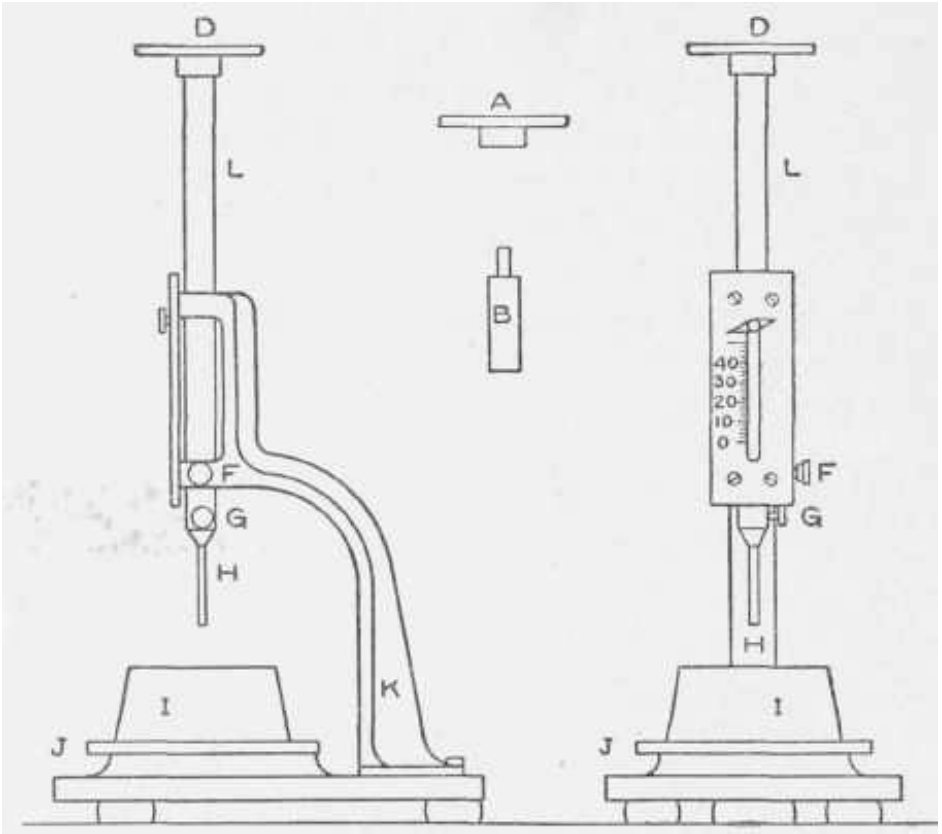


Lecture 9: Consistency

Last time

We started with a (perhaps overwrought) treatment of the different interpretations of probability -- We spent time on this in part because it's a good story and in part because it helps you understand the motivation behind some of the inferential constructions we'll be working with

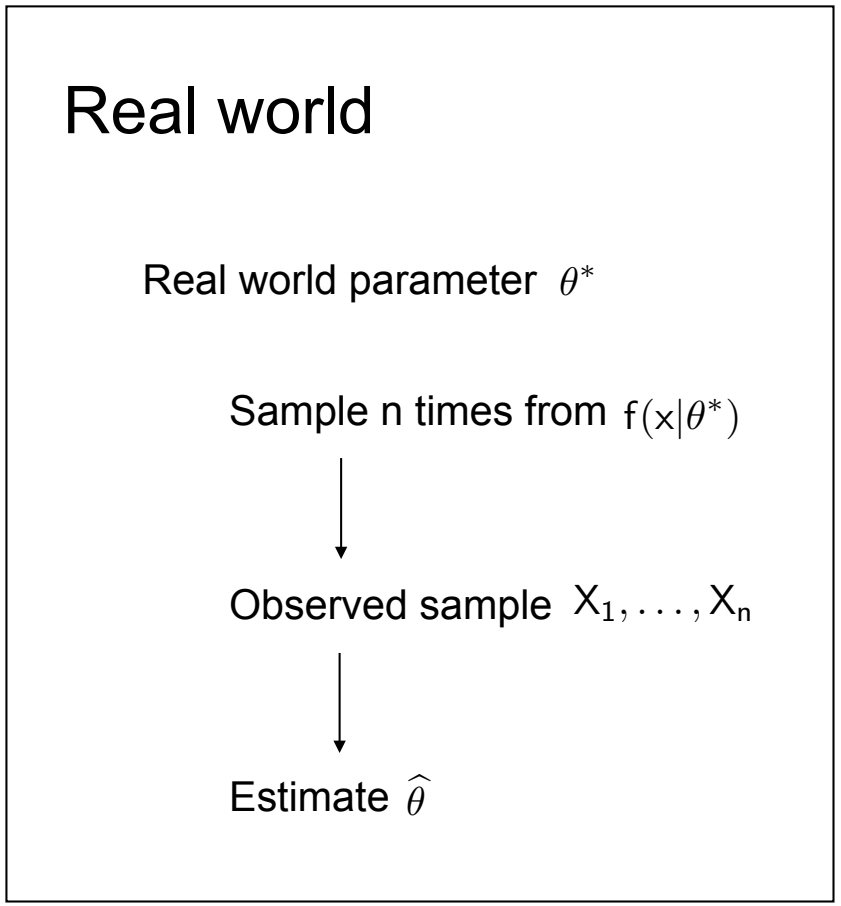
We then started to talk about properties of estimators -- Our primary "probe" for assessing the behavior of an estimator (a function of data, an algorithm) was the sampling distribution

Properties of estimators

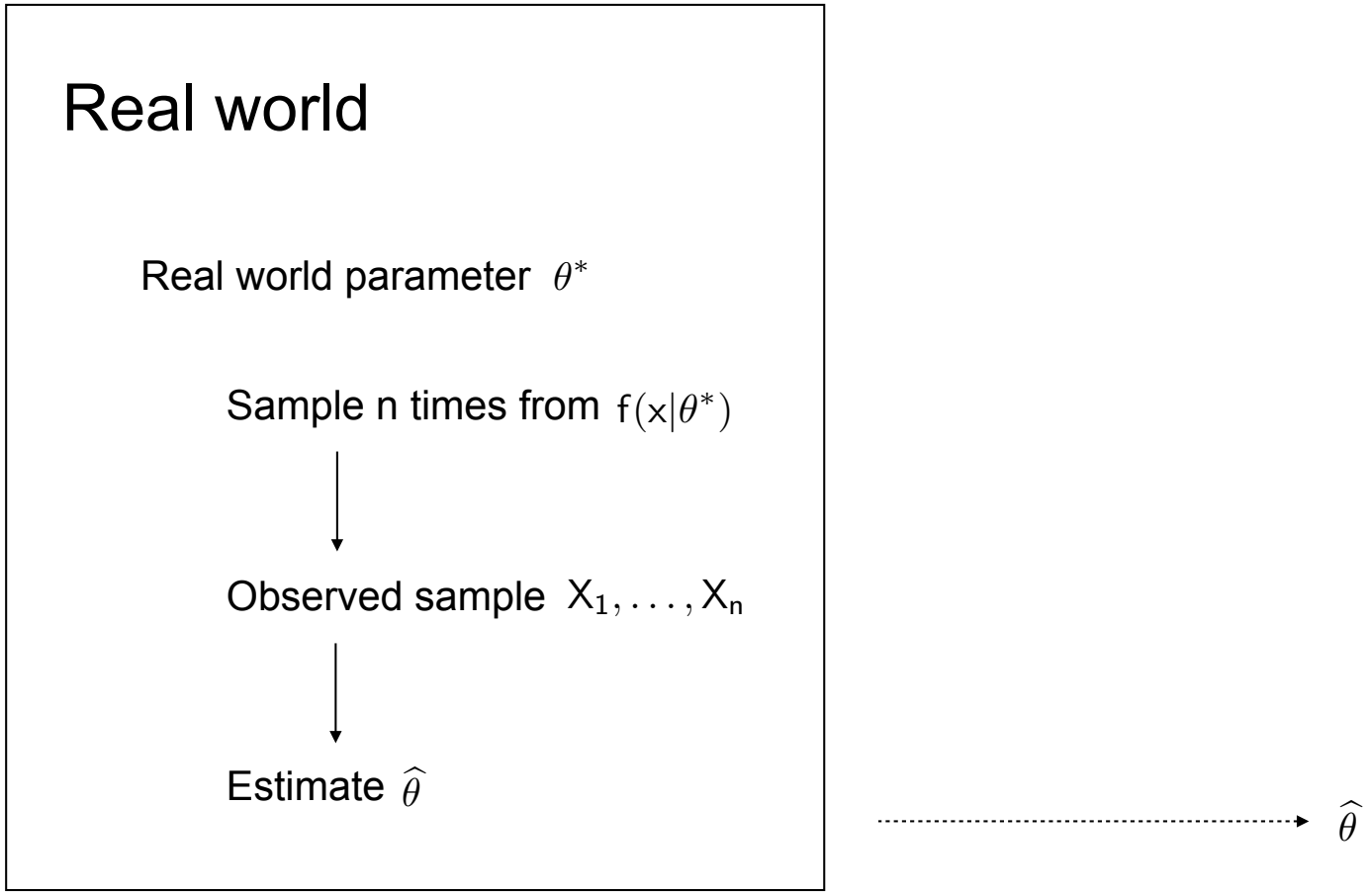
Suppose we are given a sample $X_1, \ldots, X_n$ of size n that are independent draws from some distribution $f(x|\theta^*)$ that's part of a parametric family $f(x|\theta)$

An estimate $\widehat{\theta}$ of θ^* is just some function of $X_1, \ldots, X_n$ (an algorithm, if you will, with $X_1, \ldots, X_n$ as input), or in symbols $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ -- We view $\widehat{\theta}$ as a random variable in the sense that each time we repeat our experiments, we would collect another sample of data, producing a different estimate

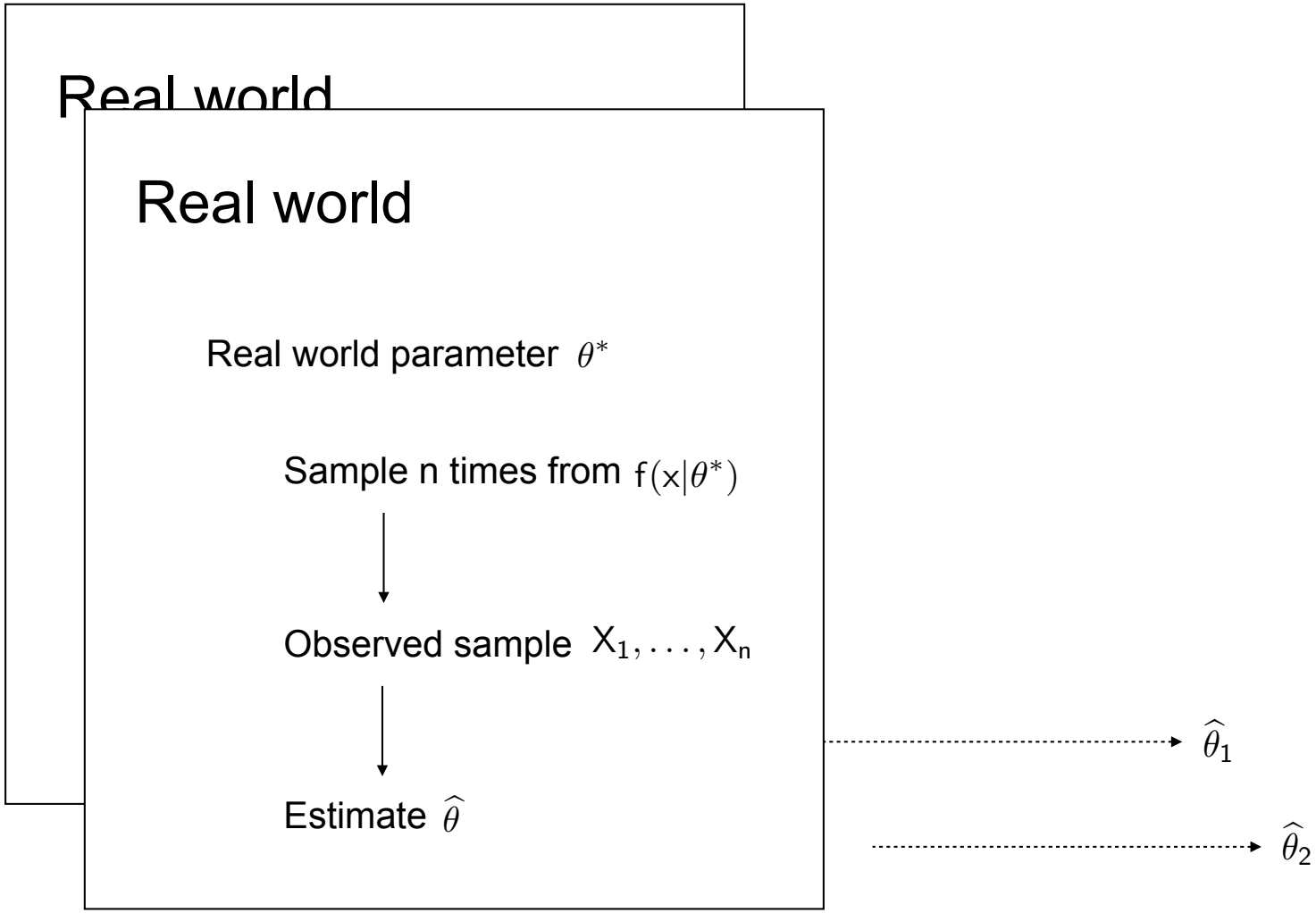
We refer to the distribution of $\widehat{\theta}$ over these repeated experiments as its **sampling distribution --** A frequentist tool for "calibrating" our expectations of what happens when the estimate is applied to data



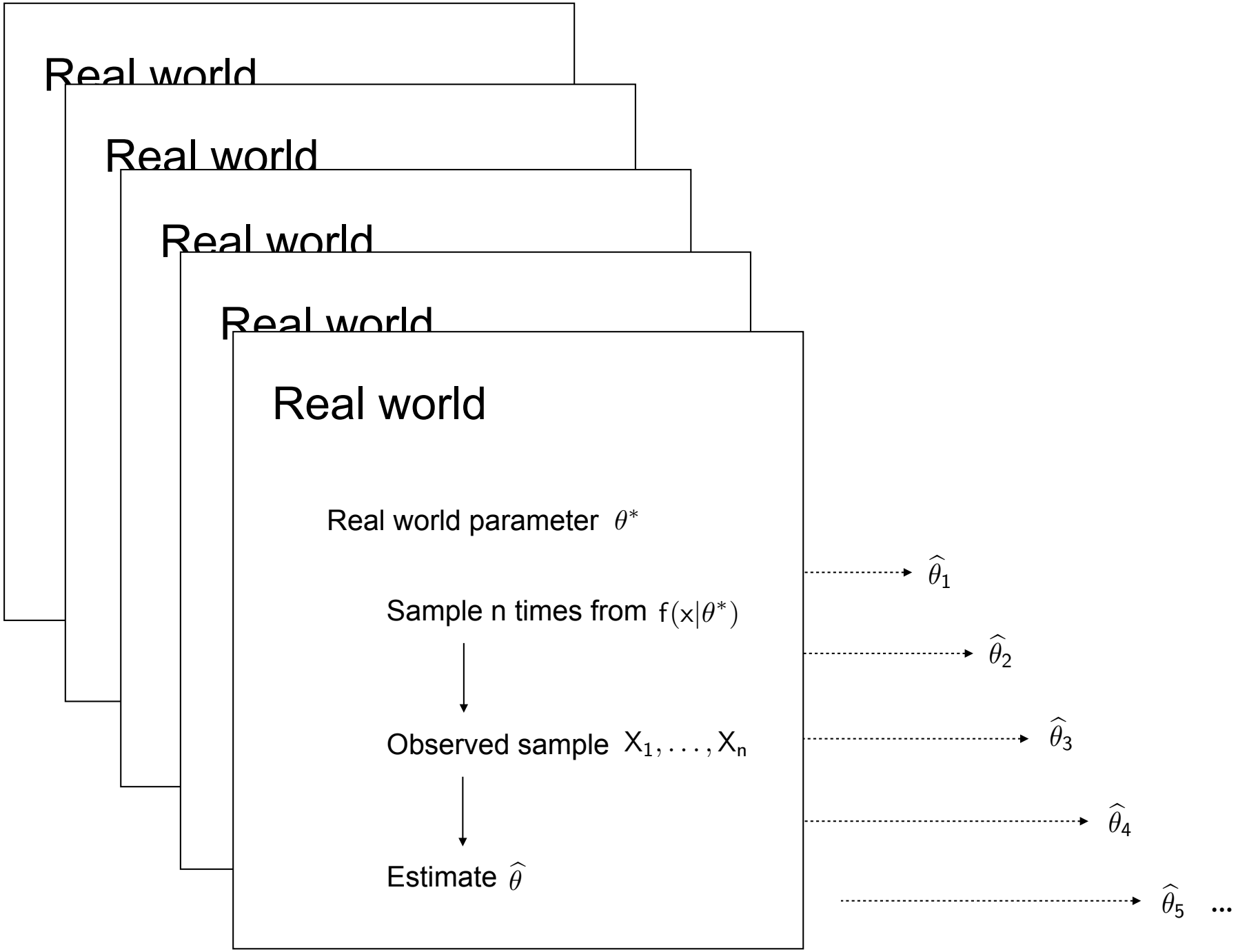
A sketch of a single experiment



A sketch of a single experiment and an estimate



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

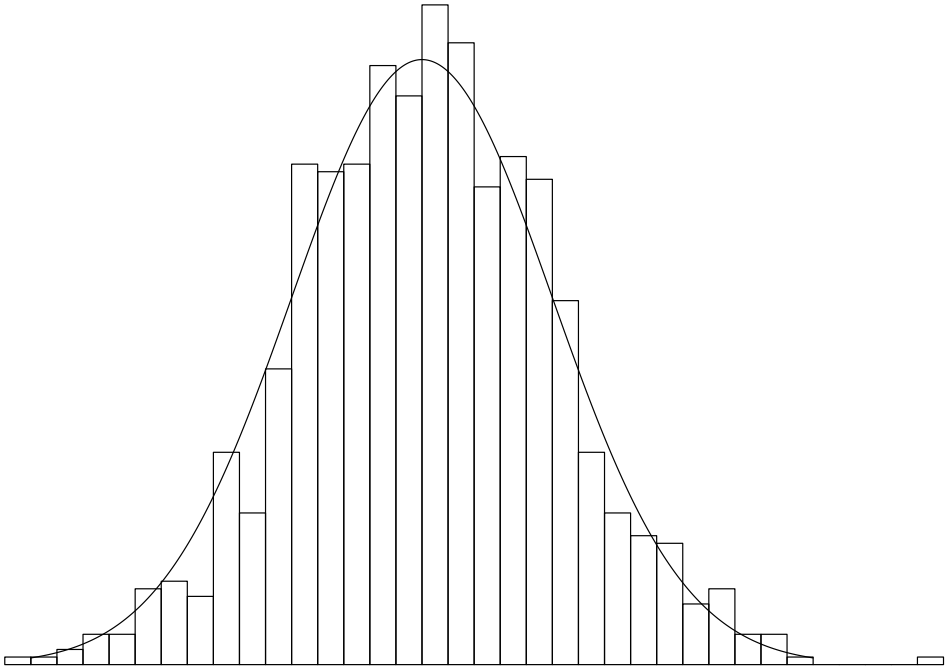


Repeating the experiment produces new data and a new estimate...

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

The mean of this distribution tell us, for example, how far on average our estimate is from the value of the true parameter that generated the data

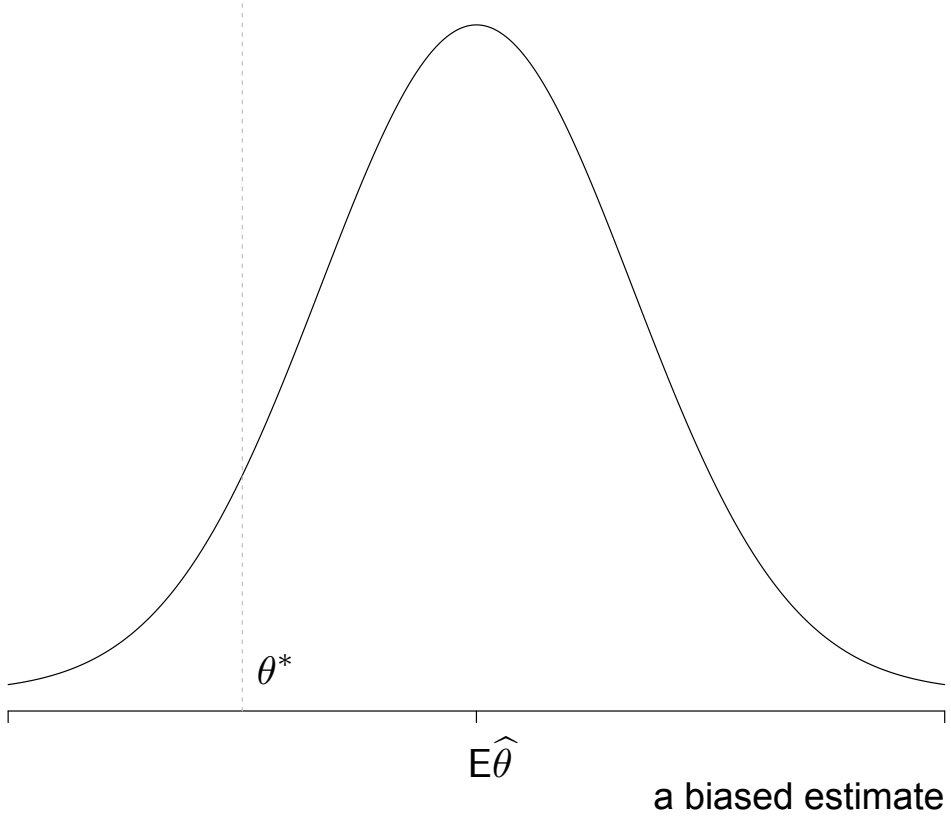


The sampling distribution

The bias of an estimate is defined as the difference between the mean of the sampling distribution, denoted $\mathsf{E}\widehat{\theta}$, and the parameter that generated the data θ^*

$$\mathsf{bias}(\widehat{\theta}) = \mathsf{E}\widehat{\theta} - \theta^*$$

We say that an estimate is unbiased if its sampling distribution is centered on θ^*

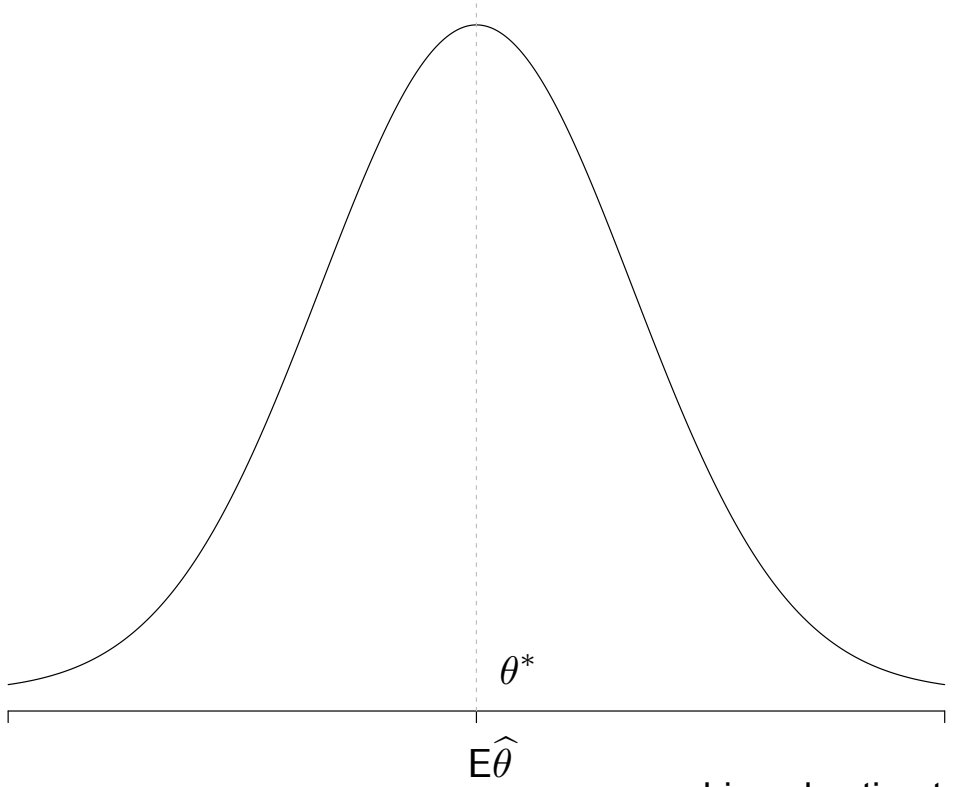


The sampling distribution

The bias of an estimate is defined as the difference between the mean of the sampling distribution, denoted $\mathsf{E}\widehat{\theta}$, and the parameter that generated the data θ^*

$$\mathsf{bias}(\widehat{\theta}) = \mathsf{E}\widehat{\theta} - \theta^*$$

We say that an estimate is unbiased if its sampling distribution is centered on θ^*



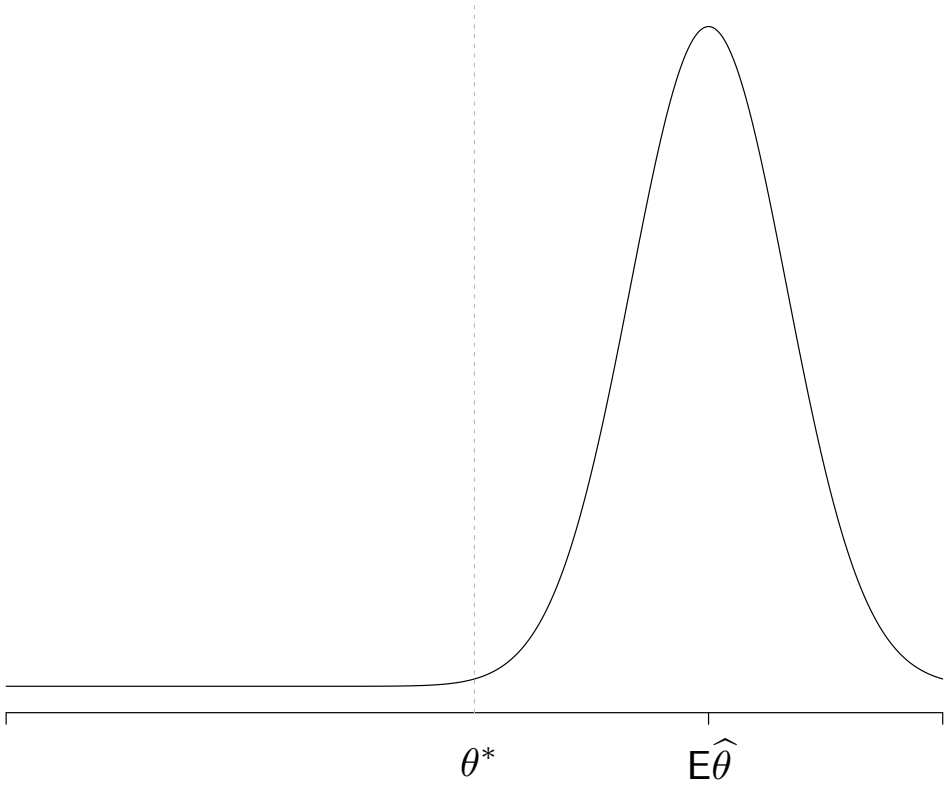
an unbiased estimate

The sampling distribution

The spread of the sampling distribution records the variability we expect to see across repeated experiments

The standard deviation of $\widehat{\theta}$ across repetitions of our experiment is known as the standard error of

$$\mathsf{se}(\widehat{\theta}) = \sqrt{\mathsf{var}(\widehat{\theta})}$$



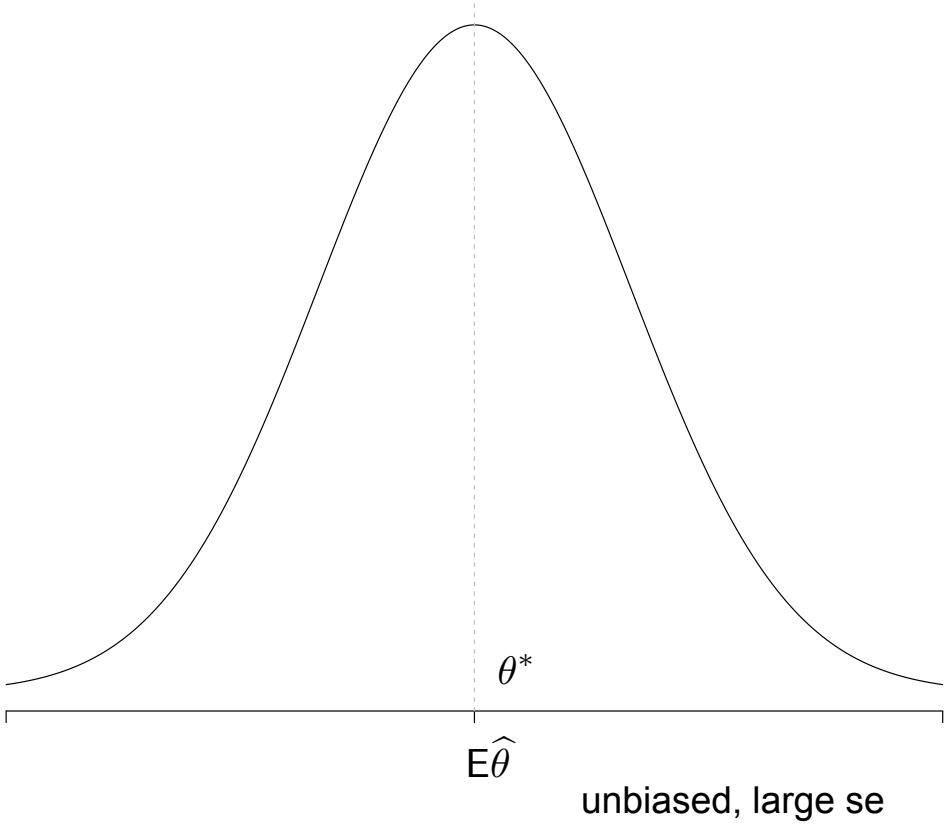
biased, small se

The sampling distribution

The spread of the sampling distribution records the variability we expect to see across repeated experiments

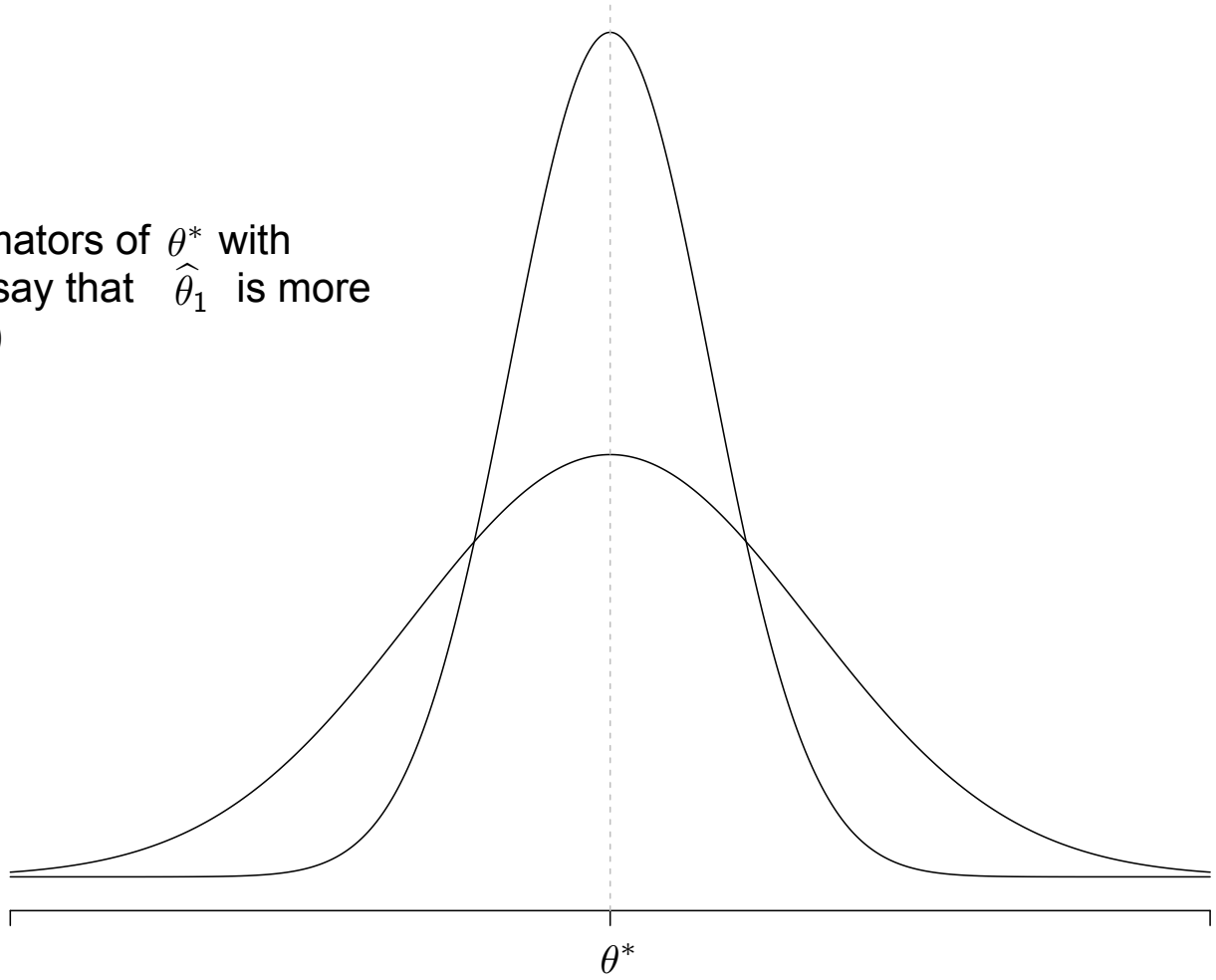
The standard deviation of $\widehat{\theta}$ across repetitions of our experiment is known as the standard error of

$$\mathsf{se}(\widehat{\theta}) = \sqrt{\mathsf{var}(\widehat{\theta})}$$



Efficiency

If $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are two unbiased estimators of θ^* with variances $\mathrm{var}(\widehat{\theta}_1)$ and $\mathrm{var}(\widehat{\theta}_2)$, we say that $\widehat{\theta}_1$ is more efficient than $\widehat{\theta}_2$ if $\mathrm{var}(\widehat{\theta}_1) < \mathrm{var}(\widehat{\theta}_2)$



Mean squared error

When we judge the reasonableness of an estimator, we often combine both (squared) bias and variance into one measure, the mean squared error

$$\mathrm{MSE} = \mathrm{E}(\widehat{\theta} - \theta^*)^2$$

To see how the center and spread of the sampling distribution come into play, we can write down a little algebra

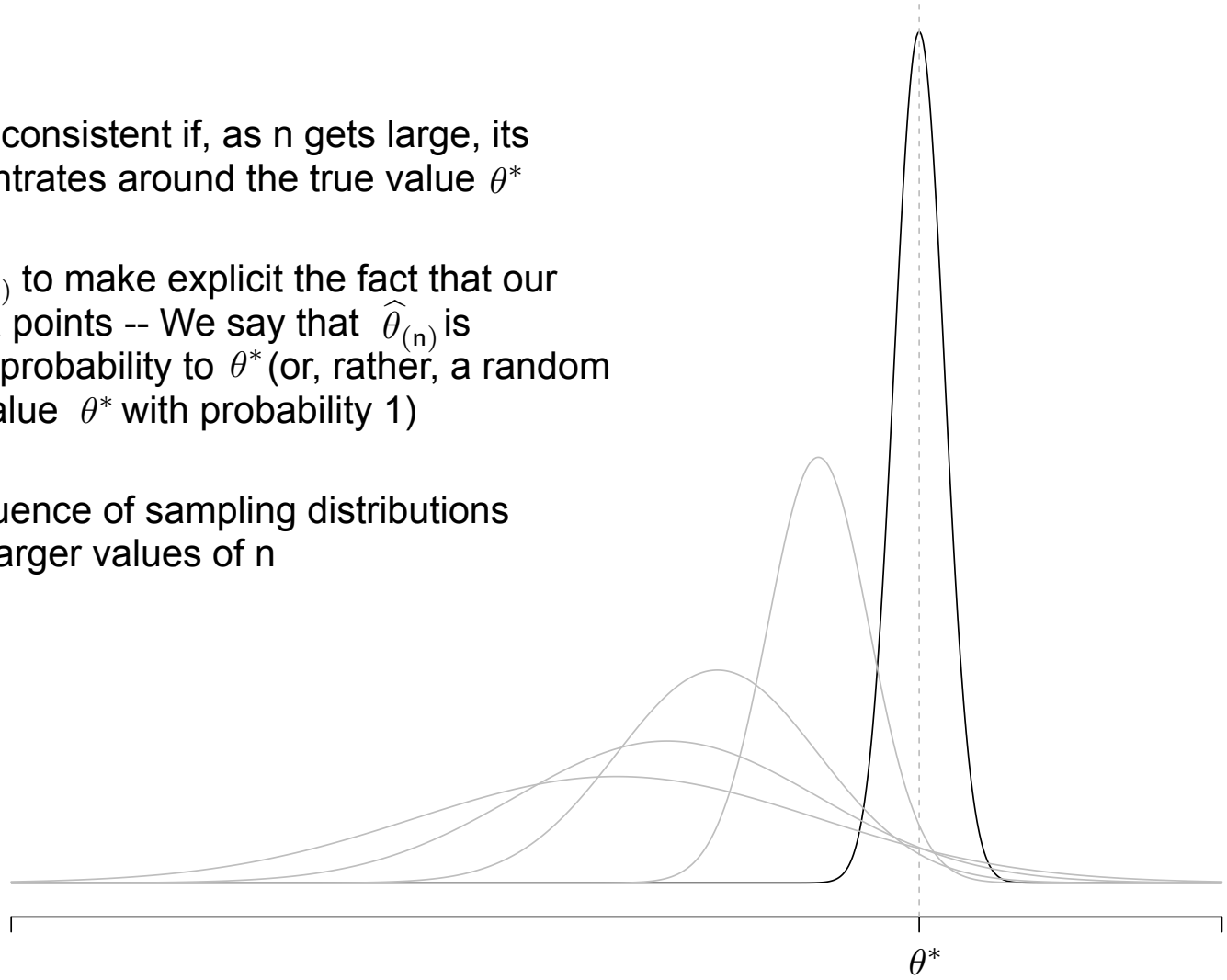
$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}(\widehat{\theta} - \theta^*)^2 \\
&= \mathrm{E}(\widehat{\theta} - \mathrm{E}\widehat{\theta} + \mathrm{E}\widehat{\theta} - \theta^*)^2 \\
&= \mathrm{E}(\widehat{\theta} - \mathrm{E}\widehat{\theta})^2 + (\mathrm{E}\widehat{\theta} - \theta^*)^2 + 2(\mathrm{E}\widehat{\theta} - \theta^*)\mathrm{E}(\widehat{\theta} - \mathrm{E}\widehat{\theta}) \\
&= \mathrm{var}(\widehat{\theta}) + \mathrm{bias}^2(\widehat{\theta})
\end{aligned}$$

The sampling distribution

We say that an estimator is consistent if, as n gets large, its sampling distribution concentrates around the true value θ^*

To be precise, let's write $\widehat{\theta}_{(n)}$ to make explicit the fact that our estimate depends on n data points -- We say that $\widehat{\theta}_{(n)}$ is consistent if it converges in probability to θ^* (or, rather, a random variable that takes on the value θ^* with probability 1)

At the right, we show a sequence of sampling distributions associated with larger and larger values of n



Convergence in probability

Recall the following from your probability course -- A sequence of random variables $Z_1, Z_2, Z_3, \ldots,$ is said to converge in probability to another random variable Z, written $Z_i \xrightarrow{P} Z$, if, for every $\epsilon > 0$

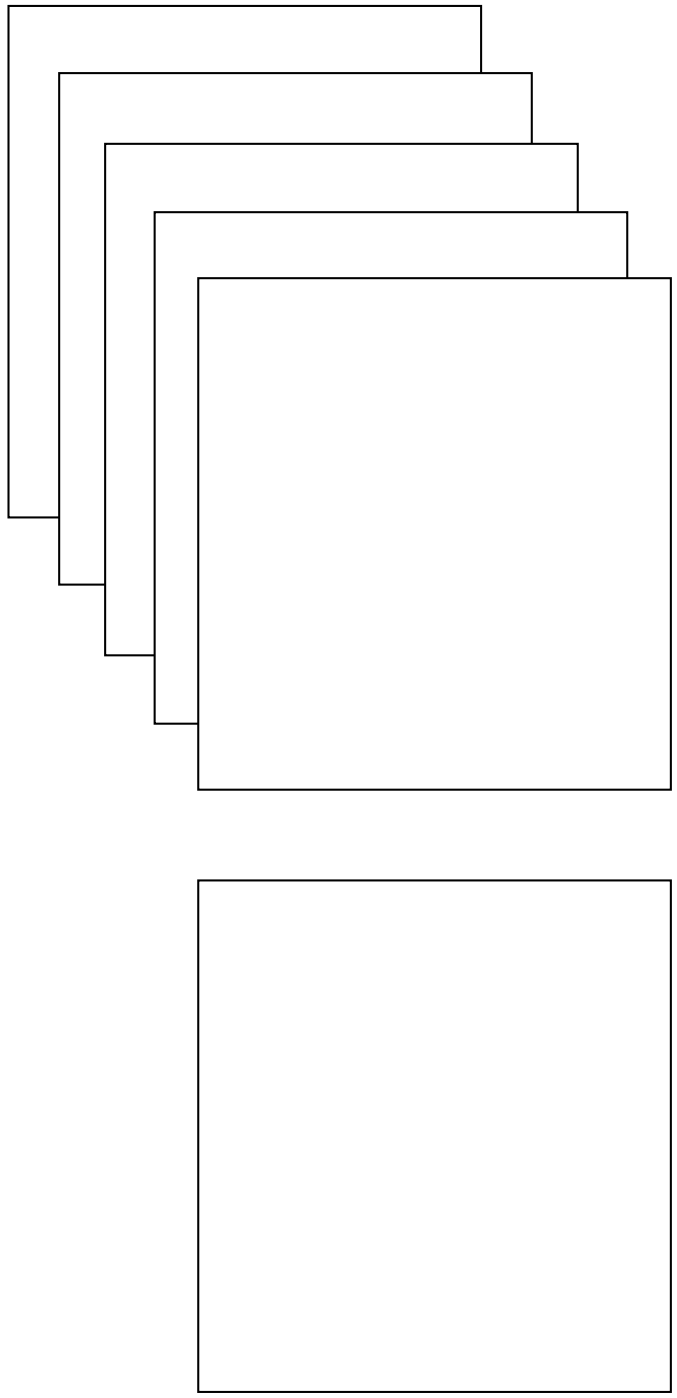
$$P(|Z_i - Z| > \epsilon) \to 0$$

as $i \to \infty$

The sampling distribution

The sampling distribution is, at the moment, a theoretical construction -- It consists of all the possible outcomes for experiments we could have run

In practice, we run only a single experiment and would like to assess the properties associated with our estimator -- Is it biased? How big is the bias? What is the standard error associated with our estimate?



The bootstrap

For the remainder of the term, our main tool for assessing the sampling variability of an estimate will be a computation procedure known as “the bootstrap” -- The term comes from the expression to “pull yourself up by your bootstraps”*

Like our re-randomizations for the A/B testing, this procedure again has us “analyze as we randomized,” mimicking the stochastic mechanism employed to generate the data in the first place

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

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 $\widehat{\theta}$

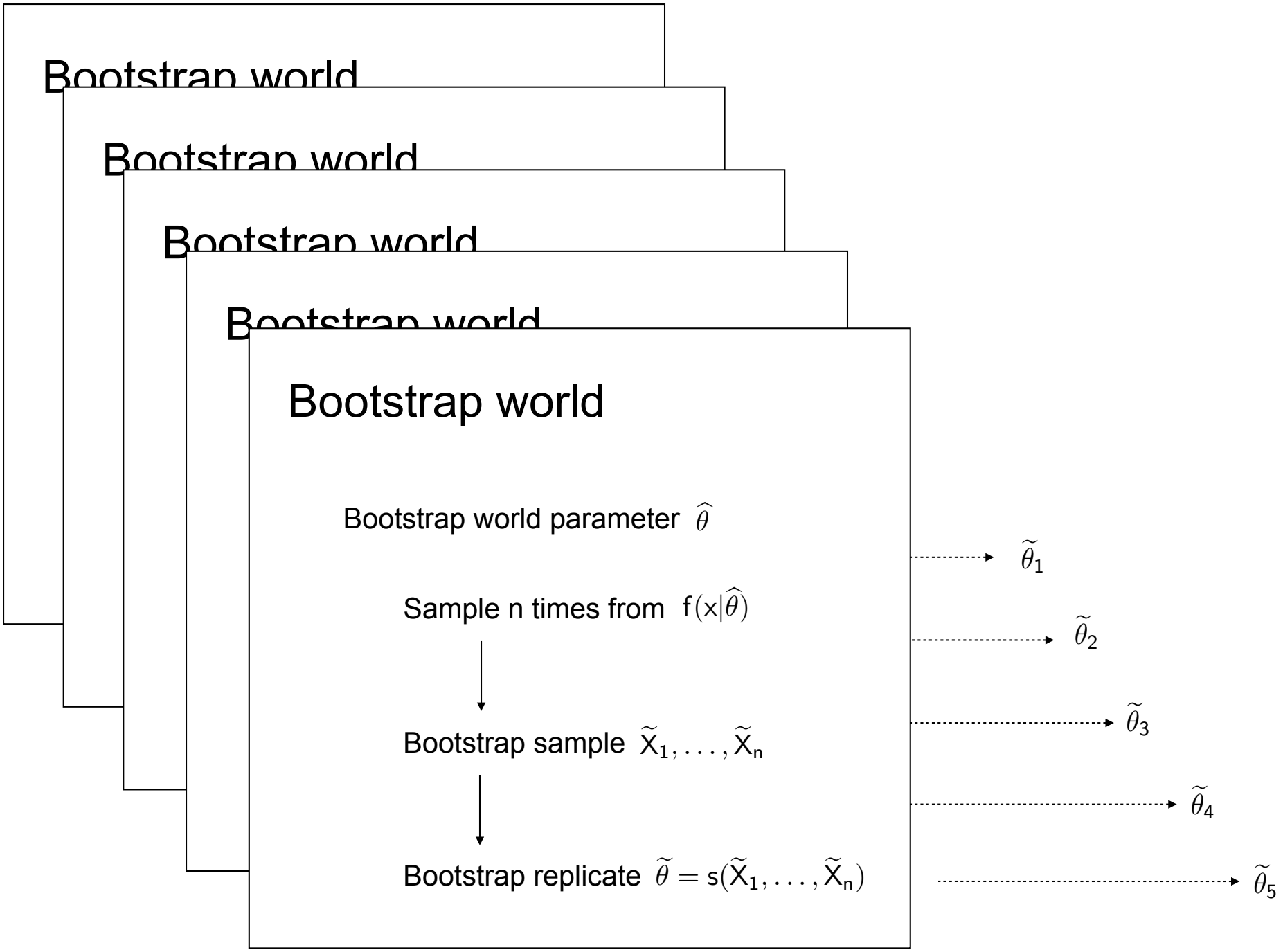
Sample n times from $f(x|\widehat{\theta})$

↓

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 $\overline{\widetilde{\theta}}$ (horrible notation) denote the mean of the B bootstrap samples

$$\overline{\widetilde{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\widetilde{\theta}_b$$

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

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

$$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widetilde{\theta}_b - \overline{\widetilde{\theta}})^2}$$

Nest box 8

Recall, for example, our binomial trails involving the presence or absence of a bird in Nest Box 8 -- We had 15 observations from a binomial distribution with m=14 (the study lasted two weeks) and success probability p, the parameter we wish to estimate

Here, again, are the data

11 10 7 7 9 10 10 11 12 10 8 8 0 8 12

which sum to 133 -- Our estimate of p was then $\widehat{p} = 133/(14 * 15) = 133/210 = 0.63$

Nest box 8

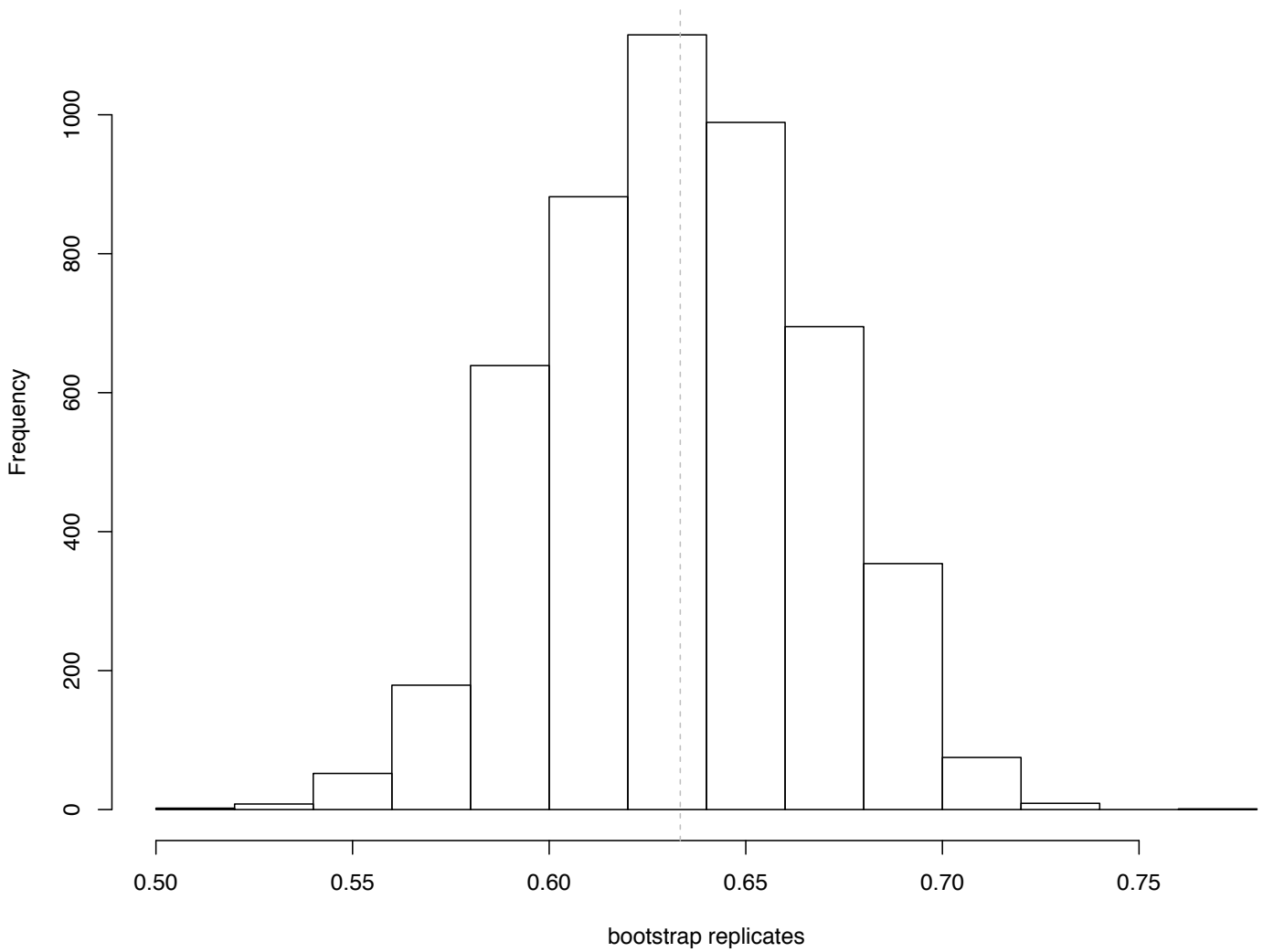
We then regenerate our data using $\hat{p} = 0.63$ -- So, for example, we can draw 15 new numbers using `rbinom()` in R

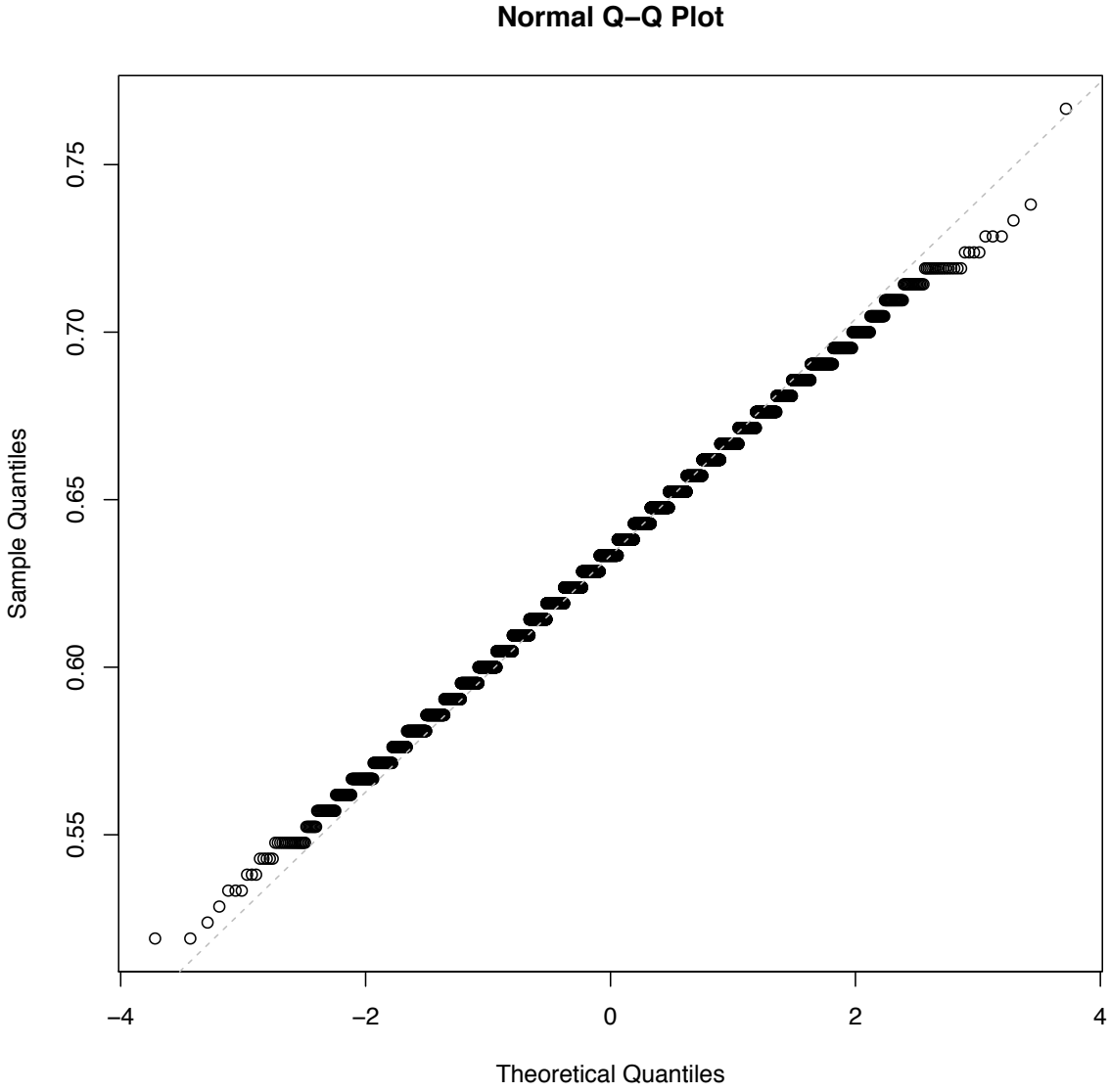
```
> rbinom(15,size=14,prob=133/210)
 [1] 11 10  8 11 10 10  9 11  8  9  5  8 11  9  9

> rbinom(15,size=14,prob=133/210)
 [1]  9 10 10  6  7  8  9  9  9  9 12  9  7  7 10

> rbinom(15,size=14,prob=133/210)
 [1]  8 11 11 10  9  6  9  9  6  7  9 10  8  8  7
```

Nest box 8

Or, putting things in a loop, we can generate B=5,000 bootstrap replicates of our estimate, again using $\hat{p} = 0.63$

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

> for(i in 1:5000){

    # generate new data
    b = rbinom(15,size=14,prob=133/210)

    # form an estimate based on that data
    boot[i] = sum(b)/210
 }

> mean(boot)
[1] 0.6338705

> sd(boot)
[1] 0.03323094

> hist(boot)
> abline(v=133/210)

> qqnorm(boot)
> qqline(boot)
```


Histogram of B=5,000 bootstrap replicates
Frequency
0
200
400
600
800
1000
0.50
0.55
0.60
0.65
0.70
0.75
bootstrap replicates


Normal Q–Q Plot
Sample Quantiles
0.55
0.60
0.65
0.70
0.75
Theoretical Quantiles
−4
−2
0
2
4

The bootstrap

We call this scheme the **“parametric” bootstrap** because it relies on substituting or “plugging in” an estimate of the parameter of interest to make statements about the sampling distribution

This is the simplest example of the so-called **“plug-in” principle** (yes, sadly, that’s what it’s called) -- In the next lecture we will see a much more elaborate version where we relax the parametric origin story for the data and instead “plug-in” our sample as a replacement for an unknown population

We should mention that the bootstrap is not the only tool for estimating the sampling distribution and that decades of work in statistics has produced “**analytical” approaches (based, for example on the CLT)** that we will review shortly

We spend time with the bootstrap because it **will work when formula don’t exist, but agree with them when they do** -- Not a bad trick!

The bootstrap

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

While bias and standard error are extremely informative, another use for the sampling distribution is the construction of a so-called **confidence interval for our unknown parameter**

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

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

Confidence intervals

Suppose we knew the sampling distribution exactly -- That is, we knew θ^* and could either work out the math to give us the sampling distribution for $\widehat{\theta}$ or literally repeat the exact experiment multiple times, relying on simulation for an approximation

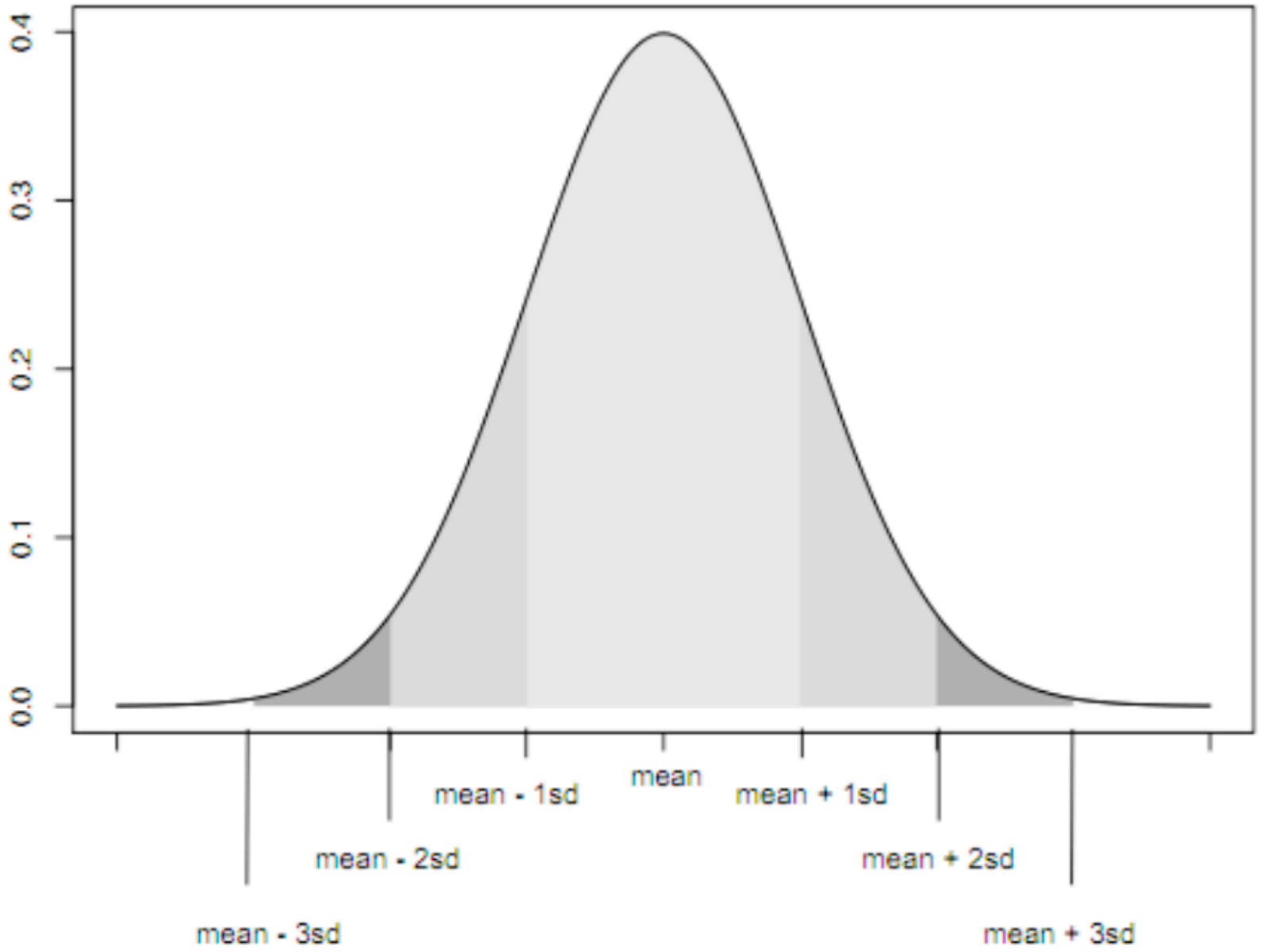
A 90% confidence interval, say, is constructed by an algorithm or rule that ensures that, if we apply this rule across all replications of our experiment, 9**0% of the intervals would contain the population parameter** θ^*

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

A simple example

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

Recall that for any normal distribution...

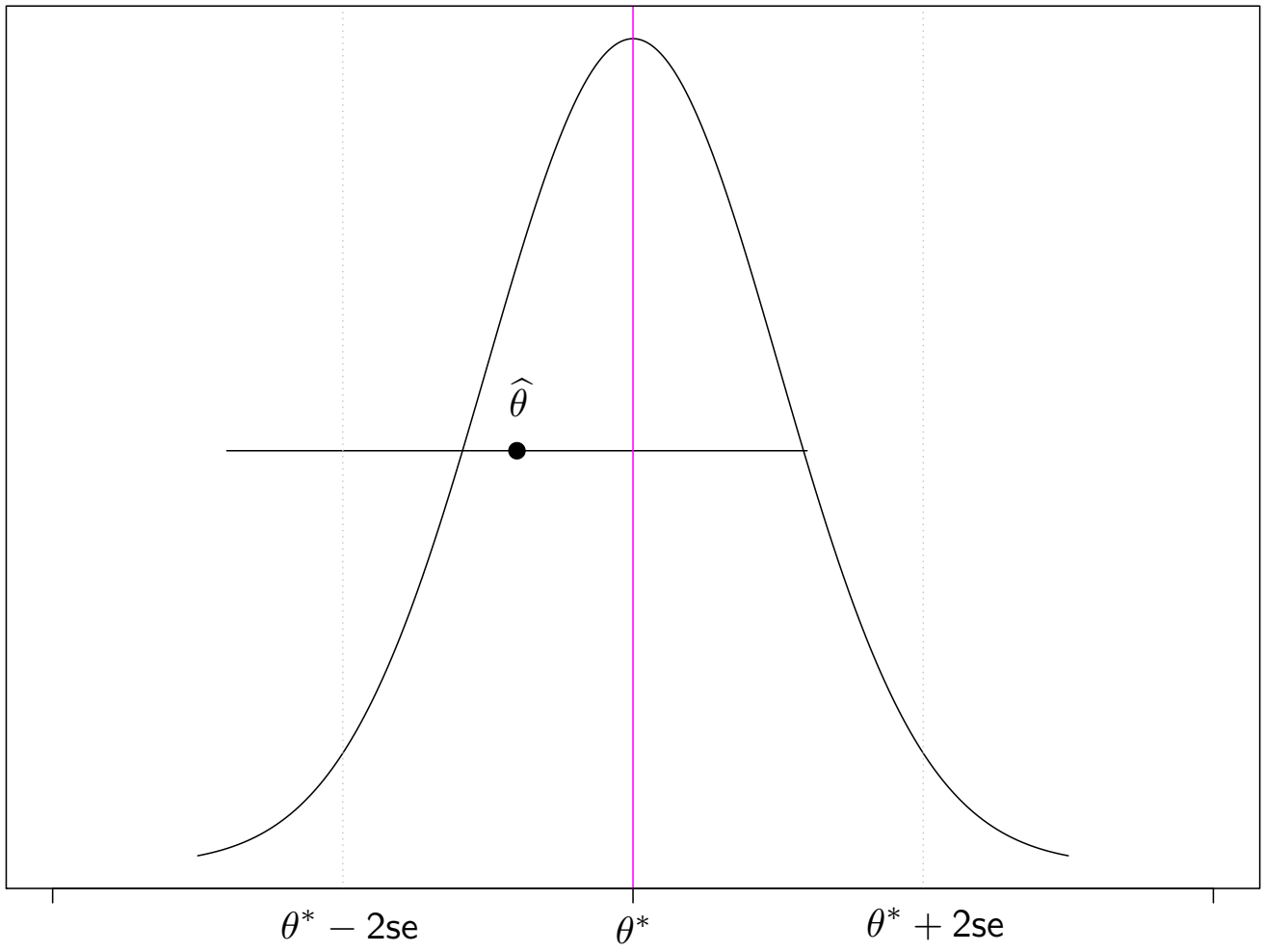

0.0
0.1
0.2
0.3
0.4
mean - 3sd
mean - 2sd
mean - 1sd
mean
mean + 1sd
mean + 2sd
mean + 3sd

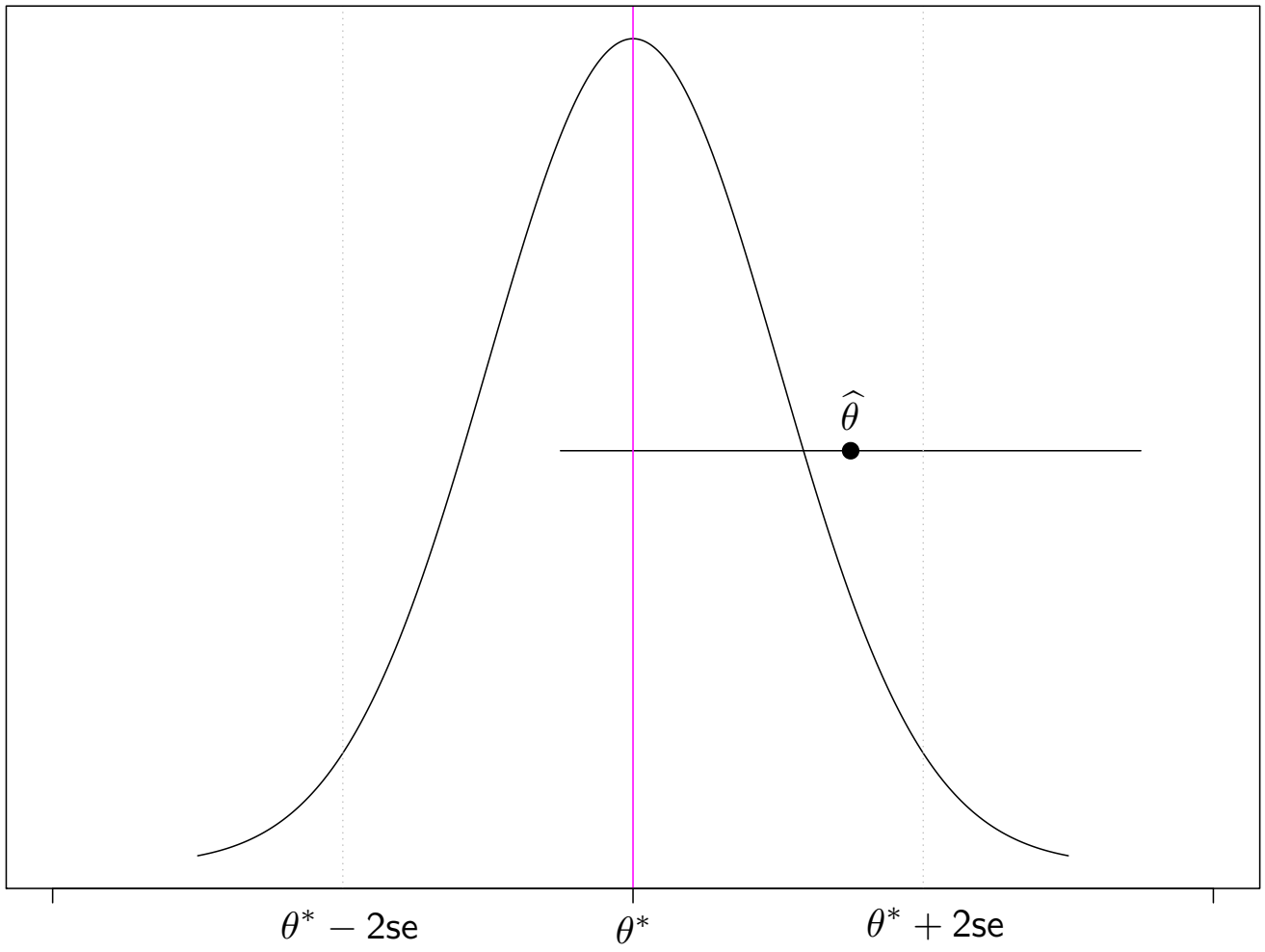
A simple rule

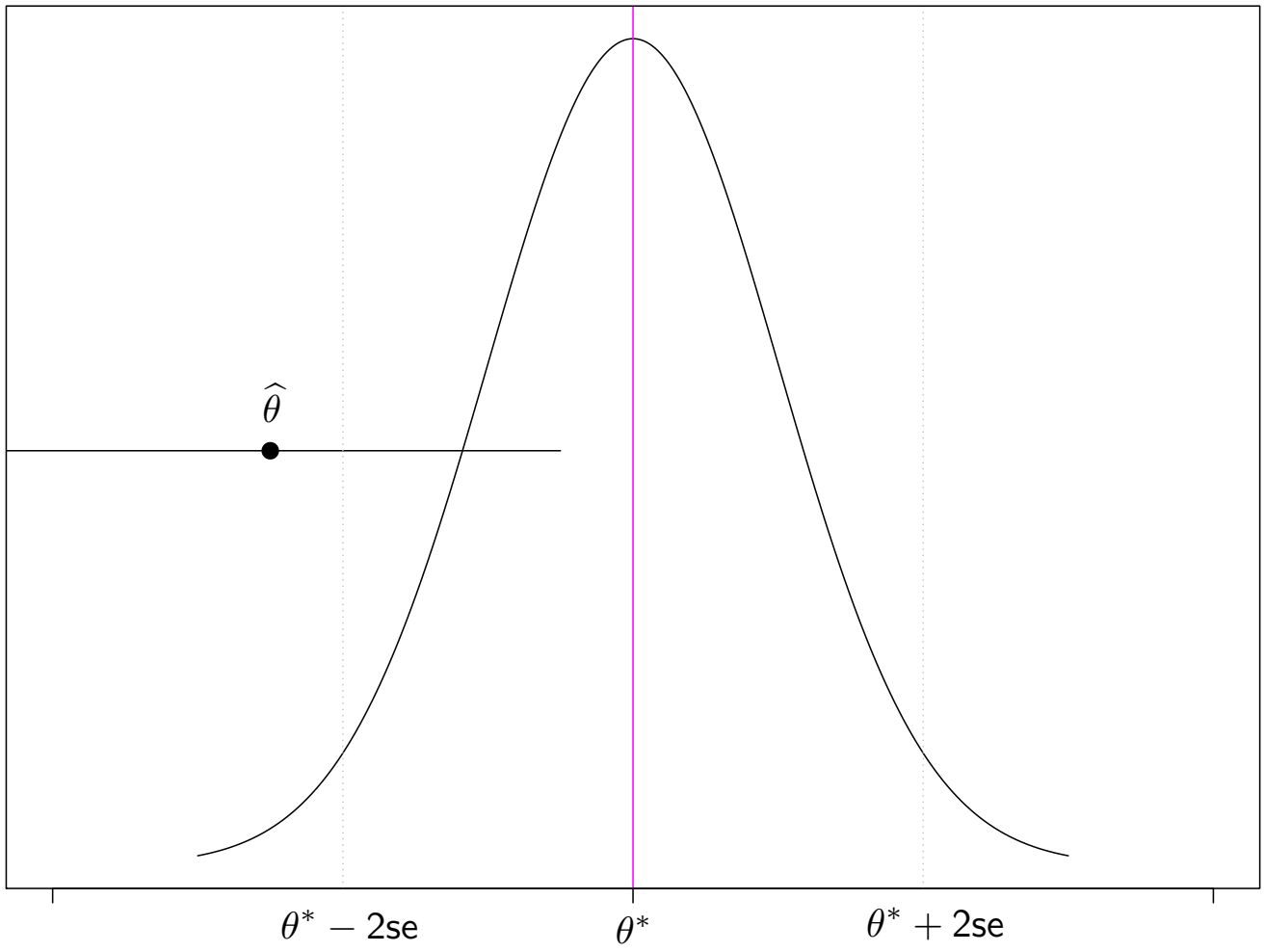
If our sampling distribution is centered on the population parameter θ^* (our estimate is unbiased), this rule implies that 95% of the samples we could take produces estimates $\widehat{\theta}$ that are within two standard errors of θ^* (finessing 1.96 v. 2.0)

Here's a simple rule: For any estimate, form the interval $\widehat{\theta} \pm 2\,\text{se}$ -- Let's see how this might perform

NOTE: For the next couple of slides we are going to assume that the standard error of our estimate is known -- That rarely happens but it's easier to describe the concept if we assume $\widehat{\theta}$ has a sampling distribution that is centered on θ^* and has a spread se


$\widehat{\theta}$
$\theta^* - 2\text{se}$
θ^*
$\theta^* + 2\text{se}$


$\widehat{\theta}$
$\theta^* - 2\text{se}$
θ^*
$\theta^* + 2\text{se}$


$\widehat{\theta}$
$\theta^* - 2\text{se}$
θ^*
$\theta^* + 2\text{se}$

A simple rule

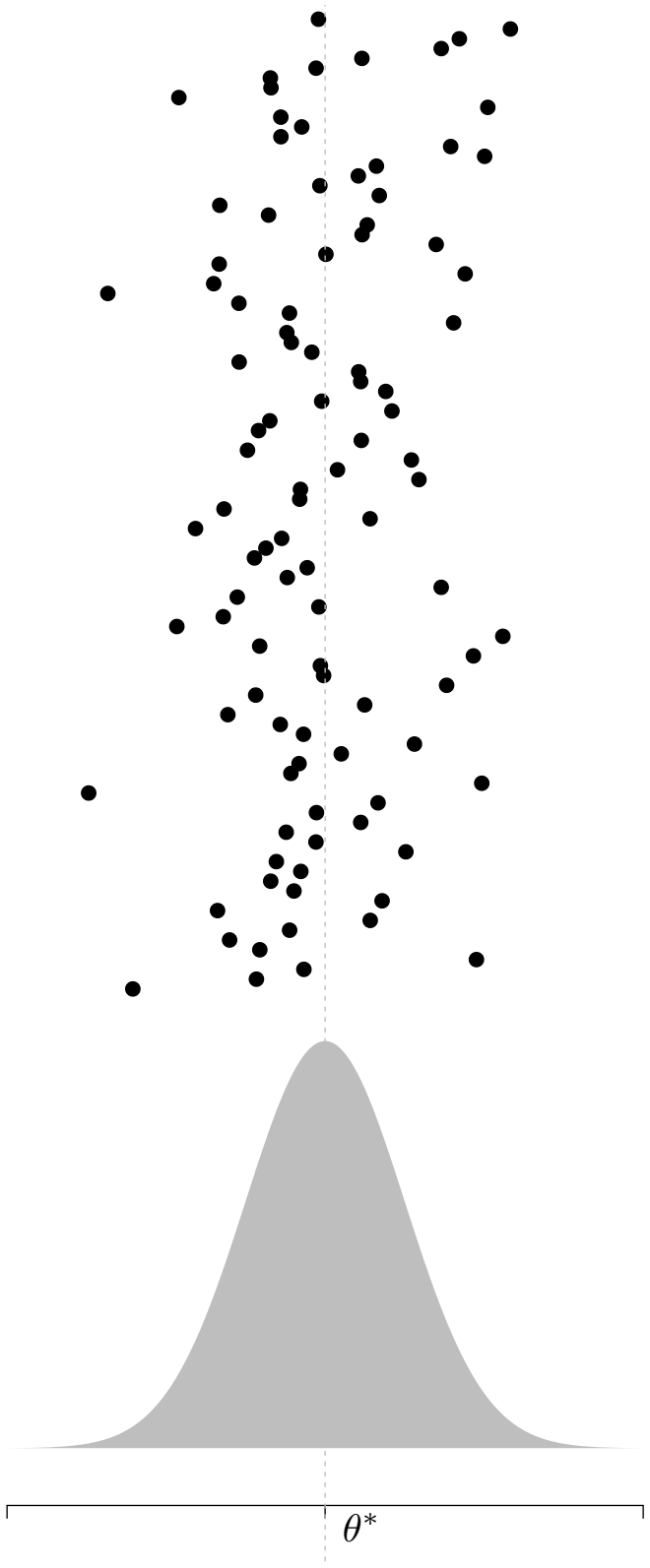
The curves on the previous slides represent what happens when we consider the sampling distribution, the set of all possible experimental results we could have seen -- Each $\widehat{\theta}$ represents one possible outcome and 95% of them are associated with intervals that contain our population parameter

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

Confidence intervals

To make this concrete, at the top each black dot represents an experimental result -- We generated a sample $X_1, \ldots, X_n$ from a parametric model $f(x|\theta^*)$ and estimated $\widehat{\theta}$

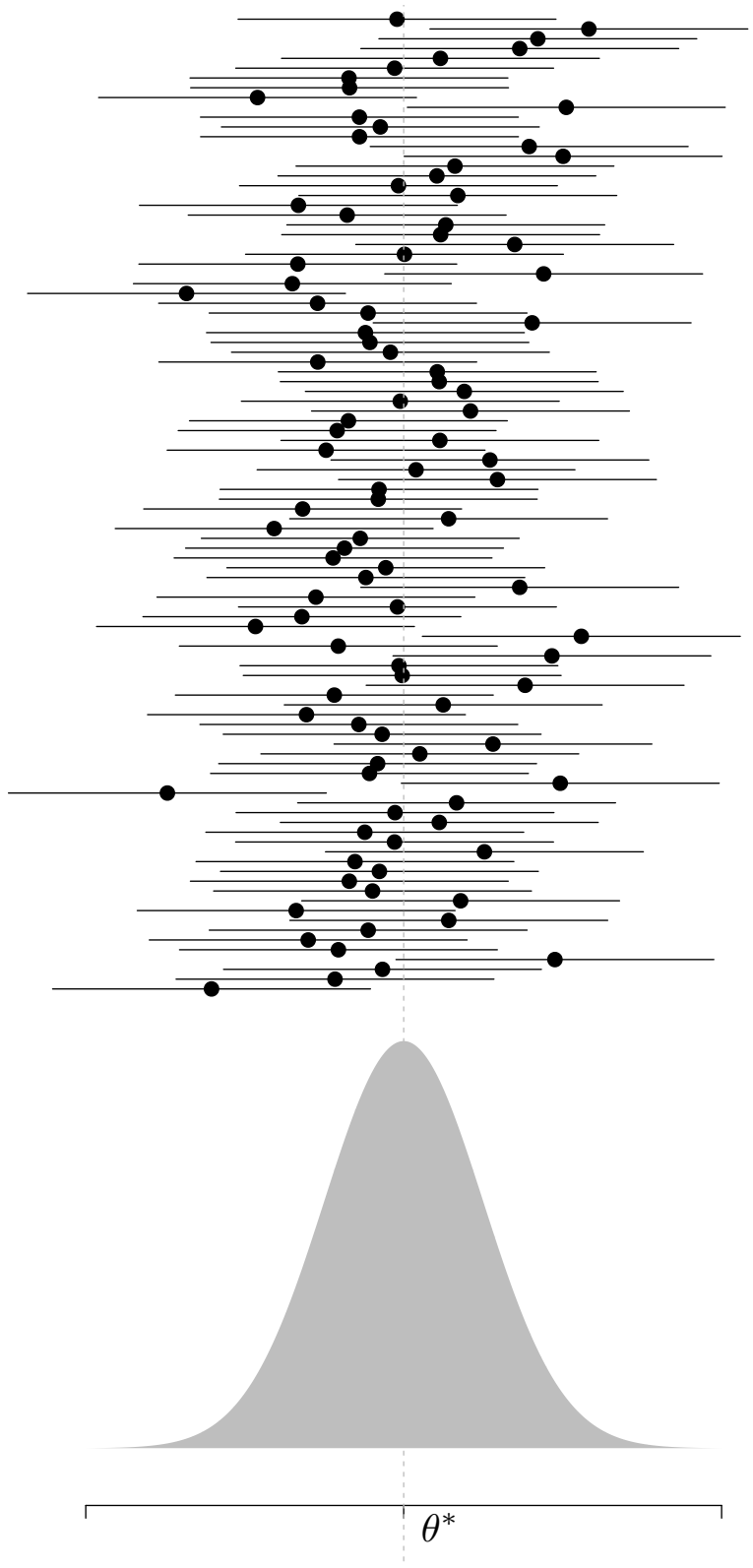
There are 100 black dots, representing 100 different sets of experimental outcomes -- The black dots are observations then from the sampling distribution and we can think of them as $\widehat{\theta}_1, \ldots, \widehat{\theta}_{100}$



Confidence intervals

For each estimate, or, rather, each time we perform our experiment, we can then form a 95% confidence interval $\widehat{\theta} \pm 2\,\mathrm{se}$

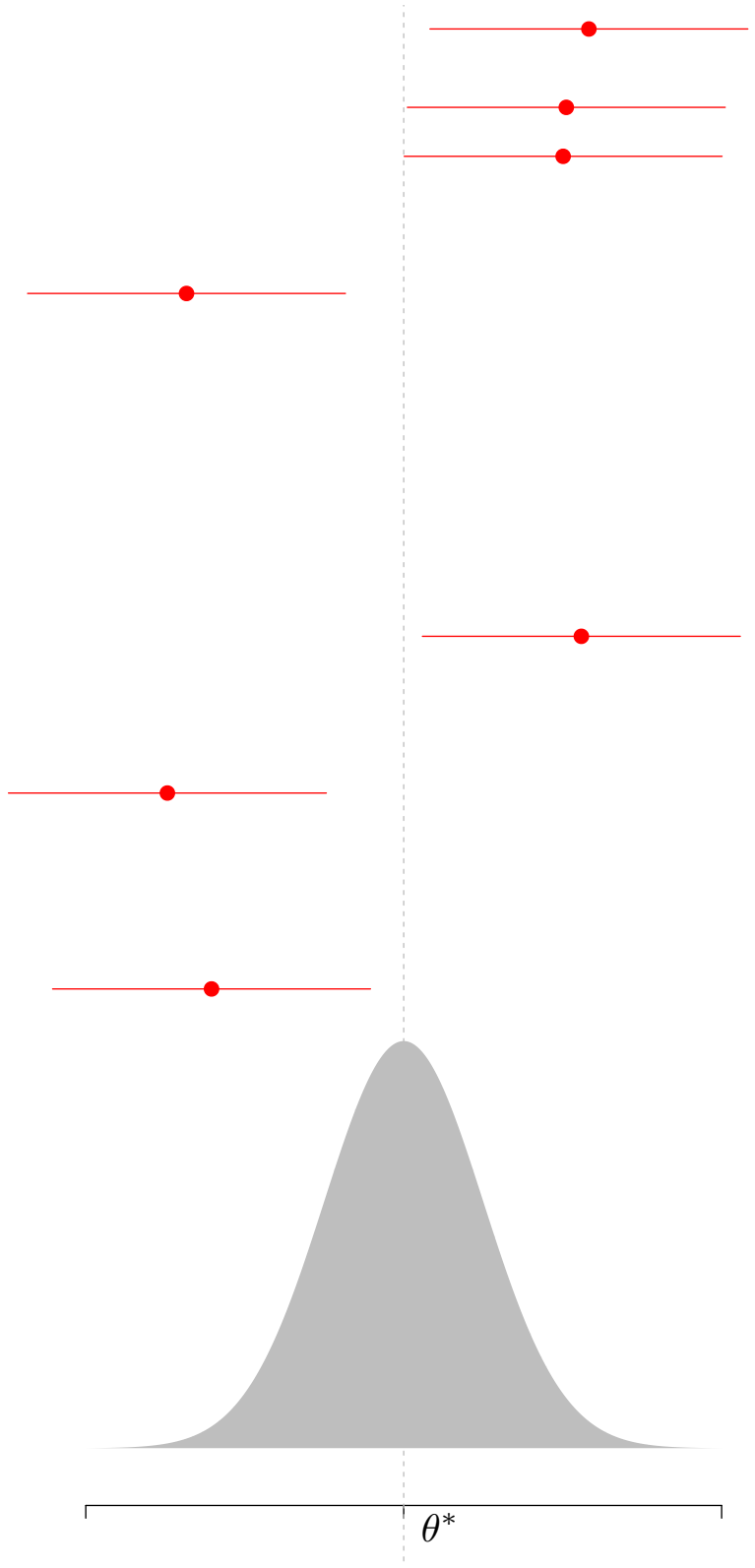
Given their construction, 95% of these intervals should cover the true parameter θ^* -- What do you think?



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 θ^* that generated the data



A simple rule

The last few slides were developed knowing that our estimate was unbiased and knowing the se -- **Without this information, we can rely on the bootstrap to help us create confidence intervals**

We can use the bootstrap, for example, to tell us about the bias (whether the sampling distribution is centered on θ^*) and we can estimate the standard error

For Nest Box 8, our estimate was $\widehat{\mathrm{p}} = 0.63$ and we saw no evidence of bias -- Our estimated standard error was 0.03

$$[0.63 - 2 * 0.03,\ 0.63 + 2 * 0.03] = [0.57,\ 0.69]$$

Interpretation

This interval is just one of many that we could have formed depending on the results of our experiment -- We have no idea whether or not the true parameter p is in the interval, just that across repeated sampling we would expect to see it in similar intervals often

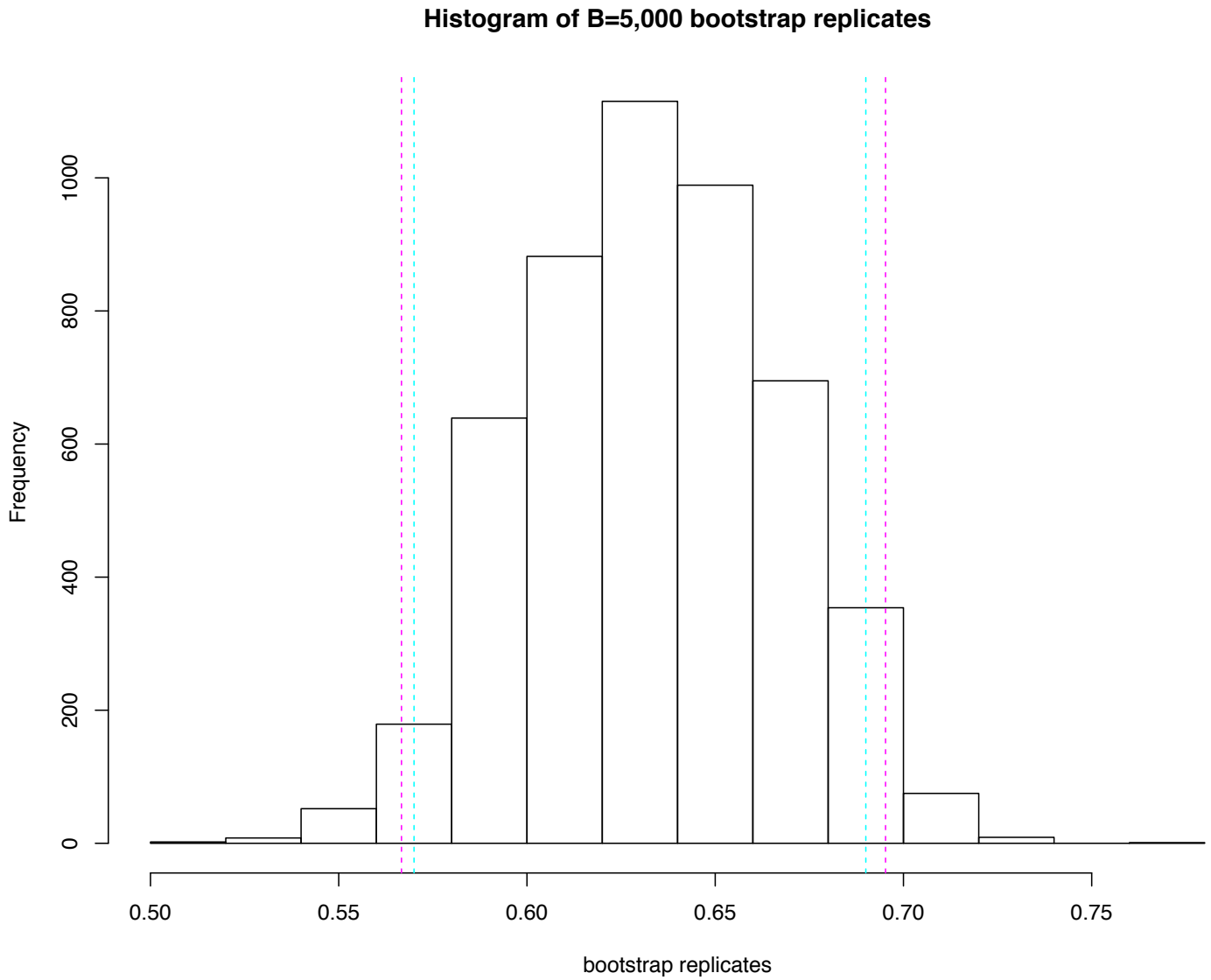
It is not correct to say that there is a 95% chance that the true data generating p is in the interval [0.57,0.69] -- The true p is fixed and unknown and this interval is fixed now that we've conducted our experiment

The true p is either in the interval or it is not -- The notion of randomness here is with respect to the imagined collection of all possible experimental results, of which our trial is just one observation

A more general rule

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

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



magenta: 0.025, 0.975 quantiles
cyan: $\widehat{\theta} \pm 2\,\widehat{\text{se}}$

Another approach

The bootstrap is a fairly new innovation, dating to the late 1970s -- Performing this kind of analysis before the days of a reasonably fast computer would be unthinkable

Prior to this, however, there was (and continues to be) a thriving statistics industry that works with other techniques for approximating the sampling distribution besides the boostrap

Let’s consider one of the first...

Confidence intervals

Suppose we draw n independent samples $X_1, \ldots, X_n$ from a normal distribution with mean μ and standard deviation σ

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

Therefore, the quantity

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

has (exactly) a standard normal distribution

Confidence intervals

For simplicity, let's round up 1.96: $\Pr\left(-2 \leq Z \leq 2\right) \approx 0.95$

Then, we can use our distributional results about $\overline{X}$ to write

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

Simplifying, we have

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

or rather

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

Confidence intervals

Let's word this in terms of our estimation vocabulary built up over the last few lectures

First, suppose $X_1, \ldots, X_n$ are independent draws from a normal distribution with unknown mean μ but known standard deviation σ

We recall that the MLE for μ is just $\overline{X}$, the sample mean; we also know it is an unbiased estimate of μ

Using basic facts about the normal distribution, we have shown that $(\overline{X} - \mu)/\text{se}\,(\overline{X})$ has a standard normal distribution so that $\overline{X} \pm 1.96\,\text{se}\,(\overline{X})$ is a 95% confidence interval for μ; that is

$$\Pr\big(\,\overline{X} - 1.96\,\text{se}\,(\overline{X}) \leq \mu \leq \overline{X} + 1.96\,\text{se}\,(\overline{X})\,\big) = 0.95$$

Some snags

The first obvious problem with this whole construction is that we are not often in a situation in which we know σ but not μ; instead, we are typically estimating both parameters of the normal distribution

Next, modeling our population as having a normal distribution is often a fairly restrictive assumption; even in the simple modeling contexts we have seen so far, departures from normality are common

Let's first consider relaxing the assumption on σ ...


GUINNESS®
ST. JAMES'S GATE BREWERY, DUBLIN

Some history

Guinness Brewing Company incorporated in 1886 and was soon the largest brewery in the world, delivering 1.5 million barrels a year in England, Ireland and around the world

At the time, brewers learned “meticulously the traditional practices” as apprentices; the Chairman and Managing Director at Guinness wanted to change all this **-- they wanted to make brewing scientific... and they invested heavily in the idea**

They started by **hiring top-notch chemists from Oxford and Cambridge** Universities; at a rate of one every one or two years starting in 1893 -- these chemists began projects to, for example, “identify and quantify what it was that gave hops and barley their brewing qualities”

This group **examined all aspects of production**, from the “raw materials” to how best to cultivate, fertilize, dry and store the barley

Some history

Guinness supported this group as they ran agricultural experiments (first renting, then buying farms) to acquiring an experimental malthouse to finally a separate brewery they could use to conduct their experiments

"Reading led to analysis, experiments and measurements. They began to accumulate data and, at once, ran into difficulties because their measurements varied. The effects they were looking for were not usually clear cut or consistent, as they had expected, and they had **no way of judging whether the differences they found were effects of treatment or accident**"

In October of 1899 Wiliam S. Gosset, a recent chemistry graduate from Oxford was hired into this group; he had studied some mathematics and so this group brought him their data to analyze -- Gosset wrote "It may seem strange that reasoning of this nature had not been widely made use of, but this is due, first, to the popular dread of mathematics"

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)

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

Any experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

By Student

Gosset's research was conducted while he was employed at Guinness; the Managing Director agreed that he could publish his results but added "that such publication might be made without the brewer's name appearing -- instead Guinness employees would be designated **'Pupil' or 'Student'"**

While we mentioned that Gosset studied a little mathematics, **he was by no means a mathematician**; his paper includes a couple insightful moves that have little to do with formal proof

So, it's 1905, say, and you want to study a sampling distribution and the mathematics is beyond you: What do you do? Hint: What would you do now?

LITERATURE.

(1) Frew, J. G. H. (1923). On the larval anatomy of *Chlorops taeniopus* Meig. and two related Acalyptrate Muscids, with notes on their winter host plants. (*P.Z.S.* p. 783.)

(2) Fulmek, L. (1911). Zum Auftreten der Halmfliege (*Chlorops taeniopus* Meig.) in Weizen. (*Osterreichischen Agrar-Zeitung* Nr. 30 vom 29 Juli.)

(3) Nowicki, M. (1871). Ueber die Weizenverwusterin *Chlorops taeniopus* Meig. und die Mittel zu ihrer Bekampfung. (*Verh. Zoologischbotanischen Gesellschaft in Wien.*)

(4) Ormerod, E. A. (1890). Manual of Injurious Insects and Methods of Prevention (2nd ed. London, pp. 75–79).

(5) "Mathetes" (1924). Statistical study on the effect of manuring on infestation of barley by Gout Fly. (*Ann. App. Biol.* xi, 2.)

By Student

You simulate! Reading over the **first part of Section VI**, you can see how painful this task was; it involved creating a physical version of the population, with values written on cards

Ah, but another question arises: It is 1905 and you need observations from a normal distribution; Where do you turn?

Now 50 to 1 corresponds to three times the probable error in the normal curve and for most purposes would be considered significant; for this reason I have only tabled my curves for values of n not greater than 10, but have given the $n=9$ and $n=10$ tables to one further place of decimals. They can be used as foundations for finding values for larger samples*.

The table for $n=2$ can be readily constructed by looking out $\theta=\tan^{-1}z$ in Chambers' Tables and then $\cdot5+\theta/\pi$ gives the corresponding value.

Similarly $\frac{1}{2}\sin\theta+\cdot5$ gives the values when $n=3$.

There are two points of interest in the $n=2$ curve. Here s is equal to half the distance between the two observations. $\tan^{-1}\frac{s}{s}=\frac{\pi}{4}$ so that between $+s$ and $-s$ lies $2\times\frac{\pi}{4}\times\frac{1}{\pi}$ or half the probability, i.e. if two observations have been made and we have no other information, it is an even chance that the mean of the (normal) population will lie between them. On the other hand the second moment coefficient is

$$\frac{1}{\pi}\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}}\tan^2\theta\,d\theta=\frac{1}{\pi}\Big[\tan\theta-\theta\Big]_{-\frac{\pi}{2}}^{\frac{\pi}{2}}=\infty\,,$$

or the standard deviation is infinite while the probable error is finite.

Section VI. *Practical Test of the foregoing Equations.*

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 (*Biometrika*, Vol. I. p. 219). 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 the 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, giving us the z of Section III.

This provides us with two sets of 750 standard deviations and two sets of 750 z's on which to test the theoretical results arrived at. The height and left middle finger correlation table was chosen because the distribution of both was approximately normal and the correlation was fairly high. Both frequency curves, however, deviate slightly from normality, the constants being for height $\beta_1=\cdot0026$, $\beta_2=3\cdot175$, and for left middle finger lengths $\beta_1=\cdot0030$, $\beta_2=3\cdot140$, and in consequence there is a tendency for a certain number of larger standard deviations to occur than if the distributions were normal. This, however, appears to make very little difference to the distribution of z.

* E.g. if $n=11$, to the corresponding value for $n=9$, we add $\frac{1}{8}\times\frac{9}{10}\times\frac{7}{8}\times\frac{5}{6}\times\frac{3}{4}\cos^8\theta\sin\theta$; if $n=13$ we add as well $\frac{1}{10}\times\frac{9}{12}\times\frac{7}{10}\times\frac{5}{8}\times\frac{3}{6}\times\frac{1}{4}\cos^{10}\theta\sin\theta$ and so on.

† I hope to publish the results of the correlation work shortly.

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...***

TABLE III. 3000 *Criminals.* Height (feet and inches).

Left Middle Finger (millimetres).	4′ 7″⅛—8″⅛	8″⅛—9″⅛	9″⅛—10″⅛	10″⅛—11″⅛	5′ 11″⅛—0″⅛	0″⅛—1″⅛	1″⅛—2″⅛	2″⅛—3″⅛	3″⅛—4″⅛	4″⅛—5″⅛	5″⅛—6″⅛	6″⅛—7″⅛	7″⅛—8″⅛	8″⅛—9″⅛	9″⅛—10″⅛	10″⅛—11″⅛	6′ 11″⅛—0″⅛	0″⅛—1″⅛	1″⅛—2″⅛	2″⅛—3″⅛	3″⅛—4″⅛	4″⅛—5″⅛	Totals
9·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
8	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
9	—	—	1	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
10·0	1	—	—	1	2	—	2	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	7
1	—	—	—	1	3	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	10
2	—	—	2	2	2	1	—	2	—	1	—	—	—	—	—	—	—	—	—	—	—	—	17
3	—	1	1	3	2	2	3	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	20
4	—	—	1	1	2	3	3	4	3	3	—	—	—	—	—	—	—	—	—	—	—	—	30
5	—	—	—	1	3	7	6	4	3	1	3	1	—	1	—	—	—	—	—	—	—	—	44
6	—	—	—	1	4	5	9	14	6	3	1	—	—	1	—	—	—	—	—	—	—	—	74
7	—	—	1	2	4	9	14	16	15	7	3	1	2	—	—	—	—	—	—	—	—	—	75
8	—	—	—	2	5	6	14	27	10	7	1	2	1	—	—	—	—	—	—	—	—	—	102
9	—	—	—	—	2	6	14	24	27	14	10	4	1	—	—	—	—	—	—	—	—	—	163
11·0	—	—	—	2	6	12	15	31	37	27	17	10	6	—	—	—	—	—	—	—	—	—	152
1	—	—	—	3	3	12	22	26	24	26	24	7	4	1	—	—	—	—	—	—	—	—	183
2	—	—	—	3	2	7	21	30	38	29	27	20	4	1	—	—	—	—	—	—	—	1	164
3	—	—	—	1	—	5	10	24	26	39	26	24	7	2	—	—	—	—	—	—	—	—	228
4	—	—	—	—	3	4	9	29	56	58	26	22	10	11	—	—	—	—	—	—	—	—	233
5	—	—	—	—	—	5	11	17	33	57	38	34	25	11	2	—	—	—	—	—	—	—	226
6	—	—	—	—	2	1	4	13	37	39	48	38	27	12	2	2	—	—	—	—	—	—	232
7	—	—	—	—	—	2	9	17	30	37	48	45	24	9	2	2	—	1	—	—	—	—	184
8	—	—	—	—	1	—	2	11	15	35	41	34	29	10	5	1	—	—	—	—	—	—	162
9	—	—	—	—	1	1	2	12	10	27	32	35	19	10	9	3	1	—	—	—	—	—	163
12·0	—	—	—	—	—	—	1	4	8	19	42	39	22	16	8	2	2	—	—	—	—	—	126
1	—	—	—	—	—	—	—	2	4	13	22	28	15	27	10	4	1	—	—	—	—	—	91
2	—	—	—	—	—	—	1	2	5	6	23	17	16	11	8	1	1	—	—	—	—	—	89
3	—	—	—	—	—	—	—	—	4	8	10	13	20	23	6	5	—	—	—	—	—	—	44
4	—	—	—	—	—	—	1	1	1	2	7	12	4	7	7	1	—	—	1	—	—	—	52
5	—	—	—	—	—	—	—	1	—	1	3	12	11	8	6	8	—	2	—	—	—	—	35
6	—	—	—	—	—	—	—	—	—	1	—	3	5	7	8	6	3	1	1	—	—	—	31
7	—	—	—	—	—	—	—	—	—	1	1	7	5	5	8	2	2	—	—	—	—	—	25
8	—	—	—	—	—	—	—	—	—	—	1	2	3	1	8	5	3	1	1	—	—	—	7
9	—	—	—	—	—	—	—	—	—	—	—	1	2	2	—	1	1	—	—	—	—	—	8
13·0	—	—	—	—	—	—	—	—	—	—	3	—	1	—	1	—	2	1	—	—	—	—	2
1	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	1	—	—	—	—	6
2	—	—	—	—	—	—	—	—	—	—	1	1	—	1	—	3	—	—	—	—	—	—	2
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	1	—	—	—	0
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	
Totals	1	1	6	23	48	90	175	317	393	462	458	413	264	177	97	46	17	7	4	0	0	1	3000
Means	100	103	102·8	107·0	107·8	109·4	110·6	111·8	113·3	114·8	116·5	117·7	118·6	120·1	122·2	123·9	125·9	126·4	127·7	—	—	112	

To clarify why measurements of criminals were in the "public data domain"
From the first page of the article...

PART I.

Material and Methods.

(1) The object of this memoir is threefold:

(i) To test to what extent the criminal classes diverge in physical characters from other classes of the community.

(ii) To consider how far the shorter methods recently proposed by Professor Karl Pearson for finding the variability and correlation of characters in the case of normal frequency may be applied to some of the chief anthropometric measurements now customarily made, and

(iii) To determine what is the best manner in which these measurements can be applied to the identification of criminals.

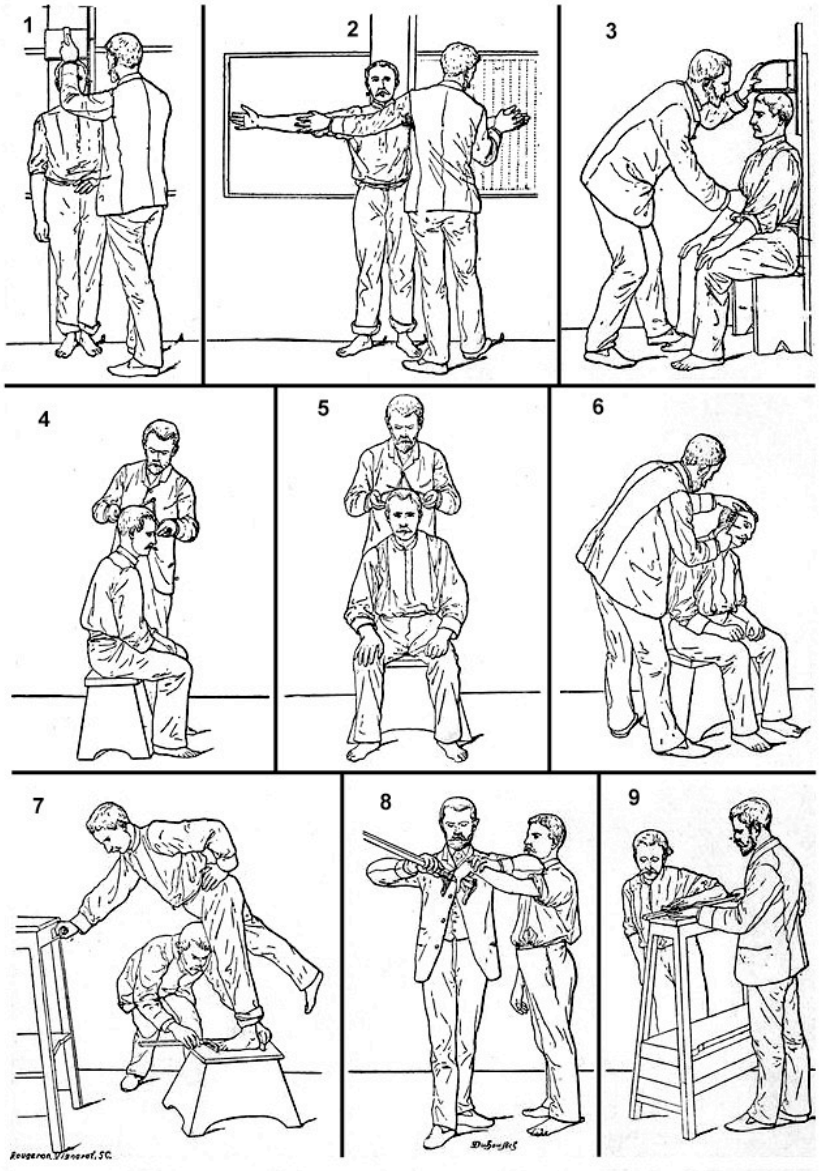
And later some conclusions...

Summing up the results of this part of the inquiry, I conclude that there is a substantial difference in stature, and in size and shape of head between the two classes; I do not assert that the source of the criminality is to be found in this difference, but only that criminals are drawn from a different section of the community. As bearing on this point it is worth noting that the mean height in Galton's middle-class measurements at the International Exhibition of 1884, viz. 67″·9, approaches our criminal mean more closely than does the Cambridge mean.

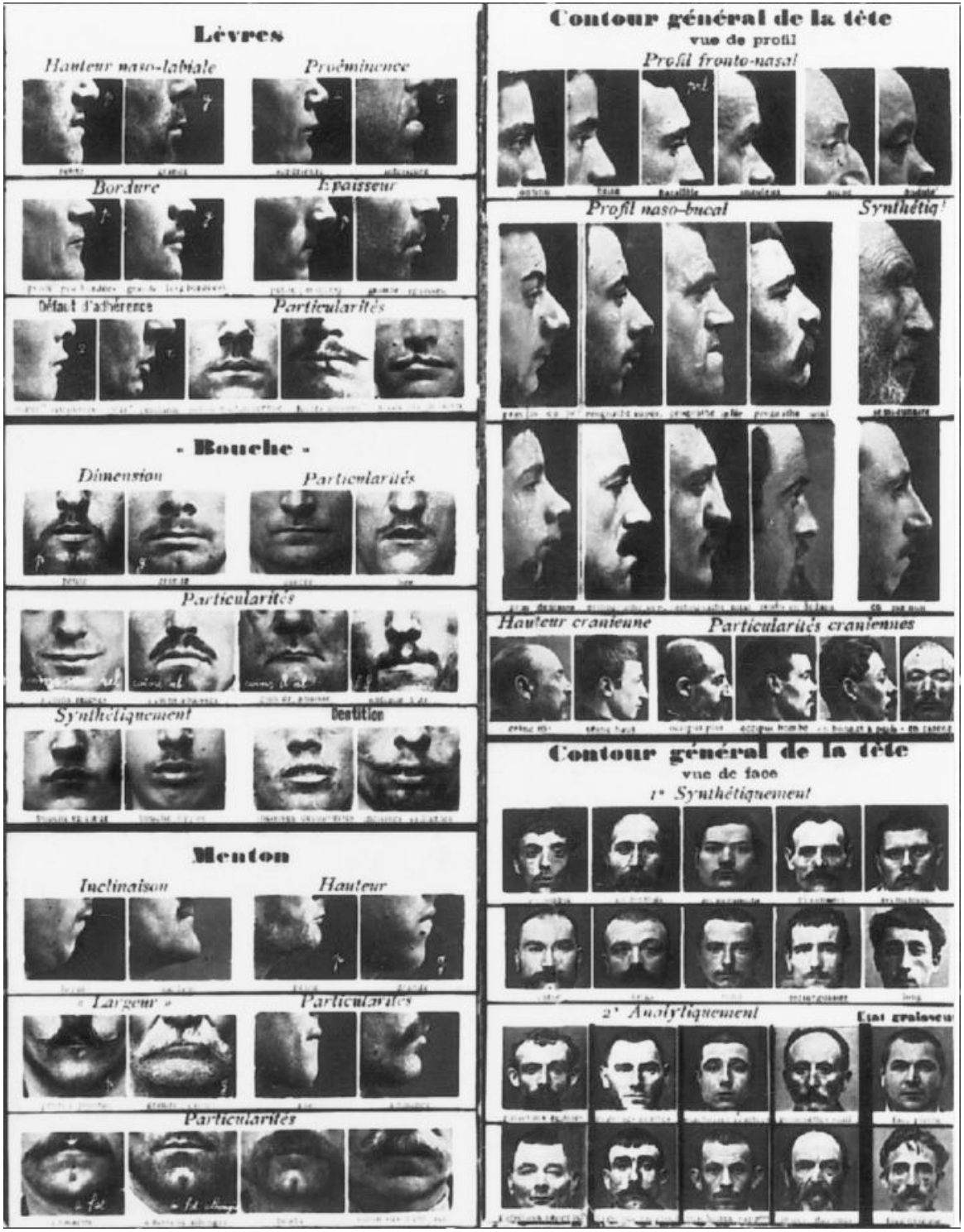
From "On Criminal anthropometry and the identification of criminals" by W. R. MacDonell, *Biometrika* 1, Vol 1, 1902

"Every measurement slowly reveals the workings of the criminal. Careful observation and patience will reveal the truth."

Alphonse Bertillon
French criminologist

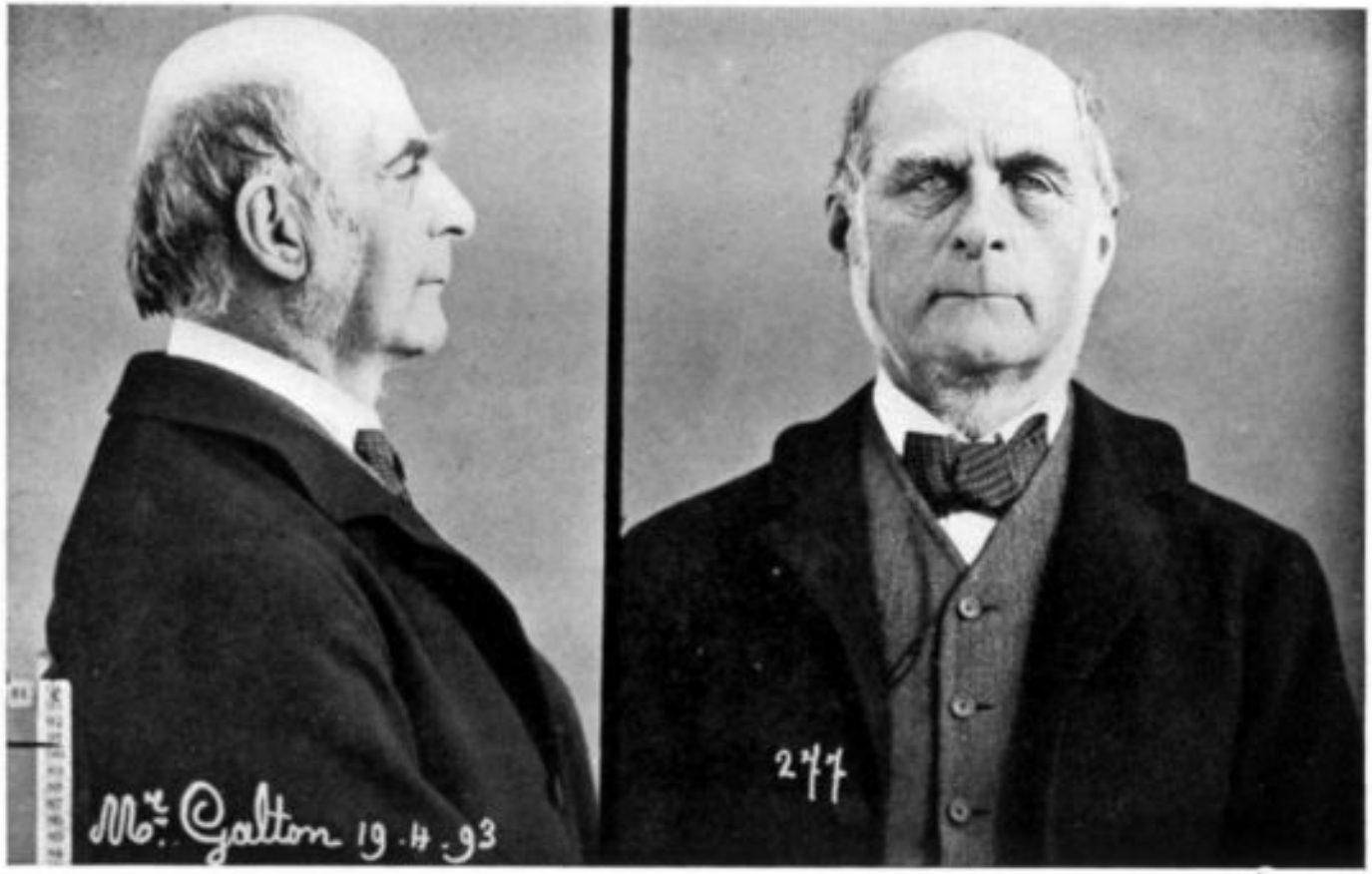


1. Height.
2. Reach.
3. Trunk
4. Length of head.
5. Width of head.
6. Right ear.
7. left foot.
8. Left middle finger.
9. Left forearm.


Lèvres
Hauteur naso-labiale
Proéminence
Bordure
Epaisseur
Défaut d'adhérence
Particularités
« Bouche »
Dimension
Particularités
Particularités
Synthétiquement
Dentition
Menton
Inclinaison
Hauteur
« Largeur »
Particularités
Particularités
Contour général de la tête
vue de profil
Profil fronto-nasal
Profil naso-bucal
Synthétiq!
Hauteur crânienne
Particularités crâniennes
Contour général de la tête
vue de face
1° Synthétiquement
2° Analytiquement
Etat graisseux

Taille 1m	Oreille dr. Tête. Long'	Pied g.	Coul' de l'iris g. N° de cl.	Agé de
Voute	Larg'	Médius g.	Aur'e	né le
Enverg 1m	Long'	Auric'e g.	Pér'e	à
Buste 0,	Larg'	Coudée g.	Part'e	dep't
				âge app't

(Réduction photographique 1/7.)

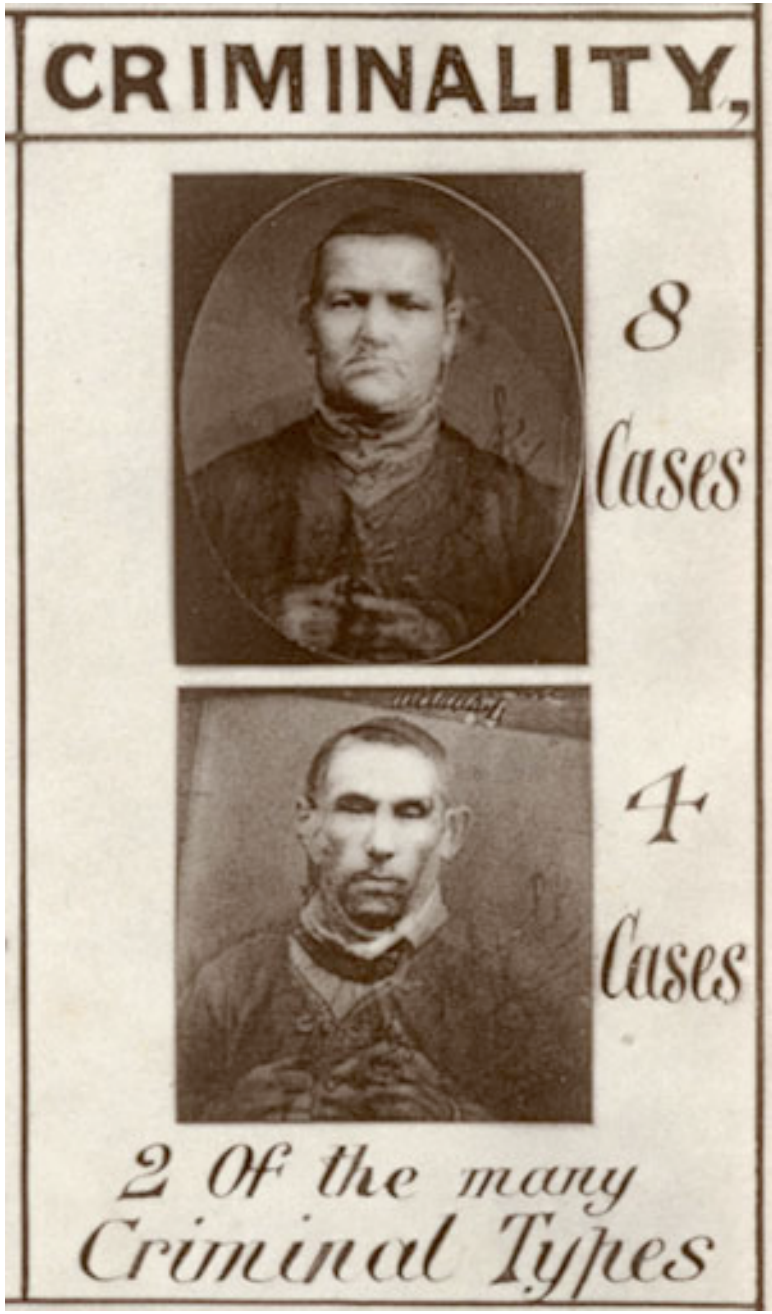


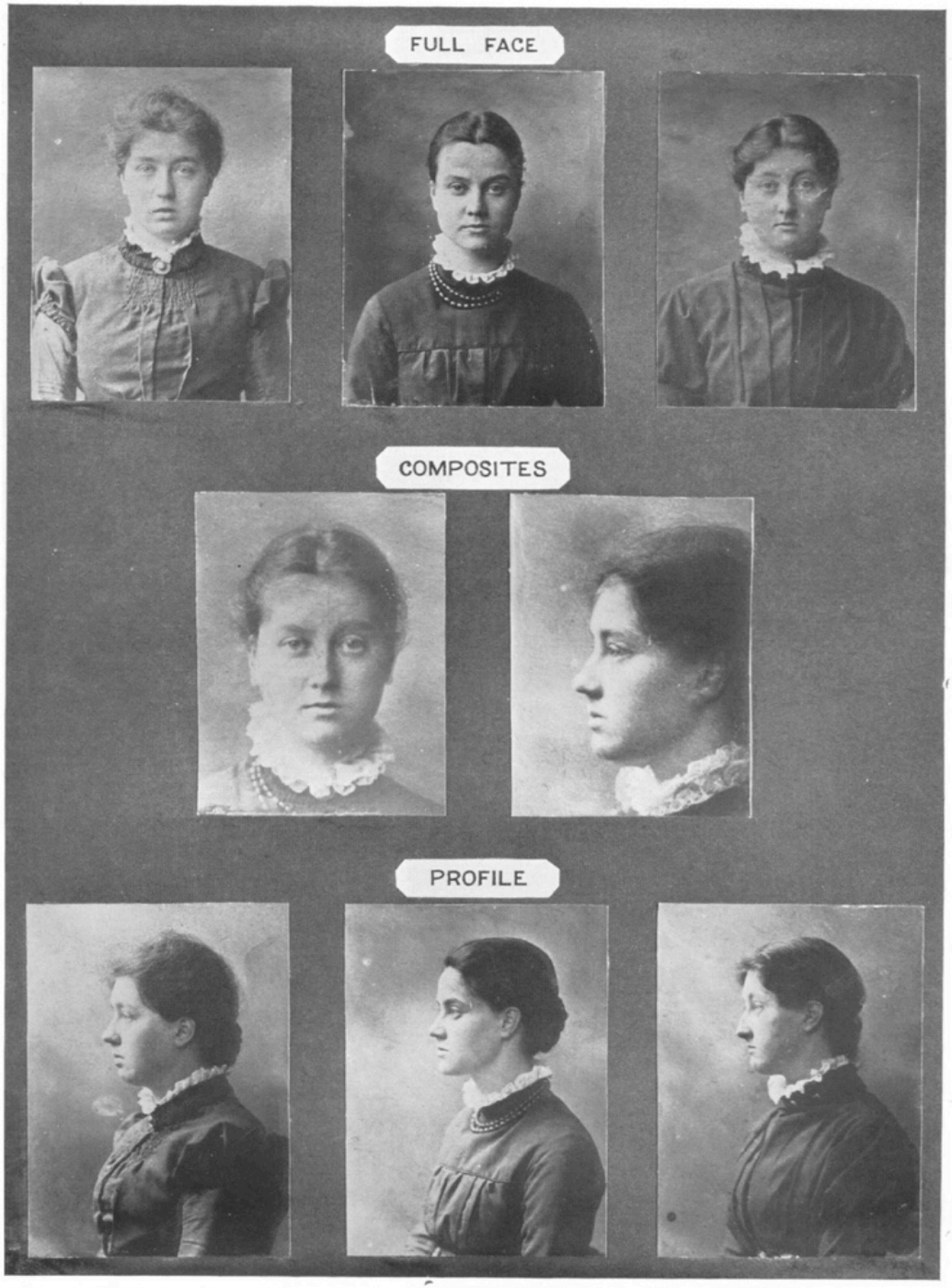
Galton

As we will see, Galton was deeply committed to the idea of **the normal curve as an important force in nature** and (as with Quetelet) thought the mean value had particular importance as **an indicator of "type"**

Quetelet was more extreme than Galton, however, in that he believed deviations from the mean were more like small errors, and **regarded the mean as something perfect or ideal**

For Galton, these types were stable from generation to generation -- You can see this in his work on fingerprints or even in his **composite photography**





Portraits of three Sisters, full face and profile, with the corresponding Composites.


5 COMPONENTS
7 COMPONENTS
4 COMPONENTS
2 COMPONENTS
PREVALENT TYPES OF FEATURES AMONG MEN CONVICTED OF LARCENY (WITHOUT VIOLENCE)

SPECIMENS OF COMPOSITE PORTRAITURE

PERSONAL AND FAMILY.

Alexander the Great From 6 Different Medals.

Two Sisters.

From 6 Members of same Family Male & Female.

HEALTH.

23 Cases. Royal Engineers. 12 Officers. 11 Privates

DISEASE.

6 Cases

9 Cases

Tubercular Disease

CRIMINALITY.

8 Cases

4 Cases

2 Of the many Criminal Types

CONSUMPTION AND OTHER MALADIES

I

20 Cases

II

36 Cases

56 Cases

Co-composite of I & II

Consumptive Cases.

100 Cases

50 Cases

Not Consumptive.

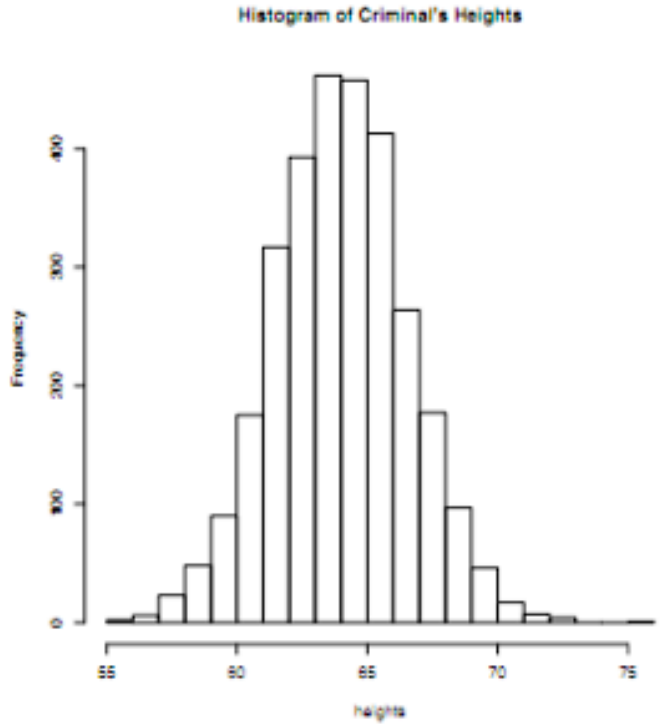
FIG. 9.—Enlarged impressions of the fore and middle finger tips of the right hand of Sir William Herschel, made in the year 1860.

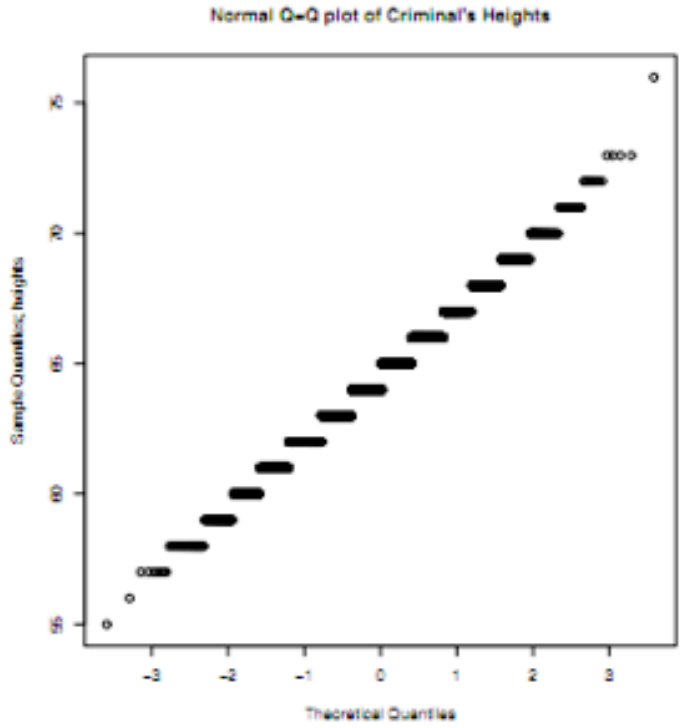
By Student

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 $\sigma = 2.6$

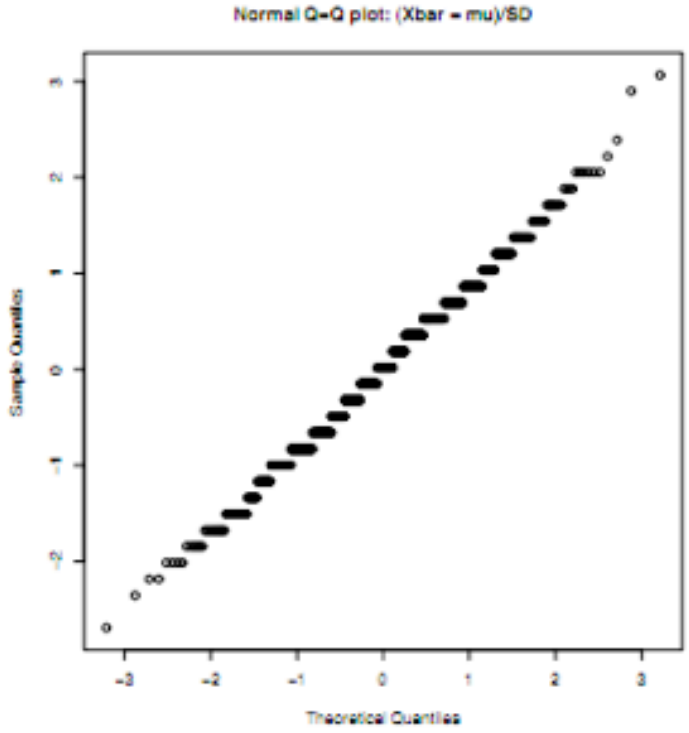
Gosset then took surveys of size *n*=4 and looked at

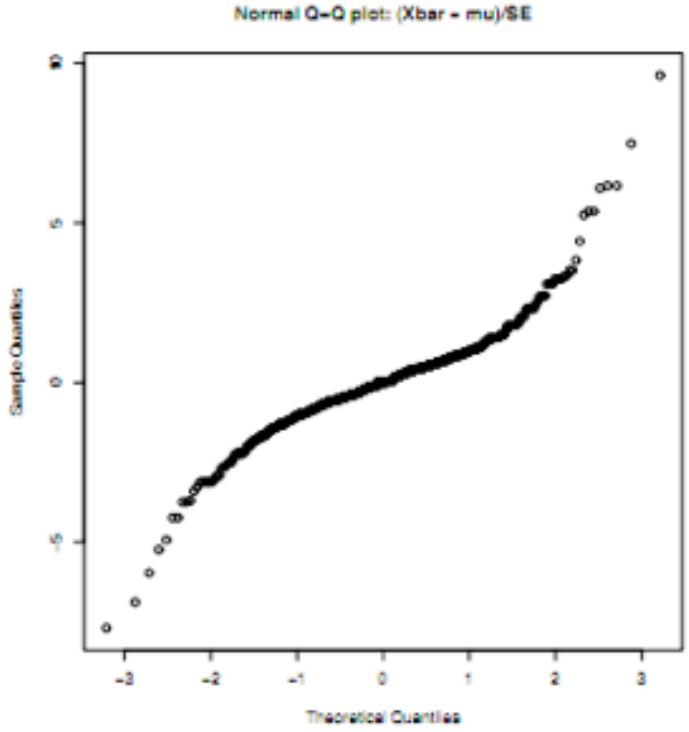
$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

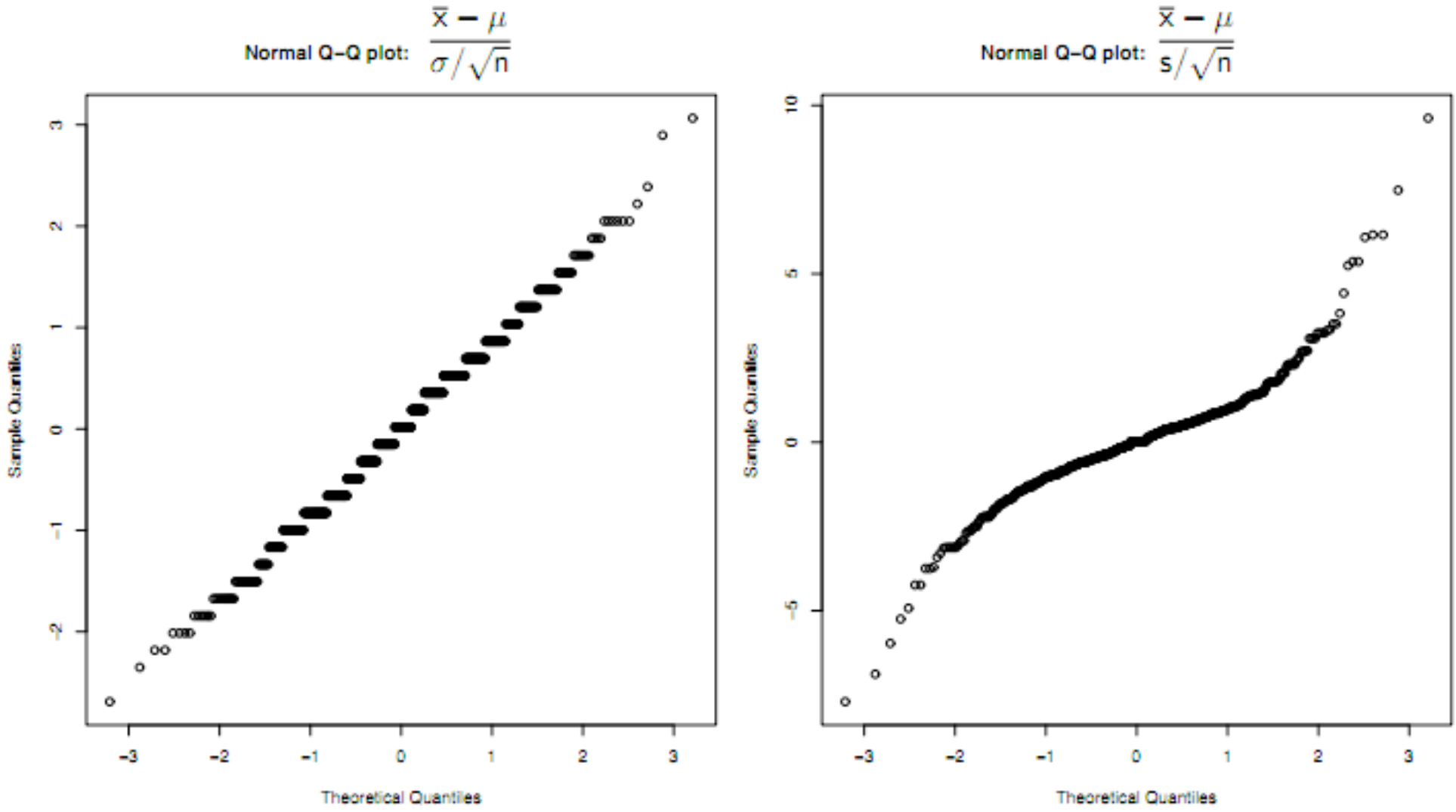
(top graph) and at

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

(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 have value somewhat 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

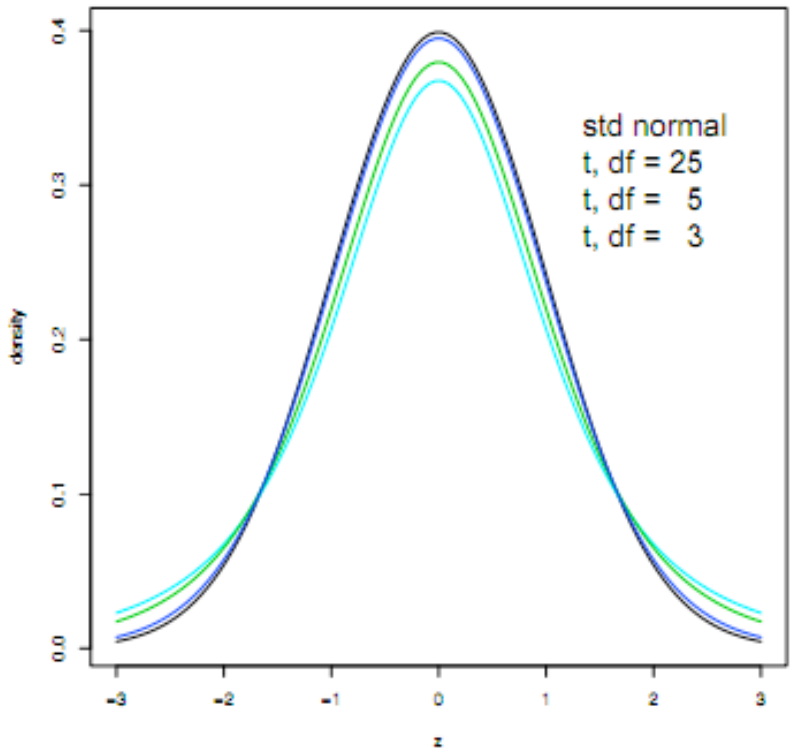
The *t*-distribution has one parameter controlling its shape; it is referred to as its *degrees of freedom*

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

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

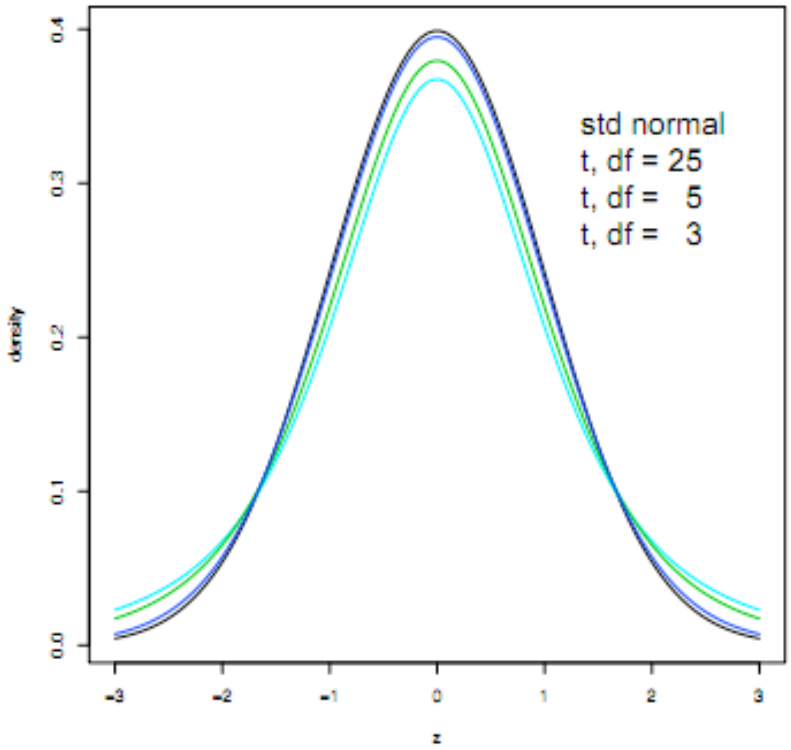
We said there were *n*-1 degrees of freedom in our estimate because 1 was used to compute $\bar{x}$



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



The *t*-distribution

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

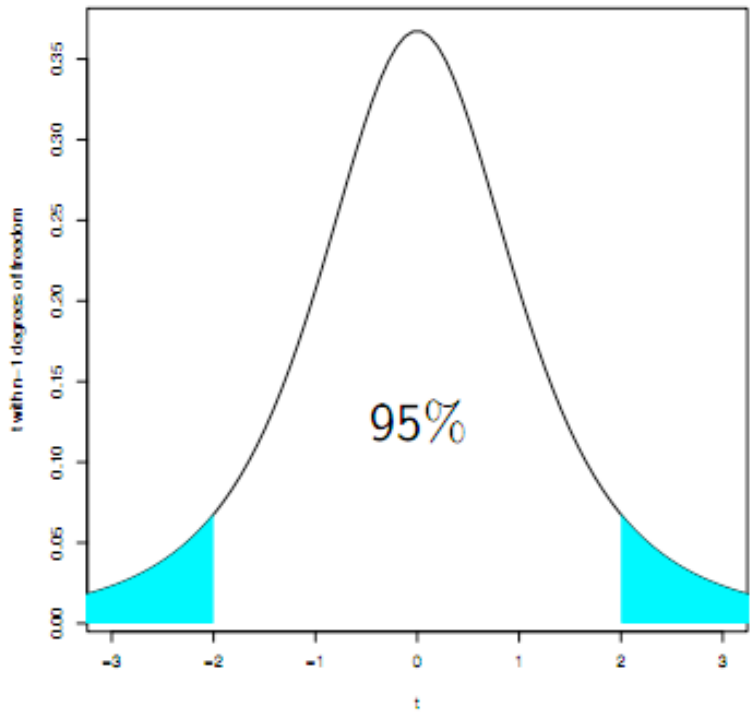
We would would need to find the point *t* such that

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

If there are *n* points in our survey, we take the degrees of freedom of the *T* to be *n*-1

We then form the confidence interval

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



Student's t-distribution

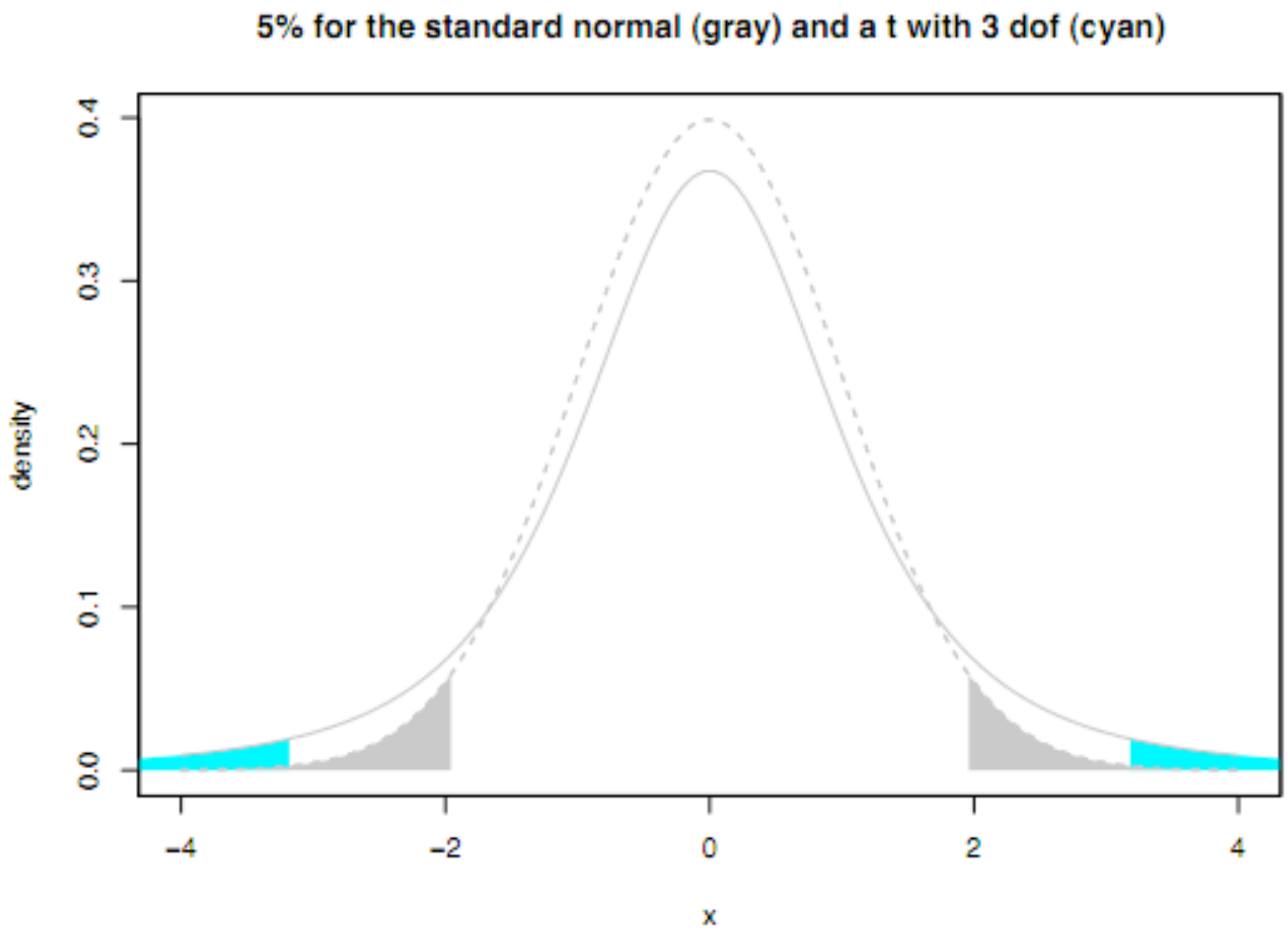
So, what does all this mean? If our data are normally distributed then the *t*-statistic

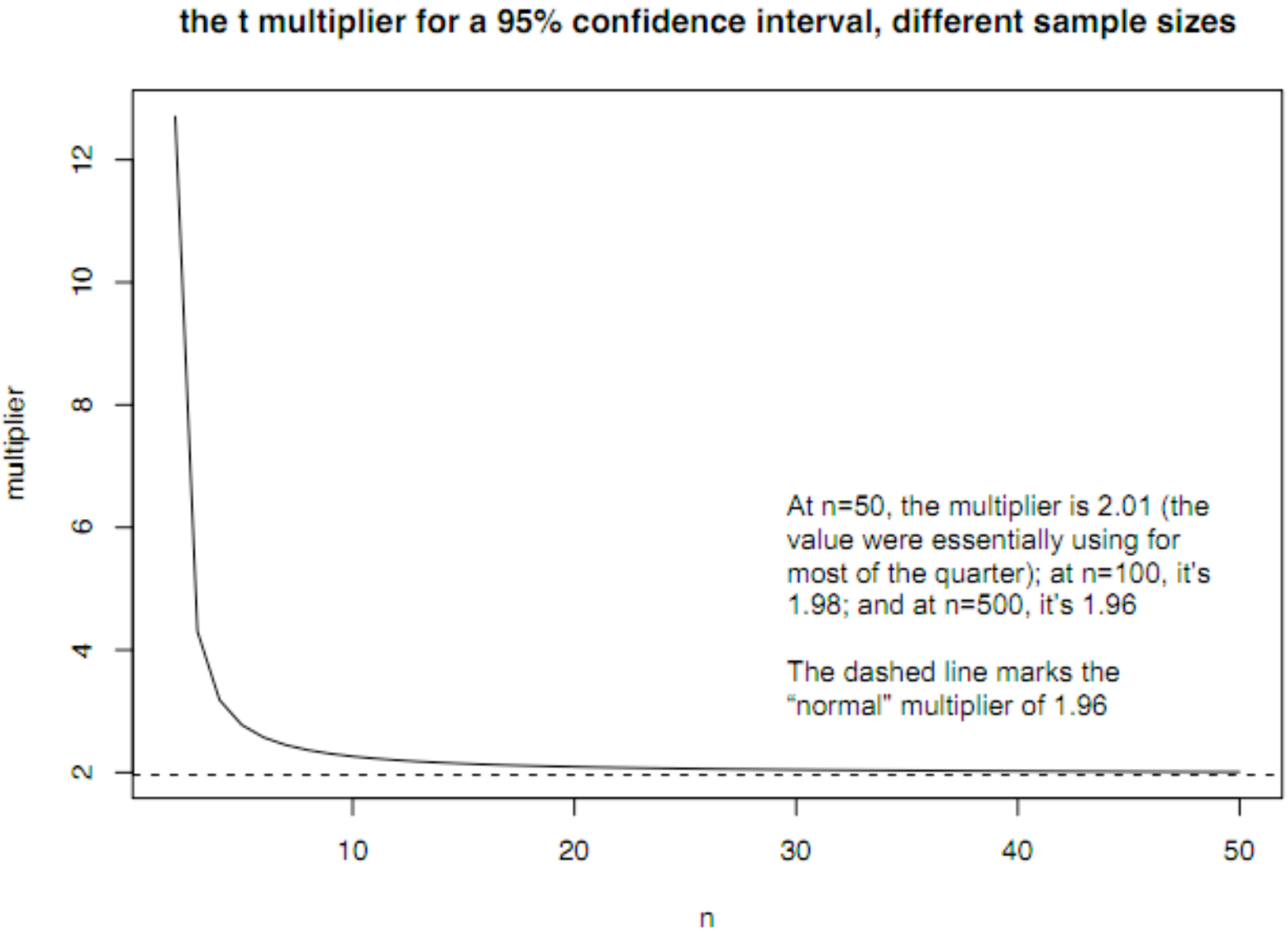
$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

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**

A remarkable fact about the t-statistic is that its distribution depends only on the sample size; this is quite a magical fact and one that we will make use of

Pivots

Notice that when our data $X_1, \ldots, X_n$ come from a normal distribution, 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, they can be “inverted” to compute confidence intervals

Pivots

Next, consider a confidence interval for σ^2 based on our estimate

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

Assuming our data $X_1, \ldots, X_n$ come from a normal distribution (μ, σ) we can show that the quantity $n\widehat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom and hence qualifies as a pivot

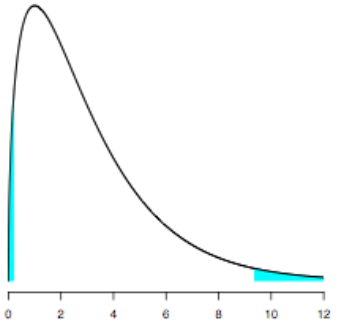
Note: The same calculations would work with s^2 (just defined) but with $(n-1)\, s^2/\sigma^2$ as the pivot

Pivots

Therefore, we can find values a and b such that

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

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



Combining these two expressions we find

$$\Pr(a \leq n\widehat{\sigma}^2/\sigma^2 \leq b) = 0.95$$

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

$$[n\widehat{\sigma}^2/b,\ n\widehat{\sigma}^2/a]$$

Note: This interval depends strongly on the normal distributional assumption -- I don't recommend it beyond it's pedagogical function

A comparison

Before you ask, it turns out that the bootstrap estimate of the standard error (the one you we computed by repeatedly sampling from the distribution) is essentially the same as $s/\sqrt{n}$ -- that is, $\text{sd}(\widehat{\theta}^*)$ (the standard deviation of our bootstrap replicates) we talked about in lecture and you will compute in lab is essentially $s/\sqrt{n}$

That means, **for large sample sizes** where we have normal sampling distributions, **every scheme you've seen so far for computing a confidence interval should agree** (that's comforting, right?)

Um, so why the bootstrap?

Rationale

For many (many many) statistics we are interested in, we don't have a formula for the standard error like we do for the sample mean; **with the bootstrap, no formula is needed** -- and in the few cases such a thing exists, it will agree with the classical formula

Instead of dealing in formulae, we rely on R or some other bootstrap enlightened software package to provide us with **a ready assessment of the precision of our estimate** -- it is fully general and quite powerful

Of course standard errors are only part of the game; we can also examine the bias of an estimate and we can generate confidence intervals that are free of the normal assumption (recall our percentile approach)