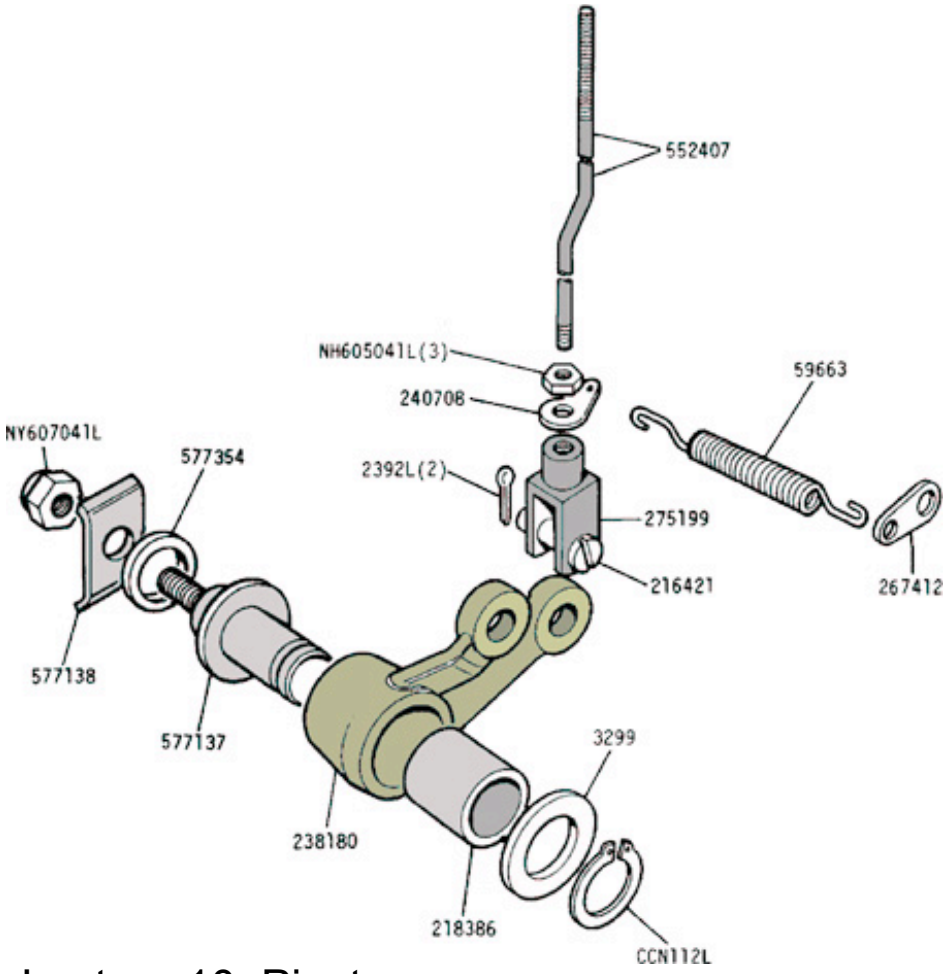


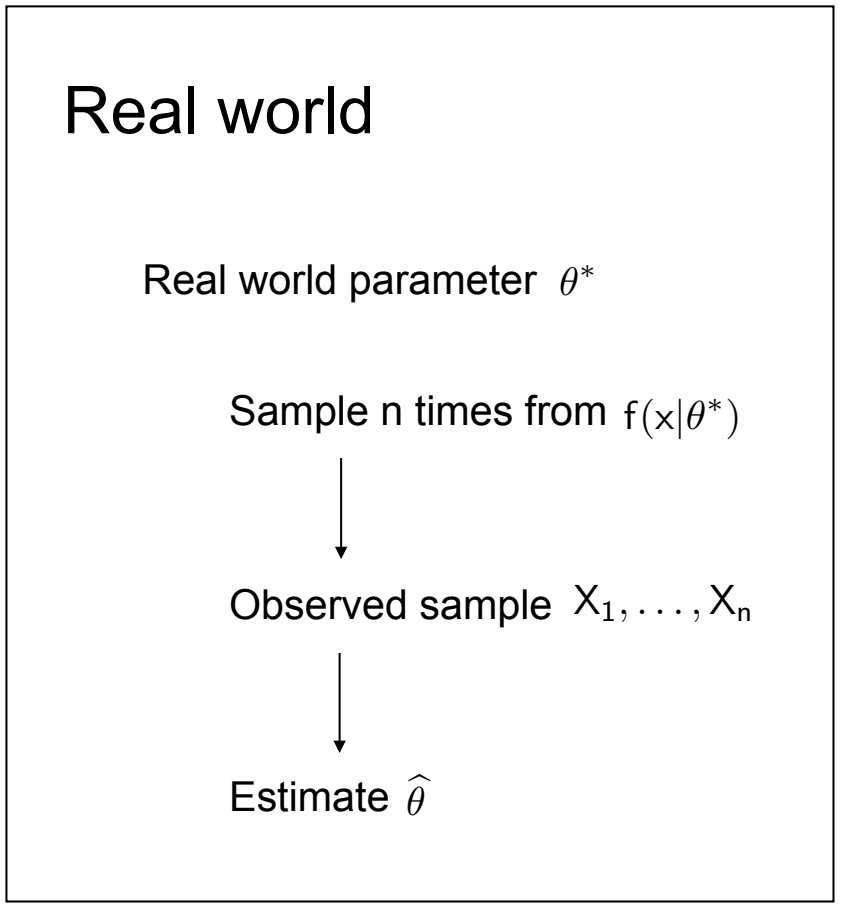
Lecture 10: Pivot

Last time

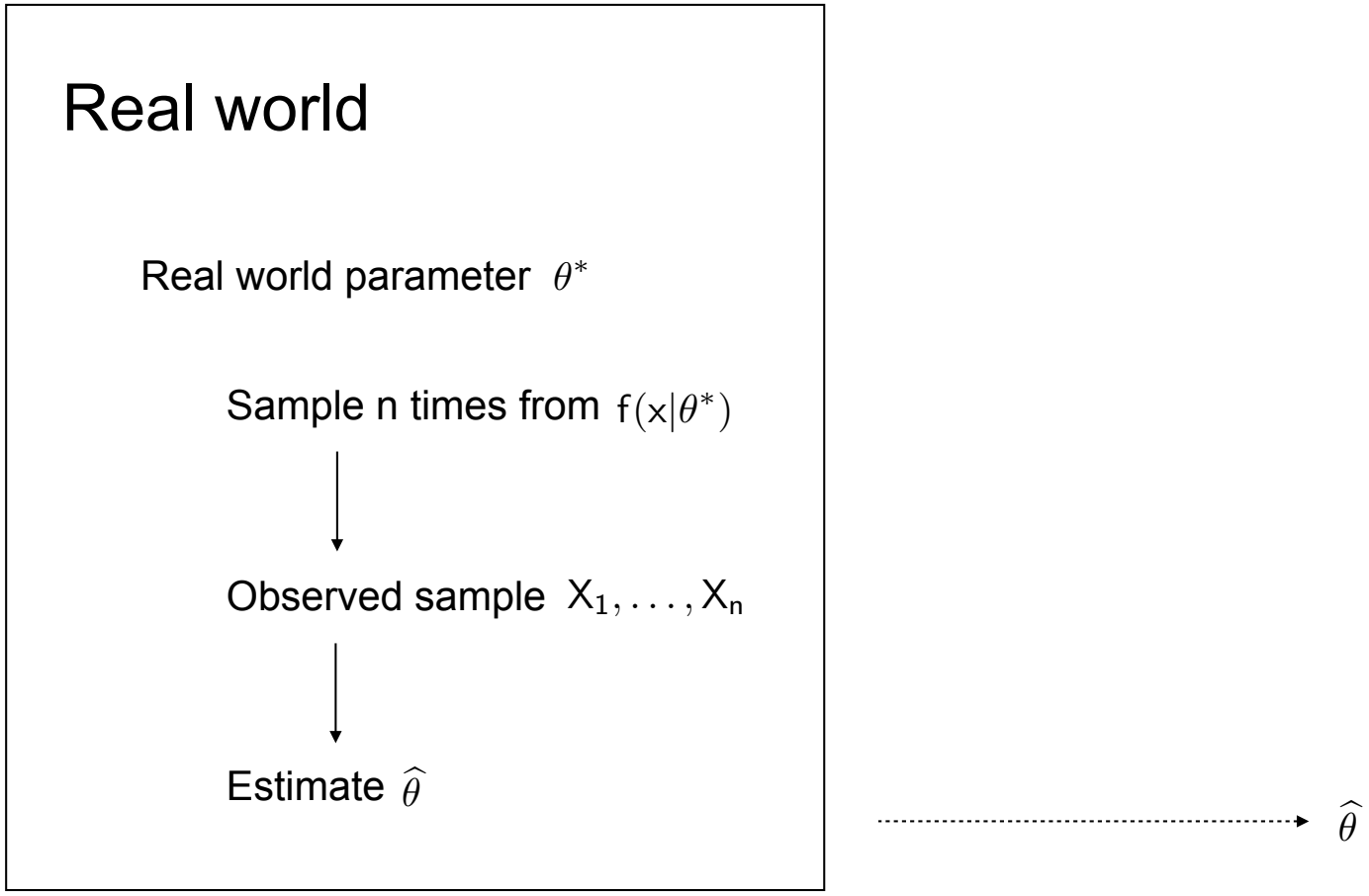
We reviewed the concept of the **sampling distribution of an estimator** -- We developed a handful of properties of estimators including their bias, standard error and mean squared error

We then introduced the notion of a **confidence interval**, a construction that let us quote not just a point estimate but range of “plausible” values for an unknown parameter -- By presenting a confidence interval, we express the uncertainty in our data about the unknown parameter

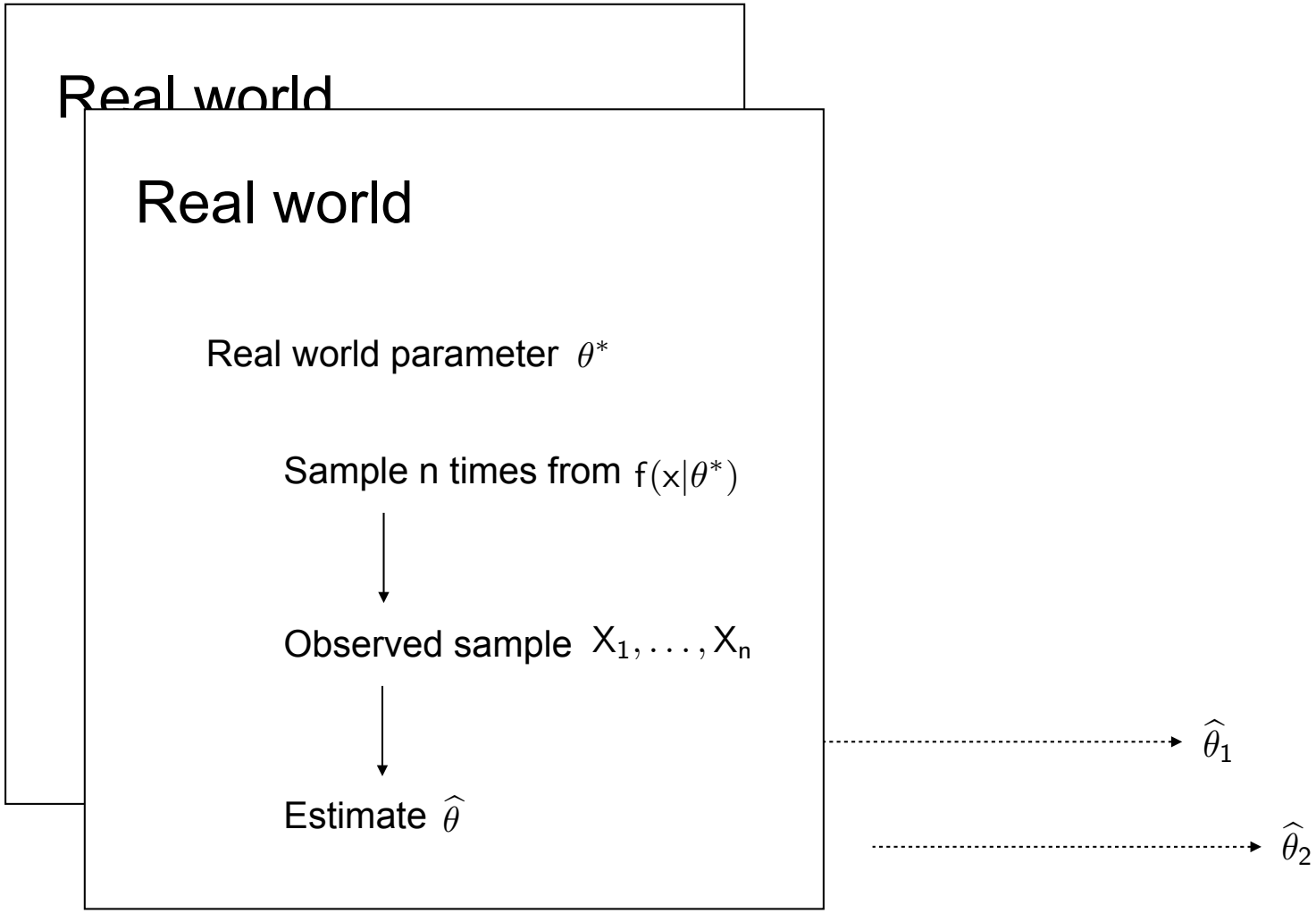
Let’s review for a moment...



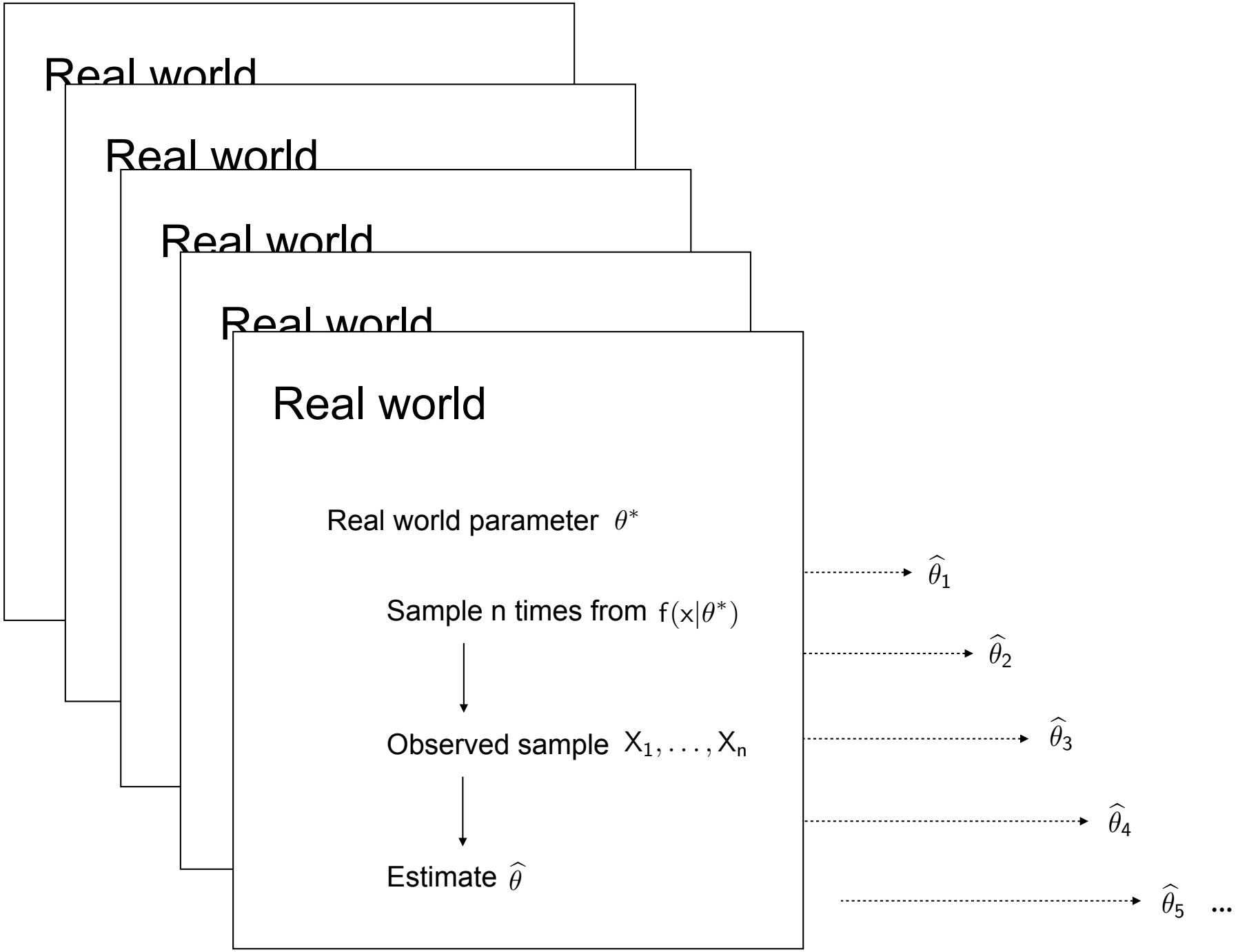
A sketch of a single experiment



A sketch of a single experiment and an estimate



Repeating the experiment produces new data and a new estimate...

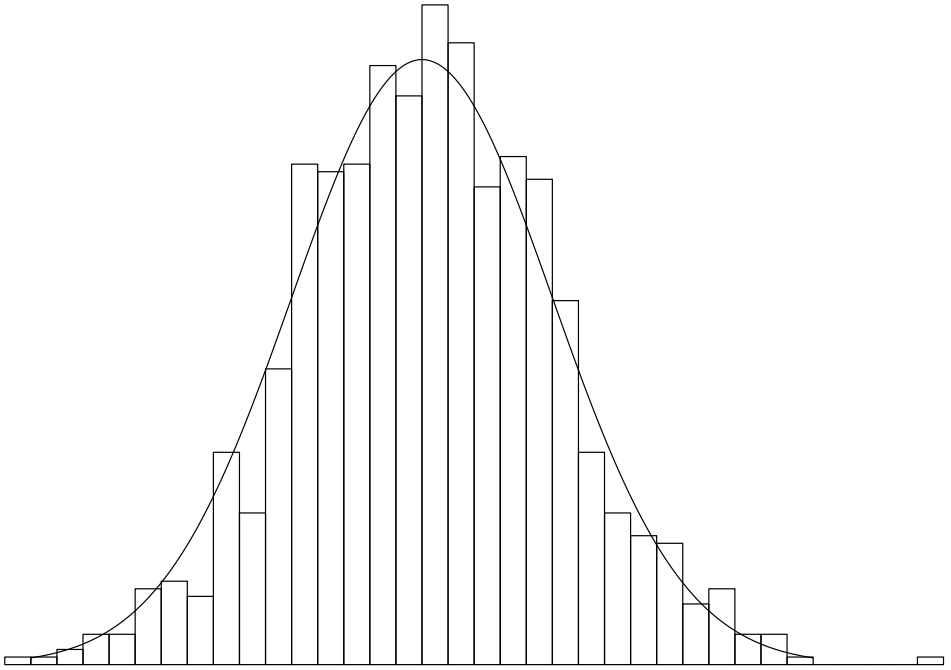


Repeating the experiment produces new data and a new estimate...

The sampling distribution

The distribution of the estimates computed from repeating our experiment multiple times is known as t**he sampling distribution** -- As a theoretical quantity, it tells us about how well our estimate is performing

Last time, we examined the mean of this distribution for bias in an estimate, used the spread to quantify the precision of an estimate, and introduced a construction that could be used to suggest “plausible” values for the unknown parameter given our data



A simple case

We ended the last lecture with a simple example -- Suppose we have n independent observations $X_1, \ldots, X_n$ from the normal distribution with mean μ and variance σ^2

You know from your probability class that the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ has a normal distribution (exactly) with mean μ and standard deviation $\sigma/\sqrt{n}$

Finally, again appealing to your probability course, we know that the quantity

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution -- Let's put this result to work!

A simple case

Because approximately 95% of the mass of the standard normal distribution is between +/-2

$$P\left(-2 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 2\right) \approx 0.95$$

we can move the downstairs in the fraction out to give

$$P\left(-2\sigma/\sqrt{n} \leq \overline{X} - \mu \leq 2\sigma/\sqrt{n}\right) \approx 0.95$$

or with one more move

$$P\left(\overline{X} - 2\sigma/\sqrt{n} \leq \mu \leq \overline{X} + 2\sigma/\sqrt{n}\right) \approx 0.95$$

A simple case

Suppose we are given n samples $X_1, \ldots, X_n$ from the normal distribution with **unknown mean** μ and **known standard deviation** σ

The MLE for μ is just the sample mean

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

From our results on the previous slides, we know the sampling distribution of our MLE μ exactly -- It is normal with mean μ and standard deviation $\sigma/\sqrt{n}$

Using our terminology from the last couple lectures, we say that $\overline{X}$ is unbiased and that its standard error is $\sigma/\sqrt{n}$

A simple case

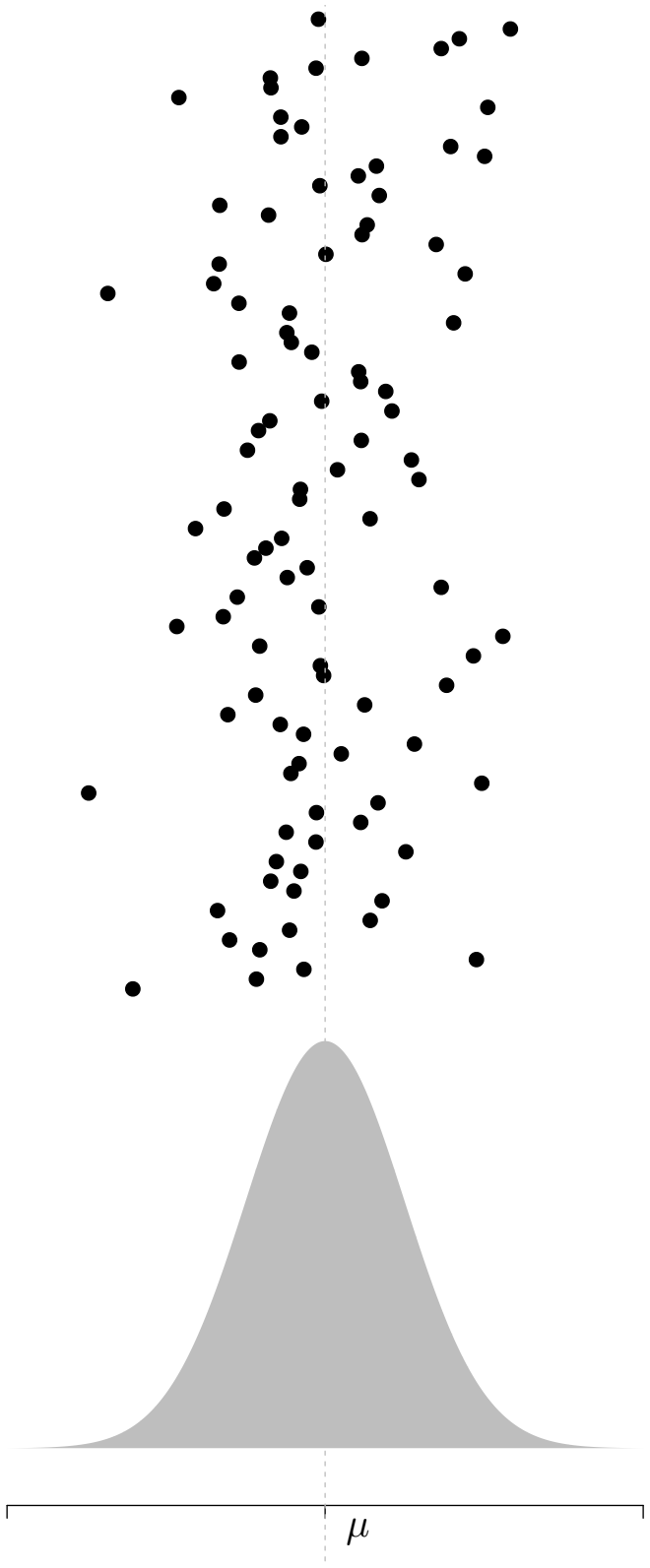
Finally, here is a 95% confidence interval for μ

$$\left[\overline{\mathsf{X}} - 2\sigma/\sqrt{\mathsf{n}},\ \overline{\mathsf{X}} + 2\sigma/\sqrt{\mathsf{n}}\right]$$

Confidence intervals

To make this concrete, at the top each black dot represents an experimental result -- We generated a sample and formed the MLE $\overline{X}$

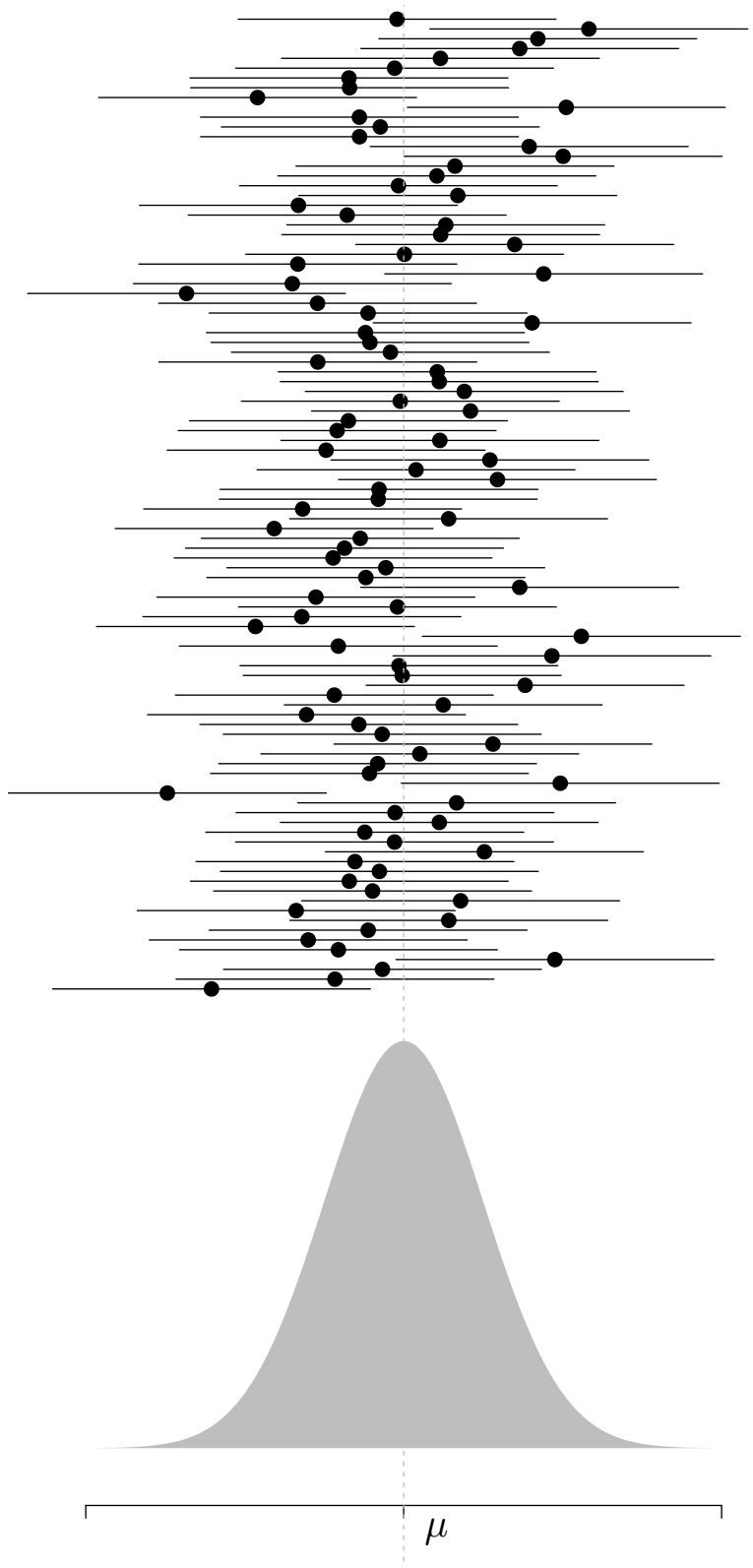
There are 100 black dots, representing 100 different sets of experimental outcomes -- The black dots are observations then from the sampling distribution and we can think of them as $\overline{X}_1, \ldots, \overline{X}_{100}$



Confidence intervals

For each estimate, or, rather, each time we perform our experiment, we can then form a 95% confidence interval $\overline{\mathrm{X}} \pm 2\sigma/\sqrt{\mathrm{n}}$

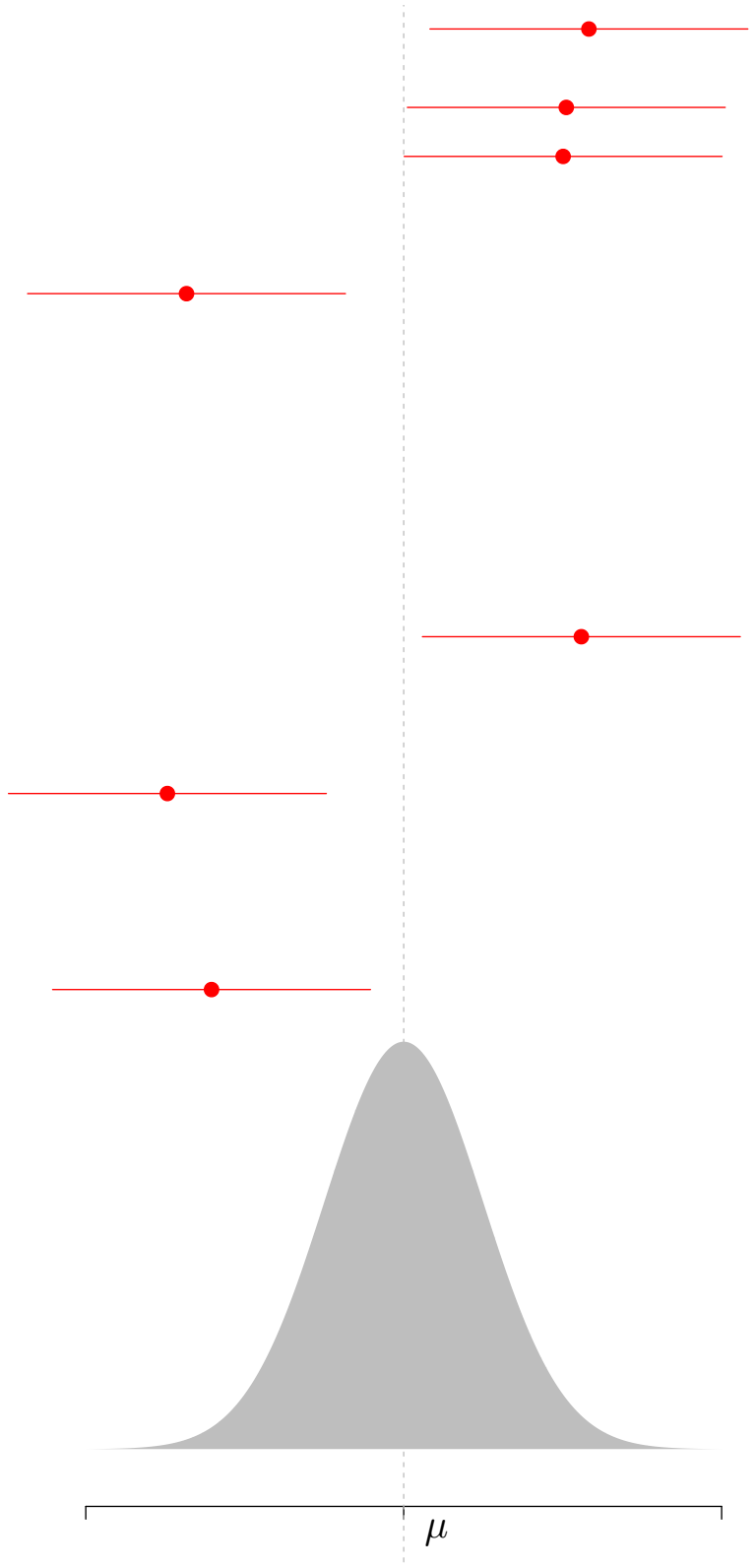
Given their construction, 95% of these intervals should cover the true parameter μ -- What do you think?



Confidence intervals

Of the 100 times we repeated our experiment, 7 (meh, about 95%) of the intervals we constructed failed to contain the true value of

This, then, is our notion of confidence -- We construct a rule based on the sampling distribution such that across repeated experiments 95% of the intervals will contain the true parameter μ that generated the data



A snag

This simple example isn't quite practical because we rarely have situations in which we know σ -- And while our interest may be in the mean, we still need to estimate to form a confidence interval

As we've seen, we can estimate σ using the MLE (or method of moments estimate)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

or the unbiased alternative

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

What impact does "plugging in" an estimate for σ have on our confidence interval?

Some history

Gosset decided to study the sampling distribution for the sample mean, or rather a "standardized" quantity

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

focusing, in particular, on cases for small values of n, but **assuming the population itself was normal**

His approach was novel; he decided to come up with **an exact expression for the sampling distribution, but under a strict assumption about the population** (one that he felt matched the experimental conditions he was seeing)

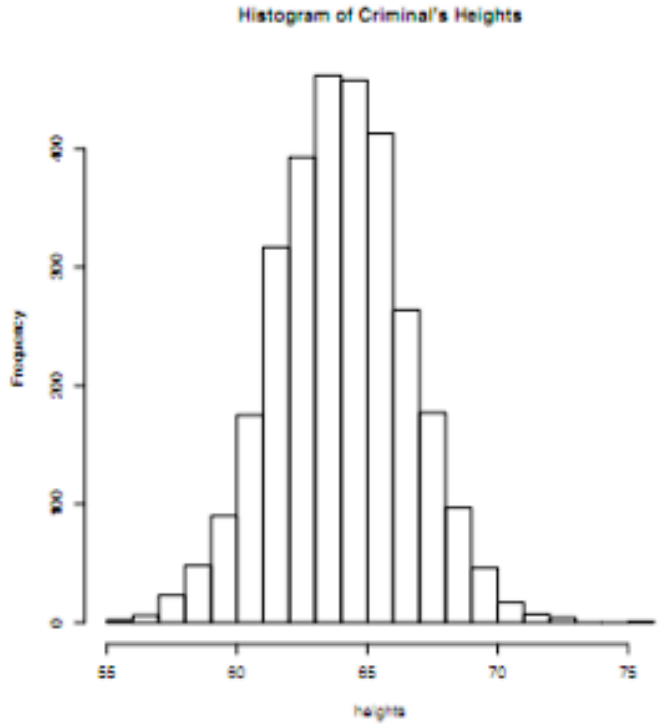
Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing ***the height and left middle finger measurements of 3000 criminals****, from a paper by W. R. Macdonell. The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. As each card was drawn its numbers were written down in a book which thus contains t****he measurements of 3000 criminals in a random order****. Finally* ***each consecutive set of 4 was taken as a sample*** *- 750 in all - and the mean, standard deviation and correlation of each sample determined.* ***The difference between the mean of each sample and the mean of the population was then divided by the standard deviation of the sample...***

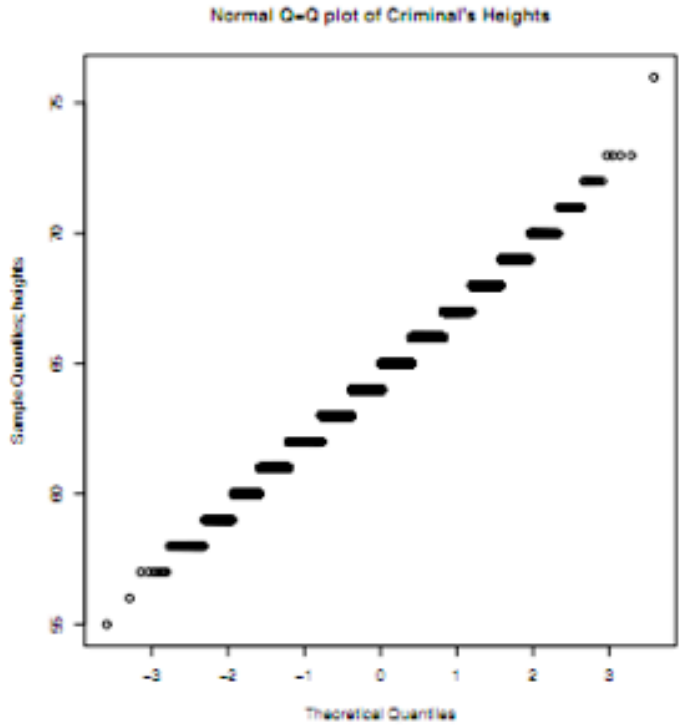
By Student

Here are Gosset's data using a couple of displays we're now very familiar with, a histogram and a normal Q-Q plot

Keep in mind these plots represent the ***entire population***; from this collection of 3,000 numbers we will draw samples (take surveys)

What do you notice?





By Student

We know the population $\mu = 64.5$ mean and the population standard deviation $\sigma = 2.6$

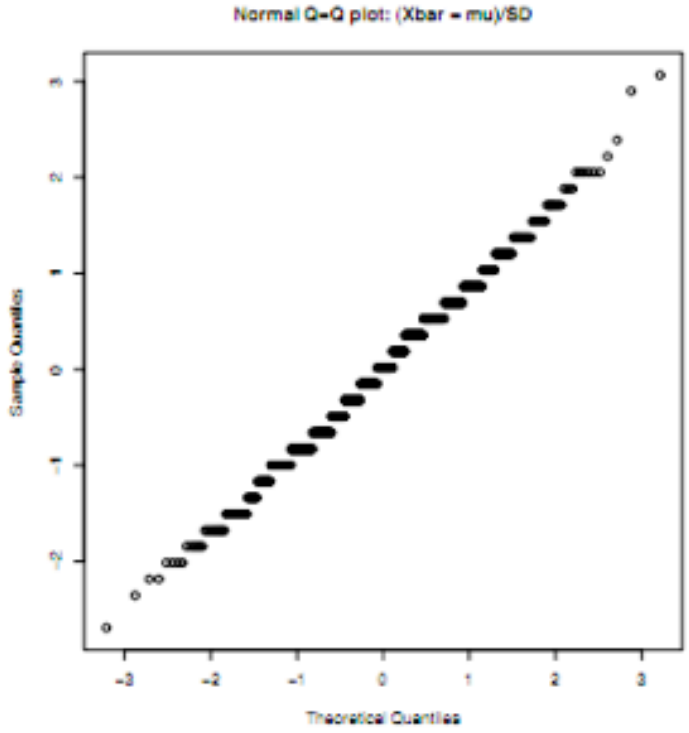
Gosset then took surveys of size *n*=4 and looked at

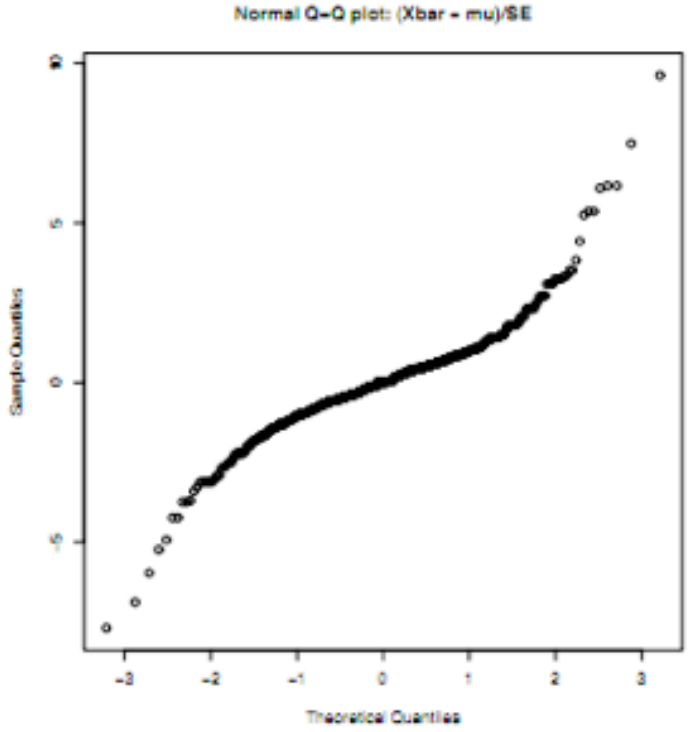
$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

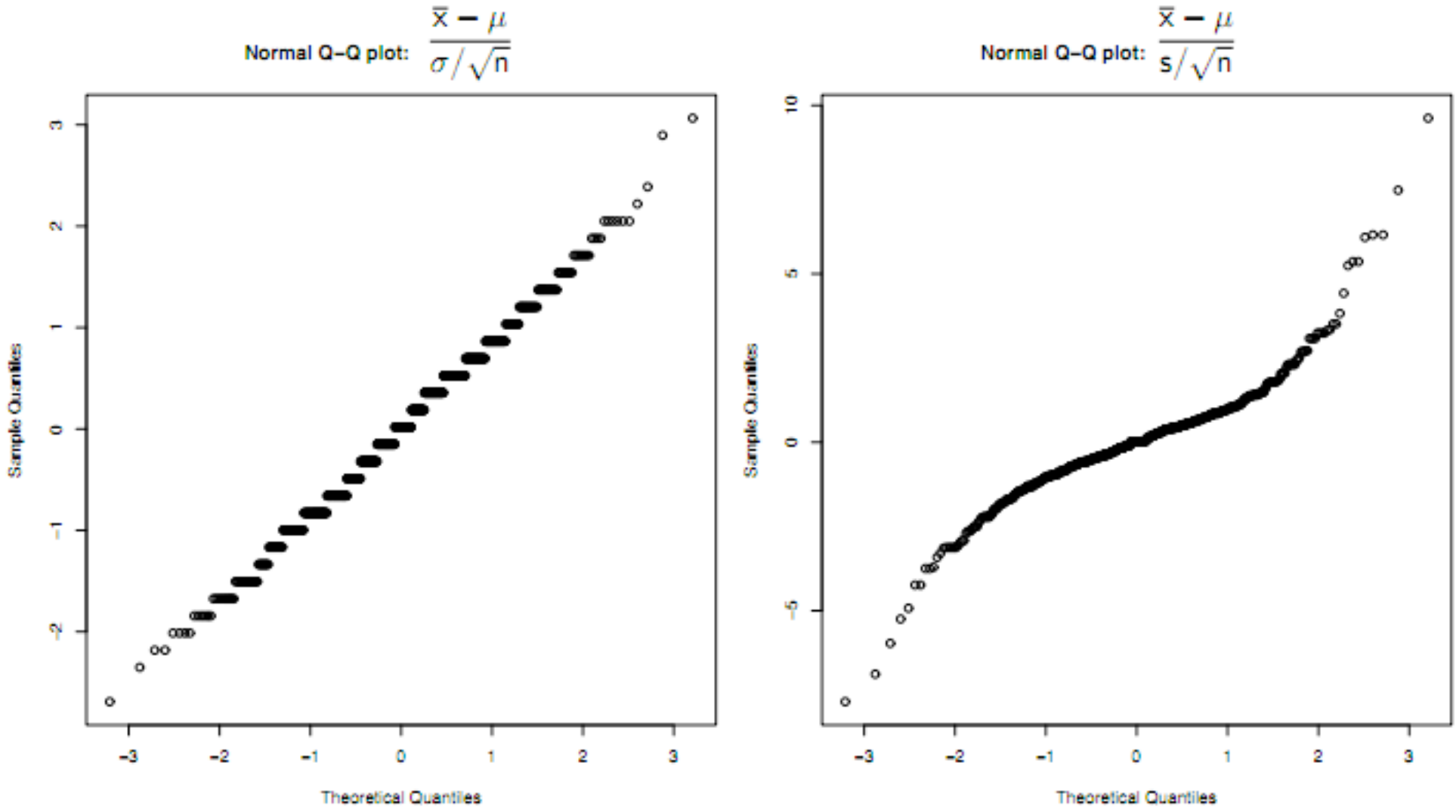
(top graph) and at

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

(bottom graph); What do you notice?







The effect of estimating σ, Gosset's simulation with sample size n=4: On the left he standardizes with the known population standard deviation and on the right he has "plugged-in" s for σ

The *t*-distribution

By having to estimate the population standard deviation in small samples, Gosset showed that the following equation has value less than 0.95

$$\text{Prob}\left(-2 < \frac{\bar{x}-\mu}{s/\sqrt{n}} < 2\right) = \text{Prob}\left(\bar{x} - 2s/\sqrt{n} < \mu < \bar{x} + 2s/\sqrt{n}\right)$$

The tails of the distribution of $(\bar{x}-\mu)/(s/\sqrt{n})$ are heavier than that of a normal; or, put another way, we see from the Q-Q plot both left and right skew

Intuitively, we have **a random quantity downstairs and this induces more spread in the distribution**

Gosset described the correct distribution when the feature of our population we're interested in has is normal looking to begin with (like heights); we refer to it as Student's *t*-distribution

The *t*-distribution

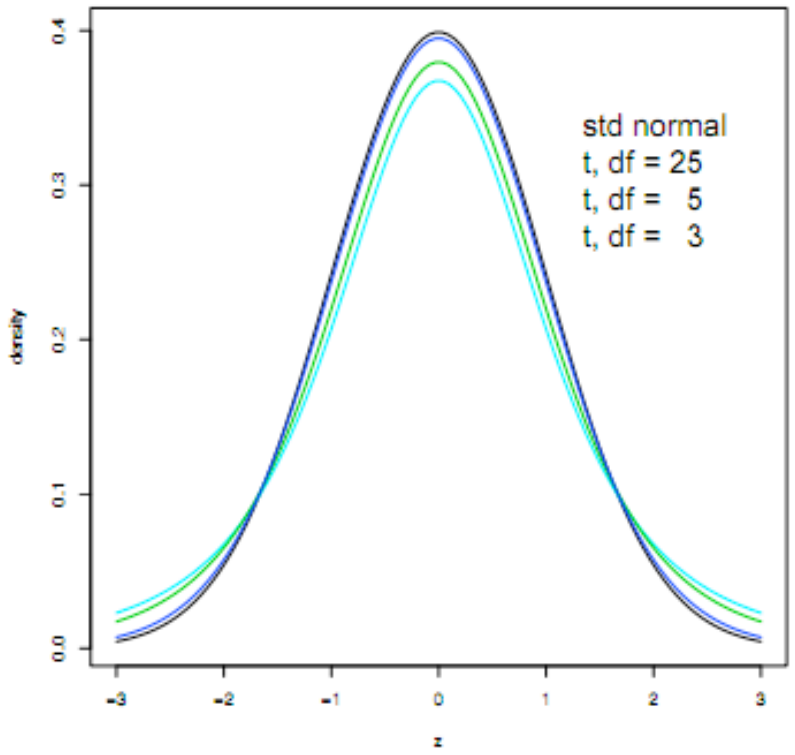
The *t*-distribution has one parameter controlling its shape; it is referred to as its *degrees of freedom*

In our context, the degrees of freedom is *n*-1, where *n* is our sample size

It comes from our original definition of the sample standard deviation

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

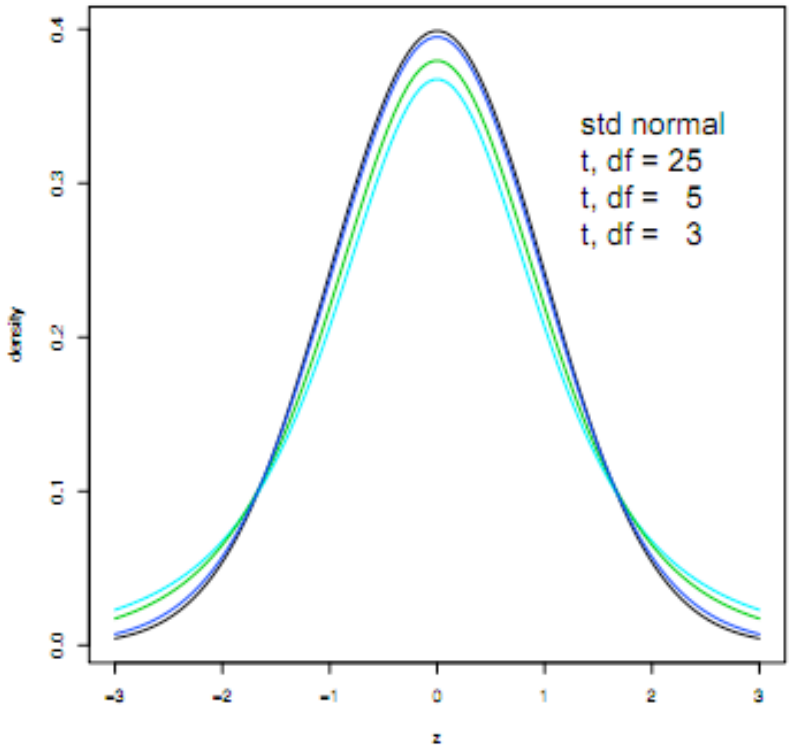
We said there were *n*-1 degrees of freedom in our estimate because 1 was used to compute $\bar{x}$

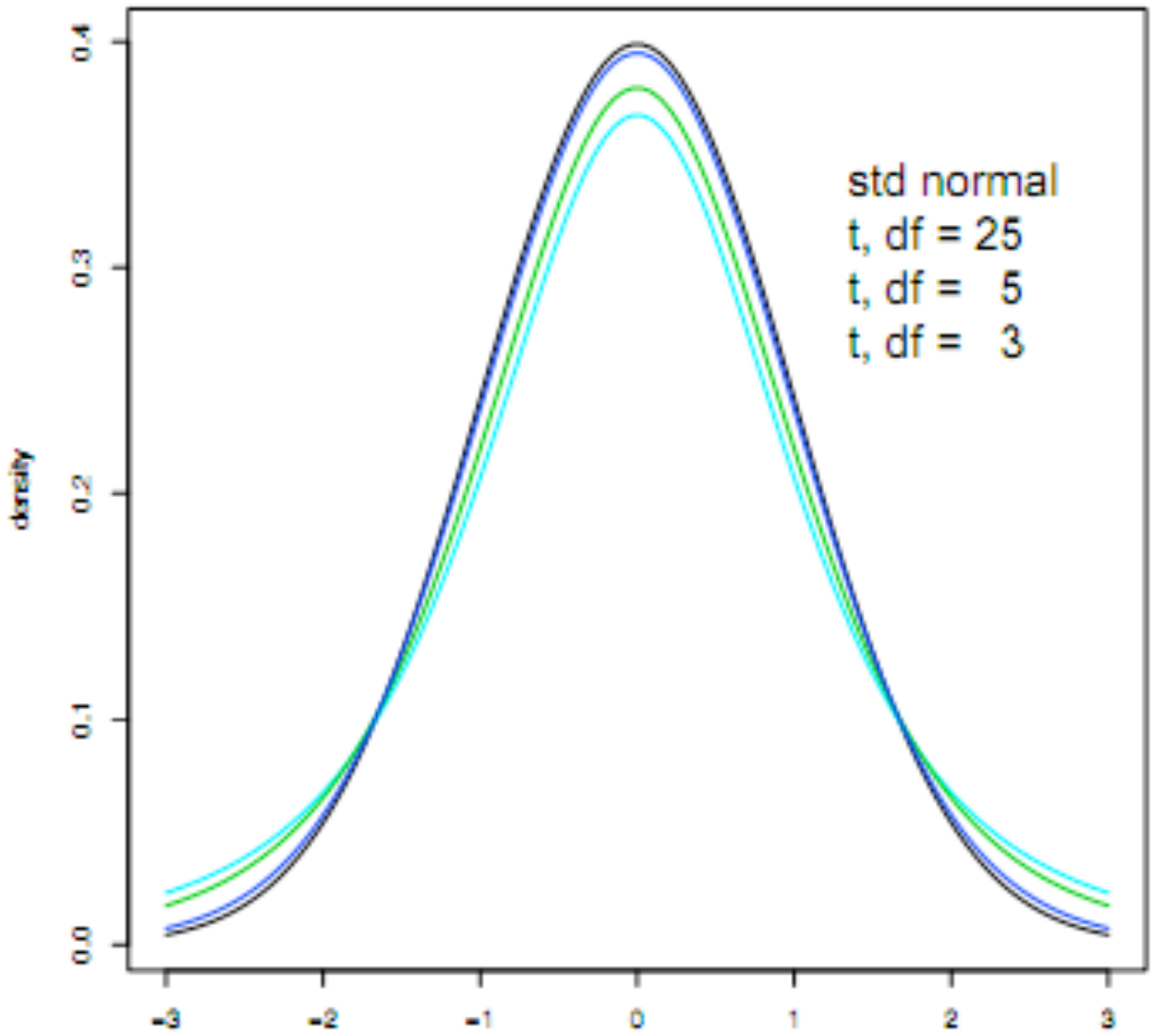


The *t*-distribution

For small samples (small degrees of freedom) s is quite variable and so we have more spread in the distribution

As we collect larger sample sizes, this variability reduces and we see that the *t*-distribution approaches the standard normal curve




std normal
t, df = 25
t, df = 5
t, df = 3
dnorm(y
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3

The *t*-distribution

Now, suppose we want a 95% interval using our estimate *s*

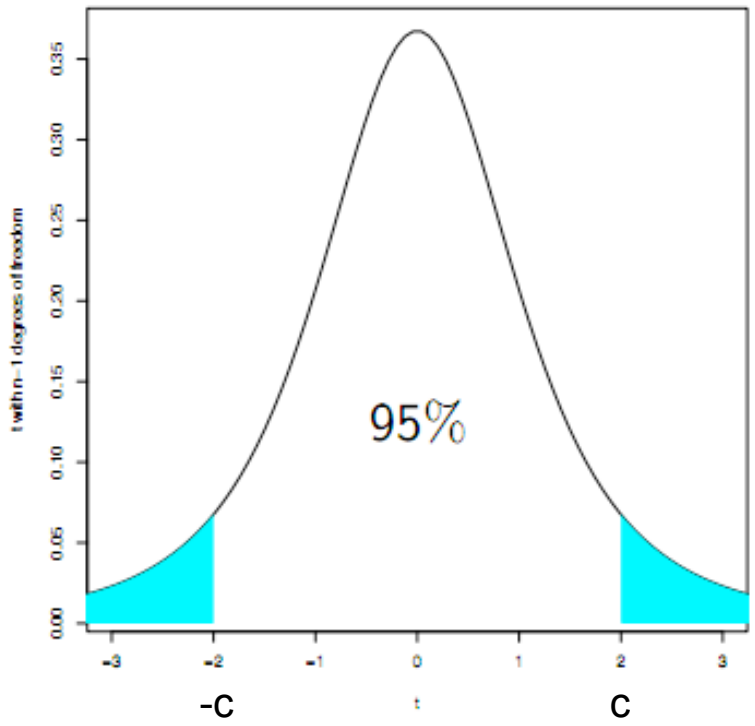
We would would need to find the point c such that

$$\text{Prob}\,(-c \leq T \leq c) = 0.95$$

(the cyan area on the right) where T has a Student's t-distribution with n-1 degrees of freedom

We then form the confidence interval

$$\bar{x} \pm c\frac{s}{\sqrt{n}}$$



Student's t-distribution

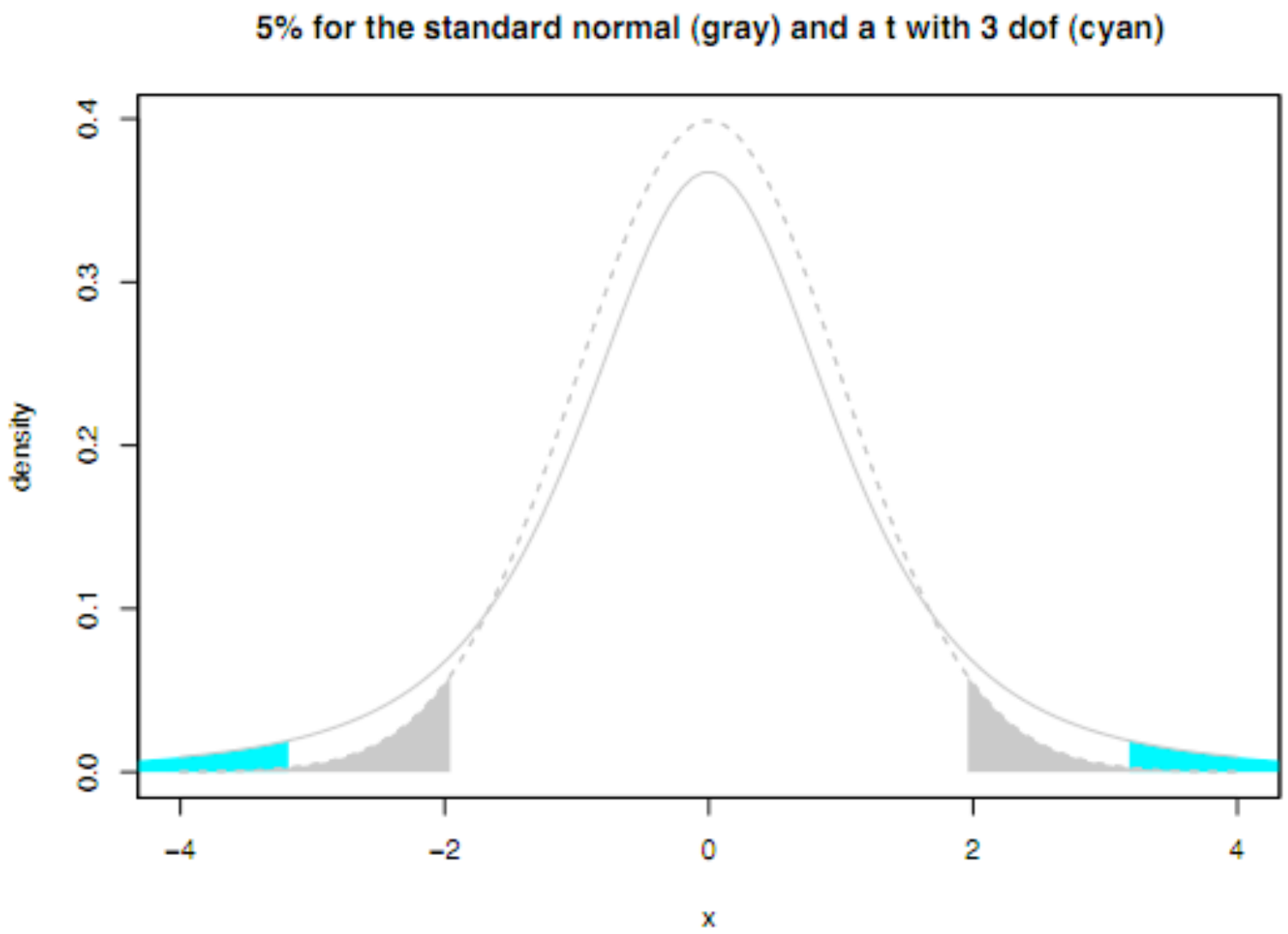
So, what does all this mean? If our data are normally distributed then the *t*-statistic

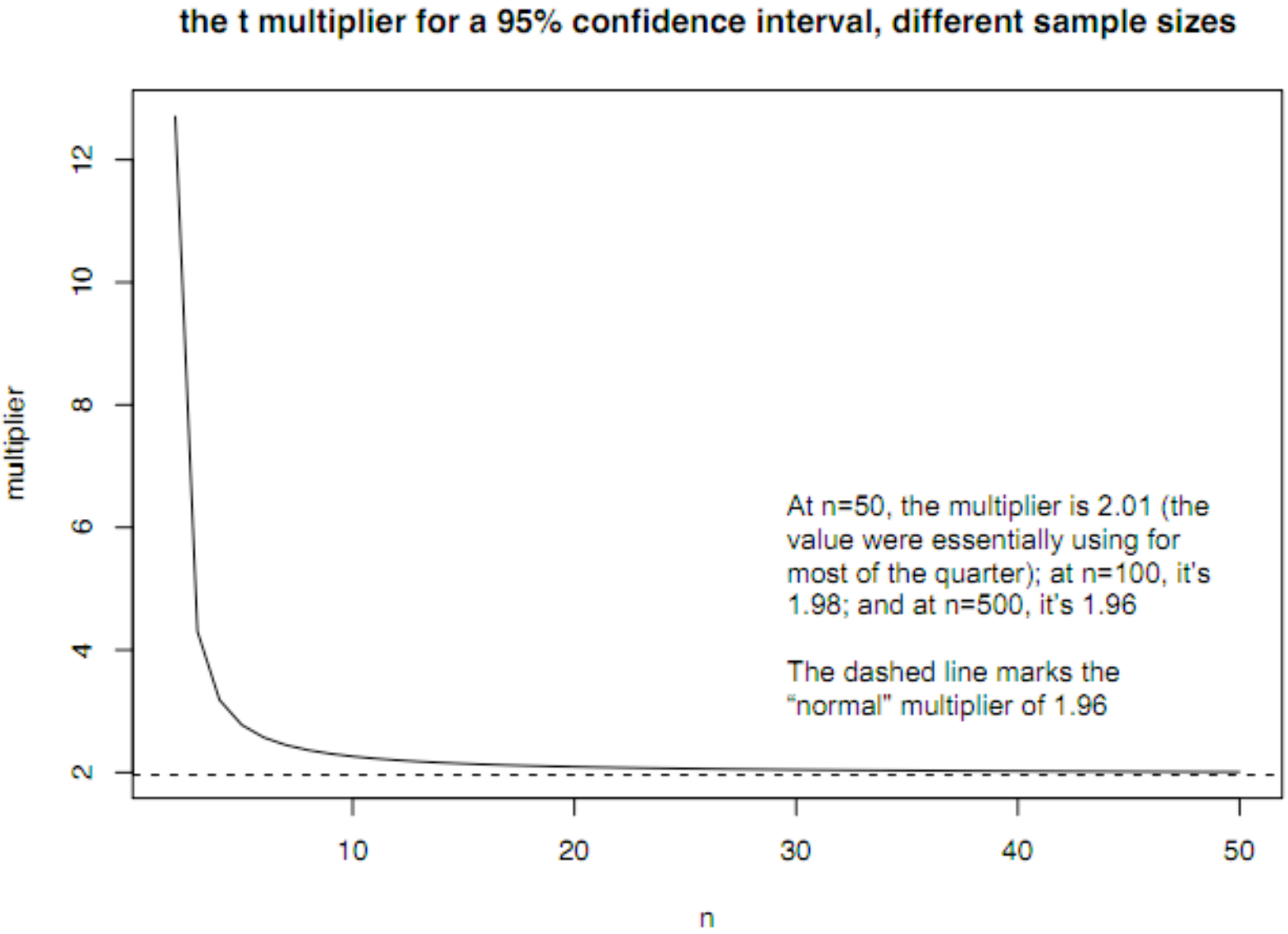
$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has a *t*-distribution with n-1 degrees of freedom

As an example, consider Gosset's original simulations with n=4 data points; we would expect 95% of his standardized differences (where 95% refers to repeated experiments) to be within plus or minus `qt(0.975,df=3)` or 3.18 (remember with n=4, we have n-1=3 degrees of freedom)

... and in this case we would use 3.18 instead of 2 (or 1.96) in the multiplier for our confidence interval $\bar{x} \pm 3.18\, s/\sqrt{n}$


5% for the standard normal (gray) and a t with 3 dof (cyan)
density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


the t multiplier for a 95% confidence interval, different sample sizes
multiplier
2
4
6
8
10
12
10
20
30
40
50
n
At n=50, the multiplier is 2.01 (the value were essentially using for most of the quarter); at n=100, it's 1.98; and at n=500, it's 1.96
The dashed line marks the "normal" multiplier of 1.96

Student's t-distribution

To sum up; Gosset worked out the sampling distribution of a standardized statistic, **the t-statistic, under the assumption that the data we've observed come from a population with a normal distribution**

Under that assumption, we can derive a confidence interval using quantiles from the t-distribution; **as our sample sizes get large, the effect of estimating σ with s diminishes and we return to the usual normal interval**

Pivots

Notice that when our data $X_1, \ldots, X_n$ come from a normal distribution, then the quantity

$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

has the same distribution no matter what values of μ and σ were used to generate the data

A quantity of this kind is known as a “pivot” -- As we have seen, pivots can be inverted to compute confidence intervals (this was the key ingredient in our chain of probability statements that pivoted our parameter of interest into the middle of our interval expression)

Another example

We have been using the following estimate for σ^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Assuming our data $X_1, \ldots, X_n$ come from a normal distribution with mean μ and standard deviation σ, we can show that the quantity

$$(n-1)S^2/\sigma^2$$

has a chi-square distribution with n-1 degrees of freedom and hence qualifies as a pivot!

(The same comment works if we use the MLE for σ^2 instead, but with n/σ^2 as the multiplier)

Pivots

Therefore, we can find values a and b such that

$$P(\chi^2_{n-1} < a) = 0.025 \quad \text{and} \quad P(\chi^2_{n-1} > b) = 0.025$$

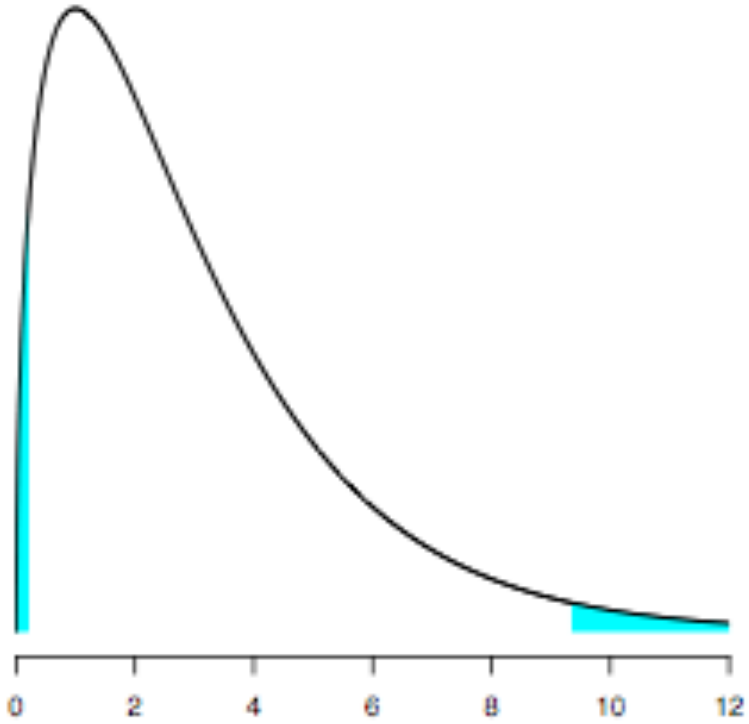
where χ^2_{n-1} has a chi-square distribution with n-1 degrees of freedom

Combining these two expressions we find

$$P\left(a \leq (n-1)\widehat{\sigma}^2/\sigma^2 \leq b\right)$$

and by inverting we derive a 95% confidence interval for σ^2

$$\left[(n-1)\widehat{\sigma}^2/b,\ (n-1)\widehat{\sigma}^2/a\right]$$



Exact v. approximate sampling distributions

In the last few slides, we have focused mainly on **exact expressions** for the sampling distribution of an estimator -- We have seen that the price of these “clean” results is a set of **strong assumptions** about how the data were generated

And even when we are willing to make strong assumptions, it’s often the case that **clean formula do not exist** for the exact sampling distribution of an estimator -- Instead we have to rely on approximations

Our starting point for these approximations comes from one of the properties we discussed last lecture, **consistency**...

Exact v. approximate sampling distributions

We let $\widehat{\theta}_n$ denote an estimator based on n samples $X_1, \ldots, X_n$ -- We said that it was consistent if $\widehat{\theta}_n$ converges to the unknown value θ^* "in probability"

Intuitively, this means that **the errors** $\widehat{\theta}_n - \theta^*$ **get small** as we collect more and more data -- As we have seen, however, there is tremendous value in not only knowing that the errors get small but in knowing how big they can be, their "typical" values

These more refined statements mean **we need to know something of the sampling distribution of our estimator** -- And, in particular, can we say something approximate "in the limit" (that is general and maybe easier to work out mathematically) as opposed to something exact

Convergence in distribution

To make this precise, we say that a sequence of random variables $Z_1, Z_2, \ldots$ and let F_n denote the cumulative distribution function of Z_n -- Let Z be a random variable with CDF F

We say that $Z_1, Z_2, \ldots$ **converges in distribution** to Z if

$$F_n(x) \to F(x)$$

at all points where F is continuous -- We write $Z_n \xrightarrow{\mathcal{D}} Z$

Convergence in distribution

We can word consistency in terms of convergence in distribution as well -- We say that an estimator is consistent if $\widehat{\theta}_n$ converges in distribution to a "constant" random variable θ^* (that is a random variable that takes on the value θ^* with probability 1)

The Central Limit Theorem is also a statement about convergence in probability -- That is, if $X_n, \ldots, X_n$ are independent and identically distributed random variables with mean μ and variance σ^2, then their sample mean $\overline{X}_n$

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} Y$$

where Y has a normal distribution with mean 0 and variance σ^2

Convergence in distribution

Informally, we say that the sample mean $\overline{X}_n$ has approximately a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ "for large n" -- Or, more concisely, we say that $\overline{X}_n$ is asymptotically normal

This is an approximate version of the exact statement we examined at the beginning of this lecture, when our data $X_n, \ldots, X_n$ really did come from a normal distribution

Results such as these are the starting point for building "approximate" or "asymptotic" expressions for the sampling distribution of an estimate

Convergence in distribution

When reasoning about convergence results, Slutsky's Theorem is a workhorse -- If g(z,y) is a function that is jointly continuous at every point of the form z,c for some fixed c, and if $Z_n \xrightarrow{\mathcal{D}} Z$ and $Y_n \xrightarrow{\mathcal{D}} c$ (converges in probability to a constant c), then

$$g(Z_n, Y_n) \xrightarrow{\mathcal{D}} g(Z, c)$$

What this means is that

$$\begin{aligned} Z_n + Y_n &\xrightarrow{\mathcal{D}} Z + c \\ Y_n Z_n &\xrightarrow{\mathcal{D}} cZ \\ Z_n / Y_n &\xrightarrow{\mathcal{D}} Z/c, \ c \neq 0 \end{aligned}$$

or, that the limit of sums is the sum of the limits, etc.

Convergence in distribution

The plug-in principle, for example, makes use of this fact -- If we have an asymptotically normal estimate $\widehat{\theta}_n$ such that

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \xrightarrow{\mathcal{D}} \text{Normal}\,(0, \tau^2)$$

and if $\widehat{\tau}_n$ is any consistent estimate of τ, then

$$\frac{\widehat{\theta}_n - \theta^*}{\widehat{\tau}_n/\sqrt{n}} \xrightarrow{\mathcal{D}} \text{Normal}\,(0, 1)$$

which we can use to “plug-in” estimates and construct confidence intervals

Plug-in

This is what we did when we considered

$$\frac{\overline{X} - \mu}{s/\sqrt{n}} \quad \text{versus} \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

as s is a consistent estimate of σ (see the end of this lecture for the details)

Notice that **Gosset worked out the exact sampling distribution** for the plug-in estimate for small sample sizes n -- As n gets large, the effect of using s diminishes and, as we see here and on a previous slide, **the whole thing approaches a normal distribution asymptotically**

Plug-in

The same can be accomplished for the binomial case -- If X comes from a binomial distribution with known n and unknown p, then the MLE for p is just

$$\widehat{p}_n = X/n$$

By the CLT, $\widehat{p}$ has an asymptotically normal distribution with mean p and variance p(1-p)/n -- Because $\widehat{p}$ is consistent for p, we can "plug-in" using Slutsky's theorem a number of times to get

$$\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}}$$

is approximately standard normal giving us a 95% confidence interval of roughly

$$\widehat{p}_n \pm 2\sqrt{\widehat{p}_n(1-\widehat{p}_n)/n}$$

which is quoted in your text

Without plugging in

Actually in this case, we can work things out without plugging in -- That is, by the CLT we know that

$$\frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}}$$

is approximately normal so that

$$P\left(-2 \leq \frac{\widehat{p}_n - p}{\sqrt{p(1-p)/n}} \leq 2\right) \approx 0.95$$

By solving a quadratic equation we come up with the interval

$$\frac{\widehat{p}_n + \frac{1}{n} \pm 2\sqrt{\frac{1}{n^2} + \frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}}{1 + \frac{4}{n}}$$

which agrees with our previous expression as n gets large (although this formula is rarely quoted in introductory texts)

Convergence in distribution: The MLE

We can also say something quite general about Maximum Likelihood Estimates -- 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 not only is $\widehat{\theta}_n$ a consistent estimate of θ^*, it is also asymptotically normal

Convergence in distribution: The MLE

To see how this might go, we can write out a Taylor expansion for l' around the true unknown parameter θ^*, which we evaluate at

$$0 = l'(\widehat{\theta}_n) = l'(\theta^*) + (\widehat{\theta} - \theta^*)l''(\theta^*) + \cdots$$

Moving things around a little, we find that

$$\sqrt{n}(\widehat{\theta}_n - \theta^*) \approx -\frac{l'(\theta^*)/\sqrt{n}}{l''(\theta^*)/n}$$

Convergence in distribution: The MLE

We can treat the limits in the upstairs and downstairs separately and apply Slutsky's theorem -- For example, the downstairs looks like an average

$$l''(\theta)/n = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial^2\theta} f(X_i|\theta)$$

This can be shown to converge to a constant known as the **expected Fisher Information** -- It is given by the expression

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial^2\theta} f(X|\theta)\right]$$

Convergence in distribution: The MLE

When the dust settles on our calculations, 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 that

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

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

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

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 $\widehat{\theta}$

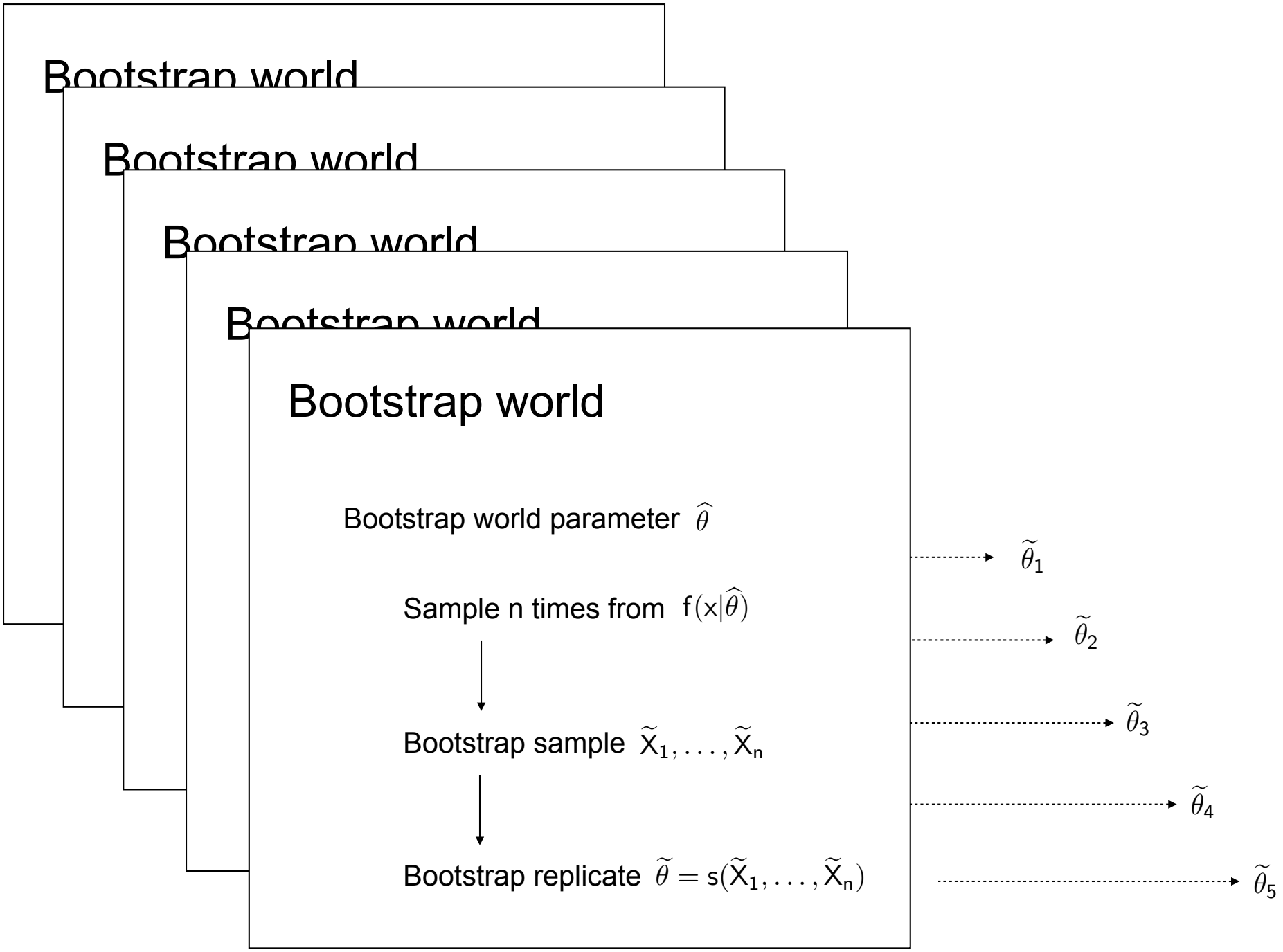
Sample n times from $f(x|\widehat{\theta})$

↓

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