Lecture 10: Release the Kraken!

Last time

We considered some simple classical probability computations, deriving the so-called binomial distribution -- We used it immediately to derive the mathematical formula for the null distribution for Arbuthnot's test of hypothesis

We then started in on the Normal distribution and examined a simple graphical device for assessing normality of a set of data, the Q-Q or quantile-quantile plot

Today

We are going to start on estimation in earnest -- We are going to extend our toolkit of computational procedures and examine a simple way to assess uncertainty in estimates

We will use as our main navigation point the sampling distribution of an estimate -- It is a fantasy that lets us think about a host of interesting questions

These questions become answerable through something called the bootstrap, a procedure that will let us asses bias, compute standard errors, RMS and even confidence intervals

Statistics: Description

As we noted in an early lecture (2 or 3?) **a statistic is something computed from a data set** -- Statistics server **different purposes**, however

So far, we have seen **descriptive statistics** (the mean and median and inter-quartile range, say) that tell us something about a particular set of data

```
> summary(enroll$tottime)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   50.0   335.0   430.0   455.9   545.0  3760.0
```

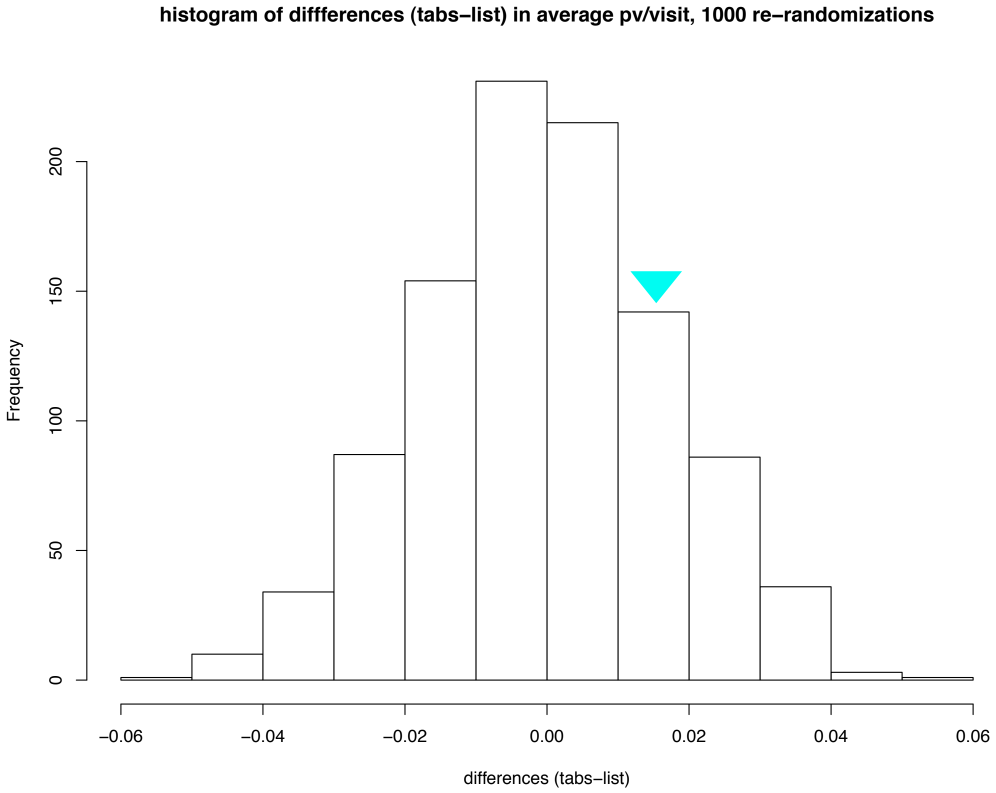
Here are a handful of descriptive statistics that helped us look at the number of minutes UCLA students spent in class per week last quarter

Statistics: Testing

Test statistics are used to judge the **adequacy of some null hypothesis** and are often compared to a reference null distribution

At the right we recall the results of the NY TImes Travel Section experiment -- The test statistic was absolute difference in average page views per visit for New York Times readers who were show the Tabs versus Lists designs

Using the experimental results, we computed a value of 0.017 which was compared to the reference null distribution (here just simple differences) to see if the value we obtained could have arisen purely by chance



Statistics: Estimation

Estimators are statistics computed from a random sample that are used to estimate the "parameters" a population distribution which we will denote generically with the greek letter θ

In our CDC BRFSS telephone survey, for example, our parameter of interest might be the fraction of the adult population in the U.S. that is obese and we could **estimate it** using a proportion computed from the CDC sample

Statistics

Note that statistics, the quantities we compute from data, **can play different roles -- **The same quantity can function as description, as the basis for a test or as an estimate of some quantity in the parent population

The source of the data, its design, and the goals of our study suggest which of these different roles are sensible or desirable

This lecture, we’ll focus on estimates!

Random sampling

Toward this end, we want to start looking at a statistical design other than random allocation into one or more treatments -- Specifically we will examine **random sampling from a population** and examine what we can infer about a population from a sample

To be a little rigorous about this, we’ll first spell out what we mean by a population...

Populations

A population (at least for the purposes of this course) consists of a number of units or cases with **three characteristics**

1. They all have or could have values for variables we are interested in;

2. We are interested in learning something about the distribution of one or more variables across the population; and

3. The population is sufficiently well-defined that we can draw from it a random sample of cases (think placing some identifier for each case in a -- possibly enormous -- hat)

There are several different kinds of populations -- **A natural population, for example, is something larger than the study you're conducting** and has "a degree of permanence to it" (the population of the U.S., all the employees of a company)

Populations

A prospective population, on the other hand, is linked to one of our previously discussed experiments -- For example, suppose we draw a random sample from a group of people suffering from a particular disease (a natural population) and we randomize them to receive either the standard or a new therapy

The two groups (standard and new therapy) are samples from different prospective populations -- Each prospective population consists of the same cases as in the associated natural population (the group of all patients suffering from the disease), but we pretend that all of them have received one or the other therapy

This explains the term “prospective” because in truth **only the patients in the sample receive the treatments** being studied

Populations

Constructed populations are as "fully defined as a natural population" but lack permanence and some exist solely to provide random samples -- For example, all of the students enrolled in the introductory courses in the psychology department of a given university might be used as a constructed population from which psych researchers draw samples for their various experiments

The constructed population, then, is one that is easily sampled -- We hope that it is similar to some natural population that you would like to study (like all the students at the university or all college students) if you had the resources

Random samples

Just like randomized allocation into treatment groups provided us with the ability to perform statistical tests, **random sampling from populations will be the underlying motivation** and justification of a collection of **inferential procedures**

Forming a random sample from a population involves mimicking in some way the act of placing an identifier for each unit in the population into **a “hat” and drawing a small sample**

The procedure used by the California Secretary of State for creating a randomized alphabet is a good mental model...

The Secretary of State shall conduct a drawing of the letters of the alphabet, the result of which shall be known as a randomized alphabet. The procedure shall be as follows:

(a) **Each letter of the alphabet shall be written on a separate slip of paper, each of which shall be folded and inserted into a capsule.** Each capsule shall be opaque and of uniform weight, color, size, shape, and texture. **The capsules shall be placed in a container, which shall be shaken vigorously in order to mix the capsules thoroughly. The container then shall be opened and the capsules removed at random one at a time.** As each is removed, it shall be opened and the letter on the slip of paper read aloud and written down. **The resulting random order of letters constitutes the randomized alphabet, which is to be used in the same manner as the conventional alphabet in determining the order of all candidates in all elections.** For example, if two candidates with the surnames Campbell and Carlson are running for the same office, their order on the ballot will depend on the order in which the letters M and R were drawn in the randomized alphabet drawing.

(b) (1) There shall be six drawings, three in each even-numbered year and three in each odd-numbered year. Each drawing shall be held at 11 a.m. on the date specified in this subdivision. The results of each drawing shall be mailed immediately to each county elections official responsible for conducting an election to which the drawing is applicable, who shall use it in determining the order on the ballot of the names of the candidates for office.

(A) The first drawing under this subdivision shall take place on the 82nd day before the April general law city elections of an even-numbered year, and shall apply to those elections and any other elections held at the same time.

(B) The second drawing under this subdivision shall take place on the 82nd day before the direct primary of an even-numbered year, and shall apply to all candidates on the ballot in that election.

(C) (i) The third drawing under this subdivision shall take place on the 82nd day before the November general election of an even-numbered year, and shall apply to all candidates on the ballot in the November general election.

(ii) In the case of the primary election and the November general election, the Secretary of State shall certify and transmit to each county elections official the order in which the names of federal and state candidates, with the exception of candidates for State Senate and Assembly, shall appear on the ballot. The elections official shall determine the order on the ballot of all other candidates using the appropriate randomized alphabet for that purpose.

(D) The fourth drawing under this subdivision shall take place on the 82nd day before the March general law city elections of each odd-numbered year, and shall apply to those elections and any other elections held at the same time.

(E) The fifth drawing under this subdivision shall take place on the 82nd day before the first Tuesday after the first Monday in June of each odd-numbered year, and shall apply to all candidates on the ballot in the elections held on that date.

(F) The sixth drawing under this subdivision shall take place on the 82nd day before the first Tuesday after the first Monday in November of the odd-numbered year, and shall apply to all candidates on the ballot in the elections held on that date.

(2) In the event there is to be an election of candidates to a special district, school district, charter city, or other local government body at the same time as one of the five major election dates specified in subparagraphs (A) to (F), inclusive, and the last possible day to file nomination papers for the local election would occur after the date of the drawing for the major election date, theprocedure set forth in Section 13113 shall apply.

(c) Each randomized alphabet drawing shall be open to the public. At least 10 days prior to a drawing, the Secretary of State shall notify the news media and other interested parties of the date, time, and place of the drawing. The president of each statewide association of local officials with responsibilities for conducting elections shall be invited by the Secretary of State to attend each drawing or send a representative. The state chairman of each qualified political party shall be invited to attend or send a representative in the case of drawings held to determine the order of candidates on the primary election ballot, the November generalelection ballot, or a special election ballot as provided for in subdivision (d).

Random samples

Reasoning like a classical probabilist, this means that if we have N objects in our population, the first selection assigns probability 1/N to each -- After the first item is identified, we have N-1 remaining, and select each with probability 1/(N-1)

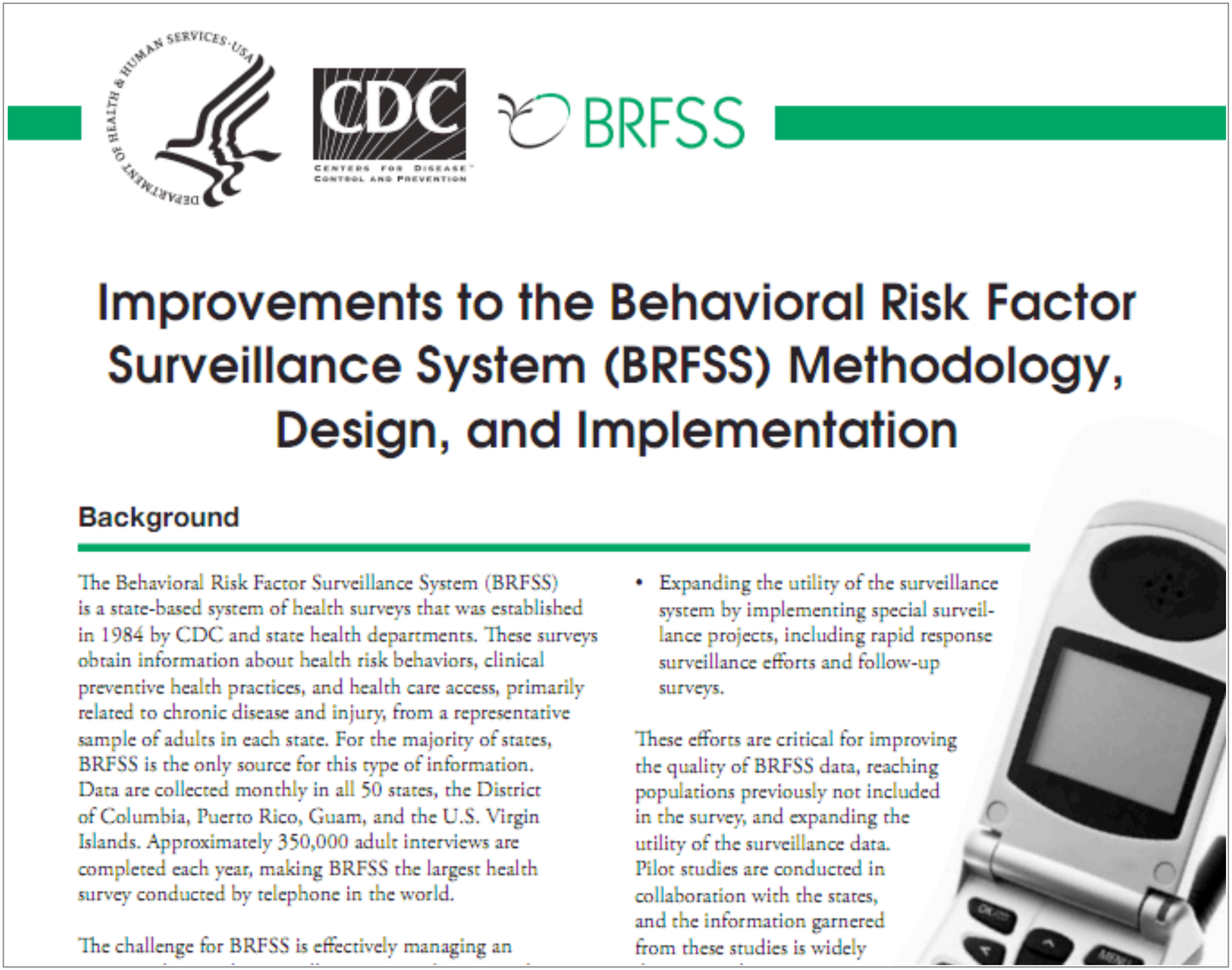
We have seen that the `sample()` command in R can be used to emulate this process using pseudo-random numbers

Random sampling

Sometimes **random selection is just this easy** -- At the right, we have 10 calls to sample in R, each producing a `sample()` of 10 items from the population, the numbers from 1 to 100

We have already seen examples where sampling is much harder -- **The CDC, for example, employs random digit dialing for the BRFSS** (and an expanding set of techniques) to try to sample randomly from the adult U.S. population

```
> sample(1:100,10)
 [1]  3 35 49 63 65 17 44 31 42 14
> sample(1:100,10)
 [1] 91 22 81 93 75 23 99 89 36 37
> sample(1:100,10)
 [1] 91 50 88 12 28 94 73 20 23 31
> sample(1:100,10)
 [1] 10 82 42 26 79 41 29 40 57  4
> sample(1:100,10)
 [1] 77 83  7 35 33 87 44 47  3 70
> sample(1:100,10)
 [1] 88  5 64 15 79 65 16 81 28 24
> sample(1:100,10)
 [1] 85 19 62 79 61 13 84 71 36 51
> sample(1:100,10)
 [1] 54 92 55  2 36 25 32 77 94 50
> sample(1:100,10)
 [1] 27 75 15 71 70 90 47 64 26 16
> sample(1:100,10)
 [1] 10  6 34  3 37 62 20 82 68 91
```



DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

BRFSS

Improvements to the Behavioral Risk Factor Surveillance System (BRFSS) Methodology, Design, and Implementation

Background

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based system of health surveys that was established in 1984 by CDC and state health departments. These surveys obtain information about health risk behaviors, clinical preventive health practices, and health care access, primarily related to chronic disease and injury, from a representative sample of adults in each state. For the majority of states, BRFSS is the only source for this type of information. Data are collected monthly in all 50 states, the District of Columbia, Puerto Rico, Guam, and the U.S. Virgin Islands. Approximately 350,000 adult interviews are completed each year, making BRFSS the largest health survey conducted by telephone in the world.

The challenge for BRFSS is effectively managing an

- Expanding the utility of the surveillance system by implementing special surveillance projects, including rapid response surveillance efforts and follow-up surveys.

These efforts are critical for improving the quality of BRFSS data, reaching populations previously not included in the survey, and expanding the utility of the surveillance data. Pilot studies are conducted in collaboration with the states, and the information garnered from these studies is widely

The CDC again

So let’s pick up there for the moment...

You looked at data from the BRFSS in lab and computed the BMI for the people in the sample and today we’ll focus on the average BMI -- In R, we compute this to be 26.3 (25 is the threshold for “Overweight”)

```
> source("http://www.stat.ucla.edu/~cocteau/stat13/data/cdc.R")
> dim(cdc)
[1] 20000    11
> bmi <- 703*(cdc$weight)/(cdc$height * cdc$height)
> mean(bmi)
[1] 26.30693
```

Our interest, of course, is not in the BMI of the 20,000 people in our sample, but instead what this number says about the average BMI computed over the adult population in the U.S. -- **What can we say?**

The frequentist approach

Recall that the frequentist view of probability relies on repeated trials -- **Probability emerges by looking at the long-run frequency** of events

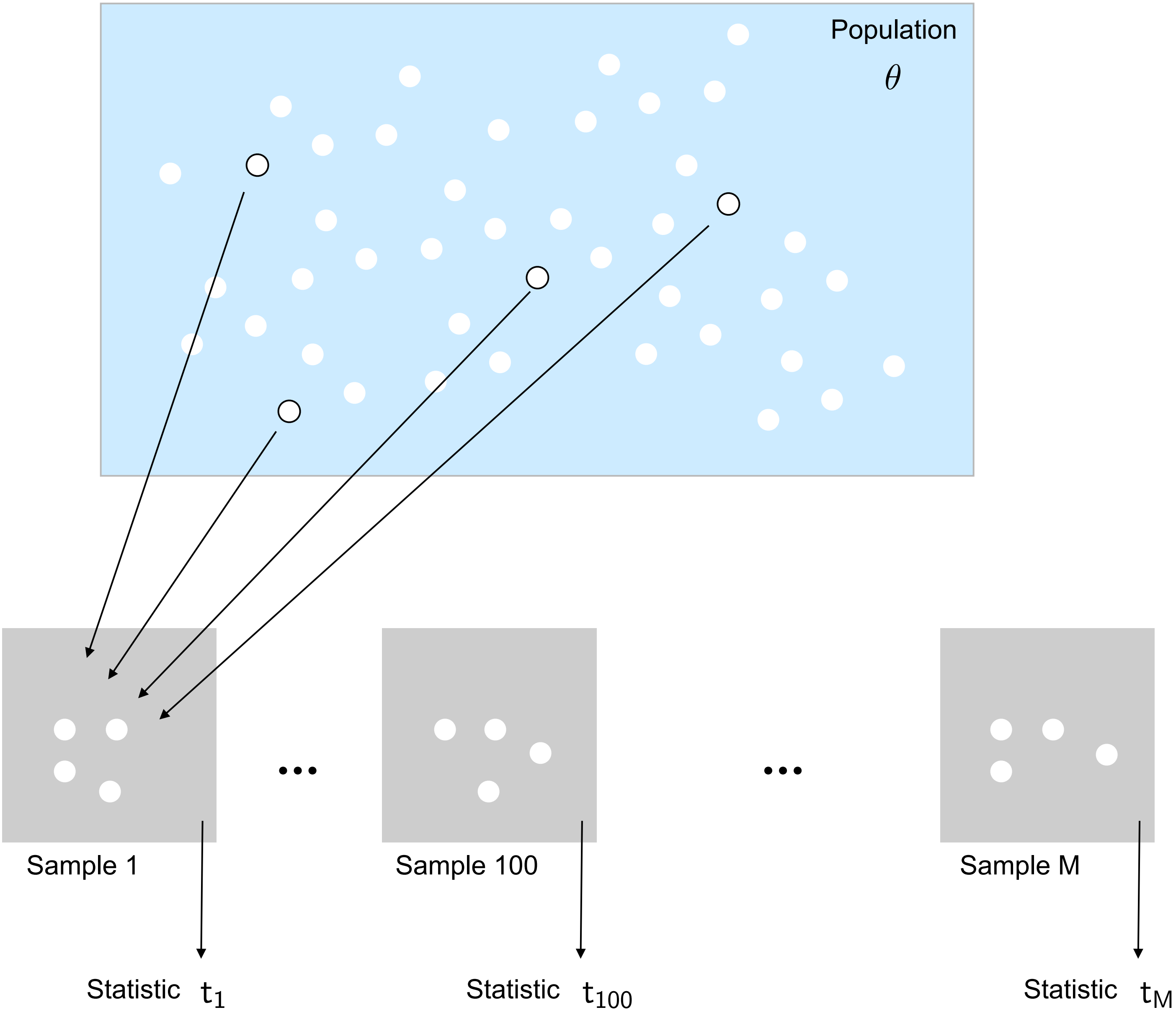
We saw this idea when we tested simple hypotheses -- The **null distribution** consisted of a range of **experimental outcomes that could have happened** under the null hypothesis

For estimation, we will again invoke this idea of repeated trials to come up with an indication of how accurate a statistic t is when estimating a population parameter θ

A repeatable process

Our approach is based on a fairly simple idea -- Rather than thinking about our specific sample and the associated statistic (the 20,000 people in our BRFSS data and their average BMI t=26.3), let’s **consider our experiment as a repeatable process**

So, we can imagine the CDC repeating its sampling, coming up with another set of 20,000 people and computing a different (almost certainly) statistic t -- In fact, we can imagine (imagining is cheap!) doing it 10 times, 100 times, 1000 times...


Population
θ
Sample 1
Sample 100
Sample M
...
...
Statistic t_1
Statistic t_{100}
Statistic t_M

Random samples

We know from the last two lectures how to count the number of samples we could draw from a given population -- That is, how many sets of size n from N elements we could form

$$M = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

For even small population sizes N, this number becomes enormous

```
> choose(100,10)

[1] 17310309456440
```

Yes, that’s 17 trillion possible samples of size 10 from a population of 100 objects!

The sampling distribution

The set of estimates $t_1, t_2, \ldots, t_M$ associated with forming all possible M samples from our population is known as **the sampling distribution** -- Call our particular estimate t_1 (the one we computed by conducting the CDC telephone surveys)

Admittedly, **this construction is a fantasy** -- First off, the number of samples M we can form from even modestly-sized populations is enormous and, second, we would never actually repeat our experiment in this way

But as a mental exercise, what does it buy us?

Bias

Consider, for example, the **center and spread of the sampling distribution,** the values $t_1, t_2, \ldots, t_M$ -- The center tells us whether or not our M estimates (again, each coming from a different sample of the population) are close to the population parameter we’re interested in --

If, for example, their average

$$\frac{1}{M}\sum_{i=1}^{M} t_i$$

is far from θ, the population parameter, we say that the estimate is **biased**

Standard error

The spread of the sampling distribution, the spread of the values $t_1, t_2, \ldots, t_M$, tells us about how our estimates change from sample to sample -- In most cases, we'll prefer having less rather than more variability when we repeat our experiments

One measure of spread that is used in this context is **the standard deviation of the values** $t_1, t_2, \ldots, t_M$ -- It is so important, actually, that it has a special name, and is called **the standard error** of our estimate

Root mean squared error

You will often see bias referred to as a measure of accuracy and the standard error as a measure of precision -- Given a single data set, our estimate might be t_1 far from the parameter we're after because of either effect

For example, t_1 may be far from θ because the sampling distribution is not centered on θ so that, on average (across all possible samples) our estimates are some distance from

It might also be far because the sampling distribution is wide -- A large spread means more variability from sample to sample

We can capture both effects with a quantity called the root mean squared error which is as much a sentence as it is a computational recipe

$$\sqrt{\frac{1}{M}\sum_{i=1}^{M}(t_i-\theta)^2}$$

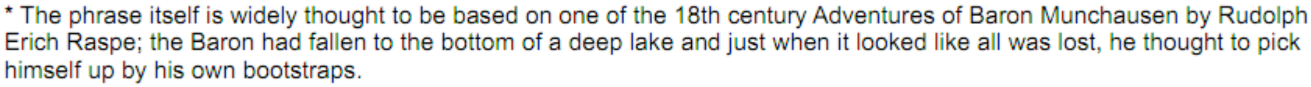
You can show with a little algebra that

$$\sqrt{\frac{1}{M}\sum_{i=1}^{M}(t_i-\theta)^2} = \sqrt{\mathrm{Bias}^2+\mathrm{SE}^2}$$

The bootstrap

We’ve been somewhat abstract for about 10 slides now, it’s time to get practical -- We are now going to try to have a look at the sampling distribution, but without calling anyone or collecting any new data

Our approach will be a fairly general methodology called **the bootstrap** -- The idea fits naturally with our strategy of **“analyze as you randomized”, which, in this case, means drawing more samples**

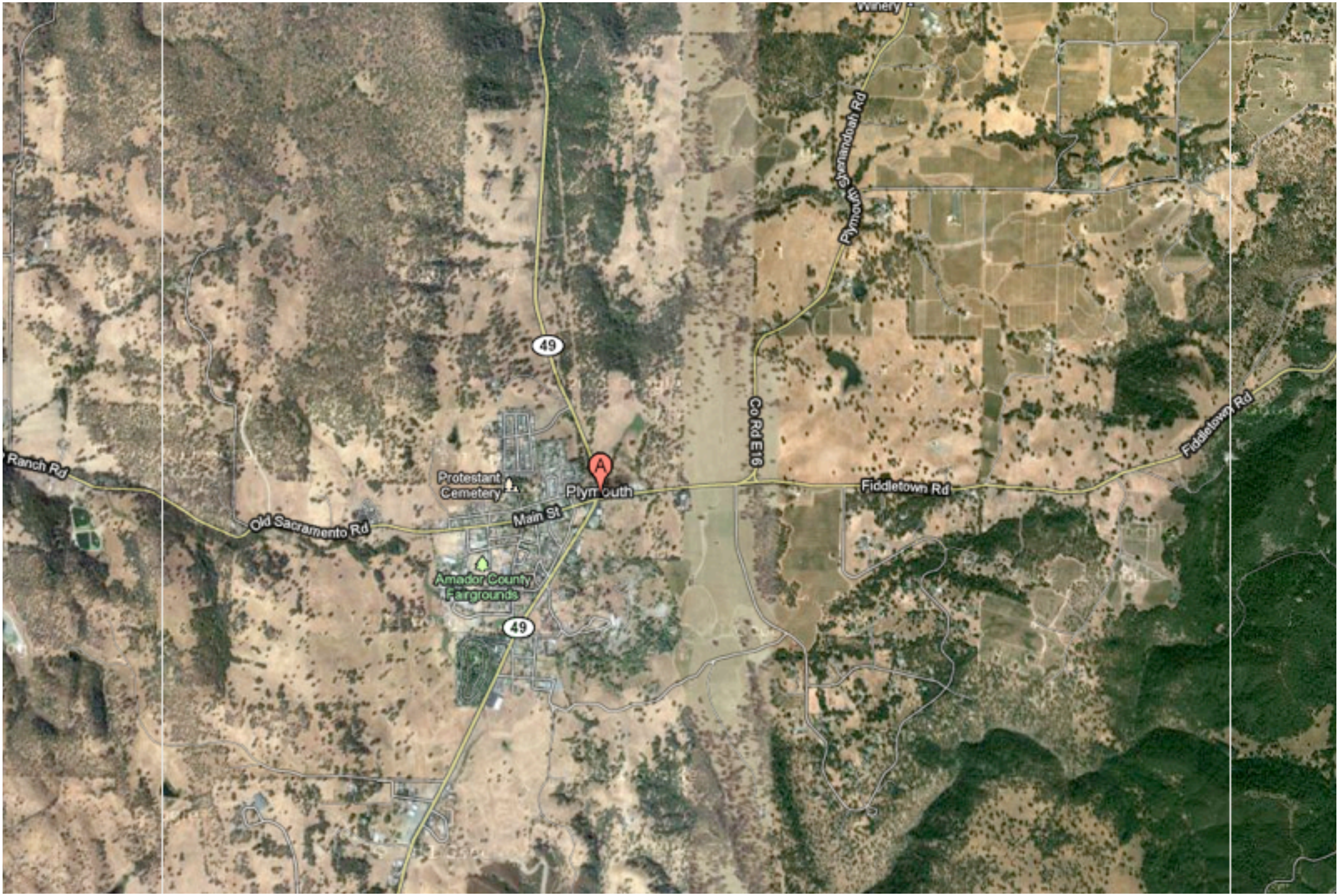


* The phrase itself is widely thought to be based on one of the 18th century Adventures of Baron Munchausen by Rudolph Erich Raspe; the Baron had fallen to the bottom of a deep lake and just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

The bootstrap world

Suppose we have **a “real world” population of 1,000 people** (say the population of Plymouth, CA) and we take a random sample of size 100 -- We form **a “bootstrap world” population** by cloning each of our 100 data points 10 times to construct a set of size 1,000

In general, if we have **a population of size N** from which we draw **a sample of size n**, then we can create a bootstrap world population by **copying each of the n items N/n times** (assuming it divides evenly, but don’t worry about this now)

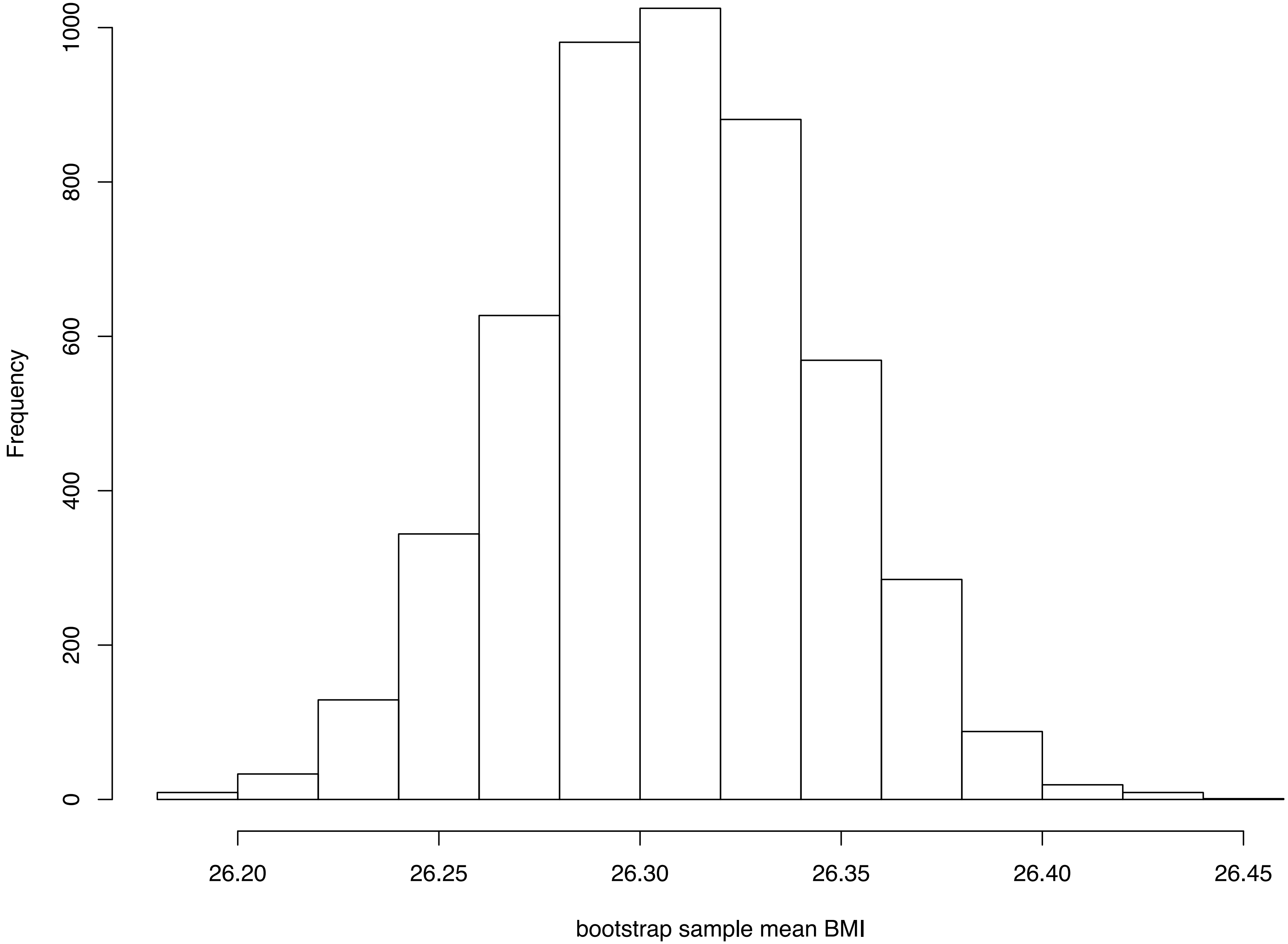

Winery
Plymouth Shenandoah Rd
49
Co Rd E16
Fiddletown Rd
Ranch Rd
Protestant Cemetery
A
Plymouth
Fiddletown Rd
Main St
Old Sacramento Rd
Amador County Fairgrounds
49

The bootstrap world

This new bootstrap world population is completely known to us and we can sample from it as often as we like -- It’s still not practical to form all samples of size n from this new collection and so instead we look at a few thousand random samples and examine the distribution of the associated estimates

Let’s see what that means for **our CDC data** -- On the next page, we have formed **5,000 samples from the bootstrap world population** and for each we have computed **the mean BMI of the people in the sample**

To get some notation here, we’ll let $t_1^*, \ldots, t_{5000}^*$ denote our 5,000 mean BMI numbers computed from each of the 5,000 bootstrap samples -- we call these **bootstrap replicates**

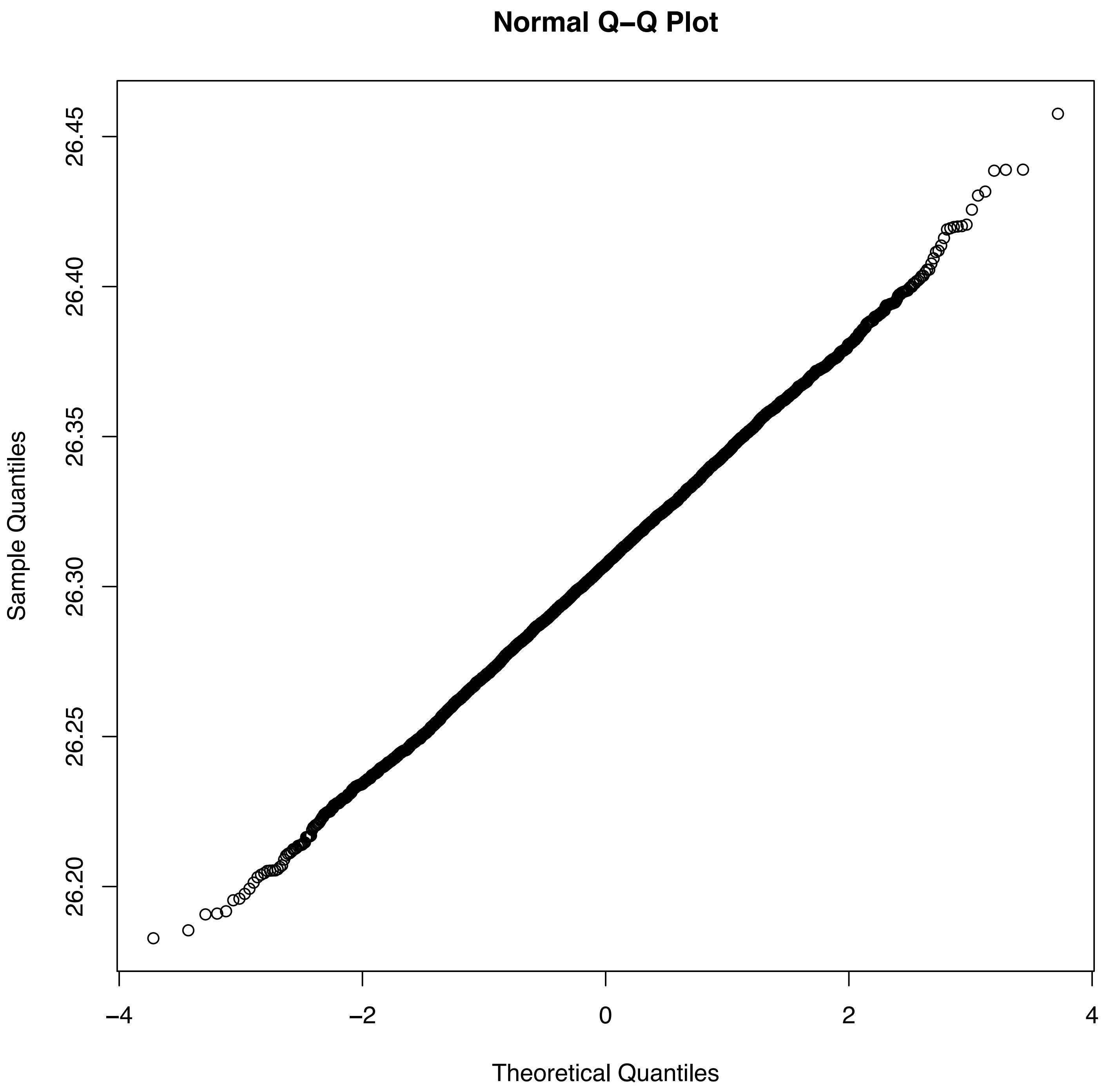


Histogram of 5,000 bootstrap replicates of the mean BMI

The bootstrap

What we are looking at on the previous page is **an approximation of the sampling distribution** associated with the estimate of mean BMI -- Again, the sampling distribution describes what would happen if we repeated our experiment many times, drawing new samples from the real world population each time

This distribution looks shockingly “normal”, a fact we’ll come back to...

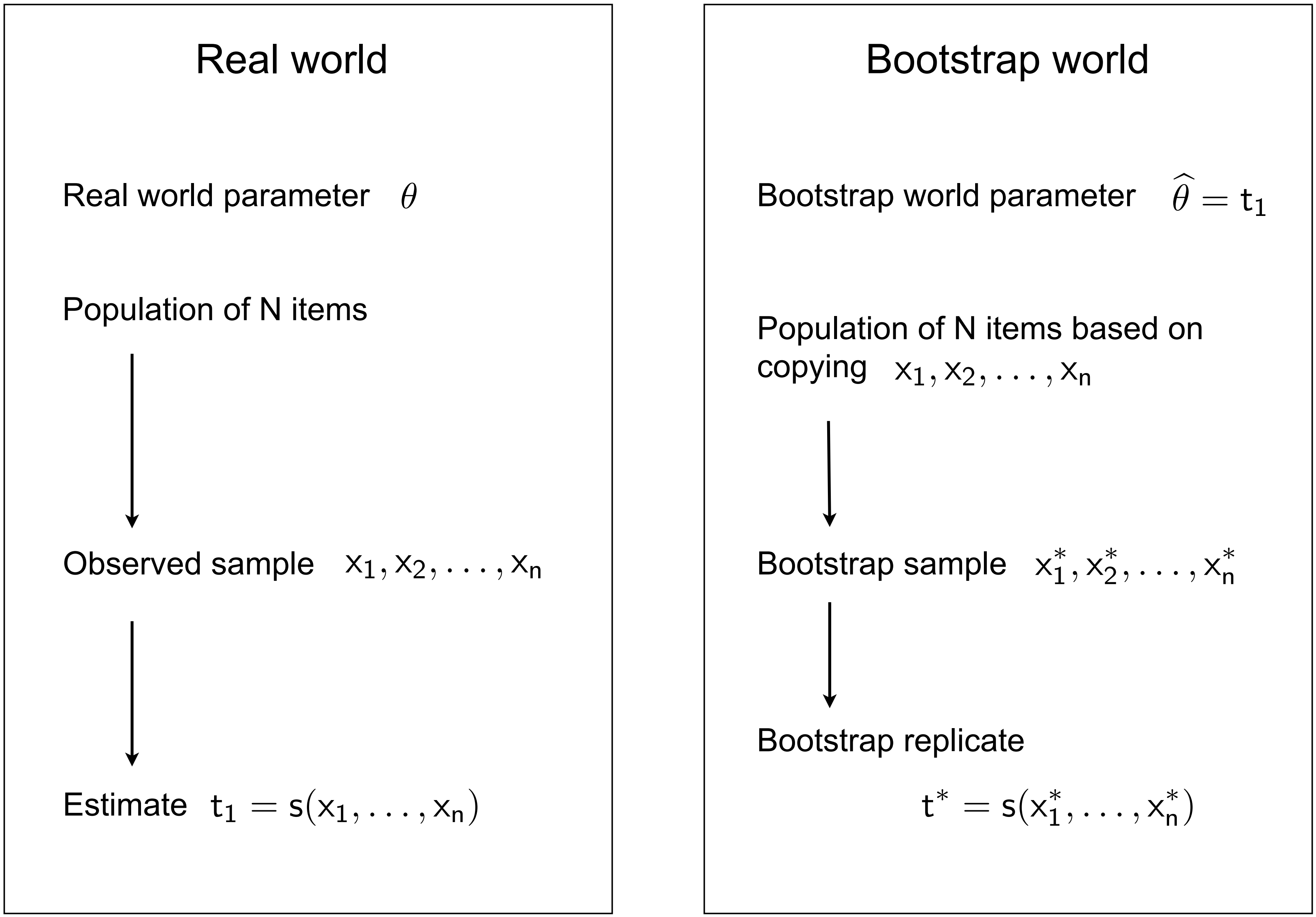

Normal Q–Q Plot
Sample Quantiles
26.20
26.25
26.30
26.35
26.40
26.45
Theoretical Quantiles
−4
−2
0
2
4

A general principle

Here is a simple sketch of the bootstrap procedure -- It falls under our general strategy of **analyze as you randomized**

In terms of notation, we have been using t_1 to refer to the estimate computed from our actual sample -- From now on, in sympathy with the population parameter θ we are going to let the symbol $\hat{\theta}$ refer to t_1

We will also use the notation s() to represent the computation of our estimate from a given set of data -- In R you can think of it like `mean()` or `median()` say



Real world

Real world parameter θ

Population of N items

Observed sample $x_1, x_2, \ldots, x_n$

Estimate $t_1 = s(x_1, \ldots, x_n)$

Bootstrap world

Bootstrap world parameter $\widehat{\theta} = t_1$

Population of N items based on copying $x_1, x_2, \ldots, x_n$

Bootstrap sample $x_1^*, x_2^*, \ldots, x_n^*$

Bootstrap replicate

$t^* = s(x_1^*, \ldots, x_n^*)$

The bootstrap

The bootstrap distribution (the distribution of the bootstrap replicates) is an **approximation to the sampling distribution** of the statistic we’re interested in -- It is an approximation in the following senses:

Shape: The bootstrap distribution approximates the shape of the sampling distribution, allowing you to check normality

Center: In most cases, the bootstrap distribution will be centered on the estimate from the original $\hat{\theta}$ sample -- If it is not, we have evidence that our estimate is biased

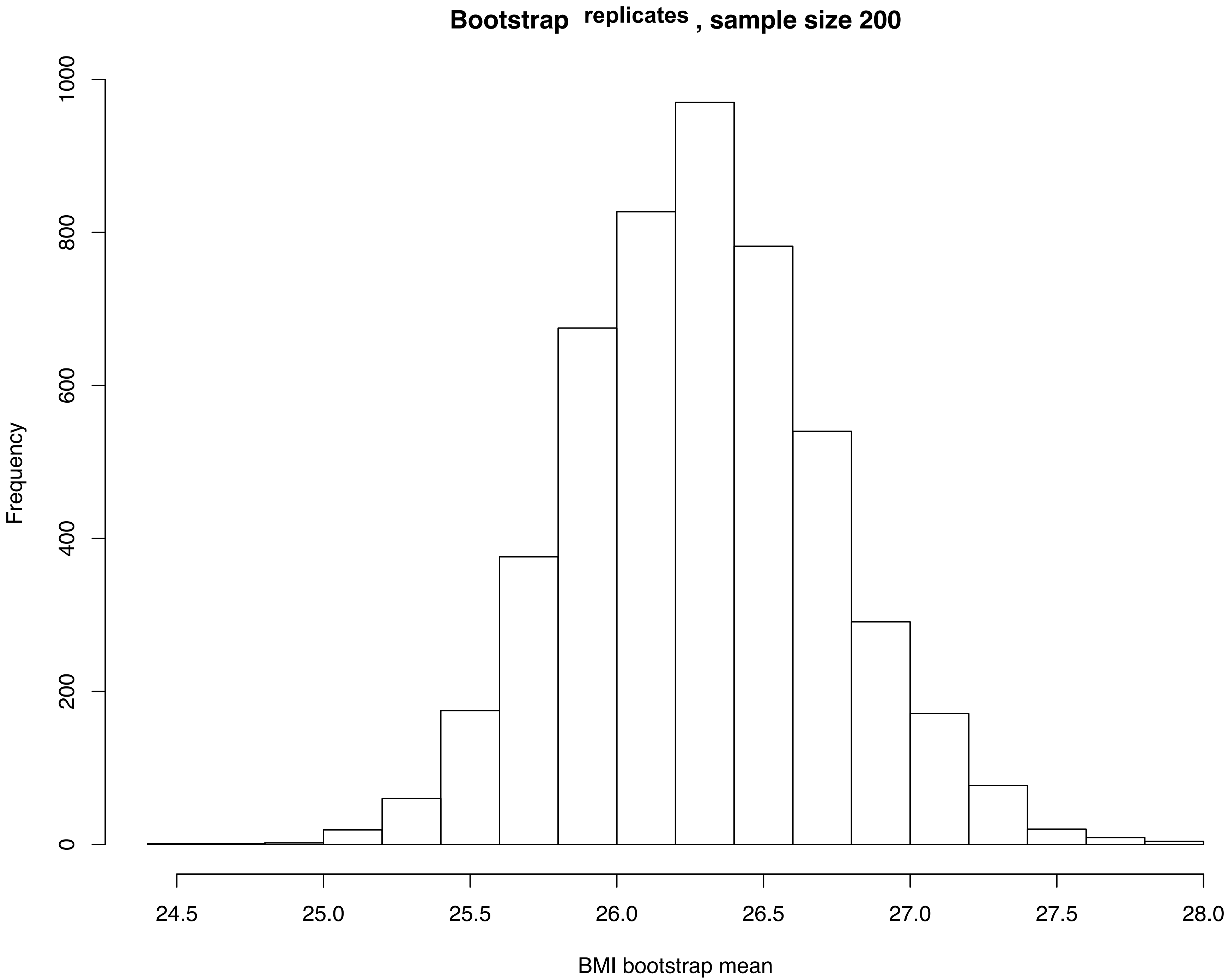
Spread: We can estimate the standard error of by $\hat{\theta}$ the standard deviation of the bootstrap distribution

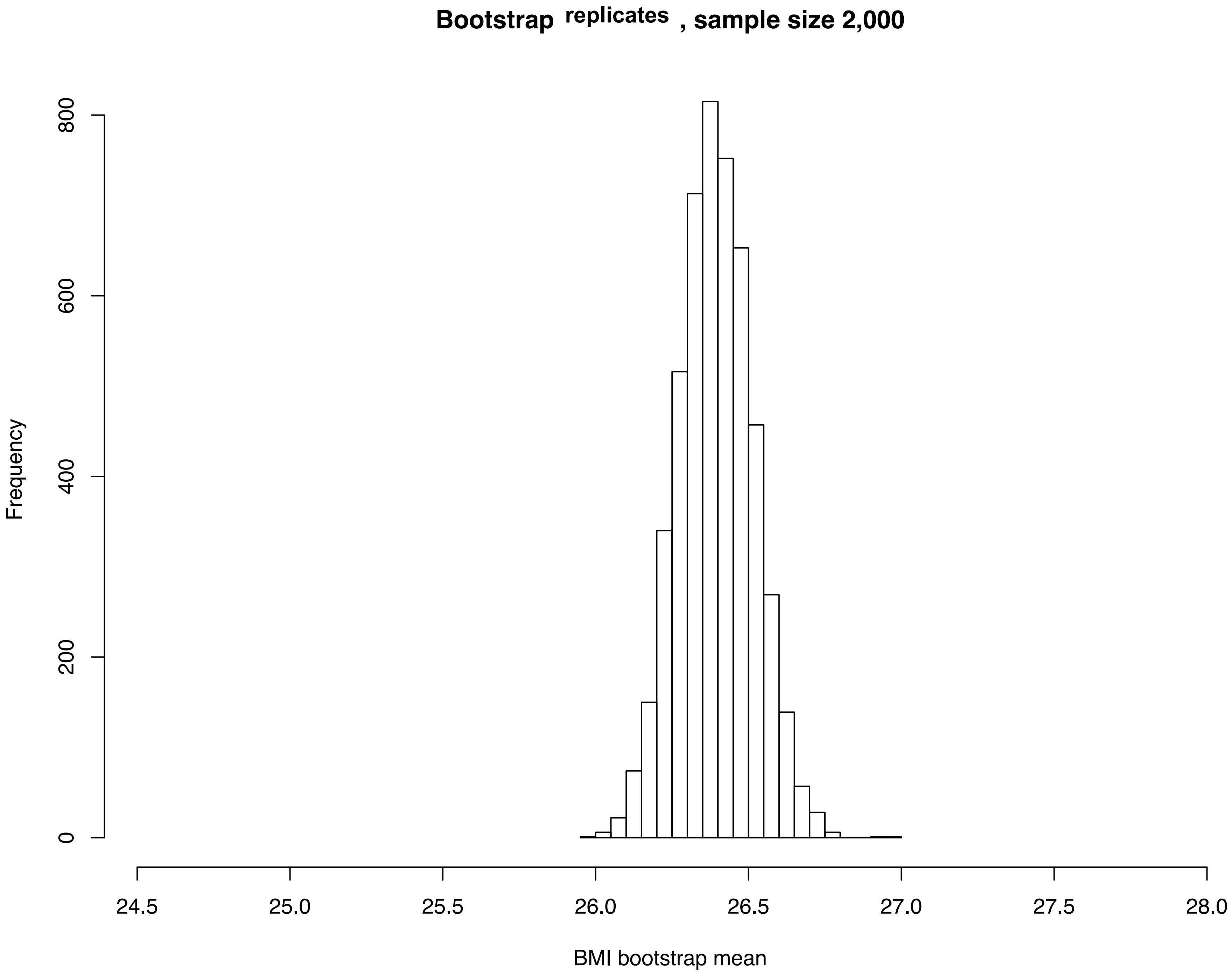
The bootstrap

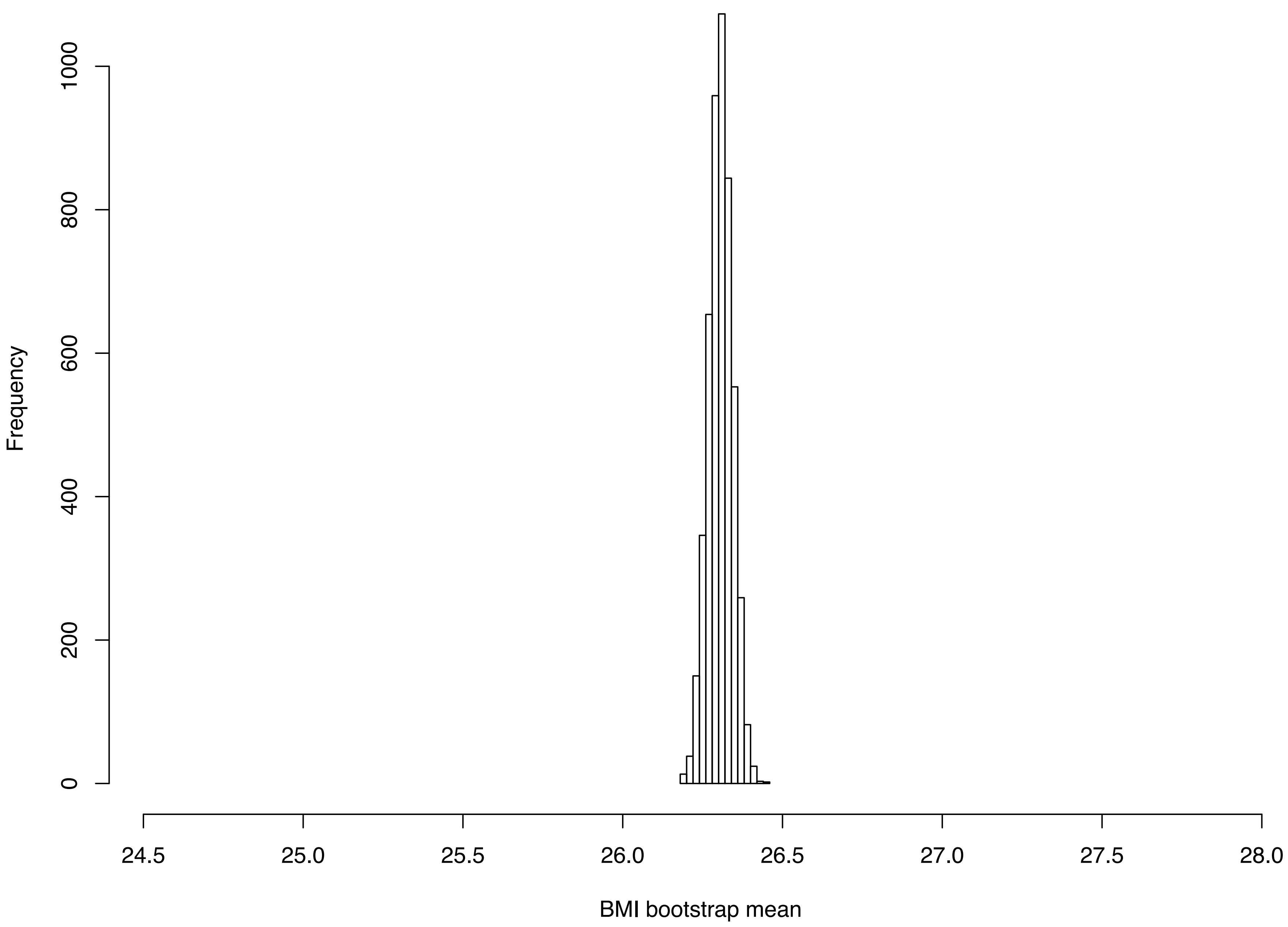
To see broadly what's going on here, suppose instead of 20,000 people, the CDC talked to 2,000 or even just 200 -- What do you expect would happen to the sampling distribution?

On the next two pages we took a subsample of the CDC data set, one of 2,000 and one of 200 -- **By ignoring the remaining data, we can pretend that the CDC only conducted surveys of 2,000 and 200 people respectively**

Then, for example, in the case of a survey of size 200, we create the bootstrap world population and construct 5,000 bootstrap samples, each of size 200 -- We then look at the average BMI for each bootstrap sample and form a histogram of the bootstrap replicates


Bootstrap replicates, sample size 200
Frequency
0
200
400
600
800
1000
24.5
25.0
25.5
26.0
26.5
27.0
27.5
28.0
BMI bootstrap mean


Bootstrap replicates , sample size 2,000
Frequency
0
200
400
600
800
24.5
25.0
25.5
26.0
26.5
27.0
27.5
28.0
BMI bootstrap mean



Bootstrap replicates|, sample size 20,000

Standard errors

What we see here is somewhat intuitive -- If the bootstrap replicates are doing a good job of approximating the sampling distribution, then our standard error is dropping as the CDC talks to more people

This makes intuitive sense in that **as we collect more data** in our original sample, we ought to be doing **a better job of estimating** the population parameter , in this case the aθverage BMI of the U.S. adult population

Bias

In addition to considering the precision of our estimates, we can also assess any bias that might be present -- Using our comparison between the bootstrap and real worlds, we can estimate the true bias

$$\frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}} \mathrm{t}_\mathrm{i} - \theta$$

using the bootstrap replicates (assuming 5,000 bootstrap iterations)

$$\frac{1}{5,000}\sum_{\mathrm{i}=1}^{5{,}000} \mathrm{t}_\mathrm{i}^{*} - \widehat{\theta}$$

In the original CDC survey (with 20,000 respondents), our mean BMI was 26.30615 and the mean of our 5,000 bootstrap replicates was 26.30693, suggesting almost no bias

Confidence intervals

The idea should be getting clear -- If there is some function of the sampling distribution that we would like to estimate, we can use the bootstrap replicates as if they were actually from a repeated experiment

While bias and standard error are extremely informative, the main use for a sampling distribution is the construction of a **confidence interval for our population parameter**

A confidence interval is a (frequentist) expression of the uncertainty we have about the population parameter θ -- Rather than report a single estimate $\hat{\theta}$, we provide an **interval of “plausible” values for** θ

It again starts with the idea of a repeatable process...

Confidence intervals

Suppose we knew the sampling distribution exactly -- That is, somehow we were given estimates associated with all M samples that we could form from our population

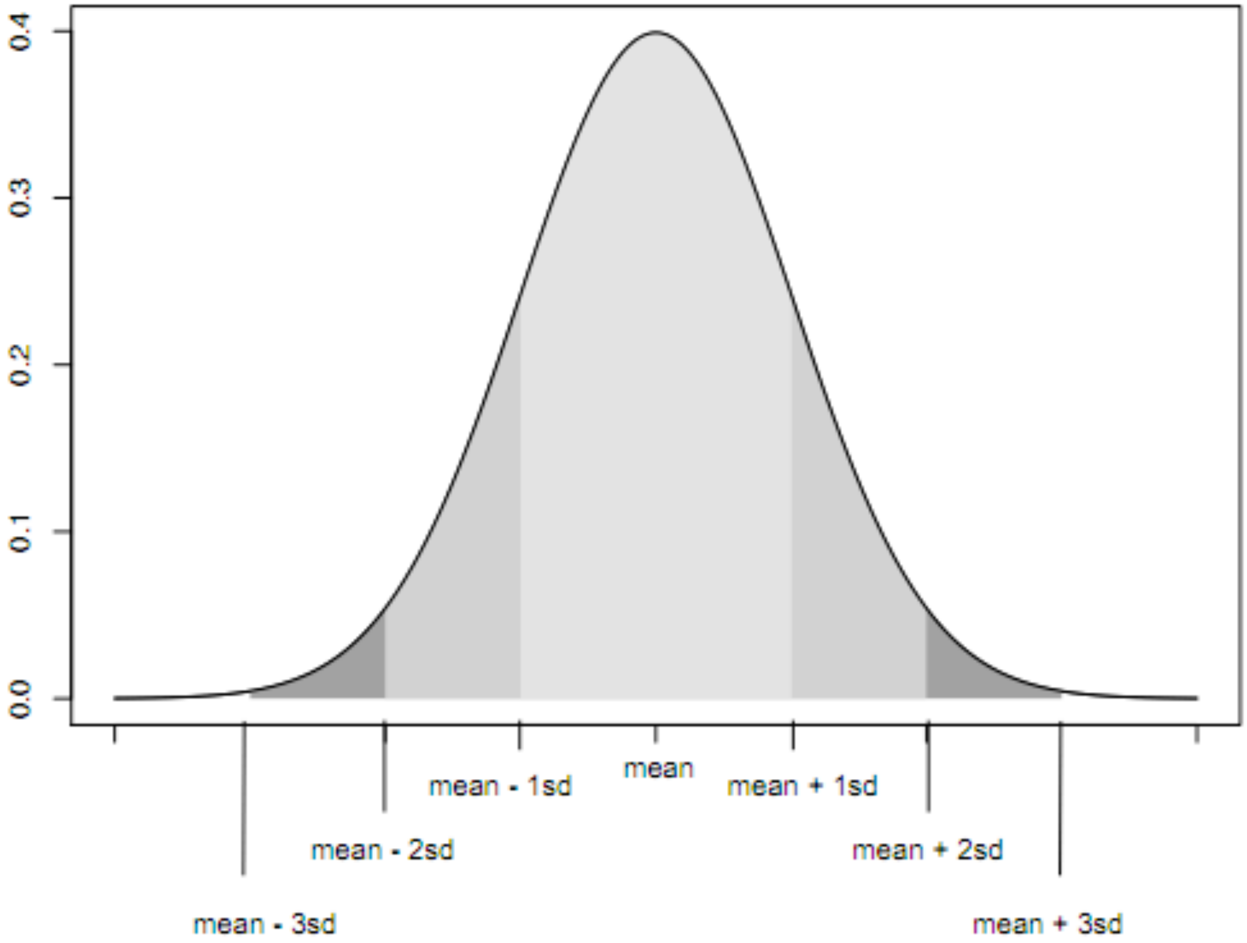
A 90% confidence interval, say, is constructed by a rule that ensures that, if we used this rule for every one of the M estimates, **90% of them would contain the population parameter** θ

This is our notion of confidence -- When we actually draw a random sample from the population and use the rule to compute an interval, we are hoping that our sample is one of the 90% of M for which the resulting interval contains the the population parameter θ

A simple example

If our sampling distribution looks normal (many, but not all, of them do), then we can come up with a simple rule for a 95% confidence interval -- Recall that for any normal distribution, 95% of the mass must be within two standard deviations of the mean

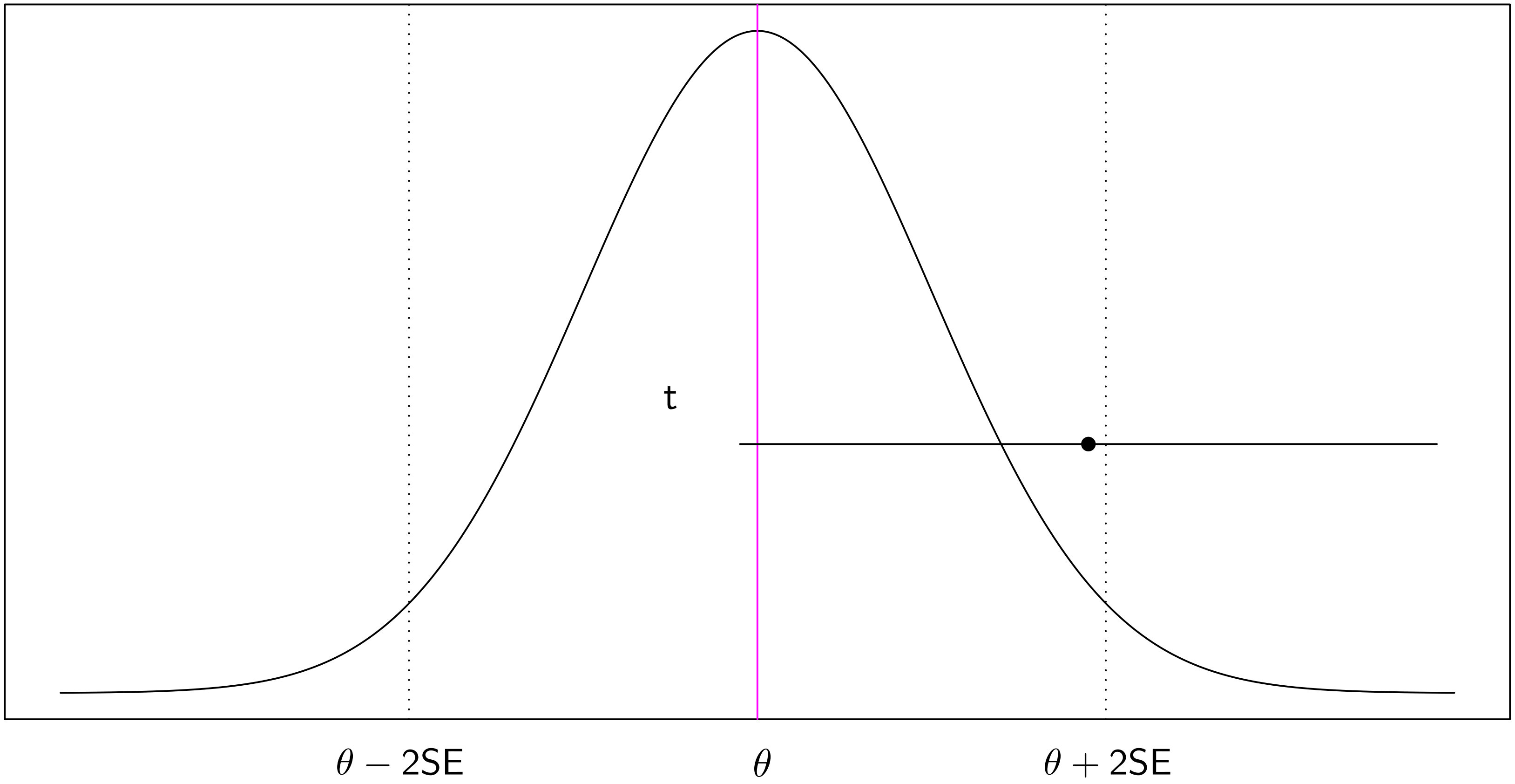
We had this picture from last lecture...

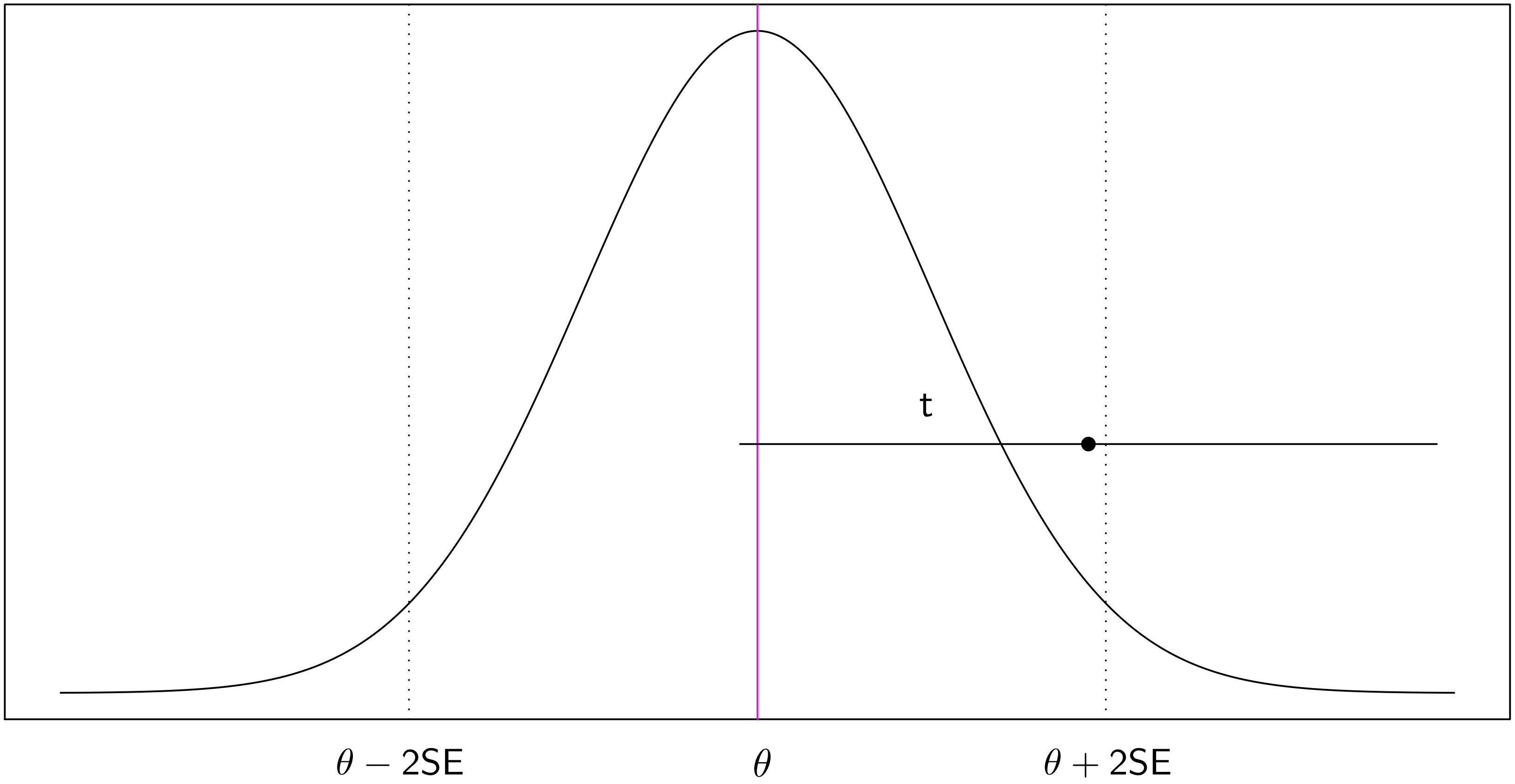

0.4
0.3
0.2
0.1
0.0
mean
mean - 1sd
mean + 1sd
mean - 2sd
mean + 2sd
mean - 3sd
mean + 3sd

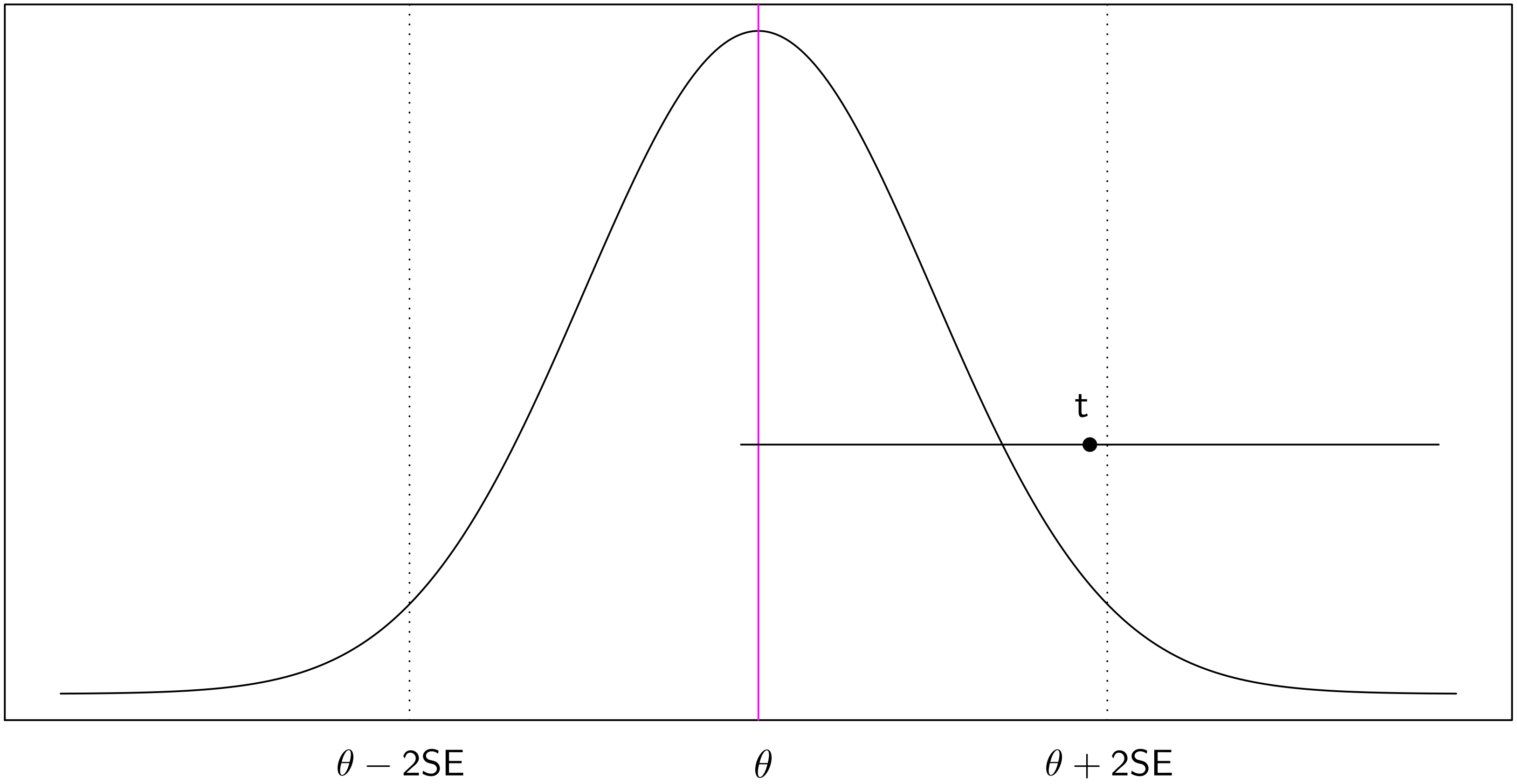
A simple rule

If our sampling distribution is centered on the population parameter θ, this rule implies that 95% of the samples we could take have values that are within two standard errors of

Here’s a simple rule: For any estimate, form the interval $t \pm 2SE$. Let’s see how this might perform


t
$\theta - 2\text{SE}$
θ
$\theta + 2\text{SE}$


t
$\theta - 2\text{SE}$
θ
$\theta + 2\text{SE}$


t
$\theta - 2\text{SE}$
θ
$\theta + 2\text{SE}$

A simple rule

The curves on the previous slides represent what happens when we consider all M possible samples we can from a population of size N -- Each t represents one of M possible outcomes and **95% of them are associated with intervals that contain our population parameter**

The notion of confidence is important to keep in mind -- When we compute an interval, **we don't know if it contains the truth or not**, we just know that **95% of the intervals we could compute do contain the truth**

A simple rule

We can use the bootstrap to tell us about the bias (whether the sampling distribution is centered on θ) and we can estimate the standard error -- That gives us all we need to compute these simple intervals

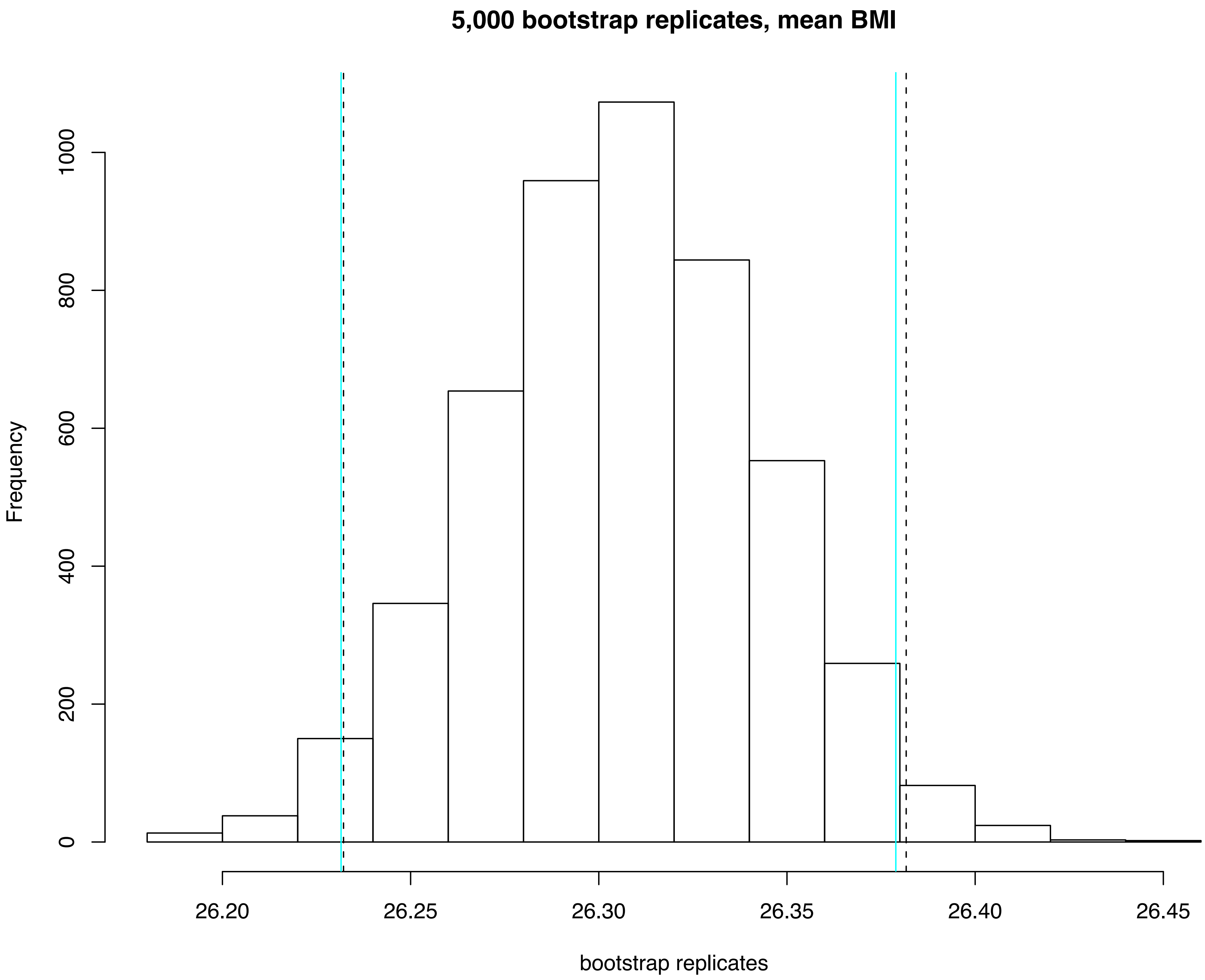
For the CDC data set with 20,000 respondents, the standard error is estimated to be 0.037 and the mean BMI was, again, 26.307 so that a 95% confidence interval is [26.233, 26.381]

A more general rule

Notice that in the normal case we are simply taking the 0.025 lower and 0.975 upper quantile of the sampling distribution -- In general, we can use our bootstrap replicates and form an interval using the 0.025 and 0.952 quantiles of the values $t^*_1, \ldots, t^*_{5,000}$

This is called the percentile bootstrap confidence interval and is pretty easy to work with -- It is intuitive and will work reasonably well even if there your bootstrap distribution suggests things are skewed

For the CDC data, because the bootstrap distribution looks fairly normal, these two approaches are about the same (black and dashed = +/- 2SE and cyan = quantiles)


5,000 bootstrap replicates, mean BMI
Frequency
0
200
400
600
800
1000
26.20
26.25
26.30
26.35
26.40
26.45
bootstrap replicates

The bootstrap

We will see that this relatively easy mechanism will provide us with the ability to estimate the bias, standard errors, RMS and confidence intervals in a wide range of problems

We've looked at the mean so far, but we could have as easily looked at the median, a trimmed mean or even the relative risk in any of our Vioxx trials!

Over the next lecture or two we'll refine these ideas and extend them to other estimation contexts