Bootstrap and resampling

Outline for today

- Estimation and hypothesis testing
- Permutation test
- Estimation
- The sampling distribution
- The bootstrap standard error
- The bootstrap confidence interval
- Comparing two groups
- Summary

Estimation and hypothesis testing

In conventional data analysis we carry out two types of statistical inference. Each is founded on a different sampling distribution.

1. **Estimation**

Uses the sampling distribution of an estimate: all the values for a parameter estimate we might obtain, when sampling from a population, and their probabilities. It is used to obtain standard errors, confidence intervals.
Most methods assume that the sampling distribution is approximately normal.

2. **Hypothesis testing**

Uses the null sampling distribution (or null distribution): the probability distribution of a test statistic if the null hypothesis is true. We frequently use the t, F, χ^2, and normal distributions to approximate null distributions, from which P-values are calculated.

Estimation and hypothesis testing

Q: What to do if the assumptions of the best method available are violated, and we cannot turn to linear or generalized linear models (because their assumptions are also violated)?

A: Computer-intensive methods.

An approach in which the power of the computer is used to generate a sampling distribution.

1. Estimation: The **bootstrap**.

2. Hypothesis testing: The **permutation test**.

Permutation test

A **permutation test** generates a null distribution for a statistic measuring association between two variables (or difference among groups) by repeatedly and randomly rearranging the values of one of the variables.

Rank tests, such as the Mann-Whitney *U*-test for two samples, are permutation tests. The data are first replaced by their ranks, and then the ranks are permuted to generate a null distribution. The exact probability distribution of the *U*-statistic is known. But replacing the data by their ranks loses information.

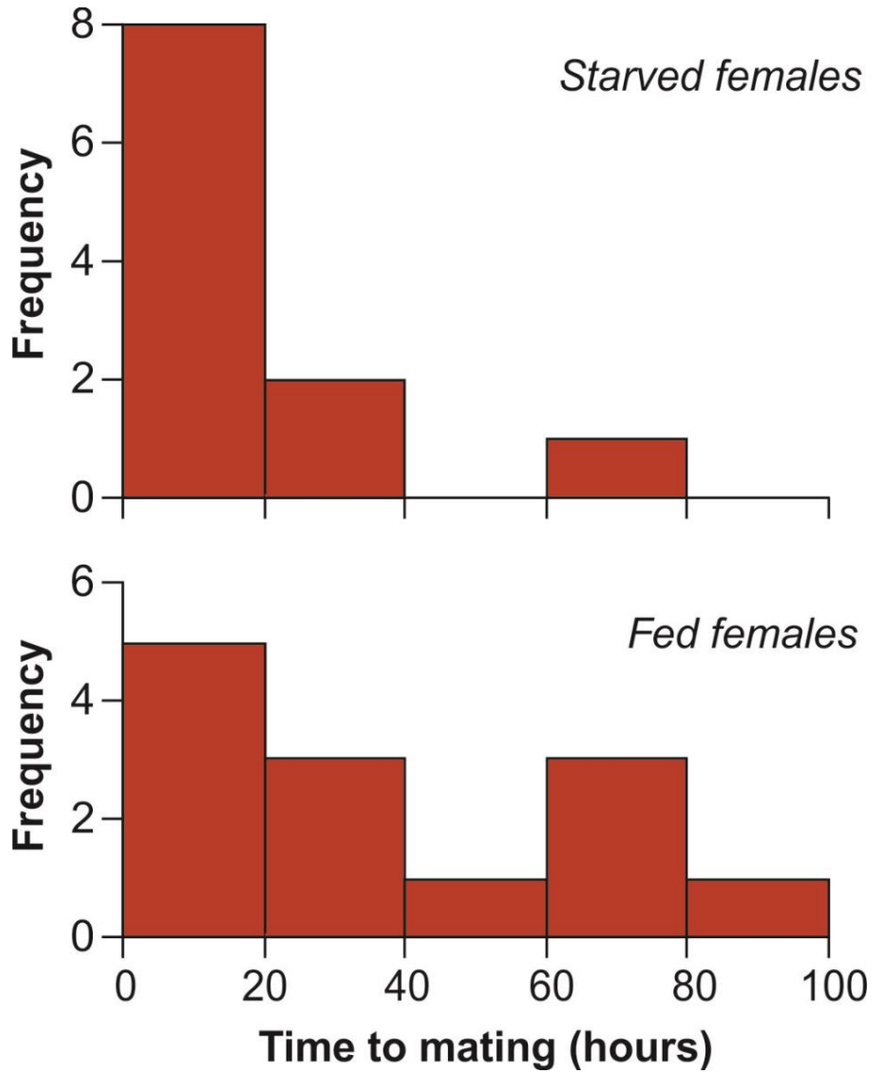
Permute the data themselves! There's no need to replace the data with the ranks. No known probability distribution is available, so we used the computer to generate a large number of permutations instead to approximate a null distribution.

Permutation test example

During mating in the sage cricket, *Cyphoderris strepitans*, the male offers his fleshy hind wings to the female to eat. Females get some nutrition from feeding on the wings, which raises the question, "Are females more likely to mate if they are hungry?" Johnson et al. (1999) addressed this question by randomly dividing 24 females into two groups:

One group of 11 females was starved for at least two days.
Another group of 13 females was fed during the same period.
Each female was put separately into a cage with a single (new) male, and the waiting time to mating was recorded.

The data are clearly not normally distribution.



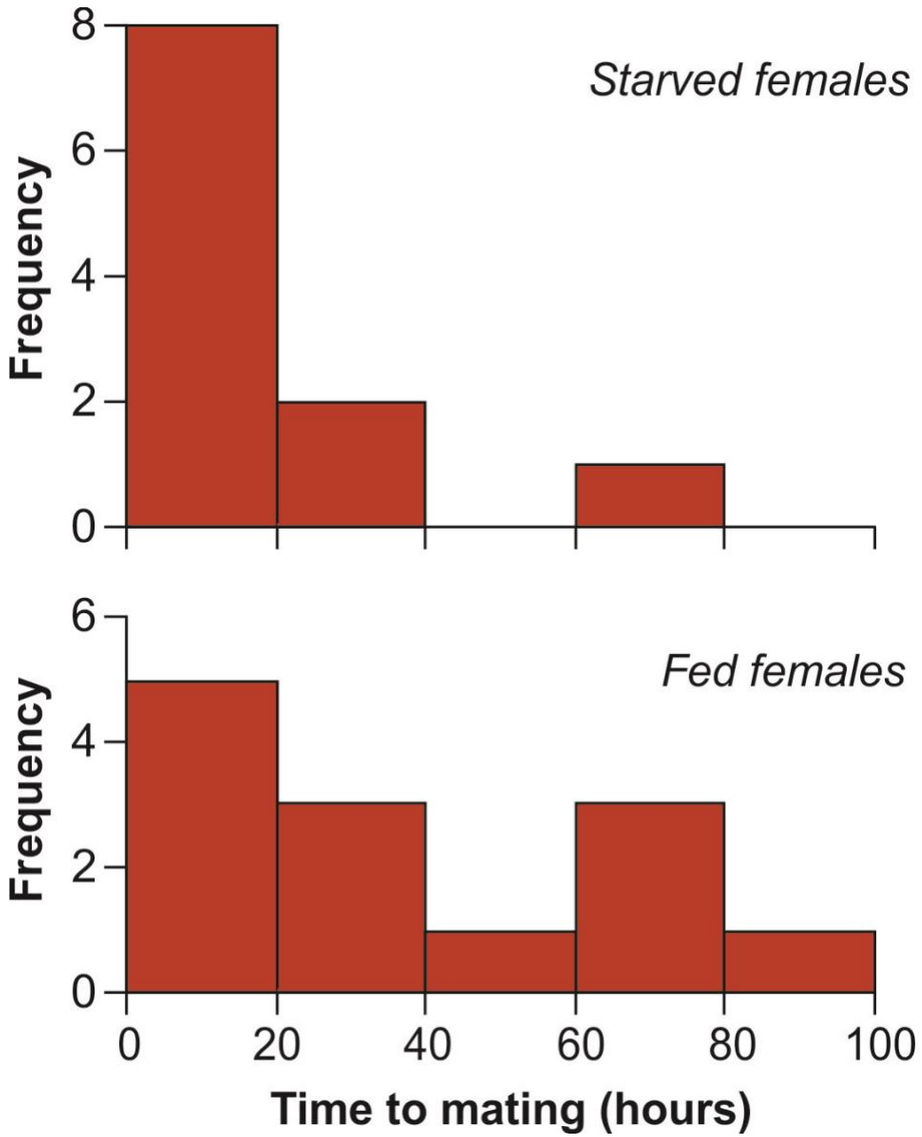
Permutation test example

Treatment	Time (hrs)	Treatment	Time (hrs)
Starved	1.9	Fed	1.5
Starved	2.1	Fed	1.7
Starved	3.8	Fed	2.4
Starved	9.0	Fed	3.6
Starved	9.6	Fed	5.7
Starved	13.0	Fed	22.6
Starved	14.7	Fed	22.8
Starved	17.9	Fed	39.0
Starved	21.7	Fed	54.4
Starved	29.0	Fed	72.1
Starved	72.3	Fed	73.6
		Fed	79.5
		Fed	88.9

H_0: Mean time to mating, $\mu_1 = \mu_2$

H_A: Mean time to mating, $\mu_1 \neq \mu_2$ (two-tailed test)

Test statistic: $\bar{Y}_1 - \bar{Y}_2 = 17.73 - 35.98 = -18.26$.



Permutation test example

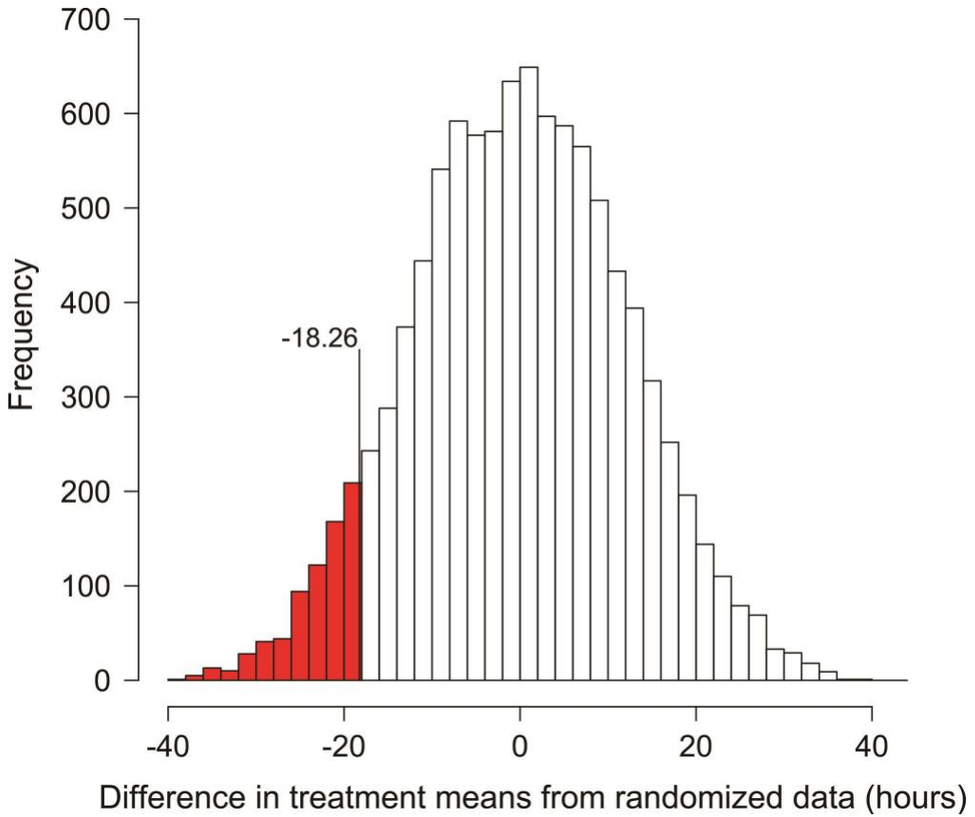
Outcome of a single permutation:

Treatment	Time (hrs)	Treatment	Time (hrs)
Starved	3.8	Fed	14.7
Starved	9.0	Fed	21.7
Starved	3.6	Fed	1.7
Starved	79.5	Fed	2.1
Starved	17.9	Fed	1.5
Starved	22.8	Fed	2.4
Starved	54.4	Fed	5.7
Starved	13.0	Fed	39.0
Starved	9.6	Fed	29.0
Starved	1.9	Fed	72.1
Starved	22.6	Fed	88.9
		Fed	72.3
		Fed	73.6

Test statistic: $\bar{Y}_1 - \bar{Y}_2 =$ 21.65 – 32.67 = –11.02.

Permutation test example

Results of 10,000 permutations: The null distribution of $\bar{Y}_1 - \bar{Y}_2$



Tail of distribution: 712/10,000 had a value of $\bar{Y}_1 - \bar{Y}_2$ less than or equal to observed value, –18.26.

P = **2** × 712/10000 = 0.1424

Permutation test assumptions

- Random samples

- To compare means or medians between groups, permutation tests assume that the distribution of the variable has the same shape in every population.

Permutation tests are robust to departures from the equal-shape assumption when sample sizes are large (more so than the Mann-Whitney *U*-test).

Permutation tests have lower power than parametric tests when the sample size is small, but they are more powerful than the Mann-Whitney *U*-test. They have similar power to parametric tests when sample size is large.

Why I don’t love permutation tests (or rank tests):

- Parametric methods provide estimates (with standard errors and confidence interval) of a useful parameter.
- Nonparametric tests, including permutations tests and rank tests, provide only a P-value. They do not provide estimates of magnitudes (effect sizes) with standard errors or confidence intervals. They perpetuate the mistake that the P-value is all you need, and that the smallness of the P-value indicates the importance of an effect.
- As our readings and discussions have stressed, the P-value in fact tells us nothing about magnitudes or biological importance.

No important conclusion in biology should ever be drawn from a *P*-value alone. Needs to be accompanied by a method for estimation (i.e., confidence intervals). Robust methods are one solution, but here I’ll describe the bootstrap.

Why I love the bootstrap:

- Used for estimation, mainly.
- Provides standard errors and confidence intervals of useful parameters (magnitudes).
- The method is nonparametric, so doesn't require normally-distributed data. It makes no assumptions about the distribution of the data.
- It can be applied to virtually any population parameter, including means, proportions, and linear model coefficients.
- It is most handy when there is no ready formula for a standard error or confidence interval (e.g., median, trimmed mean, eigenvalue).
- It even works also for estimates based on complicated sampling procedures or calculations (for example, it is used to measure uncertainty in phylogeny estimation).

To understand the bootstrap, let's review how estimation works

Estimation is the process of inferring a population parameter from sample data.

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

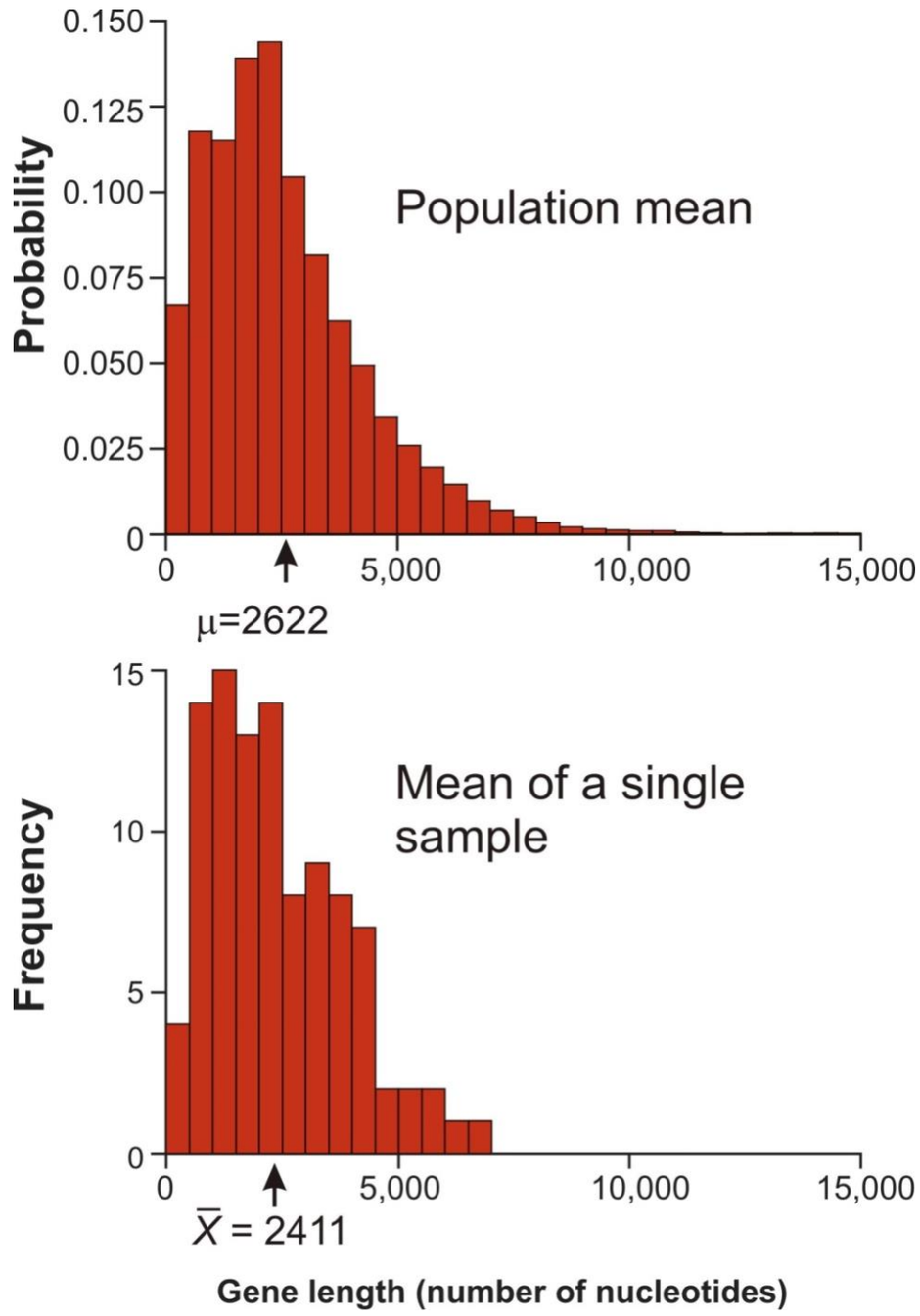
The sampling distribution of an estimate gives all the values we might have obtained from our sample, and their probabilities of occurrence.

The standard error of an estimate is the standard deviation of its sampling distribution. Standard error therefore measures the uncertainty of an estimate.

Example: Estimate a mean

What we want:
The mean of a variable in the population (e.g., the lengths of all the genes in the human genome).

What we have instead:
The sample mean (e.g., based on a random sample of n = 100 genes)

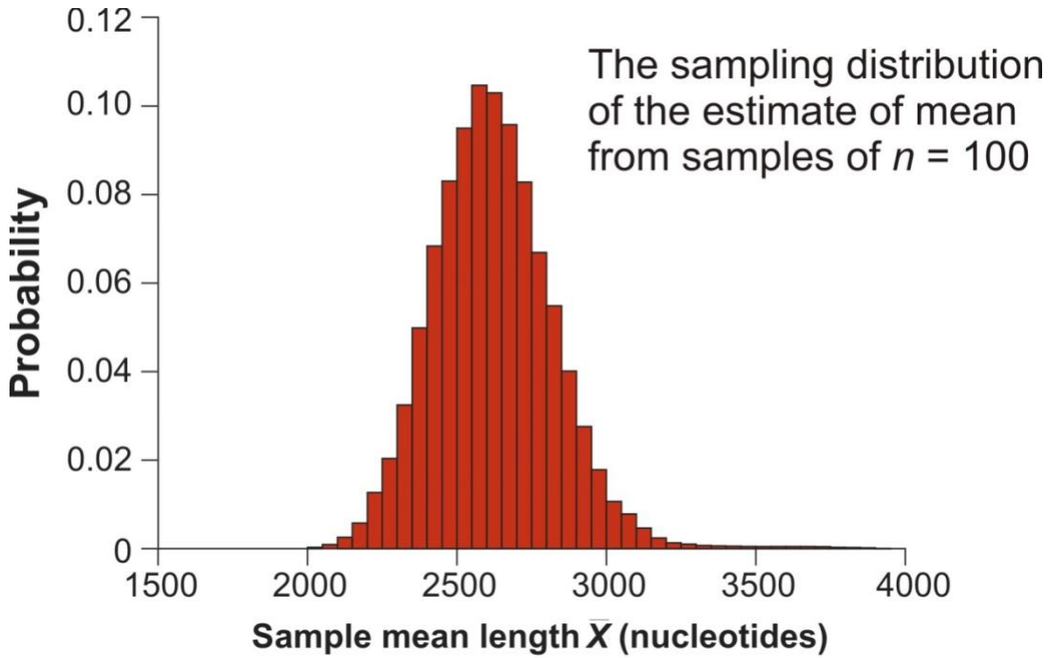


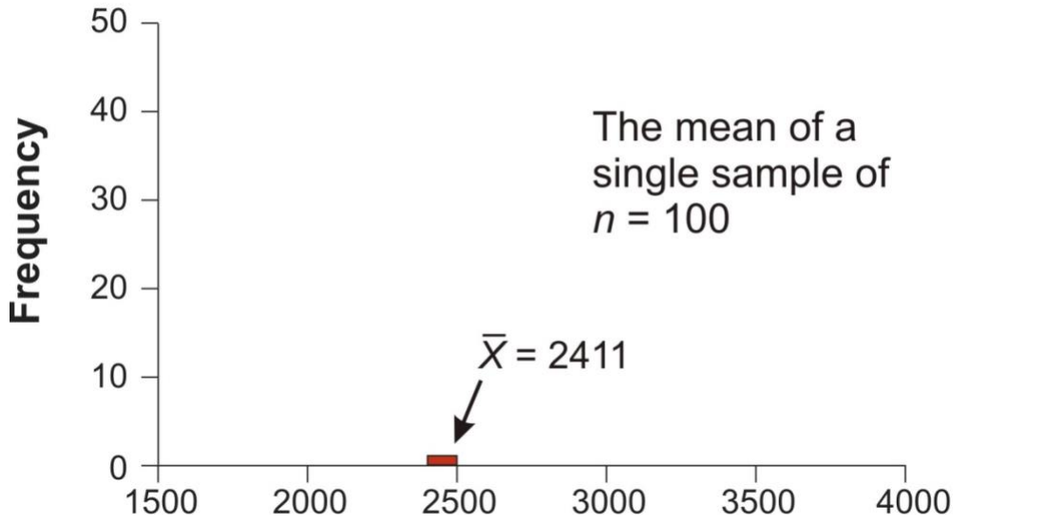
The sampling distribution

We don’t know the true mean.

So, we want an approximation of the sampling distribution, the distribution of estimates we *might* obtain from random sampling and their probabilities.

What we have instead:
Just one sample mean

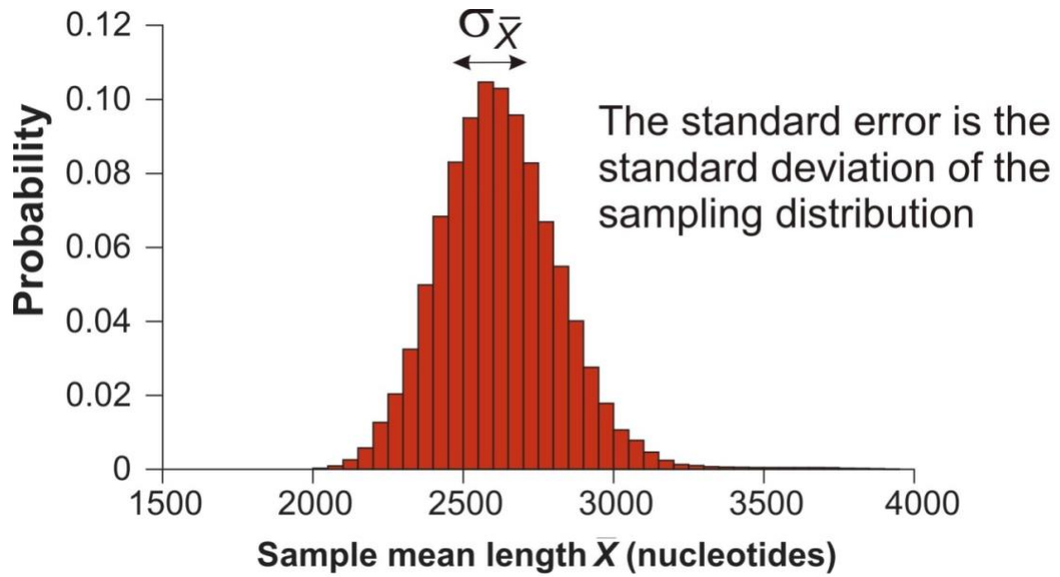


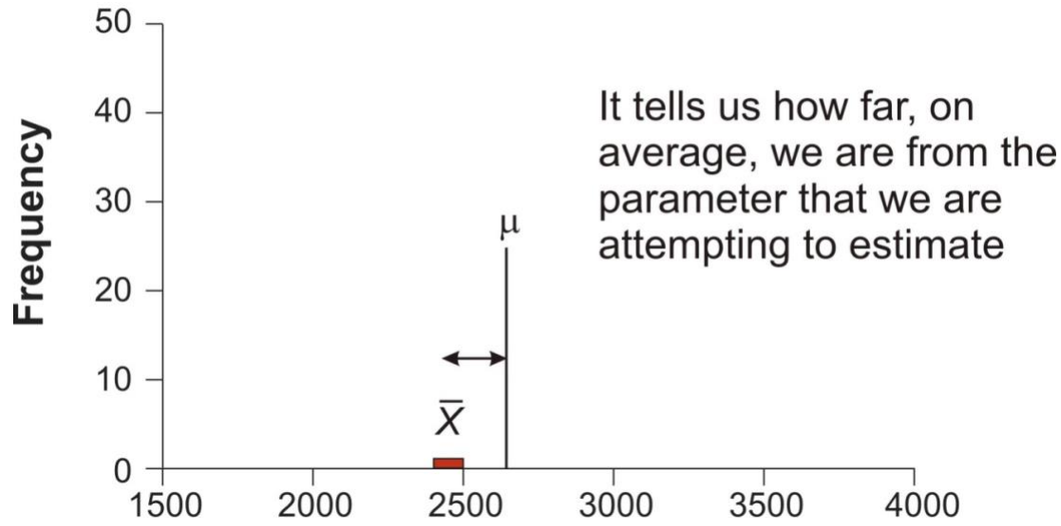


Standard error

The standard deviation of the sampling distribution (the standard error) measures the variation of sample estimates around the population parameter.

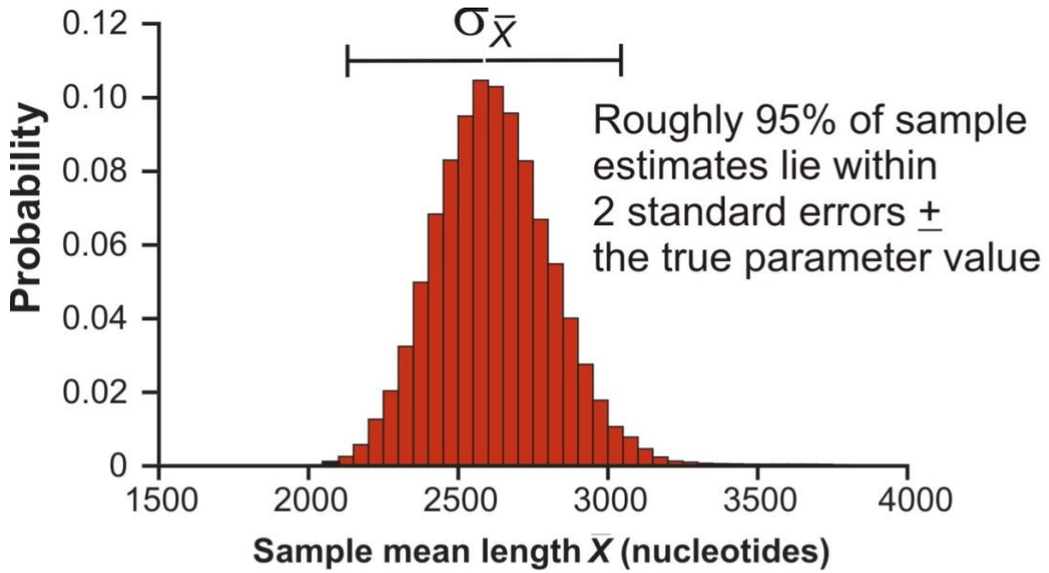
Roughly, the standard error tells us how far we are from the truth, on average.



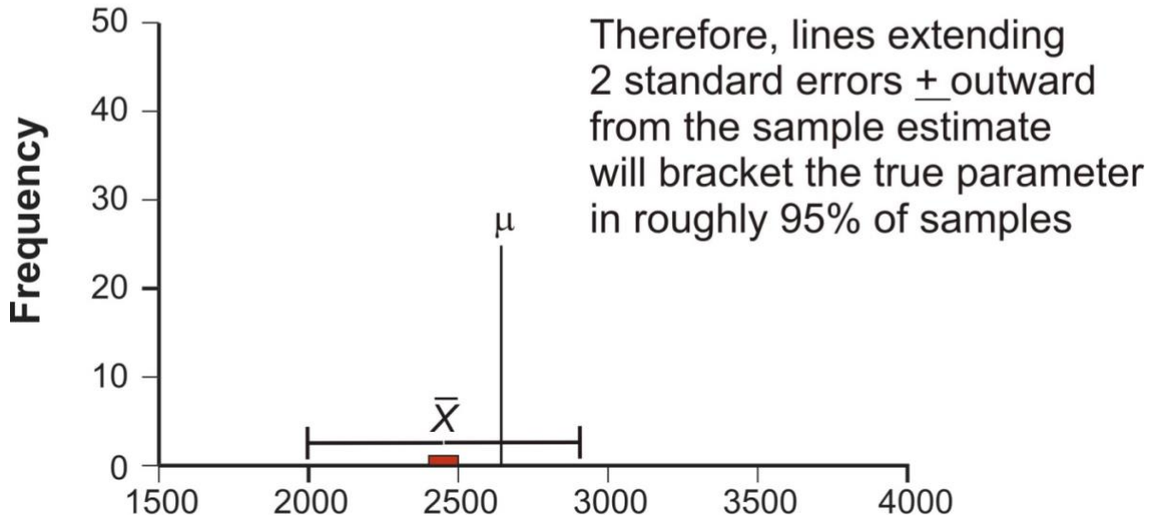


Standard error

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



Twice the SE therefore 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" (SE).

This is an unusual feature of $\bar{X}$. No assumptions about normality are required yet.

However, the assumption of normality *is* required if we use the usual formula for the 95% confidence interval of the mean.

Standard error of the sample mean has a remarkable property

Sadly, many other kinds of estimates do not have this wonderful property. What to do?

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

Method invented by Efron (1979).

The real sampling distribution

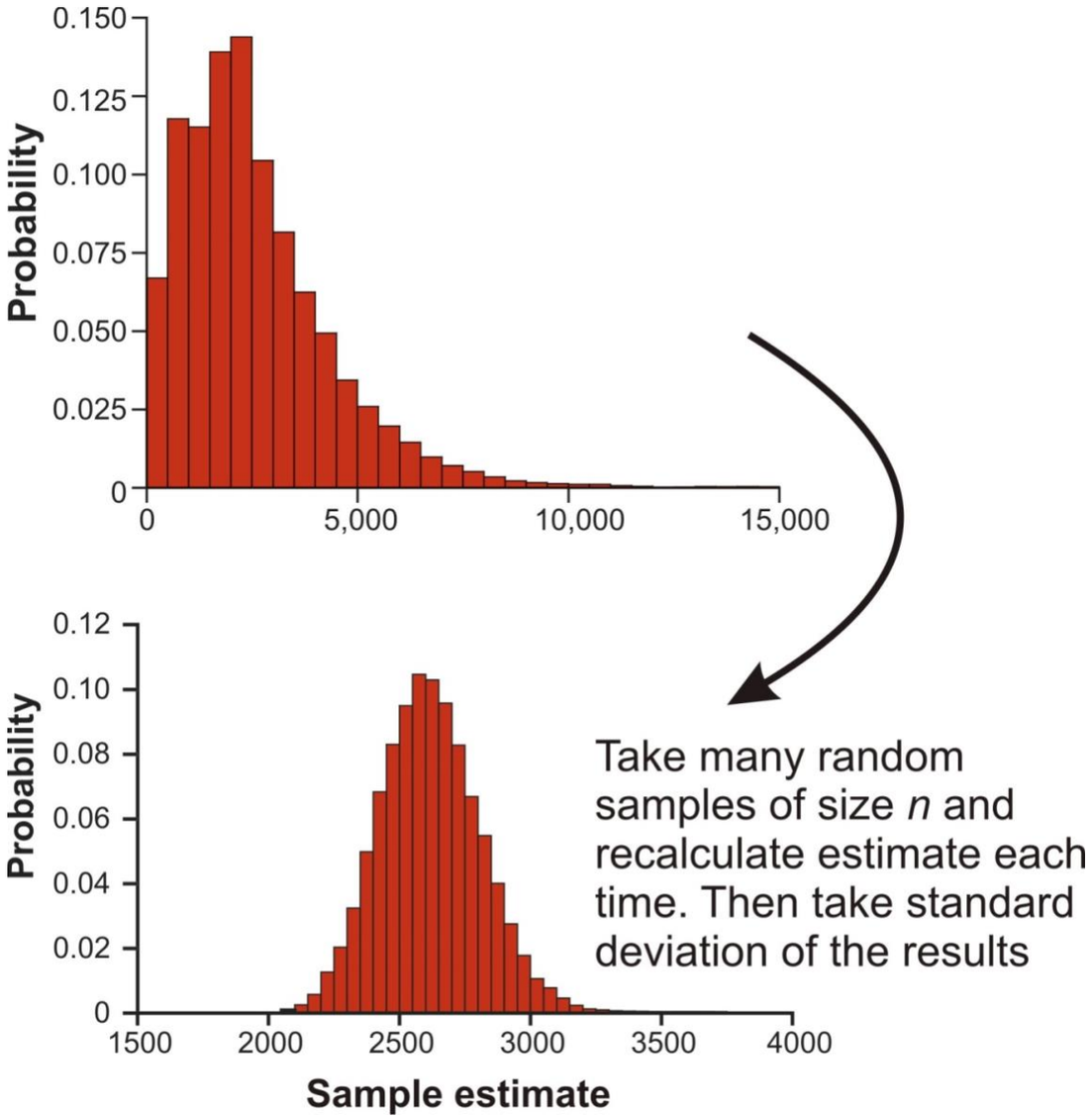
To get real sampling distribution, take many random samples from the same population and estimate the parameter each time.

Then calculate SE as the standard deviation of the resulting sampling distribution

But this is a thought experiment.

In reality we only have <u>one</u> sample, and so only <u>one</u> estimate.

Rats!

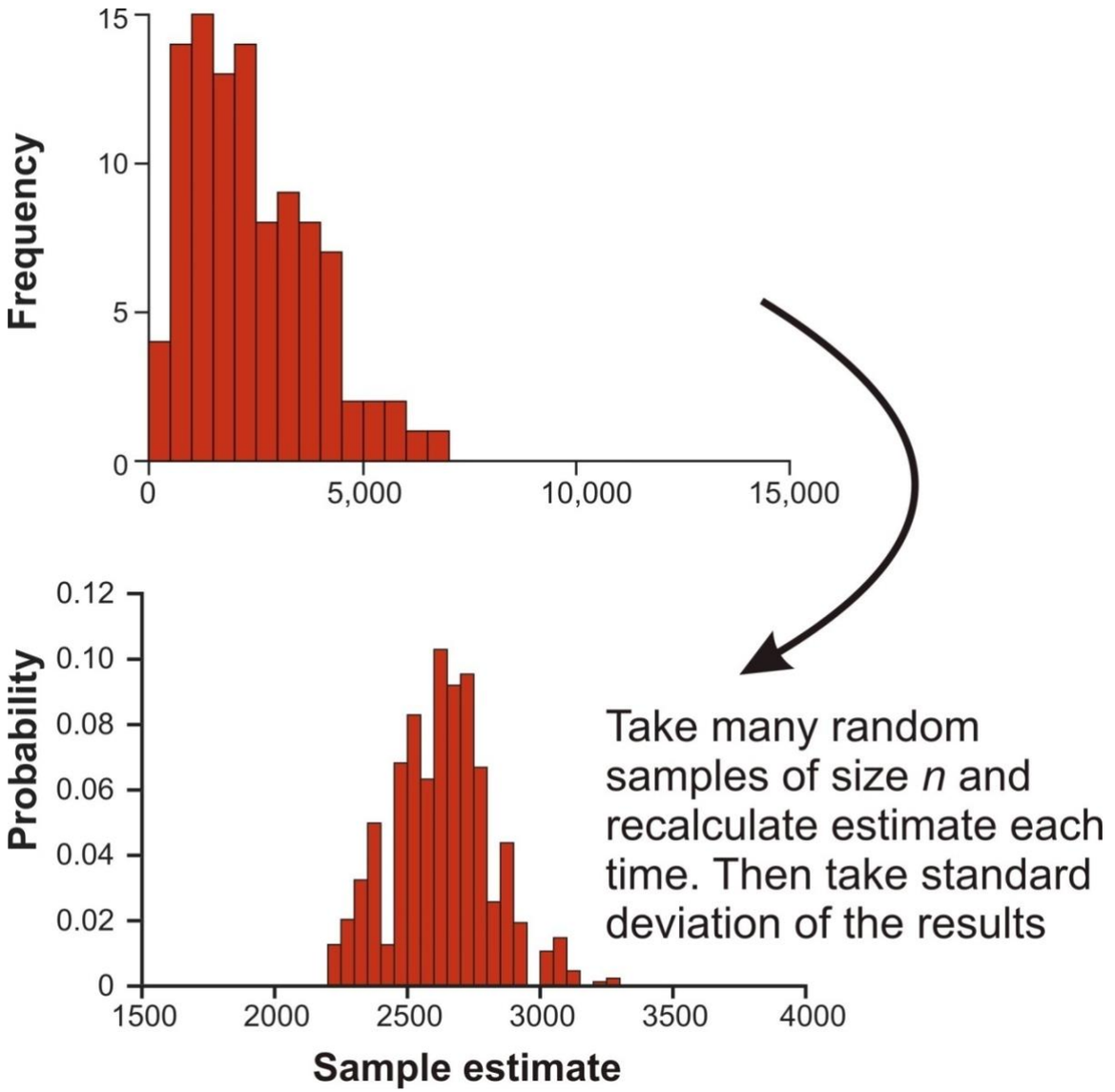


The bootstrap sampling distribution is the next best thing

Pretend the data represent the population. Sample many times from this pretend "population" instead.

Sampling is with replacement so each new bootstrap sample is missing some values from the data and has duplicates of others.

The standard deviation of resulting distribution yields the bootstrap standard error



The bootstrap algorithm

1. Use the computer to take a random sample of *n* individuals from the original data. The bootstrap sample should contain the same number of individuals as the original data: *n*. Each time an observation is chosen, it is left available in the data set to be sampled again ("sampling with replacement").
2. Calculate the statistic (estimate) of interest using the measurements in the bootstrap sample from step 1. This is the first **bootstrap replicate estimate**.
3. Repeat steps 1 and 2 many times (10,000). The frequency distribution of all bootstrap replicate estimates yields an approximation of 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**.

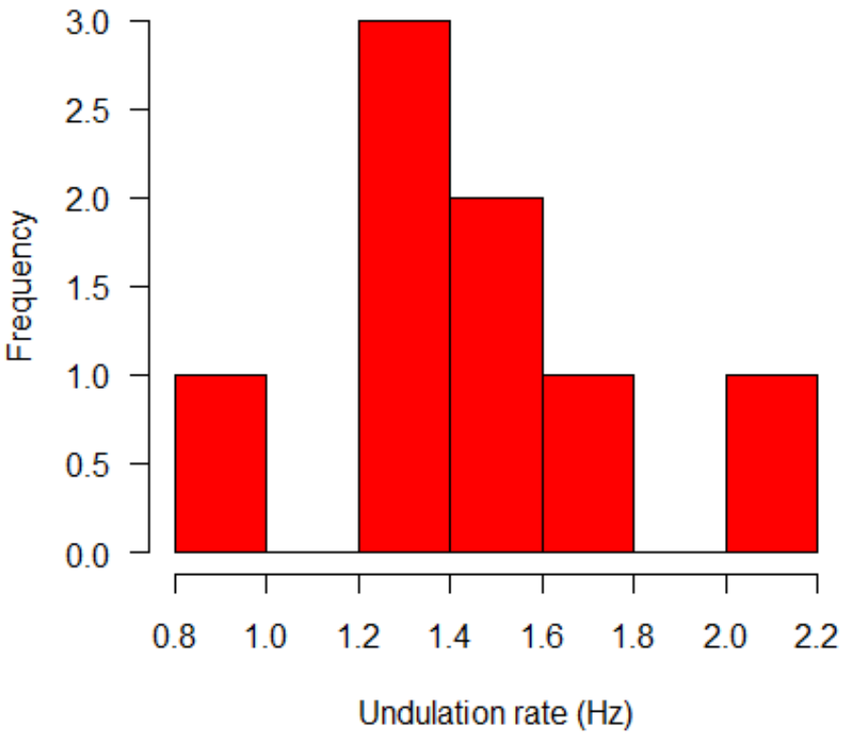
Bootstrap example: sample mean

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

$\bar{X}$ = 1.375



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

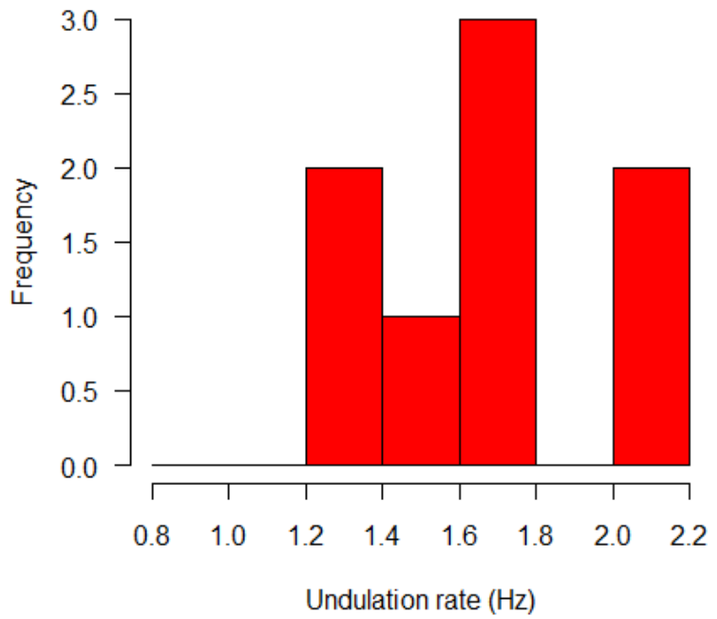
Bootstrap example: sample mean

```
hertz <- 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(hertz, replace = TRUE)
print(xboot)
[1] 2.0, 1.3, 1.6, 1.2, 1.6, 1.4, 1.6, 2.0
```

Histogram of the first bootstrap sample:



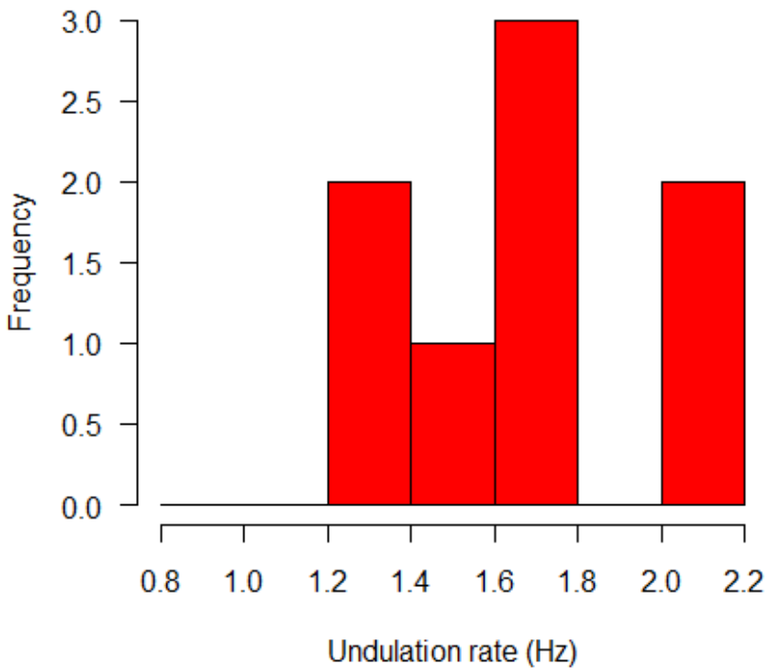
Bootstrap example: sample mean

2. Calculate the statistic (estimate) using the measurements in the bootstrap sample from step 1. This is the first bootstrap replicate estimate.

```
mean(xboot)
1.5875
```

Save the result from the first bootstrap sample:

```
z <- vector() # initialize
z[1] <- mean(xboot) # save
```



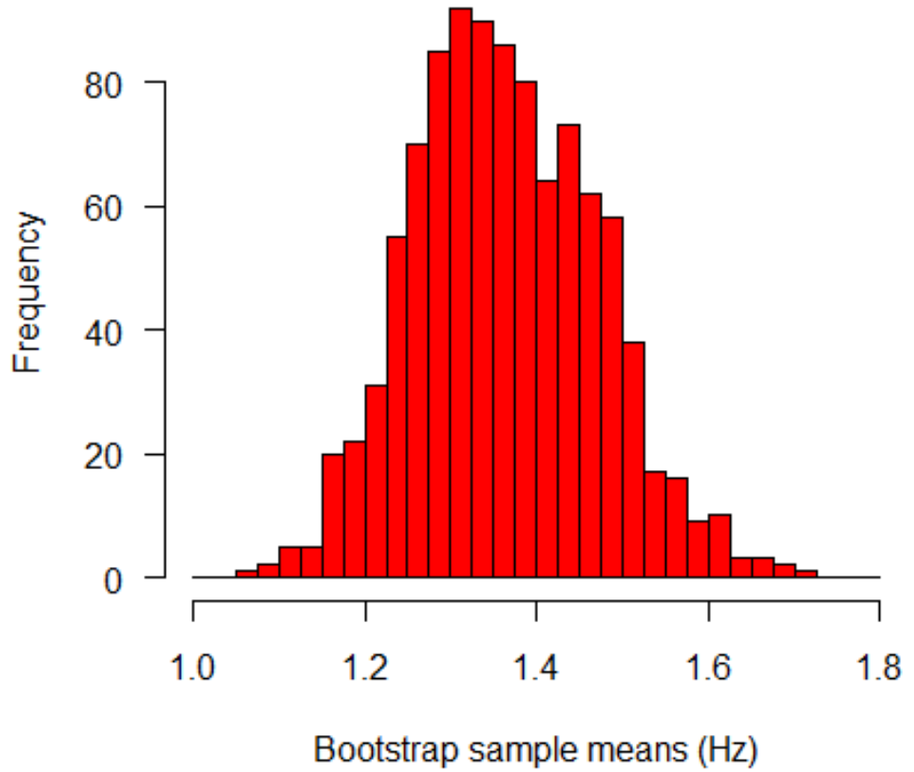
Bootstrap example: sample mean

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

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

Better idea: create a loop in R to accomplish the repeats.

Plot of the bootstrap sampling distribution:



Bootstrap example: sample mean

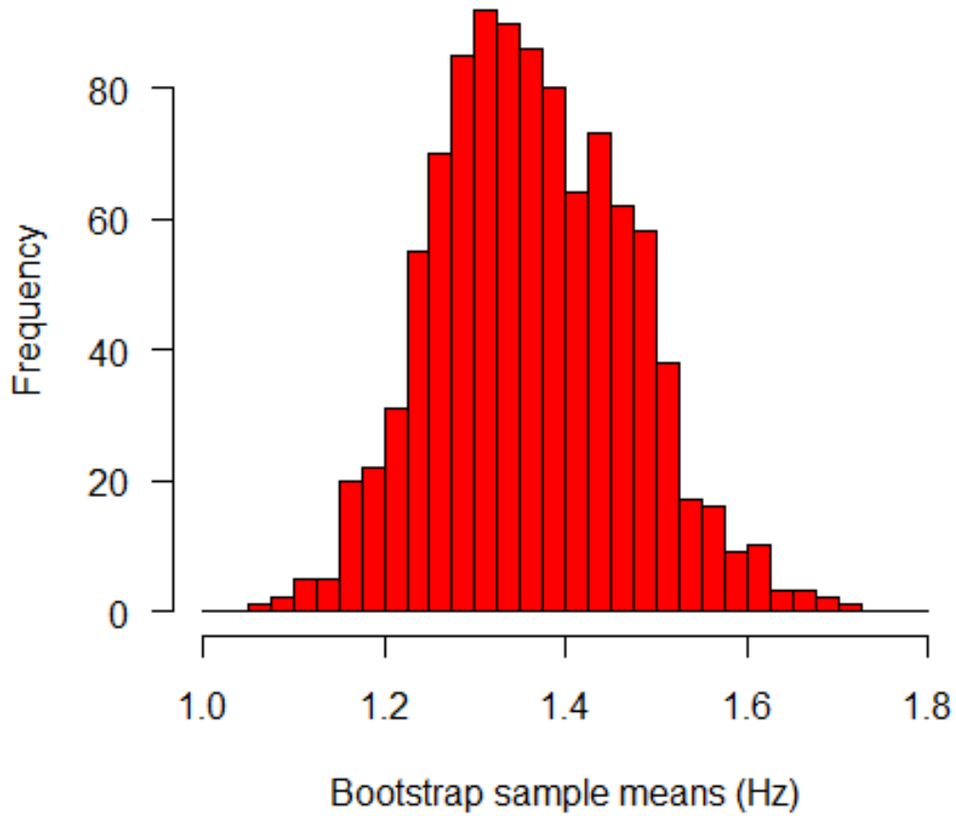
4. The bootstrap standard error is the standard deviation of all the bootstrap replicate estimates obtained in step 3.

```
sd(z)
0.1070
```

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

```
sd(hertz)/sqrt(length(hertz))
0.1146
```

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



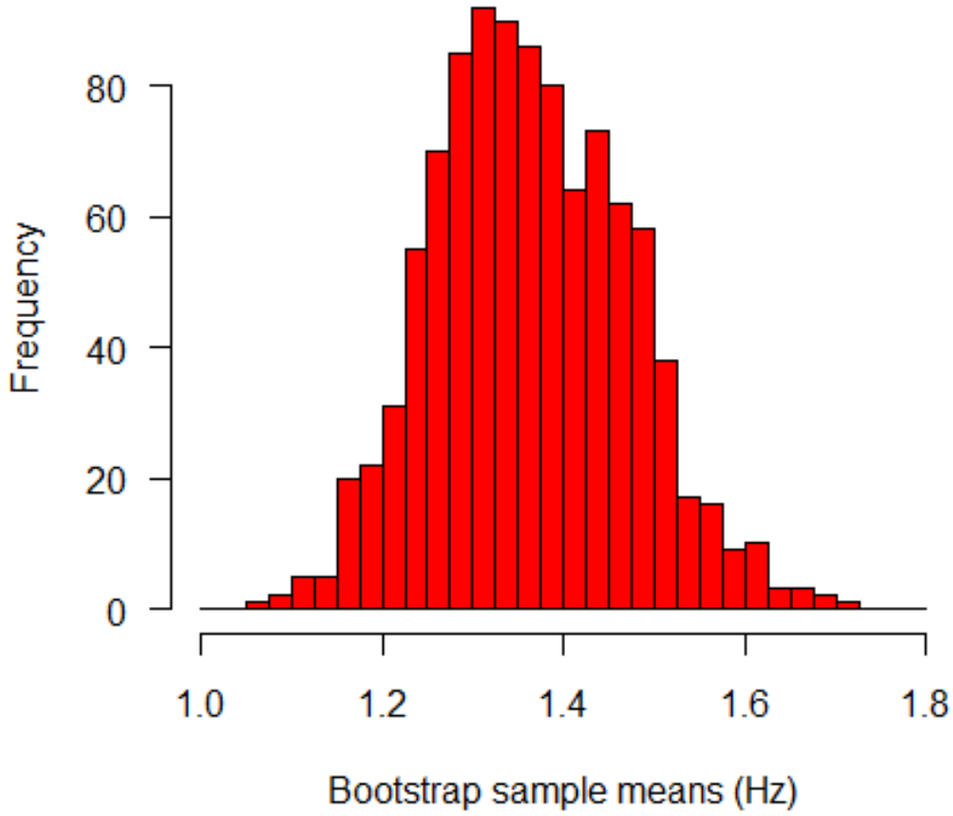
The bootstrap can also be used to calculate a confidence interval

Incredibly, the 2.5th and 97.5th percentiles of the bootstrap sampling distribution are an approximate 95% confidence interval. No transformations or normality assumptions needed.

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

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

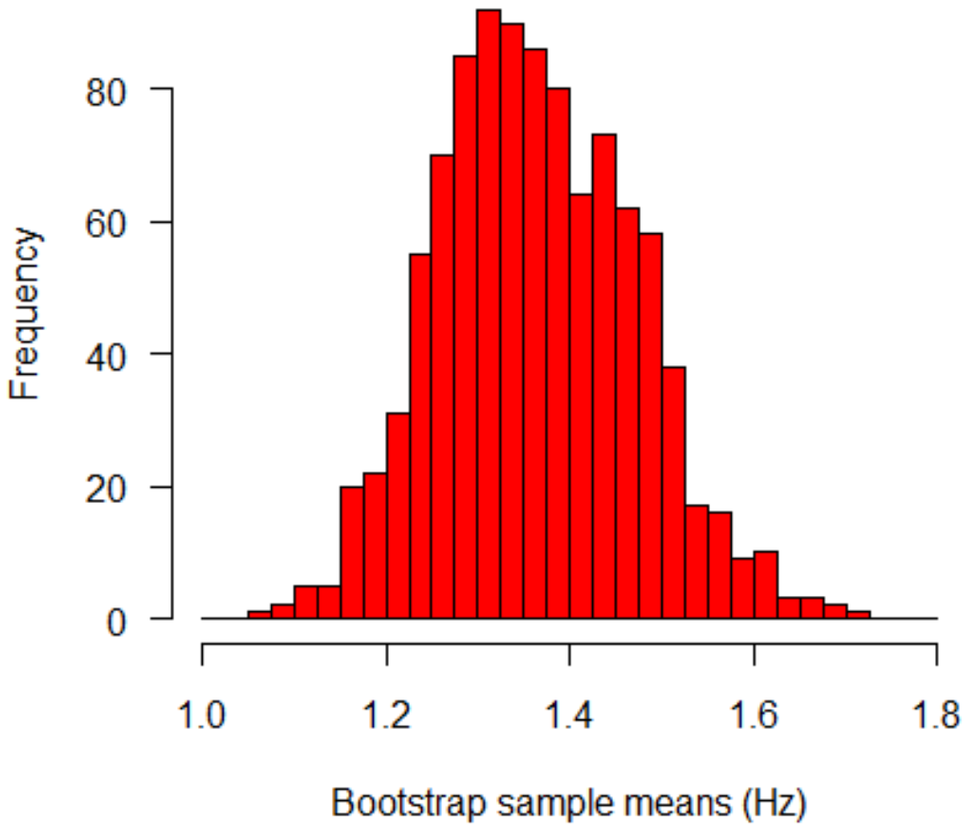
```
95 percent confidence interval:
 1.104 1.646
```



Bootstrap confidence intervals

This “percentile” method of obtaining bootstrap confidence intervals works well if the sampling distribution is symmetric and unbiased.

Improved, bias-corrected and accelerated (BCa) confidence intervals improve accuracy by correcting for bias and skew in the bootstrap sampling distribution.



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 (with replacement, same sample sizes) from each group.
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

5th instar *Manduca sexta* caterpillars were trained to associate a mild electrical shock with a specific odor (ethyl acetate; EA). Then they were 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"
(recall: log odds is how we modeled a proportion with `glm()`)

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: (thirty two 1’s and nine 0’s)
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

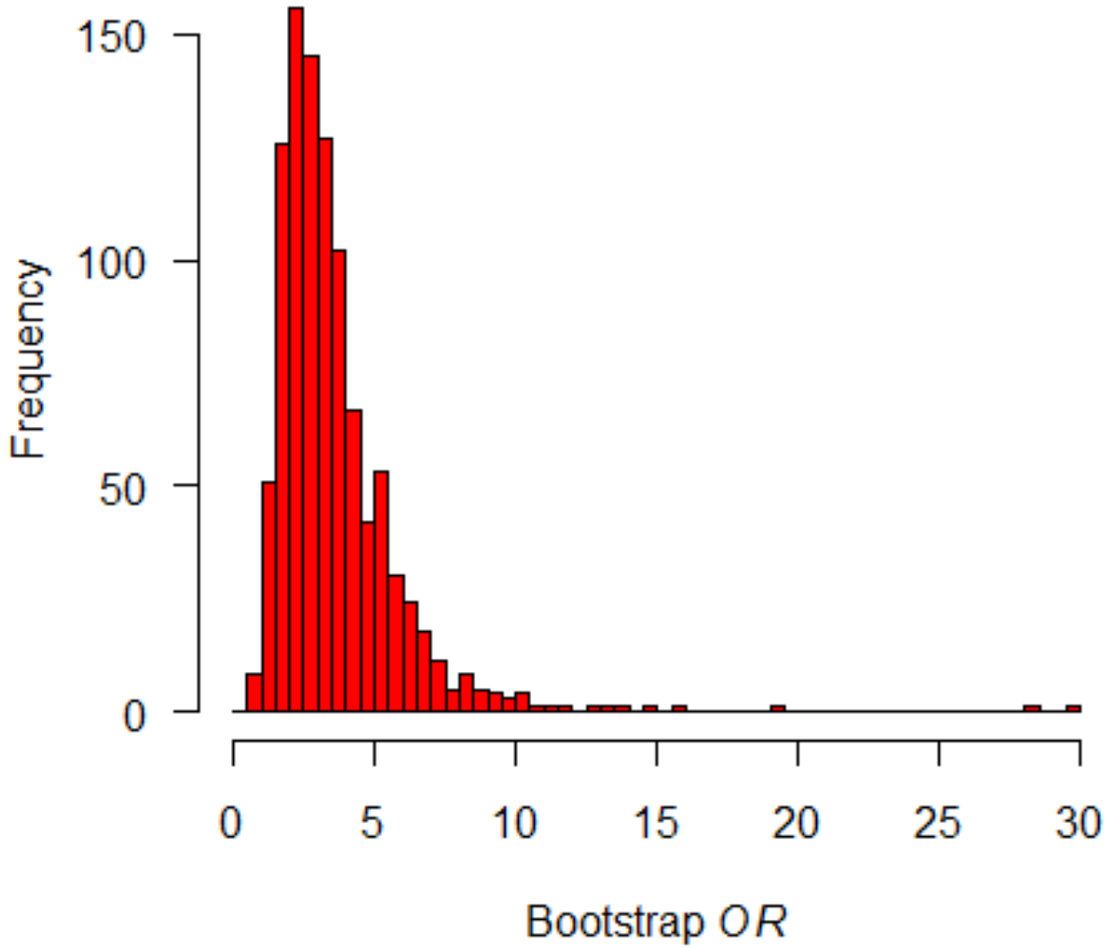
Control group: (twenty five 1’s and twenty one 0’s)
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Bootstrap procedure:
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

Bootstrap 95% CI using the percentile method:

2.5% 97.5%

1.21 8.67

Bootstrap BC_a (bias corrected and accelerated)

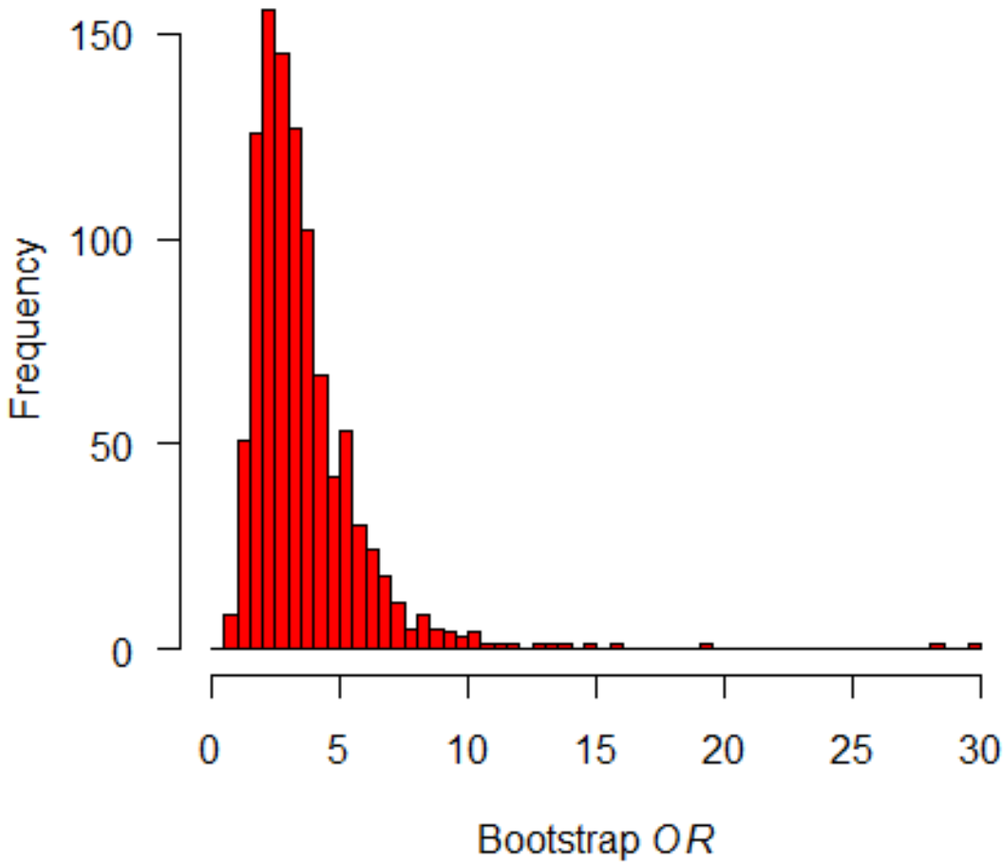
2.5% 97.5%

1.14 7.93

Compare with conventional approximate CI for odds ratio using formula

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 (if *n* not too small).
- It is approximate, though performs almost as well as parametric methods when assumptions of the parametric methods are met.
- It can also be used for hypothesis testing, though I have not discussed this.
- Permutation tests are useful for obtaining *P*-value, but that's all.
- Use the bootstrap to estimate magnitudes.

Third assignment: Model selection

See the **Homework** tab at the course web site.

Due Friday, April 12, 2024.

Discussion paper for next week:

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

Download from **Handouts** tab on course web site.

Presenters: Clare & Lauren

Moderators: Heather & _____