Lecture 11: Plug-in

Last time

We examined how to approach sampling distributions mathematically -- We started with situations in which we could determine them exactly, often starting by identifying a so-called pivotal quantity

We then considered approximations to the sampling distributions that leveraged the Central Limit Effect, the tendency of averages of independent random variables to tend to a normal distribution -- We introduced the notion of convergence in distribution and provided enough tools to motivate the so-called plug-in principle

We finished by discussing properties of Maximum Likelihood Estimates...

Parametric problems

Under our parametric framework, we are given a family of probability distributions $f(x|\theta)$ indexed by a parameter θ

We are then given n independent observations $X_1, \ldots, X_n$ from one of these densities, say, $f(x|\theta^*)$ -- Our task was to form an estimate $\widehat{\theta}_n$ of θ^* and to make some comment on its accuracy and precision

Maximum likelihood

Recall the likelihood and log-likelihood functions

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(X_i|\theta) \quad \text{and} \quad l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

If we let $\widehat{\theta}_n$ be the MLE (and there are technical conditions we won't fuss about now) then it has the following properties

1. The MLE $\widehat{\theta}_n$ is a consistent estimate of θ^* -- The errors $\widehat{\theta}_n - \theta^*$ get smaller as you collect more and more data or, rather, n gets large

2. The MLE is equivariant -- If $\widehat{\theta}_n$ is the MLE for θ^*, then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta^*)$

3. The MLE is asymptotically normal -- Its mean is θ^* and its standard deviation is $1/\sqrt{nI(\theta^*)}$ (where I is the expected Fisher information)

4. The MLE is asymptotically efficient

Fisher information

In a previous lecture, we discussed the so-called observed Fisher information -- It is based on the second derivative of the log-likelihood function

$$-\mathrm{I}''(\theta) = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial^2}{\partial^2 \theta} \log \mathrm{f}(\mathrm{X}_\mathrm{i}|\theta)$$

We evaluated this expression at $\widehat{\theta}_\mathrm{n}$ to determine the peakiness of the log-likelihood and found that that peakiness increased with sample size -- In short, we had more "information" about a parameter as we collected more data

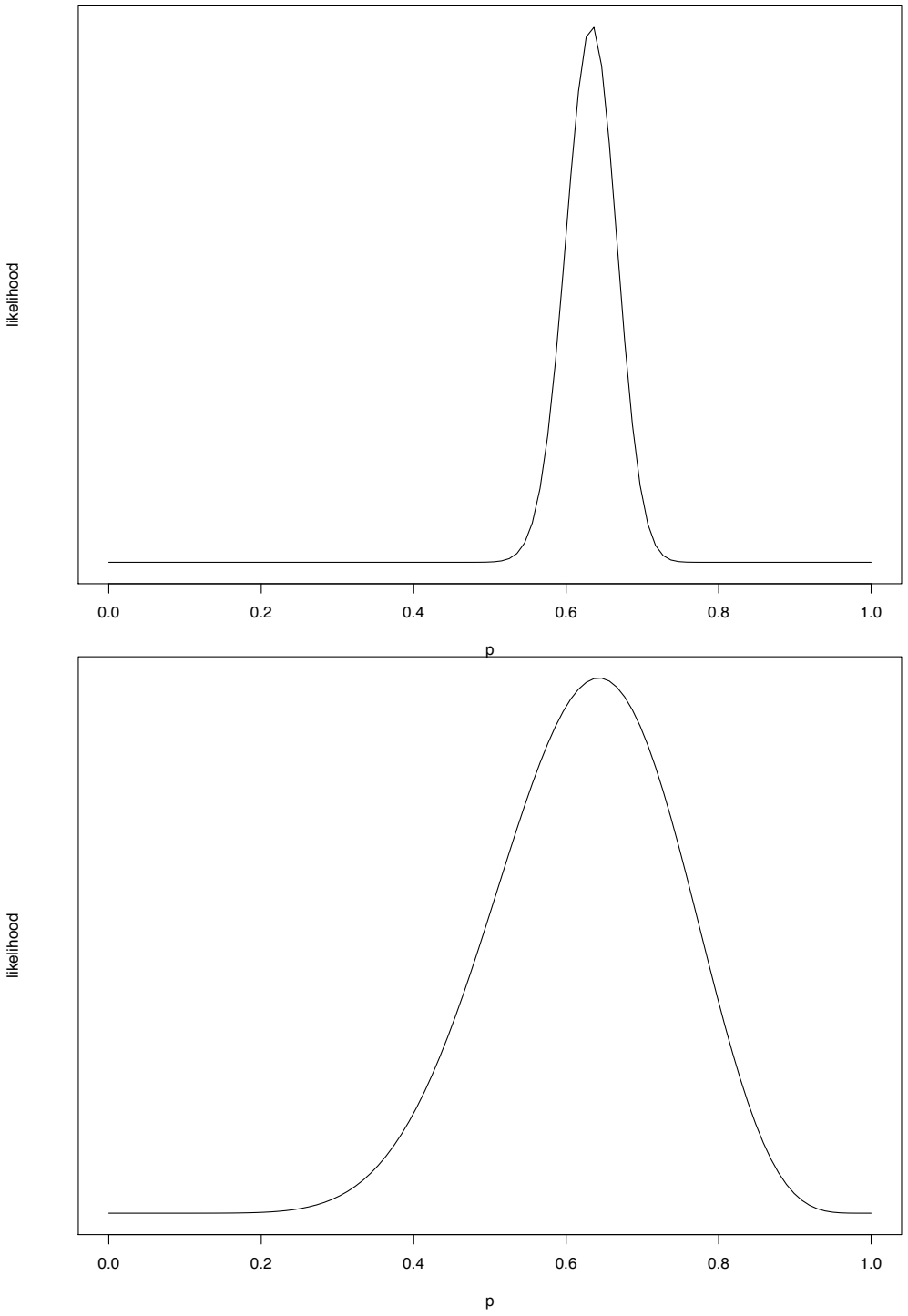
Recall our binomial example

At the top right we have the likelihood associated with our n=15 observations from binomial (14,p)

Below we have the likelihood associated with just a single of our 15 observations, $X_5 = 9$

So the top plot represents a “learning” problem when we have n=15 observations and the bottom is the same problem but using just one observation

What do you notice?



Fisher information

The observed Fisher information is a random quantity in the sense that if we repeat our experiment, we will get new data $X_1, \ldots, X_n$ and a new value of the observed information

Notice, however, that it is also a sum of independent, identically distributed random variables -- Therefore, we might expect its average to converge to something

$$-\frac{1}{n} l''(\widehat{\theta}_n) = -\frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial^2}{\partial^2 \theta} \log f(X_i|\theta) \right]_{\theta=\widehat{\theta}_n}$$

Fisher information

We define the so-called observed Fisher information as an expected value, written here as a function of θ

$$\mathrm{I}(\theta) = -\mathsf{E}_\theta\left[\frac{\partial^2}{\partial^2\theta}\log \mathsf{f}(\mathsf{X}|\theta)\right] = -\int\left[\frac{\partial^2}{\partial^2\theta}\log \mathsf{f}(\mathsf{X}|\theta)\right]\mathsf{f}(\mathsf{x}|\theta)\mathsf{dx}$$

The limit we seek is $\mathrm{I}(\theta^*)$

You can think of $\mathrm{I}(\theta^*)$ as the information associated with a single observation X and then $\mathsf{n}\mathrm{I}(\theta^*)$ as the information in n observations $\mathsf{X}_1, \ldots, \mathsf{X}_\mathsf{n}$

Convergence in distribution: The MLE

From our calculations last time, under certain regularity conditions and for large n, the MLE $\widehat{\theta}_n$ is asymptotically normal with mean θ^* and standard deviation $1/\sqrt{nI(\theta^*)}$ or in more familiar terms

$$\frac{\widehat{\theta}_n - \theta^*}{1/\sqrt{nI(\theta^*)}}$$

has a standard normal distribution for large n

Now, following the plug-in principle (applying Slutsky's theorem), we can substitute the estimate $\widehat{\theta}_n$ for θ^* in the standard deviation to come up with approximate 95% confidence intervals of the form

$$\widehat{\theta}_n \pm \frac{2}{\sqrt{nI(\widehat{\theta}_n)}}$$

Binomial

Assume we have a single observation (n=1) X from a binomial distribution with m and p^* -- The likelihood and log-likelihood functions are simply

$$\mathcal{L}(p) = \binom{m}{X} p^X (1-p)^{m-X} \quad \text{and} \quad l(p) = X \log p + (m-X) \log(1-p) + \log \binom{m}{X}$$

The first and second derivatives of the log-likelihood are

$$l'(p) = \frac{X}{p} - \frac{m-X}{1-p} \quad \text{and} \quad l''(p) = -\frac{X}{p^2} - \frac{m-X}{(1-p)^2}$$

Binomial

This means that the MLE is just $\widehat{p} = X/m$, and substituting in for the second derivative we find the observed Fisher information is just

$$\frac{X}{\widehat{p}^2} + \frac{m - X}{(1 - \widehat{p})^2} = \frac{m\widehat{p}}{\widehat{p}^2} + \frac{m - m\widehat{p}}{(1 - \widehat{p})^2} = \frac{m}{\widehat{p}(1 - \widehat{p})}$$

What can you say about this?

Binomial

The expected information follows the same path, except we use the fact that the expected value of X is mp^* -- In short, we get

$$\frac{m}{p^*(1-p^*)}$$

Plugging in $\hat{p}$ (and using the fact that n=1 in this case), we have an approximate 95% confidence interval of the form

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}$$

Aside

The binomial is a little slippery -- Here we are taking n=1 observations from a binomial with m trials (m independent coin tosses)

We could have just as easily viewed this as m observations from the Bernoulli distribution (a single coin toss) -- In that case, the expected Fisher information would be of the form

$$\frac{1}{p^*(1-p^*)}$$

and we would apply the MLE confidence interval formula from a few slides back with n=m

Binomial

Admittedly, this is a VERY long way to go to derive that result, and we could have done it directly (as you are doing in homework) reasoning from $\widehat{p} = X/m$ and using the known distribution for X

The point is that there are situations in which we cannot just write down the distribution for X but, instead, have to rely on formula like those on the previous few slides

A bit more

In point of fact, the asymptotic variance is about as good as you can get -- That is, any unbiased estimate $\widetilde{\theta}_n$ of θ^* based on the n observations $X_1, \ldots, X_n$ has variance (square of its standard error) bounded from below by

$$\text{var}\,\widetilde{\theta}_n \geq \frac{1}{nI(\theta^*)}$$

This means that at least for large n, the MLE is about as good as you can get in terms of efficiency

Our approach

I present this material mainly for pedagogical reasons -- This way you see how confidence intervals are derived analytically, pushing through various limit theorems to establish “large n” approximate results

Instead of dealing in formulae, we will rely on R or some other bootstrap-enlightened software package to provide us with ready assessments of precision or confidence intervals computationally

For the most part, when a formula exists, the bootstrap will agree with it, making it (perhaps) a more general tool for you as you venture out into the world...

Real world

Real world parameter θ^*

Sample n times from $f(x|\theta^*)$

↓

Observed sample $X_1, \ldots, X_n$

↓

Estimate $\widehat{\theta} = s(X_1, \ldots, X_n)$

Bootstrap world

Bootstrap world parameter $\widehat{\theta}$

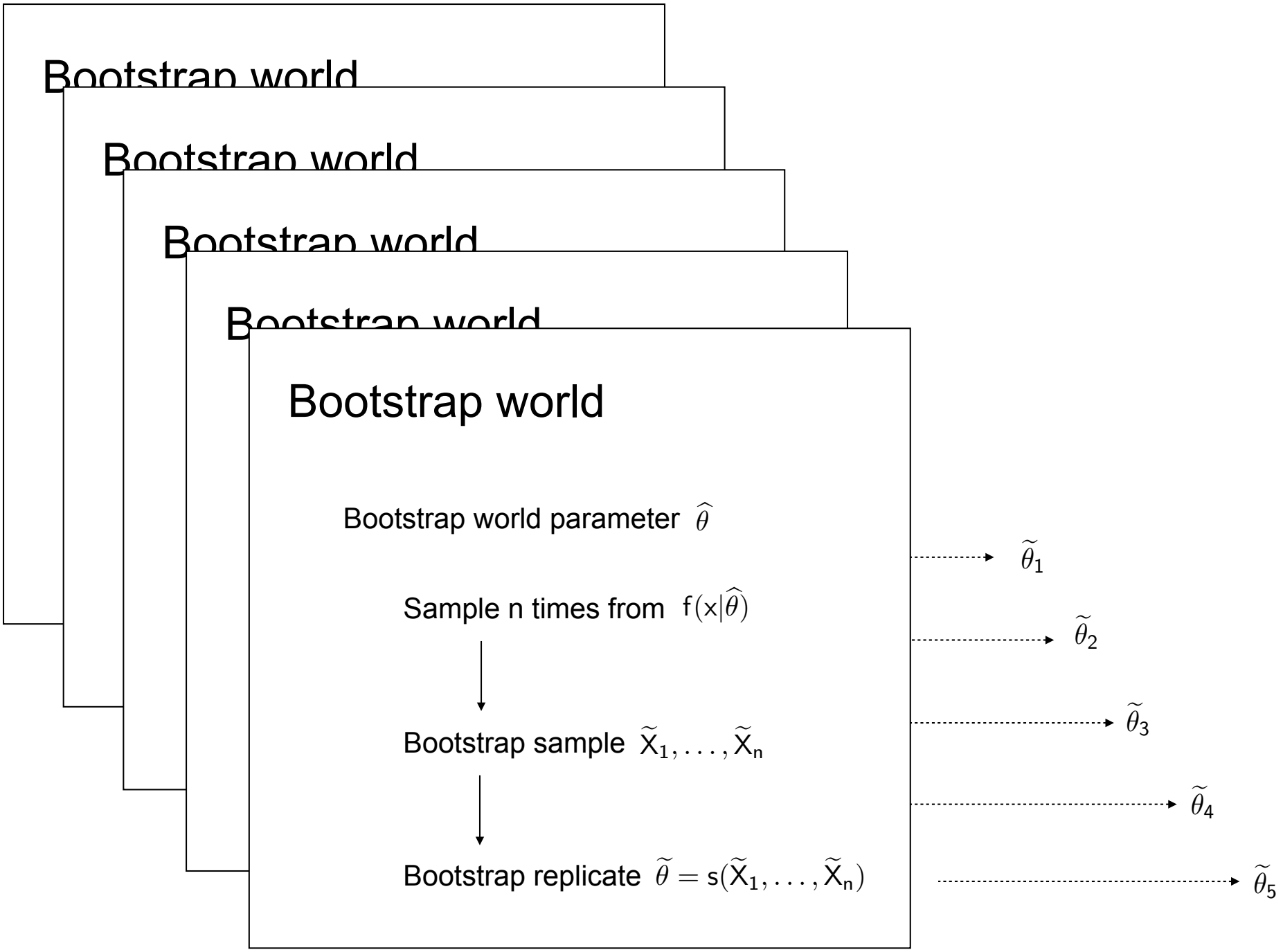
Sample n times from $f(x|\widehat{\theta})$

↓

Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$

↓

Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$


Bootstrap world
Bootstrap world
Bootstrap world
Bootstrap world
Bootstrap world
Bootstrap world parameter $\widehat{\theta}$
Sample n times from $f(x|\widehat{\theta})$
Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$
Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$
$\widetilde{\theta}_1$
$\widetilde{\theta}_2$
$\widetilde{\theta}_3$
$\widetilde{\theta}_4$
$\widetilde{\theta}_5$

The bootstrap

If we repeat this process B times, we form B bootstrap replicates from which we can estimate the sampling distribution of $\widehat{\theta}$ -- Plotting these B values (a histogram, say) gives us information about the performance of our estimator

The bootstrap

Bias: Let's let $\bar{\tilde{\theta}}$ (horrible notation) denote the mean of the B bootstrap samples

$$\bar{\tilde{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\tilde{\theta}_b$$

Recalling that $\hat{\theta}$ our estimate plays the role of θ^* in the bootstrap world, we can estimate the bias in $\hat{\theta}$ with $\bar{\tilde{\theta}} - \hat{\theta}$

Standard error: We can estimate $\mathrm{se}(\hat{\theta})$ with the sample standard deviation of the bootstrap replicates

$$\mathrm{se}_{\mathrm{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\theta}_b - \bar{\tilde{\theta}})^2}$$

The bootstrap

We can then form confidence intervals either by

$$\widehat{\theta} \pm 2\,\mathrm{se}_{\mathrm{boot}}$$

if our bootstrap replicates look reasonably normal, or by using directly **the 0.025 and 0.975 quantiles of the bootstrap replicates as our end points**

This latter scheme 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

The non-parametric bootstrap

So far, we have been wading in MLE-land, where everything spins on assuming some structure about the data generating mechanism -- We assumed our data came from a family with a particular parametric form

Today we'll end by considering how we relax this assumption and what we can say about estimators more broadly...

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
```

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

You've looked at the data from the BRFSS for homework and we computed the BMI for people in the sample -- Today we'll focus on the average BMI (recall 25-30 means overweight)

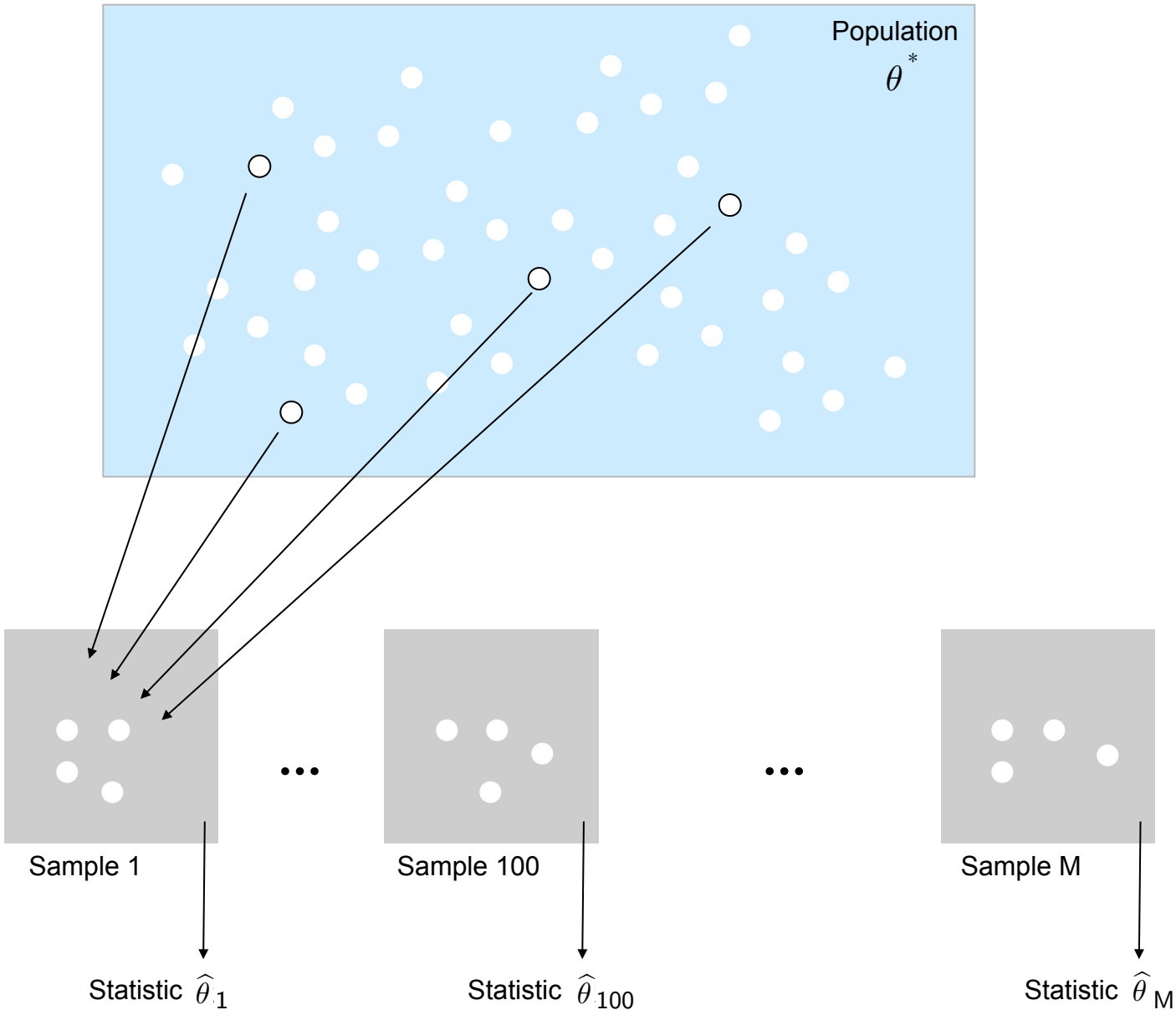
```
> bmi <- 703*cdc$weight/cdc$height^2
> mean(bmi)
[1] 26.30693
> median(bmi)
[1] 25.60354
```

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?**

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 $\widehat{\theta}$=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 $\widehat{\theta}$ -- 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 $\widehat{\theta}_1$
Statistic $\widehat{\theta}_{100}$
Statistic $\widehat{\theta}_M$

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

$$M = \left(\begin{array}{c} N \\ n \end{array} \right) = \frac{N!}{n!(N-n)!}$$

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 distribtuion

This construction is just our old friend the sampling distribution, except that in this case we can enumerate all the possible answers we might get -- There are M different groups we can form from random sampling process

As with our previous estimation setup, we don't want to actually repeat our survey a large number of times and so how might we estimate the sampling distribution here? What did we do in the parametric case?

The plug-in principle

In the parametric case, we too our estimate $\widehat{\theta}$ and substituted it for the true θ^* unknown data-generating -- This was just the plug-in principle

What do we do now?

Real world

Real world parameter θ^*

Sample n times from $f(x|\theta^*)$

↓

Observed sample $X_1, \ldots, X_n$

↓

Estimate $\widehat{\theta} = s(X_1, \ldots, X_n)$

Bootstrap world

Bootstrap world parameter $\widehat{\theta}$

Sample n times from $f(x|\widehat{\theta})$

↓

Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$

↓

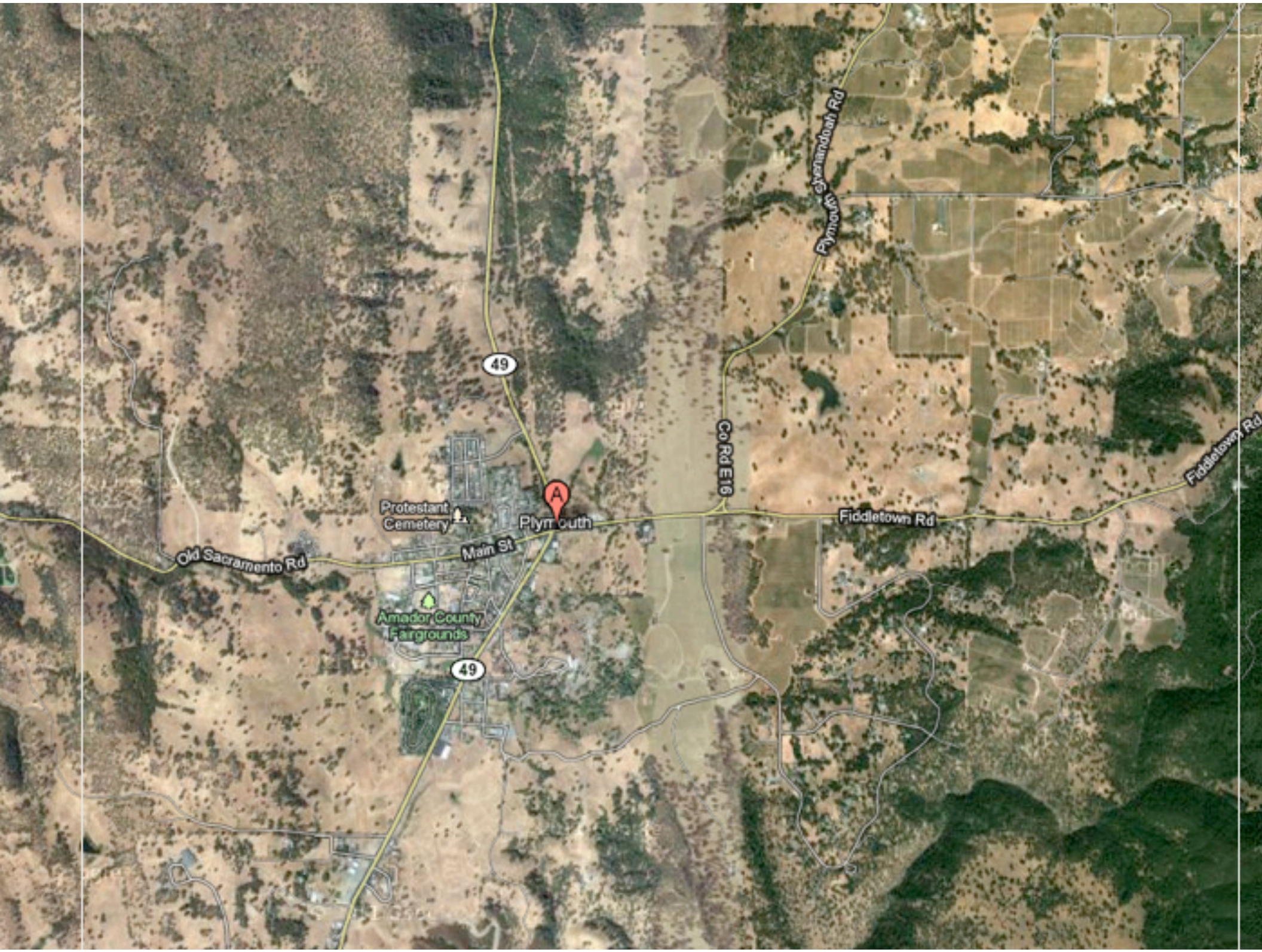
Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$

The bootstrap

Suppose we have a “real world” population of 1,000 people (say all the people living in Plymouth, CA) and we take a random sample of size 100 -- We can “plug-in” our sample and use it to estimate the true population

That is, we could form the “bootstrap world” by simply cloning our 100 data points 10 times each to construct a pseudo-population of size 1,000

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


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

Real world

Real world parameter θ^*

Sample n times from a population of size N

↓

Observed sample $X_1, \ldots, X_n$

↓

Estimate $\widehat{\theta} = s(X_1, \ldots, X_n)$

Bootstrap world

Bootstrap world parameter $\widehat{\theta}$

Sample n times from a population of size N made by copying $X_1, \ldots, X_n$

↓

Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$

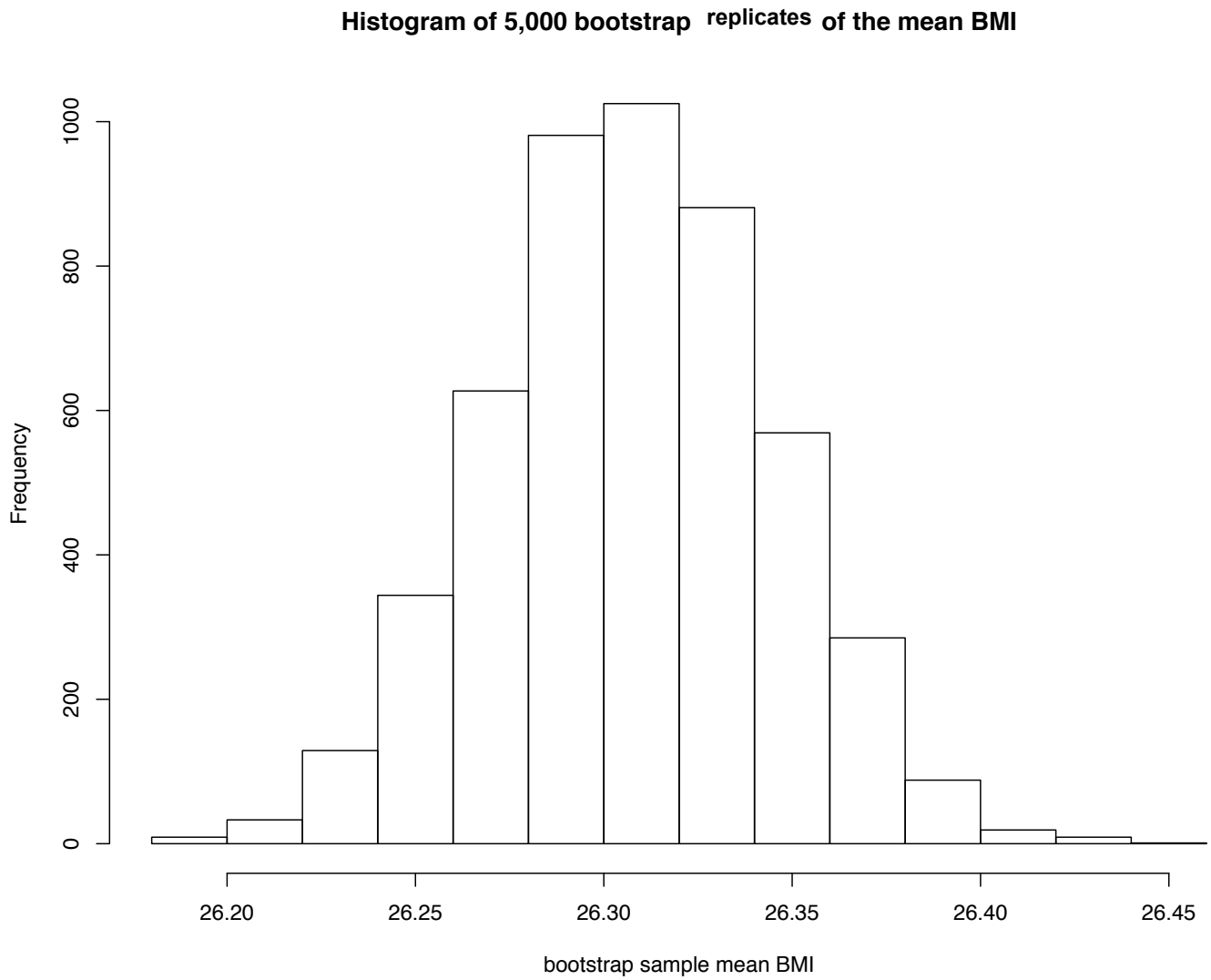
↓

Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$

The bootstrap world

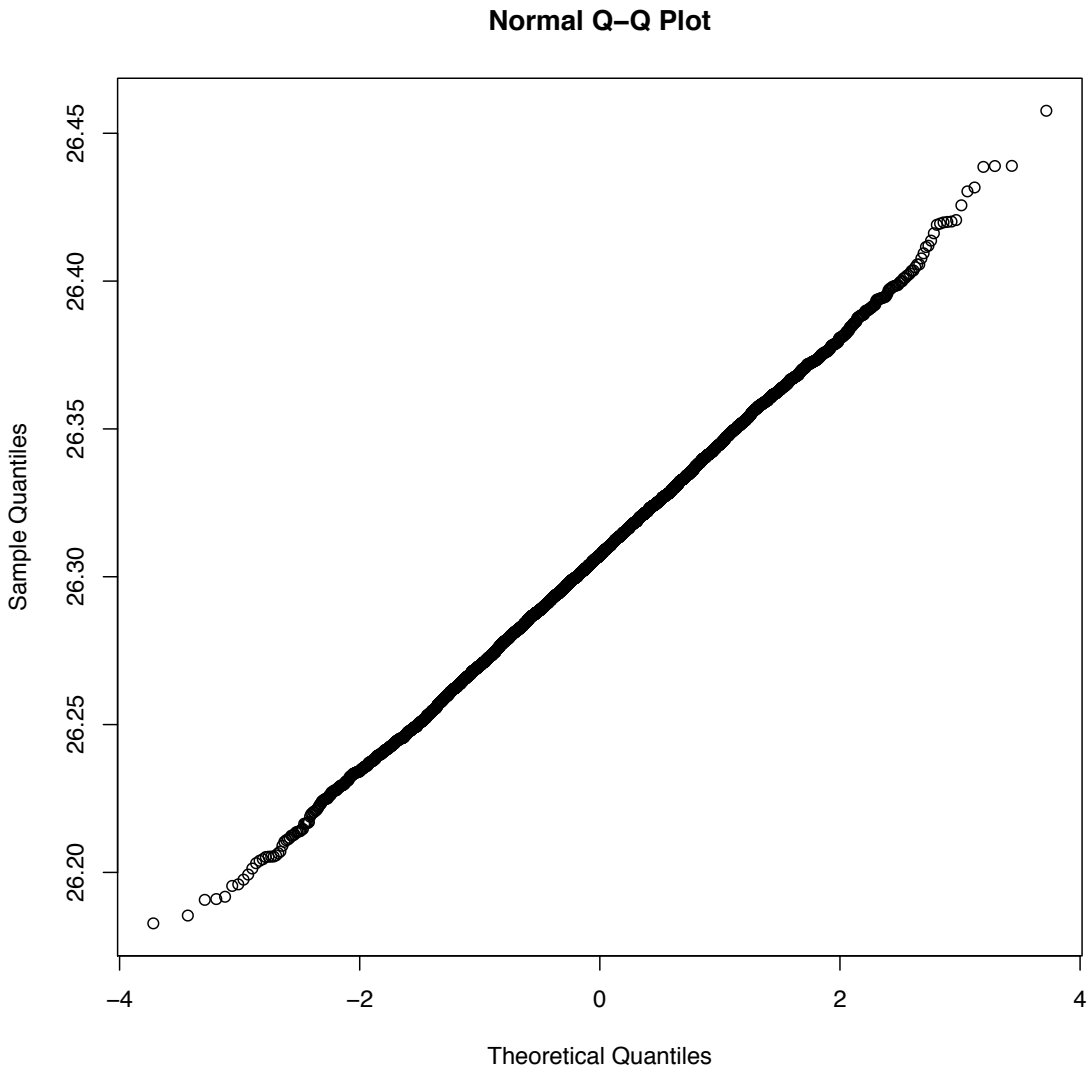
This new bootstrap world population is completely known to us and we can sample from it as often as we like -- It's still not practical to form all samples of size n from this new collection and so instead we look at a few thousand random samples and examine the distribution of the associated estimates

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**


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


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 short cut

There are two ways to pull tickets from a hat -- One is so-called sampling without replacement in which tickets are set aside after they are drawn and sampling with replacement in which they are returned to the hat after they are drawn

For large populations, these two procedures are very similar -- Under sampling with replacement, the chance that you would draw the same ticket twice (assuming you mix things up in between each draw) is very low and it’s as if you didn’t put it back in the hat at all

A shortcut

Therefore, if our real world population is fairly large (N is big), we don't actually have to copy cases at all -- Instead of copying each data point N/n times and sampling from the bootstrap world population without replacement, we can simply use our original n values we observed and form samples with replacement

This cuts down on memory (we don't have to make a data set with hundreds of millions of records if our population is adults in the U.S., say) because we can work directly from the data we have in hand

Real world

Real world parameter θ^*

Sample n times from population

↓

Observed sample $X_1, \ldots, X_n$

↓

Estimate $\widehat{\theta} = s(X_1, \ldots, X_n)$

Bootstrap world

Bootstrap world parameter $\widehat{\theta}$

Sample n times with replacement from $X_1, \ldots, X_n$

↓

Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$

↓

Bootstrap replicate $\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$

The bootstrap

Below we have some simple code that implements the bootstrap procedure here -- Notice we are just leveraging the `sample` command multiple times

```
> bmi <- 703*cdc$weight/cdc$height^2
> mean(bmi)
[1] 26.30693
> sd(bmi)
[1] 5.218105

> boot_reps <- rep(0,5000)

> for(i in 1:5000){

    boot_sample <- sample(bmi,replace=T)
    boot_reps[i] <- mean(boot_sample)
 }

> hist(boot_reps)

> mean(boot_reps)
[1] 26.30671

> sd(boot_reps)
[1] 0.03672835

> ci <- c(mean(boot_reps)-2*sd(boot_reps),mean(boot_reps)+2*sd(boot_reps))
> ci
[1] 26.23326 26.38017
```

BMI

We ended last lecture with the “classical” CLT for a sample mean -- Given that the sample mean is 26.3 and the sample standard deviation of our BMI values from the CDC is 5.2, a 95% confidence interval is

$$26.3 \pm 2 \times 5.2/\sqrt{20000} = [26.22,\ 26.37]$$

This agrees with the bootstrap interval we just computed

Rationale

Again, our goal is not to much to show that we match the old standby approach, but that we now have a tool that we can take to any statistic of the population we might want to compute, not just those for which we know a formula!

The non-parametric bootstrap

In this case, we did not make a particular parametric assumption about how the data were generated -- Instead we used the fact that our data were a random sample from some population

Sampling, then, provides us with the sole rationale for estimation “analyze as you randomized” -- We refer to this approach as the non-parametric bootstrap because it does not involve any parametric assumptions whatsoever

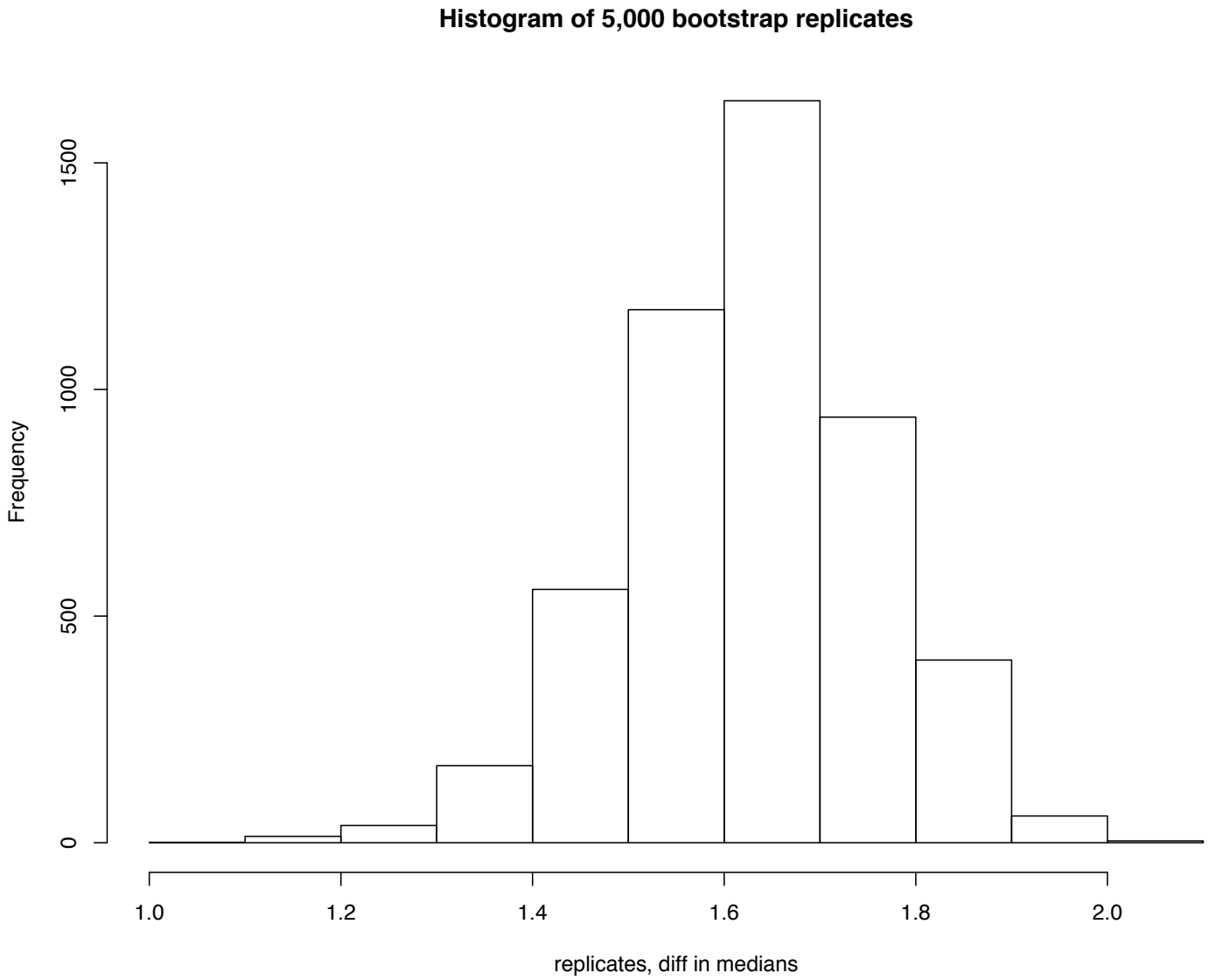
It is like the plug-in principle on steroids -- We are literally plugging in our sample to represent the population

We are also not restricted to the kind of estimate we wish to compute from the sample -- Means, medians, trimmed means, quantiles...

Men and women

Now, suppose we want to assess whether adult males in the U.S. have the same median BMI as women -- We can build a confidence interval for the difference in medians again using the bootstrap

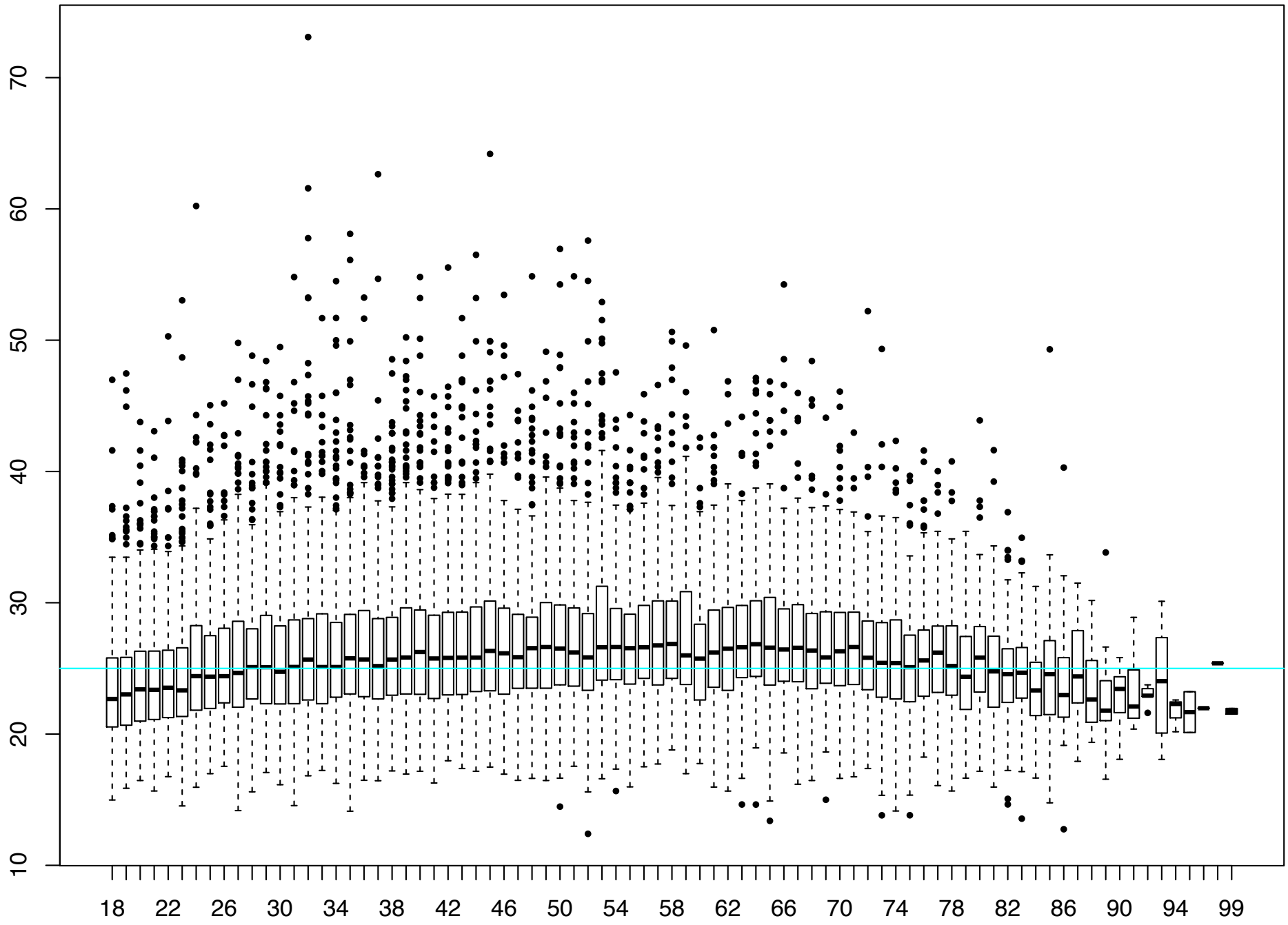
```
> cdc$bmi <- bmi
> men <- subset(cdc,gender=="m")
> women <- subset(cdc,gender=="f")

> median(men$bmi)
[1] 26.28313
> median(women$bmi)
[1] 24.63832

> median(men$bmi)-median(women$bmi)
[1] 1.644811

> for(i in 1:5000){

    boot_men <- sample(men$bmi,replace=T)
    boot_women <- sample(women$bmi,replace=T)

    boot_reps[i] <- median(boot_men)-median(boot_women)
 }

> quantile(boot_reps,c(0.025,0.975))
    2.5%    97.5%
1.370585 1.877693
```


Histogram of 5,000 bootstrap replicates
Frequency
0
500
1000
1500
1.0
1.2
1.4
1.6
1.8
2.0
replicates, diff in medians

A more exotic example

We might want to explore, for example, the relationship between age and BMI status -- That is, do things get harder for us old folk?

We might have a look at this relationship via a boxplot, splitting BMI by age on the x-axis -- What do you see?

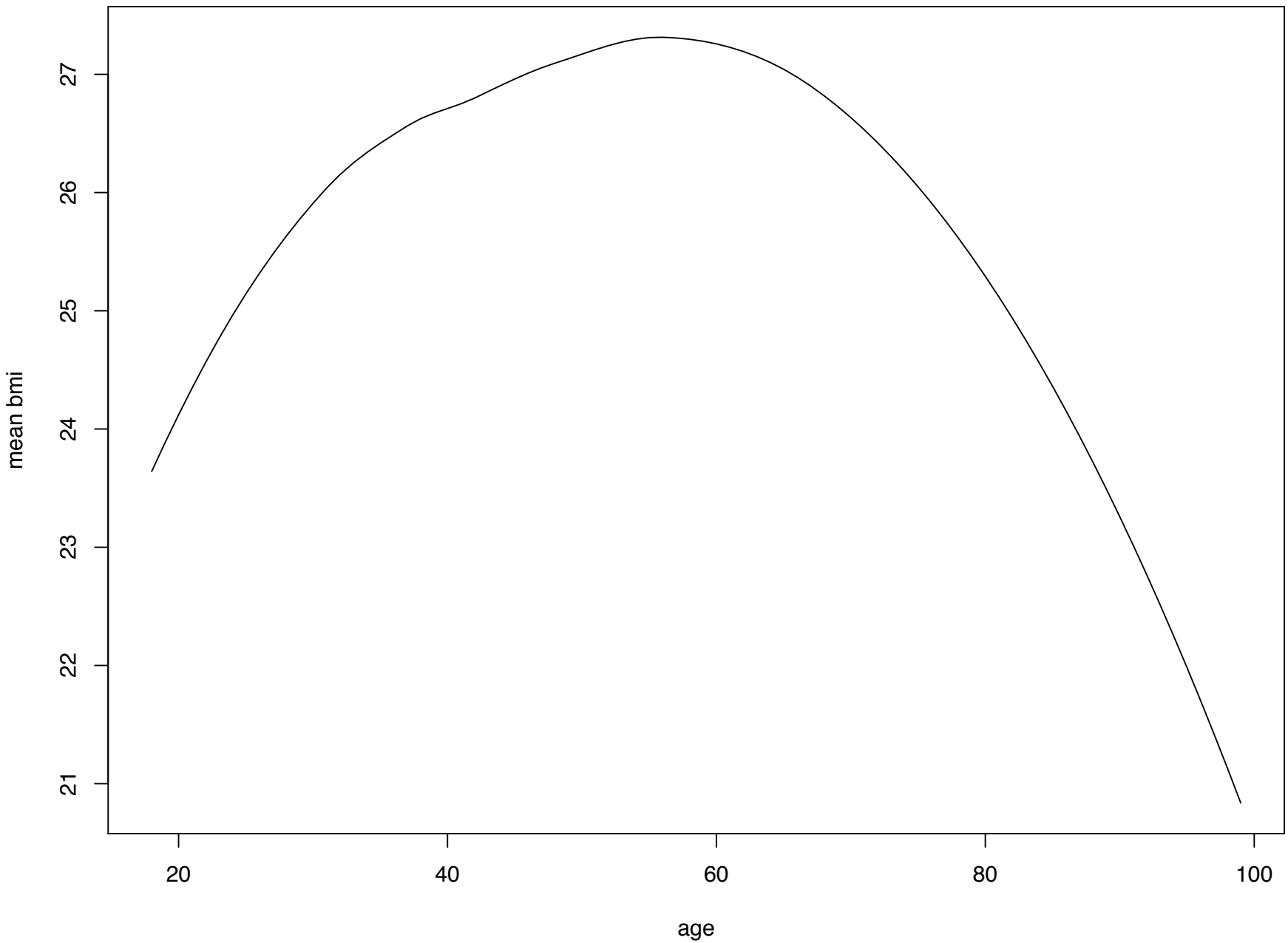

70
60
50
40
30
20
10
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
99

Smoothing

We can now formally “fit” the relationship between age and BMI using something known as a smoother -- It basically is an automated procedure for putting a smooth curve through the middle of the data

Here is what we get doing this for the BMI by age plot on the previous page -- What do you see? What does this say about the population at large?

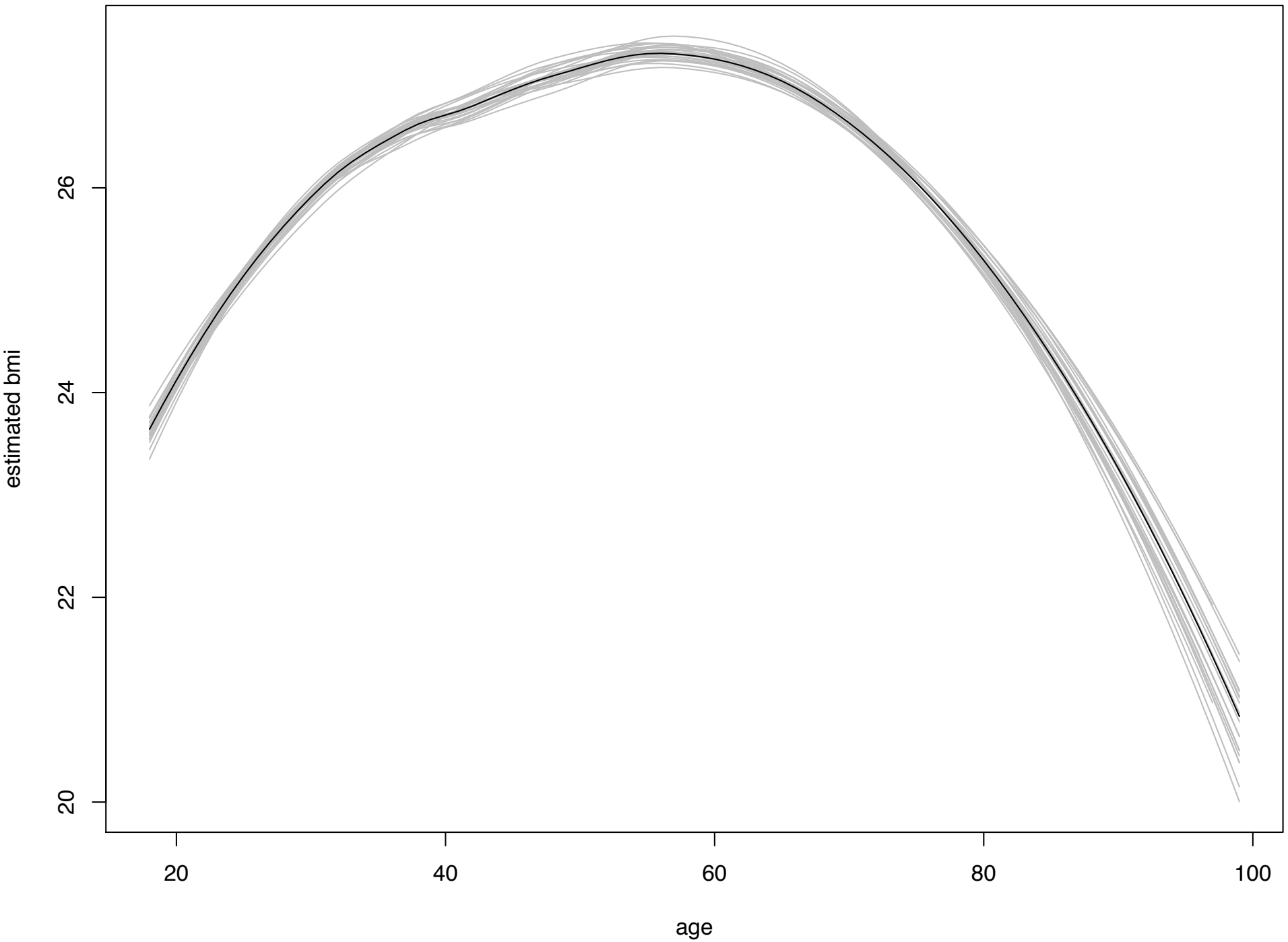
How can we assess the uncertainty here?


mean bmi
21
22
23
24
25
26
27
20
40
60
80
100
age

The bootstrap

In this case we sample cases (rows of the original CDC survey) with replacement, and each time refit the smooth curve -- With each sample we will get another curve and plotting them all at once gives us a sense of the variability present

On the next page, we ran the smoothing on 25 bootstrap samples...


estimated bmi
20
22
24
26
20
40
60
80
100
age

Another example

Let's return to the A/B testing scenario you looked at in homework -- We had a "treatment" that removed the navigation bar from the Movies Section pages at the New York Times web site

Let's compute the "rates" at which people used the search box under each treatment and then compute their ratio

```
> nobar <- subset(movies,treatment=="no bar")
> unchanged <- subset(movies,treatment=="unchanged")

> mean(nobar$outcome)
[1] 0.1031746

> mean(unchanged$outcome)
[1] 0.05040323

> mean(nobar$outcome)/mean(unchanged$outcome)
[1] 2.046984
```

As you saw from your work last week, people seeing pages without the navigation bar searched twice as often as those seeing the usual page

A/B testing

What can we say now more generally about visitors to the site and how we expect this treatment to affect them? If we implement the change and remove the navigation bar, will we expect to see twice as much use of the search box?

To answer this, we will form a confidence interval for this ratio -- Since people were selected randomly to participate in the experiment, with those selected then being randomized to see treatment or control designs

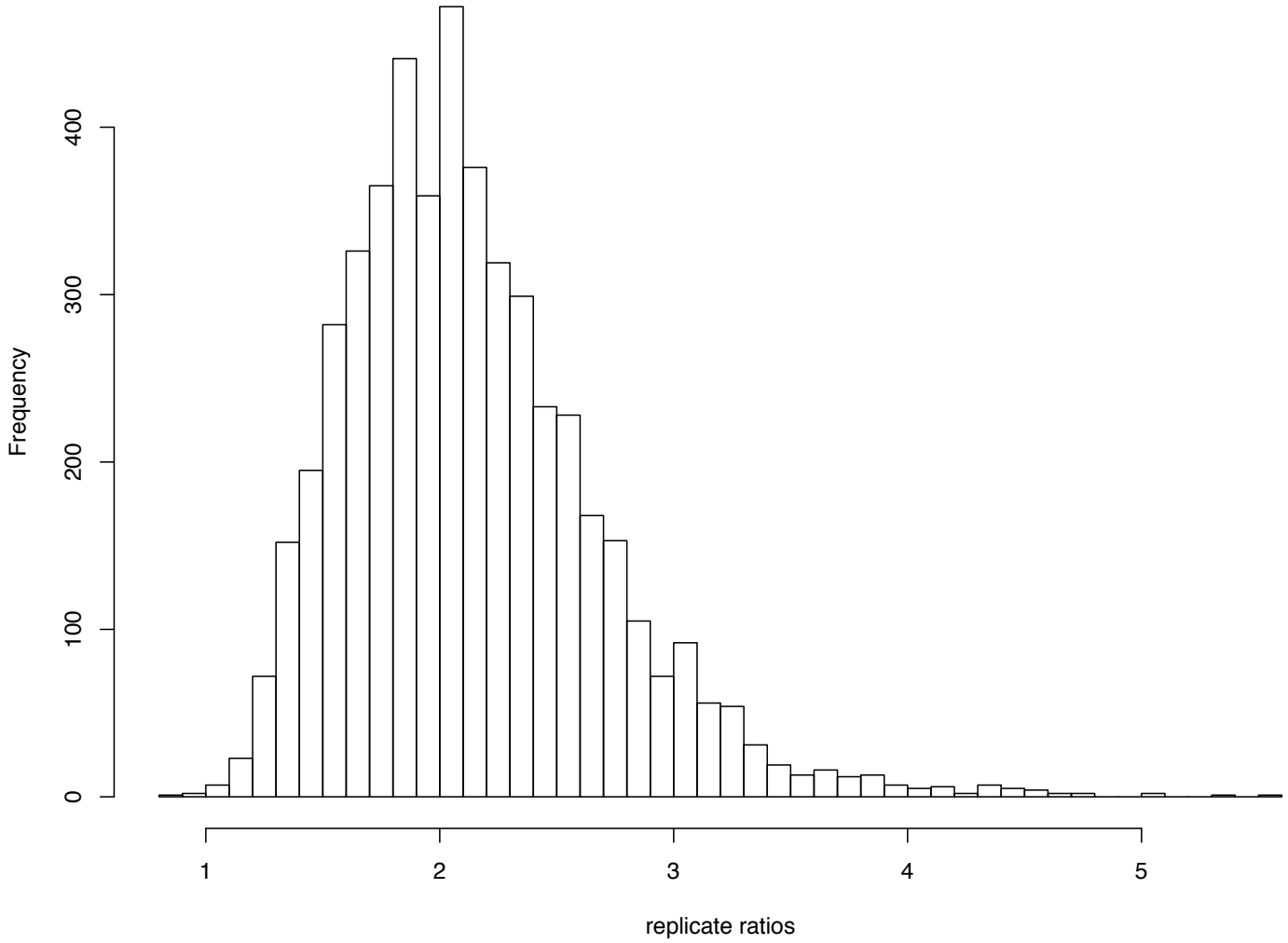
We can appeal to the nonparametric bootstrap here, viewing the two groups enrolled in the study as random samples from the population -- Here we invoke the idea of a prospective sample

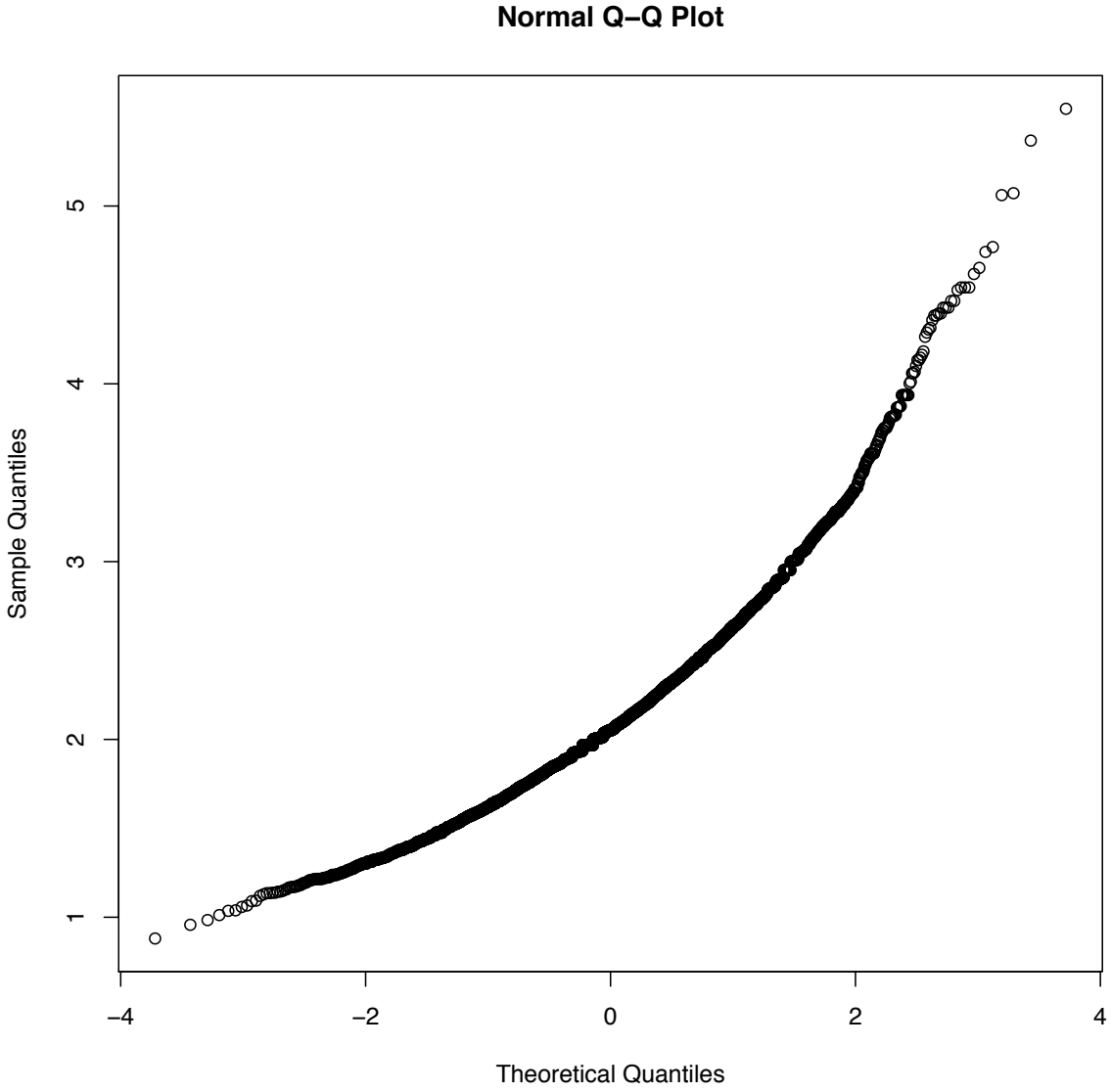
Ok, now what?

A/B testing

Our bootstrap procedure is pretty straightforward, we divide our data into the two groups (“no bar” and “unchanged”) and then resample each, forming bootstrap replicates of our ratio

```
> boot_reps <- rep(0,5000)

> for(i in 1:5000){

    boot_nobar <- sample(nobar$outcome,replace=T)
    boot_unchanged <- sample(unchanged$outcome,replace=T)
    boot_reps[i] <- mean(boot_nobar)/mean(boot_unchanged)
 }

> hist(boot_reps)
> qqnorm(boot_reps)

> mean(boot_reps)
[1] 2.132537

> sd(boot_reps)
[1] 0.5423531

# classical ci assuming approximate normality
> c(mean(boot_reps)-2*sd(boot_reps),mean(boot_reps)+2*sd(boot_reps))
[1] 1.047831 3.217243

# percentile method
> quantile(boot_reps,c(0.025,0.975))
    2.5%    97.5%
1.312169 3.366935
```


Histogram of 5,000 bootstrap replicates
Frequency
0
100
200
300
400
1
2
3
4
5
replicate ratios


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
1
2
3
4
5
−4
−2
0
2
4

Link to (parametric) hypothesis testing

In the decision-theoretic framework of Neyman and Pearson (we saw earlier in the quarter), we can easily link confidence intervals to hypothesis tests -- In short, reject your null hypothesis if the interval you've constructed doesn't contain the hypothesized value of the parameter

Suppose our null hypothesis consists of a statement about the unknown parameter θ^*; for example, we might hypothesize that Tabs and Lists have the same behavior on visitors so the ratio of click rates is 1 or that adult mean and women have the same median BMI

So, if we construct a 95% confidence interval $[\widehat{\theta}_{\text{lo}}, \widehat{\theta}_{\text{hi}}]$ for θ^* and we would like to test the hypothesis that $\theta^* = \theta_0$, then we reject if we reject when $\theta_0 \notin [\widehat{\theta}_{\text{lo}}, \widehat{\theta}_{\text{hi}}]$, we will be making a mistake only 5% of the time, a test with level 0.05!