Bootstrap

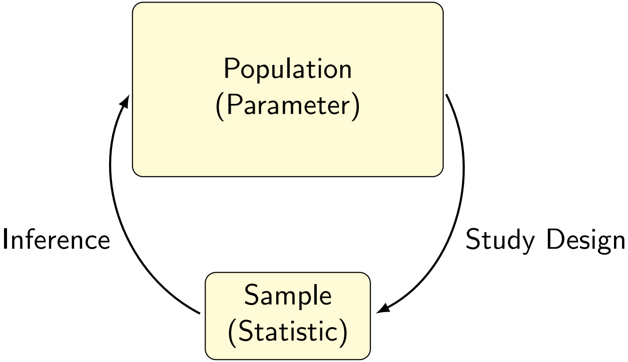
Grinnell College

November 3, 2025

Review

- **Standard deviation** (σ) and **standard error** ($\sigma/\sqrt{n}$)
- A **sampling distribution**
- Point Estimate $\pm$ Margin of Error (critical values)
- A **confidence interval** is an interval with the properties that:
  - It is constructed according to a procedure or set of rules
  - It is intended to give plausible range of values for a *parameter* based on a *statistic*
  - It has no probability; the interval either contains the true value or it does not

The Statistical Framework



Repeated Samples

The confidence intervals we constructed of the form:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

- Relied on assumptions (TBD) about our *sampling process*
- Examined what might happen if we could repeat sampling ad infinitum

There are, naturally, some limitations:

- We are limited to collecting a single sample
- Our assumptions may be tenuous, i.e., what if our statistic doesn't not follow the CLT?

It would be helpful to have a more general method of constructing intervals with similar properties we had before

Bootstrapping

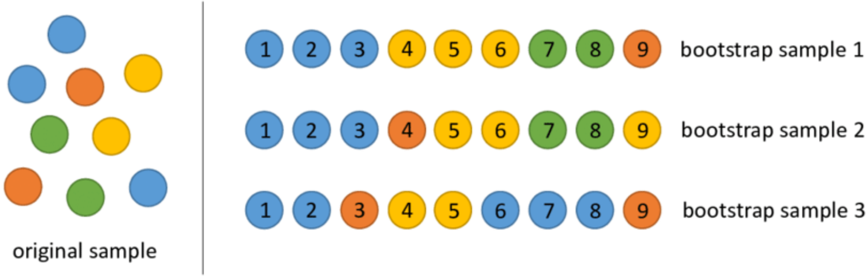
Somewhat amazingly, we can get around this problem with a resampling technique known as **bootstrapping**

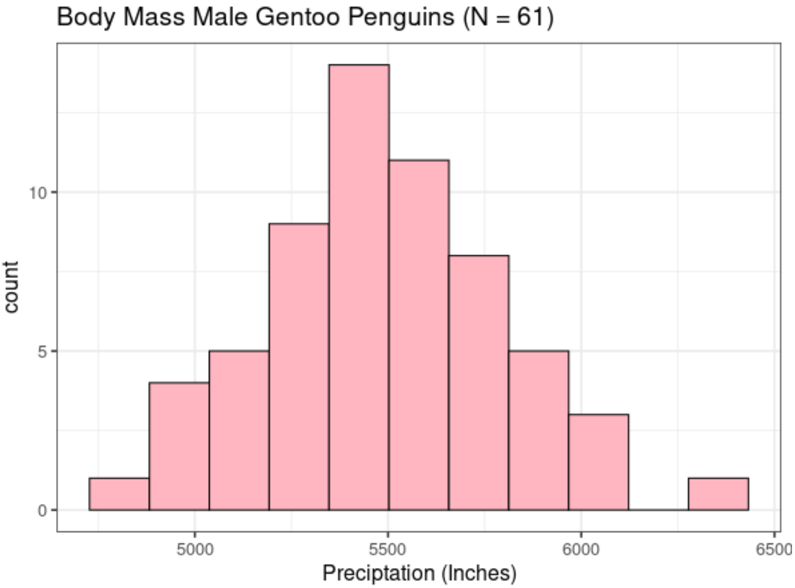
Bootstrapping refers to an algorithmic process whereby, for a sample of size n, we resample *with replacement* and compute a new statistic on the bootstrapped sample.

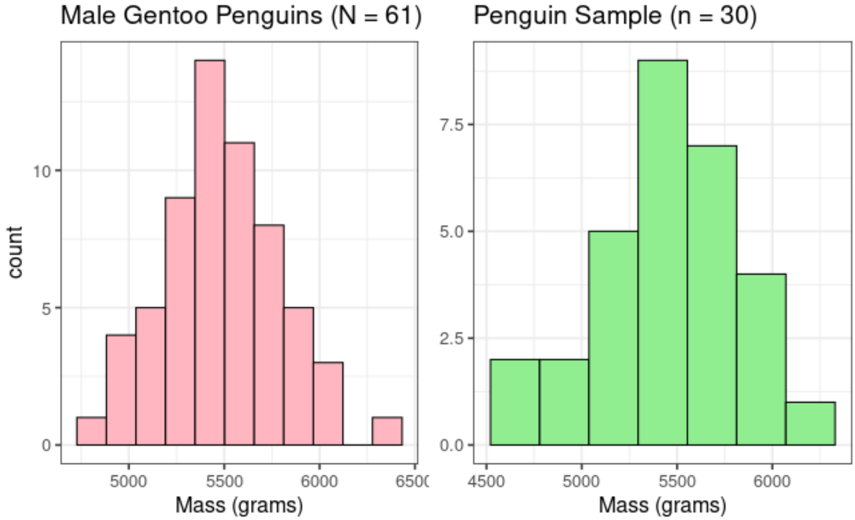
Instead of drawing more samples from our *original population*, we treat our sample as an *estimate* of the population and instead draw bootstrapped samples from our original sample

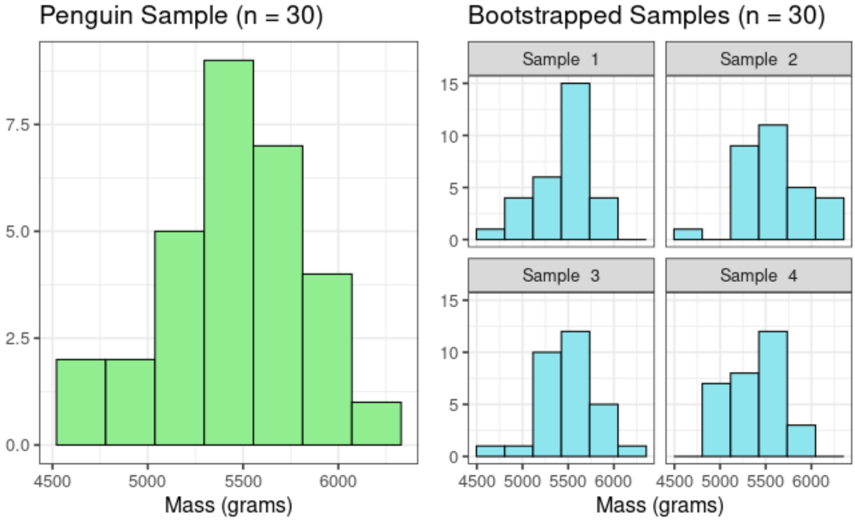
Bootstrapping

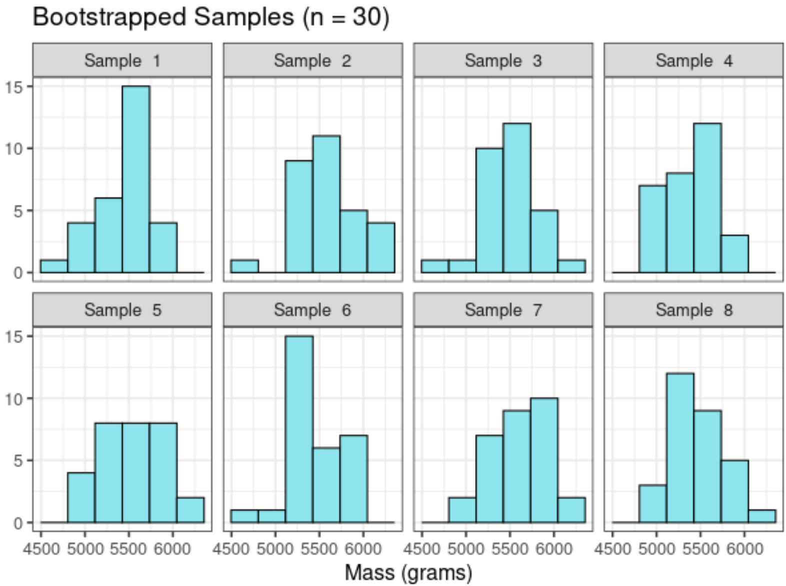
"Pick yourself up from bootstraps"



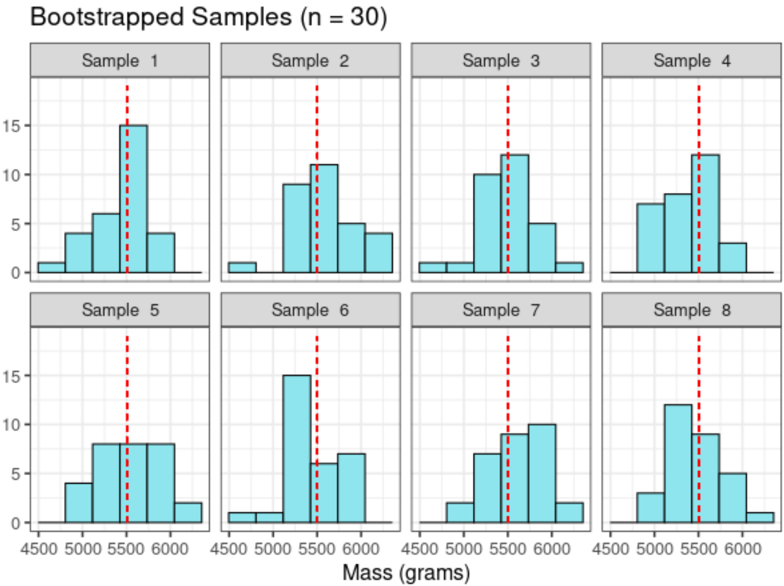

Body Mass Male Gentoo Penguins (N = 61)
count
0
5
10
5000
5500
6000
6500
Preciptation (Inches)

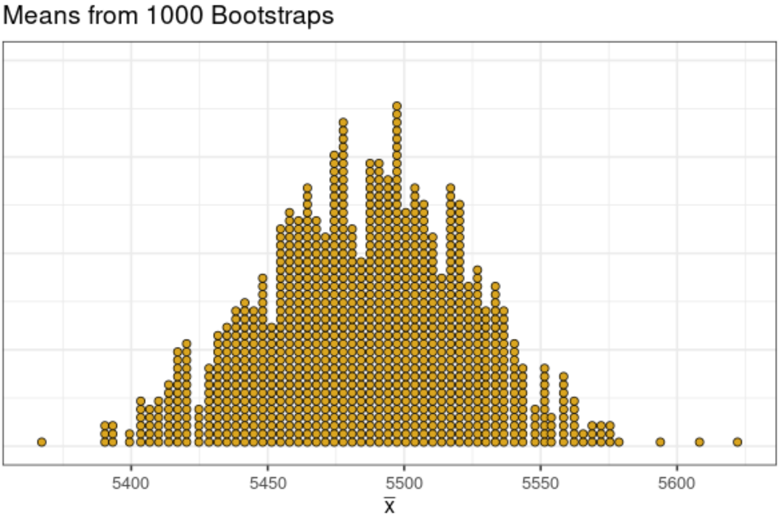

Male Gentoo Penguins (N = 61)
Penguin Sample (n = 30)
count
0
5
10
0.0
2.5
5.0
7.5
5000
5500
6000
4500
Mass (grams)
Mass (grams)


Penguin Sample (n = 30)
7.5
5.0
2.5
0.0
4500
5000
5500
6000
Mass (grams)
Bootstrapped Samples (n = 30)
Sample 1
Sample 2
Sample 3
Sample 4
15
10
5
0
4500 5000 5500 6000
Mass (grams)



Bootstrapped Samples (n = 30)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
0
5
10
15
4500 5000 5500 6000
Mass (grams)


Bootstrapped Samples (n = 30)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
0
5
10
15
4500 5000 5500 6000
Mass (grams)


Means from 1000 Bootstraps
5400
5450
5500
5550
5600
$\overline{X}$

Bootstrapped Sampling Distribution

The collection of bootstrapped statistics gives us an estimate of the *sampling distribution*

- What values did we see?
- How frequently did they appear?

In this case, because we were bootstrapping the sample mean we find that the sampling distribution looks approximately normal

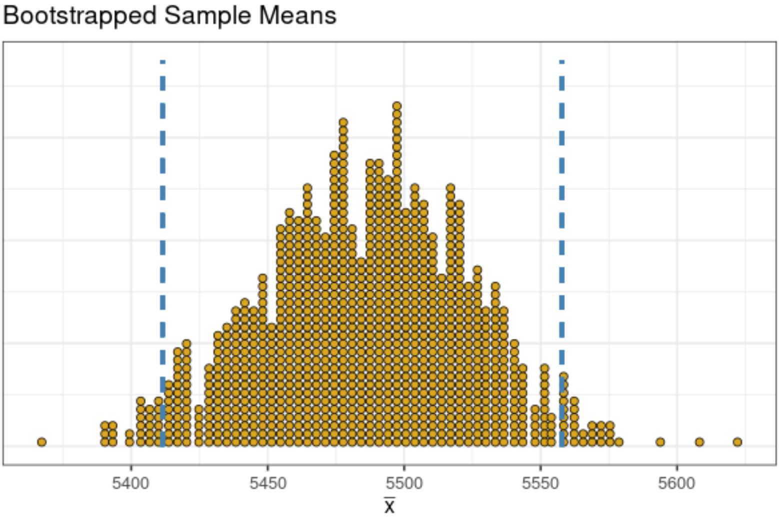
In general, we should expect that our bootstrapped sampling distribution to be theoretically identical to the *true* sampling distribution, whatever that may be

Quantiles

Recall that Monday we introduced the `qnorm()` and `qt()` functions to find us the quantiles of our sampling distribution when the distribution was known

If we don't know the distribution, we can use a vector of values, along with the `quantiles()` function to perform a similar task

```
> ## Find mean/se
> mean(mass_sample); se(mass_sample)
[1] 5484.8
[1] 40.096
>
> ## Find critical value from t
> quants95 <- c(0.025, 0.975)
> qt(quants95, df = 29)
[1] -2.0452  2.0452
>
> ## Point +/- MOE
> 5484.8 + c(-2.0452, 2.0452)*40.096
[1] 5402.8 5566.8
>
> ## Using quantile() on bootstrap
> quantile(boot_sample, probs = quants95)
  2.5%  97.5%
5408.6 5565.2
```

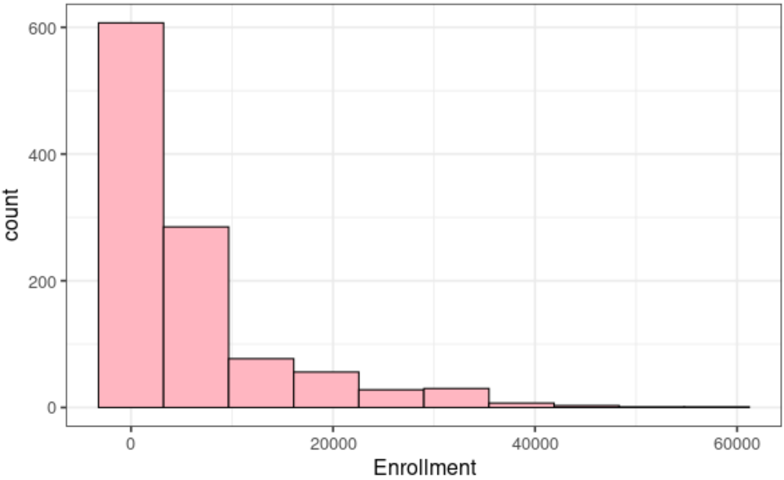

Bootstrapped Sample Means
5400
5450
5500
5550
5600
$\bar{x}$

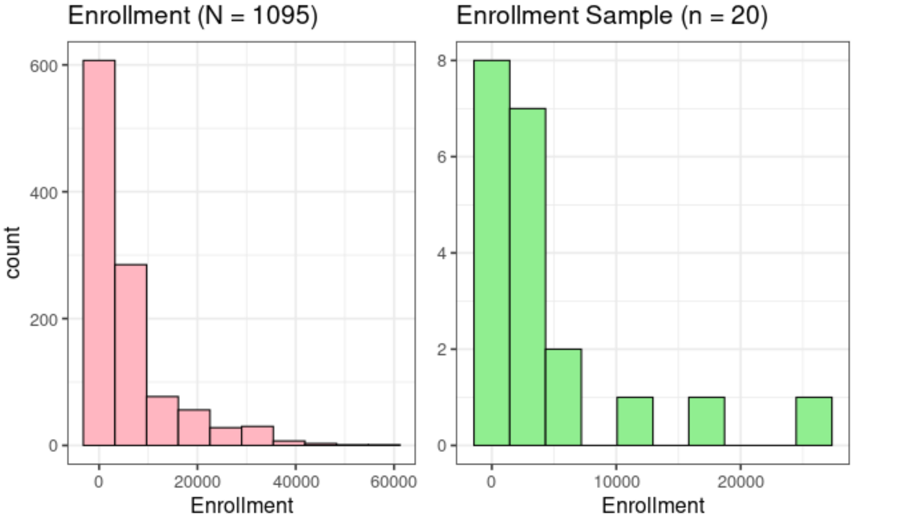
When the sampling distribution is approximately normal, the quantiles of the bootstrap should match closely with those computed using the margin of error method

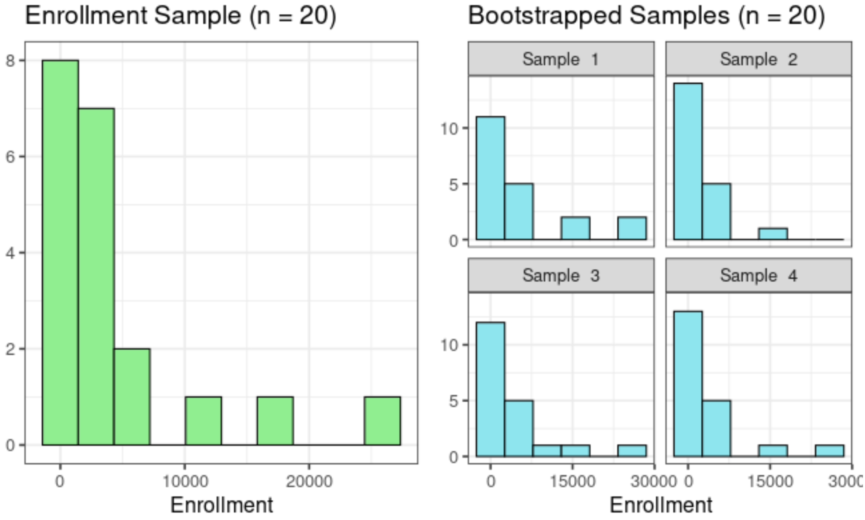
Boostrapping is especially appropriate when our sampling distribution is not normal:

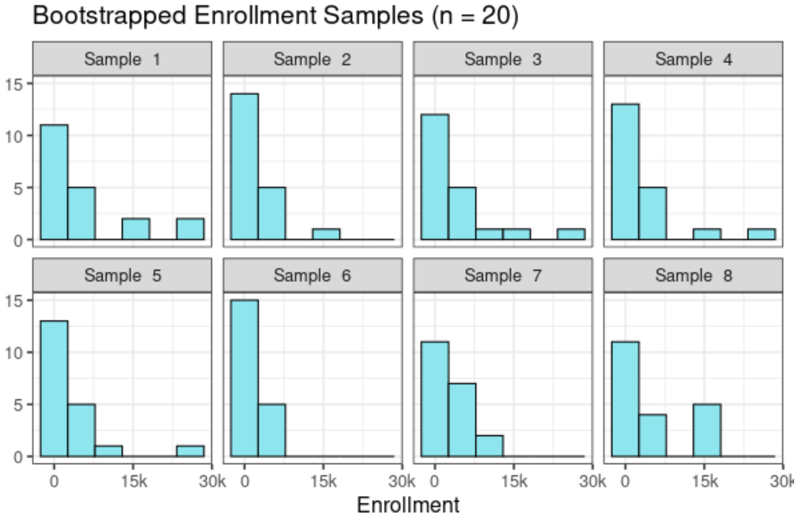
- The population variable is highly skewed
- The number of observations in our sample is not large enough for CLT approximation
- We want the sampling distribution of a statistic that is *not* normally distributed

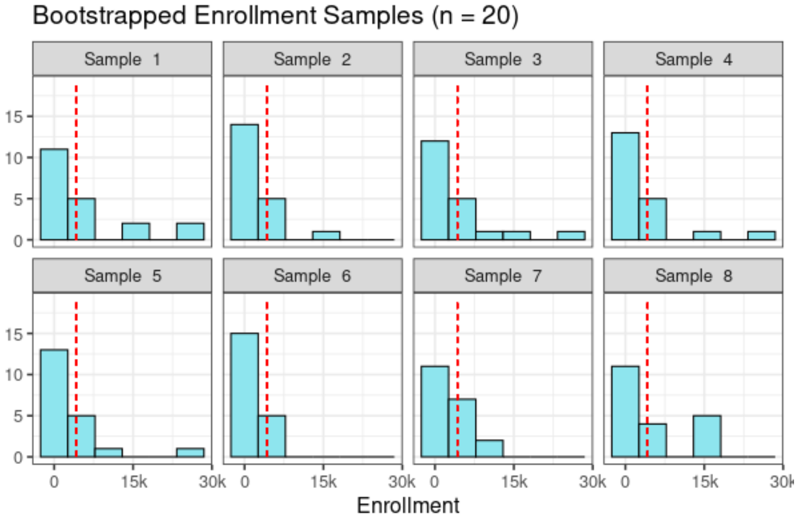
College Enrollment (N = 1095)

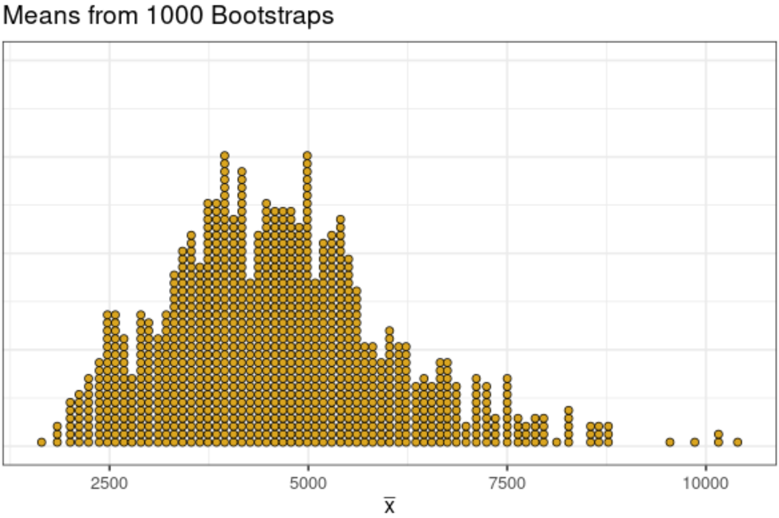



Enrollment (N = 1095)
count
600
400
200
0
0
20000
40000
60000
Enrollment
Enrollment Sample (n = 20)
8
6
4
2
0
0
10000
20000
Enrollment


Enrollment Sample (n = 20)
8
6
4
2
0
0
10000
20000
Enrollment
Bootstrapped Samples (n = 20)
Sample 1
Sample 2
Sample 3
Sample 4
10
5
0
0
15000
30000
Enrollment


Bootstrapped Enrollment Samples (n = 20)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
15
10
5
0
0
15k
30k
Enrollment


Bootstrapped Enrollment Samples (n = 20)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
0
5
10
15
0
15k
30k
Enrollment

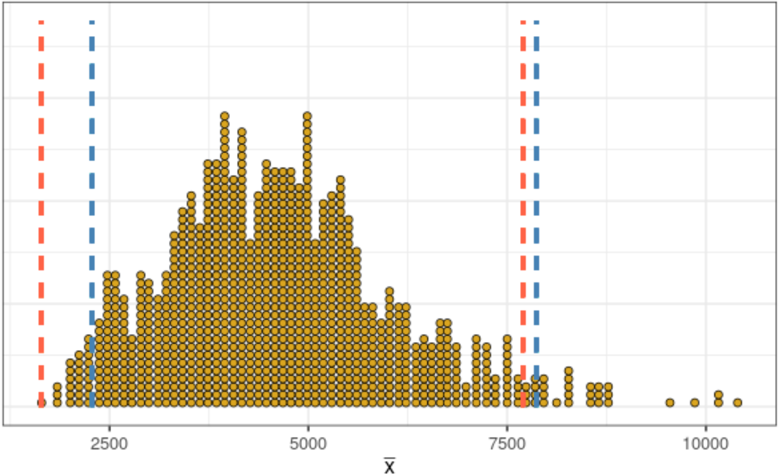


Means from 1000 Bootstraps
2500
5000
7500
10000
$\bar{x}$

Creating confidence intervals for college enrollment

```
> quants <- c(0.025, 0.975)
> qt(quants, df = 20-1)
[1] -2.093  2.093
>
> ## Point estimate +/- MOE
> 4669.4 + c(-2.093, 2.093) * 1447.3
[1] 1640.2 7698.6
>
> ## Find quantiles directly from sample distribution
> quantile(enrollment_bootstrap,  c(0.025, 0.975))
  2.5%  97.5%
2279.8 7867.2
```

Bootstrapped Sample Means



Frequently Unasked Questions

Concern: What if our sample sucks?

Concern: If we can use bootstrapping, why do we bother with CLT?

Review

Bootstrapping involves the process of *resampling with replacement* from our original sample

When we compute a statistic on our bootstrapped sample (i.e., sample mean), we have a *bootstrapped sample statistic*

Repeating this process many many times gives us an estimate of the *sampling distribution*

Quantiles can be used on this bootstrapped sampling distribution without needing any further assumptions