BISC-869, Bootstrap

March 18, 2026

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

1. **Estimation**
   Sampling distribution of an estimate. 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 has an approximately normal distribution.

2. **Hypothesis testing**
   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.

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

A: **Computationally-intensive methods**. An approach in which the power of the computer is used to generate a sampling distribution.

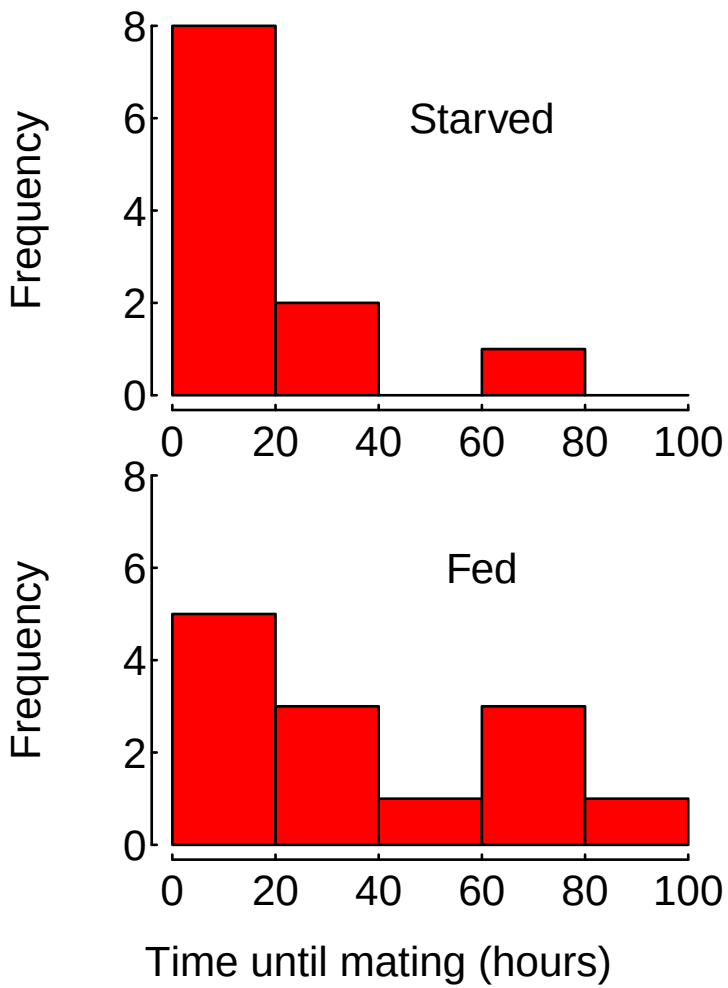
1. Estimation: The **bootstrap**.
2. Hypothesis testing: The **permutation test**.

A **permutation test** generates a null distribution for a statistic measuring the association between two variables (or difference between 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 U is known.

But there's no need to replace the data with the ranks. Permute the data themselves. No known probability distribution is available, so we can use a computer to generate a large number of permutations instead.

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:



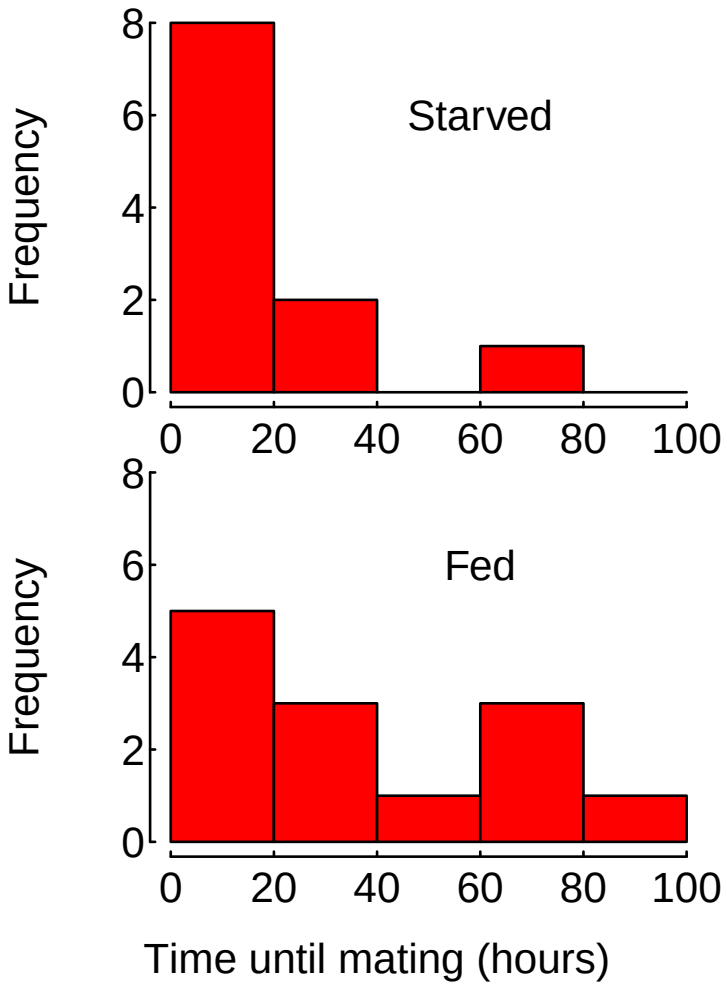
One group of 11 females was starved for at least two days. Another group of 13 females was fed during the same period.

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 distributed.

Data:

Treatment	Time	Treatment	Time
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

Test statistic:
$\bar{Y}_1 - \bar{Y}_2 = 17.73 - 35.98 = -18.26$.
H_0: Mean time to mating $\mu_1 = \mu_2$.
H_A: Mean time to mating $\mu_1 \neq \mu_2$.



Data:

Treatment	Time	Treatment	Time
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

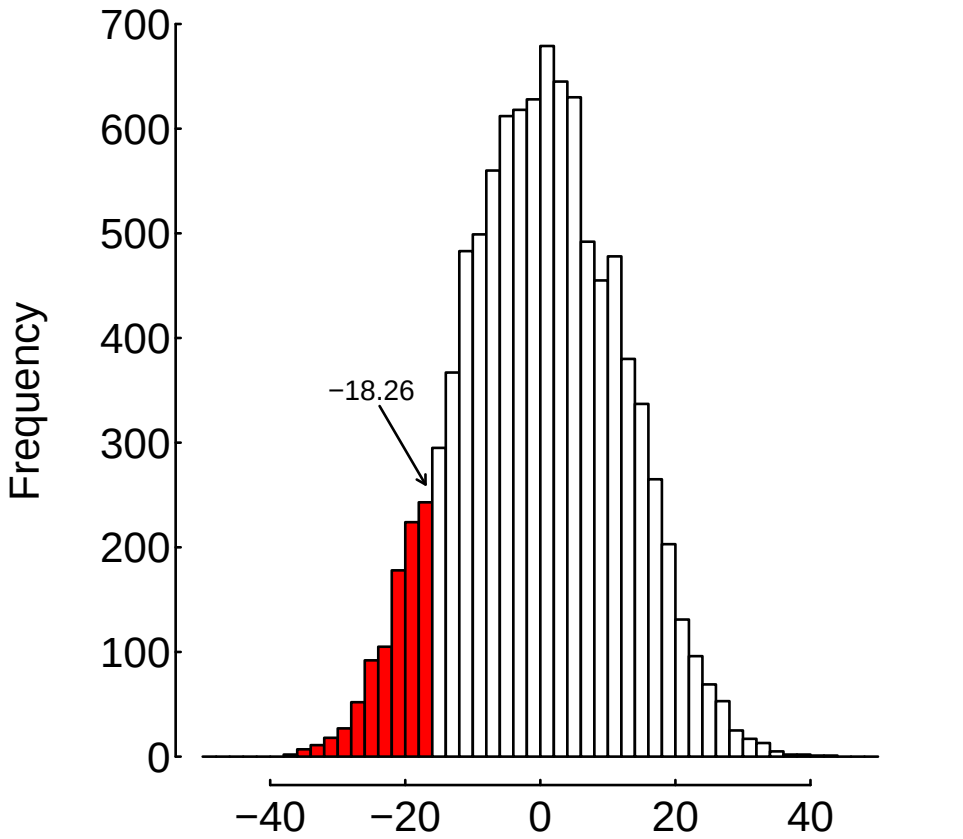
Test statistic:
$\bar{Y}_1 - \bar{Y}_2 = 17.73 - 35.98 = -18.26$.
H_0: Mean time to mating $\mu_1 = \mu_2$.
H_A: Mean time to mating $\mu_1 \neq \mu_2$.

A single permutation

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

Test statistic:
$\bar{Y}_1 - \bar{Y}_2 = 27.75 - 27.5 = 0.25$.

With $10,000$ permutations, we can create the null distribution of $\bar{Y}_1 - \bar{Y}_2$.



703/10,000 had a value $\leq$ the observed value, -18.26.

P-value$= 2 \times 0.0703 = 0.1406$.

- 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 sizes are large.

Parametric methods provide estimates (with standard error or confidence interval) of a useful parameter.

Nonparametric tests, including permutations tests and rank tests, provide only a P-value. They do not provide estimates (with standard error or confidence interval) of a useful parameter.

Nonparametric tests, including permutations tests and rank tests, perpetuate the view that the P-value is all you want from the data, and that the smallness of the P-value is an indication of the magnitude or importance of an effect.

As our readings and discussions have stressed, the P-value, in fact, tells us nothing about magnitudes of effects or biological importance. No decision should ever be made on the basis of a P-value alone.

Primarily used for estimation.

Provides standard errors and confidence intervals of useful parameters.

The method is nonparametric, so doesn't require normally-distributed data, or data having any other particular distribution.

It can be applied to virtually any 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 for estimates based on complicated sampling procedures or calculations (for example, phylogeny estimation, as we saw in today's reading).

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

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

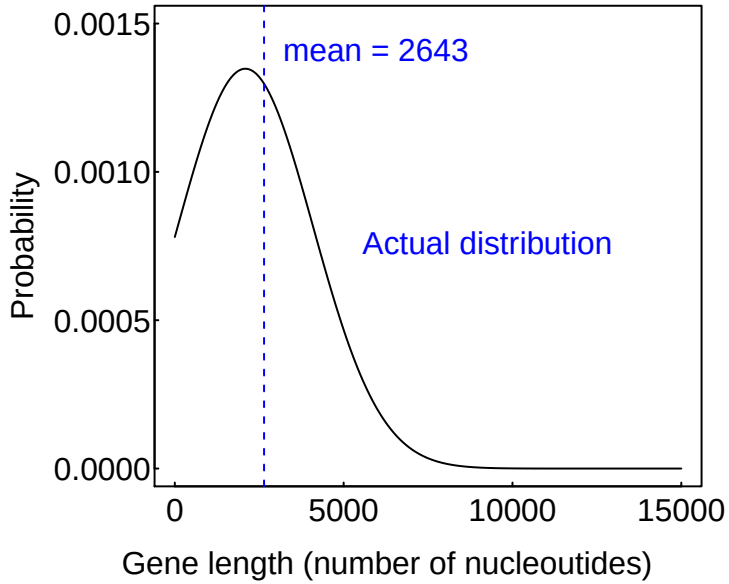
The value of a sample *estimate* is almost never the same as the *parameter* in the population because of random sampling error (chance).

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

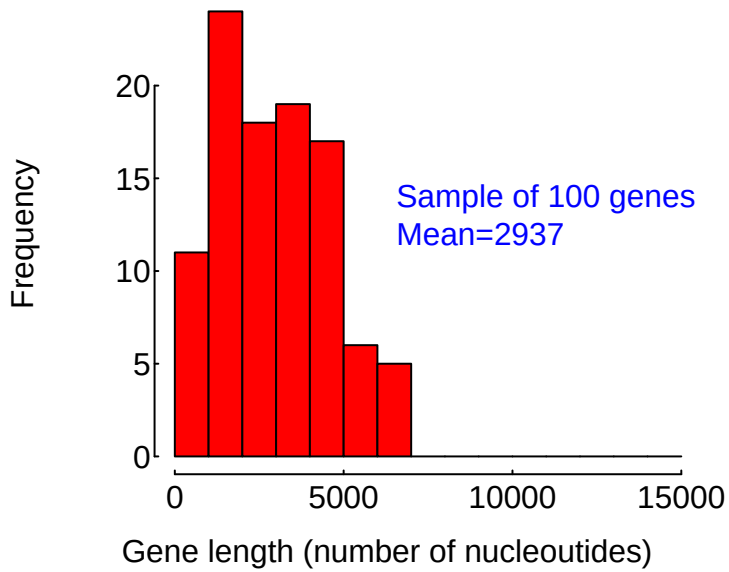
The **standard error** of an estimate is the standard deviation of its sampling distribution. No estimate is useful without it.

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

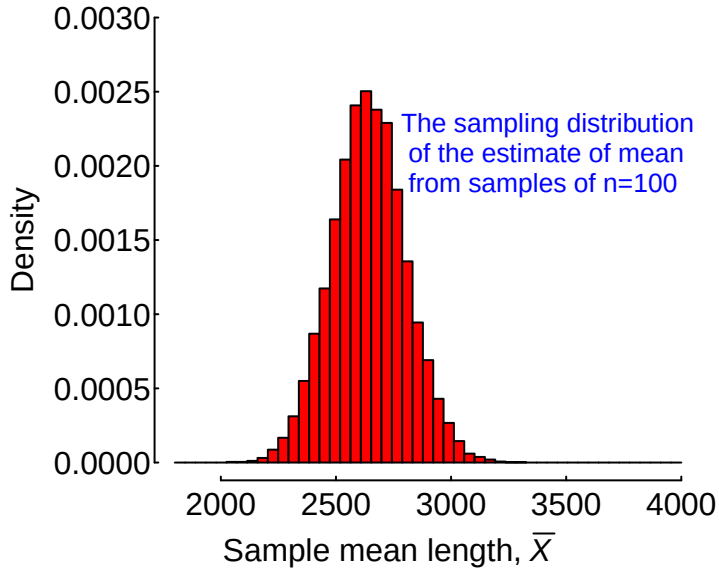
Note: simulated data.



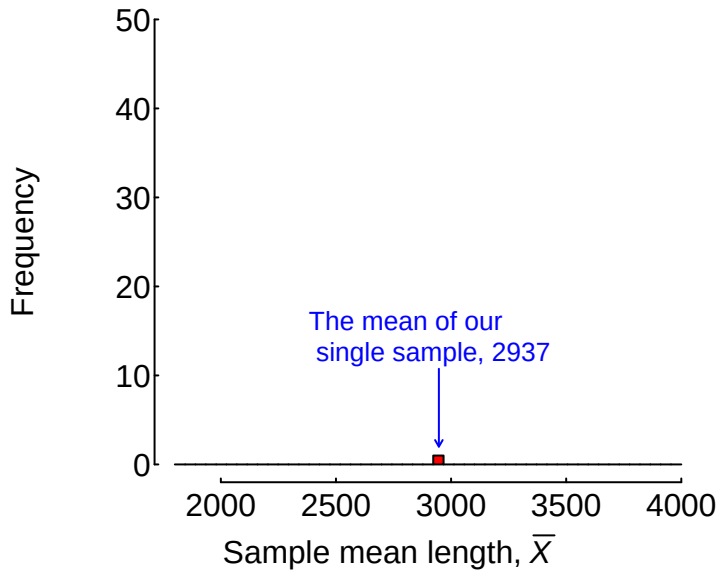
What we have instead:
The *sample* mean (e.g., based on a random sample of $n = 100$ genes).



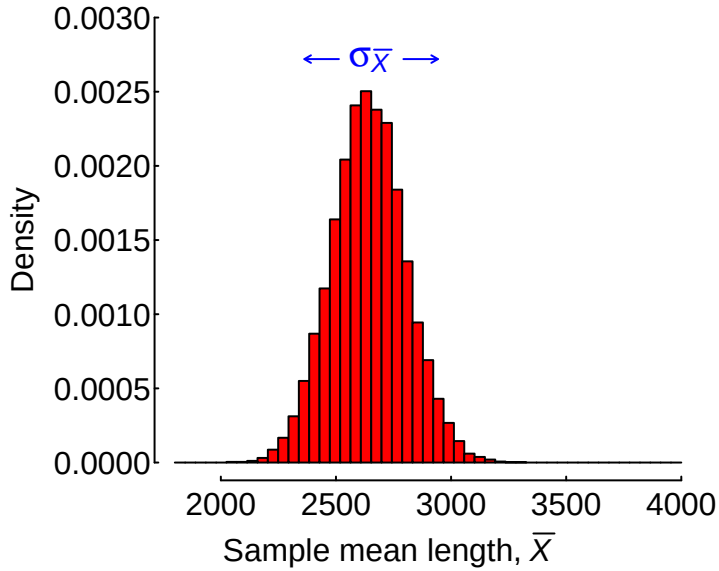
The sampling distribution:
Since we don't have the true mean, we need an approximation of the sampling distribution, giving all possible values of the estimate and their probabilities.



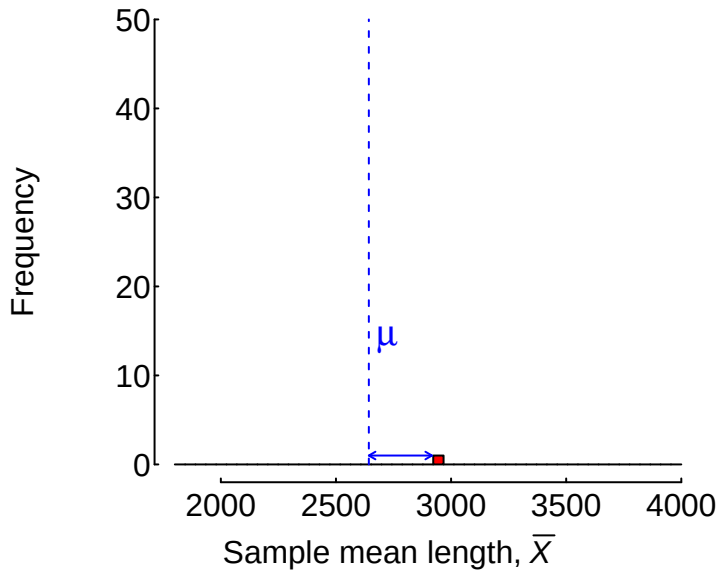
What we have:
Just one sample mean



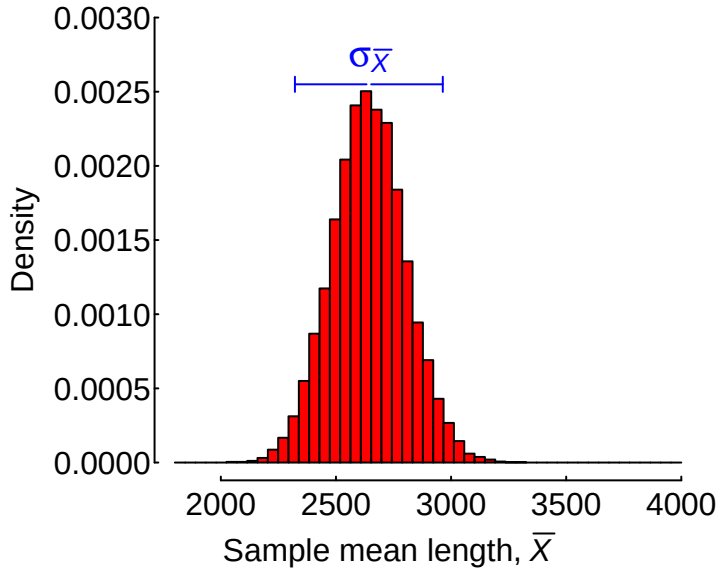
Standard error
The standard deviation of the sampling distribution (the *standard error*) measures the variation of sample estimates around the population parameter.



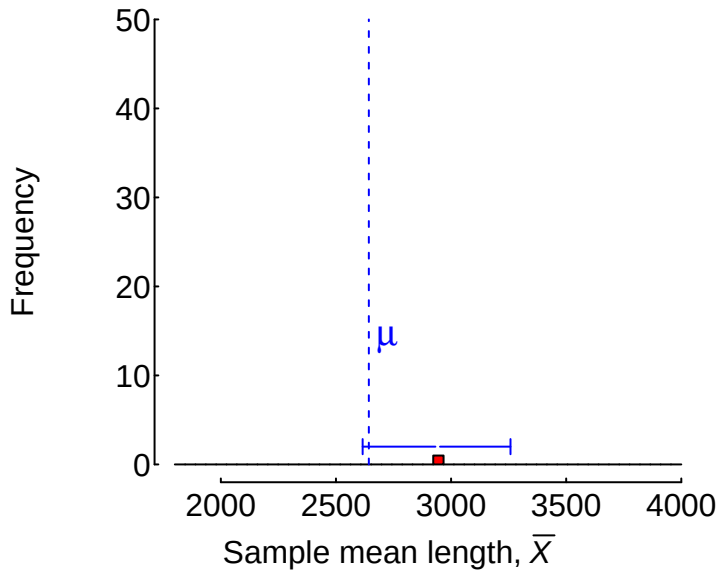
Roughly, the standard error tells us how far we are from the truth, on average.



Standard error
If the sampling distribution is roughly bell-shaped, then about 95% of estimates fall within 2 SE 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!

$$\sigma_{\bar{X}} \sim S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

$S_{\bar{X}}$ 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. The assumption of normality is nevertheless required for the 95% confidence interval.

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

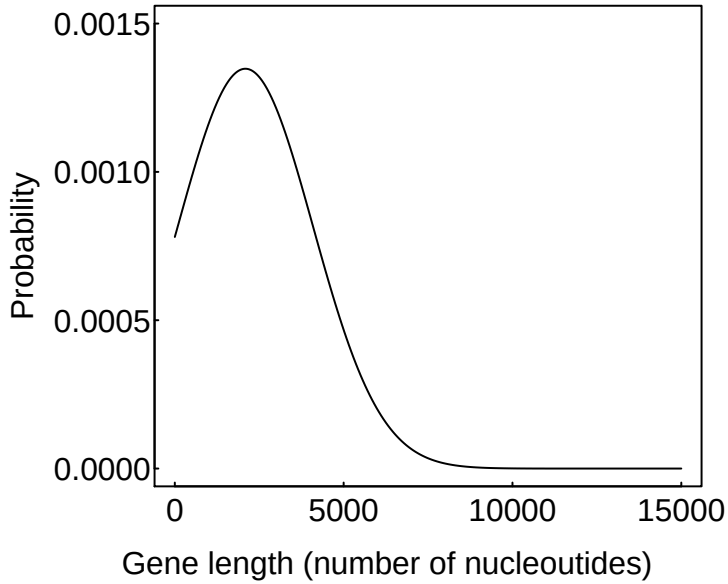
One answer: make your own sampling distribution for the estimate using the "bootstrap". Method invented by Efron (1979).

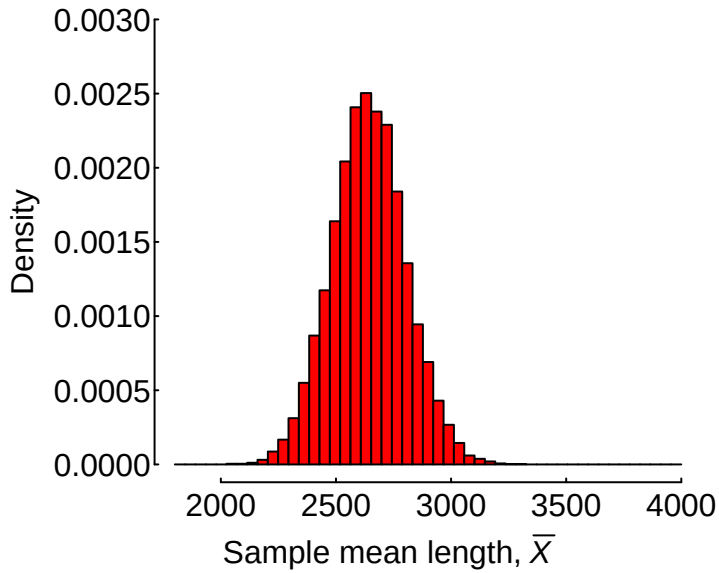
The real sampling distribution

To get the real sampling distribution, sample many times (each sample of size n), from the same population.

Calculate SE as the standard deviation of the resulting sampling distribution.

But we only have *one* sample, and so only *one* estimate!



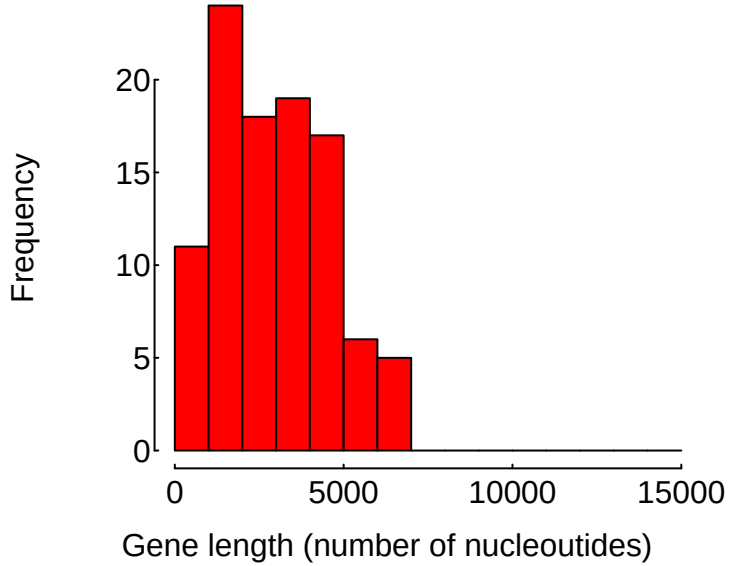


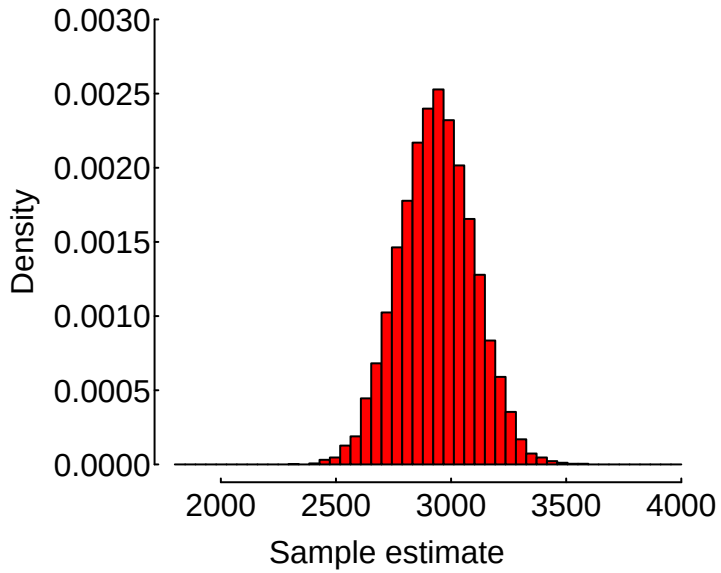
The bootstrap sampling distribution is the next best thing

Pretend the data represent the population! Sample many times from the single sample instead (each sample of size n).

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 results yields the **bootstrap standard error**.



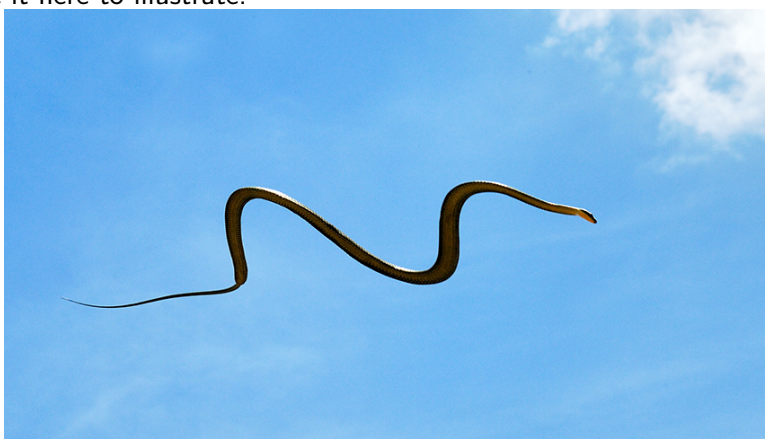


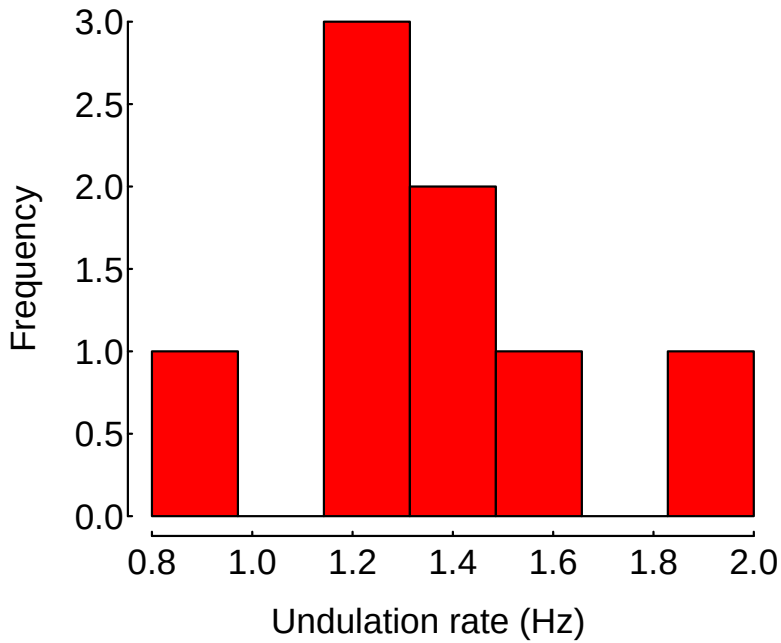
1. Use the computer to take a random sample of individuals from the original data. The bootstrap sample should contain the same number of individuals as the original data. Each time an observation is chosen, it is left available in the data set to be sampled again ("sampling with replacement").
2. Calculate the estimate of interest using the values in the bootstrap sample from step 1. This is the first **bootstrap replicate estimate**.
3. Repeat steps 1 and 2 many times (e.g., 10^4). 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**.

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 we will use it here to illustrate.



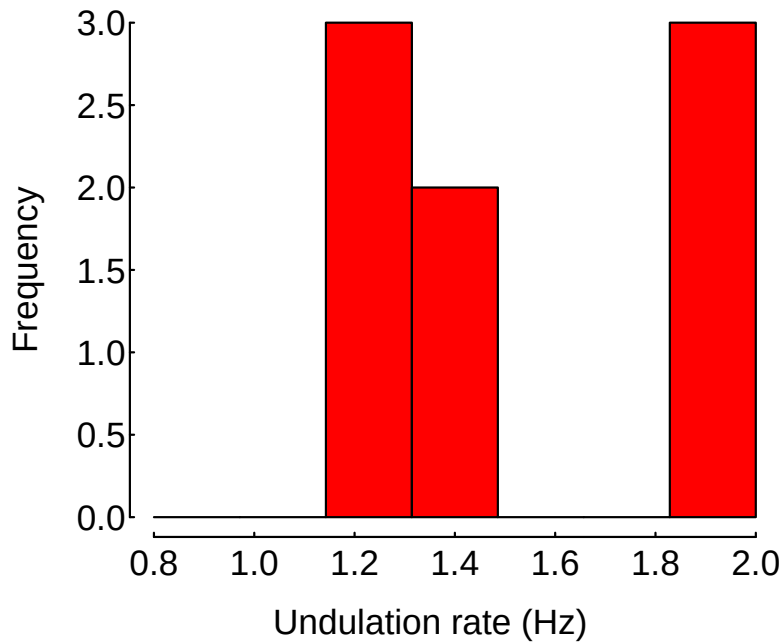


1. Use the computer to take a random sample of individuals from the original data.

```
hertz <- c(0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0)
xboot <- sample(hertz, replace=TRUE)
xboot
```

```
[1] 1.2 1.2 1.4 2.0 1.2 2.0 2.0 1.4
```

Histogram of first bootstrap sample:



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

```
mean(xboot)
1.55
```

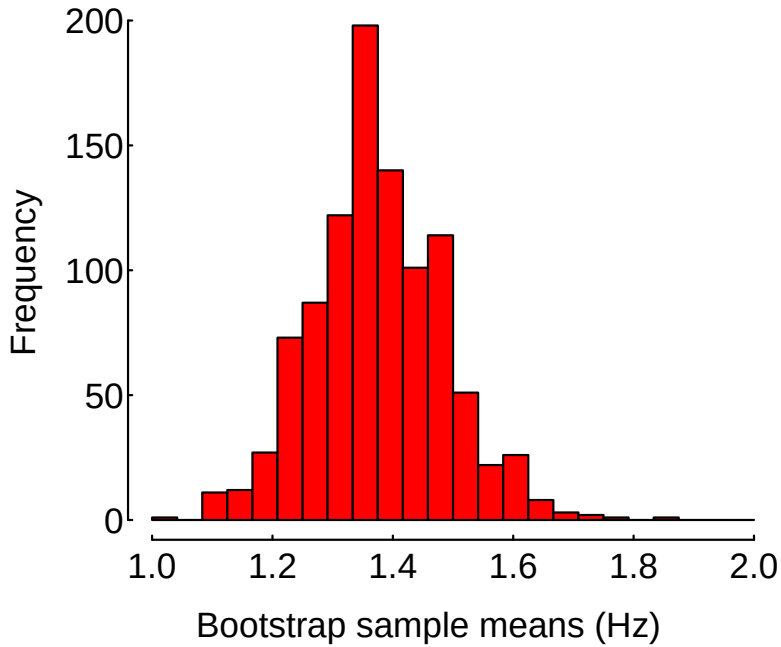
Save the result from the first bootstrap replicate:

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

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

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

Plot bootstrap sampling distribution:



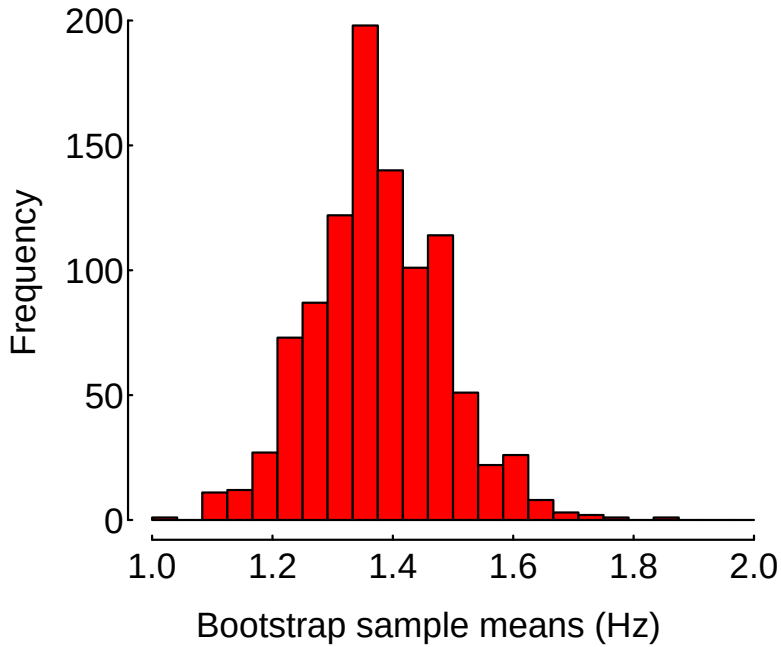
4. The bootstrap standard error is the *standard deviation* of all the bootstrap replicate estimates obtained in step 3.

```
sd(z)
```
```
0.1083
```

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

```
quantile(z, probs=c(0.025,0.975))
    2.5%    97.5%
1.175000 1.600312
```

Compare with results from using the t-distribution:

```
quantiles <- qt(c(0.025,0.975),
                length(hertz)-1)
-2.364624 2.364624
mean(hertz) + quantiles*se
```

1.104098 1.645902 where

```
se = sd(hertz)/sqrt(length(hertz))
```

as calculated on previous slide.

Pretty close!

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

Improved, bias-corrected and accelerated (BC_a) confidence intervals improve accuracy when sampling distributions are skewed and/or biased (we will see an example in the workshop).

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*.

5^{th} 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

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:

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

For the caterpillar data,

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

learned:
$p_1 = 32/41 = 0.78$
$O_1 = 0.78/0.22 = 3.56$

control:
$p_2 = 25/46 = 0.54$
$O_2 = 0.54/0.46 = 1.19$

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

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.

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

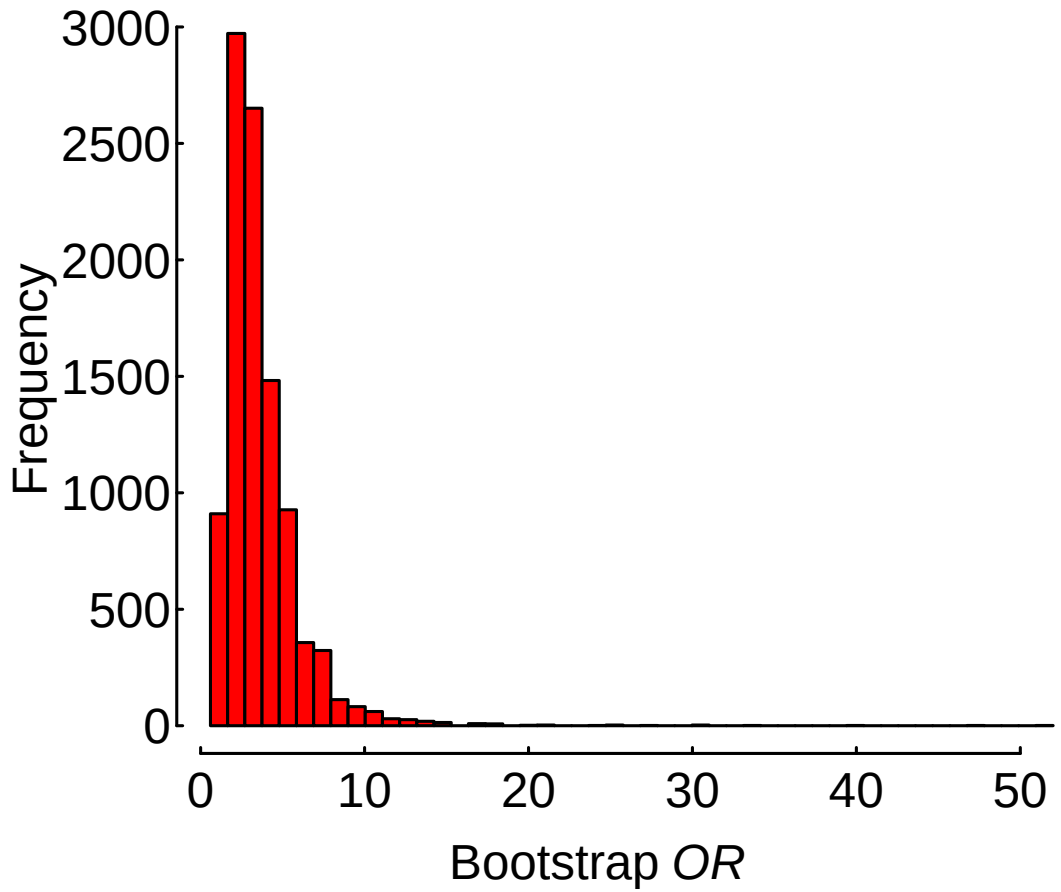
Let 1 indicate "chose clean air" and 0 indicate "chose EA air."

Learn group: 11111111111111111111111111111000000000
(thirty-two 1's and nine 0's)

Control group: 1111111111111111111111100000000000000000000
(twenty-five 1's and twenty-one 0's)

1. Take a random sample (with replacement) from each group
2. Calculate *OR*
3. Repeat 10^4 times

Bootstrap sampling distribution for *OR*:



Bootstrap SE $= 2.32$

Bootstrap 95% CI using percentiles:

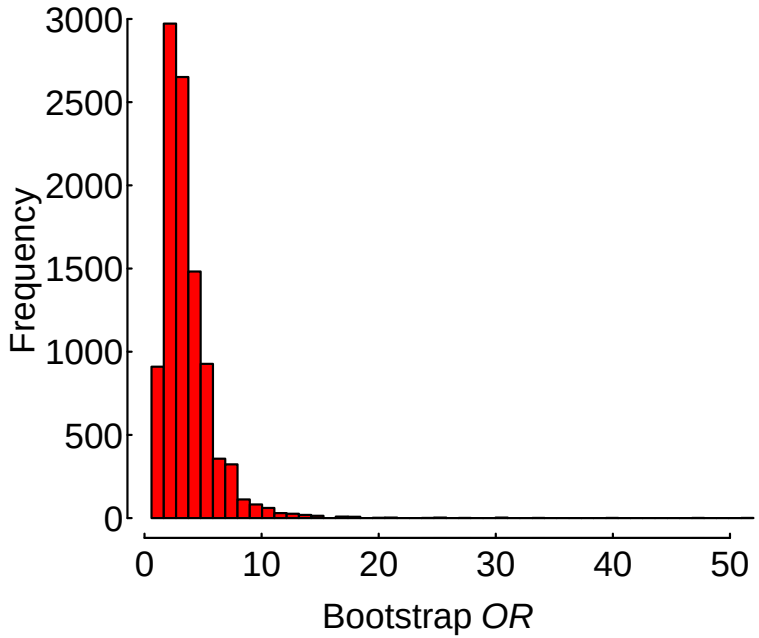
2.5%	97.5%
1.239796	9.250000

Compare with conventional approximate CI for *OR*

2.5%	97.5%
1.17	7.65

BC_a (bias corrected and accelerated)

2.5%	97.5%
1.14	7.93



BC_a corrects the percentiles for skewness in the sampling distribution, which otherwise changes the estimate; and for bias in the estimate.

The bootstrap is amazing and useful for estimation.

It works in almost any situation (be cautious when n is 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.

Permutation tests are useful for obtaining P-value but use the bootstrap to estimate magnitudes.