POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Week 7-1

Statistics for Big Data

Big Data

Prof. Hwanjo Yu
POSTECH

Hypothesis test

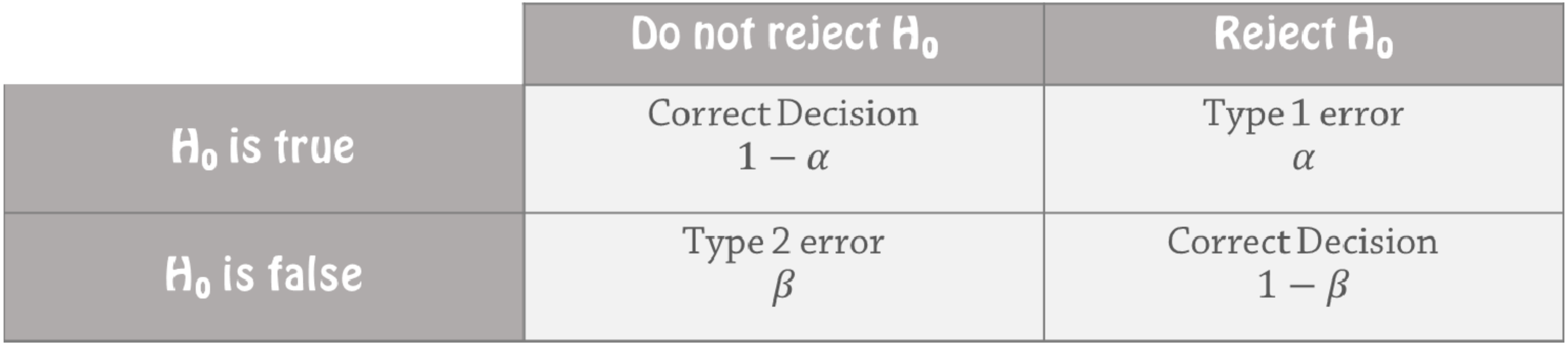
- **Background: Statistical Inference**
    - Methods for drawing conclusions about a population from sample data

Hypothesis test

- Compare an experimental group and a control group
- H_0: Null Hypothesis
  - No difference between the group
- H_A: Alternative Hypothesis
  - Statistically significant difference between the groups
- “difference” defined in terms of some test statistic
  - Different means (e.g. t-test), different variances (e.g. F-test)
- Groups defined through careful experimental design
  - randomized, blinded, double-blinded
- **Examples:**
  - “The new ad placement produces more click-throughs”
  - “This treatment produces better outcomes”

Hypothesis test

- H_0: No difference between the group
- H_A: Statistically significant difference between the groups



	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision $1-\alpha$	Type 1 error α
H_0 is false	Type 2 error β	Correct Decision $1-\beta$

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 when the null hypothesis is true?

For example,

When control group = experimental group, if we repeat the experiments again and again, how likely you will see that control group != experimental group by chance?

Hypothesis test: *one-sided or two sided?*

If the test is two-sided (or two-tailed):

- H_A is $\mu \neq \mu_0$
- P-value = $2 * \mathrm{P}(\ \mathrm{X} > |\text{population mean}|\ )$

If the test is one-sided (or one-tailed):

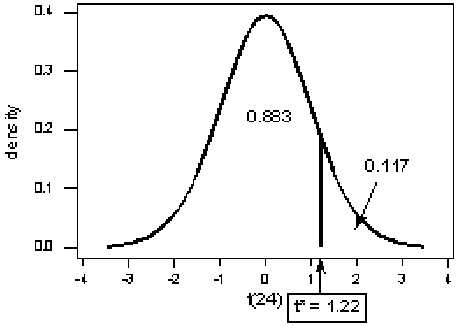
- H_A is $\mu > \mu_0$
- P-value = $\mathrm{P}(\ \mathrm{X} > \text{population mean}\ )$

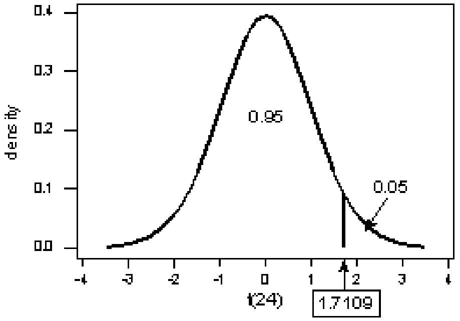
- H_A is $\mu < \mu_0$
- P-value = $\mathrm{P}(\ \mathrm{X} < \text{population mean}\ )$

One sample *t*-test: *one-sided*

$$H_0: \mu = 170$$
$$H_A: \mu > 170$$

- $\mu = 172.52, n = 25, \sigma = 10.31, SE = \frac{\sigma}{\sqrt{n}} = 2.06$
- $t = \frac{\mu - \mu_0}{SE} = 1.22$ and degree of freedom $df = n - 1 = 24,$ then p-value = 0.117
- $t > 1.7109$ (with $df = 24$) for p-value < 0.05 (to reject the nul hypothesis)
- Typical significance level $\alpha = 0.05$ (Why? No good reason)
  - ✓ Lower df => smoother pdf: Smaller sample size requires farther t-score to have lower p-value.



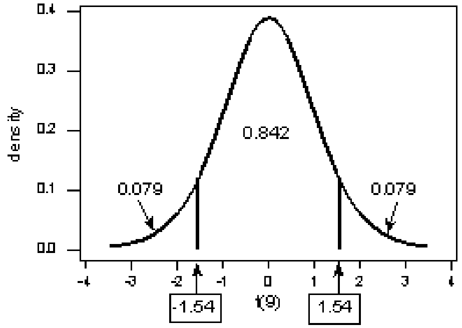


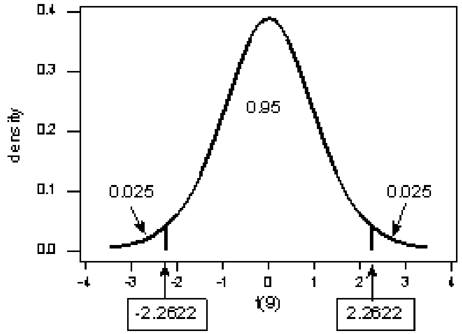
One sample *t*-test: *two-sided*

$$H_0 : \mu = 7.5$$
$$H_A : \mu \neq 7.5$$

- $\mu = 7.55, n = 10, \sigma = 0.1027, SE = \frac{\sigma}{\sqrt{n}} = 0.0325$
- $t = \frac{\mu - \mu_0}{SE} = 1.54$ and $df = n - 1 = 9$, then p-value = 0.158
- $t > 2.2622$ or $t < -2.2622$ (with $df = 9$) for p-value < 0.05 (to reject the null hypothesis)
- 95% (=1-p) confidence interval:
  - $\mu = 7.55 \pm 2.2622 * SE = 7.55 \pm 0.0735$
  - $7.4765 \leq \mu \leq 7.6235$

✓ If the confidence interval includes the null hypothesis mean, cannot reject the null hypothesis.





Independent two sample *t*-test

- $t = \frac{\mu_1 - \mu_2}{\sigma_p}$ where $\sigma_p = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- $df = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\left(s_1^2/n_1\right)^2/(n_1-1) + \left(s_2^2/n_2\right)^2/(n_2-1)}$

- Once t and df are computed, p-value can be computed.

Effect size (Cohen d)

- P-value tells whether the result is significant or not.
- Effect size tells how significant the result is or how much difference the result makes from H_0.
- Effect size = $\dfrac{\text{[Mean of experimental group] – [Mean of control group]}}{\text{standard deviation}}$ (e.g. (325 – 329) / std)
- Effect size 1 means their means differ by one standard deviation.
- small = 0.20
- medium = 0.50
- large = 0.80

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

Effect size

- Standardized Mean Difference

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

Lots of ways to estimate the pooled standard deviation

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

Glass, 1976

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

e.g. Hartung et. al., 2008

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.

Publication bias

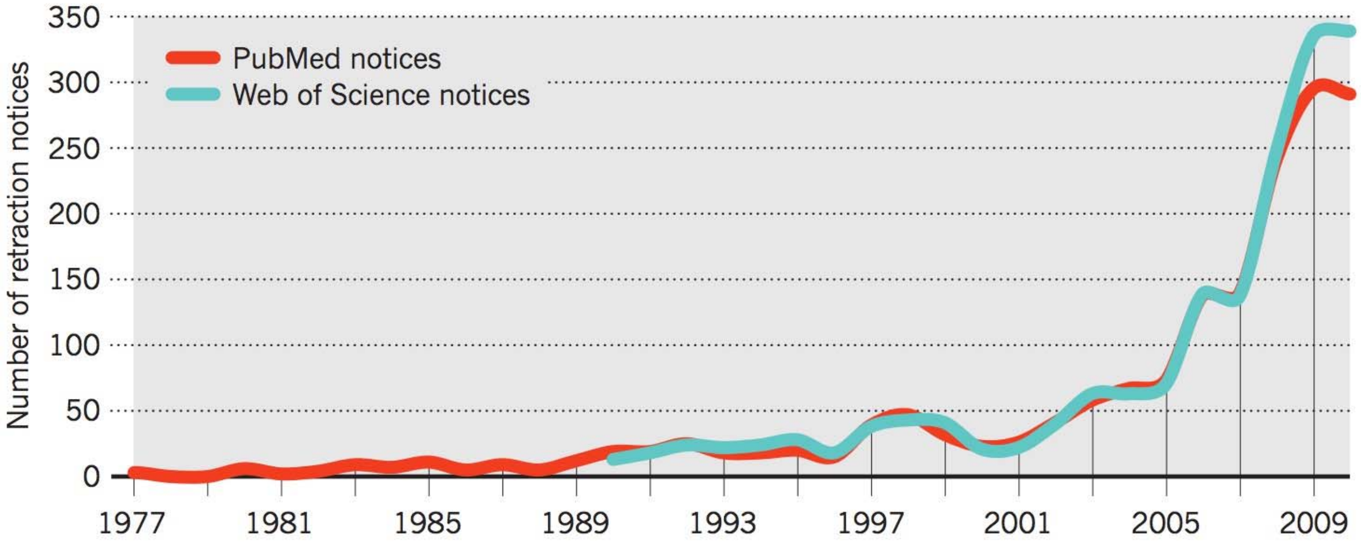
> “In the last few years, several meta-analyses have reappraised the efficacy and safety of antidepressants and concluded that the therapeutic value of these drugs may have been significantly overestimated.”

> “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.”

Publication bias

- When p-value = 0.05, 1 / 20 experiments will show H_A (positive results).
- Only positive results tend to be published.
- H_0(Null hypothesis) is not likely to be published.
- Publication bias → Type 1 error

Mistakes and fraud



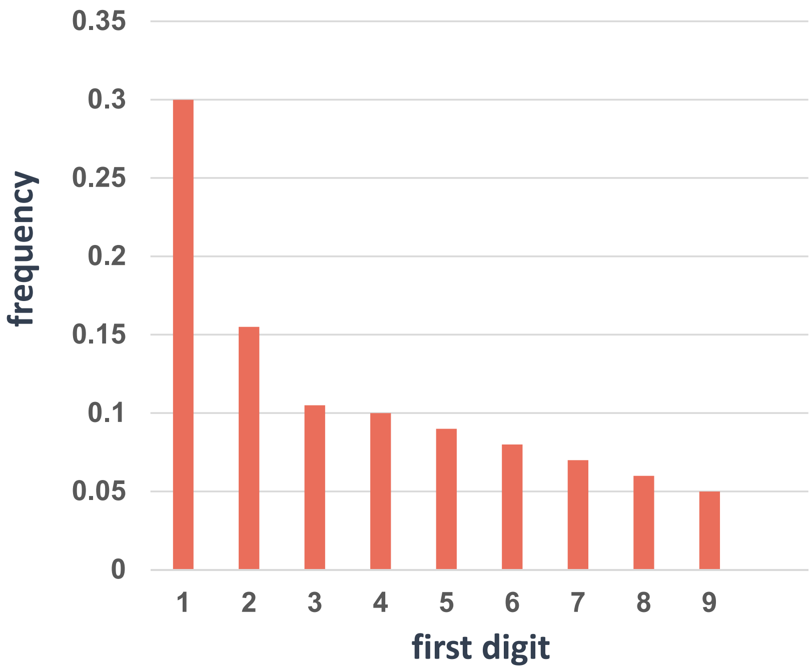
2001 – 2011:

- 10X increase in retractions
- only 1.44X increase in papers

Richard Van Noorden, The Rise of the Retractions, 2011, Nature 478
http://www.nature.com/news/2011/111005/pdf/478026a.pdf

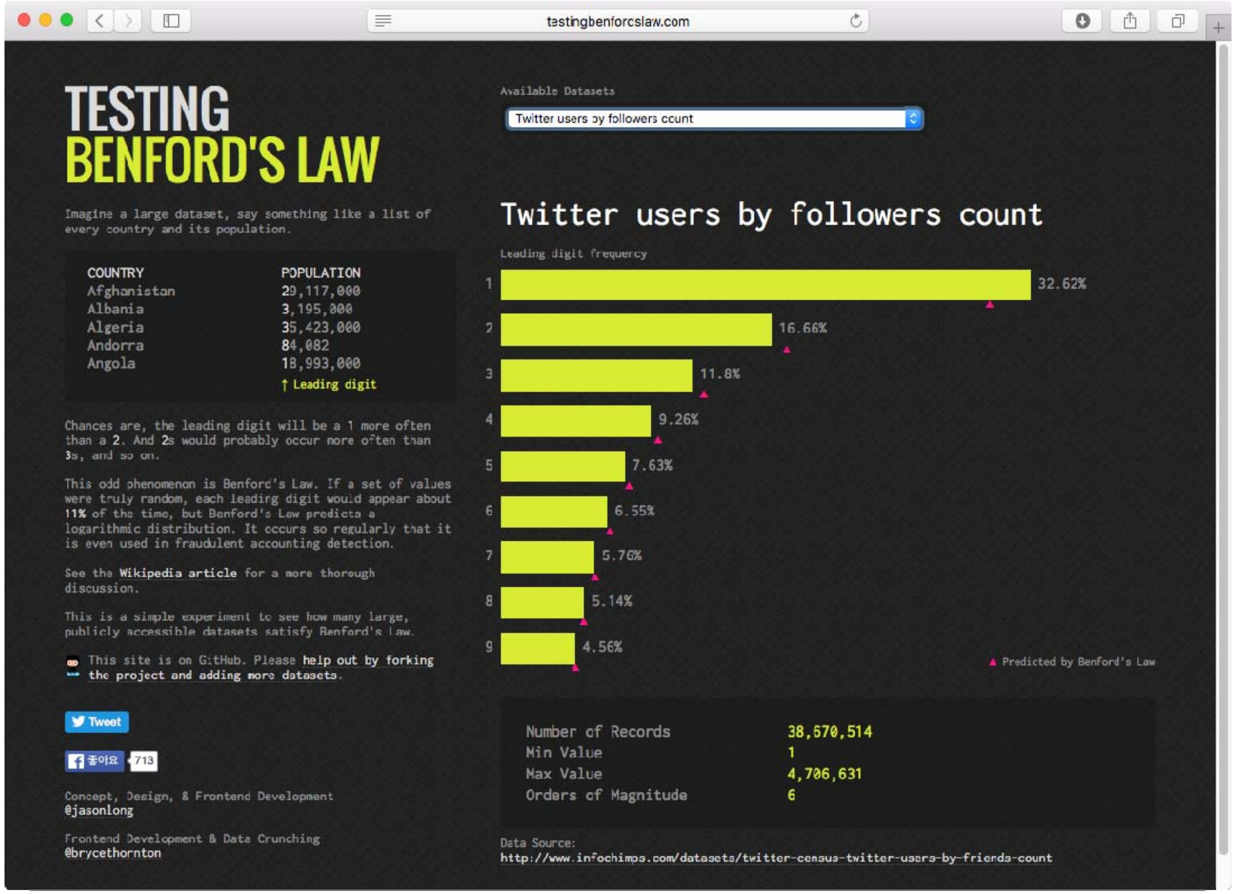
Benford's law: Potential tool for fraud detection

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



Bill Howe, coursera, "Benford's Law",
https://www.coursera.org/learn/predictive-analytics/home/week/1,(2016.06.30)

Benford’s law: Potential tool for fraud detection



TESTING BENFORD'S LAW

Imagine a large dataset, say something like a list of every country and its population.

COUNTRY	POPULATION
Afghanistan	29,117,000
Albania	3,195,000
Algeria	35,423,000
Andorra	84,082
Angola	18,993,000
	↑ Leading digit

Chances are, the leading digit will be a 1 more often than a 2. And 2s would probably occur more often than 3s, and so on.

This odd phenomenon is Benford's Law. If a set of values were truly random, each leading digit would appear about 11% of the time, but Benford's Law predicts a logarithmic distribution. It occurs so regularly that it is even used in fraudulent accounting detection.

See the Wikipedia article for a more thorough discussion.

This is a simple experiment to see how many large, publicly accessible datasets satisfy Benford's Law.

This site is on GitHub. Please help out by forking the project and adding more datasets.

Concept, Design, & Frontend Development
@jasonlong

Frontend Development & Data Crunching
@brycethornton

Available Datasets
Twitter users by followers count

Twitter users by followers count

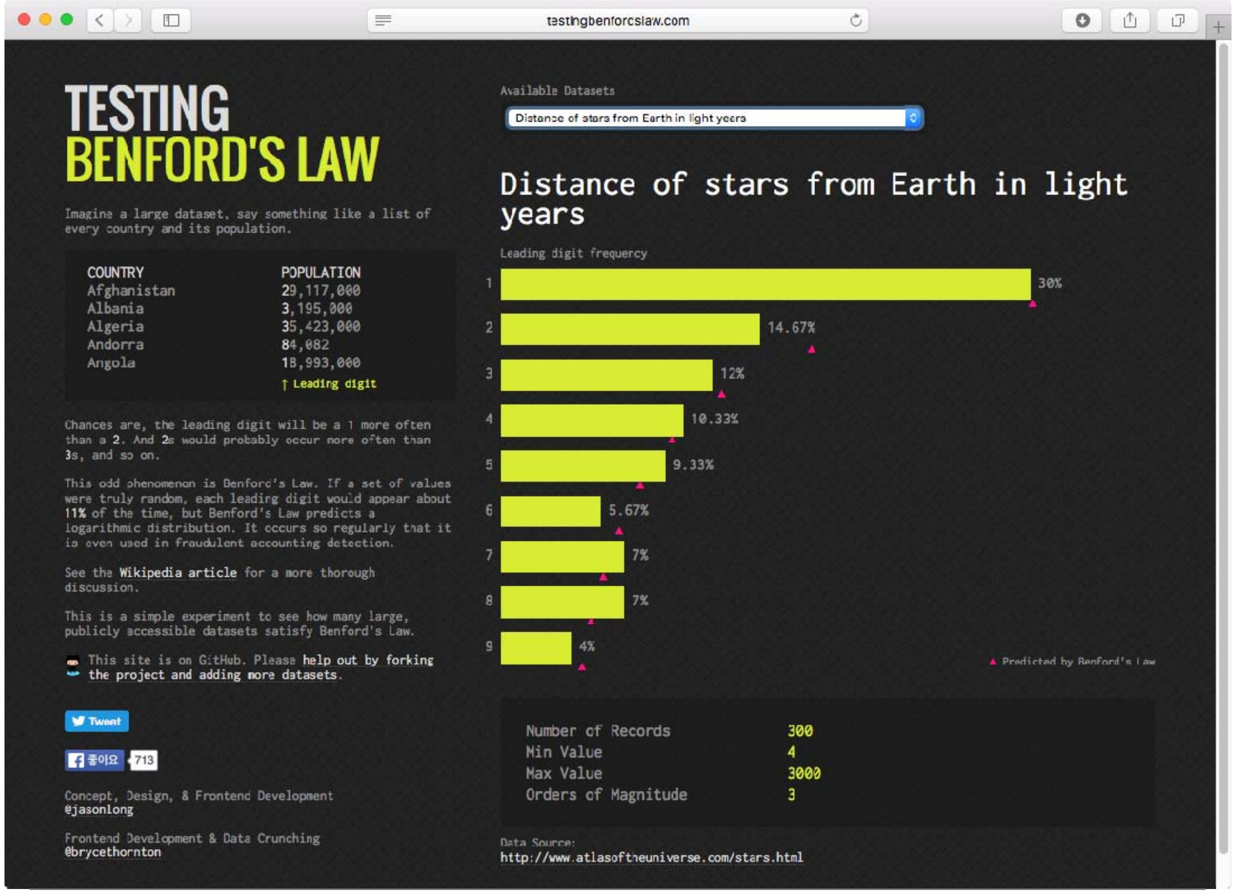
Leading digit frequency

Number of Records	38,670,514
Min Value	1
Max Value	4,706,631
Orders of Magnitude	6

Data Source:
http://www.infochimps.com/datasets/twitter-census-twitter-users-by-friends-count

testingbenfordslaw.com/x, http://testingbenfordslaw.com/twitter-users-by-followers-count ((July 14, 2016)

Benford’s law: Potential tool for fraud detection



TESTING BENFORD'S LAW

Imagine a large dataset, say something like a list of every country and its population.

COUNTRY	POPULATION
Afghanistan	29,117,000
Albania	3,195,000
Algeria	35,423,000
Andorra	84,082
Angola	18,993,000
	↑ Leading digit

Chances are, the leading digit will be a 1 more often than a 2. And 2s would probably occur more often than 3s, and so on.

This odd phenomenon is Benford's Law. If a set of values were truly random, each leading digit would appear about 11% of the time, but Benford's Law predicts a logarithmic distribution. It occurs so regularly that it is even used in fraudulent accounting detection.

See the Wikipedia article for a more thorough discussion.

This is a simple experiment to see how many large, publicly accessible datasets satisfy Benford's Law.

This site is on GitHub. Please help out by forking the project and adding more datasets.

Concept, Design, & Frontend Development
@jasonlong

Frontend Development & Data Crunching
@brycethornton

Available Datasets

Distance of stars from Earth in light years

Distance of stars from Earth in light years

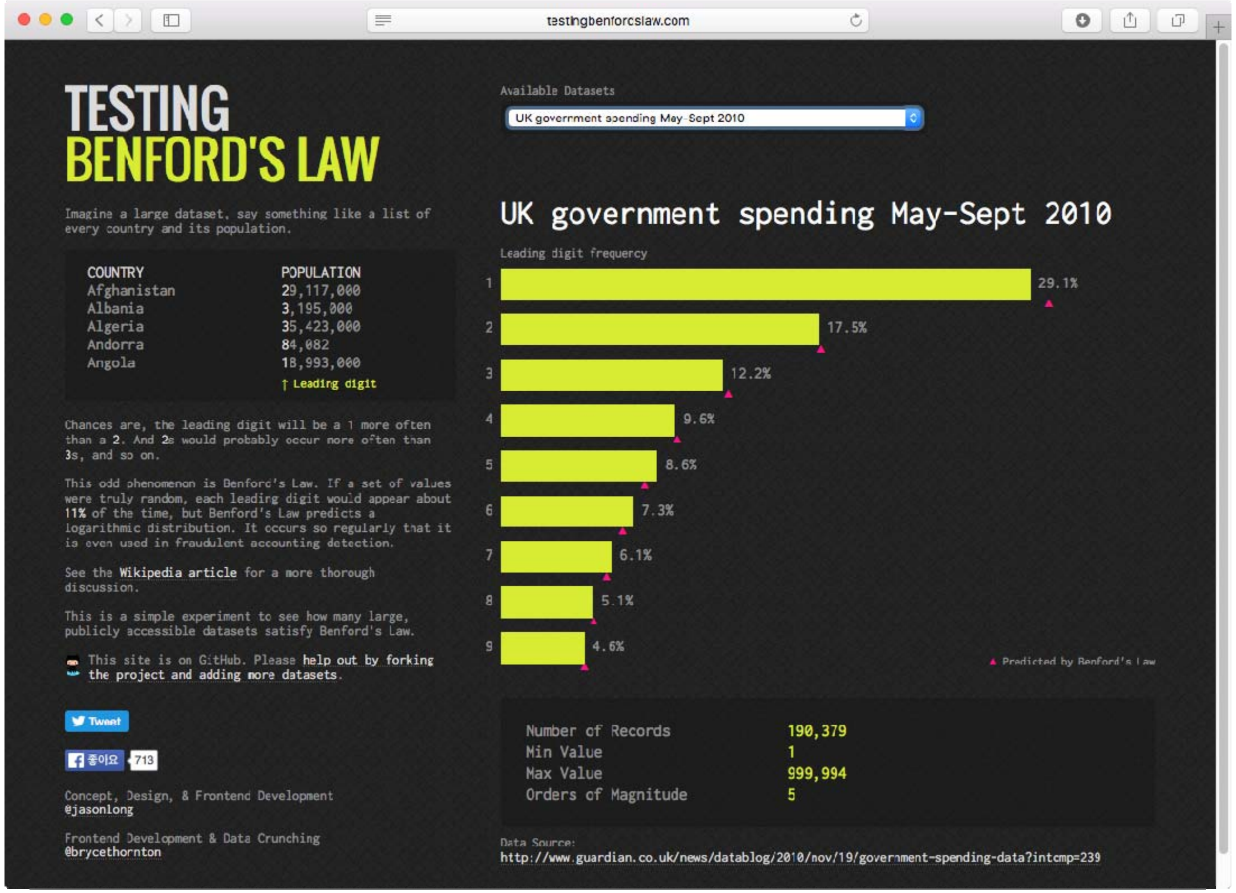
Leading digit frequency

Number of Records	300
Min Value	4
Max Value	3000
Orders of Magnitude	3

Data Source:
http://www.atlasoftheuniverse.com/stars.html

testingbenfordslaw.com/x, http://testingbenfordslaw.com/distance-of-stars-from-earth-in-light-years ((July 14, 2016)

Benford’s law: Potential tool for fraud detection



TESTING BENFORD'S LAW

Imagine a large dataset, say something like a list of every country and its population.

COUNTRY	POPULATION
Afghanistan	29,117,000
Albania	3,195,000
Algeria	35,423,000
Andorra	84,082
Angola	18,993,000

↑ Leading digit

Chances are, the leading digit will be a 1 more often than a 2. And 2s would probably occur more often than 3s, and so on.

This odd phenomenon is Benford's Law. If a set of values were truly random, each leading digit would appear about 11% of the time, but Benford's Law predicts a logarithmic distribution. It occurs so regularly that it is even used in fraudulent accounting detection.

See the Wikipedia article for a more thorough discussion.

This is a simple experiment to see how many large, publicly accessible datasets satisfy Benford's Law.

This site is on GitHub. Please help out by forking the project and adding more datasets.

Concept, Design, & Frontend Development
@jasonlong

Frontend Development & Data Crunching
@brycethornton

UK government spending May-Sept 2010

Number of Records	190,379
Min Value	1
Max Value	999,994
Orders of Magnitude	5

Data Source:
http://www.guardian.co.uk/news/datablog/2010/nov/19/government-spending-data?intcmp=239

testingbenfordslaw.com/x, http://testingbenfordslaw.com/uk-government-spending-may_to_sept-2010 ((July 14, 2016)

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 anomalous, but the second and third digits deviated from expected distributions significantly. (There are also Benford’s formula for the second or third digits.)

Andreas Diekmann, 2007, Journal of Applied Statistics, 34(3)

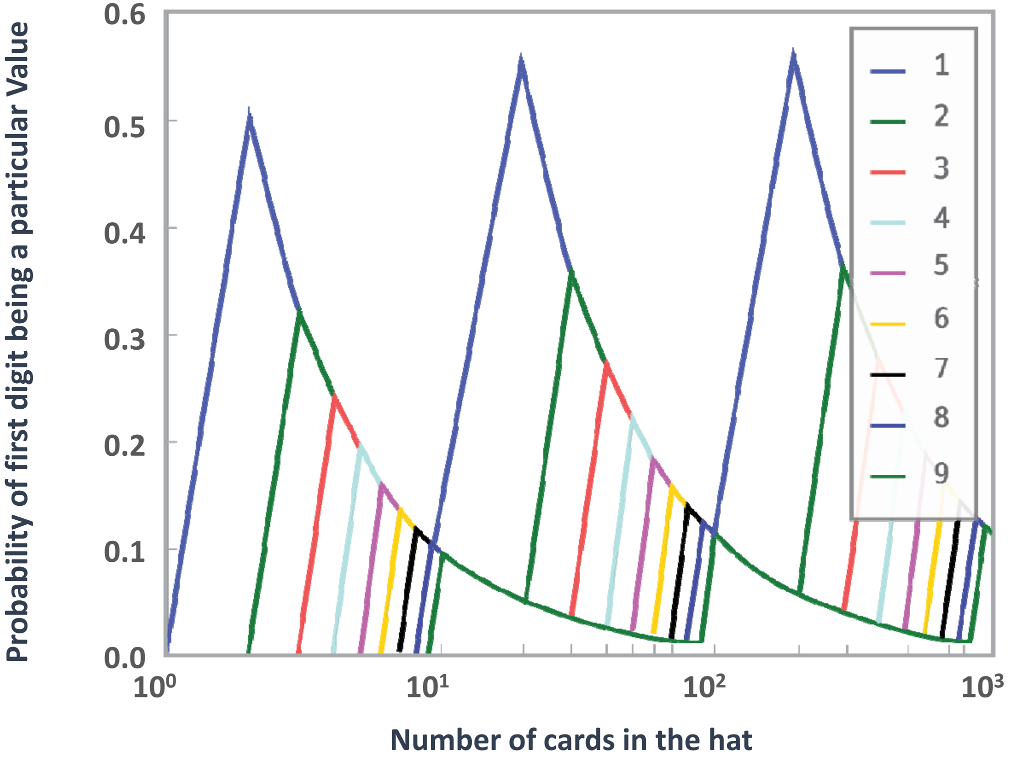
Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data

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 a card where the first digit is 1?"

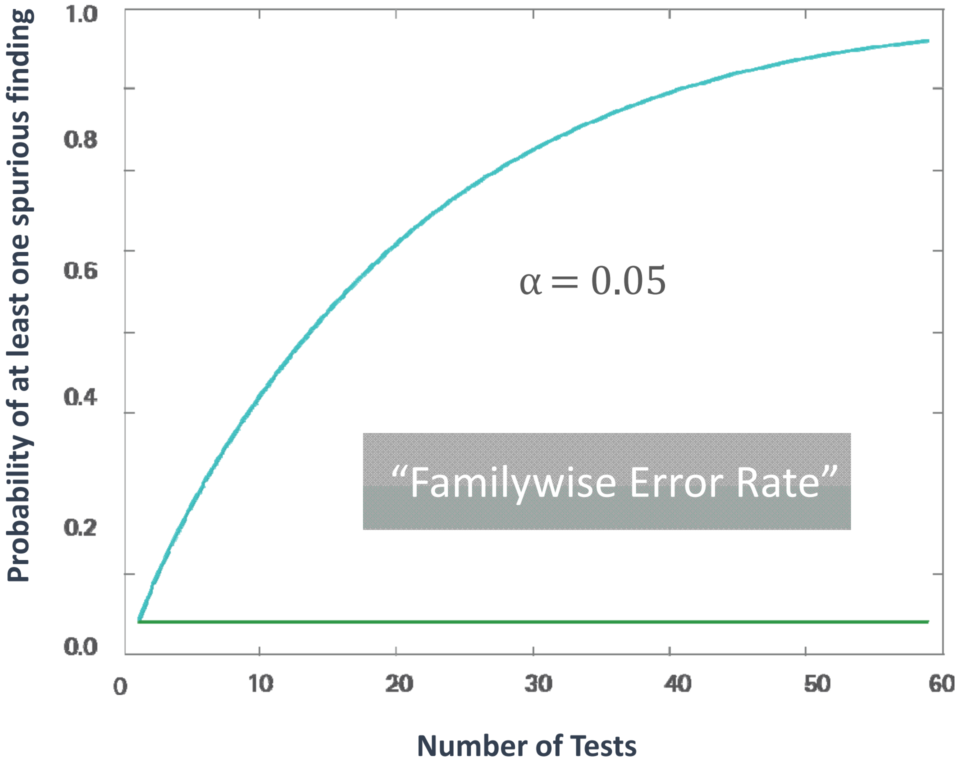
Benford’s law intuition



Bill Howe, coursera, “Benford’s Law”,
https://www.coursera.org/learn/predictive-analytics/home/week/1, 2016.06.30

Multiple hypothesis testing

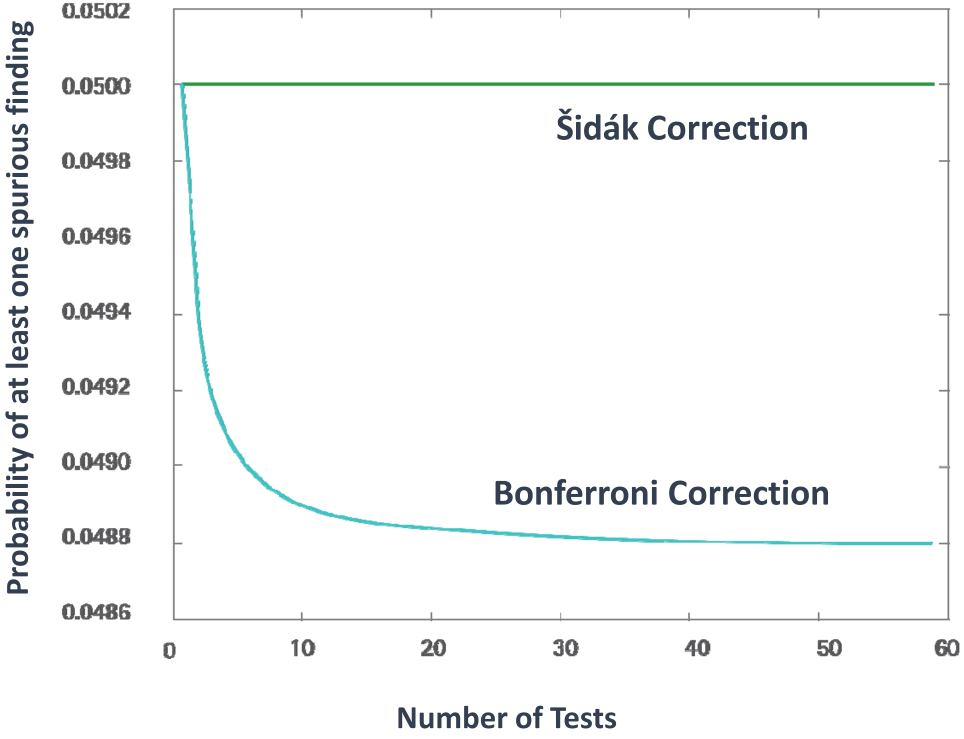
- P(detecting an effect when there is none) = $0.05(= \alpha)$
- P(not detecting an effect when there is none) = $1 - \alpha$
- P(not detecting an effect when there is none on every experiment) = $(1 - \alpha)^k$
- P(detecting an effect when there is none on at least one experiment) = $1 - (1 - \alpha)^k$



Bill Howe, coursera, "Multiple Hypothesis Testing", https://www.coursera.org/learn/predictive-analytics/home/week/1, 2016.06.30

Familywise error rate corrections

- **Bonferroni Correction**
  - Divide by the number of hypotheses
  - $\boldsymbol{\alpha_c = \frac{\alpha}{k}}$
- **Šidák Correction**
  - Asserts independence
  - $\boldsymbol{\alpha = 1 - (1 - \alpha_c)^k}$
  - $\boldsymbol{\alpha_c = 1 - (1 - \alpha)^{\frac{1}{k}}}$



Bill Howe, coursera, “Multiple Hypothesis Testing”, https://www.coursera.org/learn/predictive-analytics/home/week/1, 2016.06.30

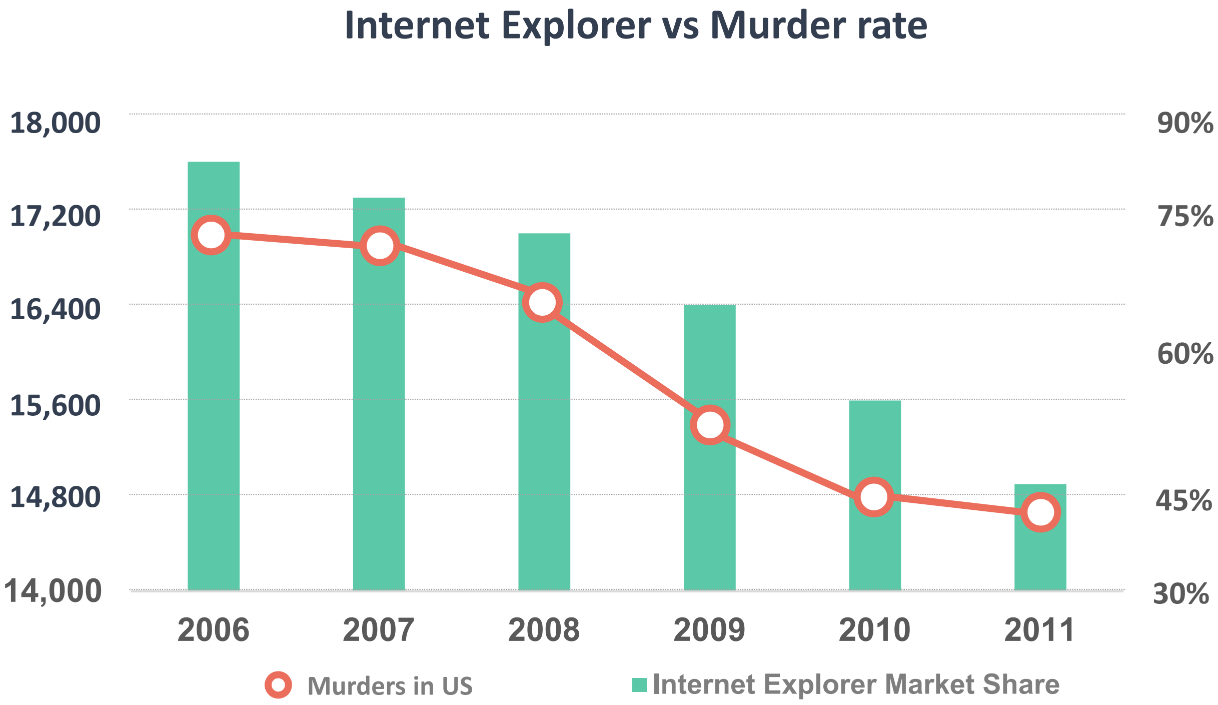
What about big data?

“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

What about big data?



Positive correlation

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

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

The “Curse” of big data?

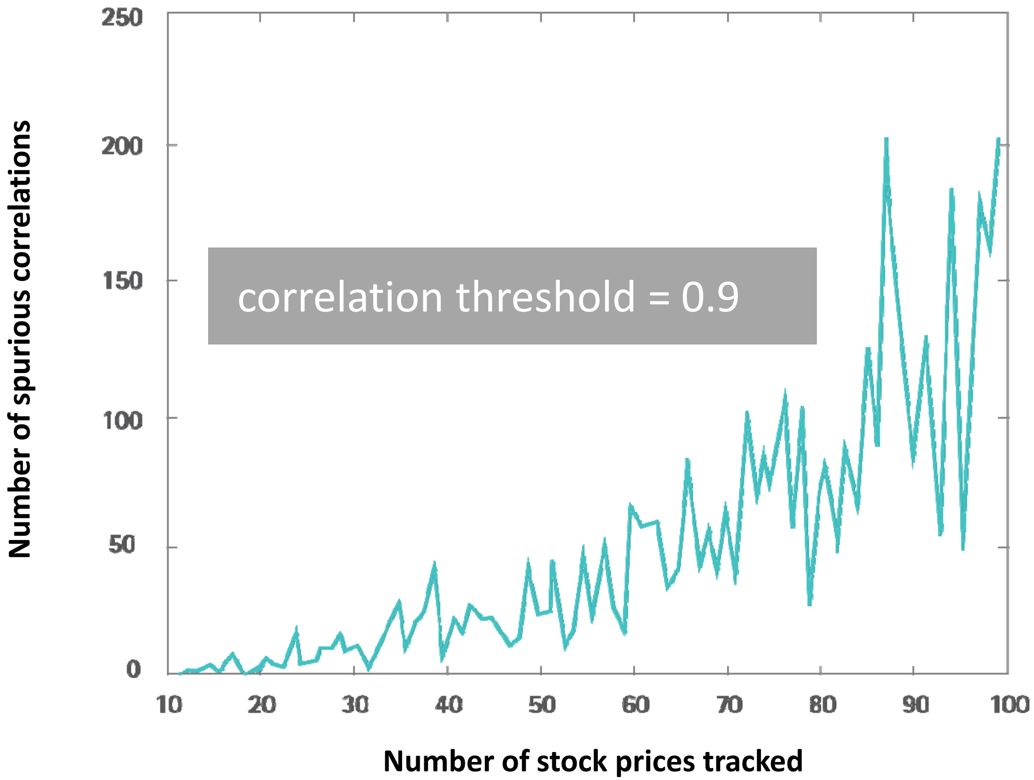
“…the curse of big data is the fact that when you search for patterns in very, very large data sets with billions or trillions of data points and thousands of metrics, **you are bound to identify coincidences that have no predictive power**.”

Vincent Granville

Vincent Granville's example

- Consider stock prices for 500 companies over a 1-month period
- Check for correlations in all pairs using Pearson's correlation.

Vincent Granville’s example



Bill Howe, coursera, “Stock Price Example”,
https://www.coursera.org/learn/predictive-analytics/home/week/1, 2016.06.30

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!
- While > N decreases variance, a wrong sampling amplifies bias
  - E.g. You log all clicks to your website to model user behavior, but this only samples current users, not the users you want to attract.
  - E.g. Using mobile data to infer buying behavior
- **Beware multiple hypothesis tests**
  - “Green jelly beans cause acne”