Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference

Lifecycle Working Group

Thomas S. Coleman

Draft November 27, 2018

[*Snow & Causal Inference*]

[Coleman]

[Nov 2018]

Why John Snow and Cholera?
I present the story of cholera in 1850s London for three reasons:

I. **Rollicking Good Tale** – full of heroism, death, and statistics

II. **Causal Inference** – template for how to marshal evidence in support of a causal explanation

III. **Statistics & Instruction** – The data are simple but the analysis demonstrates multiple data analytic tools we use today

- combining maps and data (GIS or geographic information systems)
- regression and error analysis
- difference-in-differences regression
- natural experiments and randomization

Snow's cholera work is also a humbling reminder of the sometimes meandering path towards truth: even with overwhelming evidence and strong analysis Snow failed to convince the medical establishment, the public, or the authorities

Prototype for Building a Causal Argument
David Freedman extols Snow's research methodology:

> *a success story for scientific reasoning based on nonexperimental data*

but derogates regression and statistical testing:

> *regression models are not a particularly good way of doing empirical work in the social sciences today ("Statistical Models & Shoe Leather" 1991)*

This paper:

- Endorses and expands on Snow as an example of good scientific reasoning
- Lays out Snow's approach as a template for causal inference, a prototype with valuable guidelines for practitioners
- Argues that statistics (regression in particular) must be added to Snow's analysis – without a statistical foundation the causal argument is incomplete

"Snow's work is ... a success story for scientific reasoning based on nonexperimental data." ([Freedman(1991)] p 291)

"statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings," ([Freedman(1991)] p 291)

"regression models are not a particularly good way of doing empirical work in the social sciences today" ([Freedman(1991)] p. 304).

"Snow's work exemplifies one point on a continuum of research styles; the regression examples mark another" (p. 304).

What I Do and Don't Do
What I do

- Examine Snow's effort to "prove" causal theory
- Snow's position in 1855
  - Well-developed and demonstrated theory (from 1849)
  - Wide skepticism (or plain non-interest) in his theory and evidence
  - Skeptics pushed him to stay honest, developing and examining new evidence to reject every possible alternative

What I don't do

- Examine process of building and formulating theories (as in Heckman & Singer "Abduction")

[Heckman and Singer(2017)]

Outline

Contents

1 John Snow and the Story of Cholera

1.1 Cholera, John Snow, and Waterborne Theory

Contents

Cholera – Disease of Poor Sanitation
What is Cholera?

- Vibrio Cholerae – bacterium that infects the small intestine of humans
- Causes severe diarrhea (& vomiting) that drains fluids
- Death from dehydration & organ failure
- Oral Rehydration Therapy highly succesfull (roughly 1960s)
  - In case you ever need it, here's the recipe – 1 liter boiled water, 1/2 teaspoon salt, 6 teaspoons sugar, mashed banana (potassium)

Cholera thrives in crowded cities with poor sanitation

- Transmitted through recycling (drinking) sewage
- When cholera exits one victim, needs to find a way into gut of others
- Victorian London was an ideal playground for cholera to thrive

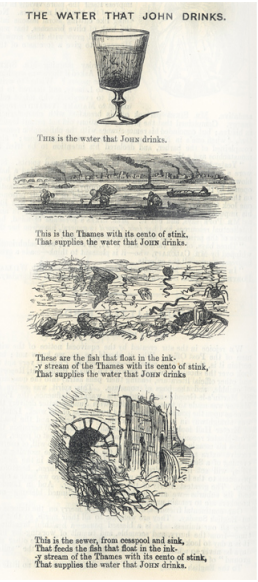
Cholera Loved Victorian London

Victorian London was an ideal playground for cholera

- Mid-1800s London was dirty, smelly place with no organized sewage treatment
- Efforts to improve sanitation made things worse
  - cesspools relatively safe – did not provide access to thousands of guts
- Public Health Act of 1848 required houses to connect to sewage lines
  - helped clean up streets, flushed filth to Thames
- By mid-1800s, cholera had easy access from the gut of one to thousands of victims

Contemporaries were aware of dirty water (*Punch* 1849)

- But water not recognized as vector for cholera



Contemporary cartoon from Punch, 1849

- http://www.choleraandthethames.co.uk/cholera-in-london/the-big-thames-clean-up/, 1849 Punch magazine, volume 17 Westminster City Archives

Solution – Construction of Bazalgette "Outfall Sewers"
Sewers that sloped towards outfalls (discharge points) lower on the Thames

- Construction started (under Bazalgette) 1859, response to 1858 "Great Stink"
- Embankments along Thames – what we see today
  - Embedded discharge pipes – still used today (?)
  - Decreased width, increased flow – scouring effect
- Moved sewage downstream, below London & water in-take

One final outbreak, 1866, limited to east London, last area unserved by sewers

John Snow's Research & Publications
Doctor – pioneer in anesthesia & medical hygiene

- Provided Queen Victoria with anesthesia during childbirth

Research and writing on Cholera

- 1849: "On the Mode of Communication of Cholera"
  - Laid out theory and evidence for waterborne transmission
- 1855: "On the Mode of Communication of Cholera"
  - Substantially expanded, additional evidence and argument
- 1856: "Cholera and the water supply in the south district of London in 1854"
  - Refined randomized analysis

John Snow's 1849 Theory & 1855 Evidence

1849: Snow developed theory of infection & transmission

- Based on medical knowledge and study of single events – Horsleydown & Albion Terrace

Fully-developed & modern theory of disease

- Infects & reproduces in the small intestine
- Exits from victim, into water supply
- Infects new victims through drinking dirty water

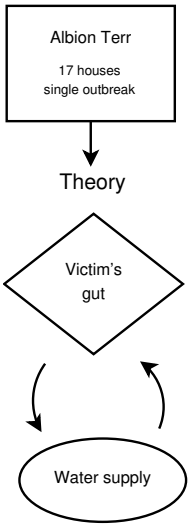
Implications for patterns of infection, across scales

- "from the membrane of the small intestine all the way up to the city itself" (Johnson)

Snow's work grounded by theory

> *Snow had a* good idea *– a causal theory about how the disease spread – that guided the gathering and assessment of evidence. (Tufte)*

1855: evidence & argument to convince skeptics



Johnson emphasizes role & importance of Snow's model:

> *The strength of his model derived from its ability to use observed phenomena on one scale to make predictions about behavior on other scales up and down the chain. ... If cholera were waterborne then the patterns of infection must correlate with the patterns of water distribution in London's neighborhoods. Snow's theory was like a ladder; each individual rung was impressive enough, but the power of it lay in ascending from bottom to top, from the membrane of the small intestine all the way up to the city itself. [Johnson(2007)] p. 148*

Tufte:

> *Most importantly, Snow had a* good idea *– a causal theory about how the disease spread – that guided the gathering and assessment of evidence. This theory developed from medical analysis and empirical observation; by mapping earlier epidemics, Snow detected a link between different water supplies and varying rates of cholera. [Tufte(1997a), Tufte(1997b)] p. 7*

The four components of Tufte's *good method*:

I. Placing the data in an appropriate context for assessing cause and effect

II. Making quantitative comparisons

III. Considering alternative explanations and contrary cases

IV. Assessment of possible errors in the numbers reported in graphics

Snow (1849):

> *The excretions of the sick at once suggest themselves as containing some material which, being accidentally swallowed, might attach itself to the mucous membrane of the small intestines, and there multiply itself by the appropriation of surrounding matter, in virtue of molecular changes going on within it, or capable of going on, as soon as it is placed in congenial circumstances. Such a mode of communication of disease is not without precedent. The ova of the intestinal worms are undoubtedly introduced in this way. The affections [sic] they induce are amongst the most chronic, whilst cholera is one of the most acute; but duration does not of itself destroy all analogy amongst organic processes. The writer, however, does not wish to be misunderstood as making this comparison so closely as to imply that cholera depends on veritable animals, or even animalcules, but rather to appeal to that general tendency to the continuity of molecular changes, by which combustion, putrefaction, fermentation, and the various processes in organized beings, are kept up. ([Snow(1849)] pp. 8-9)*

Alternative Theories
Miasma (Smells & Airborne)

- Cholera infectious & transmitted through the air
- Generally accepted in mid-1800s

Elevation, Crowding & Class, Others

- Elevation: lower elevation → more infection
- Crowding & Class: lower class & crowding → more infection

None of these absolutely crazy – correlated with cholera (and dirty water)

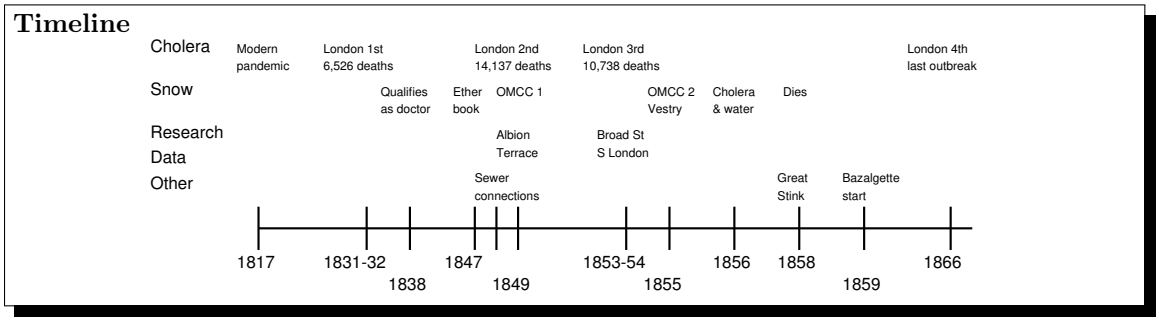
- Raw sewage associated with bad smells & dirty drinking water
- Lower class associated with crowding & poor sanitation

Other non-infectious theories (I won't seriously consider)

- Emanations from the ground
- Plague burying-pit near Broad Street pump

1.2 Data, Timeline, and Locations

Contents



History of Cholera

- Chronology of cholera http://www.choleraandthethames.co.uk
  - 1817 first outbreak from modern pandemic, India, China, Japan, ...
  - 1823 first dies down in Caucuses before reaching Europe
  - 1826-27, 2nd pandemic, moves to Russia, Poland, Europe
  - 1831-32, First England (Sunderland), 6,536 die
  - 1848-49 2nd English outbreak, 14,137 die
  - 1849 Albion Terrace
  - 1853-54 3rd English outbreak, 10,738 die
    - Aug-Sep 1854, Soho (Broad St) outbreak
  - 1866 4th & last outbreak, limited to east of London (the rest of London not strongly affected)
- John Snow chronology
  - 1838 John Snow qualifies as doctor
  - 1847 "On the Inhalation of Ether"
  - 1849 Albion Terrace outbreak – event from which Snow developed waterborne theory
  - 1849 omcc 1st ed published
  - Aug-Sep 1854, Soho (Broad St) outbreak
  - Aug-winter 1854 South London "Grand Experiment"
  - Jan 1855, omcc 2nd ed published
  - October 1856 "Cholera and the water supply in the south district of London in 1854"
  - June 1858 Snow dies
- Other events
  - 1848, (Public Health Act of 1848) sewers and cesspools must be connected to lines, dumping into Thames
  - July 1858, Great Stink that precipitated new sewer bypass
  - 1859, start of Bazalgette sewer system – Northern & Southern Outfall Sewers to discharge lower on the Thames (thus cleaning up the Thames and London's water supply

I discuss Three Strands or "Blocks" of Evidence

I. Albion Terrace

- 1849, Discovery of waterborne theory
- single event, 17 houses

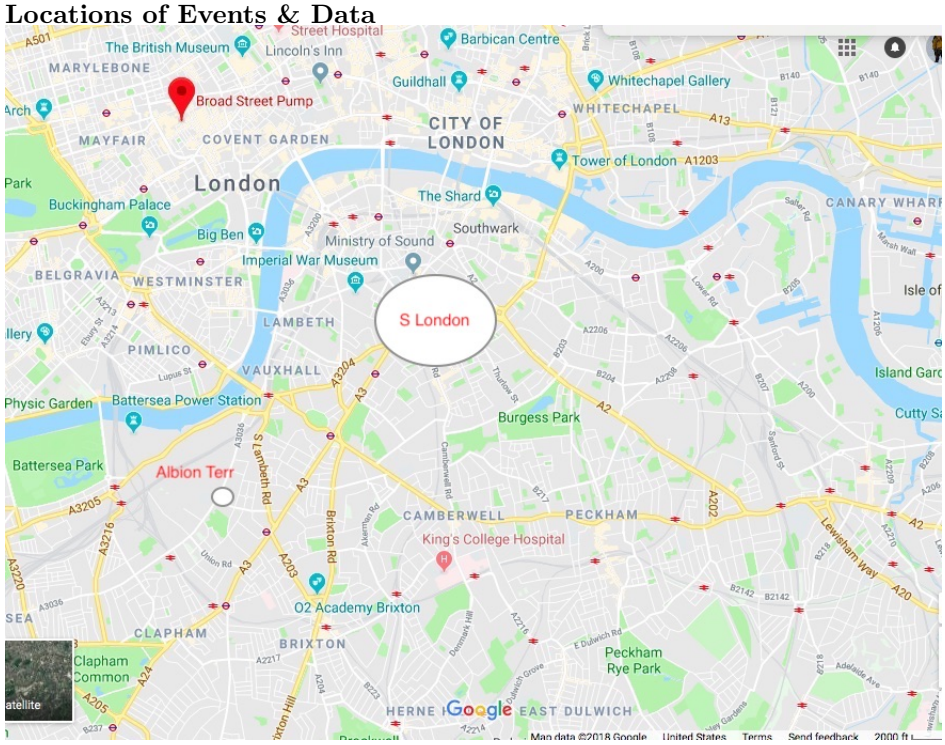
II. Broad Street Outbreak

- Aug-Sep 1854, 700 deaths over roughly 2 weeks, 10 square blocks

III. South London "Grand Experiment"

- Summer & Fall 1854, customers supplied by two water companies
- large scale, 400k mixed (quasi-random) subjects

Locations of Events & Data



2 Katz & Singer Causal Assessment Procedure

Contents

Modify Katz & Singer as "Causal Assessment Procedure"
Still tentative, based on Katz & Singer's analysis of possible Chemical & Biological Weapons attacks, 1970s-80s, "Can an Attribution Assessment Be Made for Yellow Rain?"

I. Divide evidence into blocks or types of evidence

II. Assign to each block a *veritas* rating – quality of data

III. Develop groups of hypotheses

IV. Assess each evidence block for strength of rejection for each hypothesis

- Consider *rejection* of hypotheses (refute, neutral, consistent) rather than strength of association (support of hypotheses)

V. Organize evidence blocks by hypothesis into matrix

VI. Choose hypothesis not contradicted

VII. Strongest hypothesis checked

[Katz and Singer(2007)]

What to call this?

- Causal Checklist
- Causal Engine
- Causal Matrix
- Causal Assessment Engine

Seven Step from Katz & Singer

I. Divide evidence into blocks or types of evidence

II. Assign to each block a *veritas* rating

- Degree of dubiousness (strength of evidence - "appraisal of intrinsic ambiguity or likelihood")
  - 1-3, high, moderate, minimal distortion
- Degree of fallacy ("appraisal of deception")
  - "the extent to which a piece of evidence was deceptive, misleading, or the result of unreliable reasoning"
  - 1, event probability low and evidence doubtful
  - 2, supporting information accurate but event low probability
  - 3, accepted evidence but doubted piece
  - 4, accepted all evidence as probably accurate
- But for the cholera example, I think this less relevant and need to emphasize (4) below.

III. Develop groups of hypotheses

IV. Assess each evidence block for strength of association to each hypothesis

- Strong, medium, weak
- For Snow & cholera, think about inverting and making it about contradicting or refuting hypothesis (maybe strong, neutral, fail-to-refute)
- Maybe need two dimensions?
  - Strength of evidence: strong, medium, weak (this is p-value for statistical tests)
  - Whether contradicts hypothesis or not (maybe strong, neutral, fail-to-refute)

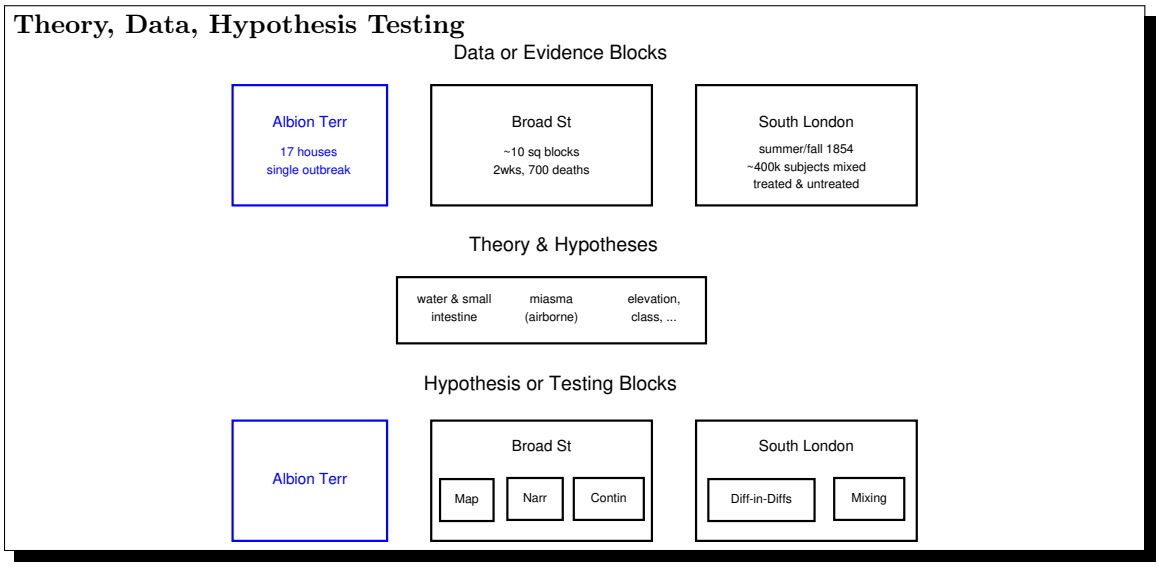
V. Organize evidence blocks by hypothesis into matrix

VI. Choose strongest hypothesis

- based on quality of evidence, quantity of evidence, strength of explanation based on evidence
- each block assigned numerical score based on coding scheme (strength of association & veritas rating)

VII. Strongest hypothesis checked

- agreed with overall state of historical and scientific knowledge
- satisfies guidelines for causation
- consistent with any definitive proof or admission (not applicable for economic problems, I think)



3 John Snow's Evidence & Causal Inference

Overview of Snow's Evidence & Argument

Model-based

> *The strength of his model derived from its ability to use observed phenomena on one scale to make predictions about behavior on other scales up and down the chain. ... If cholera were waterborne then the patterns of infection must correlate with the patterns of water*

distribution in London's neighborhoods. Snow's theory was like a ladder; each individual rung was impressive enough, but the power of it lay in ascending from bottom to top, from the membrane of the small intestine all the way up to the city itself. ([Johnson(2007)] p. 148)

Multiple evidence

I. Case study with attention to individual circumstances

II. Large-scale statistical study

It is remarkable that, of the two suppositions, first, that the air alone, and secondly, that the water more especially, was concerned in exciting the disease, whilst the former appears less and less equal to explain individual cases in proportion as these are examined more and more in detail, it is precisely in the variety and exactitude of its particular application to individual facts that the latter finds its most positive support. ([Westminster and London School of Hygiene and Tropical Medicine(1855)] p. 81)

3.1 Albion Terrace – "Discovery" of Theory

Contents

Albion Terrace Details, 1849
Terrace of 17 houses in South London (Wandsworth Road)

- Snow focused on this outbreak because no cases in surrounding houses

  there were no other cases at the time in the immediate neighbourhood; the houses opposite to, behind, and in the same line, at each end of those in which the disease prevailed, having been free from it. (Snow 1849 p 15)

Provided sharp test of how & why cholera spread

- Assistant-Surveyor for Commission of Sewers dug up and studied piping
- Storm July 26, drain burst and contaminated water for all 17 houses

  the only special and peculiar cause ... was the state of the water, which was followed by the cholera in almost every house to which it extended, whilst all the surrounding houses were quite free from the disease. (Snow 1855 p 30)

Provided Snow with final evidence that crystalized his theory

Within the last few days, however, some occurrences have come within [the author's] knowledge which seem to offer more direct proof, and have induced him to take the present course [publishing]. (Snow 1849 p 12)

Not enough to convince skeptics

Details on Albion Terrace: http://www.ph.ucla.edu/epi/snow/1859map/albionterraceoutbreak_1849_a3.html

Albion Terrace was a row of 17 houses that suffered "an extraordinary mortality from Cholera in 1849, which was the more striking as there were no other cases at the time in the immediate neighbourhood; the houses opposite to, behind, and in the same line, at each end of those in which the disease prevailed, having been free from it." Snow described in detail the outbreak (reporting findings from Mr. Grant, the Assistant-Surveyor for the Commission of Sewers), and the piping for water supply and sewage disposal from the Albion Terrace houses and the circumstances that led to contamination of their water supply but not others – how "the water got contaminated by the contents of the house-drains and cesspools; the cholera extended to nearly all the houses in which the water was thus tainted, and to no others." ([Snow(1849)] p 15 ff, also [Snow(1855)] p 30

It remains evident then, that the only special and peculiar cause connected with the great calamity which befel the inhabitants of these houses, was the state of the water, which was followed by the cholera in almost every house to which it extended, whilst all the surrounding houses were quite free from the disease. ([Snow(1855)] p 30)

3.2 Broad Street Pump

Contents

Broad Street – 2 Weeks of Horrendous Death

> *The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets ... there were upwards of five hundred fatal attacks of cholera in ten days. (Snow 1855 p. 38)*

Outbreak erupted Aug 29, lasted 2-3 weeks

- Ultimately, more than 600 dead
- Limited to small neighborhood in Soho (south of Carnaby St, east of Regent St)
- Sudden, violent, dramatic outbreak

Snow lived nearby, quickly went to neighborhood to investigate

- Walked the streets, talked with and collected data from residents

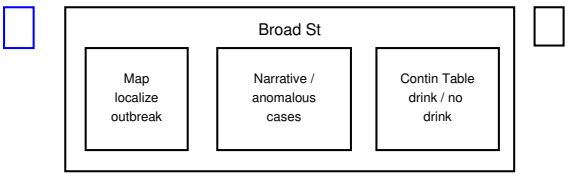
Visited last June (by happenstance)

- John Snow pub

The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there were upwards of five hundred fatal attacks of cholera in ten days. ([Snow(1855)] p. 38)

Broad Street Pump Analysis – 3 Parts
Mapping

- Discovery & explication
  - localizing outbreak
  - making visible what is hidden



- Icon: encapsulating and promoting waterborne theory
- Statistical Tests of Clustering

Narratives, Case Studies, Anomalies

- Narrative & Tracking Individual Cases
- Exceptions & Anomalies: "Snow knew that the case would be made in the exceptions from the norm." (Johnson p 140)

Contingency Table / Statistics (Whitehead, extending Snow)

- Testing Drinkers vs Non-Drinkers and Survivorship Bias

3.2.1 Mapping

Contents

Mapping – Localizing & Publicizing Source
Time-series not informative – With map, clustering around pump jumps out

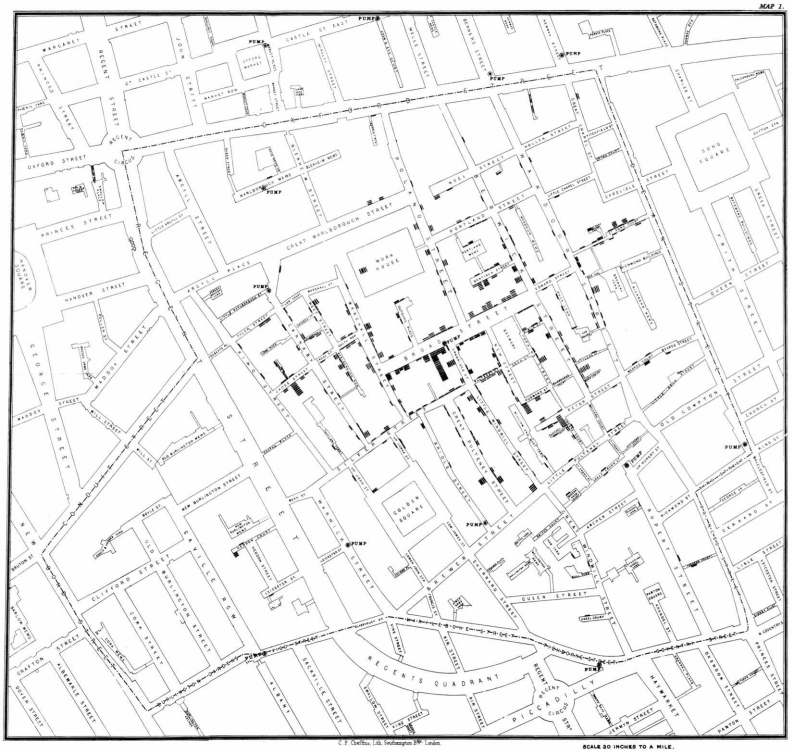
- Map has become a (well-deserved) icon of good science
- Snow recognized clustering even without map – map as tool for demonstrating – publicizing – theory

"Deaths from Cholera" produced using software from [Li()]. Map from [Snow(1855)], copied from http://www.ph.ucla.edu/epi/snow/snowmap1.pdf

> *[Snow] could see at a glance that he'd be able to demonstrate that the outbreak was clustered around the pump, yet he knew from experience that that kind of evidence, on its own, would not satisfy a miasmatist. The cluster could just as easily reflect some pocket of poisoned air that had settled over that part of Soho, something emanating from the gulley holes or cesspools – or perhaps even from the pump itself. Snow knew that the case would be made in the exceptions from the norm. Pockets of life where you could expect death, pockets of death where you would expect life. [Johnson(2007)] p. 140*
>
> *The map may not have had the impact on its immediate audience that Snow would have liked, but something about it reverberated in the culture. Like the cholera itself, it had a certain quality that made people inclined to reproduce it, and through that reproduction,*

the map spread the waterborne theory more broadly. In the long run, the map was a triumph of marketing as much as empirical science. It helped a good idea find a wide audience. ([Johnson(2007)] p. 199)

Snow was Masterful, Stripping Out Extraneous Detail

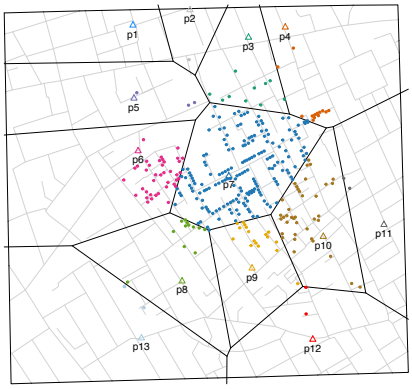


Map, Statistical Tests of Clustering

Mapping & Statistical Tests of Clustering

A fun area that many others have discussed – I only mention in passing

- Measuring clustering relative to random or underlying infection



This Voronoi partition of the map from software by Li (Github) MapsMaps

Automates and Formalizes the measurement of clustering

Voronoi partitioning, walking neighborhood, chi-squared calculation, highlighting clustering around Bridle Street and Rupert Street pumps

3.2.2 Case Studies & Narrative: Tracking Individual Cases & Anomalies

Contents

More Important than Map: Narratives & Anomalous Cases
Non-Uniform Spatial Distribution – Testing Competing Theories: "confronting the waterborne and alternative theories with evidence"

I. Those who should have died but escaped
   - Close to pump but did not die
   - Work House & Brewery (few-to-no deaths)

II. Those who should have escaped but died
   - Far from the pump but died
   - Marlborough St pump and 10 Cross St ("great drinkers of pump water")
   - Girls from the south – Ham Yard & Angel Ct – off Great Windmill St, near Bridle Street, Rupert Street, or Tichborne St pumps
   - Susannah Eley, famous "Widow in Hampstead"

III. Details on the mechanism for contamination of the pump-well
   - Index case and decaying brick-work

Story about removing pump-handle on September 7 – did not stop outbreak which was already falling quickly (see graph)

Discuss more "anecdotal" evidence

Structuring Snow's evidence as "confronting the waterborne and alternative theories with evidence"

I. Those who should have died but escaped – Those close to pump who did not die
   - Work House & Brewery

   The St. James workhouse on Poland Street had 535 inmates with only five dying. As Snow points out ([Snow(1855)] p 42) if the death rate had been as high for the workhouse as the surrounding houses more than 100 would have died. The explanation was simple: the workhouse had its own well and was also supplied by piped water (from the Grand Junction Water Works); residents did not visit the Broad Street pump.

   For the brewery (the Lion Brewery) the same problem and explanation hold: seventy workmen but no cases of cholera, but the workmen were "allowed a certain quantity of malt liquor," had access to an in-house well, and never drank from the pump.

II. Those who should have escaped but died – Those far from the pump who nonetheless died
   - Marlborough St and 10 Cross St
   - Girls from the south - Ham Yard & Angel Ct - off Great Windmill St, near Bridle Street, Rupert Street, or Tichborne St pumps
   - Susannah Eley

A fair number of deaths cluster near the Little Marlborough Street pump, nearer Marlborough than Broad Street. This should not be the case if water were the cause (and with the auxiliary hypothesis that residents drink from the closest pump). But Snow states:

> *It requires to be stated that the water of the pump in Marlborough Street, at the end of Carnaby Street, was so impure that many people avoided using it. And I found that the persons who died near this pump in the beginning of September, had water from the Broad Street pump. ([Snow(1855)] p 46)*

There is a cluster of eight deaths at 10 Cross Street, closer to Marlborough than Broad Street. Their story is told in the Vestry report: a tailor aged 50 and his 12 year-old son died September 1st, and within three days four more of his children, all "great drinkers of pump water" who often drank from the Broad Street pump.[1]

Two little girls (one from Ham Yard the other from Angel Court, both off Great Windmill Street far to the south of Broad Street) went to school in Durfours Place (off Broad Street) and drank from the Broad Street pump on the way to or from school. ([Westminster and London School of Hygiene and Tropica pp 112-113)

One of the most famous cases concerned Susannah Eley, a widow in Hampstead and her niece in Islington who died in early September when there were no other cholera deaths in those areas.[2] Snow discovered from the widow's sons, who owned a factory at 37 Broad Street near the pump, that she had lived in Soho, thought the Broad Street pump water delicious, and regularly had water from the pump brought to her in Hampstead. Both she and her niece drank Broad Street water the day before falling ill ([Snow(1855)] pp 44-45, also discussed in [Tufte(1997b), Johnson(2007)], [Hempel(2007)] p 217 ff as well as others).

III. Details on the mechanism for contamination of the pump-well – The index case and decaying brick-work

Whitehead identified a baby girl Frances Lewis at 40 Broad Street, the building next to the pump, who had fallen sick a day prior to the outbreak (and died September 2nd). Sarah Lewis, the mother, had rinsed diapers and poured the water into a cesspool at the front of the house. The existence of the cesspool was unexpected (drains were supposed to be connected to sewer lines) and further inspection showed that the cesspool was only inches from the well, there was decaying brickwork, and the ground was saturated with water from the cesspool.

Story about removing pump-handle on September 7 – did not stop outbreak which was already falling quickly (see graph)

[1] "This family were great drinkers of pump water, and used to send for it every day, but more especially to drink during the night, as they were thirsty in the warm weather, owing to the great number sleeping in one room. The children fetched the water from various pumps, but frequently from Broad Street." [Westminster and London School of Hygiene and Tropical Medicine(1855)] p 112

[2] Snow credits a Dr. David Fraser for alerting him to these anomalous deaths.

Anomalies to Test & Separate Theories

Refutation: Imre Lakatos "Core" & "Protective Belt" of Auxiliary Hypotheses

(1) **Close to pump but did not die**

	Water 1	Water 2	Miasma 1	Miasma 2
Core	Drinking	Drinking	Breathing	Breathing
Auxiliary	P[drink~ distance]	P[drink~ in-house wells]	P[breath~ distance]	P[breath~ ??]
Implication	deaths~ distance	deaths~ distance & wells	deaths~ distance	??
Core Refuted?	YES	NO	YES	??

Difficult to come up with Miasma auxiliary hypothesis to match spatial distribution

- Deaths follow drinking: Breathing pattern would need to correlate with drinking
- Could argue Snow did not search for auxiliary breathing hypothesis – but a stretch

(2) **Far from pump but did die**

	Water 1	Water 2	Miasma 1	Miasma 2
Core	Drinking	Drinking	Breathing	Breathing
Auxiliary	P[drink~ distance]	People travel to Broad St	P[breath~ distance]	Water infected by air
Implication	deaths~ distance	deaths~ taste for Broad St	deaths~ distance	deaths~ taste for Broad St
Core Refuted?	YES	NO	YES	NO

Water auxiliary: some people travel distances to Broad St pump

- Reasonable, fits naturally with known human behavior

Miasma auxiliary: water "participates in the atmospheric infection"

- To modern eyes, foolish and cooked up to support miasma
- Miasma protected by auxiliary hypothesis allowing miasma to match drinking patterns

We can only judge evidence in concert with judgement about theory

Lakatos (in [Lakatos(1980)], particularly section 1.3 p 47 ff) lays out the idea of a scientific research programme consisting of a "hard core" together with "protective belt" of auxiliary hypotheses built around the central core. Lakatos argues that anomalies or counterexamples can be accommodated by adjusting the protective belt rather than rejecting the hard core. In fact the death of Susannah Eley in the Broad Street outbreak (a widow from Hampstead discussed below) provides a near-perfect example. Her case is a strong counterexample to airborne transmission (miasma). Nonetheless the official Cholera Commission's report dismisses the anomaly by invoking a strained hypothesis about airborne influences poisoning the water – an auxiliary hypothesis that we now recognize as outlandish.

Cholera Commission's Auxiliary Hypothesis
This is really too good to pass up:

> *The water was undeniably impure with organic contamination; and ... if, at the times of epidemic invasion there was operating in the air some influence which converts putrefiable impurities into a specific poison, the water of the locality ... would probably be liable to similar poisonous conversion. Thus,* ***if the Broad Street pump did actually become a source of disease*** *to persons dwelling at a distance ...* ***this ... may have arisen****, not in its containing choleraic excrements, but* ***simply in the fact of its impure waters having participated in the atmospheric infection of the district****.*

Wonderful example of Miasma auxiliary hypothesis to protect miasma core

- Demonstrates that virtually any "core" can be protected by "auxiliary"
- An auxiliary we now recognize as foolish, cooked up to protect Miasma
- Miasma protected by auxiliary hypothesis allowing miasma to match drinking patterns

[Johnson(2007)] p 186 and [Hempel(2007)] p 242 (at greater length) quote the Cholera Commission's report as acknowledging that water was the *vehicle* of contamination, but not the ultimate cause:

> *The water was undeniably impure with organic contamination; and ... if, at the times of epidemic invasion there was operating in the air some influence which converts putrefiable impurities into a specific poison, the water of the locality ... would probably be liable to similar poisonous conversion. Thus, if the Broad Street pump did actually become a source of disease to persons dwelling at a distance ... this ... may have arisen, not in its containing choleraic excrements, but simply in the fact of its impure waters having participated in the atmospheric infection of the district.*

3.2.3 Drinkers vs Non-Drinkers and Survivorship Bias – 2x2 Contingency Analysis

Contents

Drinkers vs Non-Drinkers and Survivorship Bias
Substantive problem, recognized by Rev. Whitehead (Snow confrere)

- Snow focused on deaths, not survivors
- What if rate of drinking were similar for those who *did not* fall ill?
- Classic case of potential *survivorship bias*: need to ensure not only those who did die did drink, but those who did not die did not drink

Rev. Whitehead collected data on 497 residents of **Broad Street** & their illness and drinking history

- Found few non-drinkers fall ill
- Strong association between drinking and illness
- Water theory survived this test – Miasma did not

Drinkers vs Non-Drinkers and Survivorship Bias
Extension to Snow: Modern Statistics: 2x2 Contingency Table
Fewer non-drinkers and more drinkers fall ill than expected if independent

- Phi coefficient (Cramér's V) +0.59 – strong association drinking & illness
- Formalizing with statistics strengthens Snow's argument
- Contrary to Freedman's claim against statistics

Water Supported, Miasma Refuted by Contingency Table
Put water against data that could reject, but find strong association

- Strong water association hard for miasma theory
  - Need miasma & smells to be strongly associated with water
  - Not logically impossible, but highly improbable

Evidence so far does not prove water-borne theory, but very supportive

- Omitted (confounding) variables logically possible
  - Something *associated* with water that causes cholera
- But hard to imagine

And alternatives theories (miasma, class, elevation, ...) not looking good

3.3 South London "Grand Experiment"

Contents

"Grand Experiment" – Large-scale Mixing
Two water companies served south London, Southwark & Vauxhall Co and Lambeth Co.

- 486,936 customers in joint-service area – 300,000 **intimately mixed**
- 1852 Lambeth changed water source from dirty to clean – **treatment**

Snow recognized the value of mixing for "Grand Experiment"
Use Snow's data to demonstrate (and improve) his argument & methodology
Difference-in-differences:

- Compare 1849 vs 1854 cholera outbreaks, treated vs untreated
- Akin to a "retrospective" study

Quasi-Random Mixing (Quasi Randomized Control Trial):

- Compare Southwark vs Lambeth customers – random treatment
- Akin to a "prospective" study or clinical trial

Explanation of the south London "Grand Experiment"

I. Two water companies, Southwark & Vauxhall Co and Lambeth Co. In 1852 Lambeth changed its water source from lower Thames (polluted with sewage and subject to infection with cholera) to upper Thames (not polluted)

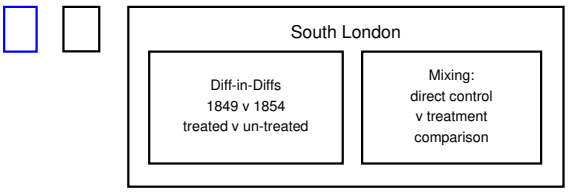
II. They served south London, part jointly, part separately (see [Snow(1855)]p. 68)

III. Seems that that Snow discovered this almost accidentally, in a note to the "Weekly Register of Births & Deaths": "a footnote in the November 26 [1849?] edition of the *Weekly Returns*. Below the cholera deaths for South London, Farr had appended this seemingly innocuous line" 'In three cases ... the same districts are supplied by two companies.'" [Johnson(2007)] p. 160

IV. Can compare 1849 vs 1854 cholera outbreaks (treated vs untreated – a form of diff-in-diff), and in 1854 compare Southwark vs Lambeth in joint area (a form of randomized treatment)

South London Analysis – 2 Parts
Aggregate, Diff-in-Diffs

- Aggregate regions
- 1849 vs 1854
- Treated (clean) vs untreated (dirty)



Mixed or quasi-random comparison

- Snow visited all houses (deaths) for seven weeks ending Aug 26
- Determined supplier – by bill or chloride test

Registration Districts & Sub-Districts – Need to keep straight

- Deaths collected weekly by Registrar-General, by Registration District & Sub-District
- In this region of South London, 32 sub-districts
  - "First 12" – Southwark & Vauxhall Water Co only – dirty water 1849 & 1854
  - "Next 16" – Joint Southwark & Vauxhall Co and Lambeth Water Co – 1849 dirty water, 1854 part dirty (Southwark) & part clean (Lambeth)
  - "Final 4" – Lambeth Water Co only – not relevant, not supplied in 1849

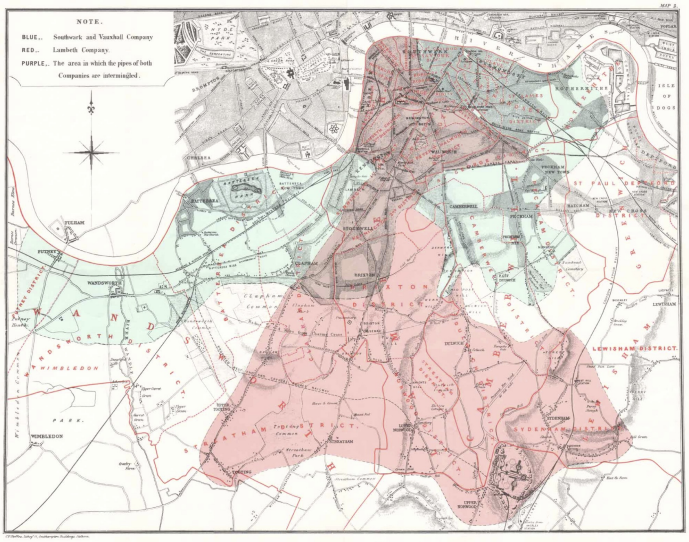
What We Can Learn From South London – Control for Omitted Variables
Mixing of water company customers allows testing for confounding or omitted variables

- Customers face same weather, elevation, topography, geography (proximity to Thames)
- Mixing by age, sex, class, income

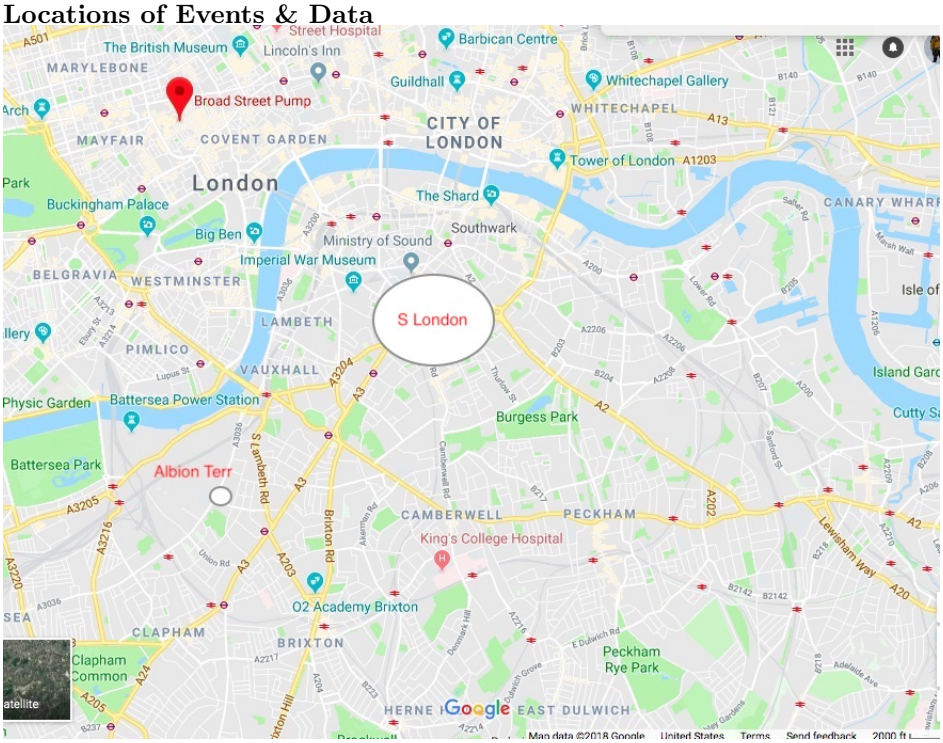
If water supply (clean vs dirty) shows big effect, hard to argue it is those other effects

- Does not prove causality, but rules out many other causes

Locations of Events & Data



Locations of Events & Data



3.3.1 Aggregate, Retrospective, Difference-in-Differences (Poisson Regression)

Contents

Difference-in-Differences (1849 vs 1854)

> *[Table XII] exhibits an increase of mortality in 1854 as compared with 1849, in the sub-districts supplied by the Southwark and Vauxhall Company only,* ***whilst there is a considerable diminution of mortality in the sub-districts partly supplied by the Lambeth Company****. (Snow p 89)*

We can sharpen, considerably, tabulating as Diff-in-Diffs (& teaching example)

- Not sure why Snow didn't express as rates here

Calculating Treatment Effect: -0.51 in logs, 1.67x

- Formalizes Snow's "considerable diminution of mortality in the sub-districts partly supplied by the Lambeth Company"
  - DiD allows control for (removing) confounding effects of region (& year)
  - In this case, both region and year effects small
- But how confident are we in estimate?
  - Contrary to Freedman, need statistics & regression

Mortality Rates from Cholera per 10,000 Persons in 1849 & 1854, Summary from Snow Table XII & Table VIII

Poisson & Neg Binom Regression for Diff-in-Diffs

We observe large variation across and within sub-districts (rates per 10,000)

- Some increased, some decreased (even for Southwark-only supply)

We want to exploit this variation to assess precision of our -0.511 estimate

- Stigler's "intercomparison" (from Galton)

Need Statistical Model that maps our problem to usable mathematical framework

- Our problem: individuals at risk of infection & death
- Statistical Model 1: probability of infection (death) generated by Poisson process (approx to Binomial)
  - Generates counts (deaths) Poisson-distributed
  - Variance = mean ⇒ Std Dev of rate ↓ as Population ↑
  - For large population, rate has little variability – not what we see diderrorbarsdiderrorbars
- Statistical Model 2: prob Poisson, but individual-level heterogeneity
  - Poisson mixture, Gamma mixing ⇒ Negative Binomial Counts (deaths)

Multiple steps to assess if statistical model adequate

I. Poisson, no FE: everyone everywhere same prob of death
   - Observed variation too much for statistical model diderrorbarsdiderrorbars

II. Poisson, sub-district FE: sub-districts vary, rate within same for all
   - Observed variation too much for statistical model

III. Neg Binom, no FE: individual heterogeneity, Poisson mixed with Gamma
   - Observed variation within bounds for statistical model

If I remember correctly, Stigler ascribes the origin of the idea of "intercomparison" to Galton

Poisson & Neg Binom Regressions for Diff-in-Diffs
Add "less Lambeth" & "more Lambeth"

- Snow said 4 sub-districts "more Lambeth"

With proper statistical framework we see:

- Large & precise "more Lambeth"
- region & time FEs not significant – little evidence of differences across regions or time

PoissonRegressionPoissonRegression	Neg Binom, two treatment
"less Lambeth" (ln)	**-0.344***
standard error	**0.250**
p-value	**17.0%**
treatment (ratio)	**1.41**
"more Lambeth" (ln)	**-1.137**
standard error	**0.356**
p-value	**0.14%**
treatment (ratio)	**3.12**
theta (Gamma mixing)	**5.48**
Resid Deviance (df, p-value)	**60.0 (51, p=18%)**
less Lambeth FE	**-0.065***
more Lambeth FE	0.058*
Time fixed effect	0.063*

* = ***not*** significant at the 10% level

Formalizes & substantially strengthens Snow's "considerable diminution of mortality in the sub-districts partly supplied by the Lambeth Company"

- Strongly contradicts Freedman's view that "regression models are not a particularly good way of doing empirical work in the social sciences today" ("Statistical Models" 1991)

3.3.2 Quasi-Random Mixing (Quasi Randomized Control Trial)

Contents

Mixing – Quasi-Randomized Control Trial
Registrar-General recorded deaths weekly by sub-district

- 16 of these (pop 300,149) mixed between Southwark Co & Lambeth Co

> *In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference in the condition or occupation of the persons receiving the water of the different companies. (Snow 1855 p 75)*

During August Snow visited every house with a death to identify supplier

- These data provide close to random mixing
- Allows control for any and all non-water characteristics

Snow needed population-at-risk

- Best he could do in 1855 was houses, aggregate, for Southwark Co vs Lambeth Co

Snow's Comparison
Using Houses for all 32 sub-districts together

- Includes "first 12" Southwark-only sub-districts (& "last 4")
- Not a clean comparison of "next 16" mixed sub-districts
- From diff-in-diffs we know not large differences between "first 12" & "next 16"

Circumstance provided Snow with an "experiment" close to an RCT

Extending Snow's Comparison
We can approximately allocate houses according to population MixHousesMixHouses

- As expected, little difference from Snow's table

Error Analysis for Snow's Comparison
We would like to assess confidence in those rates

- Use Neg Binom process estimated in Diff-in-Diffs to calculate (very approximate) confidence bands

Later publications (Simon, 1855) provided population by supplier by sub-district

- Allows Poisson & Neg Binomial regression as above PoissonRegression1856PoissonRegression1856
- Shows large & precisely-measured Lambeth effect (clean water)

Being Careful With Data
Snow was careful with data, and we need to follow his example

- Snow (1856) highlighted errors in assignment of houses & population to supplier (in 1855 Simon, Registrar-General data)
- We need to use those data carefully
- More recent authors (Koch & Denike 2006) have not been as careful

Discussion of [Koch and Denike(2006)] (preliminary)

[Koch and Denike(2006)] aim to address a problem with data present in [Snow(1856)] (for example Snow's Table IV, Koch & Denike's Figure 3). Their objective is highly worthwhile – to improve the South London data to allow for more fine-grained measurement of the geographic patterns of mortality rates – but unfortunately they do not accomplish their objective.

The problem Koch & Denike set out to address is that there were 623 deaths that were not assigned to water company supplier (Southwark & Vauxhall Company versus Lambeth Company). They state "there were 623 houses in which cholera occurred that could not be assigned reflexively to any single district nor to either of the two water supplier areas." (p. 275) The first part of this statement is simply incorrect and the second part slightly confusing. There were 623 deaths in houses where the house could not be assigned to water *supplier*, but all those death (and the houses in which they occurred) were clearly assigned to Registration District and sub-district in the original data from the Registrar General. The assignments to Registration District are shown in Snow's Table IV (Koch and Denike's Figure 3). There has never (to my knowledge) been any question about the reliability of the Registrar General's assignment of deaths to District or Sub-District (as opposed to assignment, within sub-district, to source or water *supplier*).

Koch & Denike proceed to re-allocate those 623 deaths across water supplier *and* Registration Districts (see for example their Figure 6). In doing so they move deaths across Districts, deaths that were reliably assigned to District by the Registrar General. The re-assignment across Districts is neither necessary nor justified.

[INSERT TABLES REPRODUCING SNOW'S TABLE IV AND KOCH & DENIKE'S FIGURE 6, SHOWING CHANGES IN OVERALL MORTALITY BY DISTRICT]

The problem of the 623 deaths not assigned to supplier is relatively minor, and I think can be handled simply by allocating at the sub-district level in proportion to the reported deaths for the two suppliers. This is reasonable because a) there appears to be little difference between customers of the Southwark & Vauxhall Company versus the Lambeth Company (this is the whole basis for Snow using the South London data); b) thus unassigned houses should be random relative to supplier; c) so the probability of an unassigned house belonging to one or the other supplier should be proportional to the probability of death by supplier at the sub-district level.

There is, however, an additional and more difficult problem. There are some houses that are improperly assigned to subdistricts and others that are not assigned to any subdistrict. This will bias the mortality rates, and for various reasons the biases will be more serious for the fine-grained sub-district data than for the more aggregate Registrar Districts or aggregate regions.

Regarding population improperly assigned to sub-districts, [Snow(1856)] p. 245 states: "The line has not been very accurately drawn where a street, as often happens, is partly in one district and partly in another; and thus, in the recent Report, the subdistricts of St. Saviour's, Southwark, Leather market, Bermondsey, Battersea, and Peckham, have been represented to contain a few houses supplied by the Lambeth Company although they do not contain any." From what I know of Snow's care in handling data, I believe Snow rather than the published statistics. In other words, population is being assigned to the Lambeth Company when it should not. Presumably the reverse (population incorrectly assigned to Southwark) also occurs but it is harder to identify specific instances. It is unclear whether and how these biases will offset.

Regarding houses not assigned, houses (and thus population) on streets on which no death occurred were apparently not included in the sub-district count of houses, but instead "placed all together at the end of each group of subdistricts which constitutes a district." (p. 246)[3] In other words those houses can be correctly assigned at the Registration District level (Snow's Tables III, IV, V) but not at the sub-district level (Table VI or Koch and Denike Figure 7). Snow's Table VI (and Table II) shows the population for these "unassigned" houses to be 52,267 or 12% of the total.

For all these reasons I think we need to be cautious in using the sub-district data. None of this, however, has very much impact on Snow's main objective, to demonstrate at the aggregate level a large effect of clean versus dirty water on mortality.

[3]Furthermore that are a small number, 411, of "Not identified" houses which account for 0.7% of overall population that cannot be assigned to Registration District.

What We Have Learned from Diff-in-Diffs & Mixing

I. Water shows up as big and reliable (statistically significant) effect

II. Reasonable to rule out other causes

- Weather, elevation, topography, geography (proximity to Thames) since customers face similar or same circumstance
- Age, sex, class, income since customers mixed

Does not prove causality, but

- rules out most all confounding variables I can think of
- leaves few other candidates for causal agent

A reasonable person pushed to conclude that water is the cause

Contrary view to Freedman

Strengthens Snow's conclusions by quantifying how likely or unlikely observations are. Tufte's point about *quantitative comparisons* and always ask *compared with what*?

What is the proper role of statistics in empirical work? Two-fold:

I. Allow us to "control for" observed and unobserved co-variates

(A) Standard regression: control for observed co-variates

(B) Various tools & tricks (diff-in-diffs, RCT, sample selection, IV): Control for unobserved co-variates

- But the control for unobserved co-variates is only as good as our argument (or belief) that instruments are uncorrelated with unobserved co-variates, or that fixed effects (in diff-in-diffs) appropriately capture differences between groups ("appropriately" as in the fixed effect itself does not vary with the treatment or with time)
- We can sometimes provide evidence that supports or increases the plausibility of instruments, etc. But we can never definitively rule out unobserved confounding effects

II. Discipline our temptation to believe in numbers – forces us perform quantitative comparisons and to ask "compared with what?"

But for both of these we have to use external judgement, external to the observed data, to determine whether the data and observations could be generated by an alternative hypothesis. There can almost always be *some* alternative model, some alternative world, where the observed data are generated not by our proposed causal agent but by some alternative. We have to use our knowledge of the world to judge the plausibility that such an alternative exists. I think it comes closer to "beyond a reasonable doubt" than to a formal probability statement. We cannot observe the alternatives (if we could we would include them as covariates in a regression). We have to use argument, providing multiple sources of evidence, that rules out alternatives beyond reasonable doubt.

Pedagogy & Statistics

Example of Difference-in-Differences

I. Good pedagogical example

- Linear regression versus count (Poisson or Negative Binomial)

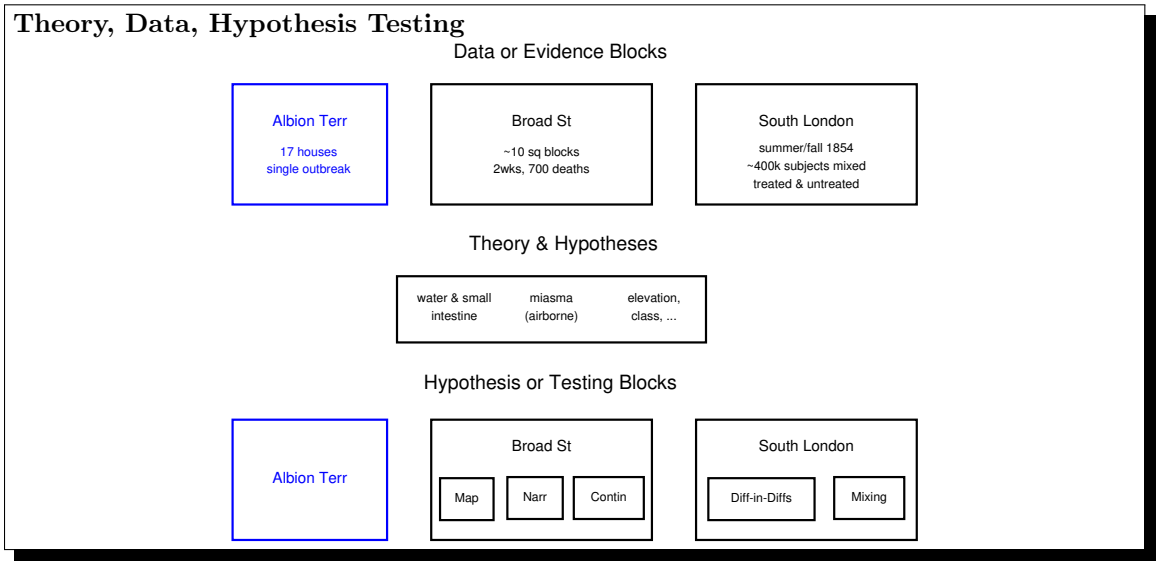
II. Regression framework allows using internal variation to ask "how likely?" (Stigler's *intercomparison*)

Example of Randomized Control Trial

Adding measurement of sampling variation (from diff-in-diffs)

4 Conclusion

Contents



Theory, Data, Hypothesis Testing

Seven Step from Katz & Singer

I. Divide evidence into blocks or types of evidence

II. Assign to each block a *veritas* rating

- Degree of dubiousness (strength of evidence - "appraisal of intrinsic ambiguity or likelihood")
  - 1-3, high, moderate, minimal distortion
- Degree of fallacy ("appraisal of deception")
  - "the extent to which a piece of evidence was deceptive, misleading, or the result of unreliable reasoning"
  - 1, event probability low and evidence doubtful
  - 2, supporting information accurate but event low probability
  - 3, accepted evidence but doubted piece
  - 4, accepted all evidence as probably accurate

- But for the cholera example, I think this less relevant and need to emphasize (4) below.

III. Develop groups of hypotheses

IV. Assess each evidence block for strength of association to each hypothesis

- Strong, medium, weak
- For Snow & cholera, think about inverting and making it about contradicting or refuting hypothesis (maybe strong, neutral, fail-to-refute)
- Maybe need two dimensions?
  - Strength of evidence: strong, medium, weak (this is p-value for statistical tests)
  - Whether contradicts hypothesis or not (maybe strong, neutral, fail-to-refute)

V. Organize evidence blocks by hypothesis into matrix

VI. Choose strongest hypothesis

- based on quality of evidence, quantity of evidence, strength of explanation based on evidence
- each block assigned numerical score based on coding scheme (strength of association & veritas rating)

VII. Strongest hypothesis checked

- agreed with overall state of historical and scientific knowledge
- satisfies guidelines for causation
- consistent with any definitive proof or admission (not applicable for economic problems, I think)

Still Much to Learn From John Snow

I. **Rollicking Good Tale** – full of heroism, death, and statistics

II. **Causal Inference**: template for how to marshal evidence in support of a causal explanation

III. **Statistics & Instruction**: The data are simple but the analysis demonstrates multiple data analytic tools we use today

- combining maps and data (GIS or geographic information systems)
- regression and error analysis
- difference-in-differences regression
- natural experiments and randomization

Snow's cholera work is also a humbling reminder of the sometimes meandering path towards truth: even with overwhelming evidence and strong analysis Snow failed to convince the medical establishment, the public, or the authorities

A Appendix Tables & Figures

A.1 Additional Broad Street Mapping

Contents

More Sophisticated Mapping – Walking Neighborhoods
MapsreturnMapsreturn

Snow Mapped Walking Neighborhood
MapsreturnMapsreturn

A.2 Broad Street Counts for Drinkers vs. Non-Drinkers

Contents

Counts for Drinkers vs Non-Drinkers

drinkersreturndrinkersreturn

Contingency Table Analysis for Drinkers vs Non-Drinkers
drinkersreturndrinkersreturn

A.3 Raw Mortality Rates for 1849 & 1854

Contents

Mortality Rates from Snow Table XII
SnowTableXIIreturnSnowTableXIIreturn

Mortality Rates from Snow Table XII
SnowTableXIIreturnSnowTableXIIreturn

A.4 Poisson Regressions 1849 & 1854

Contents

Poisson & Negative Binomial Regressions – DiDs
PoissonRegressionreturnPoissonRegressionreturn

A.5 Error Bars for Statistics Models of Difference-in-Differences (Counts)

Contents

A.6 Allocating Houses Across Regions

Contents

A.7 Poisson Regressions for "Next 16" Mixing (using data from Snow 1856)

Contents

Relative to Snow's statement, showing "the overwhelming influence which the nature of the water supply exerted over the mortality, overbearing every other circumstance which could be expected to affect the progress of the epidemic. Thus, in the crowded, dirty, and very poor subdistricts of Lambeth Church, first part, and Waterloo, first part, lying by the river side, the mortality was low in consequence of the water supply being chiefly that of the Lambeth Company; whilst in the thinly peopled, and comparatively genteel subdistricts of Clapham and Battersea the mortality was very high, in consequence of the impure water of the Southwark and Vauxhall Company." ([Snow(1856)] p. 248)

References

[Freedman(1991)] David Freedman. Statistical Models and Shoe Leather. *Sociological Methodology*, 21:291–313, 1991. ISSN 0883-4237, 2168-8745. doi: 10.2307/270939. URL `https://www-jstor-org.proxy.uchicago.edu/stable/270939`.

[Heckman and Singer(2017)] James J. Heckman and Burton Singer. Abducting Economics. *American Economic Review*, 107(5):298–302, May 2017. ISSN 0002-8282. doi: 10.1257/aer.p20171118. URL `https://www.aeaweb.org/articles?id=10.1257/aer.p20171118`.

[Hempel(2007)] Sandra Hempel. *The Strange Case of the Broad Street Pump: John Snow and the Mystery of Cholera*. University of California Press, Berkeley, first edition edition, January 2007. ISBN 978-0-520-25049-9.

[Johnson(2007)] Steven Johnson. *The Ghost Map: The Story of London's Most Terrifying Epidemic–and How It Changed Science, Cities, and the Modern World*. Riverhead Books, New York, reprint edition edition, October 2007. ISBN 978-1-59448-269-4.

[Katz and Singer(2007)] Rebecca Katz and Burton Singer. Can an Attribution Assessment Be Made for Yellow Rain? Systematic Reanalysis in a Chemical-and-Biological-Weapons Use Investigation. *Politics and the Life Sciences*, 26(1):24–42, March 2007.

[Koch and Denike(2006)] Thomas Koch and Kenneth Denike. Rethinking John Snow's South London study: A Bayesian evaluation and recalculation. *Social Science & Medicine*, 63 (1):271–283, July 2006. ISSN 0277-9536. doi: 10.1016/j.socscimed.2005.12.006. URL `http://www.sciencedirect.com/science/article/pii/S0277953605006933`.

[Lakatos(1980)] Imre Lakatos. *The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers*. Cambridge University Press, Cambridge, November 1980. ISBN 978-0-521-28031-0.

[Li()] Peter Li. R Package for Analyzing John Snow's 1854 Cholera Map. URL `https://github.com/lindbrook/cholera`.

[Snow(1849)] John Snow. *On the mode of communication of cholera*. John Churchill, London, 1849. OCLC: 14550757.

[Snow(1855)] John Snow. *On the mode of communication of cholera*. London : John Churchill, 2nd edition, 1855. URL `http://archive.org/details/b28985266`.

[Snow(1856)] John Snow. Cholera and the water supply in the south district of London in 1854. *Journal of Public Health and Sanitary Review*, 2:239–257, October 1856. URL `http://www.ph.ucla.edu/epi/snow/cholerawatersouthlondon.html`.

[Tufte(1997a)] Edward R. Tufte. *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*. Graphics Press, Cheshire, Conn, unknown edition edition, April 1997a. ISBN 978-0-9613921-3-0. https://www.edwardtufte.com/tufte/books_textb.

[Tufte(1997b)] Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, 1st edition edition edition, February 1997b. ISBN 978-1-930824-15-7. https://www.edwardtufte.com/tufte/books_visex.

[Vinten-Johansen et al.(2003)Vinten-Johansen, Brody, Paneth, Rachman, and Rip] Peter Vinten-Johansen, Howard Brody, Nigel Paneth, Stephen Rachman, and Michael Russell Rip. *Cholera, Chloroform and the Science of Medicine: A Life of John Snow*. Oxford University Press, Oxford ; New York, 1 edition edition, May 2003. ISBN 978-0-19-513544-2.

[Westminster and London School of Hygiene and Tropical Medicine(1855)] Westminster and London School of Hygiene and Tropical Medicine, editors. *Report on the cholera outbreak in the parish of St. james,Westminster, during the autumn of 1854*. J. Churchill, London, 1855.

[Zeileis(2004)] Achim Zeileis. Econometric Computing with HC and HAC Covariance Matrix Estimators | Zeileis | Journal of Statistical Software. *Journal of Statistical Software*, 11, November 2004. doi: 10.18637/jss.v011.i10. URL `https://www.jstatsoft.org/article/view/v011i10`.

Table 8: Theory & Hypotheses by Evidence Block

	T1: Water	T2: Miasma	T3: Class, Elevation, ...	Comment
Albion	Contradict: no Strength: na	Contradict: yes Strength: strong	Contradict: neut Strength: na	
Broad 1 – mapping	Contradict: no Strength: med	Contradict: no Strength: med	Contradict: yes Strength: med	
Broad 2 – cases	Contradict: no Strength: strong	Contradict: yes Strength: strong	Contradict: neut Strength: na	
Broad 3 – contin table	Contradict: no Strength: strong	Contradict: yes Strength: med	Contradict: yes Strength: med	"medium" for T2&T3: maybe could produce correlation between water & miasma
S London 1 – DiDs	Contradict: no Strength: strong	Contradict: yes Strength: med	Contradict: yes Strength: med	"medium" for T2&T3: maybe could produce correlation between water & miasma
S London 2 – Mixing	Contradict: no Strength: strong	Contradict: yes Strength: strong	Contradict: yes Strength: strong	Rules out confounders, strengthens water causality

Table 11: Mortality Rates from Cholera per 10,000 Persons in 1849 & 1854 (from Snow Table XII & using population in 1851 from Table VIII)

	Sub-Districts	Deaths rates from Cholera in 1849, per 10,000	Deaths rates from Cholera in 1854, per 10,000	Water Supplier	Degree of Lambeth Supply
1	St. Saviour, Southwark	144	188	SouthwarkVauxhall	dirty_none
2	St. Olave, Southwark	196	201	SouthwarkVauxhall	dirty_none
3	St. John, Horsleydown	169	130	SouthwarkVauxhall	dirty_none
4	St. James, Bermondsey	132	192	SouthwarkVauxhall	dirty_none
5	St. Mary Magdalen	186	175	SouthwarkVauxhall	dirty_none
6	Leather Market	148	155	SouthwarkVauxhall	dirty_none
7	Rotherhithe	198	158	SouthwarkVauxhall	dirty_none
8	Battersea	92	56	SouthwarkVauxhall	dirty_none
9	Wandsworth	115	178	SouthwarkVauxhall	dirty_none
10	Putney	15	17	SouthwarkVauxhall	dirty_none
11	Camberwell	132	135	SouthwarkVauxhall	dirty_none
12	Peckham	47	89	SouthwarkVauxhall	dirty_none
13	Christchurch, Southwark	160	71	SouthwarkVauxhall & Lambeth	more_Lambeth
14	Kent Road	147	96	SouthwarkVauxhall & Lambeth	less_Lambeth
15	Borough Road	197	170	SouthwarkVauxhall & Lambeth	less_Lambeth
16	London Road	144	52	SouthwarkVauxhall & Lambeth	more_Lambeth
17	Trinity, Newington	152	100	SouthwarkVauxhall & Lambeth	less_Lambeth
18	St. Peter, Walworth	149	130	SouthwarkVauxhall & Lambeth	less_Lambeth
19	St. Mary, Newington	102	66	SouthwarkVauxhall & Lambeth	less_Lambeth
20	Waterloo Road (1st)	137	41	SouthwarkVauxhall & Lambeth	more_Lambeth
21	Waterloo Road (2nd)	132	64	SouthwarkVauxhall & Lambeth	less_Lambeth
22	Lambeth Church (1st)	117	27	SouthwarkVauxhall & Lambeth	more_Lambeth
23	Lambeth Church (2nd)	203	72	SouthwarkVauxhall & Lambeth	less_Lambeth
24	Kennington (1st)	77	125	SouthwarkVauxhall & Lambeth	less_Lambeth
25	Kennington (2nd)	81	75	SouthwarkVauxhall & Lambeth	less_Lambeth
26	Brixton	55	33	SouthwarkVauxhall & Lambeth	less_Lambeth
27	Clapham	70	101	SouthwarkVauxhall & Lambeth	less_Lambeth
28	St. George, Camberwell	111	83	SouthwarkVauxhall & Lambeth	less_Lambeth
29	Norwood	5	25	Lambeth	all
30	Streatham	171	17	Lambeth	all
31	Dulwich	6	0	Lambeth	all
32	Sydenham	11	27	Lambeth	all
	First 12 sub-districts	135	147	first12	dirty_none
	Next 16 sub-districts	130	85	next16	some
	Last 4 sub-districts	85	19	last4	all
	TOTAL	130	104		some

Table 12: Poisson and Negative Binomial Regressions for Sub-District Difference-in-Differences 1849 vs 1854

	Poisson, Single treatment effect, robust SE	Poisson, Single treatment effect, sub-district FE, robust SE	Neg Binom, Single treatment effect	Neg Binom, two treatment effects
Treatment – "less Lambeth" (ln)	**-0.511**	**-0.511**	**-0.505**	**-0.344***
standard error	**0.212**	**0.234**	**0.248**	**0.250**
p-value	1.6%	2.9%	4.1%	17.0%
treatment (ratio)	1.67	1.67	1.66	1.41
Treatment – "more Lambeth" (ln)				**-1.137**
standard error				**0.356**
p-value				0.14%
treatment (ratio)				3.12
theta ("size" from Gamma mixing)			4.88	5.48
sub-district fixed effects	NO	YES	NO	NO
Residual Deviance (df, p-value)	1565 (52, p<1e-10)	456.8 (26, p<1e-10)	59.7 (52, p=22%)	60.0 (51, p=18%)
Single region / less Lambeth fixed effect	-0.036*		-0.033*	-0.065*
More Lambeth fixed effect				0.058*
Time fixed effect	0.084*	0.084*	0.063*	0.063*

Deaths by sub-district from 1849 and 1854 for the 28 sub-districts ("first 12" Southwark-only and "next 16" jointly-supplied) shown in [Snow(1855)] Table XII and my Table **??**, with population from Snow's Table VIII. Total 56 observations. * = ***not*** significant at the 10% level (robust errors for Poisson regression). The Poisson and Negative Binomial regressions are fitted with the R function glm (family=poisson) and the glm.nb function from the MASS package. The parameter "theta" is the size or θ for a "parametrization (1)" Negative Binomial (see appendix). Robust standard errors are calculated with the R "sandwich" package, using the default "HC3" for the adjusted variance-covariance matrix (see [Zeileis(2004)] and the R "sandwich" manual).