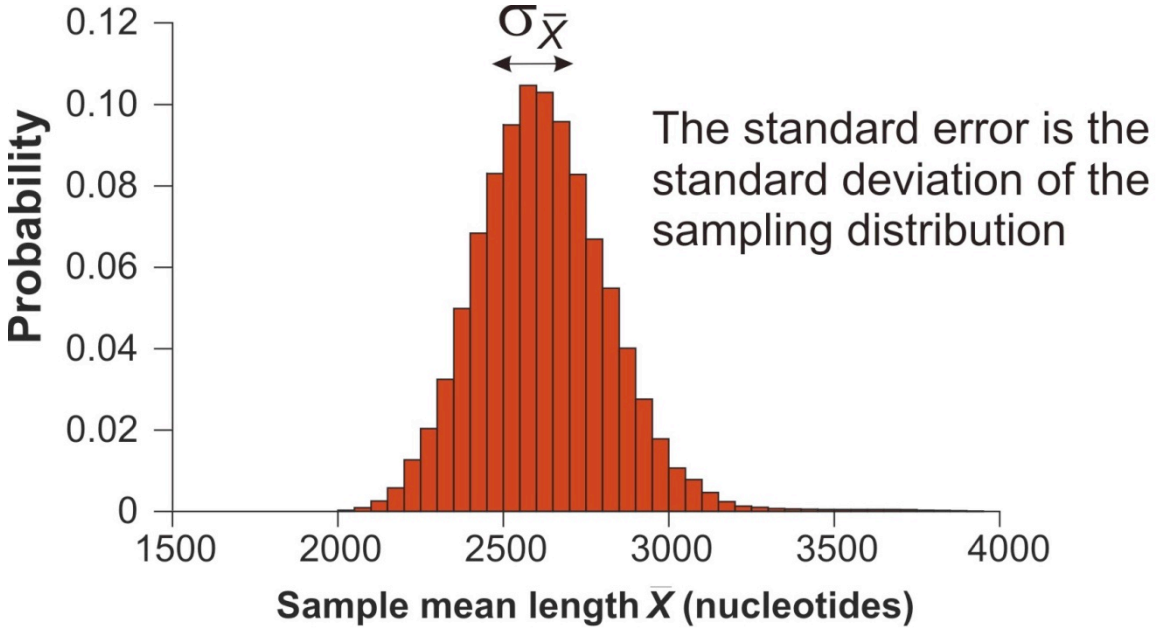

What we have instead: Just one sample mean

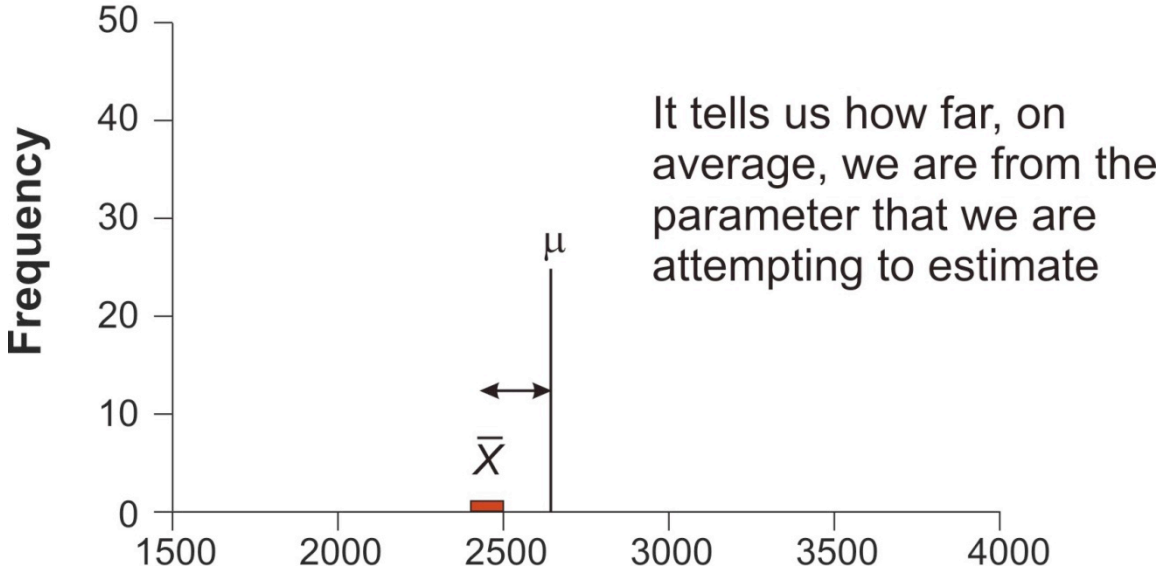


**Standard error**

Mostly we want the standard deviation of the sampling distribution (a.k.a. the standard error). It measures the variation of the sample estimates around the population parameter

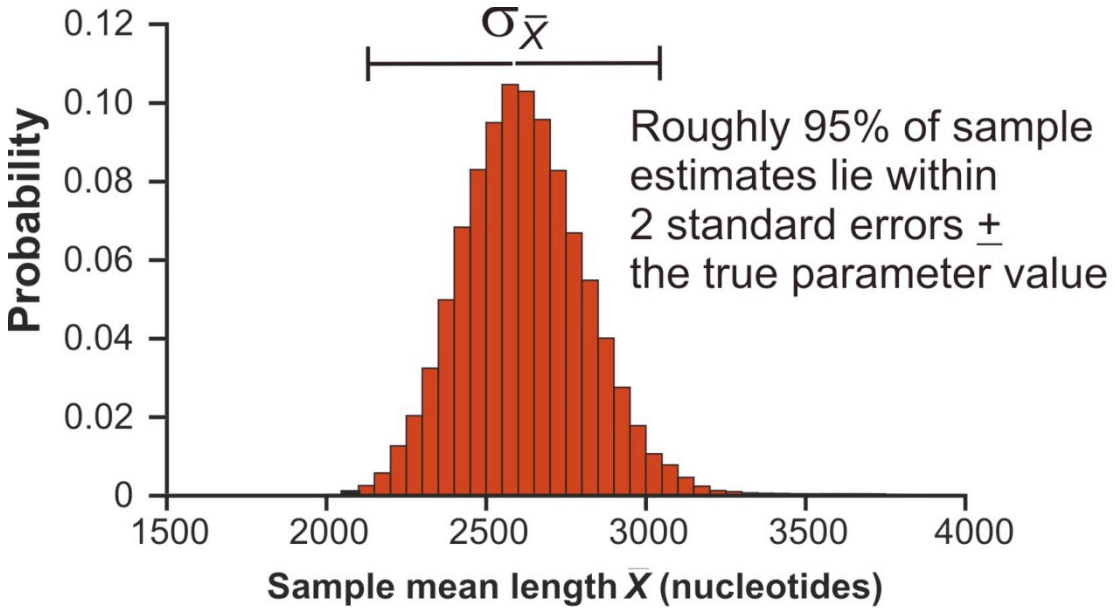
Here's why we like it.



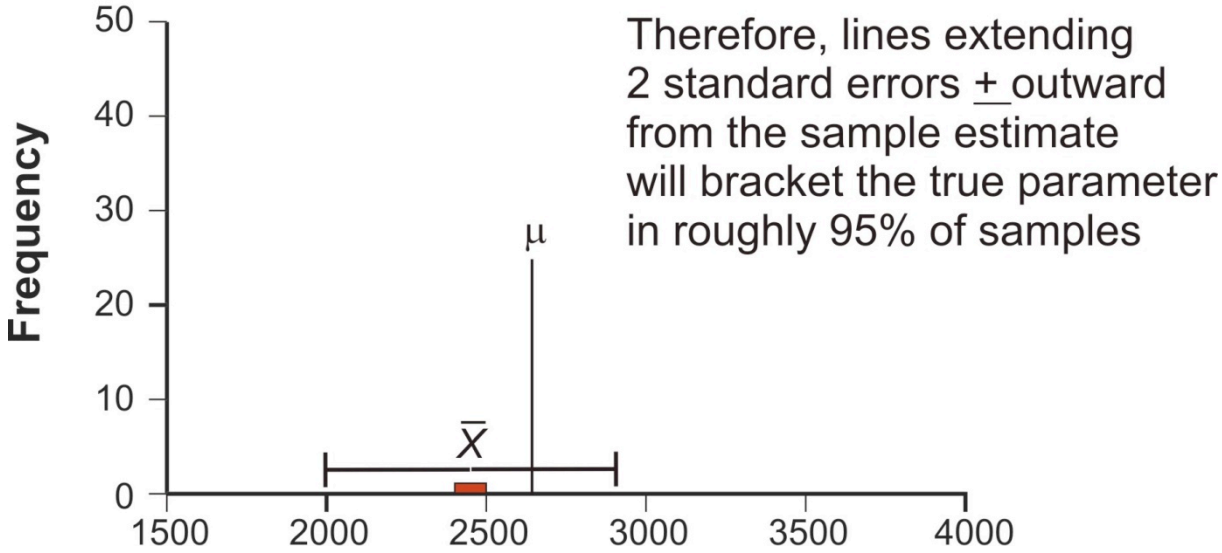



**Standard error**

If the sampling distribution is roughly bell-shaped, then about 95% of estimates fall within 2 SE's of the population parameter




Twice the SE provides an approximate 95% confidence interval for the parameter



**Standard error of the sample mean has a remarkable property**

It can be estimated from a single sample!

$$\sigma_{\bar{X}} \approx s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

$s_{\bar{X}}$ is the estimated standard error. It is usually called simply the "standard error of the mean"

This is an unusual feature of $\bar{X}$

**Standard error of the sample mean has a remarkable property**

Sadly, most other kinds of estimates do not have this amazing property.
What to do?

One answer: make your own sampling distribution for the estimate
using the “bootstrap”.

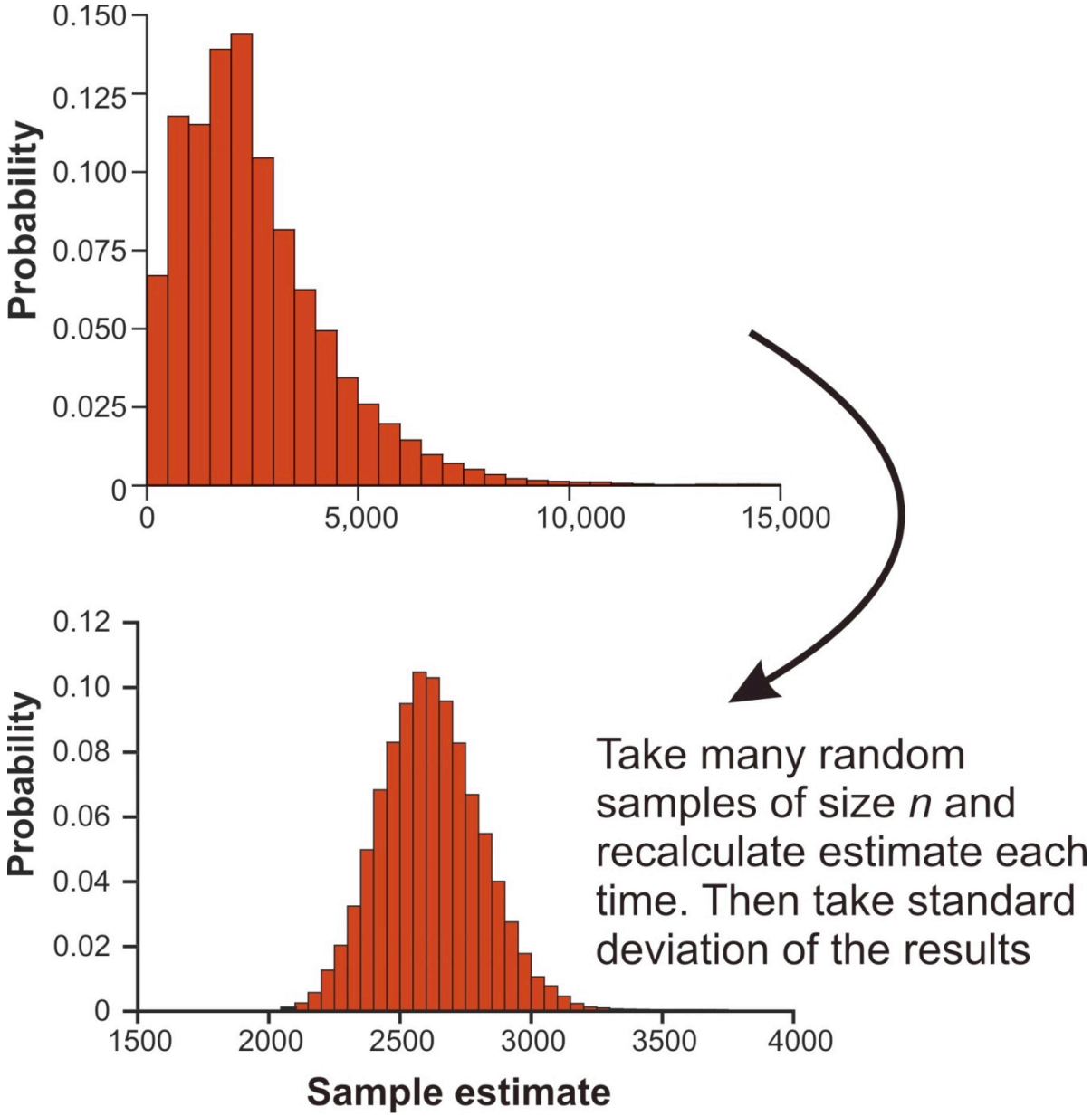
Method invented by Efron (1979).

**It is impractical to get sampling distribution by repeated sampling**

Sample many times from the same population

Calculate SE as the standard deviation of the resulting sampling distribution

But as we already noted, we only have <u>one</u> sample, and so only <u>one</u> estimate

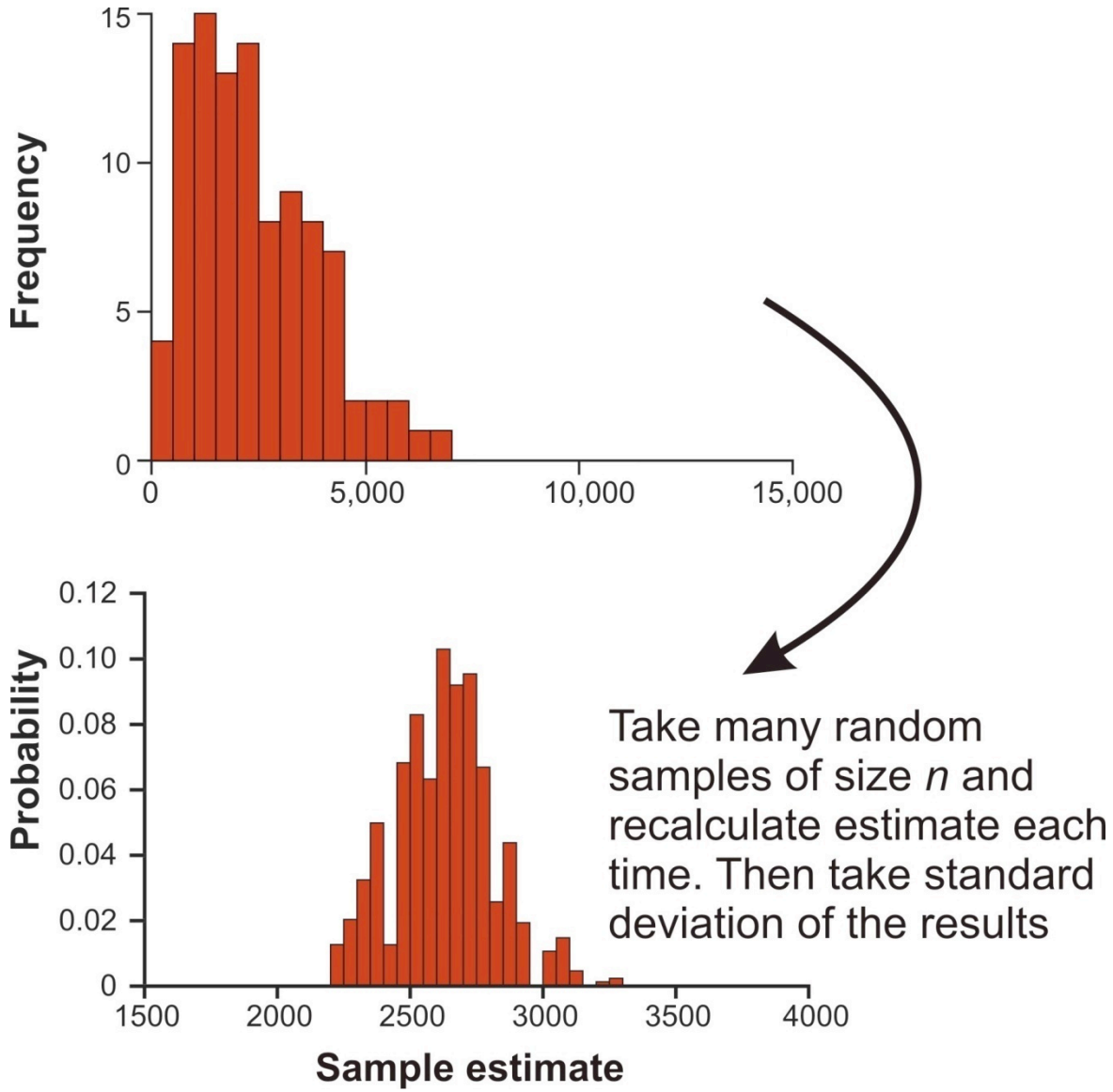


**The bootstrap sampling distribution is the next best thing**

Sample many times from the single sample instead. As if it were the population!

Sampling is “with replacement” so you don’t get the original sample each time

The standard deviation of results yields the bootstrap standard error



## The bootstrap algorithm

1. Use the computer to take a random sample of individuals from the original data. The bootstrap sample should contain the same number of individuals as the original data. Each time an observation is chosen, it is left available in the data set to be sampled again ("sampling with replacement").
2. Calculate the estimate using the measurements in the bootstrap sample from step 1. This is the first **bootstrap replicate estimate**.
3. Repeat steps 1 and 2 a large number of times (1000 times is reasonable). The frequency distribution of all bootstrap replicate estimates approximates the sampling distribution of the estimate.
4. Calculate the sample standard deviation of all the bootstrap replicate estimates obtained in step 3.

   The resulting quantity is called the **bootstrap standard error**.

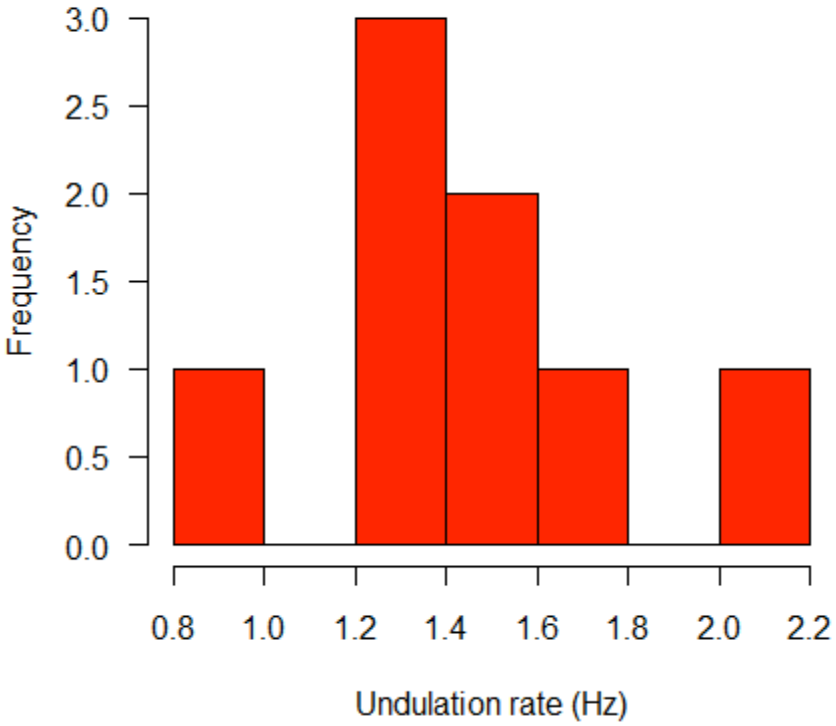
## Bootstrap example: sample mean

Data: Measurements of undulation rate (Hz) of paradise tree snakes
(Socha, J. J. 2002. Gliding flight in the paradise tree snake. Nature 418: 603–604)

*n* = 8 snakes*

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

$\overline{X}$ =1.375




*The bootstrap is not advised for sample sizes this small, but I use it here to illustrate

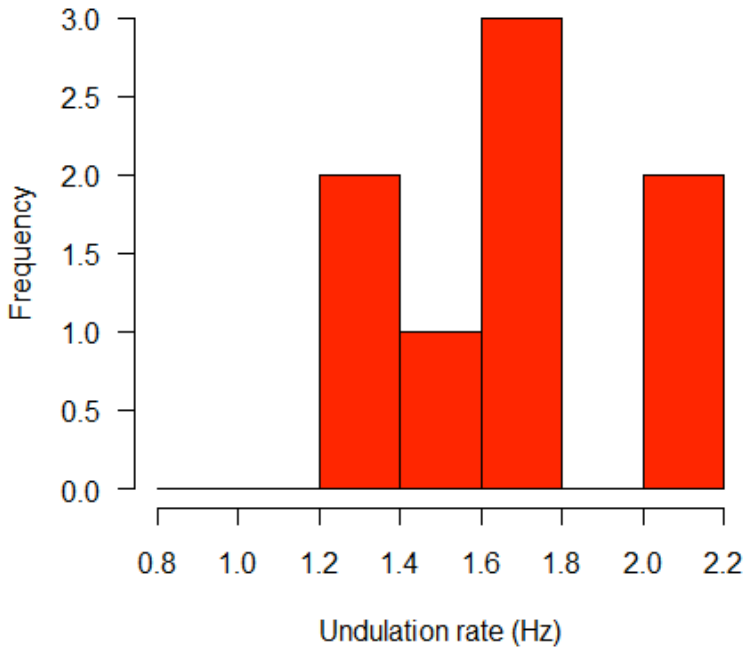
**Bootstrap example 1**

```
x<-c(0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0)
```

1. Use the computer to take a random sample of individuals from the original data

```
xboot <- sample(x, replace=TRUE)
  2.0, 1.3, 1.6, 1.2, 1.6, 1.4, 1.6, 2.0
```

Histogram of the first bootstrap sample



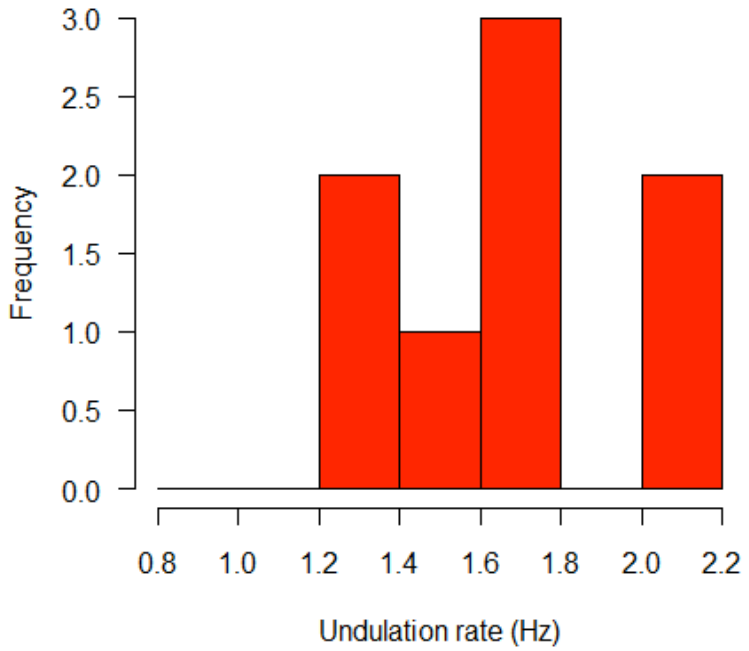
## Bootstrap example: sample mean

2. Calculate the estimate using the measurements in the bootstrap sample from step 1.

```
mean(xboot)
1.5875
```

save the result:

```
z[1]<-mean(xboot)
```

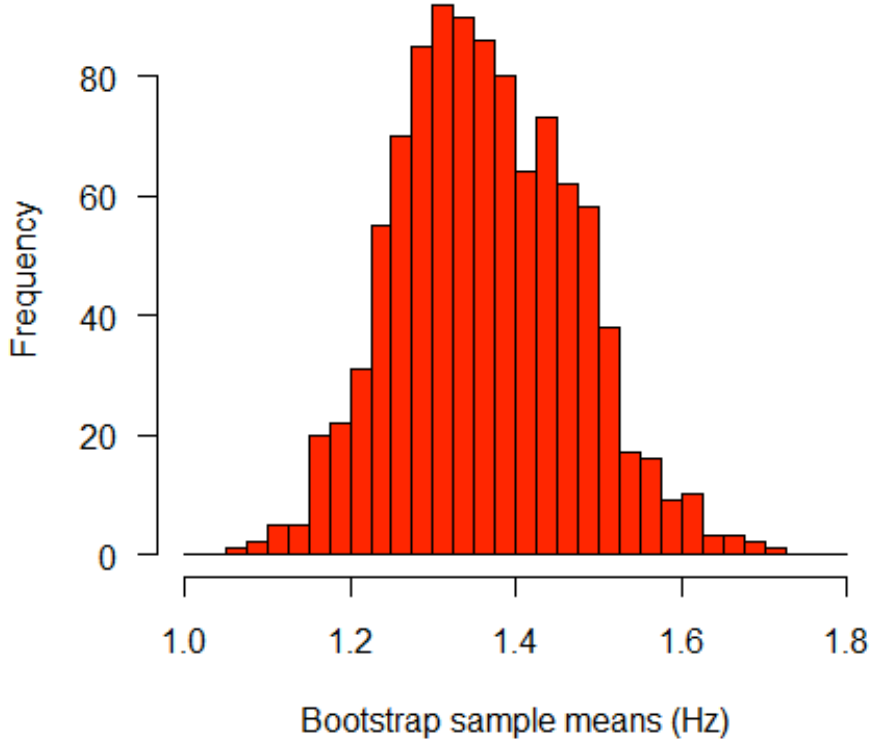


## Bootstrap: sample mean

3. Repeat steps 1 and 2 a large number of times (I used *B* = 1000).

```
xboot <- sample(x, replace=TRUE)
z[2]<-mean(xboot)
xboot <- sample(x, replace=TRUE)
z[3]<-mean(xboot)
…
z[1000]<-mean(xboot)
```

Plot of the bootstrap sampling distribution:



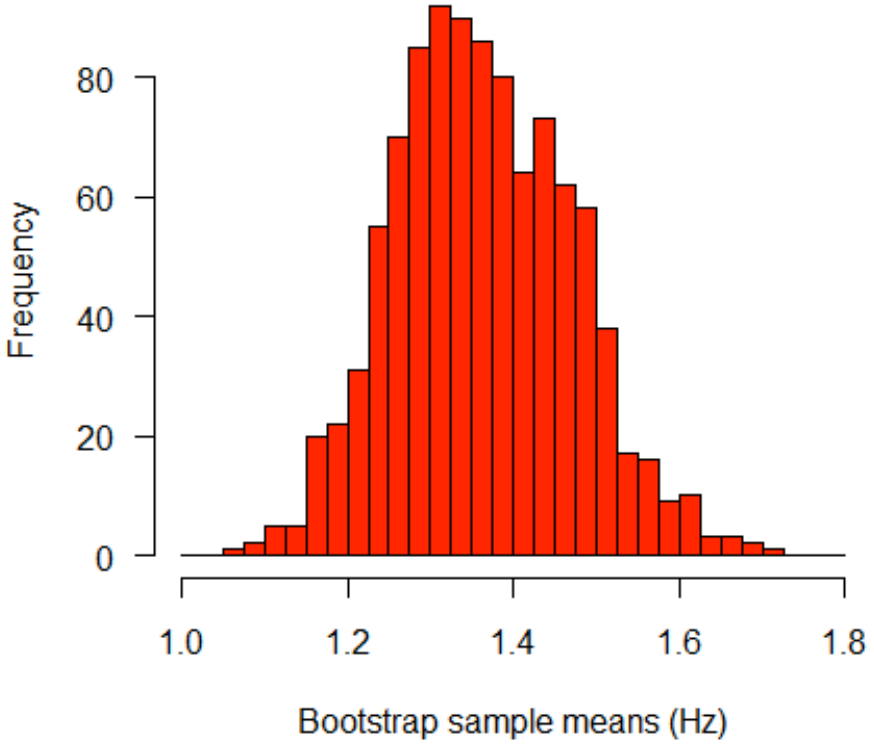
**Bootstrap example: sample mean**

4. Calculate the sample standard deviation of all the bootstrap replicate estimates obtained in step 3.

```
sd(z)
0.1070
```

How does it compare with the ordinary standard error of the mean?

```
se(x)
se(x$undulation.rate)
0.1146
```




The bootstrap SE is a little smaller (a consequence of very small sample size) but surprisingly close considering how we got it.

Pre-break factoid: the name ‘bootstrapping’

THE TRAVELS
AND SURPRISING ADVENTURES
OF
BARON MUNCHAUSEN

‘pigtailing’ turned into using his bootstraps instead.


Münchhausen O. Herrfurth pinx

All his crazy stories at:
http://homepage.ntlworld.com/forgottenfutures/munch/munch.htm

[break]

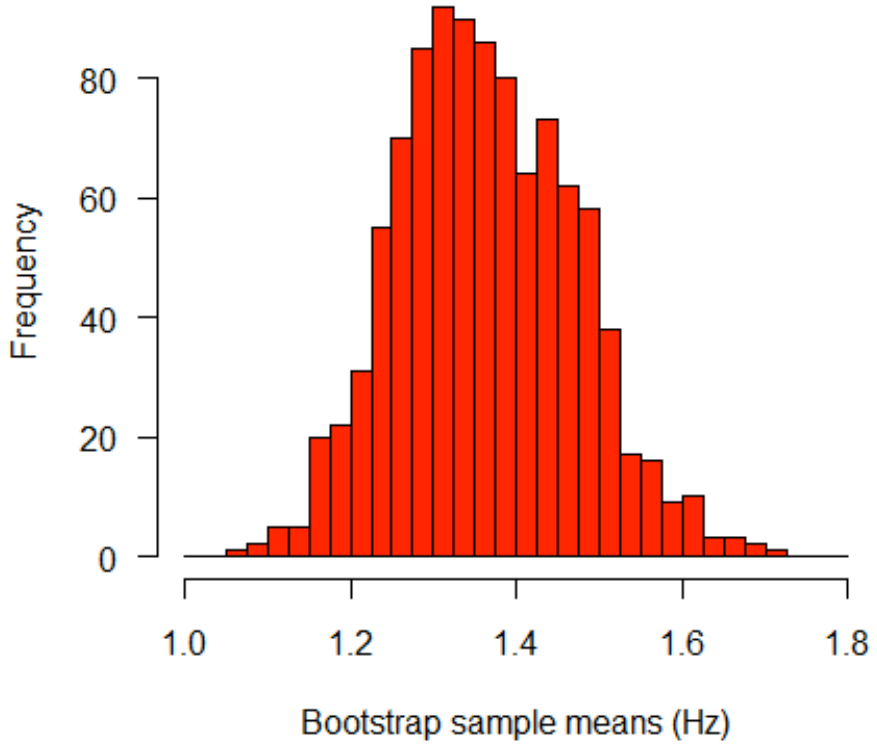
## Why the bootstrap is good

- It can be applied to almost any sample statistic, including means, proportions, correlations, regression
- Works when there is no ready formula for a standard error (e.g., median, trimmed mean, correlation, eigenvalue, etc.)
- It is nonparametric, so doesn't require normally-distributed data
- Works also for estimates based on complicated sampling procedures or calculations (for example, it is used in phylogeny estimation)

**It can even be used to calculate a confidence interval**

Incredibly, the 2.5$^{th}$ and 97.5$^{th}$ percentiles of the bootstrap sampling distribution are a passable 95% confidence interval, no transformations or normality assumptions needed

```
Level     Percentile
95%    ( 1.175,  1.600 )
```



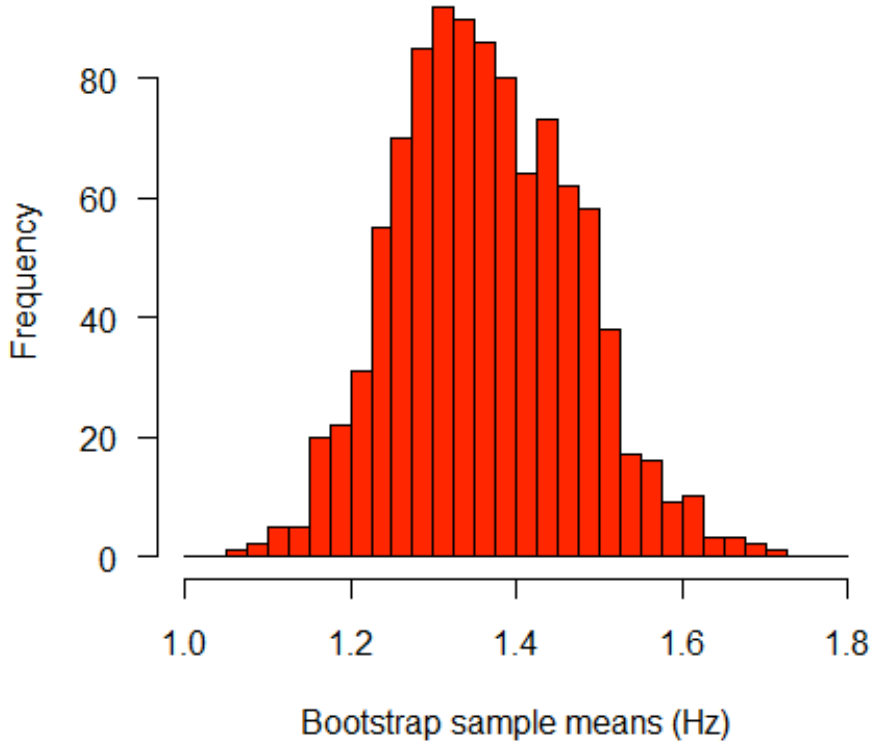

Compare with results from using the *t*-distribution:

```
95 percent confidence interval:
 1.104098 1.645902
```

## Confidence intervals

This “percentile” method works well if the sampling distribution is symmetric and unbiased

Improved, bias-corrected and accelerated (BCa) confidence intervals improve accuracy when sampling distributions are skewed and/or biased (see example in workshop)

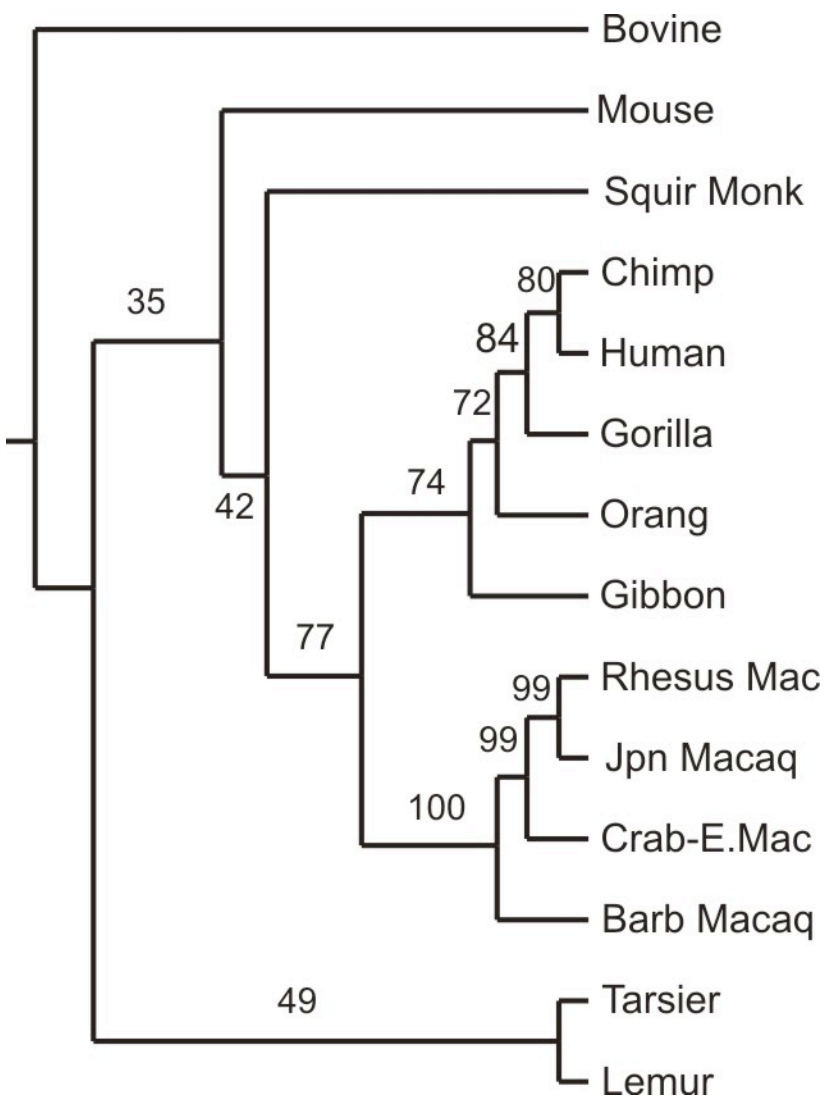


**Example of its use when estimation procedure is really complicated**

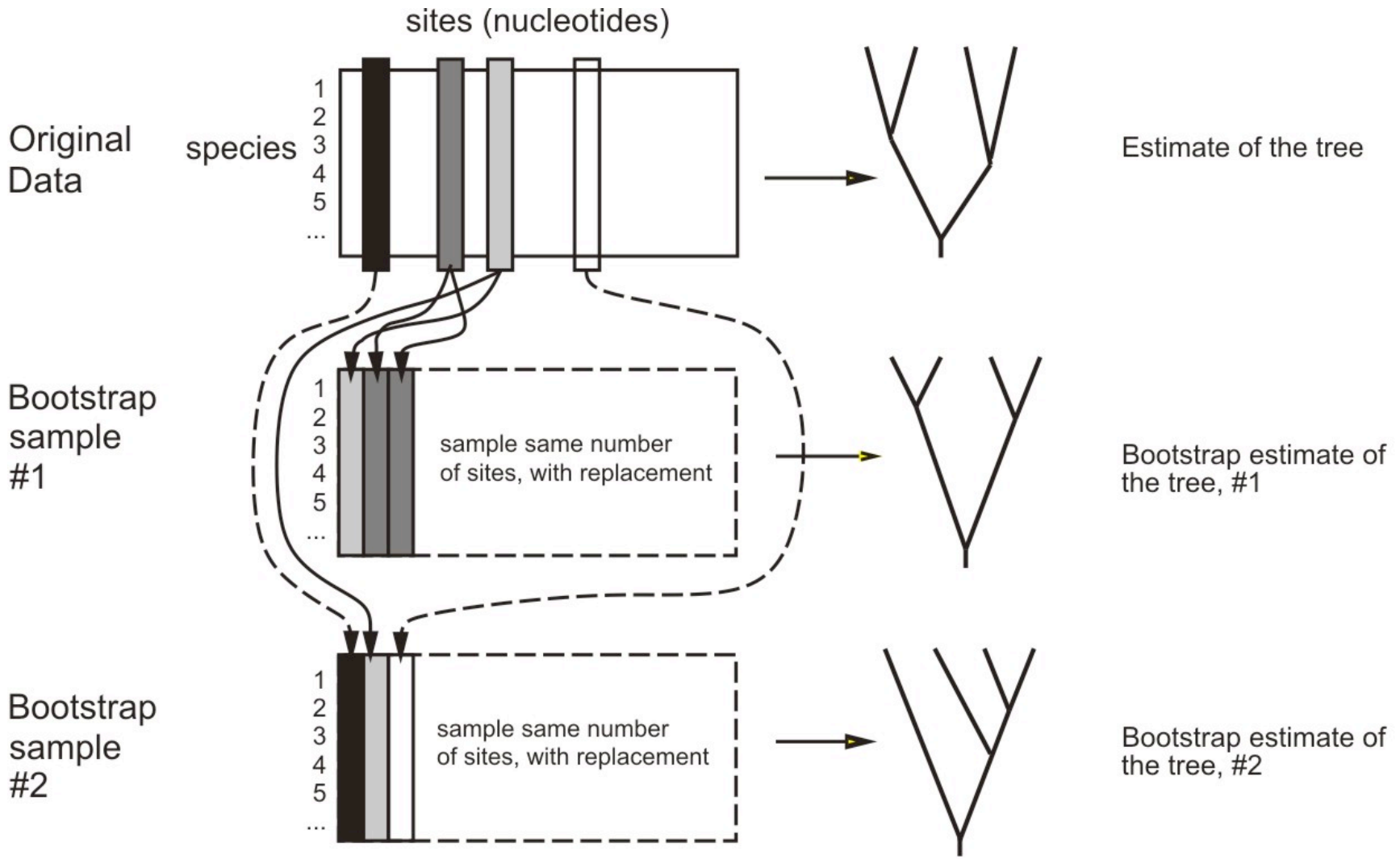
It is possible to estimate a phylogeny from gene sequence data

But how can we obtain the sampling distribution to provide a measure of uncertainty?

The bootstrap can help



# Bootstrap example: Phylogeny estimation




From Felsenstein Lecture 30 Phylogeny and bootstrap
From http://bioinfo.mbi.ucla.edu/courses/c260old/Lab/ March 2010

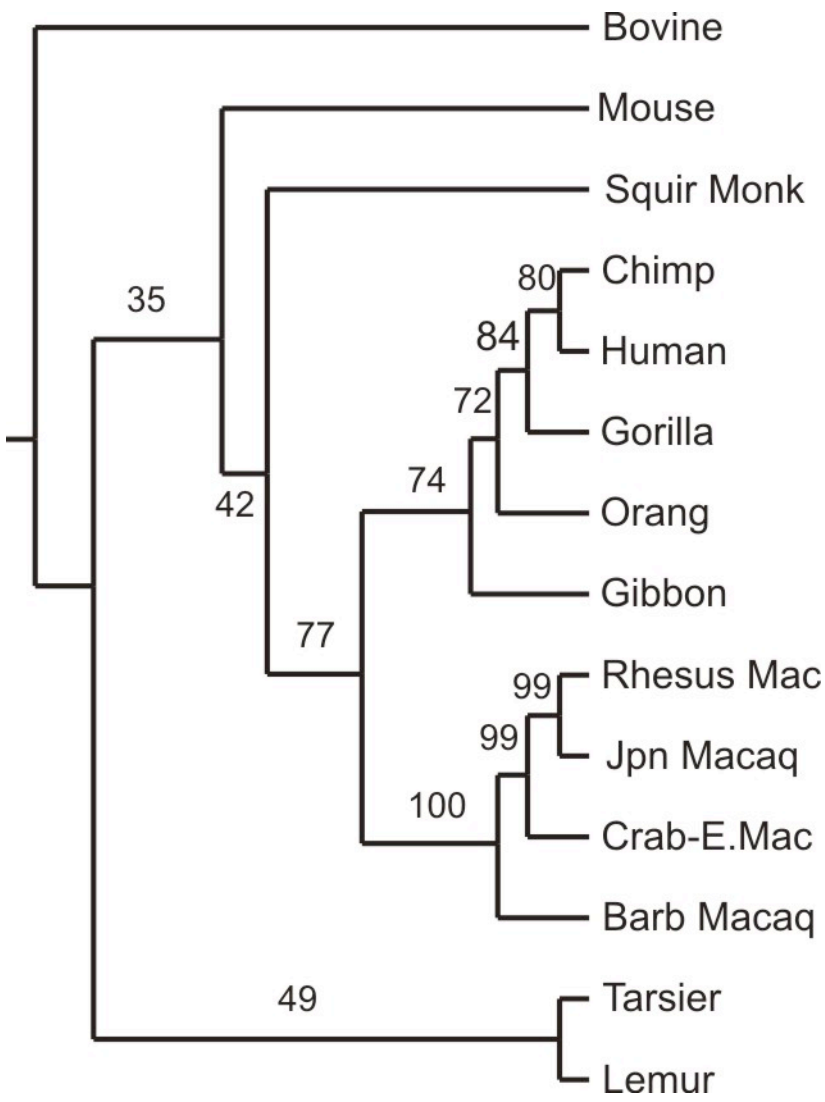
**Bootstrap example: Phylogeny estimation**

The true phylogeny (tree) plays the role of the parameter to estimate.

One ends up with a bootstrap sampling distribution of trees.

Then ask for each branch in the tree, how frequently it appears in the sampling distribution (= measure of confidence for the branch).

The final tree summarizes this for the most frequently occurring branches.



**Bootstrap example: Phylogeny estimation**

Method assumes that sites evolve independently and that sites have the same probability distribution of substitutions.

Simulations show that the fractions on the branches tend to be conservative when used to test the null hypothesis that the given branch is not present.

**Difference between two (or more) groups**

Procedure is similar, but now we resample both groups

1. Use the computer to take a random sample of the data from each group (with replacement, same sample sizes)
2. Calculate the difference between the two bootstrap samples from step 1.
3. Repeat steps 1 and 2 a very large number of times (≥1000)
4. Calculate the sample standard deviation of all the bootstrap replicate estimates obtained in step 3.

The result is the **bootstrap standard error** of the difference

## Bootstrap example: odds ratio to compare proportions

Data: Comparison of a proportion between two groups.

5th instar *Manduca sexta* caterpillars trained to associate a mild electrical shock with a specific odor (ethyl acetate; EA). Then assayed for learning in a Y-choice apparatus as larvae and again as adult moths, after metamorphosis (Blackiston et al. 2008. Retention of memory through metamorphosis: can a moth remember what it learned as a caterpillar? PLoS ONE 3: e1736)

| | **Caterpillar treatment** | |
|---|---|---|
| **Adult response** | learned | control |
| chose clean air | 32 | 25 |
| chose EA air | 9 | 21 |
| total | 41 | 46 |

**Bootstrap example: odds ratio to compare proportions**

We'll use the **odds ratio** to measure association between caterpillar treatment and adult response (difference between the proportions)

**Odds**: if we have a series of independent trials in which the probability of success in any one trial is *p*, then the odds of success is

$$O = \frac{p}{1-p}$$

If $O = 1$, the we say that the "the odds are one to one"

**Odds ratio:** Compares the odds of success under two treatments:

$$OR = \frac{O_1}{O_2}$$

## Bootstrap example: odds ratio to compare proportions

For the caterpillar data,

| | Caterpillar treatment | |
|---|---|---|
| **Adult response** | learned | control |
| chose clean air | 32 | 25 |
| chose EA air | 9 | 21 |
| total | 41 | 46 |

<u>learned:</u>

$p_1 = 32/41 = 0.78$

$O_1 = 0.78/0.22 = 3.56$

<u>control:</u>

$p_2 = 25/46 = 0.54$

$O_2 = 0.54/0.46 = 1.19$

$OR = O_1 / O_2 = 3.56/1.19 = 2.99$

The odds of choosing the clean air in a trial are about three times greater in the treatment group (learned) than in the control group.

**Bootstrap example: odds ratio to compare proportions**

We want a standard error for this estimate and a 95% confidence interval for the true (population) *OR*.

Let 1 indicate “chose clean air”, and 0 indicate “chose EA air”.

Learn group: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 ( thirty two 1’s and nine 0’s )

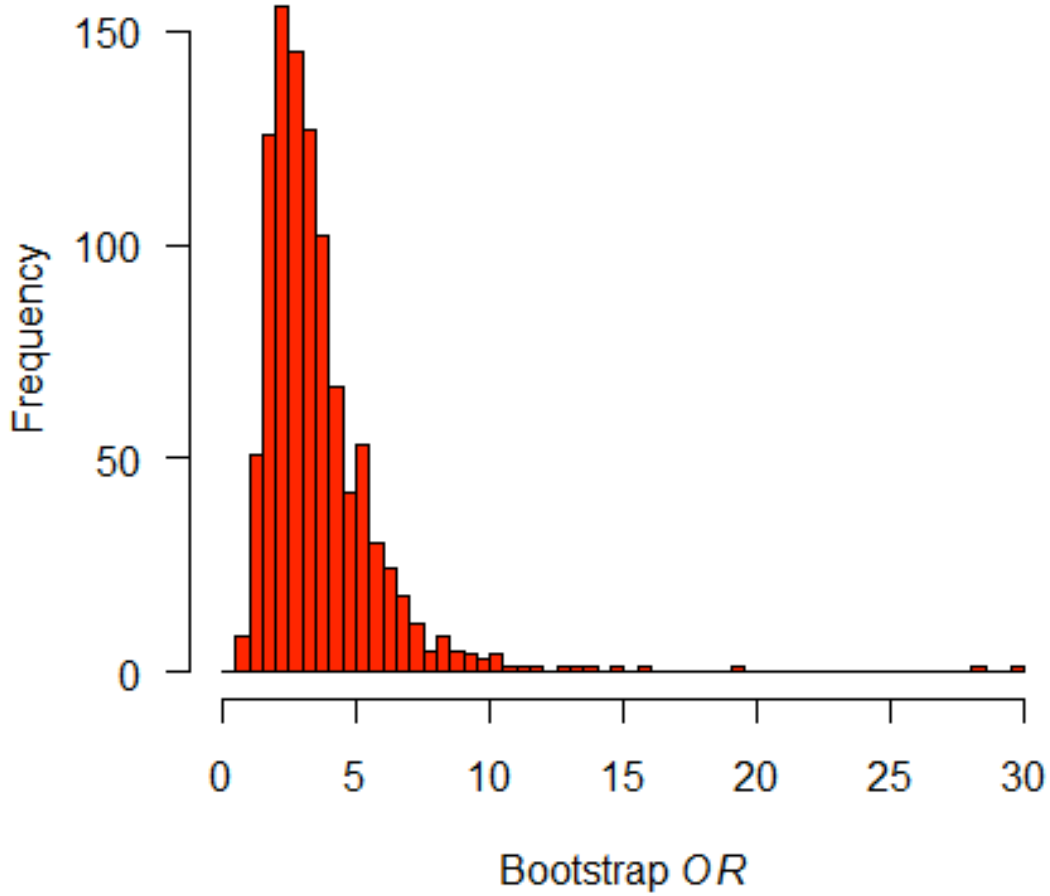
Control group: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ( twenty five 1’s and twenty one 0’s )

Step 1: take a random sample (with replacement) from each group
Step 2: calculate *OR*
Step 3: repeat 1000 times

**Bootstrap example: odds ratio to compare proportions**

Bootstrap sampling distribution for *OR*:

Bootstrap SE = 2.26



**Bootstrap example: odds ratio to compare proportions**

Crude 95% confidence interval from the percentiles of the bootstrap sampling distribution:

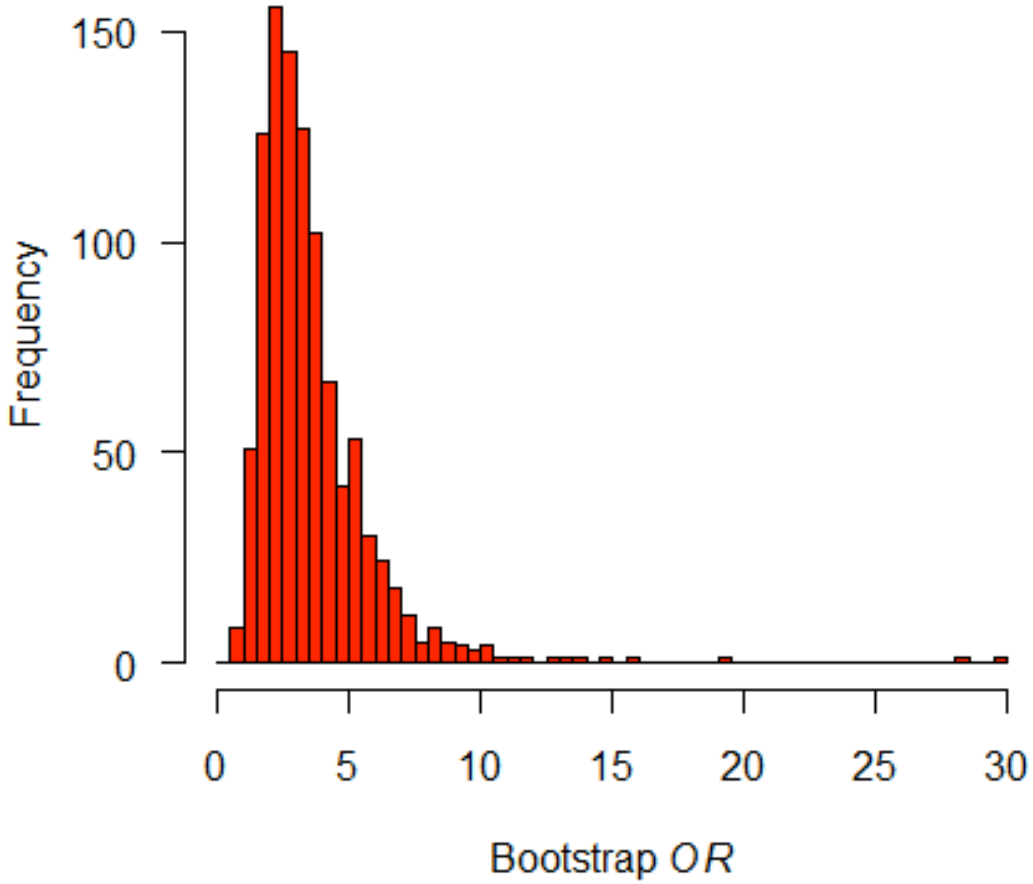
| 2.5% | 97.5% |
|---|---|
| **1.21** | **8.67** |

$BC_a$ (bias corrected and accelerated) interval

| 2.5% | 97.5% |
|---|---|
| **1.14** | **7.93** |

Compare with conventional approximate CI for *OR*

| 2.5% | 97.5% |
|---|---|
| **1.17** | **7.65** |




$BC_a$ corrects the percentiles for skewness in the sampling distribution, which otherwise results in the shape of the bootstrap sampling distribution changing with the estimate; and for bias in the estimate.

**Summary**

- The bootstrap is amazing and useful for estimation.
- It works in almost any situation (except when *n* is small).
- It is approximate, though performs almost as well as parametric methods when assumptions of the parametric methods are met.

**Hypothesis testing with the bootstrap**

The approximate 95% confidence interval obtained using the bootstrap method allows hypothesis testing: any number included within the interval is not rejected by the data.

The usual null and alternative hypothesis for *OR* is
$H_O$: $OR = 1$
$H_A$: $OR \neq 1$

Since our 95% confidence interval was

$$1.14 < OR < 7.93$$

We can reject $H_O$.

But there is a more exact method for hypothesis testing….

**Randomization test (= permutation test)**

In data analysis we use two different types of sampling distributions.

1. **Estimation**

Sampling distribution of an estimate (of a population parameter). Needed for standard error, confidence intervals.

2. **Hypothesis testing**

Null sampling distribution (simply, the null distribution). Models the distribution of a test statistic under the null hypothesis. We frequently use the $t$, $F$, $\chi^2$, and normal distributions to approximate null distributions.

If the assumptions of these standard methods are violated, the computer can be used to carry out randomization to generate null distributions.

## Null distribution

Needed to calculate
*P*-value for test

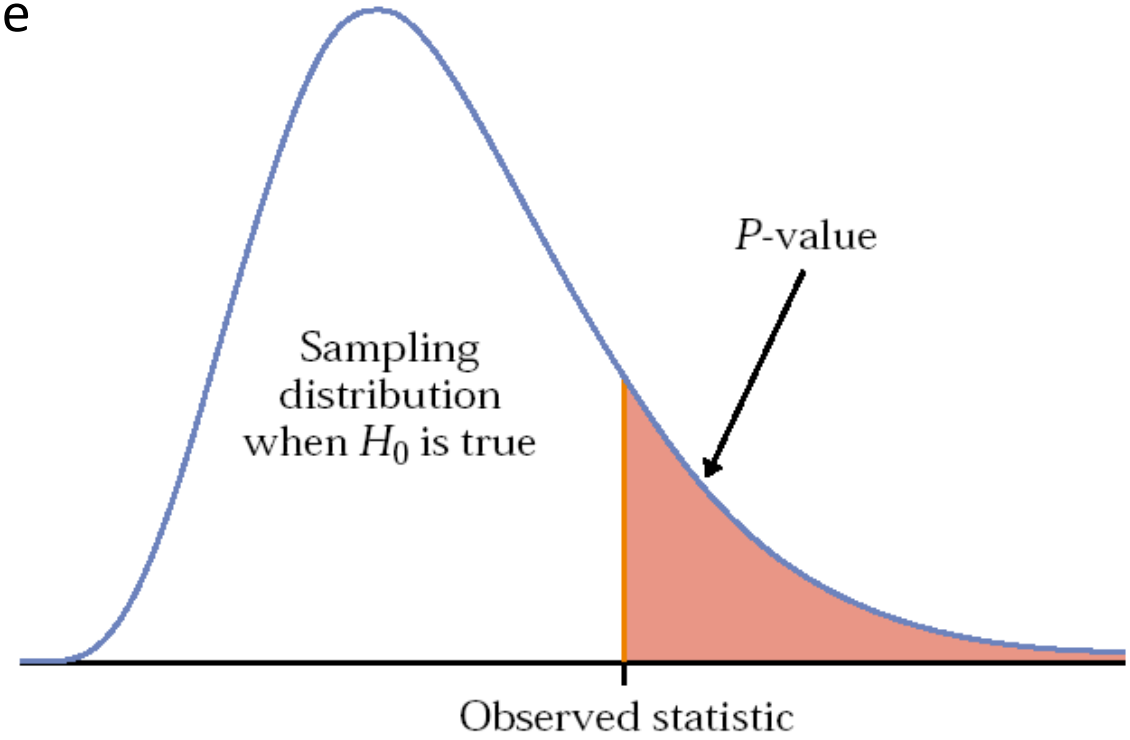



**FIGURE 14.19** The *P*-value of a statistical test is found from the sampling distribution the statistic would have if the null hypothesis were true. It is the probability of a result at least as extreme as the value we actually observed.

**Randomization test**

A **randomization test** generates a null distribution for the association between two variables (difference between groups) by repeatedly and randomly rearranging the values of one of the two variables in the data

**Example:**
Comparison of reproductive success of female pseudoscorpions randomly assigned one of two treatments: mated to two different males (DM); mated twice to the same male (SM). Experiment tested whether genetic diversity contributed to reproductive success (Newcomer et al. 1999).

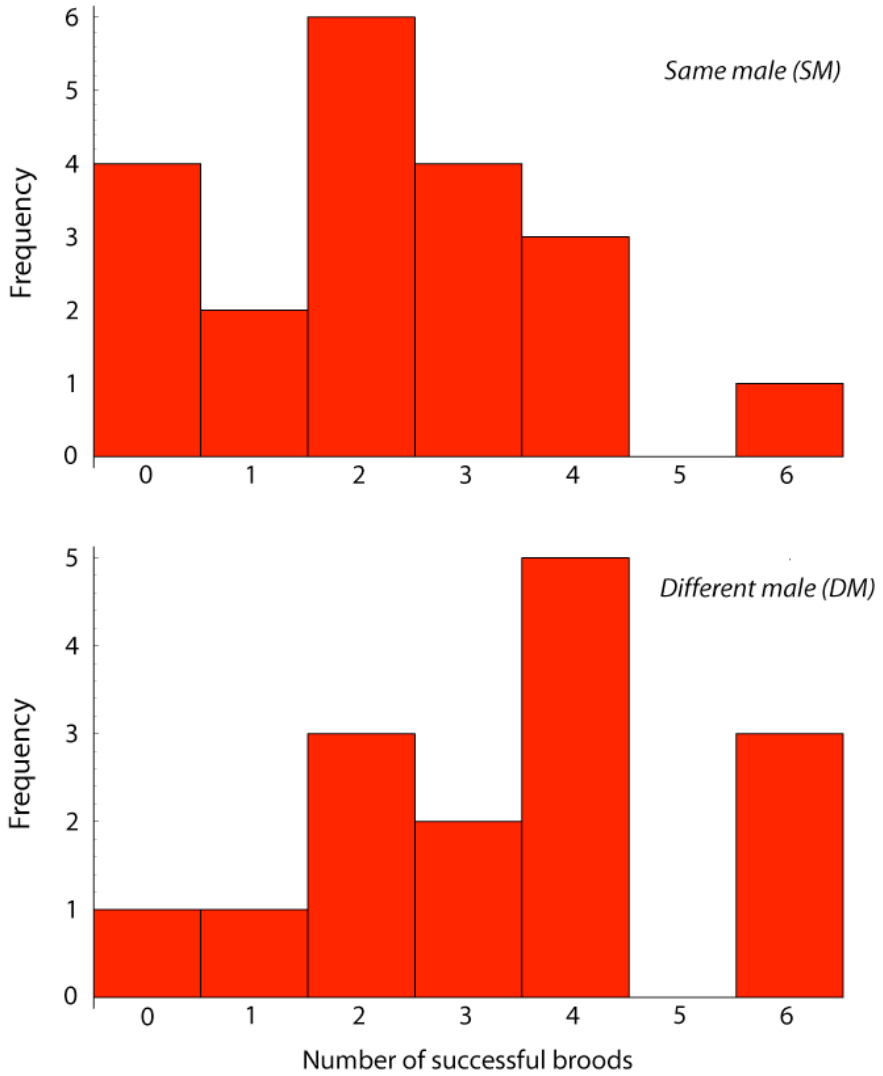
## Randomization test example 1

$H_O$: $\mu_1 = \mu_2$
$H_A$: $\mu_1 \neq \mu_2$

The randomization test requires no assumption of normality
(it is nonparametric, like bootstrap)

It is more powerful than rank tests because it uses all the data, not just their ranks

It is somewhat more robust than rank tests to departures from the assumption of equal distribution shapes



**Randomization test example 1**

Data: The number of successful broods of pseudoscorpion females that were mated twice to either a single male (SM) or two different males (DM).

SM: 4 0 3 1 2 3 4 2 4 2 0 2 0 1 2 6 0 2 3 3

DM: 2 0 2 6 4 3 4 4 2 7 4 1 6 3 6 4

$$\bar{Y}_{\text{SM}} - \bar{Y}_{\text{DM}} = 2.2 - 3.625 = -1.425$$

**Steps of the randomization test:**

1. Create a randomized data set in which the measurements are randomly reassigned (without replacement) to the two groups (color coding indicates original treatment assigned)

SM: 4 0 7 4 2 2 2 1 4 4 0 3 3 4 6 2 4 6 0 0

DM: 2 2 3 3 2 3 3 4 1 2 1 4 6 6 2 0

**Randomization test example 1**

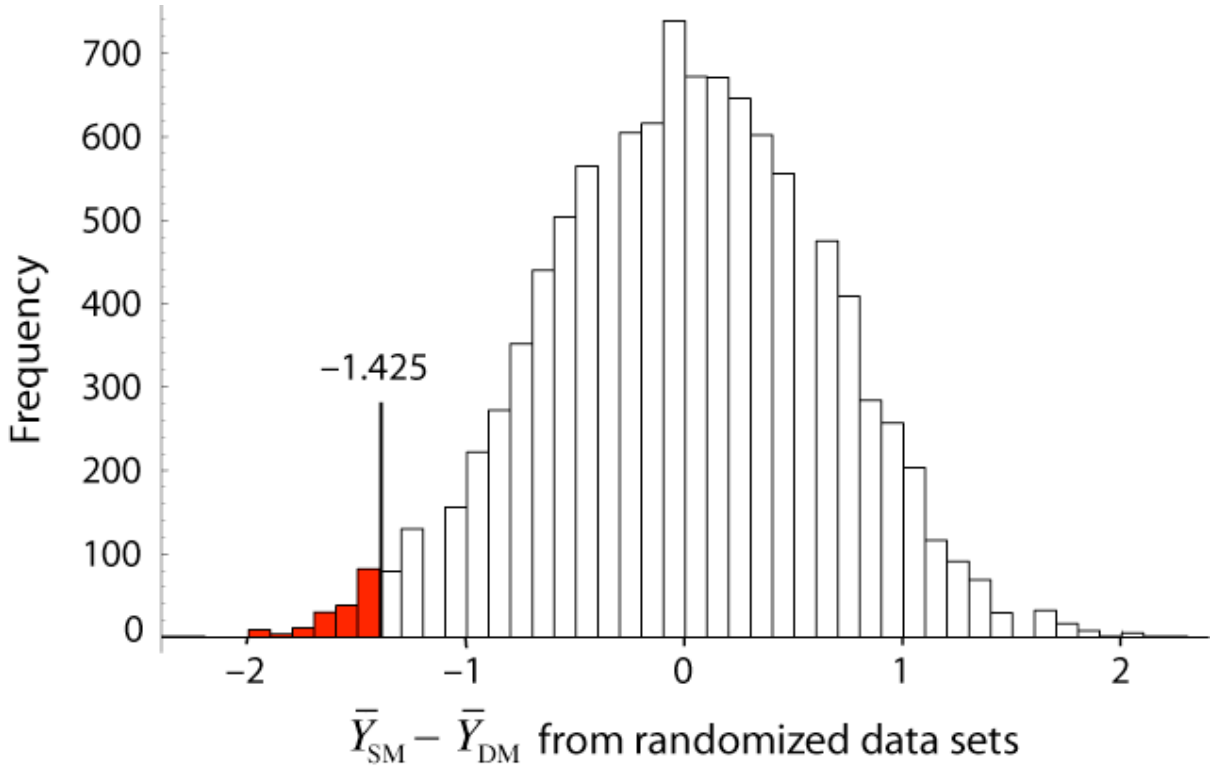
SM: 4 0 7 4 2 2 2 1 4 4 0 3 3 4 6 2 4 6 0 0
DM: 2 2 3 3 2 3 3 4 1 2 1 4 6 6 2 0

2. Calculate the test statistic measuring the association between variables (difference between group means) for the randomized sample

$$\bar{Y}_{\text{SM}} - \bar{Y}_{\text{DM}} = 2.9 - 2.75 = 0.15$$

3. Repeat steps 1 and 2 many times

## Randomization test example 1




The null distribution of the test statistic from 10,000 replicates of the randomization process applied to the pseudoscorpion data. –1.425 is the observed value from the data. The area in red is the tail beyond this observed value.

**Randomization test example**

Of these 10,000 randomizations, only 176 had a test statistic (difference between means) equal to or less than the observed value, –1.425.

Use the simulated null distribution to calculate *P*-value in the same way that we use the *t* or *F* distribution in conventional tests. The proportion of values in the null distribution that equal or lie farther in the tail than the observed value of the test statistic is 176/10000 = 0.0176. Since the test is two-tailed,

$P = 2(0.0176) = 0.0352$

We reject the null hypothesis. Mean reproductive success was detectably higher for female pseudoscorpions that mated with two different males than when they mated twice with the same male.

**Summary**

- Computer-intensive resampling methods have great advantages in many situations. They are accessible in R, as we shall see.
- The bootstrap is a recent and amazing invention for estimation
- It works in almost any situation (except when *n* is small)
- It is approximate, though performs almost as well as parametric methods when assumptions of the parametric methods are met
- Conventional nonparametric tests are randomization tests but based on ranks, which throws away the numbers. Randomization tests based on the actual data are more powerful and more robust to violations of the equal shapes assumption
- Randomization tests are less powerful than parametric tests when sample size is small

**Discussion paper for next week:**

Palmer (1999) Meta-analysis of fluctuating asymmetry and sexual selection.

Download from “**assignments**” tab on course web site.