INTRO PROBABILITY

BILL HOWE

UNIVERSITY OF WASHINGTON

HOW DIFFERENT IS DIFFERENT?

How do we know the difference in two treatments is not just due to chance?

We don't. But we can calculate the odds that it is.

This is the *p-value*

In repeated experiments at this sample size, how often would you see a result at least this extreme assuming the null hypothesis?

P-VALUE

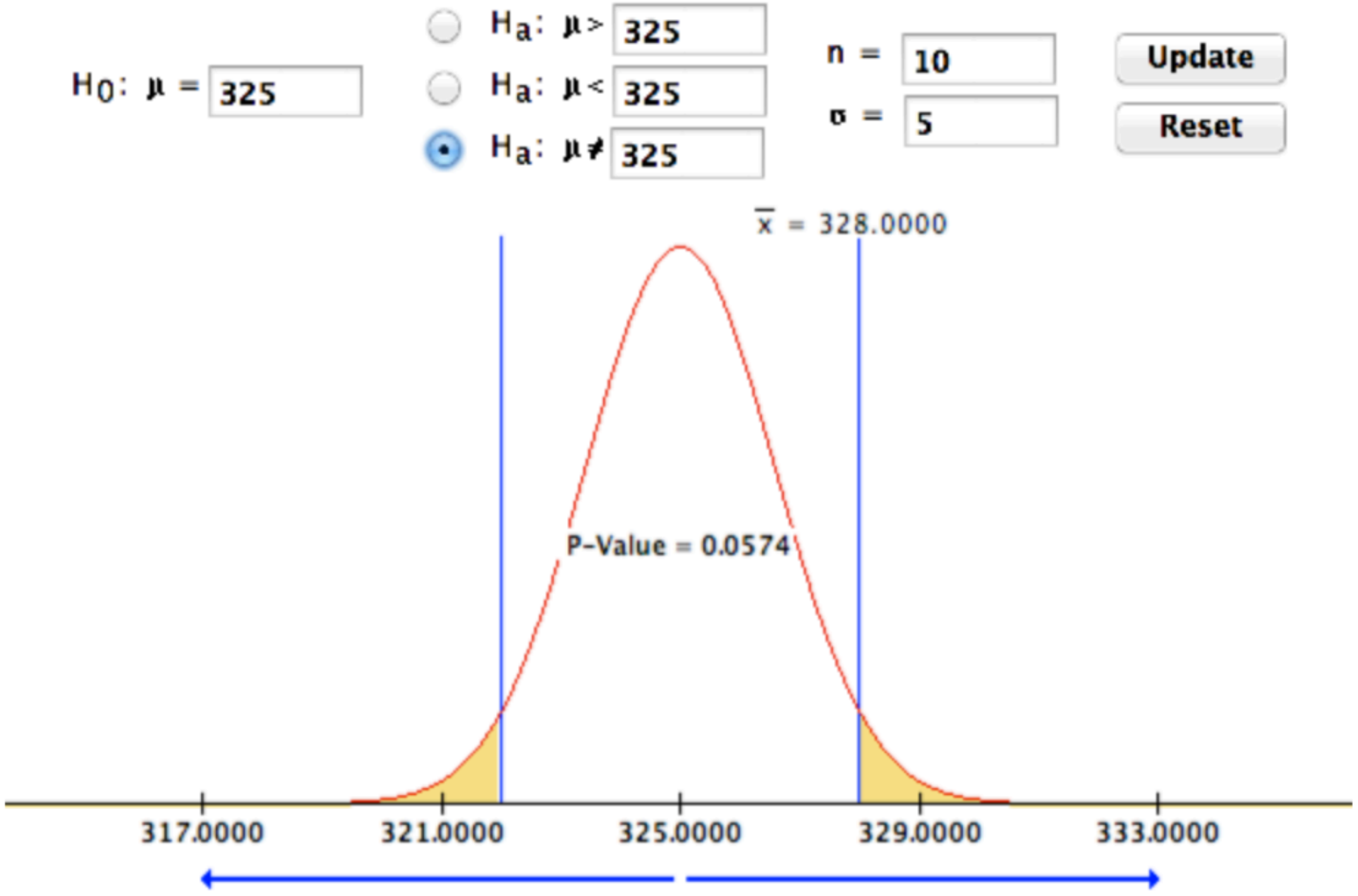
If the test is two-sided:

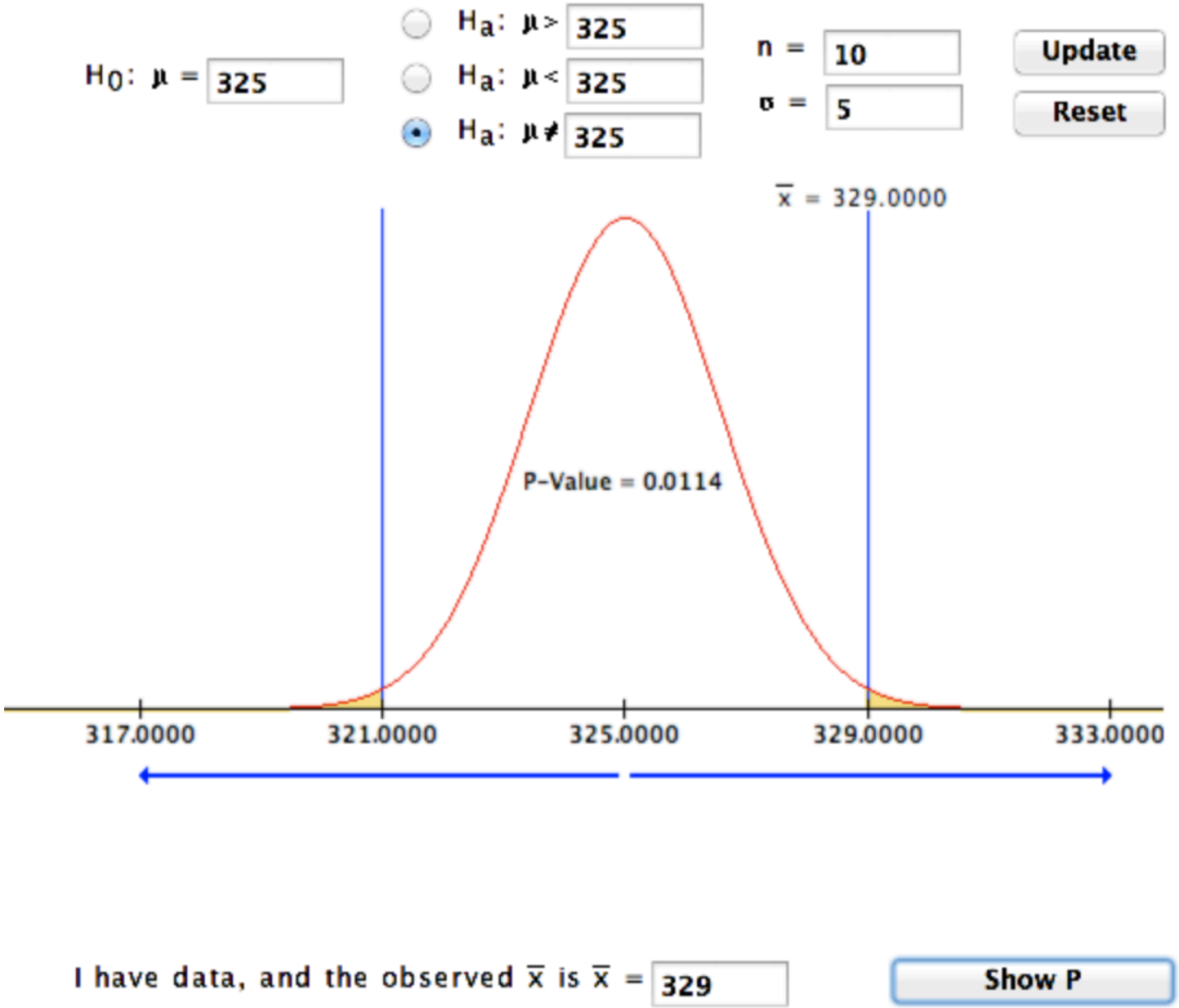
- P-value = 2 * P(X > |observed value|)

If the test is one-sided:

- H_A is $\mu > \mu_0$
- P-value = P(X > observed value)

- H_A is $\mu < \mu_0$
- P-value = P(X < observed value)


H_0: μ = 325
H_a: μ > 325
H_a: μ < 325
H_a: μ ≠ 325
n = 10
σ = 5
Update
Reset
$\bar{x}$ = 328.0000
P-Value = 0.0574
317.0000
321.0000
325.0000
329.0000
333.0000
I have data, and the observed $\bar{x}$ is $\bar{x}$ = 328
Show P


H_0: μ = 325
H_a: μ > 325
H_a: μ < 325
H_a: $\mu \neq$ 325
n = 10
σ = 5
Update
Reset
$\bar{x}$ = 329.0000
P-Value = 0.0114
317.0000
321.0000
325.0000
329.0000
333.0000
I have data, and the observed $\bar{x}$ is $\bar{x}$ = 329
Show P

JONAH LEHRER, 2010, THE NEW YORKER THE TRUTH WEARS OFF

John Davis, University of Illinois
"Davis has a forthcoming analysis demonstrating that the efficacy of antidepressants has gone down as much as threefold in recent decades."

Anders Pape Møller, 1991
"female barn swallows were far more likely to mate with male birds that had long, symmetrical feathers"
"Between 1992 and 1997, the average effect size shrank by eighty per cent."

Jonathan Schooler, 1990
"subjects shown a face and asked to describe it were much less likely to recognize the face when shown it later than those who had simply looked at it."
The effect became increasingly difficult to measure.

Joseph Rhine, 1930s, coiner of the term extrasensory perception
Tested individuals with card-guessing experiments. A few students achieved multiple low-probability streaks. But there was a "decline effect" – their performance became worse over time.

http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

REASON 1: PUBLICATION BIAS

"In the last few years, several meta-analyses have reappraised the efficacy and safety of antidepressants and concluded that the therapeutic value of these drugs may have been significantly overestimated."

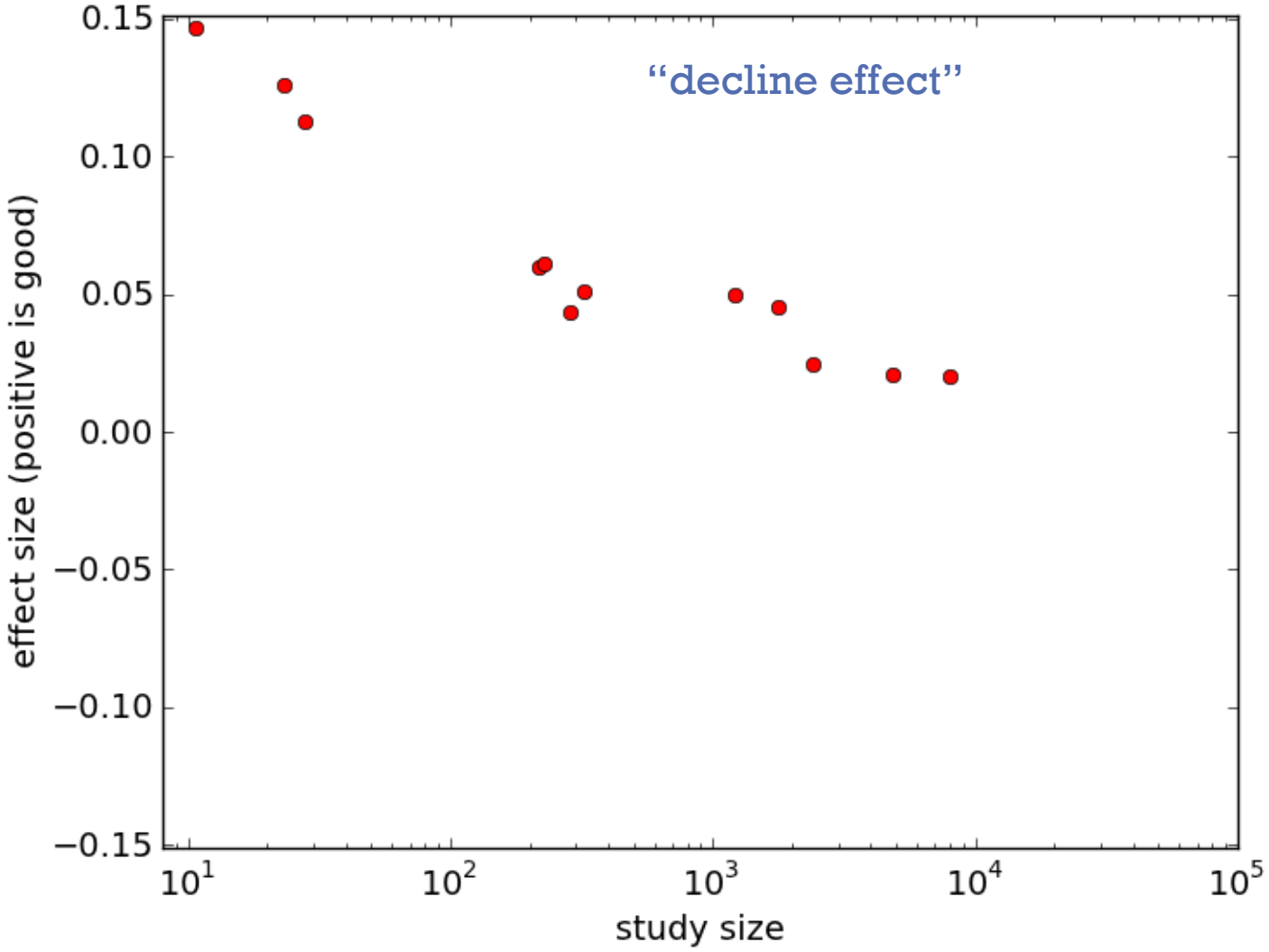
"Although publication bias has been documented in the literature for decades and its origins and consequences debated extensively, there is evidence suggesting that this bias is increasing."

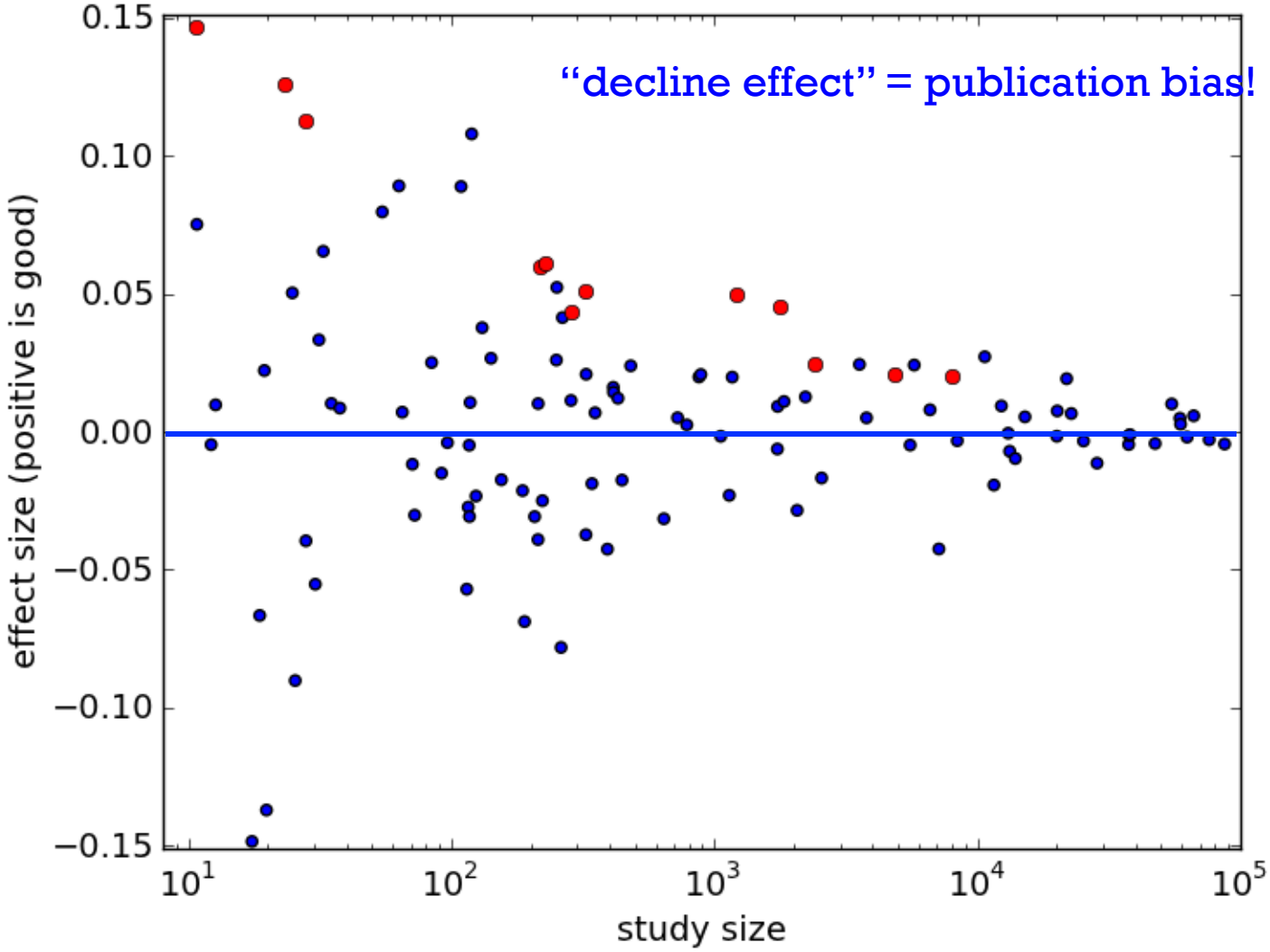
"A case in point is the field of biomedical research in autism spectrum disorder (ASD), which suggests that **in some areas negative results are completely absent**"

"… a highly significant correlation ($R^2 = 0.13$, $p < 0.001$) between impact factor and overestimation of effect sizes has been reported."

Publication bias: What are the challenges and can they be overcome?
Ridha Joober, Norbert Schmitz, Lawrence Annable, and Patricia Boksa
J Psychiatry Neurosci. 2012 May; 37(3): 149–152. doi: 10.1503/jpn.120065

PUBLICATION BIAS (2)




"decline effect" = publication bias!
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
effect size (positive is good)
10^1
10^2
10^3
10^4
10^5
study size

BACKGROUND: EFFECT SIZE

$$\text{Effect size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{standard deviation}}$$

Expressed in relevant units

Not just "significant" – *how* significant?

Used prolifically in meta-analysis to combine results from multiple studies

- But be careful – averaging results from different experiments can produce nonsense

Caveat: Other definitions of effect size exist: odds-ratio, correlation coefficient

Robert Coe, 2002, Annual Conference of the British Educational Research Association
It's the Effect Size, Stupid: What effect size is and why it is important.

EFFECT SIZE

Standardized Mean Difference

$$ES = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{pooled}}$$

Lots of ways to estimate the pooled standard deviation

$$\sigma_{pooled} = \hat{\sigma}_2$$

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}}$$

Glass, 1976

e.g., Hartung et al., 2008

META-ANALYSIS

1978: Gene V. Glass statistically aggregate the findings of 375 psychotherapy outcome studies Glass (and colleague Smith) to disprove claim that psychotherapy was useless

Glass coined the term "meta-analysis"

Earlier ideas from Fisher (1944)

- "When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance"

adapted from slide by David B. Wilson, 1999

META-ANALYSIS

Even more important in data science

- You will often be working with data you didn't collect
- "Big Data" may have become big by combining data from different sources
- When is this ok? Test for homogeneity

META-ANALYSIS: WEIGHTED AVERAGE

Idea: Average across multiple studies, but give more weight to more precise studies

Simple method: Weight by sample size

$$w_i = \frac{n_i}{\sum_j n_j}$$

Inverse-variance weight :

$$w_i = \frac{1}{se^2}$$

Lots of variants

Caveat: This is a fixed-effect model: it assumes that each individual study is measuring the same true effect. We won't talk about the random effects model.

EFFECT SIZE: COHEN'S HEURISTIC

Standardized mean difference effect size

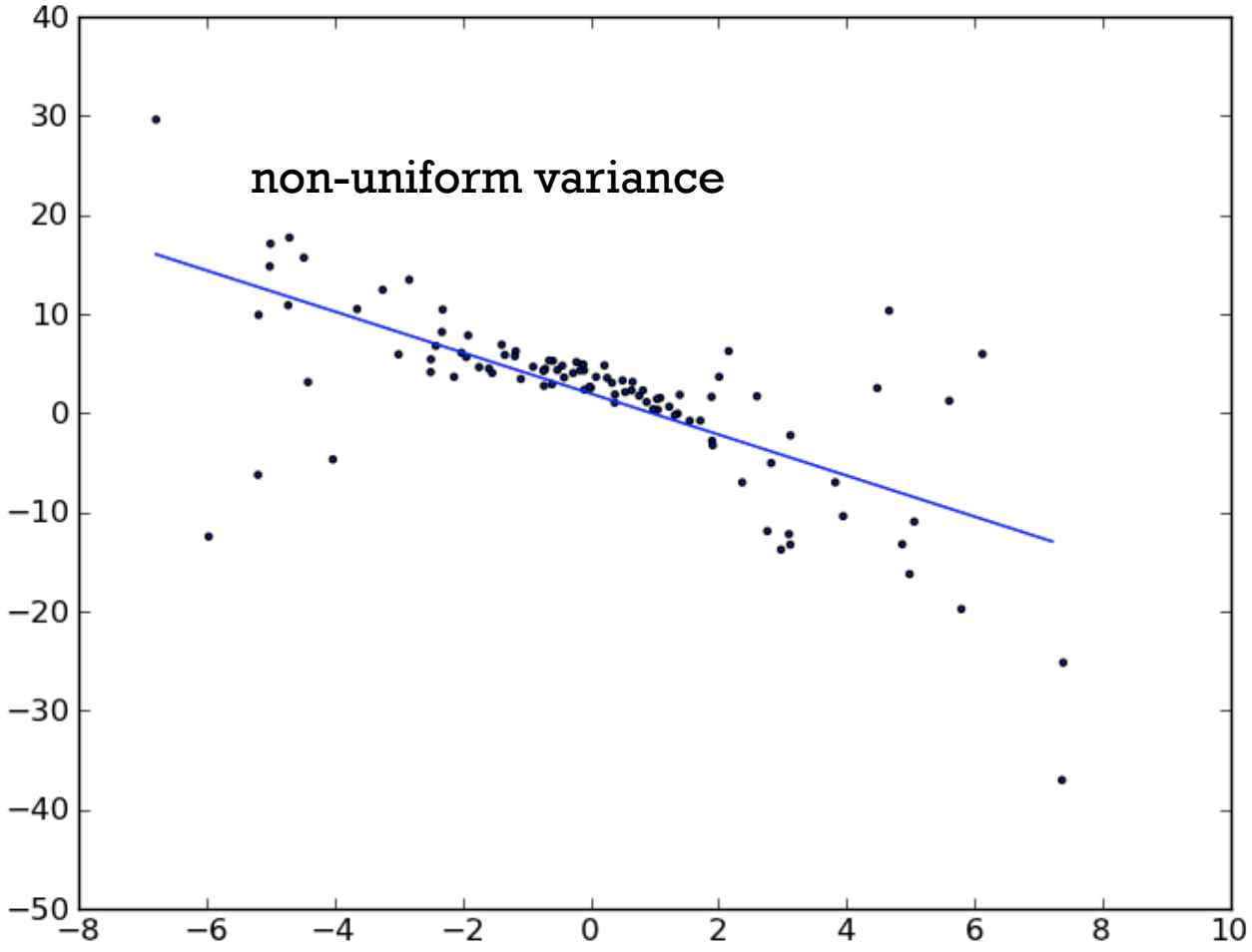
- small = 0.20
- medium = 0.50
- large = 0.80

CONFIDENCE INTERVAL (OF EFFECT SIZE)

What does a 95% confidence interval of the effect size mean?

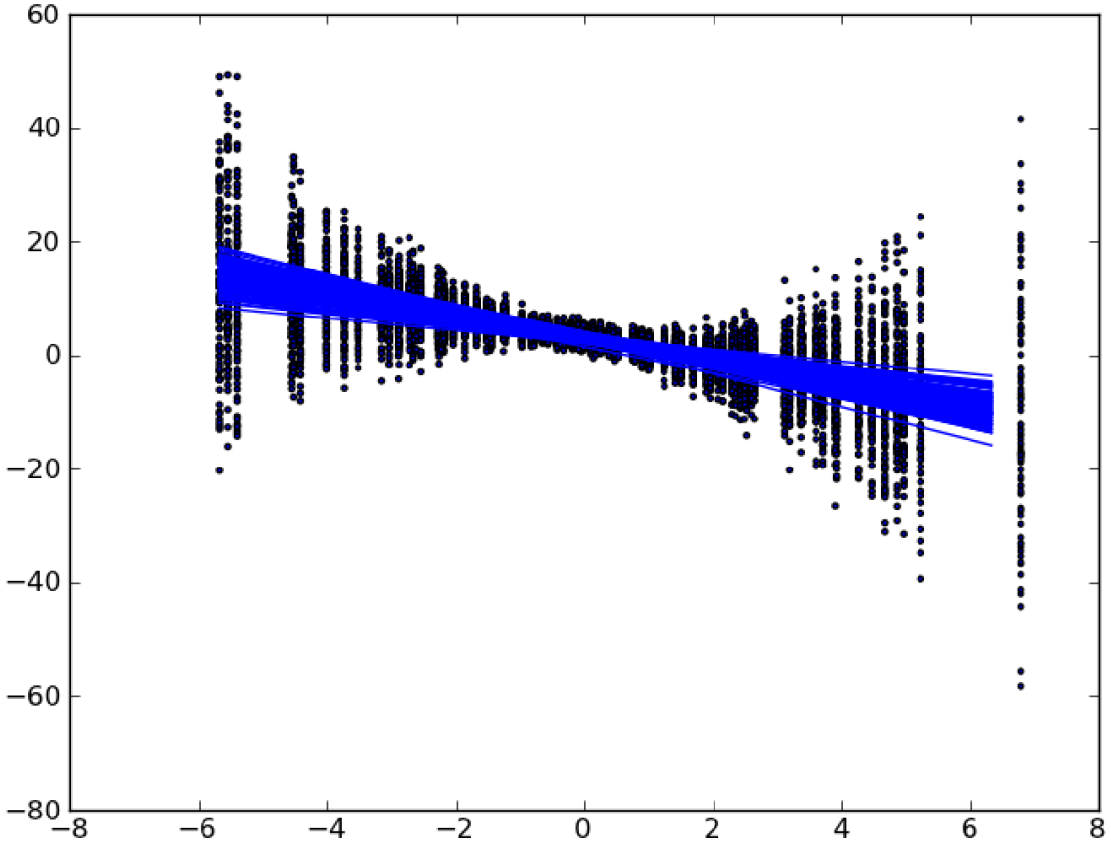
- If we repeated the experiment 100 times, we expect that the interval would include this effect size 95/100 times
- If this interval includes 0.0, that's equivalent to saying the result is not statistically significant.

ASIDE: HETEROSKEDASTICITY



ASIDE: HETEROSKEDASTICITY

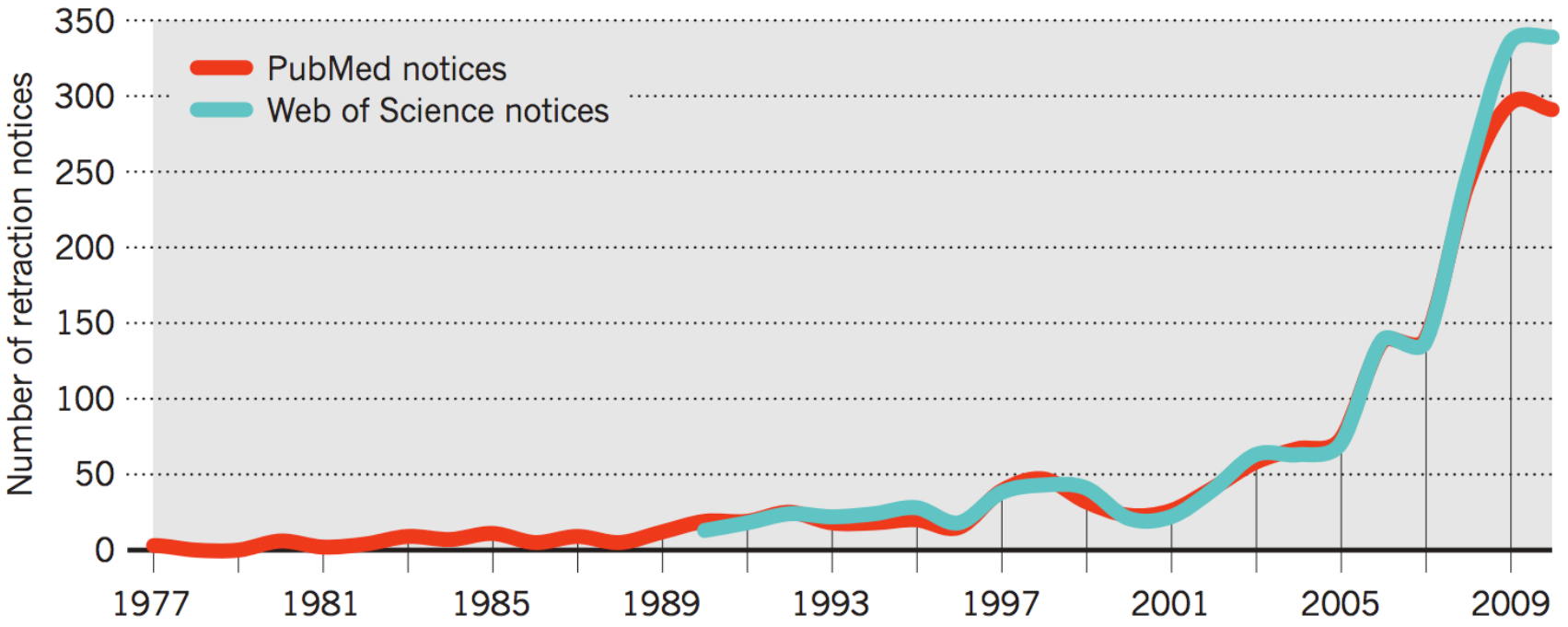
100 repetitions, same x-values, y-values drawn from the same model



- Not necessarily a problem
- Still provides an unbiased estimate
- Can increase error estimates, leading to Type 2 errors: overlooking a real effect

REASON 2: MISTAKES AND FRAUD

2001 – 2011:
- 10X increase in retractions
- only 1.44X increase in papers

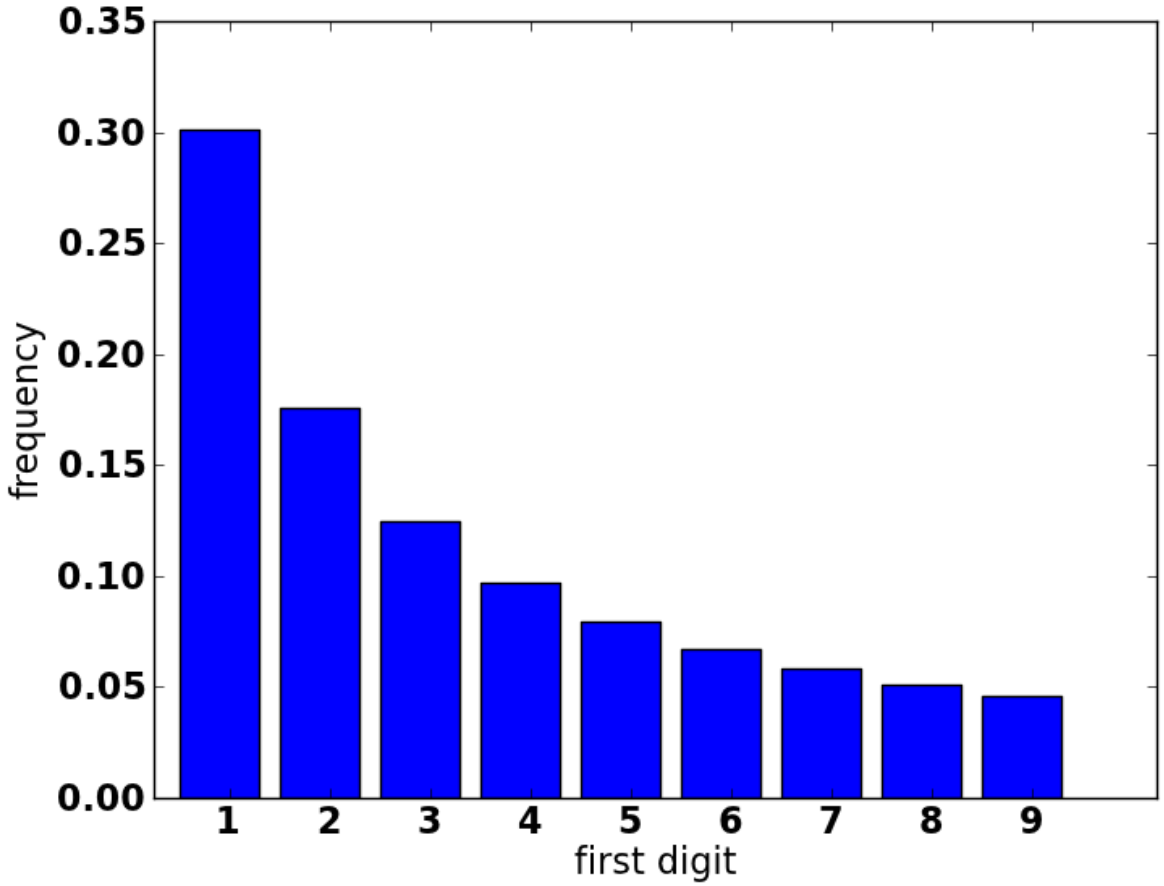


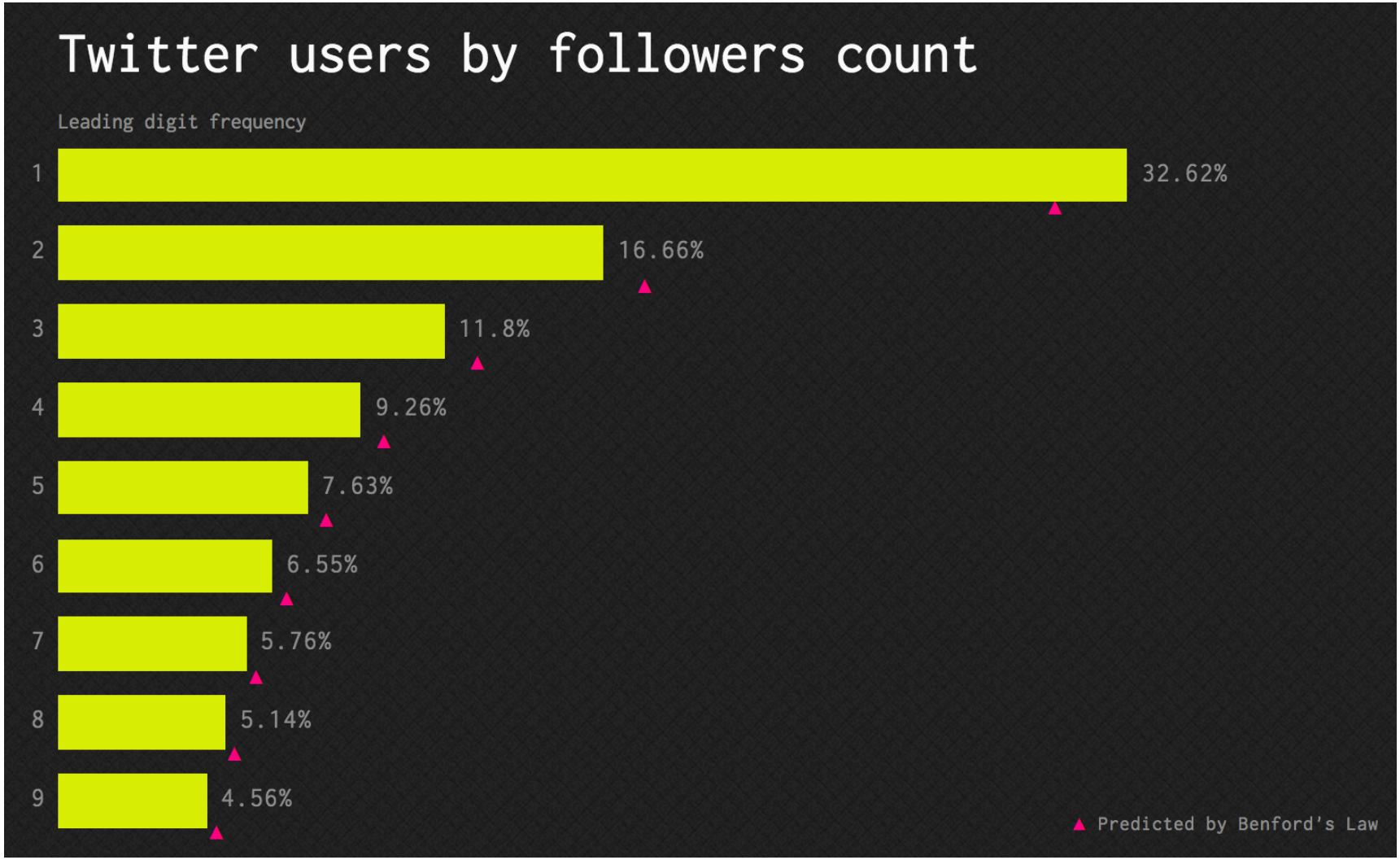
Richard Van Noorden, 2011, Nature 478
The Rise of the Retractions

http://www.nature.com/news/2011/111005/pdf/478026a.pdf

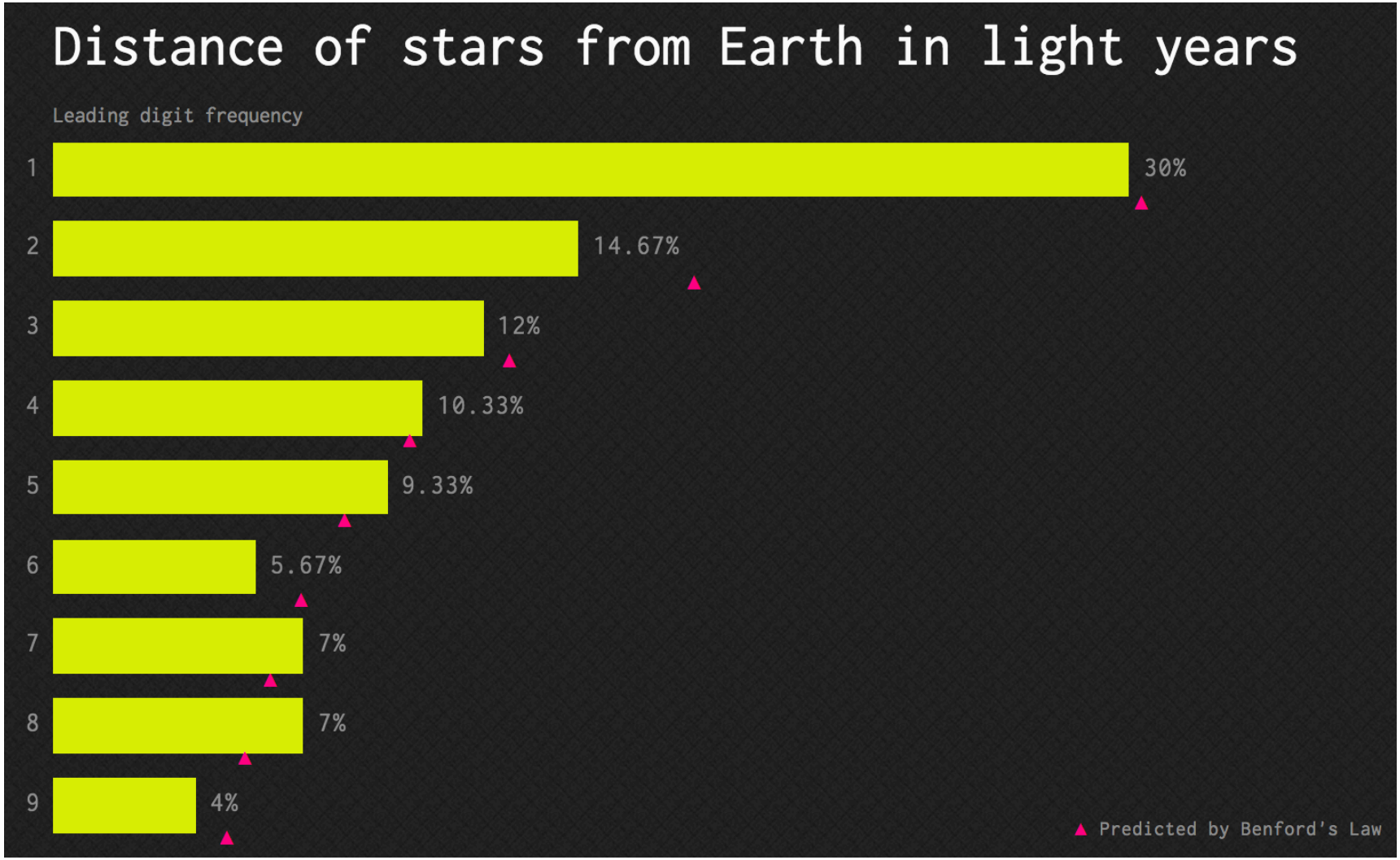
BENFORD'S LAW: POTENTIAL TOOL FOR FRAUD DETECTION

New York	8,336,697
Los Angeles	3,857,799
Chicago	2,714,856
Houston	2,160,821
Philadelphia	1,547,607
Phoenix	1,488,750
San Antonio	1,382,951
San Diego	1,338,348
Dallas	1,241,162

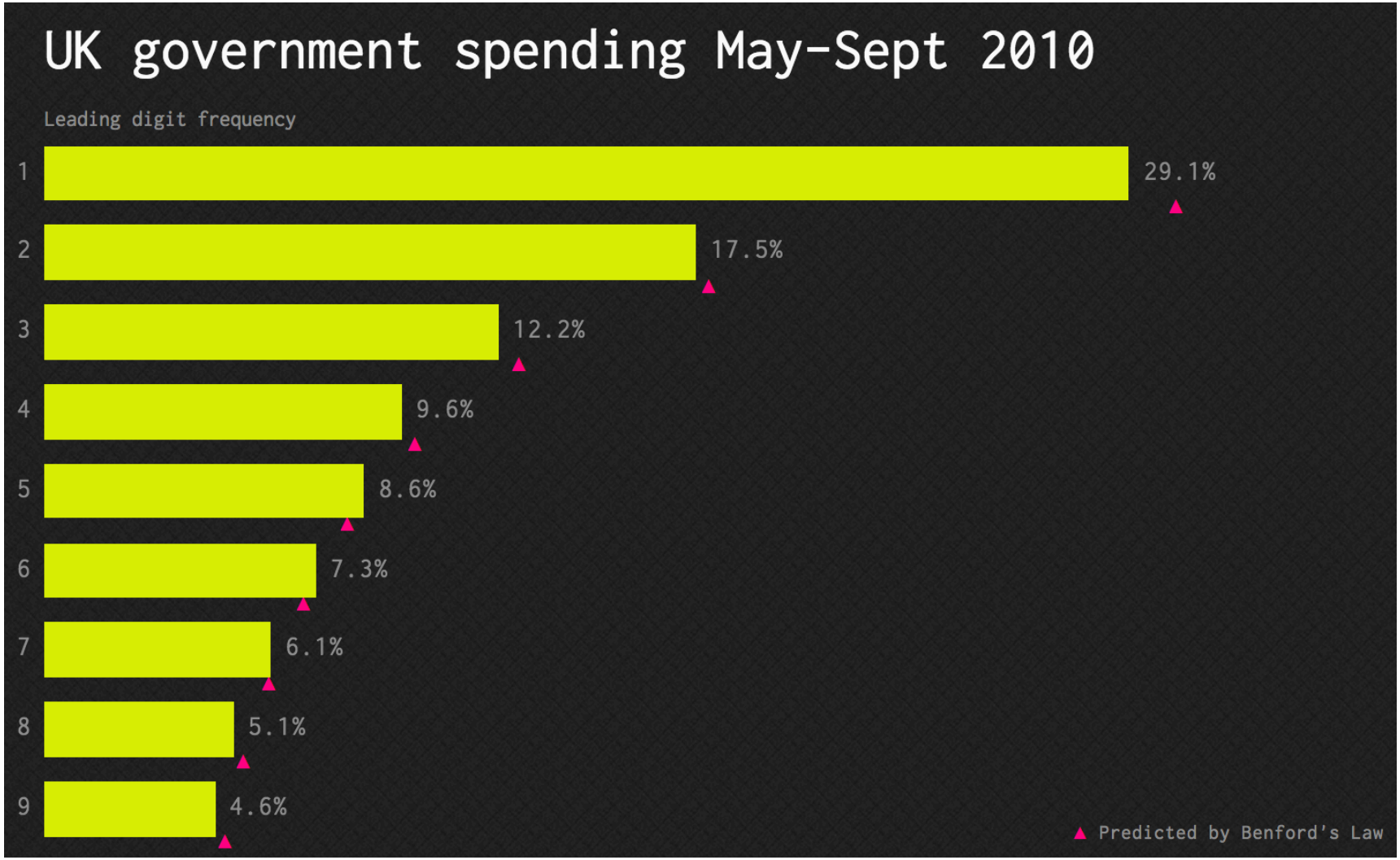




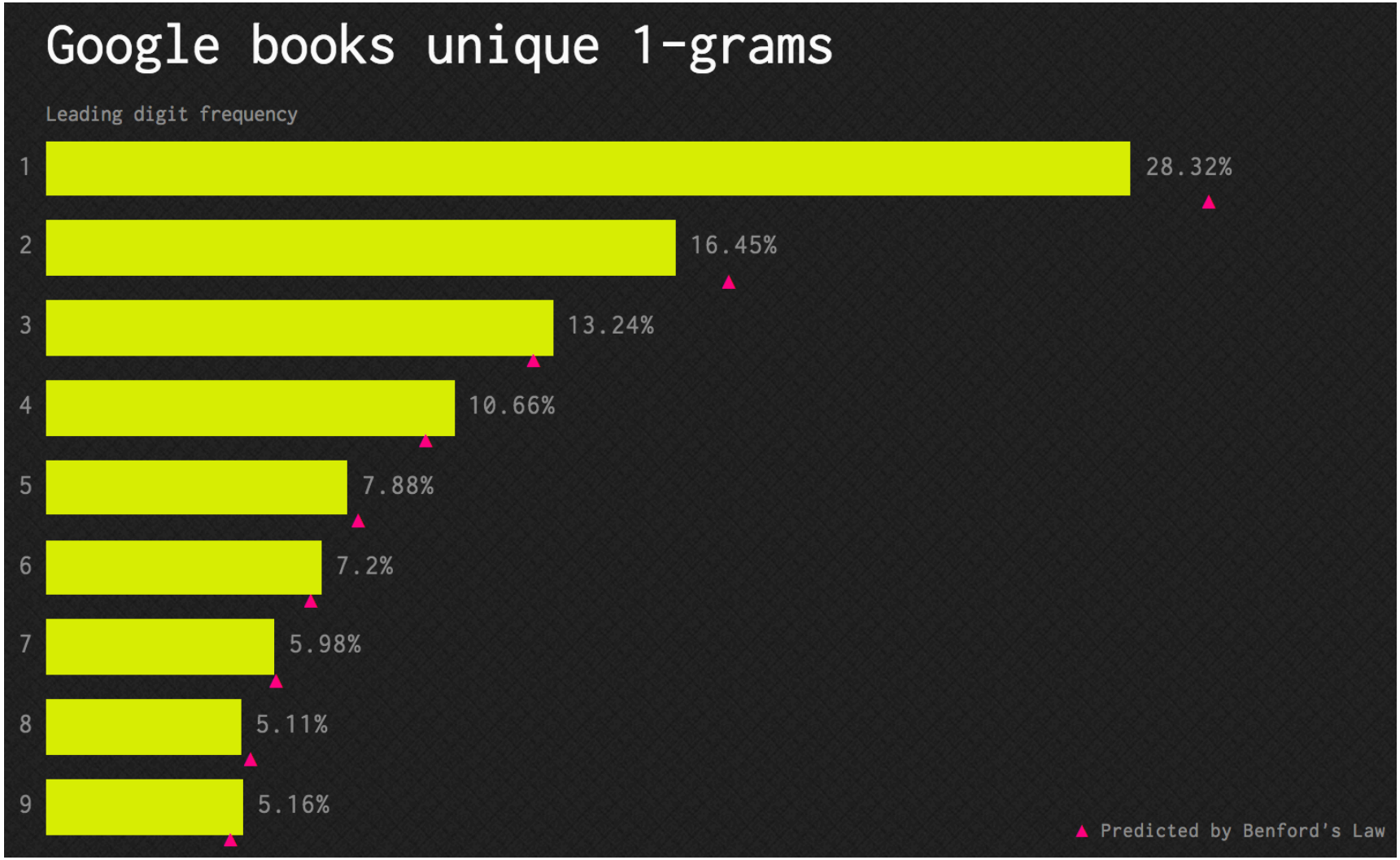
src: http://testingbenfordslaw.com/x



src: http://testingbenfordslaw.com/x



src: http://testingbenfordslaw.com/x



src: http://testingbenfordslaw.com/x

BENFORD'S LAW TO DETECT FRAUD

Diekmann, 2007

- Found that first and second digits of published statistical estimates were approximately Benford distributed
- Asked subjects to manufacture regression coefficients, and found that the first digits were hard to detect as anomolous, but the second and third digits deviated from expected distributions significantly.

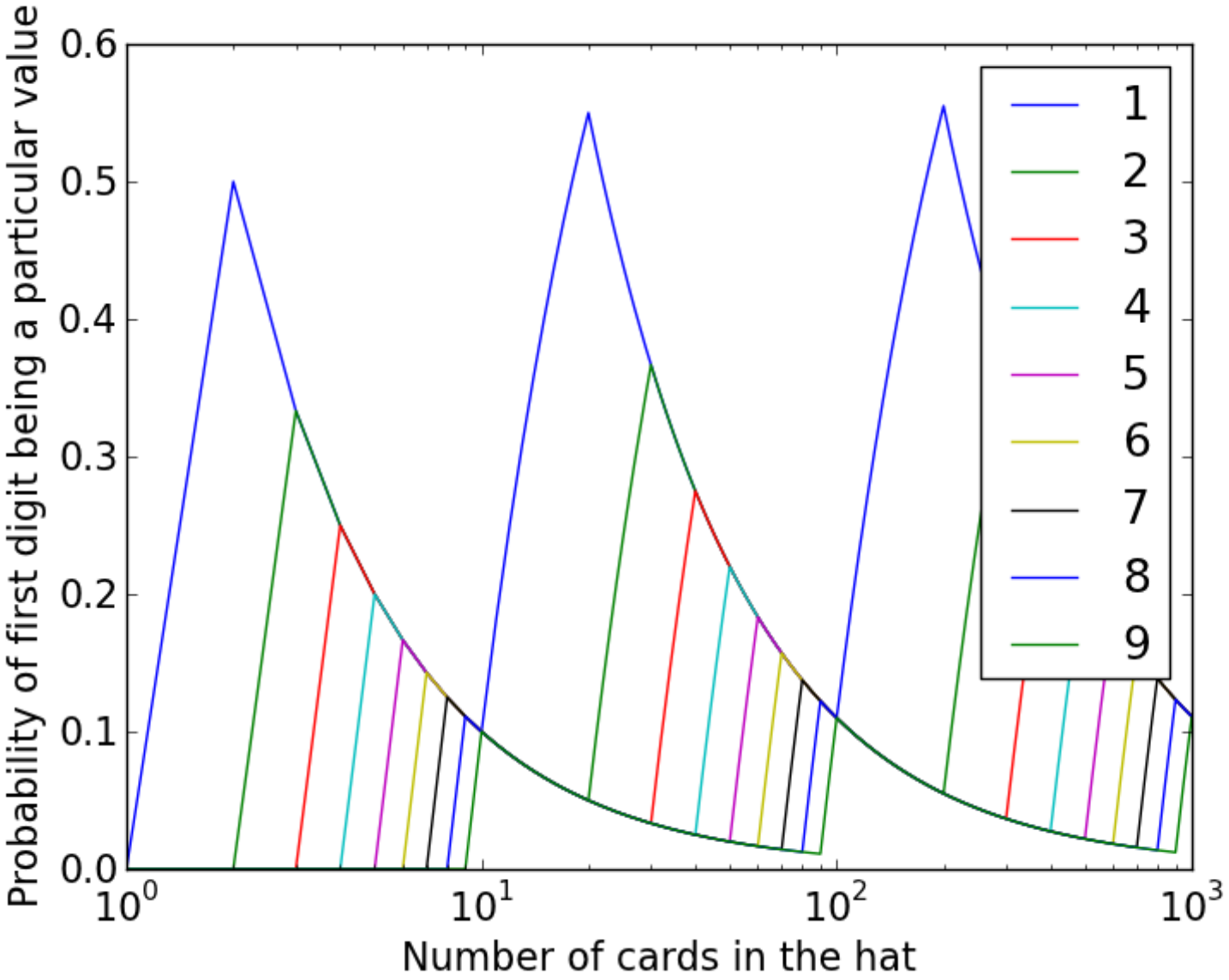
Andreas Diekmann, 2007, Journal of Applied Statistics, 34(3)
Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data

BENFORD'S LAW INTUITION

Given a sequence of cards labeled 1, 2, 3, … 999999

Put them in a hat, one by one, in order

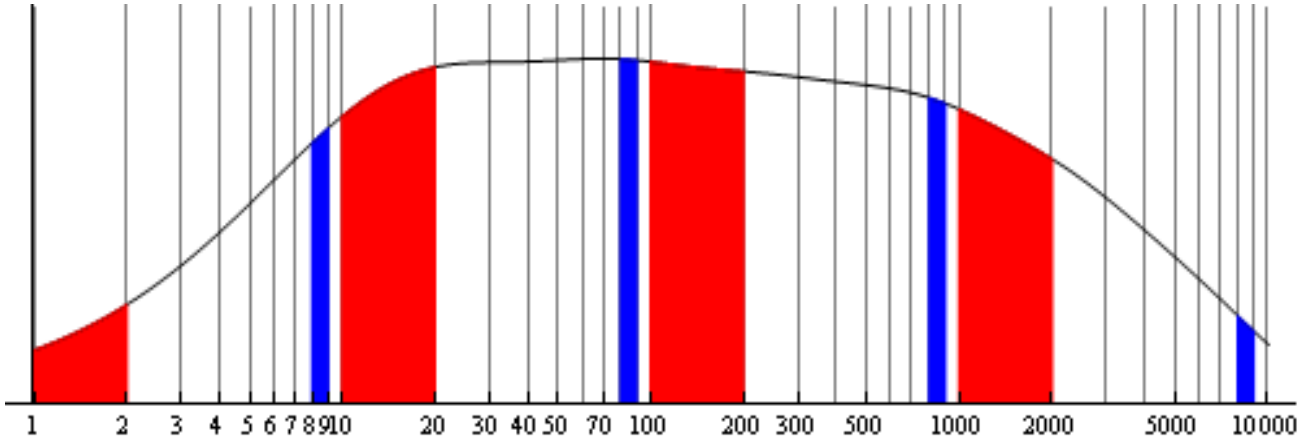
After each card, ask "What is the probability of drawing the number 1?"

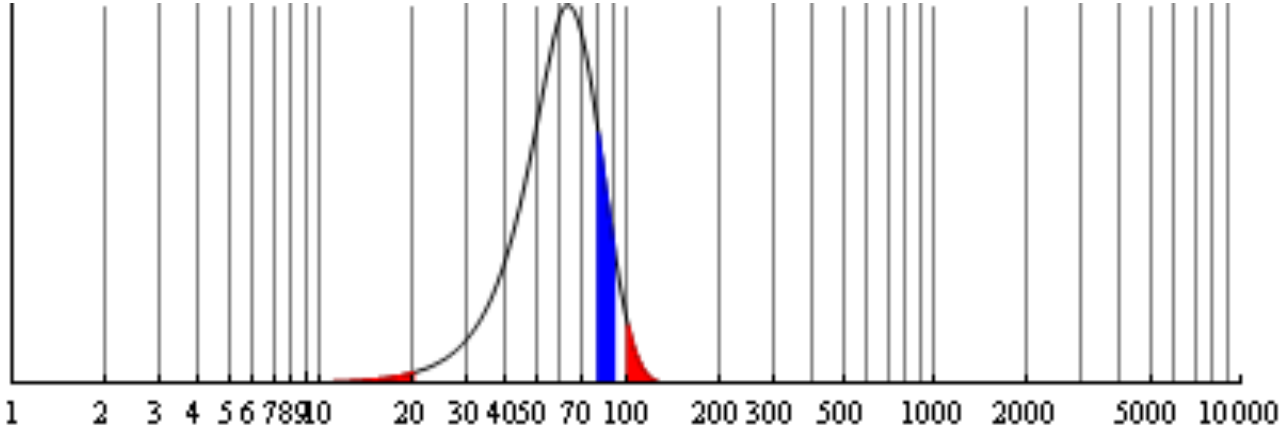

Probability of first digit being a particular value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
10^0
10^1
10^2
10^3
Number of cards in the hat
1
2
3
4
5
6
7
8
9

BENFORD'S LAW EXPLANATION

Limitations

- Data must span several orders of magnitude
- No min/max cutoffs





REASON 3: MULTIPLE HYPOTHESIS TESTING

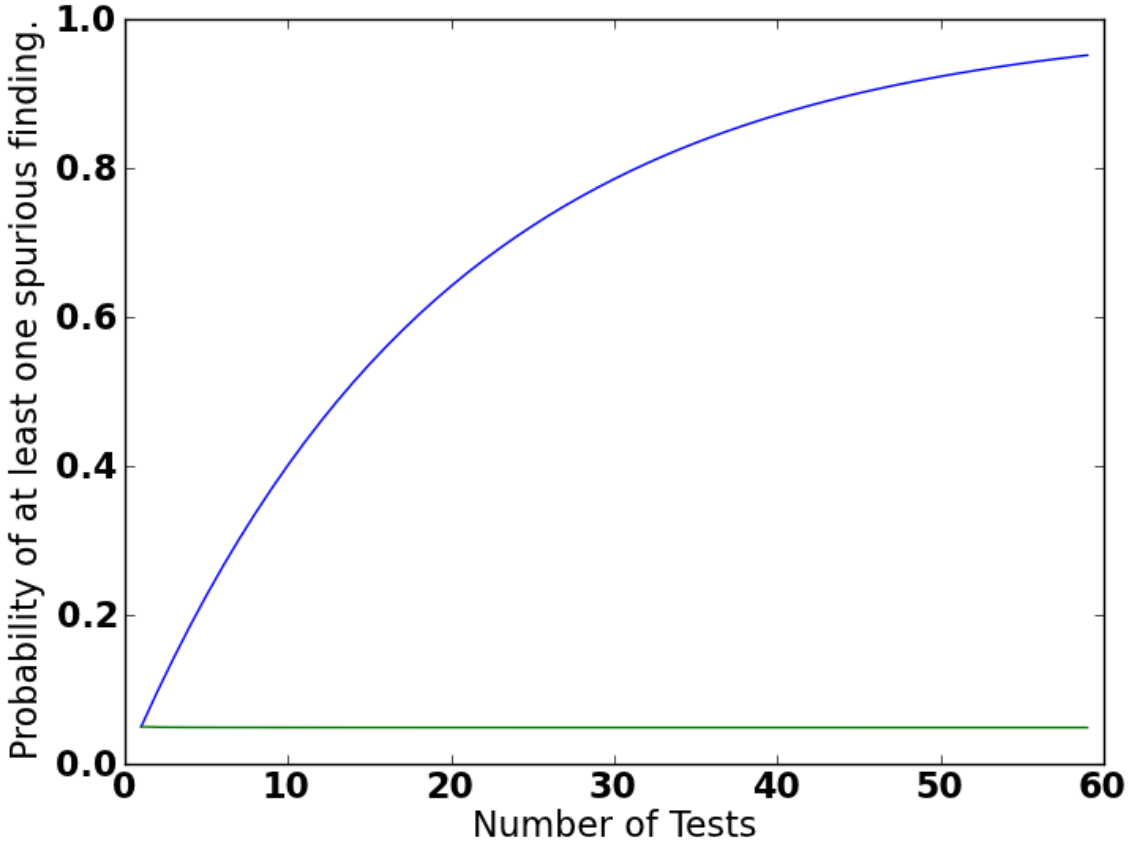
- If you perform experiments over and over, you're bound to find something
- This is a bit different than the publication bias problem: Same sample, different hypotheses
- Significance level must be adjusted down when performing multiple hypothesis tests

P(detecting an effect when there is none) = α = 0.05

P(detecting an effect when it exists) = 1 – α

P(detecting an effect when it exists on every experiment) = $(1 - \alpha)^k$

P(detecting an effect when there is none on <u>at least one</u> experiment) = $1 - (1 - \alpha)^k$



α = 0.05

"Familywise Error Rate"

FAMILYWISE ERROR RATE CORRECTIONS

Bonferroni Correction

- Just divide by the number of hypotheses

Šidák Correction

- Asserts independence

$$\alpha_c = \frac{\alpha}{k}$$

$$\alpha = 1 - (1 - \alpha_c)^k$$

$$\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}$$

FALSE DISCOVERY RATE

	Reject H0	**Do Not Reject H0**	**Total**
H0 is true	FD	TN	T
H0 is false	TD	FN	F
Total	D	N	TFDN

T/F = True/False
D/N = Discovery/Nondiscovery

$$Q = FDR = \frac{FD}{D}$$

FDR (2)

Bonferroni correction and other FWER corrections tend to wipe out evidence of the most interesting effects; they suffer from low power.

FDR control offers a way to increase power while maintaining a bound on the ratio of wrong conclusions

Intuition:

- 4 false discoveries out of 10 rejected null hypotheses

is a more serious error than

- 20 false discoveries out of 100 rejected null hypotheses.

adapted from a slide by Christopher Genovese

BENJAMINI-HOCHBERG PROCEDURE

Compute the p-value of m hypotheses

Order them in increasing order of p-value

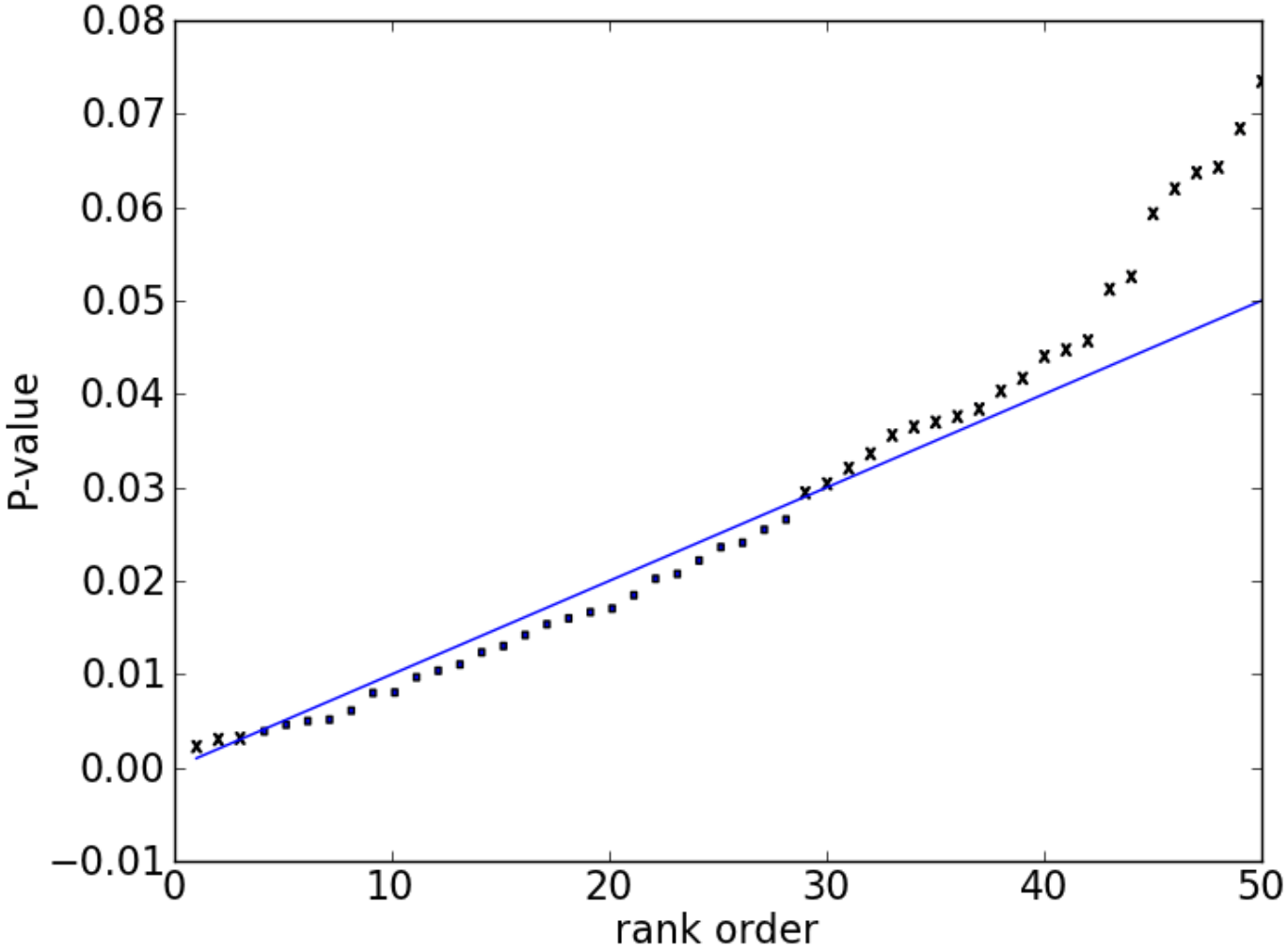
- That is, most likely hypotheses are first

$$P_i \leq \frac{i}{m}\alpha$$

$$FDR \leq \frac{T}{m}\alpha$$

i	$\frac{i}{50}\alpha$
1	0.001
2	0.002
3	0.003
…	
20	0.020
…	

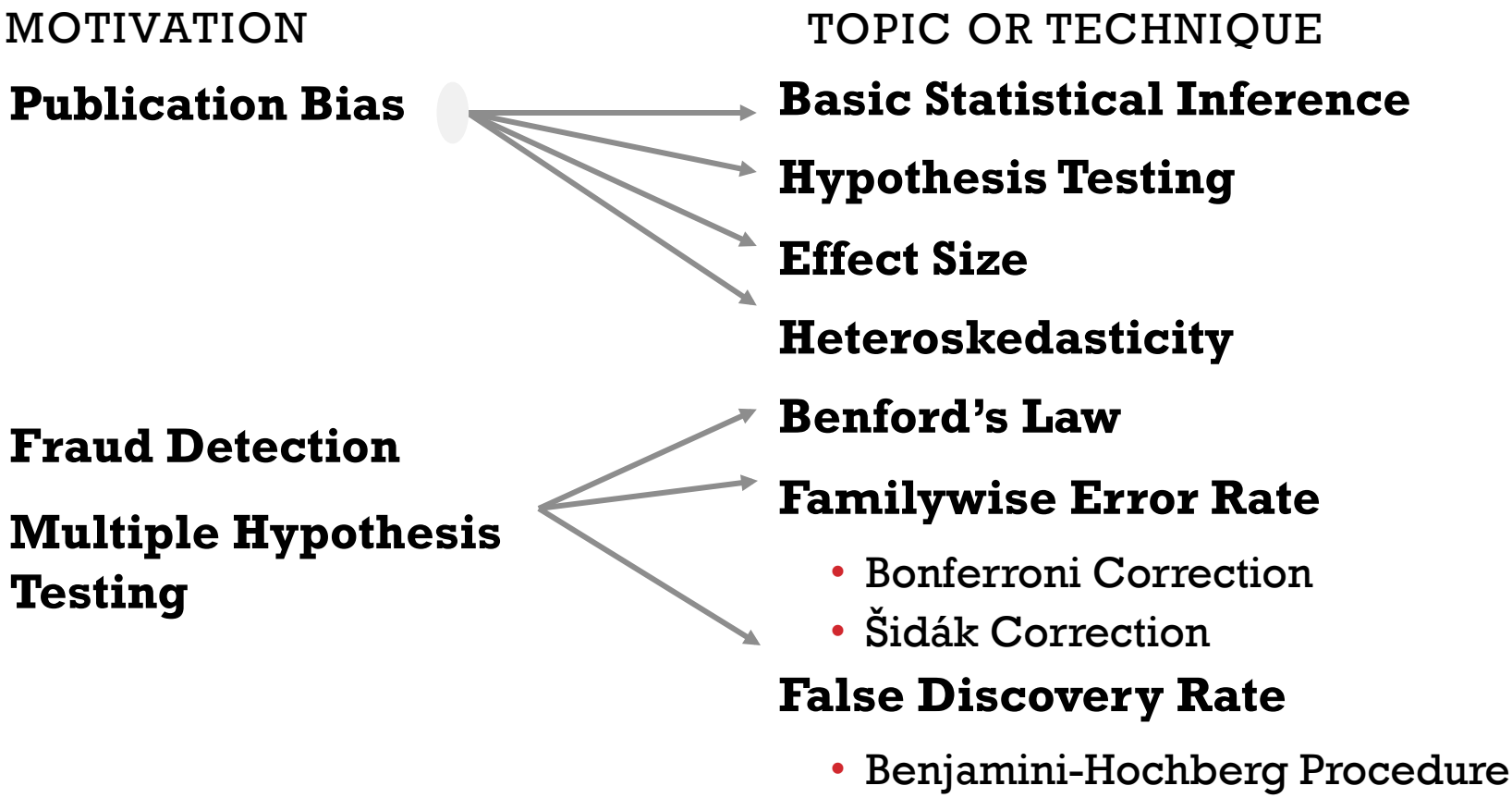
BENJAMINI-HOCHBERG PROCEDURE



$m = 50$

$\alpha = 0.05$

WHERE WE ARE


MOTIVATION
TOPIC OR TECHNIQUE
Publication Bias
Basic Statistical Inference
Hypothesis Testing
Effect Size
Heteroskedasticity
Fraud Detection
Benford's Law
Multiple Hypothesis Testing
Familywise Error Rate
Bonferroni Correction
Šidák Correction
False Discovery Rate
Benjamini-Hochberg Procedure

WHAT ABOUT BIG DATA?

"Classical statistics was fashioned for small problems, a few hundred data points at most, a few parameters."

"The bottom line is that we have entered an era of massive scientific data collection, with a demand for answers to large-scale inference problems that lie beyond the scope of classical statistics."

Bradley Efron,
Bayesians, Frequentists, and Scientists

http://www-stat.stanford.edu/~ckirby/brad/papers/2005BayesFreqSci.pdf


Internet Explorer vs Murder Rate
18,000
17,200
16,400
15,600
14,800
14,000
90%
75%
60%
45%
30%
2006 2007 2008 2009 2010 2011
Murders in US
Internet Explorer Market Share

POSITIVE CORRELATIONS

Number of police officers and number of crimes (Glass & Hopkins, 1996)

Amount of ice cream sold and deaths by drownings (Moore, 1993)

Stork sightings and population increase

(Box, Hunter, Hunter, 1978)

THE "CURSE" OF BIG DATA?

"…the curse of big data is the fact that when you search for patterns in very, very large data sets with billions or trillions of data points and thousands of metrics, you are bound to identify coincidences that have no predictive power."

Vincent Granville

http://www.analyticbridge.com/profiles/blogs/the-curse-of-big-data

VINCENT GRANVILLE'S EXAMPLE

Consider stock prices for 500 companies over a 1-month period

Check for correlations in all pairs

ASIDE: VERY BASIC TIMESERIES ANALYSIS (1)

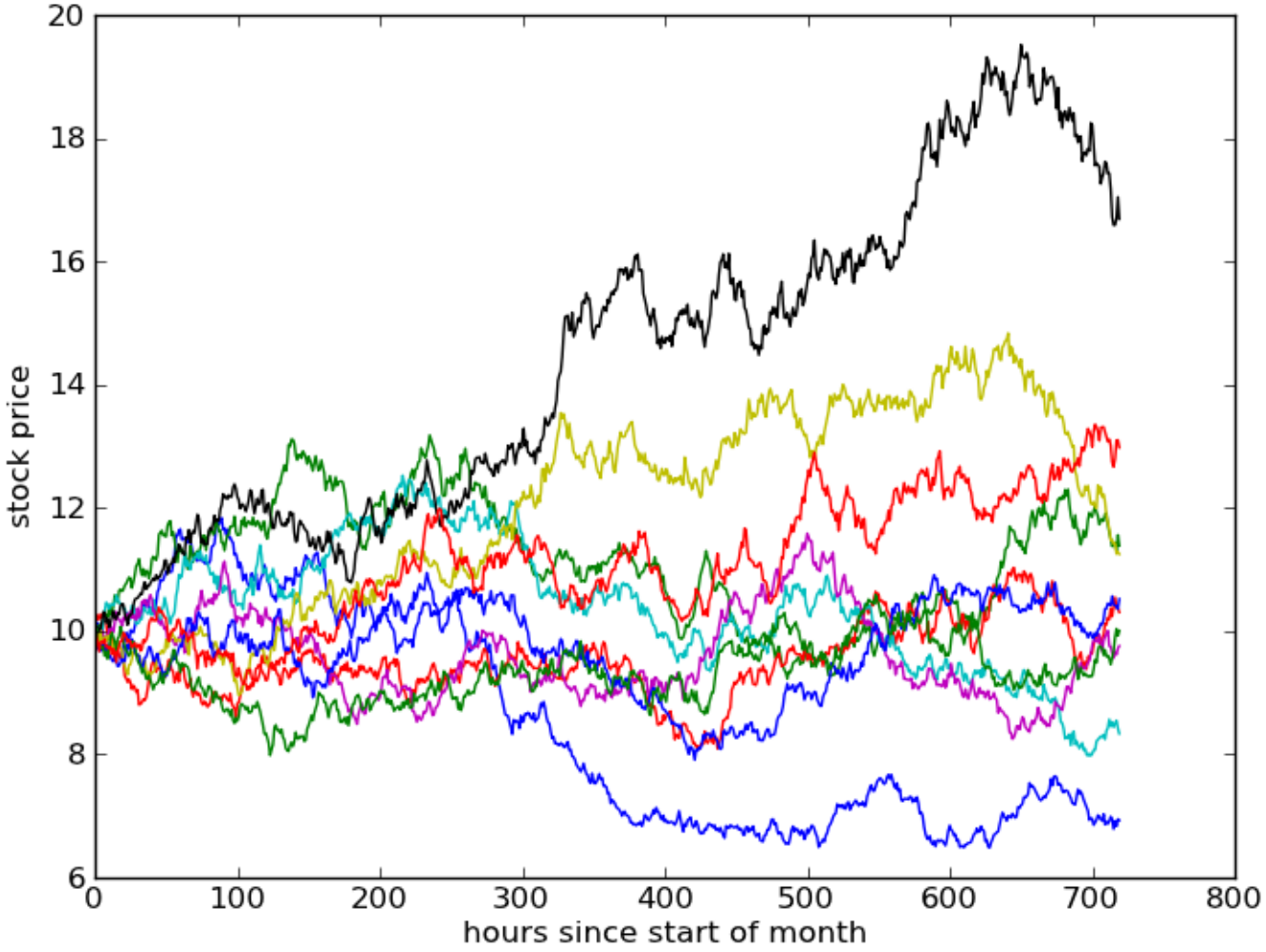
$$\mathrm{cov}(x,y) = \sum_i (x_i - u_x)(y_i - u_y)$$

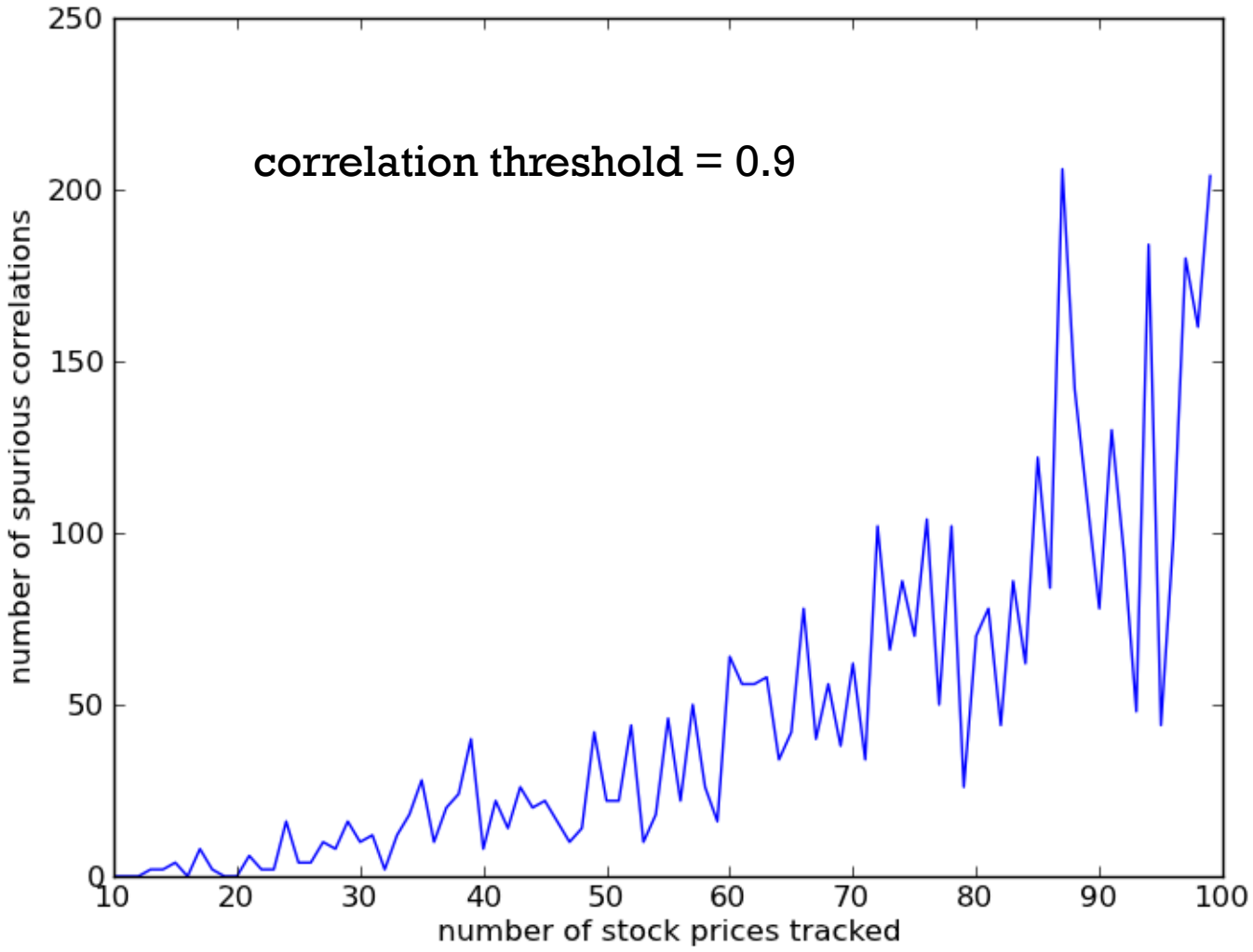
$$corr(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\sum_i (x_i - u_x)^2 (y_i - u_y)^2}}$$

standard deviation

RANDOM WALK

each step is normally distributed @ 1% of current price




correlation threshold = 0.9
number of spurious correlations
number of stock prices tracked
0
50
100
150
200
250
10
20
30
40
50
60
70
80
90
100

IS BIG DATA DIFFERENT?

Big P vs. Big N

- P = number of variables (columns)
- N = number of records

Marginal cost of increasing N is essentially zero!

But while >N decreases variance, it amplifies bias

- Ex: You log all clicks to your website to model user behavior, but this only samples current users, not the users you want to attract.
- Ex: Using mobile data to infer buying behavior

Beware multiple hypothesis tests

- "Green jelly beans cause acne"

Taleb's "Black Swan" events

- The turkey's model of human behavior

Frequentist Approach to Statistics

$P(D|H)$ Probability of seeing this data, given the (null) hypothesis

Bayesian Approach to Statistics

$P(H|D)$ Probability of a given outcome, given this data

Jeremy Fox

DIFFERENCES BETWEEN BAYESIANS AND NON-BAYESIANS WHAT IS FIXED?

FREQUENTIST

Data are a repeatable random sample

- there is a frequency

Underlying parameters remain constant during this repeatable process

Parameters are fixed

BAYESIAN

Data are observed from the realized sample.

Parameters are unknown and described probabilistically

Data are fixed

http://www.stat.ufl.edu/~casella/Talks/BayesRefresher.pdf

BAYES' THEOREM

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

A key **benefit**: The ability to incorporate prior knowledge

*A key **weakness**: The need to incorporate prior knowledge*

MATTHEWS 1998

"…different people could use Bayes's Theorem and get different results"

"Faced with the same experimental evidence for, say, ESP, true believers could use Bayes's Theorem to claim that the new results implied that telepathy is almost certainly real.

Skeptics, in contrast, could use Bayes's Theorem to insist they were still not convinced."

"Both views are possible because Bayes's Theorem shows only how to alter one's prior level of belief -- and different people can start out with different opinions."

http://junksciencearchive.com/sep98/matthews.html

MATTHEWS 1998

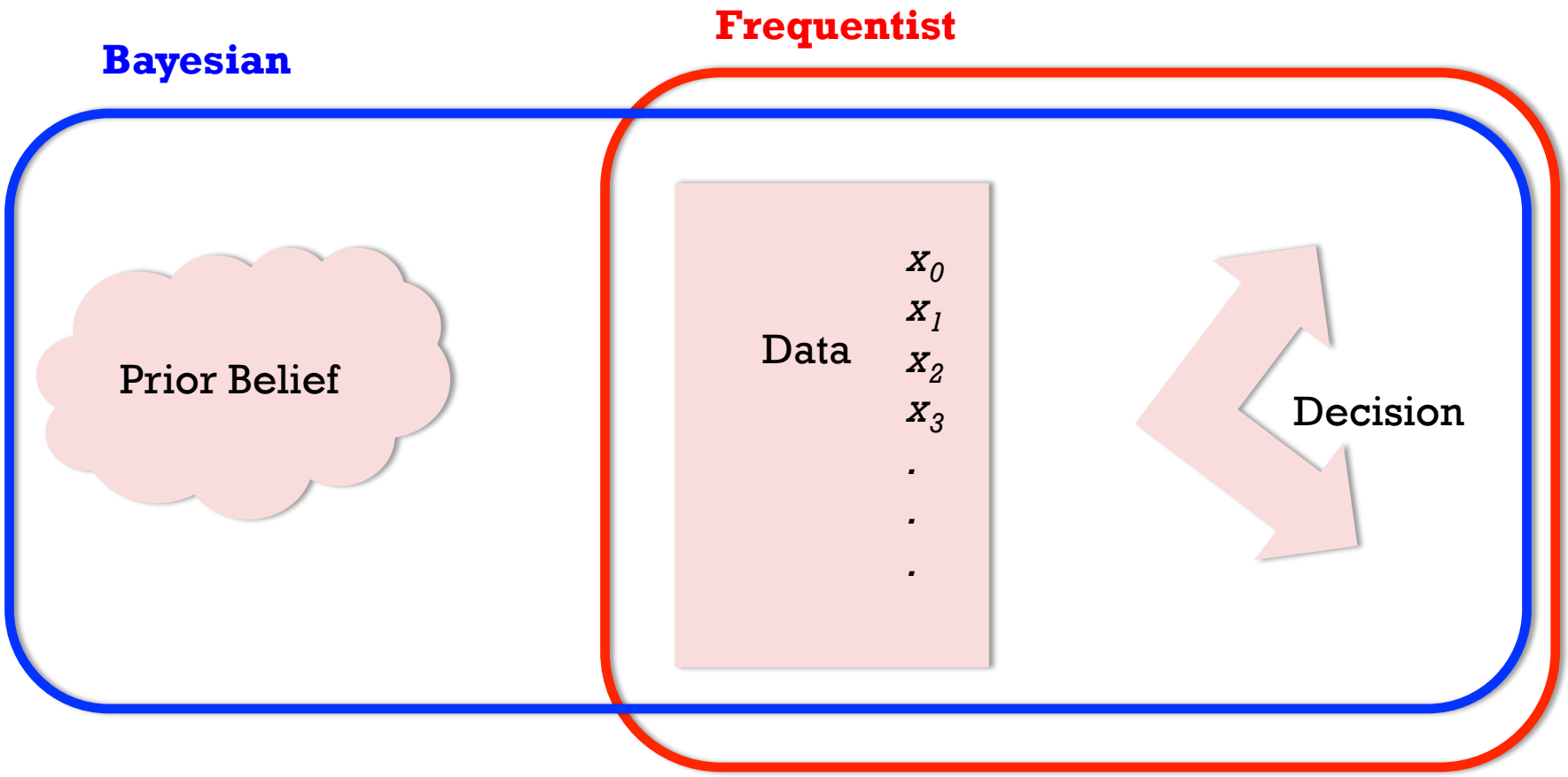
"Fisher had achieved what Bayes claimed was impossible: he had found a way of judging the "significance" of experimental data entirely objectively. That is, he had found a way that anyone could use to show that a result was too impressive to be dismissed as a fluke."

"All scientists had to do, said Fisher, was to convert their raw data into … a P-value"

"So just what were the brilliant insights that led Fisher to choose that talismanic figure of 0.05, on which so much scientific research has since stood or fallen ? Incredibly, as Fisher himself admitted, there weren't any. He simply decided on 0.05 because it was mathematically convenient."

"Professor James Berger of Purdue University - a world authority on Bayes's Theorem - published a entire series of papers again warning of the "astonishing" tendency of Fisher's P-values to exaggerate significance. Findings that met the 0.05 standard, said Berger, "Can actually arise when the data provide very little or no evidence in favour of an effect"."

http://junksciencearchive.com/sep98/matthews.html


Frequentist
Bayesian
Prior Belief
Data
x_0
x_1
x_2
x_3
.
.
.
Decision

BAYES THEOREM

Likelihood
Probability of collecting this data when our hypothesis is true

Prior
The probability of the hypothesis being true before collecting data

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

	A	$\sim A$
B	$P(A \cap B)$	
$\sim B$		

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

QUESTION

1% of women at age forty who participate in routine screening have breast cancer.

80% of women with breast cancer will get positive mammographies.

9.6% of women without breast cancer will also get positive mammographies.

A woman in this age group had a positive mammography in a routine screening.

What is the probability that she actually has breast cancer?

http://yudkowsky.net/rational/bayes

	Have Cancer (1%)	**Do Not Have Cancer (99%)**
Positive Test	True Positive: 1% * 80%	False Positive: 99% * 9.6%
Negative Test	False Negative: 1% * 20%	True Negative: 99% * 90.4%

$$P(cancer \mid positive\ test) = \frac{P(positive\ test \mid cancer)P(cancer)}{P(positive\ test)}$$

$$P(positive\ test) = P(positive\ test \mid cancer)P(cancer) + P(positive\ test \mid no\ cancer)P(no\ cancer)$$

$$P(positive\ test) = 0.8(0.01) + 0.096(0.99) = 0.103$$

$$P(cancer) = 0.01$$

$$P(positive\ test \mid cancer) = 0.8$$

$$P(cancer \mid positive\ test) = \frac{0.8(0.01)}{0.103} = 0.078$$

SPAM FILTERING WITH NAÏVE BAYES

$$P(spam|words) = \frac{P(spam)P(spam|words)}{P(words)}$$

$$P(spam|viagra, rich, ..., friend) = \frac{P(spam)P(viagra, rich, ..., friend|spam)}{P(viagra, rich, ..., friend)}$$

$$\begin{aligned} & P(spam)P(viagra, rich, ..., friend|spam) \\ & \propto P(spam)P(viagra|spam)P(rich, ..., friend|spam, viagra) \\ & \propto P(spam)P(viagra|spam)P(rich|spam, viagra)P(..., friend|spam, viagra, rich) \end{aligned}$$

...

$$P(viagra|spam)P(rich|spam)...P(friend|spam)$$

SLIDES CAN BE FOUND AT:

TEACHINGDATASCIENCE.ORG