ST 516: Foundations of Data Analytics

Sampling distributions

Recall: Sampling from a population

Imagine I'm interested in the heights of US adults. I randomly sample one person from all US adults and measure their height.

We can think of the value I measure as a random variable,

$$X = \text{height of one randomly sampled US adult}$$

X is random because before I do the study I don't know what value I will get for X.

What is this random variable's probability distribution?

Under simple random sampling, the probability distribution is simply the relative frequency of all possible heights in the population. We call this the **population distribution**.

What random variables are there?

When we take a simple random sample of more than one unit from a population, there are a number of random variables we might consider:

- A single sampled value, say the first, X_1, is a random variable, whose probability distribution is described by the population distribution.
- The collection of all the sampled values, $X_1, X_2, \ldots, X_n$, is a collection of independent random variables, each whose probability distribution is described by the population distribution.
- A statistic calculated using the sampled values (for example, a sample average, or sample median) is a random variable, and it has a probability distribution. We call this distribution the **sampling distribution** for the statistic.

What is a sampling distribution?

The **sampling distribution** represents the distribution of a statistic based on all possible samples of a fixed size from a certain population. It is useful to think of a particular statistic as being drawn from such a distribution.

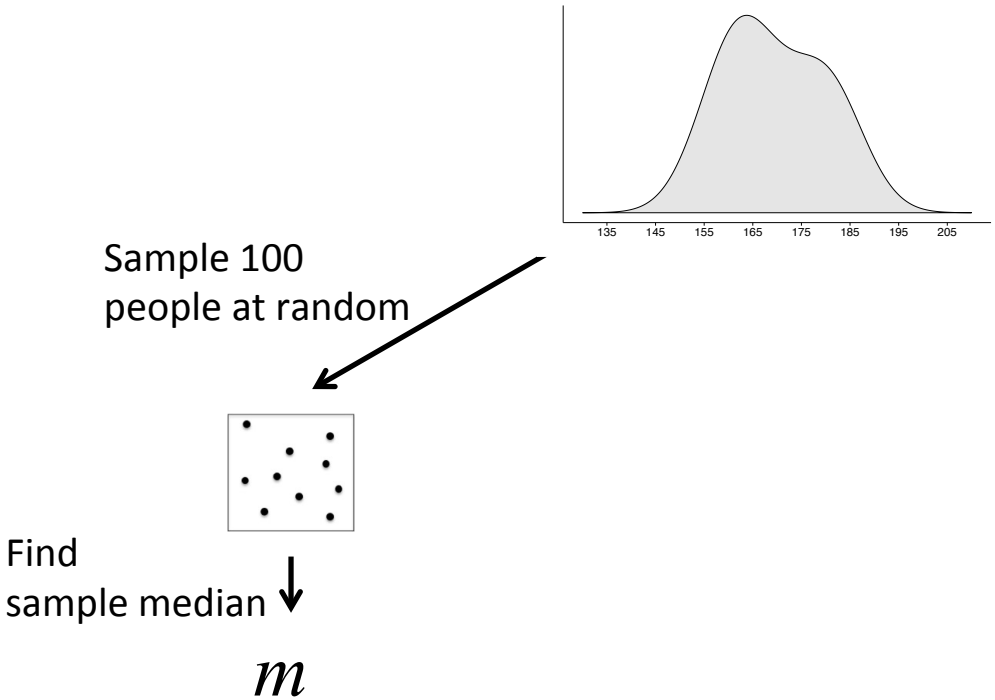
To talk about a specific sampling distribution, you must specify:

- the statistic being calculated on the sample
- the number of observations in the sample (the sample size)
- and the population from which samples are drawn

Understanding properties of the sampling distribution for a specific statistic, is the key to being able to use the sample value to generalize to the population value.

Example

Consider sampling 100 US adults at random and measuring their height. Then, calculate the median of these 100 heights.



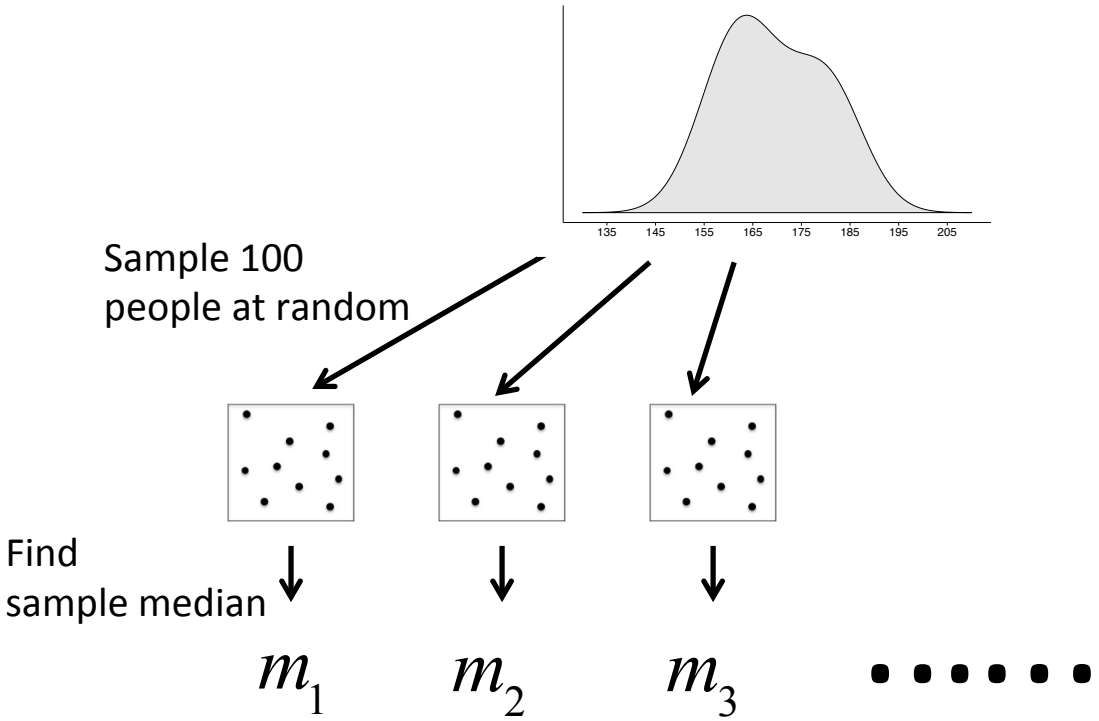
Example

The population distribution, is the distribution of all adult heights in the US.

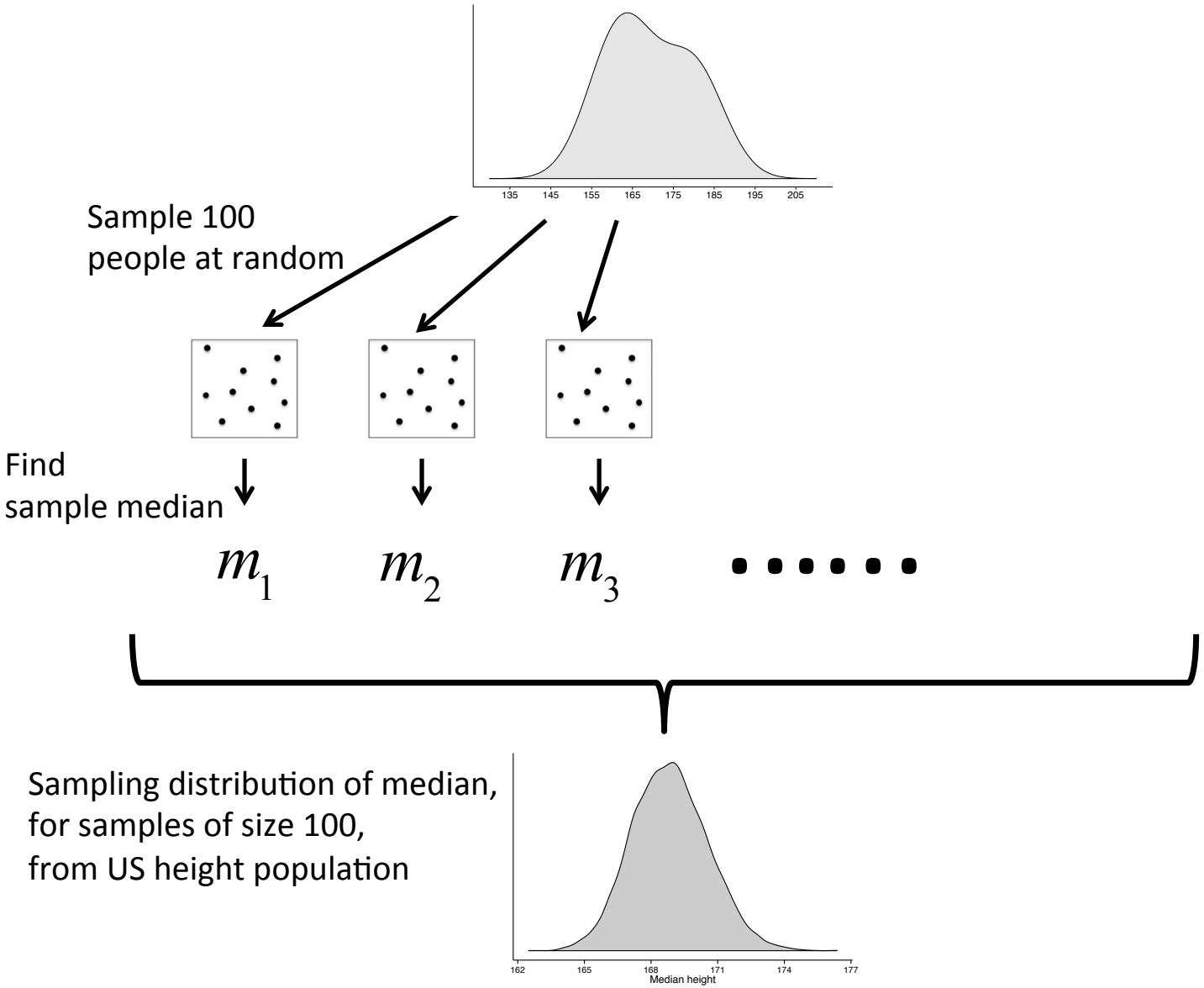
We draw a sample from the population. The sample is of size 100.

We summarize the sample by calculating the sample median.

Example: imagine repeating the sampling



Example: look at the distribution of the sample medians



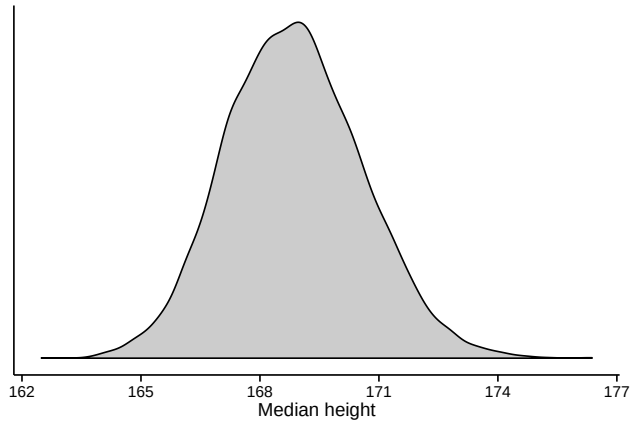
More examples

Any change would result in a different sampling distribution:

- Change the sample size: sampling distribution of the median, for samples of size 1000, from the US adult height population
- Change the statistic: sampling distribution of the standard deviation, for samples of size 1000, from the US adult height population
- Change the population: sampling distribution of the standard deviation, for samples of size 1000, from a Normal population

Generalizations about sampling distributions often take the form, "Regardless of the population distribution, for large sample sizes, the sampling distribution of T is approximately . . . " where T is a specific statistic.

Useful properties to know about a sampling distribution



The center of the sampling distribution can be summarized by the expected value. This tells us, on average over many samples, the value the statistic will take.

The spread of the sampling distribution can be summarized by the variance or standard deviation. This tells us how much the statistic will vary around it's expectation over many samples.

Exploring sampling distributions by simulation

General procedure:

1. Draw a sample of the specified size from the population
2. Calculate the summary statistic using the sample
3. Repeat 1 & 2 many times and draw a histogram of the statistics

The histogram is our simulation based approximation of the sampling distribution.

Example

If `sim_height(n)` samples n heights at random for the population of US adults, I can generate a single sample of size 100 with

```
sim_height(100)
```

To generate the sample and find the median

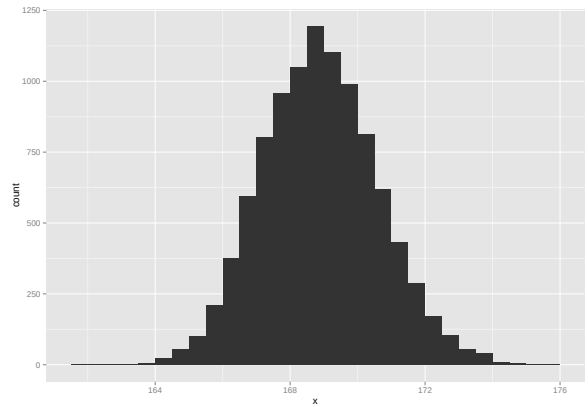
```
median(sim_height(100))
```

To repeat many times,

```
x <- replicate(10000, median(sim_height(100)))
```

Example

```
qplot(x, binwidth = 0.5)
```



Our simulation based approximation to the sampling distribution of samples of size 10 from the adult US height population.