Lecture 9: Normal

Last time

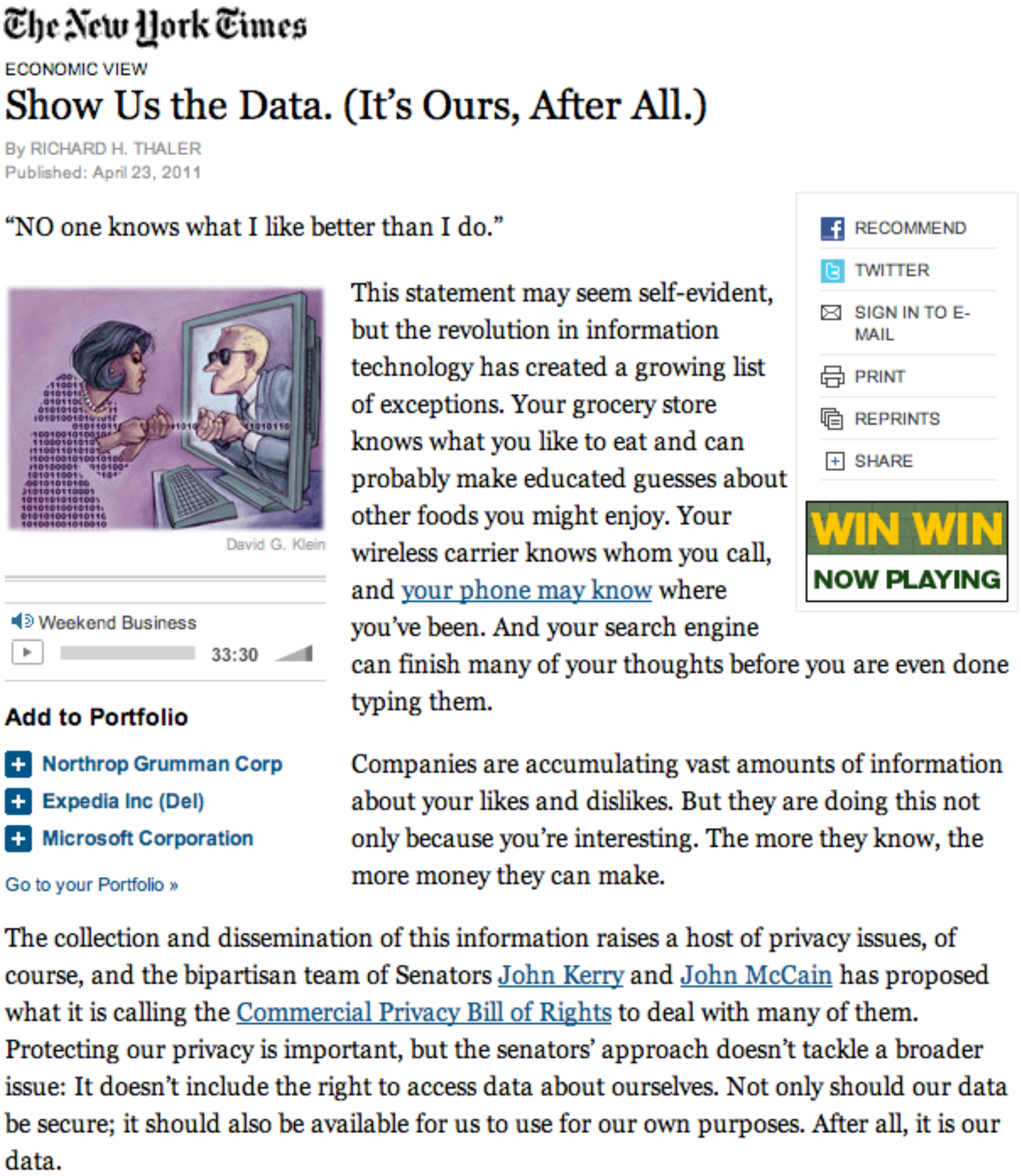
We spent some time talking about probability -- Every statistics class has to come to grips with that at some point, and last Wednesday was the day

We examined the basic axioms of probability, but by considering examining different interpretations of probability -- We saw that probability can be thought of in different ways, each with a different approach to inference

Today

We are going to finish up our probability discussion, deriving the binomial distribution -- It will give us a chance to see how the “classical” tools of probability can be applied to derive a null distribution

We’ll then spend a little time with a graphical device that can be used to decide whether or not data follow a particular theoretical model -- We’ll focus on the normal distribution



The New York Times

ECONOMIC VIEW

Show Us the Data. (It's Ours, After All.)

By RICHARD H. THALER

Published: April 23, 2011

"NO one knows what I like better than I do."

David G. Klein

Add to Portfolio

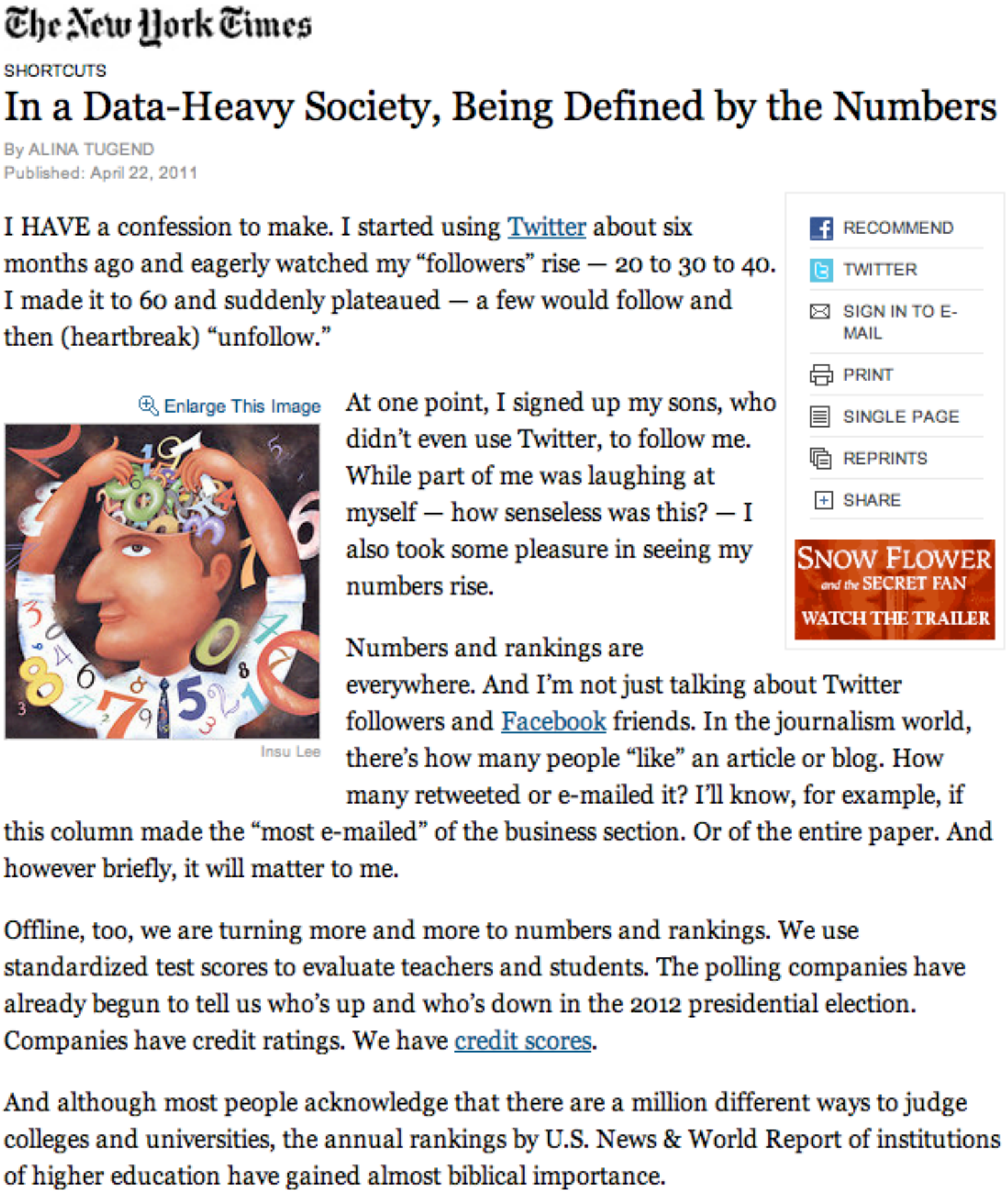
- Northrop Grumman Corp
- Expedia Inc (Del)
- Microsoft Corporation

Go to your Portfolio »

This statement may seem self-evident, but the revolution in information technology has created a growing list of exceptions. Your grocery store knows what you like to eat and can probably make educated guesses about other foods you might enjoy. Your wireless carrier knows whom you call, and your phone may know where you've been. And your search engine can finish many of your thoughts before you are even done typing them.

Companies are accumulating vast amounts of information about your likes and dislikes. But they are doing this not only because you're interesting. The more they know, the more money they can make.

The collection and dissemination of this information raises a host of privacy issues, of course, and the bipartisan team of Senators John Kerry and John McCain has proposed what it is calling the Commercial Privacy Bill of Rights to deal with many of them. Protecting our privacy is important, but the senators' approach doesn't tackle a broader issue: It doesn't include the right to access data about ourselves. Not only should our data be secure; it should also be available for us to use for our own purposes. After all, it is our data.



The New York Times

SHORTCUTS

In a Data-Heavy Society, Being Defined by the Numbers

By ALINA TUGEND

Published: April 22, 2011

I HAVE a confession to make. I started using Twitter about six months ago and eagerly watched my "followers" rise — 20 to 30 to 40. I made it to 60 and suddenly plateaued — a few would follow and then (heartbreak) "unfollow."

Enlarge This Image

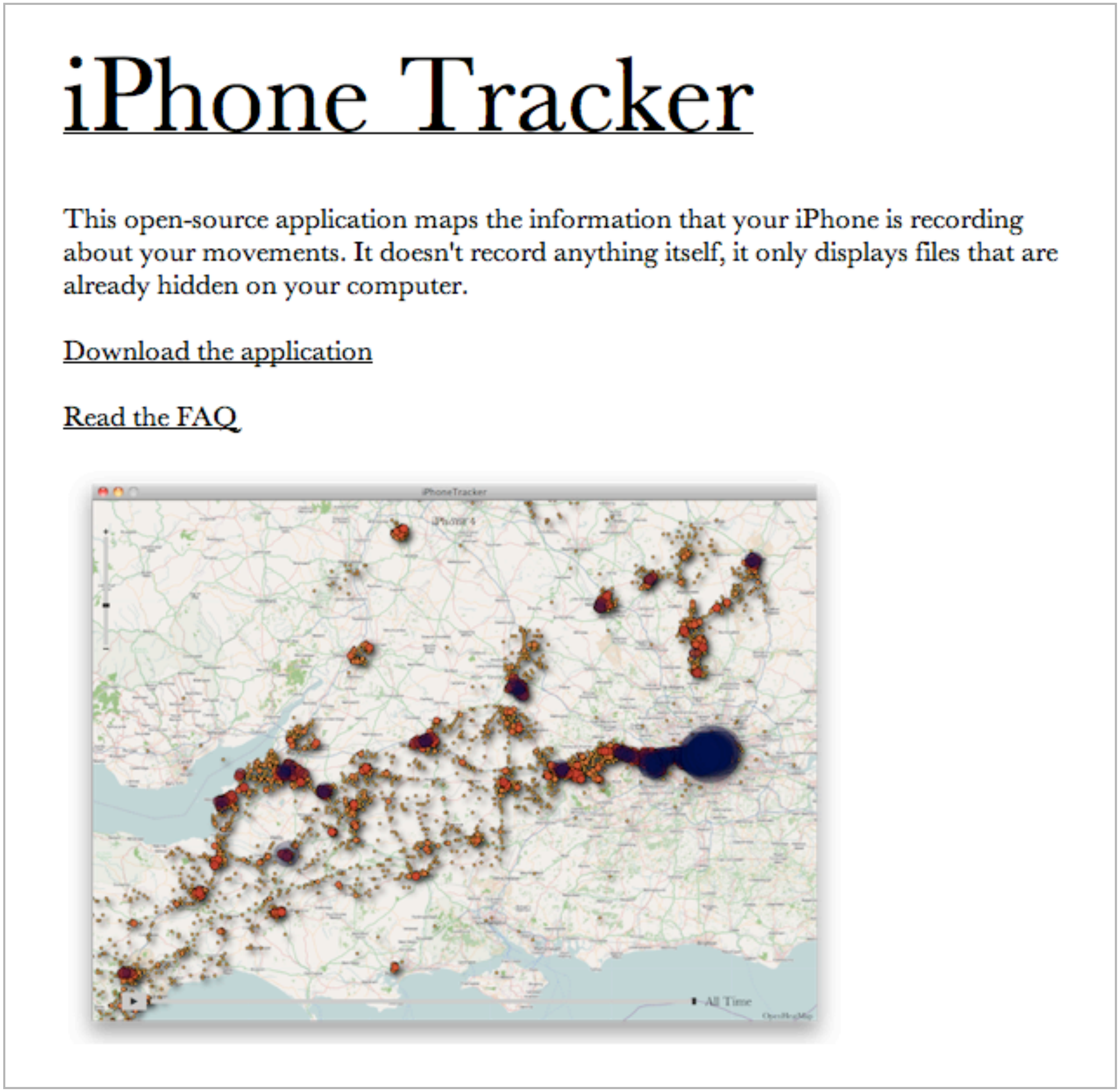
Insu Lee

At one point, I signed up my sons, who didn't even use Twitter, to follow me. While part of me was laughing at myself — how senseless was this? — I also took some pleasure in seeing my numbers rise.

Numbers and rankings are everywhere. And I'm not just talking about Twitter followers and Facebook friends. In the journalism world, there's how many people "like" an article or blog. How many retweeted or e-mailed it? I'll know, for example, if this column made the "most e-mailed" of the business section. Or of the entire paper. And however briefly, it will matter to me.

Offline, too, we are turning more and more to numbers and rankings. We use standardized test scores to evaluate teachers and students. The polling companies have already begun to tell us who's up and who's down in the 2012 presidential election. Companies have credit ratings. We have credit scores.

And although most people acknowledge that there are a million different ways to judge colleges and universities, the annual rankings by U.S. News & World Report of institutions of higher education have gained almost biblical importance.



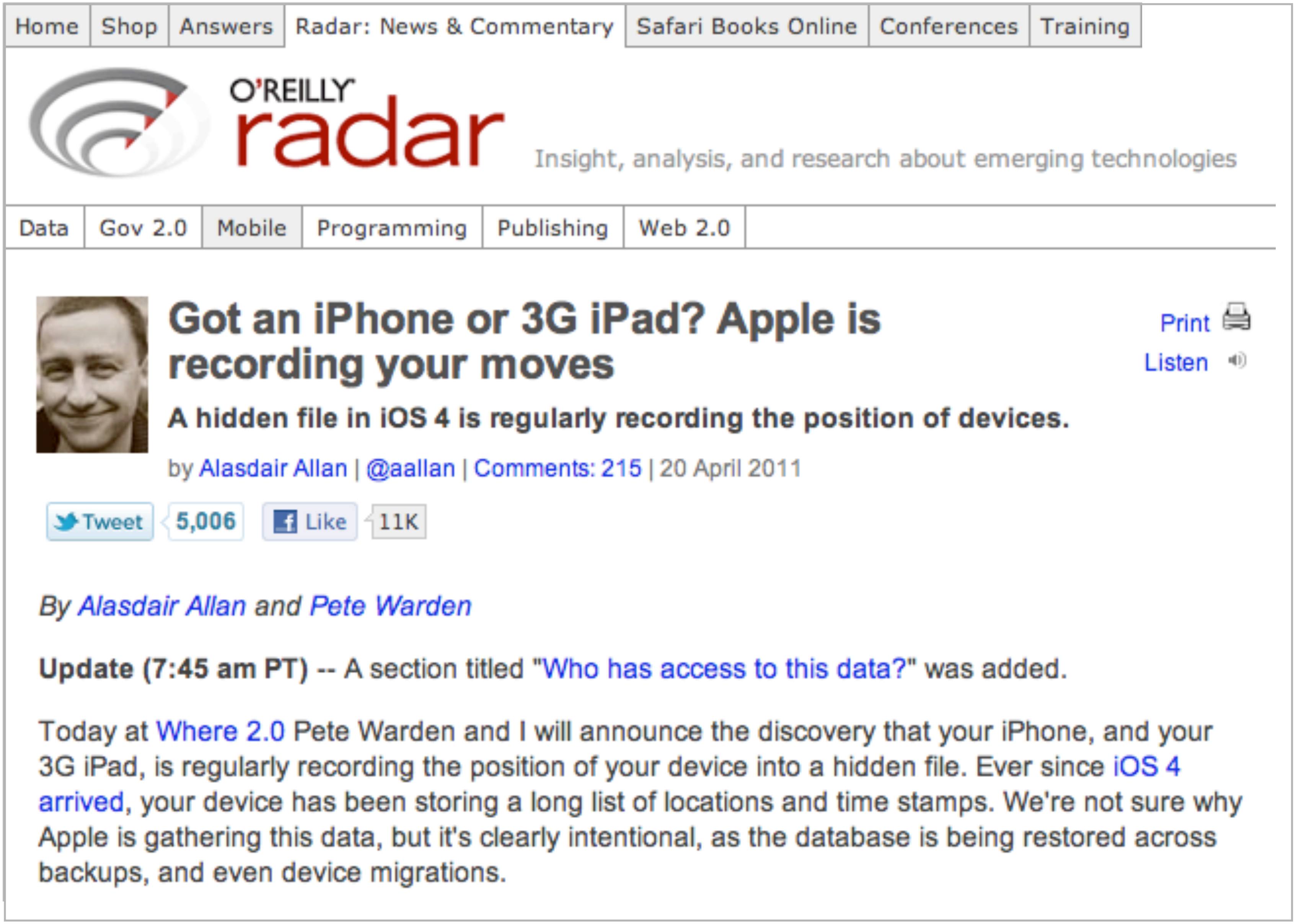
iPhone Tracker

This open-source application maps the information that your iPhone is recording about your movements. It doesn't record anything itself, it only displays files that are already hidden on your computer.

Download the application

Read the FAQ

http://petewarden.github.com/iPhoneTracker/



Got an iPhone or 3G iPad? Apple is recording your moves

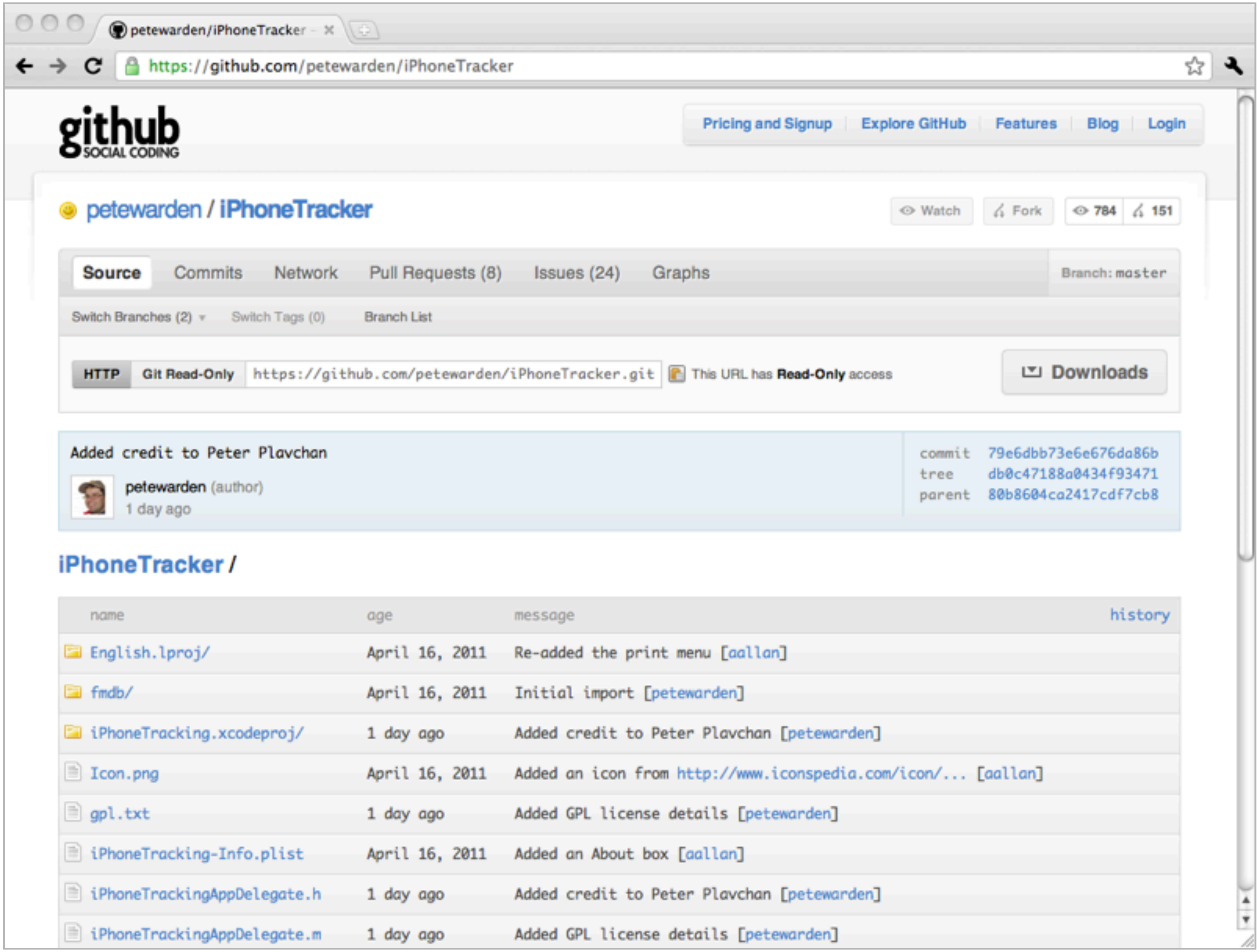
A hidden file in iOS 4 is regularly recording the position of devices.

by Alasdair Allan | @aallan | Comments: 215 | 20 April 2011

By Alasdair Allan and Pete Warden

Update (7:45 am PT) -- A section titled "Who has access to this data?" was added.

Today at Where 2.0 Pete Warden and I will announce the discovery that your iPhone, and your 3G iPad, is regularly recording the position of your device into a hidden file. Ever since iOS 4 arrived, your device has been storing a long list of locations and time stamps. We're not sure why Apple is gathering this data, but it's clearly intentional, as the database is being restored across backups, and even device migrations.



https://github.com/petewarden/iPhoneTracker.git This URL has **Read-Only** access

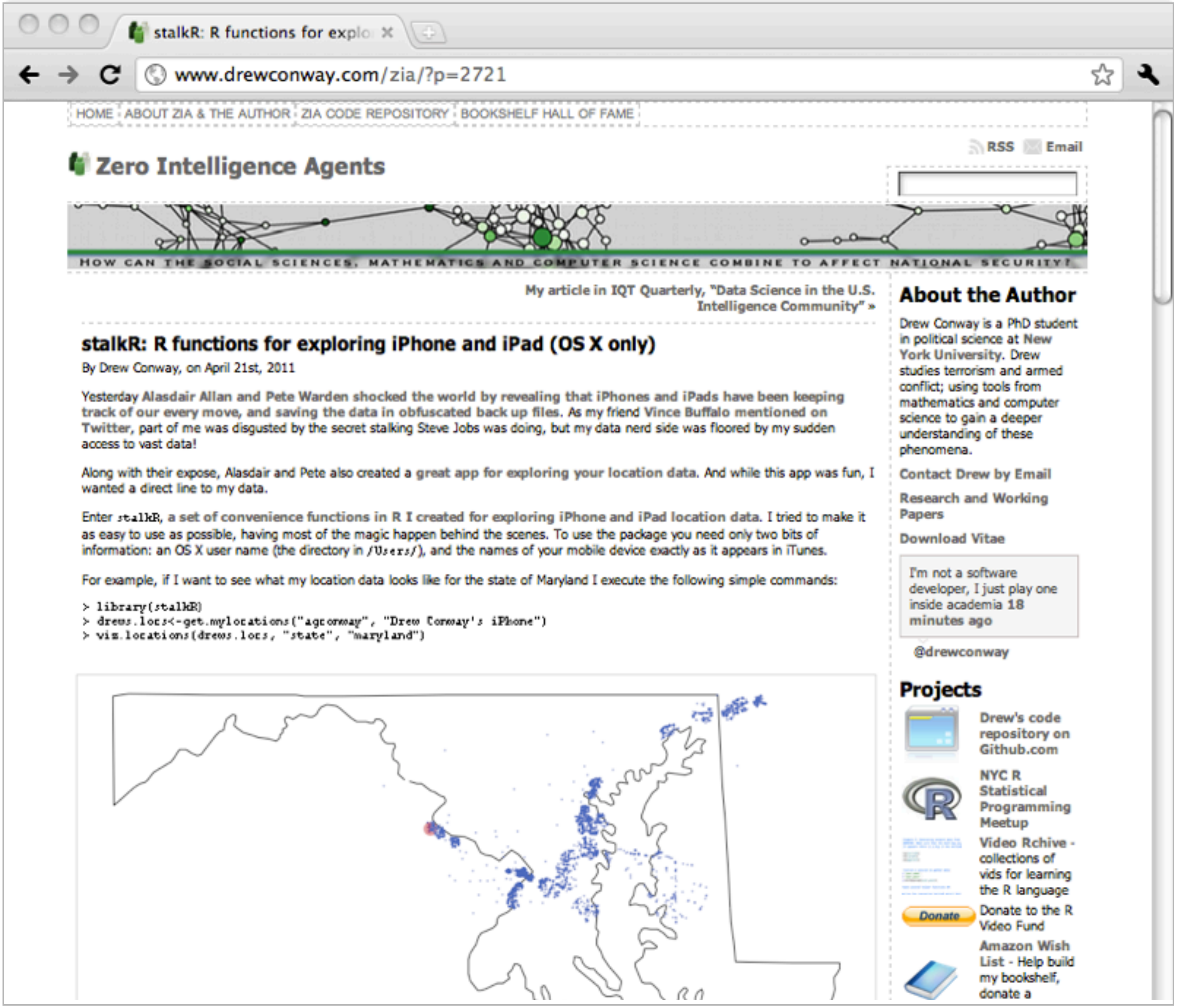
Added credit to Peter Plavchan

petewarden (author)
1 day ago

```
commit  79e6dbb73e6e676da86b
tree    db0c47188a0434f93471
parent  80b8604ca2417cdf7cb8
```

iPhoneTracker /

name	age	message
English.lproj/	April 16, 2011	Re-added the print menu [aallan]
fmdb/	April 16, 2011	Initial import [petewarden]
iPhoneTracking.xcodeproj/	1 day ago	Added credit to Peter Plavchan [petewarden]
Icon.png	April 16, 2011	Added an icon from http://www.iconspedia.com/icon/... [aallan]
gpl.txt	1 day ago	Added GPL license details [petewarden]
iPhoneTracking-Info.plist	April 16, 2011	Added an About box [aallan]
iPhoneTrackingAppDelegate.h	1 day ago	Added credit to Peter Plavchan [petewarden]
iPhoneTrackingAppDelegate.m	1 day ago	Added GPL license details [petewarden]


stalkR: R functions for explo
www.drewconway.com/zia/?p=2721
HOME
ABOUT ZIA & THE AUTHOR
ZIA CODE REPOSITORY
BOOKSHELF HALL OF FAME
RSS
Email
HOW CAN THE SOCIAL SCIENCES, MATHEMATICS AND COMPUTER SCIENCE COMBINE TO AFFECT NATIONAL SECURITY?
Donate

Zero Intelligence Agents

My article in IQT Quarterly, "Data Science in the U.S. Intelligence Community" »

stalkR: R functions for exploring iPhone and iPad (OS X only)

By Drew Conway, on April 21st, 2011

Yesterday Alasdair Allan and Pete Warden shocked the world by revealing that iPhones and iPads have been keeping track of our every move, and saving the data in obfuscated back up files. As my friend Vince Buffalo mentioned on Twitter, part of me was disgusted by the secret stalking Steve Jobs was doing, but my data nerd side was floored by my sudden access to vast data!

Along with their expose, Alasdair and Pete also created a great app for exploring your location data. And while this app was fun, I wanted a direct line to my data.

Enter stalkR, a set of convenience functions in R I created for exploring iPhone and iPad location data. I tried to make it as easy to use as possible, having most of the magic happen behind the scenes. To use the package you need only two bits of information: an OS X user name (the directory in /Users/), and the names of your mobile device exactly as it appears in iTunes.

For example, if I want to see what my location data looks like for the state of Maryland I execute the following simple commands:

```
> library(stalkR)
> drews.locs<-get.mylocations("agconway", "Drew Conway's iPhone")
> vis.locations(drews.locs, "state", "maryland")
```

About the Author

Drew Conway is a PhD student in political science at New York University. Drew studies terrorism and armed conflict; using tools from mathematics and computer science to gain a deeper understanding of these phenomena.

Contact Drew by Email

Research and Working Papers

Download Vitae

I'm not a software developer, I just play one inside academia 18 minutes ago

@drewconway

Projects

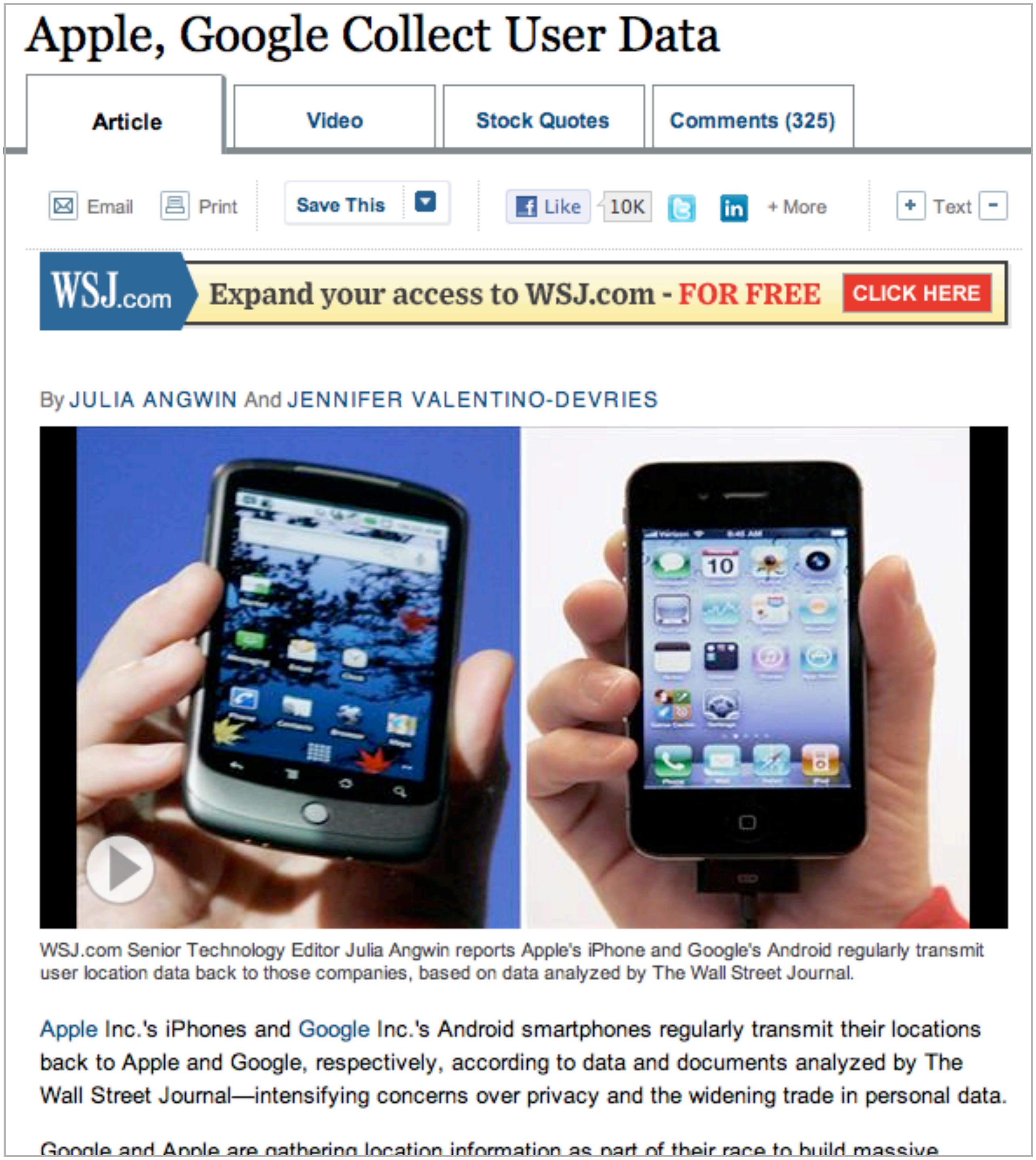
Drew's code repository on Github.com

NYC R Statistical Programming Meetup

Video Rchive - collections of vids for learning the R language

Donate to the R Video Fund

Amazon Wish List - Help build my bookshelf, donate a



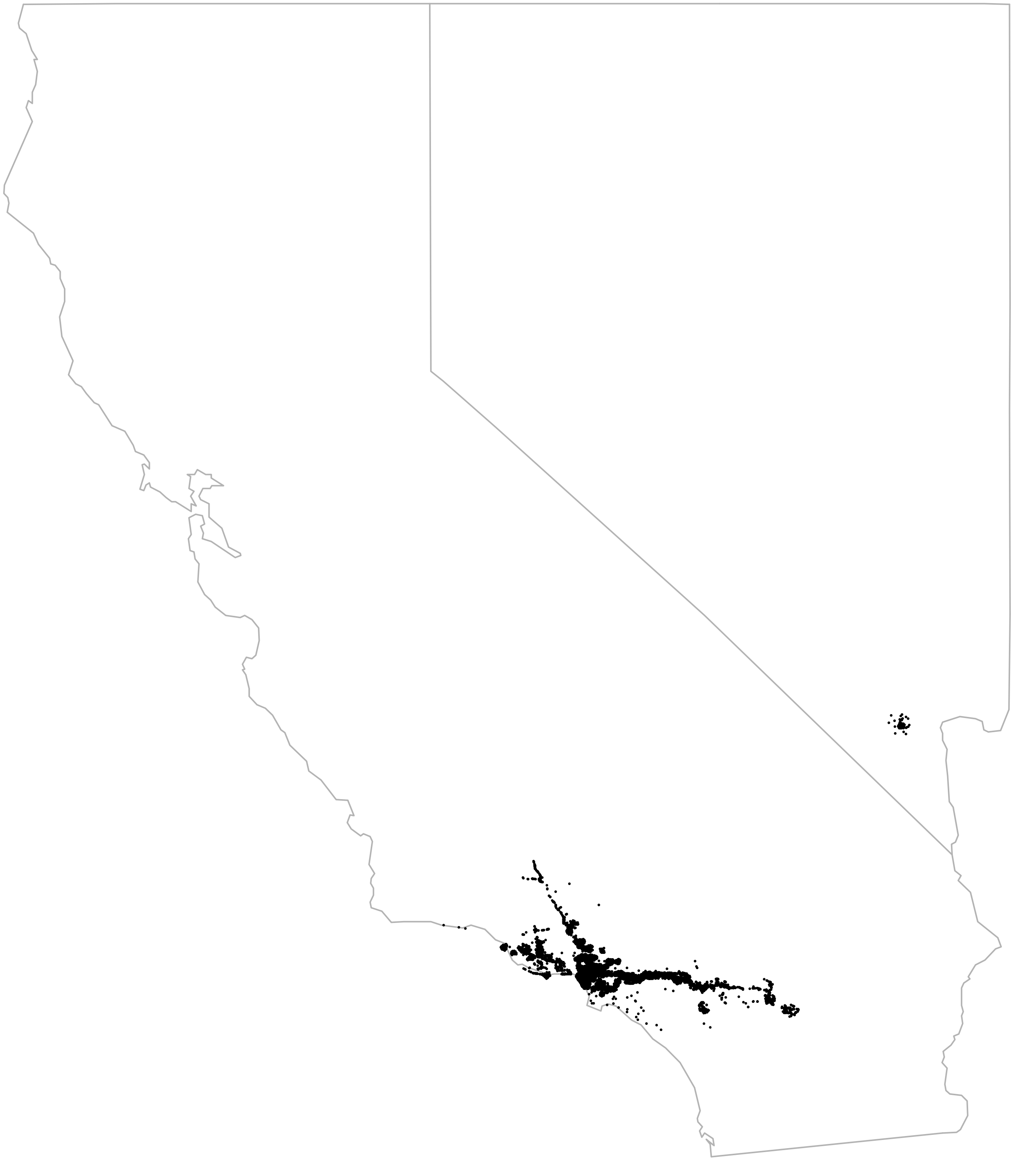
Apple, Google Collect User Data

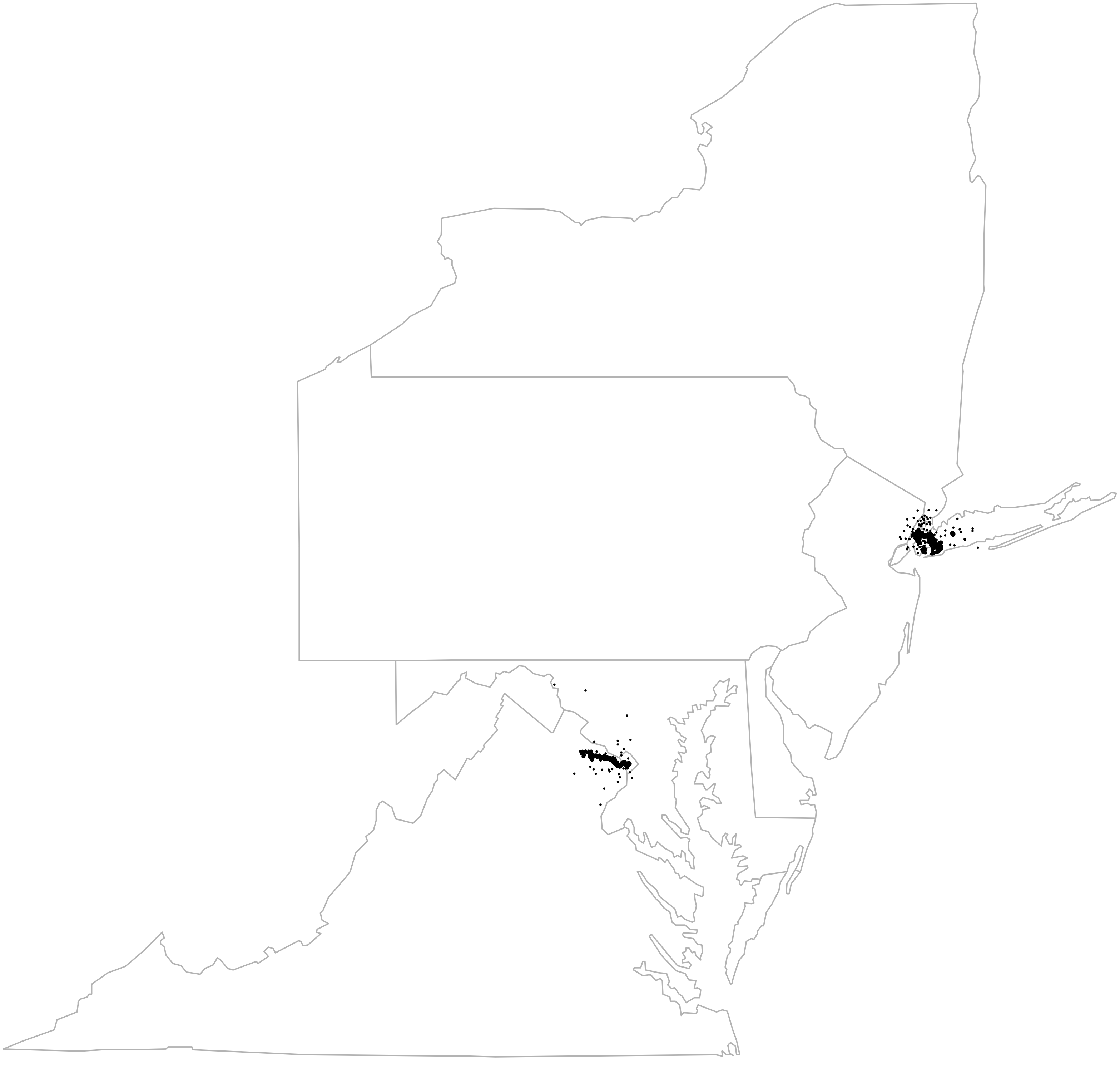
By JULIA ANGWIN And JENNIFER VALENTINO-DEVRIES

WSJ.com Senior Technology Editor Julia Angwin reports Apple's iPhone and Google's Android regularly transmit user location data back to those companies, based on data analyzed by The Wall Street Journal.

Apple Inc.'s iPhones and Google Inc.'s Android smartphones regularly transmit their locations back to Apple and Google, respectively, according to data and documents analyzed by The Wall Street Journal—intensifying concerns over privacy and the widening trade in personal data.

Google and Apple are gathering location information as part of their race to build massive





Probability

Last time we discussed some of **the basic axioms for working with probabilities** -- I am not assigning a lot of detailed homework to have you working out the chance that you see a particular kind of hand in poker or toss a certain sequence of rolls of a pair of dice (I am fairly confident you have seen this material elsewhere)

Instead, we want to focus on **how these ideas help us reckon with data** -- We saw last time how being able to count the number of sequences and sets you could from a collection of object helped us derive the exact form of the null distribution for our re-randomization tests (Hill's data, the Vioxx trials, etc.)

Let's use these ideas to go even farther back and see how we can come up with **the null distribution associated with Arbuthnot's test** -- But first, recall the axioms we motivated last time...

A more complete set of axioms

Below we list a more complete set of axioms for the calculus of probability; let $\mathcal{X}$ be the set of all possible outcomes of some experiment or trial or situation we'd like to study, let A denote an "event" or collection of outcomes from $\mathcal{X}$; and finally let P(A) be the probability of A

1. The probability of A is a number between 0 and 1, $0 \leq \mathsf{P}(\mathsf{A}) \leq 1$

2. The probability that an outcome will occur is 1, $\mathsf{P}(\mathcal{X}) = 1$

3. If A and B have no outcomes in common (they're disjoint), then their probabilities add $\mathsf{P}(\mathsf{A} \text{ or } \mathsf{B}) = \mathsf{P}(\mathsf{A}) + \mathsf{P}(\mathsf{B})$

4. Conditional probability:
$$\mathsf{P}(\mathsf{A}|\mathsf{B}) = \mathsf{P}(\mathsf{A} \text{ and } \mathsf{B})/\mathsf{P}(\mathsf{B}) \quad \text{or} \quad \mathsf{P}(\mathsf{A} \text{ and } \mathsf{B}) = \mathsf{P}(\mathsf{A}|\mathsf{B})\mathsf{P}(\mathsf{B})$$

5. The law of total probability: $\mathsf{P}(\mathsf{A}) = \mathsf{P}(\mathsf{A}|\mathsf{B}_1)\mathsf{P}(\mathsf{B}_1) + \cdots + \mathsf{P}(\mathsf{A}|\mathsf{B}_\mathsf{J})\mathsf{P}(\mathsf{B}_\mathsf{J})$ where $\mathsf{B}_1, \mathsf{B}_2, \ldots, \mathsf{B}_\mathsf{J}$ are all disjoint and their union is $\mathcal{X}$

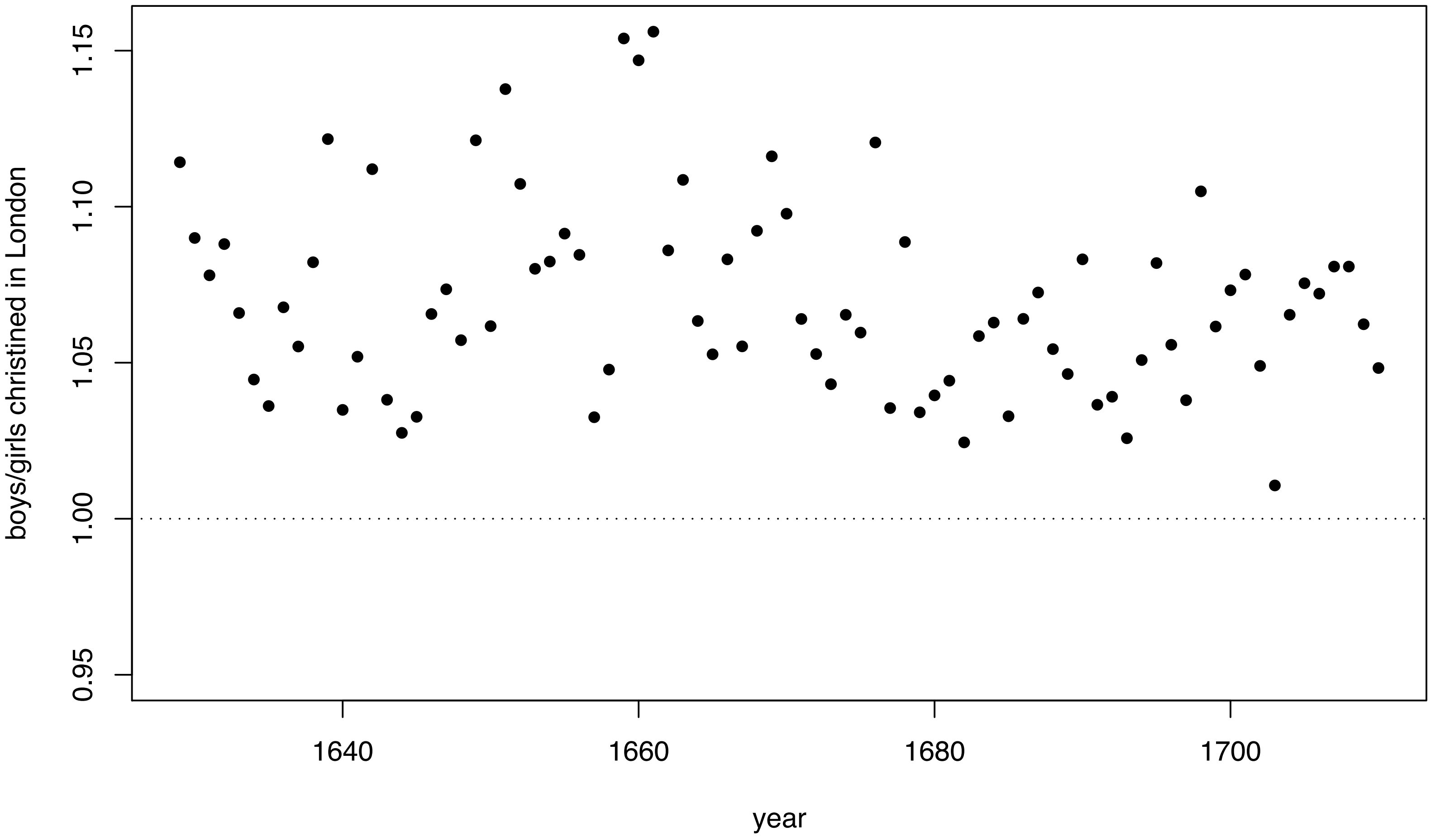
Arbuthnot

Recall that Arbuthnot was interested in **the apparent difference in the ratios of births of boy babies to girl babies** -- Specifically, he entertained **the null hypothesis that boys and girls were born with the same frequency each year**

His alternative was that there were more boys born than girls -- His test statistic was the count of years (between 1629 and 1710) for which boy births outnumbered girl births

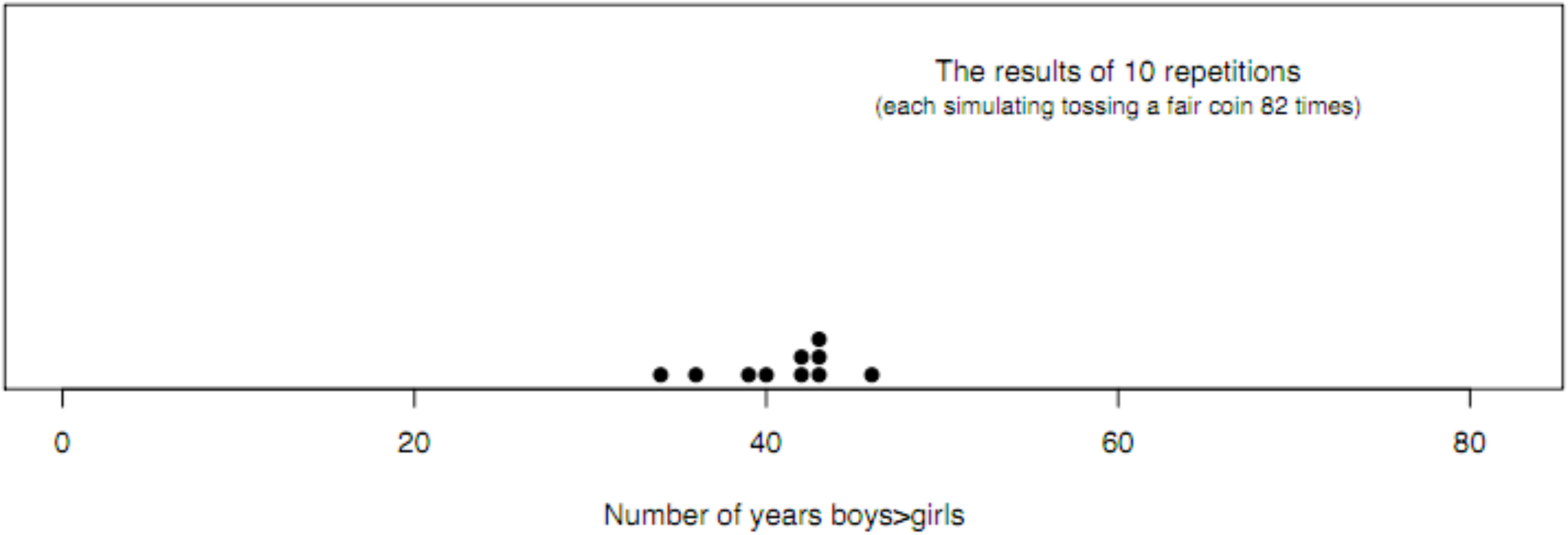
He reckoned that under the null hypothesis of equal frequencies, the outcome for each year was **the result of an independent toss of a fair coin** -- As a result, his test statistic could be thought of as **counting the number of heads we see in 82 independent tosses of a fair coin**

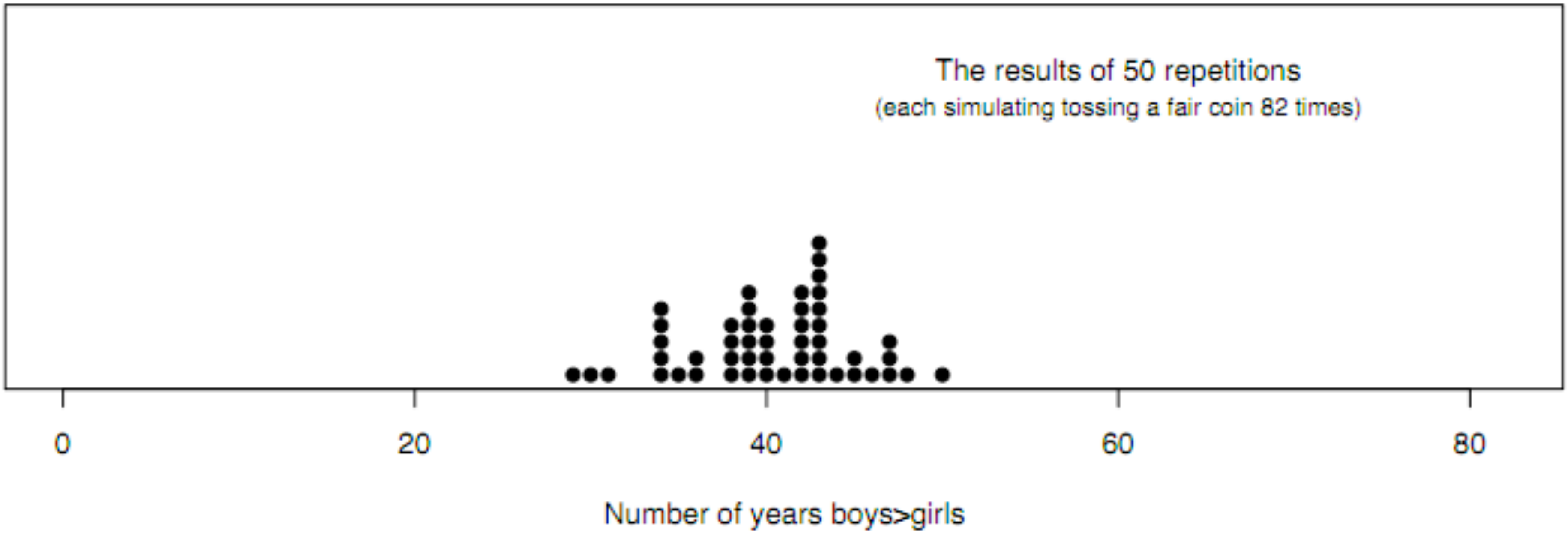
We initially relied on computer simulation to come up with the distribution but with our new found probability skills, we can work this out directly!

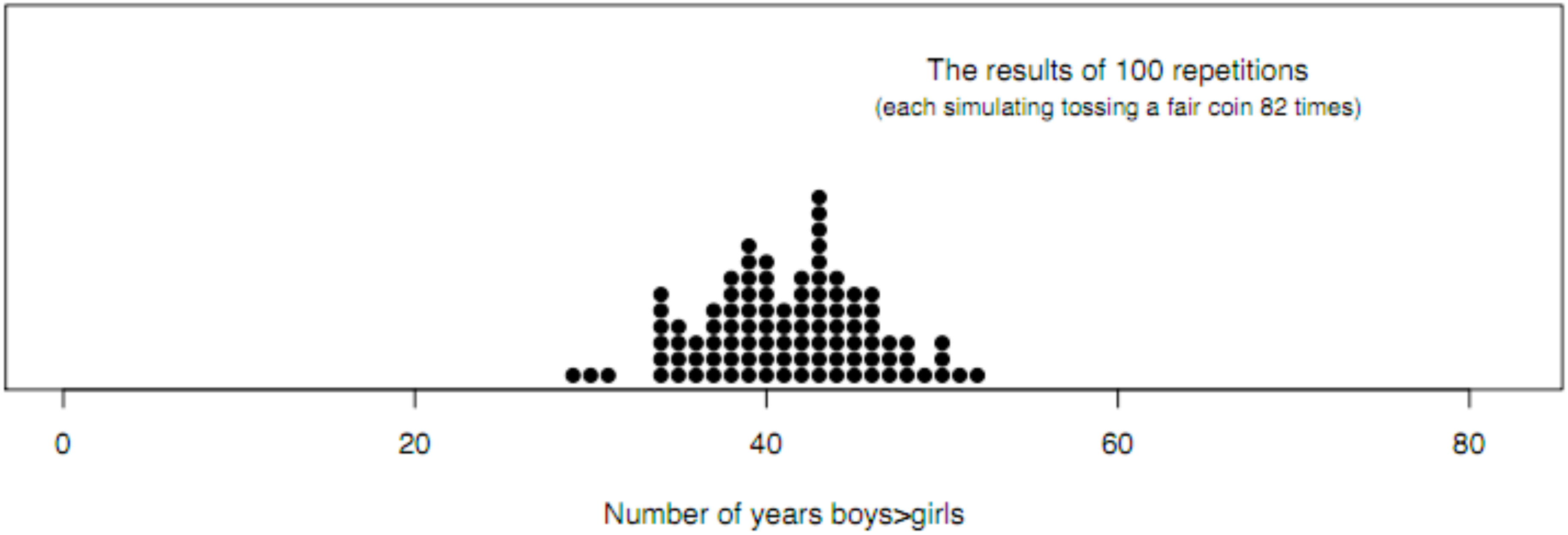


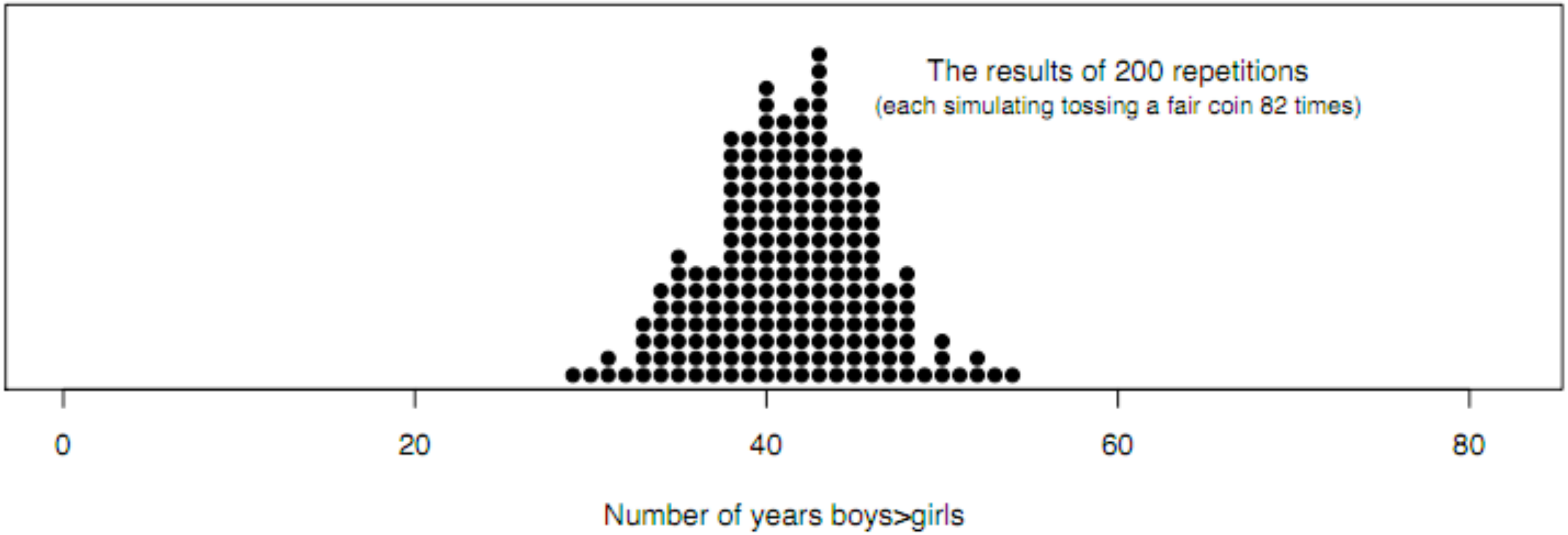
Proportion of boys to girls christined in London

(dotted line represents equal proportion between sexes)


The results of 10 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


The results of 50 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


The results of 100 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


The results of 200 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

Scratching things down

(Those images seem a bit crude now!) Let’s start by considers a particular outcome, **a sequence of 82 independent coin tosses** (re-labeled B and G for years in which boy births outnumbered girl births) where the first B is for 1629, the second is 1630 and so on

B B B B B G B B G B B B B B B B B G G B B G G B G B B B B G B B B G
G B B G B B B B B G B B B B B B G B B B G B B B G B B B B G B B B B B
B G B G B B B G G G B G B

There are 60 B’s in this sequence and 22 G’s, and since each label is the result of a fair con toss, each carries the probability 0.5 -- So the chance that we see B in each of years 1629 through 1633 is 0.5 and the chance that we see a G in 1634 is also 0.5

Because **the tosses for each year are independent** (meaning what happens in one year is assumed to have no effect on the results from other years), we compute the probability of seeing the entire sequence of B’s and G’s above by **multiplying the individual probabilities**

That means the probability for this particular sequence is $0.5 \cdot 0.5 \cdots 0.5 = 0.5^{82}$

Scratching things down

If our null hypothesis had assigned a different number, say p, to the chance that one year would see more boys than girls born, then the chance for the sequence we wrote down would instead be

$$\mathrm{p}^{60}(1-\mathrm{p})^{22}$$

because each of the 60 B’s carries a probability p and each of the 82-60=22 G’s has a probability 1-p

Now, the string of G’s and B’s on the previous page is just one outcome -- In fact, we can (using results from last lecture) work out the chance associated with any string of years with 60 B’s and 22 G’s

Scratching things down

Directly applying our **choose notation** from last lecture, we can write down the **number of different ways we can select 60 from the 82 years**

$$\left(\begin{array}{c} 82 \\ 60 \end{array} \right) = \frac{82!}{60!\,22!} = \frac{82 \cdot 81 \cdots 24 \cdot 23}{60 \cdot 59 \cdots 2 \cdot 1} = 50,825,908,693,881,536,512$$

That is, out of our 82 years, we can choose a set of 60 in a zillion different ways (a zillion being a technical term for truckloads)

Each of these carries the same probability because (whether p=0.5 or not), **the probability of any sequence just depends on the count of B's and G's**

Scratching things down

Each of these zillions of different strings represent **a different outcome** of our coin tossing (baby birthing) experiment or, in technical terms, they are disjoint outcomes -- That means to find the probability that we see 60 B's and 22 G's, **we just add the probabilities across all these occurrences**

In slightly more abstract notation, the probability of seeing 60 B's is

$$\left(\begin{array}{c}82\\60\end{array}\right) \mathsf{p}^{60}(1-\mathsf{p})^{22}$$

Scratching things down

Of course nothing is special about the number 60 and we can work out the probability of seeing k years out of 82 in which boys outnumber girls born

$$\left(\begin{array}{c} 82 \\ k \end{array} \right) p^k (1-p)^{82-k}$$

Scratching things down

Now, we can make this fully general by assuming we have **n independent trials** each with **success probability p** -- The chance that we see **exactly k successes** is then

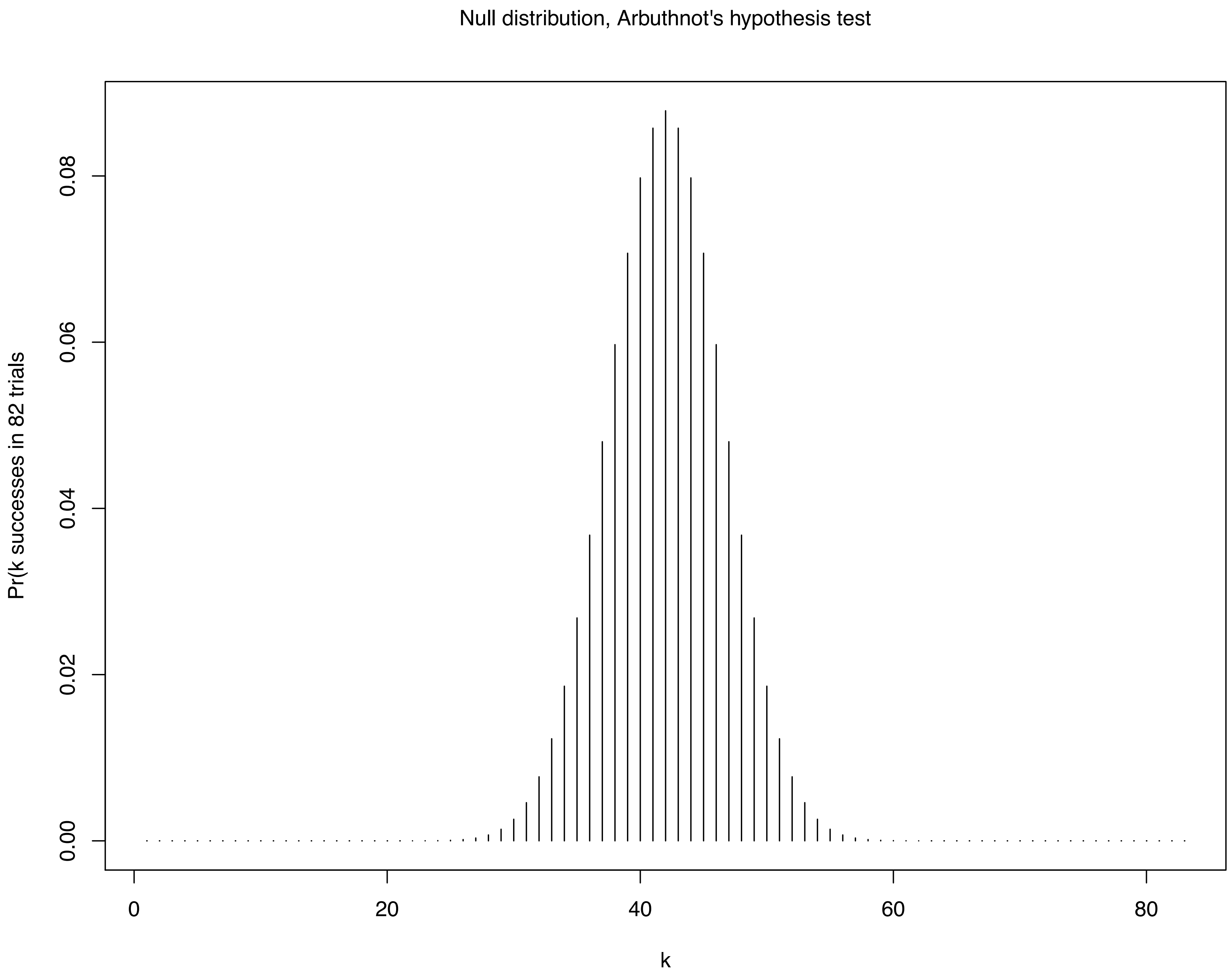
$$\left(\begin{array}{c} n \\ k \end{array} \right) p^n (1-p)^{n-k}$$

This is known as the binomial distribution -- We introduce it because it is a relatively simple model but comes up in a lot of circumstances (It also lets us kick the tires on some of the probability axioms we learned in the last lecture)

Scratching things down

We now seem very far away from Arbuthnot, but again, under his null hypothesis that boys and girls are born with equal frequency, his Bills of Mortality data represen**t 82 independent trials**, each with **probability of success 0.5**

We can use the formula on the previous page to make a barplot of the probability of seeing k=0,1,...,82 successes

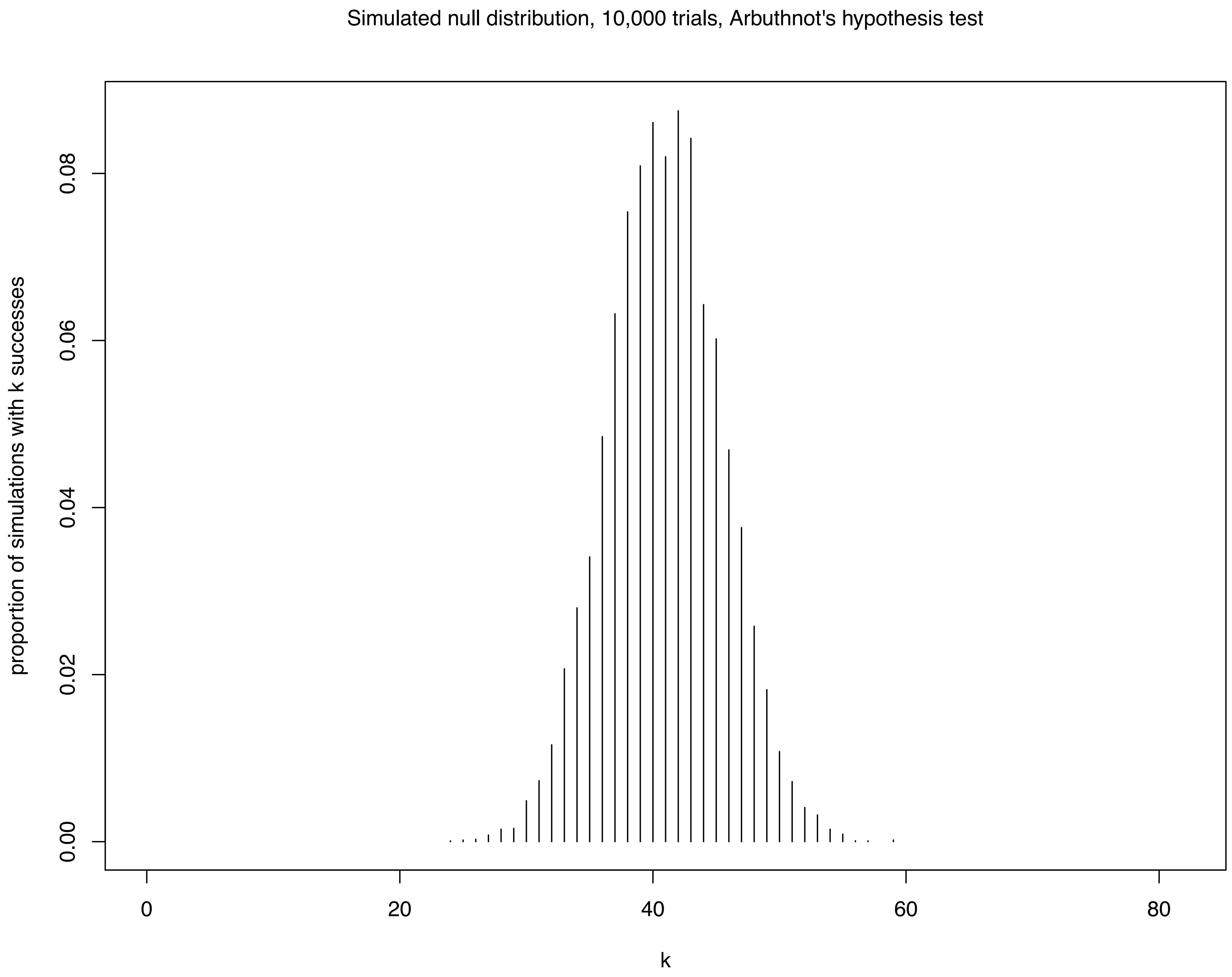

Null distribution, Arbuthnot's hypothesis test
Pr(k successes in 82 trials
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
k

Simulation

On the next page we have a similar plot associated with **10,000 simulations** of the same experiment (82 independent coin tosses) and we see they're reasonably close

In general, we will rely on the fact that we can **simulate a complex process** and examine **the proportion of trials that match the criterion we're after** -- These proportions will get closer and closer to the probability we're after (P-values, say) the longer we simulate

This limit theorem and our use of simulation is really the basis for much of statistical inference -- I don't want to burden this example with too many ideas but we will see this kind of reasoning again appear as we flush out the essential concepts behind estimation


Simulated null distribution, 10,000 trials, Arbuthnot's hypothesis test
proportion of simulations with k successes
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
k

To sum

We have presented some tools that help us **compute mathematically probabilities associated with simple chance mechanisms** -- We have also provided you with at least **a basic explanation about why our computer simulations** ought to give us similar answers

Finally, we saw that while these computations are fairly straightforward, probability as a concept is fairly diverse -- **We have seen a number of different interpretations**, each of which yield different techniques for reasoning about or with uncertain data

Now, before we leave the subject, **we're going to need to be able to reckon a bit with the so-called normal distribution** -- It is a particular continuous probability distribution that we've been more or less staring at all term...

A remarkable fact

Most of the null distributions we have looked at this term (the results of our re-randomizations) **have a common look to them** -- This is an amazingly robust fact and is, in some sense, the cornerstone of a great deal of statistical practice

The distribution of our test statistic under the null will, in many cases, have a **bell-shaped distribution** -- The explanation for this has to do with something called the **Central Limit Theorem**

Rather than complicate matters with a lot of math, we’ll simply comment that before statisticians could routinely compute the re-randomizations we are taking for granted, they were talented enough to anticipate the likely shape of the null distribution under a number of special cases

The normal distribution

The word “normal”, then, is not used in its common meaning of “ordinary or common” or its medical meaning of “not diseased” but instead the usage relates to the older meaning of **“conforming to a rule or pattern”**

As the target of the so-called Central Limit Theorem, it is the **“rule or pattern” to which many null distributions tend** -- The normal distribution is also used as a “model” for data and as such suggests very specific ways in which mass is distributed by the probability function

Before we can dig in and describe the distribution in detail, we’ll revisit mean and standard deviation

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY • IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE, AGRICULTURE, AND ENGINEERING •
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

William Youden (chemist, statistician, amateur typographer)

“In science, multiple discoveries have been found to be the rule (Merton, 1973), but multiple independent appearances of the same terminology for the same scientific object must surely be the exception. Yet, this is exactly what happened with the appearance of the word “normal” as a descriptive of the probability curve

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

by Charles S. Pierce (1873), by Francis Galton (1877) and by Wilhelm Lexis (1877). Such multiplicity of naming - in three countries and two languages - is remarkable, and surely signals a widespread simultaneously evolving conceptual understanding in the 1870’s: of populations of people, of measurements, and of their similarities.”

Normative Terminology by W.H. Kruskal

Sample mean and standard deviation

Given a set of n data points $x_1, \ldots, x_n$ we've previously defined the sample mean and standard deviation to be

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

Let's look into this a little more closely -- First note that the sum of the deviations around the mean is zero

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = (x_1 + \cdots + x_n) - n\bar{x} = 0$$

As a result, if we knew n-1 of the deviations from the mean, we could figure out what the last one is knowing that all n sum to 0 -- This means our sample standard deviation involves just n-1 independent pieces of information and not n and hence the n-1 in its definition

Sample mean and standard deviation

There is another (maybe less useful explanation) for those of you who remember your calculus -- While it is not immediately obvious, the mean $\bar{x}$ can be defined via **a minimization problem**

That is, consider the sum of squares calculated around any point b

$$(x_1 - b)^2 + (x_2 - b)^2 + \cdots + (x_n - b)^2$$

The value of b that that minimizes (makes as small as possible) this expression is $b = \bar{x}$ -- In statistical parlance we have used up a “degree of freedom” in determining the minimizer $\bar{x}$ and hence the n-1 in the definition of s

We will see a similar sum of squares formulation when we get to regression!

The normal distribution

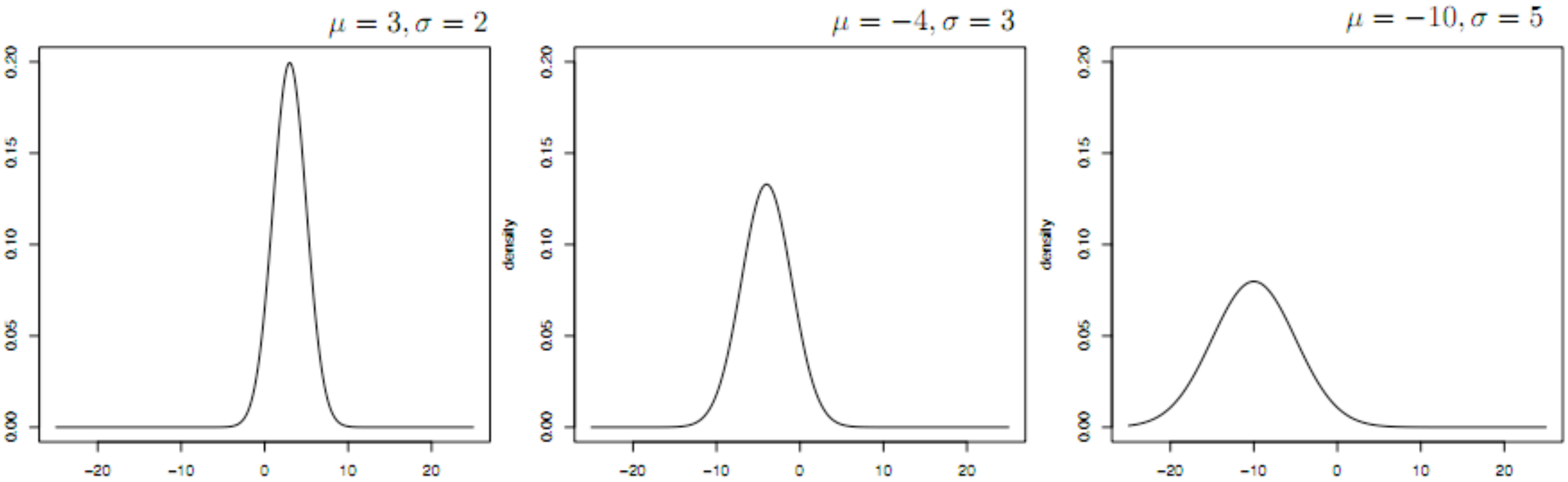
For this point in the class, we are going to use the normal distribution as **a kind of ruler against which we'll measure data** -- The normal is really **a family of shapes indexed by two quantities also called the mean μ and standard deviation σ**

Like their sample-based counter parts, these control **the center and spread of the bell curve**...

The normal distribution

Below we have a mathematical expression that demonstrates how the center of the bell curve and its spread change for different values of μ and σ

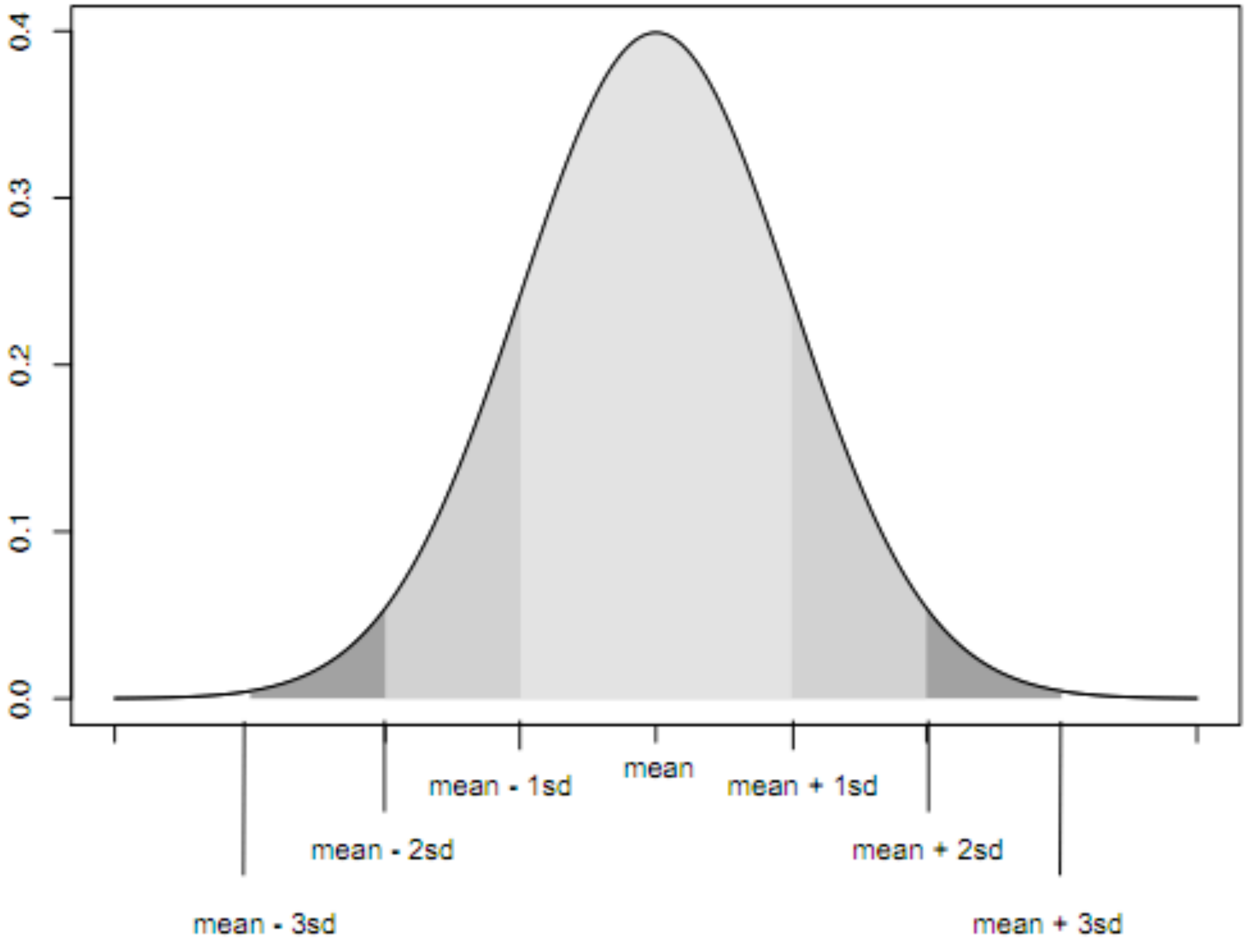
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}[\frac{x-\mu}{\sigma}]^2}$$



The normal distribution

There are rules of thumb that relate to how the normal distribution allocates its points -- These relate to a scale defined by the number of standard deviations from the mean

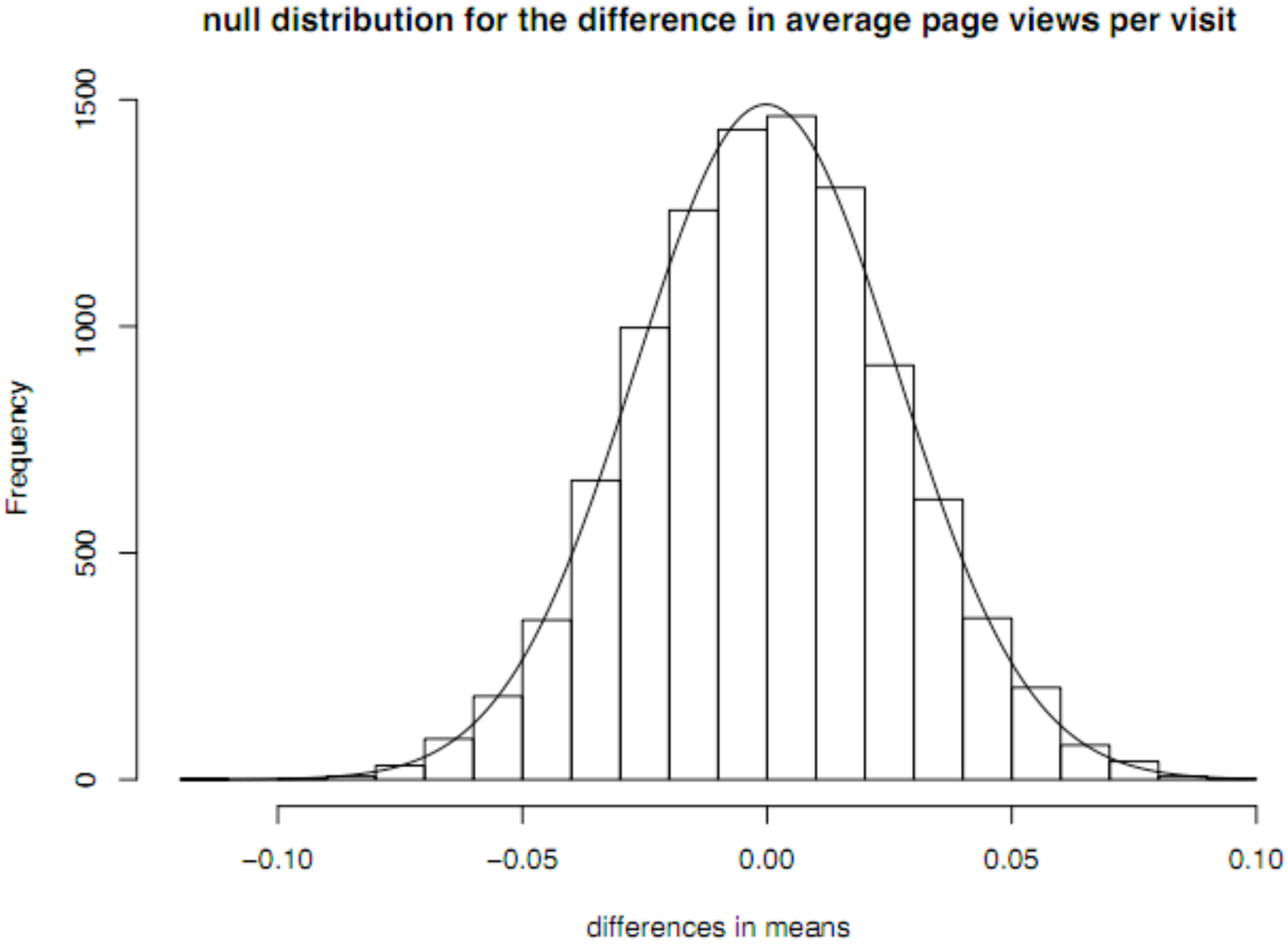
Specifically, if the distribution is normal, about 68% of the data should be within one standard deviation of its mean, about 95% should be within two standard deviations and about 99% should be within 3 -- These hold no matter what values of μ and σ are chosen


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

Measuring up

When given data, we can assess its normality with some graphical tools -- For example, we might simply overlay a bell curve on top of our data where we take the values of μ and σ to match their sample-based counterparts $\overline{\mathrm{x}}$ and s

Here is the normal curve added to the null distribution of the difference in mean numbers of page visits between the Tabs and Lists designs for the NYT Travel Section

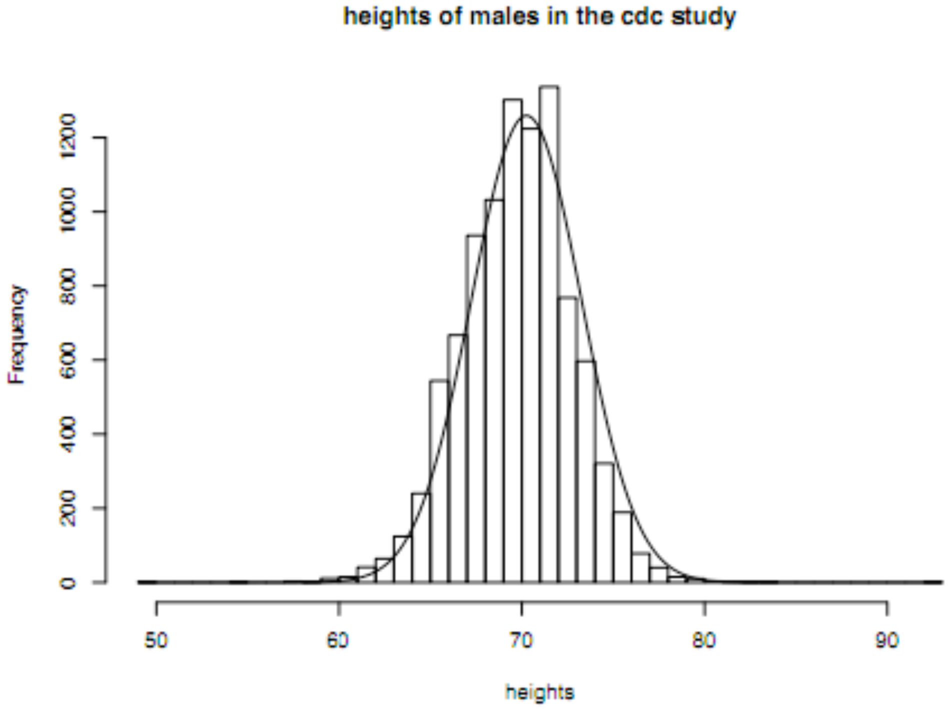

null distribution for the difference in average page views per visit
Frequency
0
500
1000
1500
−0.10
−0.05
0.00
0.05
0.10
differences in means

Assessing normality

The example on the previous page seems to "work" visually because the curve tracks the data pretty well -- At the right we present the heights of males in the CDC study

Here we have some misfit, but we're forced to reason about the graphic in a somewhat imprecise way

Ultimately, what we want is an image that let's us immediately judge where the sample data seem to mass up according to the proposed model and where they fall short

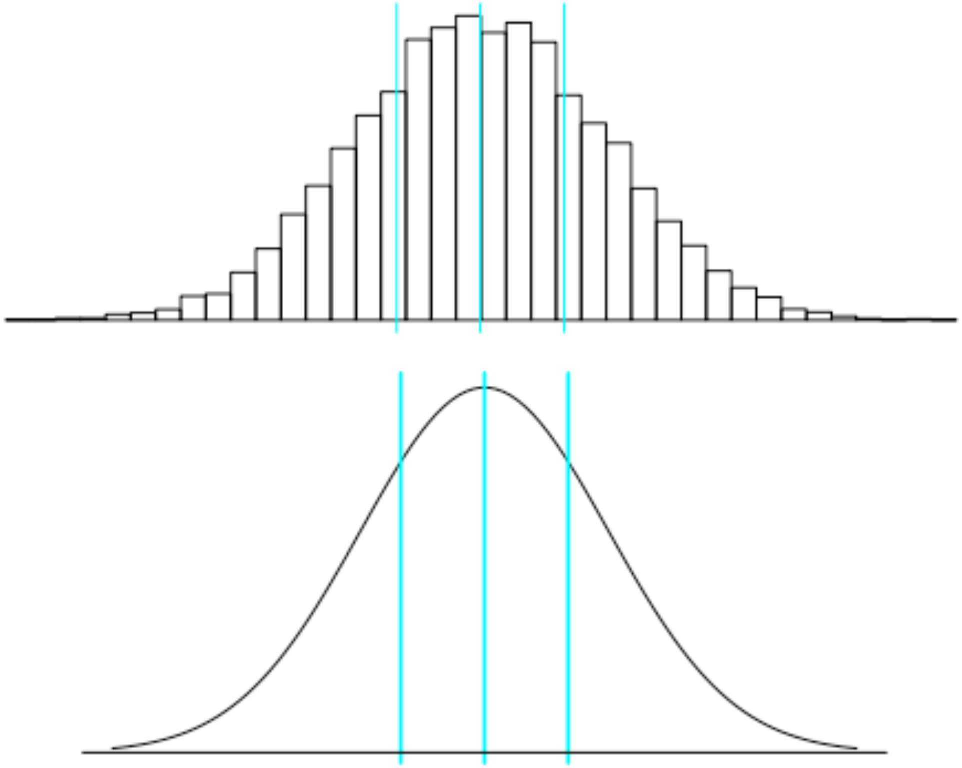


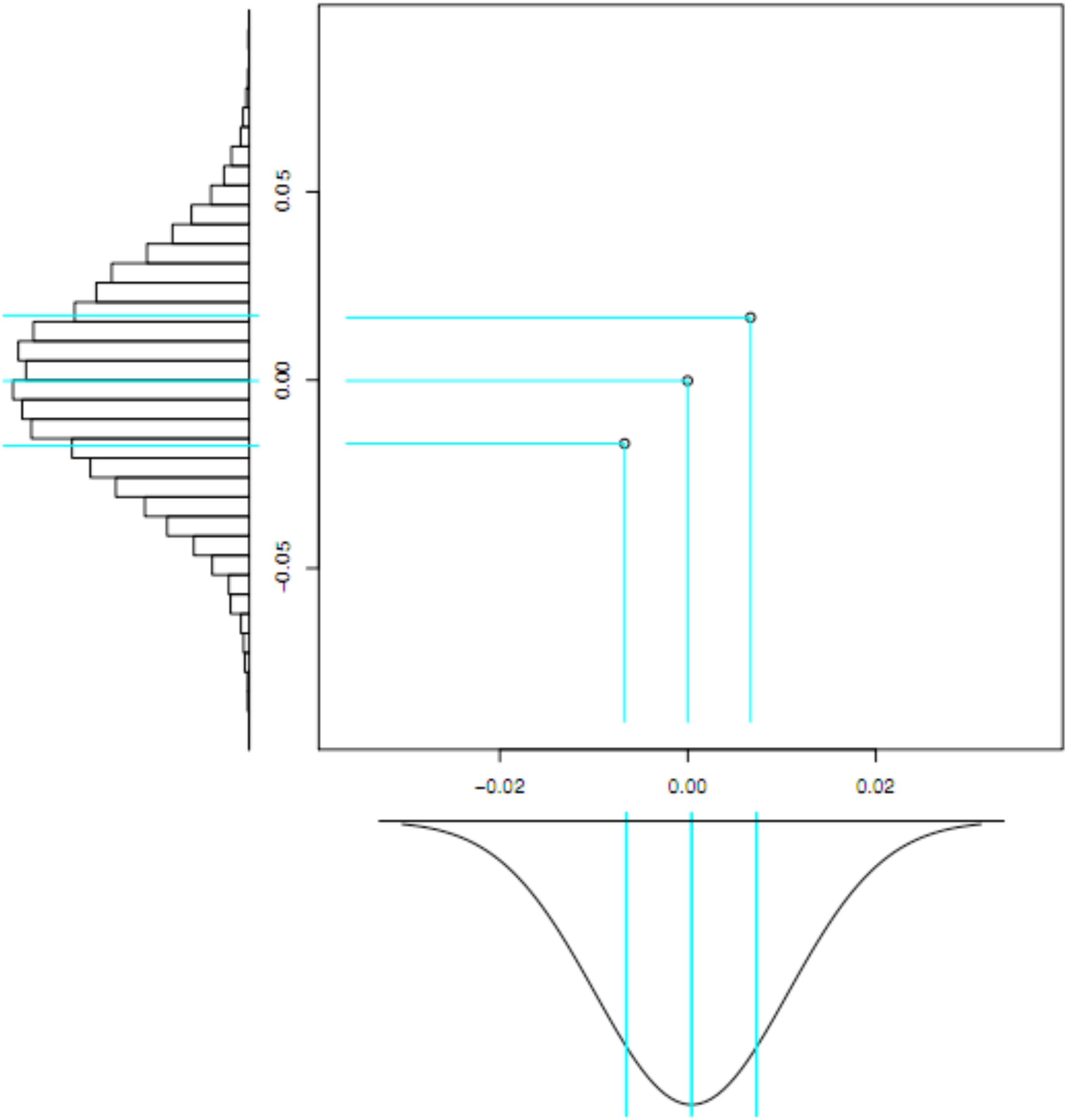
Assessing normality

A normal probability plot compares the way the normal curve distributes probability to the way our sample has arranged its points

Let's start by dividing each into four pieces; for our sample, this means dividing the data using the quartiles we defined for the box plot; for the normal density this means finding regions that divide the total area under the curve into four pieces

To make a more direct comparison, we can try plotting these points against each other...

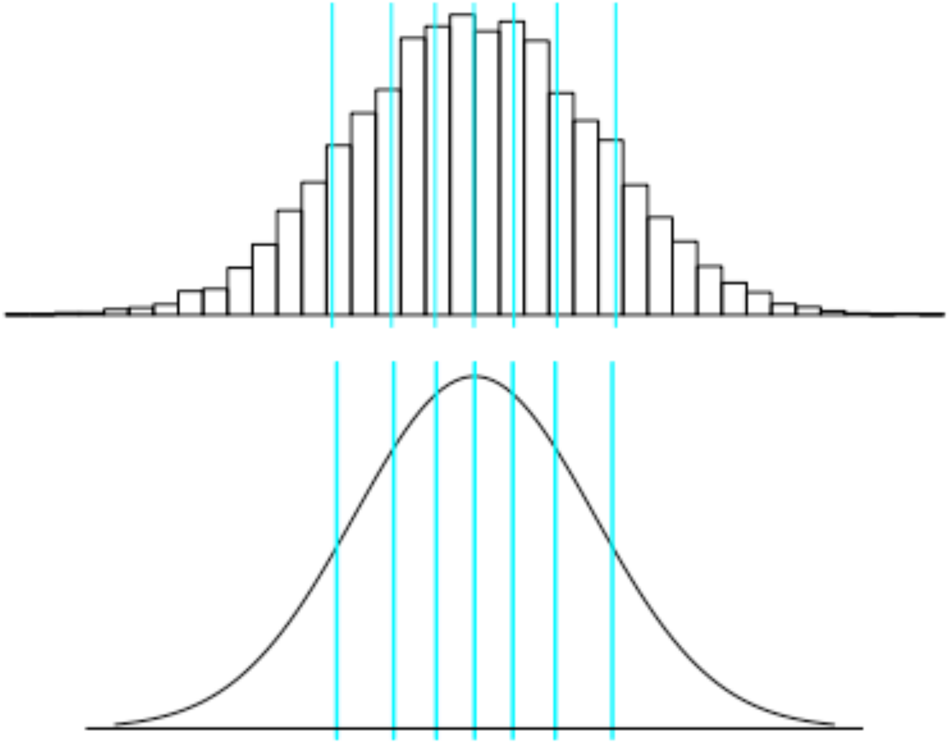


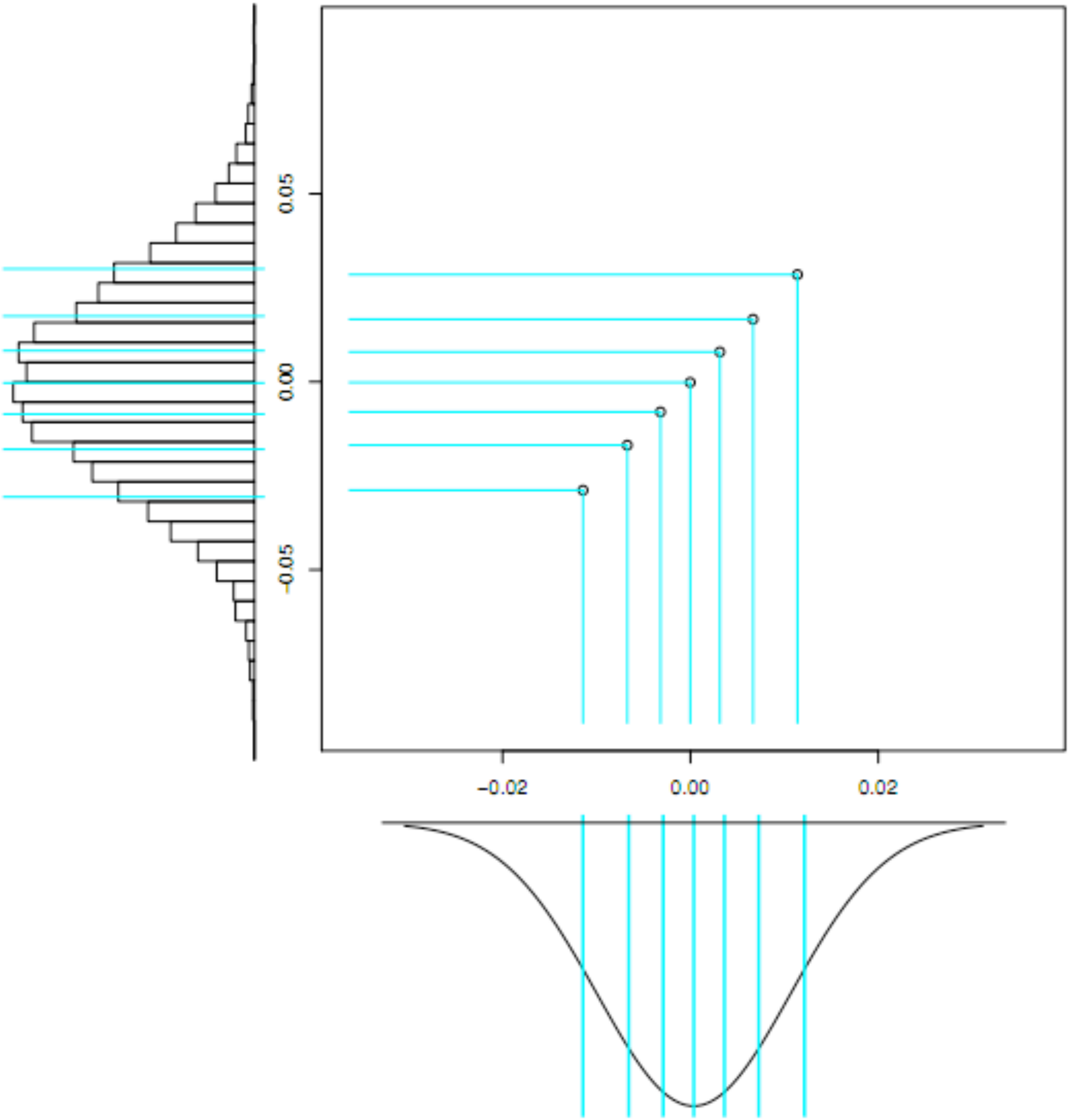

0.05
0.00
−0.05
−0.02
0.00
0.02

Assessing normality

We can continue, this time, dividing the data into 8 pieces (or taking each of the four and dividing them in half)

And again, to make a more direct comparison, we can try plotting these points against each other...



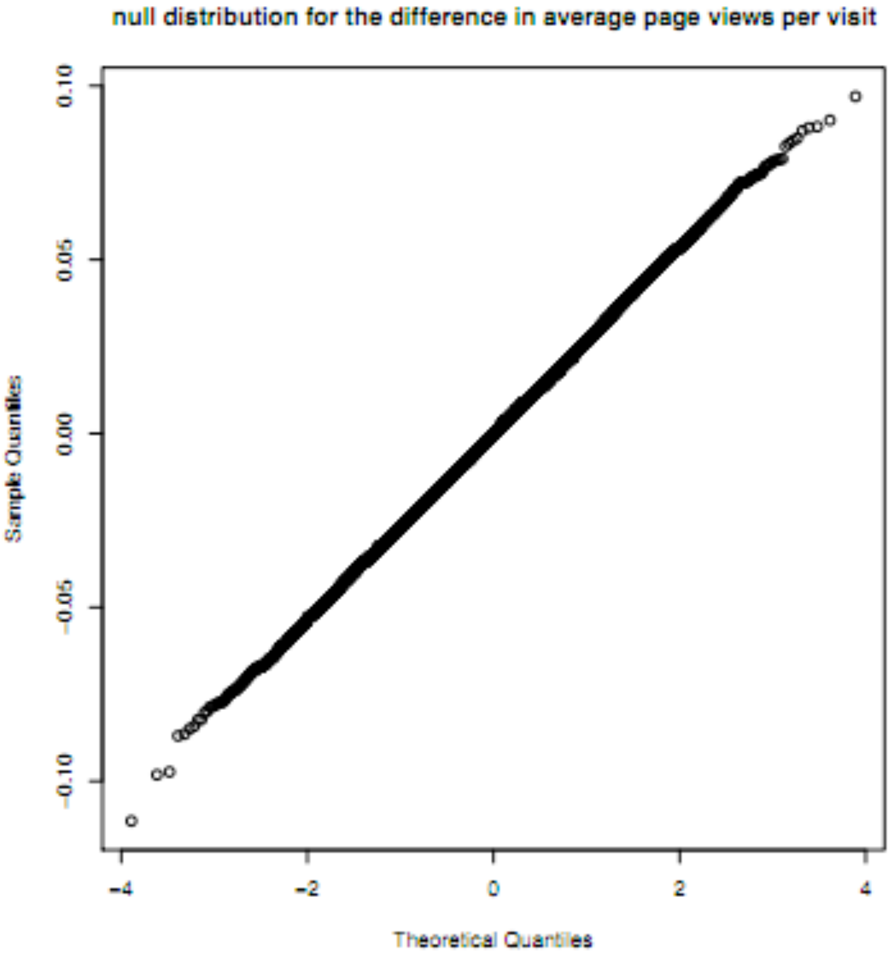

0.05
0.00
−0.05
−0.02
0.00
0.02

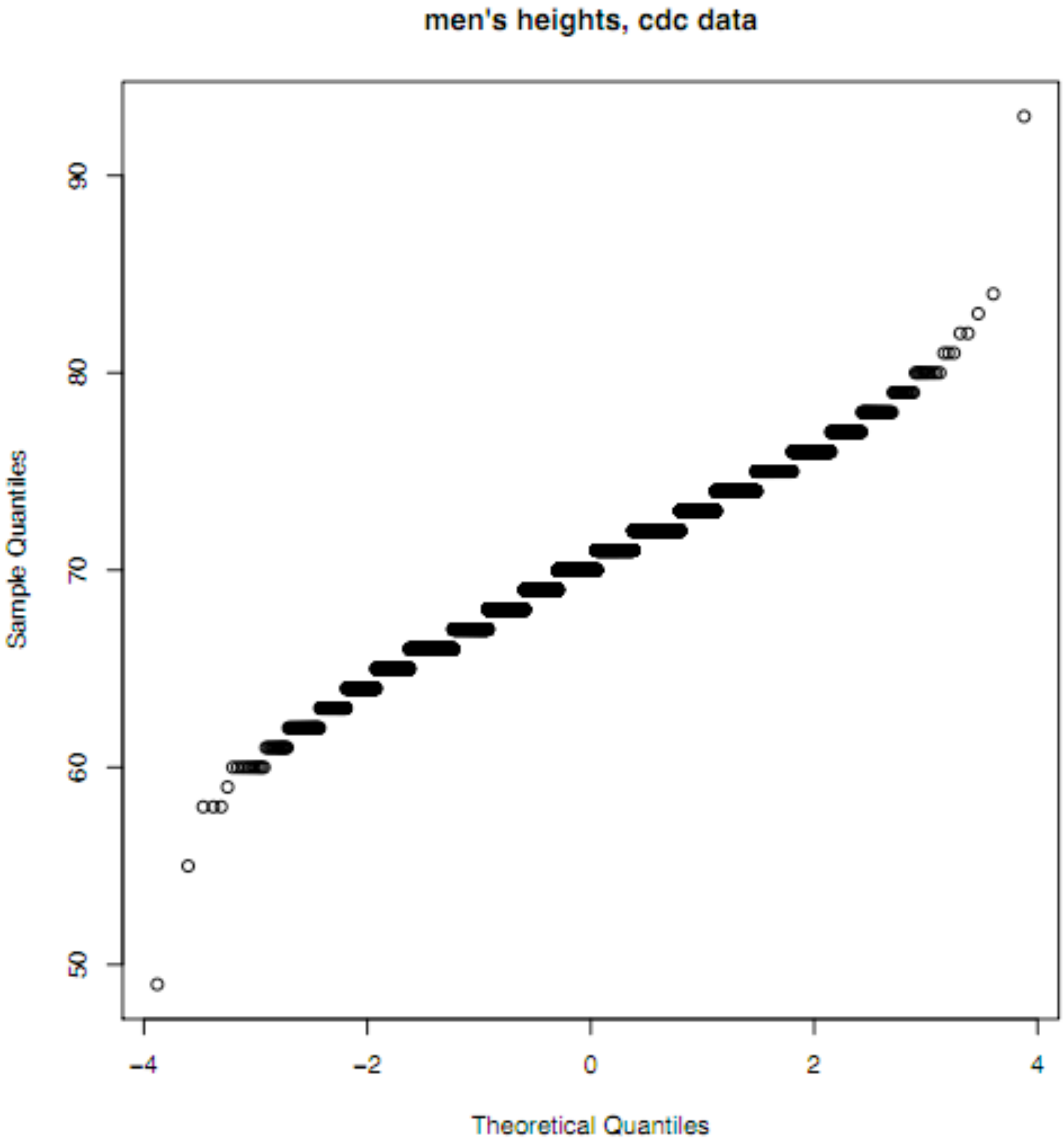
Assessing normality

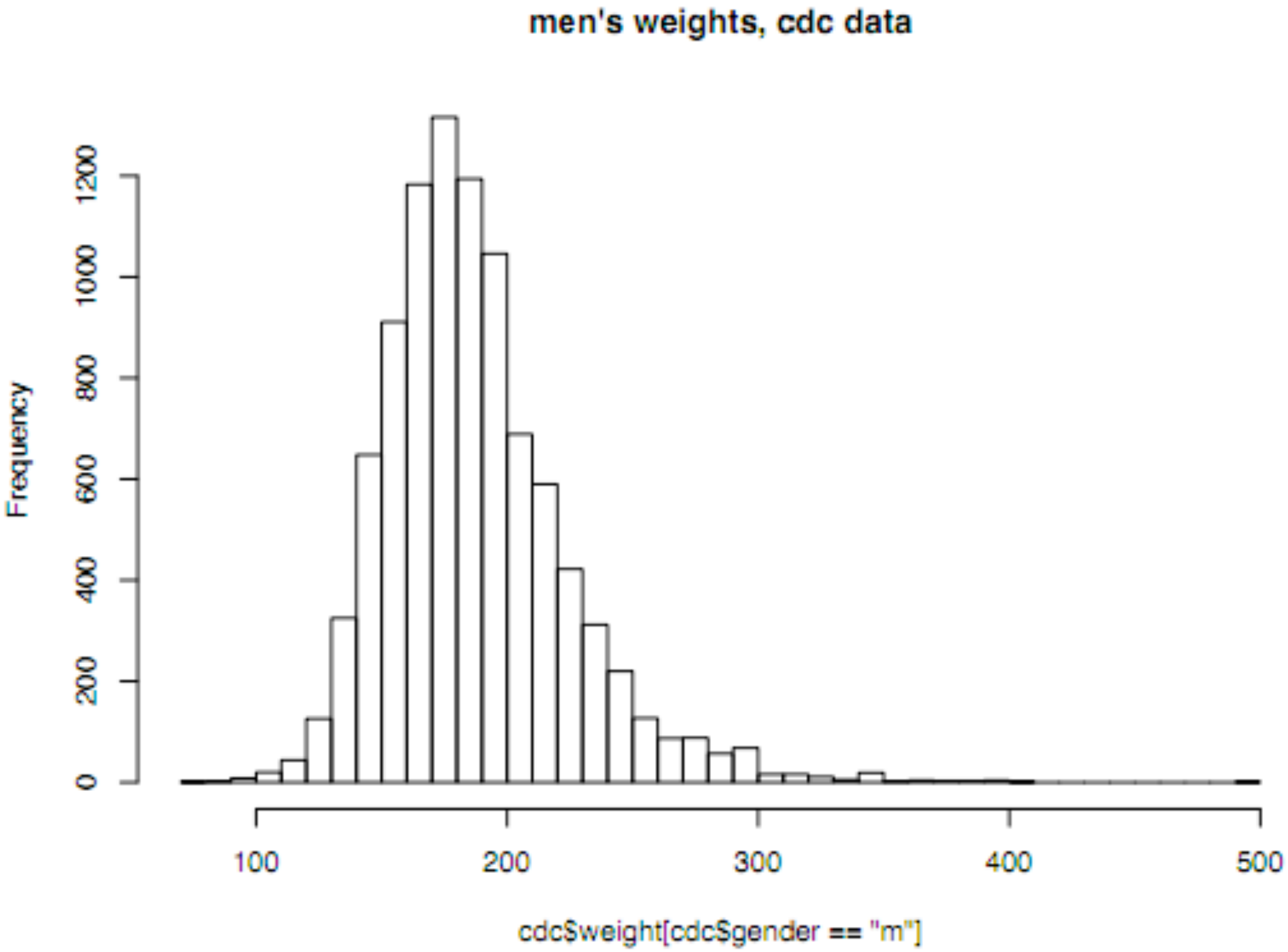
Continuing in this way, we can continue dividing, adding points until we get a plot like the one on the right

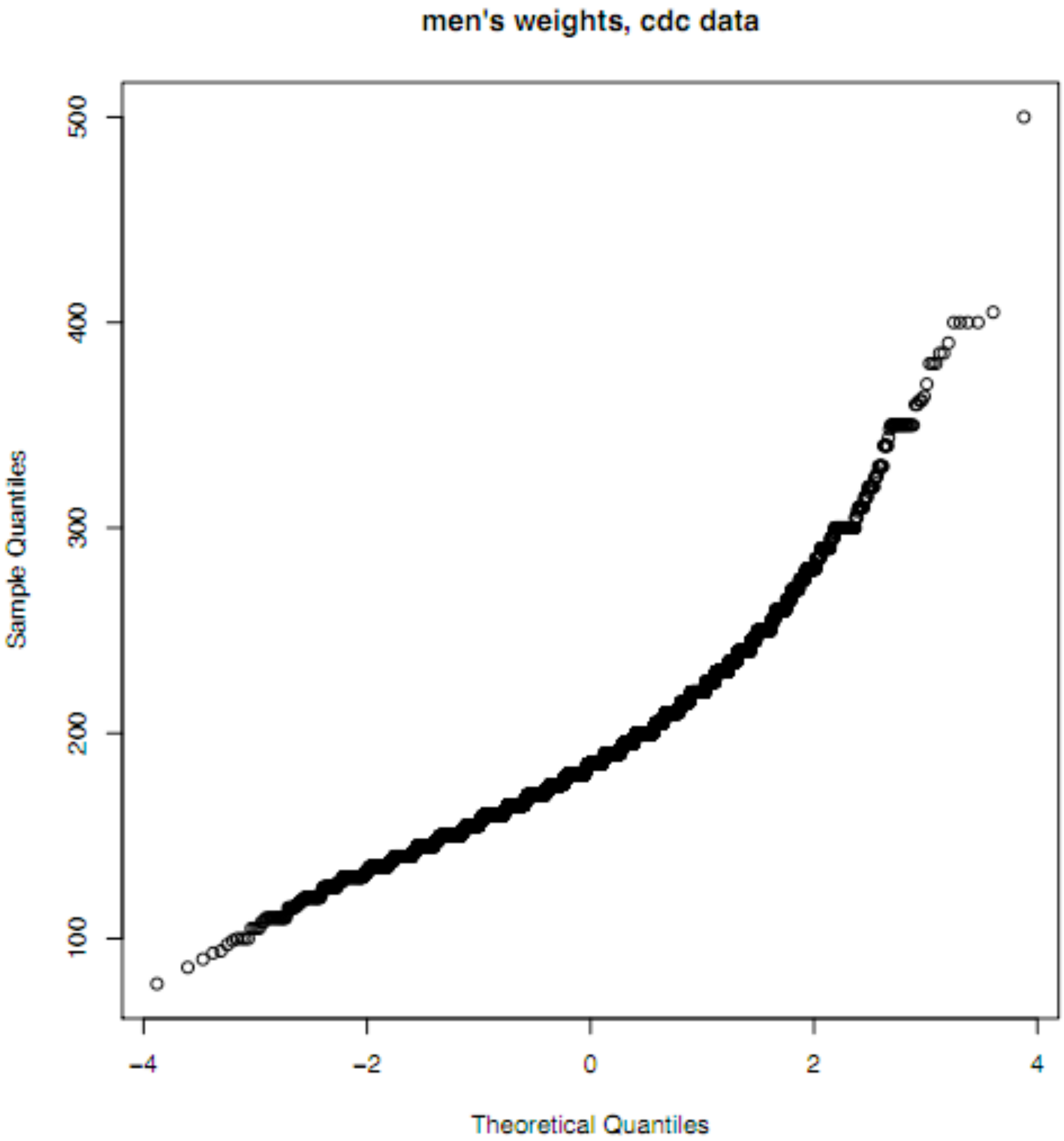
The good thing about this kind of plot is that departures from normality are seen as deviations from a straight line; this is, visually speaking, a HUGE advance

Consider the probability plot for men's heights...




men's heights, cdc data
Sample Quantiles
90
80
70
60
50
Theoretical Quantiles
−4
−2
0
2
4


men's weights, cdc data
Frequency
0
200
400
600
800
1000
1200
100
200
300
400
500
cdc$weight[cdc$gender == "m"]


men's weights, cdc data
Sample Quantiles
100
200
300
400
500
Theoretical Quantiles
−4
−2
0
2
4