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

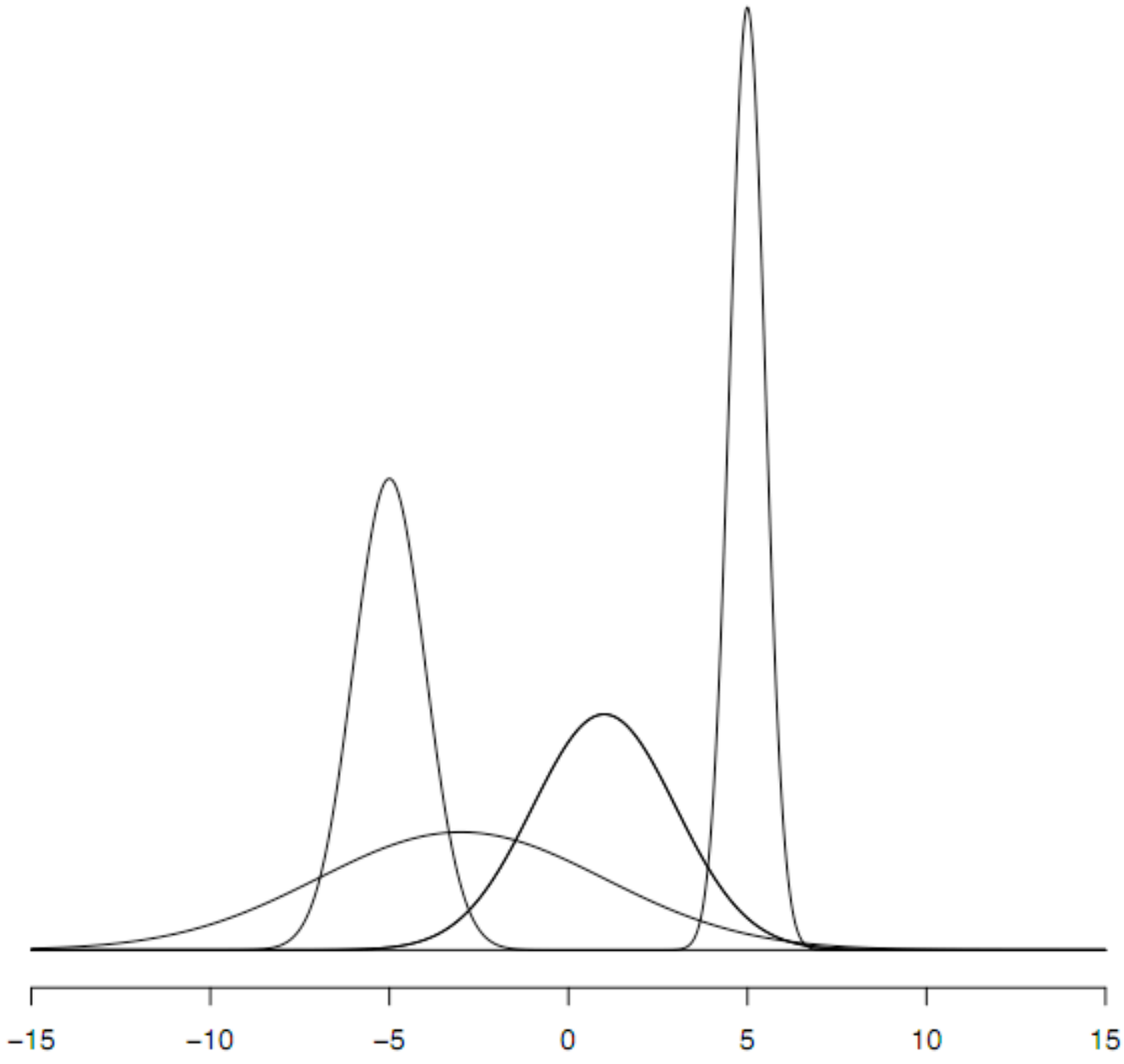
The normal distribution

The so-called bell curve is a very special “probability distribution” — it’s a theoretical model or description of data. While the term is often used informally, it has a very precise mathematical formula that governs its shape. It also has a remarkable history.

> “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 there 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 1870s: of populations of people, of measurements and of their similarities.”
>
> *Normative Terminology* by W. H. Kruskal

The normal distribution

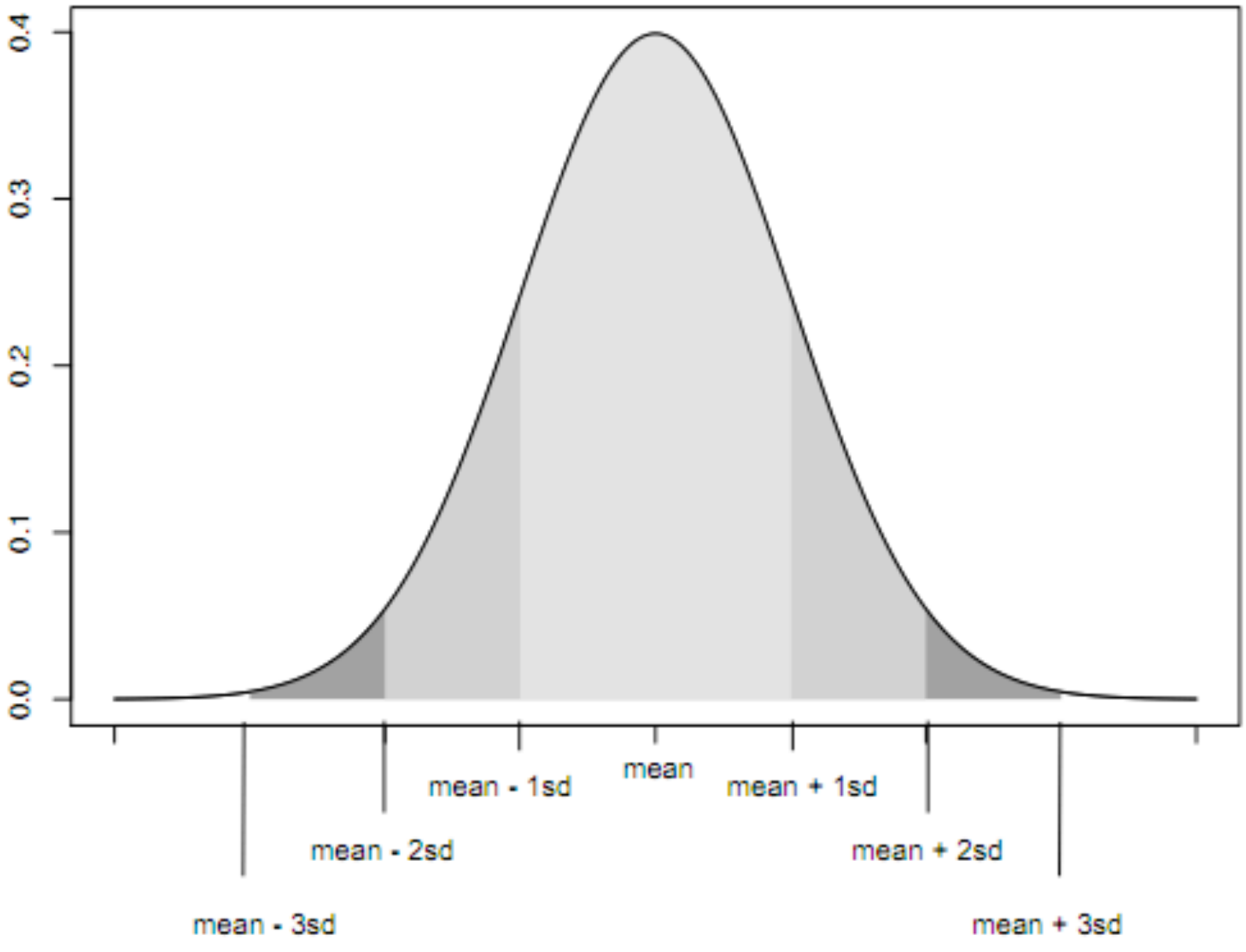
The normal distribution is not just one curve, it's a family of curves — in technical terms it's a location-scale family. That is, the center of the bell is called the mean and its width, or spread, is governed by its standard deviation.

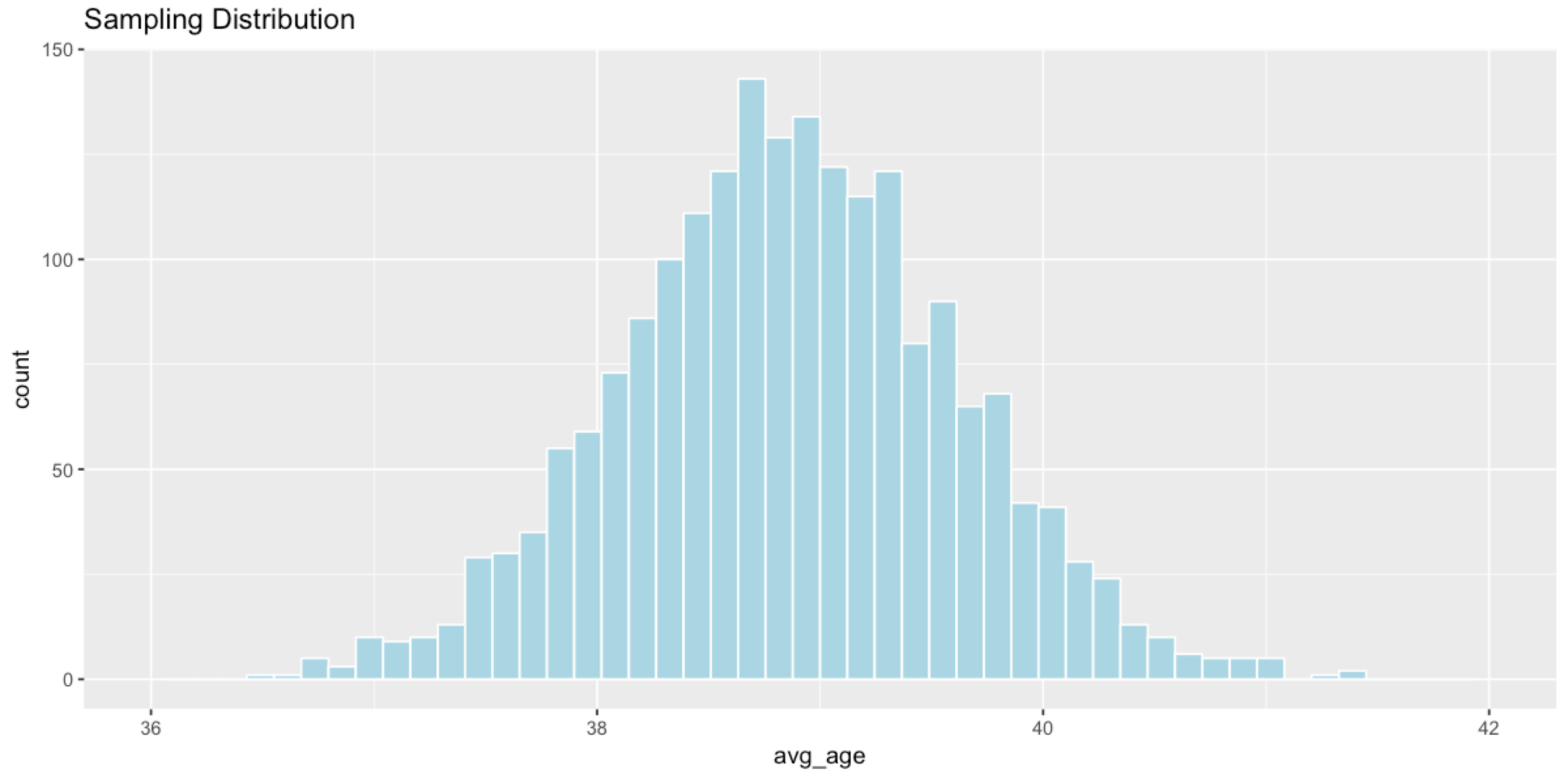


The normal distribution

The mathematical expression describing these curves define for us how we expect to see “normal” data arranged. There are a few simple rules of thumb, for example, that dictate what makes a bell curve a bell. Its shape is really a relationship between the its center and spread, its mean and standard deviation.

If our data follow a normal distribution, the math suggests that about 68% should be within one standard deviation of the mean, about 95% should be within two standard deviations of the mean and about 99% should be within 3. These rules hold for every member of the normal “family” no matter how we choose the mean and standard deviation.


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


Sampling Distribution
150
100
50
0
36
38
40
42
avg_age

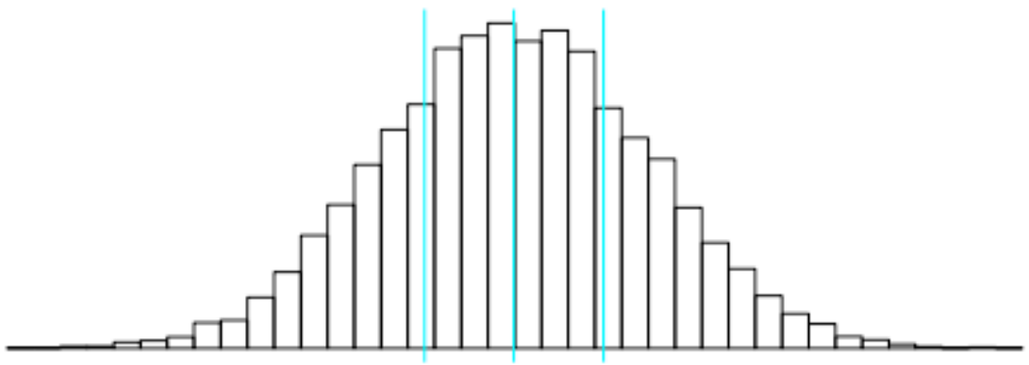
Assessing normality

The normal distribution is a “model” we might use to describe a set of data, a model that depends on two “parameters,” the mean and standard deviation — data that follow a bell curve come up in a variety of places that we’ll see shortly

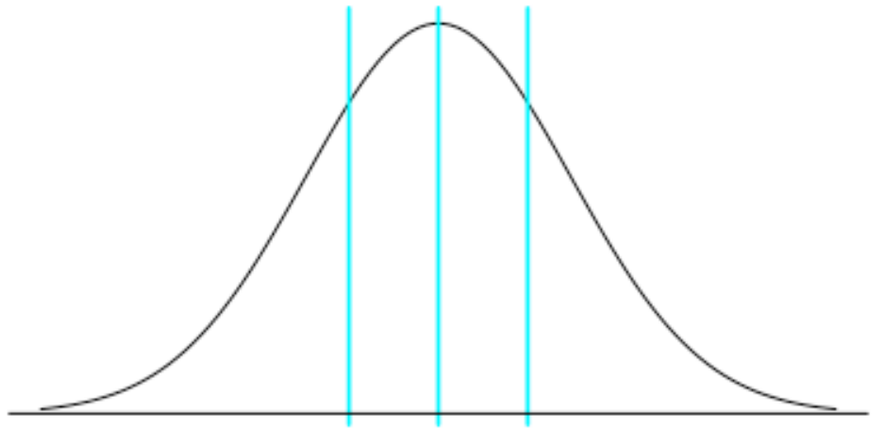
A simple device to judge the normality is called a normal quantile-quantile plot, or just Q-Q plot for short…

Assessing normality

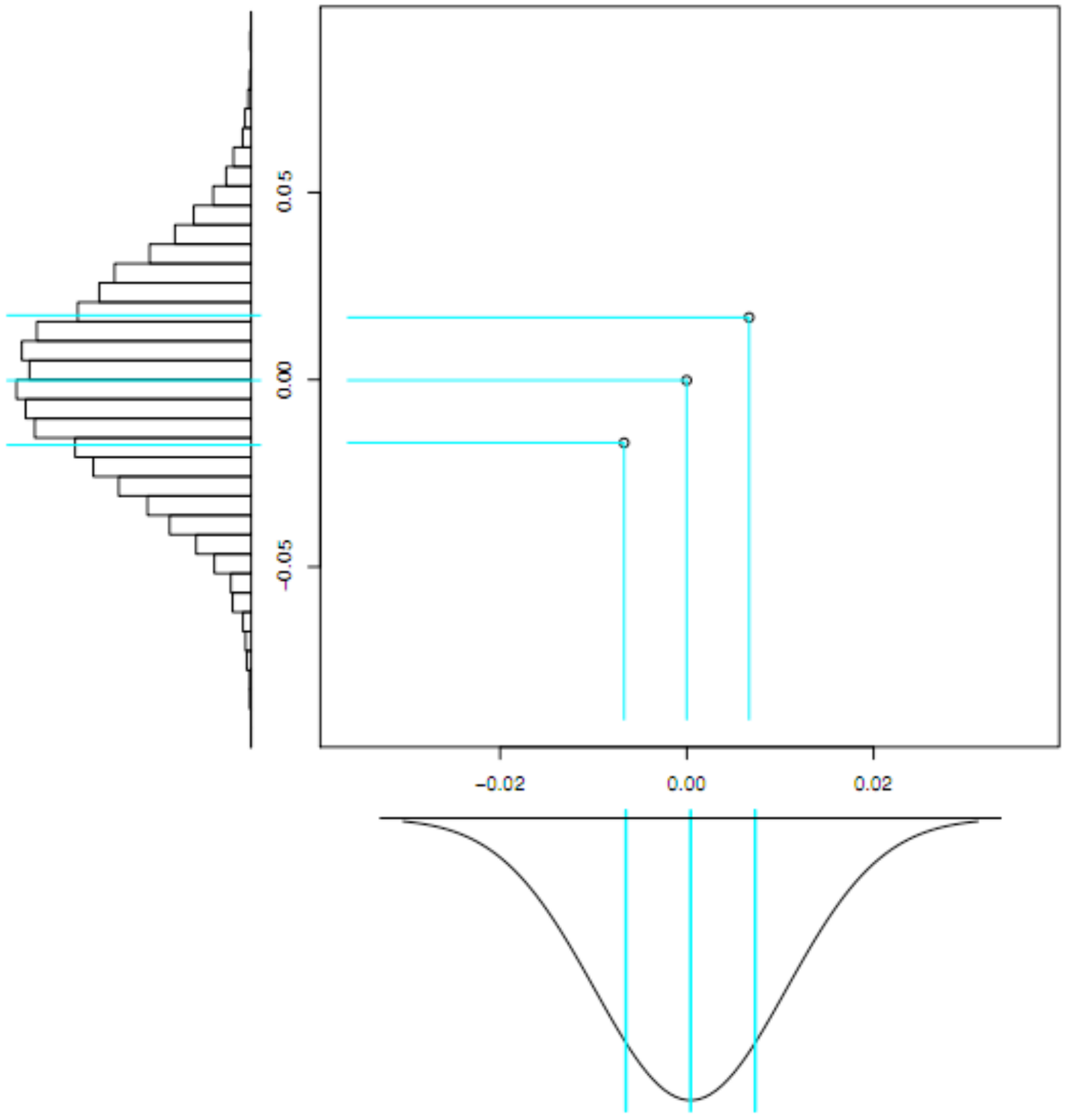
A normal probability plot compares the way the normal curve distributes probability to the way our sample has arranged its points



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

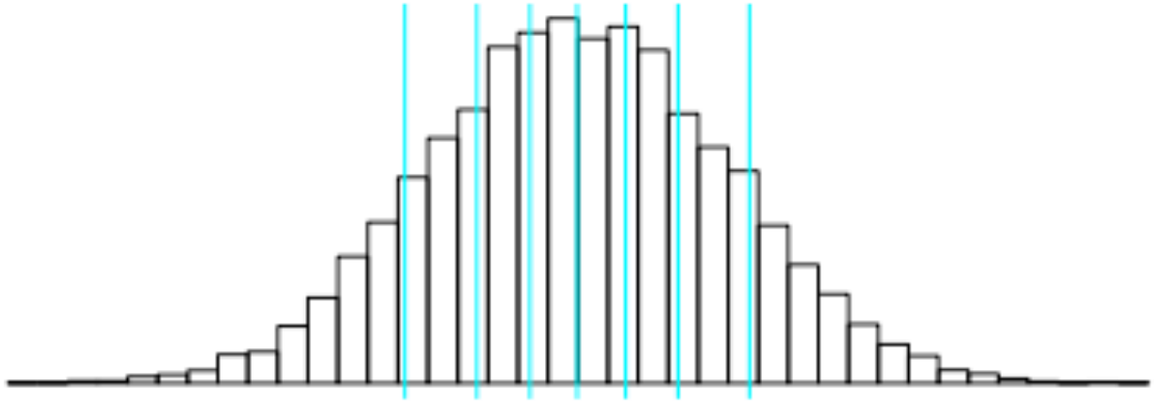


To make a more direct comparison, we can try plotting these points against each other...

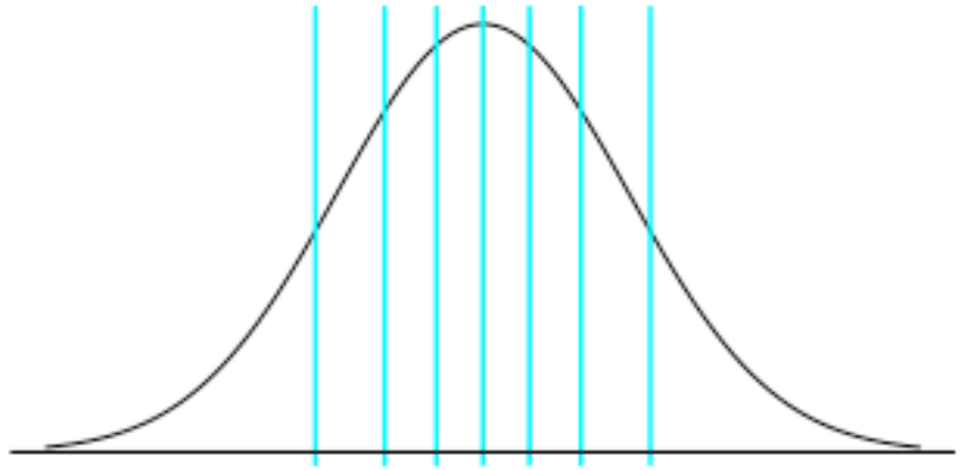

0.05
0.00
−0.05
−0.02
0.00
0.02

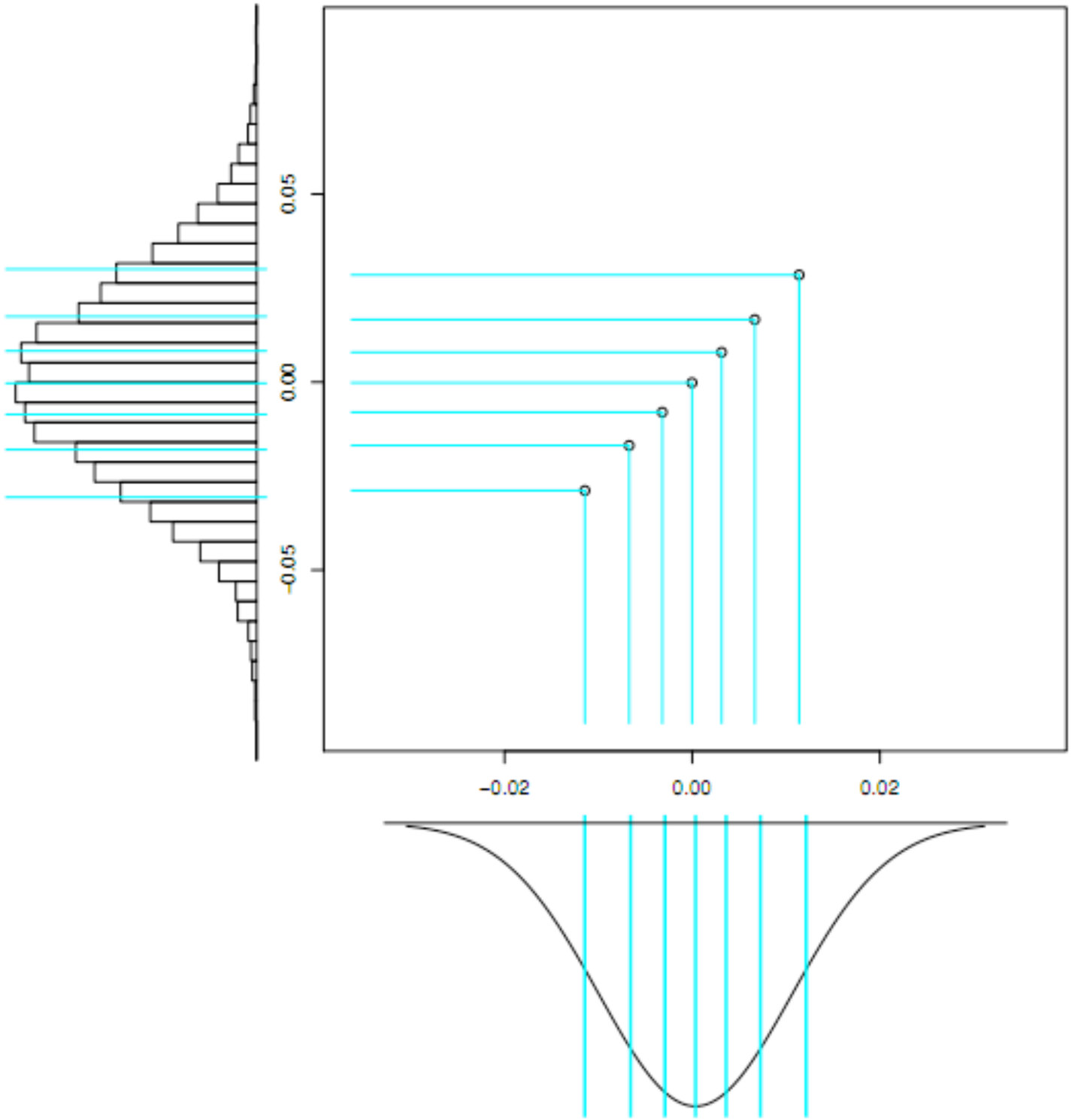
Assessing normality

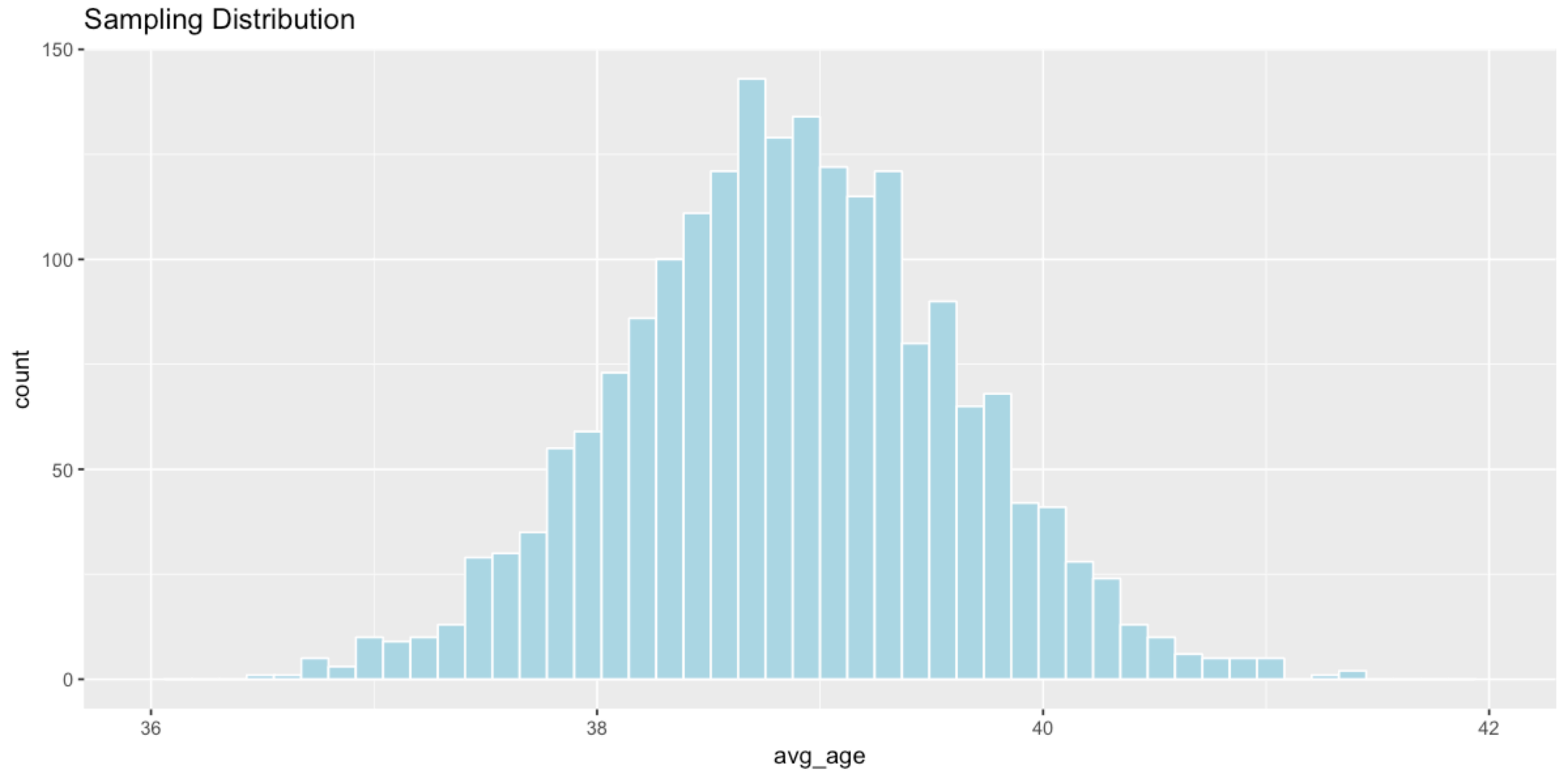
We can continue, this time, dividing the data into 8 pieces (or taking each of the four and dividing them in half)

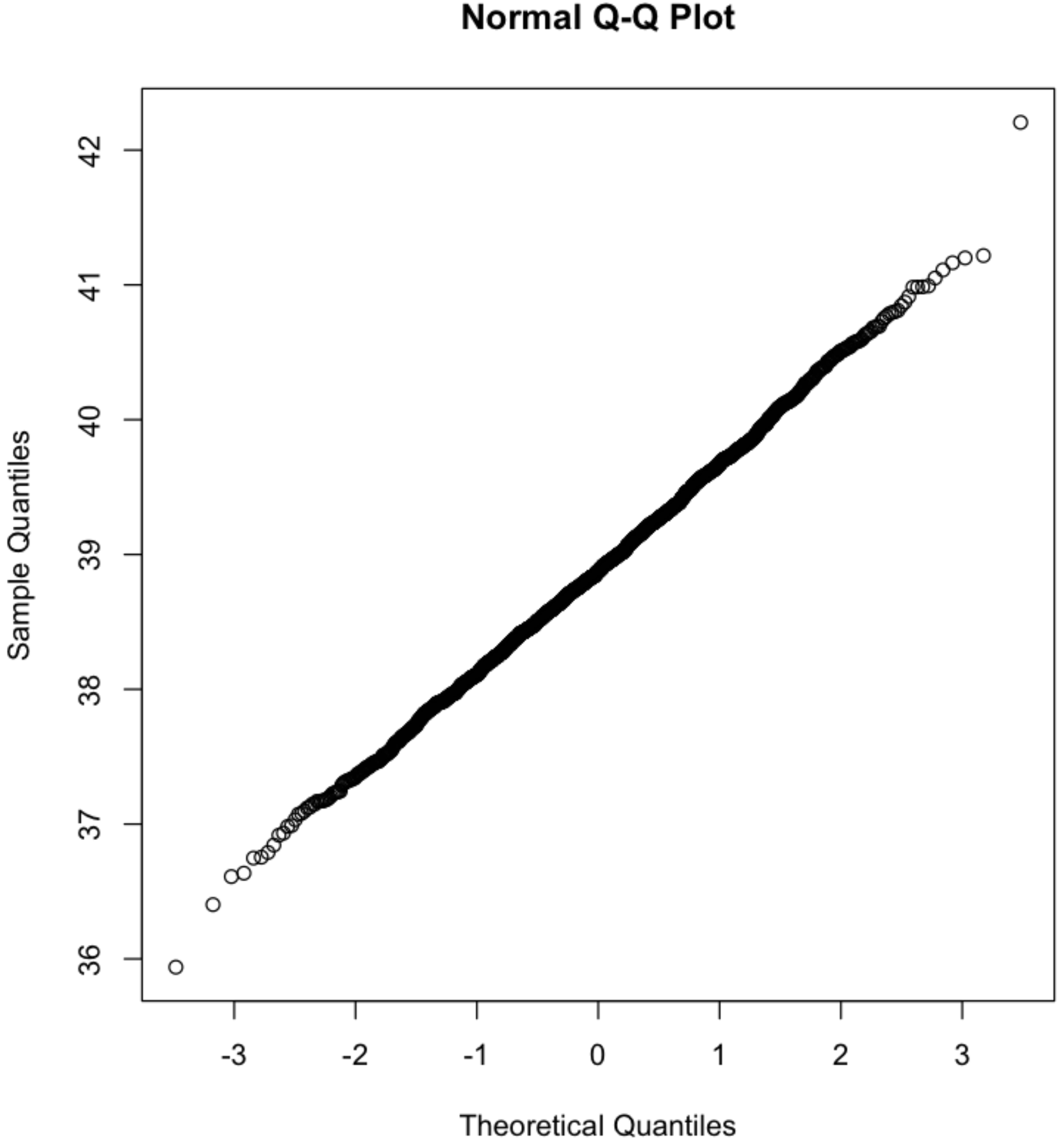


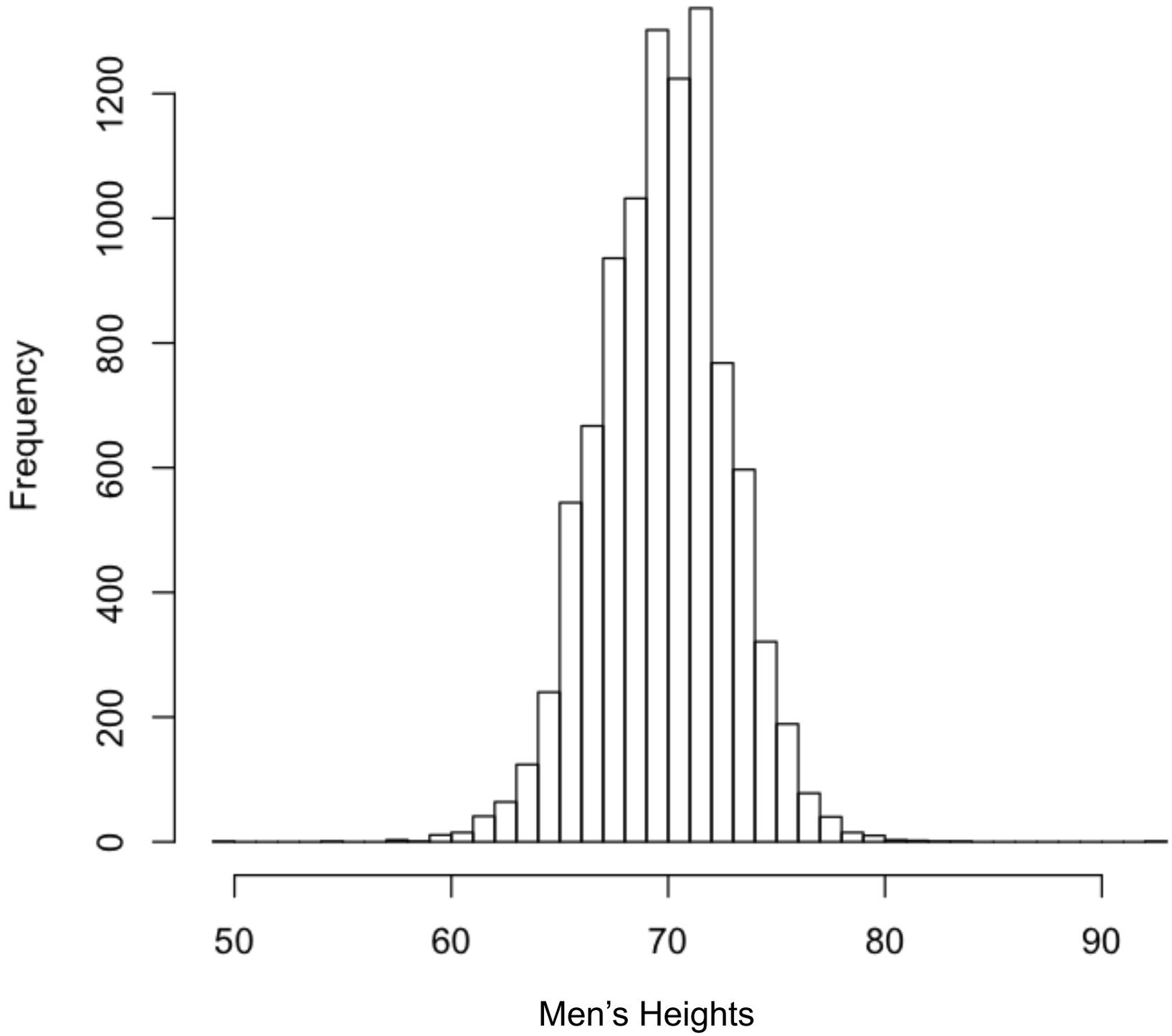
And again, to make a more direct comparison, we can try plotting these points against each other...

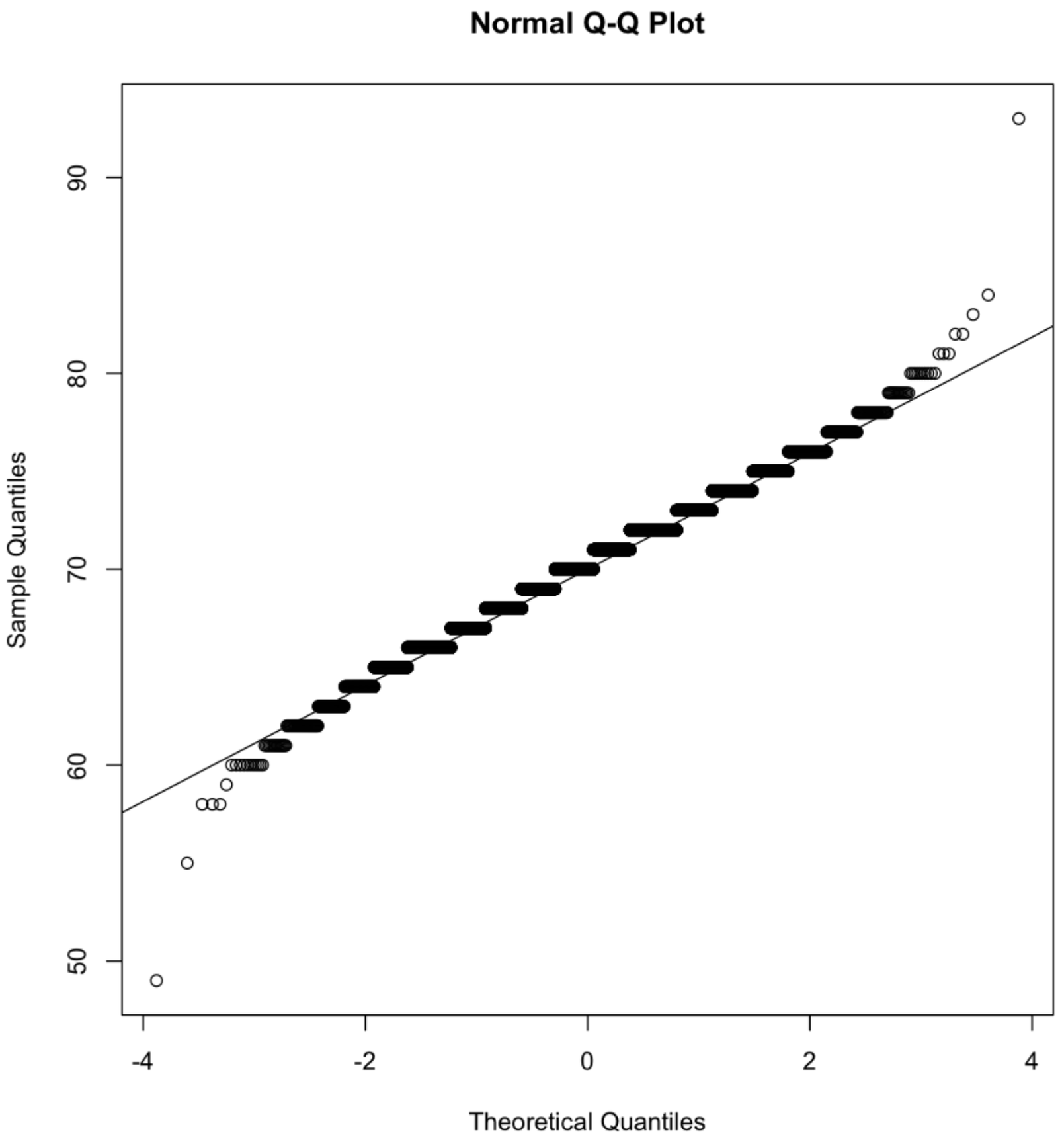


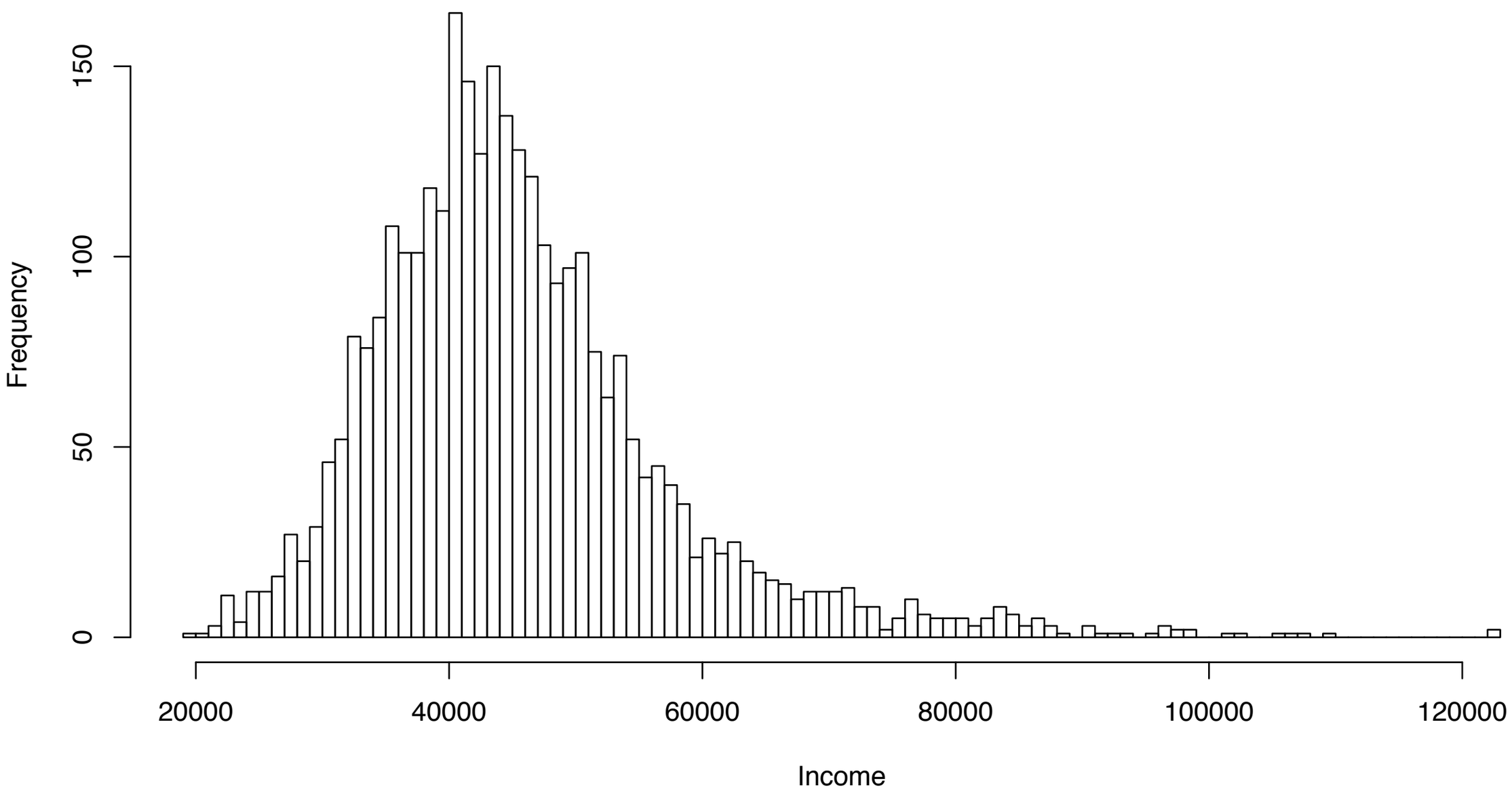

0.05
0.00
−0.05
−0.02
0.00
0.02

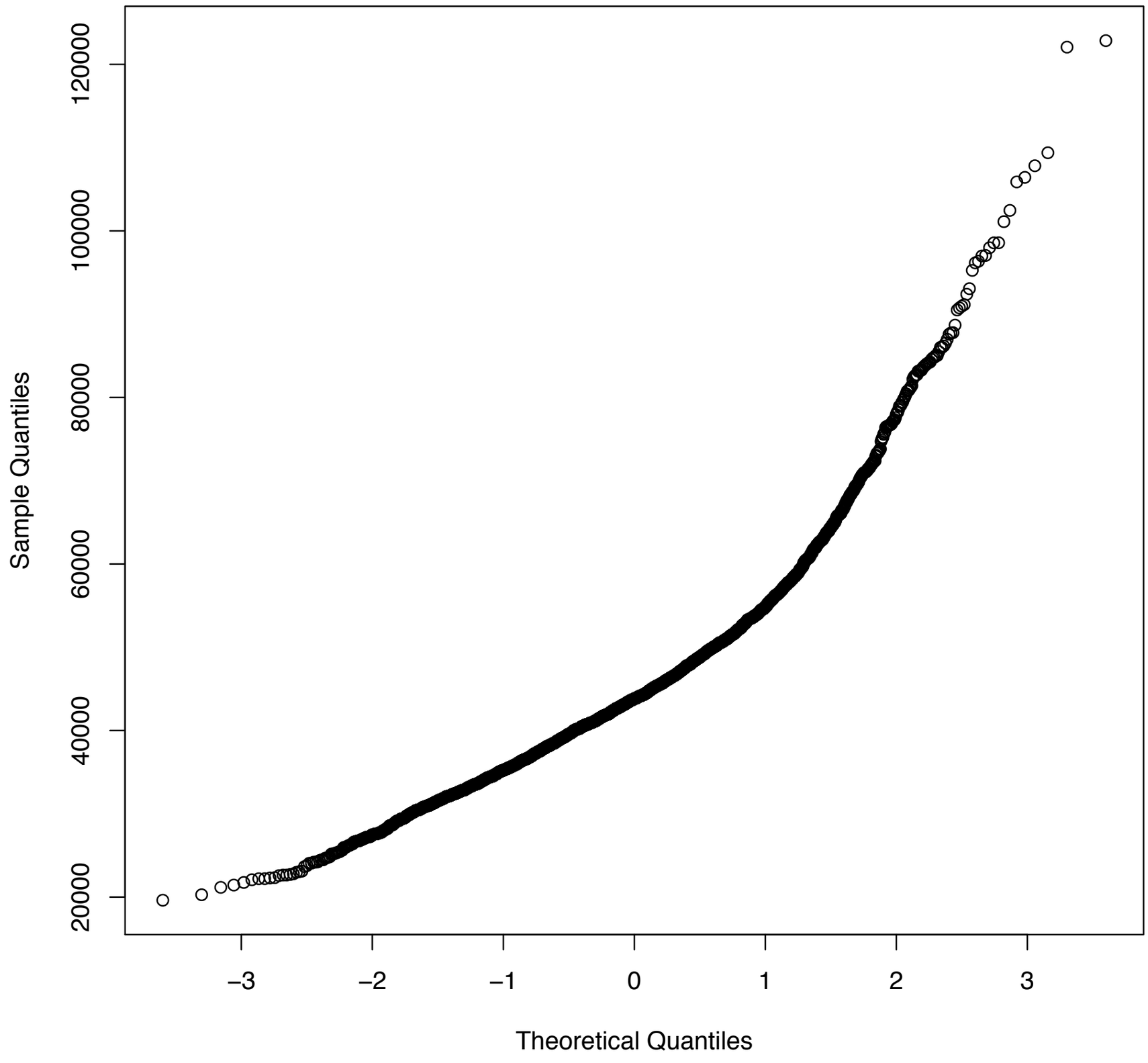

Sampling Distribution
150
100
50
0
36
38
40
42
avg_age


Normal Q-Q Plot
Sample Quantiles
36
37
38
39
40
41
42
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Histogram of Men's Heights, BRFSS
Frequency
0
200
400
600
800
1000
1200
50
60
70
80
90
Men's Heights


Normal Q-Q Plot
Sample Quantiles
90
80
70
60
50
-4
-2
0
2
4
Theoretical Quantiles


Histogram of income
Frequency
0
50
100
150
20000
40000
60000
80000
100000
120000
Income


QQ plot of Income
Sample Quantiles
20000
40000
60000
80000
100000
120000
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Numerical descriptions

A statistic is something you compute from data — so far we have been focusing on graphical descriptions of data, but now it's time to look at summary or descriptive statistics

There is a lot of history behind these that help motivate their use — we'll start with the humble sample mean and sample standard deviation

As we have seen, the normal distribution is a model for data and it's two parameters are the center and spread, the mean and standard deviation — given data that follow a normal, we might think of the sample mean and sample standard deviation as "estimates" of the parameters in this model

We will have a lot to say about estimation in the weeks ahead, but for now, it's sufficient to keep in your head the link between the normal distribution and these quantities — it will provide us clues for when they might not be such great descriptions of a data set

Sample mean and standard deviation

Given a set of n data points $x_1, \ldots, x_n$ we compute the sample mean and standard deviation

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

Why n-1?

Let's look into this a little more closely -- First note that the sum of the deviations around the mean is zero. Since $n\bar{x} = x_1 + \cdots + x_n$

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

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

The sample mean

This also gives us an interpretation of the mean — it is the “balance point” of our data. Archimedes worked out that if you place weights on a lever, the sum of the products of the weights and their distances to the left to the left of the balance point equals the sum of the products of weights and their distances to the right of the balance point.

The sample mean

We can also think of the mean as the solution to an optimization problem -- Suppose we want to find the point b that minimizes the sum of squares

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

We can rewrite this expression as

$$(x_1 - \bar{x} + \bar{x} - b)^2 + (x_2 - \bar{x} + \bar{x} - b)^2 + \cdots + (x_n - \bar{x} + \bar{x} - b)^2$$

which we can simplify to

$$\begin{aligned}(x_1 - \bar{x})^2 + (\bar{x} - b)^2 + (x_2 - \bar{x})^2 + (\bar{x} - b)^2 + \cdots + (x_n - \bar{x})^2 + (\bar{x} - b)^2 \\ +2(x_1 - \bar{x})(\bar{x} - b) + 2(x_2 - \bar{x})(\bar{x} - b) + \cdots + 2(x_n - \bar{x})(\bar{x} - b)\end{aligned}$$

but the bottom row is really just

$$2(\bar{x} - b)\left[(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x})\right] = 0$$

The sample mean

In the end, we can rewrite the sum of squares to be

$$(x_1 - \bar{x})^2 + (\bar{x} - b)^2 + (x_2 - \bar{x})^2 + (\bar{x} - b)^2 + \cdots + (x_n - \bar{x})^2 + (\bar{x} - b)^2$$

and clearly if we take $b = \bar{x}$ we get a minimum (any other value of b gives an error that's bigger!)

Sample mean and standard deviation

The sample mean, then, is the point b that is “central” in the sense that it is associated with the minimum error, the sum of squares

We can then interpret the sample standard deviation as the smallest error you can get if you try to approximate all your data points by a constant b -- The smaller the standard deviation, the better you can do, or equivalently, the more tightly clumped your data are

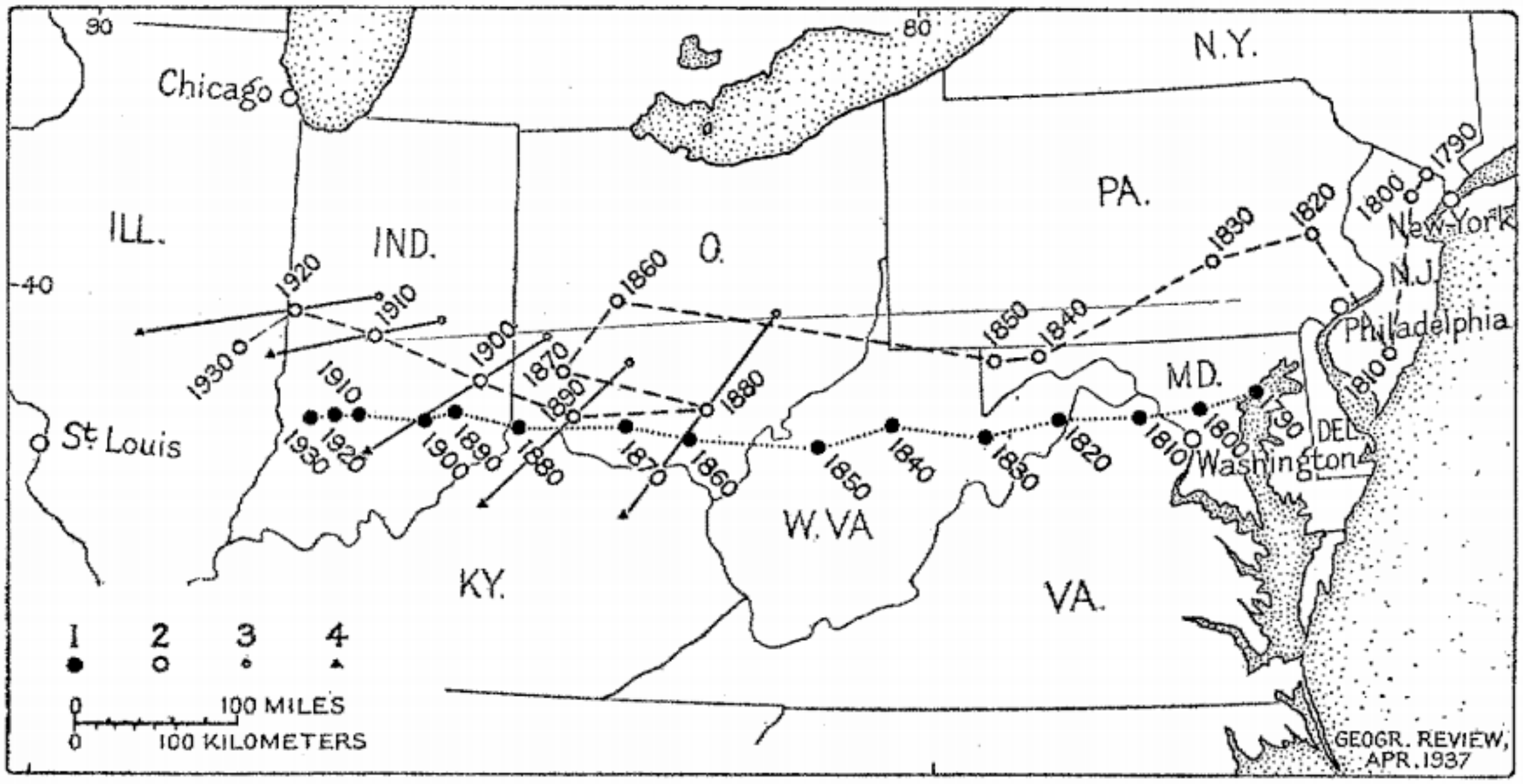


FIG. 5—Centrogram showing movements of the centers of population and higher education in the United States of America from 1790 to 1930. Key: centers of 1, general population; 2, higher educational population (universities and colleges); 3, higher educational population, men; 4, higher educational population, women. (From a study made by Walter C. Eells for publication in the forthcoming Mendeleev Memorial Volume of the Centrographical Laboratory in Leningrad.)

CANADA

1 : 30,000,000

0 300 MILES

0 300 KILOMETERS

1931

Sault Ste Marie

Sudbury

Quebec

Montreal

1851

Ottawa

Toronto

50

45

85 80 75 70

UNITED STATES

1 : 30,000,000

Chicago

1930

Cincinnati

Baltimore

New York

Philadelphia

1790

Washington

45

40

85 80 75

EUROPE

1 : 30,000,000

0 300 MILES

0 300 KILOMETERS

Prague

Passau

1720

1930

Munich

Linz

Vienna

Budapest

Bern

Milan

R. Danube

50

45

10 15 20

GEOGR. REVIEW, APR. 1937

U.S.S.R.

1 : 30,000,000

Moscow

Tambov

1858

Penza

Samara

Saratov

1926

Kharkov

R. Don

R. Volga

R. Ural

Stalingrad

55

50

35 40 45 50 55

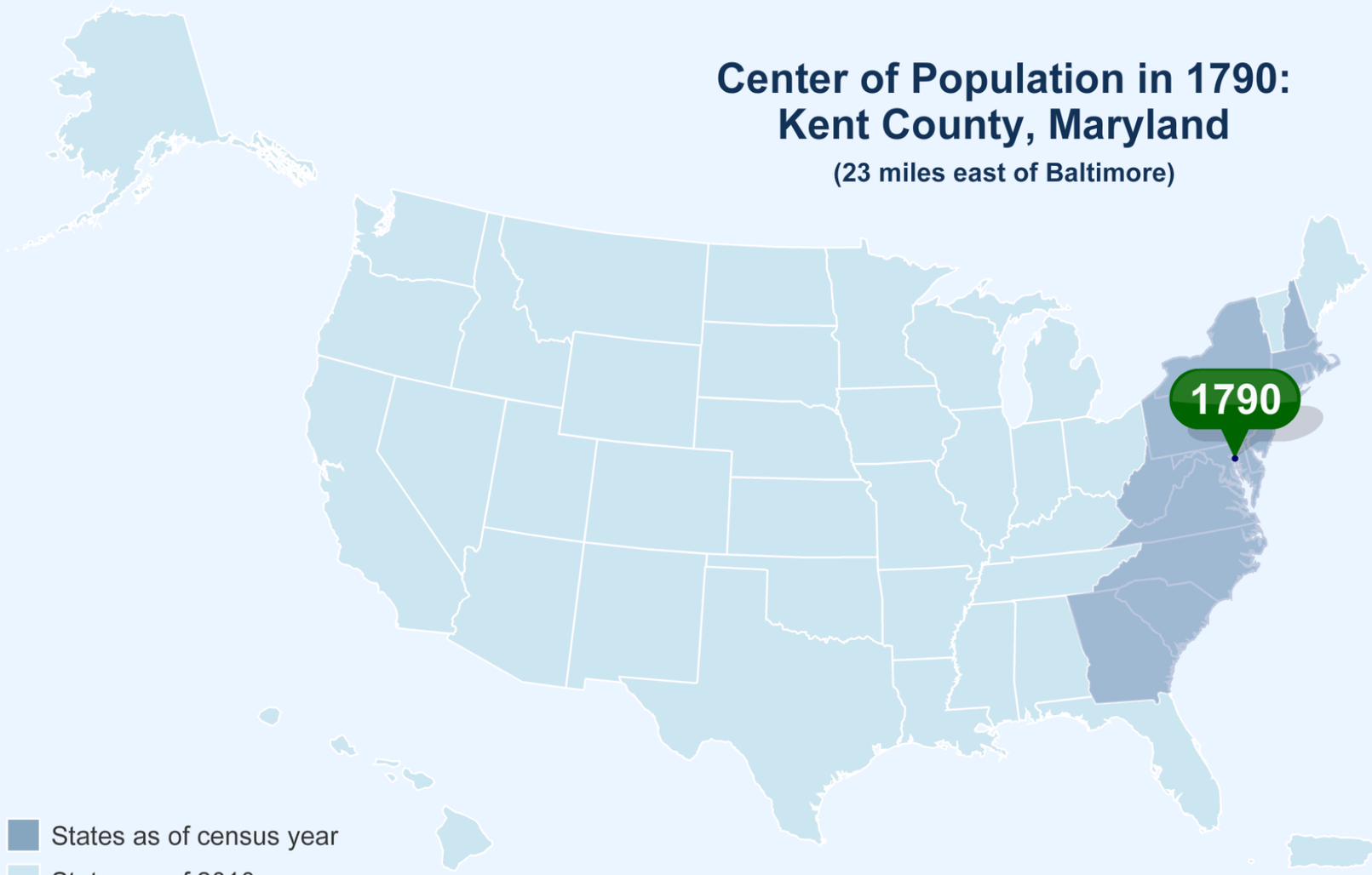
HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.

Center of Population in 1790: Kent County, Maryland

(23 miles east of Baltimore)



Kent County, Maryland

LAT: 39.27500 LONG: 76.18667

Each decade, after it tabulates the decennial census, the Census Bureau calculates the center of population. Historically, it has followed a trail that reflects the sweep of the nation's brush stroke across America's population canvas—the settling of the frontier, waves of immigration and the migration west and south. Since 1790, the location has moved in a westerly, then a more southerly pattern.

1790 2010

United States Census Bureau

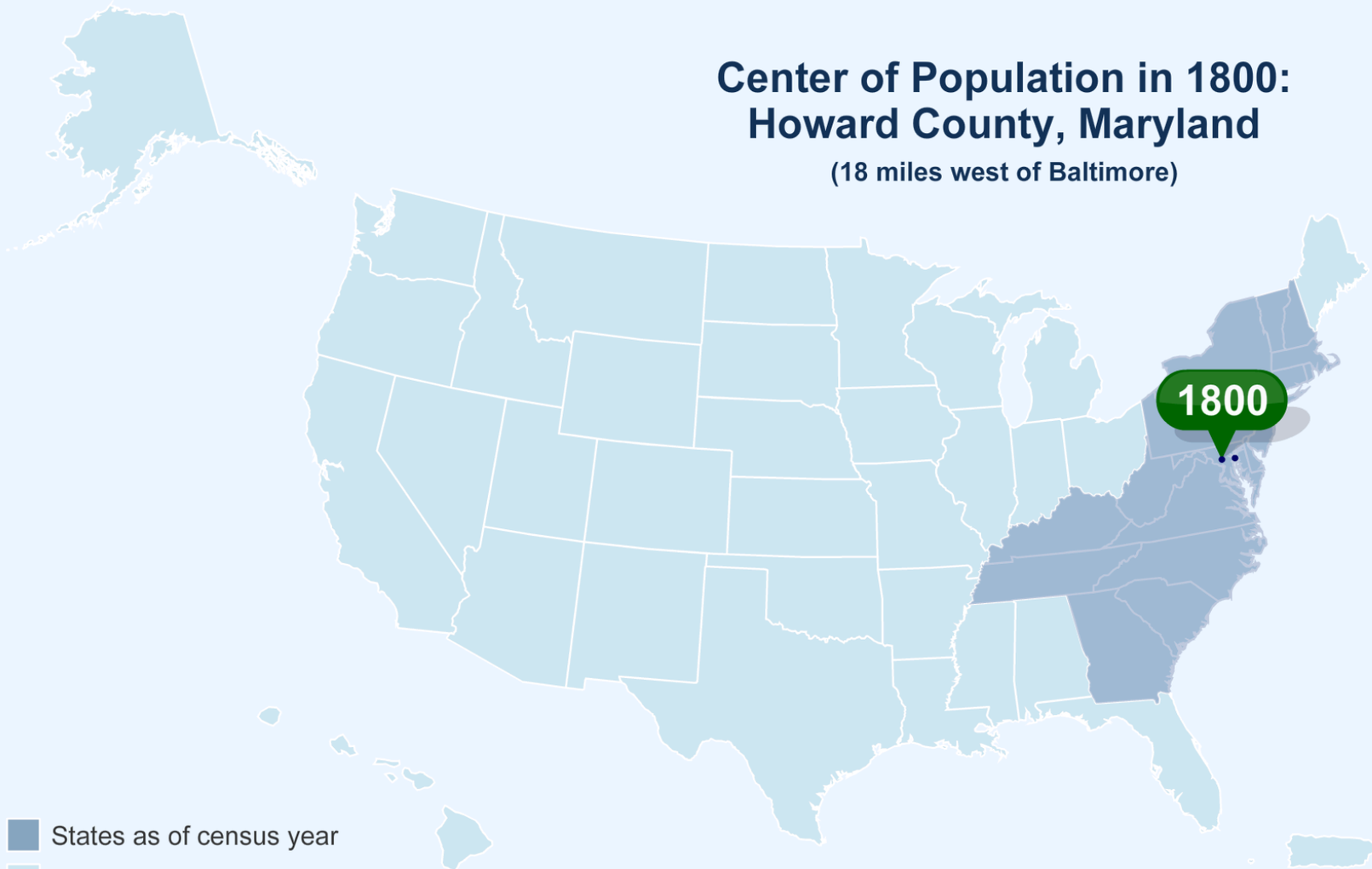
HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.

Center of Population in 1800: Howard County, Maryland

(18 miles west of Baltimore)



Howard County, Maryland
LAT: 39.26833 LONG: 76.94167

United States Census Bureau

1790 2010

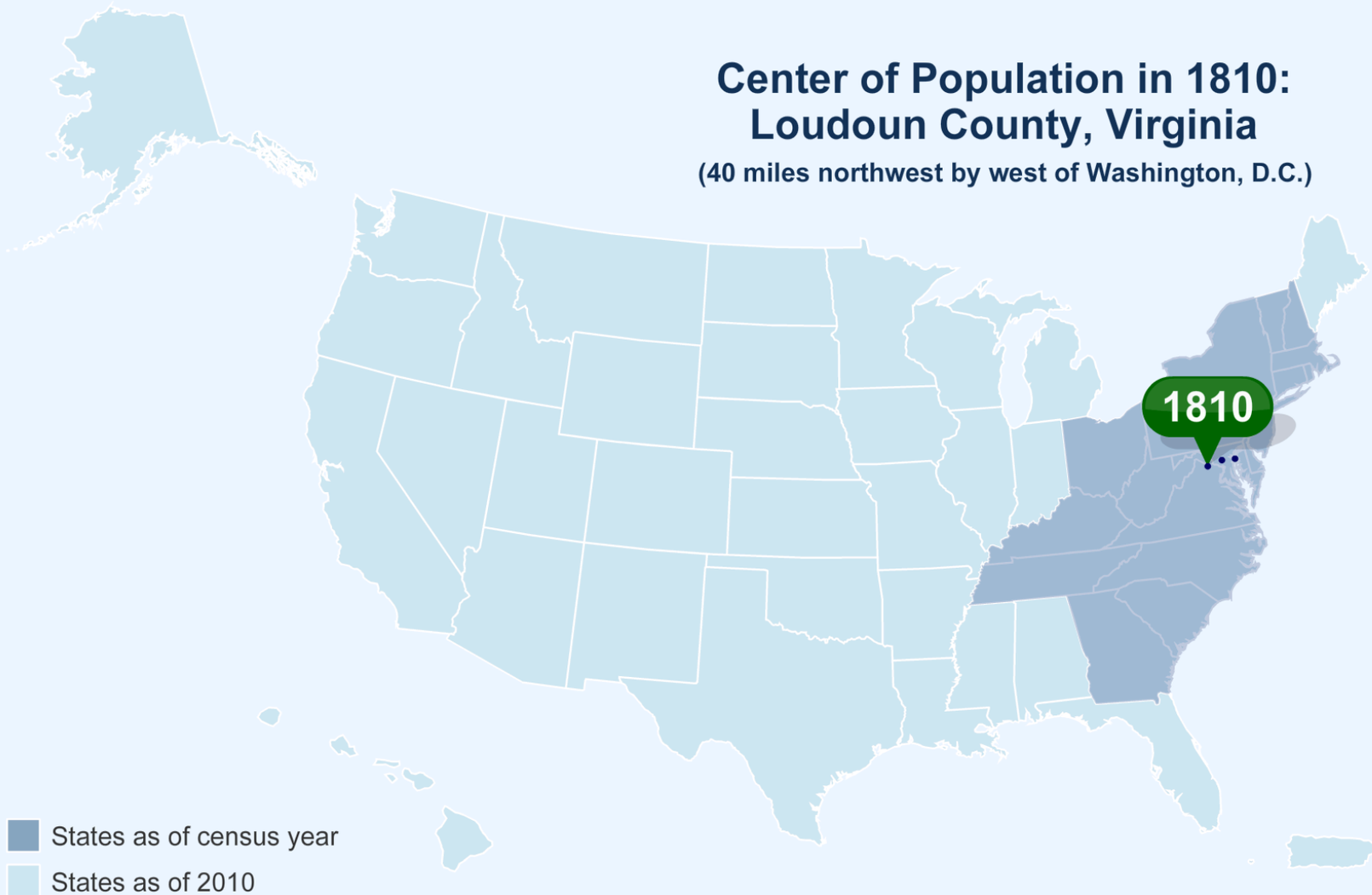
HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.

Center of Population in 1810: Loudoun County, Virginia

(40 miles northwest by west of Washington, D.C.)



Loudoun County, Virginia
LAT: 39.19167 LONG: 77.62000

1790 2010

United States Census Bureau

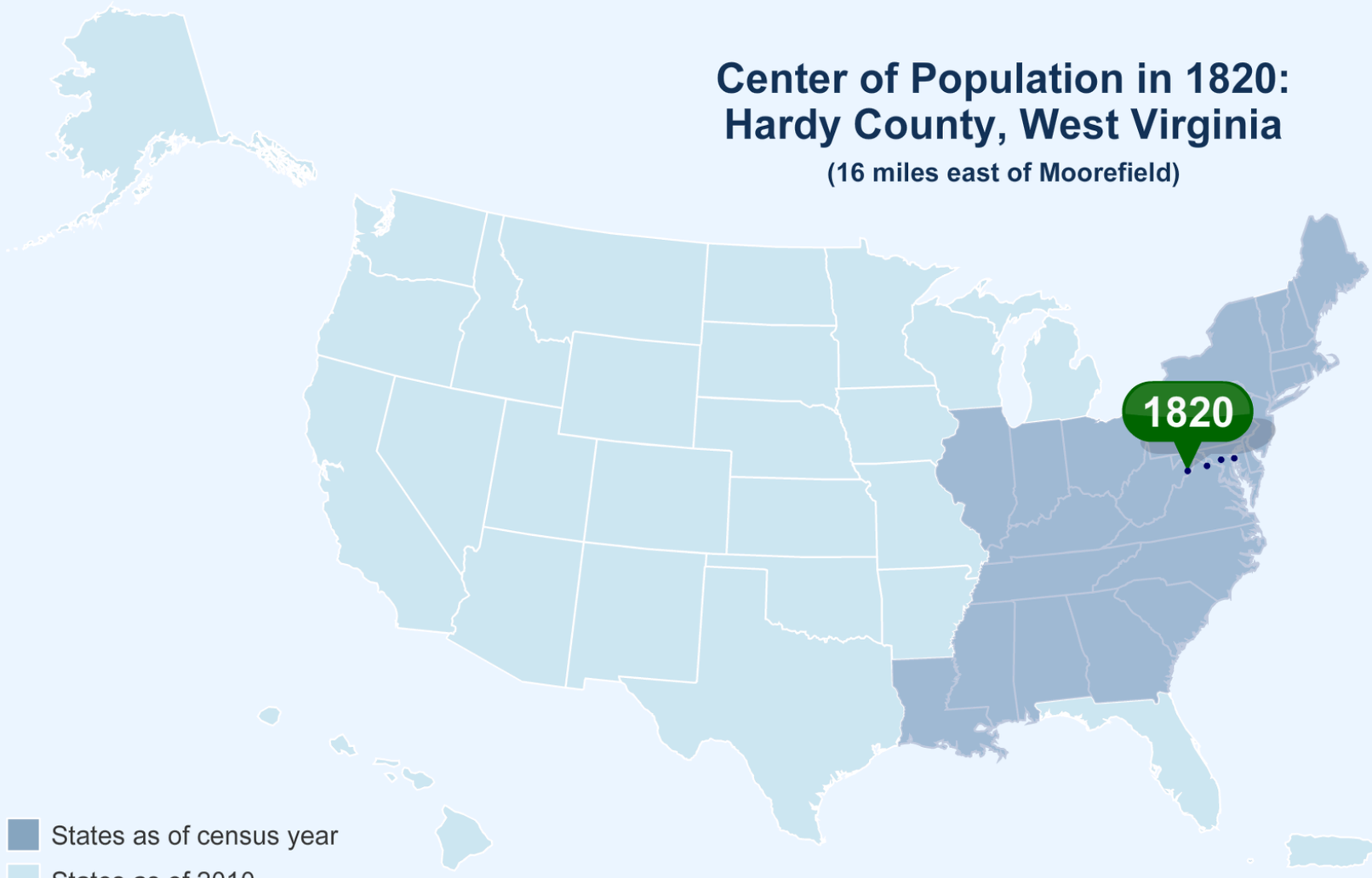
HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.

Center of Population in 1820: Hardy County, West Virginia

(16 miles east of Moorefield)



Hardy County, West Virginia
LAT: 39.09500 LONG: 78.55000

United States Census Bureau

1790 2010

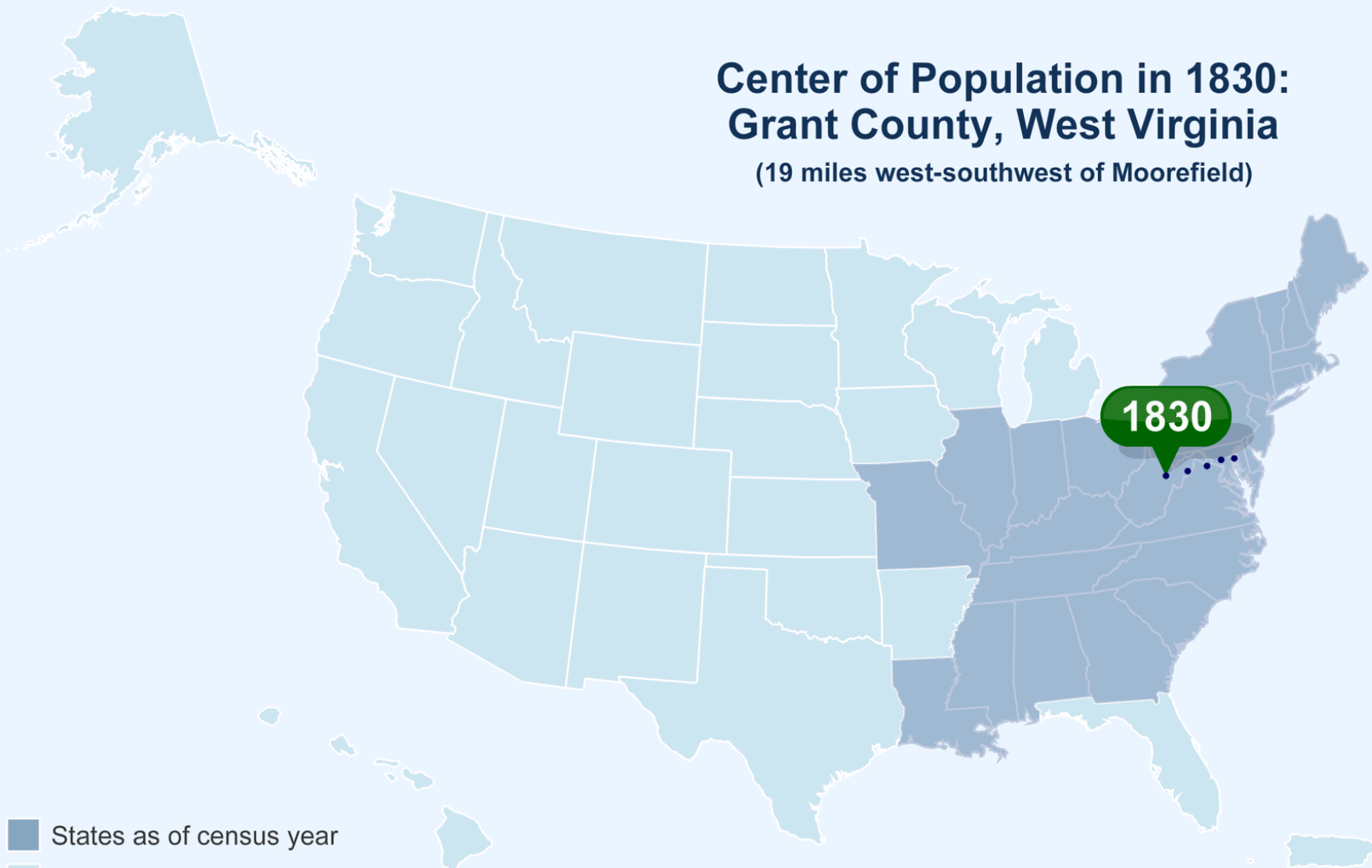
HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.

Center of Population in 1830: Grant County, West Virginia

(19 miles west-southwest of Moorefield)



Grant County, West Virginia
LAT: 38.96500 LONG: 79.28167

United States Census Bureau

1790 2010

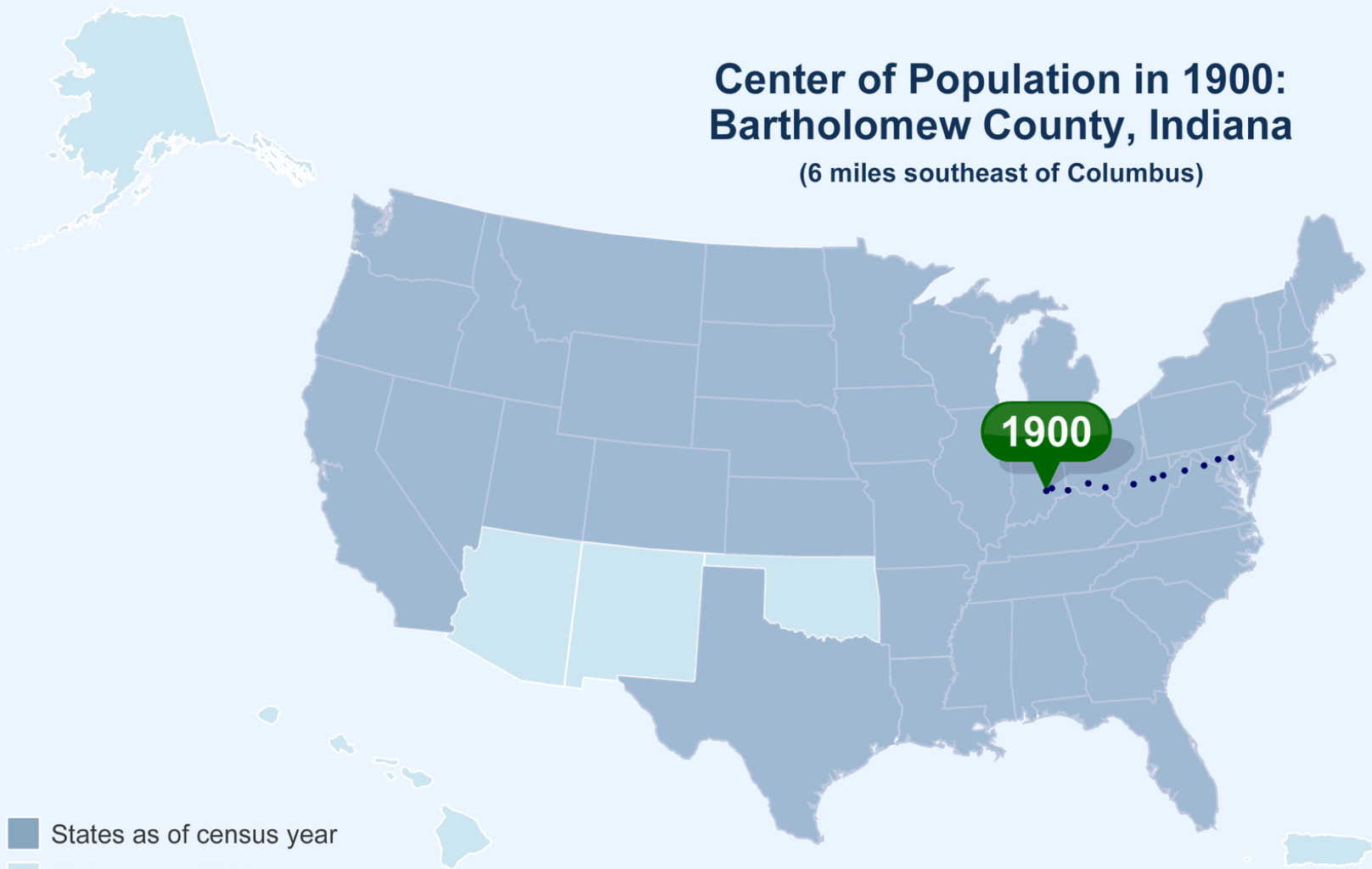
HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.

Center of Population in 1900: Bartholomew County, Indiana

(6 miles southeast of Columbus)



Bartholomew County, Indiana
LAT: 39.16000 LONG: 85.81500

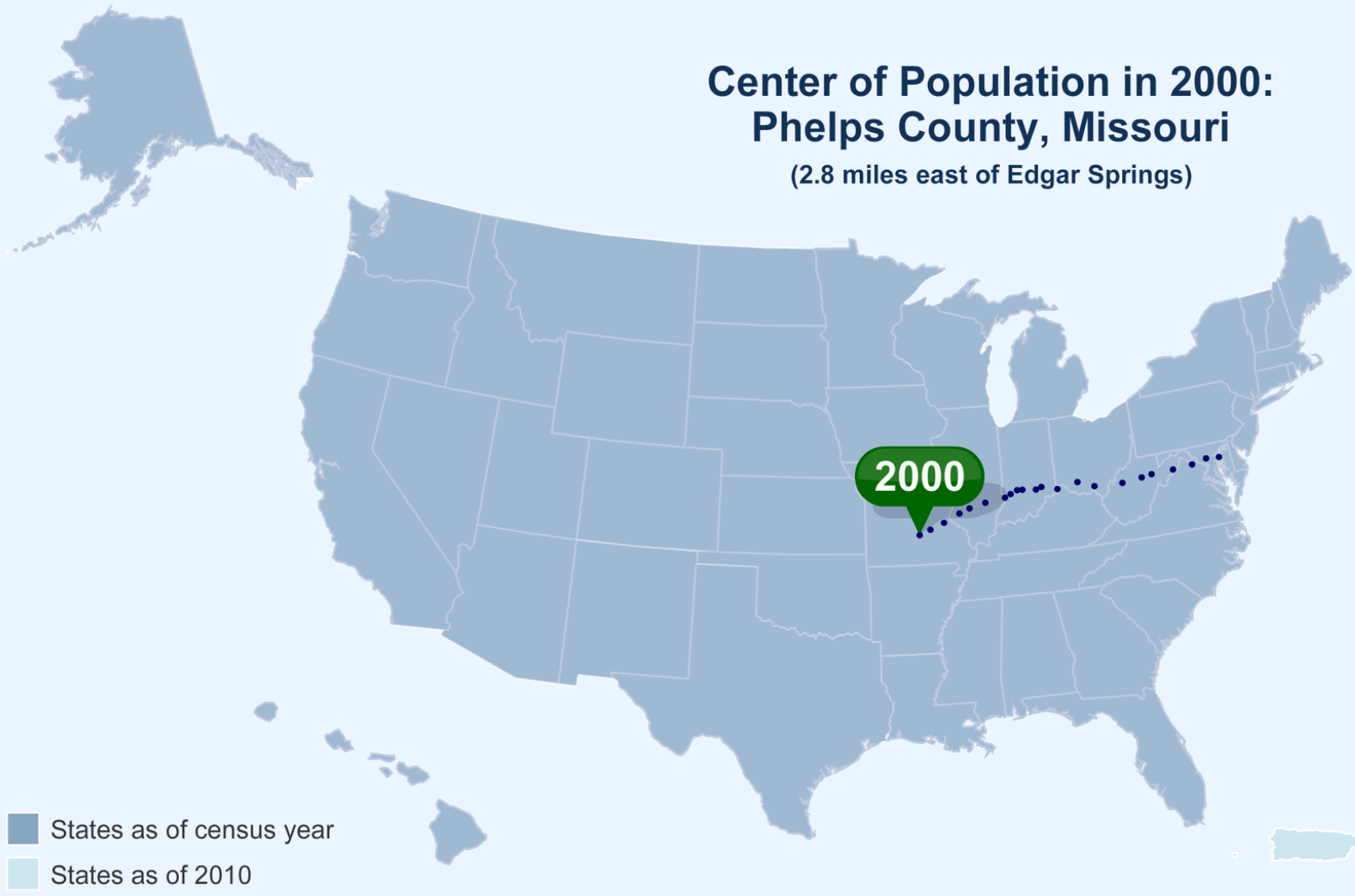
United States Census Bureau

1790 2010

HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.



Center of Population in 2000: Phelps County, Missouri

(2.8 miles east of Edgar Springs)

Phelps County, Missouri

LAT: 37.69699 LONG: 91.80957

In 2000, the new center of population was more than 1,000 miles from the first center in 1790.

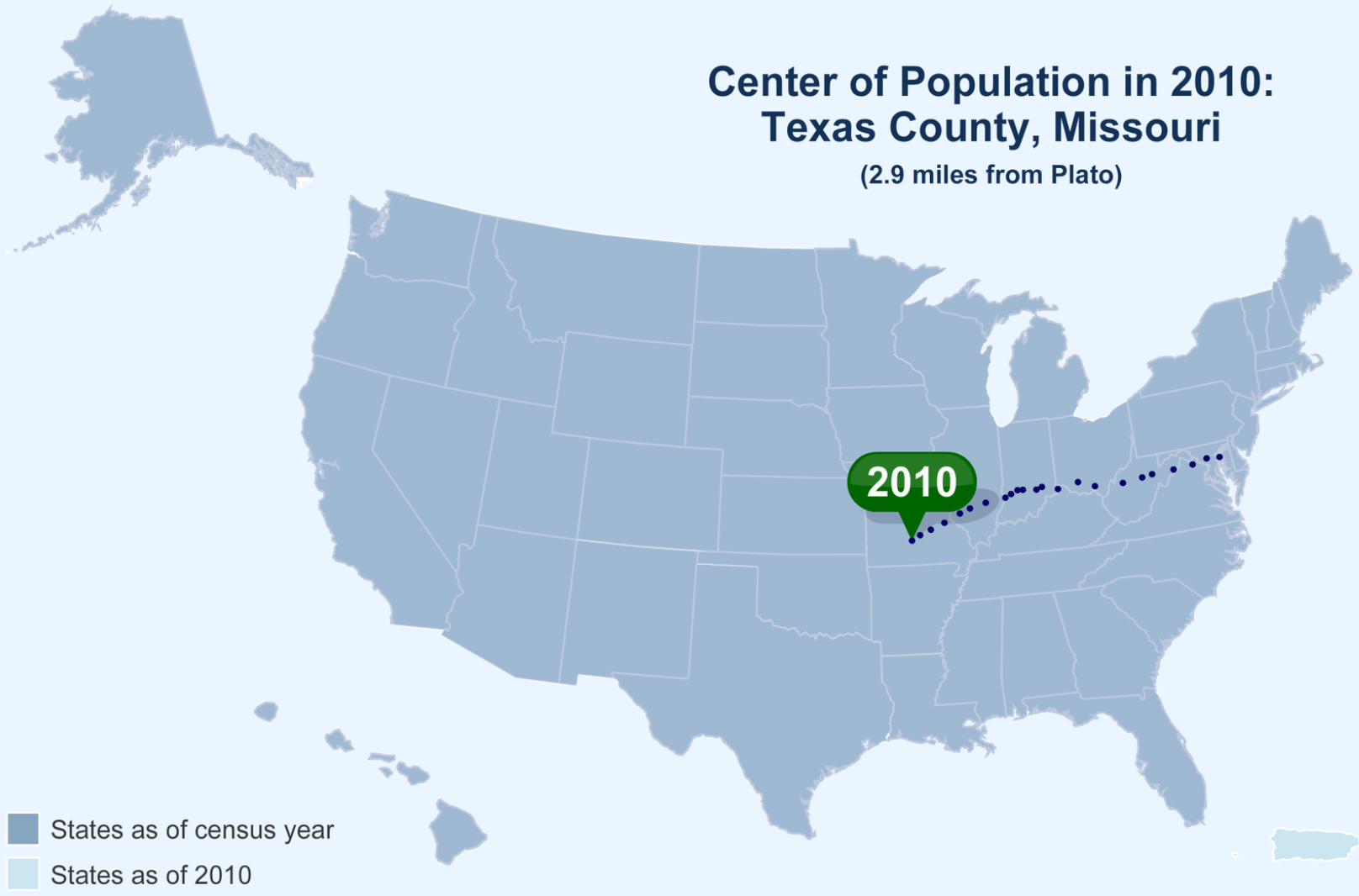
United States Census Bureau

1790 2010

HIDE FULL SCREEN

Center of Population, 1790 – 2010

The mean center of population is determined as the place where an imaginary, flat, weightless and rigid map of the United States would balance perfectly if all residents were of identical weight. View all 2010 Census data on American Factfinder.



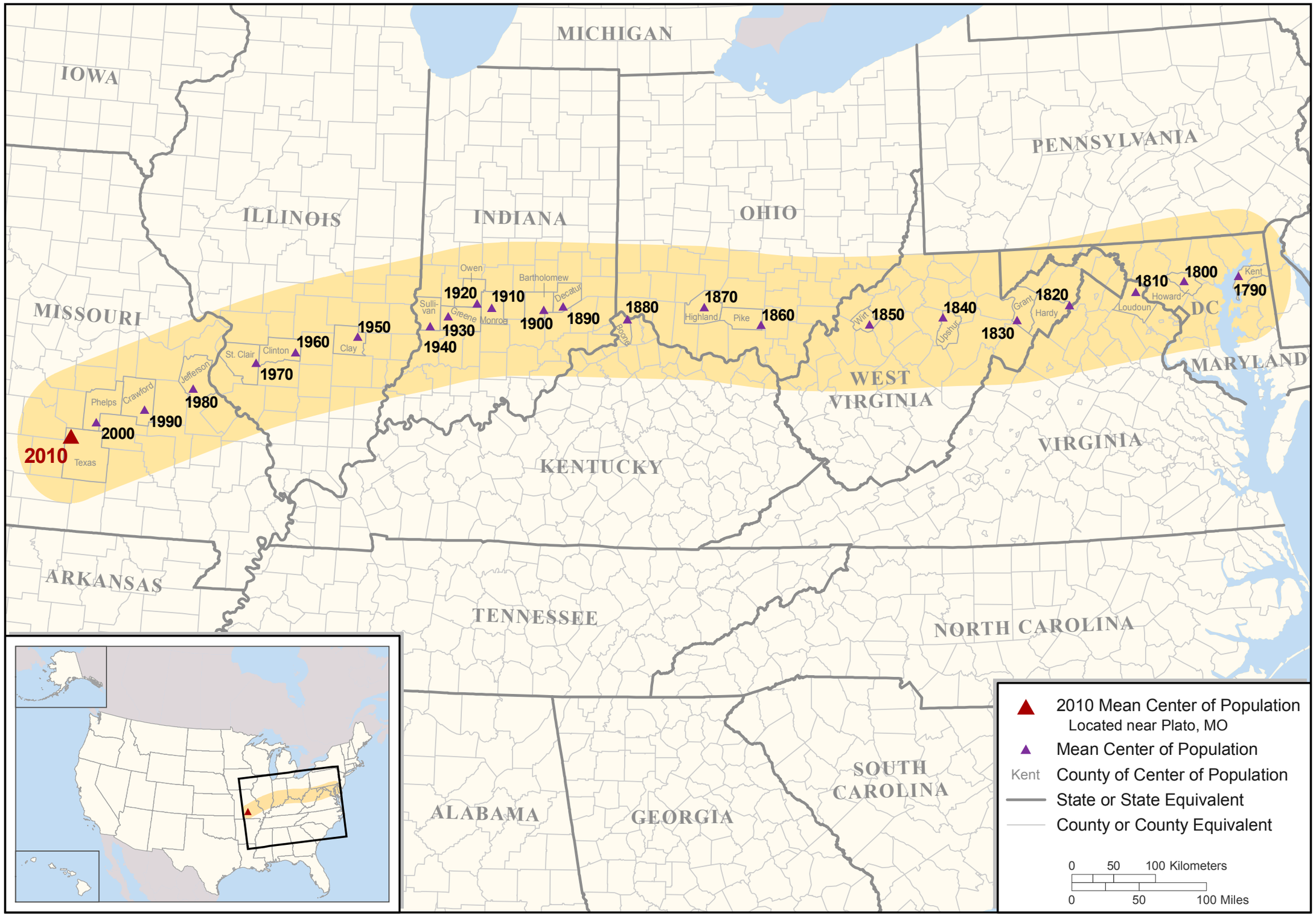
Texas County, Missouri

LAT: 37.51753 LONG: 92.17310

The center moved in a more southerly direction than in previous decades. The distance moved—23.4 miles—is the shortest distance since 1970. This southerly drift and shorter distance can be attributed to a strong pull on the center by population growth in the Southeast—Georgia, Florida, and the Carolinas—as well as growth in Texas.

1790 2010

United States Census Bureau



U.S. DEPARTMENT OF COMMERCE Economics and Statistics Administration U.S. Census Bureau

Prepared by the Geography Division

Adolphe Quetelet (1796-1874)

Quetelet

Adolphe Quetelet is considered a founder of the social sciences — He was a “tireless promoter of statistical data collection” with one of his main analytical tools being the normal distribution (the Gaussian distribution or the bell curve)

> [After graduation with a doctorate from the University of Gent, Quetelet was] recruited to teach mathematics at the Athenée in Brussels... and on a trip to Paris he met Joseph Fourier (1768–1830), Siméon Poisson (1781–1840) and Pierre Laplace (1749–1827), became impassioned by the subject of probability, and went on to put it to practical use in the study of the human body, **a subject in which he had developed an early interest as a painter and sculptor.** In doing so, he became the first to develop height and weight tables to study their relationships, and a pioneer in the application of mathematical analysis to the study of man...
>
> Quetelet's interest in the subject was kindled when on his return from Paris, he got involved in a national population census of the Netherlands and argued that a random sample from a representative diversified group could be used to estimate the total population. **His subsequent conceptual evolution in the study of man evolved from the study of averages (physical characteristics), to rates (birth, marriage, growth) and ultimately distributions (around an average, over time, between regions and countries).** The latter was the basis of one of his contributions to statistics; the demonstration that the normal Gaussian distribution, typical throughout nature, applied equally to physical attributes of humans, including body parts, derived from large-scale population studies. **After that, he saw bell-shaped curves everywhere he looked, including in social phenomena and the variables that determine character and aptitudes...**

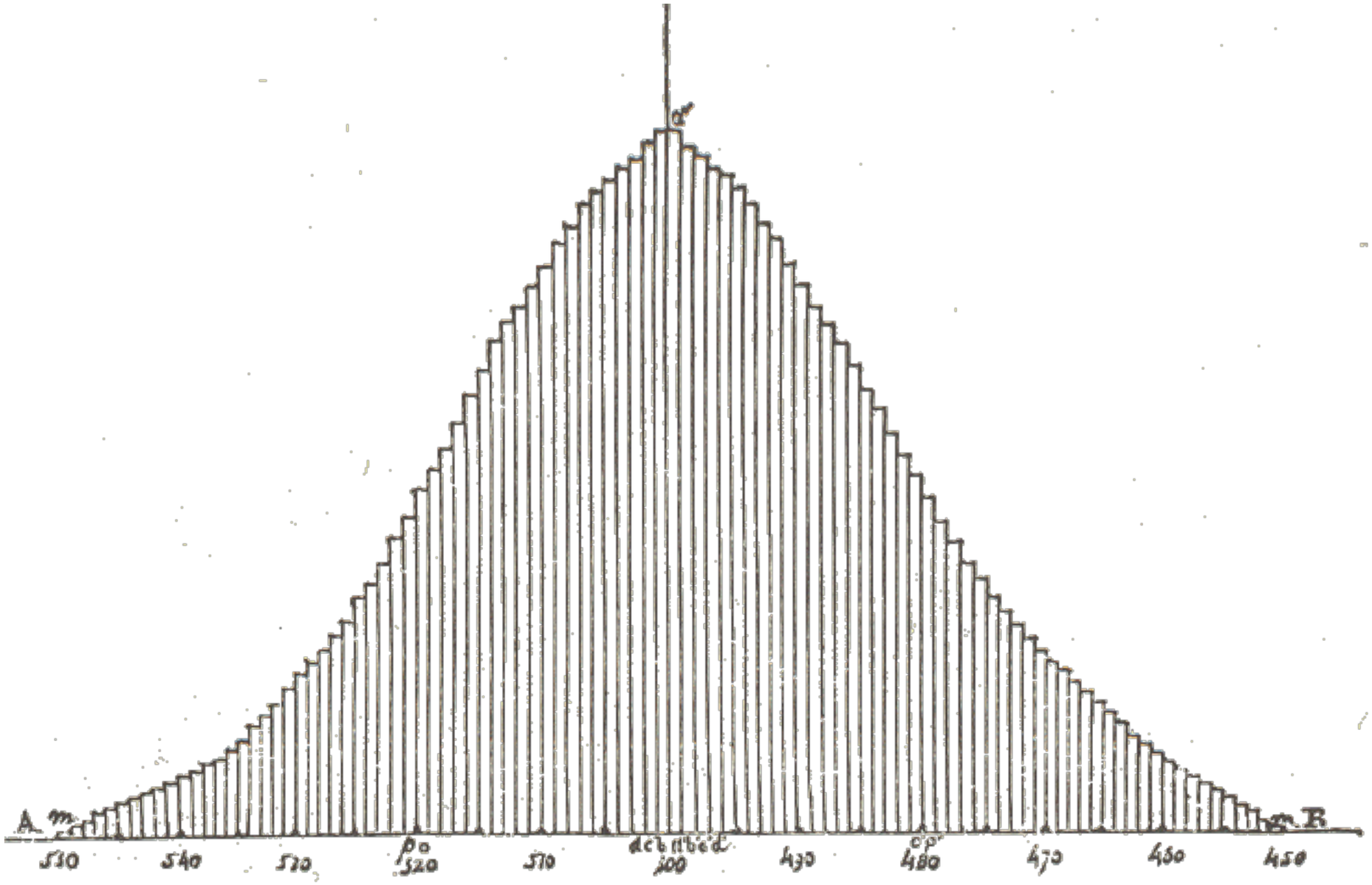
Adolphe Quetelet (1796–1874)—the average man and indices of obesity
`http://ndt.oxfordjournals.org/content/23/1/47.full`

The error distribution

Quetelet made extensive use of the bell curve or normal distribution in his work — He saw that it described the distribution of physical characteristics (not to mention “moral” characteristics) of people in the various data sets he examined

Interestingly, his understanding of the normal came from astronomy where it was applied as an error distribution -- Broadly, if you add up a number small mistakes, the resulting distribution has a bell shape

Here’s Quetelet’s illustration of the normal...

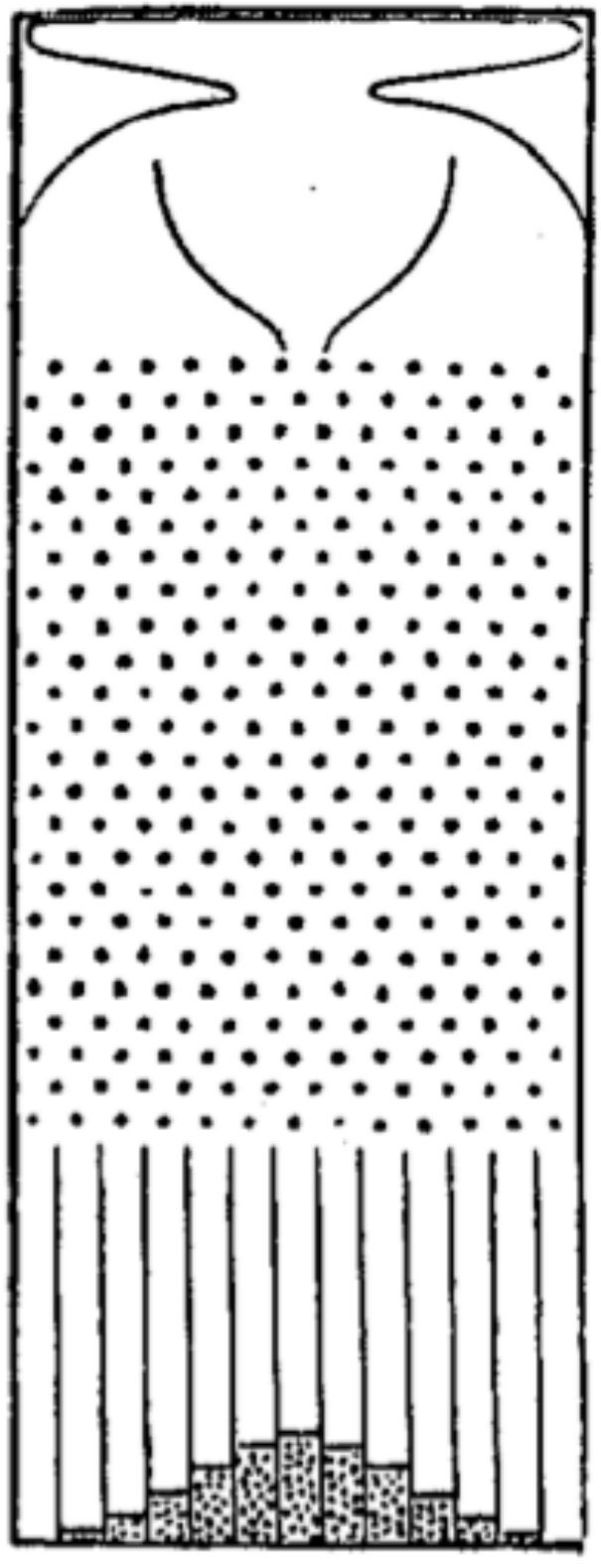

A
B
510
540
520
520
570
430
480
470
450
450

Galton (a look ahead)

In 1873, Francis Galton had a machine built which he christened the quincunx -- The name comes from the similarity of the pin pattern to the arrangement of fruit trees in English agriculture (quincunxial because it was based on a square of four trees with a fifth in the center)

The machine was originally devised to illustrate the central limit theorem and how a number of independent events might add up to produce a normal distribution -- Lead shot were dropped at the top of the machine and piled up according to the “binomial” coefficients at the bottom

This left-or-right mechanism makes physical the idea of adding up a number of small errors, some +1 some -1, all independent -- In the end, their accumulated effect is a bell curve



PEOPLE'S EDITION.

A TREATISE ON MAN

AND THE DEVELOPMENT OF HIS FACULTIES.

By M. A. QUETELET,

PERPETUAL SECRETARY OF THE ROYAL ACADEMY OF BRUSSELS, CORRESPONDING MEMBER OF THE INSTITUTE OF FRANCE, ETC.

NOW FIRST TRANSLATED INTO ENGLISH.

Stable frequencies

Quetelet saw stability in tables like the one on the right — This kind of stability felt law like to him, making an analogy with the laws of physics

He saw this in births and deaths and marriages, in the ratio of boys to girls born, and even in the crime rate...

Periods.	Average Number of Marriages.	Average Number of Baptisms.	Average Number of Deaths.	Baptisms. to one Marriage.
1693 to 1697, -	5,747	19,715	14,862	3·43
1698 to 1702, - -	9,070	24,112	14,474	3·97
1703 to 1708, -	6,082	26,896	16,430	4·42
1709 to 1711, - -	5,835	18,833	85,955	3·23
1712 to 1716, -	4,965	21,603	11,948	4·35
1717 to 1721, - -	4,324	21,396	12,039	4·95
1722 to 1726, -	4,719	21,452	12,863	4·55
1727 to 1731, - -	4,808	20,559	12,825	4·28
1732 to 1735, -	5,424	22,692	15,475	4·18
1736 to 1737, - -	5,522	20,304	25,425	3·69
1738 to 1742, -	5,582	22,099	15,255	3·96
1743 to 1746, - -	5,469	25,275	15,117	4·62
1747 to 1751, -	6,423	28,235	17,272	4·40
1752 to 1756, - -	5,599	28,392	19,154	5·07
1816 to 1823, -	109,237	480,632	307,113	5·40 *
1827, - -	106,270	524,062	368,578	4·93 †

States and Provinces.	Males to 100 Females.
Russia,	108·91
The province of Milan,	107·61
Mecklenburg,	107·07
France,	106·59
Belgium and Holland,	106·44
Brandenburg and Pomerania,	106·27
Kingdom of the Two Sicilies,	106·18
Austrian Monarchy,	106·10
Silesia and Saxony,	106·05
Prussian States (*en masse*),	105·94
Westphalia and Grand Dutchy of the Rhine,	105·86
Kingdom of Wurtemburg,	105·69
Eastern Prussia and Dutchy of Posen,	105·66
Kingdom of Bohemia,	105·38
Great Britain,	104·75
Sweden,	104·62
Average for Europe,	106.

Interpretation: L’homme Moyen

That the bell curve was used as a kind of error distribution is important because it frames how Quetelet interprets the statistical work he’s done -- Noticing that human physical and “moral” characteristics have bell-shaped distributions led him to a very particular interpretation of the mean

In short, Quetelet interprets the average as a kind of ideal or a state of perfection...

Interpretation: L’homme Moyen

It is a statistical conception of the universe possessing qualities of poetic and artistic beauty. **Everything is to be viewed s varying about a normal state in a a manner to be accurately described by beautiful bell-shaped curves of perfect symmetry but of varying amplitude.** Thus it is that the individual varies about his normal self; thus members of a group vary about their average; thus the men of a nation, viewed as individuals, vary about the average man of the nation; thus a nation varies about its normal state; and finally, inasmuch as the qualities of the average man change from time to time and place to place in obedience to general causes, to follow the course of the average man in the whole series of nations would give us, in Quetelet’s view, the principles of a social physics, the true mechanics of human history.

Adolphe Quetelet as Statistician, by Frank Hamilton Hankins

It's been said that around 1830 there was a paradigm shift from the enlightened or rational man to the average man. Could you elaborate a little on the history of the term "the average man" and what is at stake in this shift—from the man of reason to man as quantifiable and subject to quantification?

The enlightened man was also quantifiable, in fact was more obviously quantifiable, than the average man, because enlightenment meant rationality, and rational decisions should reflect a kind of mathematical ordering. The rise of statistics in the early 19th century attests to a loss of faith in the power of individual reason, or at least to a new anxiety about the masses. The average man embodied a form of political activity that could no longer be understood as the rationality of probabilistic calculation. In place of individualistic rationality, the age of mass society gave us a collective order of statistical averages. The term "average man" was invented by the Belgian Adolphe Quetelet in the 1830s to support his new quantitative science, a science of the whole population. The average was a way of embracing the whole of society and of giving it an individualized representation. Quetelet's idea was to found a science of society on the basis of this emblematic figure of the average man, which was taken as typical or representative.

But Quetelet often refers to the average man as a fiction, as a construction. How does "the average man" relate to the life of the living individual?

It is a fiction in the sense that no real person will have all the characteristics of the average man. In some ways you might think of the average man as the basis for a quantitative model, a person whom the ambitious social scientist uses to abstract away from concrete individuality. This was a fiction anchored in reality, and Quetelet even argued that statistics could help novelists and poets delineate characters who would be appropriate for their age. And there are ways in which he made the average man more real than actual, flesh-and-blood human beings. He portrayed the average man as the center of the symmetrical distribution that we know as the normal or bell curve. For Quetelet, this role was comparable to that of an average in the physical science of astronomy. Here the average stands for the true value of something that you're trying to measure, and the spread of measures around the mean is just a distribution of errors. From the standpoint of science, the statistical, emblematic figure of the average man more closely represents society as a whole than any flesh-and-blood individual ever can. Quetelet's average, though fictional, becomes more real than the actual people who make up the distribution.

Is the average man the ideal? Should one aspire to be average?

It's an interesting question because now we tend to see the average as mediocre rather than exemplary. Quetelet wasn't totally unaware of this aspect. Yet there are a couple of ways he managed to idealize the average man nonetheless. One was by celebrating the stability of the average. While any individual will have highs and lows, being sometimes desperately in love, sometimes obsessed with philosophy or poetry or natural science, wavering perhaps between radicalism and conservatism in politics, the average smoothes over these fluctuations and is therefore more stable. Quetelet was acutely conscious of living in a revolutionary age, not just in the wake of the French Revolution of 1789, but as one who saw his scientific ambitions gravely threatened by the Belgian Revolution in 1830. One reason for idealizing the average is just its stability, its equanimity and its calmness, when the world is raging with passion. A second perspective, which he drew from Aristotle, is to see extremes as vice and the average as virtuous balance. Instead of thinking of a continuum from low ability to high ability, he imagined a spectrum, say, of passion or of politics, where both extremes are dangerous and the average means stability, the foundation of steady progress. I would say, finally, that Quetelet's average man fits into a system of social teleology, representing not any concrete production of nature, but nature's intention.

Life on the Bell Curve: An Interview with Theodore Porter
http://www.cabinetmagazine.org/issues/15/fleming2.php

The bell curve

Quetelet's average man is the center of the bell curve -- Here he applies the reasoning to physical characteristics, but he reasoned similarly with a range of "moral" data as well

> Suppose... that one wished to make a thousand copies of a statue, say the Gladiator. Like astronomical observations of a single object, these copies would be subject to a variety of errors -- in measuring the various dimensions, in workmanship and so on. The independent errors are like terms of a binomial, and combine in a characteristic fashion. Hence the variation among the copies would be governed by a profound regularity, the error law or normal curve, with the dimensions of the original Gladiator at the mean. But this is an impossible experiment.
>
> How did Quetelet know what the result would be? "I shall perhaps astonish you very much by stating that the experiment has been already made. Yes, surely, more than a thousand copies have been measured of a statue, which I do not assert to be that of the Gladiator, but which in all cases differences little from it. These copies were living ones... " (1846, p. 136)
>
> from The empire of chance

Thinking of variation through errors, Quetelet considered the average man as a kind of ideal -- He wrote **"an individual who epitomized in himself, at a given time, all the qualities of the average man, would represent at once all the greatness, beauty and goodness of that being"**

STATEMENT of the Sizes of Men in different Counties of Scotland, taken from the Local Militia.

1st Argyll Regiment of Local Militia, of 686 Men.

Heights.				Number of Men.	Thickness round the Chest, in Inches.															
					33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Ft.	Inch.	Ft.	Inch.																	
5	4 &	5	5	79			1	3	13	20	20	11	8	1	2					
5	6	5	7	221		1	0	11	16	45	49	46	36	14	3					
5	8	5	9	245				2	6	17	42	57	55	36	20	10				
5	10	5	11	107						5	14	19	23	23	14	7	2			
6	0	6	1	34							1	3	8	5	7	8	1	0	1	

2d Argyll Regiment of Local Militia, of 736 Men.

					33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
5	4 &	5	5	77			1	3	12	21	17	13	8	2						
5	6	5	7	288			2	7	37	48	67	59	46	19	3					
5	8	5	9	234				1	10	21	30	47	58	48	12	6	1			
5	10	5	11	109						1	6	19	26	27	16	7	6	1		
6	0	6	1	28									6	10	3	6	3			

Annan and Eskdale Regiment of Local Militia, of 493 Men.

Height.				Number of Men.	Thickness round the Chest, in Inches.															
					33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Ft.	Inch.	Ft.	Inch.																	
5	4 &	5	5	58			2		4	14	9	14	10	3	1	1				
5	6	5	7	193			2	3	8	30	35	37	33	26	14	4	1			
5	8	5	9	141				1	1	12	23	30	28	28	13	3	1		1	
5	10	5	11	78						2	11	7	20	17	13	7	1			
6	0	6	1	23							1	2	9	5	3	3				

2d Fifeshire Regiment of Local Militia, of 621 Men.

					33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
5	4 &	5	5	121	1	1	6	17	17	25	20	22	9	2	0	1				
5	6	5	7	218		1	3	5	15	36	61	40	29	23	2	2				
5	8	5	9	213				2	3	17	40	34	35	40	27	12	3			
5	10	5	11	36				1	1	4	7	6	5	6	2	3	0	2		
6	0	6	1	33						3	2	8	5	8	3	3	1			

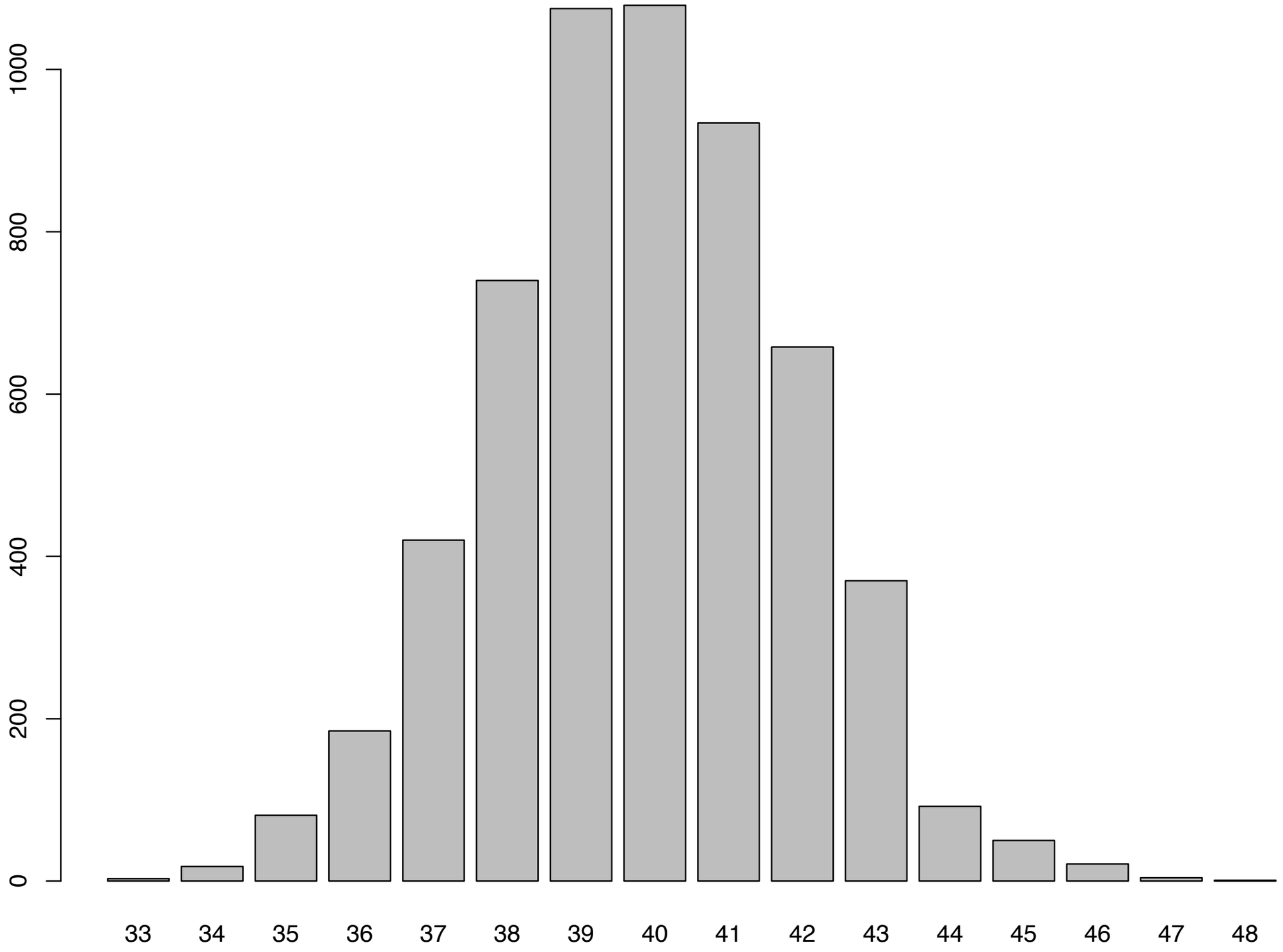
2d Edinburgh Regiment of Local Militia, of 506 Men.

					33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
5	4 &	5	5	37				1	5	12	11	6	2							
5	6	5	7	173	1	4	12	18	32	46	32	22	5	1						
5	8	5	9	192		1	4	9	15	45	41	36	22	11	3	3	2			
5	10	5	11	76		1	2	3	5	10	15	16	9	5	6	2	1	1		
6	0	6	1	28					4	1	7	7	4	1	1	2	1			

Peebles-shire Regiment of Local Militia, of 224 Men.

					33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
5	4 &	5	5	10			2	1	1	2	1	3								
5	6	5	7	69		2	2	9	5	14	21	10	2	4						
5	8	5	9	90				3	5	8	21	17	14	14	3	3	2			
5	10	5	11	35					2	4	7	9	7	2	3	0	1			
6	0	6	1	20					2	1	1	3	4	5	2	1	0	1		

MESURES de la POITRINE.	NOMBRE d'hommes.	NOMBRE PROPORTIONNEL.	PROBABILITÉ d'après L'OBSERVATION.	RANG dans LA TABLE.	RANG d'après le CALCUL.	PROBABILITÉ d'après LA TABLE.	NOMBRE D'OBSERVATIONS calculé.
Pouces.							
33	3	5	0,5000			0,5000	7
34	18	31	0,4905	52	50	0,4903	29
35	81	141	0,4964	42,5	42,5	0,4964	110
36	185	322	0,4823	33,5	34,5	0,4854	323
37	420	732	0,4501	26,0	26,5	0,4531	732
38	749	1305	0,3769	18,0	18,5	0,3799	1333
39	1075	1867	0,2464	10,5	10,5	0,2466	1838
			0.0597	2,5	2,5	0,0628	
40	1079	1882	0,1285	5,5	5,5	0,1359	1987
41	934	1628	0,2913	13	13,5	0,3034	1675
42	658	1148	0,4061	21	21,5	0,4130	1096
43	370	645	0,4706	30	29,5	0,4690	560
44	92	160	0,4866	35	37,5	0,4911	221
45	50	87	0,4953	41	45,5	0,4980	69
46	21	38	0,4991	49,5	53,5	0,4996	16
47	4	7	0,4998	56	61,8	0,4999	3
48	1	2	0,5000			0,5000	1
	5738	1,0000					1,0000


0
200
400
600
800
1000
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48

LETTER XX.

TO DISCOVER WHETHER THE ARITHMETICAL MEAN IS THE TRUE MEAN-TYPE OF THE HUMAN SIZE.

The Gladiator is certainly one of the most beautiful works of ancient sculpture. It is with reason that artists have studied its free and noble forms, and have often measured the principal dimensions of the head and of the body to obtain its proportions and its harmony.

To measure a statue is not so easy an operation as might at first appear, particularly if it be desired to obtain very precise results. In measuring ten times in succession the circumference of the chest, we are not sure of finding two results identically the same. It almost always happens that the values obtained are more or less distant from that sought; and I even suppose the most favourable circumstances those where there is no tendency to make the measurements either too small or too great.

If we had the courage to recommence a thousand times, we should in the end have a series of numbers differing from one another, according to the degree of precision exercised in their collection. The mean of all these numbers would certainly differ very little from the true value. Moreover, in classing all the measurements in order of magnitude, we should be not a little astonished to find the groups succeed one another with the greatest regularity. The measurements which differed the least from the general mean would compose the largest group; and the other groups would be so much the smaller as they contained measurements differing the more from this same mean. If the succession of groups were traced by a line, this line would be the curve of possibility: this result might in fact have been foreseen. So that unskilfulness, or chance (if to gratify our self-love we prefer this

Reasoning today

While we might find the elevation of the average to an ideal quaint, we routinely reason from mathematical or statistical models — If a model reproduces patterns in nature, we are tempted to make the assumption that nature must follow the same rules as our simple model

Quetelet saw the normal as an accumulation of errors, meaning we are all flawed copies of some ideal average — Today, this kind of reasoning appears in the new “normal”...

Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

MORE "NORMAL" THAN NORMAL: SCALING DISTRIBUTIONS AND COMPLEX SYSTEMS

Walter Willinger

AT&T Labs-Research
180 Park Ave., Room B207
Florham Park, NJ 07932, U.S.A.

David Alderson, John C. Doyle, Lun Li

Department of Control and Dynamical Systems
California Institute of Technology
Pasadena, CA 91125, U.S.A.

ABSTRACT

One feature of many naturally occurring or engineered complex systems is tremendous variability in event sizes. To account for it, the behavior of these systems is often described using power law relationships or scaling distributions, which tend to be viewed as "exotic" because of their unusual properties (e.g., infinite moments). An alternate view is based on mathematical, statistical, and data-analytic arguments and suggests that scaling distributions should be viewed as "more normal than Normal". In support of this latter view that has been advocated by Mandelbrot for the last 40 years, we review in this paper some relevant results from probability theory and illustrate a powerful statistical approach for deciding whether the variability associated with observed event sizes is consistent with an underlying Gaussian-type (finite variance) or scaling-type (infinite variance) distribution. We contrast this approach with traditional model fitting techniques and discuss its implications for future modeling of complex systems.

1 INTRODUCTION

A common research theme in the study of complex systems is the pursuit of universal properties that transcend specific system details. It is in exactly what those properties are, and the theories to explain and exploit them, where sharp differences arise. One aspect of many complex systems that has received considerable attention is a tendency toward tremendous variability in event sizes, such that they can be reasonably represented by a so-called "power law" relationship. That is, the cumulative probability $P(X > l)$ of observing events greater than a given size l is given by $P(X > l) \approx l^{-\alpha}$ and manifests itself as a straight line of slope $-\alpha$ in a $\log(P)$ vs. $\log(l)$ plot (for large values of l, and for $\alpha > 0$). For example, consider the relative sizes of the largest disaster events during the 20^{th} Century (Figure 1). Simple inspection of the data shows a striking relationship between the size and frequency of large events, namely that they are reasonably approximated by a power law having $\alpha = 1$. Power law relationships have been observed within many naturally occurring and man made systems, including species within plant genera (Yule 1925); mutants in old bacterial populations (Luria and Delbrück 1943); a number of applications in the social sciences (Simon 1955), including economics (income distributions, city populations) and linguistics (word frequencies) (Mandelbrot 1997); forest fires (Malamud, Morein, and Turcotte 1998); Internet traffic (flow sizes, file sizes, Web documents) (Crovella and Bestavros 1997) and Internet topology (node degrees of physical or virtual connectivity graphs) (Faloutsos, Faloutsos, and Faloutsos 1999); and metabolic networks (Barabasi and Oltavi 2004). That such a diversity of systems exhibits similar scaling features has prompted many researchers to ask whether or not there are universal drivers of these phenomena.

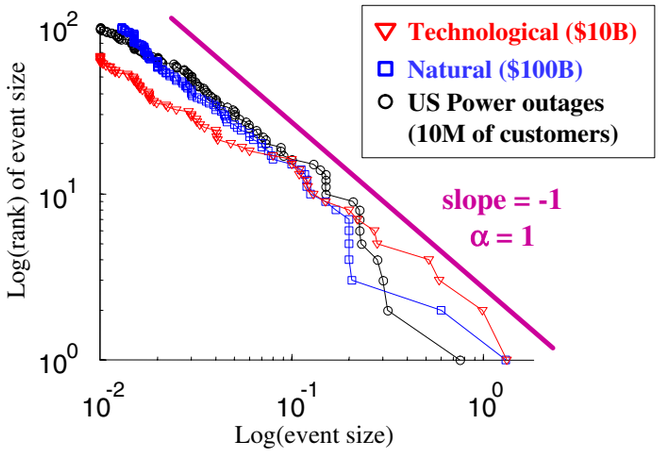


Figure 1: Log-log Plot of Event Size Versus Event Rank (100 Largest Disasters of the 20^{th} Century)

Given the discovery of such "emergent" properties of complex systems and the ability to describe them with power law-type relationships or scaling distributions, a fundamental issue underlying the attempts by complex systems researchers to understand and explain these highly variable event sizes has been the extent to which such high variability should be viewed as "exotic" (in the sense of

the distributions of the X_i's need not follow the scaling distribution exactly. In fact, a result from extreme value theory (see for example Resnick 1987) identifies the invariant distributions as the Frechet distributions and characterizes the distributions of the X_i's that are in the domain of attraction of the Frechet distribution. The Frechet distribution is defined by $P[M > x] = 1 - \exp(-x^{-\alpha}),\ x > 0$ and is clearly scaling for large x. As a consequence, the individual X_i's must be so close to being scaling distributions as to be scaling for all practical purposes. In this sense, scaling distributions are the only distributions that are invariant under the transformation of maximization. In particular, Gaussian distributions lack this invariance property.

2.2.3 Weighted Mixtures

As before, assume that $X_1, X_2, \ldots, X_n$ are n independent random variables with scaling distribution functions F_i, all with the same tail index parameter α, but possibly with different scale coefficients $c_i > 0$. Consider the *weighted mixture* W_n of the X_i's, and denote by p_i the probability that $W_n = X_i$. It is easy to show that

$$P[W_n > x] = \sum p_i P[X_i > x] \approx c_{W_n} x^{-\alpha},$$

where $c_{W_n} = \sum p_i c_i$ is the *weighted average* of the separate scale coefficients c_i. Thus, the distribution of the weighted mixture of scaling distributions is also scaling, with the same tail index α, but a different scale coefficient than the individual X_n's.

Mathematically, the converse (i.e., W_n is scaling only if the X_i's are scaling) holds only to a first approximation. In fact, in the limit as $n \to \infty$, the invariant "distributions" for W are of the form $P[W > x] = cx^{-\alpha}, x \geq 0$, which are improper distribution functions because they yield an infinite total probability. However, for all practical purposes, the X_i's are typically restricted by some relation of the form $0 < a \leq x$ which results in perfectly well-defined (conditional) distribution functions of the scaling type. With these qualifications, scaling distributions are the only distributions that are invariant under the transformation

2.3 Scaling Distributions: More Normal than Normal

Aggregation, mixture, maximization, and marginalization are transformations that occur frequently in natural and engineered systems and are inherently part of many measured observations that are collected about them. For example, aggregate incomes are more widely collected and reported than each type of income separately; the flow or file/document sizes transmitted across the Internet and observed at a particular link within the network are naturally a mixture of distributions of the different file/document sizes residing on the various file/Web servers; for historical data such as natural or technological disasters (i.e., droughts, floods, earthquakes, hurricanes, blackouts, nuclear accidents), the fully recorded and commonly available observations reflect a maximizing choice and correspond to the exceptional (i.e., largest, or most catastrophic) events; and the marginalization transformation is relevant for dealing with a variety of multidimensional economic quantities. In turn, these invariance properties suggest that the presence of scaling distributions in data obtained from complex natural or engineered systems should be considered the norm rather than the exception and should not require "special" explanations.

However, there is an implicit tradeoff between Gaussians being the norm for low variability data and scaling distributions being the norm for high variability data. In the former case, the (traditional) CLT imposes only minimal conditions on the distribution of the individual constituent (i.e., finite variance), but as a result, invariance properties can only be obtained for aggregation and marginalization. In contrast, for high variability data, the relevant CLT requires the (right) tail of the distribution of the individual constituents to decay at a certain rate, and as a result of this more restrictive assumption, the individual constituents are not only invariant under aggregation and marginalization, but also under maximization and weighted mixtures. The pragmatic approach to dealing with high variability data advocated in this paper then consists of viewing Gaussians as the natural null hypothesis for low variability data, where

Complexity and robustness

J. M. Carlson*† and John Doyle‡

*Department of Physics, University of California, Santa Barbara, CA 93106; and ‡Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125

Highly optimized tolerance (HOT) was recently introduced as a conceptual framework to study fundamental aspects of complexity. HOT is motivated primarily by systems from biology and engineering and emphasizes, (*i*) highly structured, nongeneric, self-dissimilar internal configurations, and (*ii*) robust yet fragile external behavior. HOT claims these are the most important features of complexity and not accidents of evolution or artifices of engineering design but are inevitably intertwined and mutually reinforcing. In the spirit of this collection, our paper contrasts HOT with alternative perspectives on complexity, drawing on real-world examples and also model systems, particularly those from self-organized criticality.

A vision shared by most researchers in complex systems is that certain intrinsic, perhaps even universal, features capture fundamental aspects of complexity in a manner that transcends specific domains. It is in identifying these features that sharp differences arise. In disciplines such as biology, engineering, sociology, economics, and ecology, individual complex systems are necessarily the objects of study, but there often appears to be little common ground between their models, abstractions, and methods. Highly optimized tolerance (HOT) (1–6) is one recent attempt, in a long history of efforts, to develop a general framework for studying complexity. The HOT view is motivated by examples from biology and engineering. Theoretically, it builds on mathematics and abstractions from control, communications, and computing. In this paper, we retain the motivating examples but avoid theories and mathematics that may be unfamiliar to a nonengineering audience. Instead, we aim to make contact with the models, concepts, and abstractions that have been loosely collected under the rubric of a "new science of complexity" (NSOC) (7) or "complex adaptive systems" (CAS), and particularly the concept of self-organized criticality (SOC) (8, 9). SOC is only one element of NSOC/CAS but is a useful representative, because it has a well-developed theory and broad range of claimed applications.

In Table 1, we contrast HOT's emphasis on design and rare configurations with the perspective provided by NSOC/CAS/SOC, which emphasizes structural complexity as "emerging between order and disorder," (*i*) at a bifurcation or phase transition in an interconnection of components that is (*ii*) otherwise largely random. Advocates of NSOC/CAS/SOC are inspired by critical phenomena, fractals, self-similarity, pattern formation, and self-organization in statistical physics, and bifurcations and deterministic chaos from dynamical systems. Motivating examples vary from equilibrium statistical mechanics of interacting spins on a lattice to the spontaneous formation of spatial patterns in systems far from equilibrium. This approach suggests a unity from apparently wildly different examples, because details of component behavior and their interconnection are seen as largely irrelevant to system-wide behavior.

Table 1 shows that SOC and HOT predict not just different but exactly opposite features of complex systems. HOT suggests that random interconnections of components say little about the complexity of real systems, that the details can matter enormously, and that generic (e.g., low codimension) bifurcations and phase transitions play a peripheral role. In principle, Table 1 could have a separate column for Data, by which we mean the observable features of real systems. Because HOT and Data turn out to be identical for these features, we can collapse the table as shown. This is a strong claim, and the remainder of this paper is devoted to justifying it in as much detail as space permits.

Table 1. Characteristics of SOC, HOT, and data

	Property	SOC	HOT and Data
1	Internal configuration	Generic, homogeneous, self-similar	Structured, heterogeneous, self-dissimilar
2	Robustness	Generic	Robust, yet fragile
3	Density and yield	Low	High
4	Max event size	Infinitesimal	Large
5	Large event shape	Fractal	Compact
6	Mechanism for power laws	Critical internal fluctuations	Robust performance
7	Exponent α	Small	Large
8	α vs. dimension d	$\alpha \approx (d-1)/10$	$\alpha \approx 1/d$
9	DDOFs	Small (1)	Large (∞)
10	Increase model resolution	No change	New structures, new sensitivities
11	Response to forcing	Homogeneous	Variable

What Do We Mean By Complexity?

To motivate the theoretical discussion of complex systems, we briefly discuss concrete and hopefully reasonably familiar examples and begin to fill in the "Data" part of Table 1. We start with biological cells and their modern technological counterparts such as very large-scale integrated central processing unit (CPU) chips. Each is a complex system, composed of many components, but is also itself a component in a larger system of organs or laptop or desktop personal computers or embedded in control systems of vehicles such as automobiles or commercial jet aircraft like the Boeing 777. These are again components of the even larger networks that make up organisms and ecosystems,

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Self-Organized Complexity in the Physical, Biological, and Social Sciences," held March 23–24, 2001, at the Arnold and Mabel Beckman Center of the National Academies of Science and Engineering in Irvine, CA.

Abbreviations: NSOC, new science of complexity; CAS, complex adaptive systems; SOC, self-organized criticality; HOT, highly optimized tolerance; CPU, central processing unit; DC, data compression; DDOF, design degree of freedom; CF, California brushfires; FF, U.S. Fish and Wildlife Service land fires; PLR, probability-loss-resource; WWW, World Wide Web.

†To whom reprint requests should be addressed. E-mail: carlson@physics.ucsb.edu.

www.pnas.org/cgi/doi/10.1073/pnas.012582499

Complexity and robustness

J. M. Carlson*† and John Doyle‡

*Department of Physics, University of California, Santa Barbara, CA 93106; and ‡Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125

Highly optimized tolerance (HOT) was recently introduced as a conceptual framework to study fundamental aspects of complexity. HOT is motivated primarily by systems from biology and engineering and emphasizes, (*i*) highly structured, nongeneric, self-dissimilar internal configurations, and (*ii*) robust yet fragile external behavior. HOT claims these are the most important features of complexity and not accidents of evolution or artifices of engineering design but are inevitably intertwined and mutually reinforcing. In the spirit of this collection, our paper contrasts HOT with alternative perspectives on complexity, drawing on real-world examples and also model systems, particularly those from self-organized criticality.

A vision shared by most researchers in complex systems is that certain intrinsic, perhaps even universal, features capture fundamental aspects of complexity in a manner that transcends specific domains. It is in identifying these features that sharp differences arise. In disciplines such as biology, engineering, sociology, economics, and ecology, individual complex systems are necessarily the objects of study, but there often appears to be little common ground between their models, abstractions, and methods. Highly optimized tolerance (HOT) (1–6) is one recent attempt, in a long history of efforts, to develop a general framework for studying complexity. The HOT view is motivated by examples from biology and engineering. Theoretically, it builds on mathematics and abstractions from control, communications, and computing. In this paper, we retain the motivating examples but avoid theories and mathematics that may be unfamiliar to a nonengineering audience. Instead, we aim to make contact with the models, concepts, and abstractions that have been loosely collected under the rubric of a "new science of complexity" (NSOC) (7) or "complex adaptive systems" (CAS), and particularly the concept of self-organized criticality (SOC) (8, 9). SOC is only one element of NSOC/CAS but is a useful representative, because it has a well-developed theory and broad

Table 1. Characteristics of SOC, HOT, and data

	Property	SOC	HOT and Data
1	Internal configuration	Generic, homogeneous, self-similar	Structured, heterogeneous, self-dissimilar
2	Robustness	Generic	Robust, yet fragile
3	Density and yield	Low	High
4	Max event size	Infinitesimal	Large
5	Large event shape	Fractal	Compact
6	Mechanism for power laws	Critical internal fluctuations	Robust performance
7	Exponent α	Small	Large
8	α vs. dimension d	$\alpha \approx (d-1)/10$	$\alpha \approx 1/d$
9	DDOFs	Small (1)	Large (∞)
10	Increase model resolution	No change	New structures, new sensitivities
11	Response to forcing	Homogeneous	Variable

mously, and that generic (e.g., low codimension) bifurcations and phase transitions play a peripheral role. In principle, Table 1 could have a separate column for Data, by which we mean the observable features of real systems. Because HOT and Data turn out to be identical for these features, we can collapse the table as shown. This is a strong claim, and the remainder of this paper is devoted to justifying it in as much detail as space permits.

What Do We Mean By Complexity?

across long distances as an adaptive, emergent, self-organizing, far-from-equilibrium, nonlinear phenomenon. What is both an attraction and a potential weakness of this perspective is that it could be applied to a tornado as well. The HOT view is quite different. Our examples all have high performances and yields and high densities of interconnection (Table 1.3), as well as robustness and reliability. We want to sharpen the distinction not only between the likely short and fatal ride of a tornado with the much faster but relatively boring 777 experience but, more importantly between the 777 design and alternatives that might have worse, or even better, performance, robustness, and reliability.

Are All These Features of Complexity *Necessary*? That is, must systems be broadly robust if they are to successfully function and persist, must this robustness entail highly structured internal complexity, and is fragility a necessary risk? In this sense, is complexity in engineering and biological systems qualitatively the same? We believe that the answer to these questions is largely affirmative, and the examples briefly examined in this section support this, as do numerous other studies of this issue (e.g., see ref. 10). The remainder of this paper offers a thin slice through the HOT theoretical framework that is emerging to systematically address these questions. The concept of HOT was introduced to focus attention on exactly these issues. *Tolerance* emphasizes that robustness in complex systems is a constrained and limited quantity that must be carefully managed and protected. *Highly optimized* emphasizes that this is achieved by highly structured, rare, nongeneric configurations that are products either of deliberate design or evolution. The characteristics of HOT systems are high performance, highly structured internal complexity, and apparently simple and robust external behavior, with the risk of hopefully rare but potentially catastrophic cascading failure events initiated by possibly quite small perturbations.

Power Laws and Complexity

Recently a great deal of attention has been given to the fact that statistics of events in many complex interconnected systems share a common attribute: the distributions of sizes are described by power laws. Several examples are illustrated in Fig. 1, where we plot the cumulative probability $\mathcal{P}(l \geq l_i)$ of events greater than or equal to a given size l_i. Power laws $\mathcal{P}(l \geq l_i) \sim \{l_i\}^{-\alpha}$ are associated with straight lines of slope $-\alpha$ in a $\log(\mathcal{P})$ vs. $\log(l)$ plot, and describe all of the data sets reasonably well, with the exception of data compression (DC), which is exponential.

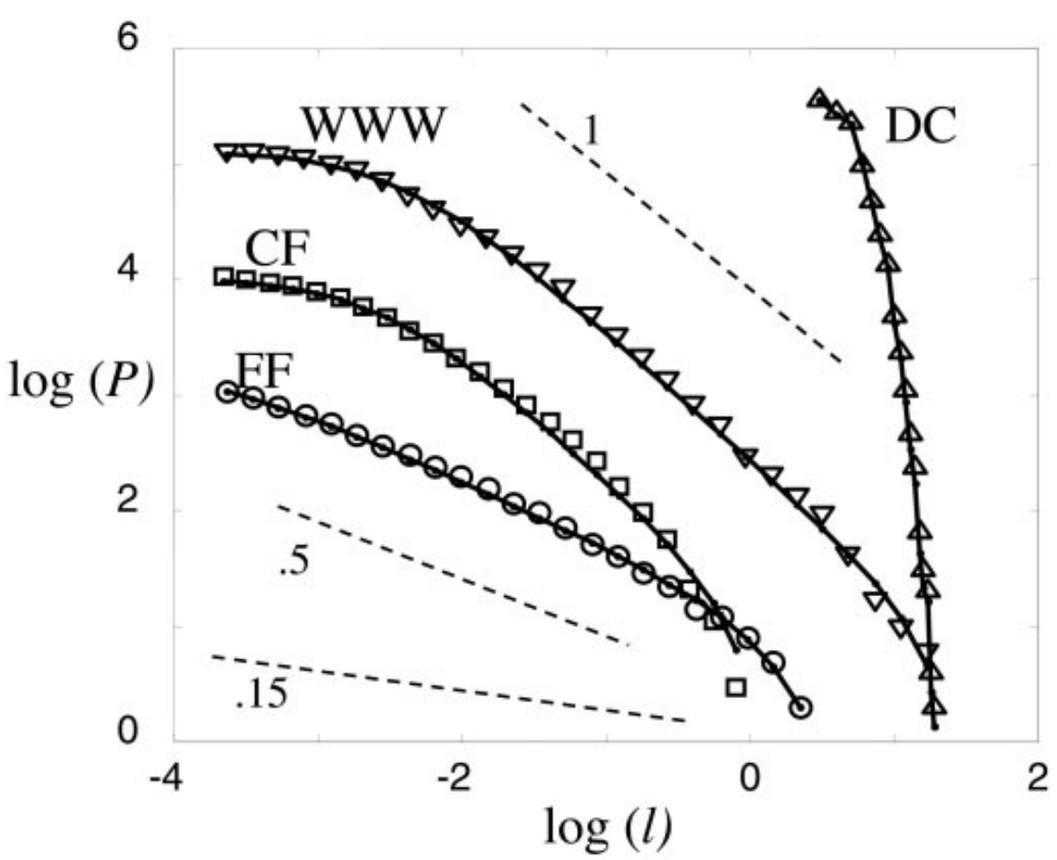


Fig. 1. Log–log (base 10) comparison of DC, WWW, CF, and FF data (symbols) with PLR models (solid lines) (for $\beta = 0, 0.9, 0.9, 1.85$, or $\alpha = 1/\beta = \infty, 1.1, 1.1, 0.054$, respectively) and the SOC FF model ($\alpha = 0.15$, dashed). Reference lines of $\alpha = 0.5$, 1 (dashed) are included. The cumulative distributions of frequencies $\mathcal{P}(l \geq l_i)$ vs. l_i describe the areas burned in the largest 4,284 fires from 1986 to 1995 on all of the U.S. Fish and Wildlife Service Lands (FF) (17), the >10,000 largest California brushfires from 1878 to 1999 (CF) (18), 130,000 web file transfers at Boston University during 1994 and 1995 (WWW) (19), and code words from DC. The size units [1,000 km^2 (FF and CF), megabytes (WWW), and bytes (DC)] and the logarithmic decimation of the data are chosen for visualization.

would be one signature of an internal self-sustaining critical state. The details associated with the initiation of events would be a statistically inconsequential factor in determining their size. Large events would be the result of chance random internal fluctuations characteristic of the self-similar onset of systemwide connectivity at the critical state. In contrast, for HOT power law, statistics are just one symptom of "robust, yet fragile," which we suggest is central to complexity. Heavy tails reflect tradeoffs in systems characterized by high densities and throughputs, where many internal variables have been tuned to favor small losses in common events, at the expense of large losses when subject to rare or unexpected perturbations, even if the perturbations are infinitesimal.

The Forest Fire Models

comments on this story

Stories by subject

- Physics

This article elsewhere

Blogs linking to this article

Add to Connotea

Add to Digg

Add to Facebook

Add to Newsvine

Add to Del.icio.us

Add to Twitter

Published online 17 March 2000 | Nature | doi:10.1038/news000323-2

News

Too darn hot

Why do carefully engineered systems sometimes fail catastrophically? Physicists have attempted to explain why no design is ever perfect, Philip Ball reports.

Philip Ball

In June 1996, the French rocket Ariane 5 made its ignominious maiden flight. Thirty seconds after lift-off, the on-board computer controlling the rocket's trajectory went haywire, attempting such an abrupt change in course that the rocket simply fell apart. It later emerged that the computer software was taken over from an earlier design, and had not been tested under the launch conditions of the Ariane 5 rocket because it was deemed too expensive. The dilemma is one familiar to engineers: how to balance thorough testing and optimization against the practical demands of getting the job done?

Now two US physicists claim to have figured out how it is that complex systems, from computer networks to rainforests, can be robust in the face of anticipated disturbances, yet too fragile to cope with unexpected events. They call this state 'highly optimized tolerance' (HOT). A HOT system, they say, can cope with all the 'slings and arrows' that it is designed to withstand, but can fail catastrophically if presented with some unforeseen challenge.

In the case of the Ariane crash, the fatal problem was agonizingly trivial: in effect, the rounding-off of a number used in the

most recent | commented

- Corrections
  09 November 2011
- Seth Stein: The quake killer
  09 November 2011
- Seven days: 4–10 November 2011
  09 November 2011
- Fresh dispute about MMR 'fraud'
  09 November 2011
- The pollinator crisis: What's best for bees
  09 November 2011

You Know That Space-Time Thing? Never Mind

By George Johnson
Published: June 09, 2002

A NEW KIND OF SCIENCE

By Stephen Wolfram.

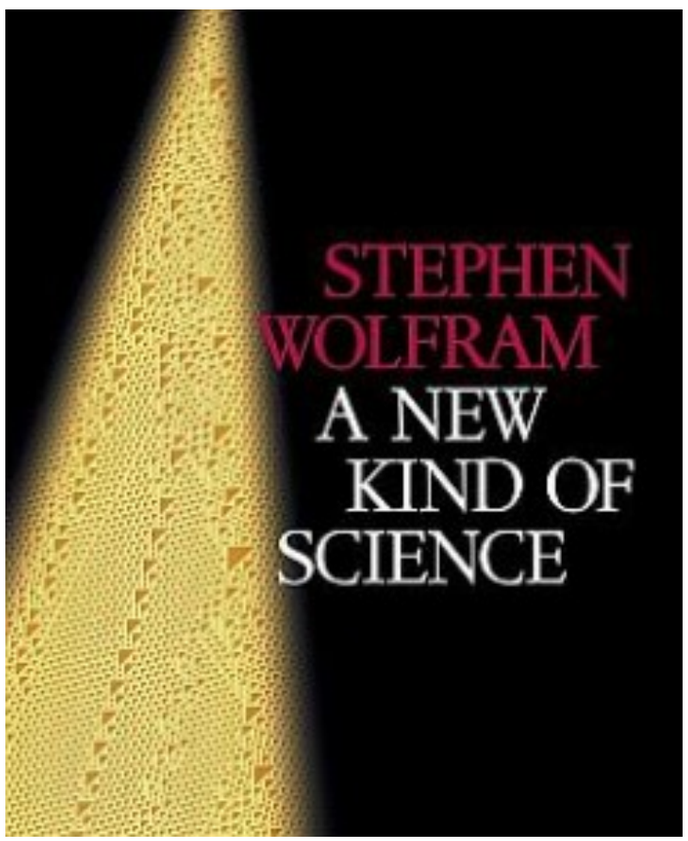
Illustrated. 1,263 pp. Champaign, Ill.:

Wolfram Media. $44.95.

AMONG a small group of very smart people, the publication of "A New Kind of Science," by Stephen Wolfram, has been anticipated with the anxiety aroused in literary circles by, say, Jonathan Franzen's recent novel, "The Corrections." For more than a decade, Wolfram, a theoretical physicist turned millionaire software entrepreneur, has been laboring in solitude on a work that, he has promised, will change the way we see the world. Adding to the suspense, the book has been announced and withdrawn as the artist returned to his garret to tinker, ignoring the bad vibes and hexes cast by jealous colleagues hoping to see him fall flat on his face.

Now, weighing in at 1,263 pages (counting a long, unpaginated index) and 583,313 words, the book could hardly be more intimidating. But that is the price one pays for a first-class intellectual thrill. While experimenting with a simple computer program 20 years ago, Wolfram stumbled on something rather eerie: "the beginning of a crack in the very foundations of existing science." Ever since, he has been following it deeper as it widens into a crevasse.

My approach in investigating issues like the Second Law is in effect to use simple programs as metaphors for physical systems. But can such programs in fact be more than that? And for example is it conceivable that at some level physical systems actually operate directly according to the rules of our simple program? ... At first the laws might seem much too complicated to correspond to any simple program. But one of the crucial discoveries of this book is that even programs with very simple underlying rules can yield great complexity... And so it could be with fundamental physics. Underneath the laws of physics as we know them today it could be that there lies a ver simple program from which all the known laws — and ultimately all the complexity we see in the universe -- emerges.



Variants on the mean

Given the natural friendship between the mean/standard deviation and the normal distribution, these measures can break down when the data are badly skewed or have "outliers" — there have been various attempts to patch things up

Mid-Mean - computes a mean using the data between the 25th and 75th percentiles.

Trimmed Mean - similar to the mid-mean except different percentile values are used. A common choice is to trim 5% of the points in both the lower and upper tails, i.e., calculate the mean for data between the 5th and 95th percentiles.

Winsorized Mean - similar to the trimmed mean. However, instead of trimming the points, they are set to the lowest (or highest) value. For example, all data below the 5th percentile are set equal to the value of the 5th percentile and all data greater than the 95th percentile are set equal to the 95th percentile.

Numerical descriptions

The median is the middle point — the point where half the data are above and half below. It has a history behind it that uses language about “democracy” and “one data point, one vote”

It is insensitive to outliers (often called robust) and minimizes a distance calculation as well, this time using the absolute value and not the squared difference.

It’s partner in crime is the interquartile range (IQR) for measuring spread

Vox Populi

“In these democratic days, any investigation into the trustworthiness and peculiarities of popular judgements is of interest… According to the democratic principle of ‘one vote one value,’ the middlemost estimate expresses the Vox populist, ever other estimate being condemned as too low or too high by a majority of the voters.”

17°·0 at Moyeni, Basutoland, on August 23. The mean yearly value of the absolute maxima was 86°·9, and of the corresponding minima 41°·6. The mean temperature for the year was 0°·9 below the average. The stormiest month was October, and the calmest was April.

We have also received the official meteorological year-books for South Australia (1904) and Mysore (1905). Both of these works contain valuable means for previous years.

Forty Years of Southern New Mexico Climate.—Bulletin No. 59 of the New Mexico College of Agriculture contains the meteorological data recorded at the experimental station from 1892 to 1905 inclusive, together with results of temperature and rainfall observations at other stations in the Mesilla Valley for most of the years between 1851 and 1890, published some years ago by General Greely in a “Report on the Climate of New Mexico.” The station is situated in lat. 32° 15′ N., long. 106° 45′ W., and is 3868 feet above sea-level. The data have a general application to those portions of southern New Mexico with an altitude less than 4000 feet. The mean annual temperature for the whole period was 61°·6, mean maximum (fourteen years) 76°·8, mean minimum 41°·4, absolute maximum 106° (which occurred several times), absolute minimum 1° (December, 1895). The mean annual rainfall was 8·8 inches; the smallest yearly amount was 3·5 inches, in 1873, the largest 17·1 inches, in 1905. Most of the rain falls during July, August, and September. The relative humidity is low, the mean annual amount being about 51 per cent. The bulletin was prepared by J. D. Tinsley, vice-director of the station.

Meteorological Observations in Germany.—The results of the observations made under the system of the Deutsche Seewarte, Hamburg, for 1905, at ten stations of the second order, and at fifty-six storm-warning stations, have been received. This is the twenty-eighth yearly volume published by the Seewarte, and forms part of the series of German meteorological year-books. We have frequently referred to this excellent series, and the volume in question is similar in all respects to its predecessors; it contains most valuable data relating to the North Sea and Baltic coasts. We note that the sunshine at Hamburg was only 29 per cent. of the possible annual amount, and that there were 103 sunless days; the rainfall was 25·9 inches, the rainy days being 172 in number.

VOX POPULI.

IN these democratic days, any investigation into the trustworthiness and peculiarities of popular judgments is of interest. The material about to be discussed refers to a small matter, but is much to the point.

A weight-judging competition was carried on at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox having been selected, competitors bought stamped and numbered cards, for 6*d*. each, on which to inscribe their respective names, addresses, and estimates of what the ox would weigh after it had been slaughtered and “dressed.” Those who guessed most successfully received prizes. About 800 tickets were issued, which were kindly lent me for examination after they had fulfilled their immediate purpose. These afforded excellent material. The judgments were unbiassed by passion and uninfluenced by oratory and the like. The sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best. The competitors included butchers and farmers, some of whom were highly expert in judging the weight of cattle; others were probably guided by such information as they might pick up, and by their own fancies. The average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among the voters to judge justly was probably much the same in either case.

After weeding thirteen cards out of the collection, as being defective or illegible, there remained 787 for discussion. I arrayed them in order of the magnitudes of the estimates, and converted the *cwt.*, *quarters*, and *lbs.* in which they were made, into lbs., under which form they will be treated.

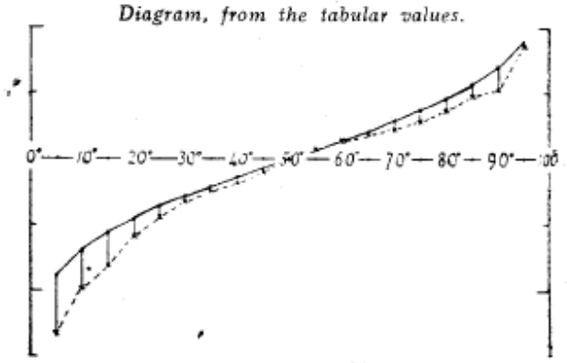
Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e. =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q_1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
m 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q_3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	−10
85	1254	+ 47	+57	−10
90	1267	+ 52	+70	−18
95	1293	+ 86	+90	− 4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

According to the democratic principle of “one vote one value,” the middlemost estimate expresses the *vox populi*, every other estimate being condemned as too low or too high by a majority of the voters (for fuller explanation see “One Vote, One Value,” NATURE, February 28, p. 414). Now the middlemost estimate is 1207 lb., and the weight of the dressed ox proved to be 1198 lb.; so the *vox populi* was in this case 9 lb., or 0·8 per cent. of the whole weight too high. The distribution of the estimates about their middlemost value was of the usual type, so far that they clustered closely in its neighbourhood and became rapidly more sparse as the distance from it increased.

Diagram, from the tabular values.



The continuous line is the normal curve with p.e. = 37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

But they were not scattered symmetrically. One quarter of them deviated more than 45 lb. above the middlemost (3·7 per cent.), and another quarter deviated more than 29 lb. below it (2·4 per cent.), therefore the range of the two middle quarters, that is, of the middlemost half, lay within those limits. It would be an equal chance that the estimate written on any card picked at random out of the collection lay within or without those limits. In other words, the “probable error” of a single observation may be reckoned as ½(45+29), or 37 lb. (3·1 per cent.). Taking this for the p.e. of the normal curve that is best adapted for comparison with the observed values, the results are obtained which appear in above table, and graphically in the diagram.

One Vote, One Value

Galton focuses on two simple cases — a town council has to estimate how much money to allocate on a given project, and a jury is tasked with assessing damages after a trial.

Galton wrote that each voter in these cases should have “equal authority with each of his colleagues” and that this property invalidates the mean of the individual estimates as it gives “voting power to ‘cranks’ in proportion to their crankiness.”

He clarifies — “One absurdly large or small estimate would leave a greater impress on the result than one of reasonable amount and the more an estimate diverges from the bulk of the rest, the more influence it would exert”

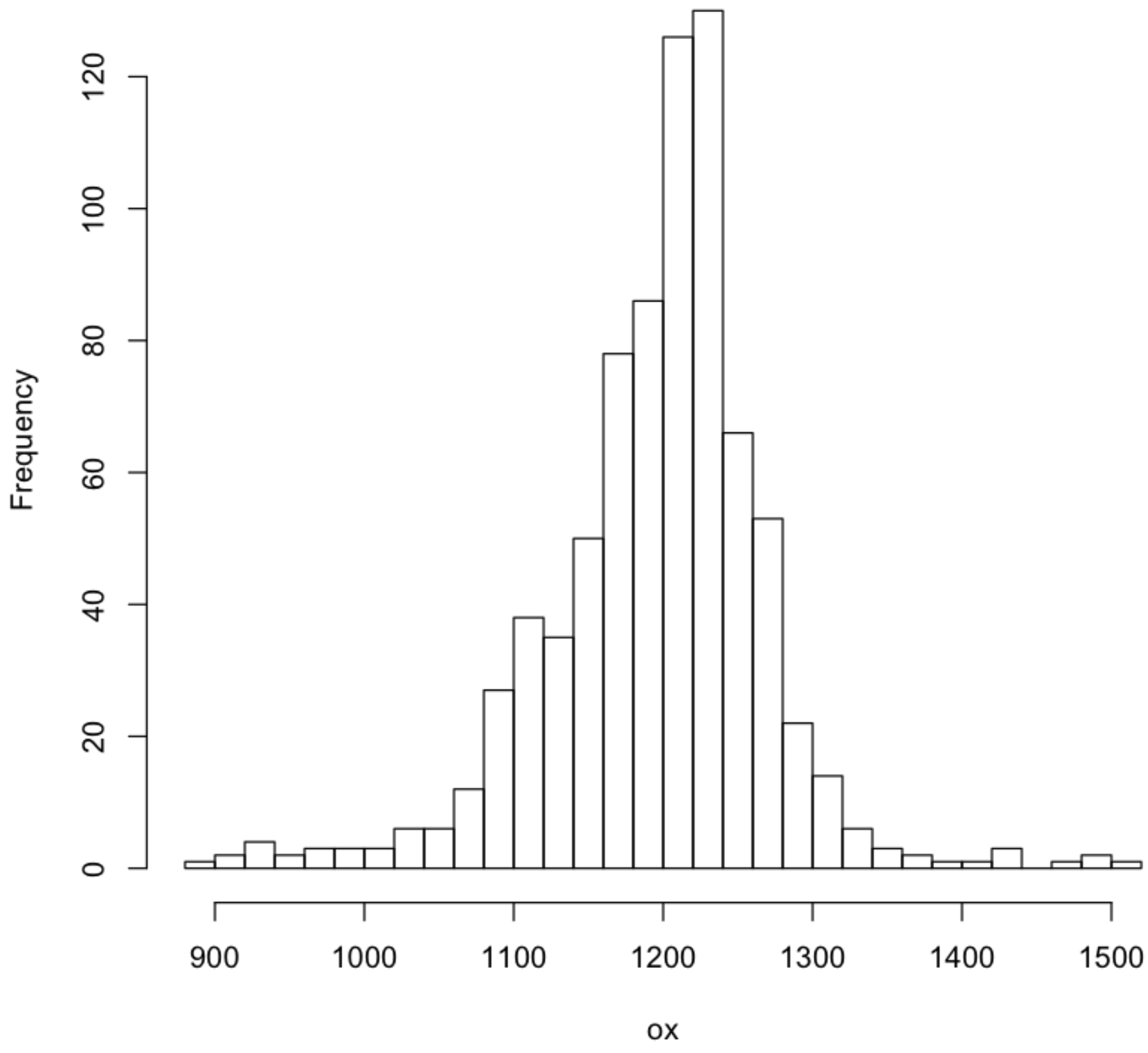
LETTERS TO THE EDITOR.

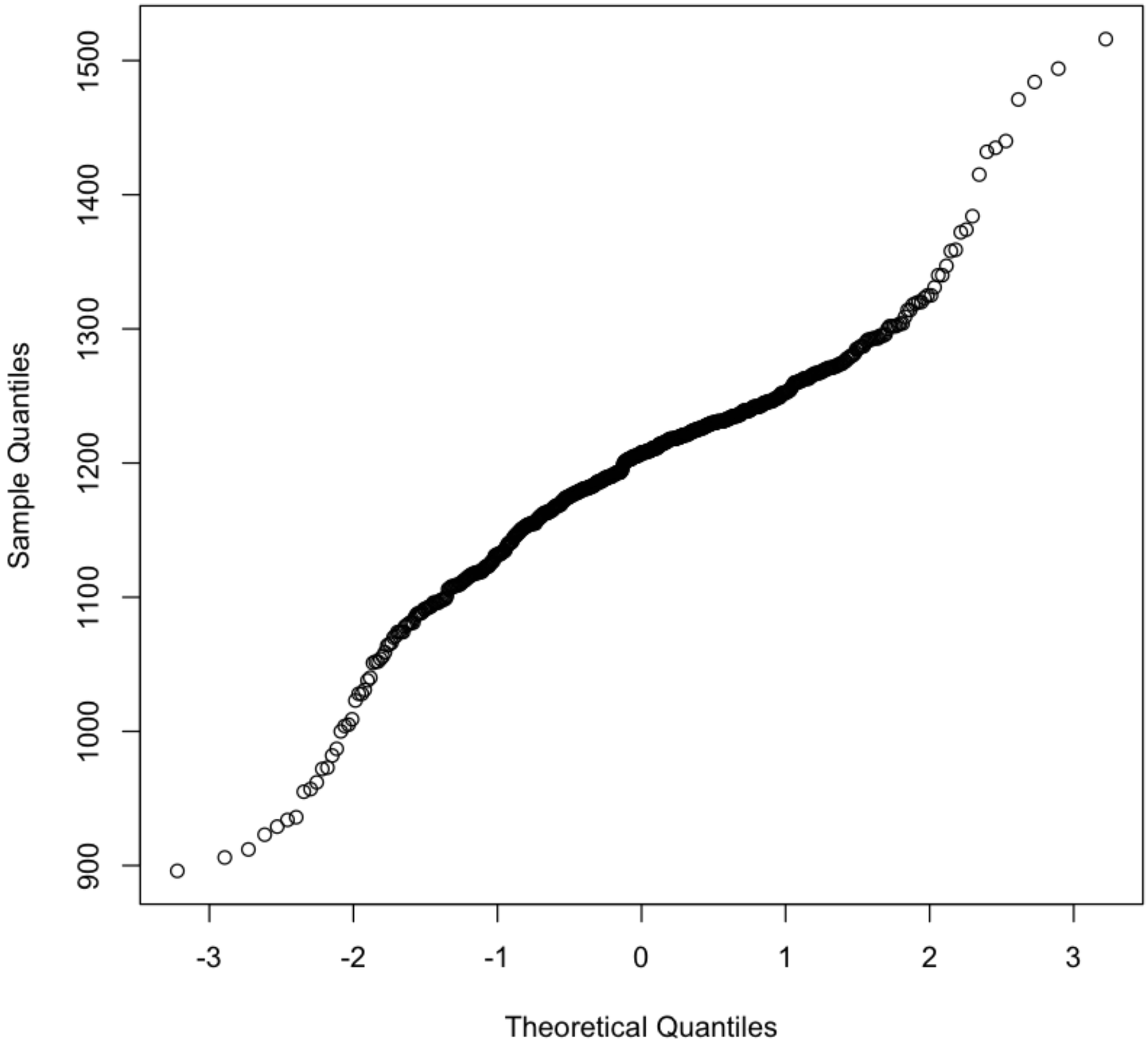
One Vote, One Value.

A CERTAIN class of problems do not as yet appear to be solved according to scientific rules, though they are of much importance and of frequent recurrence. Two examples will suffice. (1) A jury has to assess damages. (2) The council of a society has to fix on a sum of money, suitable for some particular purpose. Each voter, whether of the jury or of the council, has equal authority with each of his colleagues. How can the right conclusion be reached, considering that there may be as many different estimates as there are members? That conclusion is clearly *not* the *average* of all the estimates, which would give a voting power to “cranks” in proportion to their crankiness. One absurdly large or small estimate would leave a greater impress on the result than one of reasonable amount, and the more an estimate diverges from the bulk of the rest, the more influence would it exert. I wish to point out that the estimate to which least objection can be raised is the *middlemost* estimate, the number of votes that it is too high being exactly balanced by the number of votes that it is too low. Every other estimate is condemned by a majority of voters as being either too high or too low, the middlemost alone escaping this condemnation. The number of voters may be odd or even. If odd, there is one middlemost value; thus in 11 votes the middlemost is the 6th; in 99 votes the middlemost is the 50th. If the number of voters be even, there are two middlemost values, the mean of which must be taken; thus in 12 votes the middlemost lies between the 6th and the 7th; in 100 votes between the 50th and the 51st. Generally, in $2n-1$ votes the middlemost is the nth; in $2n$ votes it lies between the nth and the $(n+1)$th.

I suggest that the process for a jury on their retirement should be (1) to discuss and interchange views; (2) for each juryman to write his own independent estimate on a separate slip of paper; (3) for the foreman to arrange the slips in the order of the values written on them; (4) to take the average of the 6th and 7th as the verdict, which might be finally approved as a substantive proposition. Similarly as regards the resolutions of councils, having regard to the above $(2n-1)$ and $2n$ remarks.

FRANCIS GALTON.


Histogram of ox
Frequency
0
20
40
60
80
100
120
900
1000
1100
1200
1300
1400
1500
ox


Normal Q-Q Plot
Sample Quantiles
900
1000
1100
1200
1300
1400
1500
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Today — Inference

Inevitably, while you are reporting a story, you are going to come across the artifacts of statistical inference. Maybe it's a P-value, maybe it's a confidence interval. Maybe it's something sexy like a Bayes Factor. In the next couple of breakfasts, we are going to give you a sense of the style of reasoning that statisticians employ when learning from the world.

In some sense, statistics is about the clever deployment of randomness to learn something about the world. While this is usually portrayed in text books as a stable enterprise, there's plenty of disagreements in the statistical community about methods for learning, for making inferences from data.

Some of the most popular tools have proven to be the most tricky to understand, largely because statistics is taught badly. We'll use a couple of simple examples to hopefully demystify these tools.

First, a kind of story that's pretty common…

The New York Times

Alternatives to Opioids for Pain Relief

By **NICHOLAS BAKALAR** NOV. 8, 2017

A combination of Tylenol and Advil worked just as well as opioids for relief of pain in the emergency room, a randomized trial has found.

Researchers studied 416 men and women who arrived in the E.R. with moderate to severe pain in their arms or legs from sprains, strains, fractures or other injuries. They randomly assigned them to an oral dose of acetaminophen (Tylenol) with either ibuprofen (Advil) or the opioids oxycodone, hydrocodone or codeine. Two hours later, they questioned them using an 11-point pain scale.

The average score was 8.7 before taking medicine. That score decreased 4.3 points with ibuprofen and Tylenol, 4.4 with oxycodone and Tylenol, 3.5 with hydrocodone and Tylenol, and 3.9 with codeine and Tylenol. In other words, there was no significant difference, either statistically or clinically, among any of the four regimens. The study is in JAMA.

JAMA | Original Investigation

Effect of a Single Dose of Oral Opioid and Nonopioid Analgesics on Acute Extremity Pain in the Emergency Department
A Randomized Clinical Trial

Andrew K. Chang, MD, MS; Polly E. Bijur, PhD; David Esses, MD; Douglas P. Barnaby, MD, MS; Jesse Baer, MD

IMPORTANCE The choice of analgesic to treat acute pain in the emergency department (ED) lacks a clear evidence base. The combination of ibuprofen and acetaminophen (paracetamol) may represent a viable nonopioid alternative.

OBJECTIVES To compare the efficacy of 4 oral analgesics.

DESIGN, SETTINGS, AND PARTICIPANTS Randomized clinical trial conducted at 2 urban EDs in the Bronx, New York, that included 416 patients aged 21 to 64 years with moderate to severe acute extremity pain enrolled from July 2015 to August 2016.

INTERVENTIONS Participants (104 per each combination analgesic group) received 400 mg of ibuprofen and 1000 mg of acetaminophen; 5 mg of oxycodone and 325 mg of acetaminophen; 5 mg of hydrocodone and 300 mg of acetaminophen; or 30 mg of codeine and 300 mg of acetaminophen.

MAIN OUTCOMES AND MEASURES The primary outcome was the between-group difference in decline in pain 2 hours after ingestion. Pain intensity was assessed using an 11-point numerical rating scale (NRS), in which 0 indicates no pain and 10 indicates the worst possible pain. The predefined minimum clinically important difference was 1.3 on the NRS. Analysis of variance was used to test the overall between-group difference at $P = .05$ and 99.2% CIs adjusted for multiple pairwise comparisons.

RESULTS Of 416 patients randomized, 411 were analyzed (mean [SD] age, 37 [12] years; 199 [48%] women; 247 [60%] Latino). The baseline mean NRS pain score was 8.7 (SD, 1.3). At 2 hours, the mean NRS pain score decreased by 4.3 (95% CI, 3.6 to 4.9) in the ibuprofen and acetaminophen group; by 4.4 (95% CI, 3.7 to 5.0) in the oxycodone and acetaminophen group; by 3.5 (95% CI, 2.9 to 4.2) in the hydrocodone and acetaminophen group; and by 3.9 (95% CI, 3.2 to 4.5) in the codeine and acetaminophen group ($P = .053$). The largest difference in decline in the NRS pain score from baseline to 2 hours was between the oxycodone and acetaminophen group and the hydrocodone and acetaminophen group (0.9; 99.2% CI, −0.1 to 1.8), which was less than the minimum clinically important difference in NRS pain score of 1.3. Adverse events were not assessed.

CONCLUSIONS AND RELEVANCE For patients presenting to the ED with acute extremity pain, there were no statistically significant or clinically important differences in pain reduction at 2 hours among single-dose treatment with ibuprofen and acetaminophen or with 3 different opioid and acetaminophen combination analgesics. Further research to assess adverse events and other dosing may be warranted.

TRIAL REGISTRATION clinicaltrials.gov Identifier: NCT02455518

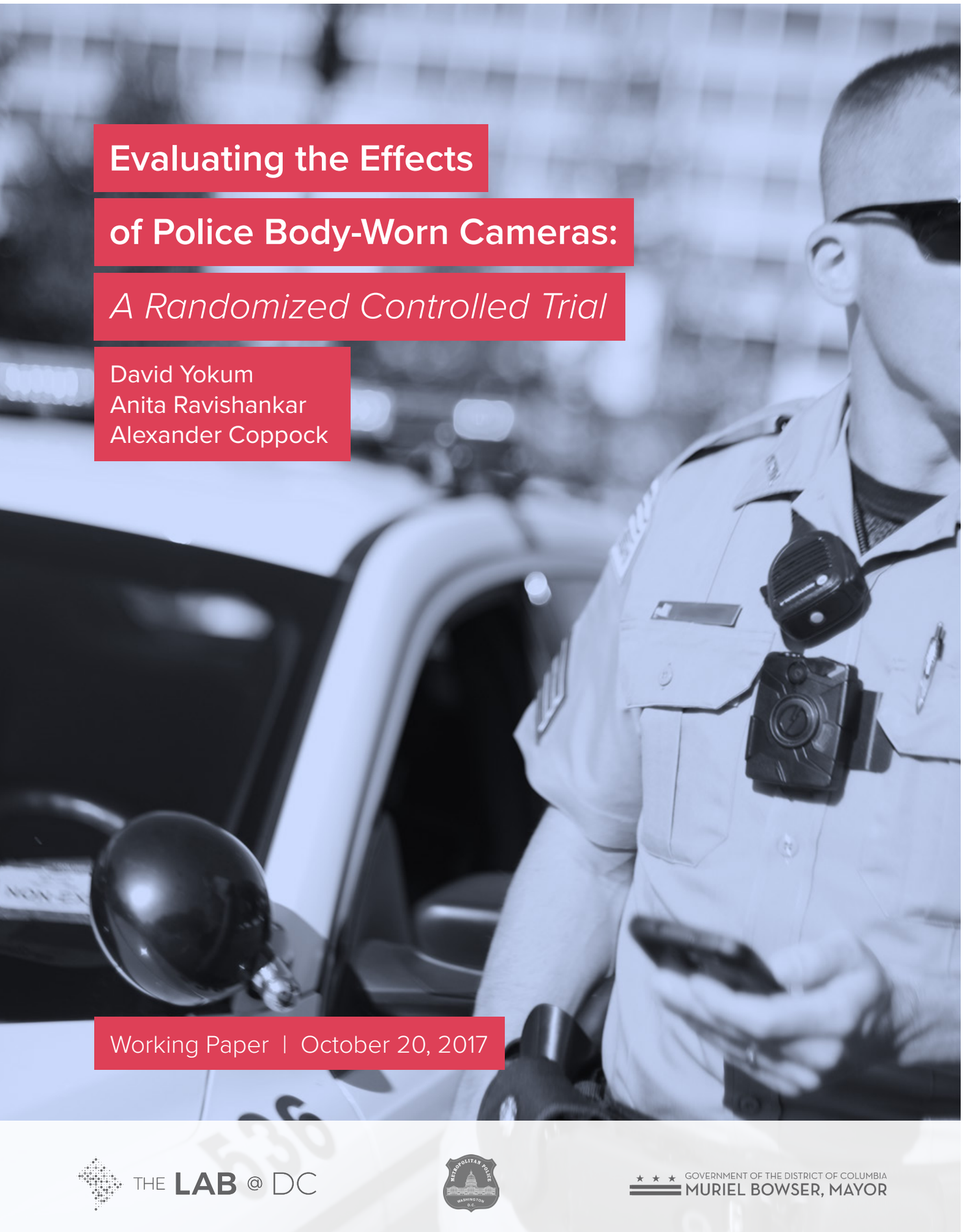
Editorial page 1655

Supplemental content

Author Affiliations: Department of Emergency Medicine, Albany Medical College, Albany, New York (Chang); Department of Emergency Medicine, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, New York (Bijur, Esses, Barnaby, Baer).

Corresponding Author: Andrew K. Chang, MD, MS, Department of Emergency Medicine, Albany Medical College, 16 New Scotland Ave, MC-139, Albany, NY 12208 (achang3@yahoo.com).

JAMA. 2017;318(17):1661-1667. doi:10.1001/jama.2017.16190



Evaluating the Effects of Police Body-Worn Cameras: *A Randomized Controlled Trial*

David Yokum
Anita Ravishankar
Alexander Coppock

Working Paper | October 20, 2017

THE LAB @ DC

GOVERNMENT OF THE DISTRICT OF COLUMBIA
MURIEL BOWSER, MAYOR

JAMA | Original Investigation

Effect of a Single Dose of Oral Opioid and Nonopioid Analgesics on Acute Extremity Pain in the Emergency Department
A Randomized Clinical Trial

Andrew K. Chang, MD, MS; Polly E. Bijur, PhD; David Esses, MD; Douglas P. Barnaby, MD, MS; Jesse Baer, MD

IMPORTANCE The choice of analgesic to treat acute pain in the emergency department (ED) lacks a clear evidence base. The combination of ibuprofen and acetaminophen (paracetamol) may represent a viable nonopioid alternative.

OBJECTIVES To compare the efficacy of 4 oral analgesics.

DESIGN, SETTINGS, AND PARTICIPANTS Randomized clinical trial conducted at 2 urban EDs in the Bronx, New York, that included 416 patients aged 21 to 64 years with moderate to severe acute extremity pain enrolled from July 2015 to August 2016.

INTERVENTIONS Participants (104 per each combination analgesic group) received 400 mg of ibuprofen and 1000 mg of acetaminophen; 5 mg of oxycodone and 325 mg of acetaminophen; 5 mg of hydrocodone and 300 mg of acetaminophen; or 30 mg of codeine and 300 mg of acetaminophen.

MAIN OUTCOMES AND MEASURES The primary outcome was the between-group difference in decline in pain 2 hours after ingestion. Pain intensity was assessed using an 11-point numerical rating scale (NRS), in which 0 indicates no pain and 10 indicates the worst possible pain. The predefined minimum clinically important difference was 1.3 on the NRS. Analysis of variance was used to test the overall between-group difference at $P = .05$ and 99.2% CIs adjusted for multiple pairwise comparisons.

RESULTS Of 416 patients randomized, 411 were analyzed (mean [SD] age, 37 [12] years; 199 [48%] women; 247 [60%] Latino). The baseline mean NRS pain score was 8.7 (SD, 1.3). At 2 hours, the mean NRS pain score decreased by 4.3 (95% CI, 3.6 to 4.9) in the ibuprofen and acetaminophen group; by 4.4 (95% CI, 3.7 to 5.0) in the oxycodone and acetaminophen group; by 3.5 (95% CI, 2.9 to 4.2) in the hydrocodone and acetaminophen group; and by 3.9 (95% CI, 3.2 to 4.5) in the codeine and acetaminophen group ($P = .053$). The largest difference in decline in the NRS pain score from baseline to 2 hours was between the oxycodone and acetaminophen group and the hydrocodone and acetaminophen group (0.9; 99.2% CI, −0.1 to 1.8), which was less than the minimum clinically important difference in NRS pain score of 1.3. Adverse events were not assessed.

CONCLUSIONS AND RELEVANCE For patients presenting to the ED with acute extremity pain, there were no statistically significant or clinically important differences in pain reduction at 2 hours among single-dose treatment with ibuprofen and acetaminophen or with 3 different opioid and acetaminophen combination analgesics. Further research to assess adverse events and other dosing may be warranted.

TRIAL REGISTRATION clinicaltrials.gov Identifier: NCT02455518

Editorial page 1655

Supplemental content

Author Affiliations: Department of Emergency Medicine, Albany Medical College, Albany, New York (Chang); Department of Emergency Medicine, Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, New York (Bijur, Esses, Barnaby, Baer).

Corresponding Author: Andrew K. Chang, MD, MS, Department of Emergency Medicine, Albany Medical College, 16 New Scotland Ave, MC-139, Albany, NY 12208 (achang3@yahoo.com).

JAMA. 2017;318(17):1661-1667. doi:10.1001/jama.2017.16190

Figure. Flow of Patients Through Acute Extremity Pain Trial

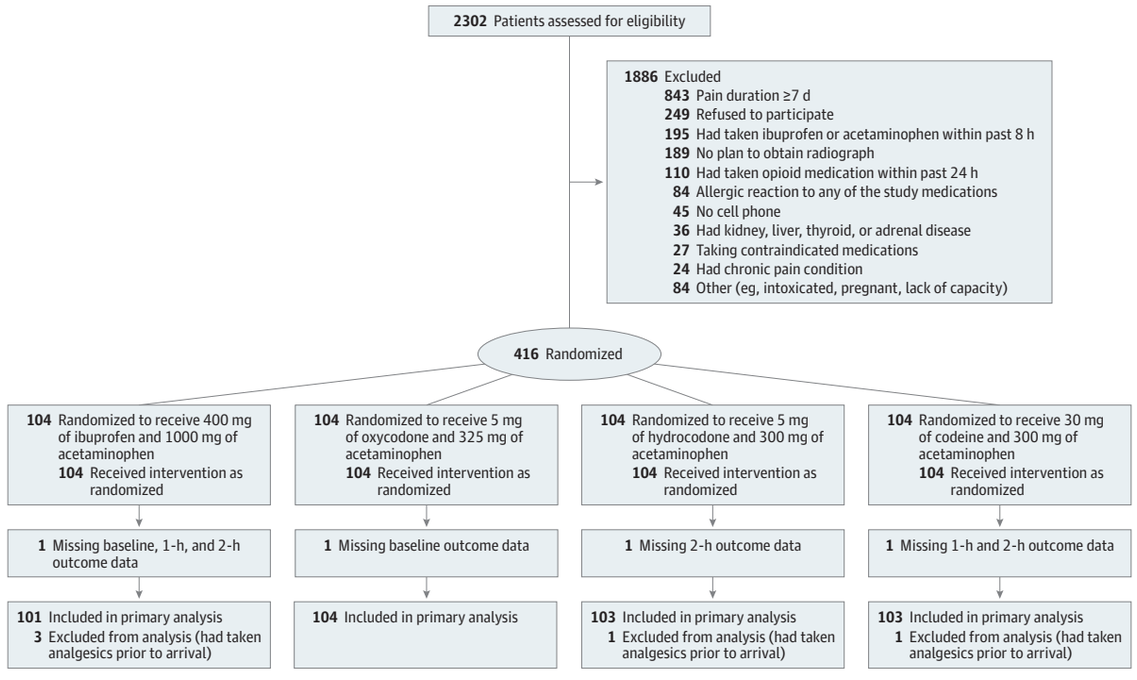


Table 1. Patient Characteristics

	Ibuprofen and Acetaminophen[a]	Oxycodone and Acetaminophen[b]	Hydrocodone and Acetaminophen[c]	Codeine and Acetaminophen[d]
No. of patients	101	104	103	103
Female sex, No. (%)	54 (54)	50 (48)	51 (50)	44 (43)
Age, mean (SD), y	37 (11)	37 (12)	37 (13)	37 (12)
Diagnosis, No. (%)				
Sprain or strain	64 (63)	66 (64)	59 (57)	67 (65)
Extremity fracture	21 (21)	23 (22)	21 (20)	24 (23)
Muscle pain	8 (8)	9 (9)	12 (12)	7 (7)
Contusion	4 (4)	3 (3)	7 (7)	2 (2)
Other	4 (4)	3 (3)	4 (4)	3 (3)
Nonpharmacological ED interventions, No. (%)				
Elastic bandage	39 (39)	37 (36)	23 (22)	36 (35)
Splint	12 (12)	20 (19)	18 (18)	10 (10)
Cast	10 (10)	14 (14)	6 (6)	11 (11)
Ice	7 (7)	11 (11)	10 (10)	4 (4)
Other	11 (11)	5 (5)	15 (15)	16 (16)

Abbreviation: ED, emergency department.

[a] Patients received 400 mg of ibuprofen and 1000 mg of acetaminophen.

[b] Patients received 5 mg of oxycodone and 325 mg of acetaminophen.

[c] Patients received 5 mg of hydrocodone and 300 mg of acetaminophen.

[d] Patients received 30 mg of codeine and 300 mg of acetaminophen.

in the codeine and acetaminophen group. The overall test of the null hypothesis that there is no difference in change in pain by treatment group from baseline to 2 hours (the primary outcome measure) was not statistically significant ($P = .053$). There was also no significant difference at 1 hour ($P = .13$) (Table 2).

Table 3 shows the comparisons in mean change in pain between each pair of analgesics. None of the differences between analgesics was statistically significant or met the a priori definition of a minimally clinically important difference in mean NRS pain score of 1.3.

Seventy-three patients (17.8%) received rescue analgesics within the 2-hour period (**Table 4**). The distribution of receipt of rescue analgesia was not statistically significant, but the estimates varied by as much as 9% (oxycodone and acetaminophen vs codeine and acetaminophen). Results of the analysis with multiple imputations of the NRS pain scores for

Table 2. Numerical Rating Scale (NRS) Pain Scores and Decline in Pain Scores by Treatment Group

	NRS Pain Score, Mean (95% CI)[a]				
	Ibuprofen and Acetaminophen[b]	Oxycodone and Acetaminophen[c]	Hydrocodone and Acetaminophen[d]	Codeine and Acetaminophen[e]	*P* Value[f]
No. of patients[g]	101	104	103	103	
Primary end point: decline in score to 2 h	4.3 (3.6 to 4.9)	4.4 (3.7 to 5.0)	3.5 (2.9 to 4.2)	3.9 (3.2 to 4.5)	.053
Baseline score	8.9 (8.5 to 9.2)	8.7 (8.3 to 9.0)	8.6 (8.3 to 9.0)	8.6 (8.2 to 8.9)	.47
Score at 1 h	5.9 (5.3 to 6.6)	5.5 (4.9 to 6.2)	6.2 (5.6 to 6.9)	5.9 (5.2 to 6.5)	.25
Score at 2 h	4.6 (3.9 to 5.3)	4.3 (3.6 to 5.0)	5.1 (4.5 to 5.8)	4.7 (4.0 to 5.4)	.13
Decline in score to 1 h	2.9 (2.4 to 3.5)	3.1 (2.6 to 3.7)	2.4 (1.8 to 3.0)	2.7 (2.1 to 3.3)	.13

[a] Pain intensity was assessed using an 11-point NRS in which a score of 0 indicates no pain and a score of 10 indicates the worst possible pain.
[b] Patients received 400 mg of ibuprofen and 1000 mg of acetaminophen.
[c] Patients received 5 mg of oxycodone and 325 mg of acetaminophen.
[d] Patients received 5 mg of hydrocodone and 300 mg of acetaminophen.
[e] Patients received 30 mg of codeine and 300 mg of acetaminophen.
[f] Calculated using analysis of variance.
[g] One patient in each group had imputed NRS data.

Table 3. Between-Group Difference in Mean Change in Numerical Rating Scale (NRS) Pain Scores

Comparison	Between-Group Difference in Mean Change in NRS Pain Score (99.2% CI)[a]	
	From Baseline to 1 h	From Baseline to 2 h
Ibuprofen and acetaminophen vs oxycodone and acetaminophen	−0.2 (−1.0 to 0.6)	−0.1 (−1.0 to 0.8)
Ibuprofen and acetaminophen vs hydrocodone and acetaminophen	0.5 (−0.3 to 1.3)	0.8 (−0.2 to 1.7)
Ibuprofen and acetaminophen vs codeine and acetaminophen	0.2 (−0.6 to 1.0)	0.4 (−0.6 to 1.3)
Oxycodone and acetaminophen vs hydrocodone and acetaminophen	0.7 (−0.1 to 1.5)	0.9 (−0.1 to 1.8)
Oxycodone and acetaminophen vs codeine and acetaminophen	0.4 (−0.4 to 1.2)	0.5 (−0.4 to 1.4)
Hydrocodone and acetaminophen vs codeine and acetaminophen	−0.3 (−1.1 to 0.5)	−0.4 (−1.3 to 0.6)

[a] Indicates mean change in pain of first analgesic minus mean change in pain from second analgesic. Pain intensity was assessed using an 11-point NRS in which a score of 0 indicates no pain and a score of 10 indicates the worst possible pain.

Table 4. Rescue Analgesic and Total Morphine Equivalent Units Received Within 2 Hours

	Ibuprofen and Acetaminophen	Oxycodone and Acetaminophen	Hydrocodone and Acetaminophen	Codeine and Acetaminophen	*P* Value
No. of patients	101	104	103	103	
Received rescue analgesic, No. (%)	18 (17.8)	14 (13.5)	18 (17.5)	23 (22.3)	.42
Type of rescue analgesic received, No. (%)					
Oxycodone	17 (16.8)	13 (12.5)	17 (16.5)	22 (21.4)	.55
Morphine	1 (1.0)	0	0	1 (1.0)	
Tramadol	0	1 (1.0)	1 (1.0)	0	
Analgesic dose in morphine equivalent units, mean (SD)[a]					
Initial	0 (0)	7.5 (0)	5.0 (0)	4.5 (0)	NA[b]
Rescue	1.6 (3.5)	1.1 (2.7)	1.7 (3.2)	2.0 (3.4)	.27
Total	1.6 (3.5)	8.6 (2.7)	6.7 (3.2)	6.5 (3.4)	<.001

[a] Calculated based on the US Centers for Medicare & Medicaid Services Opioid Oral Morphine Milligram Equivalent conversion factor table: 1.5 for oxycodone; 1.0 for hydrocodone; 0.15 for codeine; 0.1 for tramadol; and 3.0 for intravenous morphine.
[b] Statistical test cannot be calculated.

patients who received rescue analgesics were nearly identical to the analysis without imputation (eTable 1 and eTable 2 in Supplement 2). There were no clinically important or statistically significant differences in efficacy when these post hoc analyses were performed.

The amount of rescue analgesia received in morphine equivalent units was not significantly different across groups (Table 4). The total amount of opioid was significantly associated with treatment group. One patient in the ibuprofen and acetaminophen group received 6 mg of intravenous morphine and 1 patient in the codeine and acetaminophen group received 4 mg of intravenous morphine.

We conducted a post hoc subset analysis to assess whether any analgesic was more effective for severe pain among patients who either (1) rated their initial pain as a score of 10 on the NRS or (2) had a documented fracture on radiological imaging. The results were similar to those from the entire sample. There were no statistically significant or clinically important between-group differences (eTable 3 in Supplement 2).

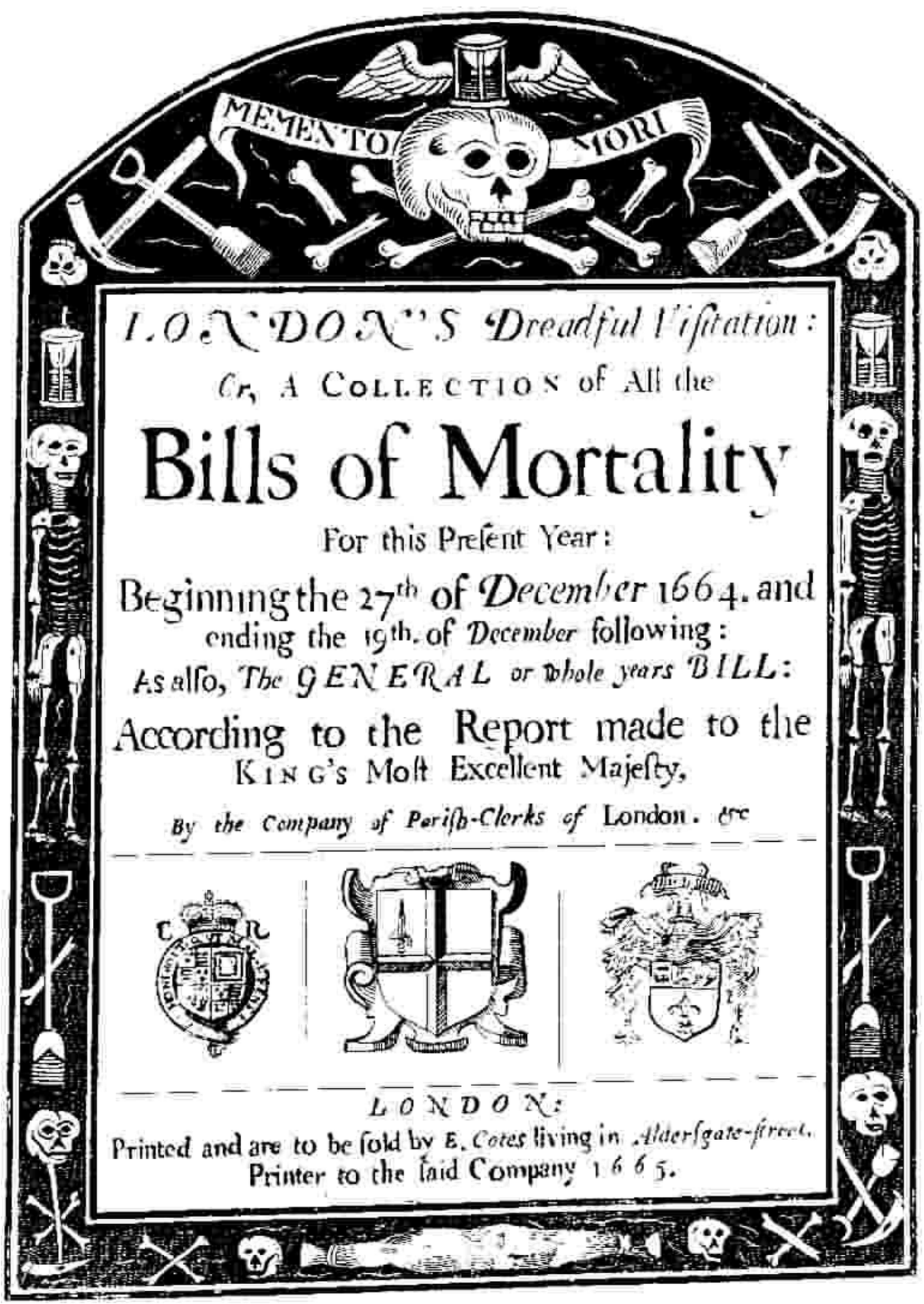
Discussion

Among patients presenting to the ED with acute extremity pain, none of 4 different combination analgesics, 1 of which was opioid-free, resulted in greater pain relief after 2 hours. The largest difference in decline in mean NRS pain score between any 2 treatments was 0.9 at the 2-hour time point, a difference that

In an effort to monitor the incidence of the plague, an injunction issued in 1538 on behalf of Henry VIII required the registration of all burials and christenings in every English Parish

The weekly Bills of Mortality were compiled from these registers, and were initially circulated only to government officials

The Bills were made available to the public in 1594, but were discontinued the next year when the plague abated; publication of the Bills resumed in 1603 when the plague broke out again


MEMENTO MORI

LONDON'S Dreadful Visitation:

Or, A COLLECTION of All the

Bills of Mortality

For this Present Year:

Beginning the 27th of December 1664. and ending the 19th. of December following:

As also, The GENERAL or Whole years BILL:

According to the Report made to the KING's Most Excellent Majesty,

By the Company of Parish-Clerks of London. &c

LONDON:

Printed and are to be sold by E. Cotes living in Aldersgate-street. Printer to the said Company 1665.

John Arbuthnot, a physician to Queen Anne, used the christening records in the London Bills to support an argument for the existence of "Divine Providence"

While Arbuthnot's larger point is certainly beyond the scope of this course, the article is interesting for us because it is widely regarded as the first published statistical test of significance

It also lets us consider (in a more contemporary setting) issues of data representation and the social implications of data collection

II. *An Argument for Divine Providence, taken from the conſtant Regularity obſerv'd in the Births of both Sexes. By Dr.* John Arbuthnott, *Phyſitian in Ordinary to Her Majeſty, and Fellow of the College of Phyſitians and the Royal Society.*

AMong innumerable Footſteps of Divine Providence to be found in the Works of Nature, there is a very remarkable one to be obſerved in the exact Ballance that is maintained, between the Numbers of Men and Women; for by this means it is provided, that the Species may never fail, nor periſh, ſince every Male may have its Female, and of a proportionable Age. This Equality of Males and Females is not the Effect of Chance but Divine Providence, working for a good End, which I thus demonſtrate:

Let there be a Die of Two ſides, M and F, (which denote Croſs and Pile), now to find all the Chances of any determinate Number of ſuch Dice, let the Binome M+F be raiſed to the Power, whoſe Exponent is the Number of Dice given; the Coefficients of the Terms will ſhew all the Chances ſought. For Example, in Two Dice of Two ſides M+F the Chances are $M^2+2\,MF+F^2$, that is, One Chance for M double, One for F double, and Two for M ſingle and F ſingle; in Four ſuch Dice there are Chances $M^4+4\,M^3F+6\,M^2F^2+4\,MF^3+F^4$,

In his argument for "Divine Providence," Arbuthnot considers the gender of babies born in London

While reflecting on the lives of men and women in 1710 England, he notes that men are subject to various "external Accidents" as they "must seek their Food with danger"

For Arbuthnot, these external accidents meant that to maintain a balance between men and women, Divine Providence would arrange for the birth of a larger proportion of boys than girls

Therefore, for Arbuthnot, to demonstrate that boys and girls were not born in equal proportion was to argue in favor of the existence of Divine Providence

the middle Term will not exactly give A's Chances, but his Chances will take in ſome of the Terms next the middle one, and will lean to one ſide or the other. But it is very improbable (if mere Chance govern'd) that they would never reach as far as the Extremities: But this Event is wiſely prevented by the wiſe Oeconomy of Nature; and to judge of the wiſdom of the Contrivance, we muſt obſerve that the external Accidents to which are Males ſubject (who muſt ſeek their Food with danger) do make a great havock of them, and that this loſs exceeds far that of the other Sex, occaſioned by Diſeaſes incident to it, as Experience convinces us. To repair that Loſs, provident Nature, by the Diſpoſal of its wiſe Creator, brings forth more Males than Females; and that in almoſt a conſtant proportion. This appears from the annexed Tables, which contain Obſervations for 82 Years of the Births in *London*. Now, to reduce the Whole to a Calculation, I propoſe this.

Problem. A lays againſt B, that every Year there ſhall be born more Males than Females: To find A's Lot, or the Value of his Expectation.

It is evident from what has been ſaid, that A's Lot for each Year is leſs than $\frac{1}{2}$; (but that the Argument may be ſtronger) let his Lot be equal to $\frac{1}{2}$ for one Year. If he undertakes to do the ſame thing 82 times running, his Lot will be $\overline{\frac{1}{2}}|^{82}$, which will be found eaſily by the Table of Logarithms to be $\frac{1}{4836\ 0000\ 0000\ 0000\ 0000\ 0000}$.

To make his case, Arbuthnot starts with a simple probability model in which the sex of a baby is determined by the toss of a fair coin; that is, we see an “M” with probability 0.5 and “F” with probability 0.5*

Because the underlying mechanism is assumed to be stochastic**, you expect to see fluctuations from year to year in the proportion of boys to girls; some years you will see more boys, in others, more girls

But because the gender of each is determined by the toss of a fair coin, Arbuthnot reasoned that for any given year, the probability that boys outnumbered girls was again 0.5

Arbuthnot then uses the christening records to “test” the hypothesis that boys and girls are born in equal proportion; or, rather that boys outnumber girls in a given year based on the toss of a fair coin

So, what do the data say?

* Arbuthnot actually refers to “a Die of Two sides, M and F
** Stochastic, from the Greek "Στόχος" which means "aim, guess", means of, relating to, or characterized by conjecture and randomness”

Christened.		
Anno.	*Males.*	*Females.*
1629	5218	4683
30	4858	4457
31	4422	4102
32	4994	4590
33	5158	4839
34	5035	4820
35	5106	4928
36	4917	4605
37	4703	4457
38	5359	4952
39	5366	4784
40	5518	5332
41	5470	5200
42	5460	4910
43	4793	4617
44	4107	3997
45	4047	3919
46	3768	3395
47	3796	3536

Christened.		
Anno.	*Males.*	*Females.*
1648	3363	3181
49	3079	2746
50	2890	2722
51	3231	2840
52	3220	2908
53	3196	2959
54	3441	3179
55	3655	3349
56	3668	3382
57	3396	3289
58	3157	3013
59	3209	2781
60	3724	3247
61	4748	4107
62	5216	4803
63	5411	4881
64	6041	5681
65	5114	4858
66	4678	4319

Christened.		
Anno.	*Males.*	*Females.*
1667	5616	5322
68	6073	5560
69	6506	5829
70	6278	5719
71	6449	6061
72	6443	6120
73	6073	5822
74	6113	5738
75	6058	5717
76	6552	5847
77	6423	6203
78	6568	6033
79	6247	6041
80	6548	6299
81	6822	6533
82	6909	6744
83	7577	7158
84	7575	7127
85	7484	7246
86	7575	7119
87	7737	7214
88	7487	7101

Christened.		
Anno.	*Males.*	*Females.*
1689	7604	7167
90	7909	7302
91	7662	7392
92	7602	7316
93	7676	7483
94	6985	6647
95	7263	6713
96	7632	7229
97	8062	7767
98	8426	7626
99	7911	7452
1700	7578	7061
1701	8102	7514
1702	8031	7656
1703	7765	7683
1704	6113	5738
1705	8366	7779
1706	7952	7417
1707	8379	7687
1708	8239	7623
1709	7840	7380
1710	7640	7288

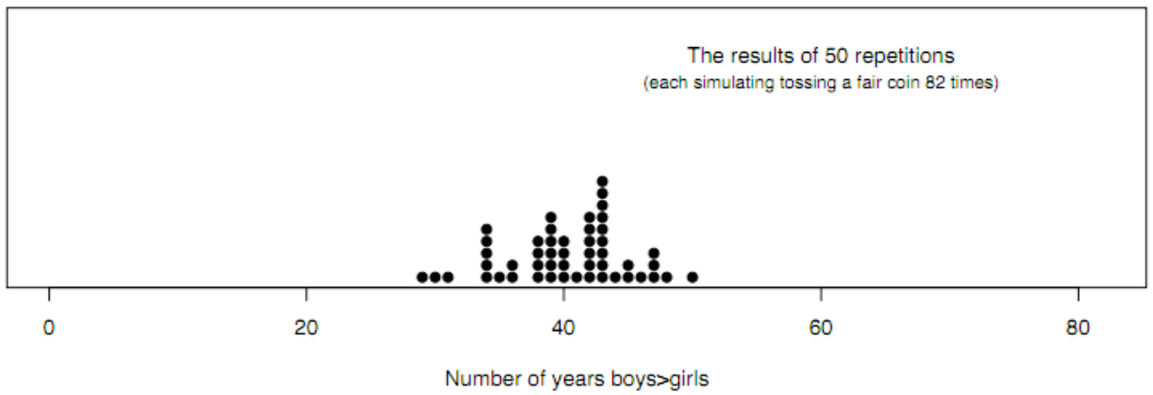
Arbuthnot noticed that in every of the 82 years from 1629 to 1710, there were more boys christened than girls; while this might seem like a compelling enough observation on its own, Arbuthnot takes it farther

His idea was to compare this observation to the probability model he hypothesized for the data; that is, if boys outnumber girls in a given year based on the toss of a fair coin, what is the chance that we see 82 heads in 82 tosses?

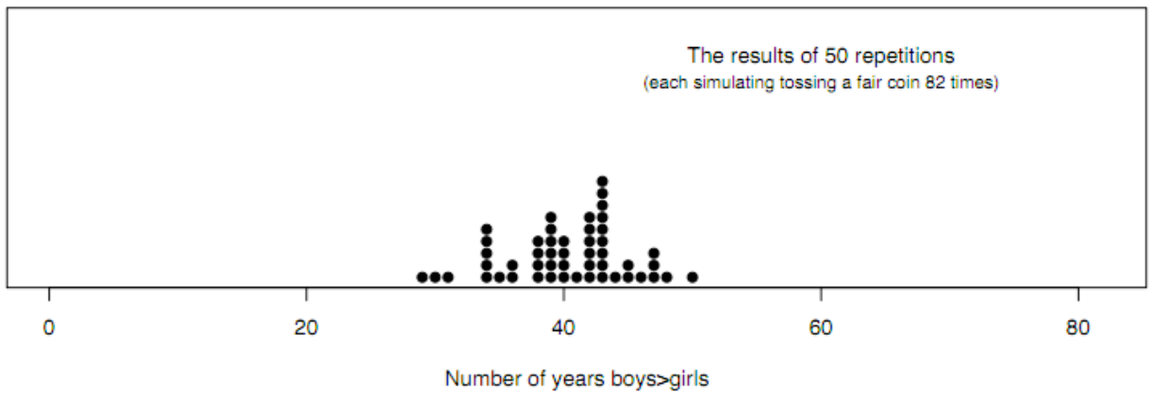
For that matter, what is the chance that we would see any large number, say 70 or 80 heads out of 82 tosses?

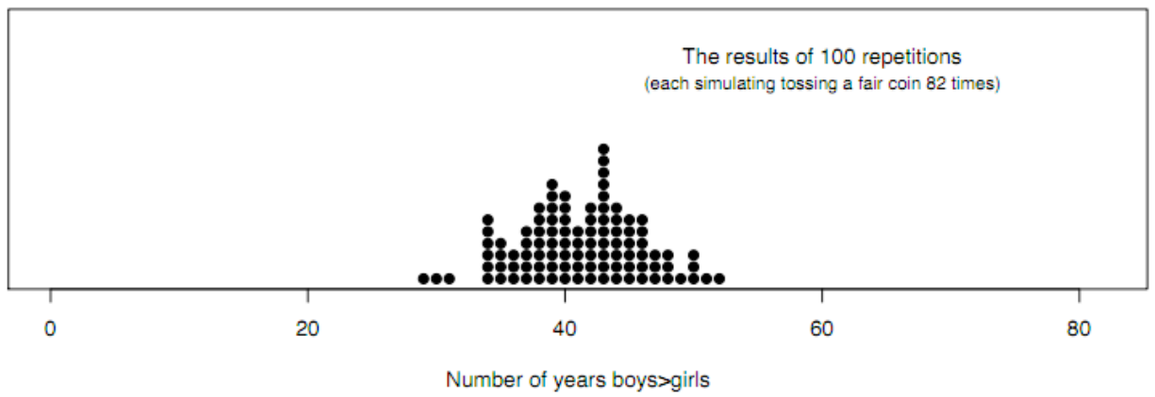

The results of 10 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


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

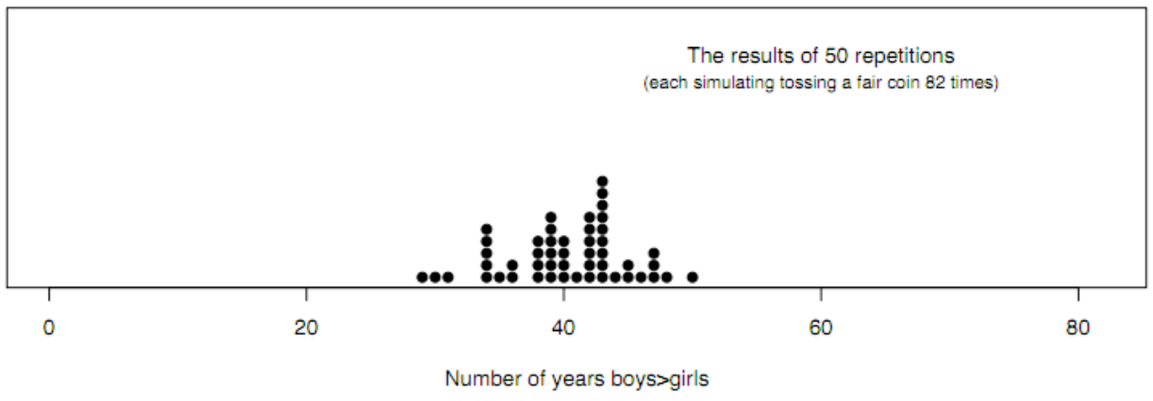

The results of 50 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

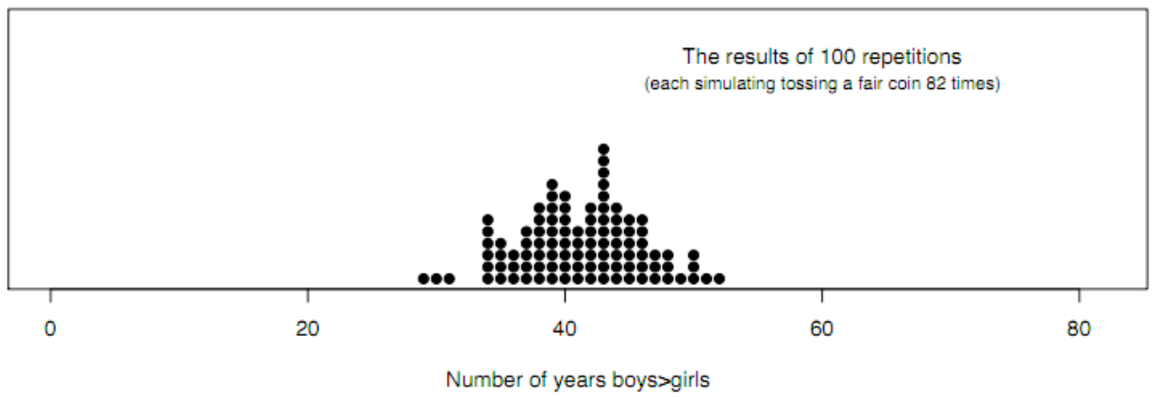

The results of 10 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

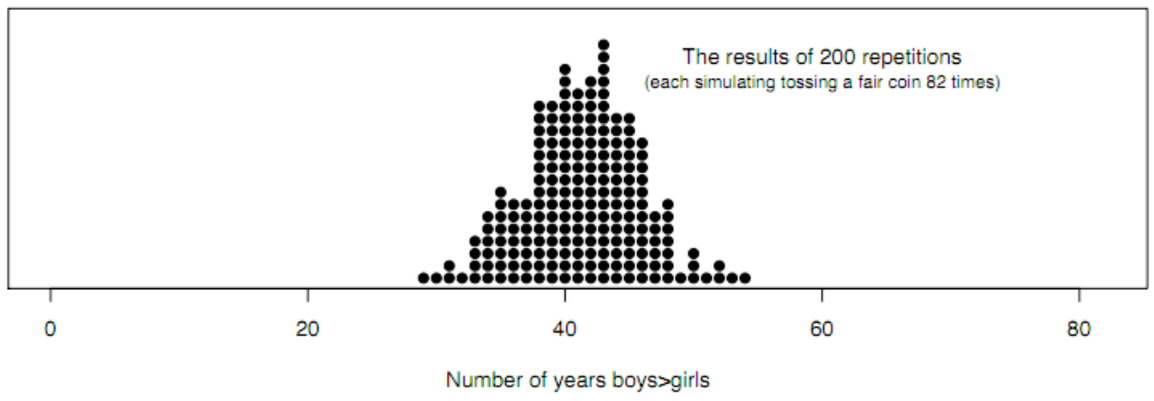

The results of 50 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


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 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

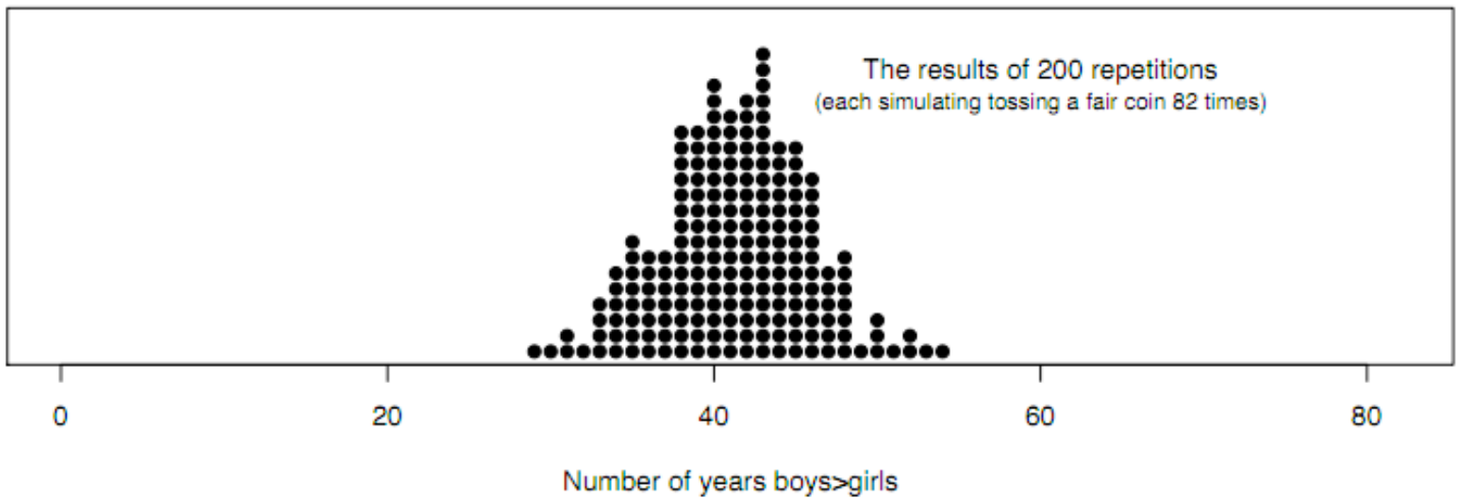

The results of 200 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

In the previous slides, each dot is a different repetition, a different set of 82 simulated coin tosses; looking at these data we can make a few simple observations

1. The simulated data appear to be centered around 41 (since everything is decided by a fair coin toss, we might expect to see fluctuations around 42, half of the 2 tosses we simulated in each repetition)

2. The data we simulated is concentrated primarily between 31 and 51 (that is, an interval with endpoints 41 +/- 10), and values outside this range seem far less likely

What does this say about the actual christening data, with its run of 82 out of 82 heads? What about 80 or even 70 out of 82 heads?

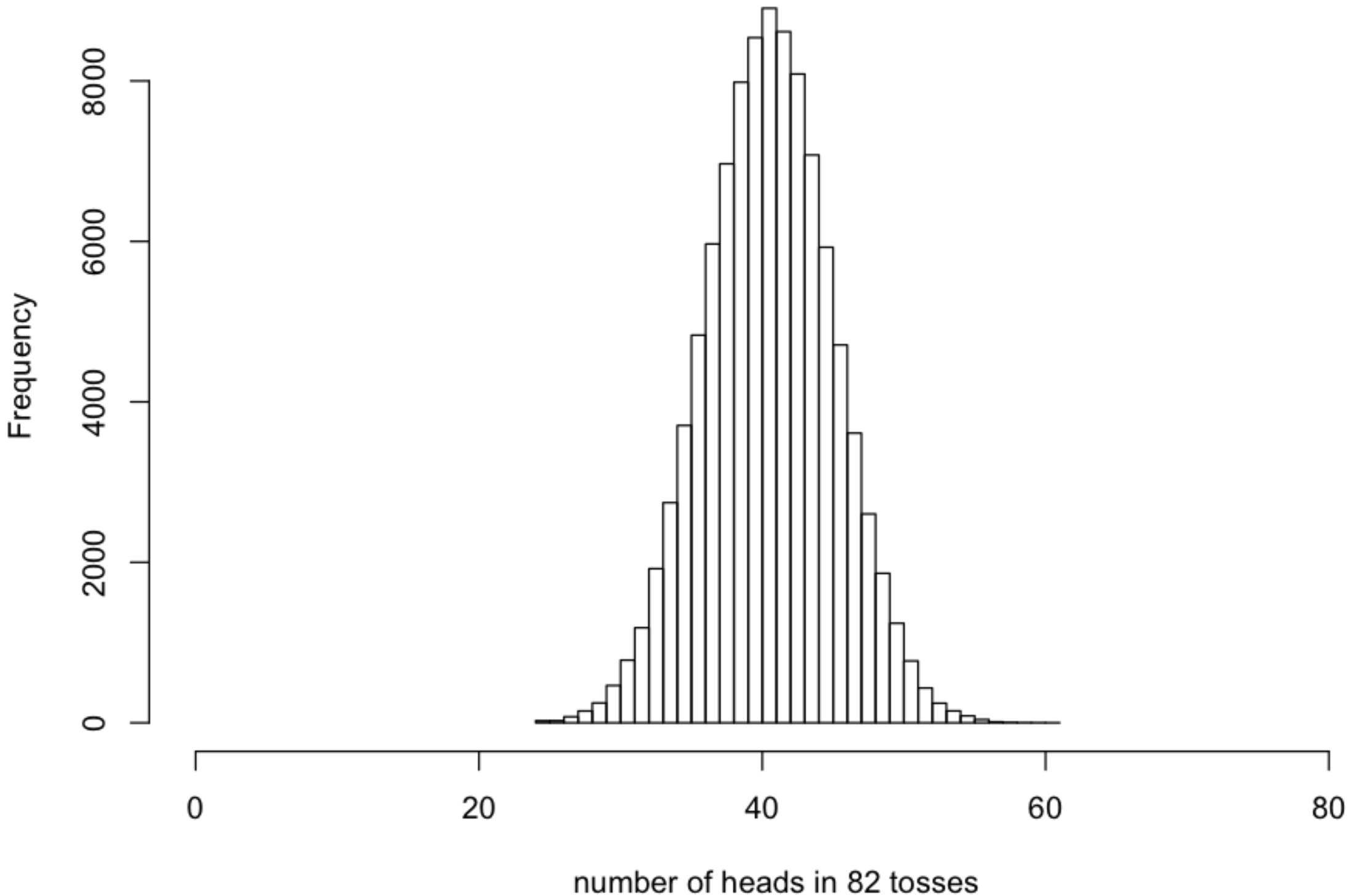


Assuming that boys and girls are born in the same proportion implies the simple ‘coin toss’ probability model; that is, the number of years that boys outnumber girls is the same as counting the number of heads in 82 independent tosses of a fair coin

Therefore, if we believe that the christening data could have been generated by the same mechanism we used for our simulation (82 coin tosses), we would expect that it would lie within the cluster of points we simulated

In the 200 simulations we plotted above, about half (99 out of 200) have 42 or more years with more boys than girls; only about 5% (8 out of 200) have as many as 50 years with more boys than girls; and none have boys outnumbering girls in 60 or more years

We could go on simulating in this way and see if the mound of data we have changes; instead I could cheat a little and use the binomial distribution to write down a mathematical expression for the chance that we see n years or more where boys outnumber girls


100,000 experiments, each tossing 82 fair coins
Frequency
0
2000
4000
6000
8000
0
20
40
60
80
number of heads in 82 tosses

Arbuthnot showed that the chance of seeing boys outnumbering girls year after year is extremely unlikely, approximately 1 in 4,836,000,000,000,000,000,000,000

The idea, then, is that if the hypothesis that boys and girls are born in equal proportion is correct, then the christening data is extraordinarily unlikely; to give you some perspective, the odds of winning the New York Lotto is 1 in 45,000,000

With this calculation in hand, we would feel comfortable abandoning our hypothesis that boys and girls are born in equal proportion

the middle Term will not exactly give A's Chances, but his Chances will take in ſome of the Terms next the middle one, and will lean to one ſide or the other. But it is very improbable (if mere Chance govern'd) that they would never reach as far as the Extremities: But this Event is wiſely prevented by the wiſe Oeconomy of Nature; and to judge of the wiſdom of the Contrivance, we muſt obſerve that the external Accidents to which are Males ſubject (who muſt ſeek their Food with danger) do make a great havock of them, and that this loſs exceeds far that of the other Sex, occaſioned by Diſeaſes incident to it, as Experience convinces us. To repair that Loſs, provident Nature, by the Diſpoſal of its wiſe Creator, brings forth more Males than Females; and that in almoſt a conſtant proportion. This appears from the annexed Tables, which contain Obſervations for 82 Years of the Births in *London*. Now, to reduce the Whole to a Calculation, I propoſe this.

Problem. A lays againſt B, that every Year there ſhall be born more Males than Females: To find A's Lot, or the Value of his Expectation.

It is evident from what has been ſaid, that A's Lot for each Year is leſs than $\frac{1}{2}$; (but that the Argument may be ſtronger) let his Lot be equal to $\frac{1}{2}$ for one Year. If he undertakes to do the ſame thing 82 times running, his Lot will be $\overline{\frac{1}{2}}|^{82}$, which will be found eaſily by the Table of Logarithms to be $\frac{1}{4836\ 0000\ 0000\ 0000\ 0000\ 0000}$.

Significance Testing

With this example, we have the basic ingredients of how significance testing works.

We establish a **null hypothesis,** plausible statement (a model or scenario) which might explain some pattern in a given set of data. This hypothesis is made for the purposes of argument — a good null hypothesis is a statement that would be interesting to reject. Think of it as a kind of devil's advocate (or maybe straw man is a better reference as the test was about divine intervention, after all).

We then define **a test statistic**, some quantity calculated from our data that is used to evaluate how compatible the results are with those expected under the null hypothesis (if the hypothesized statement - or model or scenario - was true)

We then simulate the values of the test statistic using the null hypothesis. In our analysis of Arbuthnot's hypothesis, that meant simulating a series of data sets assuming the null hypothesis is true and there is a 50/50 chance of boys outnumbering girls in a given year. For each data set we compute the test statistic. The ensemble of simulated test statistics is often called a **null distribution**.

Finally, we compare the value of the test statistic we computed for our data to the values we obtained by simulation — If they are very different, we have evidence that the null hypothesis is wrong. The chance that we see a value of the test statistic in simulations as or more extreme than what we computed from our data is referred to as the **P-value** of the test.

R.A. Fisher proposed this measure to express the weight of evidence against a null hypothesis — the smaller the value, the stronger the evidence. Fisher, however, believed that it should be combined with other sources of information as you reason about the phenomenon you were studying.

Significance Testing

P-values and significance testing comes from so-called **frequentist statistics**. Under this framework, probability reveals itself through repeated experiments.

For example, if we want to know the probability of a coin landing “heads,” we could toss it many, many times and see what fraction of times we see heads. In the long run, the probability we’re after will emerge.

This reliance on the idea of repeated experiments can be a problem — researchers who make decisions based on their data can break this framework. Choices made that seem obvious with one set of data might be made differently with a different set of outcomes.

This is just one of many ways of thinking about probability — the basic mathematics of probability remains the same, the interpretation of what the basic quantity means can be different.

There are a few obvious questions facing practitioners, the first of which involves evaluating the information provided by a P-value — I**s there a rule which helps you decide when you should "reject" the null hypothesis**, or, rather, decide that it's not true?

Fisher wrote: If [the P-value] is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. **We shall not often be astray if we draw a conventional line at 0.05**...." (Fisher 1950) — and certainly in his own work on agricultural field trials, used thresholds of 0.05 and 0.01 as guides to "reject" a null hypothesis

Still, Fisher believed that t**he individual researcher should interpret a P-value** (a value of 0.05 might not lead to either belief or disbelief in the null, but to a decision to conduct another experiment); **he wrote that the rigid use of thresholds was the "result of applying mechanically rules laid down in advance; no thought is given to the particular case, and the tester's state of mind, or his capacity for learning, is inoperative."** (Fisher 1955, p.73-4).

“No test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint.

Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong.”

Why P=0.05?

The standard level of significance used to justify a claim of a statistically significant effect is 0.05. For better or worse, the term *statistically significant* has become synonymous with $P \leq 0.05$.

There are many theories and stories to account for the use of P=0.05 to denote statistical significance. All of them trace the practice back to the influence of R.A. Fisher. In 1914, Karl Pearson published his *Tables for Statisticians & Biometricians*. For each distribution, Pearson gave the value of P for a series of values of the random variable. When Fisher published *Statistical Methods for Research Workers* (SMRW) in 1925, he included tables that gave the value of the random variable for specially selected values of P. SMRW was a major influence through the 1950s. The same approach was taken for Fisher's *Statistical Tables for Biological, Agricultural, and Medical Research*, published in 1938 with Frank Yates. Even today, Fisher's tables are widely reproduced in standard statistical texts.

Fisher's tables were compact. Where Pearson described a distribution in detail, Fisher summarized it in a single line in one of his tables making them more suitable for inclusion in standard reference works*. However, Fisher's tables would change the way the information could be used. While Pearson's tables provide probabilities for a wide range of values of a statistic, Fisher's tables only bracket the probabilities between coarse bounds.

The impact of Fisher's tables was profound. Through the 1960s, it was standard practice in many fields to report summaries with one star attached to indicate $P \leq 0.05$ and two stars to indicate $P \leq 0.01$, Occasionally, three starts were used to indicate $P \leq 0.001$.

Still, why should the value 0.05 be adopted as the universally accepted value for statistical significance? Why has this approach to hypothesis testing not been supplanted in the intervening three-quarters of a century?

It was Fisher who suggested giving 0.05 its special status. Page 44 of the 13th edition of SMRW, describing the standard normal distribution, states

> The value for which P=0.05, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available. Small effects will still escape notice if the data are insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty.

Similar remarks can be found in Fisher (1926, 504).

> ... it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials."...
>
> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this

level of significance.

However, Fisher's writings might be described as inconsistent. On page 80 of SMRW, he offers a more flexible approach

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. Belief in the hypothesis as an accurate representation of the population sampled is confronted by the logical disjunction: *Either* the hypothesis is untrue, *or* the value of χ^2 has attained by chance an exceptionally high value. The actual value of P obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis. A value of χ^2 exceeding the 5 per cent. point is seldom to be disregarded.

These apparent inconsistencies persist when Fisher dealt with specific examples. On page 137 of SMRW, Fisher suggests that values of P slightly less than 0.05 are are not conclusive.

> [T]he results of *t* shows that P is between .02 and .05.
>
> The result must be judged significant, though barely so; in view of the data we cannot ignore the possibility that on this field, and in conjunction with the other manures used, nitrate of soda has conserved the fertility better than sulphate of ammonia; the data do not, however, demonstrate this point beyond the possibility of doubt.

On pages 139-140 of SMRW, Fisher dismisses a value greater than 0.05 but less than 0.10.

> [W]e find...*t*=1.844 [with 13 df, P = 0.088]. The difference between the regression coefficients, though relatively large, cannot be regarded as significant. There is not sufficient evidence to assert that culture B was growing more rapidly than culture A.

while in Fisher [19xx, p 516] he is willing pay attention to a value not much different.

> ...P=.089. Thus a larger value of χ^2 would be obtained by chance only 8.9 times in a hundred, from a series of values in random order. There is thus some reason to suspect that the distribution of rainfall in successive years is not wholly fortuitous, but that some slowly changing cause is liable to affect in the same direction the rainfall of a number of consecutive years.

Yet *in the same paper* another such value is dismissed!

> [paper 37, p 535] ...P=.093 from Elderton's Table, showing that although there are signs of association among the rainfall distribution values, such association, if it exists, is not strong enough to show up significantly in a series of about 60 values.

Part of the reason for the apparent inconsistency is the way Fisher viewed P values. When Neyman and Pearson proposed using P values as absolute cutoffs in their style of fixed-level testing, Fisher disagreed strenuously. Fisher viewed P values more as measures of the evidence against a hypotheses, as reflected

in the quotation from page 80 of SMRW above and this one from Fisher (1956, p 41-42)

> The attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to hypothetical frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who "rejects" a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is incorrect he will never be mistaken in rejection. This inequality statement can therefore be made. However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. Further, the calculation is based solely on a hypothesis, which, in the light of the evidence, is often not believed to be true at all, so that the actual probability of erroneous decision, supposing such a phrase to have any meaning, may be much less than the frequency specifying the level of significance.

Still, we continue to use P values nearly as absolute cutoffs but with an eye on rethinking our position for values close to 0.05**. Why have we continued doing things this way? A procedure such as this has an important function as a gatekeeper and filter--it lets signals pass while keeping the noise down. The 0.05 level guarantees the literature will be spared 95% of potential reports of effects where there are none.

For such procedures to be effective, it is essential ther be a tacit agreement among researchers to use them in the same way. Otherwise, individuals would modify the procedure to suit their own purposes until the procedure became valueless. As Bross (1971) remarks,

> Anyone familiar with certain areas of the scientific literature will be well aware of the need for curtailing language-games. Thus if there were no 5% level firmly established, then some persons would stretch the level to 6% or 7% to prove their point. Soon others would be stretching to 10% and 15% and the jargon would become meaningless. Whereas nowadays a phrase such as *statistically significant difference* provides some assurance that the results are not merely a manifestation of sampling variation, the phrase would mean very little if everyone played language-games. To be sure, there are always a few folks who fiddle with significance levels--who will switch from two-tailed to one-tailed tests or from one significance test to another in an effort to get positive results. However such gamesmanship is severely frowned upon and is rarely practiced by persons who are *native speakers* of fact-limited scientific languages--it is the mark of an amateur.

Bross points out that the continued use of P=0.05 as a convention tells us a good deal about its practical value.

> The continuing usage of the 5% level is indicative of another important practical point: it is a feasible level at which to do research work. In other words, if the 5% level is used, then in most experimental situations it is feasible (though not necessarily easy) to set up a study which will have a fair chance of picking up those effects which are large enough to be of scientific interest. If past experience in actual applications had not shown this feasibility, the convention would not have been useful to scientists and it would not have stayed in their languages. For suppose that the 0.1% level had been proposed. This level is rarely attainable in biomedical experimentation. If it were made a prerequisite for reporting positive results,

> there would be very little to report. Hence from the standpoint of communication the level would have been of little value and the evolutionary process would have eliminated it.

The fact that many aspects of statistical practice in this regard *have* changed gives Bross's argument additional weight. Once (mainframe) computers became available and it was possible to calculate precise P values on demand, standard practice quickly shifted to reporting the P values themselves rather than merely whether or not they were less than 0.05. The value of 0.02 suggested by Fisher as a *strong indication that the hypothesis fails to account for the whole of the facts has been replaced by 0.01. However, science has seen fit to continue letting 0.05 retain its special status denoting statistical significance.*

*Fisher may have had additional reasons for developing a new way to table commonly used distribution functions. Jack Good, on page 513 of the discussion section of Bross (1971), says, "Kendall mentioned that Fisher produced the tables of significance levels to save space and to avoid copyright problems with Karl Pearson, whom he disliked."

**It is worth noting that when researchers worry about P values close to 0.05, they worry about values slightly greater than 0.05 and why they deserve attention nonetheless. I cannot recall published research downplaying P values less than 0.05. Fisher's comment cited above from page 137 of SMRW is a rare exception.

References

- *Bross IDJ (1971), "Critical Levels, Statistical Language and Scientific Inference," in Godambe VP and Sprott (eds) Foundations of Statistical Inference. Toronto: Holt, Rinehart & Winston of Canada, Ltd.*
- *Fisher RA (1956), Statistical Methods and Scientific Inference New York: Hafner*
- *Fisher RA (1926), "The Arrangement of Field Experiments," Journal of the Ministry of Agriculture of Great Britain, 33, 503-513.*
- *Fisher RA (19xx), "On the Influence of Rainfall on the Yield of Wheat at Rothamstead,"*

Gerard E. Dallal
Last modified: 10/19/2003 18:06:34.

TABLES FOR STATISTICIANS AND BIOMETRICIANS

EDITED BY

KARL PEARSON, F.R.S.

GALTON PROFESSOR, UNIVERSITY OF LONDON

ISSUED WITH ASSISTANCE FROM THE GRANT MADE BY
THE WORSHIPFUL COMPANY OF DRAPERS TO THE
BIOMETRIC LABORATORY
UNIVERSITY COLLEGE
LONDON

Cambridge :
at the University Press
1914

TABLE V. *Probable Errors of Means and Standard Deviations.*

n	χ_1	χ_2	n	χ_1	χ_2	n	χ_1	χ_2
151	·05489	·03881	*201*	·04757	·03364	*251*	·04257	·03010
152	·05471	·03868	*202*	·04746	·03356	*252*	·04249	·03004
153	·05453	·03856	*203*	·04734	·03347	*253*	·04240	·02998
154	·05435	·03843	*204*	·04722	·03339	*254*	·04232	·02993
155	·05418	·03831	*205*	·04711	·03331	*255*	·04224	·02987
156	·05400	·03819	*206*	·04699	·03323	*256*	·04216	·02981
157	·05383	·03806	*207*	·04688	·03315	*257*	·04207	·02975
158	·05366	·03794	*208*	·04677	·03307	*258*	·04199	·02969
159	·05349	·03782	*209*	·04666	·03299	*259*	·04191	·02964
160	·05332	·03771	*210*	·04654	·03291	*260*	·04183	·02958
161	·05316	·03759	*211*	·04643	·03283	*261*	·04175	·02952
162	·05299	·03747	*212*	·04632	·03276	*262*	·04167	·02947
163	·05283	·03736	*213*	·04622	·03268	*263*	·04159	·02941
164	·05267	·03724	*214*	·04611	·03260	*264*	·04151	·02935
165	·05251	·03713	*215*	·04600	·03253	*265*	·04143	·02930
166	·05235	·03702	*216*	·04589	·03245	*266*	·04136	·02924
167	·05219	·03691	*217*	·04579	·03238	*267*	·04128	·02919
168	·05204	·03680	*218*	·04568	·03230	*268*	·04120	·02913
169	·05188	·03669	*219*	·04558	·03223	*269*	·04112	·02908
170	·05173	·03658	*220*	·04547	·03216	*270*	·04105	·02903
171	·05158	·03647	*221*	·04537	·03208	*271*	·04097	·02897
172	·05143	·03637	*222*	·04527	·03201	*272*	·04090	·02892
173	·05128	·03626	*223*	·04517	·03194	*273*	·04082	·02887
174	·05113	·03616	*224*	·04507	·03187	*274*	·04075	·02881
175	·05099	·03605	*225*	·04497	·03180	*275*	·04067	·02876
176	·05084	·03595	*226*	·04487	·03173	*276*	·04060	·02871
177	·05070	·03585	*227*	·04477	·03166	*277*	·04053	·02866
178	·05056	·03575	*228*	·04467	·03159	*278*	·04045	·02860
179	·05041	·03565	*229*	·04457	·03152	*279*	·04038	·02855
180	·05027	·03555	*230*	·04447	·03145	*280*	·04031	·02850
181	·05013	·03545	*231*	·04438	·03138	*281*	·04024	·02845
182	·05000	·03535	*232*	·04428	·03131	*282*	·04017	·02840
183	·04986	·03526	*233*	·04419	·03125	*283*	·04009	·02835
184	·04972	·03516	*234*	·04409	·03118	*284*	·04002	·02830
185	·04959	·03507	*235*	·04400	·03111	*285*	·03995	·02825
186	·04946	·03497	*236*	·04391	·03105	*286*	·03988	·02820
187	·04932	·03488	*237*	·04381	·03098	*287*	·03981	·02815
188	·04919	·03478	*238*	·04372	·03092	*288*	·03974	·02810
189	·04906	·03469	*239*	·04363	·03085	*289*	·03968	·02806
190	·04893	·03460	*240*	·04354	·03079	*290*	·03961	·02801
191	·04880	·03451	*241*	·04345	·03172	*291*	·03954	·02796
192	·04868	·03442	*242*	·04336	·03066	*292*	·03947	·02791
193	·04855	·03433	*243*	·04327	·03060	*293*	·03940	·02786
194	·04843	·03424	*244*	·04318	·03053	*294*	·03934	·02782
195	·04830	·03415	*245*	·04309	·03047	*295*	·03927	·02777
196	·04818	·03407	*246*	·04300	·03041	*296*	·03920	·02772
197	·04806	·03398	*247*	·04292	·03035	*297*	·03913	·02767
198	·04793	·03389	*248*	·04283	·03029	*298*	·03907	·02763
199	·04781	·03381	*249*	·04274	·03022	*299*	·03901	·02758
200	·04769	·03372	*250*	·04266	·03016	*300*	·03894	·02754

TABLE V. *Probable Errors of Means and Standard Deviations.*

n	χ_1	χ_2	n	χ_1	χ_2	n	χ_1	χ_2
301	·03888	·02749	*351*	·03600	·02546	*401*	·03368	·02382
302	·03881	·02744	*352*	·03595	·02542	*402*	·03364	·02379
303	·03875	·02740	*353*	·03590	·02538	*403*	·03360	·02376
304	·03868	·02735	*354*	·03585	·02535	*404*	·03356	·02373
305	·03862	·02731	*355*	·03580	·02531	*405*	·03352	·02370
306	·03856	·02726	*356*	·03575	·02528	*406*	·03347	·02367
307	·03850	·02722	*357*	·03570	·02524	*407*	·03343	·02364
308	·03843	·02718	*358*	·03565	·02521	*408*	·03339	·02361
309	·03837	·02713	*359*	·03560	·02517	*409*	·03335	·02358
310	·03831	·02709	*360*	·03555	·02514	*410*	·03331	·02355
311	·03825	·02704	*361*	·03550	·02510	*411*	·03327	·02353
312	·03819	·02700	*362*	·03545	·02507	*412*	·03323	·02350
313	·03812	·02696	*363*	·03540	·02503	*413*	·03319	·02347
314	·03806	·02692	*364*	·03535	·02500	*414*	·03315	·02344
315	·03800	·02687	*365*	·03530	·02496	*415*	·03311	·02341
316	·03794	·02683	*366*	·03526	·02493	*416*	·03307	·02338
317	·03788	·02679	*367*	·03521	·02490	*417*	·03303	·02336
318	·03782	·02675	*368*	·03516	·02486	*418*	·03299	·02333
319	·03776	·02670	*369*	·03511	·02483	*419*	·03295	·02330
320	·03771	·02666	*370*	·03507	·02479	*420*	·03291	·02327
321	·03765	·02662	*371*	·03502	·02476	*421*	·03287	·02324
322	·03759	·02658	*372*	·03497	·02473	*422*	·03283	·02322
323	·03753	·02654	*373*	·03492	·02469	*423*	·03279	·02319
324	·03747	·02650	*374*	·03488	·02466	*424*	·03276	·02316
325	·03741	·02646	*375*	·03483	·02463	*425*	·03272	·02313
326	·03736	·02642	*376*	·03478	·02460	*426*	·03268	·02311
327	·03730	·02637	*377*	·03474	·02456	*427*	·03264	·02308
328	·03724	·02633	*378*	·03469	·02453	*428*	·03260	·02305
329	·03719	·02629	*379*	·03465	·02450	*429*	·03256	·02303
330	·03713	·02625	*380*	·03460	·02447	*430*	·03253	·02300
331	·03707	·02621	*381*	·03456	·02443	*431*	·03249	·02297
332	·03702	·02618	*382*	·03451	·02440	*432*	·03245	·02295
333	·03696	·02614	*383*	·03446	·02437	*433*	·03241	·02292
334	·03691	·02610	*384*	·03442	·02434	*434*	·03238	·02289
335	·03685	·02606	*385*	·03438	·02431	*435*	·03234	·02287
336	·03680	·02602	*386*	·03433	·02428	*436*	·03230	·02284
337	·03674	·02598	*387*	·03429	·02424	*437*	·03227	·02281
338	·03669	·02594	*388*	·03424	·02421	*438*	·03223	·02279
339	·03663	·02590	*389*	·03420	·02418	*439*	·03219	·02276
340	·03658	·02587	*390*	·03415	·02415	*440*	·03216	·02274
341	·03653	·02583	*391*	·03411	·02412	*441*	·03212	·02271
342	·03647	·02579	*392*	·03407	·02409	*442*	·03208	·02269
343	·03642	·02575	*393*	·03402	·02406	*443*	·03205	·02266
344	·03637	·02571	*394*	·03398	·02403	*444*	·03201	·02263
345	·03631	·02568	*395*	·03394	·02400	*445*	·03197	·02261
346	·03626	·02564	*396*	·03389	·02397	*446*	·03194	·02258
347	·03621	·02560	*397*	·03385	·02394	*447*	·03190	·02256
348	·03616	·02557	*398*	·03381	·02391	*448*	·03187	·02253
349	·03610	·02553	*399*	·03377	·02388	*449*	·03183	·02251
350	·03605	·02549	*400*	·03372	·02385	*450*	·03180	·02248

Original Investigation

Evolution of Reporting *P* Values in the Biomedical Literature, 1990-2015

David Chavalarias, PhD; Joshua David Wallach, BA; Alvin Ho Ting Li, BHSc; John P. A. Ioannidis, MD, DSc

IMPORTANCE The use and misuse of *P* values has generated extensive debates.

OBJECTIVE To evaluate in large scale the *P* values reported in the abstracts and full text of biomedical research articles over the past 25 years and determine how frequently statistical information is presented in ways other than *P* values.

DESIGN Automated text-mining analysis was performed to extract data on *P* values reported in 12 821 790 MEDLINE abstracts and in 843 884 abstracts and full-text articles in PubMed Central (PMC) from 1990 to 2015. Reporting of *P* values in 151 English-language core clinical journals and specific article types as classified by PubMed also was evaluated. A random sample of 1000 MEDLINE abstracts was manually assessed for reporting of *P* values and other types of statistical information; of those abstracts reporting empirical data, 100 articles were also assessed in full text.

MAIN OUTCOMES AND MEASURES *P* values reported.

RESULTS Text mining identified 4 572 043 *P* values in 1 608 736 MEDLINE abstracts and 3 438 299 *P* values in 385 393 PMC full-text articles. Reporting of *P* values in abstracts increased from 7.3% in 1990 to 15.6% in 2014. In 2014, *P* values were reported in 33.0% of abstracts from the 151 core clinical journals (n = 29 725 abstracts), 35.7% of meta-analyses (n = 5620), 38.9% of clinical trials (n = 4624), 54.8% of randomized controlled trials (n = 13 544), and 2.4% of reviews (n = 71 529). The distribution of reported *P* values in abstracts and in full text showed strong clustering at *P* values of .05 and of .001 or smaller. Over time, the "best" (most statistically significant) reported *P* values were modestly smaller and the "worst" (least statistically significant) reported *P* values became modestly less significant. Among the MEDLINE abstracts and PMC full-text articles with *P* values, 96% reported at least 1 *P* value of .05 or lower, with the proportion remaining steady over time in PMC full-text articles. In 1000 abstracts that were manually reviewed, 796 were from articles reporting empirical data; *P* values were reported in 15.7% (125/796 [95% CI, 13.2%-18.4%]) of abstracts, confidence intervals in 2.3% (18/796 [95% CI, 1.3%-3.6%]), Bayes factors in 0% (0/796 [95% CI, 0%-0.5%]), effect sizes in 13.9% (111/796 [95% CI, 11.6%-16.5%]), other information that could lead to estimation of *P* values in 12.4% (99/796 [95% CI, 10.2%-14.9%]), and qualitative statements about significance in 18.1% (181/1000 [95% CI, 15.8%-20.6%]); only 1.8% (14/796 [95% CI, 1.0%-2.9%]) of abstracts reported at least 1 effect size and at least 1 confidence interval. Among 99 manually extracted full-text articles with data, 55 reported *P* values, 4 presented confidence intervals for all reported effect sizes, none used Bayesian methods, 1 used false-discovery rates, 3 used sample size/power calculations, and 5 specified the primary outcome.

CONCLUSIONS AND RELEVANCE In this analysis of *P* values reported in MEDLINE abstracts and in PMC articles from 1990-2015, more MEDLINE abstracts and articles reported *P* values over time, almost all abstracts and articles with *P* values reported statistically significant results, and, in a subgroup analysis, few articles included confidence intervals, Bayes factors, or effect sizes. Rather than reporting isolated *P* values, articles should include effect sizes and uncertainty metrics.

JAMA. 2016;315(11):1141-1148. doi:10.1001/jama.2016.1952
Corrected on May 12, 2016.

Editorial page 1113

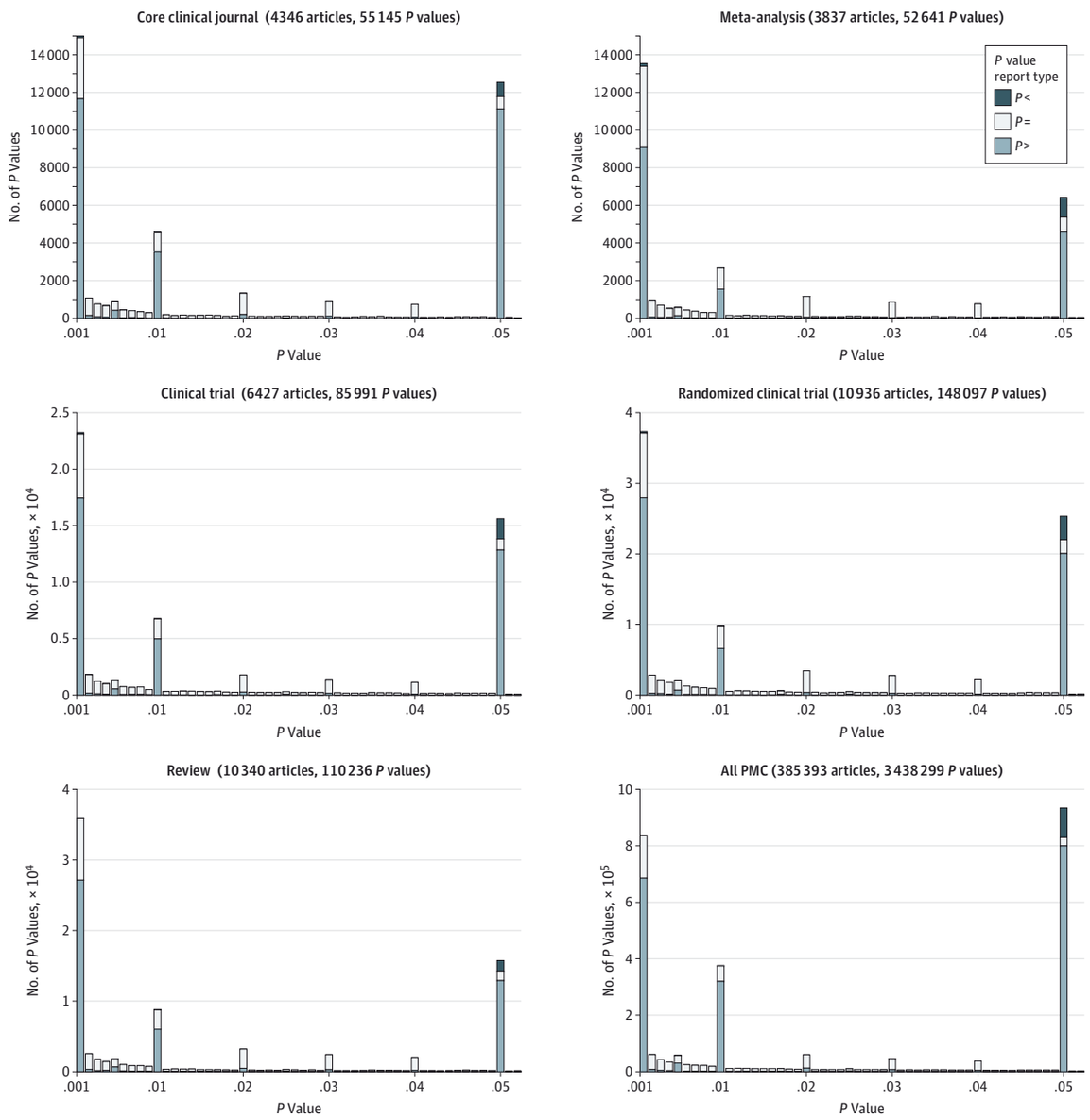
Supplemental content at jama.com

CME Quiz at jamanetworkcme.com

Author Affiliations: Centre d'Analyse et de Mathématiques Sociales (CAMS), EHESS–CNRS UMR8557 and Complex Systems Institute of Paris Île-de-France (ISC-PIF, UPS3611), Paris, France (Chavalarias); Departments of Health Research and Policy and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California (Wallach); Department of Epidemiology and Biostatistics, Western University, London, Ontario, Canada (Li); Departments of Medicine, Health Research and Policy, and Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California (Ioannidis).

Corresponding Author: John P. A. Ioannidis, MD, DSc, Departments of Medicine, Health Research and Policy, and Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, 1265 Welch Rd, MSOB X306, Stanford, CA 94305 (jioannid@stanford.edu).

Figure 2. Distribution of *P* Values in 385 393 PMC Full-Text Articles That Have Abstracts



Numerical values not shown (>.05) represent 17.41% of the total (598 611 *P* values). There are 50 bins shown, each with width .001.

(corresponding to *P* = .02-.03). Thus, even the "worst" reported *P* values still remained mostly within the range of nominally statistically significant results (*P* < .05). In the PMC full-text articles, at the end of the study period the average $-\log_{10}$ best *P* value reached approximately 2.57 overall, ie, *P* = .0027 (eFigure 7 in the Supplement).

In addition, the proportion of *P* values reported in MEDLINE abstracts as inequalities (eg, "*P* <" or "*P* ≤") decreased over time (a larger percentage of "*P* =" values were reported, eFigure 8 in the Supplement). When analyses were limited to precise *P* values ("*P* ="), at the end of the study period, across MEDLINE abstracts the mean $-\log_{10}$ best reported *P* value was 2.2 (corresponding to *P* = .006) and the mean $-\log_{10}$ worst reported *P* value was 1.45 (corresponding to *P* = .035), whereas the mean $-\log_{10}$ best reported *P* value in PMC full-text articles was 2.42 (corresponding to *P* = .004) (eFigures 9-11 in the Supplement).

Frequency of Reporting of at Least 1 *P* Value of .05 or Less

Across the 1 608 736 MEDLINE abstracts with any *P* value reported, 96.0% reported at least 1 *P* value that was .05 or less, with a slight decrease over time from 97.9% in 1990 to 95.0% in 2014 (**Figure 3**A). Similarly high proportions of *P* values of .05 or less

ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice "emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean."

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

"Over time it appears the *p*-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as '*p*-hacking' and 'data dredging' that emphasize the search for small *p*-values over other statistical and scientific reasoning."

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

The statement has short paragraphs elaborating on each principle.

In light of misuses of and misconceptions concerning *p*-values, the statement notes that statisticians often supplement or even replace *p*-values with other approaches. These include methods "that emphasize estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretic modeling and false discovery rates."

"The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades," Utts said. "But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues."

"The issues involved in statistical inference are difficult because inference itself is challenging," Wasserstein said. He noted that more than a dozen discussion papers are being published in the ASA journal *The American Statistician* with the statement to provide more perspective on this broad and complex topic. "What we hope will follow is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research."

About the American Statistical Association

The ASA is the world's largest community of statisticians and the oldest continuously operating professional science society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical

Title: Redefine Statistical Significance

Authors: Daniel J. Benjamin[1]*, James O. Berger[2], Magnus Johannesson[3]*, Brian A. Nosek[4,5], E.-J. Wagenmakers[6], Richard Berk[7, 10], Kenneth A. Bollen[8], Björn Brembs[9], Lawrence Brown[10], Colin Camerer[11], David Cesarini[12, 13], Christopher D. Chambers[14], Merlise Clyde[2], Thomas D. Cook[15,16], Paul De Boeck[17], Zoltan Dienes[18], Anna Dreber[3], Kenny Easwaran[19], Charles Efferson[20], Ernst Fehr[21], Fiona Fidler[22], Andy P. Field[18], Malcolm Forster[23], Edward I. George[10], Richard Gonzalez[24], Steven Goodman[25], Edwin Green[26], Donald P. Green[27], Anthony Greenwald[28], Jarrod D. Hadfield[29], Larry V. Hedges[30], Leonhard Held[31], Teck Hua Ho[32], Herbert Hoijtink[33], James Holland Jones[39,40], Daniel J. Hruschka[34], Kosuke Imai[35], Guido Imbens[36], John P.A. Ioannidis[37], Minjeong Jeon[38], Michael Kirchler[41], David Laibson[42], John List[43], Roderick Little[44], Arthur Lupia[45], Edouard Machery[46], Scott E. Maxwell[47], Michael McCarthy[48], Don Moore[49], Stephen L. Morgan[50], Marcus Munafó[51, 52], Shinichi Nakagawa[53], Brendan Nyhan[54], Timothy H. Parker[55], Luis Pericchi[56], Marco Perugini[57], Jeff Rouder[58], Judith Rousseau[59], Victoria Savalei[60], Felix D. Schönbrodt[61], Thomas Sellke[62], Betsy Sinclair[63], Dustin Tingley[64], Trisha Van Zandt[65], Simine Vazire[66], Duncan J. Watts[67], Christopher Winship[68], Robert L. Wolpert[2], Yu Xie[69], Cristobal Young[70], Jonathan Zinman[71], Valen E. Johnson[72]*

Affiliations:

[1]Center for Economic and Social Research and Department of Economics, University of Southern California, Los Angeles, CA 90089-3332, USA.

[2]Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA.

[3]Department of Economics, Stockholm School of Economics, SE-113 83 Stockholm, Sweden.

[4]University of Virginia, Charlottesville, VA 22908, USA.

[5]Center for Open Science, Charlottesville, VA 22903, USA.

[6]University of Amsterdam, Department of Psychology, 1018 VZ Amsterdam, The Netherlands.

[7]University of Pennsylvania, School of Arts and Sciences and Department of Criminology, Philadelphia, PA 19104-6286, USA.

[8]University of North Carolina Chapel Hill, Department of Psychology and Neuroscience, Department of Sociology, Chapel Hill, NC 27599-3270, USA.

[9]Institute of Zoology - Neurogenetics, Universität Regensburg, Universitätsstrasse 31 93040 Regensburg, Germany.

[71]Department of Economics, Dartmouth College, Hanover, NH 03755-3514, USA.

[72]Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

*Correspondence to: Daniel J. Benjamin, daniel.benjamin@gmail.com; Magnus Johannesson, magnus.johannesson@hhs.se; Valen E. Johnson, vejohnson@exchange.tamu.edu.

One Sentence Summary: We propose to change the default *P*-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

Main Text:

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on "statistically significant" findings. There has been much progress toward documenting and addressing several causes of this lack of reproducibility (e.g., multiple testing, P-hacking, publication bias, and under-powered studies). However, we believe that a leading cause of non-reproducibility has not yet been adequately addressed: Statistical standards of evidence for claiming new discoveries in many fields of science are simply too low. Associating "statistically significant" findings with $P < 0.05$ results in a high rate of false positives *even in the absence of other experimental, procedural and reporting problems*.

For fields where the threshold for defining statistical significance for new discoveries is $P < 0.05$, we propose a change to $P < 0.005$. This simple step would immediately improve the reproducibility of scientific research in many fields. Results that would currently be called "significant" but do not meet the new threshold should instead be called "suggestive." While statisticians have known the relative weakness of using $P \approx 0.05$ as a threshold for discovery and the proposal to lower it to 0.005 is not new (*1, 2*), a critical mass of researchers now endorse this change.

We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (e.g., genomics and high-energy physics research; see Potential Objections below).

We also restrict our recommendation to studies that conduct null hypothesis significance tests. We have diverse views about how best to improve reproducibility, and many of us believe that other ways of summarizing the data, such as Bayes factors or other posterior summaries based on clearly articulated model assumptions, are preferable to *P*-values. However, changing the *P*-value threshold is simple, aligns with the training undertaken by many researchers, and might quickly achieve broad acceptance.

“Ronald Fisher understood that the choice of 0.05 was arbitrary when he introduced it (14). Since then, theory and empirical evidence have demonstrated that a lower threshold is needed. A much larger pool of scientists are now asking a much larger number of questions, possibly with much lower prior odds of success.

For research communities that continue to rely on null hypothesis significance testing, reducing the P-value threshold for claims of new discoveries to 0.005 is an actionable step that will immediately improve reproducibility. We emphasize that this proposal is about standards of evidence, not standards for policy action nor standards for publication. Results that do not reach the threshold for statistical significance (whatever it is) can still be important and merit publication in leading journals if they address important research questions with rigorous methods. This proposal should not be used to reject publications of novel findings with $0.005 < P < 0.05$ properly labeled as suggestive evidence. We should reward quality and transparency of research as we impose these more stringent standards, and we should monitor how researchers’ behaviors are affected by this change. Otherwise, science runs the risk that the more demanding threshold for statistical significance will be met to the detriment of quality and transparency.

Journals can help transition to the new statistical significance threshold. Authors and readers can themselves take the initiative by describing and interpreting results more appropriately in light of the new proposed definition of “statistical significance.” The new significance threshold will help researchers and readers to understand and communicate evidence more accurately.”

Much of what you will do in your reporting practice is to find ways to “look” at data or create computational models that let us see aspects of the data — What about Arbuthnot in 1710? What views of data were popular or possible three centuries ago?

Far from being tools developed thousands of years ago by some unnamed or long-forgotten inventor, it is believed that statistical graphics began with William Playfair in the late 1700s*; do we have any evidence that Arbuthnot saw more than his tables of christenings?

Let’s answer this question by first asking what you’d do... how would you look at the christening data?

New Open Save Print Import Copy Paste Format Undo Redo AutoSum

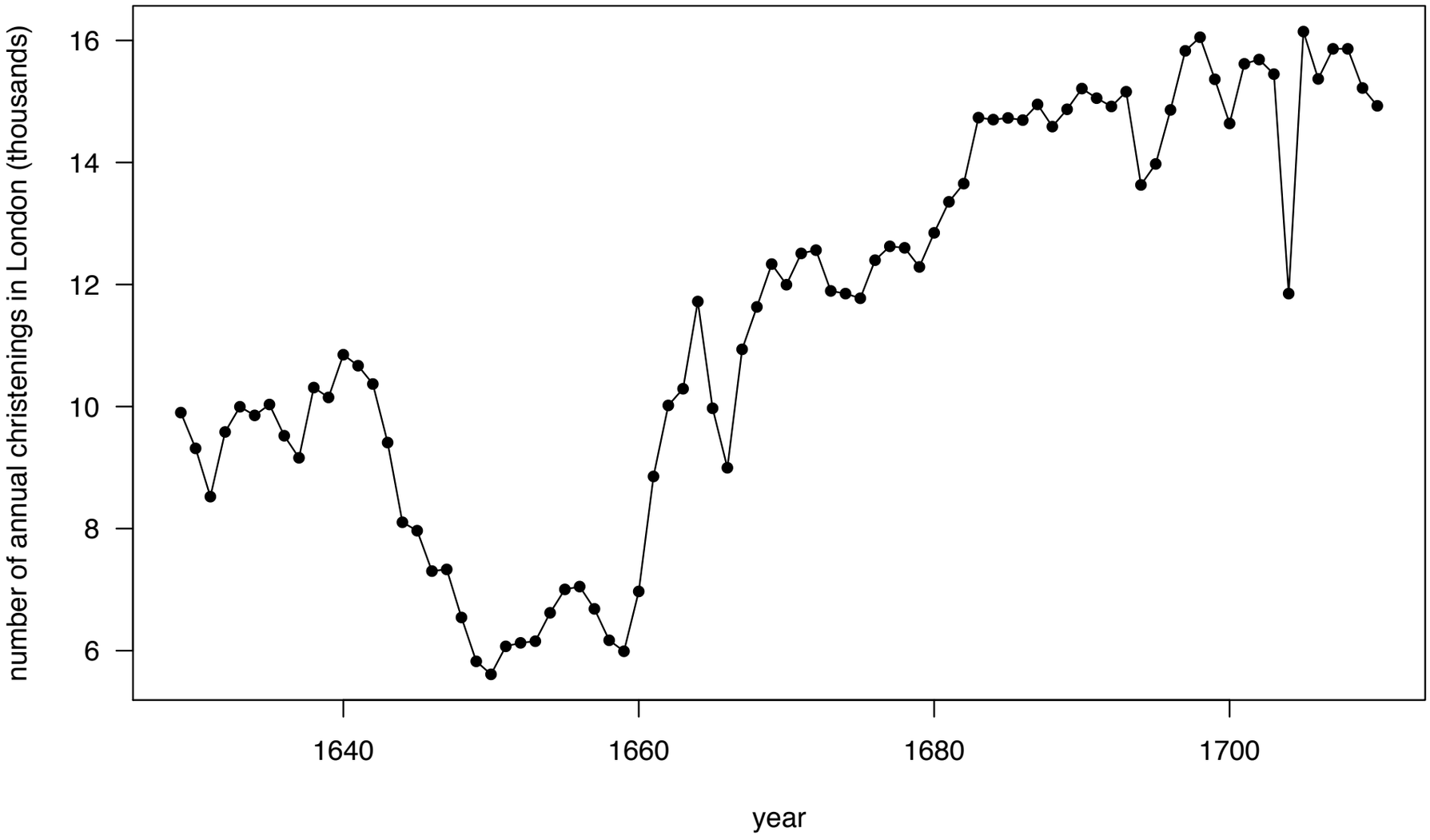
Sheets Charts SmartArt Graphics WordArt

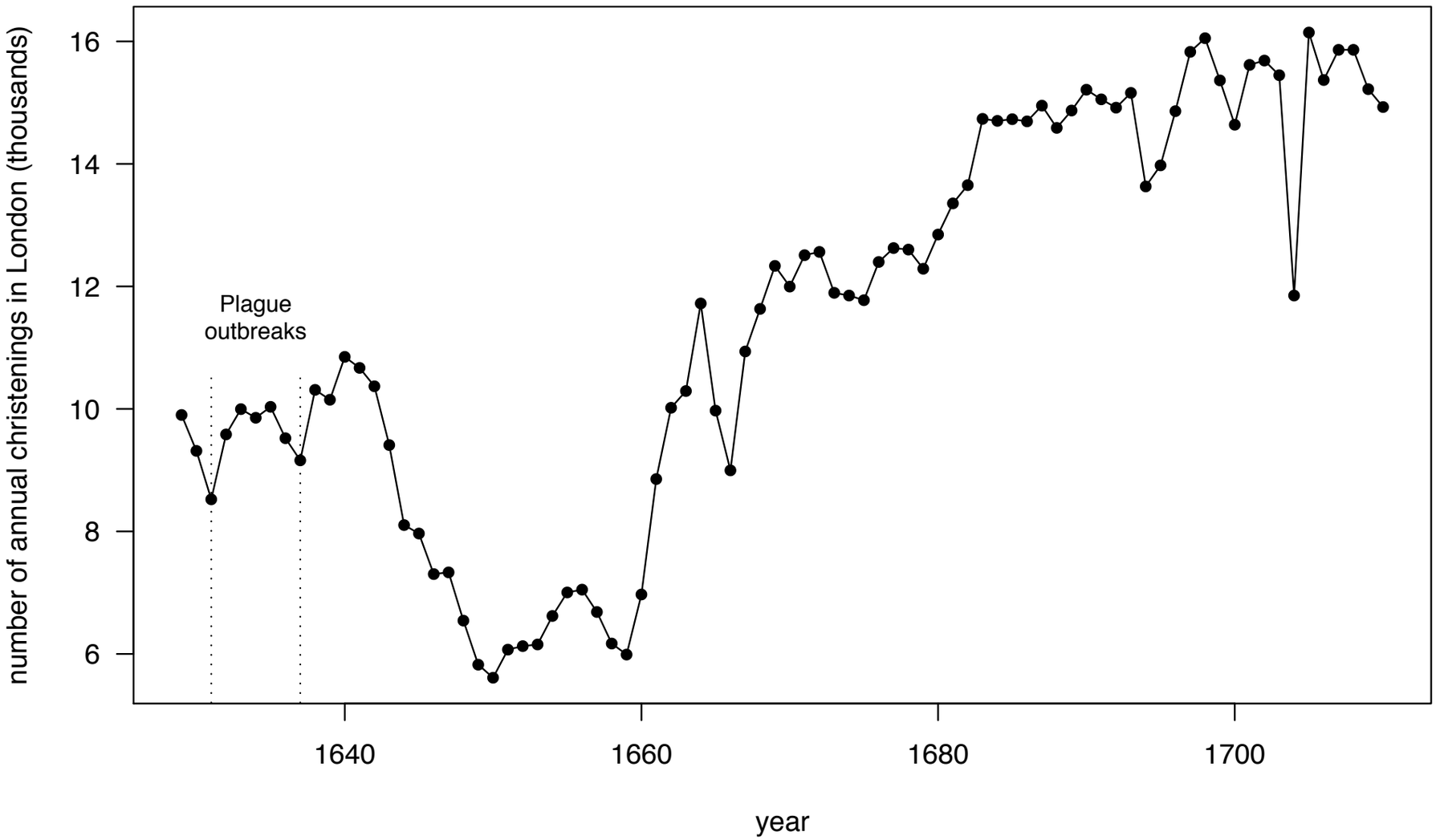
	A	B	C	D	E	F
1	year	boys	girls			
2	1629	5218	4683			
3	1630	4858	4457			
4	1631	4422	4102			
5	1632	4994	4590			
6	1633	5158	4839			
7	1634	5035	4820			
8	1635	5106	4928			
9	1636	4917	4605			
10	1637	4703	4457			
11	1638	5359	4952			
12	1639	5366	4784			
13	1640	5518	5332			
14	1641	5470	5200			
15	1642	5460	4910			
16	1643	4793	4617			
17	1644	4107	3997			
18	1645	4047	3919			
19	1646	3768	3536			
20	1647	3796	3536			
21	1648	3363	3181			
22	1649	3079	2746			
23	1650	2890	2722			
24	1651	3231	2840			
25	1652	3220	2908			
26	1653	3196	2959			
27	1654	3441	3179			
28	1655	3655	3349			
29	1656	3668	3382			
30	1657	3396	3289			
31	1658	3157	3013			
32	1659	3209	2781			
33	1660	3724	3247			
34	1661	4748	4107			
35	1662	5216	4803			
36	1663	5411	4881			
37	1664	6041	5681			
38	1665	5114	4858			
39	1666	4678	4319			
40	1667	5616	5322			
41	1668	6073	5560			
42	1669	6506	5829			
43	1670	6278	5719			
44	1671	6449	6061			

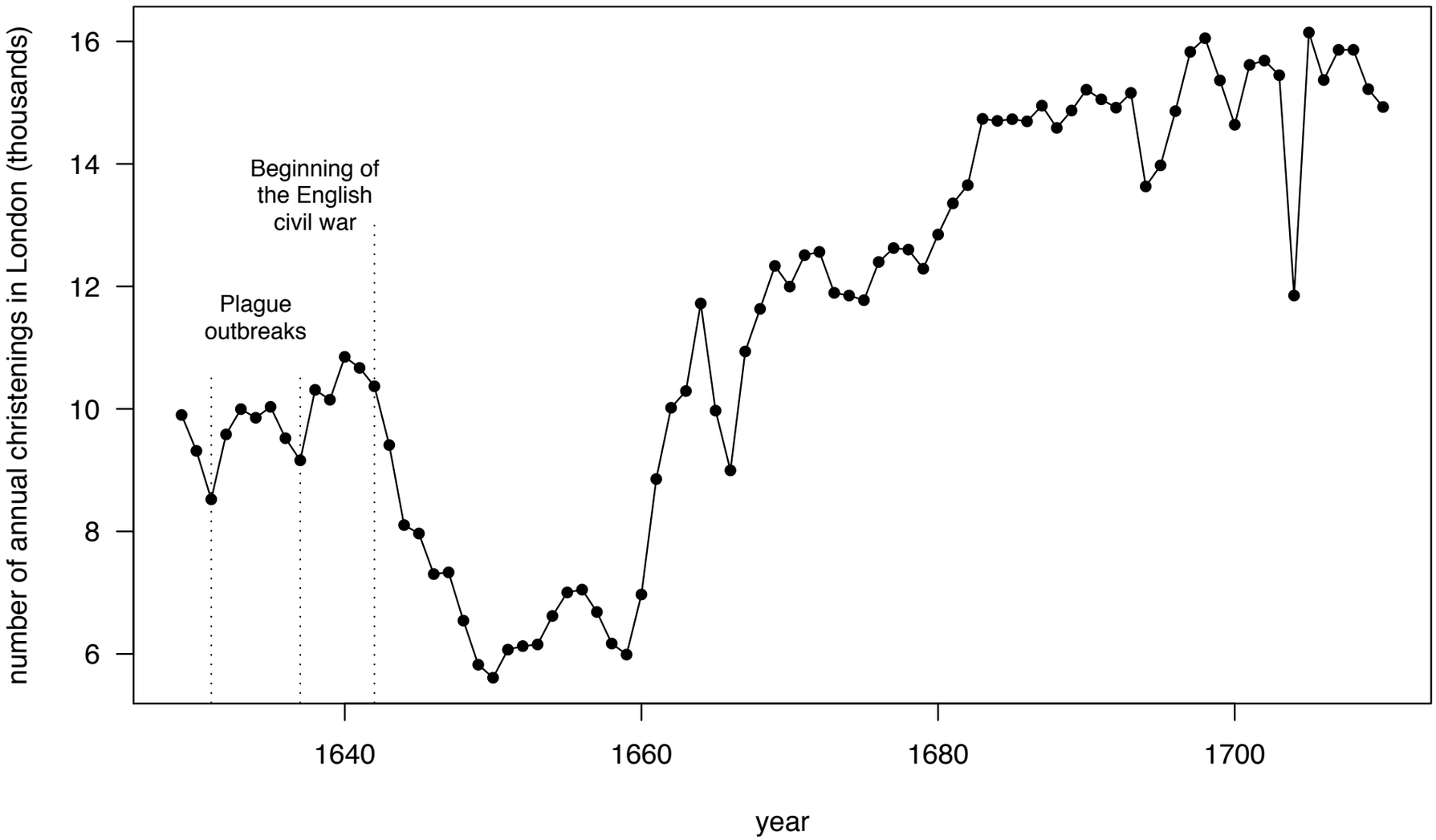

Christenings in London
(girls, solid; boys, dotted)
number of christenings (thousands)
3
4
5
6
7
8
1640
1660
1680
1700
year

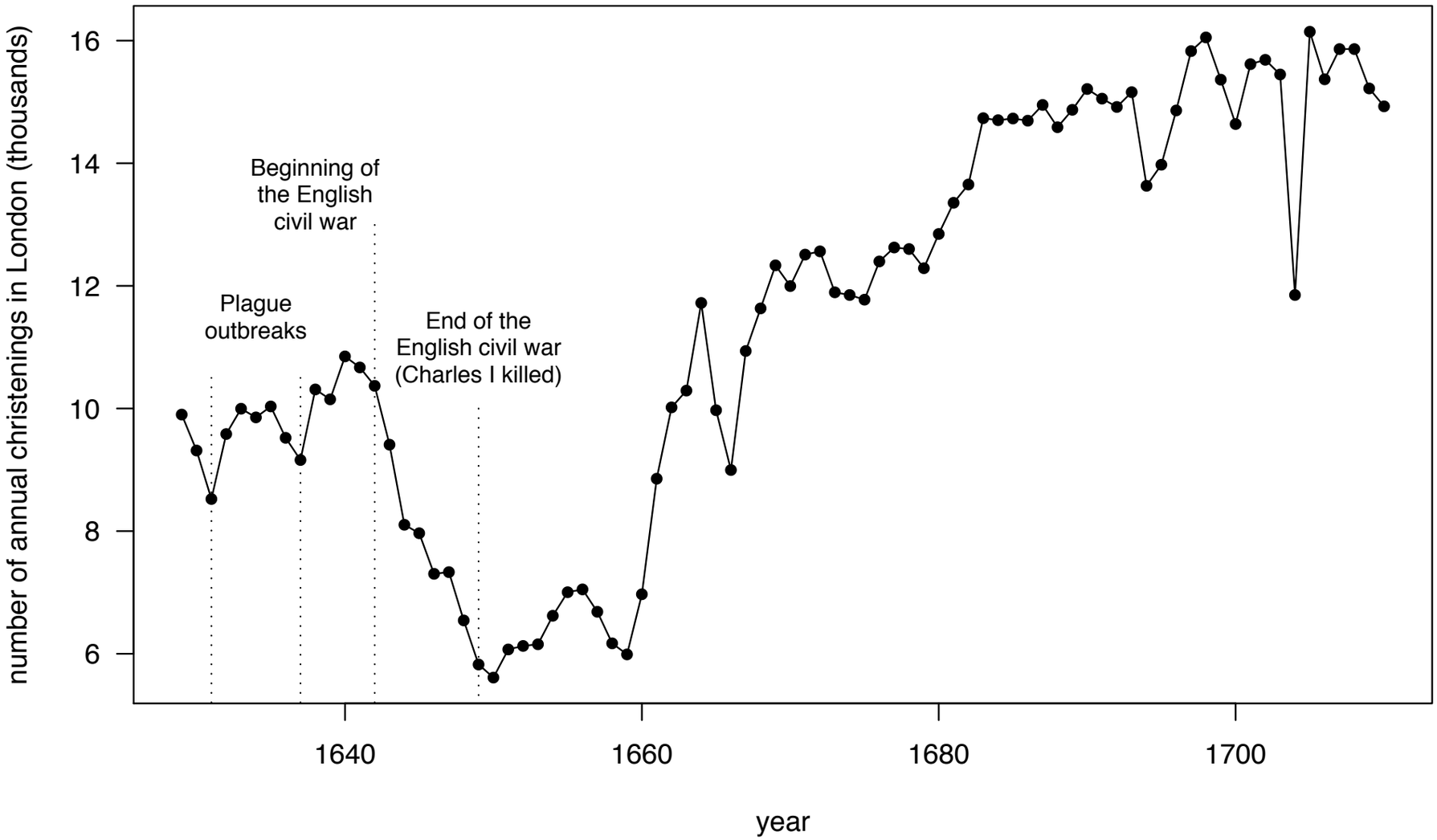
What does the “time series” plot of christenings, broken down by sex, show us? There is certainly a lot of structure to the graph, with periods of downturn in the total number of christenings, superimposed on an overall increase in the birthrate

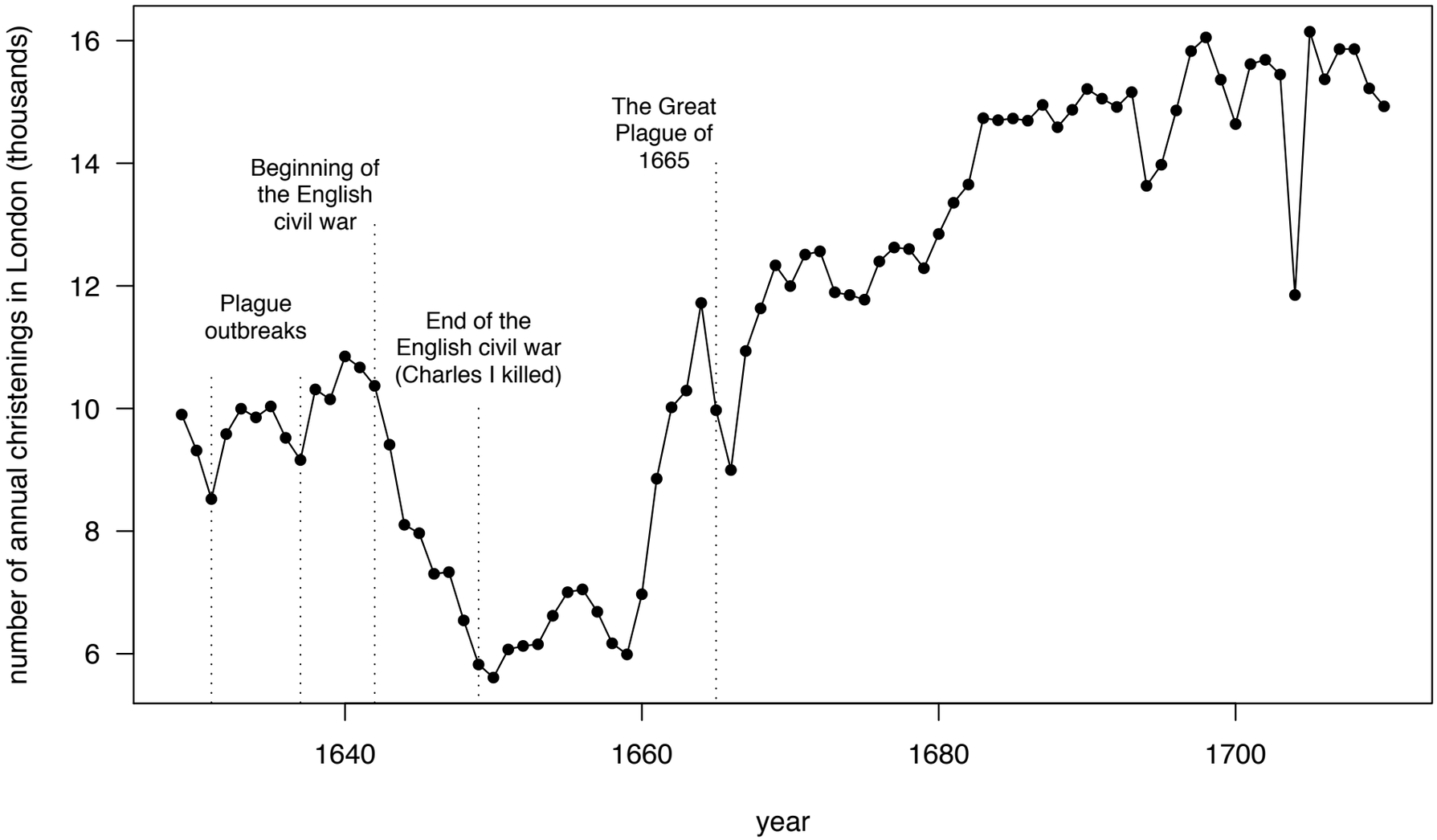
In a modern treatment of these data, Wainer (2005) started to identify peaks and valleys with specific historical events like wars and plagues; let’s see what that amounts to...

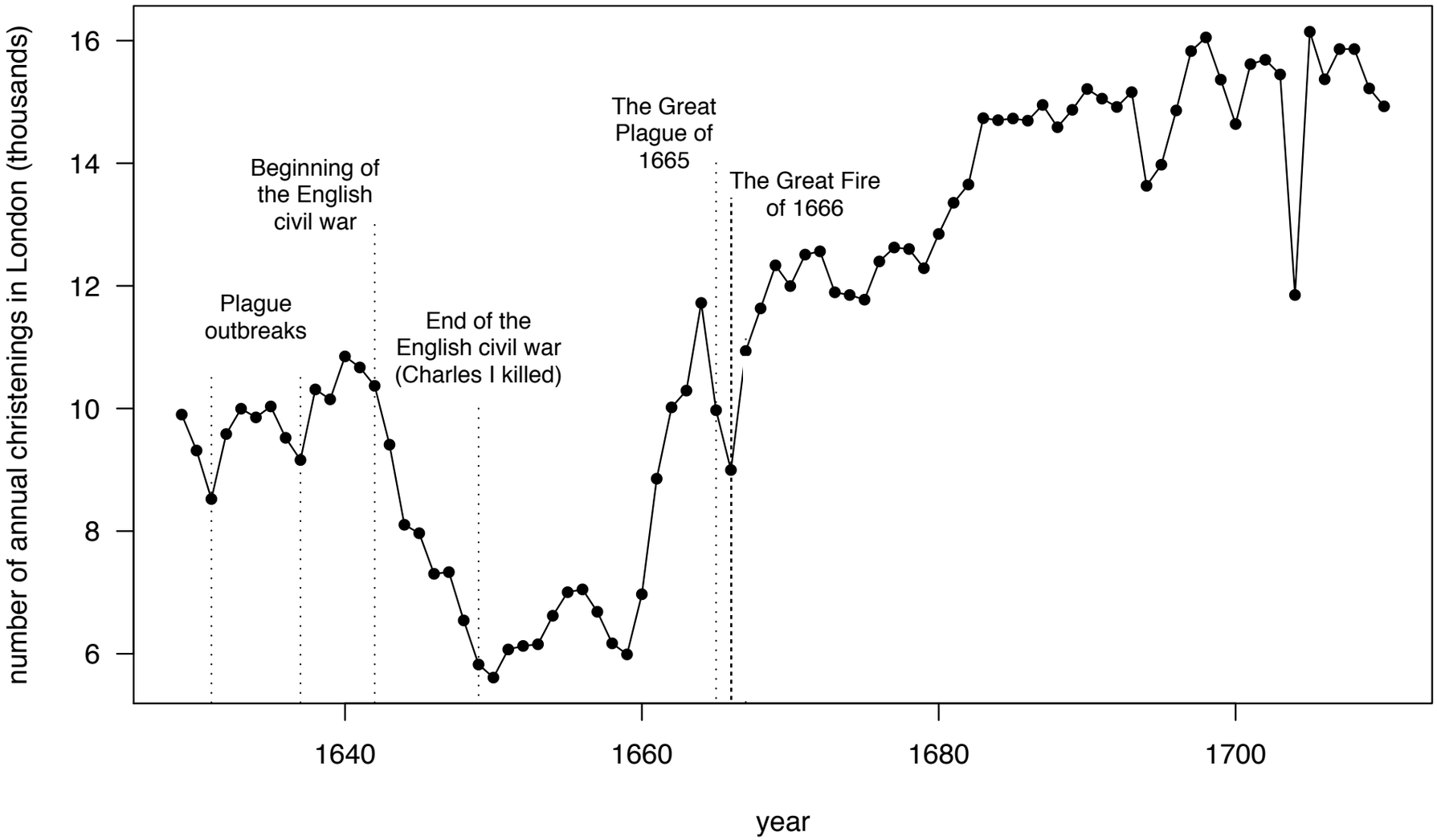

number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year

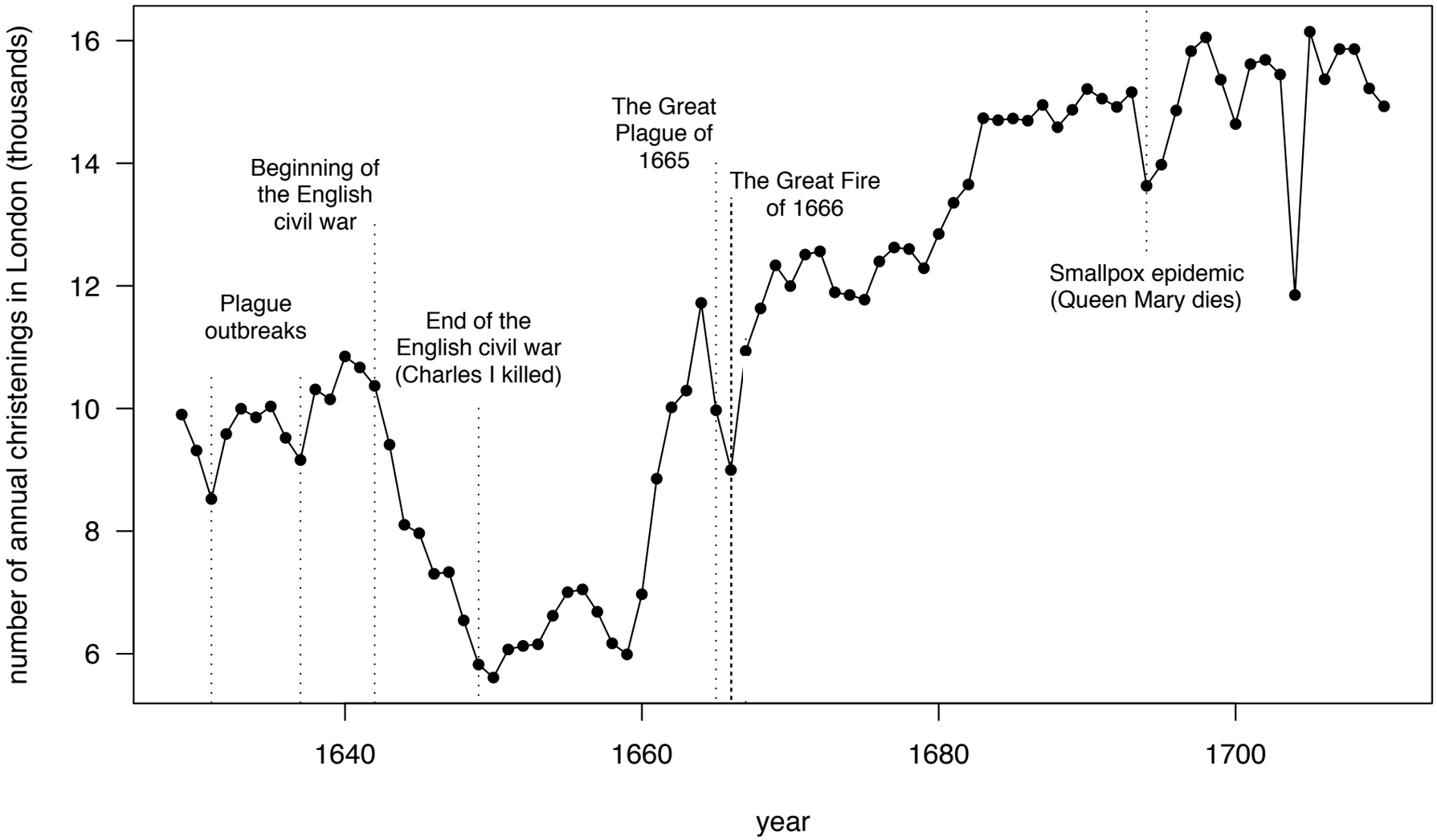

number of annual christenings in London (thousands)
16
14
12
10
8
6
Plague
outbreaks
1640
1660
1680
1700
year

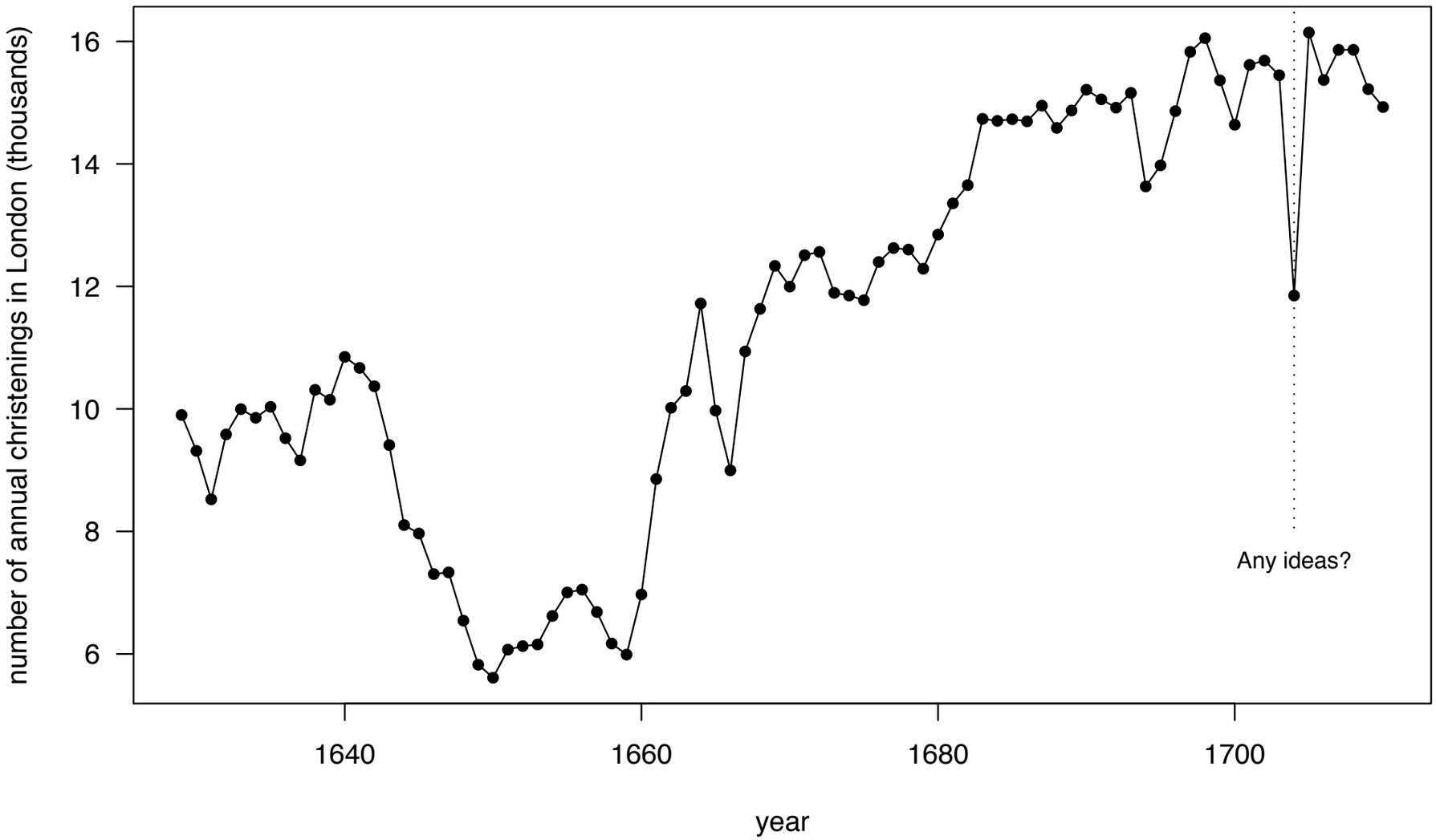

number of annual christenings in London (thousands)
16
14
12
10
8
6
Beginning of
the English
civil war
Plague
outbreaks
1640
1660
1680
1700
year


number of annual christenings in London (thousands)
16
14
12
10
8
6
Beginning of
the English
civil war
Plague
outbreaks
End of the
English civil war
(Charles I killed)
1640
1660
1680
1700
year


number of annual christenings in London (thousands)
16
14
12
10
8
6
Beginning of the English civil war
Plague outbreaks
End of the English civil war (Charles I killed)
The Great Plague of 1665
1640
1660
1680
1700
year

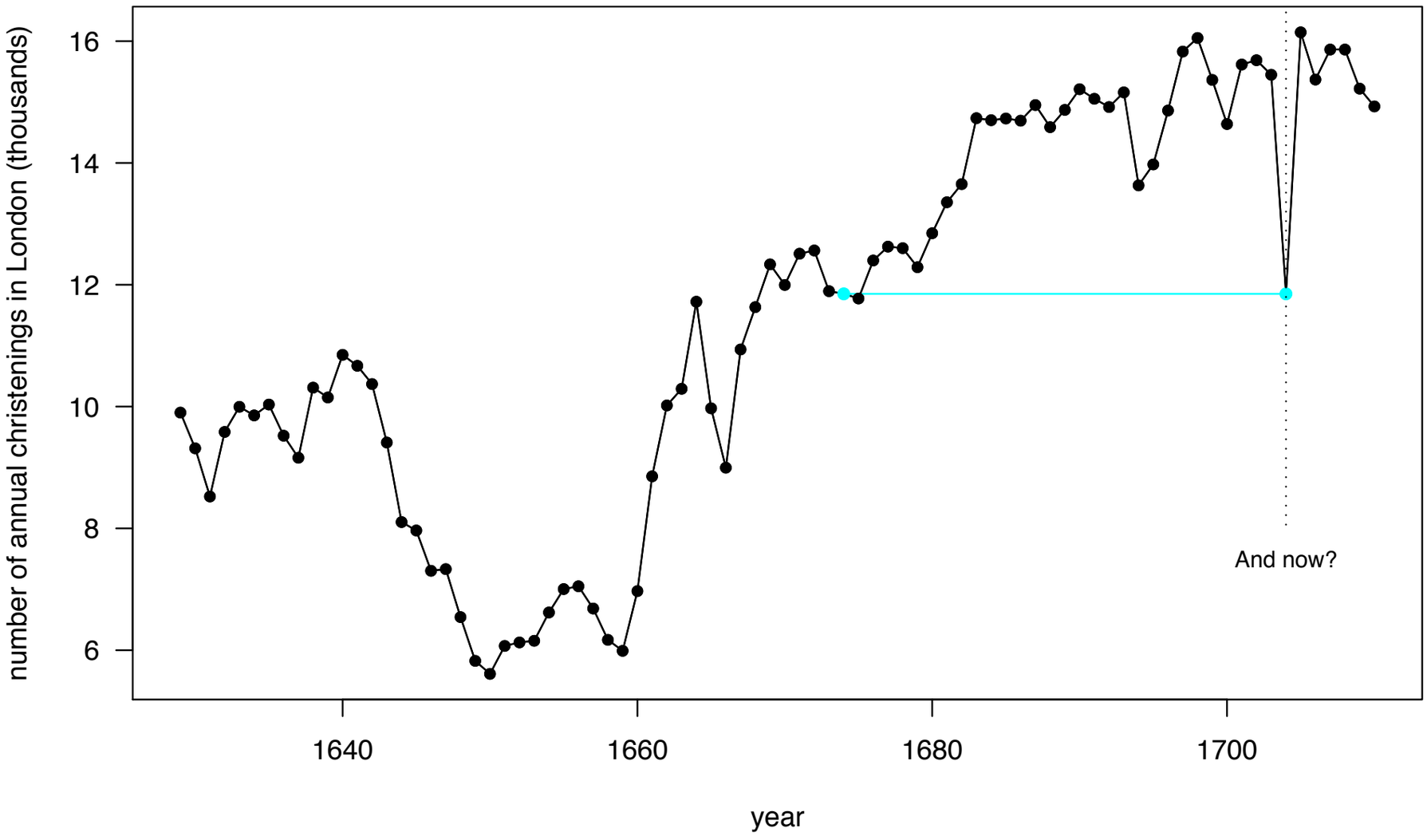

number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year
Beginning of the English civil war
Plague outbreaks
End of the English civil war (Charles I killed)
The Great Plague of 1665
The Great Fire of 1666


number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year
Beginning of the English civil war
Plague outbreaks
End of the English civil war (Charles I killed)
The Great Plague of 1665
The Great Fire of 1666
Smallpox epidemic (Queen Mary dies)


number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year
Any ideas?

The drop at 1704 is a little harder to explain; it is the single largest one-year drop in the entire data set (a decrease of about 30% from 1703), and certainly the event that prompted it would have been significant (“The great disaster x”)

Looking at the drop a little closer, we get a clue as to its origin, and, in so doing, see why some believe that Arbuthnot could not have looked at a plot of his data, even a simple line plot of the sort we’ve made here


number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year
And now?

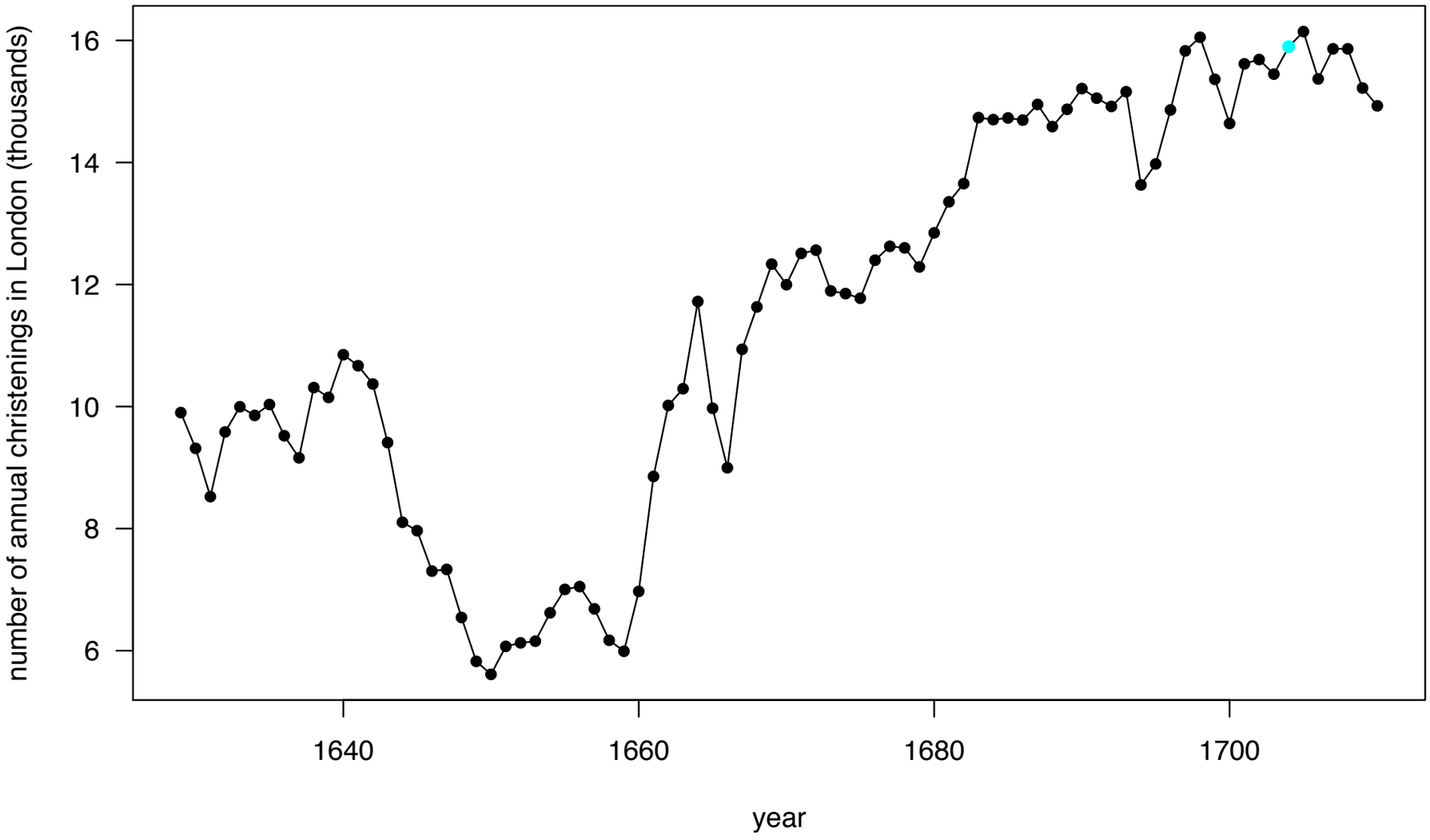
Anno.	Christened. Males.	Females.
1629	5218	4683
30	4858	4457
31	4422	4102
32	4994	4590
33	5158	4839
34	5035	4820
35	5106	4928
36	4917	4605
37	4703	4457
38	5359	4952
39	5366	4784
40	5518	5332
41	5470	5200
42	5460	4910
43	4793	4617
44	4107	3997
45	4047	3919
46	3768	3395
47	3796	3536

Anno.	Christened. Males.	Females.
1648	3363	3181
49	3079	2746
50	2890	2722
51	3231	2840
52	3220	2908
53	3196	2959
54	3441	3179
55	3655	3349
56	3668	3382
57	3396	3289
58	3157	3013
59	3209	2781
60	3724	3247
61	4748	4107
62	5216	4803
63	5411	4881
64	6041	5681
65	5114	4858
66	4678	4319

Anno.	Christened. Males.	Females.
1667	5616	5322
68	6073	5560
69	6506	5829
70	6278	5719
71	6449	6061
72	6443	6120
73	6073	5822
74	6113	5738
75	6058	5717
76	6552	5847
77	6423	6203
78	6568	6033
79	6247	6041
80	6548	6299
81	6822	6533
82	6909	6744
83	7577	7158
84	7575	7127
85	7484	7246
86	7575	7119
87	7737	7214
88	7487	7101

Anno.	Christened. Males.	Females.
1689	7604	7167
90	7909	7302
91	7662	7392
92	7602	7316
93	7676	7483
94	6985	6647
95	7263	6713
96	7632	7229
97	8062	7767
98	8426	7626
99	7911	7452
1700	7578	7061
1701	8102	7514
1702	8031	7656
1703	7765	7683
1704	6113	5738
1705	8366	7779
1706	7952	7417
1707	8379	7687
1708	8239	7623
1709	7840	7380
1710	7640	7288

error replaced with correct value
(things look better now)



Arbuthnot's example is the first known test of significance; interestingly, these data are the site of a number of important "firsts"

60 years earlier, in 1662, John Graunt, a successful London shopkeeper (who also had a taste for scholarship), published *Observation upon the Bills of Mortality*, in which he reported systematic observations about the population in and around London

His observations were, in effect, generalizations about the patterns and regularities in births, deaths and migration in England

The Diseases, and Casualties this year being 1632.

Abortive, and Stilborn ..	445	Grief	11
Affrighted	1	Jaundies	43
Aged	628	Jawfaln	8
Ague	43	Impostume	74
Apoplex, and Meagrom	17	Kil'd by several accidents..	46
Bit with a mad dog.....	1	King's Evil.....	38
Bleeding	3	Lethargie	2
Bloody flux, scowring, and flux	348	Livergrown	87
Brused, Issues, sores, and ulcers, ...	28	Lunatique	5
Burnt, and Scalded.....	5	Made away themselves.....	15
Burst, and Rupture.....	9	Measles	80
Cancer, and Wolf.....	10	Murthered	7
Canker	1	Over-laid, and starved at nurse	7
Childbed	171	Palsie	25
Chrisomes, and Infants.....	2268	Piles.....	1
Cold, and Cough.....	55	Plague.....	8
Colick, Stone, and Strangury	56	Planet	13
Consumption	1797	Pleurisie, and Spleen.....	36
Convulsion	241	Purples, and spotted Feaver	38
Cut of the Stone.....	5	Quinsie	7
Dead in the street, and starved	6	Rising of the Lights.....	98
Dropsie, and Swelling.....	267	Sciatica	1
Drowned	34	Scurvey, and Itch.....	9
Executed, and prest to death	18	Suddenly	62
Falling Sickness.....	7	Surfet	86
Fever	1108	Swine Pox	6
Fistula	13	Teeth	470
Flocks, and small Pox.....	531	Thrush, and Sore mouth...	40
French Pox.....	12	Tympany	13
Gangrene	5	Tissick	34
Gout	4	Vomiting	1
		Worms	27

Christened	Males....4994 Females..4590 In all....9584	Buried	Males....4932 Females..4603 In all....9535	Whereof, of the Plague.8

Increased in the Burials in the 122 Parishes, and at the Pest-house this year..... 993

Decreased of the Plague in the 122 Parishes, and at the Pest-house this year..... 266

His innovation was to apply **the scientific method to the study of populations;** in Graunt's time science was largely limited to observations and descriptions of "naturally" occurring events

It is an early example of what has been termed **"political arithmetic,"** a practice that hoped to ground "official policy... in an understanding of the land and its inhabitants"

"Implicit in the use by political arithmeticians of social numbers was the belief that the wealth and strength of the state depended strongly on the number and character of its subjects"

Using these data, Graunt also constructed the first known "life table," a numerical device summarizing mortality in terms of the number, percent and probability of living or dying throughout a lifetime

Cut of the Stone	47	40	30	308	444	3
Dropſy, and Tympany	8	17	2	27	49	
Drowned			29			
Exceſſive drinking	3	2		43	24	
Executed	139	400	2			
Fainted in a Bath	6	6	1190	3		
Falling-Sickneſs	18	29	9	184	8	
Flox, and ſmall pox	4	4	15		18	
Found dead in the Streets	9	5	1			
French-Pox	12	13	12			
Frighted	11	10	16			
Gout		1	13			
Grief	57	35	11			
Hanged, and made-away themſelves	1	1	39			
Head-Ach	75	61				
Jaundice		1				
Jaw-faln	27	57				
Impoſtume	27	26				
Itch	3	4				
Killed by ſeveral Accidents						
King's Evil						
Lethargy	53		46			
Leproſy	12		11			
Livergrown, Spleen, and Rickets						
Lunatique						

The Conclusion.

IT may be now asked, to what purpose tends all this laborious buzzling, and groping? To know,

1. The number of the People?
2. How many Males, and Females?
3. How many Married, and single?
4. How many Teeming Women?
5. How Many of every Septenary, orDecad of years in age?
6. How many Fighting Men?
7. How much London is, and by what steps it hath increased?
8. In what time the housing is replenished after a Plague?
9. What proportion die of each general and perticular Casualties?
10. What years are Fruitfull, and Mortal, and in what Space, and Intervals, they follow each other?
11. In what proportion Men neglect the Orders of the Church, and Sects have increased?
12. The disproportion of Parishes?
13. Why the Burials in London exceed the Christnings, when the contrary is visible in the Country?

To this I might answer in general by saying, that those, who cannot apprehend the reason of these Enquiries, are unfit to trouble themselves to ask them.

Graunt, John, **Natural and Political Observations mentioned in a following Index, and made upon the Bills of Mortality**, *1665*

Types of studies

In the health and life sciences, we are faced with two kinds of studies that differ in terms the conditions under which data are collected

- In an **experimental study,** we impose some **change or treatment** and measure the result or response

- In an **observational study,** we simply **observe and measure something that has taken place or is taking place** (while trying not to cause any changes by our presence)

The kinds of inference you can make will depend on the type of study you conduct, **as well as its overall design or program for how data are to be collected** -- These kinds of considerations will lead to a range of (admittedly more technical) questions you should ask of a data set

In the last two lectures, we've talked about two data sets, the CDC Behavioral Risk Surveillance System and a list of courses from the registrar -- What kinds of "studies" do these represent

Optimizely

The BIG BOOK of Experimentation

35+ Customer Stories

Truila
Hotwire
Secret Escapes
Teletext Holidays
Smarter Travel
FSA Store
Auto Scout 24
Spread Shirt
Sony
Home 24
Marc O'Polo
Evans Cycles
Jawbone
MVMT
Verivox
Brooks
Bigstock
Vergelijk.nl
Lostmy.name
Liftopia
BluDot
Revolve
EA
WWF
Obama 2012
Movember
charity: water
Code
The Guardian
Upworthy
Trinity Mirror
Ringier
eHow
Discovery
Asana
Citrix
Iron Mountain

ORIGINAL:

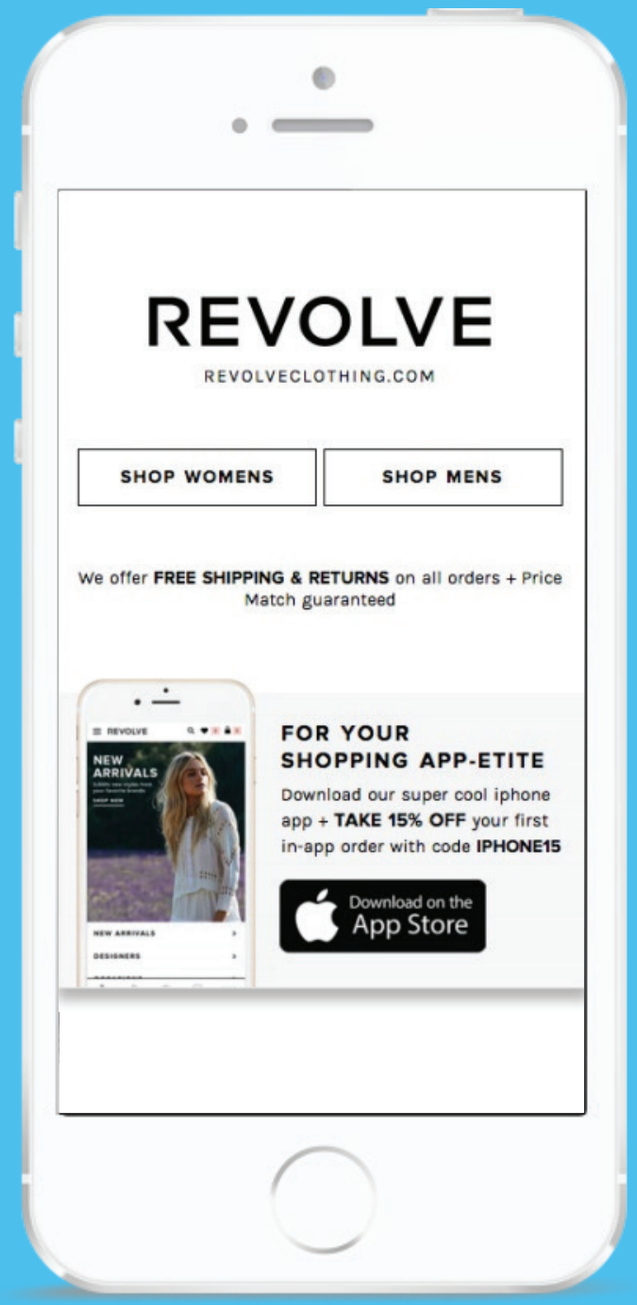
The original page prominently featured the pre-order offer at the top of the page.

VARIATION B:

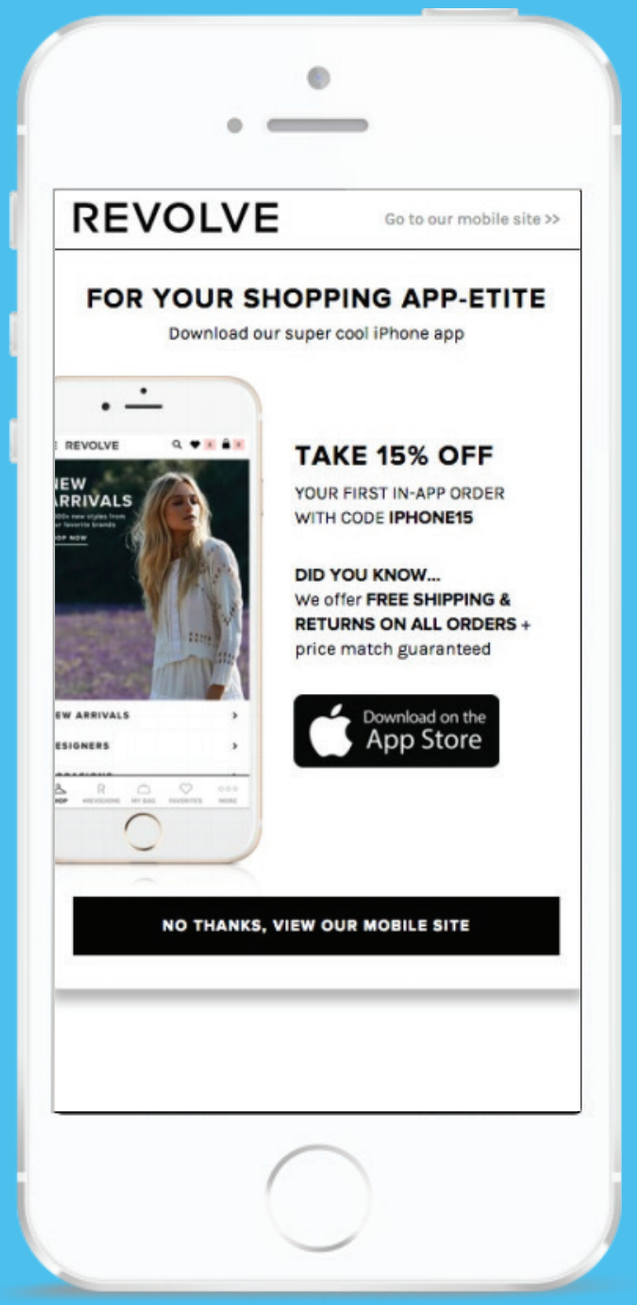
The winning variation completely removed the pre-order offer and drove a 43.4% lift in revenue.

ORIGINAL VS. VARIATION:

The winning variation featured a bold, aggressive style and message.

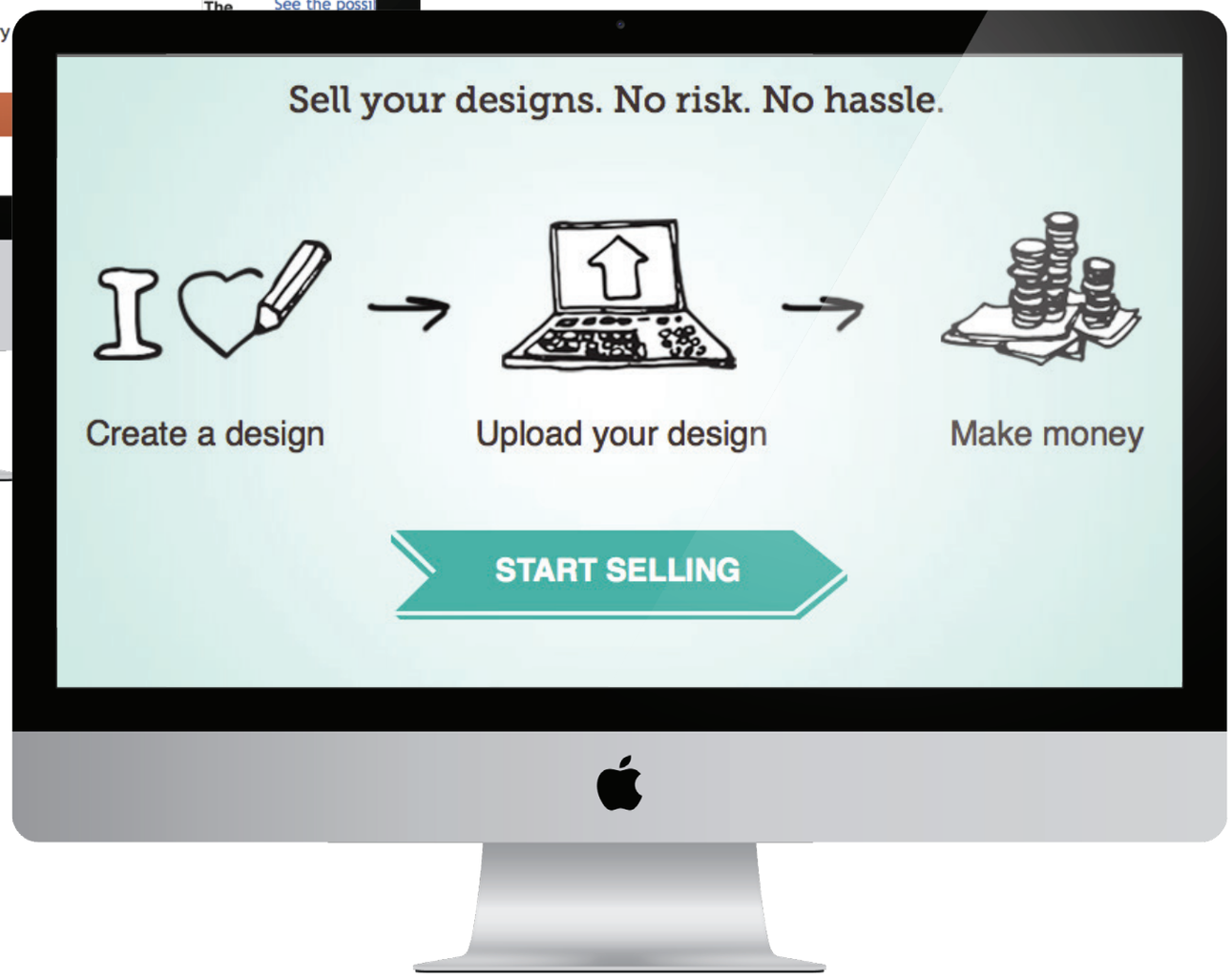


ORIGINAL:



VARIATION:

Text and graphic heavy original with multiple CTAs.



Streamlined winning variation with focused CTA.

IRON MOUNTAIN
INCORPORATED

INDUSTRY:
B2B

EXPERIMENT:
Improve Quality of Website Leads

140%

Improvement in Lead Quality

Iron Mountain Improves the Quality of Inbound Leads by Optimizing Sales Contact Form

As a large B2B company with over $3 billion in annual revenue, Iron Mountain is constantly striving to drive sales and improve lead conversion on it's website.

With the help of their digital performance agency iProspect, Iron Mountain identified multiple challenges with their Sales lead form, which was contributing to poor lead quality, including incomplete or irrelevant inquiries.

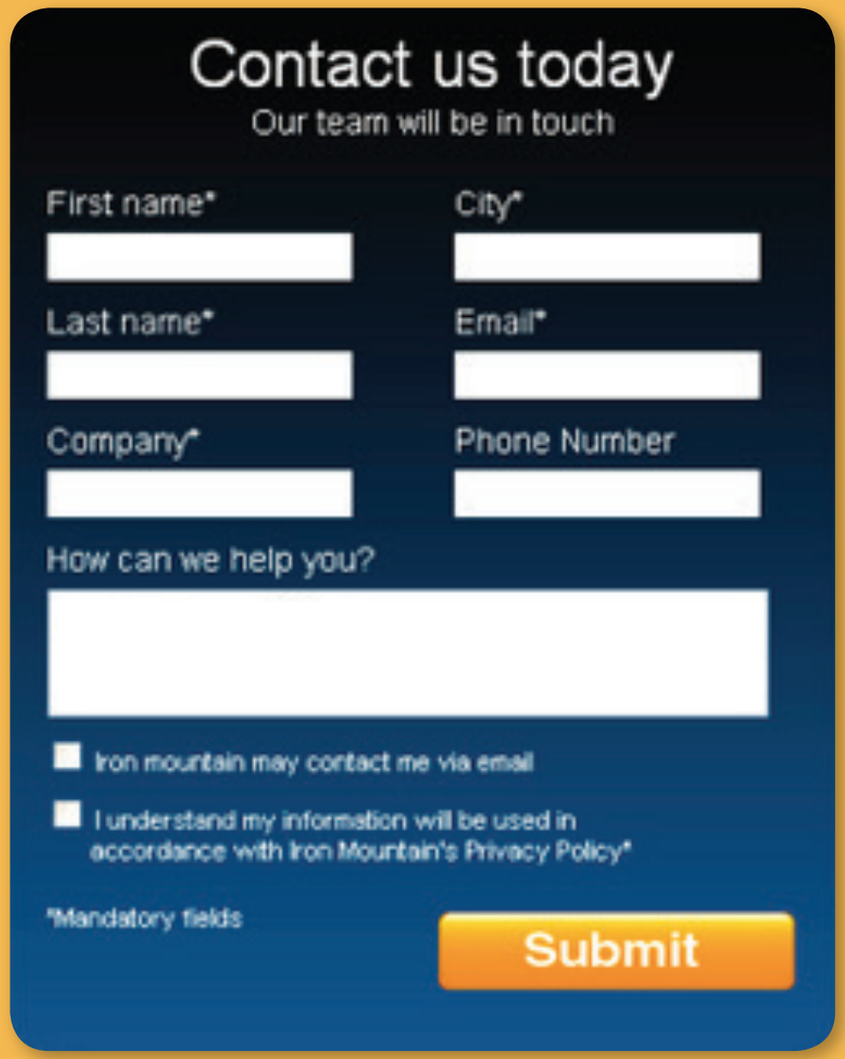
Together they designed an alternative variation of the form to address each of the key challenges they had identified, and tested the changes with Optimizely. Their new variation clarified the form's headline, made the phone number field compatible with their database, and added clearer validation of correct inputs.

The variation form resulted in a 140% improvement in lead quality, representing the highest lead quality Iron Mountain's sales team had ever seen.

"Ensuring that sales teams have solid leads to follow is fundamental in the B2B marketplace. We wanted a way to deliver better leads without burdening our busy internal IT team." —*Nimesh Parmar, Online Manager, Iron Mountain*

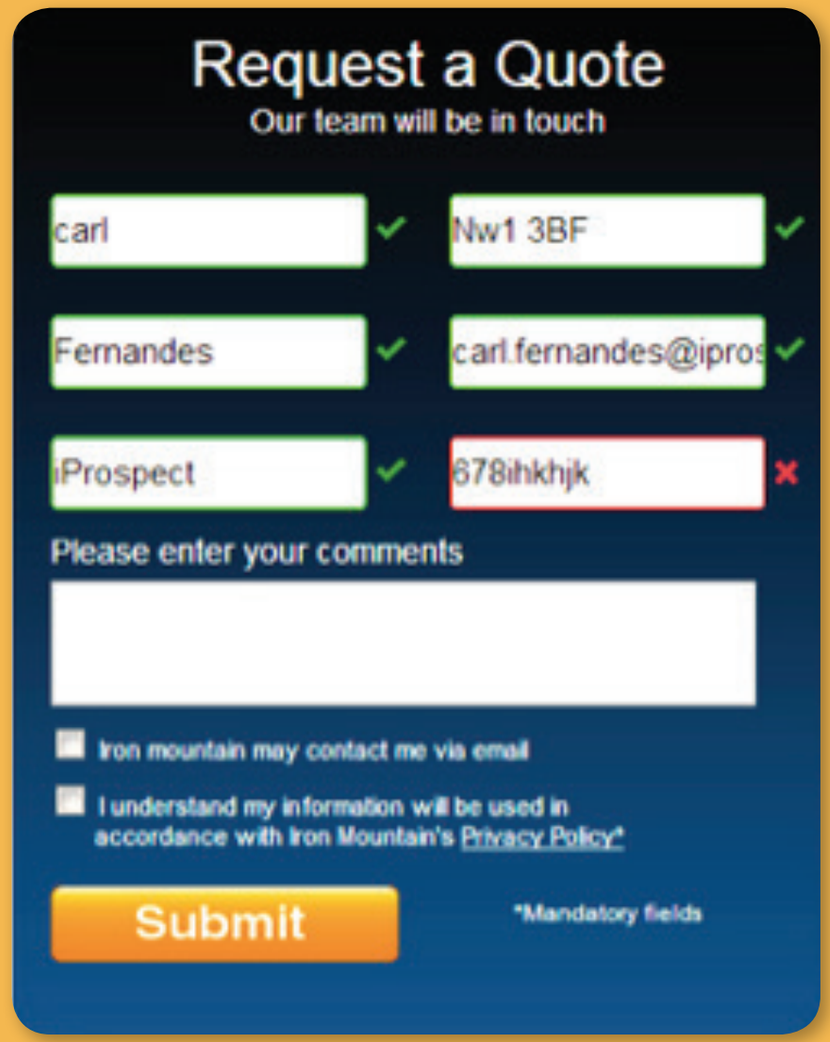
VARIATION A:

Default lead form often resulting in poor lead quality

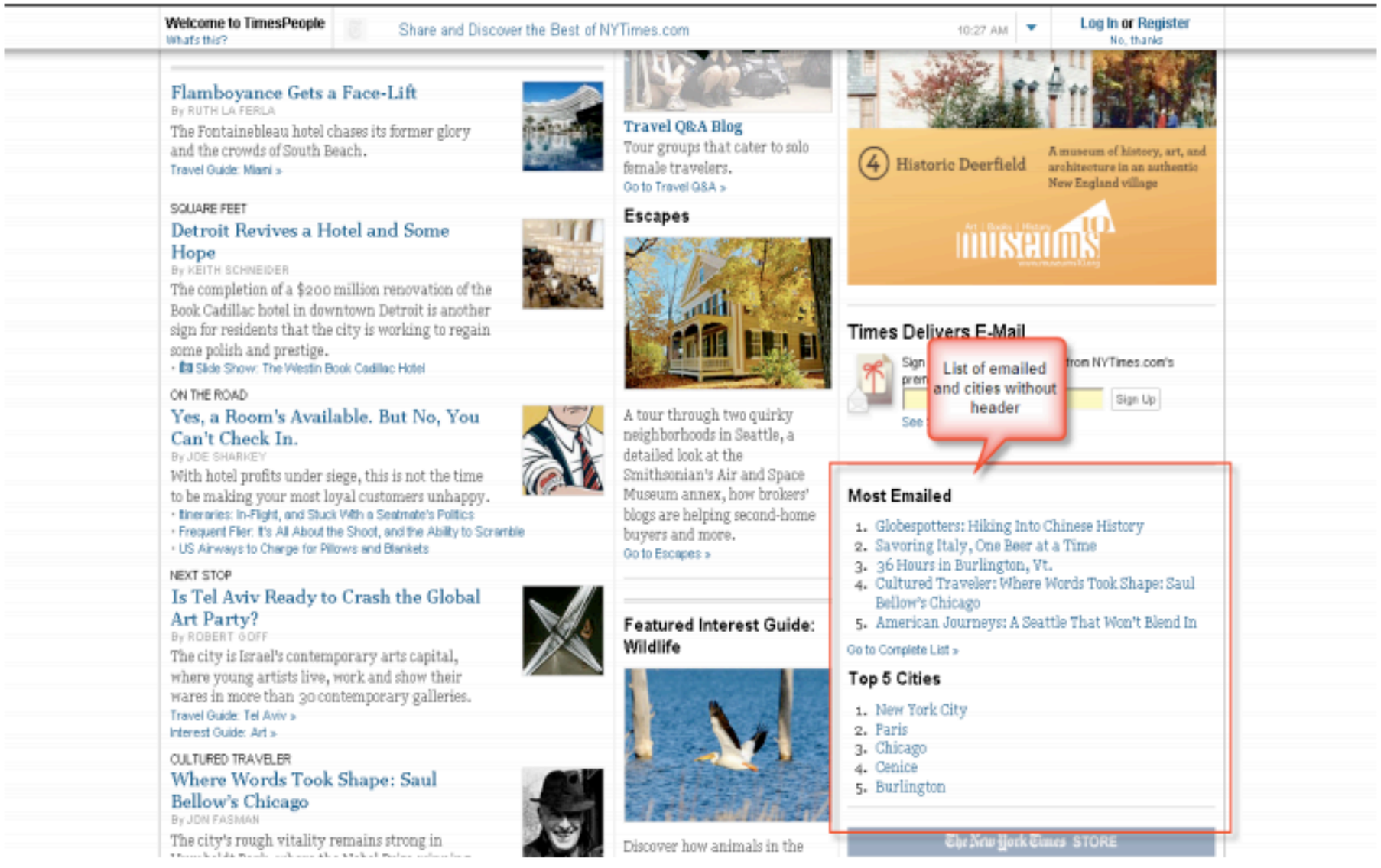


VARIATION B:

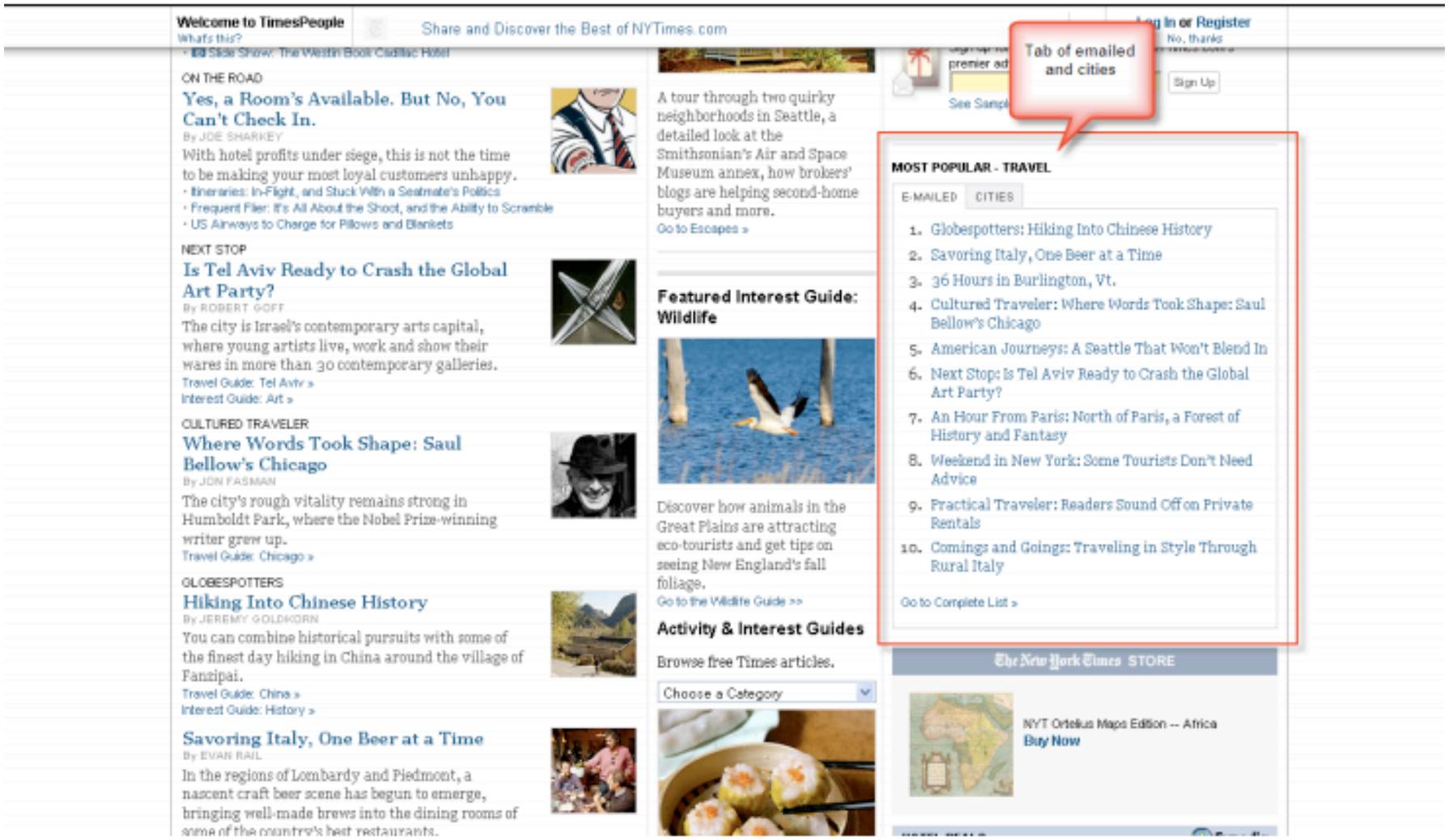
Improved lead form with clarified call-to-action, field validation, and backend compatibility.



List: Variation 10858



Tabs: Variation 10859



Soul-Searching in Baltimore, a Year After Freddie Gray’s Death

Baltimore After Freddie Gray: The 'Mind-Set Has Changed' (+1,677%)

Nuclear Plant’s Neighbors Confront Regulators on Safety

Staff ‘Overwhelmed’ at Nuclear Plant, but U.S. Won’t Shut It (+407%)

Tax Bill Clears Senate Panel as Support Widens Among G.O.P.

Republicans Who Wavered on Tax Bill Now Signal Support (+410%)

From Ali Tehrani’s 2018 MS J-School Thesis

Clinical trials

A clinical trial is simply an experimental study in which two or more treatments are assigned to human subjects -- Experimental studies in all areas of biology have been greatly informed by procedures used in clinical trials

The clinical trial, however, has evolved considerably -- It was not always the "gold standard" of experimental designs -- Richard Doll (a well known epidemiologist who studied lung cancer) noted that before 1946

... new treatments were almost always introduced on the grounds that in the hands of professor A or in the hands of a consultant at one of the leading teaching hospitals, the results in a small series of patients (seldom more than 50) had been superior to those recorded by professor B (or some other consultant) or by the same investigator previously. ***Under these conditions variability of outcome, chance, and the unconscious (leave alone the conscious) in the selection of patients brought about apparently important differences in the results obtained****; consequently, there were many competing new treatments*

Clinical trials

In an attempt to improve the evaluation of different treatments, **Austin Bradford Hill began advocating a more systematic approach to designing clinical trials**; like Doll, he was frustrated with the quality of research at the time, going so far as to question the ethics of the existing system

Hill was the son of a distinguished physiologist; his hope of a medical career was thwarted by the onset of tuberculosis in 1917, and instead, while an invalid, he completed a degree in economics by correspondence

In 1927 Hill moved to the London School of Hygiene and Tropical Medicine and during the 1930s he researched mainly in occupational epidemiology; his renown in medical statistics started in 1937 with the publication of his textbook, *Principles of Medical Statistics*, based on a series of articles in the Lancet

Clinical Trials

Hill's work emphasizes the **practical snags and difficulties of applying statistics** in a clinical setting rather than theoretical minutiae -- It seems that his advice, while often statistically sound, was motivated by practical concerns

In terms of clinical trials, Hill argued for **well-specified study aims or outcomes**, and the consistent use of controls -- Patients were to be divided into two groups: **the "treatment" group would receive a new drug or procedure**, while **the "control" group would be prescribed the standard therapy**

Upon completion of the trial, researchers would examine the differences between the two groups, measuring outcomes, and determine if the proposed treatment is superior to the existing therapy

With his very practical approach to clinical work, Hill took a special interest in how patients were divided into the treatment and control groups -- **Left solely to physicians, he felt there could be a problem**

What was he worried about?

Clinical Trials

To remove the subjective bias of physicians in making assignments, some clinicians (including HIll, initially) had recommended the so-called **alternation method --** That is, as patients appear at a clinic or study center, researchers alternately assign them to treatment or control

Other similar schemes include the assignment of a patient based on his or her initials or even their birthdate -- Taking Hill's very practical stance, do these methods completely remove potential bias?

Clinical trials

In 1948, Hill published a groundbreaking study on the effectiveness of streptomycin (an antibiotic) in treating pulmonary tuberculosis; here is how he assigned patients to the treatment and control groups

Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on ***random sampling numbers drawn up for each sex*** *at each centre by Professor Bradford Hill; the details of the series were unknown to any of the investigators or to the co-ordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and number. After acceptance of a patient by the panel, and before admission to the streptomycin centre,* ***the appropriate numbered envelope was opened at the central office: the card inside told if the patient was to be an S or C case****, and this information was then given to the medical officer of the centre. Patients were not told before admission that they were to get special treatment; C patients did not know throughout their stay in hospital that they were control patients in a special study; they were in fact treated as they would have been in the past, the sole difference being that they had been admitted to the centre more rapidly than was normal. Usually they were not in the same wards as S patients, but the same regimen was maintained."*

Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. BMJ 1948; 2: 769-782.

An aside: Some history

Following the **immense success of penicillin**, there was a great deal of research activity around detecting other potential antibiotics

Also, tuberculosis was the **"most important cause of death"** of young adults in Europe and North America at the time

Considerable laboratory work and some early experiments on patients suggested that Streptomycin would be an effective treatment for pulmonary tuberculosis

The MRC randomized trial of streptomycin and its legacy: a view from the clinical front line, J. Crofton
http://jrsm.rsmjournals.com/cgi/reprint/99/10/531

Clinical Trials

The tuberculosis study was the first time **randomization of treatments** was used in a clinical trial; after its publication, Hill wrote a series of articles describing its use

> *In these articles, I had set out the need for controlled experiments in clinical medicine with groups chosen at random. At the outset, I think I pleaded that trials should be made using alternate cases. I suspect if (and its a very large IF) if that, in fact, were done strictly they would be random.* ***I deliberately left out the words "randomization" and "random sampling numbers" at that time, because I was trying to persuade the doctors to come into controlled trials in the very simplest form and I might have scared them off.*** *I think the concepts of "randomization" and "random sampling numbers" are slightly odd to the layman, or, for that matter, to the lay doctor, when it comes to statistics. I thought it would be better to get doctors to walk first, before I tried to get them to run.*
>
> *Memories of the British streptomycin trial in tuberculosis: The first randomized clinical trial*, Sir Austin Bradford Hill

Clinical Trials

Through randomization (and the blinding of the physicians), Hill achieved his goal of reducing bias by allocating "the patients to the 'treatment' and 'control' groups in such a way that the two groups are initially equivalent in all respects relevant to the inquiry" -- He writes

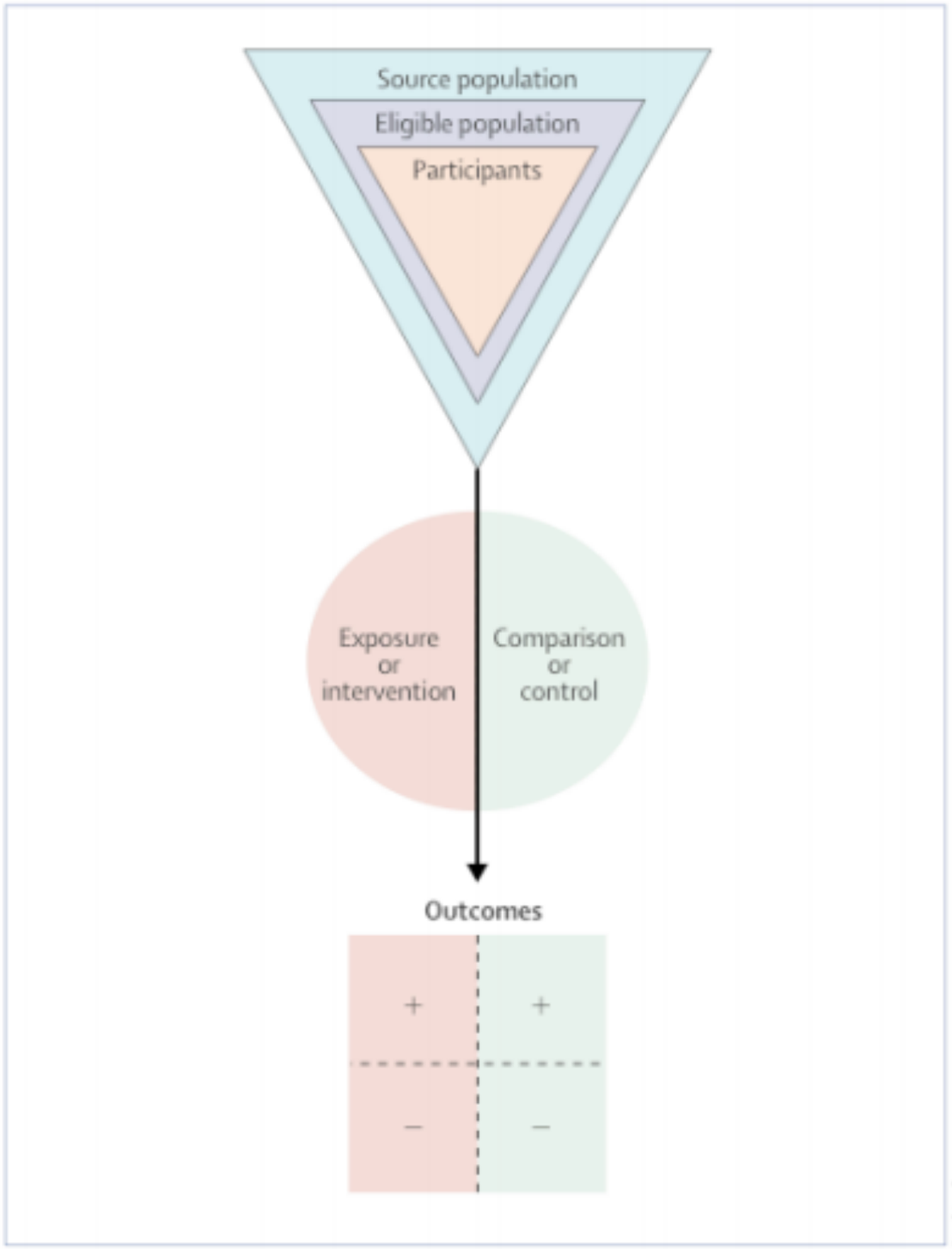
> It ensures that neither our personal idiosyncrasies (our likes or dislikes consciously or unwittingly applied) nor our lack of balanced judgement has entered into the construction of the different treatment groups—the allocation has been outside our control and the groups are therefore unbiased;
>
> ... it removes the danger, inherent in an allocation based on personal judgement, that believing we may be biased in our judgements we endeavour to allow for that bias, to exclude it, and that in doing so we may overcompensate and by thus 'leaning over backward' introduce a lack of balance from the other direction;
>
> ... and, having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity.

Memories of the British streptomycin trial in tuberculosis: the first randomized clinical trial. *Control Clin Trials* 1990;11:77–7

Randomized controlled trials

As an experiment then, the design is straightforward: participants are assigned randomly to either receive a treatment under study or a “control,” perhaps a placebo or a standard therapy

At the end of the study, an outcome is recorded for each participant; in some cases, scientists are evaluating whether a drug helps with a particular condition, say



Fisher and randomization

It would be incorrect to suggest that the idea of randomization is due to Hill; Hill was working in the 1940's and 50's and became an advocate of randomization on fairly practical grounds (reducing bias)

In the 1920's and 1930's, R A Fisher (who we met in the first lecture, leaning thoughtfully over his calculator) was promoting the idea of randomization from a technical perspective; **to Fisher, randomization gave rise to valid statistical procedures**

Fisher and randomization

"*The theory of estimation presupposes a process of random sampling. All our conclusions within that theory rest on this basis; without it our tests of significance would be worthless. ... In controlled experimentation it has been found not difficult to introduce explicit and objective randomisation in such a way that the tests of significance are demonstrably correct. In other cases we must still act in faith that Nature has done the randomisation for us....* ***We now recognise randomisation as a postulate necessary to the validity of our conclusions, and the modern experimenter is careful to make sure that this postulate is justified.***"

Fisher RA. *Development of the theory of experimental design*. Proceedings of the International Statistical Conferences 1947;3:434–39

Another aside: Fisher and Hill

There is, in fact, an interesting story that connects these two researchers; both were active in roughly the same time period and they were certainly aware of each other's work

They exchanged correspondence starting in 1929, **"Dear Sir"**; and then in 1931 **"Dear Fisher"** and **"Dear Bradford Hill";** and then in 1940 **"My dear Fisher"** and **"My dear Bradford Hill"**; and then by 1952 **"My dear Ron"** and **"My dear Tony"** (Hill went by Tony)

But by 1958 they were back to **"Dear Fisher"** and **"Dear Bradford Hill"** as the two (Doll, a significant co-investigator with Hill) were on opposites sides in a dispute as to **whether or not smoking caused lung cancer**

From the point of view of our discussion, one of Fisher's main criticisms of the studies suggesting that smoking caused lung cancer was the fact tha**t they were entirely observational** -- He wanted a "properly randomized experiment" (which of course would be difficult as you can't force people to start smoking)

We will speak more about causation and what you can conclude from different types of studies over the next couple of lectures

BRITISH MEDICAL JOURNAL

LONDON SATURDAY OCTOBER 30 1948

STREPTOMYCIN TREATMENT OF PULMONARY TUBERCULOSIS

A MEDICAL RESEARCH COUNCIL INVESTIGATION

The following gives the short-term results of a controlled investigation into the effects of streptomycin on one type of pulmonary tuberculosis. The inquiry was planned and directed by the Streptomycin in Tuberculosis Trials Committee, composed of the following members: Dr. Geoffrey Marshall (chairman), Professor J. W. S. Blacklock, Professor C. Cameron, Professor N. B. Capon, Dr. R. Cruickshank, Professor J. H. Gaddum, Dr. F. R. G. Heaf, Professor A. Bradford Hill, Dr. L. E. Houghton, Dr. J. Clifford Hoyle, Professor H. Raistrick, Dr. J. G. Scadding, Professor W. H. Tytler, Professor G. S. Wilson, and Dr. P. D'Arcy Hart (secretary). The centres at which the work was carried out and the specialists in charge of patients and pathological work were as follows:

Brompton Hospital, London.—Clinician: Dr. J. W. Crofton, Streptomycin Registrar (working under the direction of the honorary staff of Brompton Hospital); Pathologists: Dr. J. W. Clegg, Dr. D. A. Mitchison.

Colindale Hospital (L.C.C.), London.—Clinicians: Dr. J. V. Hurford, Dr. B. J. Douglas Smith, Dr. W. E. Snell; Pathologists (Central Public Health Laboratory): Dr. G. B. Forbes, Dr. H. D. Holt.

Harefield Hospital (M.C.C.), Harefield, Middlesex.—Clinicians: Dr. R. H. Brent, Dr. L. E. Houghton; Pathologist: Dr. E. Nassau.

Bangour Hospital, Bangour, West Lothian.—Clinician: Dr. I. D. Ross; Pathologist: Dr. Isabella Purdie.

Killingbeck Hospital and Sanatorium, Leeds.—Clinicians: Dr. W. Santon Gilmour, Dr. A. M. Reevie; Pathologist: Professor J. W. McLeod.

Northern Hospital (L.C.C.), Winchmore Hill, London.—Clinicians: Dr. F. A. Nash, Dr. R. Shoulman; Pathologists: Dr. J. M. Alston, Dr. A. Mohun.

Sully Hospital, Sully, Glam.—Clinicians: Dr. D. M. E. Thomas, Dr. L. R. West; Pathologist: Professor W. H. Tytler.

The clinicians of the centres met periodically as a working subcommittee under the chairmanship of Dr. Geoffrey Marshall; so also did the pathologists under the chairmanship of Dr. R. Cruickshank. Dr. Marc Daniels, of the Council's scientific staff, was responsible for the clinical co-ordination of the trials, and he also prepared the report for the Committee, with assistance from Dr. D. A. Mitchison on the analysis of laboratory results. For the purpose of final analysis the radiological findings were assessed by a panel composed of Dr. L. G. Blair, Dr. Peter Kerley, and Dr. Geoffrey S. Todd.

Introduction

When a special committee of the Medical Research Council undertook in September, 1946, to plan clinical trials of streptomycin in tuberculosis the main problem faced was that of investigating the effect of the drug in pulmonary tuberculosis. This antibiotic had been discovered two years previously by Waksman (Schatz, Bugie, and Waksman, 1944); in the intervening period its power of inhibiting tubercle bacilli *in vitro*, and the results of treatment in experimental tuberculous infection in guinea-pigs, had been reported; these results were strikingly better than those with any previous chemotherapeutic agent in tuberculosis. Preliminary results of trials in clinical tuberculosis had been published (Hinshaw and Feldman, 1945; Hinshaw, Feldman, and Pfuetze, 1946; Keefer *et al.*, 1946); the clinical results in pulmonary tuberculosis were encouraging but inconclusive.

The natural course of pulmonary tuberculosis is in fact so variable and unpredictable that evidence of improvement or cure following the use of a new drug in a few cases cannot be accepted as proof of the effect of that drug. The history of chemotherapeutic trials in tuberculosis is filled with errors due to empirical evaluation of drugs (Hart, 1946); the exaggerated claims made for gold treatment, persisting over 15 years, provide a spectacular example. It had become obvious that, in future, conclusions regarding the clinical effect of a new chemotherapeutic agent in tuberculosis could be considered valid only if based on adequately controlled clinical trials (Hinshaw and Feldman, 1944). The one controlled trial of gold treatment (and the only report of an adequately controlled trial in tuberculosis we have been able to find in the literature) reported negative therapeutic results (Amberson, McMahon, and Pinner, 1931). In 1946 no controlled trial of streptomycin in pulmonary tuberculosis had been undertaken in the U.S.A. The Committee of the Medical Research Council decided then that a part of the small supply of streptomycin allocated to it for research purposes would be best employed in a rigorously planned investigation with concurrent controls.

The many difficulties of planning and conducting a trial of this nature are important enough to warrant a full description here of the methods of the investigation.

Plan and Conduct of the Trial

Type of Case

A first prerequisite was that all patients in the trial should have a similar type of disease. To avoid having to make allowances for the effect of forms of therapy other than bed-rest, the type of disease was to be one not suitable for other forms of therapy. The estimated chances of spontaneous regression must be small. On the other hand, the type of lesion should be such as to offer some prospect of action by an effective chemotherapeutic agent; for this reason old-standing disease, and disease with thick-walled

Hill's tuberculosis trial

Here are Hill's original results from his 1948 paper — what do you see?

Results at End of Six Months

Four of the 55 S patients (7%) and 14 of the 52 C patients (27%) died before the end of six months. The difference between the two series is statistically significant ; the probability of it occurring by chance is less than one in a hundred.

Assessment of condition at the end of the six-months period should be based on a judicious combination of changes in the radiological picture, changes in general condition, temperature, weight, sedimentation rate, and bacillary content of the sputum. We have not attempted a numerical evaluation of the relative importance of each of these, and changes in them will be reported in turn. Appreciation of the clinical effects of the drug have not been lacking in the many reports published within the past two years. So far as possible, the analysis in this report will deal with the more readily measurable data only.

The following preliminary analysis is based on changes in the radiological picture alone, this being in our opinion the most important single factor to consider ; it will be seen later that in the great majority of cases clinical and radiological changes followed similar trends.

TABLE II.—*Assessment of Radiological Appearance at Six Months as Compared with Appearance on Admission*

Radiological Assessment	Streptomycin Group		Control Group	
Considerable improvement ..	28	*51%*	4	*8%*
Moderate or slight improvement	10	*18%*	13	*25%*
No material change	2	*4%*	3	*6%*
Moderate or slight deterioration	5	*9%*	12	*23%*
Considerable deterioration ..	6	*11%*	6	*11%*
Deaths	4	*7%*	14	*27%*
Total	55	*100%*	52	*100%*

Some analysis with Hill's data

Here we create a 2x2 table for Hill's data; we will focus on whether or not patients survived to the end of the trial

Status	Treatment: C	Treatment: S	
Survived	38	51	89
Died	14	4	18
	52	55	107

Some analysis with Hill's data

Here When you read about these kinds of trials in the medical literature, it is not uncommon **to work with a single figure of merit** — Rather than look at the two conditional proportions, it is customary to look at their fraction

In this case, the ratio of the proportion of patients that died in the Streptomycin group (7.3%) to those that died in the Control group (27%) is 0.27 — Streptomycin reduced the rate of mortality by nearly a quarter

This ratio is often called **the relative risk** — The language comes from epidemiological studies where "treatment" is really exposure to some toxic substance and the outcome is not that you get better but that something horrible happens to you

Some analysis with Hill's data

On the face of it, things look promising for Streptomycin relative to the standard therapy, bed rest, but is that where our analysis stops?

How do we judge the size of an effect? In particular, could these results have occurred "by pure chance"?

And what is the model for chance here?

Significance Testing

With this example, we have the basic ingredients of how significance testing works.

We establish a **null hypothesis,** plausible statement (a model or scenario) which might explain some pattern in a given set of data. This hypothesis is made for the purposes of argument — a good null hypothesis is a statement that would be interesting to reject. Think of it as a kind of devil's advocate (or maybe straw man is a better reference as the test was about divine intervention, after all).

We then define **a test statistic**, some quantity calculated from our data that is used to evaluate how compatible the results are with those expected under the null hypothesis (if the hypothesized statement - or model or scenario - was true)

We then simulate the values of the test statistic using the null hypothesis. In our analysis of Arbuthnot's hypothesis, that meant simulating a series of data sets assuming the null hypothesis is true and there is a 50/50 chance of boys outnumbering girls in a given year. For each data set we compute the test statistic. The ensemble of simulated test statistics is often called a **null distribution**.

Finally, we compare the value of the test statistic we computed for our data to the values we obtained by simulation — If they are very different, we have evidence that the null hypothesis is wrong. The chance that we see a value of the test statistic in simulations as or more extreme than what we computed from our data is referred to as the **P-value** of the test.

R.A. Fisher proposed this measure to express the weight of evidence against a null hypothesis — the smaller the value, the stronger the evidence. Fisher, however, believed that it should be combined with other sources of information as you reason about the phenomenon you were studying.

Hill's study

So let's talk about each of these components in the context of Hill's randomized trial -- When testing the efficacy of a new medical procedure, **the natural null hypothesis is that it offers no improvement over the standard therapy**

Under this "model" we assume that the two treatments are the same, so that patients would have had **the same chance of survival under either** -- Put another way, **their outcome, whether they lived or died, would have been the same regardless of which group they were placed in**

Under this hypothesis, the table we see is merely the result of random assignment -- That is, 18 people would have died regardless of what group we assigned them to, and **the fact that we saw 4 in the Streptomycin group and 14 in the control group was purely the result of chance**

Hill's study

Therefore, under the null hypothesis, if we had chosen a different random assignment of patients, **we would still have 18 people who died and 89 who survived, but they would appear in different cells of the table**

We can simulate under this "model" pretty easily -- That is, we take the 18 people who died and the 89 who survived and we re-randomize, **assigning 52 of them to the control group and 55 to the treatment group**

Let's see what that produces...

Simulating random assignments

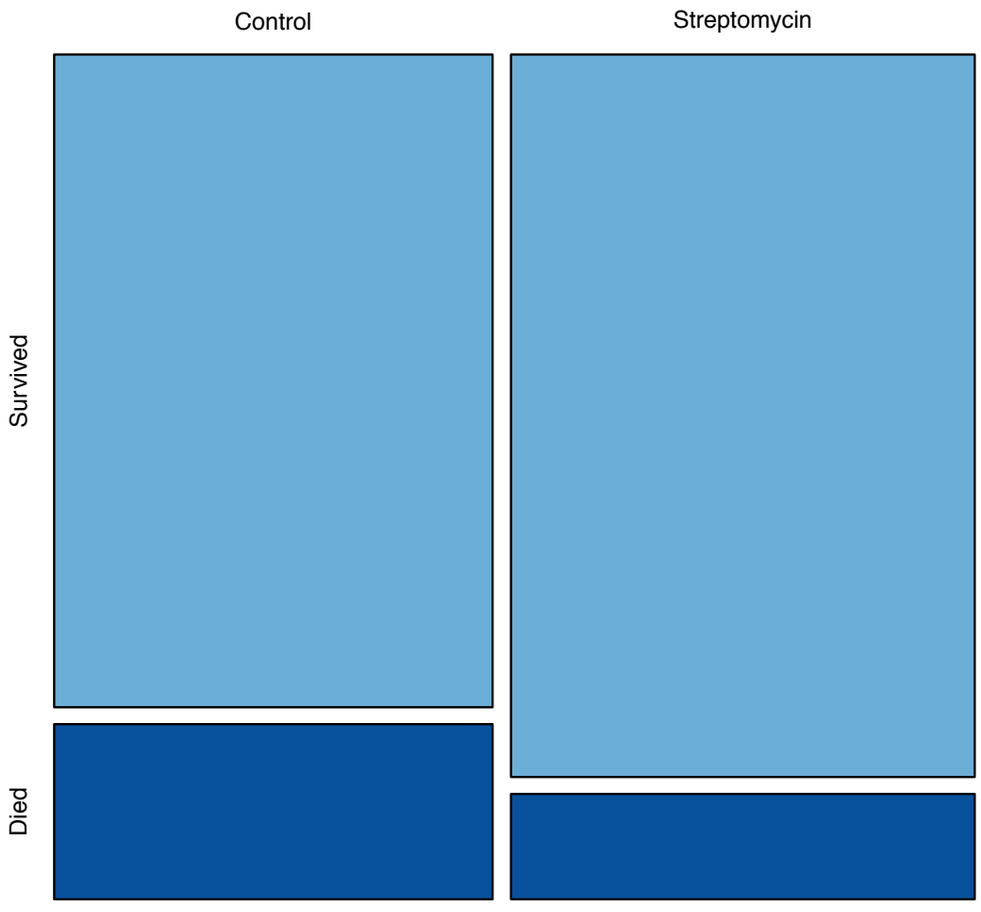
In this simulated table, we have 11/52 or 21% chance of dying under the control, and a 7/55 or 12% chance under Streptomycin; the treatment reduced the mortality rate among the participants by nearly 60%

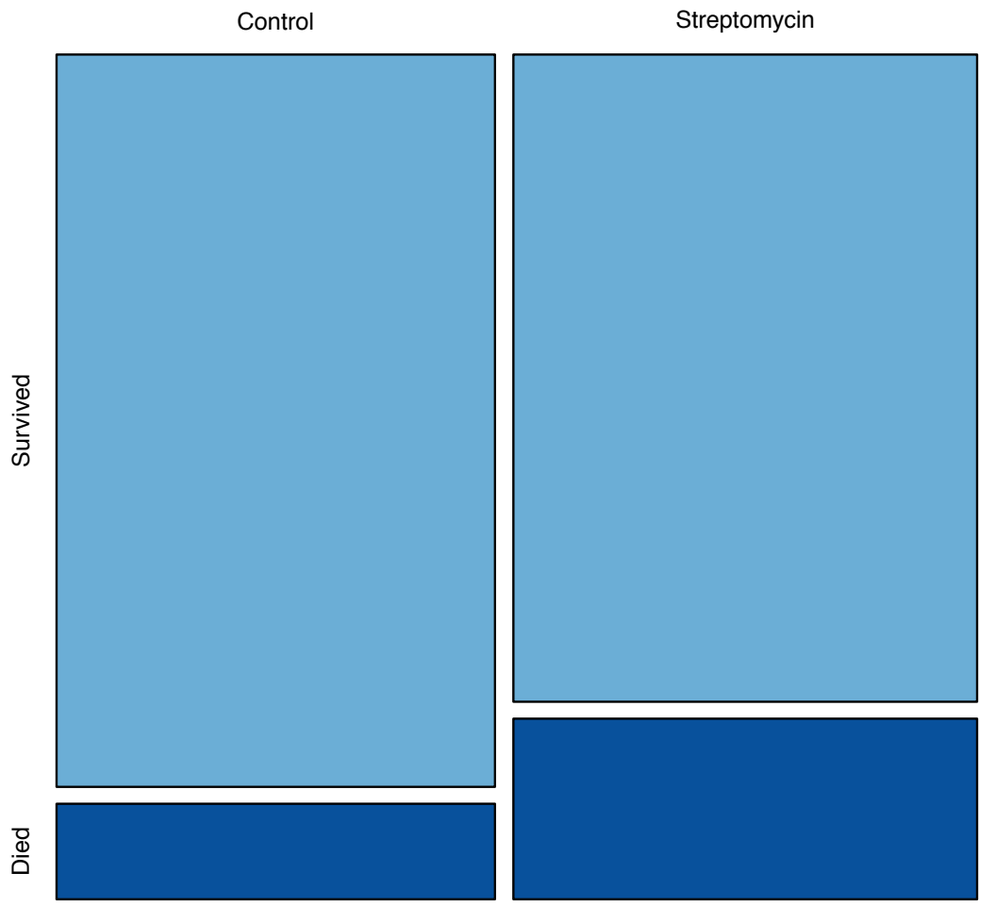
Status		Treatment: C	Treatment: S	
	Survived	41	48	89
	Died	11	7	18
		52	55	107

Simulating random assignments

In this simulated table, we have the opposite, with 6/52 or 12% chance of dying under the control, and a 12/55 or 22% chance under Streptomycin; the treatment almost doubled the mortality rate among the participants

Status	Treatment: C	Treatment: S	
Survived	46	43	89
Died	6	12	18
	52	55	107


Simulated data
Control
Streptomycin
Survived
Died


Simulated data
Control
Streptomycin
Survived
Died

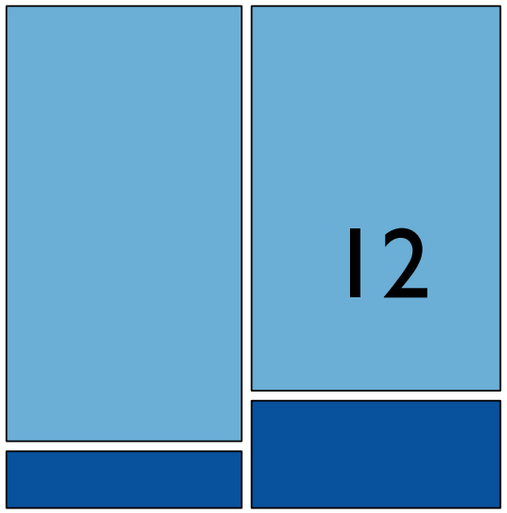
Simulating random assignments

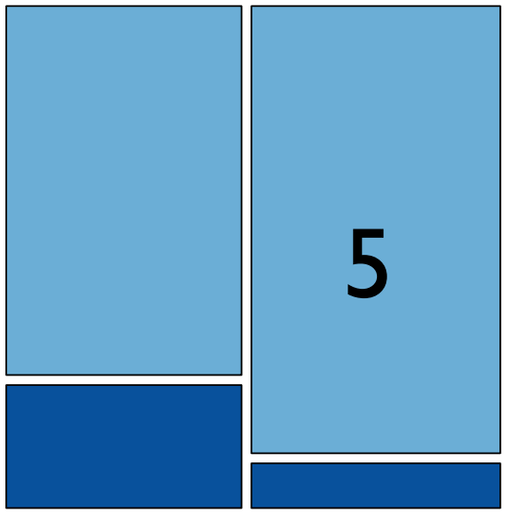
Notice that we only need to record **one piece of information for each trial, the number of deaths under Streptomycin** -- Knowing that we know all the other entries in the table

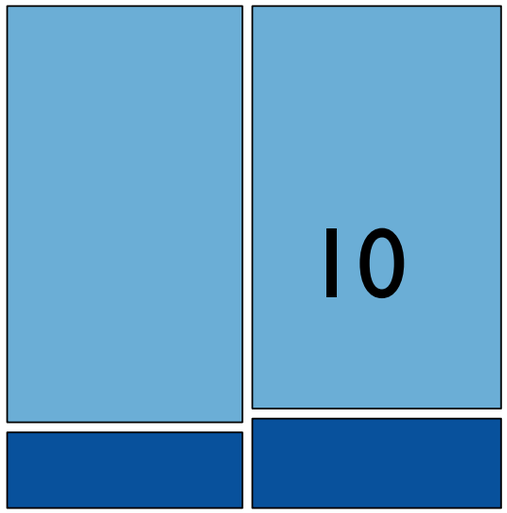
Using the language of significance testing, we will take **the number of patients in the Streptomycin group that died as our test statistic**

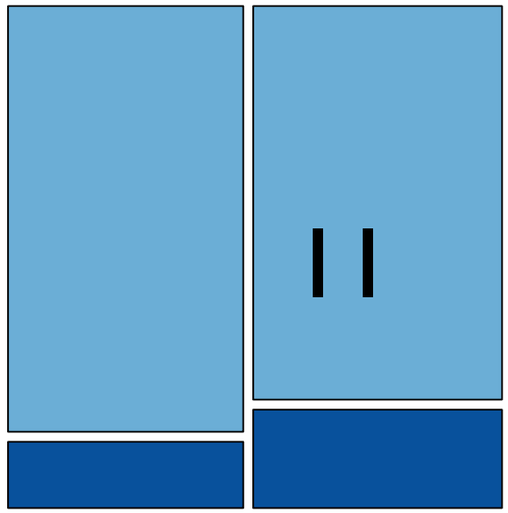
Therefore, the question becomes, under the random assignment patients to treatments, **how common is it for us to see 4 or fewer deaths in the Streptomycin group?**

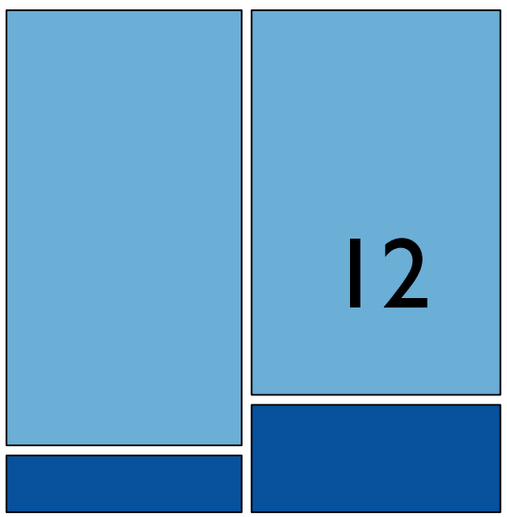
How would we figure this out?

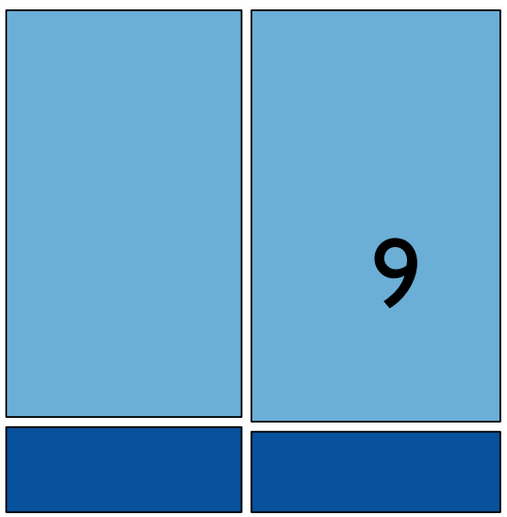

12

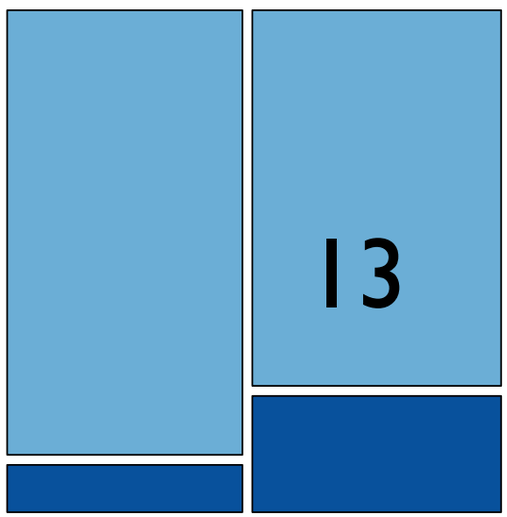

5

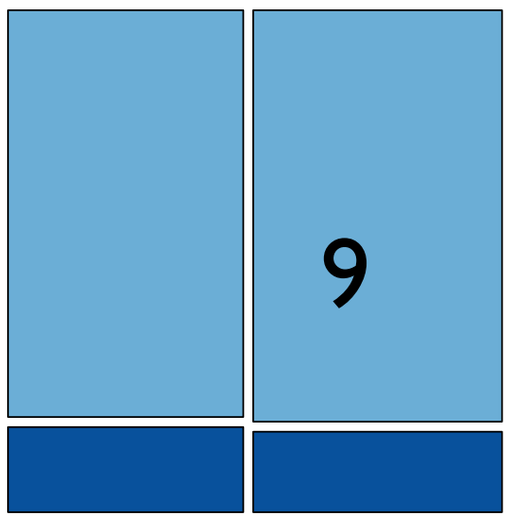

10

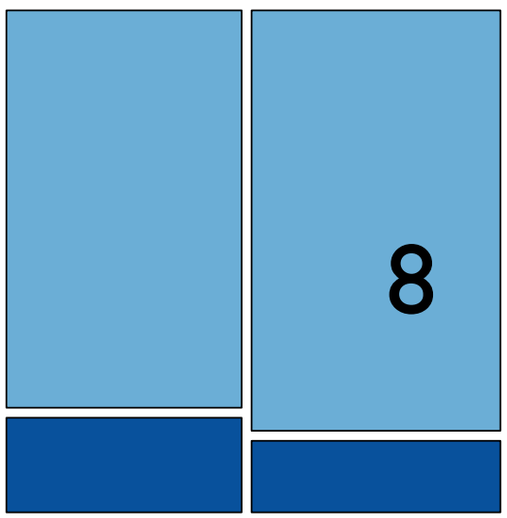

11

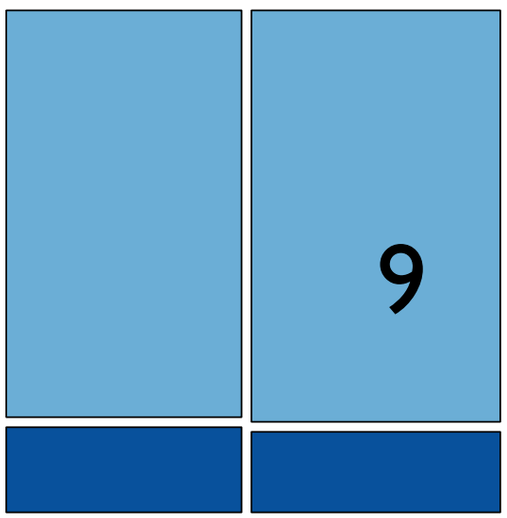

12

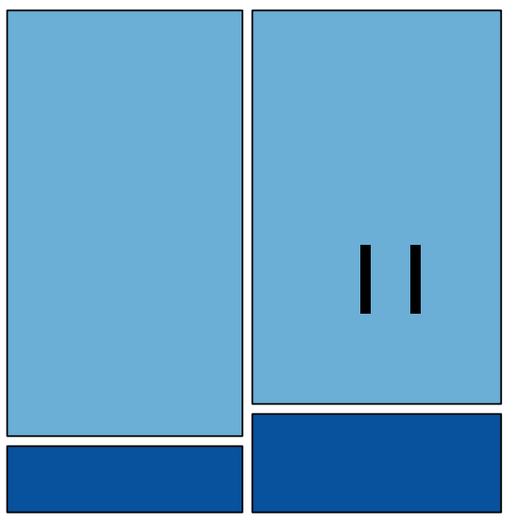

9

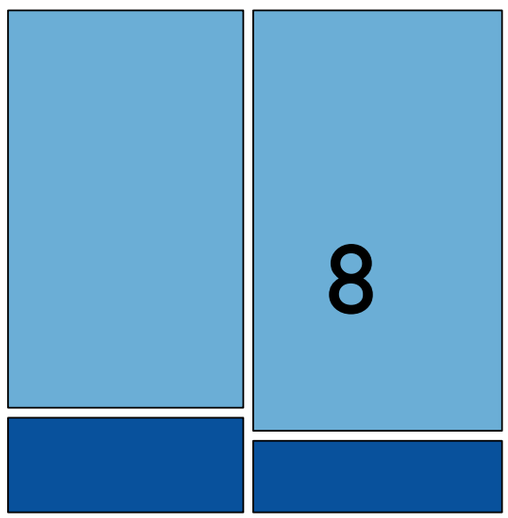

13


9

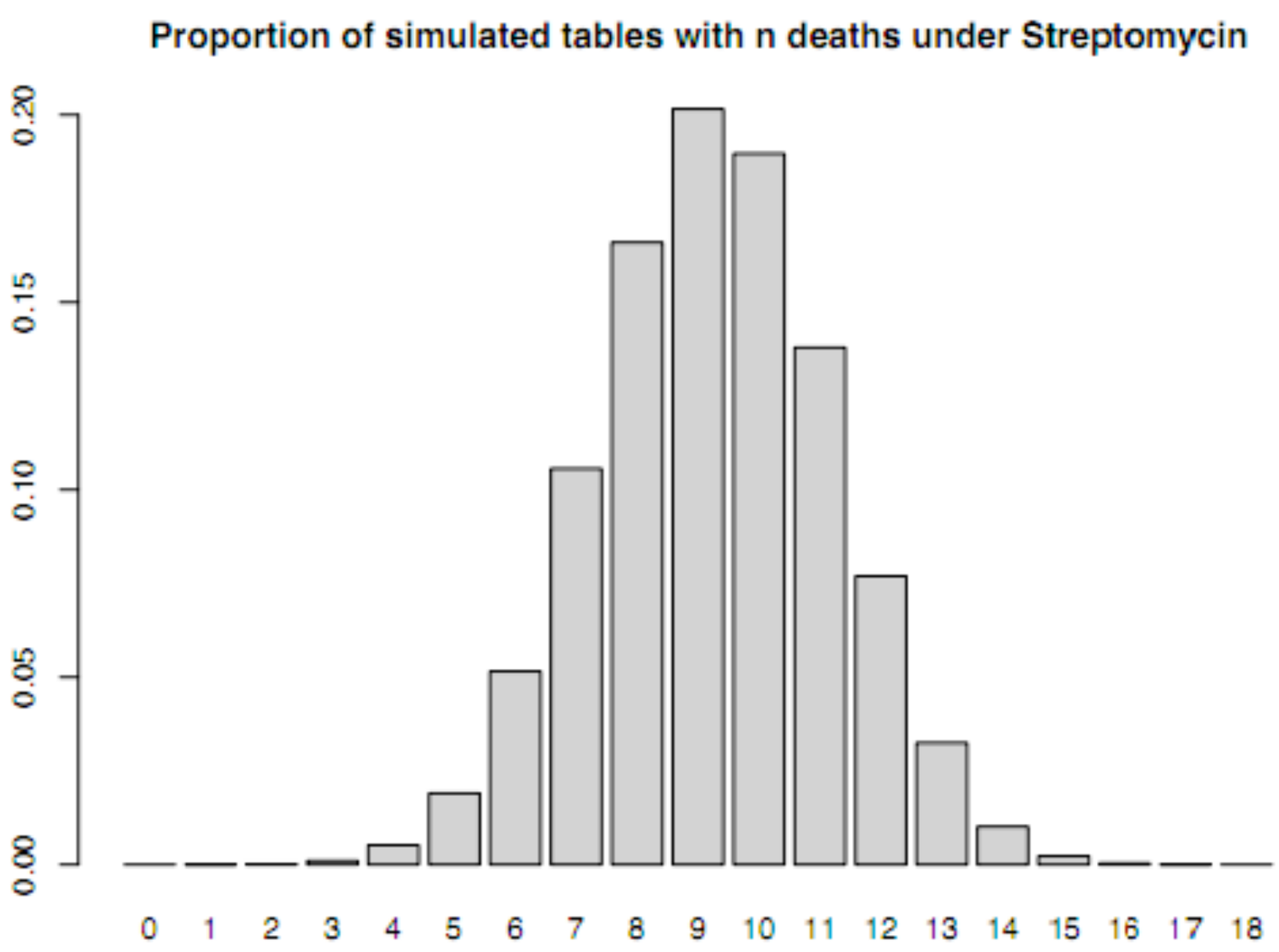

8


9


11


8

12	5	10	11
12	9	13	9
8	9	11	8

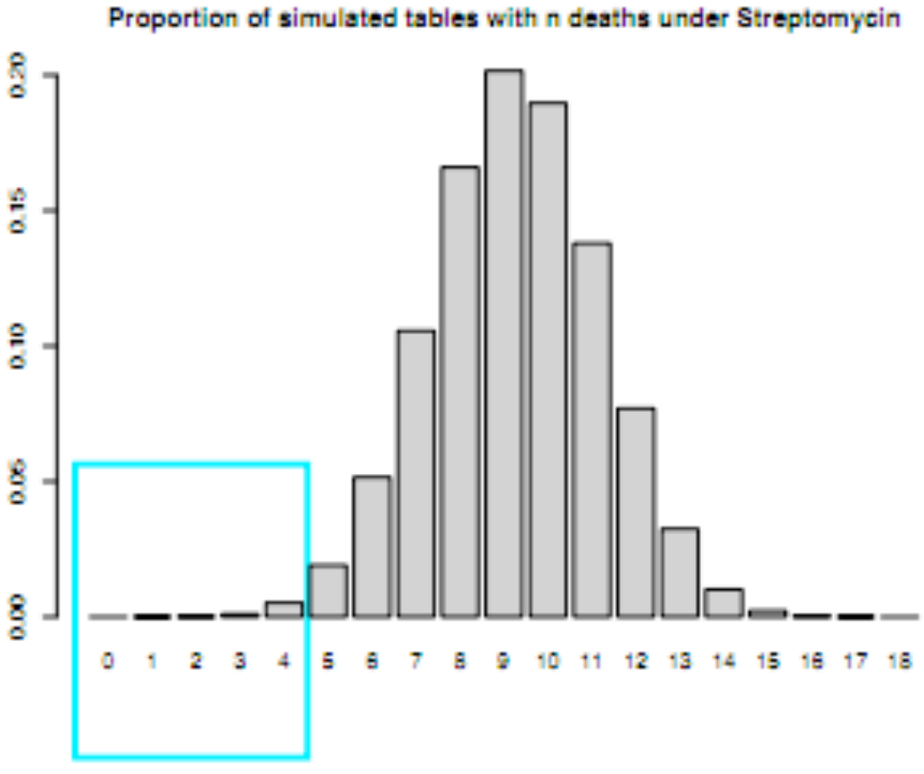

Proportion of simulated tables with n deaths under Streptomycin
0.00
0.05
0.10
0.15
0.20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Simulating random assignments

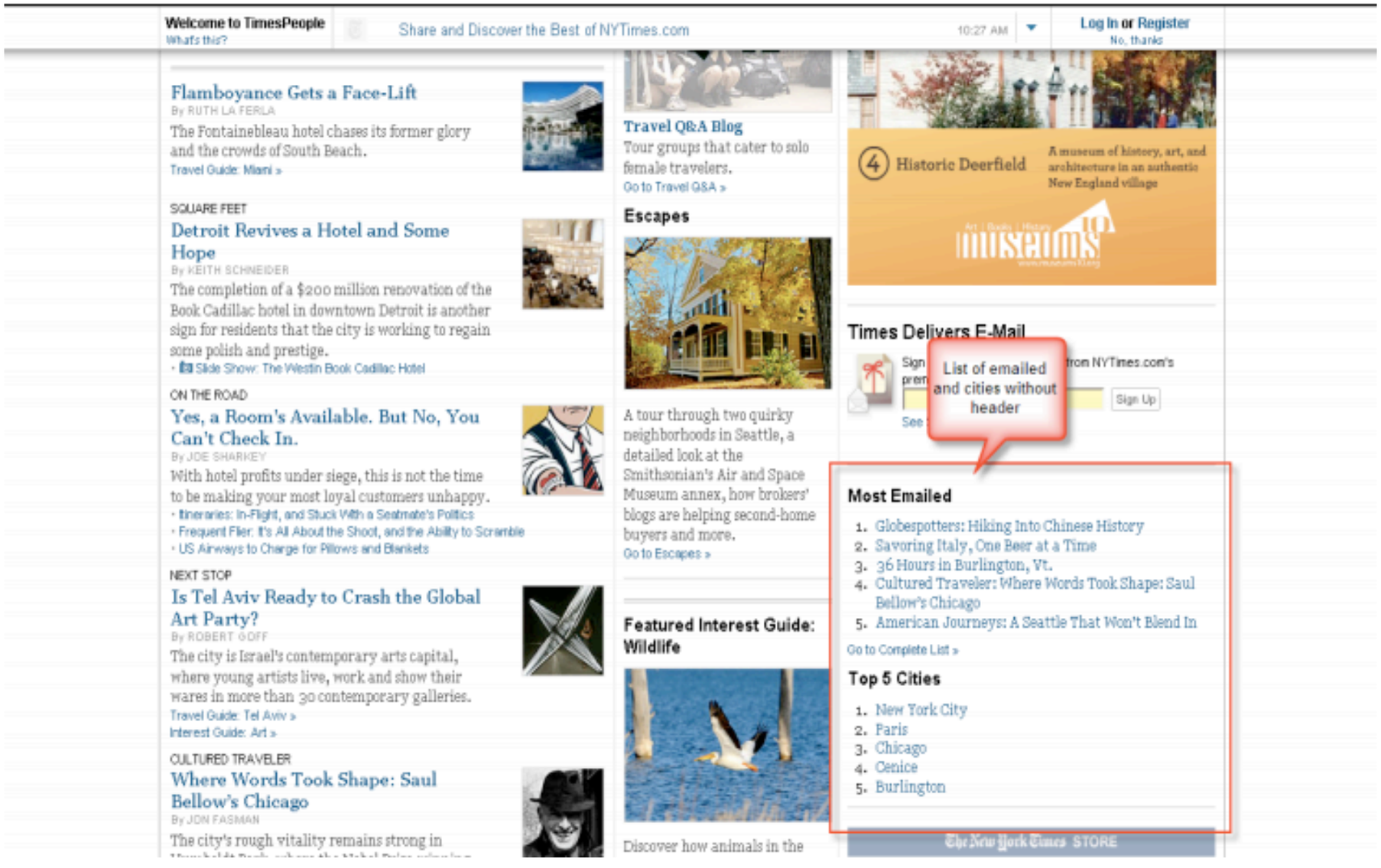
In this plot we see that a value as small or smaller than four is fairly rare; to be precise, only 0.6% of the tables have 4 or fewer deaths in the Streptomycin group

This, then, provides us with evidence that there is something more at work here than random assignment

If we believed the null hypothesis, that there was no difference between Streptomycin and bed rest, the results HIII observed would have been extremely rare, coming up a very small fraction of the time



List: Variation 10858



Tabs: Variation 10859