Lecture 11: Risk

Last time

We started to discuss some of the basic ideas behind on estimation -- We extended our toolkit of computational procedures and examined a simple way to assess uncertainty in estimates

We used 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

We then saw that these questions become answerable through something called the bootstrap, a procedure that let us asses bias, compute standard errors, RMS and even confidence intervals associated with estimates

Today

We'll use these ideas in an extended example involving relative risk -- We will exercise the full set of tools we can derive from the bootstrap distribution

We will then return to a meta-analysis we mentioned in a previous lecture -- Next time we will examine other "classical" and decidedly mathematical ways to approximate the sampling distribution

We will emphasize the bootstrap in applications as it can be used in a variety of estimation contexts -- Our connection to the mathematical material is mainly historical and will be useful vocabulary when you use your newfound skills in other classes

Frequentist statistics

We've commented several times that the frequentist view of statistics depends on **the notion of repeated trials --** We interpret the outcome from our given experiment as just one of a large number of experiments that could have taken place

In this setting, we need to think about an estimate, something we've computed from a random sample, as more of **an algorithm that can be repeated for any particular outcome of the experiment**

So to make sense of an estimate that we've computed on a given set of data from a particular experiment, we start by thinking about **how an estimator would perform if applied to all the possible experimental outcomes**

Terminology

By the end of last lecture, we had been referring to the population parameter (a quantity we'd like to infer something about given a sample) as θ and an estimate as $\widehat{\theta}$

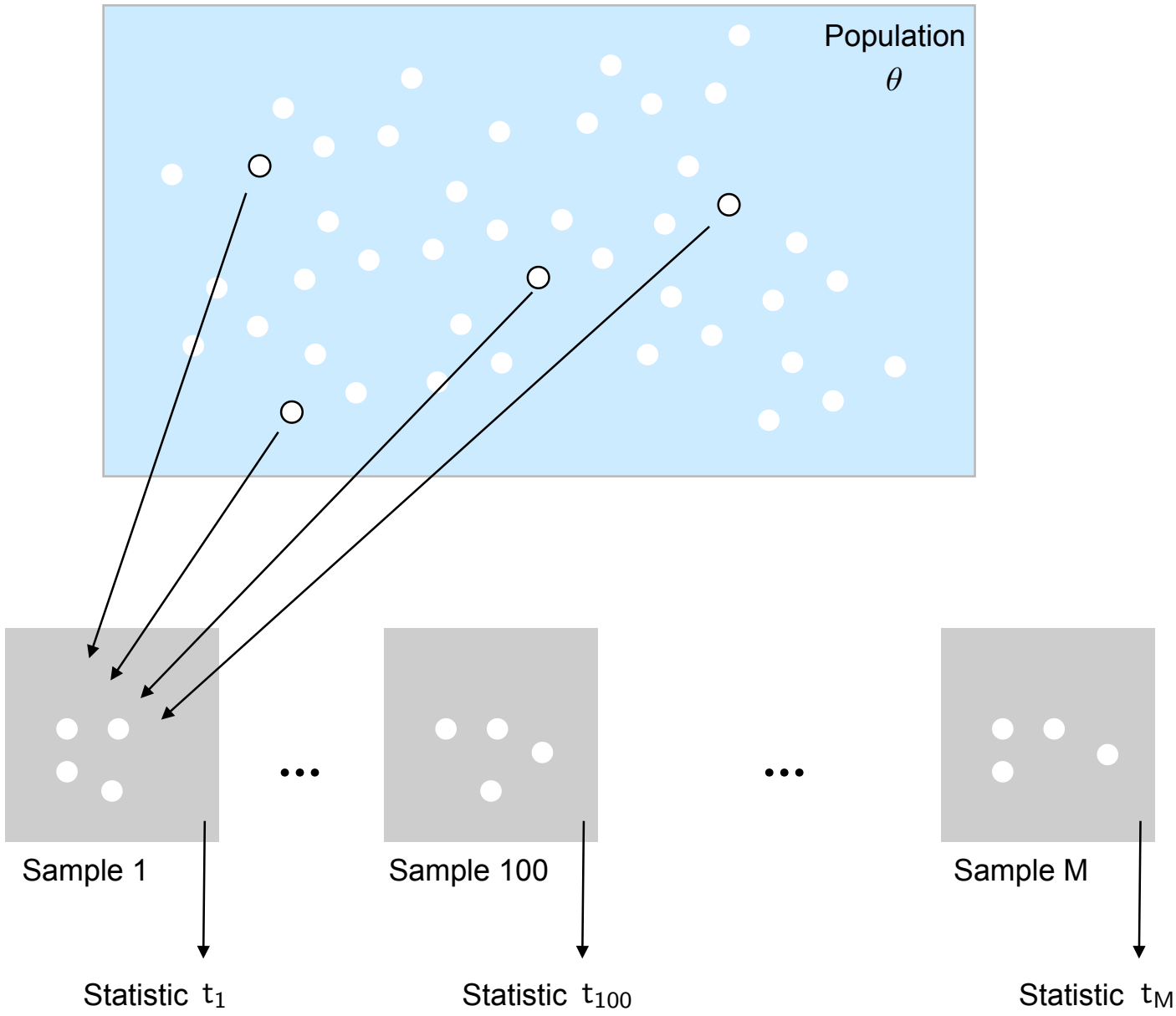
The population parameter θ is unknown to us -- All we have to work with is a single sample of data and whatever we can compute from it

Simple random sampling

To ground this idea, we focused on **simple random sampling from a population** and considered the distribution of estimates that could be computed for any possible sample we could have drawn

The collection is known as the sampling distribution and is really more of an intellectual exercise or a theoretical construction as no one would ever form all the samples possible from a population

Still, it is a powerful tool for assessing the performance of an estimator...


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

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

We then considered how this distribution could be used to assess the accuracy of any single estimate, say our t_1

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 t_1 might be 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

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

Confidence intervals

The conceptually trickiest notion we introduced last time was that of a confidence interval -- You can think of it as an algorithm that, given a data set, produces an interval of “plausible” values for the population parameter of interest

The notion of confidence is derived from the sampling distribution -- A 95% confidence interval, say, is an algorithm that, when applied to all of the possible samples we could form, produces an interval that contains the population parameter in 95% of the cases

Confidence intervals

At the right we have a sampling distribution -- Again, it is made up statistics or estimates that we construct from all the possible samples (of a given size) we could draw from the population of interest

We've drawn it not totally smooth to underscore the fact that at the moment we are thinking of it consisting of M points, since there are M different samples we could draw

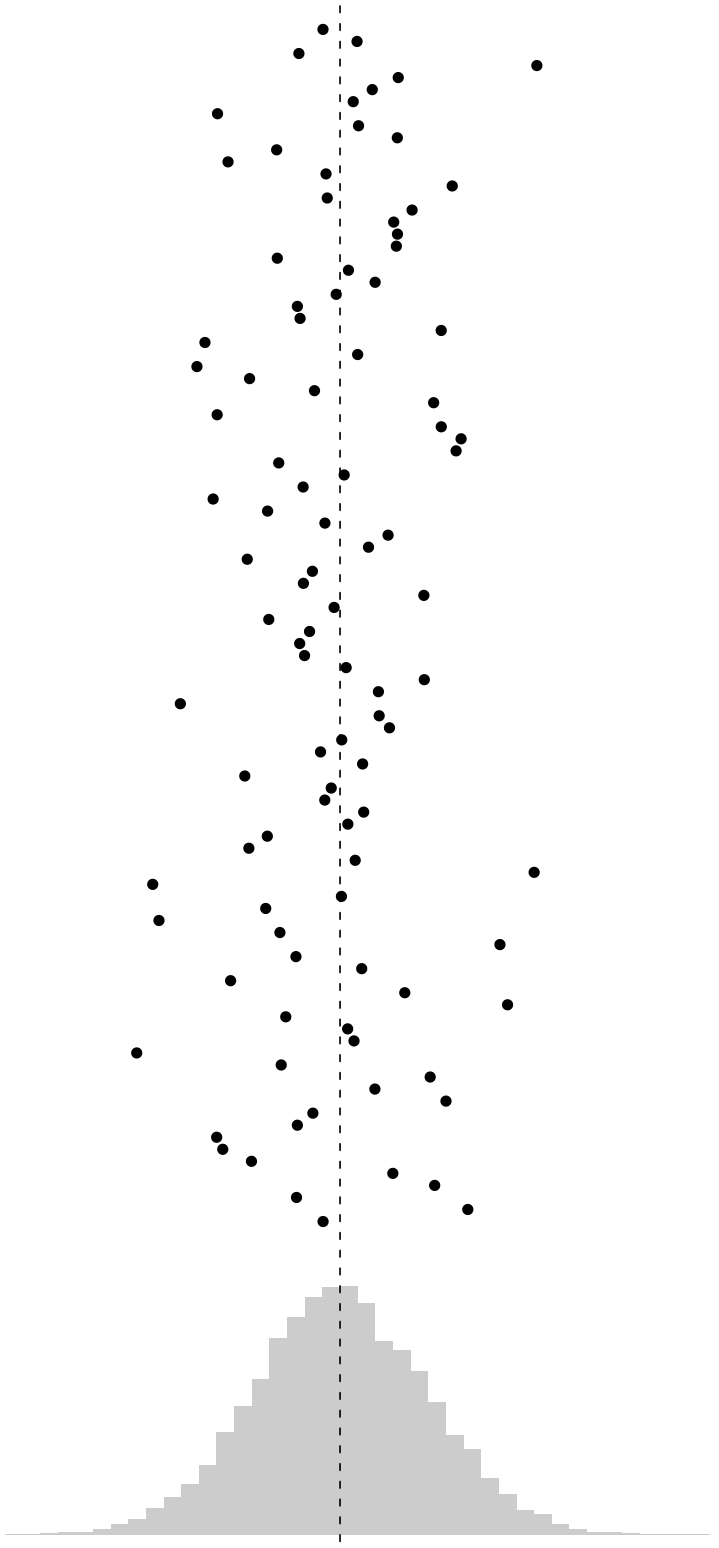
The vertical dashed line is the value of the population parameter -- In this case, the sampling distribution seems to be centered on this value, meaning there is little bias

Confidence intervals

At the right we have a sampling distribution -- Again, it is made up statistics or estimates that we construct from all the possible samples (of a given size) we could draw from the population of interest

Now, we've added a few of the M estimates -- We've taken 100 of the M samples and used each data set to form an estimate; these are the 100 dots

If you squint your eyes you can see that they are somewhat more concentrated in the middle than toward the edges and don't seem inconsistent with the distribution at the bottom

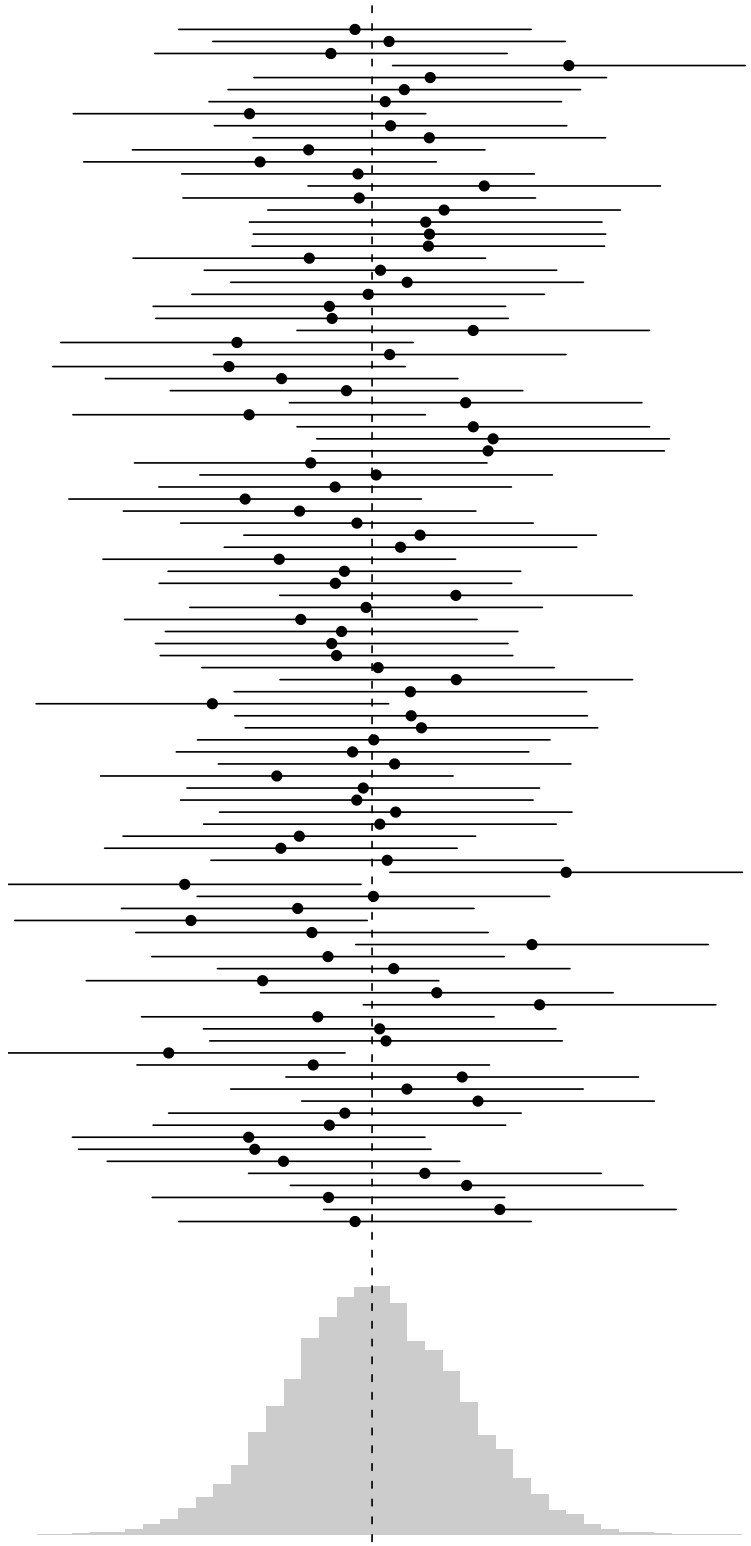


Confidence intervals

At the right we have a sampling distribution -- Again, it is made up statistics or estimates that we construct from all the possible samples (of a given size) we could draw from the population of interest

For each of the 100 samples on the right, we also form a confidence interval -- **This interval uses only the data in the sample** (think of it as the bootstrap estimate from last time or maybe just the estimate +/- 2 standard errors)

These are 95% confidence intervals meaning they have been calibrated so that...

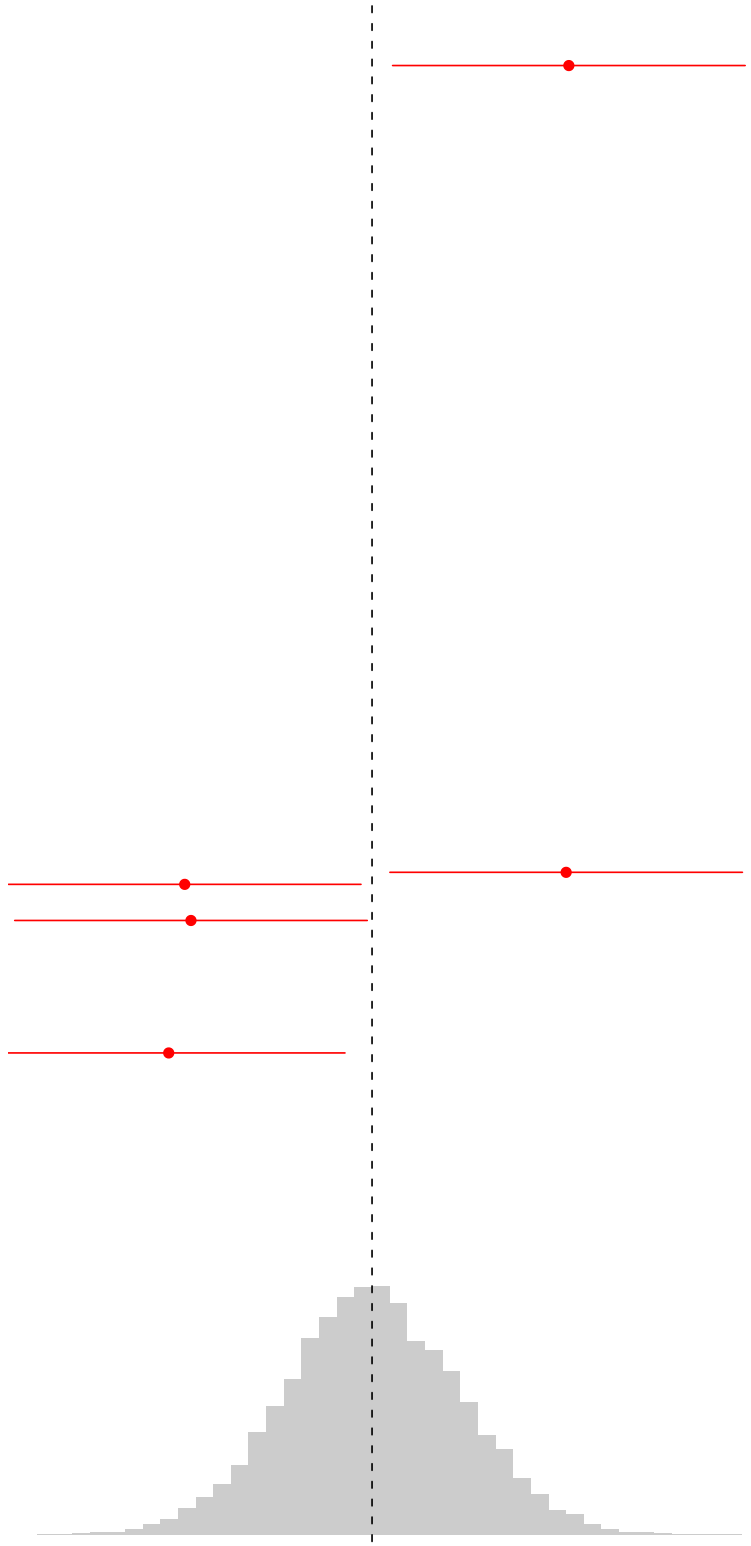


Confidence intervals

... all but about 5% of them actually contain the true parameter value

This is our frequentist notion of confidence -- We can talk about the properties of the algorithm we use to form confidence intervals but that conversation is in reference to all the possible samples we could have drawn

We are only given one sample and it's not clear to us whether our confidence intervals is one of the 5 on the right -- 5% we will be in that place and we won't know it



Confidence intervals

The classic confusion about confidence intervals is, well, this crazy notion of confidence -- To say we are 95% confident (in frequentist lingo) is to refer to 95% of the M samples we could have taken

We reason indirectly about our own experimental results -- We would have to be unlucky for the population parameter to not be contained in our interval or, put another way, we'd have to have picked one of the 5% bum M samples

We're used to thinking this way from hypothesis testing -- It's just important to keep it up here

The bootstrap

While the sampling distribution is a theoretical object, there are various ways to approximate it -- Statisticians at the beginning of the last century attacked it mathematically, depending on the central limit effect to suggest when it will look bell shaped

Our technique will build on the aphorism “analyze as you randomized” -- The bootstrap sees us repeating the sampling procedure many times to generate an approximation to the sampling distribution

The sampling, however, is not from the original population, but a bootstrap world version of it...

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

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

A shortcut

We were vague last time about the copying process to construct the bootstrap world population -- If the original population is large (all the employees of a company, all the students at UCLA), then we don't literally have to copy data

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 like 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

The bootstrap

Once computed, 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 $\widehat{\theta}$ from the original sample -- If it is not, we have evidence that our estimate is biased

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

Vioxx (the really really last time, really)

The anti-inflammatory drug was introduced to the market in the late 1990s and was prescribed for the treatment of arthritis and acute pain

In 2000, the New England Journal of Medicine published the results from a large randomized controlled trial designed to examine whether patients receiving rofecoxib (Vioxx) would have fewer upper gastrointestinal events (perforations, ulcers, bleeding) than those taking naproxen (marketed as Aleve)

Recall the study involved 8,076 patients suffering from rheumatoid arthritis, each randomized into treatment or control (Vioxx or Aleve, respectively)

The New England

COMPARISON OF UPPER GASTROIN
AND NAPROXEN IN PATIENTS

CLAIRE BOMBARDIER, M.D., LOREN LAINE, M.D.,
RUBEN BURGOS-VARGAS, M.D., BARRY DAVIS, M.D., PH.D
CHRISTOPHER J. HAWKEY, M.D., MARC C
AND THOMAS J. SCHNITZER, M.D., F

ABSTRACT

Background Each year, clinical upper gastrointestinal events occur in 2 to 4 percent of patients who are taking nonselective nonsteroidal antiinflammatory drugs (NSAIDs). We assessed whether rofecoxib, a selective inhibitor of cyclooxygenase-2, would be associated with a lower incidence of clinically important upper gastrointestinal events than is the nonselective NSAID naproxen among patients with rheumatoid arthritis.

Methods We randomly assigned 8076 patients who were at least 50 years of age (or at least 40 years of age and receiving long-term glucocorticoid therapy) and who had rheumatoid arthritis to receive either 50 mg of rofecoxib daily or 500 mg of naproxen twice daily. The primary end point was confirmed clinical upper gastrointestinal events (gastroduodenal perforation or obstruction, upper gastrointestinal bleeding, and symptomatic gastroduodenal ulcers).

Results Rofecoxib and naproxen had similar efficacy against rheumatoid arthritis. During a median follow-up of 9.0 months, 2.1 confirmed gastrointestinal events per 100 patient-years occurred with rofecoxib, as compared with 4.5 per 100 patient-years with naproxen (relative risk, 0.5; 95 percent confidence interval, 0.3 to 0.6; $P<0.001$). The respective rates of com-

Vioxx

In addition to GI problems, the researchers considered a variety of possible side-effects from taking rofecoxib (R) or naproxen (N); here we present a two-by-two table of patients who experienced cardiovascular adverse events (CE)

Status	N	R	
no CE	4010	4002	8012
CE	19	45	64
	4029	4047	

Treatment

Vioxx

Here we see that in the naproxen group, 19/4029 or 0.5% patients experienced cardiovascular adverse events, while under rofecoxib 45/4047 or 1.1% of patients had problems; the chance that a patient develop CE under rofecoxib is over twice as high

Status	Treatment N	Treatment R	
no CE	4010	4002	8012
CE	19	45	64
	4029	4047	

Vioxx

Initially, we examined these data to assess whether **the difference in cardiac adverse events could have been the result of the randomization** employed during the study design

At that point, we commented that our analysis didn't say much about how **Vioxx would operate among a larger group of patients** -- For that, we needed to know that the people in the study were in some way representative of the larger population

This is where random sampling come in -- Typically, **a randomized controlled trial also starts with a random selection of patients**, the population being, in this case, all people diagnosed with rheumatoid arthritis

Vioxx

This means we can apply our new-found inferential skills to assess the relative risk of experiencing a heart attack on Vioxx versus Aleve **among all patients with rheumatoid arthritis**

To do this we invoke the idea of **prospective populations** -- We imagine having treated all the patients with rheumatoid arthritis with Vioxx, and our study's Vioxx group is a random sample from this population

The same thought process can be followed with the Aleve group -- We imagine having treated all the patients with rheumatoid arthritis with Aleve, and our study's Aleve group is just a random sample form this population

This is just a mental exercise as the only patients receiving any treatment in our study are the 8,000 patients enrolled in the trial -- It is in this sense that the two populations are said to be prospective

```
# load the data

> source("http://www.stat.ucla.edu/~cocteau/stat13/data/trial.R")
> head(trial)

  response        treatment
1        0 naproxen (aleve)
2        0 naproxen (aleve)
3        0 naproxen (aleve)
4        0 naproxen (aleve)
5        0 naproxen (aleve)
6        0 naproxen (aleve)

# and see that it has the right counts of treatment and
# response cells in our table

> table(trial)
        treatment
response naproxen (aleve) rofecoxib (vioxx)
       0             4010              4002
       1               19                45

# risk of heart attach under vioxx
> 45/(45+4002)
[1] 0.01111935

# and under aleve
> 19/(4010+19)
[1] 0.00471581

# and the relative risk
> (45/(45+4002))/(19/(4010+19))
[1] 2.357887
```

Vioxx

The relative risk computed from our experimental data is 2.36, which means that patients taking Vioxx were over twice as likely to have a heart attack than those taking Aleve

Can we now assign some notion of the accuracy of our estimate? What does it say about the performance of Aleve and Vioxx in the larger populations? To answer these questions, we appeal to **the bootstrap**

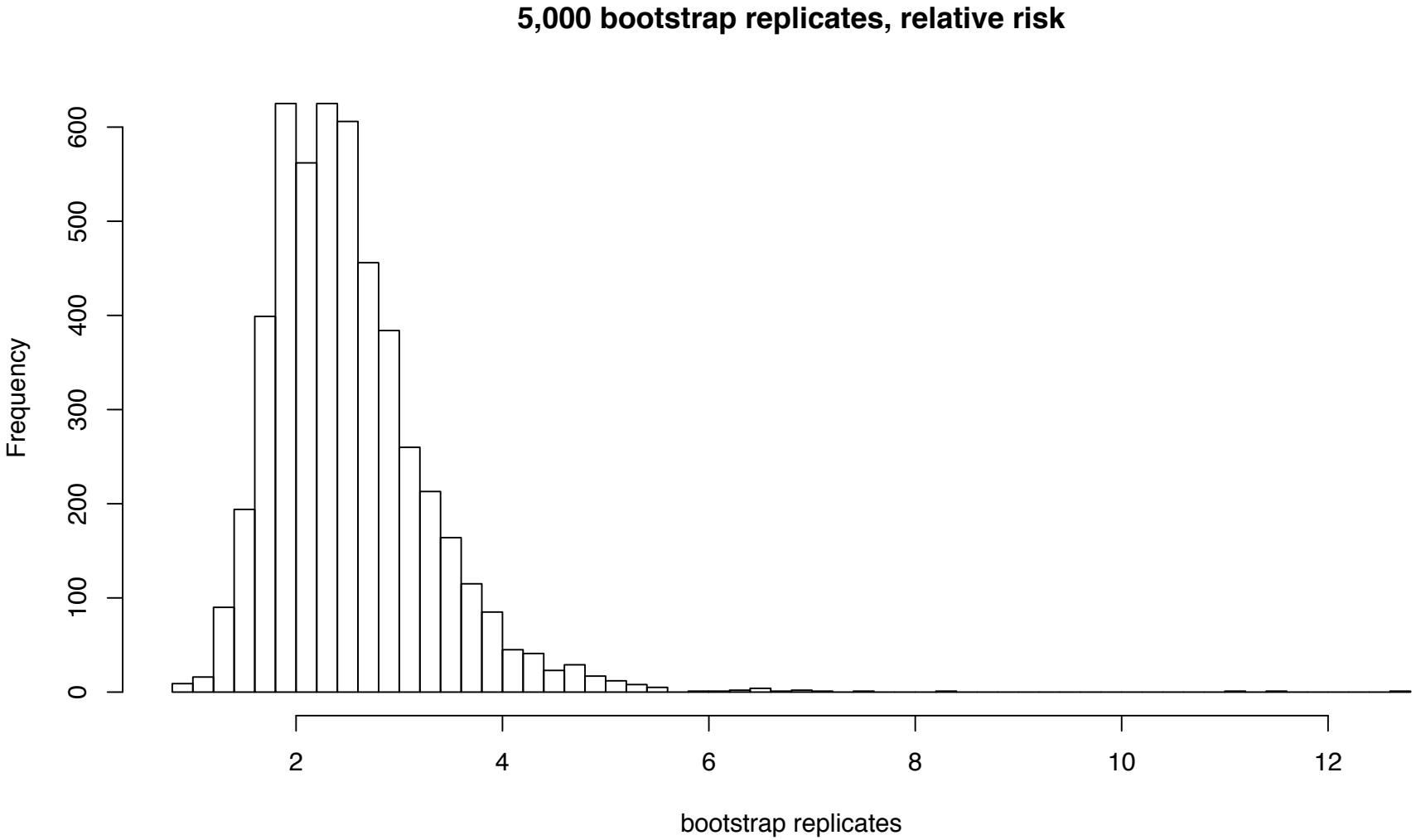
Because the collection of people suffering from rheumatoid arthritis is so large, we can play the sampling with replacement trick -- We start by forming two groups of patients from which we sample repeatedly, drawing bootstrap samples

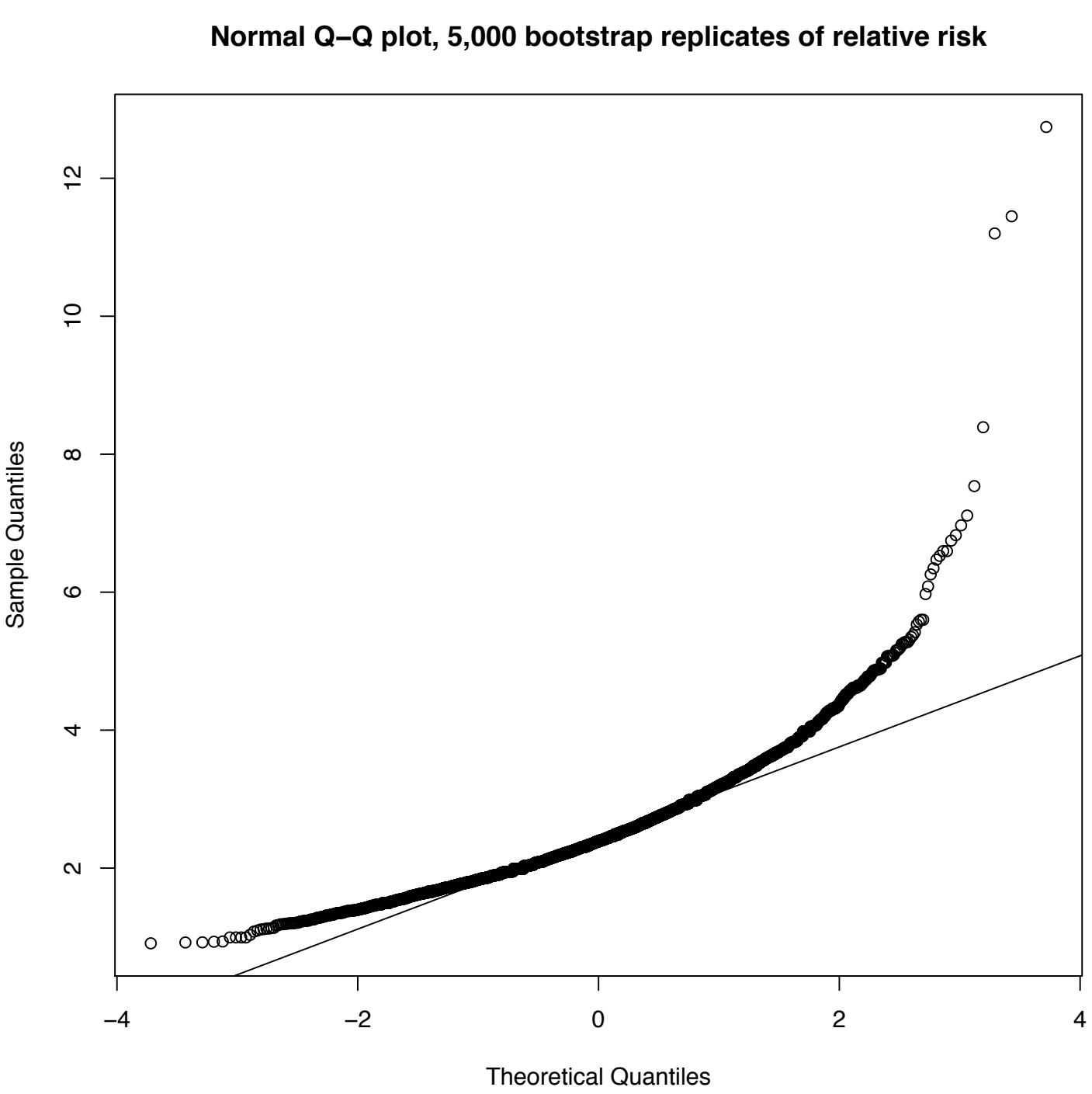
We will generate **a number (in this case, say, 5,000) bootstrap samples and for each we will form the relative risk -- This will produce 5,000 bootstrap replicates which will form an approximation to the sampling distribution** of the relative risk

Let's get dirty and program a little!

```
# form the bootstrap world “populations”

> popvioxx <- subset(trial$response,trial$treatment=="rofecoxib (vioxx)")
> popaleve <- subset(trial$response,trial$treatment=="naproxen (aleve)")

# and check that the counts are ok

> table(popvioxx)
popvioxx
   0    1
4002   45

> table(popaleve)
popaleve
   0    1
4010   19

# now the bootstrap!
# this vector will hold our 5,000 bootstrap replicates of the rel risk

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

```
# then iterate!

> for(i in 1:5000){

    # draw a bootstrap sample

    bootvioxx <- sample(popvioxx,replace=TRUE)
    bootaleve <- sample(popaleve,replace=TRUE)

    # compute the chances of having a heart attack on
    # each medication

    pvioxx <- mean(bootvioxx)
    paleve <- mean(bootaleve)

    # and store the relative risk

    r[i] <- pvioxx/paleve
 }

# plot the bootstrap distribution

> hist(r,breaks=50)

# compare it to a normal distribution
# what do you see?

> qqnorm(r,breaks=50)
> qqline(r)
```


5,000 bootstrap replicates, relative risk
Frequency
0
100
200
300
400
500
600
2
4
6
8
10
12
bootstrap replicates


Normal Q–Q plot, 5,000 bootstrap replicates of relative risk
Sample Quantiles
Theoretical Quantiles
2
4
6
8
10
12
−4
−2
0
2
4

The bootstrap

Again, what we have is an approximation to the sampling distribution for our estimate of relative risk -- We can use it to assess **the shape of the sampling distribution**, derive **the standard error** and **root mean squared error**, examine **any possible bias** and derive **confidence intervals** for the population relative risk

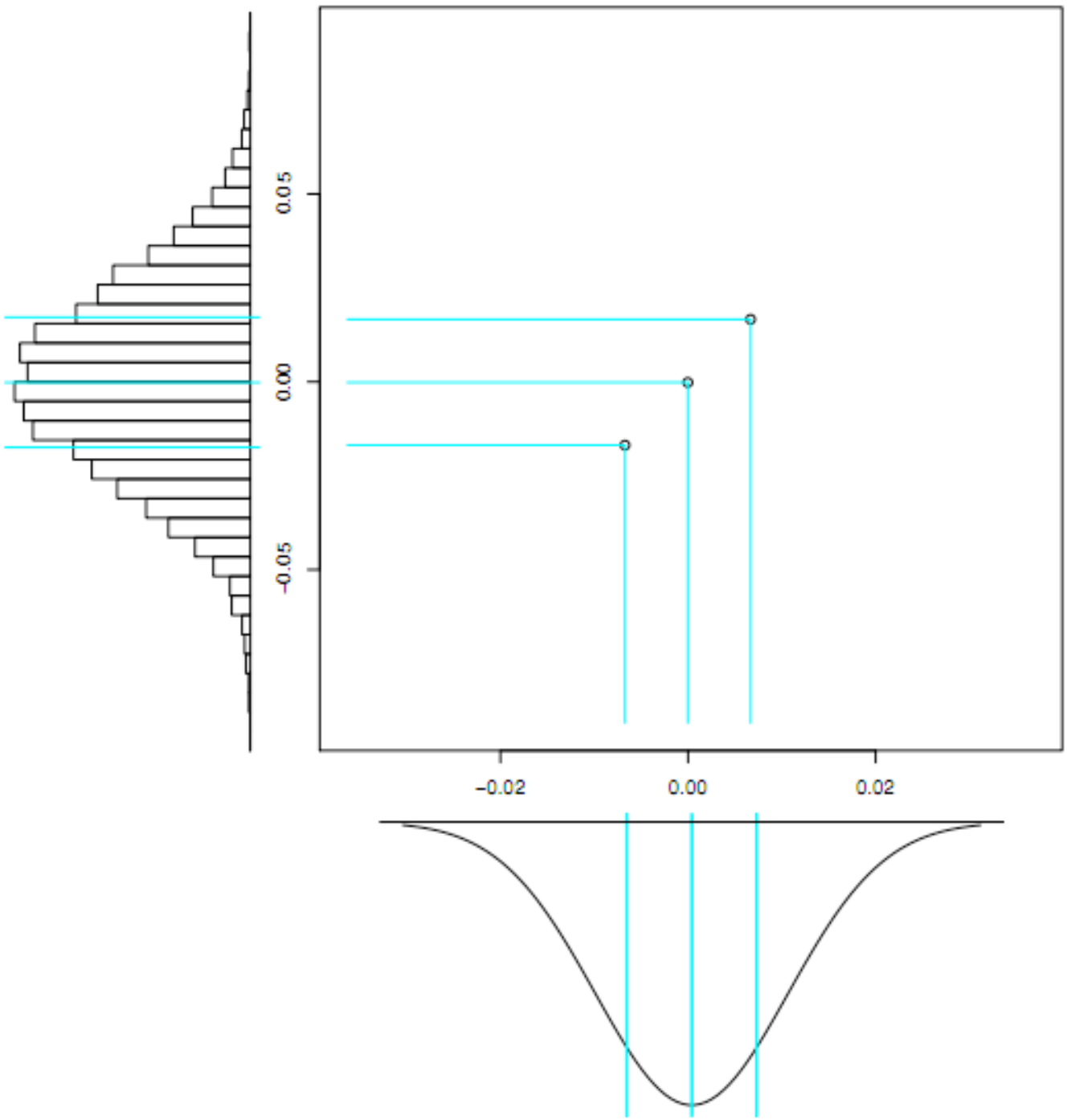
In short, the distribution of bootstrap samples tells us quite a lot about the performance of our estimator...

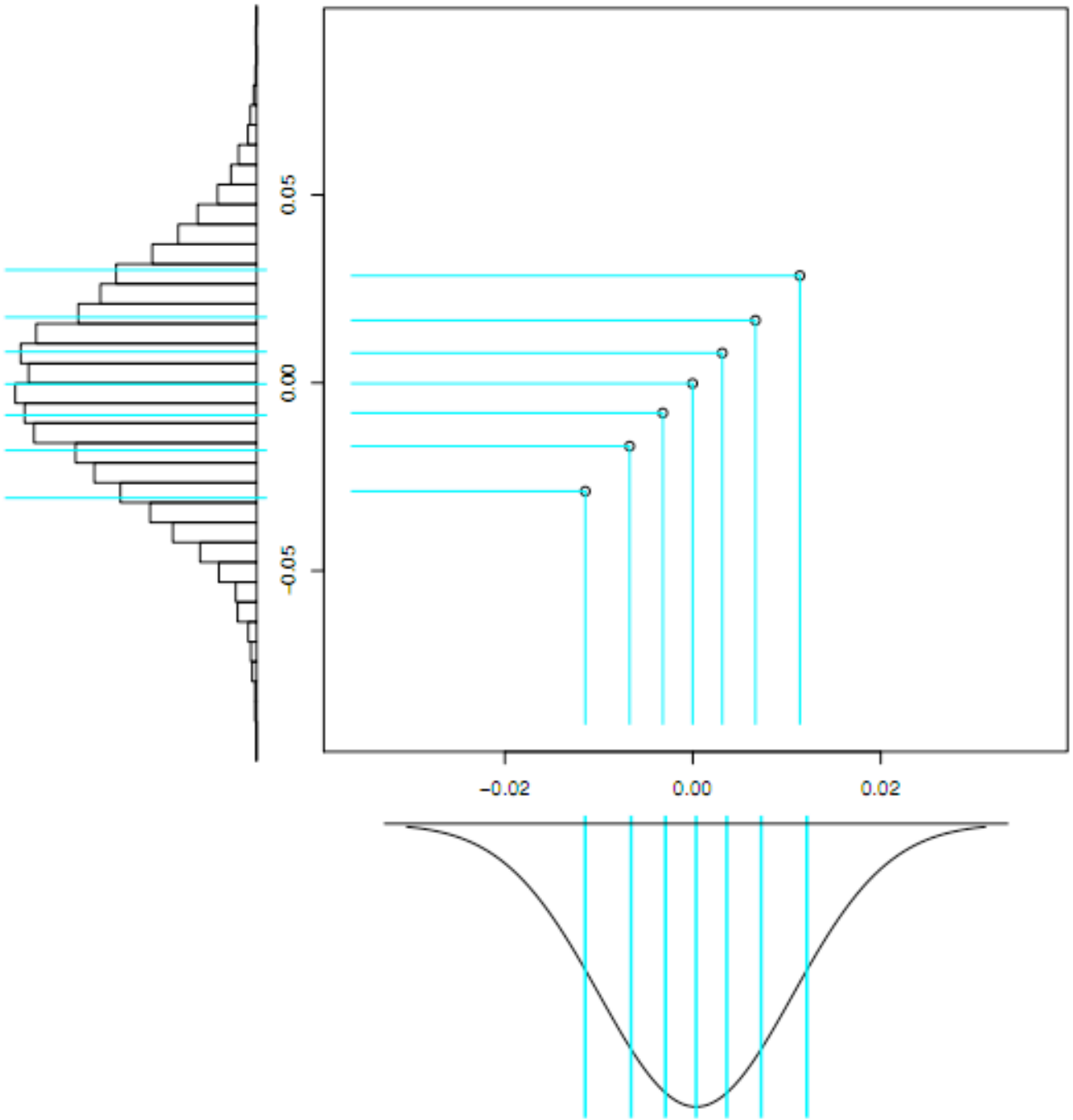
Aside: Normal quantile-quantile plots

A few lectures ago we saw how these were constructed and how they let us compare the “shape” of a collection of values (actual data, bootstrap replicates, anything) to the standard normal bell curve (mean zero and standard deviation 1) -- The construction is quite general and can be used to compare data to any known distribution

We saw that we could think about extrapolating a simple process whereby we compared quantiles computed from our data set against the corresponding quantiles for the standard normal distribution

We motivated the whole process by dividing the data into pieces and comparing it to regions of equal area under the normal curve...


0.05
0.00
−0.05
−0.02
0.00
0.02

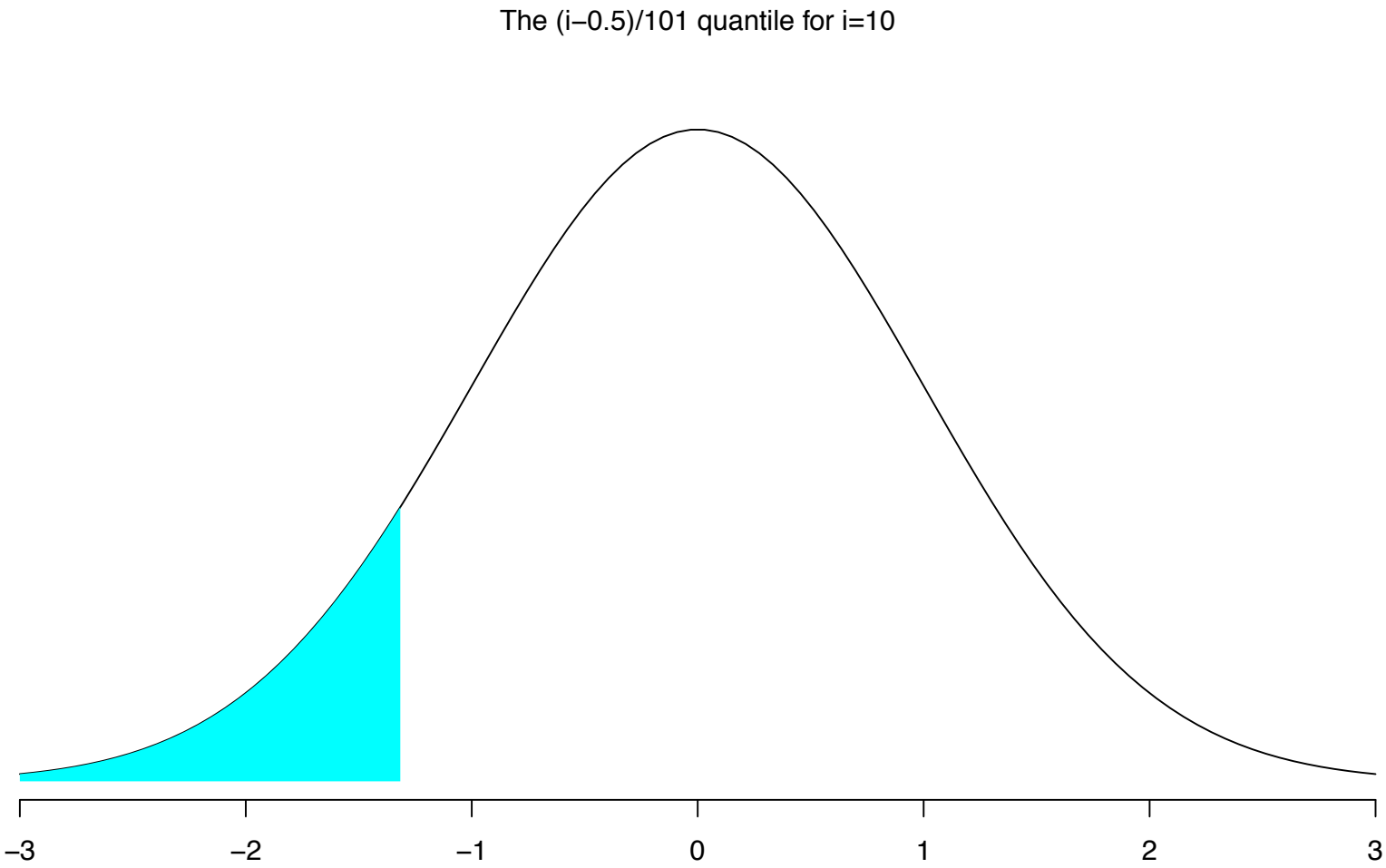

0.05
0.00
−0.05
−0.02
0.00
0.02

Normal quantile-quantile plots

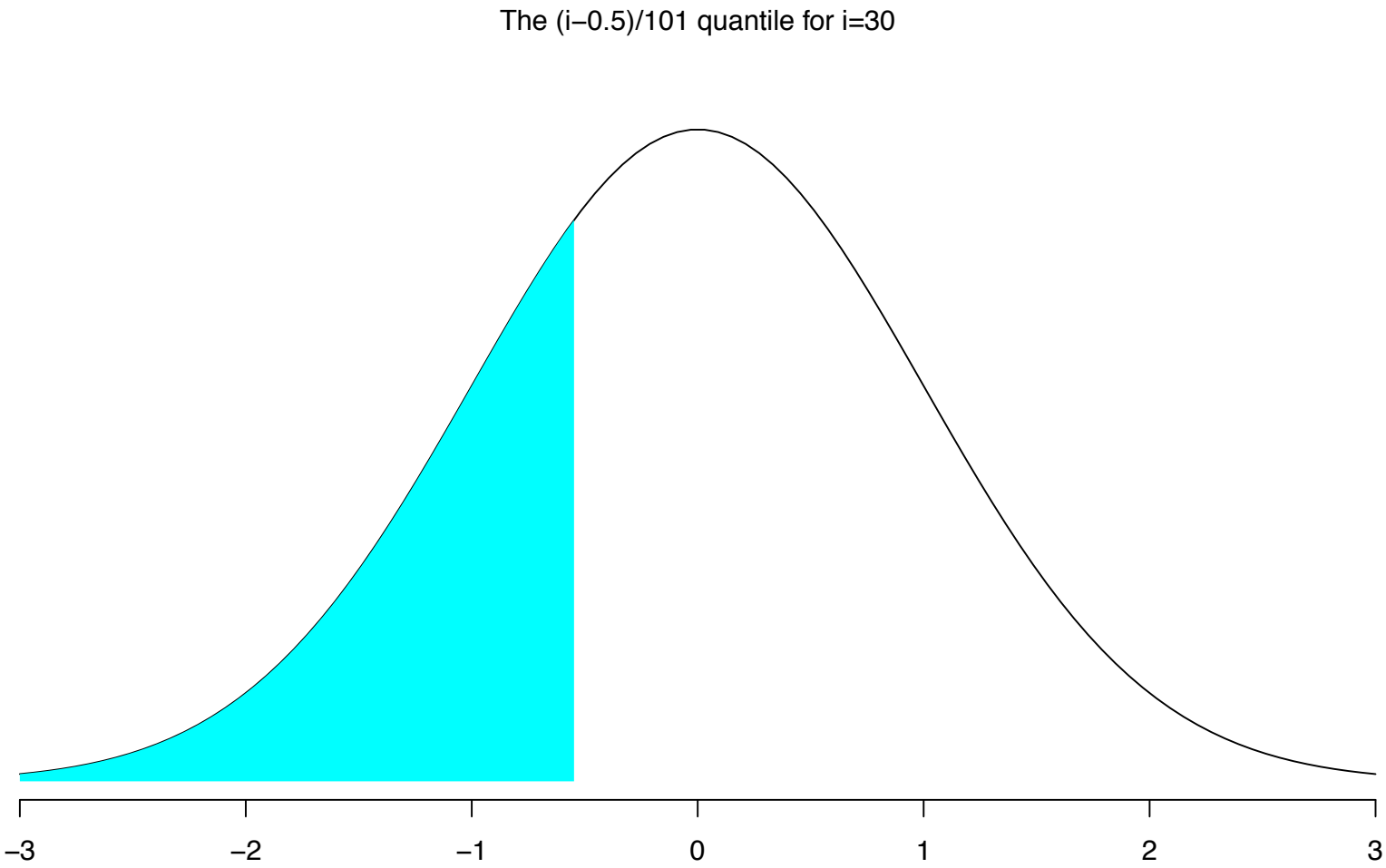
Technically, when R forms this plot using a data set with n values, it compares the **ith largest data point to the normal (i-0.5)/n quantile** -- If there are an odd number of points then the middle point corresponds to the center point of the standard normal distribution or 0

As an example, if n=101, say, then the 51st largest point is compared to the (51-0.5)/101 = 5.05/101 = 0.5 quantile of the standard normal, the point below which we see half of the area, or 0

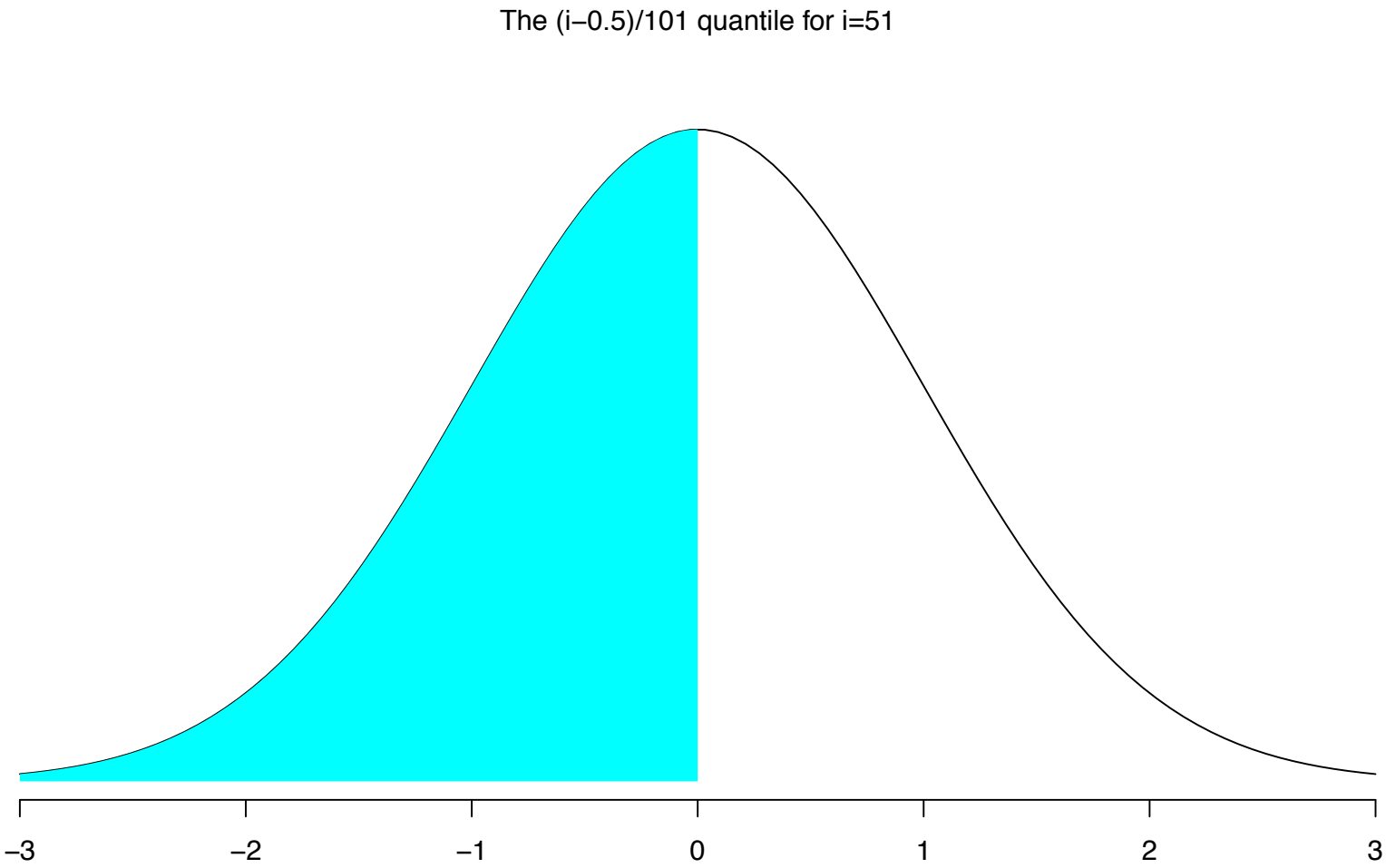
On the next few pages we make this clear...



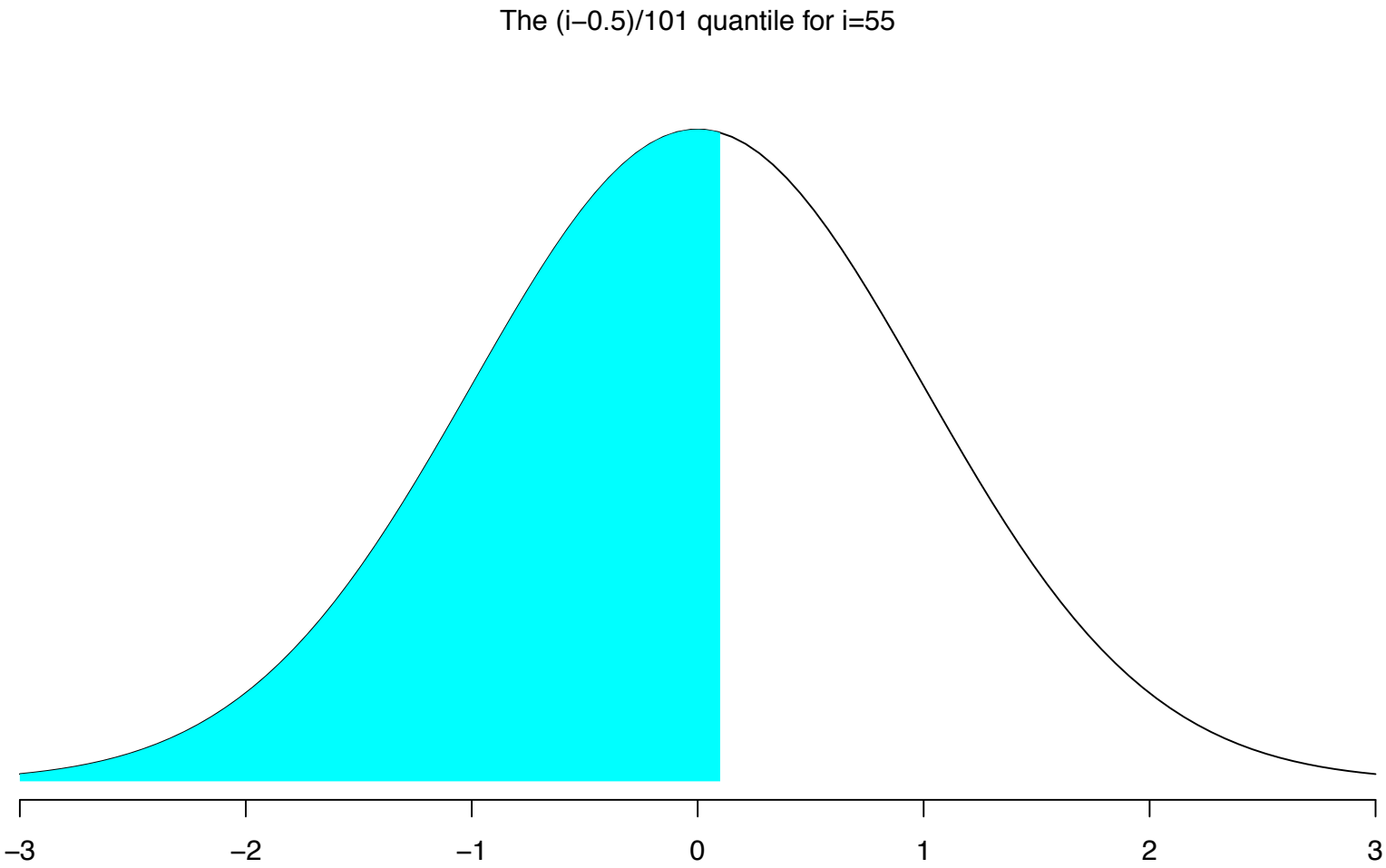
$(10 - 0.5)/101 = 0.094$ and $q_{0.094} = -1.32$



$(30 - 0.5)/101 = 0.292$ and $q_{0.292} = -0.55$



$(51 - 0.5)/101 = 0.5$ and $q_{0.5} = 0$

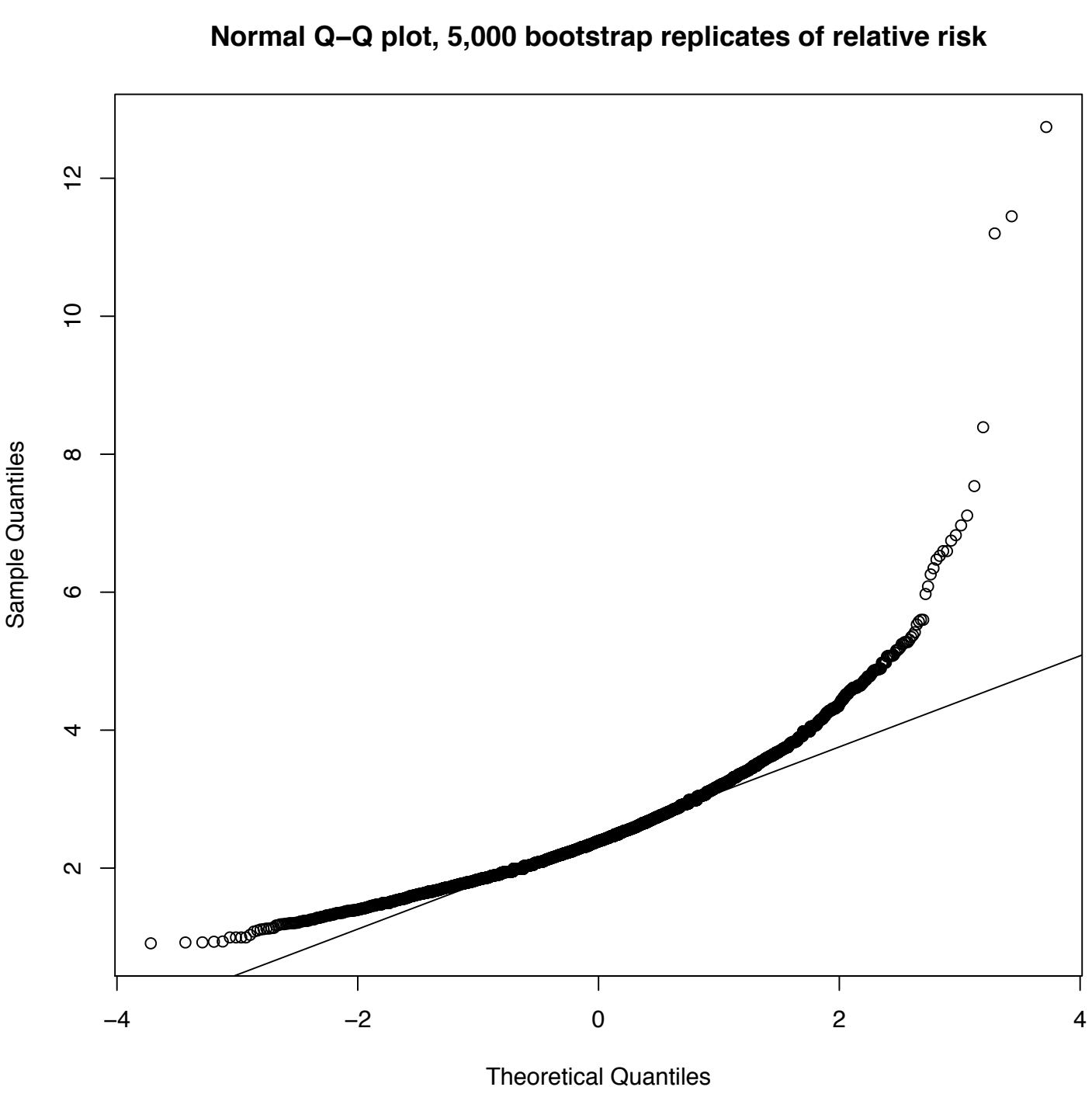


$(55 - 0.5)/101 = 0.540$ and $q_{0.540} = 0.099$

Normal quantile-quantile plots

At the time we commented that the nice thing about this display is that **we just have to compare it to a line** -- A straight line means that the data are distributed in a way that agrees with the bell curve

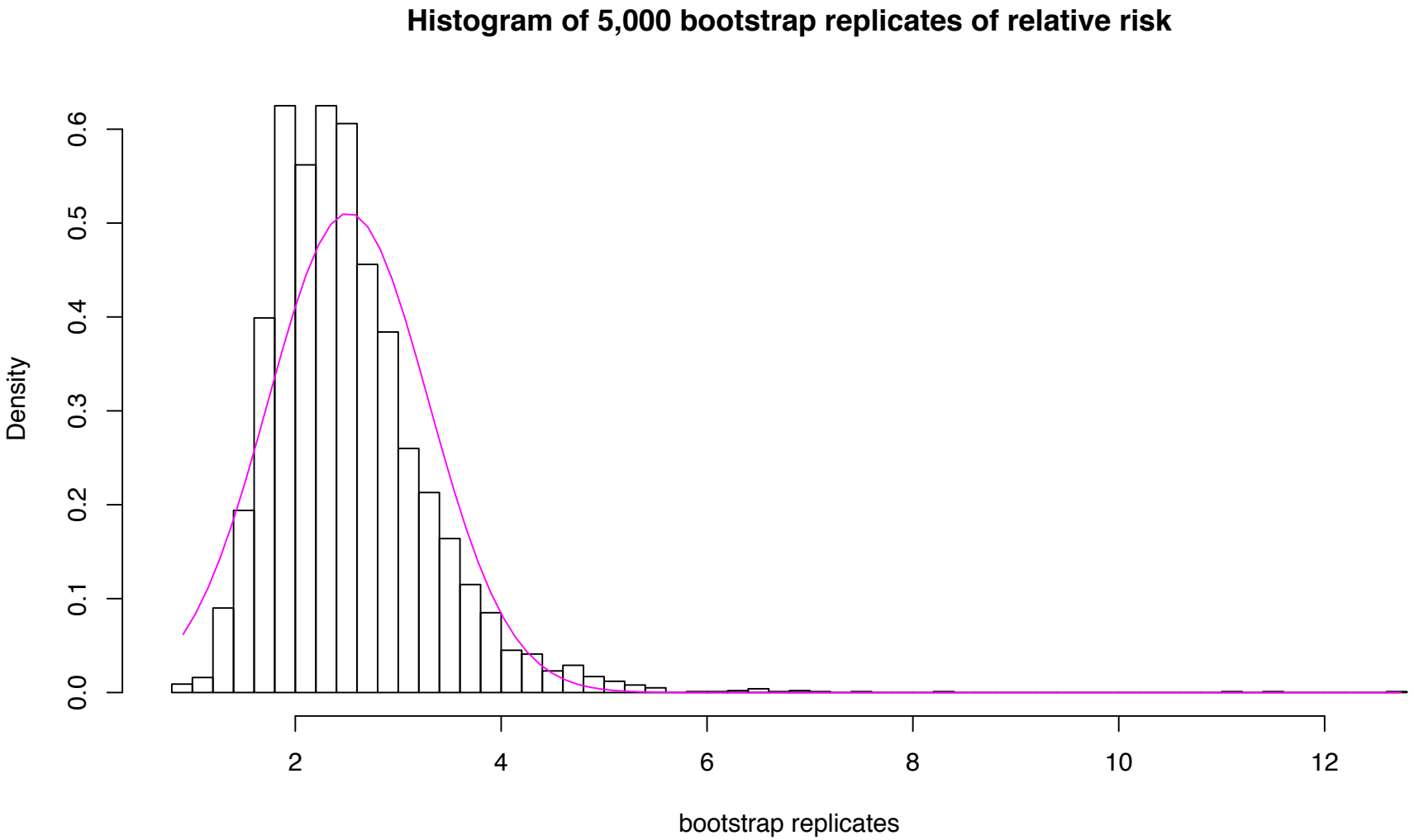
The line we've added here (with the function `qqline`) is drawn between the first and third quantiles...


Normal Q–Q plot, 5,000 bootstrap replicates of relative risk
Sample Quantiles
Theoretical Quantiles
2
4
6
8
10
12
−4
−2
0
2
4

Aside: Why just the standard normal?

So far, we've only been using the standard normal distribution as our "ruler" and not other members of the normal family (with different means and standard deviations)

When we added a normal curve over the top of a histogram, we used the normal curve with matching mean and variance...

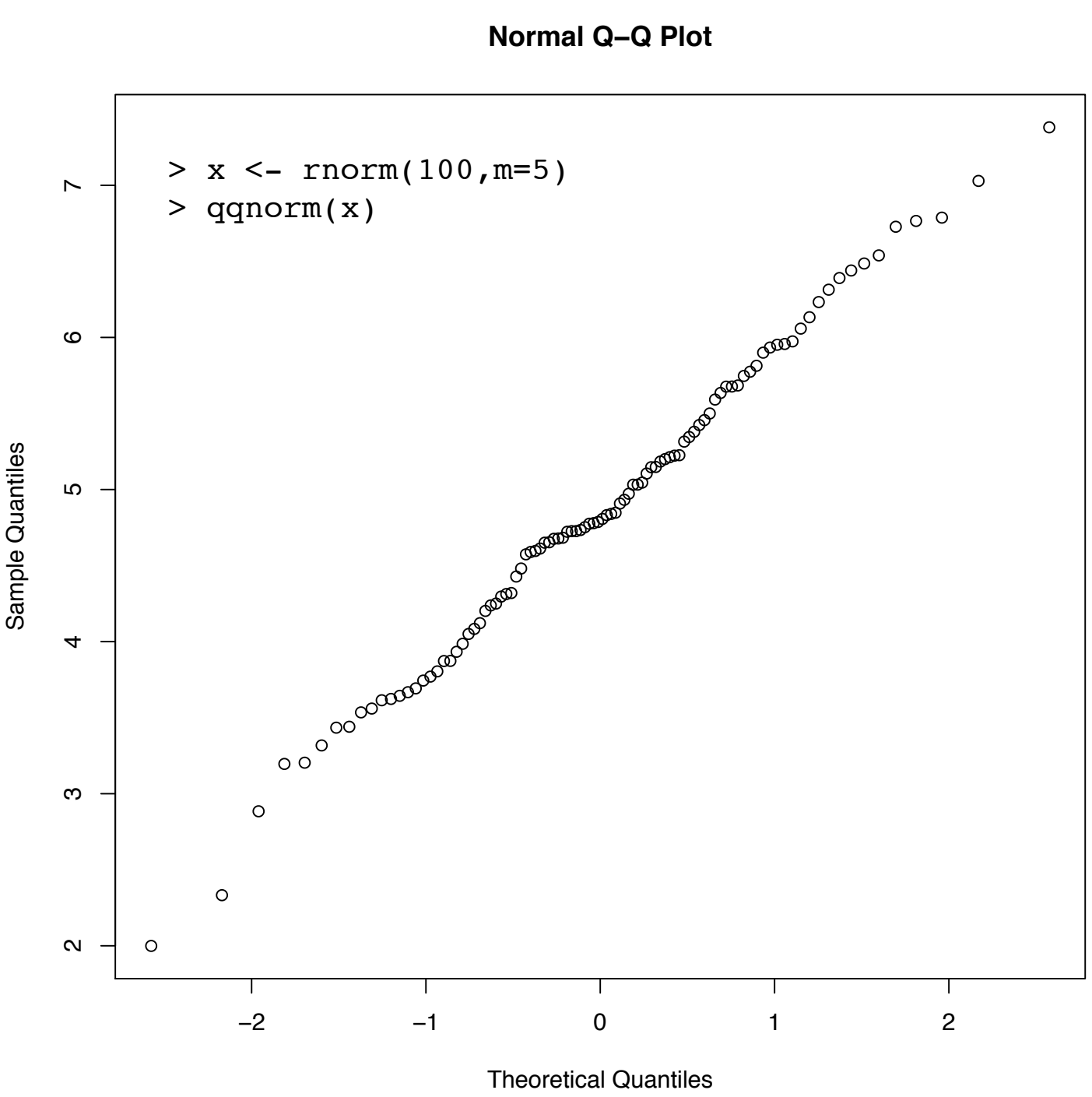

Histogram of 5,000 bootstrap replicates of relative risk
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2
4
6
8
10
12
bootstrap replicates

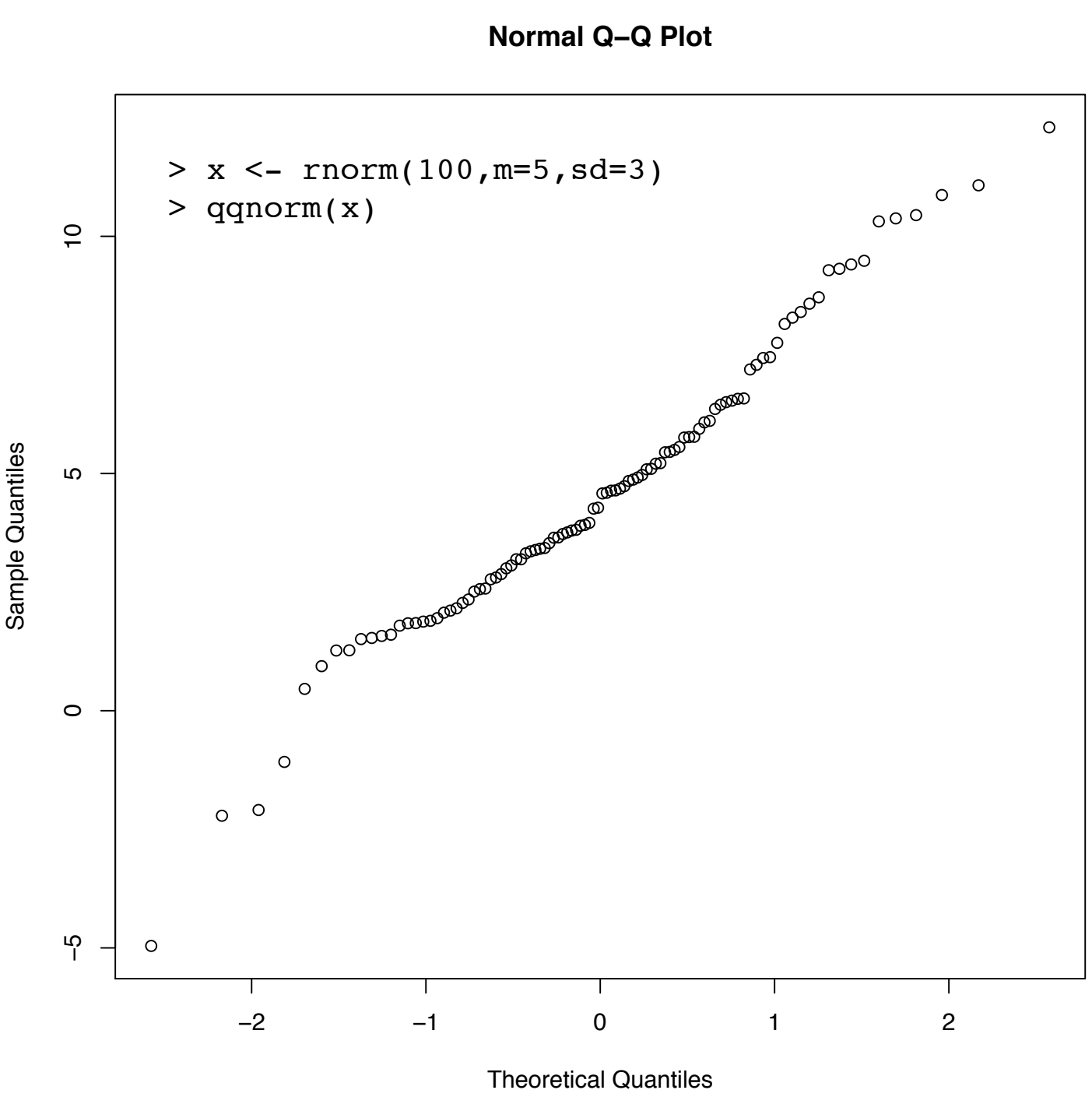
Aside: Why just the standard normal?

For a normal Q-Q plot, we don't have to compare our data to a normal with matched means and variances -- If our data were really normally distributed, then our **Q-Q points would follow a line with intercept at their mean and slope approximately their standard deviation**

Here's a couple examples -- The first is a Q-Q plot of 101 points sampled from the normal distribution with mean 5 and the second is sampled from the normal with mean 5 and standard deviation 3

R's scaling of these plots makes it hard to see the change in slope and intercept, but look at these plots closely...


Normal Q–Q Plot
> x <- rnorm(100,m=5)
> qqnorm(x)
Sample Quantiles
Theoretical Quantiles
2
3
4
5
6
7
–2
–1
0
1
2


Normal Q–Q Plot
> x <- rnorm(100,m=5,sd=3)
> qqnorm(x)
Sample Quantiles
Theoretical Quantiles
−5
0
5
10
−2
−1
0
1
2

Aside: Why just the standard normal?

To see why this is happening, consider the process we talked about in a previous lecture -- That is, taking a data set and “standardizing it”...

Aside: Why just the standard normal?

We saw two lectures ago that we could "standardize" a data set so that it had mean zero and standard deviation 1 -- To be precise, if our data values are denoted $x_1, \ldots, x_n$, then we can create new data $z_1, \ldots, z_n$ where

$$z_i = \frac{x_i - \bar{x}}{sd(x)}$$

where

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} \quad \text{and} \quad sd(x) = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

Now, the new data values $z_1, \ldots, z_n$ have mean zero and standard deviation 1 -- To see this, well, it's algebra...

A little algebra

To show the $z_1, \ldots, z_n$ have mean 0, we just follow our nose(s)

$$\begin{aligned}
\frac{z_1 + \cdots + z_n}{n} &= \frac{\frac{x_1 - \bar{x}}{sd(x)} + \cdots + \frac{x_n - \bar{x}}{sd(x)}}{n} \\
&= \frac{(x_1 - \bar{x}) + \cdots + (x_n - \bar{x})}{n\ sd(x)} \\
&= \frac{(x_1 + \cdots + x_n) - n\bar{x}}{n\ sd(x)} \\
&= \frac{0}{n\ sd(x)} \\
&= 0
\end{aligned}$$

A little algebra

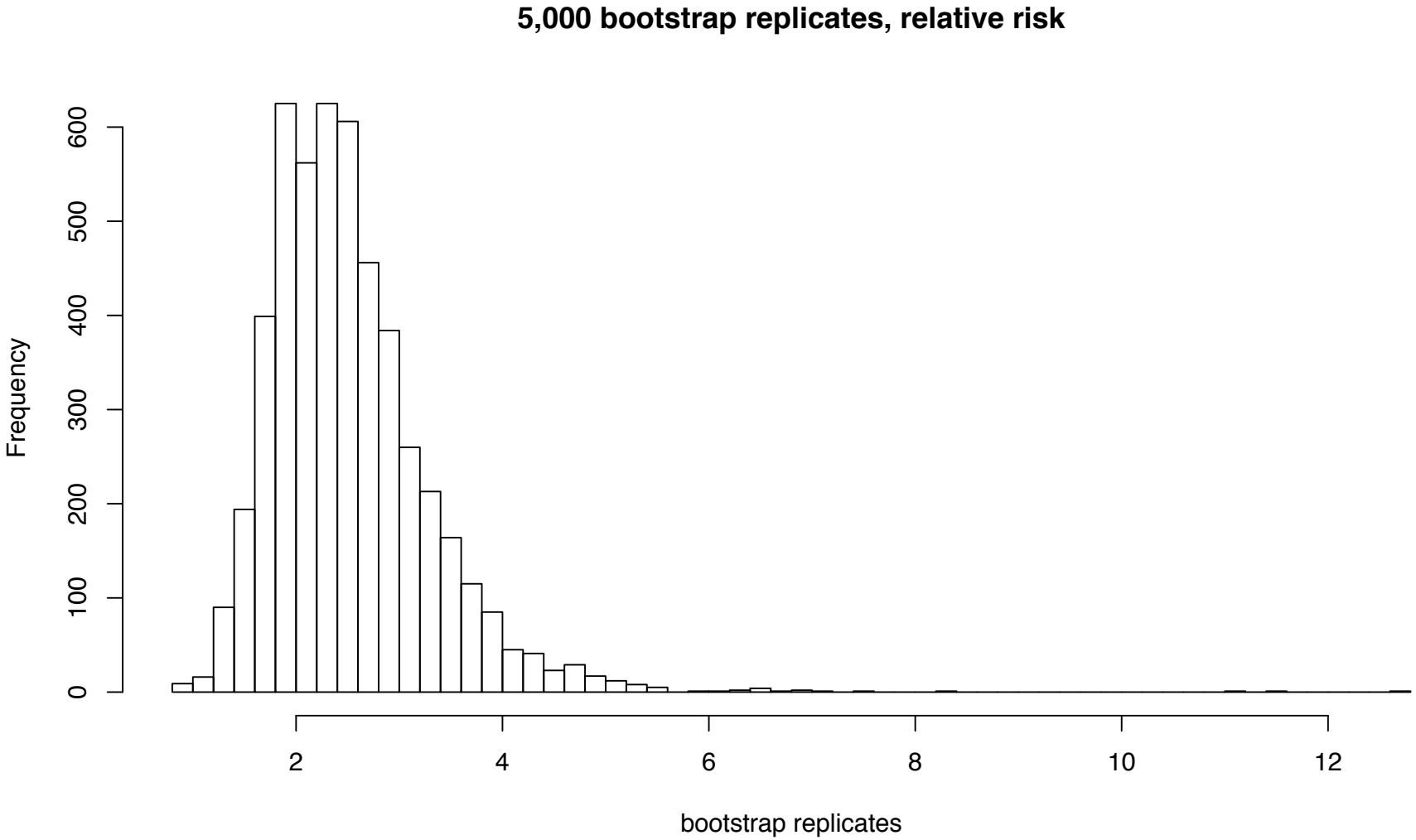
To show the $z_1, \ldots, z_n$ have standard deviation 1, we keep following...

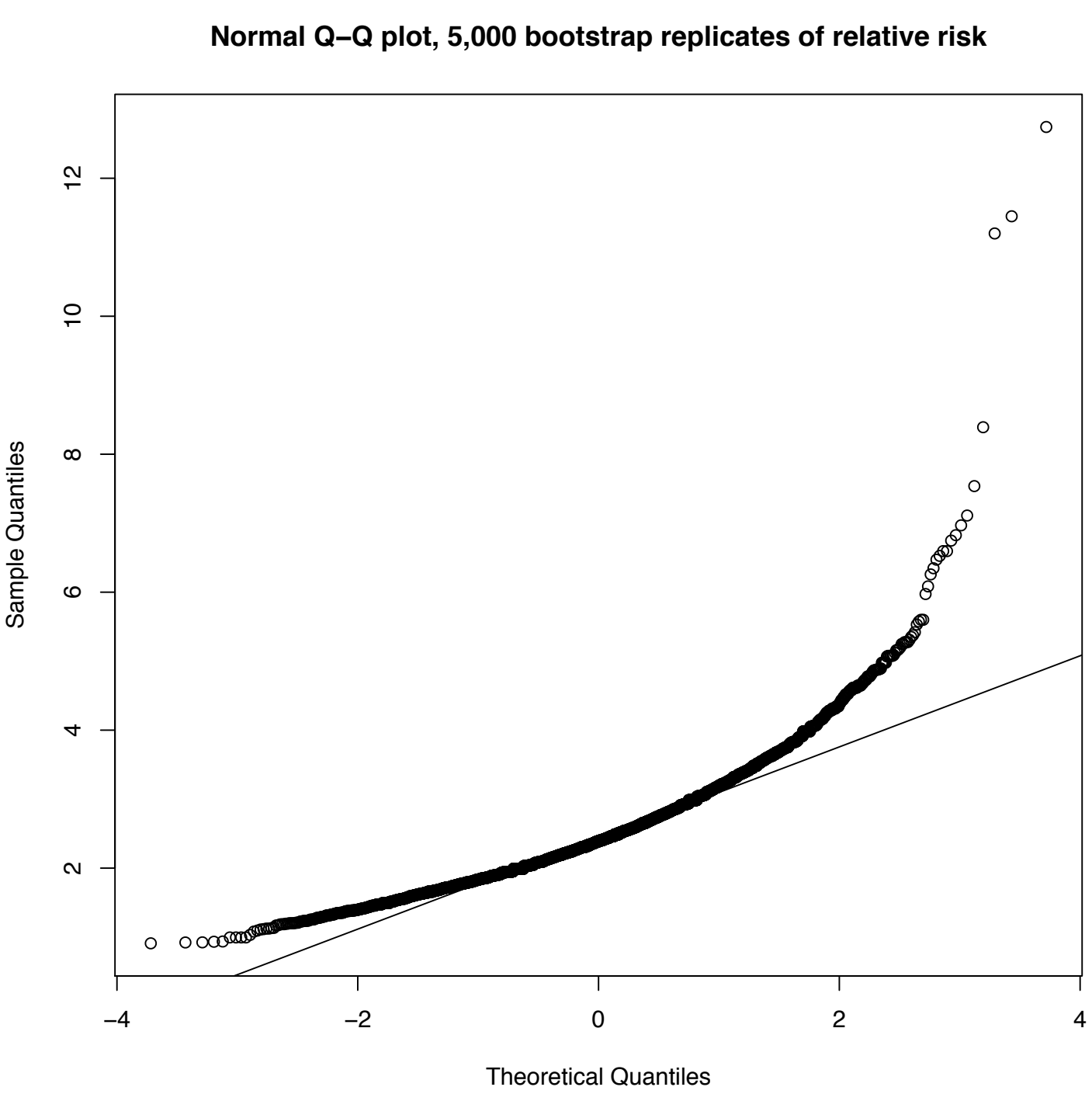
$$\begin{aligned}
\frac{z_1^2 + \cdots + z_n^2}{n-1} &= \frac{\left(\frac{x_1 - \bar{x}}{sd(x)}\right)^2 + \cdots + \left(\frac{x_n - \bar{x}}{sd(x)}\right)^2}{n-1} \\
&= \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{(n-1)\ sd^2(x)} \\
&= \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{(n-1)\ \left[\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}\right]} \\
&= 1
\end{aligned}$$

Aside: Why just the standard normal?

To sum, then, using the standard normal distribution as the “ruler” for normality implies that the intercept and slope of the implied Q-Q points (when they follow a line) tell us about the mean and standard deviation of our data

But let’s get back to our relative risks...


5,000 bootstrap replicates, relative risk
Frequency
0
100
200
300
400
500
600
2
4
6
8
10
12
bootstrap replicates


Normal Q–Q plot, 5,000 bootstrap replicates of relative risk
Sample Quantiles
Theoretical Quantiles
2
4
6
8
10
12
−4
−2
0
2
4

The bootstrap: Bias and SE

The distribution on the previous two slides tells us a lot about how we expect our estimation procedure to perform -- The **standard deviation** of the 5,000 bootstrap replicates provides an estimate of the measure of accuracy known as **the standard error**

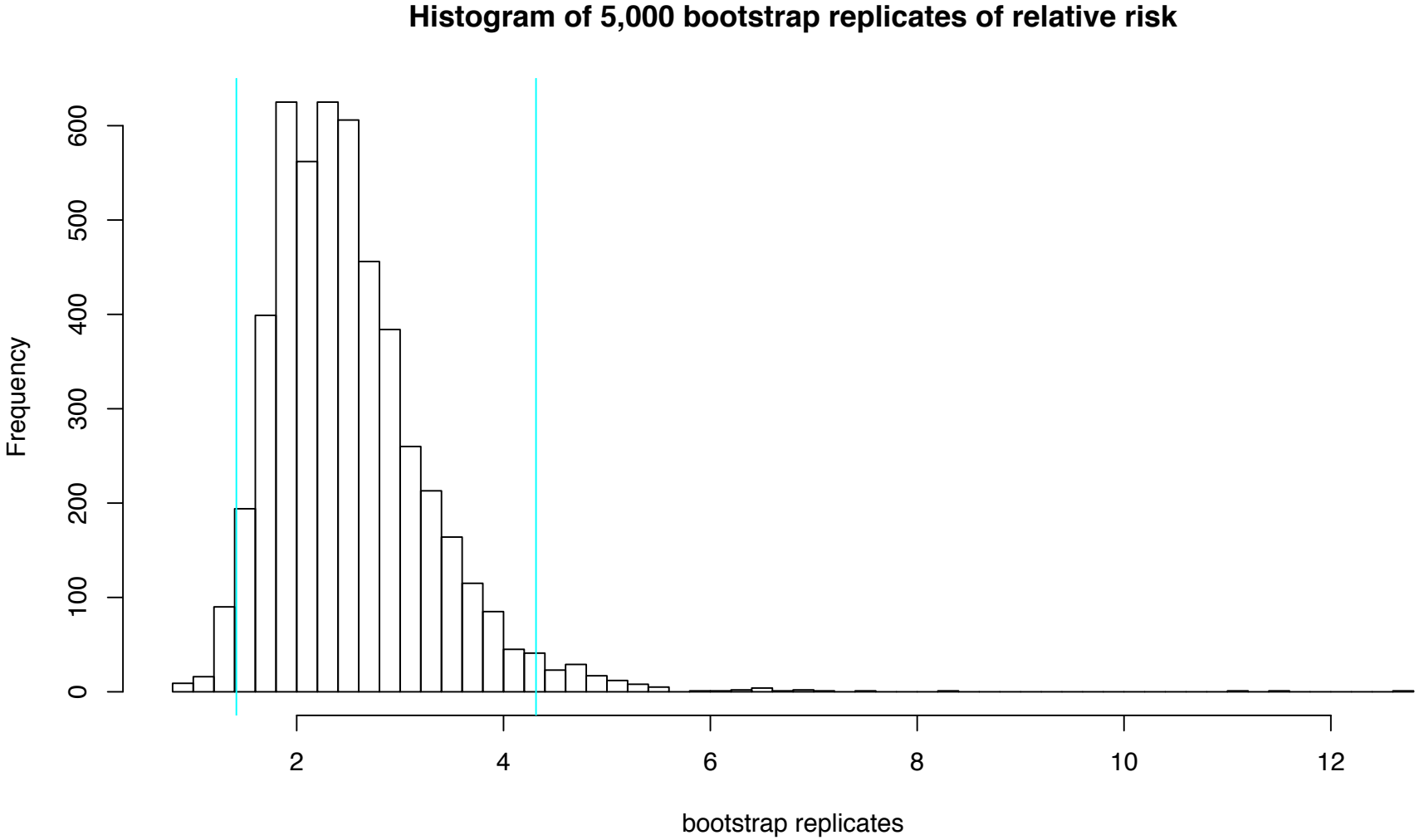
The mean of the 5,000 bootstrap replications can be compared to our real-world estimate of the relative risk to give us a sense of **the bias in our estimate** -- A rule of thumb is that if our bias is less than about a quarter of the standard error, we don't worry about it

```
# plot the bootstrap distribution

> hist(r,breaks=50)

# compare it to a normal distribution
# what do you see?

> qqnorm(r,breaks=50)
> qqline(r)

# the standard error of our estimate for relative risk

> sd(r)
[1] 0.7813014

# we get a sense of the bias in our estimate for relative
# risk by comparing the relative risk computed on our
# real world data to the mean of the bootstrap replicates

> mean(r)
[1] 2.51532

> (45/(45+4002))/(19/(4010+19))
[1] 2.357887

# the bias
> mean(r)-(45/(45+4002))/(19/(4010+19))
[1] 0.1574333
```

The bootstrap: Confidence intervals

Finally, we can form a 95% confidence interval for the population-based relative risk -- As we did last time, this can be formed by simply looking up the 0.025 and 0.975 quantiles of the bootstrap replicates...

```
# plot the bootstrap distribution

> hist(r,breaks=50)

# lower limit

> lo <-quantile(r,0.025)
> lo
    2.5%
1.416655

# upper limit

> hi <- quantile(r,0.975)
> hi
  97.5%
4.31406

# our 95% confidence interval is [1.4,4.3]

# add these to the plot

> abline(v=c(lo,hi),col=5)
```


Histogram of 5,000 bootstrap replicates of relative risk
Frequency
0
100
200
300
400
500
600
2
4
6
8
10
12
bootstrap replicates

The bootstrap: Confidence intervals

The confidence interval for the relative risk of heart attach on Vioxx versus Aleve among the population of patients with rheumatoid arthritis is [1.4,4.3] -- Importantly, this interval does not contain 1 as one of its “plausible” values

As we will see, confidence intervals are intimately related to hypothesis tests -- The hypothesis that Vioxx and Aleve carry the same risk of heart attack in the population of patients with rheumatoid arthritis is the same as testing if the relative risk is 1 in this population

We’ll see the relationship between testing and confidence intervals more formally in a later lecture -- For now, it’s enough to see that **we tend to look for notable values when we form a confidence interval**, and that other information like the their width tells us about how uncertain we are about the population parameter given our data

The bootstrap

The process, then, is somewhat straightforward -- **The computational procedure is intuitive**, with the performance of our estimator being judged by **simple summaries of the bootstrap distribution**

In this case, we are facing **considerable skew in the bootstrap replicates** -- This is because, as an estimate, **the relative risk sees variation in both its numerator and denominator**, with small values of the latter producing large relative risks

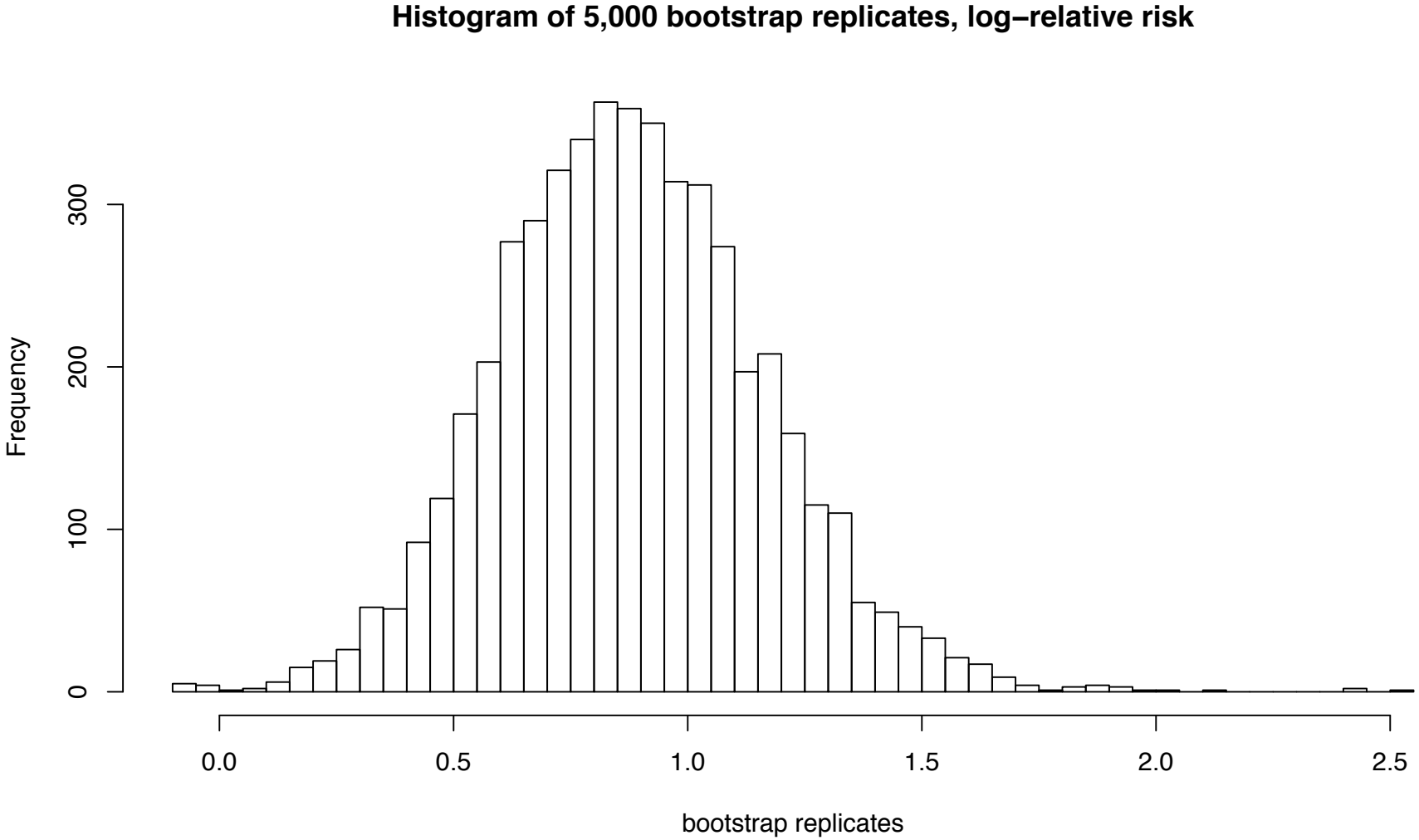
We can “improve” its performance by considering a transformation -- In this case, the transformation is motivated by sampling properties of the estimator

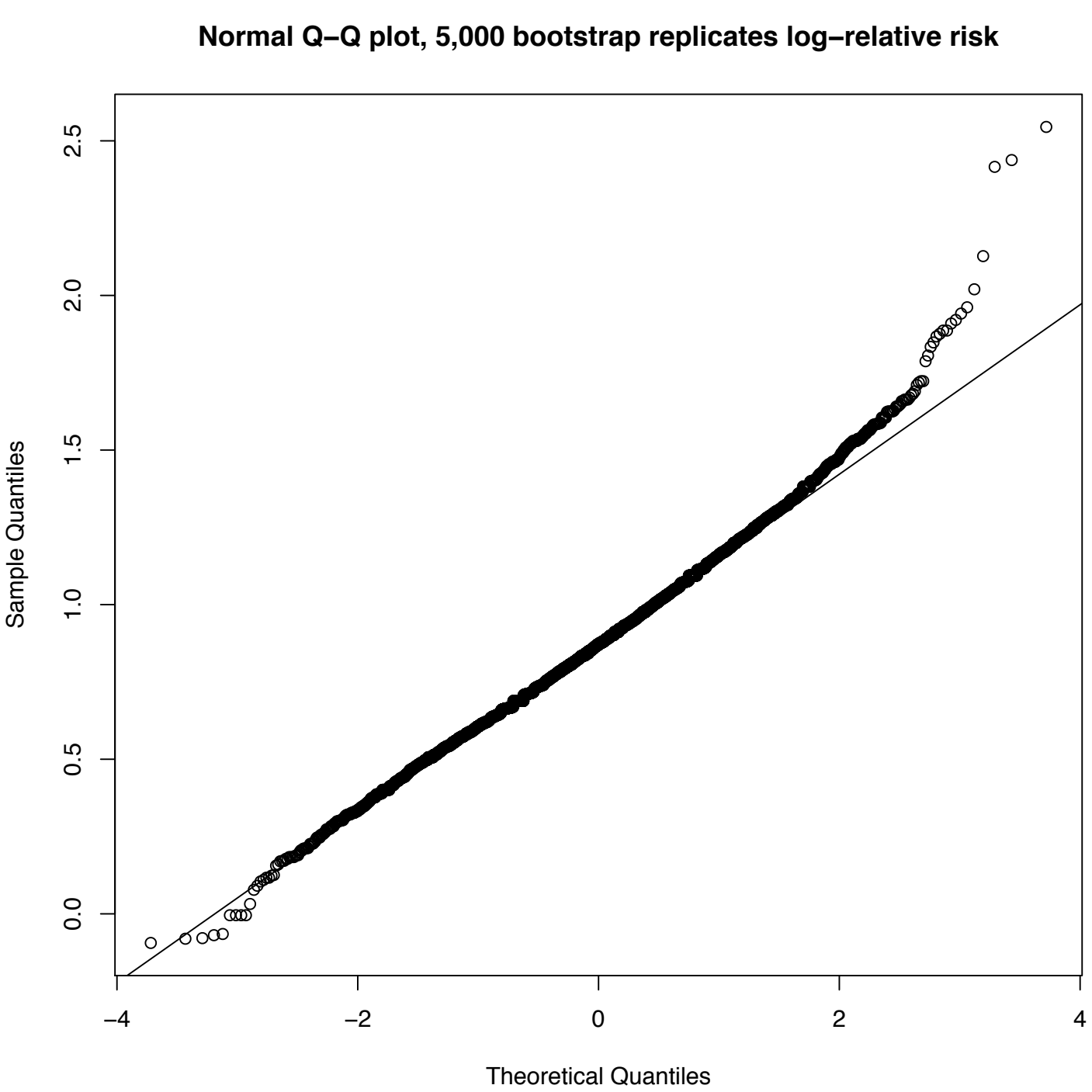
A transformation

So instead of the relative risk, let's look at **the logarithm of the relative risk** -- It is now an estimate of the logarithm of the relative risk of heart attack from Vioxx versus Aleve among rheumatoid arthritis patients (keep in mind that we are free to estimate with whatever statistic we like and the logarithm is invertible so we can move from logged to unlogged values easily)

We can use our same 5,000 bootstrap samples from the previous slides, but now look at their logarithms rather than the relative risks alone

```
# plot the bootstrap distribution
> hist(log(r),breaks=50)

# lower limit

> lo <-quantile(log(r),0.025)
> lo
     2.5%
0.3482982

# upper limit

> hi <- quantile(log(r),0.975)
> hi
   97.5%
1.461879

# our 95% confidence interval is [0.35,1.46]

# the bias -- comparing avg log-relative risk in bootstrap
# samples to log-relative risk from the experiment

> mean(log(r))
[1] 0.8806665

> log((45/(45+4002))/(19/(4010+19)))
[1] 0.8577659

# it's almost gone!

# and the standard error
> sd(log(r))
[1] 0.2842146
```


Histogram of 5,000 bootstrap replicates, log−relative risk
Frequency
0
100
200
300
0.0
0.5
1.0
1.5
2.0
2.5
bootstrap replicates


Normal Q–Q plot, 5,000 bootstrap replicates log–relative risk
Sample Quantiles
Theoretical Quantiles
0.0
0.5
1.0
1.5
2.0
2.5
−4
−2
0
2
4

The bootstrap: Transformations

Notice that by transforming, we have reduce the bias in our estimator -- We have reduced the effect of small values in the denominator by tucking them closer to the main body of the sampling distribution

This time, because our bootstrap distribution looks fairly normal (except for some wobble in the left tail), our quantile approach to confidence intervals agrees with adding and subtracting two standard errors from the relative risk in our experiment -- In this case, it's

```
# the experiment's value of relative risk
> log((45/(45+4002))/(19/(4010+19)))
[1] 0.8577659

# and minus/plus two standard errors

> log((45/(45+4002))/(19/(4010+19))) -2*sd(log(r))
[1] 0.2893367
> log((45/(45+4002))/(19/(4010+19))) +2*sd(log(r))
[1] 1.426195

# These agree closely with the quantile approach
```

The bootstrap: Transformations

Recall that transformations like the logarithm are monotone so that quantiles of our logged relative risk values will be the logarithm of the original-scale quantiles -- That means **the endpoints for our confidence intervals based on the transformed data are just the logarithms of the original intervals**

As we will see next time, the “classical” statistical approach makes extensive use of the fact that a sampling distribution is normal -- When they’re not, statisticians will try to find a transform that makes them normal so the classical techniques can be used

The bootstrap, on the other hand, is a general purpose tool that you can use even if the sampling distributions are not normal -- The bootstrap distribution is an extremely informative tool for analysis and the procedure can be consistently applied across a large set of problems

To sum

The bootstrap helps us assess the accuracy of our estimate by exhibiting an approximation to the sampling distribution

By examining its shape, its center and spread and specific quantiles, we can construct a number of summaries that help us express what we know about “plausible values” for the population parameter

Let’s see how this construction is used beyond this single Vioxx study...

Vioxx (epilogue)

In lectures and labs, we considered a number of Vioxx studies, some of which seemed to suggest that it was dangerous (causing various kinds of heart problems), while others were inconclusive

As an exercise in interpreting confidence intervals, let's consider an attempt to cull the knowledge from all these studies...

DRUG THERAPY

Rofecoxib, a New Cyclooxygenase 2 Inhibito Shows Sustained Efficacy, Comparable With Other Nonsteroidal Anti-inflammatory Drug

A 6-Week and a 1-Year Trial in Patients With Osteoarthritis

Kenneth Saag, MD, MSc; Desiree van der Heijde, MD; Chester Fisher, MD; Adil Samara, MD; Lisa D James Bolognese, MS; Rhoda Sperling, MD; Brian Daniels, MD; for the Osteoarthritis Studies Group

Introduction: Rofecoxib, a cyclooxygenase 2 inhibitor (sometimes known as a specific cyclooxygenase 2 inhibitor or Coxib), is used in osteoarthritis (OA). Published information indicates rofecoxib's improved gastrointestinal safety profile over nonselective nonsteroidal anti-inflammatory agents (NSAIDs).

Objective: To evaluate the efficacy and safety of rofecoxib in treating OA in 2 studies.

id, parallel- knee or hip ria and end rial in 736 coxib once laily, and a f rofecoxib

once daily with 50 mg of diclofen 693 patients.

Results: Rofecoxib, at 12.5 and 25 efficacy clinically comparable with by 3 primary end points according parability criteria. Both rofecoxib d provided significantly greater effica all primary end points at 6 weeks. B and diclofenac showed similar effic treatments were well tolerated.

Conclusions: Rofecoxib is effective once-daily dosing for 6 weeks and 1 generally safe and well-tolerated in OA and 1 year.

Arch Fam Med. 2000;9:1124-1134

ROFECOXIB, A SPECIFIC INHIBITOR OF CYCLOOXYGENASE 2, WITH CLINICAL EFFICACY COMPARABLE WITH THAT OF DICLOFENAC SODIUM

Results of a One-Year, Randomized, Clinical Trial in Patients with Osteoarthritis of the Knee and Hip

GRANT W. CANNON, JACQUES R. CALDWELL, PETER HOLT, BARRY McLEAN, BETH SEIDENBERG, JAMES BOLOGNESE, ELLIOT EHRICH, SUARABH MUKHOPADHYAY, and BRIAN DANIELS, for the ROFECOXIB PHASE III PROTOCOL 035 STUDY GROUP

Objective. **To compare the clinical efficacy of rofecoxib, a specific inhibitor of cyclooxygenase 2 (COX-2), with that of diclofenac in patients with osteoarthritis (OA) and to evaluate the safety and tolerability of rofecoxib.**

Methods. **We performed a randomized, double-blind, active comparator–controlled trial in 784 adults with OA of the knee or hip. Patients were randomized to cording to predefined statistical criteria, to that of 150 mg of diclofenac per day in this 1-year study. Specific inhibition of COX-2 provided therapeutic efficacy in OA.**

Nonsteroidal antiinflammatory drugs (NSAIDs) are widely used in the treatment of osteoarthritis (OA) (1,2). Although NSAIDs effectively control mild-to-moderate joint pain associated with OA, their use is

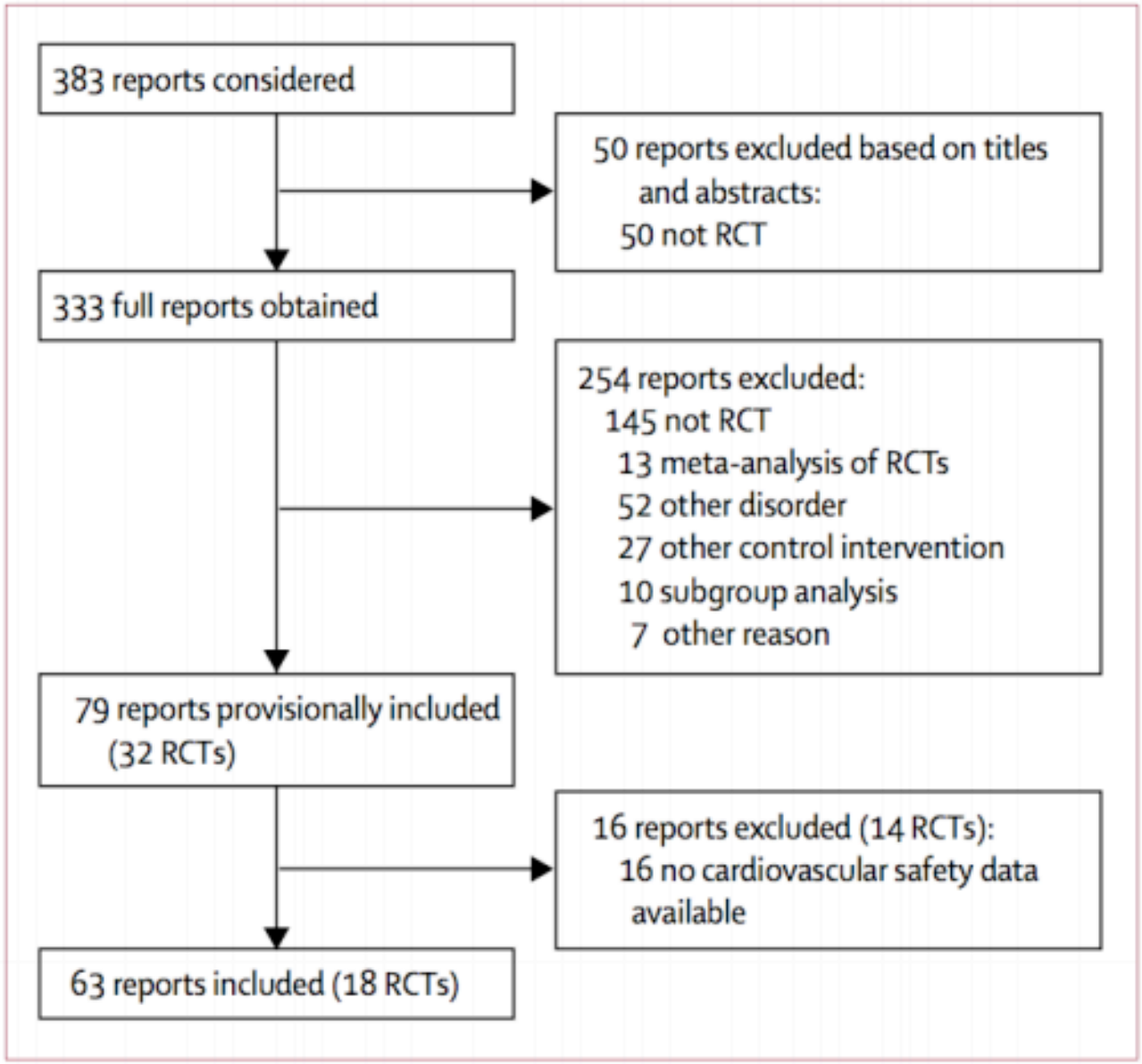
Meta-analysis

Meta-analysis is a tool for combining results across several related studies; the idea goes back to Karl Pearson (the gentleman who took on the roulette wheels of Monte Carlo) who was looking for a way to boost the power of several small studies (remember, power has a lot to do with sample size)

Cumulative meta-analysis applies this idea iteratively, folding in data whenever a new trial becomes available for inclusion; some researchers argue that cumulative meta-analysis can retrospectively identify the point in time **when a treatment effect first reached conventional levels of significance**

The next few pages present results from a paper entitled "Risk of cardiovascular events and rofecoxib: cumulative meta-analysis" by Peter Jüni, et a that appeared in The Lancet in 2004

(In setting up confidence intervals, we commented that you never actually have access to repeated experiments -- with meta-analysis, we find a way to achieve this replication, at least in spirit)


383 reports considered
50 reports excluded based on titles and abstracts:
50 not RCT
333 full reports obtained
254 reports excluded:
145 not RCT
13 meta-analysis of RCTs
52 other disorder
27 other control intervention
10 subgroup analysis
7 other reason
79 reports provisionally included (32 RCTs)
16 reports excluded (14 RCTs):
16 no cardiovascular safety data available
63 reports included (18 RCTs)

	Protocol number	Submitted to FDA (year)	Treated disorder (number of patients)	Intervention (number of patients)		Duration (weeks)
				Rofecoxib	Control	
Ehrich et al (1999)[13]	010	1998	Osteoarthritis (n=145)	Rofecoxib 25 mg (n=73)	Placebo (n=72)	6
Laine et al (1999)[14]	044	1998	Osteoarthritis (n=742)	Rofecoxib 25 mg (n=195)	Placebo (n=177)	24
				Rofecoxib 50 mg (n=186)	Ibuprofen 2400 mg (n=184)	
Schnitzer et al (1999)[14]	068	2001	Rheumatoid arthritis (n=500)	Rofecoxib 25 mg (n=171)	Placebo (n=168)	8
				Rofecoxib 50 mg (n=161)		
Extension of Schnitzer et al (1999)[15]	068-P2	2001	Rheumatoid arthritis (n=544)	Rofecoxib 25 mg (n=235)	Naproxen 1000 mg (n=86)	44
				Rofecoxib 50 mg (n=223)		
Bombardier et al (2000)[9]	088c	2000	Rheumatoid arthritis (n=8076)	Rofecoxib 50 mg (n=4047)	Naproxen 1000 mg (n=4029)	Up to 56
Cannon et al (2000)[16]	035	1998	Osteoarthritis (n=784)	Rofecoxib 12·5 mg (n=259)	Diclofenac 150 mg (n=268)	52
				Rofecoxib 25 mg (n=257)		
Day et al (2000)[17]	040	1998	Osteoarthritis (n=809)	Rofecoxib 12·5 mg (n=244)	Placebo (n=74)	6
				Rofecoxib 25 mg (n=242)	Ibuprofen 2400 mg (n=249)	
Hawkey et al (2000)[15]	045	1998	Osteoarthritis (n=775)	Rofecoxib 25 mg (n=195)	Placebo (n=194)	24
				Rofecoxib 50 mg (n=193)	Ibuprofen 2400 mg (n=193)	
Saag et al (2000)[18]	033	1998	Osteoarthritis (n=736)	Rofecoxib 12·5 mg (n=219)	Placebo (n=69)	6
				Rofecoxib 25 mg (n=227)	Ibuprofen 2400 mg (n=221)	
Saag et al (2000 A)[19]	034	1998	Osteoarthritis (n=693)	Rofecoxib 12·5 mg (n=231)	Diclofenac 150 mg (n=230)	52
				Rofecoxib 25 mg (n=232)		
Ehrich et al (2001)[20]	029	1998	Osteoarthritis (n=523)	Rofecoxib 12·5 mg (n=144)	Placebo (n=145)	6
				Rofecoxib 25 mg (n=137)		
				Rofecoxib 50 mg (n=97)		
Unpublished extension of Ehrich et al (2001)[20]	029-10	1998	Osteoarthritis (n=438)	Rofecoxib 12·5 mg (n=102)	Diclofenac 150 mg (n=90)	26
				Rofecoxib 25 mg (n=146)		
				Rofecoxib 50 mg (n=100)		
Geba et al (2001)[21]	090	2000	Osteoarthritis (n=978)	Rofecoxib 12·5 mg (n=390)	Placebo (n=196)	6
					Nabumetone 1000 mg (n=392)	
Truitt et al (2001)[22]	058	1998	Osteoarthritis (n=341)	Rofecoxib 12·5 mg (n=118)	Placebo (n=52)	6
				Rofecoxib 25 mg (n=56)	Nabumetone 1500 mg (n=115)	
Truitt et al (2001 A)[23]	096	2001	Rheumatoid arthritis (n=909)	Rofecoxib 12·5 mg (n=148)	Placebo (n=301)	12
				Rofecoxib 25 mg (n=311)	Naproxen 1000 mg (n=149)	
Unpublished extension of Truitt et al (2001 A)[23]	096-P2	2001	Rheumatoid arthritis (n=673)	Rofecoxib 25 mg (n=335)	Naproxen 1000 mg (n=224)	40
				Rofecoxib 50 mg (n=114)		
Geusens et al (2002)[24]	097	2001	Rheumatoid arthritis (n=1058)	Rofecoxib 25 mg (n=315)	Placebo (n=299)	12
				Rofecoxib 50 mg (n=297)	Naproxen 1000 mg (n=147)	
Unpublished extension of Geusens et al (2002)[24]	097-P2	2001	Rheumatoid arthritis (n=893)	Rofecoxib 25 mg (n=253)	Naproxen 1000 mg (n=248)	40
				Rofecoxib 50 mg (n=392)		
Hawkey et al (2003)[25]	098/103	-	Rheumatoid arthritis (n=660)	Rofecoxib 50 mg (n=219)	Placebo (n=221)	12
					Naproxen 1000 mg (n=220)	
Katz et al (2003)[26]	-	-	Chronic low back pain (n=690)	Rofecoxib 25 mg (n=233)	Placebo (n=228)	4
				Rofecoxib 50 mg (n=229)		
Lisse et al (2003)[11]	102	2000	Osteoarthritis (n=5586)	Rofecoxib 25 mg (n=2799)	Naproxen 1000 mg (n=2787)	12
Kivitz et al (2004)[12]	085	2000	Osteoarthritis (n=1042)	Rofecoxib 12·5 mg (n=424)	Placebo (n=208)	6
					Nabumetone 1000 mg (n=410)	

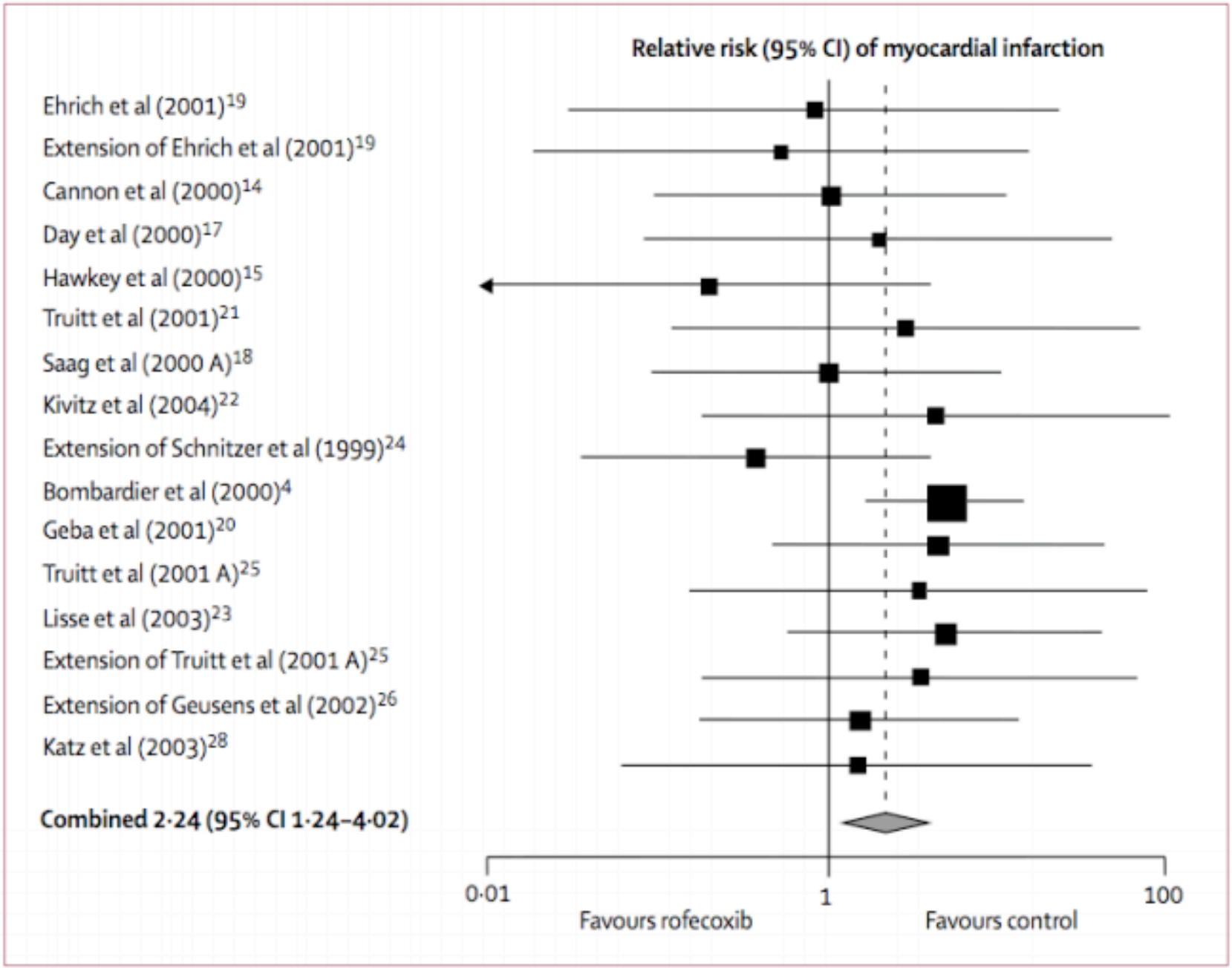
Table 1: Characteristics of randomised controlled trials and extensions of trials of therapeutic doses of rofecoxib in chronic musculoskeletal disorders

Confidence intervals

On the next page, we present Jüni's figure exhibiting the estimated relative risk and the associated confidence intervals for the population quantities

The size of the boxes reflect the number of patients enrolled in the study; the horizontal lines sweep out confidence intervals; for the relative risk (relative to the control group) what special value should we be looking at?

What do you think about the patterns you see on the next page?


Relative risk (95% CI) of myocardial infarction
Ehrich et al (2001)[19]
Extension of Ehrich et al (2001)[19]
Cannon et al (2000)[14]
Day et al (2000)[17]
Hawkey et al (2000)[15]
Truitt et al (2001)[21]
Saag et al (2000 A)[18]
Kivitz et al (2004)[22]
Extension of Schnitzer et al (1999)[24]
Bombardier et al (2000)[4]
Geba et al (2001)[20]
Truitt et al (2001 A)[25]
Lisse et al (2003)[23]
Extension of Truitt et al (2001 A)[25]
Extension of Geusens et al (2002)[26]
Katz et al (2003)[28]
Combined 2·24 (95% CI 1·24–4·02)
0·01
1
100
Favours rofecoxib
Favours control

Some issues

In theory, aggregating data from multiple trials should provide us with **pooled estimates that are better in terms of their accuracy and precision** (reducing bias and variability), but meta-analysis is not free of criticism

1. Combining data requires a leap of faith, **requiring the assumption that differences among studies are primarily due to chance**; in fact, differences in the direction or size of treatment effects may be caused by other factors, including subtle differences in treatments, populations, outcome measures, study design, and study quality

2. Also, it has been observed that studies with statistically significant outcomes (remember P-values and the 0.05 threshold?) **are more likely to be published than non-significant studies** (for randomized trials, significant results are three times more likely to be published!); also, negative studies can take significantly longer to appear in print, meaning cumulative analysis will inevitably generate an inflated and unduly precise estimate of a treatment's effects

That said...

We will have a look at the cumulative meta-analysis performed by Jüni et al; again, the idea is that we aggregate study results as they are published

What does the Jüni's figure suggest?

Year	Patients	Events	p
1997	523	1	0·916
1998	615	2	0·736
	1399	5	0·828
	2208	6	0·996
	2983	8	0·649
	3324	9	0·866
1999	4017	12	0·879
	5059	13	0·881
2000	5193	16	0·855
	13 269	40	0·070
	14 247	44	0·034
	15 156	46	0·025
	20 742	52	0·010
2001	20 742	58	0·007
	20 742	63	0·007
	21 432	64	0·007