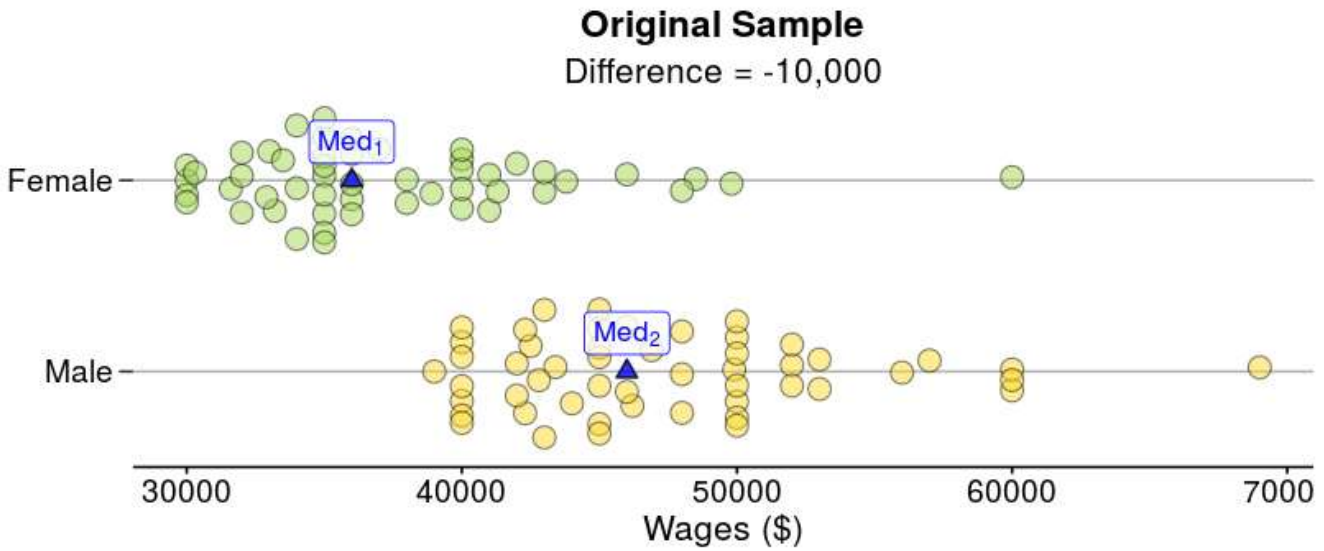
BCS1520: Statistics

Lecture 6.5: Bootstrapping

Maastricht University | Department of Advanced Computing Sciences

Measuring Uncertainty: Difference in Averages



We know wage is a thick-tailed distribution.

So mean is a difficult quantity to use to judge the distribution.

We want instead to look at the median.

Measuring Uncertainty: Difference in Averages

Descriptive Statistics:

Group	Sample Size	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
Female	51	37,388	5,893	30,000	33,750	36,000	40,000	60,000
Male	51	47,339	6,342	39,000	42,650	46,000	50,000	69,000

From descriptive statistics, we see that there is a difference in the medians of ~10,000. Can we assess the margin of error on this?

Measuring Uncertainty: Gender Wage Gap

Descriptive Statistics:

Group	Sample Size	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
Female	51	37,388	5,893	30,000	33,750	36,000	40,000	60,000
Male	51	47,339	6,342	39,000	42,650	46,000	50,000	69,000

IF we can assume the population distribution looks like the sample distribution then we could randomly sample 102 individuals, with replacement several times to get a sense for the margin of error.

This is called *bootstrapping.*

Measuring Uncertainty in Gender Wage Gap

Definition

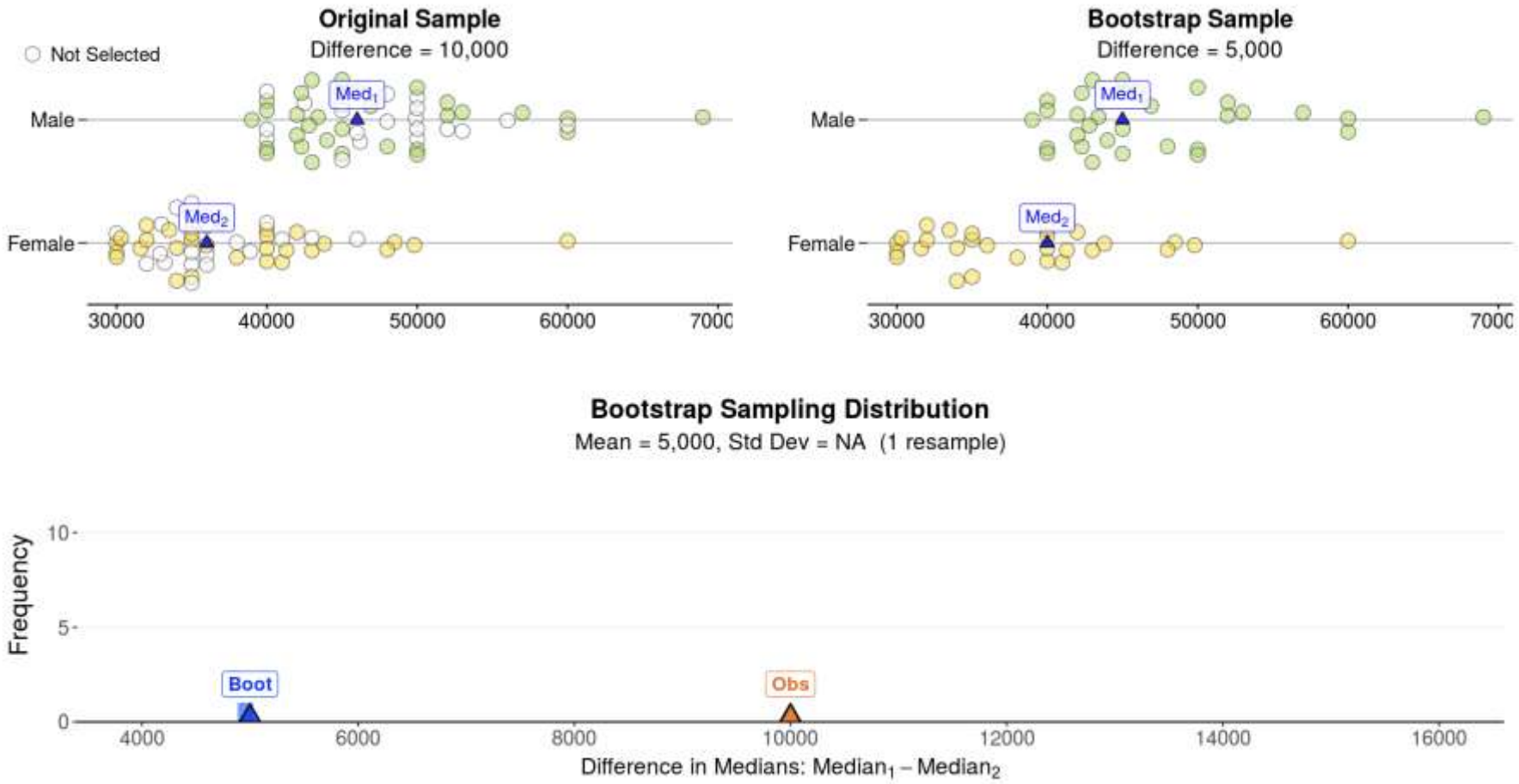
Bootstrapping (the process of pulling yourself up by your boot straps) refers to the process of using the available samples to try to approximate the population distribution of a statistic.

- Under fairly broad assumptions, this is quite successful at giving us a distribution for our statistic of interest.
- The alternative would require us to assume a parametric form (like the bell curve) for the population distribution (i.e. something to tell us how the sampling distribution will look), which is a strong assumption, but can also be useful.

Note: Sampling Distribution vs Data Distribution

- Sampling distribution is a probability distribution of a statistic derived from the sample, i.e. the data.
- The sample itself has some distribution, this derives from the population, or if biased, is altered from the population distribution in some way – may be referred to as *data distribution.*
- We will come back to this later.

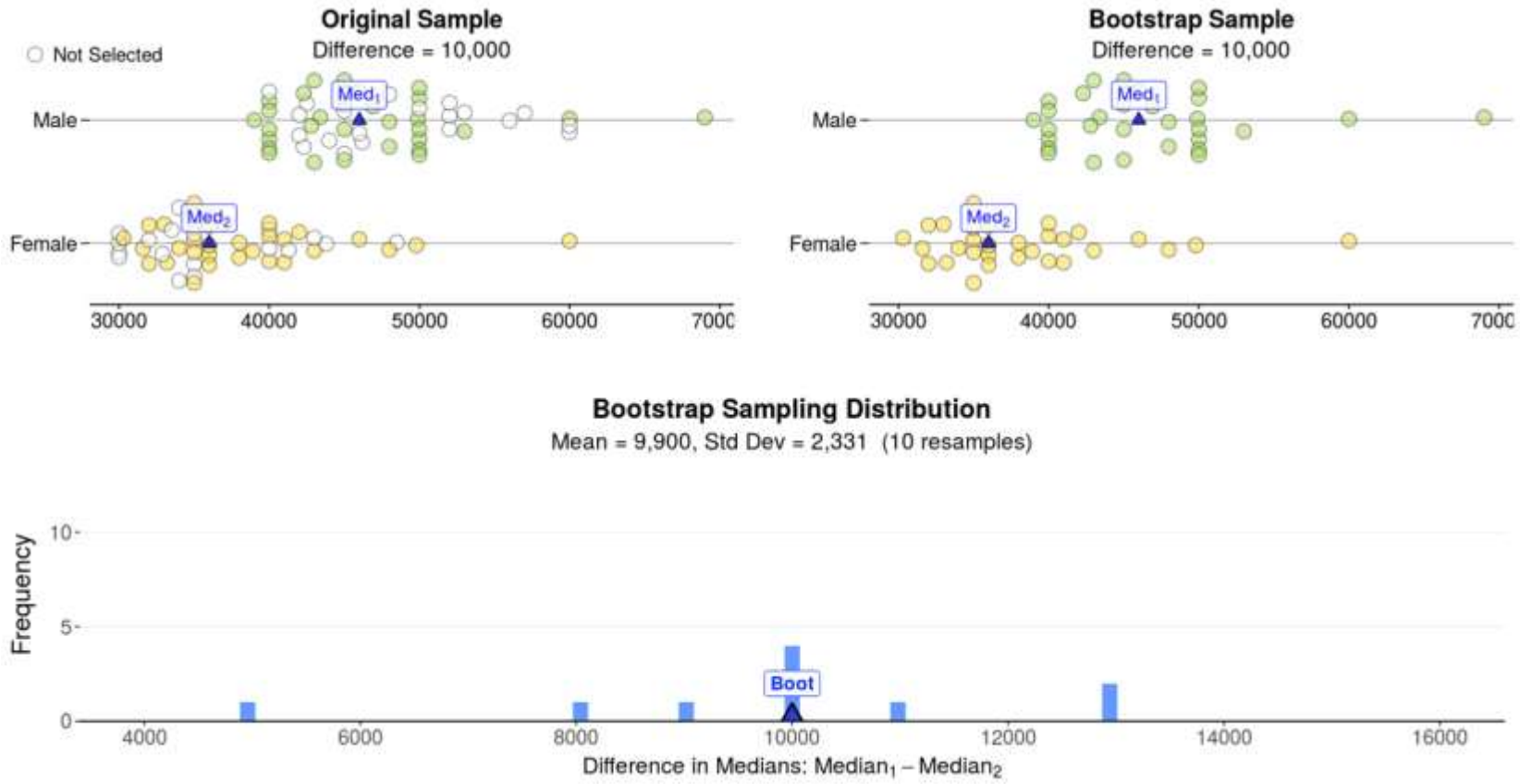
Measuring Uncertainty in Gender Wage Gap



https://istats.shinyapps.io/Boot2samp/

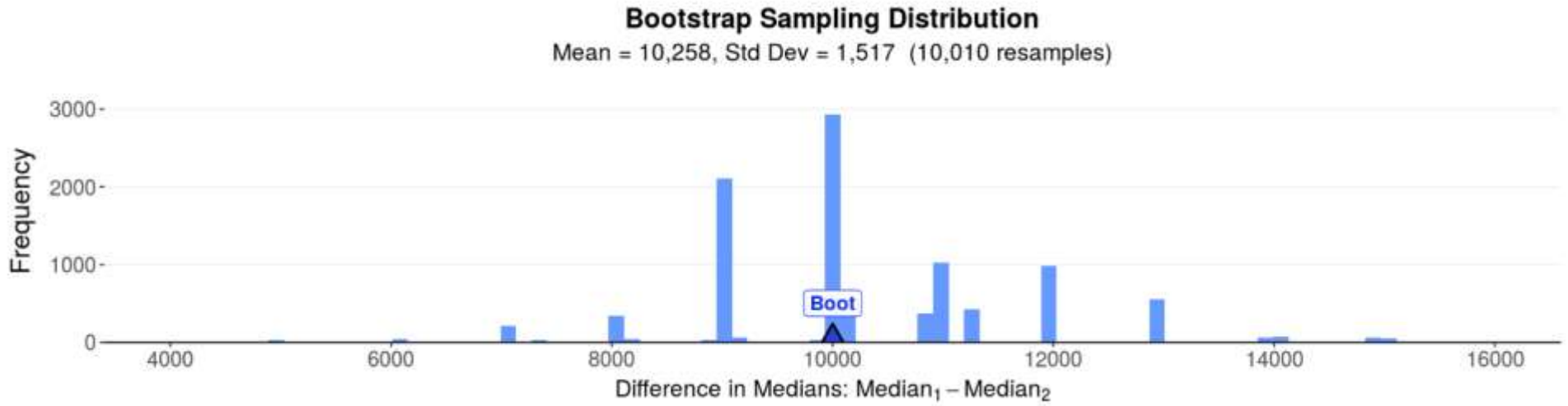
Measuring Uncertainty in Gender Wage Gap

- The sampling distribution after 10 resamples is forming but kind of random.



https://istats.shinyapps.io/Boot2samp/

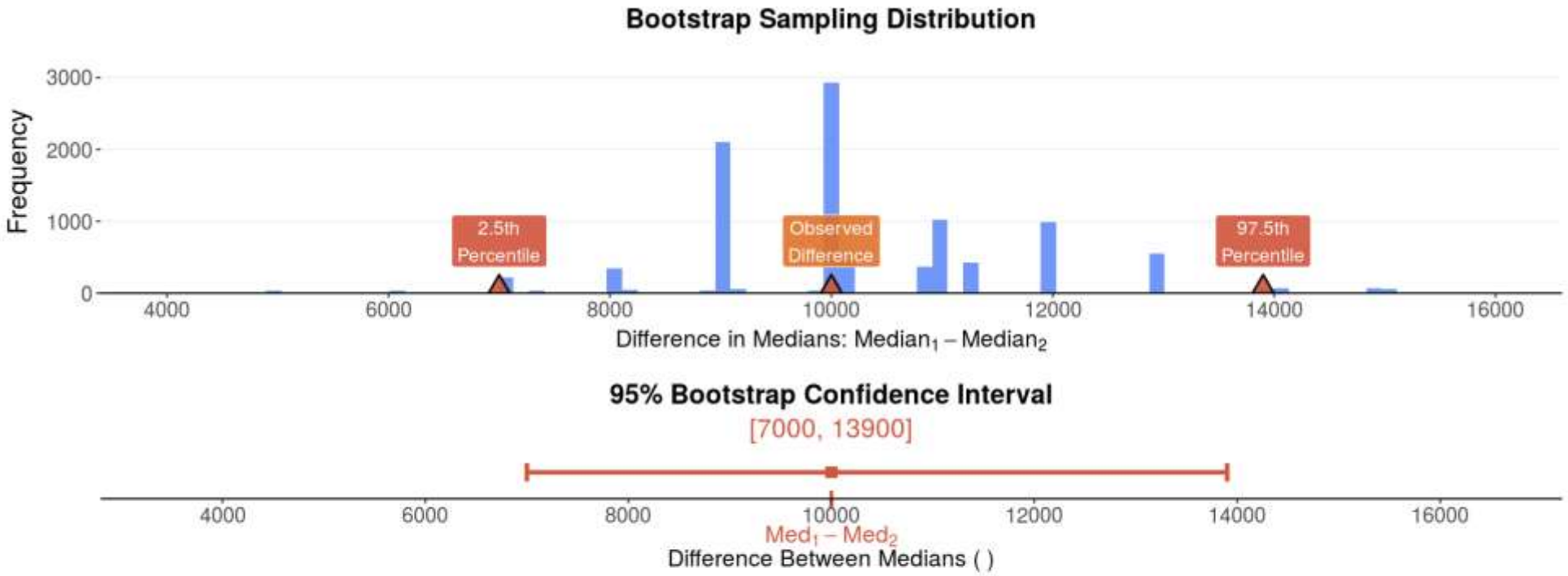
Measuring Uncertainty in Gender Wage Gap



- Repeated 10000 times!
- Now we have an actual distribution on the medians.
- Curiously, it looks like a bell curve.

https://istats.shinyapps.io/Boot2samp/

Measuring Uncertainty in Gender Wage Gap



Bootstrap Percentile Confidence Interval:

Population Parameter	Point Estimate	Lower Bound	Upper Bound	Confidence Level
Difference Between Medians	10,000	7,000	13,900	95%

https://istats.shinyapps.io/Boot2samp/

How to run the bootstrap?

Bootstrap Pseudo-Algorithm:

1. Draw n samples from the data for one bootstrap resample.
2. Assess the statistic of interest (difference in medians for example) on the new sample.
3. Repeat (1 and 2) B times.
4. Generate the distribution of the statistic and assess the $100 * \alpha/2$ and $100 * (1 - \alpha/2)$ percentiles to get the $1 - \alpha$ Confidence Interval.

   (e.g. If $B = 200$, then you can sort the estimated values of your statistic across 200 simulated statistics and take the 5th and 195th values to get a 95% CI.)

B is the number of times you resample the data – many numbers are quoted as being reasonable here. To be on the safe side I tend to lean towards doing 1000 or more resamples, but this may simply be an abundance of caution.

Others suggest even $B = 100$ will be sufficient. Depends on your dataset!

As an aside: **Never trust arbitrary numbers as an adjudicator of truth.**

Measuring Uncertainty in Gender Wage Gap

- Through bootstrapping we were able to get a measure on the uncertainty of the difference in median wages.
- We did not make any assumptions about the distribution and indeed, for the median, it can be difficult to establish a distribution.

Measuring Uncertainty

- Summary statistics have some uncertainty.
- Can measure uncertainty using bootstrapping.
- Sampling distributions of summary statistics need not look like the original distribution (remember that wages were heavy-tailed!)

Using Bootstrapping on Linear Regression

- What if we wanted a measure of uncertainty on the parameters b_0 and b_1 in the following equation?

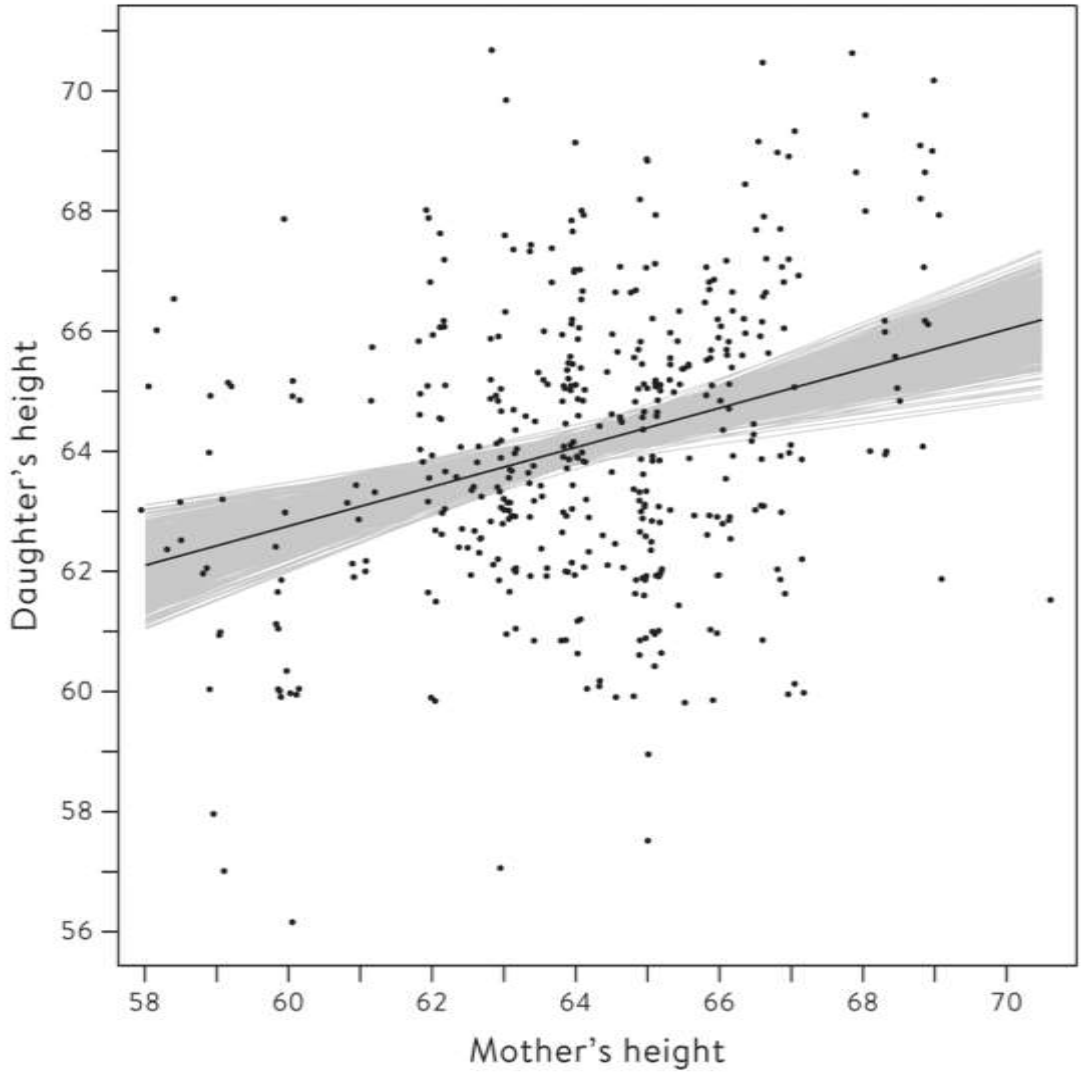
- $Daughter's\ height = b_0 + b_1 * Mother's\ Height$

Using Bootstrapping on Linear Regression

$$Daughter's\ height = b_0 + b_1 * Mother's\ Height$$

- We can bootstrap!
- We will resample the data of mother and daughter heights $(x_1, y_1), (x_2, y_2)$ … and refit the best-fit line for each resample.
- From this we essentially have a sampling distribution of lines!

Using Bootstrapping on Linear Regression



Summary

- A summary statistic is meaningless without a sense for the uncertainty.
- Bootstrapping is a powerful technique that gives us a measure of uncertainty on the summary statistic without making strong assumptions of the population distribution.
- Can apply bootstrapping in a large range of cases!