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:

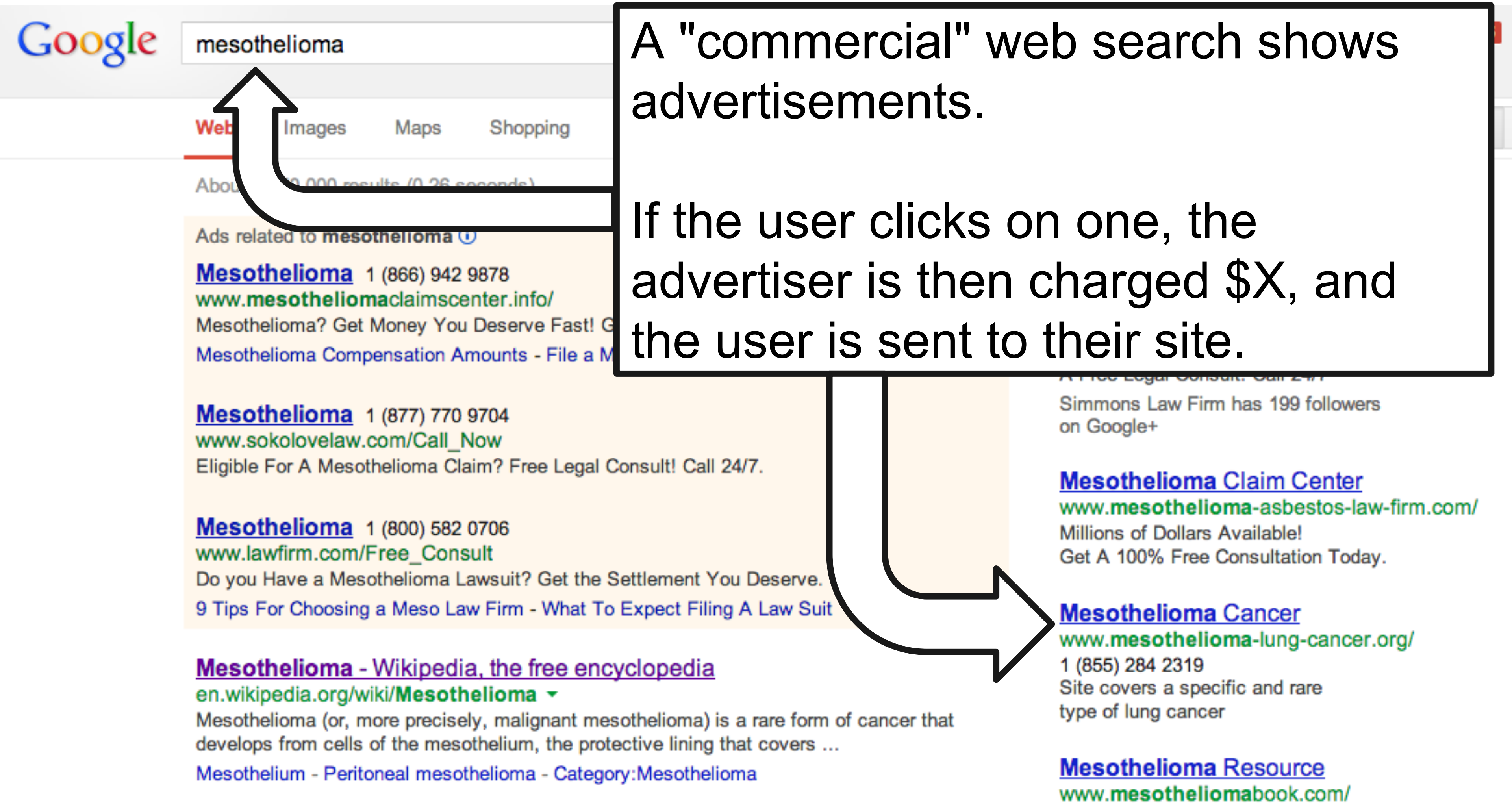
- Daryl Pregibon
- Hal Varian
- Matt Cary
- Chris Neff

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

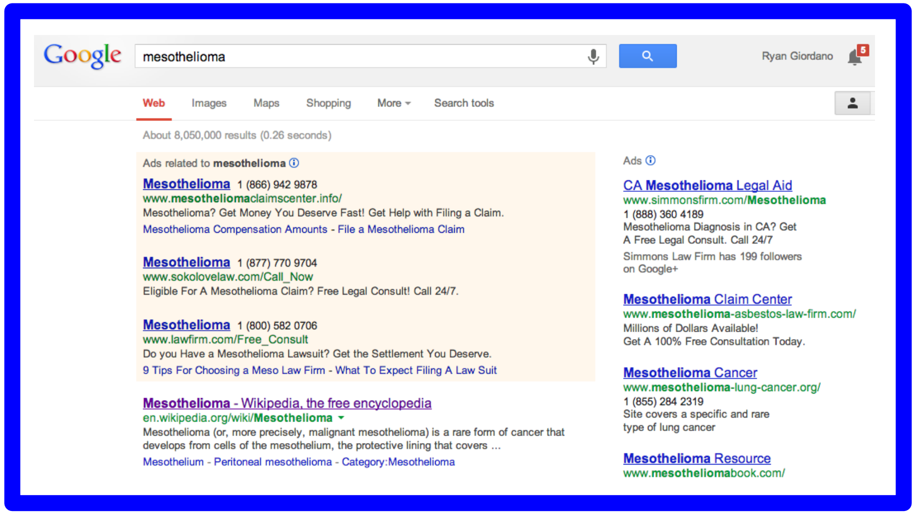


A "commercial" web search shows advertisements.

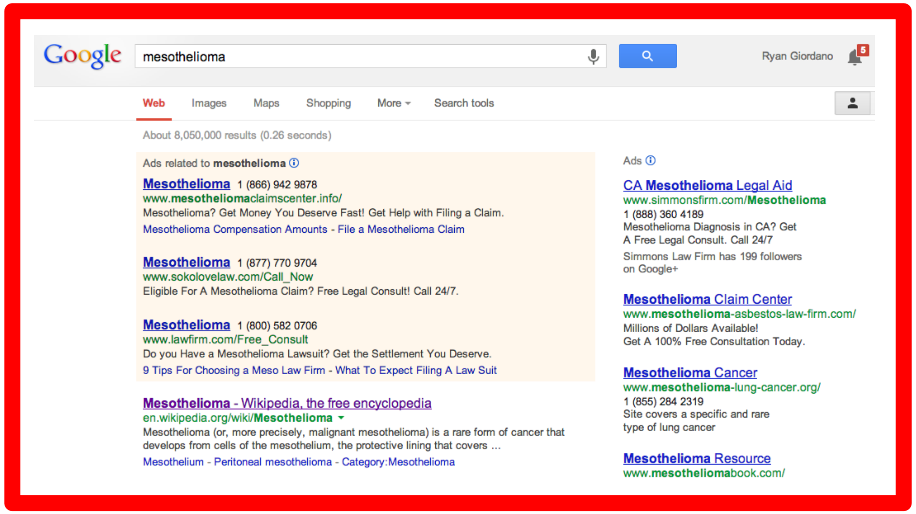
If the user clicks on one, the advertiser is then charged $X, and the user is sent to their site.

Motivating Example: Internet Advertising

Google Version 1
(Control)



Google Version 2
(Treatment)



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,

$X = CPC$ (**C**ost **P**er **C**lick)

$S = Spend$

$N = Clicks$

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

$X_{2i} = \frac{S_{2i}}{N_{2i}}$ (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

$$X_{1i} = S_{1i}/N_{1i} \qquad X_{2i} = S_{2i}/N_{2i}$$

$$E(X_{1i}) = \mu_{1i} \qquad E(X_{2i}) = \mu_{2i}$$

$\mu_{.i}$ 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:

$$\frac{1}{N}\sum_i (Y_i - Z_i) \to E(Y_i) - E(Z_i) = \mu_Y - \mu_Z$$ **Difference**

$$\frac{1}{N}\sum_i Y_i/Z_i \to E(Y_i/Z_i) \neq \mu_Y/\mu_Z$$ **Ratio**

Formal Problem Statement

N paired observations with independent mean-zero noise:

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

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

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

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

Paired observations

Independent

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:

- $\theta > 0$
- $\mu_i > 0$
- $var(X_{1i}) < \infty, var(X_{2i}) < \infty$
- 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

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

$$\frac{S_2}{N_2} = \ldots \qquad \ldots = \sum_i w_{2i} X_{2i}$$

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,    w_11 = 10%`	`x_11 = $9,      w_11 = 90%`
Adv. 2	`x_12 = $1,     w_12 = 90%`	`x_12 = $0.9,    w_12 = 10%`
Totals:	`x_1  = $1.9`	`x_2  = $8.19`

θ = \$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

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

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

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

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

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

Problem:

Linearity of expectations and

Sparse data (or zeroes)

Things You Might Try #2b: Average of Logs

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

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

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

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

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

Problem:

Exactly the same!

Sparse data

(or zeroes)

Things You Might Try #3: Random Effects Model

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

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

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

$$\Rightarrow$$

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

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

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

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

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

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

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

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

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

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 $n \to \infty$

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

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

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

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

=> 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 <= i <= 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_2i | m_i)
```

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

Non-uniform Changes

Suppose

```
t_i ~ 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 ~ 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

```
        n_1i n_2i
w_i = -------------,    x_1i = s_1i / n_1i
       n_1i + n_2i
w_i ~ n
m_i w_i ~ Spend
```

Result: a spend-weighted average proportional change.