February 28, 2018

CS 361: Probability & Statistics

Samples and populations

Inference

Inference

We're going to begin the section on inference. Inference at its most broad is the process of drawing principled conclusions from data

In many of the settings we will consider we will only have data and not necessarily know about the underlying probability distributions that generated the data

In the 4 upcoming chapters we will look at: how to draw general conclusions from a specific subset of data, how to assess the significance of the evidence that exists against a hypothesis about our data, how to tell if two groups are experiencing different effects, and how to infer a probability model from a dataset

Samples and Populations

Samples and populations

Sometimes we only have a subset of the data that we could have.

If we want to know the average heights of students, we probably won't have the means to measure the height of every student. If we want to predict how the public is going to vote, we probably won't be able to afford to ask every single potential voter

The terminology we will use to talk about the set of data items we can actually measure is a **sample**. And the term for the entire dataset to which we may not have access is called the **population**

So the question becomes under what conditions can we use our sample to say things about the population at large, like what the variance and mean of some measure of the population is, for instance.

Samples and populations

For the next few slides we will suppose that our sample is of size N and that our population is of size N_p. And we will suppose that N is much smaller than N_p

How did we get the sample that we got? To model sampling, we will assume that each item of the sample is chosen independently and fairly from the population.

A classic way of thinking about this is by imagining that each potential data item in the population corresponds to a ticket and that we are drawing these tickets one by one from a jar, writing down what we draw, and then putting the ticket back in the jar. This is sometimes called the urn model (replace jar with urn in the description)

This is also known as **sampling with replacement**

Samples and populations

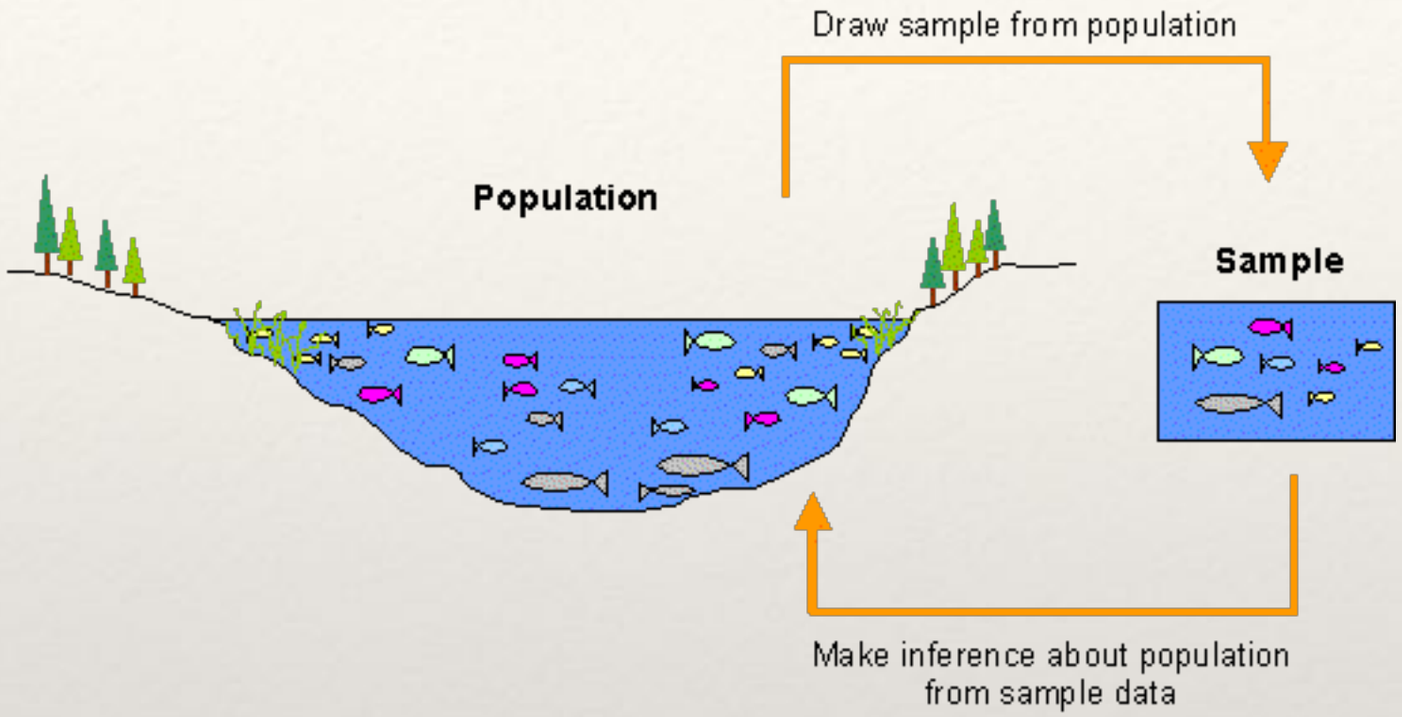
With this model for sampling we will see whether the sample can tell us anything about the population. In particular, we will focus on the mean of the population and estimating it by looking at our random sample

The **population mean**, which we can't actually observe, is indicated with the notation popmean({x}). If we could look at the entire population we would calculate this directly using our formula for the mean of a dataset from Chapter 2

Instead, we have a sample of size N. The mean of our sample is called the **sample mean**

The **sample mean** is a random variable (it could be different depending on which N items we get when we randomly sample). We will write $X^{(N)}$ for this random variable and determine its value for a given sample according to our formula for the mean of a dataset

Example: 3 means, 3 variances



Source: http://labs.geog.uvic.ca/geog226/frLab3.html

Weight of fish in the sample and population both have a mean and variance

The sample mean is a random variable which means it has a distribution, i.e. an expected value (or mean) and a random variable variance.

To analyze whether the sample mean is a good estimator of the population mean, we will have to analyze the expectation and variance of the random variable which describes the randomness according to which we got our sample

Samples and populations

Let's try and relate the sample mean to the population mean

$X^{(N)}$ is the random variable that outputs the sample mean of a sample of size N

The value of the sample mean for a given sample is given by

$$X^{(N)} = \frac{1}{N}(X_1 + X_2 + \ldots + X_N)$$

If we take expectations we get

$$E[X^{(N)}] = \frac{1}{N}(E[X^{(1)}] + E[X^{(1)}] + \ldots + E[X^{(1)}])$$

Or

$$E[X^{(N)}] = E[X^{(1)}]$$

Samples and populations

But we can reason about the expected value of $X^{(1)}$ on the basis of the way we've defined how we are sampling

$X^{(N)}$ is the random variable that outputs the sample mean of a sample of size N

Using the definition of expectation—that we just sum over every possible value of a random variable and multiply by its probability—we have

$$E[X^{(1)}] = \sum_{i \in 1,\ldots,N_p} x_i p(i)$$

every member of the population

We've assumed we draw fairly from the jar, so this can be rewritten

$$E[X^{(1)}] = \sum_{i \in 1,\ldots,N_p} x_i \frac{1}{N_p}$$

Samples and populations

$$E[X^{(1)}] = \sum_{i \in 1,\ldots,N_p} x_i \frac{1}{N_p}$$

Rewriting

$$E[X^{(1)}] = \frac{\sum_{i \in 1,\ldots,N_p} x_i}{N_p}$$

Which is the formula for the mean of the population, i.e. $E[X^{(1)}] = \text{popmean}(\{x\})$

We showed earlier that $E[X^{(N)}] = E[X^{(1)}]$ thus the we've shown the following

$$E[X^{(N)}] = \text{popmean}(\{x\})$$

Estimators

$$E[X^{(N)}] = \text{popmean}(\{x\})$$

This was a nice result. In general when the we find a random variable that has as its expectation something we are trying to estimate, we call the random variable an **unbiased estimator** of that quantity

Something we will note but not show here is that the variance of the sample dataset with the variance formula applied to our sample

$$\frac{\sum_{i=1}^{N}(x_i - X^{(N)})^2}{N}$$

gives us a random variable, but this random variable is <u>not</u> an unbiased estimator of the population variance. Ie the expected value of the above does not equal the population variance

Estimators

The quantity given by

$$\frac{\sum_{i=1}^{N}(x_i - X^{(N)})^2}{N-1}$$

however, is an unbiased estimate.

In other words, the above quantity is a random variable (it is different depending on exactly which sample of N items we happen to draw) with expectation equal to the population variance

When we refer to the **sample variance** or the **sample standard deviation** we will be referring to this quantity or, respectively, its square root. For a large enough N there won't be much difference between this quantity and the variance formula applied to the sample dataset

Sample mean - variance

So we know that the expected value for the sample mean is equal to the population mean, but we might randomly get a sample that is far from this expected value and thus get a bad estimate for the population mean. How can we analyze the likelihood of this happening? By analyzing the *variance* of the random variable $X^{(N)}$

If we do a little algebra we find that the variance and standard deviation of the random variable $X^{(N)}$ are: (Note this is not the variance of the sample or the population)

$$\text{var}[X^{(N)}] = \frac{\text{popsd}(\{X\})^2}{N}$$

$$\text{std}(X^{(N)}) = \frac{\text{popsd}(\{X\})}{\sqrt{N}}$$

where popsd is the the standard deviation of the population, which we cannot measure

Sample mean - variance

$$\text{std}\left(X^{(N)}\right) = \frac{\text{popsd}\left(\{X\}\right)}{\sqrt{N}}$$

The standard deviation of the estimate of the mean is sometimes called **the standard error** of our estimate. Note that we cannot calculate the above exactly. Though it still tell us some interesting things

A larger sample size reduces the error of our estimate of the popmean but that this reliability grows less than linearly in the sample size

It also says that if we are trying to estimate the mean of a population which itself has a large amount of variance, we will need a larger number of samples

Note that our error does not depend on the size of the population. Whether the population we are trying to estimate for is 1000 or 10,000,000 this is our variance

Urn model, samples, and populations

If we have managed to sample the population with a procedure that corresponds to this urn model — independent samples taken fairly from the population—then we see that estimating the mean is fairly straightforward

If our samples are not independent or aren't taken fairly from the population, however, determining what our sample tells us about the population is more difficult.

If I want to estimate the average height of people in Champaign and I do my sampling at the local daycare or at the university's basketball practice, my sample mean is not going to be a good estimate of the population mean, no matter how large it is!

Interval estimates

Interval estimates

So far we have dealt with point estimation or point inference, where we look at a dataset and get out a number estimating some quantity of interest

It can also be desirable to give a range for a quantity we are trying to estimate, along with the probability that the true value of the quantity lies somewhere in the range

If we have an interval that contains the true value with some probability we call the interval and probability a **confidence interval**

Distribution of the sample mean

The sample mean is a random variable. In particular it is a random variable that we get by summing together a bunch of independent and identically distributed random variables—our samples. From the central limit theorem, then, we can expect that our estimate—the sample mean—has a normal distribution.

That is, if we took a sample of size N, recorded the sample mean, then did this again and again the numbers we write down would be normal data

We have just shown that the mean of this normal random variable is popmean({x}) and the standard deviation is $\frac{\text{popsd}(\{X\})}{\sqrt{N}}$

Recall: Normal data

Data that is drawn from a normal distribution tends to be close to the mean in terms of the number of standard deviations. Integrating the standard normal density, we see, that if we observe a bunch of standard normal data:

$$\frac{1}{\sqrt{2\pi}}\int_{-1}^{1}\exp\left(-\frac{x^2}{2}\right)dx \approx .68$$

Around 68% of the data will be within 1 standard deviation of the mean

$$\frac{1}{\sqrt{2\pi}}\int_{-2}^{2}\exp\left(-\frac{x^2}{2}\right)dx \approx .95$$

Around 95% of the data will be within 2 standard deviations of the mean

$$\frac{1}{\sqrt{2\pi}}\int_{-3}^{3}\exp\left(-\frac{x^2}{2}\right)dx \approx .99$$

Around 99% of the data will be within 3 standard deviations of the mean

Confidence intervals for samples

Suppose we have a population and are sampling from that population

Recall that the sample mean $X^{(N)}$ is a normal random variable with expectation popmean and standard deviation $\frac{\text{popsd}(\{x\})}{\sqrt{N}}$

Given what we know about normal random variables. This tells us that for about 68% of the samples we will have

$$\text{popmean} - \frac{popsd}{\sqrt{N}} \leq X^{(N)} \leq \text{popmean} + \frac{popsd}{\sqrt{N}}$$

Confidence intervals for samples

It turns out that the variance for our estimate of popsd can be assumed to be small for reasonable sized *N*. So we can replace popsd in the interval with the sample standard deviation

And say that the true value of the population mean is bounded by

Sample standard deviation

$$X^{(N)} - \frac{sd(\{x\})}{\sqrt{N}} \leq \text{popmean} \leq X^{(N)} + \frac{sd(\{x\})}{\sqrt{N}}$$

in 68% of samples. Or in 95% of samples we have

$$X^{(N)} - 2\frac{sd(\{x\})}{\sqrt{N}} \leq \text{popmean} \leq X^{(N)} + 2\frac{sd(\{x\})}{\sqrt{N}}$$

We call these intervals the 68% and 95% confidence interval, respectively

So what?

So what does all that mean? It means we have a recipe for how we can report an interval when estimating the population mean as opposed to what we learned earlier today in the context of point estimates where we just reported a single number, the sample mean

First we calculate the sample mean $X^{(N)}$

Then we calculate the sample standard deviation using our unbiased estimator from before

$$\frac{\sum_{i=1}^{N}(x_i - X^{(N)})^2}{N-1}$$

Then we form our standard error by dividing the sample standard deviation by the square root of N

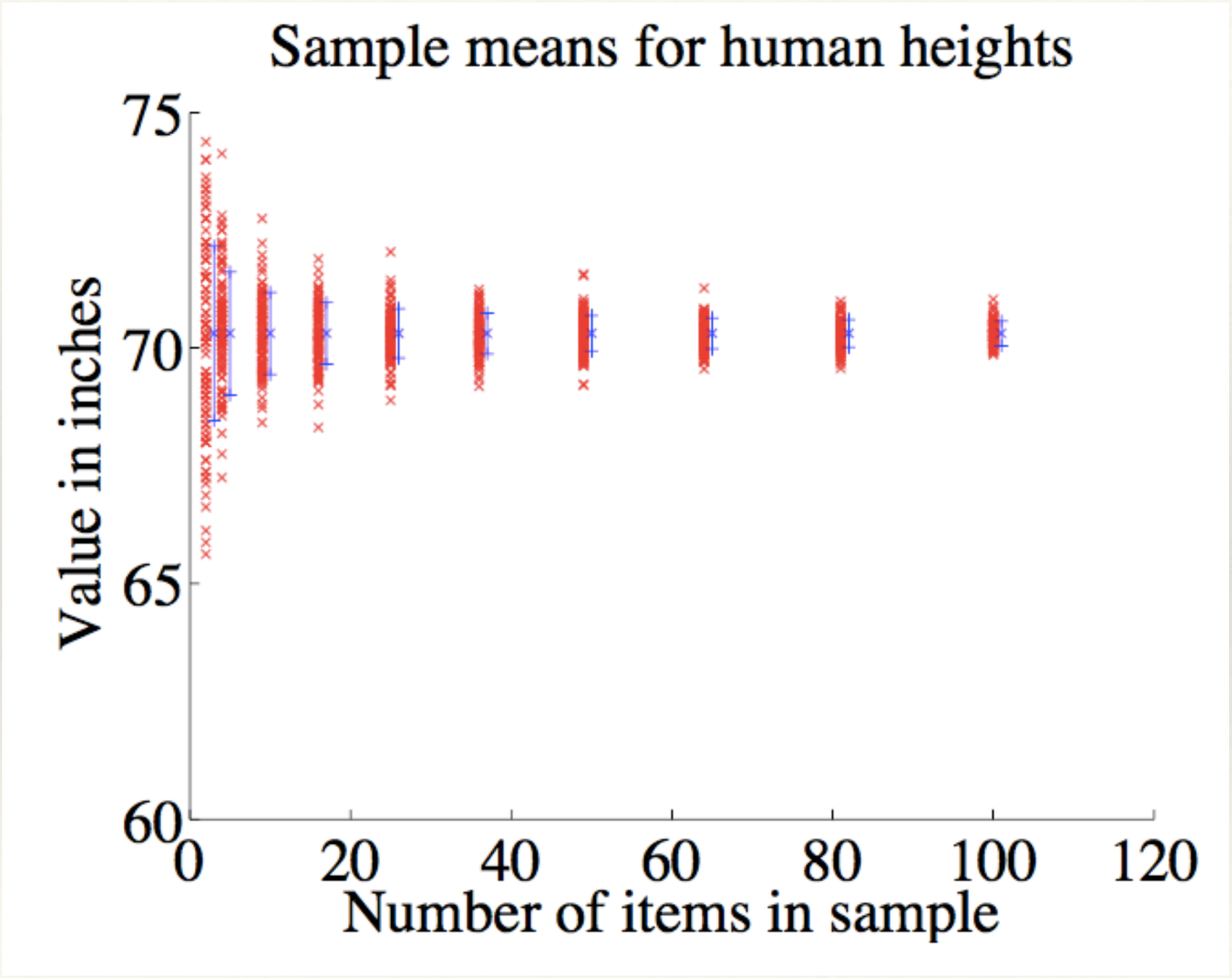
Then we report the sample mean plus or minus the standard error for a 68% confidence interval or 2*stderr for a 95% confidence interval

Example

Suppose we have 100 heights in inches of students at UIUC and the sample mean is 66 inches and the sample variance is 16.

Our sample standard deviation is thus 4. So our standard error is 4/sqrt(100) or 0.4. Thus with 95% confidence the value for the mean of heights of all students at UIUC is between 65.2 and 66.8 inches

Simulation



251 human heights in the population and various sample estimates of the mean

popmean and 1 popsd standard error bars in blue. Estimates from simulating the sampling process in red

Caveat

If our sample size N isn't sufficiently large, the sample mean doesn't exactly follow a normal distribution. We will talk a bit later about what distribution we can use to better describe the sample mean for a smaller N

Standard error

What was nice about the population mean was that our estimate for it—the sample mean—was a normal random variable whose standard deviation—the standard error—was easy to reason about.

This meant that determining a confidence interval involved invoking what we know about how much of a normal dataset is going to be within *k* standard deviations of its mean

In general, we might consider any **statistic** of a sample and want to come up with an interval of estimates for what would happen if we could measure that statistic in the whole population

Unfortunately it isn't always straightforward to calculate the standard error of our estimate. If we were trying to estimate the median, for example, the standard error of the sample median is a bit more difficult to reason about

The bootstrap

Now if we had multiple samples, we would probably want to combine them into one larger sample in order to estimate some population parameter. But if we had multiple samples we could also compute the statistic for each sample and maybe use the range of values we get to come up with an interval estimate for the statistic of interest

The bootstrap is a process to generate synthetic datasets from a single dataset to do just that.

Suppose we have a sample of N items from a population. We compute a **bootstrap replicate** of the sample by sampling N items from our original sample, uniformly and with replacement. We can do this as many times as we wish.

Since we are sampling with replacement, in general our bootstrap replicates won't just be a copy or a simple permutation of our original data.

Example

Suppose we have the following sample of heights of UIUC students in inches: {60, 58, 54, 62, 64, 73, 65, 62}. We sample from this dataset uniformly and with replacement to get the following bootstrap replicates of our dataset

{64, 60, 65, 54, 65, 54, 64, 64}
{60, 73, 60, 62, 62, 73, 58, 60}
{54, 62, 65, 54, 73, 64, 65, 64}
{64, 58, 60, 60, 60, 54, 65, 62}

Estimating the standard error

If we compute *r* different bootstrap replicates, we can use the following to get a good approximation of the standard deviation of our estimate

Compute S({**x**}_i) — that is the value of the statistic of interest on the i-th replicate, for each replicate. Use all *r* values we get this way to calculate:

$$\overline{S} = \frac{\sum_i S(\{\mathbf{x}\}_i)}{r}$$

Calculate

$$\text{stderr}\left(\{S\}\right) = \sqrt{\frac{\sum_i \left[S(\{\mathbf{x}\}_i) - \overline{S}\right]}{r-1}}$$

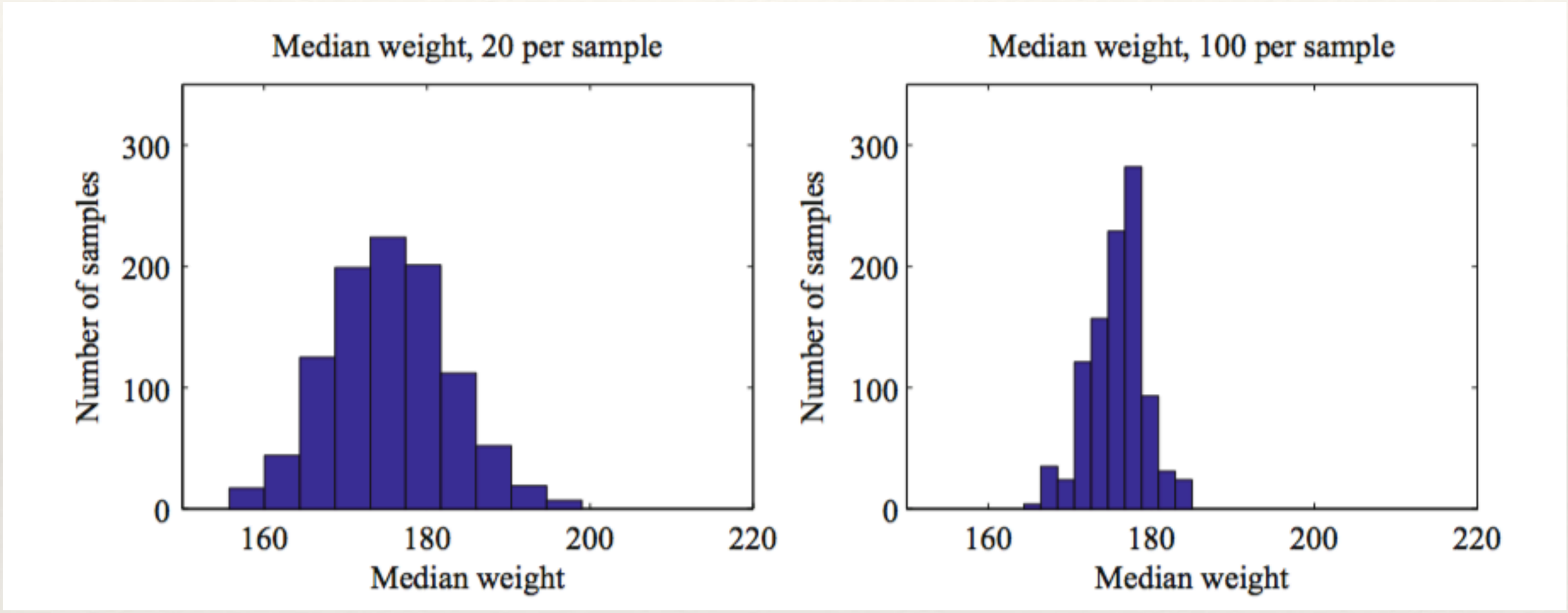
Standard error

Now we can construct a confidence interval by looking at the value of our statistic for each of the bootstrap replicates and giving a range that contains, say, 95% of the data

Or if we have reason to believe that our estimate follows a normal distribution, we can use the standard error to bound the data. A 95% confidence interval would be given by reporting

$$\overline{S} \pm 2\text{stderr}$$

Example


Median weight, 20 per sample
Number of samples
300
200
100
0
160
180
200
220
Median weight
Median weight, 100 per sample
Number of samples
300
200
100
0
160
180
200
220
Median weight