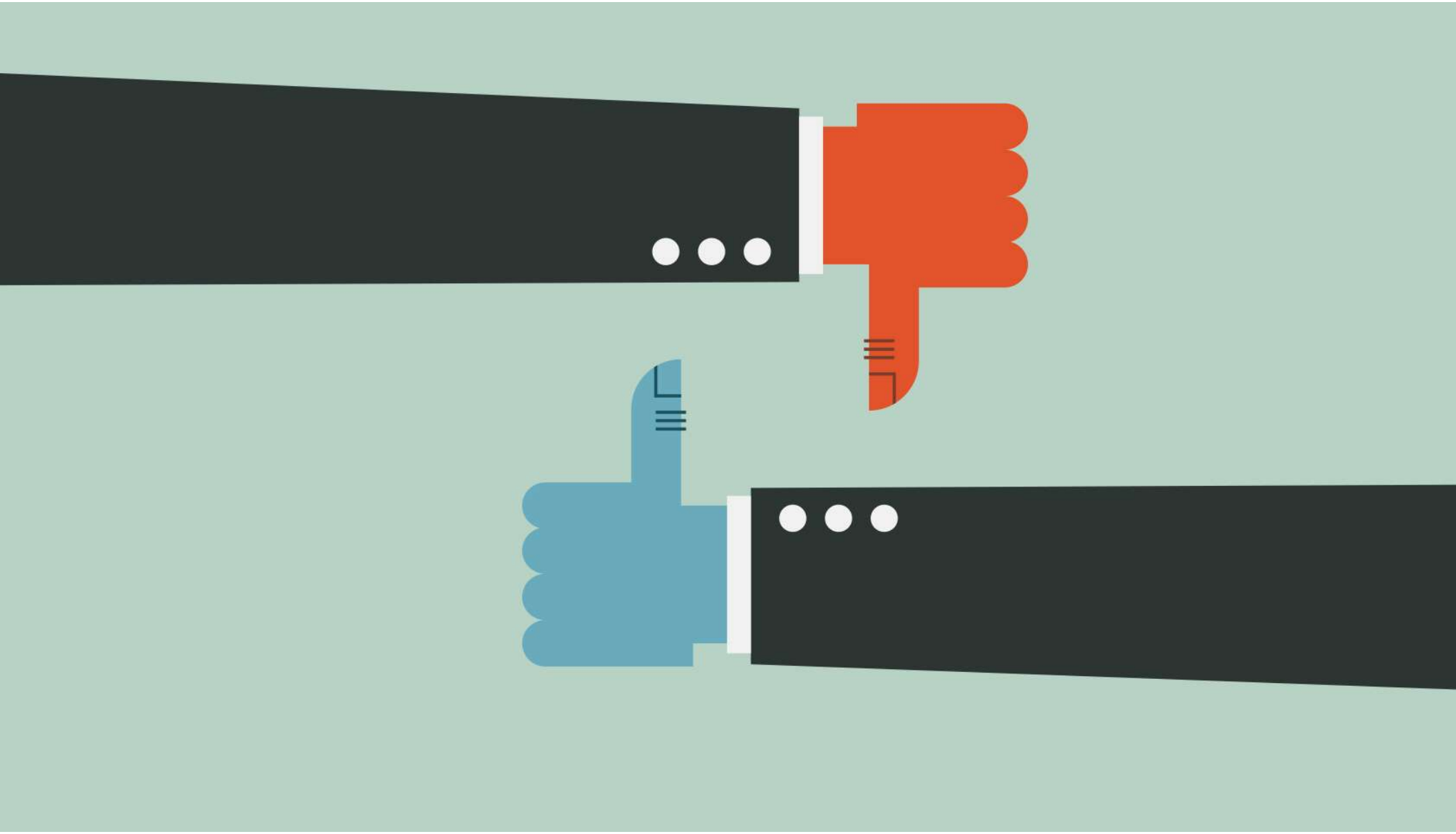


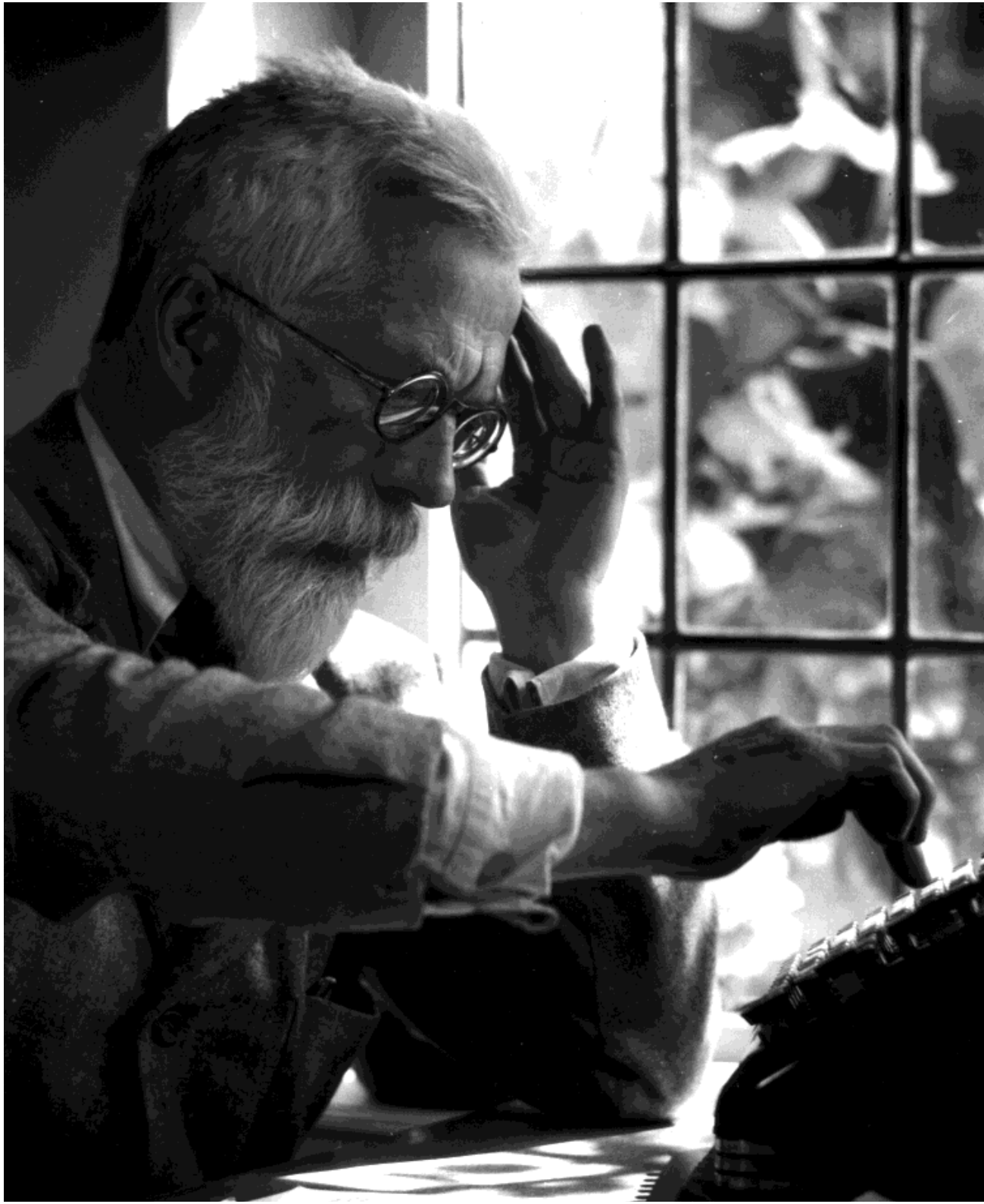
Lecture 2

Introduction to Statistical Inference — RCT's and A/B testing

R. A. Fisher is often credited as the single most important figure in 20th century statistics -- He is responsible for creating a mathematical framework for statistics (before Fisher, statistics has been characterized as an **“ingenious collection of ad hoc devices”**)

Fisher once commented that “he had learned all he knew” over his hand calculator; as we will see, these computations can reveal structures in data

Fisher’s data were often from designed experiments and were typically small in observation counts



John Tukey is another pioneer of statistical theory and practice, working through the later half of the 20th century -- He promoted the idea of looking at data, of exploratory analysis

In fact, he literally wrote the book on the subject; in *Exploratory Data Analysis* (1977), Tukey creates graphical tools for exploring features in data

His style is iterative, advocating many different analyses in an approach that is graphically and computationally intensive -- His unit of measure for the amount of analysis you'd performed was the side-inch (he would pour over stacks of computer output)

For Tukey, **"[w]hat the statistician often needs is enough different "looks'** to have a good chance to learn about this real world."

NORTH AMERICAN REVIEW

No. DLXX.

MAY, 1904.

THE COLLEGE OF JOURNALISM.

A Review of Criticisms and Objections—Reflections Upon the Power, the Progress and the Prejudices of the Press—Why Specialized Concentration and Education at College Would Improve the Character and Work of Journalists and So Promote the Welfare of the Republic.

> *"The man who writes, the man who month in and month out, week in and week out, day in and day out, furnishes the material which is to shape the thoughts of our people, is essentially the man who more than any other determines the character of the people and the kind of government this people shall possess."*
> —PRESIDENT ROOSEVELT, *April 7, 1904.*

BY JOSEPH PULITZER.

THE editor of the NORTH AMERICAN REVIEW has asked me to reply to an article recently printed in its pages criticising the College of Journalism which it has been my pleasure to found and permanently to endow in Columbia University. In complying with his request I have enlarged the scope of the reply to include all other criticisms and misgivings, many honest, some shallow, some based on misunderstanding, but the most representing only prejudice and ignorance. If my comment upon these criticisms shall seem to be diffuse and perhaps repetitious, my apology is that—alas!—I am compelled to write by voice, not by pen, and to revise the proofs by ear, not by eye—a somewhat difficult task.

Some of my critics have called my scheme "visionary." If it

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. 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.

Types of studies

In many branches of science, we are faced with two kinds of studies that differ in terms of the conditions under which the 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 care 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 — tThese kinds of considerations will lead to a range of (admittedly more technical0 questions you should sk of a data set

CDC - BRFSS

https://www.cdc.gov/brfss/

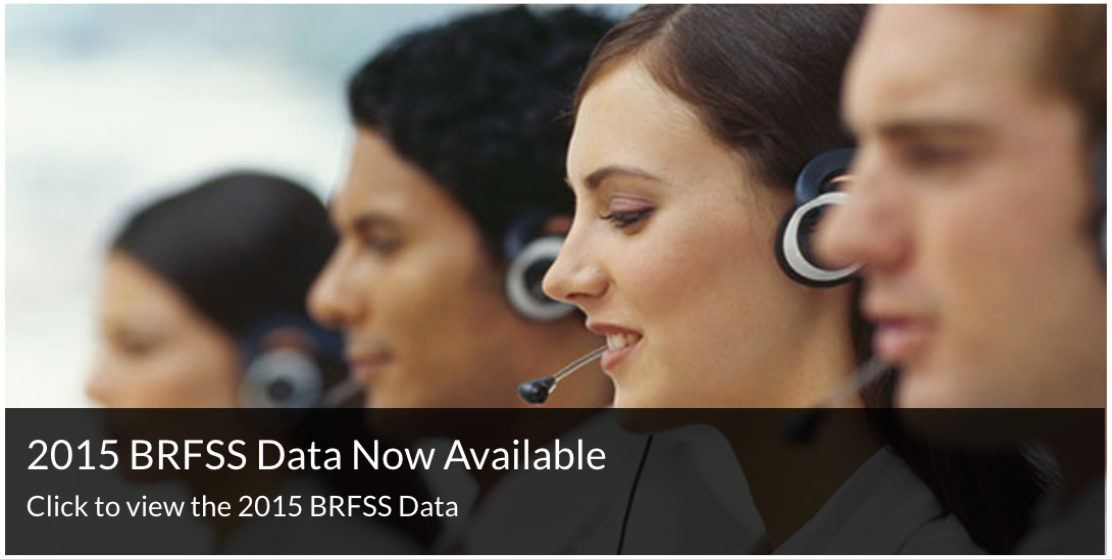
Apps ha Kernel

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

SEARCH

CDC A-Z INDEX

Behavioral Risk Factor Surveillance System



2015 BRFSS Data Now Available

Click to view the 2015 BRFSS Data

BRFSS™

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. See More.

BRFSS
About BRFSS
BRFSS Today
BRFSS History
BRFSS FAQs
Prevalence Data and Data Analysis Tools
Survey Data and Documentation
Questionnaires
Publications and Resources
State Information
Fact Sheets

About BRFSS

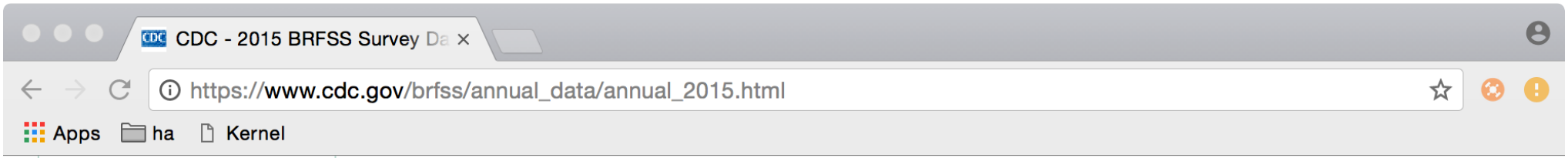
The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

By collecting behavioral health risk data at the state and local level, BRFSS has become a powerful tool for targeting and building health promotion activities. As a result, BRFSS users have increasingly demanded more data and asked for more questions on the survey. Currently, there is a wide sponsorship of the BRFSS survey, including most divisions in the CDC National Center for Chronic Disease Prevention and Health Promotion; other CDC centers; and federal agencies, such as the Health Resources and Services Administration, Administration on Aging, Department of Veterans Affairs, and Substance Abuse and Mental Health Services Administration.

In addition, countries eager to develop similar surveillance systems have requested technical assistance from BRFSS staff.

These countries include:

- Australia
- Brazil



1990 Data
1989 Data
1988 Data
1987 Data
1986 Data
1985 Data
1984 Data
Asthma Call-back Survey Data +
GIS Maps Data +
SMART: City and County Survey Data +
Statistical Briefs
Questionnaires
Publications and Resources +
State Information +
Fact Sheets

Data Files

There are 441, 456 records for 2015. More information on participation is available in the states conducting surveillance, by year table. The data files are provided in ASCII and SAS Transport formats.

2015 BRFSS Data (ASCII) [ZIP -64.3 MB]

Data released August 2016

This file for the combined landline and cell phone data set is in ASCII format. It has a fixed record length of 2155 positions.

2015 BRFSS Data (SAS Transport Format)

[ZIP - 96.5 MB]

Data released August 2016

This file for the combined landline and cell phone data set was exported from SAS V9.3 in the XPT transport format. This file contains 330 variables. This format can be imported into SPSS or STATA. Please note: some of the variable labels get truncated in the process of converting to the XPT format so they may be slightly different from what is on the SASOUT14.SAS program.

Variable Layout

Format information on variable name by column position.

The combined Landline and Cellular Telephone Survey

Resident of State

LandLine: 0.23 Land Line Introduction **Type:** Num

Column: 66 **SAS Variable Name:** STATERES

Prologue: Variable only on the land line survey

Description: Do you reside in ____(state)____?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes—Go to CELLFON3	254,643	100.00	100.00
BLANK	Not asked or Missing Notes: QSTVER >= 20	186,813		

Cellular Telephone

LandLine: 0.24 Land Line Introduction **Type:** Num

Column: 67 **SAS Variable Name:** CELLFON3

Prologue: Variable only on the land line survey

Description: Is this a cellular telephone? (Telephone service over the internet counts as landline service (includes Vonage, Magic Jack and other home-based phone services).)[Read only if necessary: “By cellular (or cell) telephone we mean a telephone that is mobile and usable outside of your neighborhood.”]

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Not a cellular phone	115,830	45.49	44.42
2	Yes—Terminate Phone Call	138,816	54.51	55.58
BLANK	Missing Notes: QSTVER >= 20	186,810		

Are you 18 years of age or older?

LandLine: 0.25 Land Line Introduction **Type:** Num

Column: 68 **SAS Variable Name:** LADULT

Prologue: Variable only on the land line survey

Description: Are you 18 years of age or older?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes, Male Respondent—Go to Survey Introduction	18	40.00	28.87
2	Yes, Female Respondent—Go to Survey Introduction	27	60.00	71.13
BLANK	Missing Notes: QSTVER >= 20	441,411		

General Health

Section: 1.1 Health Status **Type:** Num

Column: 90 **SAS Variable Name:** GENHLTH

Prologue:

Description: Would you say that in general your health is:

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Excellent	76,032	17.22	18.68
2	Very good	145,065	32.86	31.78
3	Good	136,975	31.03	31.59
4	Fair	58,962	13.36	13.06
5	Poor	23,175	5.25	4.60
7	Don’t know/Not Sure	799	0.18	0.18
9	Refused	446	0.10	0.11
BLANK	Not asked or Missing	2		

Number of Days Physical Health Not Good

Section: 2.1 Healthy Days — Health Related Quality of Life **Type:** Num

Column: 91-92 **SAS Variable Name:** PHYSHLTH

Prologue:

Description: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 30	Number of days	157,570	35.69	35.20
88	None	274,143	62.10	62.77
77	Don’t know/Not sure	7,664	1.74	1.51
99	Refused	2,078	0.47	0.52
BLANK	Not asked or Missing	1		

Number of Days Mental Health Not Good

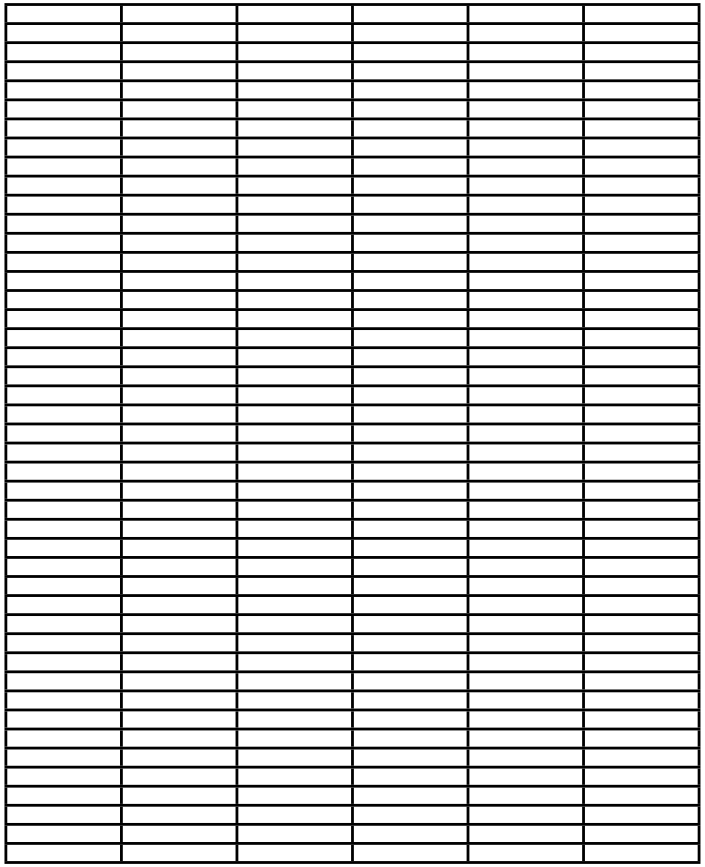
Section: 2.2 Healthy Days — Health Related Quality of Life **Type:** Num

Column: 93-94 **SAS Variable Name:** MENTHLTH

Prologue:

Description: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?

Value	Value Label	Frequency	Percentage	Weighted Percentage
1 - 30	Number of days Notes: _ _ Number of days	132,972	30.12	33.58
88	None	301,076	68.20	64.74
77	Don’t know/Not sure	5,204	1.18	1.11
99	Refused	2,204	0.50	0.56



	state	genhlth	physhlth	exerany	hlthplan	smoke100	height	weight	wtdesire	age	gender
1	Louisiana	good	0	0	1	0	70	175	175	77	m
2	Massachusetts	good	30	0	1	1	64	125	115	33	f
3	California	good	2	1	1	1	60	105	105	49	f
4	California	good	0	1	1	0	66	132	124	42	f
5	Ohio	very good	0	0	1	0	61	150	130	55	f
6	Pennsylvania	very good	0	1	1	0	64	114	114	55	f
7	California	very good	0	1	1	0	71	194	185	31	m
8	Texas	very good	1	0	1	0	67	170	160	45	m
9	California	good	2	0	1	1	65	150	130	27	f
10	Texas	good	3	1	1	0	70	180	170	44	m
11	Connecticut	excellent	4	1	1	1	69	186	175	46	m
12	Michigan	fair	30	1	1	1	69	168	148	62	m
13	New Mexico	excellent	0	1	0	1	66	185	220	21	m
14	Massachusetts	excellent	0	1	1	1	70	170	170	69	m
15	New Jersey	fair	3	1	0	0	69	170	170	23	m
16	California	good	0	1	1	1	73	185	175	79	m
17	California	good	0	0	0	1	67	156	150	47	m
18	Nebraska	fair	30	0	1	1	71	185	185	76	m
19	Alabama	good	0	1	1	1	75	200	190	43	m
20	Nevada	very good	0	1	1	0	67	125	120	33	f

RealClear Politics

Polls

Top 2018 Senate Races | Top 2018 Governor Races | Find Any Poll | Try 'State,' or 'Candidate' | Quick Poll/Map Links

Latest Polls

Senate Polls | Governor Polls | House Polls | Generic Ballot | State of Union Polls | All Election Polls

Thursday, August 9

Race/Topic (Click to Sort)	Poll	Results	Spread
President Trump Job Approval	Reuters/Ipsos	Approve 45, Disapprove 52	Disapprove +7
2018 Generic Congressional Vote	Reuters/Ipsos	Democrats 41, Republicans 39	Democrats +2

Wednesday, August 8

Race/Topic (Click to Sort)	Poll	Results	Spread
Maine Senate - Brakey vs. King	Suffolk*	King 52, Brakey 25	King +27
Maine Governor - Moody vs. Mills	Suffolk*	Moody 39, Mills 39	Tie
President Trump Job Approval	Rasmussen Reports	Approve 47, Disapprove 51	Disapprove +4
President Trump Job Approval	Economist/YouGov	Approve 44, Disapprove 51	Disapprove +7
2018 Generic Congressional Vote	Economist/YouGov	Democrats 44, Republicans 41	Democrats +3
2018 Generic Congressional Vote	IBD/TIPP	Democrats 45, Republicans 45	Tie

TheUpshot

THE NEW HEALTH CARE

Workplace Wellness Programs Don't Work Well. Why Some Studies Show Otherwise.

Randomized controlled trials, despite their flaws, remain a powerful tool.

By Aaron E. Carroll

Aug. 6, 2018

The gold standard of medical research, the randomized controlled trial, has been taking a bit of a beating lately.

An entire issue of the journal Social Science and Medicine was recently devoted to it, with many articles pointing to shortcomings. Others have argued that randomized controlled trials often can't address the questions that patients and physicians most want answered. I recently wrote about the limitations of the method in studying effectiveness, which is what we care about in real-world situations.

But the randomized controlled trial remains a powerful tool. It's still, perhaps, the best method for conducting explanatory research. In past articles, I have recounted numerous times when hypotheses from observational studies, those based solely on observations of particular groups, have failed to be confirmed by a controlled trial.

The Illinois Workplace Wellnes ×

Not Secure | www.nber.org/workplacewellness/

The Illinois Workplace Wellness Study

HOME BACKGROUND RESULTS DOWNLOADS MEDIA SUPPORT TEAM CONTACT NBER

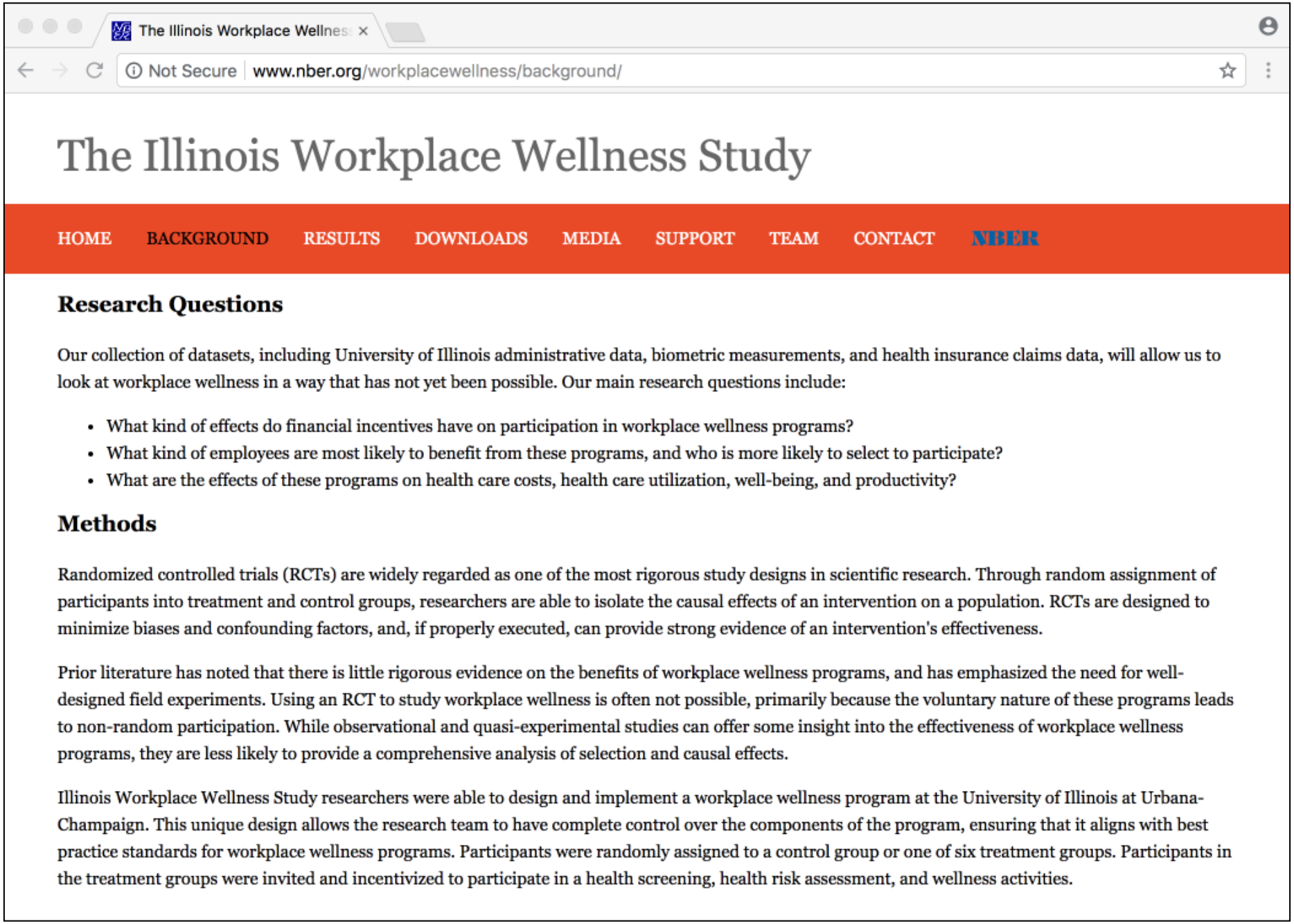
What Do Workplace Wellness Programs Do?

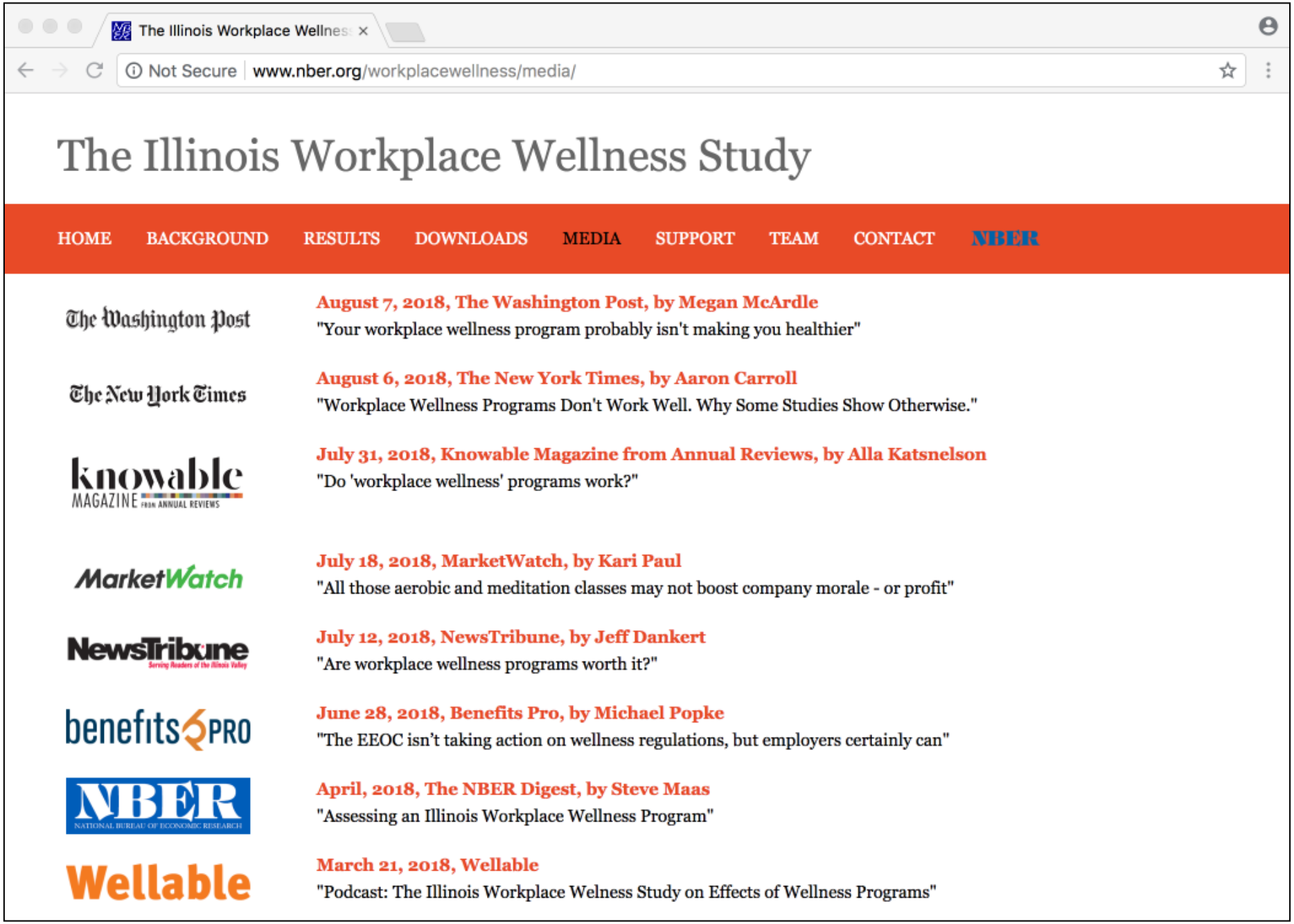
We designed a large-scale, randomized controlled trial (RCT) to evaluate their effectiveness

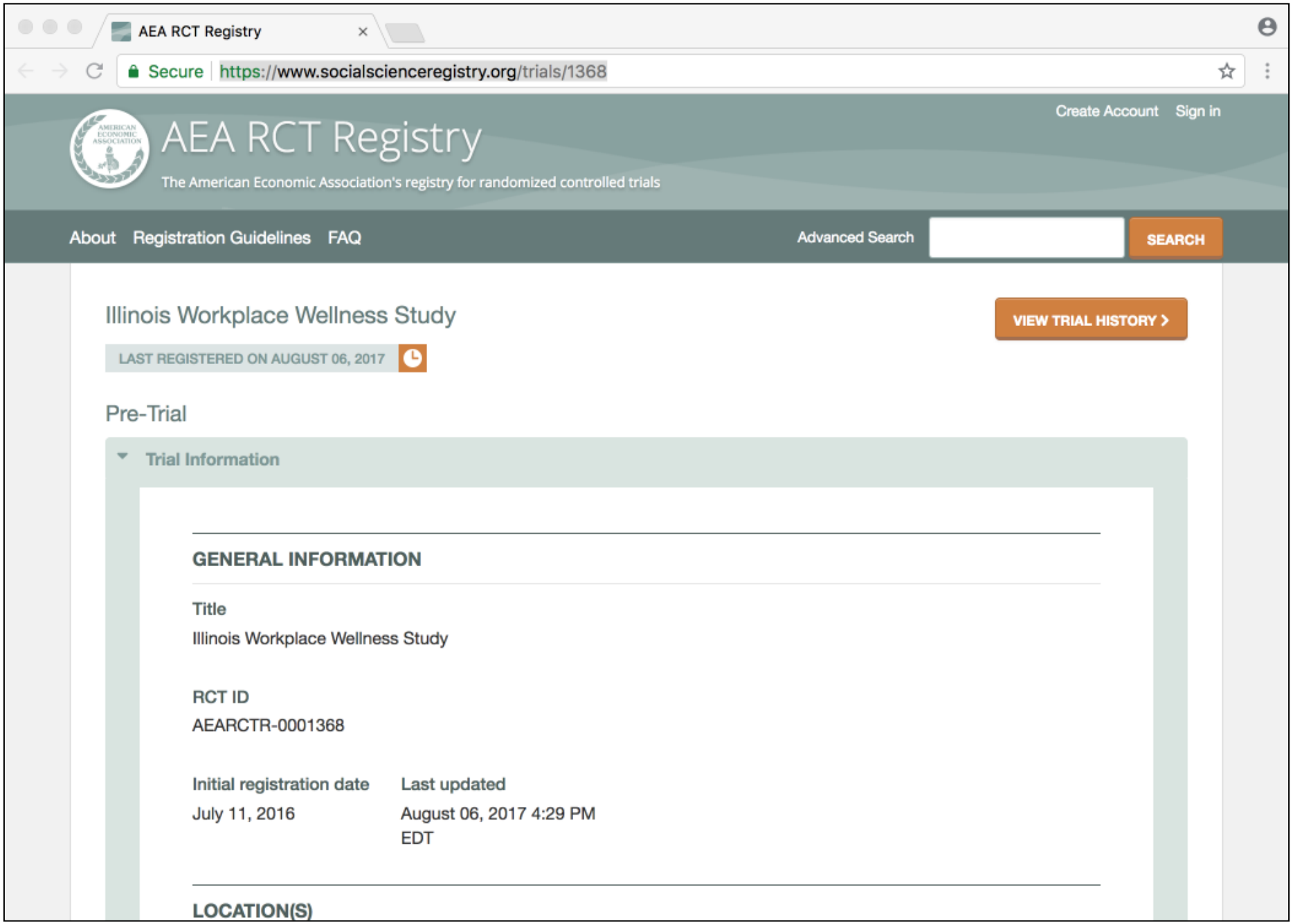
The Illinois Workplace Wellness Study is a large, randomized controlled trial of a comprehensive wellness program at the University of Illinois at Urbana-Champaign. The study is designed to:

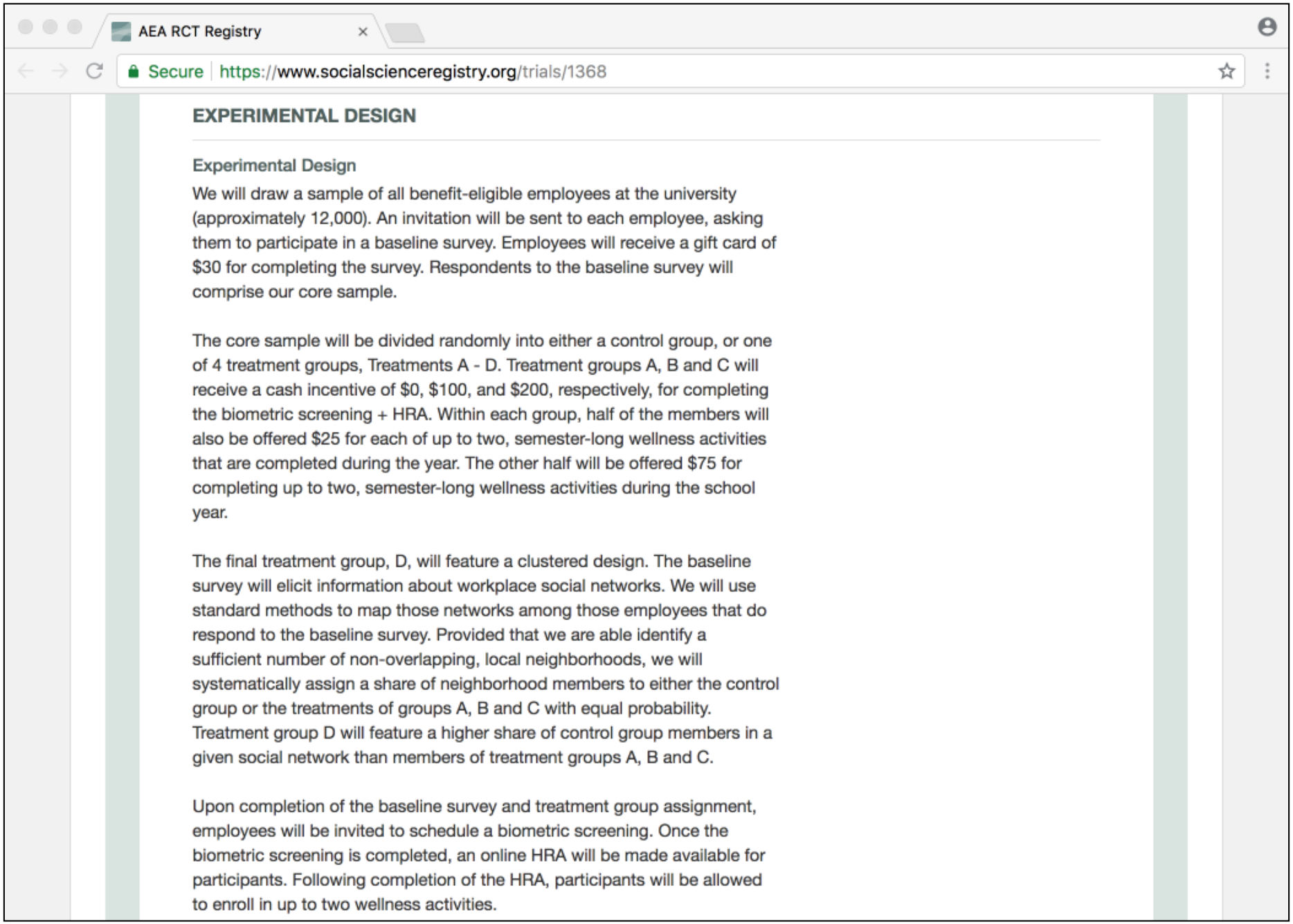
- Examine the effects of financial incentives on workplace wellness participation
- Investigate who benefits from workplace wellness programs
- Estimate the causal effect of workplace wellness on employee health care costs, health behaviors, well-being, and productivity
- Test for peer effects in wellness program participation

The Illinois Workplace Wellness Study aims to inform the national conversation surrounding workplace wellness, drawing from strong scientific evidence and an innovative study design. The study's findings will empower employers, public health professionals, and policymakers to make more informed decisions regarding the implementation of workplace wellness programs throughout the United States. Below we summarize the first set of results from the multi-year study.


The Illinois Workplace Wellnes
Not Secure
www.nber.org/workplacewellness/background/
The Illinois Workplace Wellness Study
HOME
BACKGROUND
RESULTS
DOWNLOADS
MEDIA
SUPPORT
TEAM
CONTACT
NBER
Research Questions
Our collection of datasets, including University of Illinois administrative data, biometric measurements, and health insurance claims data, will allow us to look at workplace wellness in a way that has not yet been possible. Our main research questions include:
What kind of effects do financial incentives have on participation in workplace wellness programs?
What kind of employees are most likely to benefit from these programs, and who is more likely to select to participate?
What are the effects of these programs on health care costs, health care utilization, well-being, and productivity?
Methods
Randomized controlled trials (RCTs) are widely regarded as one of the most rigorous study designs in scientific research. Through random assignment of participants into treatment and control groups, researchers are able to isolate the causal effects of an intervention on a population. RCTs are designed to minimize biases and confounding factors, and, if properly executed, can provide strong evidence of an intervention's effectiveness.
Prior literature has noted that there is little rigorous evidence on the benefits of workplace wellness programs, and has emphasized the need for well-designed field experiments. Using an RCT to study workplace wellness is often not possible, primarily because the voluntary nature of these programs leads to non-random participation. While observational and quasi-experimental studies can offer some insight into the effectiveness of workplace wellness programs, they are less likely to provide a comprehensive analysis of selection and causal effects.
Illinois Workplace Wellness Study researchers were able to design and implement a workplace wellness program at the University of Illinois at Urbana-Champaign. This unique design allows the research team to have complete control over the components of the program, ensuring that it aligns with best practice standards for workplace wellness programs. Participants were randomly assigned to a control group or one of six treatment groups. Participants in the treatment groups were invited and incentivized to participate in a health screening, health risk assessment, and wellness activities.


The Illinois Workplace Wellnes
Not Secure
www.nber.org/workplacewellness/media/
The Illinois Workplace Wellness Study
HOME
BACKGROUND
RESULTS
DOWNLOADS
MEDIA
SUPPORT
TEAM
CONTACT
NBER
The Washington Post
August 7, 2018, The Washington Post, by Megan McArdle
"Your workplace wellness program probably isn't making you healthier"
The New York Times
August 6, 2018, The New York Times, by Aaron Carroll
"Workplace Wellness Programs Don't Work Well. Why Some Studies Show Otherwise."
knowable MAGAZINE FROM ANNUAL REVIEWS
July 31, 2018, Knowable Magazine from Annual Reviews, by Alla Katsnelson
"Do 'workplace wellness' programs work?"
MarketWatch
July 18, 2018, MarketWatch, by Kari Paul
"All those aerobic and meditation classes may not boost company morale - or profit"
NewsTribune
July 12, 2018, NewsTribune, by Jeff Dankert
"Are workplace wellness programs worth it?"
benefitspro
June 28, 2018, Benefits Pro, by Michael Popke
"The EEOC isn't taking action on wellness regulations, but employers certainly can"
NBER
NATIONAL BUREAU OF ECONOMIC RESEARCH
April, 2018, The NBER Digest, by Steve Maas
"Assessing an Illinois Workplace Wellness Program"
Wellable
March 21, 2018, Wellable
"Podcast: The Illinois Workplace Welness Study on Effects of Wellness Programs"


AEA RCT Registry
Secure https://www.socialscienceregistry.org/trials/1368
Create Account Sign in
AEA RCT Registry
The American Economic Association's registry for randomized controlled trials
About Registration Guidelines FAQ
Advanced Search
SEARCH
Illinois Workplace Wellness Study
VIEW TRIAL HISTORY >
LAST REGISTERED ON AUGUST 06, 2017
Pre-Trial
Trial Information
GENERAL INFORMATION
Title
Illinois Workplace Wellness Study
RCT ID
AEARCTR-0001368
Initial registration date
July 11, 2016
Last updated
August 06, 2017 4:29 PM EDT
LOCATION(S)



EXPERIMENTAL DESIGN

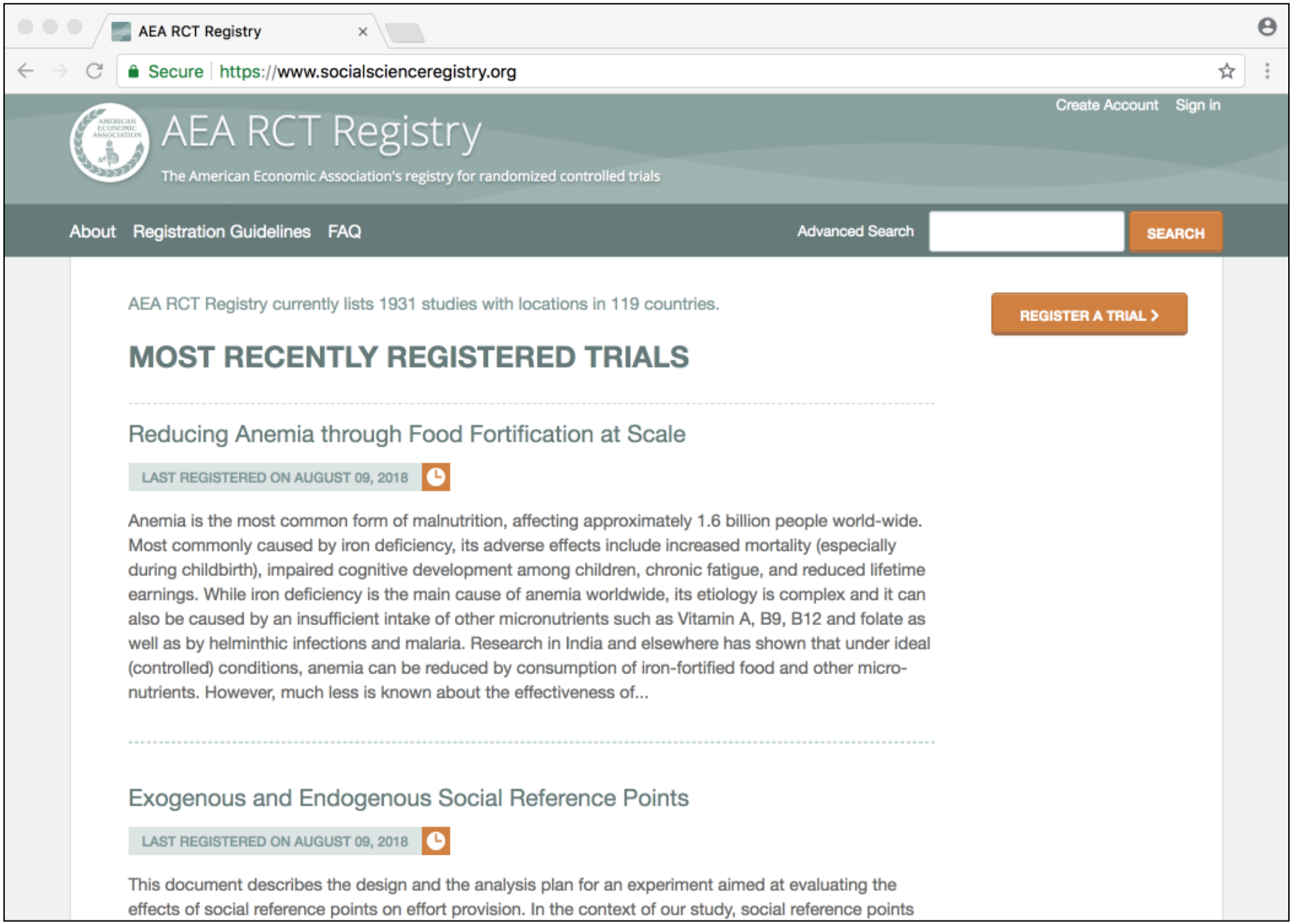
Experimental Design

We will draw a sample of all benefit-eligible employees at the university (approximately 12,000). An invitation will be sent to each employee, asking them to participate in a baseline survey. Employees will receive a gift card of $30 for completing the survey. Respondents to the baseline survey will comprise our core sample.

The core sample will be divided randomly into either a control group, or one of 4 treatment groups, Treatments A - D. Treatment groups A, B and C will receive a cash incentive of $0, $100, and $200, respectively, for completing the biometric screening + HRA. Within each group, half of the members will also be offered $25 for each of up to two, semester-long wellness activities that are completed during the year. The other half will be offered $75 for completing up to two, semester-long wellness activities during the school year.

The final treatment group, D, will feature a clustered design. The baseline survey will elicit information about workplace social networks. We will use standard methods to map those networks among those employees that do respond to the baseline survey. Provided that we are able identify a sufficient number of non-overlapping, local neighborhoods, we will systematically assign a share of neighborhood members to either the control group or the treatments of groups A, B and C with equal probability. Treatment group D will feature a higher share of control group members in a given social network than members of treatment groups A, B and C.

Upon completion of the baseline survey and treatment group assignment, employees will be invited to schedule a biometric screening. Once the biometric screening is completed, an online HRA will be made available for participants. Following completion of the HRA, participants will be allowed to enroll in up to two wellness activities.


AEA RCT Registry
Secure https://www.socialscienceregistry.org
Create Account Sign in
AEA RCT Registry
The American Economic Association's registry for randomized controlled trials
About Registration Guidelines FAQ
Advanced Search
SEARCH
AEA RCT Registry currently lists 1931 studies with locations in 119 countries.
REGISTER A TRIAL >
MOST RECENTLY REGISTERED TRIALS
Reducing Anemia through Food Fortification at Scale
LAST REGISTERED ON AUGUST 09, 2018
Anemia is the most common form of malnutrition, affecting approximately 1.6 billion people world-wide. Most commonly caused by iron deficiency, its adverse effects include increased mortality (especially during childbirth), impaired cognitive development among children, chronic fatigue, and reduced lifetime earnings. While iron deficiency is the main cause of anemia worldwide, its etiology is complex and it can also be caused by an insufficient intake of other micronutrients such as Vitamin A, B9, B12 and folate as well as by helminthic infections and malaria. Research in India and elsewhere has shown that under ideal (controlled) conditions, anemia can be reduced by consumption of iron-fortified food and other micro-nutrients. However, much less is known about the effectiveness of...
Exogenous and Endogenous Social Reference Points
LAST REGISTERED ON AUGUST 09, 2018
This document describes the design and the analysis plan for an experiment aimed at evaluating the effects of social reference points on effort provision. In the context of our study, social reference points

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

Figure. Flow of Patients Through Acute Extremity Pain Trial

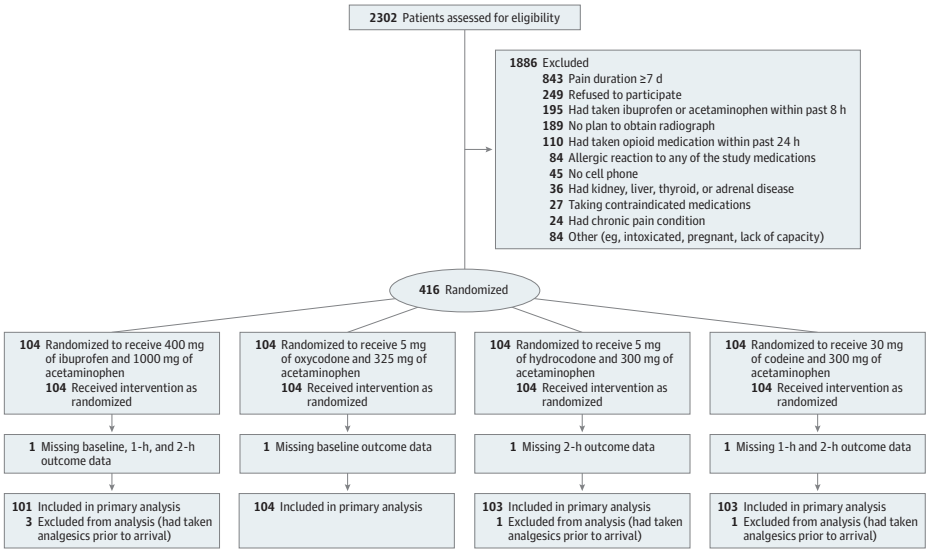


Table 1. Patient Characteristics

	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

	Between-Group Difference in Mean Change in NRS Pain Score (99.2% CI)[a]	
Comparison	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.

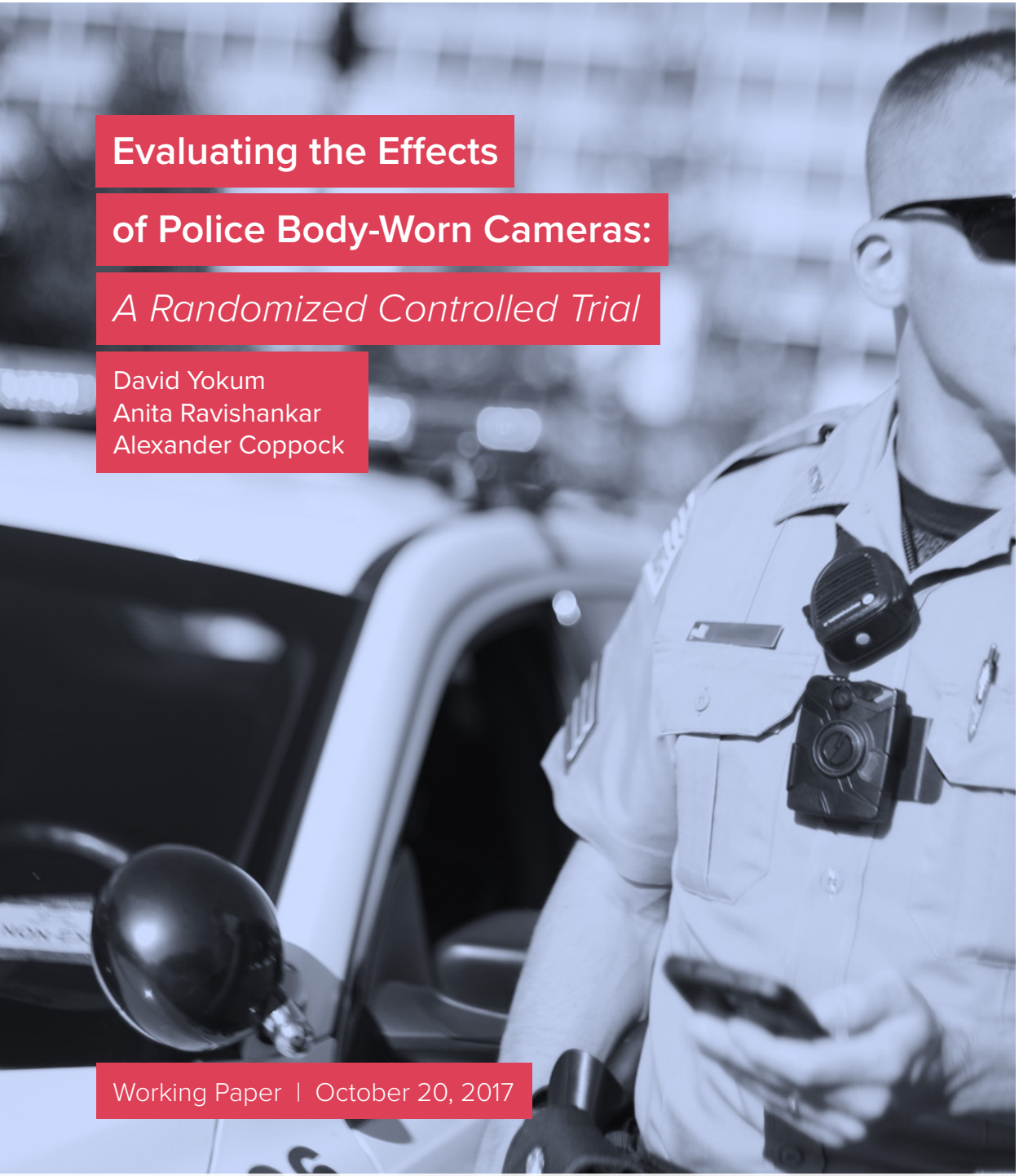
patients who received rescue analgesics were nearly identical to the analysis without imputation (eTable 1 and eTable 2 in Supplement 2). There were no clinically important or statistically significant differences in efficacy when these post hoc analyses were performed.

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



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

David Yokum
Anita Ravishankar
Alexander Coppock

Working Paper | October 20, 2017

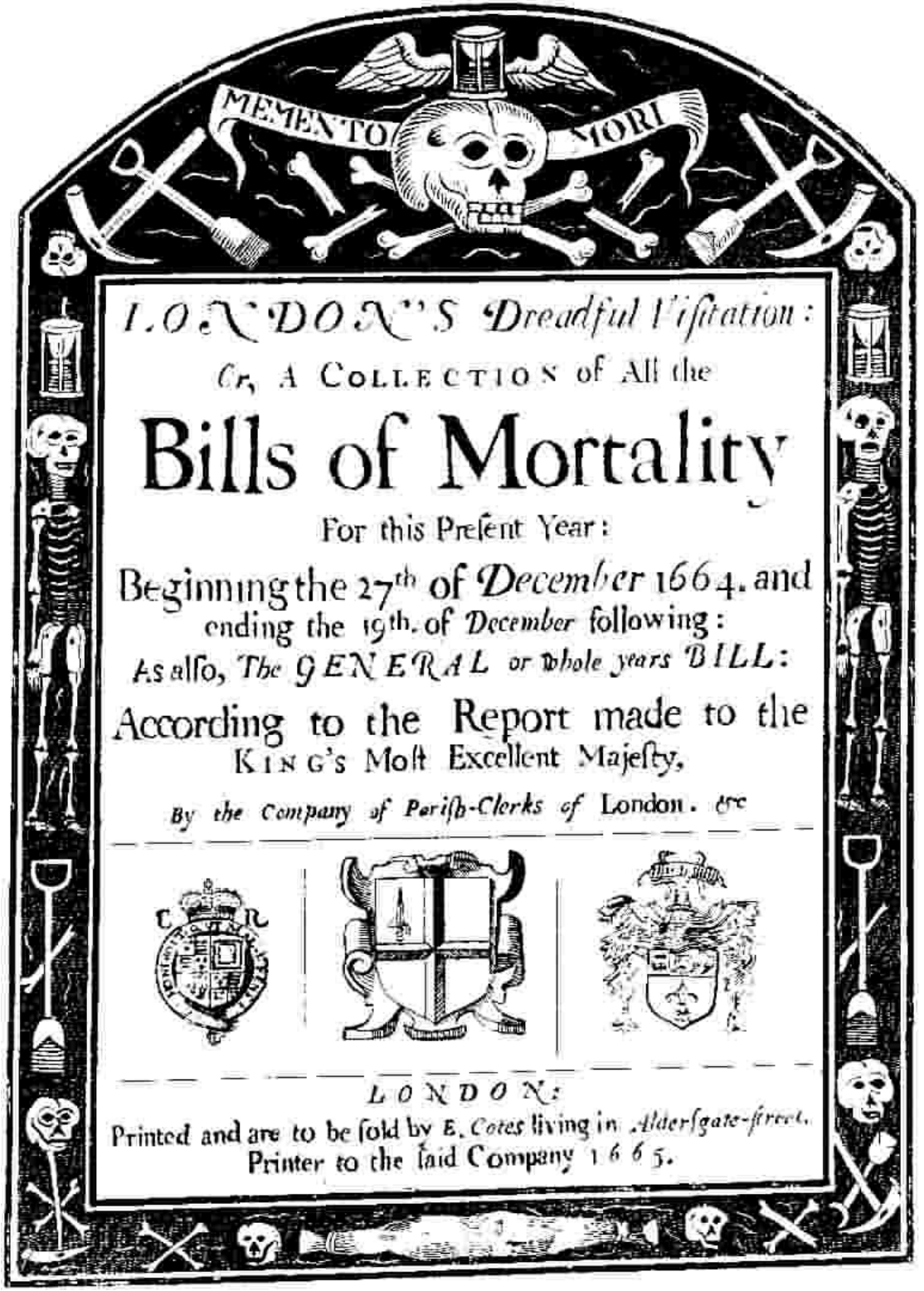
THE LAB @ DC

GOVERNMENT OF THE DISTRICT OF COLUMBIA
MURIEL BOWSER, MAYOR

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



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

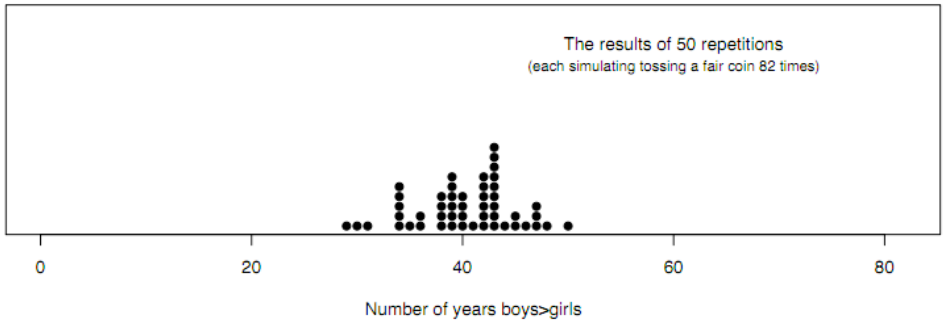
Arbuthnot noticed that in every of the 82 years from 1629 to 1710, there were more boys christened than girls; while this might seem like a compelling enough observation on its own, Arbuthnot takes it farther

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?

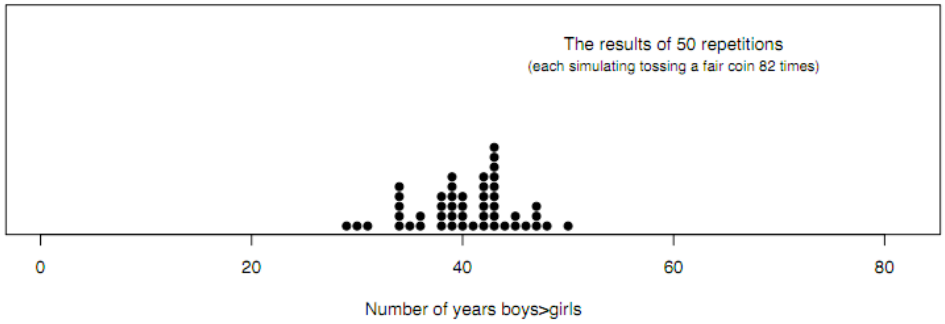
For that matter, what is the chance that we would see any large number, say 70 or 80 heads out of 82 tosses?

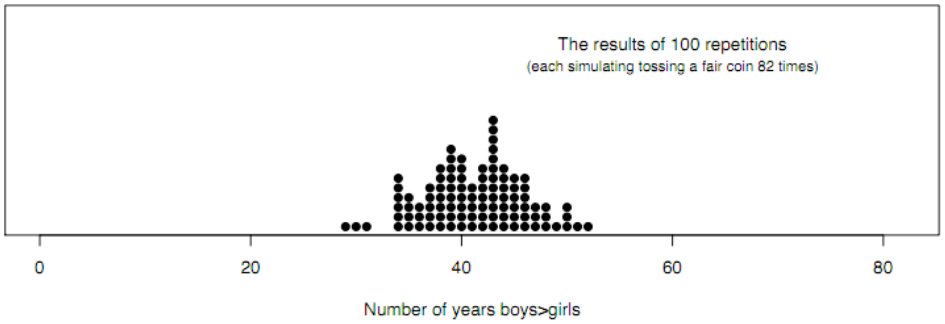

The results of 10 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


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

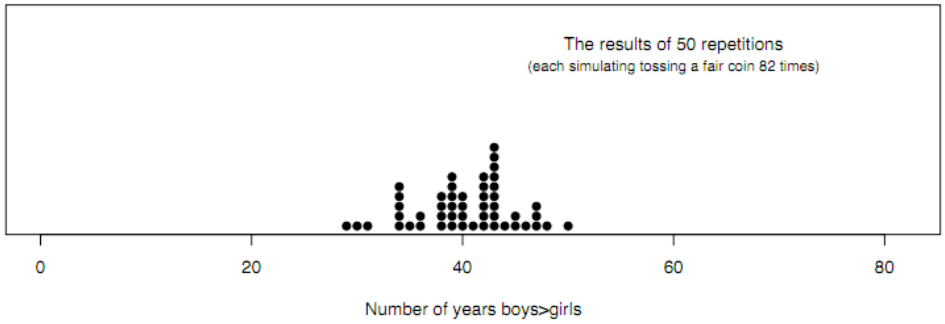

The results of 50 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

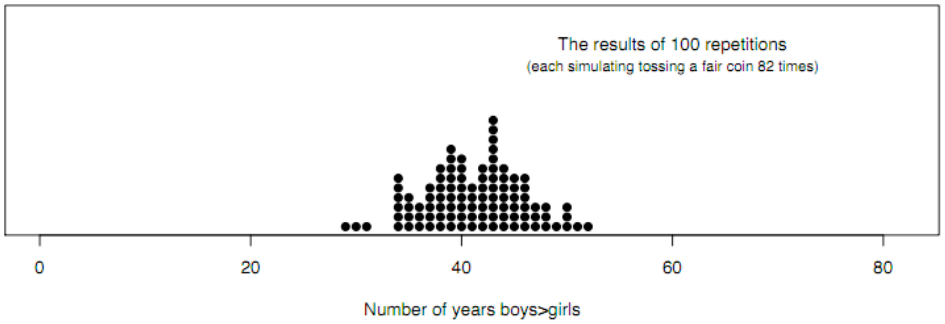

The results of 10 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

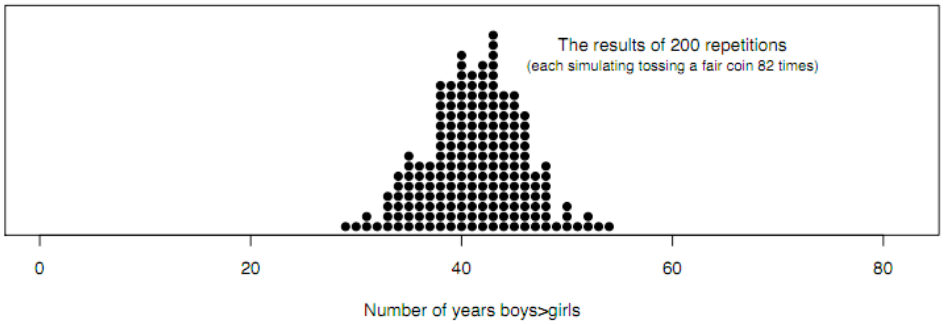

The results of 50 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls


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

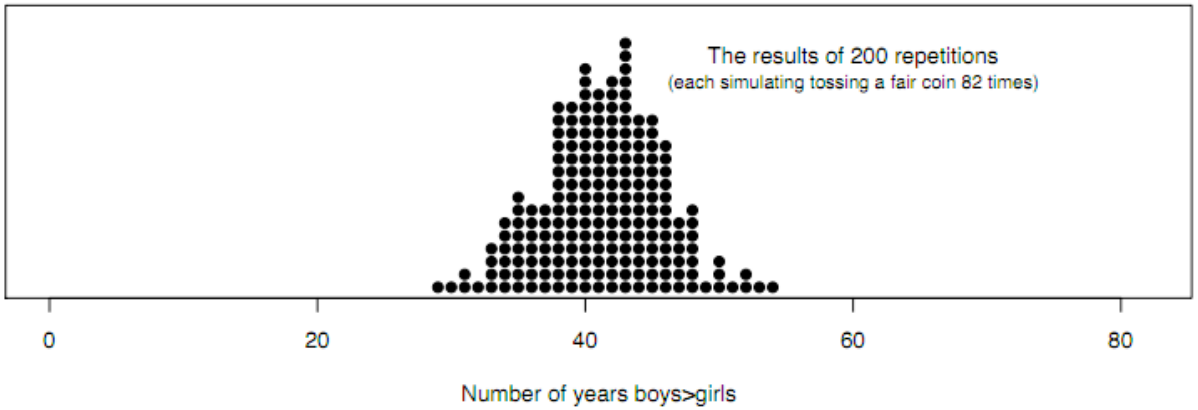

The results of 200 repetitions
(each simulating tossing a fair coin 82 times)
0
20
40
60
80
Number of years boys>girls

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?

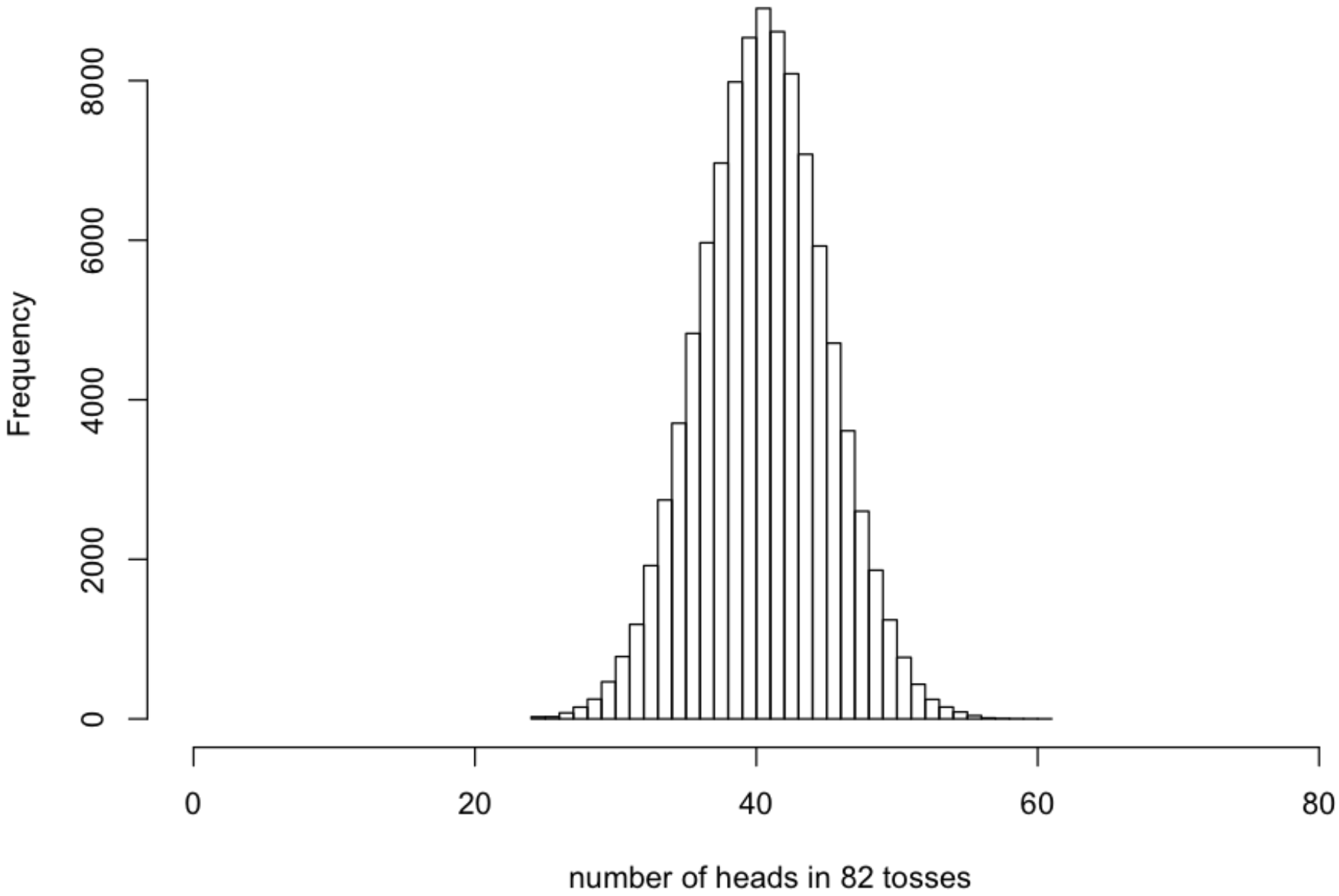


Assuming that boys and girls are born in the same proportion implies the simple ‘coin toss’ probability model; that is, the number of years that boys outnumber girls is the same as counting the number of heads in 82 independent tosses of a fair coin

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 $\frac{1}{4836\ 0000\ 0000\ 0000\ 0000\ 0000}$.

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

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

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

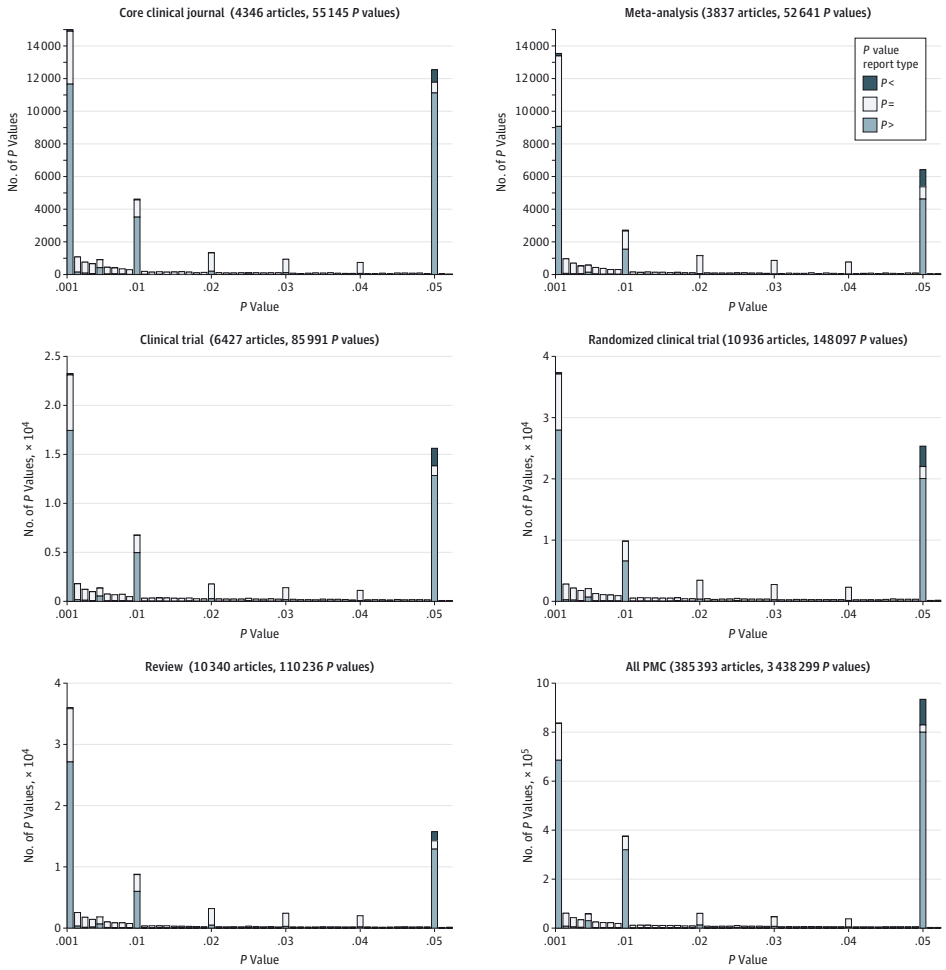
Supplemental content at jama.com

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 Munafò[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.

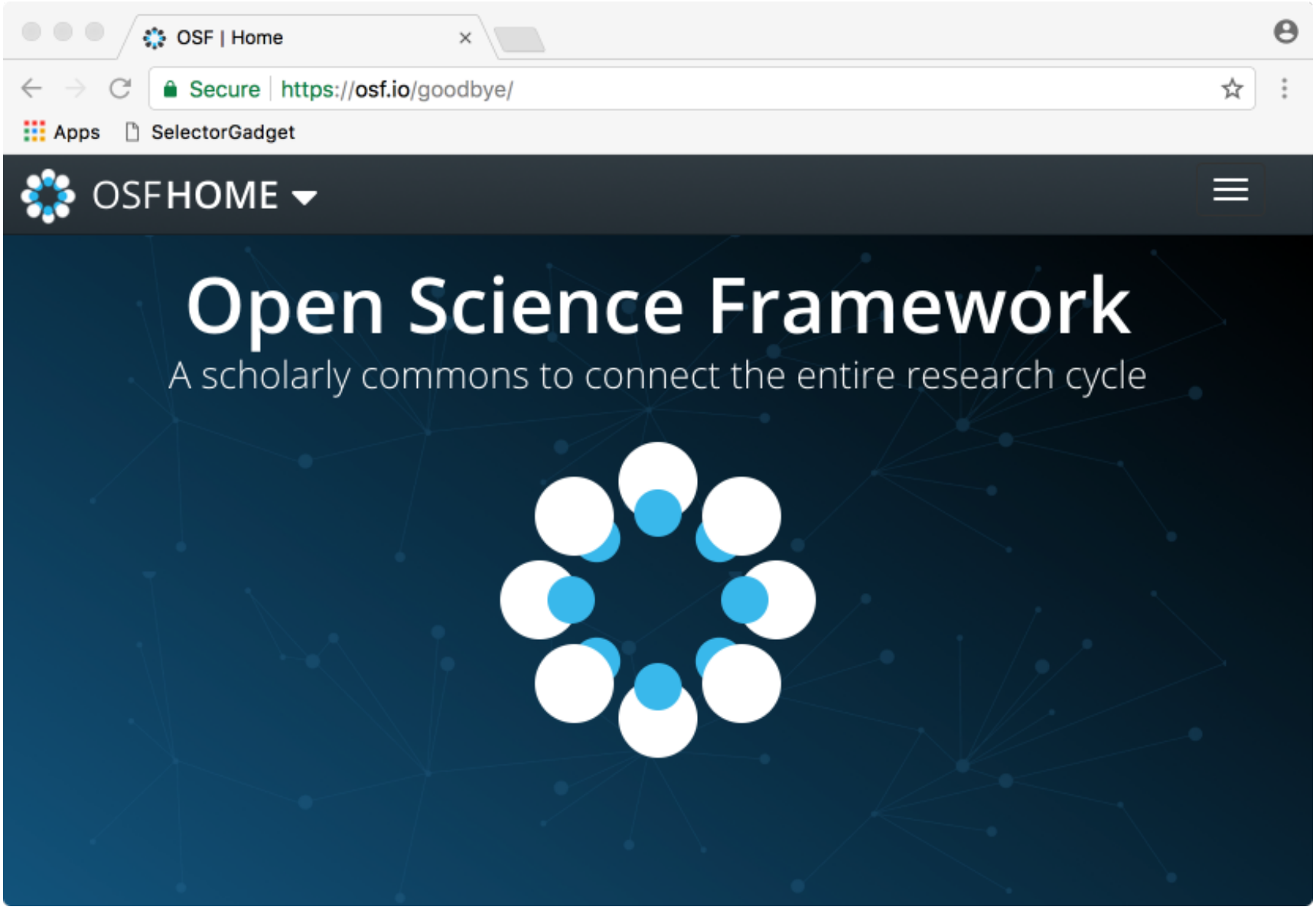
We restrict our recommendation to claims of discovery of new effects. We do not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent standards (e.g., 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.

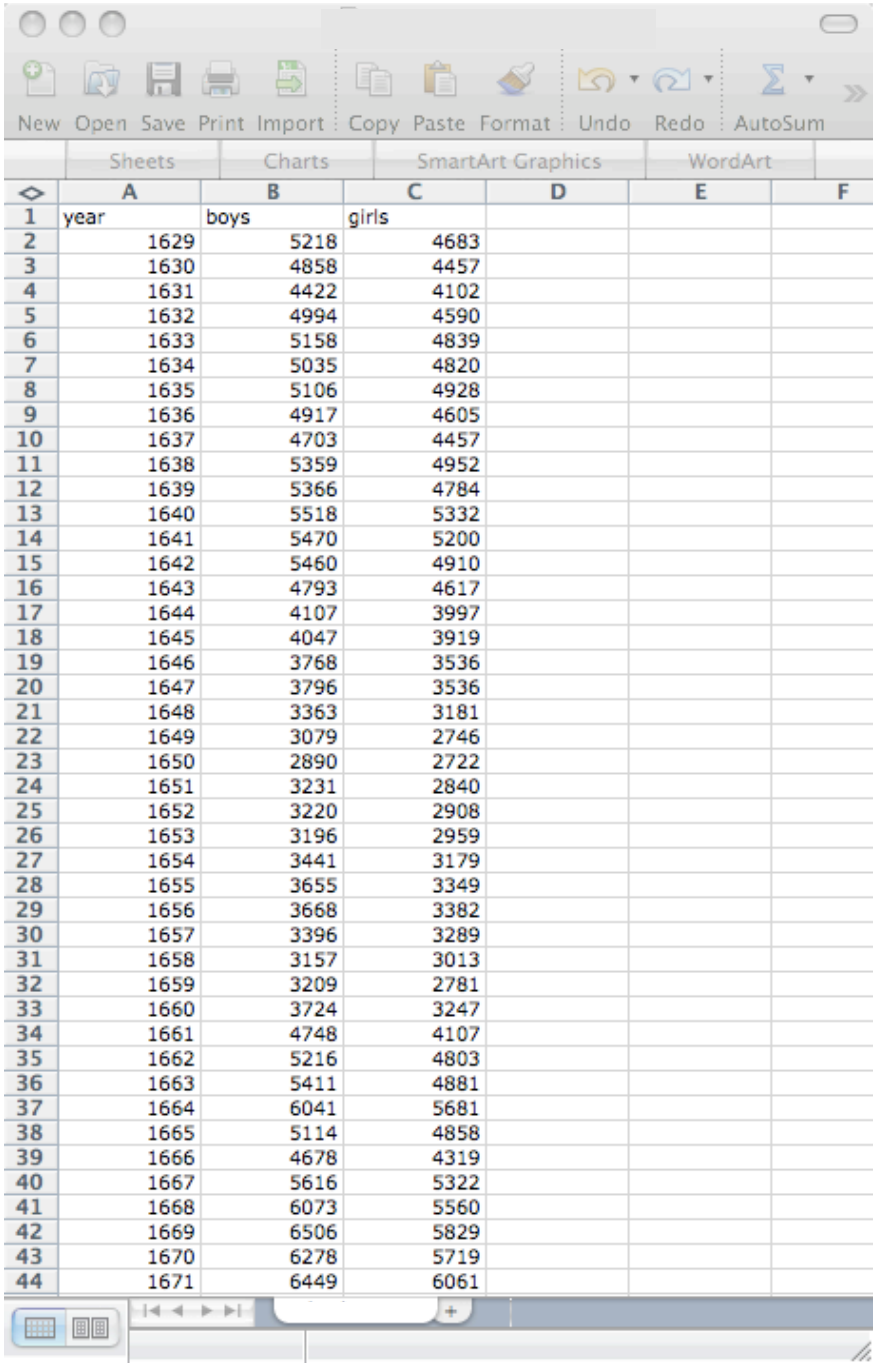
Journals can help transition to the new statistical significance threshold. Authors and readers can themselves take the initiative by describing and interpreting results more appropriately in light of the new proposed definition of “statistical significance.” The new significance threshold will help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


OSF | Home
Secure
https://osf.io/goodbye/
Apps
SelectorGadget
OSF HOME
Open Science Framework
A scholarly commons to connect the entire research cycle

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?

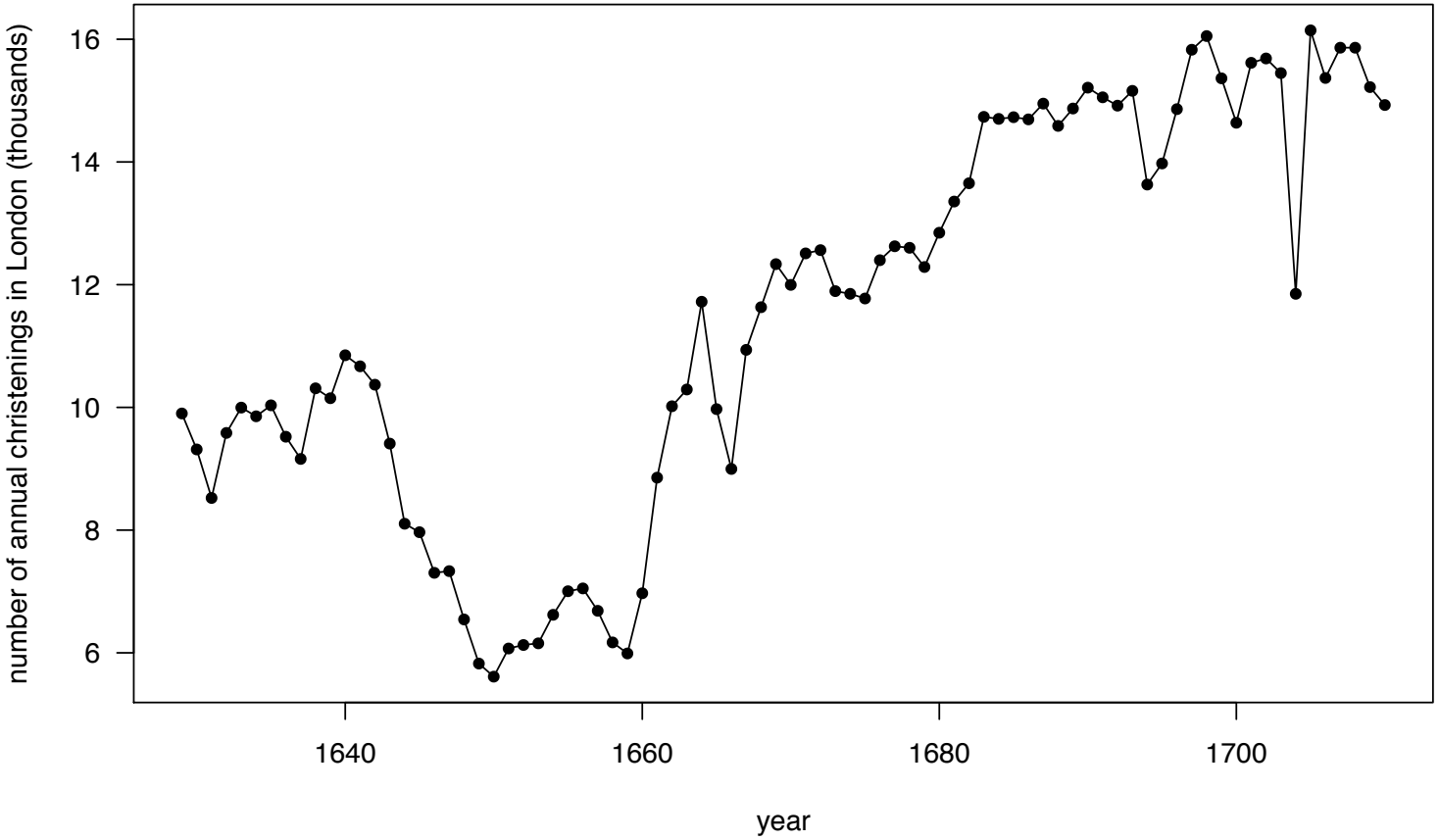


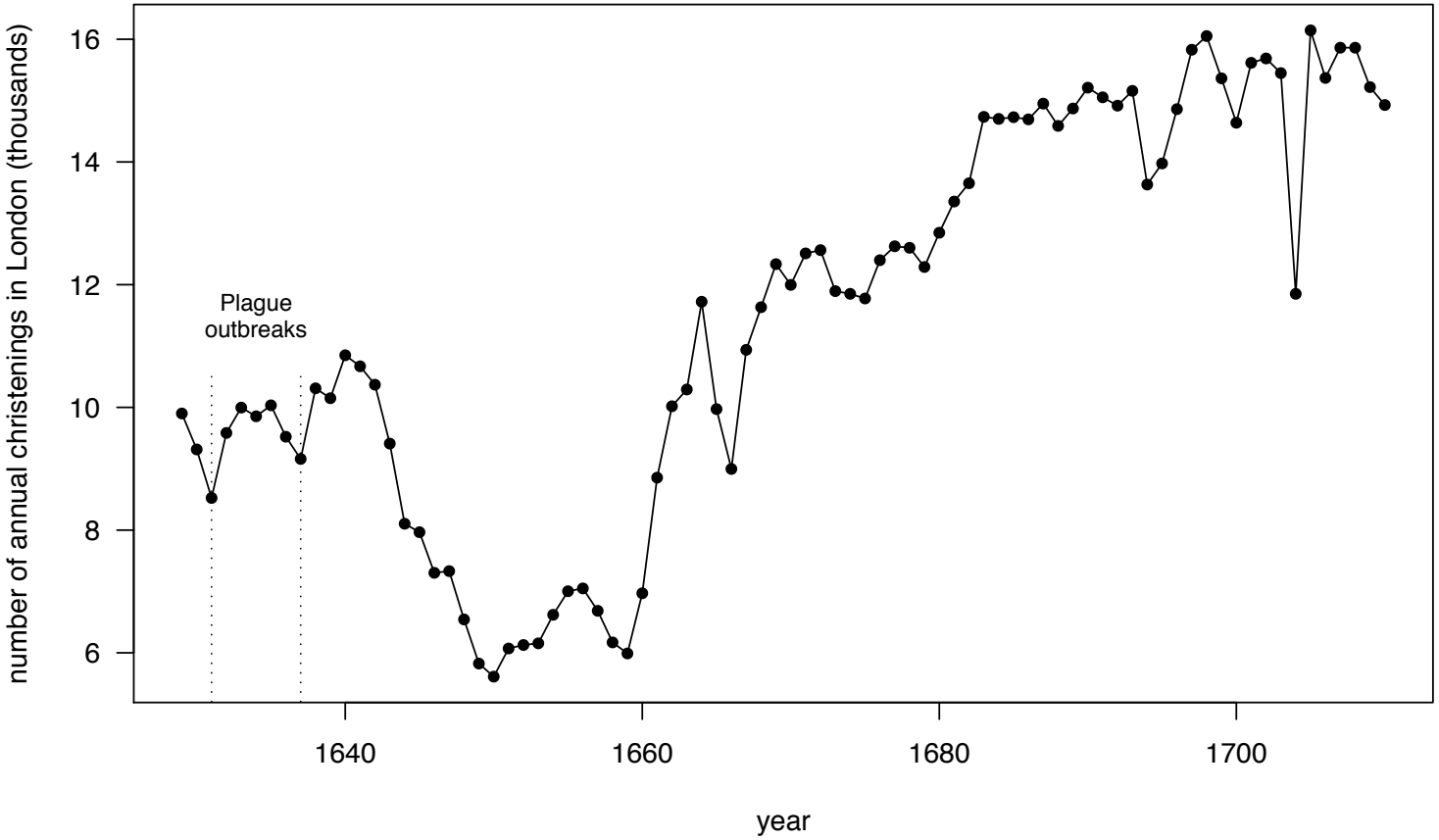
	A	B	C	D	E	F
1	year	boys	girls			
2	1629	5218	4683			
3	1630	4858	4457			
4	1631	4422	4102			
5	1632	4994	4590			
6	1633	5158	4839			
7	1634	5035	4820			
8	1635	5106	4928			
9	1636	4917	4605			
10	1637	4703	4457			
11	1638	5359	4952			
12	1639	5366	4784			
13	1640	5518	5332			
14	1641	5470	5200			
15	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			

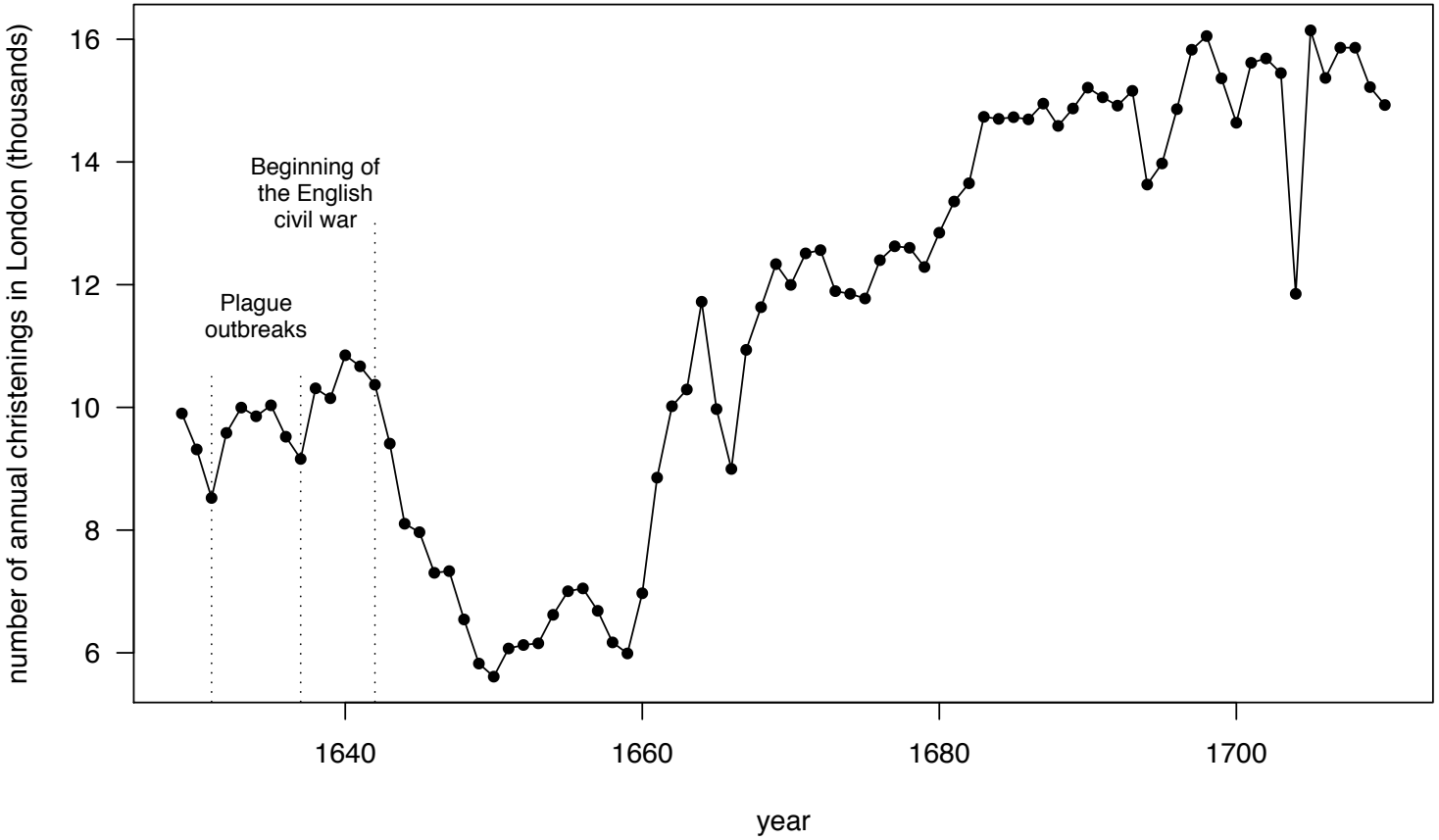

Christenings in London
(girls, solid; boys, dotted)
number of christenings (thousands)
3
4
5
6
7
8
1640
1660
1680
1700
year

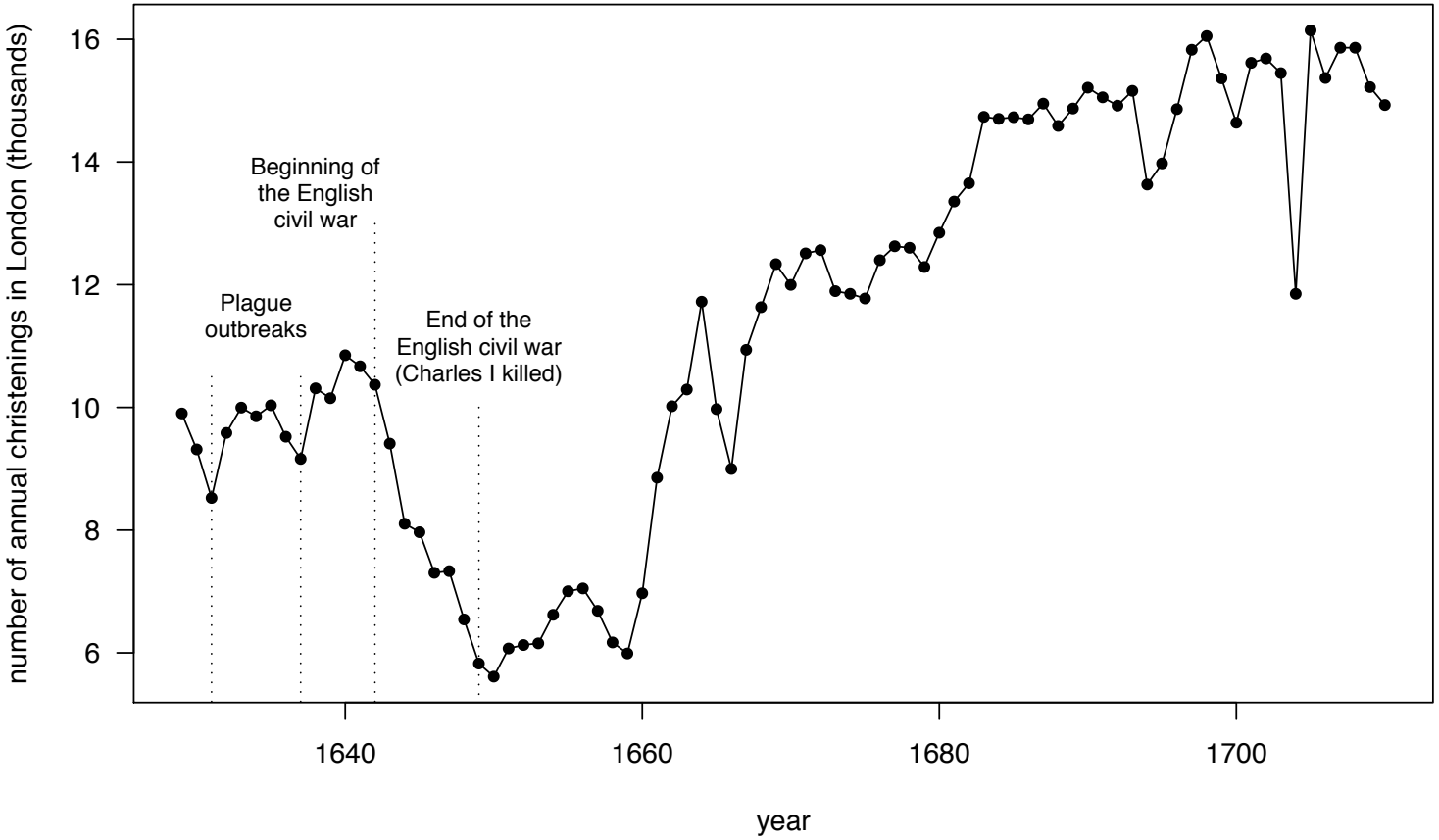
What does the “time series” plot of christenings, broken down by sex, show us? There is certainly a lot of structure to the graph, with periods of downturn in the total number of christenings, superimposed on an overall increase in the birthrate

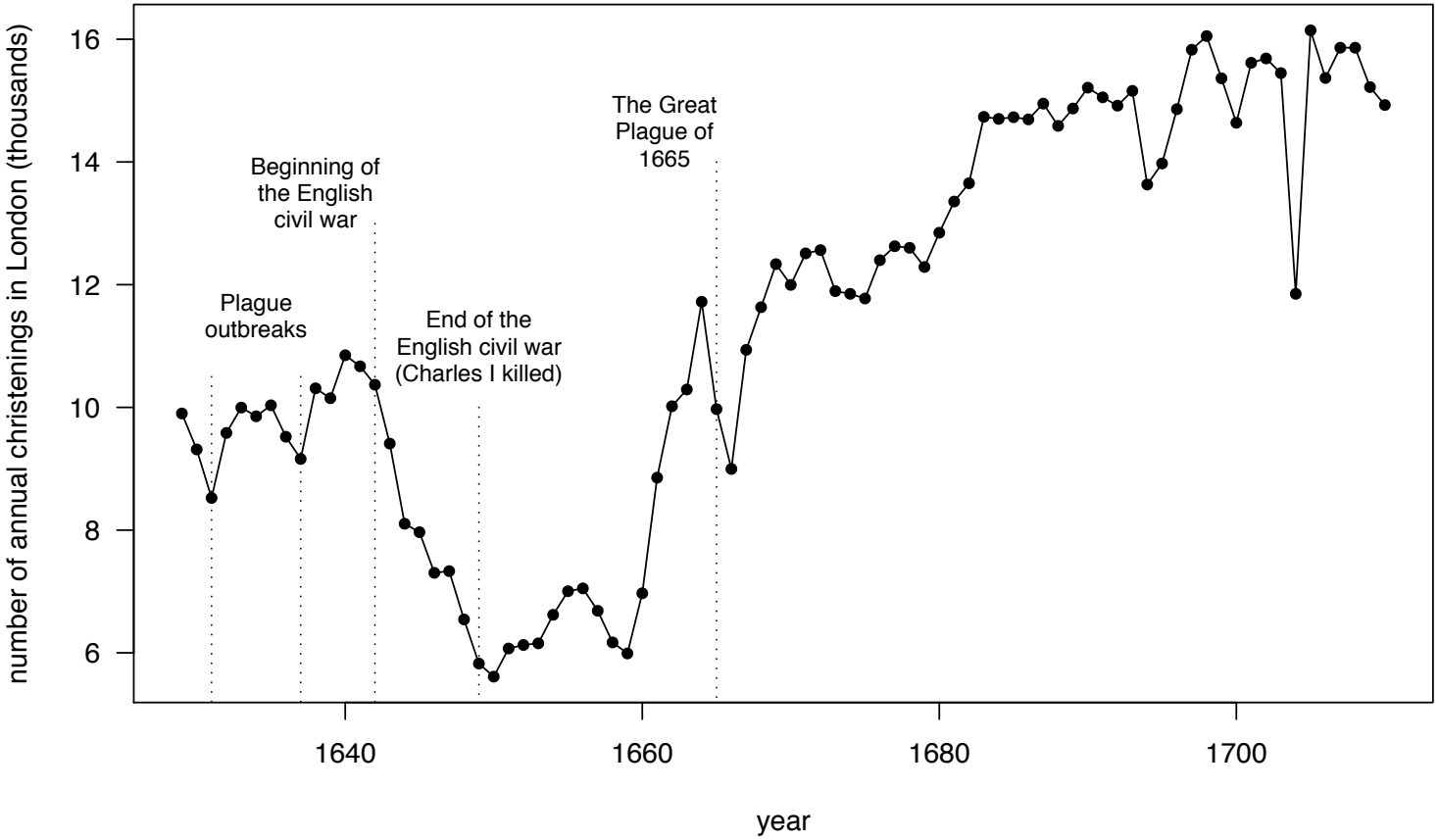
In a modern treatment of these data, Wainer (2005) started to identify peaks and valleys with specific historical events like wars and plagues; let’s see what that amounts to...

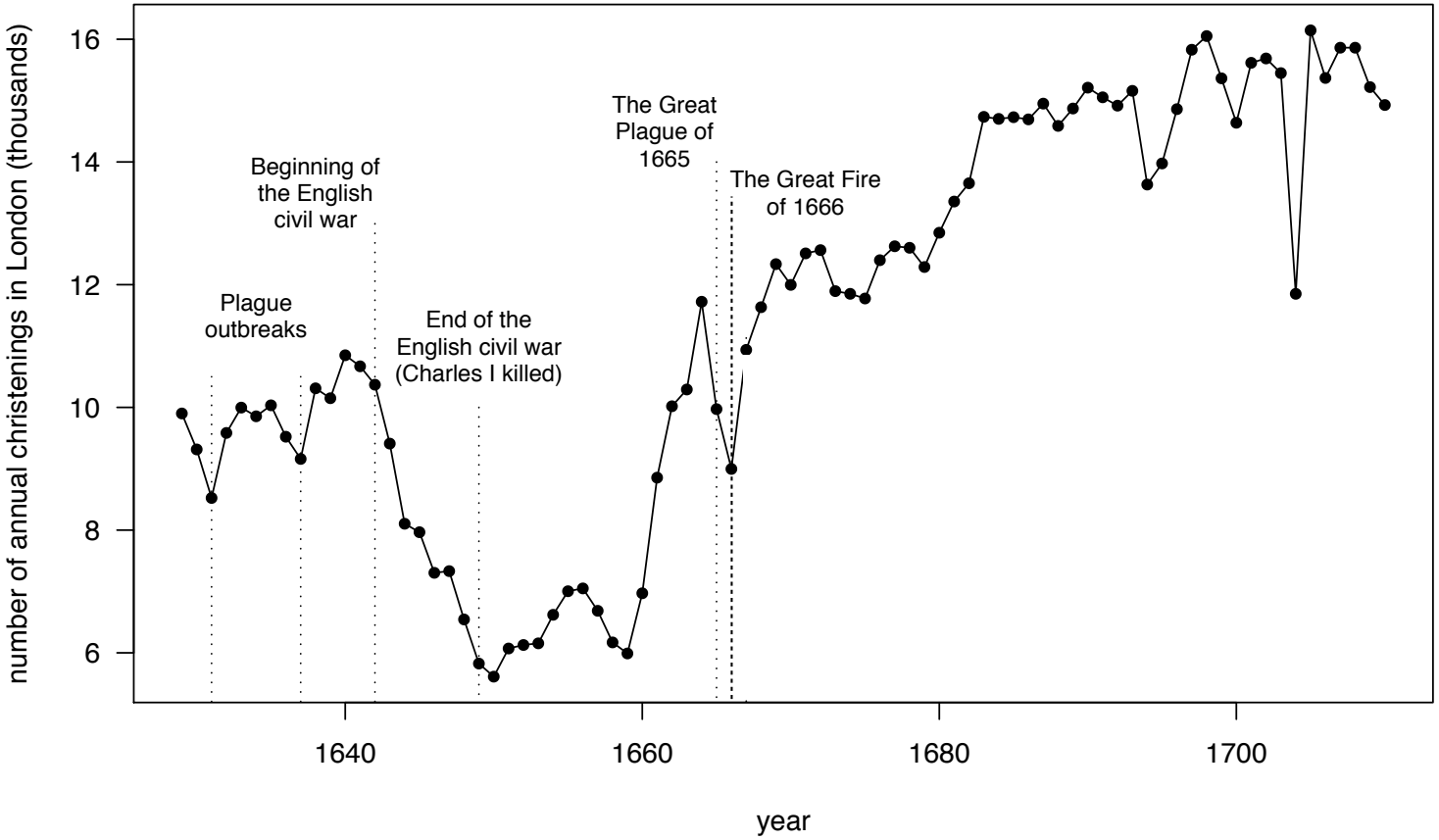

number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year

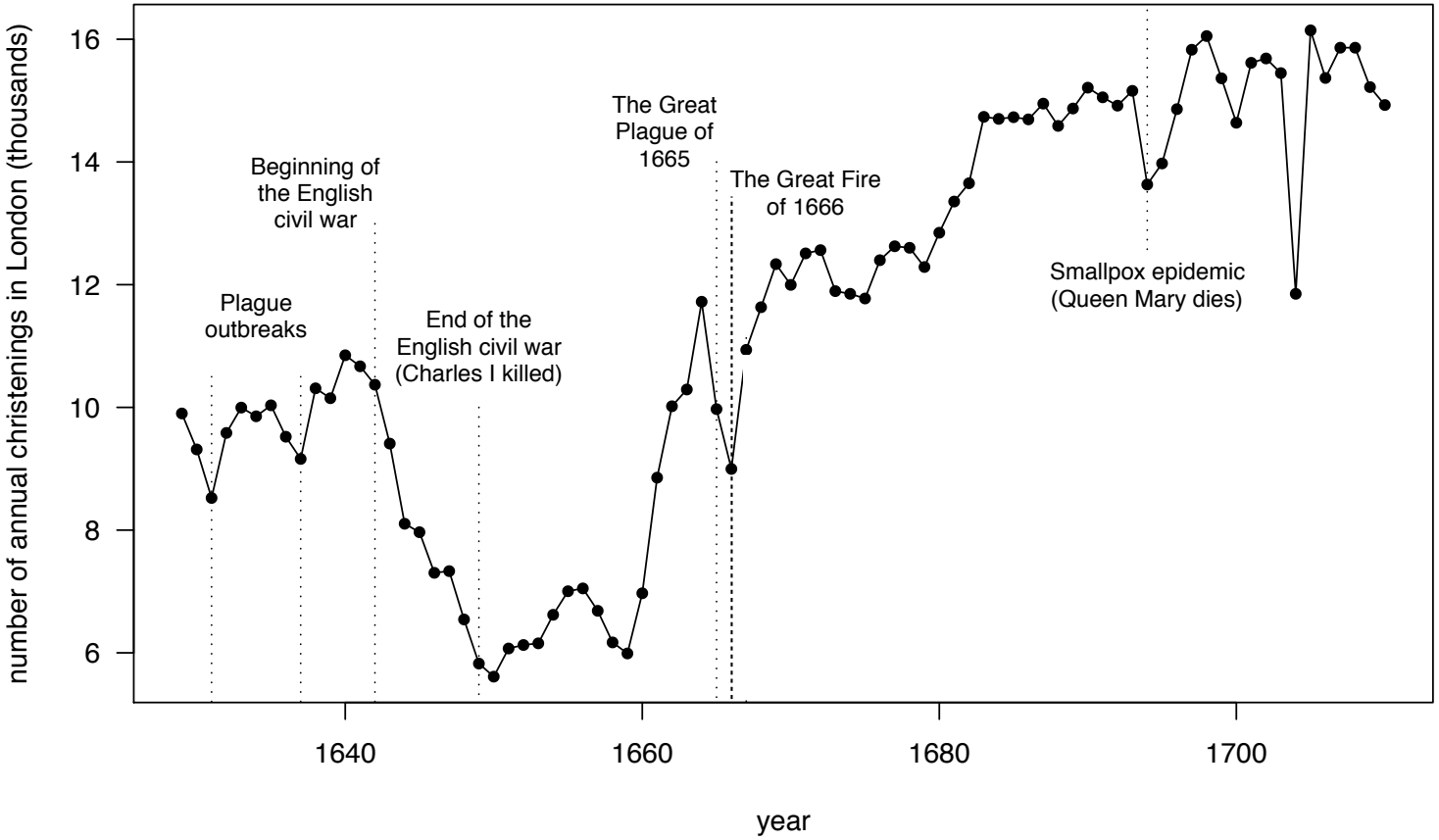

number of annual christenings in London (thousands)
16
14
12
10
8
6
Plague
outbreaks
1640
1660
1680
1700
year

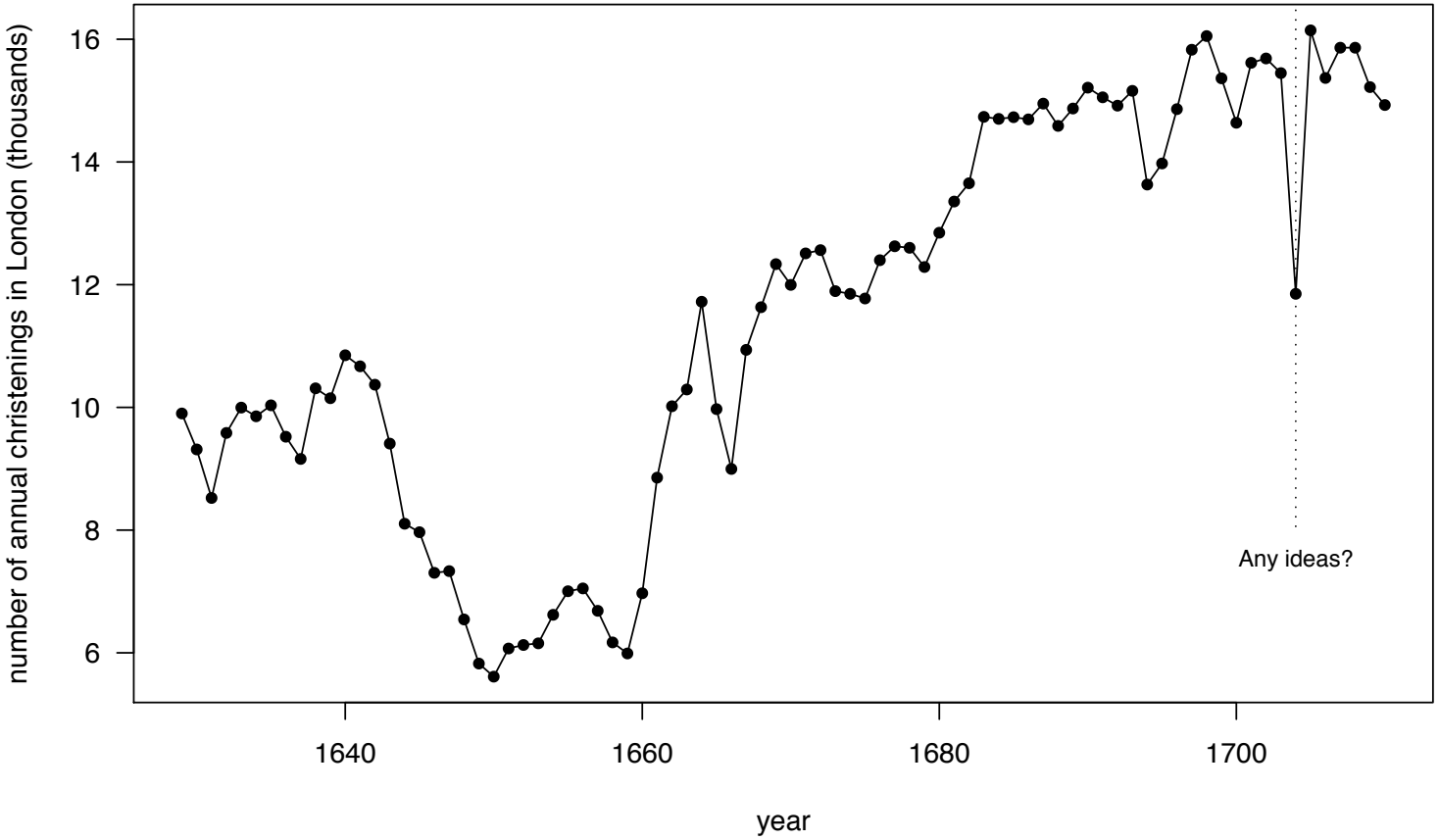

number of annual christenings in London (thousands)
16
14
12
10
8
6
Beginning of
the English
civil war
Plague
outbreaks
1640
1660
1680
1700
year


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
1640
1660
1680
1700
year
Plague outbreaks
Beginning of the English civil war
End of the English civil war (Charles I killed)
The Great Plague of 1665

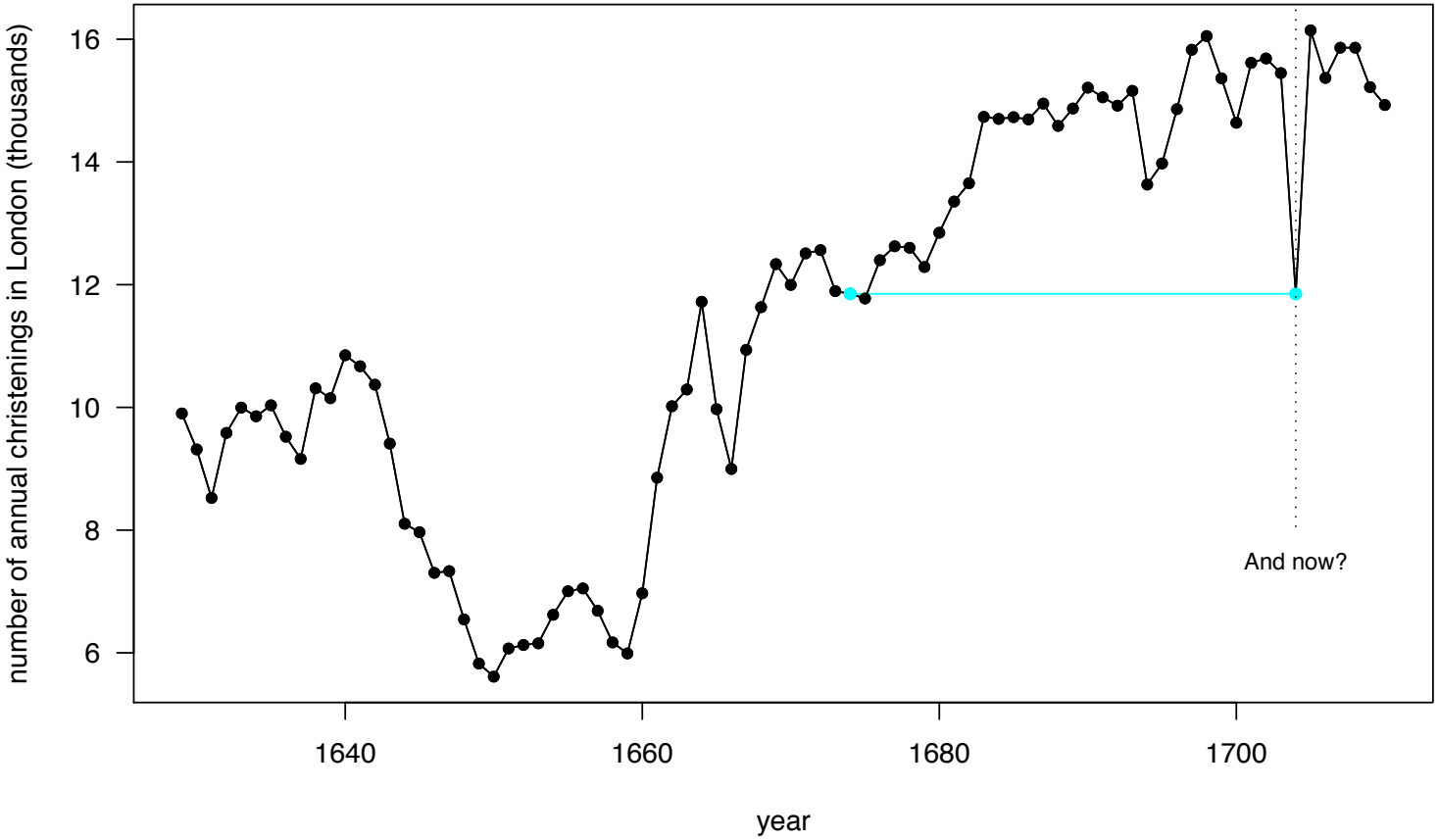

number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year
Plague
outbreaks
Beginning of
the English
civil war
End of the
English civil war
(Charles I killed)
The Great
Plague of
1665
The Great Fire
of 1666


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

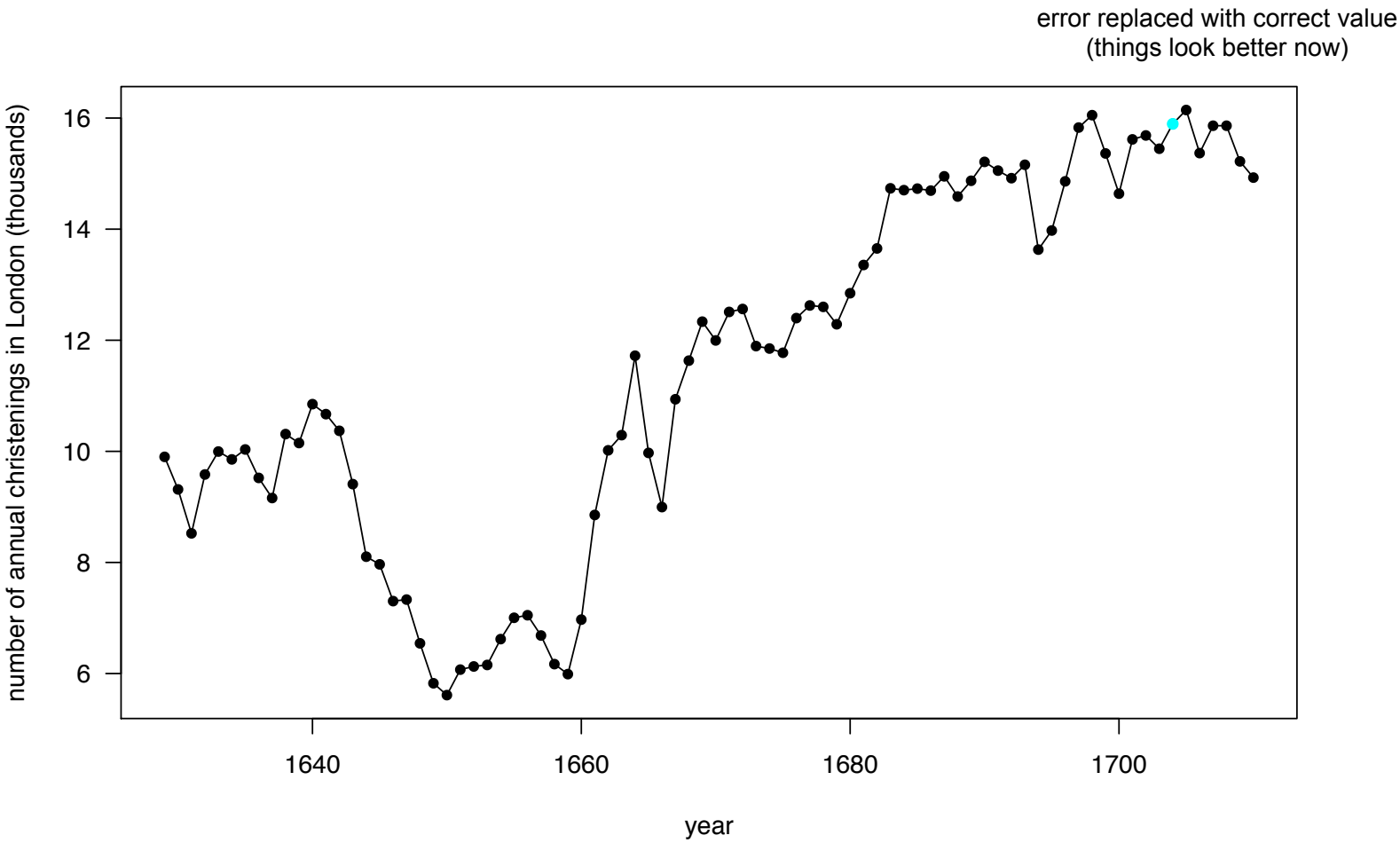

number of annual christenings in London (thousands)
16
14
12
10
8
6
1640
1660
1680
1700
year
And now?

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

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

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

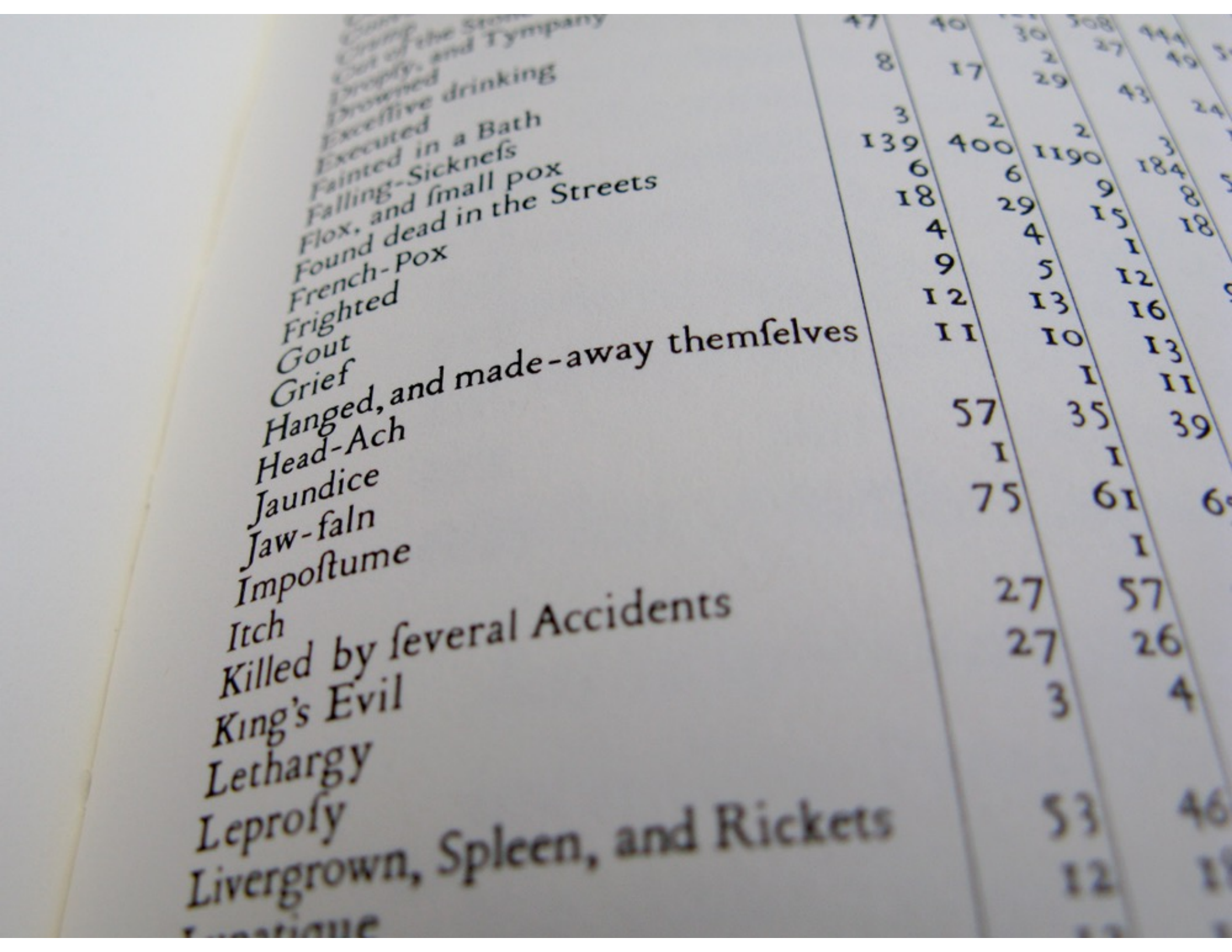
Decreased of the Plague in the 122 Parishes, and at the Pest-house this year 266

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


Dropſy, and Tympany
Drowned
Excellive drinking
Executed
Fainted in a Bath
Falling-Sickneſs
Flox, and ſmall pox
Found dead in the Streets
French-Pox
Frighted
Gout
Grief
Hanged, and made-away themſelves
Head-Ach
Jaundice
Jaw-faln
Impoſtume
Itch
Killed by ſeveral Accidents
King's Evil
Lethargy
Leproſy
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.

NBER WORKING PAPER SERIES

WHAT DO WORKPLACE WELLNESS PROGRAMS DO? EVIDENCE FROM THE ILLINOIS WORKPLACE WELLNESS STUDY

Damon Jones
David Molitor
Julian Reif

Working Paper 24229
http://www.nber.org/papers/w24229

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2018, Revised June 2018

This research was supported by the National Institute on Aging of the National Institutes of Health under award number R01AG050701; the National Science Foundation under Grant No. 1730546; the Abdul Latif Jameel Poverty Action Lab (J-PAL) North America U.S. Health Care Delivery Initiative; Evidence for Action (E4A), a program of the Robert Wood Johnson Foundation; and the W.E. Upjohn Institute for Employment Research. This study was pre-registered with the American Economics Association RCT Registry (AEARCTR-0001368). We are grateful to Andy de Barros for thoroughly replicating our analysis and to J-PAL for coordinating this replication effort. We thank our co-investigator Laura Payne for her vital contributions to the study, Lauren Geary for outstanding project management, Michele Guerra for excellent programmatic support, and Illinois Human Resources for invaluable institutional support. We are also thankful for comments from Kate Baicker, Jay Bhattacharya, Tatyana Deryugina, Joseph Doyle, Amy Finkelstein, Eliza Forsythe, Drew Hanks, Bob Kaestner, David Meltzer, Michael Richards, Richard Thaler, and seminar participants at AHEC, Harvard, Junior Health Economics Summit, MHEC, NBER Summer Institute, Ohio State University, University of Chicago AFE Conference, University of Zurich, UPenn Behavioral Economics and Health Symposium, SEA, and SIEPR. The findings and conclusions expressed are solely those of the authors and do not represent the views of the National Institutes of Health, any of our funders, the University of Illinois, or the National Bureau of Economic Research.

frequently asked questions, and contact information for participant support.

2.3 Data

Our analysis employs a combination of self-reported survey data and a number of administrative data sources, all merged together at the individual level. We briefly describe each data source below. Appendix Table A.7 provides a definition for each variable used in our analysis. Additional details are provided in Appendix D.2.

2.3.1 University Administrative Data

We obtained university administrative data on 12,486 employees who as of June 2016 were (1) working at the Urbana-Champaign campus of the University of Illinois and (2) eligible for part-time or full-time employee benefits from the Illinois Department of Central Management Services. We excluded 27 people who did not have a university email address or who were substantially involved with our study, yielding a final sample size of 12,459 employees.

The initial denominator file includes the employee's name, university identification number, contact information (email and home mailing address), date of birth, sex, race, salary, and employee class (faculty, academic staff, or civil service). We used the email and home mailing address to invite employees to participate in our study, and we used the sex, race, date of birth, salary, and employee class variables to generate the strata for random sampling.

A second file includes employment history information as of July 31, 2017. This provides two employee productivity outcomes that are measured over the first 12 months of our study: job termination and salary raises. All employees in our sample were eligible for a mid-year, merit-based salary increase that occurred in February 2017.

A third file provides data on sick leave. The number of sick days taken is available at the monthly level for Civil Service employees. For academic faculty and staff, the number of sick days taken is available biannually, on August 15 and May 15. We first calculate the total number of sick days taken during our pre-period (August 2015 - July 2016) and post-period

(August 2016 - July 2017) for each employee. We then normalize by the number of days employed to make this measure comparable across employees. All specifications that include sick days taken as an outcome variable are weighted by the number of days employed.

A fourth file contains data on exact attendance dates for the university's gym and recreational facilities. Entering one of these facilities requires swiping an ID card, which creates a database record linked to the individual's university ID. We calculate the total number of visits per year for the pre-period (August 2015 - July 2016) and the post-period (August 2016 - July 2017).

2.3.2 Online Survey Data

As described in Section 2.2, all study participants took a 15-minute online survey in July 2016 as a condition of enrollment in the study. The survey covered topics including health status, health care utilization, job satisfaction, and productivity.

Our survey software recorded that, out of the 12,459 employees invited to take the survey, 7,468 employees clicked on the link to the survey, 4,918 employees began the survey, and 4,834 employees completed the survey. Although participants were allowed to skip questions, response rates for the survey were very high: 4,822 out of 4,834 participants (99.7 percent) answered every one of the questions used in our analysis. To measure the reliability of the survey responses, we included a question about age at the end of the survey and compared participants' self-reported ages with the ages available in the university's administrative data. Of the 4,830 participants who reported an age, only 24 (<0.5 percent) reported a value that differed from the university's administrative records by more than one year.

All study participants were also invited via postcard and email to take a one-year, follow-up survey online in July 2017.[9] In addition to the questions asked on the baseline survey, the follow-up survey included additional questions on productivity, presenteeism, and job satisfaction. A total of 3,568 participants (74 percent) successfully completed the 2017

[9]Invitations to the follow-up survey were sent regardless of current employment status with the university.

follow-up survey. The completion rates for the control and treatment groups were 75.4 and 73.1 percent, respectively. This difference in completion rates is marginally significant ($p = 0.079$). The full texts of our 2016 baseline and 2017 follow-up online surveys are available on the study website and as part of our supplementary materials.[10]

2.3.3 Health Insurance Claims Data

We obtained health insurance claims data for the time period January 1, 2015, through July 31, 2017, for the 67 percent of employees who subscribe to the university's most popular insurance plan. We use the total payment due to the provider to calculate average total monthly spending. We also use the place of service code on the claim to break total spending into four major subcategories: pharmaceutical, office, hospital, and other.[11] Our spending measures include all payments from the insurer to providers, as well as any deductibles or copays paid by individuals. We merged these data at the individual level with our other datasets for those employees who consented to participate in our study. In addition, we have access to anonymized panel data on health claims for non-participating employees who subscribe to this same plan.

Employees choose their health plan annually during the month of May, and plan changes become effective July 1. Participants were informed of their treatment assignment on August 9, 2016. We therefore define baseline medical spending to include all allowed amounts with dates of service corresponding to the 13-month time period July 1, 2015, through July 31, 2016. We define spending in the post period to correspond to the 12-month time period August 1, 2016, through July 31, 2017.

In our health claims sample, 11 percent of employees are not continuously enrolled

[10] Interactive examples of the surveys administered for the study are available at http://www.nber.org/workplacewellness.

[11] Pharmaceutical and office-based spending each have their own place of service codes. Hospital spending is summed across the following four codes: "Off Campus - Outpatient Hospital," "Inpatient Hospital," "On Campus - Outpatient Hospital," and "Emergency Room - Hospital." All remaining codes are assigned to "other" spending, which serves as the omitted category in our analysis. We did not pre-specify subcategories of spending in our pre-analysis plan.

throughout the 13-month pre-period, and 9 percent are not continuously enrolled throughout the 12-month post-period. This is primarily due to job turnover. Because measures of average monthly spending are less noisy for employees with more months of claims data, we weight our regressions by the number of covered months whenever the outcome variable is average spending.

2.3.4 Illinois Marathon/10K/5K Data

The Illinois Marathon is a running event held annually in Champaign, Illinois. The individual races offered include a marathon, a half marathon, a 5K, and a 10K. When registering for a race, a participant must provide her name, age, sex, and hometown. That information, along with the results of the race, are published online after the races have concluded. We downloaded those data for the 2014-2017 races and matched it to individuals in our dataset using name, age, sex, and hometown.

2.4 Baseline Summary Statistics and Balance Tests

Tables 1a and 1b provide summary statistics at baseline for the employees in our sample. Columns (2)-(8) report means for those who were assigned to our control group and to each of our six treatment groups. Column (1) additionally reports summary means for employees not enrolled in our study, where available. The variables are grouped into four panels, based on the source and type of data. Panel A presents means of the university administrative data variables used in our stratified randomization, Panel B presents means of variables from our 2016 baseline survey, Panel C presents means of medical spending variables from our health insurance claims data for the July 2015 - July 2016 time period, and Panel D presents baseline means of administrative data variables used to measure health behaviors and employee productivity.

Our experimental framework relies on the random assignment of study participants to the treatment and control groups. To evaluate the validity of this assumption, we first compare

Limitless Worker Surveillance

Ifeoma Ajunwa,* Kate Crawford,** and Jason Schultz***

From the Pinkerton private detectives of the 1850s, to the closed-circuit cameras and email monitoring of the 1990s, to new apps that quantify the productivity of workers, and to the collection of health data as part of workplace wellness programs, American employers have increasingly sought to track the activities of their employees. Starting with Taylorism and Fordism, American workers have become accustomed to heightened levels of monitoring that have only been mitigated by the legal counterweight of organized unions and labor laws. Thus, along with economic and technological limits, the law has always been presumed as a constraint on these surveillance activities. Recently, technological advancements in several fields—big data analytics, communications capture, mobile device design, DNA testing, and biometrics—have dramatically expanded capacities for worker surveillance both on and off the job. While the cost of many forms of surveillance has dropped significantly, new technologies make the surveillance of workers even more convenient and accessible, and labor unions have become much less powerful in advocating for workers. The American worker must now contend with an all-seeing Argus Panoptes built from technology that allows for the trawling of employee data from the Internet and the employer collection of productivity data and health data, with the ostensible consent of the worker. This raises the question of whether the law still remains a meaningful avenue to delineate boundaries for worker surveillance.

DOI: https://dx.doi.org/10.15779/Z38BR8MF94

* Fellow, Berkman Klein Center at Harvard University; Assistant Professor, Cornell Industrial and Labor Relations (ILR) School; Associate Faculty, Cornell Law School.

** Visiting Professor, MIT Center for Civic Media; Principal Researcher, Microsoft Research; Senior Fellow, NYU Information Law Institute.

*** Professor of Clinical Law, NYU School of Law. First, the authors wish to thank the editors of the *California Law Review* for their capable and fastidious editing assistance. The authors also wish to thank the attendees of the 2016 Privacy Law Scholars Conference at George Washington University and the 2016 Law and Society Association Conference in New Orleans. Special thanks to Professors Andrew G. Ferguson, Pauline Kim, and Brett Frischmann. We also thank Microsoft Research New York for funding Professor Ajunwa's initial research on these topics.

In this Article, we start from the normative viewpoint that the right to privacy is not an economic good that may be exchanged for the opportunity for employment. We then examine the effectiveness of the law as a check on intrusive worker surveillance, given recent technological innovations. In particular, we focus on two popular trends in worker tracking—productivity apps and worker wellness programs—to argue that current legal constraints are insufficient and may leave American workers at the mercy of 24/7 employer monitoring. We consider three possible approaches to remedying this deficiency of the law: (1) a comprehensive omnibus federal information privacy law, similar to approaches taken in the European Union, which would protect all individual privacy to various degrees regardless of whether or not one is at work or elsewhere and without regard to the sensitivity of the data at issue; (2) a narrower, sector-specific Employee Privacy Protection Act (EPPA), which would focus on prohibiting specific workplace surveillance practices that extend outside of work-related locations or activities; and (3) an even narrower sector and sensitivity-specific Employee Health Information Privacy Act (EHIPA), which would protect the most sensitive type of employee data, especially those that could arguably fall outside of the Health Insurance Portability and Accountability Act's (HIPAA) jurisdiction, such as wellness and other data related to health and one's personhood.

Probability

We'll now move from these data to probability -- It turns out that modern probability is born around the time of Graunt and his work

In fact, his work will provide inspiration to some of the early probabilists, struggling for a way to connect mathematics to the workings of the real world

Calculation versus interpretation

In what follows, we are going to try to keep distinct the mathematical rules for calculating with probabilities from their interpretation -- Probability is **a branch of mathematics with it's own rules** and most definitions of probability adhere to these rules

Quite aside from computation, however, we have the application of probability, **the interpretation of probability in our daily lives** -- What does it mean to say that event will occur with probability 1/3?

The mathematical framework lets us solve textbook problems (rolling dice or pulling cards from a well-shuffled deck), but the interpretation, what we mean by the term probability can be a different animal entirely

If statistics uses the language of probability to describe events in our lives, then **our interpretation of probability can influence how we reason about the world**

The emergence of probability

Ian Hacking, a historian and philosopher, believes that probability was "born" in 1660 -- That's precisely the time of **John Graunt and his work on the London bills**, work that will feed into the emergence of probability

Hacking notes that since that time, probability has had two faces: In one it is an **explanation for "stable" frequencies seen in the world**, while in another it is **a relation between a hypothesis and the evidence for it**

He notes that prior to 1660, a statement was said **to be probable if it could be attested to by an authority** and a kind of expert testimony was involved -- Over time that changed and slowly **Nature became a kind of expert**

He writes *"It is here that we find* **the old notion of probability as testimony conjoined with that of frequency***. It is here that stable and law-like regularities become both observable and worthy of observation.* ***They are part of the technique of reading the true world****."*

The emergence of probability

He continues: "*A proposition was now probable, as we should say, if there was evidence for it, but in those days it was probable because it was testified to by the best authority.* ***Thus: to call something probable was still to invite the recitation of authority. But: since the authority was founded on natural signs, it was usually of a sort that was only "often to be trusted"****. Probability was communicated by what we should now call law-like regularities and frequencies.* ***Thus the connection of probability, namely testimony, with stable law-like frequencies is a result of the way in which the new concept of internal evidence came into being****."*

And so probability emerges with two faces -- **One related to hypothesis and evidence, and another related to stable frequencies seen in data about the world** (think about Graunt's birth and death records)

We'll see that these two views of probability, these two interpretations of the word, still exist today in modern statistical practice (Hacking says that **we've been swinging on a pendulum for centuries**) -- But first, let's go to the birth of probability, circa 1660..


Blaise Pascal

Classical probability

It is often said that the era of mathematical probability (or, rather, the view of probability as a branch of mathematics) **started with an exchange of letters** between Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1657) that took place in 1654

Their correspondence began with a question posed by Chevalier de Méré, “a gambler and a philosopher” known as the **“The problem of Points,”** one of a large class of so-called “division problems”

Their calculations applied **combinatorics** to questions about repeated gaming, and provided a framework for the so-called **classical approach to interpreting probability**

Classical probability

Here is de Méré's question:

> *Suppose two people, A and B, agree to play a series of fair games (think of tossing a coin) until one person has won a fixed number of games (say 6). They each have wagered the same amount of money, the intention being that the winner will be awarded the entire pot. But, suppose, for whatever reason, the series is prematurely terminated at which point A needs **a** more games to win, and B needs **b**. How should the stakes be divided?*

Seeing this problem for the first time, it's difficult approach, and to be fair, questions like it had been discussed for over a 100 years without a mathematical solution

If a=b, then it seems clear that the players should just divide the pot; but what if a=2 and b=3?

http://www.york.ac.uk/depts/maths/histstat/pascal.pdf

Classical probability

The solution hit upon by Pascal and Fermat is less about the history of the game as it was played before being interrupted, but instead considered **all the possible ways the game might have continued** if it had not been interrupted

Pascal and Fermat reasoned that the game would be over in a+b-1 further plays (possibly sooner); and that there were a total of 2^{a+b-1} possible outcomes (why?)

To figure A's share, you should count how many of these outcomes see A winning and divide by the total -- This fraction is **the probability that A would have eventually won the game**

Classical probability

For example, suppose the game was meant to be played until either A or B had won 6 times; but suppose play is interrupted with A having won 4 games and B three (or a = 6 - 4 = 2 and b = 6 - 3 = 3)

Play could have continued for at most $2^{2+3-1} = 2^4 = 16$ more games, and all the possible outcomes are

AAAA AAAB AABA AABB ABAA ABAB ABBA ABBB
BAAA BAAB BABA BABB **BBAA** BBAB BBBA BBBB

Of these, 11 out of 16 favor A (bolded), so A should take 11/16 of the total and B should take 5/16

Classical probability

The answer was a significant conceptual advance; in addressing this problem, Pascal and Fermat provided **a recipe for calculating probabilities**, one that involves combinatorics (counting outcomes)

In their letters, they address other games of chance with a similar kind of approach, each time taking the **probability of an event as the number of possible outcomes that make up that event** (the number of outcomes that have B winning, for example) **divided by the total number of outcomes**

In forming his solution, Pascal makes use of his famous **triangle of binomial coefficients**, a fact we'll come back to shortly..

Classical probability

Classical probability (so-named because of it's "early and august pedigree") assigns probabilities equally to the possible outcomes that could occur in a given problem, so that **the classical probability of an event is simply the fraction of the total number of outcomes in which the event occurs**

To add yet another heavy-hitter to our list of distinguished mathematicians, Pierre-Simon Laplace (1749-1827) clearly describes the idea as follows (a statement which was later termed the **Principle of Indifference**)

> ***The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible****, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply* ***a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.*** (1814, 1951 6-7)

This approach is well-adapted to games of chance (throwing dice, pulling cards from a well-shuffled deck) and is the basis for most of the probability problems in your textbook -- In these cases, **symmetry or the open physical process of drawing from a hat make the basic equally likely outcomes intuitively clear**

Some axioms

With the classical view of probability, you can establish the basic mathematical framework or calculus for probability; let $\mathcal{X}$ be the set of **all possible outcomes** of some experiment or trial or situation we'd like to study, let A denote **an "event" or collection of outcomes** from $\mathcal{X}$; and finally let P(A) be the probability of A

1. The probability of A is a number between 0 and 1, $0 \leq \mathrm{P}(\mathrm{A}) \leq 1$
2. The probability that an outcome will occur is 1, $\mathrm{P}(\mathcal{X}) = 1$
3. If A and B have no outcomes in common (they're disjoint), then their probabilities add $\mathrm{P}(\mathrm{A} \text{ or } \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$

Classical probability

While the classical model has all the ingredients for computing with probabilities, that is, it provides us with the basic calculus for probability; **as an interpretation of the concept of probability, the classical approach leaves a bit to be desired**

Ian Hacking puts it this way:

> *"The problems of real life are less tractable. We have stable mortality statistics, but who can ever tell the numbers of diseases? Who can enumerate the parts of the body that are attacked by disease? ... We have statistical regularities but no [fundamental set of outcomes]"*

The framework has limited scope -- It's not always appropriate to assign equal probabilities in more complex settings, and there are cases where the various outcomes are not obviously some known finite number

In addition, **many have criticized the underlying reasoning as circular** -- That is, saying events are "equipossible" already assumes equal probability

The first limit theorem

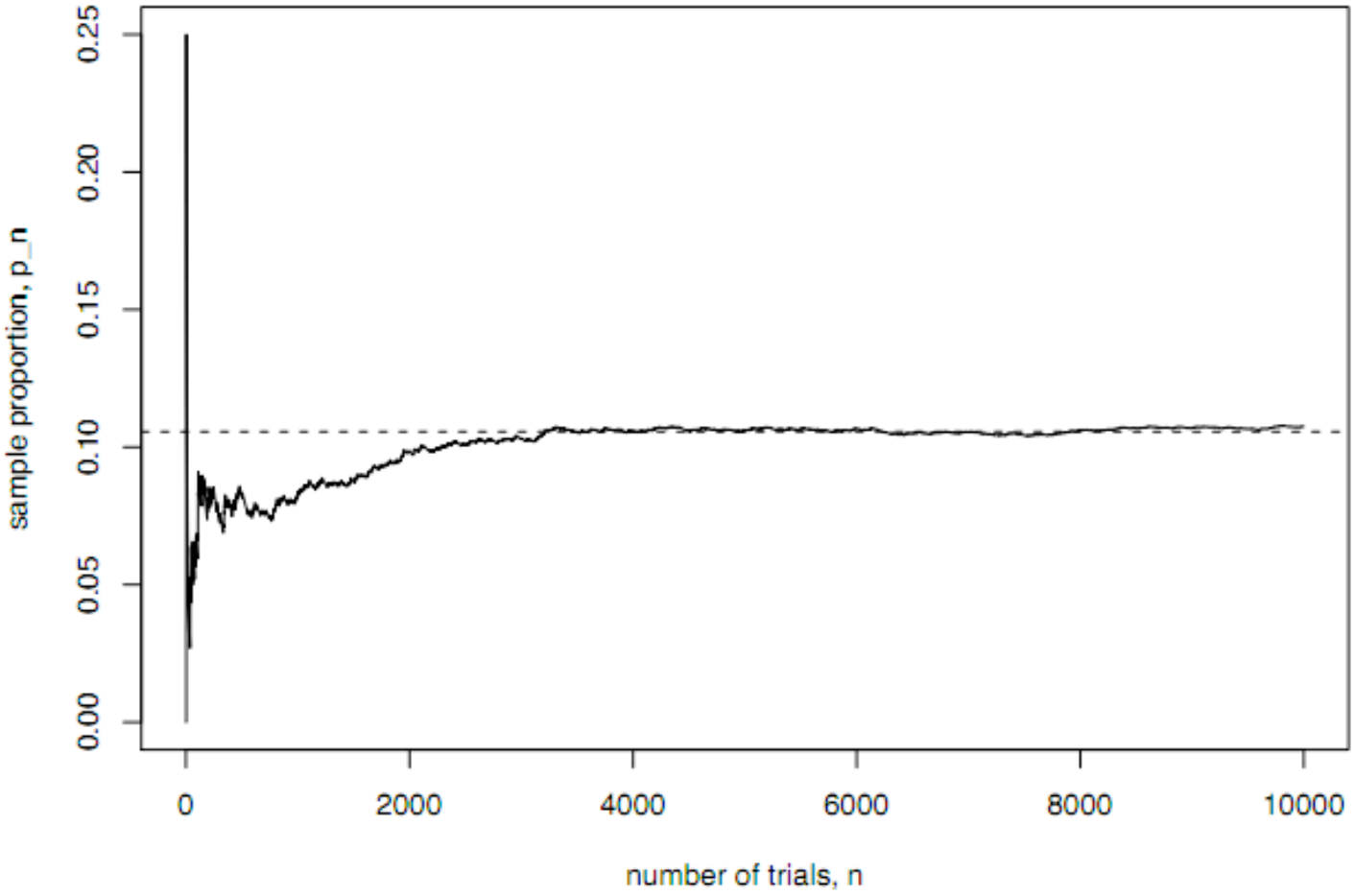
After the exchange between Pascal and Fermat was published, probability as a mathematical discipline really took off; starting in 1684, for example, Jakob Bernoulli (1654-1705) worked on a text entitled "*Ars Conjectandi*" (or, the *Art of Conjecture*)


JACOBI BERNOULLI,
Profeſſ. Baſil. & utriuſque Societ. Reg. Scientiar.
Gall. & Pruſſ. Sodal.
MATHEMATICI CELEBERRIMI,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et EPISTOLA Gallicè ſcripta
DE LUDO PILÆ
RETICULARIS.
BASILEÆ,
Impenſis THURNISIORUM, Fratrum.
cIↃ IↃcc xIII.

For example, working with the axioms of probability, Bernoulli derived the first "limit theorem," a mathematical result often called **the law of averages or the (weak) law of large numbers**

The result is related to repeated trials; namely, if on each trial you have the same probability of success p, then the proportion of successes in n trials p_n "tends to" p

Keep in mind that this is a mathematical result, and idealization, a model; later when we talk about the binomial distribution we'll see that this is not so hard to prove


sample proportion, p_n
0.00
0.05
0.10
0.15
0.20
0.25
0
2000
4000
6000
8000
10000
number of trials, n

The first limit law

The previous slides dealt mainly with mathematical abstraction, the calculus of probability, and its counterpart in simulation software; we now turn to **what all this means in terms of the interpretation of probability**

While much can be said about what precisely Bernoulli proved and what it meant for statistical inference in general, from the point of view of this lecture, his law helped people at the time make sense of **all the stable statistical frequencies that people were observing at the end of the 17th and the beginning of the 18th centuries**

Remember, this is when John Graunt was looking at the **Bills of Mortality,** observing a number of regularities in birth and death and marriage statistics, and even computing life tables proportions of people who died with different ailments -- This is also about the time that Arbuthnot made his own interesting comments on the Bills and the sex ratio

Bernoulli's theorem, a purely mathematical result, **seemed to tie the classical view of probability with the stable frequencies** observed by Graunt and Arbuthnot and others -- **If a random mechanism like coin flipping was at work in the world, then frequencies will be stable**

The frequentist view of probability

Which leads us to another view of probability -- In the frequentist view, we would say, for example, that an event has probability 1/3 if the event occurs about 1/3 of the time **in a long sequence of repetitions done under more or less ideal circumstances**

Of course the relationship between relative frequencies (proportions over many trials) and probability will come as **no surprise to people who actually play games of chance** -- Certainly people have been assigning odds on games for a long time, long before Bernoulli, Pascal and Fermat

And frankly, if you ask most practicing statisticians what we mean by probability and they aren't ready for a long conversation, this is the kind of answer you'll get -- **Certainly, over the years there have been many attempts to "verify" this interpretation of probability**, to extract some "objective" sense of probability by repeating trials a large number of times...

Passing time

In his “A Treatise on Probability,” the British Economist John Maynard Keynes discusses several attempts to verify the conclusions of Bernoulli’s Theorem -- He writes “**I record them because they have a good deal of historical and psychological interest, and because they satisfy a certain idle curiosity from which few students of probability are altogether free.”**

The French naturalist Count Buffon (1707-1788), who “assisted by a child tossing a coin into the air” recorded 2048 heads in 4040 flips (for a relative frequency of 0.507)

A similar experiment was carried out by **a student of the British mathematician De Morgan** (1806-1871) “for his own satisfaction” involving 4092 tosses, 2048 of which were heads (relative frequency of 0.500

Passing time

The Belgian mathematician/astronomer/statistician/sociologist Adolphe Quetelet (1796-1874) drew 4096 balls from an urn, replacing them each time, and recorded the result at different stages; in all, he drew 2066 white balls and 2030 black balls (relative frequency of 0.504)

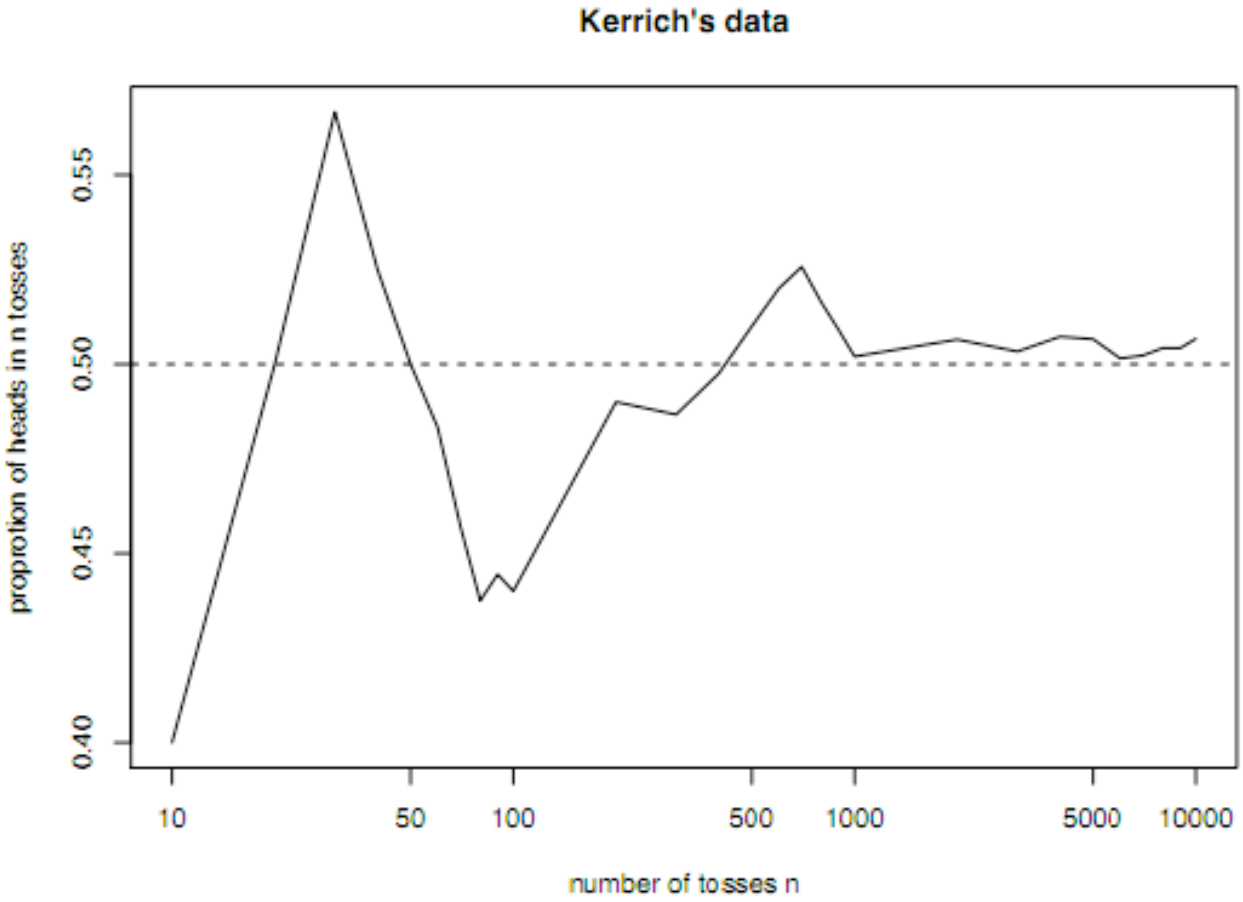
English economist W S Jevons (1835-1882) made 2048 throws of ten coins at a time; in all, he saw 20,480 tosses out of which 10,353 were heads (relative frequency of 0.506, although this is not quite the same kind of trial)

Around 1900, the English statistician Karl Pearson (1857-1936) made two heroic studies; the first involved 12,000 tosses (relative frequency of heads 0.52) and the second 24,000 times (12,012 of which landed heads for a relative frequency of 0.501)

While imprisoned by the Germans during World War II, **the South African mathematician John Kerrich** tossed a coin 10,000 times, 5067 of them heads (this gives a relative frequency of 0.5067 -- while interned, he also recorded a monograph **“An Experimental Introduction to the Theory of Probability”**

Passing time

Kerrich's data show the same pattern in relative frequencies that we observed from our computer simulation; as we repeat the trial over and over, the proportion of successes "settles down"



Passing time

Prisoners of war and 19th century intellectuals are not the only people to have tested the frequency notion of statistics; as it is the dominant interpretation of probability covered in introductory statistics textbooks, **students of statistics are routinely forced to participate**

At the right we have the results from several semesters of an introductory statistics class taught by Robin Lock at St. Lawrence University; he actually has his students record data on flips, spins and tips

"I have my students do a lab on this each semester. They do 100 flips, around 70 spins and 50 tips each - so I've accumulated lots of data, but I'm never completely sure about the reliability of the data. They do the trials on their own outside of class so I can't monitor how carefully they follow the instructions"

Flip (H)	Trials	prop	Semester
1079	2100	0.51	Fall 97A
1121	2260	0.50	Fall 97B
1071	2200	0.49	Spring 98
1093	2200	0.50	Fall 98A
1041	2000	0.52	Fall 98B
1232	2400	0.51	Fall 98C
802	1500	0.53	Spring 99
1002	2005	0.50	Fall 99A
1070	2200	0.49	Fall 99B
1000	2000	0.50	Spring 00
1021	2050	0.50	Fall 00A
984	1900	0.52	Fall 00B
1036	1900	0.55	Fall 01A
1157	2300	0.50	Fall 01B
14709	**29015**	**0.507**	**Combined**

www.dartmouth.edu/~chance/chance_news/for_chance_news/ChanceNews10.03/Robin.Lock.data.htm

The frequentist view of probability

At a technical level, strict frequentists view probability as arising from a sequence of identical trials; there is a problem lurking, however, in determining how long we should go

Sure, probabilities may seem stable after 10,000 trials, but who is to say that they won't change farther in the series? In some sense, the frequentists are led to imagining not just a long sequence of trials, but an infinitely long one

The frequentist view also provides us with no real ability to reason about singular events like the election of Trump or whether or not a particular person will have some medical condition — there is no imaginary infinite sequence of trials here

Frequentist statistics

So far, the inferential procedures we have studied (re-randomization and P-values) are based on **the frequentist notion of probability** --They refer to an (**imaginary) set of possible alternative outcomes that could have happened had we repeated the experiment many times**

D.R. Cox puts it this way

> *In the first so-called frequentist approach, we ... use probability as representing a long-run frequency... [W]e measure uncertainty via procedures such as confidence limits and significance levels (P-values), whose behaviour ... is assessed by* ***considering hypothetically how they perform when used repeatedly under the same conditions.*** *The performance may be studied analytically or by computer simulation.*
>
> *In that, the procedure is calibrated by what happens when it is used, it is no different from other measuring devices.*

The subjective view

The third view of probability we will talk about is more in line with Hacking's use of the term probability as a "relation between a hypothesis and the evidence for it"

It has its roots in a result by the Rev. T. Bayes, published in 1763 after his death; Bayes Theorem is a simple fact about conditional probabilities, that we will define next

The probability of any event is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon its happening.

The subjective view

As we mentioned at the beginning of the lecture, probability moves between two poles, as a framework for reasoning and as a "stable law" for frequencies

The subjective view rejects the interpretation of probability as a physical feature of the world and interprets probability as a statement about an individual's state of knowledge; Persi Diaconis at Stanford says "Coins don't have probabilities, people have probabilities"

While we could provide a fairly long history of how the first idea developed, we'll focus instead on one of the main proponents, Bruno de Finetti (1906-1985); he began his *"Theory of Probability"* with the statement "Probability does not exist"

The subjective view

To de Finetti, the definition of probability and its evaluation are two different things; he takes issue with the frequentists and the classical probabilists who seem to conflate these two and in so doing embrace a "rigid" attitude toward probability

By contrast, subjectivism maintains **a distinction between definition and evaluation;** probability is defined as the degree of belief "as actually held by someone, on the ground of his whole knowledge, experience, information" regarding an event whose outcome is uncertain

De Finetti writes:

> *The subjective theory... does not content that the opinions about probability are uniquely determined and justifiable. Probability does not correspond to a self-proclaimed "rational" belief but to the effective personal belief of anyone...*

He contends that

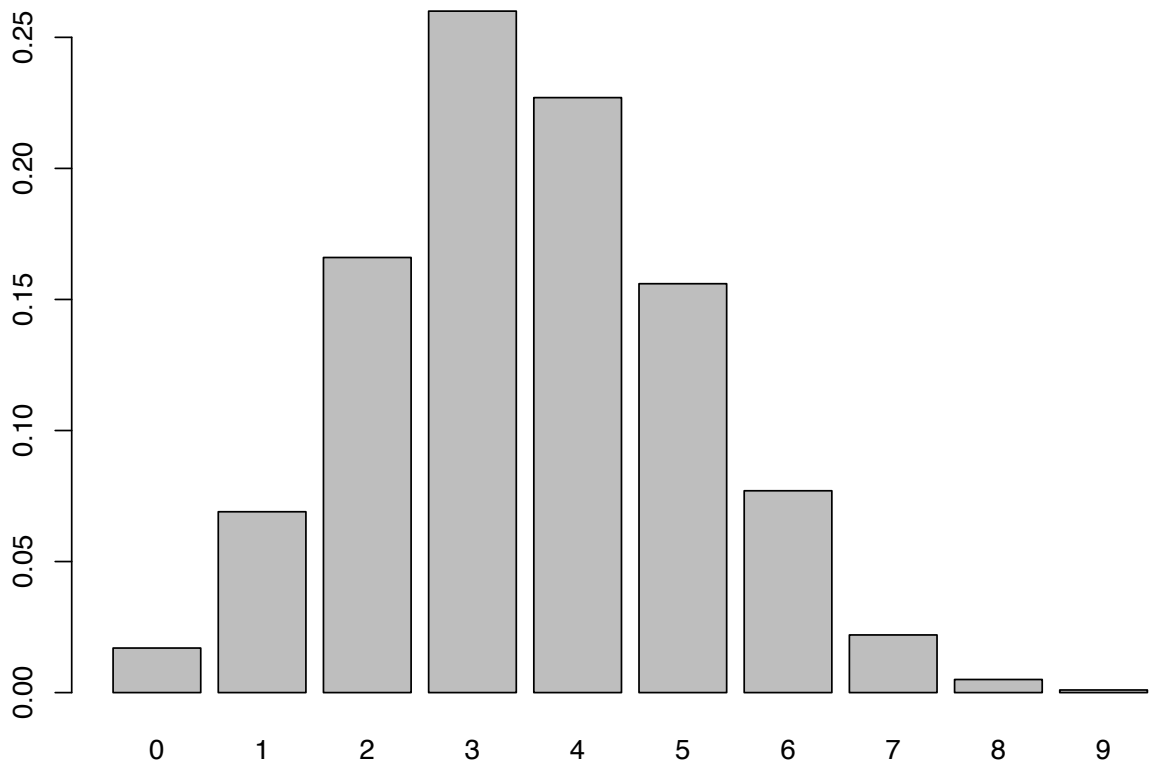
> *"every probability evaluation essentially depends on two components: (1) the objective component, consisting of the evidence of known data and facts; and (2) the subjective component, consisting of the opinion concerning unknown facts based on known evidence*

A simple example

Suppose we want to test a new therapy for some disease, **the standard treatment for which has a historical success rate of 35%** -- To examine the effectiveness of the new therapy we prescribe it to 10 patients and see whether they improve or not

In a frequentist framework, we would consider the “null hypothesis” that the new therapy is the same (or at least, not better) as the traditional treatment -- **Under the null our experimental results should look like 10 coin tosses with success probability (a patient getting better) being 0.35**

We can simulate (a la Arbuthnot) or use a mathematical result about the Binomial distribution to compute the distribution for the number of patients seeing improvement under this model



mathematical table using pbinom in R (more later)

0	1	2	3	4	5	6	7	8	9	10
0.013	0.072	0.176	0.252	0.238	0.154	0.069	0.021	0.004	0.001	0.000

The frequentist approach

At the end of the experiment, suppose we see 7 patients improve -- We can use our simulations or the mathematical table to compute the probability of seeing 7 or more successes if the chance of a success is p=0.35

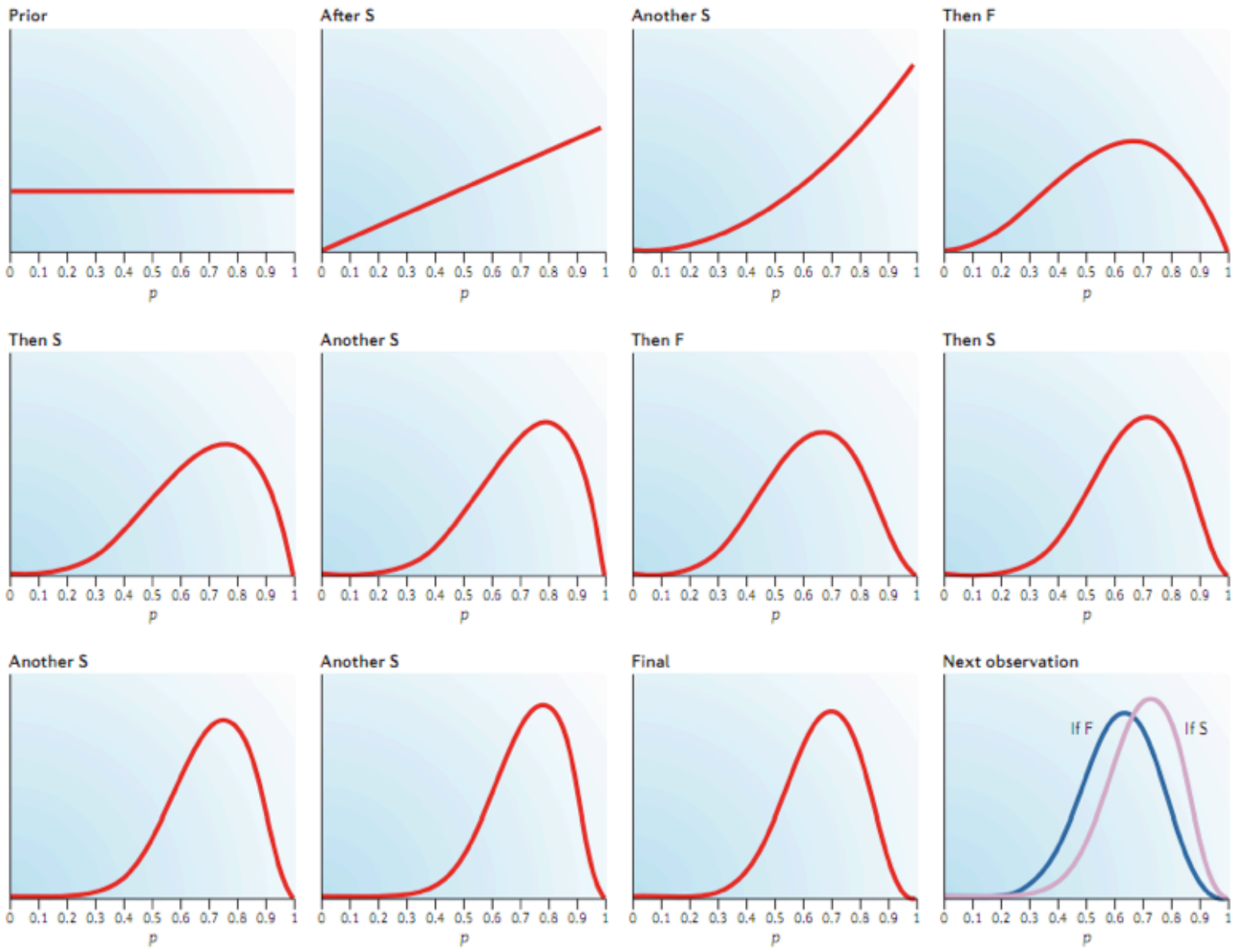
In this case, we sum up 0.021+0.004+0.001 = 0.026 to compute our P-value and we would reject the null hypothesis at the 0.05 level -- All pretty standard at this point

The Bayesian approach

In the Bayesian framework, **we use probability to express our uncertainty about unknown quantities, in this case the probability p** that someone improves on the new therapy

Our “prior” assessment might be one of complete ignorance -- We have no idea what value p might be except that it is in the interval [0,1], leading us to the uniform distribution

Starting from here, we can introduce data and update our beliefs about p -- Suppose that the experiment resulted in a sequence of successes and failures SSFSSFSSSF (again, 7 successes and 3 failures, but now listed in order of occurrence), then on the next page, **we update our beliefs sequentially**


Prior
After S
Another S
Then F
Then S
Another S
Then F
Then S
Another S
Another S
Final
Next observation
If F
If S
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
p

The Bayesian approach

While the mechanics of this are still opaque, the idea is clear -- **As we collect data, our notion of what values p can take become more and more sharp**, in this case coalescing on 0.7 (because we observed 7 out of 10 successes)

If we wanted to evaluate whether the new therapy was no better than the existing treatment, we would simply **compute the weight our current beliefs assign to the region to the left of 0.35** -- Literally we would find the area under the curve in the “Final” box to the left of 0.35... it is just 0.014

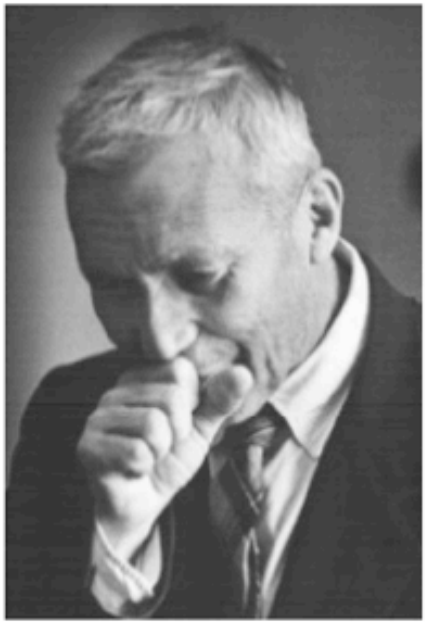
This number functions a bit like a P-value, but notice that the interpretation is very very different -- **The P-value references an imaginary set of experiments** while this approach **attempts to assess the evidence (data and prior beliefs) to support the idea** that p is less than or equal to 0.35

A computational view

There is one last view of probability I'd like to mention; at the left we have two pictures of Andrey Kolmogorov (1903-1987); he was a Russian who in the 1930s produced an axiomization of probability theory, a mathematical framework grounded in "measure theory"

Early in his career, he was a frequentist, believing that one should interpret probabilities as the result of long-run proportions of events in identical and independent trials

Toward the end of his life, he had a change of heart, and wanted to be able to speak about probability in finite terms; this led him to a somewhat remarkable line of reasoning



A computational view

Let's consider programs that "print" strings of numbers (let's say 0's and 1's for simplicity) -- Now, given a string, let's think about **the shortest program that would print the string**

If the string has a high degree of regularity

0101010101010101010101010101010101010101

then it can be printed with a short program (you just need to say **print "01" 20 times**) -- If the string has less structure,

10110110011111011100011101000100101000 11

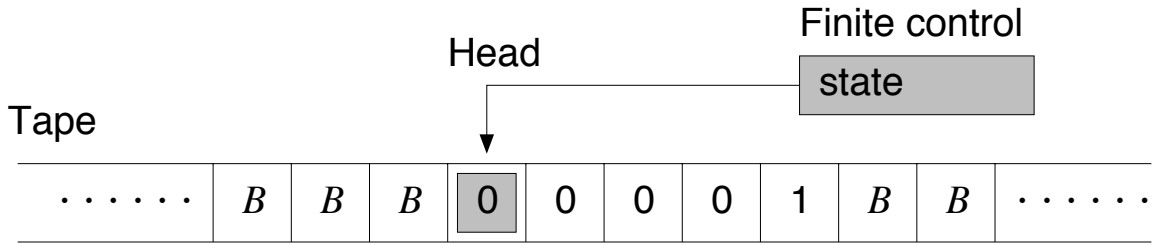
we might require a longer program

You can always have at least one program that does the job if you provide the actual string to the program and say

print "10110110011111011100011101000100101000 11"

A computational view

This feels very abstract and there are lots of questions to answer about the computer language you're using and what the commands might look like and so on -- Kolmogorov worked with **the mathematical abstraction of a computer known as a Turing machine**



This machine reads from a tape divided into cells and containing data (0 or 1 or a blank) and then takes action depending on its current state (taken from a set of states Q) -- The actions include moving the head, writing something to the tape or changing its state

The actions are specified in **a transition function** which you can think of as a program -- In 1936 A. M. Turing proposed the Turing machine as a model of **``any possible computation."**

A computational view

Kolmogorov linked this idea to a notion of what he called **universal probability** -- Simply, if we let L(s) be the length of the shortest program to print a string s (some pattern of 0's and 1's say), then $2^{-L(s)}$ can be thought of as its probability

Interestingly, this definition **embeds "Occam's razor",** the idea that the simplest explanation for an event is likely correct -- **Simple strings are much more probable than more complex strings and the most complex strings are considered "random"**

A computational view

The ideas behind this computational view are embedded in a number of statistical tools that try to choose between models for data -- They are a bit beyond the scope of this course, but the general framework should be intuitive

I bring this up now just to hint at the fact that probability is a fairly rich topic on its own and there are interesting approaches to the subject, each of which come with notions of inference and “randomness”

A more complete set of axioms

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

1. The probability of A is a number between 0 and 1, $0 \leq \mathrm{P}(\mathrm{A}) \leq 1$
2. The probability that an outcome will occur is 1, $\mathrm{P}(\mathcal{X}) = 1$
3. If A and B have no outcomes in common (they're disjoint), then their probabilities add $\mathrm{P}(\mathrm{A} \text{ or } \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B})$
4. Conditional probability:
$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \mathrm{P}(\mathrm{A} \text{ and } \mathrm{B})/\mathrm{P}(\mathrm{B}) \quad \text{or} \quad \mathrm{P}(\mathrm{A} \text{ and } \mathrm{B}) = \mathrm{P}(\mathrm{A}|\mathrm{B})\mathrm{P}(\mathrm{B})$$
5. The law of total probability: $\mathrm{P}(\mathrm{A}) = \mathrm{P}(\mathrm{A}|\mathrm{B}_1)\mathrm{P}(\mathrm{B}_1) + \cdots + \mathrm{P}(\mathrm{A}|\mathrm{B}_J)\mathrm{P}(\mathrm{B}_J)$ where $\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_J$ are all disjoint and their union is $\mathcal{X}$

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

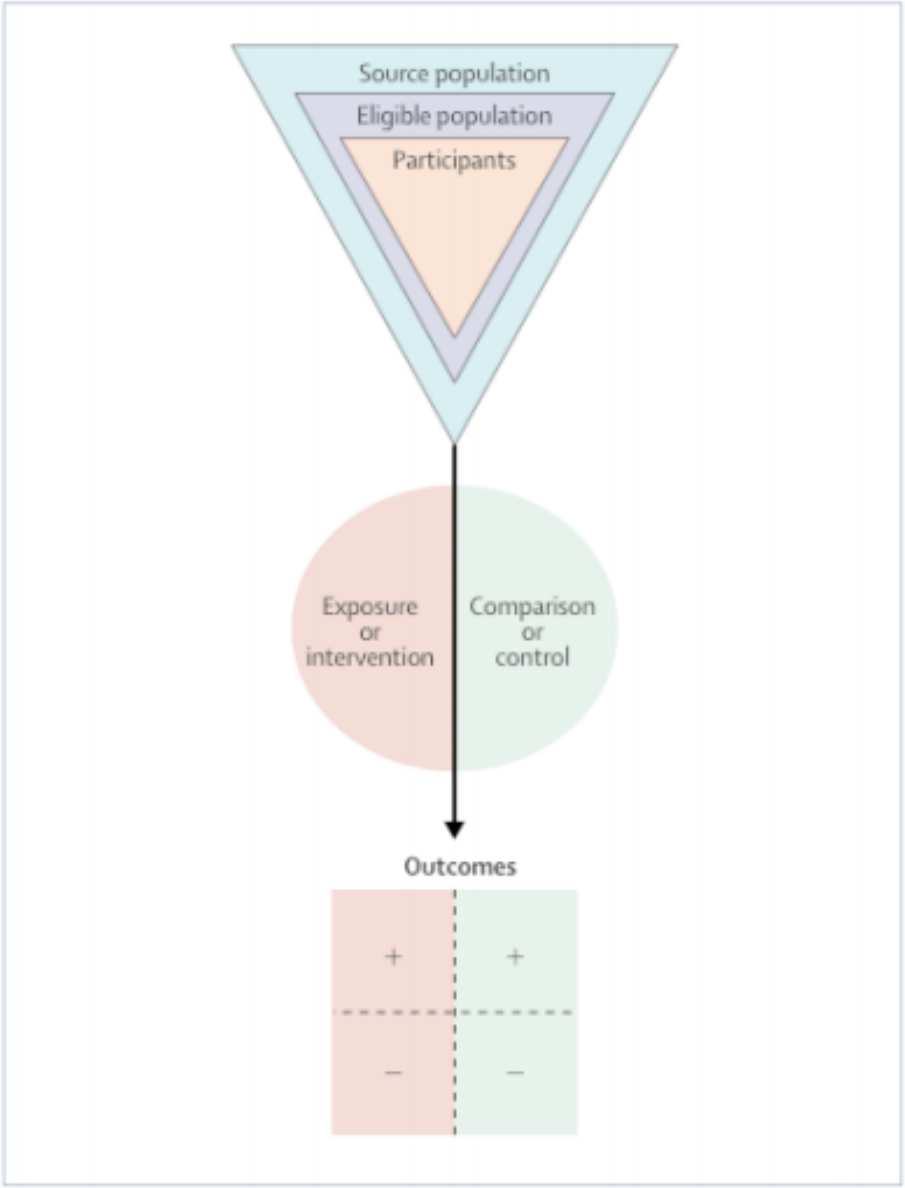
> It ensures that neither our personal idiosyncrasies (our likes or dislikes consciously or unwittingly applied) nor our lack of balanced judgement has entered into the construction of the different treatment groups—the allocation has been outside our control and the groups are therefore unbiased;
>
> ... 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.

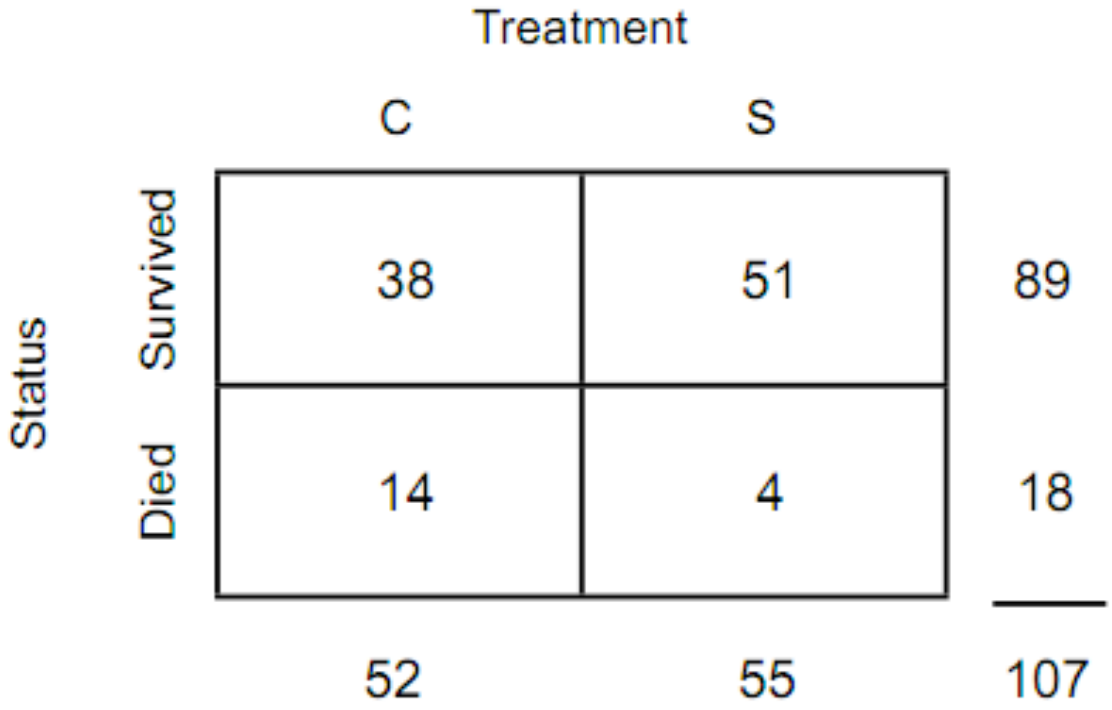
The following preliminary analysis is based on changes in the radiological picture alone, this being in our opinion the most important single factor to consider; it will be seen later that in the great majority of cases clinical and radiological changes followed similar trends.

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



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?

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

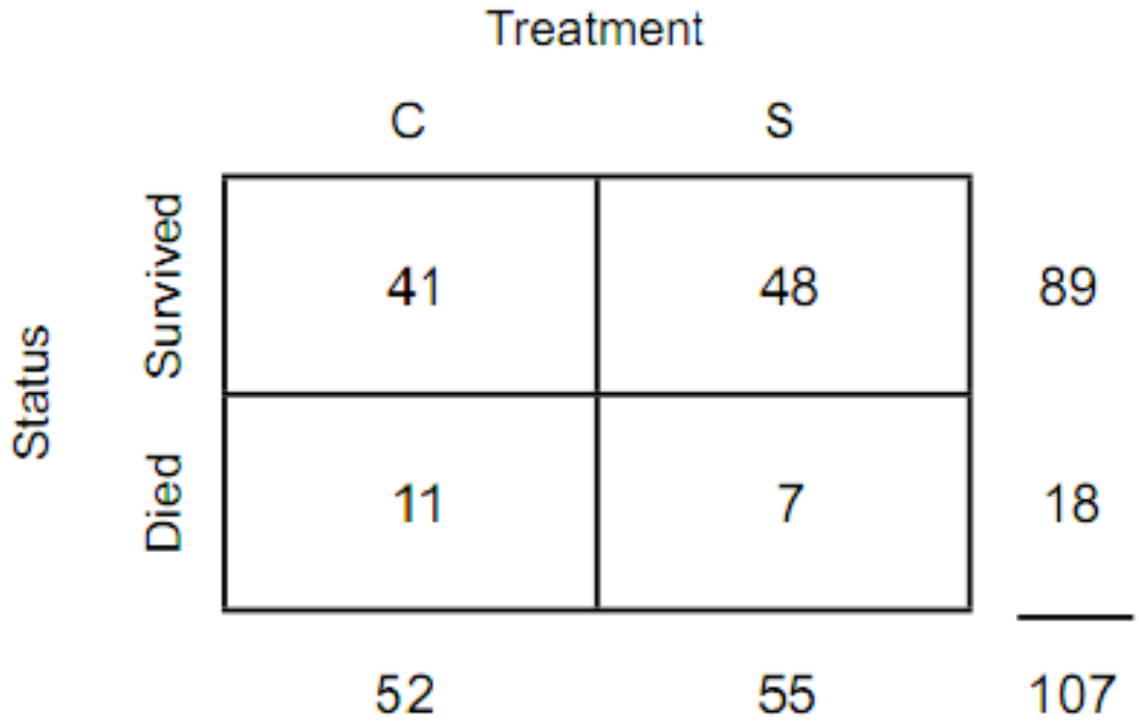
Therefore, under the null hypothesis, if we had chosen a different random assignment of patients, **we would still have 18 people who died and 89 who survived, but they would appear in different cells of the table**

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

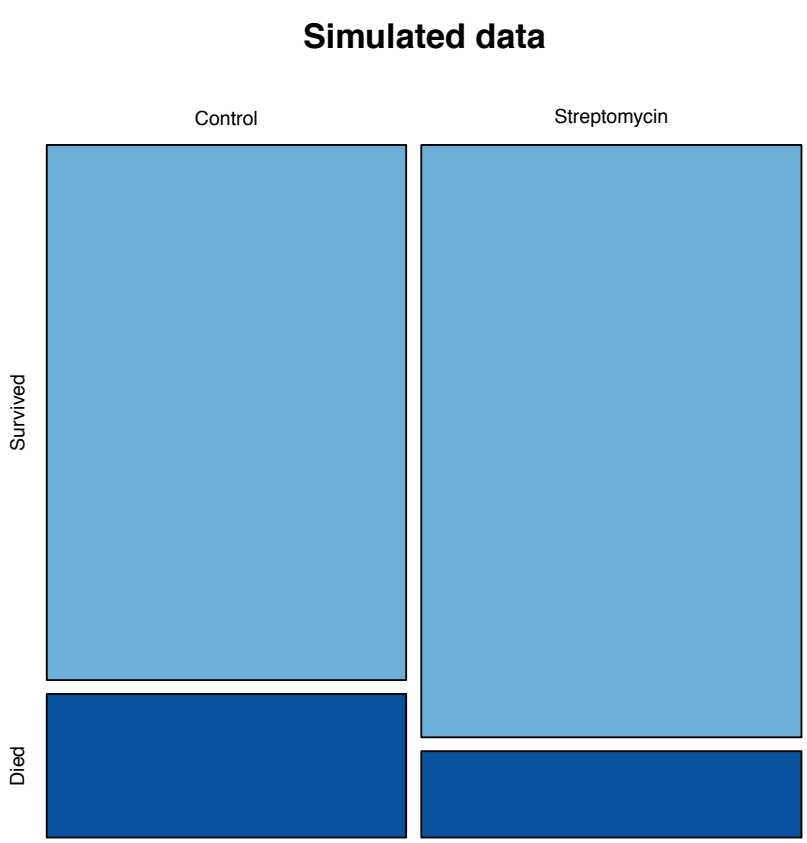
In this simulated table, we have 11/52 or 21% chance of dying under the control, and a 7/55 or 12% chance under Streptomycin; the treatment reduced the mortality rate among the participants by nearly 60%

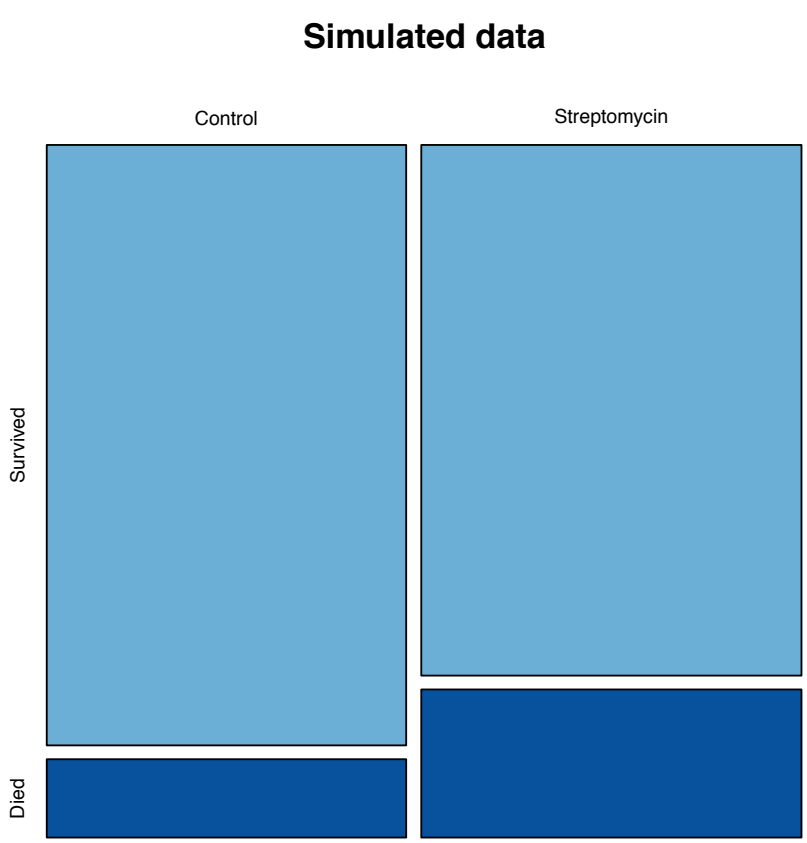


Simulating random assignments

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

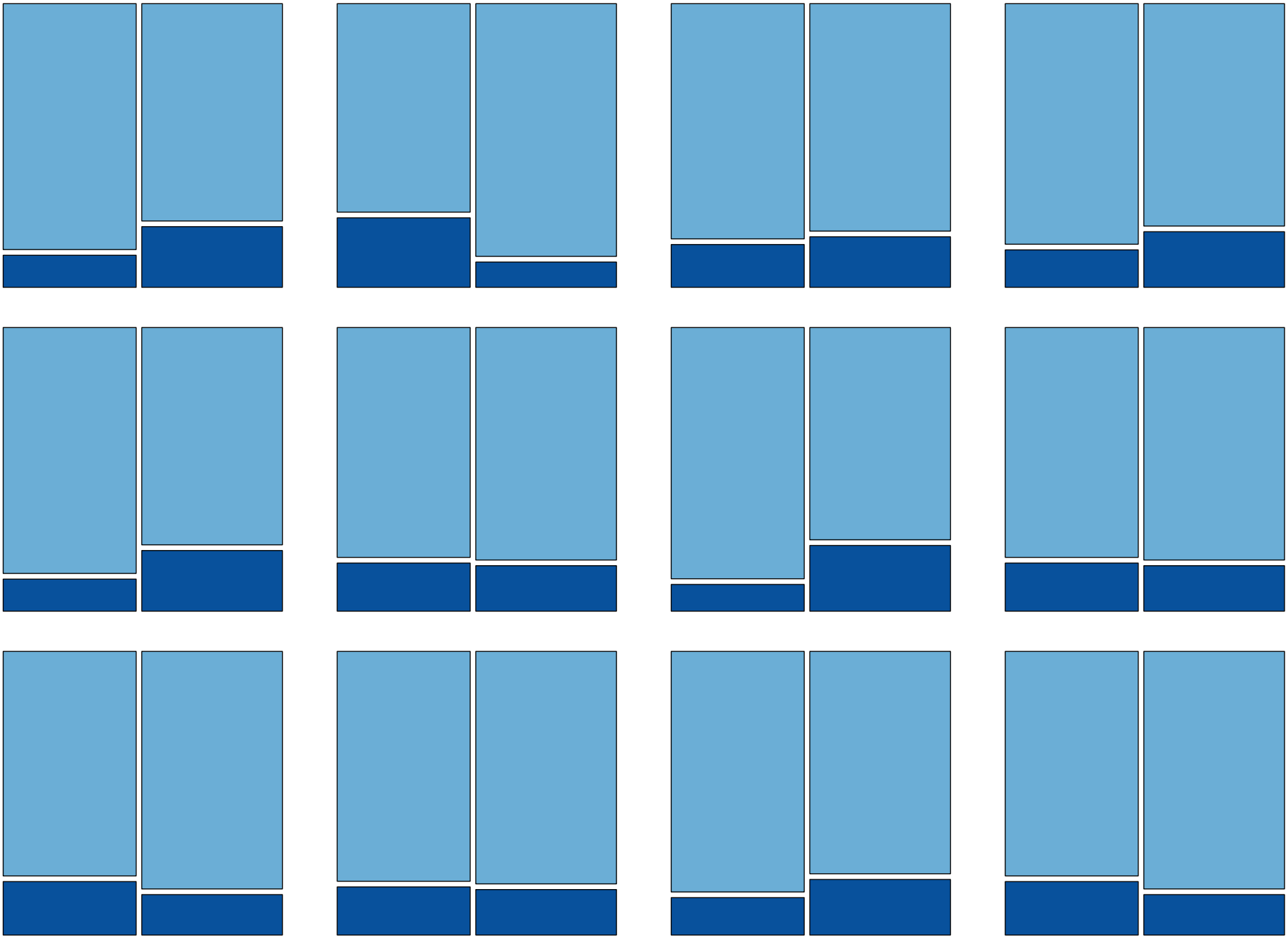
Simulating random assignments

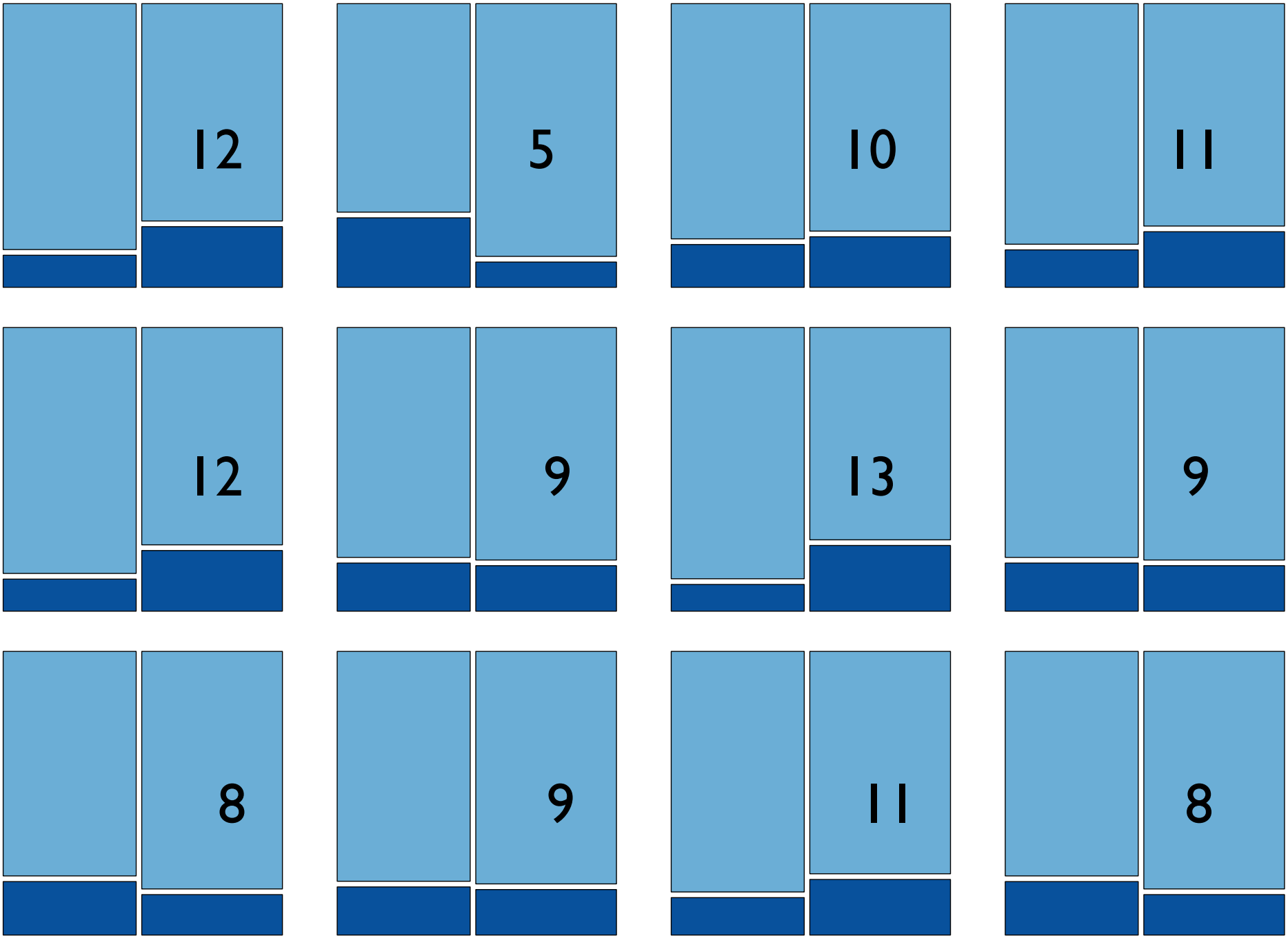
Notice that we only need to record **one piece of information for each trial, the number of deaths under Streptomycin** -- Knowing that we know all the other entries in the table

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

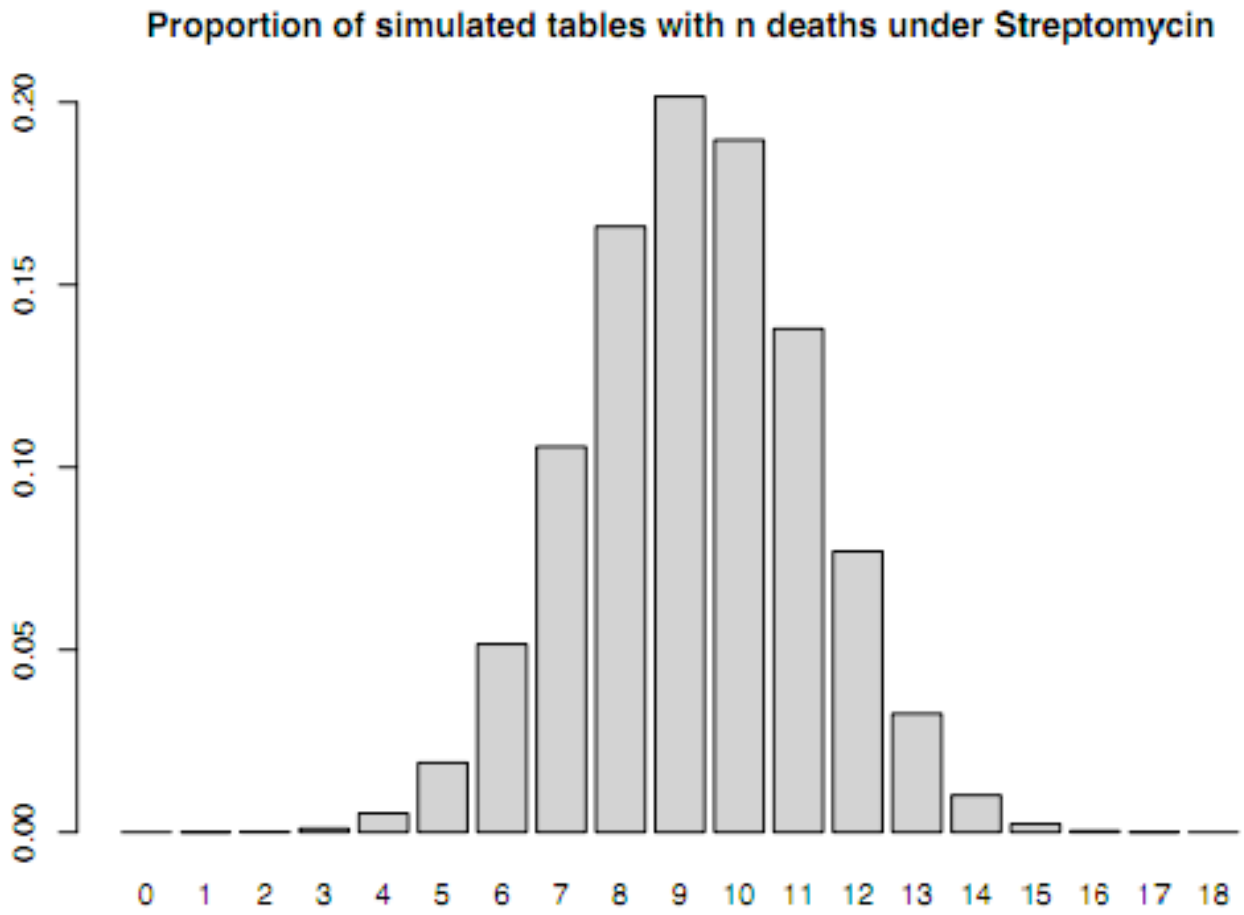
Therefore, the question becomes, under the random assignment patients to treatments, **how common is it for us to see 4 or fewer deaths in the Streptomycin group?**

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

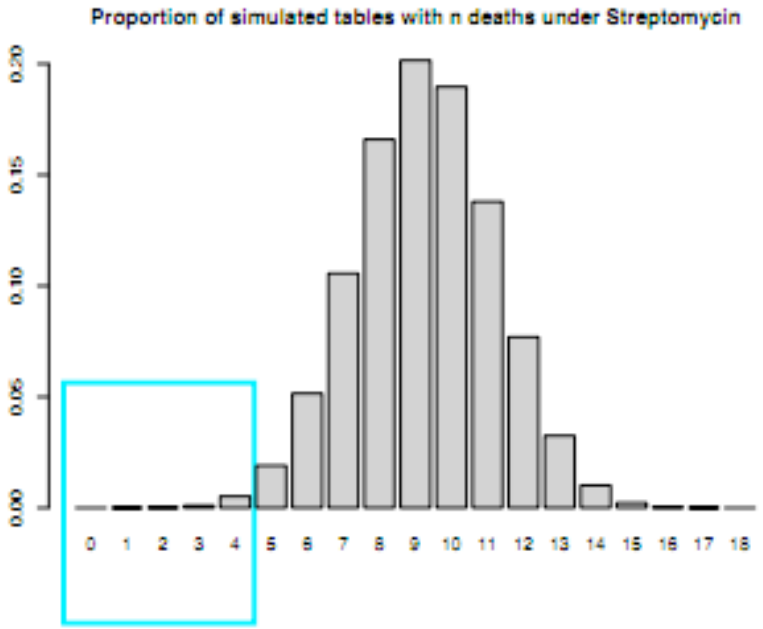

Proportion of simulated tables with n deaths under Streptomycin
0.00
0.05
0.10
0.15
0.20
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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

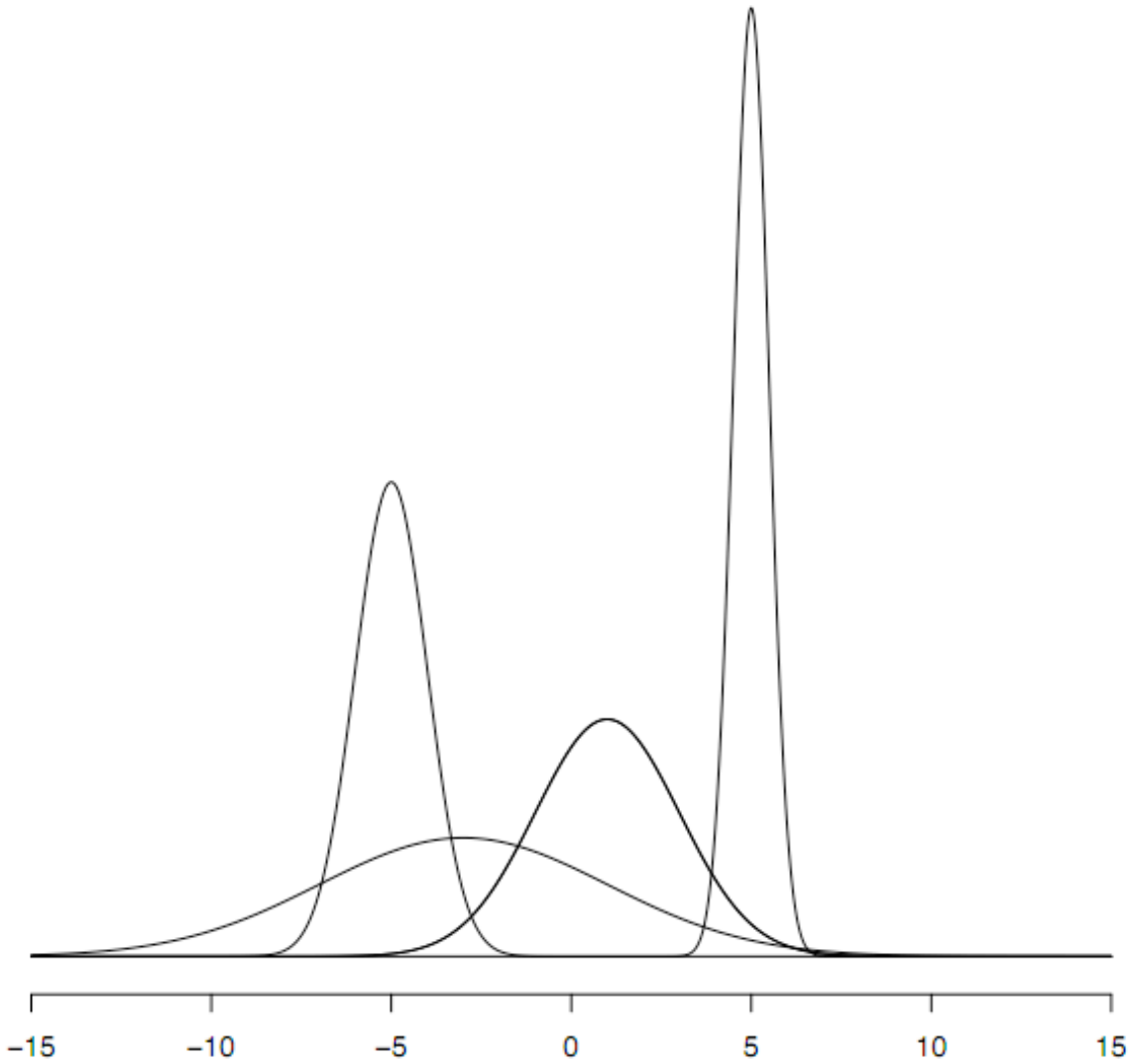
This, then, provides us with evidence that there is something more at work here than random assignment

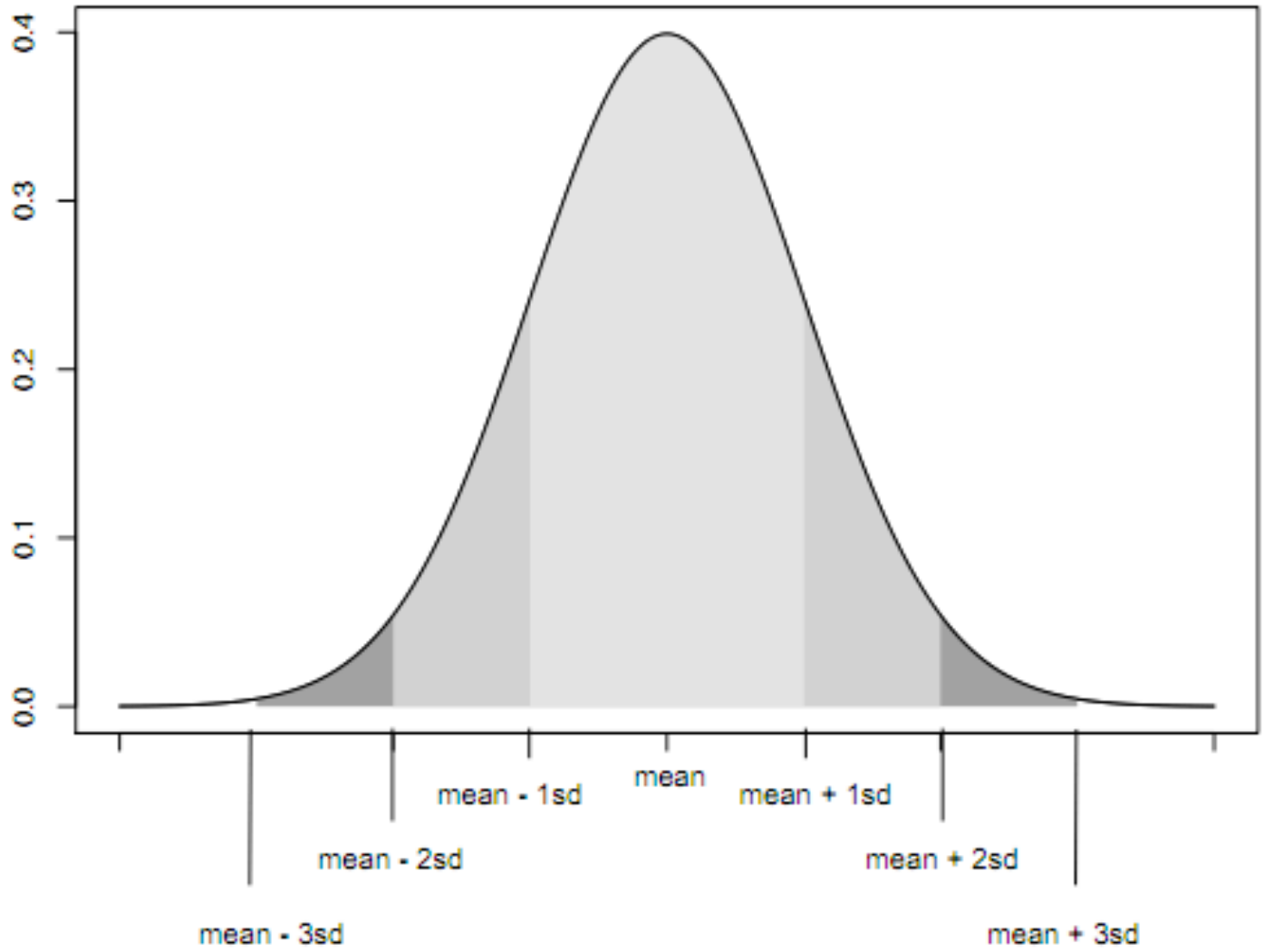
If we believed the null hypothesis, that there was no difference between Streptomycin and bed rest, the results HIII observed would have been extremely rare, coming up a very small fraction of the time



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.




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

“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.”

Generalities

Notice that it was Hill's **random assignment that gives our analysis it's validity** -- The way we collect data dictates the kinds of inferences we are allowed to make

However, **we have not said anything about what the effect of Streptomycin might be on patients outside the study** -- For that, we have to make more assumptions about how people were recruited into the trial (more on this next time)

This means that in addition to all the questions I had you ask about where data come from, you can now add a few technical ones -- **We are going to start paying attention to the role that randomness plays and, in particular, we will analyze as we randomized!**

Significance testing

Our discussion of P-values and our examination of the null distribution are in line with the methodology advocated by Fisher throughout his career; **the null hypothesis plays the role of devil's advocate, and a P-value provides evidence against the null** -- this is often called **significance testing**

There are a few obvious questions facing practitioners, the first of which involves evaluating the evidence provided by a P-value -- **Is 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 **the 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).

Vioxx

Randomized controlled trials are common in medical research; let's have a look at a more recent case

Vioxx, an anti-inflamatory agent was introduced to the market in the late 1990s and was prescribed for the treatment of arthritis and acute pain

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

8,076 patients suffering from rheumatoid arthritis were randomized into two treatment groups: One received a twice daily dose of 50 mg of rofecoxib while the other received 500 mg of naproxen

The New England

COMPARISON OF UPPER GASTROIN
AND NAPROXEN IN PATIENTS

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

ABSTRACT

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

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

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

Vioxx

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

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

Vioxx

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

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

Vioxx

As with Hill's data, this number seems convincing; and yet, we should ask whether or not these results could be produced by pure chance

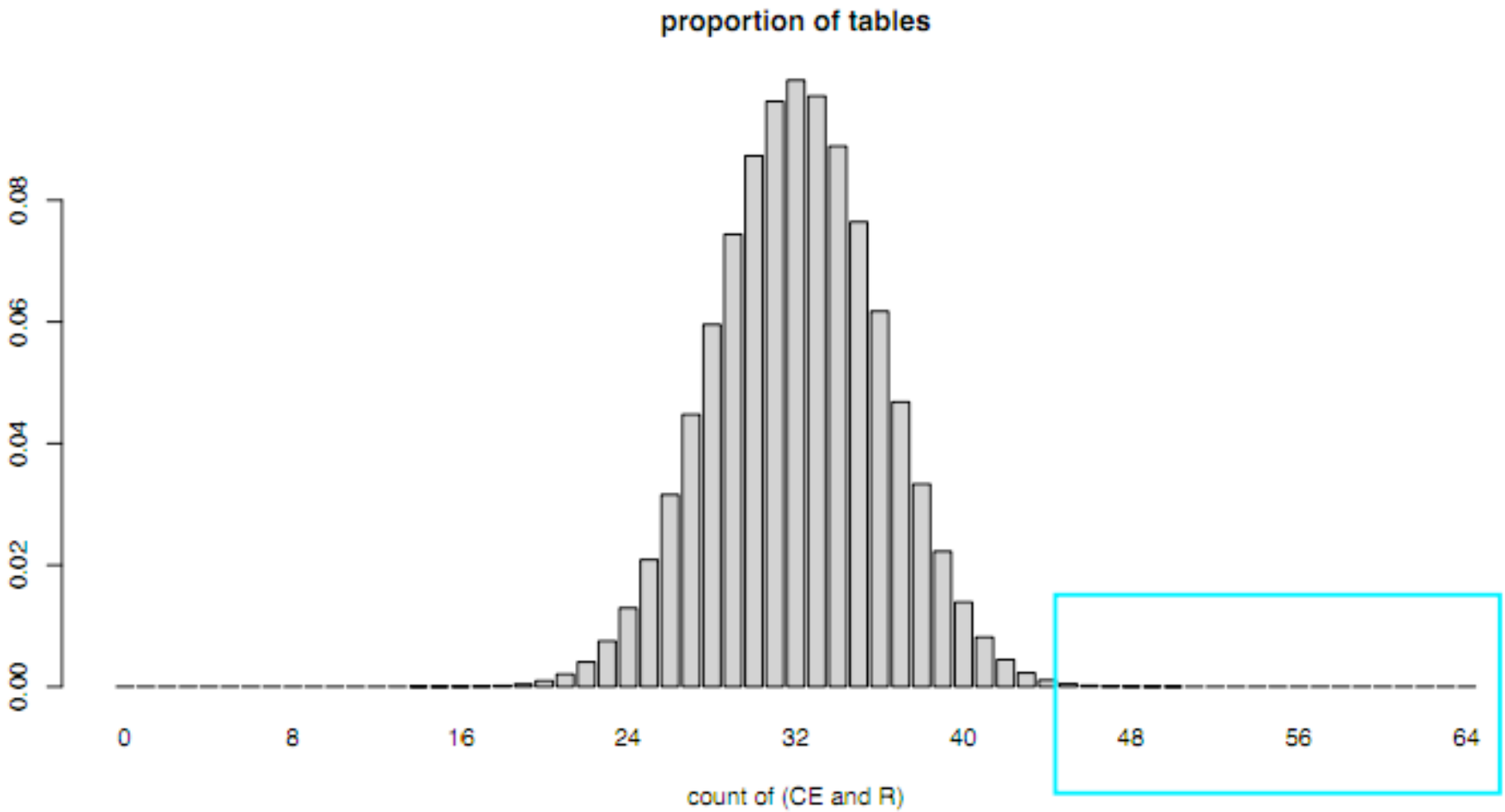
The same analysis framework holds up here; under the null hypothesis that patients are equally likely to have a CE under rofecoxib as naproxen, we can repeat the assignment of patients to treatment groups and examine the resulting tables

We're specifically interested in seeing how many tables yield results that are as strong or stronger than what we observed; in this case, under different randomizations of the patients, how often do we see tables with 45 or more deaths under Vioxx?

Well, we can simulate..

Vioxx

In this case, the data provide strong evidence that the proportions we observed were not due to the random assignment; the P-value here is 0.0000004



Vioxx

The evidence here is seems very strong; Merck, the manufacturer of Vioxx, however, concluded instead that **it wasn't that Vioxx was responsible for more adverse events, it was that naproxen helped reduce them**

As the research community debated the meaning of this particular trial, several ongoing trials began to exhibit similar problems, this time with placebos as the control treatment; eventually, **in 2004, Merck withdrew the drug from the market, citing increased risk of cardiovascular adverse events**

The debate then turned from whether Vioxx was harmful to what Merck scientists and senior management knew of these hazards, and **when they knew it**

A problem with thresholds

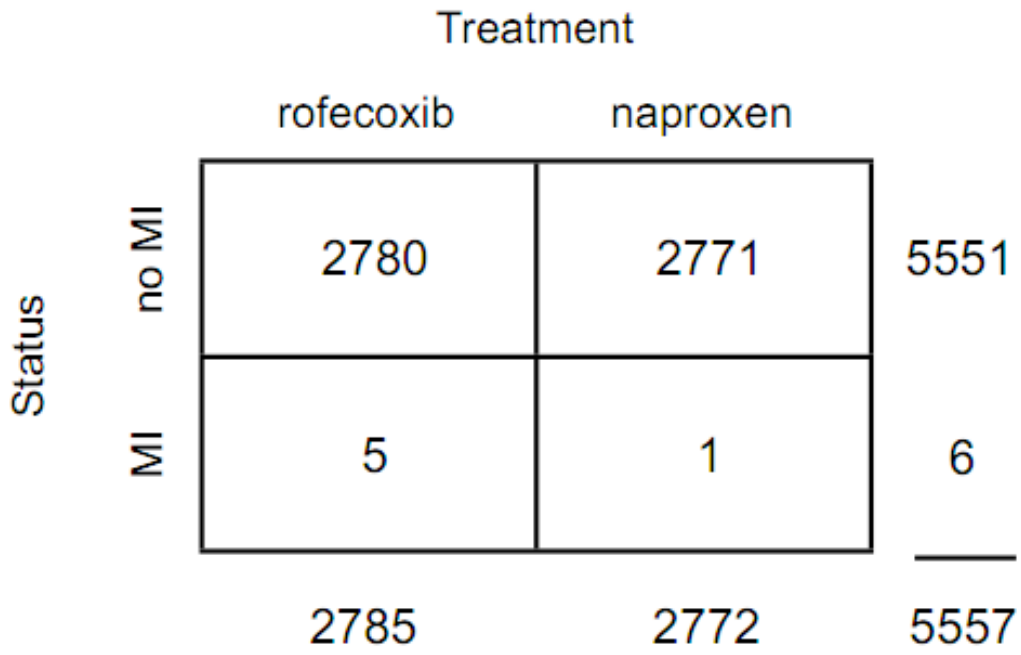
In 2003, before Merck pulled Vioxx from the shelves, Lisse et al published a report in the Annals of Internal Medicine; **with Merck funding, this group again compared rofecoxib to naproxen** for relief of osteoarthritis pain

The bulk of the paper was concerned with “gastrointestinal tolerability” of rofexocib; but they did mention some cardiovascular problems; specifically, **five people died of myocardial infarction (heart attack) while on Vioxx, while only 1 did from the naproxen group**

In the study they say that “a Fisher exact test was used to compare incidence of confirmed... cardiovascular events” **but that the results were not significant** (evidently appealing to a 5% cutoff level for the P-value)

Lisse et al (2003)

Here are the results in tabular form -- While small, the conditional proportion of people having MI under Vioxx is five times that for naproxen



Evidence in Vioxx Suits Shows Intervention by Merck Officials

By **ALEX BERENSON** APRIL 24, 2005

In 2000, amid rising concerns that its painkiller Vioxx posed heart risks, Merck overruled one of its own scientists after he suggested that a patient in a clinical trial had probably died of a heart attack.

In an e-mail exchange about Vioxx, the company's most important new drug at the time, a senior Merck scientist repeatedly urged the researcher to change his views about the death "so that we don't raise concerns." In later reports to the Food and Drug Administration and in a paper published in 2003, Merck listed the cause of death as "unknown" for the patient, a 73-year-old woman.

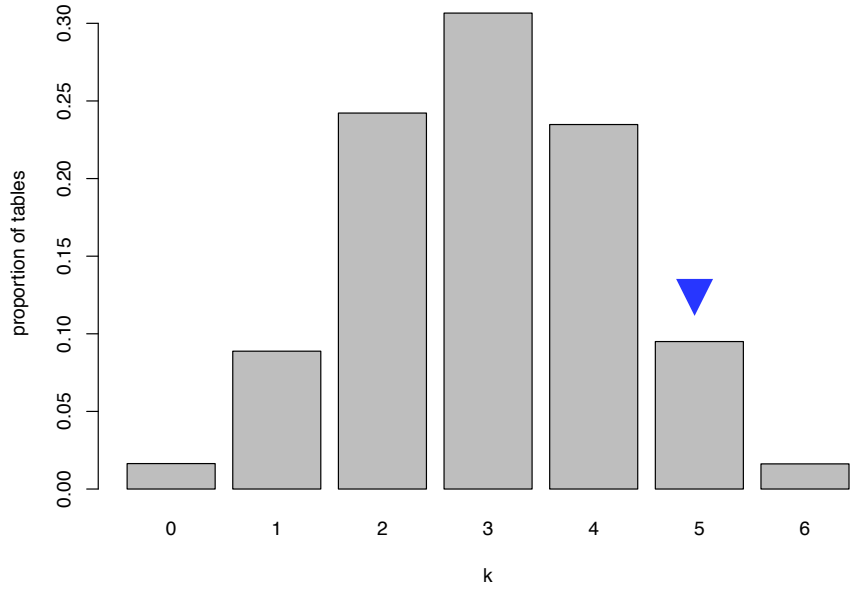
The discussion of the death is contained in several previously undisclosed Merck records, including e-mail messages from Dr. Edward M. Scolnick, Merck's top scientist from 1985 until 2002, and from Dr. Alise S. Reicin, a vice president for clinical research, that indicate Merck's concerns about data contradicting its view that Vioxx was safe.

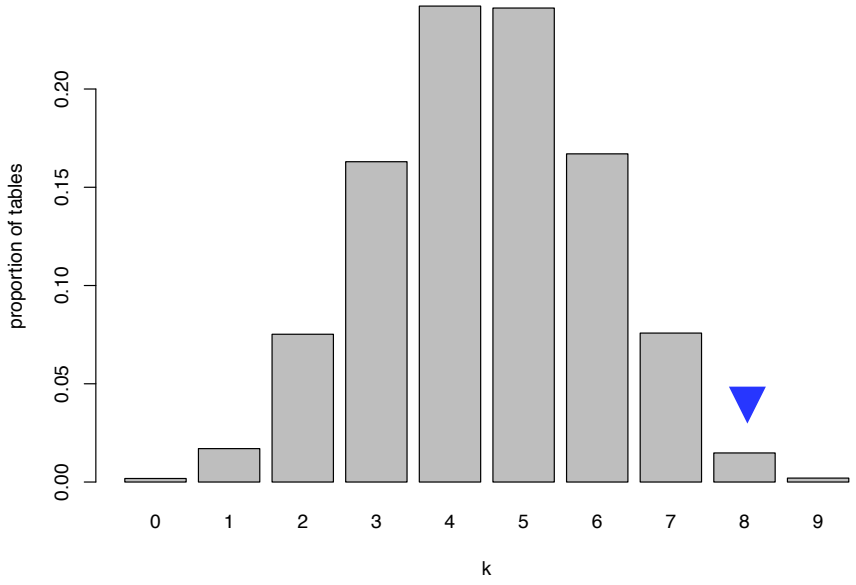
The problem with thresholds

At the right (top) we show the null distribution associated with Lisse's data; clearly the results are not very extreme given the randomness involved (a P-value of 0.11)

The problem, comes, however, as part of a Federal investigation into Vioxx, **Merck was forced to disclose three more deaths in the Vioxx group**; this changes the distribution to that shown in lower panel

In this case, the P-value is 0.02, smaller than the 0.05 cutoff and now significant





The problem with thresholds

We appeal to statistics because we want some kind of a simple procedure for telling truth from fiction -- This leads to cutoffs on P-values

The problem is that the objective security may blind us from important results, or have us fixate on effects that are statistically significant but uninteresting; either way, **many disciplines have felt the sting of having researchers incentivized to be on one side or another of a very hard threshold**

Over the last few decades, there have been many attempts to improve how scientific results are reported, how evidence is presented; in the next few lectures we will come across constructions like confidence intervals that many insist are more sensible summaries than P-values or a significance test

Random number generation

So far, we have emphasized the use of **graphics and simple simulation** (re-randomization) to analyze a data set -- We will circle back to talk about probability more formally next lecture (probably), but before that we should (probably) **put simulation on firmer footing**

For example, we seem to be trusting that R can generate “randomizations” for us, dividing or redividing patients into treatment and control -- In short, this means that **we can depend on the computer to toss coins for us**

How does it do this?

Random number generation

In Hill's trial, the samples were small enough that you cold rely on actual "random" mechanisms like **drawing tickets from a hat** (and we'll see lots of examples of the heroic work behind pre-computer simulation!)

Even **Francis Galton** (right, and we'll see a fair bit from him later too!) in the late 1880s recognized the need for statisticians to have access to simulated data (and suggested various physical mechanisms)

Besides randomized trials, toward what other purposes might we apply a sequence of random numbers?

DICE FOR STATISTICAL EXPERIMENTS.

EVERY statistician wants now and then to test the practical value of some theoretical process, it may be of smoothing, or of interpolation, or of obtaining a measure of variability, or of making some particular deduction or inference. It happened not long ago, while both a friend and myself were trying to find appropriate series for one of the above purposes, that the same week brought me letters from two eminent statisticians asking if I knew of any such series suitable for their own respective and separate needs. The assurance of a real demand for such things induced me to work out a method for supplying it, which I have already used frequently, and finding it to be perfectly effective, take this opportunity of putting it on record.

The desideratum is a set of values taken at random out of a series that is known to conform strictly to the law of frequency of error, the probable error of any single value in the series being also accurately known. We have (1) to procure such a series, and (2) to take random values out of it in an expeditious way.

Suppose the axis of the curve of distribution (whose ordinates at 100 equidistant divisions are given in my "Natural Inheritance," p. 205) to be divided into n equal parts, and that a column is erected on each of these, of a + or a − height as the case may be, equal to the height of the ordinate at the middle of each part. Then the values of these heights will form a series that is strictly conformable to the law of frequency when n is infinite, and closely conformable when n is fairly large. Moreover the probable error of any one of these values irrespectively of its sign, is 1.

As an instrument for selecting at random, I have found nothing superior to dice. It is most tedious to shuffle cards thoroughly between each successive draw, and the method of mixing and stirring up marked balls in a bag is more tedious still. A tee-totum or some form of roulette is preferable to these, but dice are better than all. When they are shaken and tossed in a basket, they hurtle so variously against one another and against the ribs of the basket-work that they tumble wildly about, and their positions at the outset afford no perceptible clue to what they will be after even a single good shake and toss. The chances afforded by a die are more various than are commonly supposed; there are 24 equal possibilities, and not only 6, because each face has four edges that may be utilized, as I shall show.

I use cubes of wood $1\frac{1}{4}$ inch in the side, for the dice.

... at one time in the not too distant past, this problem was addressed in a very direct way!

A MILLION
Random Digits

WITH

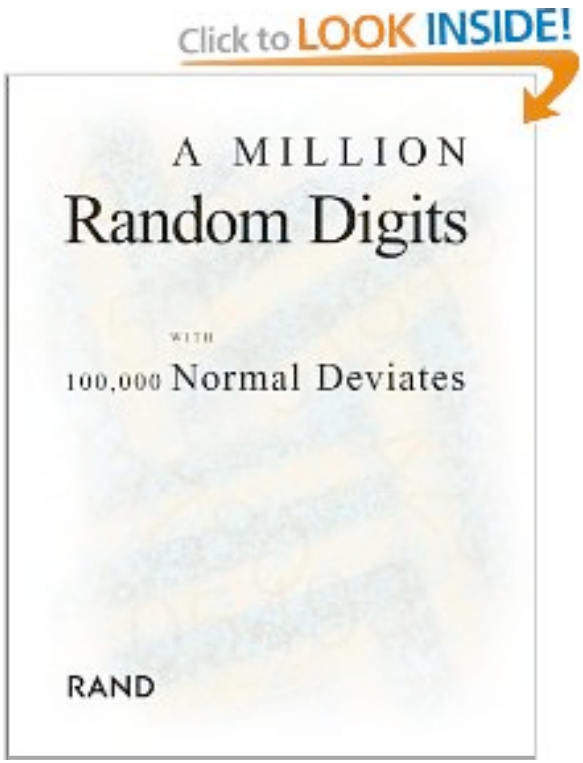
100,000 Normal Deviates

RAND

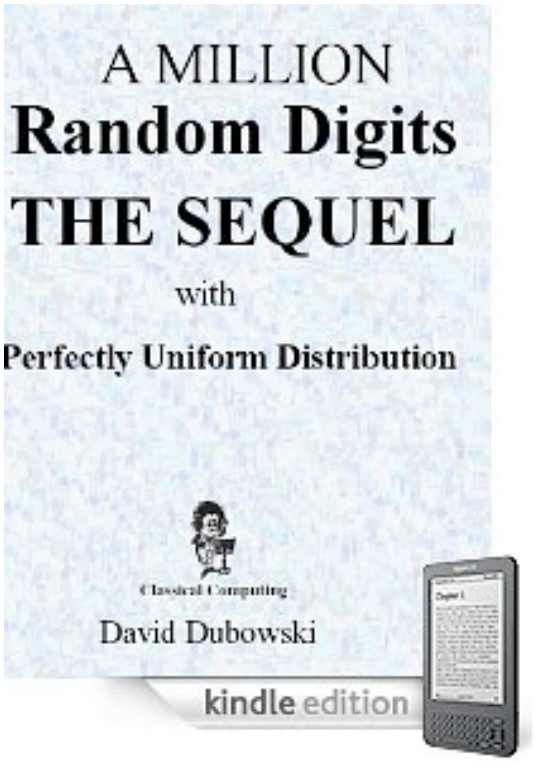
00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
00026	15474	45266	95270	79953	59367	83848	82396	10118	33211	59466
00027	94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
00028	42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
00029	23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
00030	04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00031	00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
00032	35963	15307	26898	09354	33351	35462	77974	50024	90103	39333
00033	59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
00034	46058	85236	01390	92286	77281	44077	93910	83647	70617	42941
00035	32179	00597	87379	25241	05567	07007	86743	17157	85394	11838
00036	69234	61406	20117	45204	15956	60000	18743	92423	97118	96338
00037	19565	41430	01758	75379	40419	21585	66674	36806	84962	85207
00038	45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
00039	94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
00040	98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
00041	33185	16232	41941	50949	89435	48581	88695	41994	37548	73043
00042	80951	00406	96382	70774	20151	23387	25016	25298	94624	61171
00043	79752	49140	71961	28296	69861	02591	74852	20539	00387	59579
00044	18633	32537	98145	06571	31010	24674	05455	61427	77938	91936
00045	74029	43902	77557	32270	97790	17119	52527	58021	80814	51748
00046	54178	45611	80993	37143	05335	12969	56127	19255	36040	90324
00047	11664	49883	52079	84827	59381	71539	09973	33440	88461	23356
00048	48324	77928	31249	64710	02295	36870	32307	57546	15020	09994
00049	69074	94138	87637	91976	35584	04401	10518	21615	01848	76938

00050	09188	20097	32825	39527	04220	86304	83389	87374	64278	58044
00051	90045	85497	51981	50654	94938	81997	91870	76150	68476	64659
00052	73189	50207	47677	26269	62290	64464	27124	67018	41361	82760
00053	75768	76490	20971	87749	90429	12272	95375	05871	93823	43178
00054	54016	44056	66281	31003	00682	27398	20714	53295	07706	17813
00055	08358	69910	78542	42785	13661	58873	04618	97553	31223	08420
00056	28306	03264	81333	10591	40510	07893	32604	60475	94119	01840
00057	53840	86233	81594	13628	51215	90290	28466	68795	77762	20791
00058	91757	53741	61613	62269	50263	90212	55781	76514	83483	47055
00059	89415	92694	00397	58391	12607	17646	48949	72306	94541	37408
00060	77513	03820	86864	29901	68414	82774	51908	13980	72893	55507
00061	19502	37174	69979	20288	55210	29773	74287	75251	65344	67415
00062	21818	59313	93278	81757	05686	73156	07082	85046	31853	38452
00063	51474	66499	68107	23621	94049	91345	42836	09191	08007	45449
00064	99559	68331	62535	24170	69777	12830	74819	78142	43860	72834
00065	33713	48007	93584	72869	51926	64721	58303	29822	93174	93972
00066	85274	86893	11303	22970	28834	34137	73515	90400	71148	43643
00067	84133	89640	44035	52166	73852	70091	61222	60561	62327	18423
00068	56732	16234	17395	96131	10123	91622	85496	57560	81604	18880
00069	65138	56806	87648	85261	34313	65861	45875	21069	85644	47277
00070	38001	02176	81719	11711	71602	92937	74219	64049	65584	49698
00071	37402	96397	01304	77586	56271	10086	47324	62605	40030	37438
00072	97125	40348	87083	31417	21815	39250	75237	62047	15501	29578
00073	21826	41134	47143	34072	64638	85902	49139	06441	03856	54552
00074	73135	42742	95719	09035	85794	74296	08789	88156	64691	19202
00075	07638	77929	03061	18072	96207	44156	23821	99538	04713	66994
00076	60528	83441	07954	19814	59175	20695	05533	52139	61212	06455
00077	83596	35655	06958	92983	05128	09719	77433	53783	92301	50498
00078	10850	62746	99599	10507	13499	06319	53075	71839	06410	19362
00079	39820	98952	43622	63147	64421	80814	43800	09351	31024	73167
00080	59580	06478	75569	78800	88835	54486	23768	06156	04111	08408
00081	38508	07341	23793	48763	90822	97022	17719	04207	95954	49953
00082	30692	70668	94688	16127	56196	80091	82067	63400	05462	69200
00083	65443	95659	18288	27437	49632	24041	08337	65676	96299	90836
00084	27267	50264	13192	72294	07477	44606	17985	48911	97341	30358
00085	91307	06991	19072	24210	36699	53728	28825	35793	28976	66252
00086	68434	94688	84473	13622	62128	98408	12843	82590	09815	93146
00087	48908	15877	54745	24591	35700	04754	83824	52692	54130	55160
00088	06913	45197	42672	78601	11883	09528	63011	98901	14974	40344
00089	10455	16019	14210	33712	91342	37821	88325	80851	43667	70883
00090	12883	97343	65027	61184	04285	01392	17974	15077	90712	26769
00091	21778	30976	38807	36961	31649	42096	63281	02023	08816	47449
00092	19523	59515	65122	59659	86283	68258	69572	13798	16435	91529
00093	67245	52670	35583	16563	79246	86686	76463	34222	26655	90802
00094	60584	47377	07500	37992	45134	26529	26760	83637	41326	44344
00095	53853	41377	36066	94850	58838	73859	49364	73331	96240	43642
00096	24637	38736	74384	89342	52623	07992	12369	18601	03742	83873
00097	83080	12451	38992	22815	07759	51777	97377	27585	51972	37867
00098	16444	24334	36151	99073	27493	70939	85130	32552	54846	54759
00099	60790	18157	57178	65762	11161	78576	45819	52979	65130	04860

00100	03991	10461	93716	16894	66083	24653	84609	58232	88618	19161
00101	38555	95554	32886	59780	08355	60860	29735	47762	71299	23853
00102	17546	73704	92052	46215	55121	29281	59076	07936	27954	58909
00103	32643	52861	95819	06831	00911	98936	76355	93779	80863	00514
00104	69572	68777	39510	35905	14060	40619	29549	69616	33564	60780
00105	24122	66591	27699	06494	14845	46672	61958	77100	90899	75754
00106	61196	30231	92962	61773	41839	55382	17267	70943	78038	70267
00107	30532	21704	10274	12202	39685	23309	10061	68829	55986	66485
00108	03788	97599	75867	20717	74416	53166	35208	33374	87539	08823
00109	48228	63379	85783	47619	53152	67433	35663	52972	16818	60311
00110	60365	94653	35075	33949	42614	29297	01918	28316	98953	73231
00111	83799	42402	56623	34442	34994	41374	70071	14736	09958	18065
00112	32960	07405	36409	83232	99385	41600	11133	07586	15917	06253
00113	19322	53845	57620	52606	66497	68646	78138	66559	19640	99413
00114	11220	94747	07399	37408	48509	23929	27482	45476	85244	35159
00115	31751	57260	68980	05339	15470	48355	88651	22596	03152	19121
00116	88492	99382	14454	04504	20094	98977	74843	93413	22109	78508
00117	30934	47744	07481	83828	73788	06533	28597	20405	94205	20380
00118	22888	48893	27499	98748	60530	45128	74022	84617	82037	10268
00119	78212	16993	35902	91386	44372	15486	65741	14014	87481	37220
00120	41849	84547	46850	52326	34677	58300	74910	64345	19325	81549
00121	46352	33049	69248	93460	45305	07521	61318	31855	14413	70951
00122	11087	96294	14013	31792	59747	67277	76503	34513	39663	77544
00123	52701	08337	56303	87315	16520	69676	11654	99893	02181	68161
00124	57275	36898	81304	48585	68652	27376	92852	55866	88448	03584
00125	20857	73156	70284	24326	79375	95220	01159	63267	10622	48391
00126	15633	84924	90415	93614	33521	26665	55823	47641	86225	31704
00127	92694	48297	39904	02115	59589	49067	66821	41575	49767	04037
00128	77613	19019	88152	00080	20554	91409	96277	48257	50816	97616
00129	38688	32486	45134	63545	59404	72059	43947	51680	43852	59693
00130	25163	01889	70014	15021	41290	67312	71857	15957	68971	11403
00131	65251	07629	37239	33295	05870	01119	92784	26340	18477	65622
00132	36815	43625	18637	37509	82444	99005	04921	73701	14707	93997
00133	64397	11692	05327	82162	20247	81759	45197	25332	83745	22567
00134	04515	25624	95096	67946	48460	85558	15191	18782	16930	33361
00135	83761	60873	43253	84145	60833	25983	01291	41349	20368	07126
00136	14387	06345	80854	09279	43529	06318	38384	74761	41196	37480
00137	51321	92246	80088	77074	88722	56736	66164	49431	66919	31678
00138	72472	00008	80890	18002	94813	31900	54155	83436	35352	54131
00139	05466	55306	93128	18464	74457	90561	72848	11834	79982	68416
00140	39528	72484	82474	25593	48545	35247	18619	13674	18611	19241
00141	81616	18711	53342	44276	75122	11724	74627	73707	58319	15997
00142	07586	16120	82641	22820	92904	13141	32392	19763	61199	67940
00143	90767	04235	13574	17200	69902	63742	78464	22501	18627	90872
00144	40188	28193	29593	88627	94972	11598	62095	36787	00441	58997
00145	34414	82157	86887	55087	19152	00023	12302	80783	32624	68691
00146	63439	75363	44989	16822	36024	00867	76378	41605	65961	73488
00147	67049	09070	93399	45547	94458	74284	05041	49807	20288	34060
00148	79495	04146	52162	90286	54158	34243	46978	35482	59362	95938
00149	91704	30552	04737	21031	75051	93029	47665	64382	99782	93478


Click to LOOK INSIDE!
A MILLION
Random Digits
WITH
100,000 Normal Deviates
RAND


Click to LOOK INSIDE!
A Small Book
Of
Random Numbers
James McNalley


A MILLION
Random Digits
THE SEQUEL
with
Perfectly Uniform Distribution
David Dubowski
kindle edition

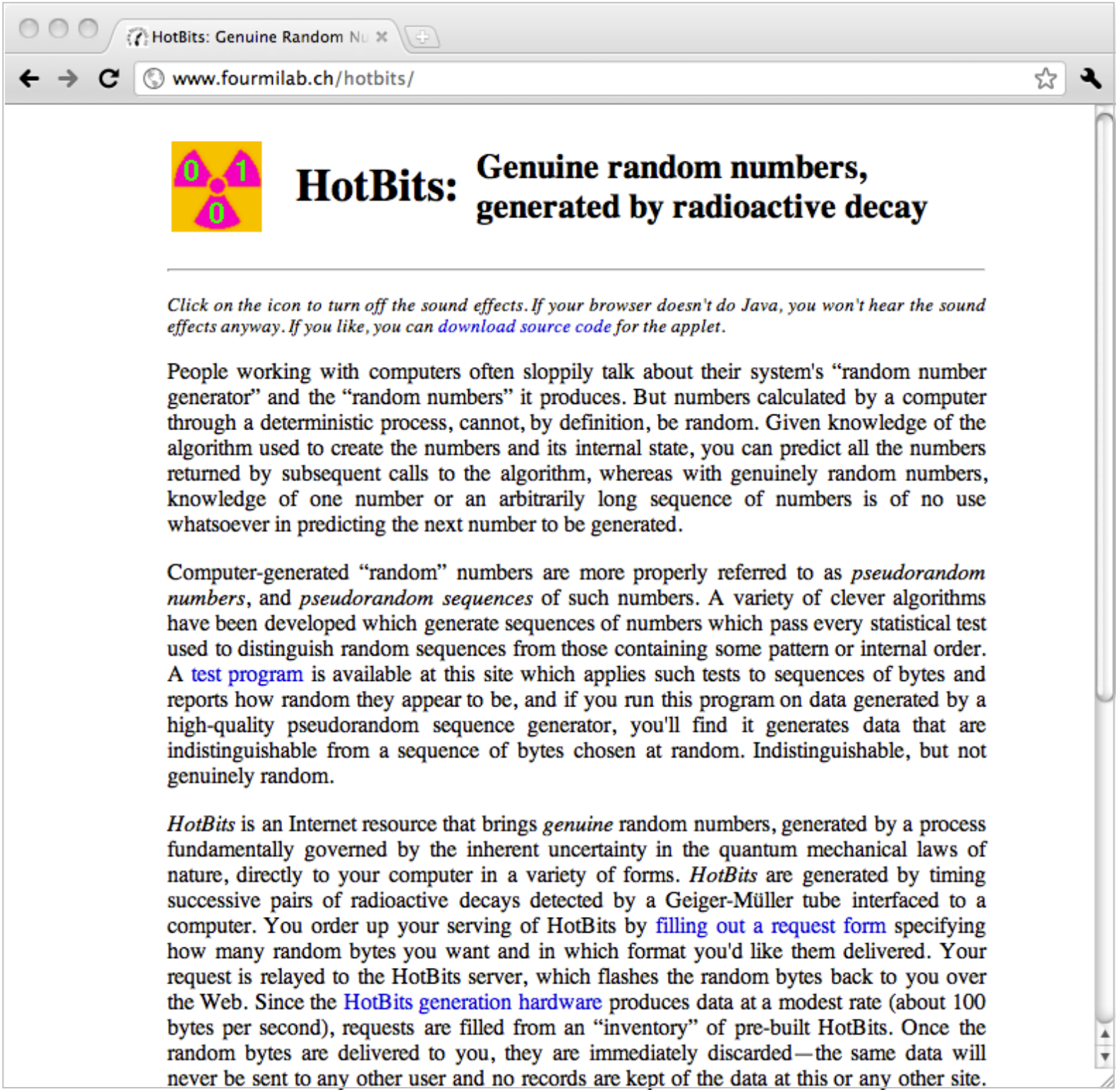
Random number generation

These days, there are two dominant techniques for generating random numbers

One is not really random according to any romantic notions of the word and are the result of **a mathematical formula** which is entirely predictable and repeatable -- These are often called **pseduo-random numbers**

The second, on the other hand, is often touted as **“true” random numbers** and are generated by **observing some physical process** -- You can think of a small coin-tossing device attached to your computer although the physical phenomena used tend to be more exotic

Let’s have a look at both, starting with the latter as it will give us to talk about data and the publication of data (a theme for this course)

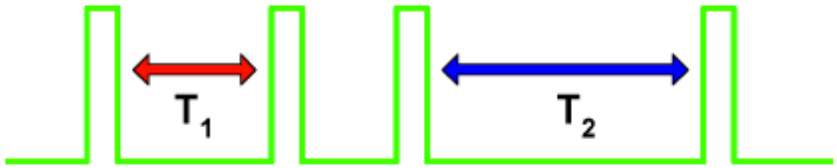

HotBits: Genuine Random Nu
www.fourmilab.ch/hotbits/
0
1
0
HotBits: Genuine random numbers, generated by radioactive decay
Click on the icon to turn off the sound effects. If your browser doesn't do Java, you won't hear the sound effects anyway. If you like, you can download source code for the applet.
People working with computers often sloppily talk about their system's "random number generator" and the "random numbers" it produces. But numbers calculated by a computer through a deterministic process, cannot, by definition, be random. Given knowledge of the algorithm used to create the numbers and its internal state, you can predict all the numbers returned by subsequent calls to the algorithm, whereas with genuinely random numbers, knowledge of one number or an arbitrarily long sequence of numbers is of no use whatsoever in predicting the next number to be generated.
Computer-generated "random" numbers are more properly referred to as pseudorandom numbers, and pseudorandom sequences of such numbers. A variety of clever algorithms have been developed which generate sequences of numbers which pass every statistical test used to distinguish random sequences from those containing some pattern or internal order. A test program is available at this site which applies such tests to sequences of bytes and reports how random they appear to be, and if you run this program on data generated by a high-quality pseudorandom sequence generator, you'll find it generates data that are indistinguishable from a sequence of bytes chosen at random. Indistinguishable, but not genuinely random.
HotBits is an Internet resource that brings genuine random numbers, generated by a process fundamentally governed by the inherent uncertainty in the quantum mechanical laws of nature, directly to your computer in a variety of forms. HotBits are generated by timing successive pairs of radioactive decays detected by a Geiger-Müller tube interfaced to a computer. You order up your serving of HotBits by filling out a request form specifying how many random bytes you want and in which format you'd like them delivered. Your request is relayed to the HotBits server, which flashes the random bytes back to you over the Web. Since the HotBits generation hardware produces data at a modest rate (about 100 bytes per second), requests are filled from an "inventory" of pre-built HotBits. Once the random bytes are delivered to you, they are immediately discarded—the same data will never be sent to any other user and no records are kept of the data at this or any other site.

Bit from It

This inherent randomness in decay time has profound implications, which we will now exploit to generate random numbers—HotBits. For if there's no way to know when a given Cæsium-137 nucleus will decay then, given an collection of them, there's no way to know when the *next* one of them will shoot its electron bolt and settle down to a serene eternity as Barium. That's uncertainty, with its origins in the deepest and darkest corners of creation—precisely what we're looking for to make genuinely random numbers.

If we knew the precise half-life of the radioactive source driving our detector (and other details such as the solid angle to which our detector is sensitive, the energy range of decay products and the sensitivity of the detector to them, and so on), we could generate random bits by measuring whether the time between a pair of beta decays was more or less than the time expected based on the half-life. But that would require our knowing the average beta decay detection time, which depends on a large number of parameters which can only be determined experimentally. Instead, we can exploit the inherent uncertainty of decay time in a parameter-free fashion which requires less arm waving and fancy footwork.

The trick I use was dreamed up in a conversation in 1985 with John Nagle, who is doing some fascinating things these days with artificial animals. Since the time of any given decay is random, then the *interval* between two consecutive decays is also random. What we do, then, is measure a pair of these intervals, and emit a zero or one bit based on the relative length of the two intervals. If we measure the same interval for the two decays, we discard the measurement and try again, to avoid the risk of inducing bias due to the resolution of our clock.



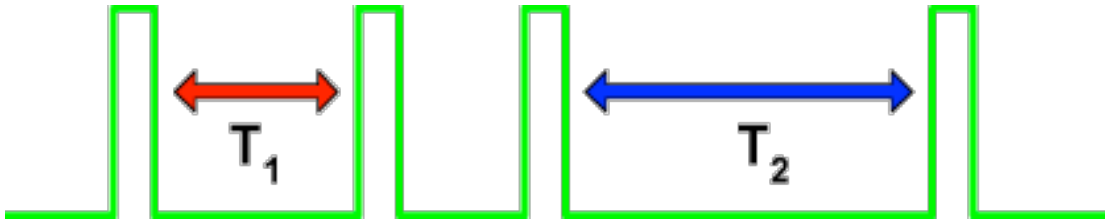
To create each random bit, we wait until the first count occurs, then measure the time, T_1, until the next. We then wait for a second pair of pulses and measure the interval T_2 between them, yielding a pair of durations. If they're the same, we throw away the measurement and try again. Otherwise if T_1 is less than T_2 we emit a zero bit; if T_1 is greater than T_2, a one bit. In practice, to avoid any residual bias resulting from non-

Bits

A “bit” stands for a “binary digit” and takes on the value 0 or 1 -- You can think of it as a coin toss where we map “heads” to 1, say, and “tails” to 0 (The term “bit” was actually coined by our man John Tukey; he also came up with the term “software”)

HotBits uses radioactive decay as a means for generating physically random or “true” random bits (coin tosses) -- **You can imagine listening to a Geiger counter and group the ticks into pairs**

If $T_1 > T_2$ they produce a 1, otherwise they generate a 0 -- **Voila! “true” random numbers!**



Bits

Another project, this run through the web site `random.org` uses atmospheric noise, suitably filtered, to accomplish the same task

Our interest in this site was because of the mechanism they used to “publish” their random bits, **offering a service that scientists around the world** -- They have implemented **a public API (application programming interface)** that let’s programs (like R) request data (random bits)


RANDOM.ORG – True Rando
www.random.org
Home Games Numbers Lists & More Drawings Web Tools Statistics Testimonials Learn More Login
RANDOM.ORG
Search RANDOM.ORG
Search
True Random Number Service
What's this fuss about true randomness?
Perhaps you have wondered how predictable machines like computers can generate randomness. In reality, most random numbers used in computer programs are pseudo-random, which means they are a generated in a predictable fashion using a mathematical formula. This is fine for many purposes, but it may not be random in the way you expect if you're used to dice rolls and lottery drawings.
RANDOM.ORG offers true random numbers to anyone on the Internet. The randomness comes from atmospheric noise, which for many purposes is better than the pseudo-random number algorithms typically used in computer programs. People use RANDOM.ORG for holding drawings, lotteries and sweepstakes, to drive games and gambling sites, for scientific applications and for art and music. The service has existed since 1998 and was built and is being operated by Mads Haahr of the School of Computer Science and Statistics at Trinity College, Dublin in Ireland.
As of today, RANDOM.ORG has generated 958.5 billion random bits for the Internet community.
True Random Number Generator
Min: 1
Max: 100
Generate
Result:
Powered by RANDOM.ORG
Like RANDOM.ORG?
Get the Newsletter
FREE services
Games and Gambling
Lottery Quick Pick is perhaps the Internet's most popular with over 130 lotteries
Keno Quick Pick for the popular game played at many casinos
Coin Flipper will give you heads or tails in many currencies
Dice Roller does exactly what it says on the tin
Playing Card Shuffler will draw cards from multiple shuffled decks
Birdie Fund Generator will create birdie holes for golf courses
PAID service
Random Drawings
Q3.1 in the FAQ explains how to pick a winner for your giveaway for FREE
Third-Party Draw Service is the premier solution to holding random drawings online
Step by Step Guide explains how to hold a drawing with the Third-Party Draw Service
Step by Step Video shows how to hold a drawing with the Third-Party Draw Service
Price Calculator tells exactly how much your drawing will cost
Drawing FAQ answers common questions about holding drawings

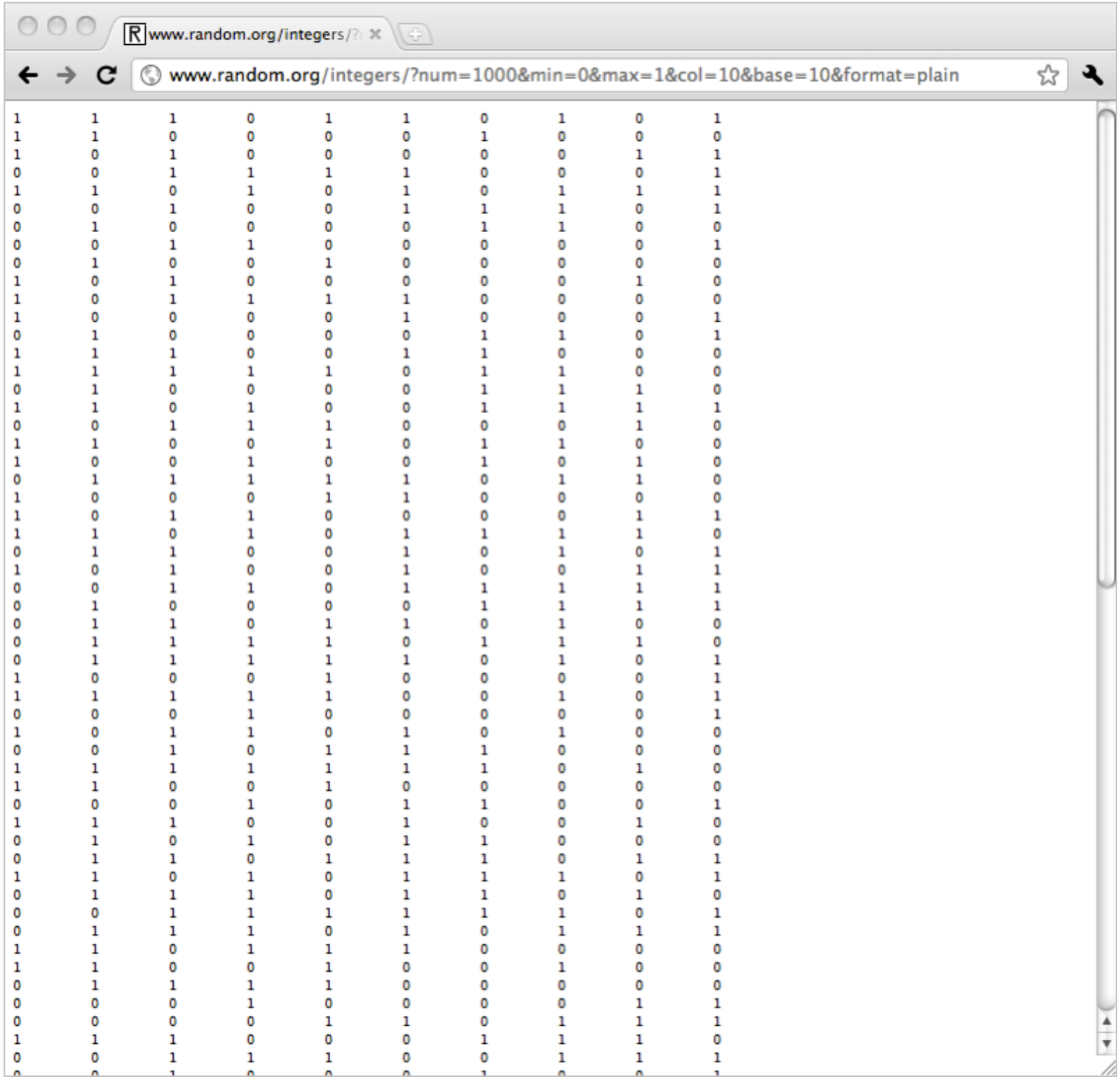

RANDOM.ORG – HTTP Interf
www.random.org/clients/http/
Home Games Numbers Lists & More Drawings Web Tools Statistics Testimonials Learn More Login
RANDOM.ORG
Search RANDOM.ORG
Search
True Random Number Service
HTTP Interface Description
RANDOM.ORG is a true random number service that generates randomness via atmospheric noise. This page explains how to interface to the service via the Hyper-Text Transfer Protocol (HTTP). There is also the HTTP Client Archive, which contains clients that other people have written.
Important note!
If you access RANDOM.ORG via an automated client, please make sure you observe the Guidelines for Automated Clients or your computer may be banned.
If you are writing a general-purpose client, please make sure it is easy for your users to run it in accordance with the guidelines.
This page contains documentation for the Integer Generator, the Sequence Generator, the String Generator and the Quota Checker, which allows you to examine your current bit allowance.
All the interfaces on this page return HTTP status code 503 (Service Unavailable) in the case of errors and code 200 (OK) when successful. Not all languages allow you to access the HTTP status codes in a straightforward manner. A reasonable workaround is to look for the string "Error:" (don't forget the colon) as the first line of the response. This will work for all the generators on this page, including the String Generator (which could by chance produce the string "Error" in a successful response, but which cannot produce the colon character).
Please note that the old CGI scripts (randbyte, randnum, etc.) are no longer supported and you should use the ones described below instead. In particular, the old scripts do not return the 503 status code in case of errrors (they return the 200 response code in all cases), so please use the new ones instead.
Integer Generator
The Integer Generator will generate truly random integers in configurable intervals. It is pretty easy to write a client to access the integer generator. The integer generator accepts only HTTP GET requests, so parameters are passed via encoding in the URL.
Parameters
Name
Possible Values
Description

Integer Generator

The Integer Generator will generate truly random integers in configurable intervals. It is pretty easy to write a client to access the integer generator. The integer generator accepts only HTTP GET requests, so parameters are passed via encoding in the URL.

Parameters

Name	Possible Values	Description
num	[1,1e4]	The number of integers requested.
min	[-1e9,1e9]	The smallest value allowed for each integer.
max	[-1e9,1e9]	The largest value allowed for each integer.
col	[1,1e9]	The number of columns in which the integers will be arranged. The integers should be read (or processed) left to right across columns.
base	**2** | **8** | **10** | **16**	The base that will be used to print the numbers, i.e., binary, octal, decimal or hexadecimal.
format	**html** | **plain**	Determines the return type of the document that the server produces as its response. If **html** is specified, the server produces a nicely formatted XHTML document (MIME type text/html), which will display well in a browser but which is somewhat cumbersome to parse. If **plain** is specified, the server produces as minimalistic document of type plain text (MIME type text/plain) document, which is easy to parse. If you are writing an automated client, you probably want to specify **plain** here.
rnd	**new** | **id.identifier** | **date.iso-date**	Determines the randomization to use to generate the numbers. If **new** is specified, then a new randomization will created from the truly random bitstream at RANDOM.ORG. This is probably what you want in most cases. If **id.identifier** is specified, the identifier is used to determine the randomization in a deterministic fashion from a large pool of pregenerated random bits. Because the numbers are produced in a deterministic fashion, specifying an id basically uses RANDOM.ORG as a pseudo-random number generator. The third (**date.iso-date**) form is similar to the second; it allows the randomization to be based on one of the daily pregenerated files. This form must refer to one of the dates for which files exist, so it must be the current day (according to UTC) or a day in the past. The date must be in ISO 8601 format (i.e., YYYY-MM-DD) or one of the two shorthand strings **today** or **yesterday**.



1 1 1 0 1 1 0 1 0 1
1 1 0 0 0 0 1 0 0 0
1 0 1 0 0 0 0 0 1 1
0 0 1 1 1 1 0 0 0 1
1 1 0 1 0 1 0 1 1 1
0 0 1 0 0 1 1 1 0 1
0 1 0 0 0 0 1 1 0 0
0 0 1 1 0 0 0 0 0 1
0 1 0 0 1 0 0 0 0 0
1 0 1 0 0 0 0 0 1 0
1 0 1 1 1 1 0 0 0 0
1 0 0 0 0 1 0 0 0 1
0 1 0 0 0 0 1 1 0 1
1 1 1 0 0 1 1 0 0 0
1 1 1 1 1 0 1 1 0 0
0 1 0 0 0 0 1 1 1 0
1 1 0 1 0 0 1 1 1 1
0 0 1 1 1 0 0 0 1 0
1 1 0 0 1 0 1 1 0 0
1 0 0 1 0 0 1 0 1 0
0 1 1 1 1 1 0 1 1 0
1 0 0 0 1 1 0 0 0 0
1 0 1 1 0 0 0 0 1 1
1 1 0 1 0 1 1 1 1 0
0 1 1 0 0 1 0 1 0 1
1 0 1 0 0 1 0 0 1 1
0 0 1 1 0 1 1 1 1 1
0 1 0 0 0 0 1 1 1 1
0 1 1 0 1 1 0 1 0 0
0 1 1 1 1 0 1 1 1 0
0 1 1 1 1 1 0 1 0 1
1 0 0 0 1 0 0 0 0 1
1 1 1 1 1 0 0 1 0 1
0 0 0 1 0 0 0 0 0 1
1 0 1 1 0 1 0 1 0 0
0 0 1 0 1 1 1 0 0 0
1 1 1 1 1 1 1 0 1 0
1 1 0 0 1 0 0 0 0 0
0 0 0 1 0 1 1 0 0 1
1 1 1 0 0 1 0 0 1 0
0 1 0 1 0 1 1 0 0 0
0 1 1 0 1 1 1 1 0 1
1 1 0 1 0 1 1 1 0 1
0 1 1 1 0 1 1 1 0 1
0 0 1 1 1 1 1 1 0 1
0 1 1 1 0 1 0 1 1 1
1 1 0 1 1 1 0 0 0 0
1 1 0 0 1 0 0 1 0 0
0 1 1 1 1 0 0 0 0 0
0 0 0 1 0 0 0 0 1 1
0 0 0 0 1 1 0 1 1 1
1 1 1 0 0 0 1 1 1 0
0 0 1 1 1 0 0 1 1 1

APIs

I mention this because many organizations offer their data via web services like this -- Specific HTTP requests yield not web pages designed for human consumption (reading), but instead **structured data that can be processed easily in an autonomous way by computers**

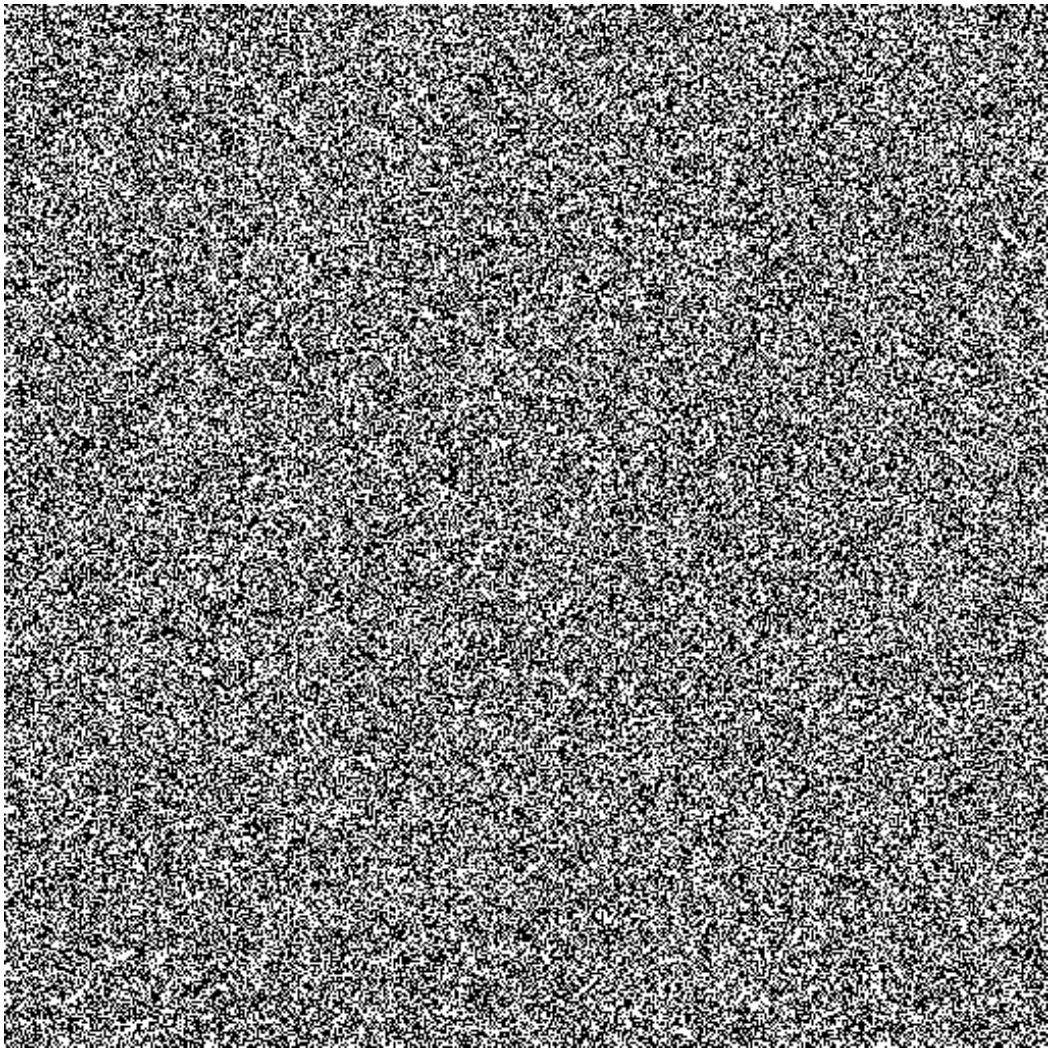
In this way, services build “mash-up” style, with data flowing easily between different organizations -- This is really an amazing development that opens up incredible possibilities for statistics, for computer science, for machine learning for the data geeks out there!

Testing?

Just because a service advertises random bits (and they have a good story to go along with it) doesn't mean that it works -- To get a little technical, **we probably shouldn't think about a random number in isolation (is 1 random?) but instead talk about a sequence of random numbers**

Even this is a little vague -- What properties would we expect from a sequence of random bits (random coin tosses)? Intuitively, **what do you expect to see as you look across and down the web page on the previous slide?**

On the next slide we mapped each 1 to a black pixel and each 0 to a white one -- We then asked for 512*512 = 262,144 random bits from `random.org` and displayed them as a 512 by 512 image (putting the first bit in the upper lefthand corner, the second bit just below it and so on, filling up each column one at a time from left to right)



Testing?

Formally, there are a set of **classical statistical tests (yes, tests of hypothesis!) that could help us assess if a random number generator (true or otherwise) is performing as expected**

In this case, we can use the mathematics of probability to determine the null distribution for test statistics like **the fraction of 1's in the sequence (it should be about a half) or the length of "runs" of bits of the same kind (we shouldn't see long runs of 1s or 0s)**

These mathematical results help us avoid **a chicken and egg problem** -- If we needed simulation to test a null hypothesis, how could we ever test the simulator?!

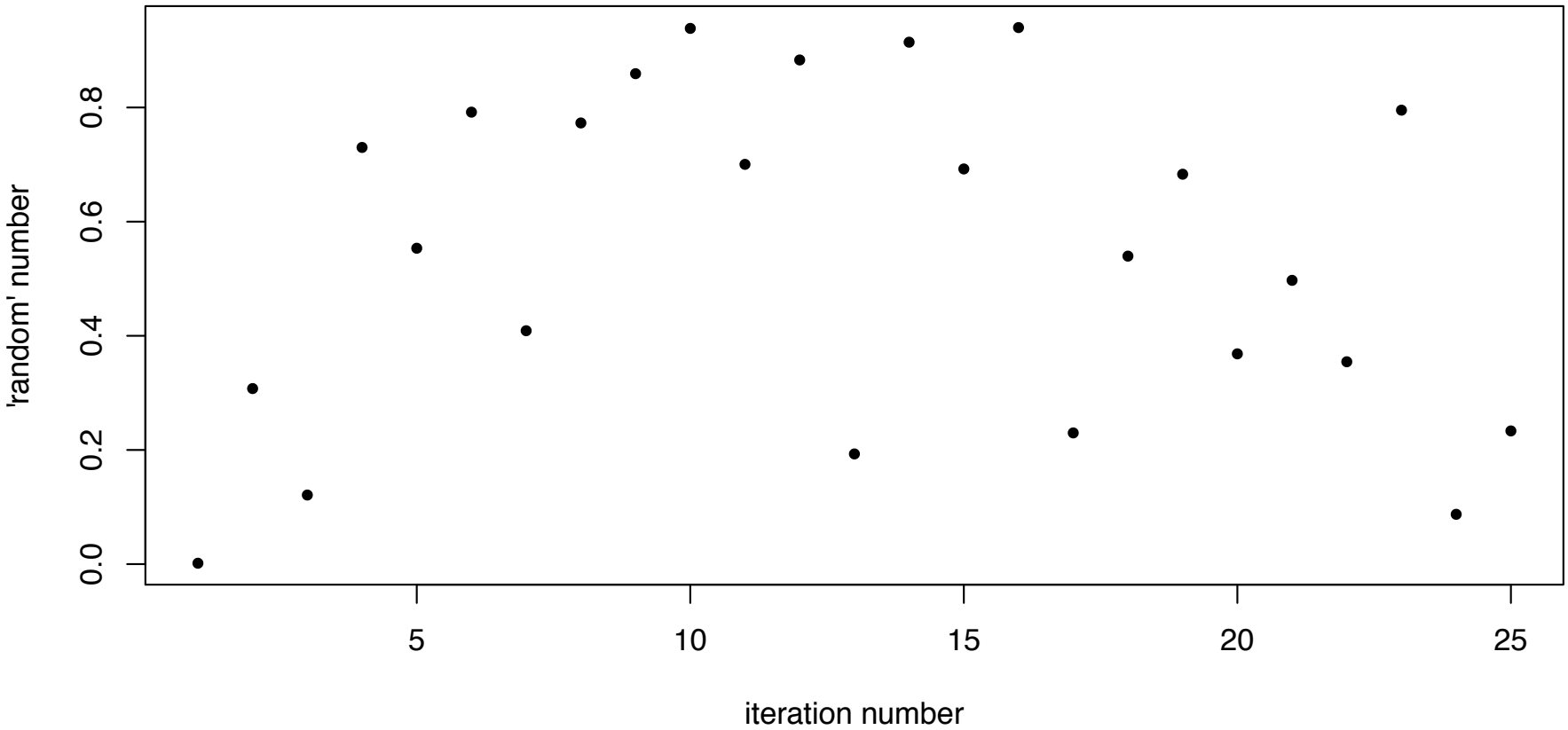
Random number generation

The second kind of random number generation comes from **a mathematical formula, a deterministic algorithm** that produces a repeatable, predictable series of numbers -- These are called **pseudo-random numbers**

Here is a snippet of code that implements a “classical” example of one of these algorithms -- We start by setting the variable `seed` to some number we choose (here I picked 200)

```
# initialize

> seed <- 200

# we then update the seed and generate
# a “random” number

> a <- 16807
> m <- 2147483647
> seed <- (a*seed)%%m
> random <- seed/m

# the values random will be in the
# interval (0,1)
```

```
# iteration 1
> seed <- 200
> (a*seed)
[1] 3361400
> (a*seed)%%m
[1] 3361400
> ((a*seed)%%m)/m
[1] 0.00156527385188512 2

# iteration 2
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 56495049800
> (a*seed)%%m
[1] 660474978
> ((a*seed)%%m)/m
[1] 0.3075576286332484

# iteration 3
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 11100602955246
> (a*seed)%%m
[1] 259983903
> ((a*seed)%%m)/m
[1] 0.1210644390066454

# iteration 4
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 4369549457721
> (a*seed)%%m
[1] 1567719723
> ((a*seed)%%m)/m
[1] 0.7300263846898575

# iteration 5
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 26348665384461
> (a*seed)%%m
[1] 1188519418
> ((a*seed)%%m)/m
[1] 0.5534474824338 44

# iteration 6
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 19975445858326
> (a*seed)%%m
[1] 1700457579
> ((a*seed)%%m)/m
[1] 0.7918372656180697

# iteration 7
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 28579590530253
> (a*seed)%%m
[1] 878155977
> ((a*seed)%%m)/m
[1] 0.408923242897225

# iteration 8
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 14759167505439
> (a*seed)%%m
[1] 1659883255
> ((a*seed)%%m)/m
[1] 0.77294337366379

# iteration 9
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 27897657866785
> (a*seed)%%m
[1] 1845292255
> ((a*seed)%%m)/m
[1] 0.859281167322435

# iteration 10
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 31013826929785
> (a*seed)%%m
[1] 2015583458
> ((a*seed)%%m)/m
[1] 0.938579188165525

# iteration 11
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 33875911178606
> (a*seed)%%m
[1] 1504130828
> ((a*seed)%%m)/m
[1] 0.700415497971892

# iteration 12
> seed <- ((a*seed)%%m)
> (a*seed)
[1] 25279926826196
> (a*seed)%%m
[1] 1896817359
> ((a*seed)%%m)/m
[1] 0.883274413590913
```


'random' number
0.0
0.2
0.4
0.6
0.8
5
10
15
20
25
iteration number

Uniform random numbers

This procedure is also known by the mouthful of a name **“prime modulus multiplicative linear congruential generator”** (and often shortened to the equally difficult PMMLCG)

Technically the algorithm leaves the constants (`a` and `m`) unspecified, but our choice can be shown to have good properties relative to the statistical tests I alluded to

By construction, the numbers we highlighted in red are all between 0 and 1 -- They can be used anywhere we might need observations from the so-called **uniform distribution** on the interval [0,1]

Uniform random numbers

As its name suggests, we expect to see observations from the uniform distribution distributed, well, uniformly over [0,1] -- To be a little more precise, if we have a sample of 1200 such observations, we'd expect about 600 to be less than 0.5

Going farther, if we divided [0,1] into four equally sized subintervals (from 0 to 0.25, 0.25 to 0.5, 0.5 to 0.75 and 0.75 to 1) we would expect to see 300 observations of the 1200 in each bin (or so)

In general, under the uniform distribution, we expect **the proportion of our sample that falls in some subinterval we specify to be equal to the length of that interval** (the uniform distribution is a mathematical construction that we will examine more closely when we discuss probability in a future lecture)

So, using the algorithm two slides back, let's create some random bits -- We'll generate 512*512 = 262,144 numbers in [0,1] using this algorithm and **coloring a square black if the associated number is larger than 0.5 and color it white if it is less than or equal to 0.5**...

Uniform random numbers

Here is the sequence of seed values we get

```
 [1]    3361400  660474978  259983903 1567719723 1188519418 1700457579
 [7]  878155977 1659883255 1845292255 2015583458 1504130828 1896817359
[13]  414612998 1963706518 1486760930 2018717665  493656702 1158862153
[19] 1467010828  791234989 1067717899  761374161 1707955101  187473058
[25]  501175657
```

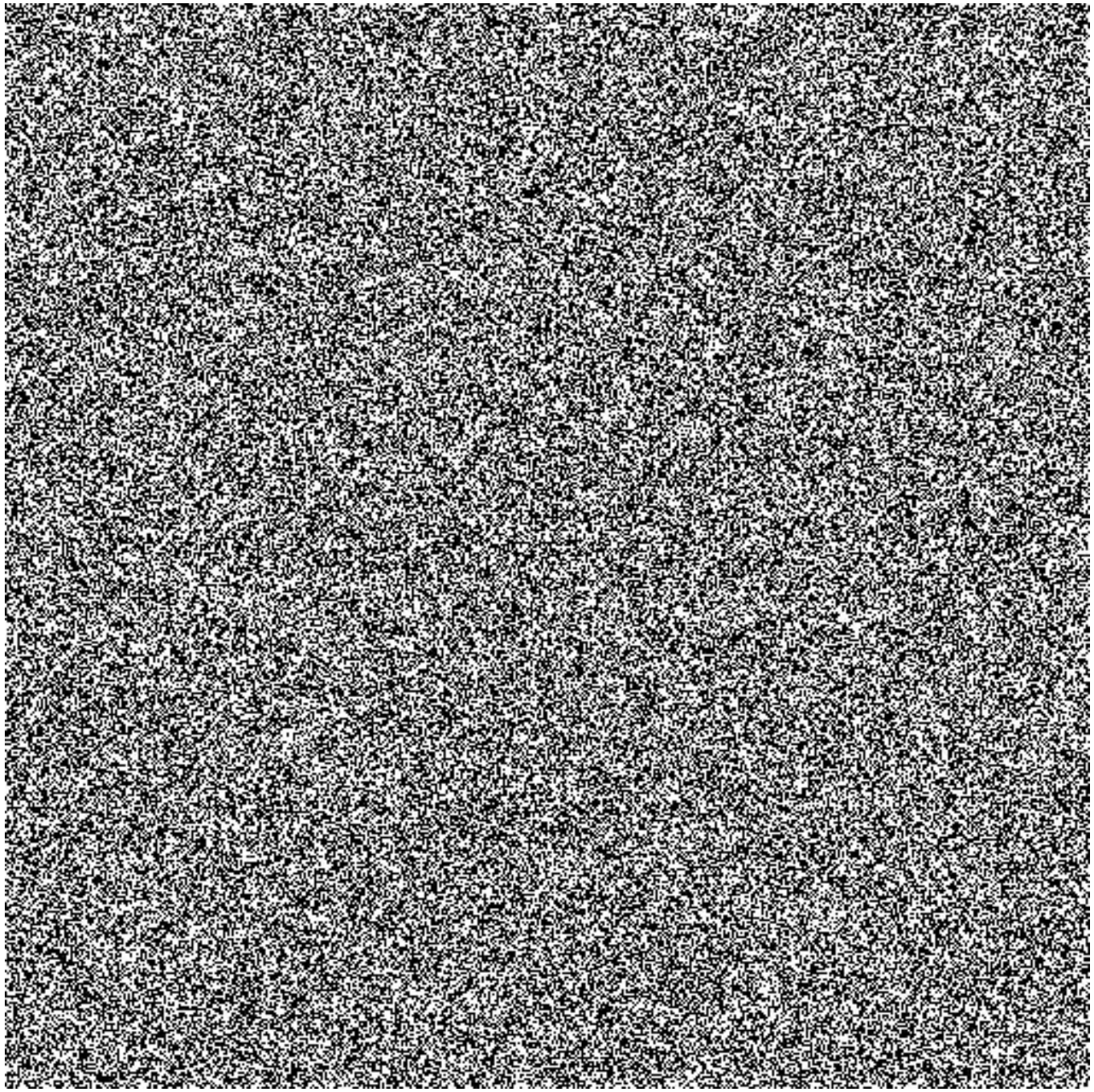
which when divided by `2147483647 = 2^31-1` gives the pseudo-random uniform observations

```
 [1] 0.001565273851885122 0.307557628633248425 0.121064439006645430
 [4] 0.730026384689857477 0.553447482433844118 0.791837265618069663
 [7] 0.408923242897225203 0.772943373663790179 0.859281167322434980
[10] 0.938579188165524547 0.700415497971892176 0.883274413590912855
[13] 0.193069222473105984 0.914422105492289194 0.692327008905041508
[16] 0.940038667032513153 0.229876815448457755 0.539637242229486502
[19] 0.683130150978979778 0.368447503712236668 0.497194891561379138
[22] 0.354542472099206640 0.795328571365833570 0.087298945564450212
[25] 0.233378101714597136
```

and the random bits

```
 [1] 0 0 0 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 0 0 0 1 0 0
```

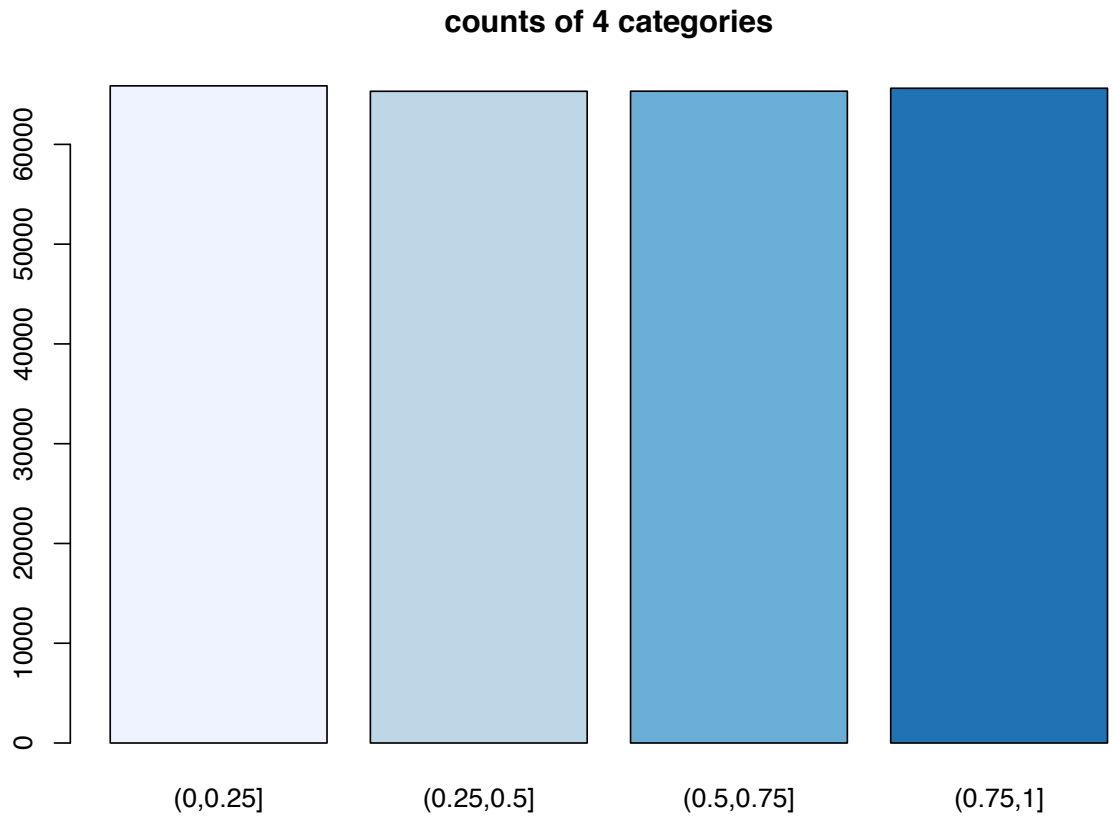
which, when continued for the full 512*512 values and arranged in an image gives us...

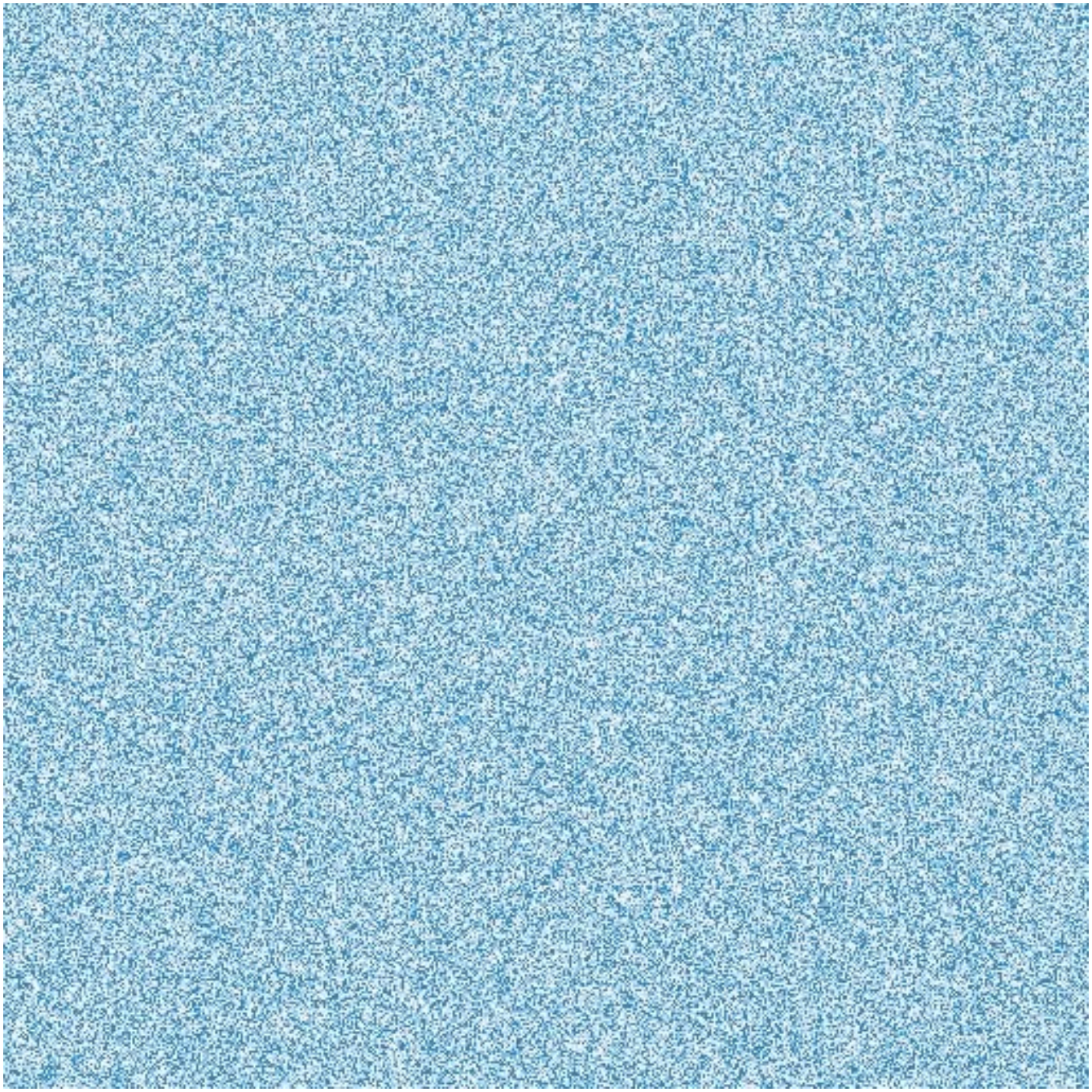


Uniform random numbers

The proportion of 1's constructed this way is 0.49955 -- We can take this farther and **divide the data into four pieces** (those between 0 and 0.25, between 0.25 and 0.5, between 0.5 and 0.75 and then larger than 0.75) and below we have a barplot of counts in each interval (and yes, it looks essentially flat)

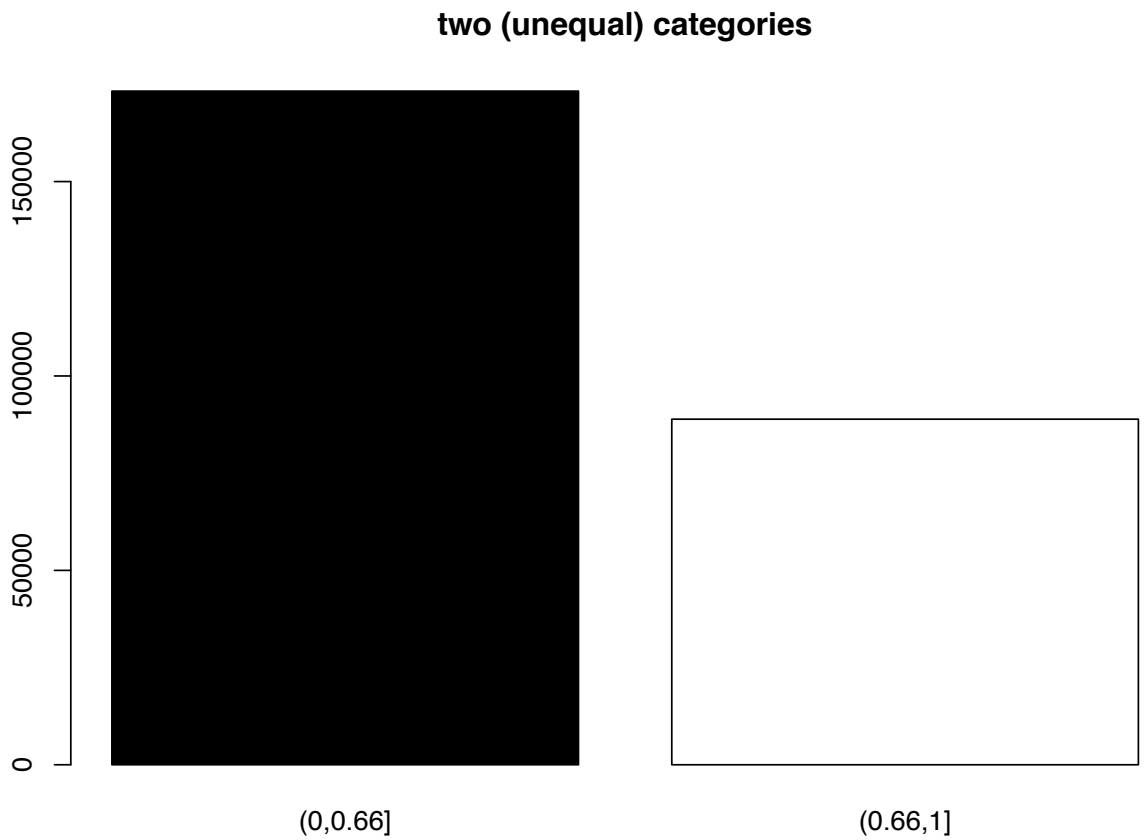
Essentially, an equal number of points falls in each interval -- We can color points falling into the four pieces using different colors (the colors of the four bars on the right) and pack them into an image again

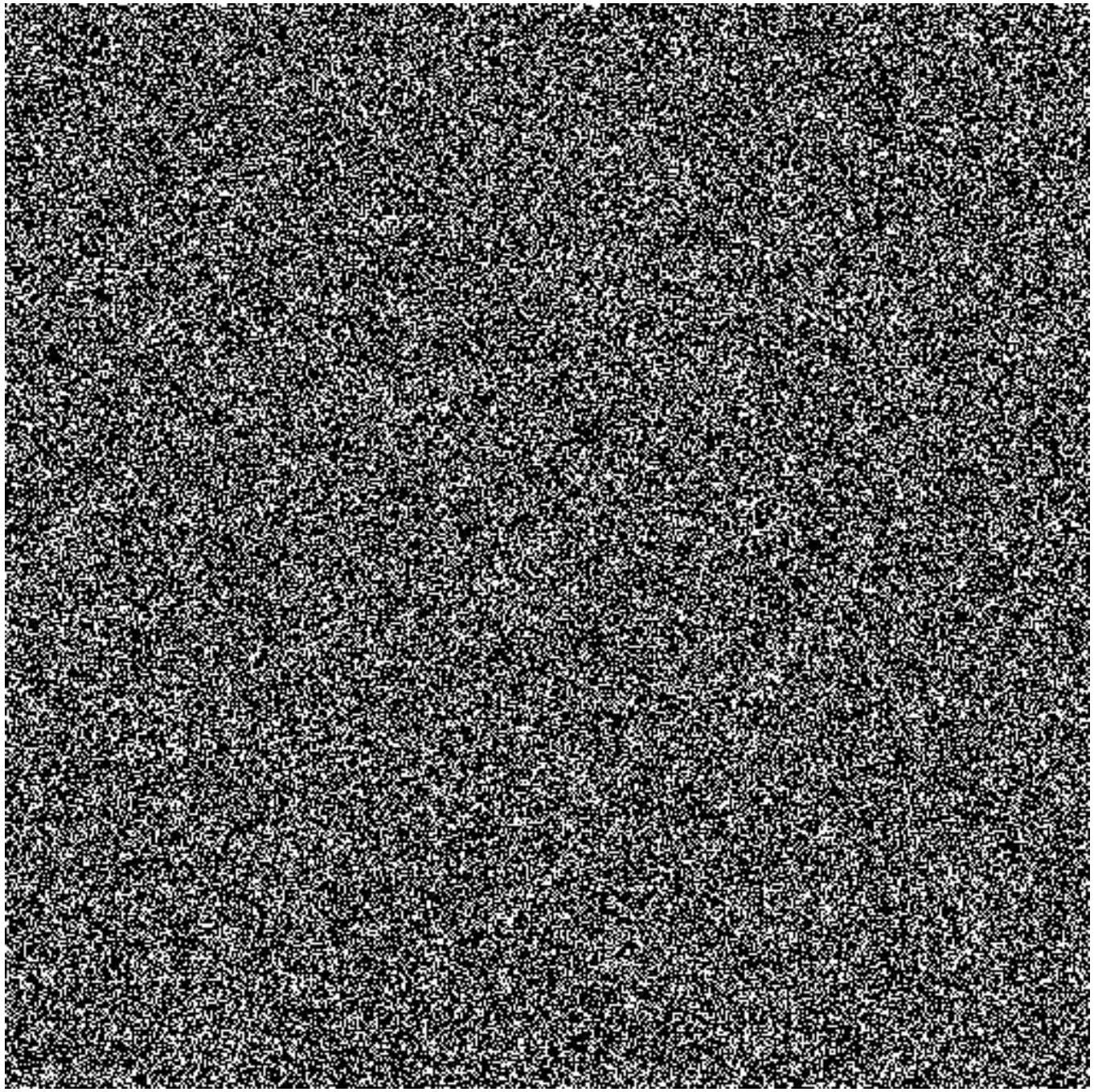




Uniform random numbers

Dividing the data using the intervals [0,0.66) and [0.66,1], we see essentially 2/3 of the data fall into the first category and 1/3 in the second -- On the next slide we play the image game again





To sum

There are **deterministic mathematical algorithms that allow us to generate observations that have many of the characteristics we'd expect to see from truly random data** (where our expectations are set by statistical tests like proportions falling into intervals or "runs" of particular kinds)

Starting with the uniform distribution, we can simulate a host of random phenomena, from permuting our data as in the case of our re-randomization analysis to tossing coins (a la Arbuthnot) to making observations from the normal (bell-shaped) distribution (later)

This idea might take some getting used to, but **the use of pseudo-random numbers is common practice** and even R depends on one such algorithm (the default is the so-called Mersenne-Twister)

Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.

John von Neumann

Some advantages

While pseudo-random numbers are entirely deterministic, there are some advantages for scientific uses

Chief among them is reproducibility! If our analysis depends on simulation (like our re-randomization procedures for inference) we would like to be able to reproduce our results exactly (this comes in hand, say, when you want to debug more complex algorithms)

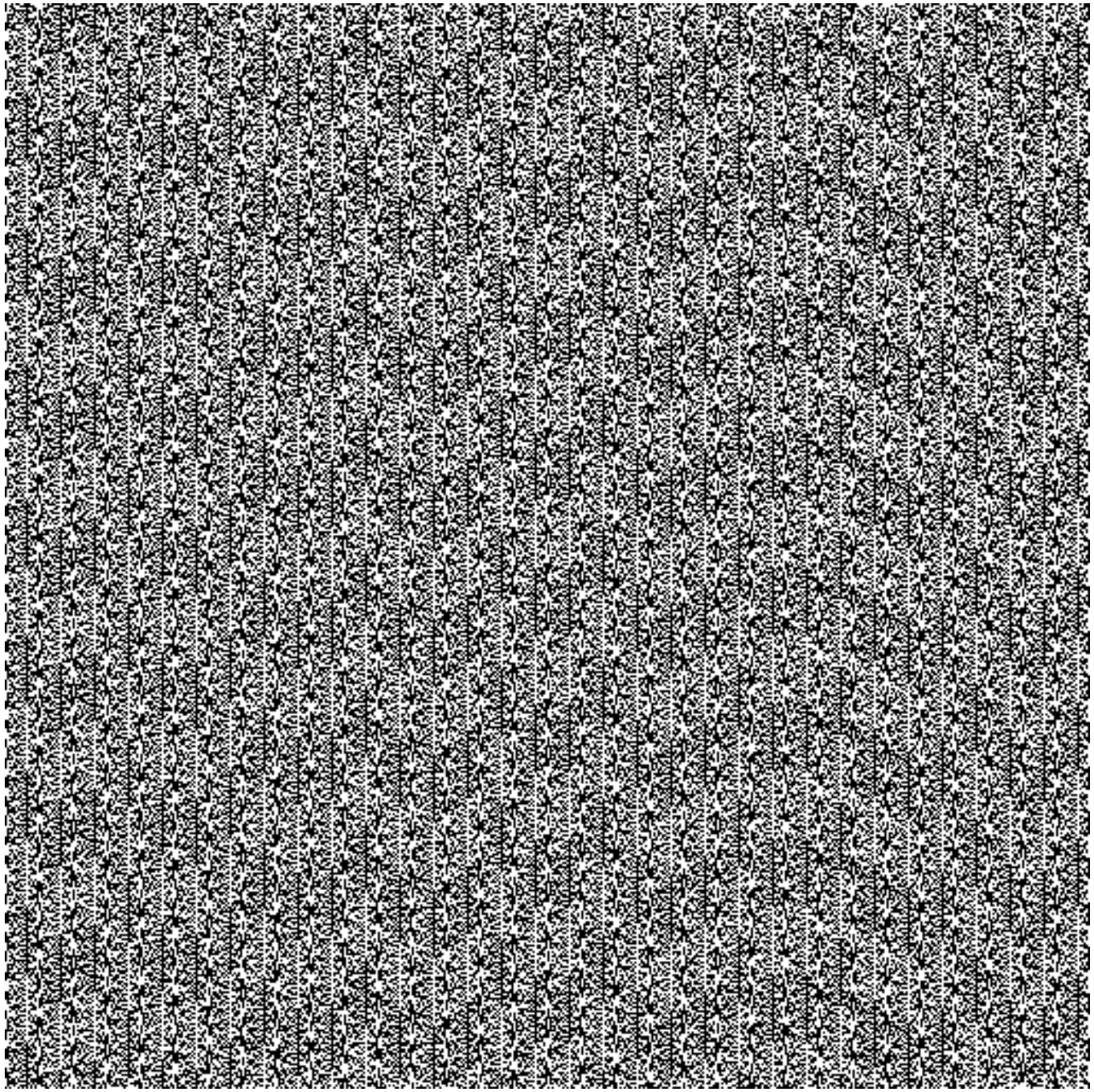
In R you can use the function `set.seed()` at any point to reset your sequence of random numbers (R does not use the algorithm described here, but it shares the properties of being a mathematical formula, deterministic and predictable)

```
> treat
 [1] "T" "T" "T" "T" "T" "C" "C" "C" "C" "C"

# treat holds a division into treatment and
# control -- let's use sample() to permute
# them or re-randomize

> set.seed(1000)
> sample(treat)
 [1] "T" "C" "T" "T" "C" "C" "T" "T" "C" "C"
> sample(treat)
 [1] "T" "C" "T" "C" "T" "T" "T" "C" "C" "C"
> sample(treat)
 [1] "T" "C" "T" "T" "T" "C" "C" "C" "C" "T"

# we can repeat our rerandomizations by
# resetting the seed

> set.seed(1000)
> sample(treat)
 [1] "T" "C" "T" "T" "C" "C" "T" "T" "C" "C"
> sample(treat)
 [1] "T" "C" "T" "C" "T" "T" "T" "C" "C" "C"
> sample(treat)
 [1] "T" "C" "T" "T" "T" "C" "C" "C" "C" "T"
```

Testing?

One final note -- **Not every service or system or program that advertises random numbers is any good!**

Here is the same bit picture for a combination of the programming language **PHP running on a Windows machine** -- Evidently this was the result of a bug that has since been corrected, but it does serve as a cautionary example

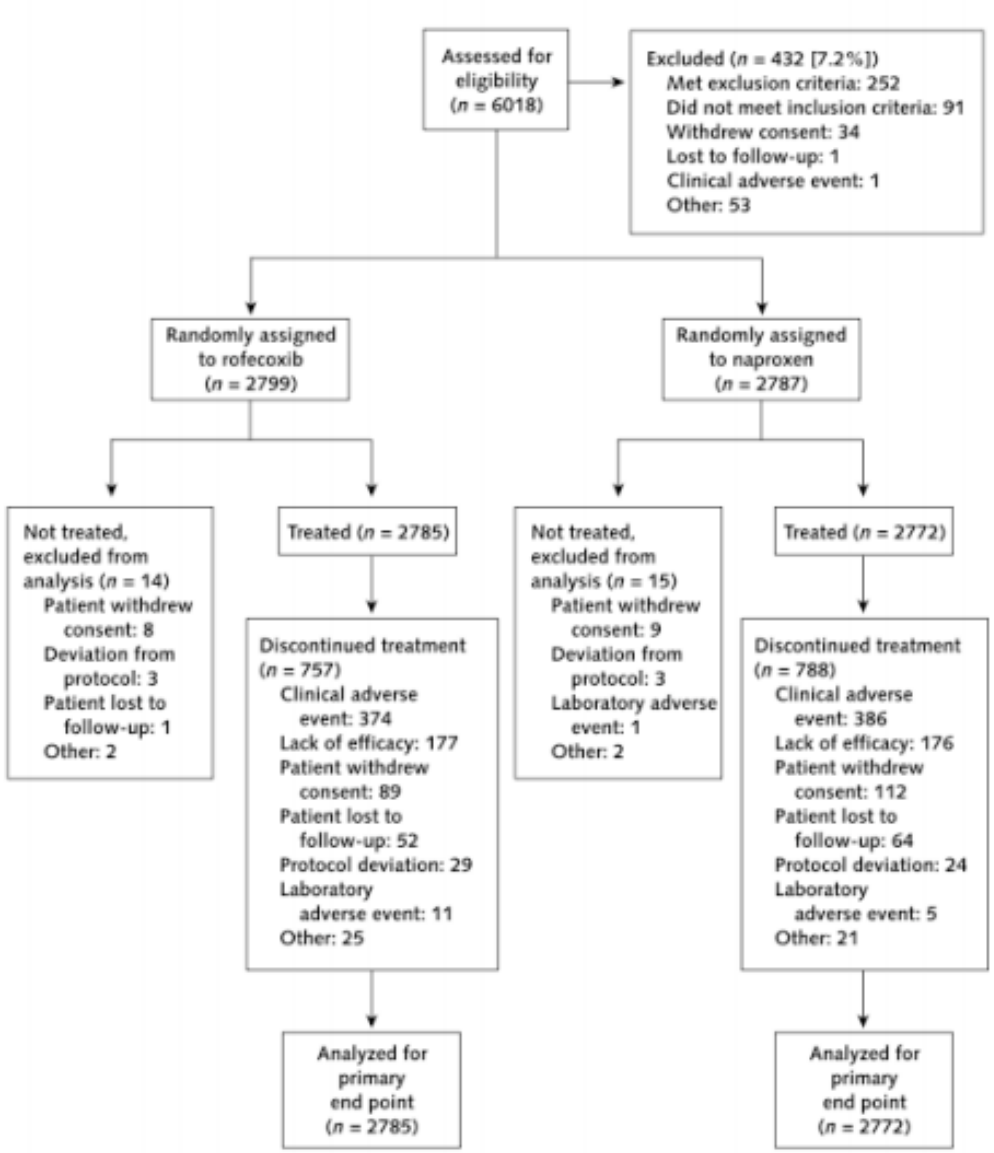


Other randomized experiments

So far, we have examined mainly experimental setups in health studies; the randomized controlled trial is a big deal and nicely highlights many of the features of hypothesis testing

But randomization in experimentation is extremely common; as we mentioned previously, Fisher was an aggressive advocate of its use in general

We'll now consider a more modern application of randomization, in experiments that "optimize" the operation of web sites

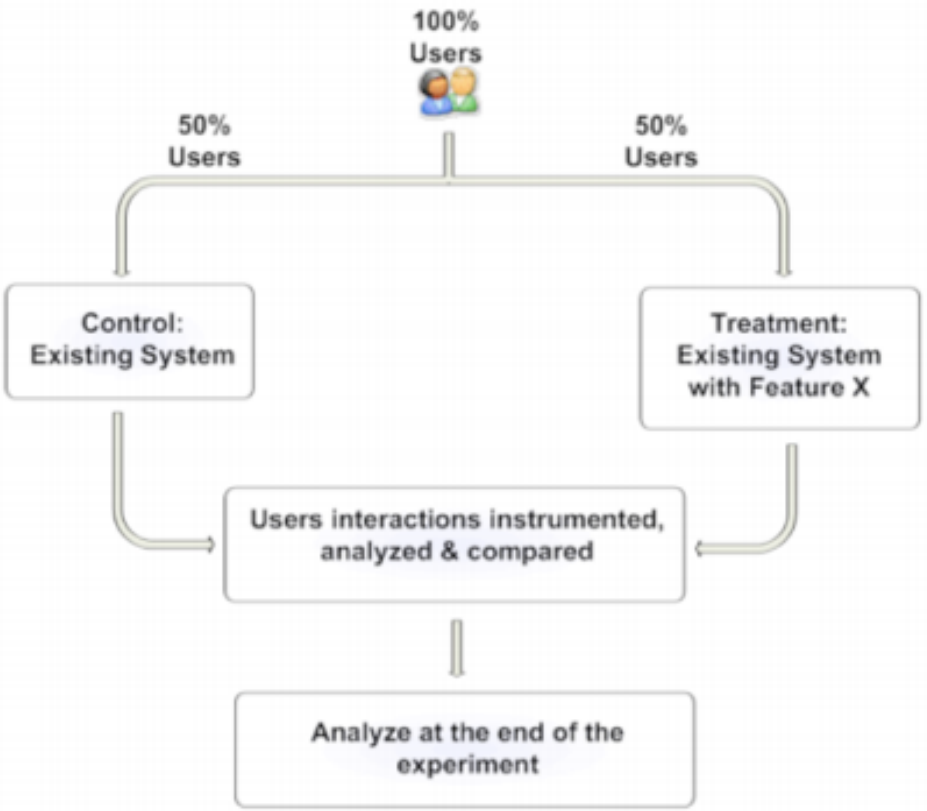


A/B testing

At the right we present a diagram for conducting an experiment on a web site; the structure is not that dissimilar from that for the ADVANTAGE trial (minus all the comments about “exclusions” and “discontinued” treatments

In an A/B test, visitors are presented with two different versions of the site (in some cases we literally have a treatment and control); sometimes the differences have to do with **the content on the page**, while in other cases they have to do with **the location of or visual arrangement of the content**

The outcomes that are measured depend on the site’s overall objectives; an informational site might be interested in **the number of pages visitors read**, while commerce sites would be concerned with the **number of transactions made**

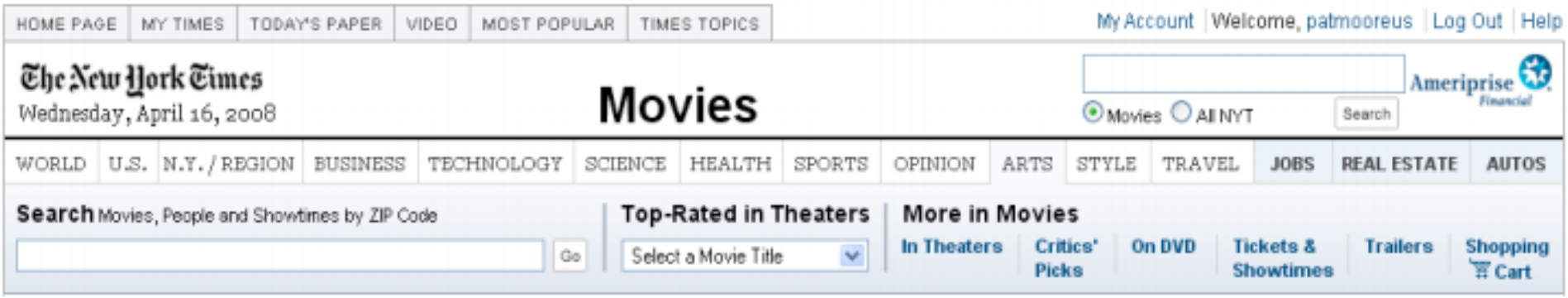


Kohavi, Henne, and Sommerfield, *"Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO,"* KDD 2007

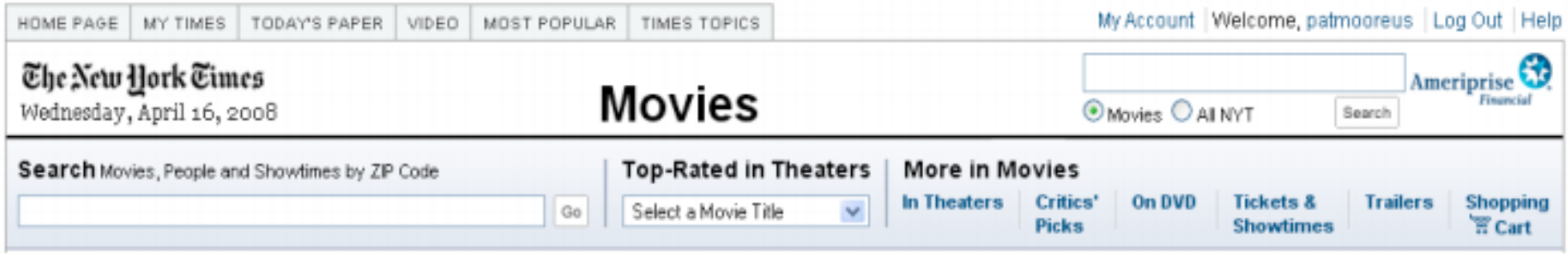
An example

Here is a simple experiment conducted by the New York Times web site (`nytimes.com`) in 2008 -- What is the difference between these two pages and what differences in visits might you be interested in comparing?

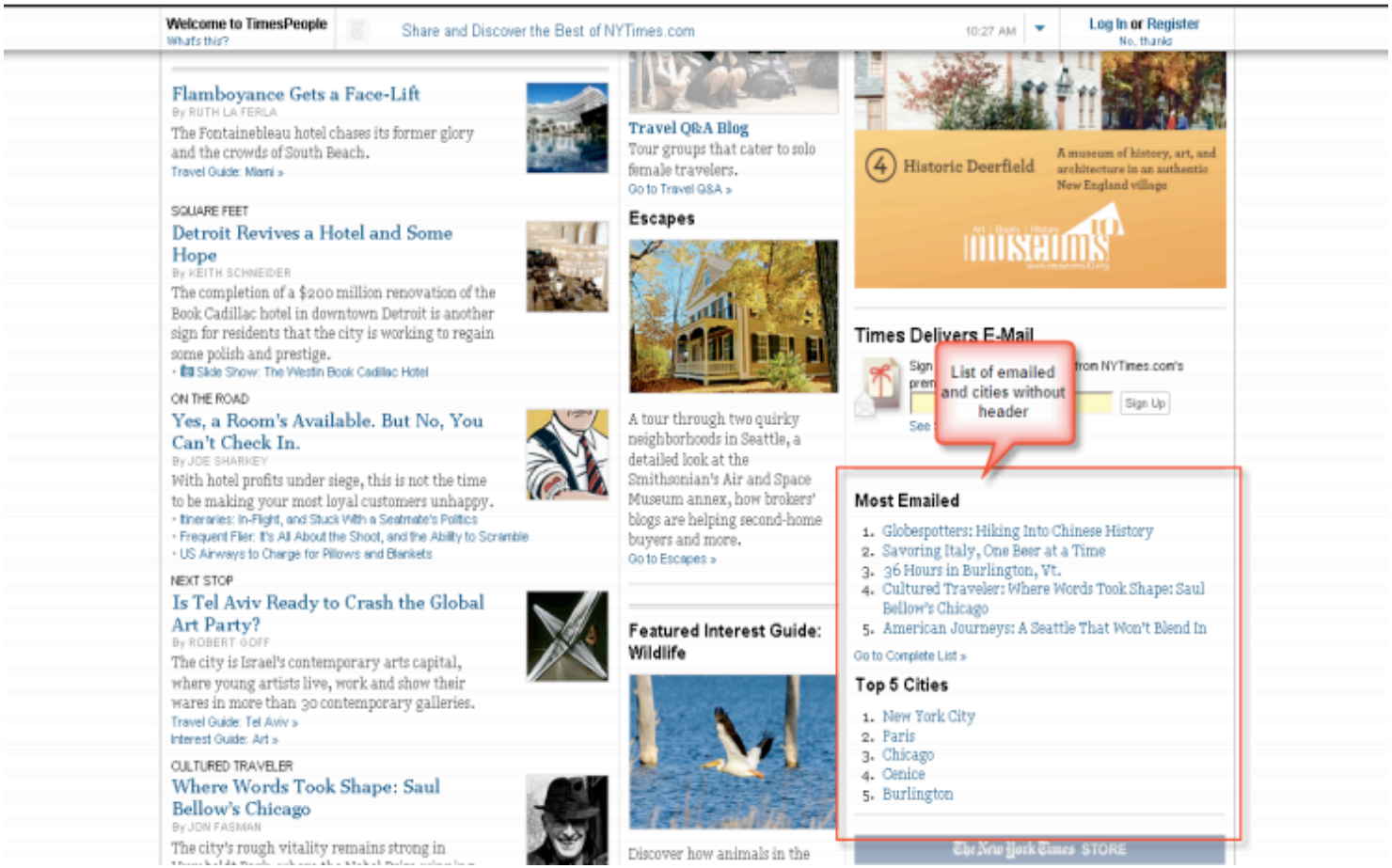
A: Control



B: Test



List: Variation 10858



Tabs: Variation 10859