

Estimating Average Proportional Changes in Large, Sparse Data

**Ryan Giordano
& many others at Google**

Credits

Many people at Google have thought / worked on the problems described here, especially:

- Daryl Pregibon
- Hal Varian
- Matt Cary
- Chris Neff

All errors in this presentation are my own, and any views herein are not necessarily Google's.

Outline

- The problem statement
 - A motivating example from internet advertising
 - The general problem statement
 - The nature of our "big" data
- Some techniques that won't work (and why)
 - Ratio of ratios
 - Average of logs
 - Random effects
- The Mantel-Haenszel ("MH") estimator
 - Classical form
 - Generalization

Motivating Example: Internet Advertising

Google mesothelioma

Web Images Maps Shopping

About 10,000 results (0.26 seconds)

Ads related to mesothelioma

Mesothelioma 1 (866) 942 9878
www.mesotheliomaclaimscenter.info/
Mesothelioma? Get Money You Deserve Fast! G
Mesothelioma Compensation Amounts - File a M

Mesothelioma 1 (877) 770 9704
www.sokolovelaw.com/Call_Now
Eligible For A Mesothelioma Claim? Free Legal Consult! Call 24/7.

Mesothelioma 1 (800) 582 0706
www.lawfirm.com/Free_Consult
Do you Have a Mesothelioma Lawsuit? Get the Settlement You Deserve.
9 Tips For Choosing a Meso Law Firm - What To Expect Filing A Law Suit

Mesothelioma - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mesothelioma ▾
Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Mesothelium - Peritoneal mesothelioma - Category:Mesothelioma

Simmons Law Firm has 199 followers on Google+

Mesothelioma Claim Center
www.mesothelioma-asbestos-law-firm.com/
Millions of Dollars Available!
Get A 100% Free Consultation Today.

Mesothelioma Cancer
www.mesothelioma-lung-cancer.org/
1 (855) 284 2319
Site covers a specific and rare type of lung cancer

Mesothelioma Resource
www.mesotheliomabook.com/

A "commercial" web search shows advertisements.

If the user clicks on one, the advertiser is then charged \$X, and the user is sent to their site.

Motivating Example: Internet Advertising

Google Version 1
(Control)

A screenshot of a Google search for "mesothelioma" showing organic search results and a sidebar of ads. The ads include "CA Mesothelioma Legal Aid", "Mesothelioma Claim Center", "Mesothelioma Cancer", and "Mesothelioma Resource".

Google mesothelioma Ryan Giordano

Web Images Maps Shopping More Search tools

About 8,050,000 results (0.26 seconds)

Ads related to mesothelioma

Mesothelioma 1 (866) 942 9878
www.mesotheliomacclaimcenter.info/
Mesothelioma? Get Money You Deserve Fast! Get Help with Filing a Claim.
Mesothelioma Compensation Amounts - File a Mesothelioma Claim

Mesothelioma 1 (877) 770 9704
www.sokolovlaw.com/Call_Now
Eligible For A Mesothelioma Claim? Free Legal Consult! Call 24/7.

Mesothelioma 1 (800) 582 0706
www.lawfirm.com/Free_Consult
Do you Have a Mesothelioma Lawsuit? Get the Settlement You Deserve.
9 Tips For Choosing a Meso Law Firm - What To Expect Filing A Law Suit

Mesothelioma - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mesothelioma

Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Mesothelium - Peritoneal mesothelioma - Category:Mesothelioma

Ads

CA Mesothelioma Legal Aid
www.simmonsfirm.com/Mesothelioma
1 (888) 350 4189
Mesothelioma Diagnosis in CA? Get A Free Legal Consult. Call 24/7
Simmons Law Firm has 199 followers on Google+

Mesothelioma Claim Center
www.mesothelioma-asbestos-law-firm.com/
Millions of Dollars Available!
Get A 100% Free Consultation Today.

Mesothelioma Cancer
www.mesothelioma-lung-cancer.org/
1 (855) 284 2319
Site covers a specific and rare type of lung cancer

Mesothelioma Resource
www.mesotheliomabook.com/

Google Version 2
(Treatment)

A screenshot of a Google search for "mesothelioma" showing organic search results and a sidebar of ads. The ads include "CA Mesothelioma Legal Aid", "Mesothelioma Claim Center", "Mesothelioma Cancer", and "Mesothelioma Resource".

Google mesothelioma Ryan Giordano

Web Images Maps Shopping More Search tools

About 8,050,000 results (0.26 seconds)

Ads related to mesothelioma

Mesothelioma 1 (866) 942 9878
www.mesotheliomacclaimcenter.info/
Mesothelioma? Get Money You Deserve Fast! Get Help with Filing a Claim.
Mesothelioma Compensation Amounts - File a Mesothelioma Claim

Mesothelioma 1 (877) 770 9704
www.sokolovlaw.com/Call_Now
Eligible For A Mesothelioma Claim? Free Legal Consult! Call 24/7.

Mesothelioma 1 (800) 582 0706
www.lawfirm.com/Free_Consult
Do you Have a Mesothelioma Lawsuit? Get the Settlement You Deserve.
9 Tips For Choosing a Meso Law Firm - What To Expect Filing A Law Suit

Mesothelioma - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Mesothelioma

Mesothelioma (or, more precisely, malignant mesothelioma) is a rare form of cancer that develops from cells of the mesothelium, the protective lining that covers ...
Mesothelium - Peritoneal mesothelioma - Category:Mesothelioma

Ads

CA Mesothelioma Legal Aid
www.simmonsfirm.com/Mesothelioma
1 (888) 350 4189
Mesothelioma Diagnosis in CA? Get A Free Legal Consult. Call 24/7
Simmons Law Firm has 199 followers on Google+

Mesothelioma Claim Center
www.mesothelioma-asbestos-law-firm.com/
Millions of Dollars Available!
Get A 100% Free Consultation Today.

Mesothelioma Cancer
www.mesothelioma-lung-cancer.org/
1 (855) 284 2319
Site covers a specific and rare type of lung cancer

Mesothelioma Resource
www.mesotheliomabook.com/

Imagine a randomized A/B study with two different "versions" of Google. An advertiser's average "cost per click" (CPC) may change.

Motivating Example

For advertiser i ,

$$X = CPC \quad (\underline{\mathbf{C}}\text{ost } \underline{\mathbf{P}}\text{er } \underline{\mathbf{C}}\text{lick})$$

$$S = \text{Spend}$$

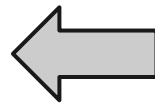
$$N = \text{Clicks}$$

$$X_{1i} = \frac{S_{1i}}{N_{1i}} \quad (\text{Control})$$

$$X_{2i} = \frac{S_{2i}}{N_{2i}} \quad (\text{Treatment})$$

These are random due to:

- User behavior
 - Random searches
 - Random clicks
- Random allocation in the A/B study
- State of Google's systems



Motivating Example

$$X_{1i} = S_{1i}/N_{1i}$$

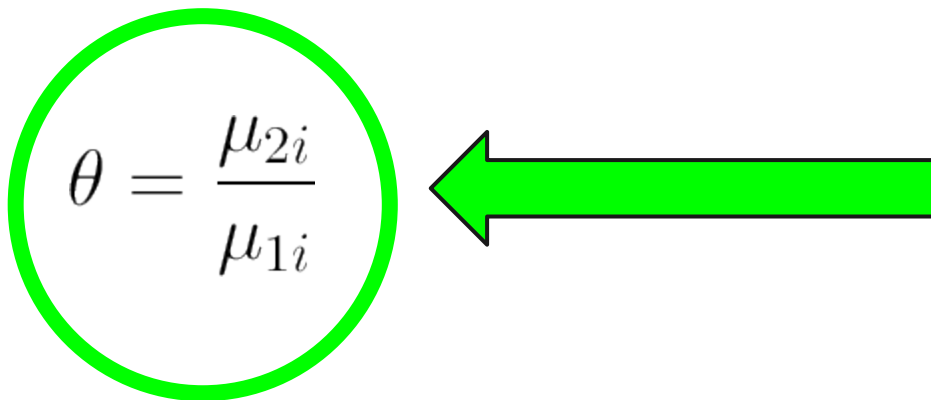
$$E(X_{1i}) = \mu_{1i}$$

$$X_{2i} = S_{2i}/N_{2i}$$

$$E(X_{2i}) = \mu_{2i}$$

$\mu_{.i}$ can range from pennies to hundreds of dollars (!)

=> We care about the *ratio* of the means, not the difference.


$$\theta = \frac{\mu_{2i}}{\mu_{1i}}$$

This is what we're after. For this (short) presentation we'll (mostly) assume it's the same for each i .

Reminder: Ratios Are Harder Than Differences

The difference of means is easier to estimate than the ratio:

$$\frac{1}{N} \sum_i (Y_i - Z_i) \rightarrow E(Y_i) - E(Z_i) = \mu_Y - \mu_Z \quad \text{Difference}$$

$$\frac{1}{N} \sum_i Y_i / Z_i \rightarrow E(Y_i / Z_i) \neq \mu_Y / \mu_Z \quad \text{Ratio}$$

Formal Problem Statement

N paired observations with independent mean-zero noise:

$$(X_{1i}, X_{2i}), 1 \leq i \leq N$$

Paired observations

$$P(X_{1i}, X_{2i} | \mu_i) = P(X_{1i} | \mu_i) P(X_{2i} | \mu_i)$$

Independent

$$E(X_{1i}) = \mu_i$$

$$E(X_{2i}) = \theta \mu_i$$

Different means
...and a proportional
change

We want to know θ .

The Data

We're interested in cases where:

- N is large (40m), data is large (~20Gb+)
- Each pair has little data (zeroes or large variance)
- Simpson's paradox may occur (more later)

We'll (sloppily) require:

- $\theta > 0$
- $\mu_i > 0$
- $\text{var}(X_{1i}) < \infty, \text{var}(X_{2i}) < \infty$
- Sane regularity conditions that will be obvious

Things You Might Try: Outline

Method	Positives	Problem
Ratio of ratios (compare totals)	Easy to calculate, very simple	Simpson's Paradox
Average of logs	Intuitive (logs are for proportions)	Sparse data
Random effects model	Theoretically sound	Data is too big

Things You Might Try #1: Ratio of Ratios

$$X_1 = \frac{S_1}{N_1} = \frac{\sum_i S_{1i}}{\sum_i N_{1i}}$$

Total (unpaired)
CPC in the control

$$X_2 = \frac{S_2}{N_2} = \frac{\sum_i S_{2i}}{\sum_i N_{2i}}$$

Total (unpaired) CPC
in the treatment

$$\hat{\theta} = X_2 / X_1$$

...and their ratio

Problem: **Simpson's Paradox**

Simpson's Paradox Formally

$$\frac{S_1}{N_1} = \frac{\sum_i S_{1i}}{\sum_i N_{1i}} = \sum_i \frac{S_{1i}}{\sum_j S_{1j}} \frac{S_{1i}}{N_{1i}} = \sum_i w_{1i} X_{1i}$$
$$\frac{S_2}{N_2} = \dots \dots = \sum_i w_{2i} X_{2i}$$

The ratios can change with changes in the weights alone (e.g. in the distribution of clicks).

The can mask, simulate, or counteract changes in the X .

Simpson's Paradox Example

Two advertisers:

...one expensive (Adv 1)

...and one cheap (Adv. 2)

	Control (1)	Treatment (2)
Adv. 1	$x_{11} = \$10, \quad w_{11} = 10\%$	$x_{11} = \$9, \quad w_{11} = 90\%$
Adv. 2	$x_{12} = \$1, \quad w_{12} = 90\%$	$x_{12} = \$0.9, \quad w_{12} = 10\%$
Totals:	$x_1 = \$1.9$	$x_2 = \$8.19$

$$\theta = \$9 / \$10 = \$0.9 / \$1 = 0.9$$

But the average goes from \$1.9 in the control to \$8.19 in the treatment because of the change in w (click distribution).

Things You Might Try #2a: Average of Ratios

$$(X_{1i}, X_{2i}), 1 \leq i \leq N$$

$$E(X_{1i}) = \mu_i$$

$$E(X_{2i}) = \theta_i \mu_i$$

$$\hat{\theta} = \frac{1}{N} \sum_i \frac{X_{2i}}{X_{1i}}$$

$$\hat{\theta} \rightarrow E \left[\frac{X_{2i}}{X_{1i}} \right] \neq \frac{E[X_{2i}]}{E[X_{1i}]}$$

Problem:

Linearity of expectations and

Sparse data

(or zeroes)

Things You Might Try #2b: Average of Logs

$$(X_{1i}, X_{2i}), 1 \leq i \leq N$$

$$E(X_{1i}) = \mu_i$$

$$E(X_{2i}) = \theta_i \mu_i$$

$$\hat{\theta} = \frac{1}{N} \sum_i [\log(X_{2i}) - \log(X_{1i})]$$

$$E [\log(X_{2i}) - \log(X_{1i})] \neq \log(E[X_{2i}]) - \log(E[X_{1i}])$$

Problem:

Exactly the same!

**Sparse data
(or zeroes)**

Things You Might Try #3: Random Effects Model

$$X_{1i} \sim N(\mu_i, \sigma_{1i}^2)$$

$$X_{2i} \sim N(\theta \mu_i, \sigma_{2i}^2)$$

$$\mu_i \sim F(\mu_i; \gamma)$$

\Rightarrow

$$E(X_{1i} | \mu_i) = \mu_i$$

$$E(X_{2i} | \mu_i) = \theta \mu_i$$

Use MLE to
estimate θ, γ

Problem:

Requires multiple
passes through
the data.

Data is too big

Classical Mantel Haenszel Estimator

2x2 contingency tables

Unit i	Success	Trials
Control	S_{1i}	N_{1i}
Treatment	S_{2i}	N_{2i}

$$S_{1i} \sim \text{Poisson}(\mu_i \cdot N_{1i})$$

$$S_{2i} \sim \text{Poisson}(\theta \cdot \mu_i \cdot N_{2i})$$

Assume $\theta \approx 1$ to derive MLE of θ :

$$\hat{\theta} = \frac{\sum_i w_i \cdot X_{2i}}{\sum_i w_i \cdot X_{1i}} \quad w_i = \frac{N_{1i} N_{2i}}{N_{1i} + N_{2i}}$$

Classical Mantel Haenszel Estimator

MH:

$$\hat{\theta} = \frac{\sum_i w_i \cdot X_{2i}}{\sum_i w_i \cdot X_{1i}}$$

Ratio of Ratios:

$$\frac{X_2}{X_1} = \frac{\sum_i w_{2i} \cdot X_{2i}}{\sum_i w_{1i} \cdot X_{1i}}$$

Note the formal similarity to the ratio of ratios, but with no Simpson because we've made w the same in the numerator and denominator.

Generalized "MH" Estimator

$$\hat{\theta} = \frac{\sum_i w_i \cdot X_{2i}}{\sum_i w_i \cdot X_{1i}} \rightarrow_p \theta$$

The precise weights don't matter as long as:

- They are the same in the numerator and denominator
- $E[X_{2i}|w_i] = \theta \cdot \mu_i = \theta \cdot E[X_{1i}|w_i]$
- The weights don't do something stupid as $n \rightarrow \infty$

Example, Revisited

Step 1) Group the advertisers into rows:

$$(S_{1i}, N_{1i}, S_{2i}, N_{2i})$$

Step 2) For each row, calculate w_i, X_{1i}, X_{2i}

Step 3) Keep running totals of $w_i \cdot X_{1i}$ and $w_i \cdot X_{2i}$

Step 4) Divide the two totals to get

$$\hat{\theta} = \frac{\sum_i w_i \cdot X_{2i}}{\sum_i w_i \cdot X_{1i}}$$

This is "embarrassingly parallel" (except for step 1, which you'll probably have to do anyway, or you get Simpson).

Beyond the Scope

- Variance is straightforward (e.g. your favorite online bootstrap algorithm)
- Often approximately normal (classical hypothesis tests have good coverage)
- Robust to non-uniformity of the effect
- ...and custom weights give you a weighted average of your choice.

Some Shortcomings

- Can't easily drop into a regression context
- The denominator must be far from zero with high probability
- Potentially inefficient if you have more information

Summary

- Estimating a ratio of averages can be tricky due to:
 - Simpson's paradox (ratio of ratios)
 - Sparse data (average of logs or ratios)
 - Big data (random effects)
- Generalized MH resolves these issues:
 - Very parallelizable
 - Robust to misspecification
 - Robust to Simpson's Paradox
 - Easy to understand

Questions?



Contact Information

Ryan Giordano
rgiordan@gmail.com

Extra Slides

which I probably won't have time to present

Generalized "MH" Estimator

$$\hat{\theta} = \frac{\sum_i w_i \cdot X_{2i}}{\sum_i w_i \cdot X_{1i}} \quad w_i = \frac{N_{1i}N_{2i}}{N_{1i} + N_{2i}}$$

$$E \left[\sum_i w_i \cdot X_{2i} | w_j, m_j, \forall j \right] = \theta \sum_i w_i \cdot m_i$$

$$E \left[\sum_i w_i \cdot X_{1i} | w_j, m_j, \forall j \right] = \sum_i w_i \cdot m_i$$

=> As a LLN kicks in,

$$\hat{\theta} \rightarrow_p \theta$$

Average Proportional Changes

Now suppose that t is not constant:

$$(x_{1i}, x_{2i}), 1 \leq i \leq N$$

$$E(x_{1i}) = m_i$$

$$E(x_{2i}) = t_i * m_i$$

$$P(x_{1i}, x_{2i} | m_i) = P(x_{2i} | m_i) P(x_{1i} | m_i)$$

We want to know the average t_i , weighted by some attribute of the pair, i .

Non-uniform Changes

Suppose

$$t_i \sim f_i(t)$$

Then defining

$$W_i = m_i w_i / \sum_i m_i w_i$$

$$E_i(t_i * W_i)$$

Usually, $m_i * w_i \sim s_i$

Non-uniform Changes

Suppose we don't want spend weighting, but click weighting instead.

Use historical (out-of-sample) data to get

$$x_h = n_h / s_h$$

and use

$$w'_i = x_h * w_i$$

Example with non-uniform changes

$$w_i = \frac{n_{1i} n_{2i}}{n_{1i} + n_{2i}}, \quad x_{1i} = s_{1i} / n_{1i}$$

$$w_i \sim n$$

$$m_i w_i \sim \text{Spend}$$

Result: a spend-weighted average proportional change.