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

UChicago Spatial Study Group & City of Chicago Dep't of Public Health

Thomas S. Coleman

February 18 & 21 2020 Draft February 17, 2020

[*Snow & Causal Inference*]

[Coleman]

[Feb 2020]

**What Causes Cholera? Hugely Important in 1850s London**


Horrendous way to die – dehydration, convulsions, blue skin, die within hours

Scourge of mid-1800s London – 1831-32 6,526 dead; 1849 14,137; 1853-54 10,738
Massive uncertainty as to cause

- Bad air (miasma); bad breeding (poverty); bad ground (plague pits)

**Huge public health & policy question** – and one man knew the answer:

- John Snow & bad water – effort to prove contaminated water as causal agent

Image:

- https://wellcomeimages.org/indexplus/obf_images/6e/39/38b7c365914565dbd94fdf33b9ba.jpg
- Gallery: https://wellcomeimages.org/indexplus/image/V0010484.html
- Wellcome Collection gallery (2018-03-21): https://wellcomecollection.org/works/acghney7 CC-BY-4.0

**Why John Snow and 1850s Cholera?**
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

[Angrist and Pischke(2008)] p. ? credit Snow with the first application of Differences-in-Differences.

[Angrist and Pischke(2014)] discusses Snow p. 205 ff and reproduces Snow's Table XII.

[Greene()] p 228 calls this IV but I think it's not best thought of as IV.

**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).

**Outline**

# Contents

# 1 Overview: John Snow and the Story of Cholera

## 1.1 Cholera, John Snow, and Waterborne Theory

**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

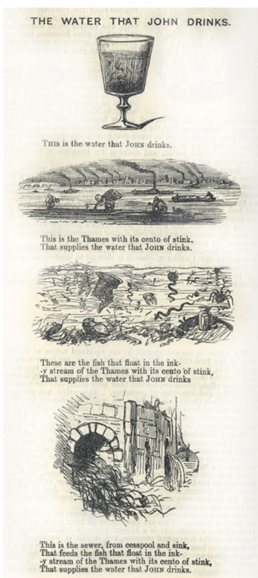
**Cholera Loved Victorian London**

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

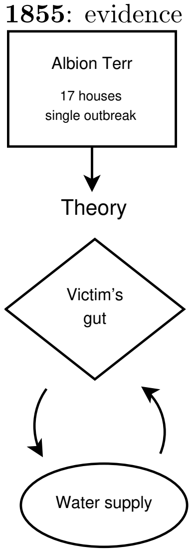
Implications for patterns of infection, across scales

- "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)

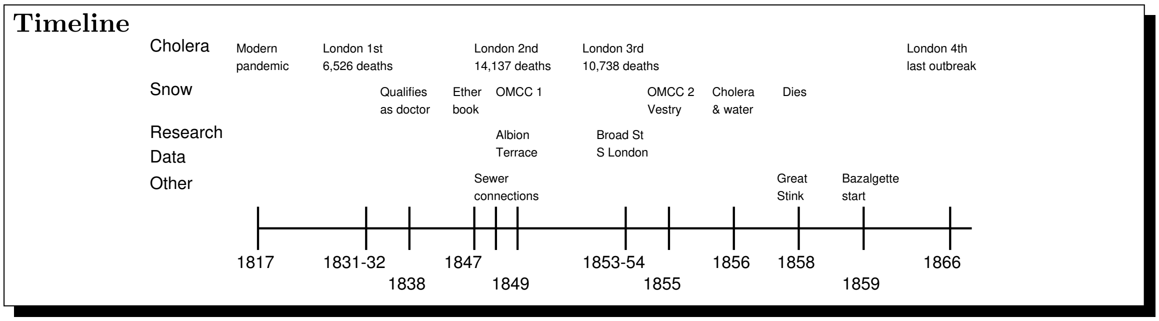
- Raw sewage associated with bad smells & dirty drinking water
- Lower class associated with crowding & poor sanitation

**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

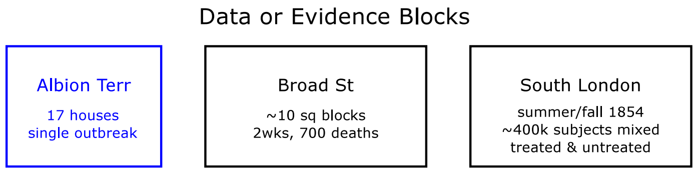
- 1849, Discovery of waterborne theory
- single event, 17 houses

II. Broad Street Outbreak

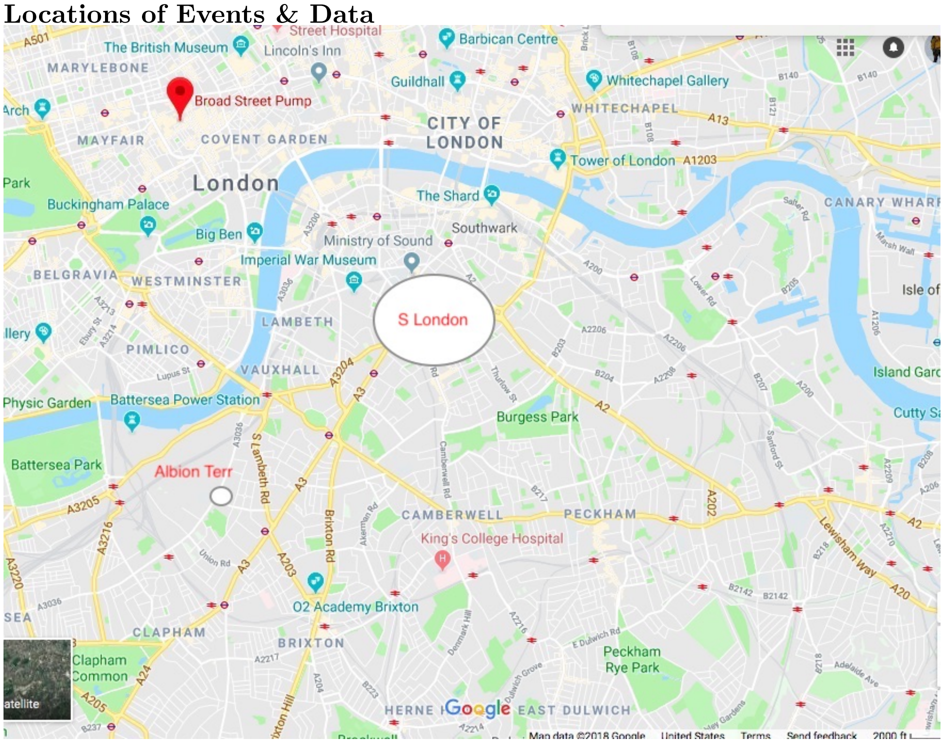
- Aug-Sep 1854, 700 deaths over roughly 2 weeks, 10 square blocks

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 John Snow's Evidence & Causal Inference

## 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

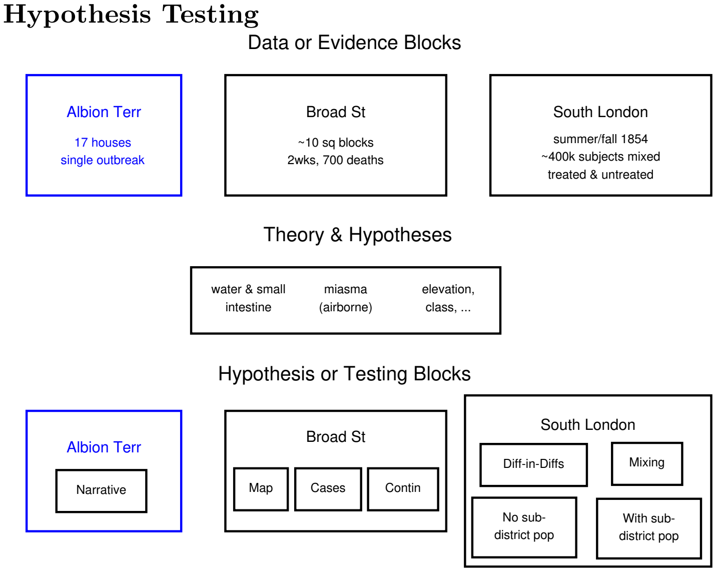
VI. Choose strongest hypothesis

- 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)

**Theory, Data, Hypothesis Testing**




## 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 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

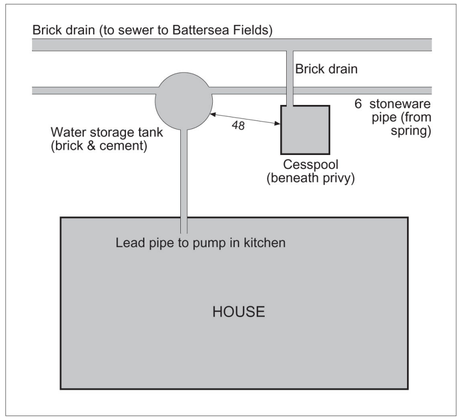
Details on Albion Terrace: http://www.ph.ucla.edu/epi/snow/1859map/albionterraceoutbreak_1849_a3.html

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)*

**Schematic of Cesspools & Water Tanks**

17 houses sharing common water source




from "Cholera, Chloroform, and the Science of Medicine", Vinten-Johansen et al.

- Storms July 26 & Aug 2nd, burst pipes and mixed cesspool with drinking water
- All 17 shared same water source, so all contaminated
- No surrounding houses affected

# 4 Broad Street Pump – Famous for "The Map"

## 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

- 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)*

**Tufte – The Classic Story of Snow's Map**
Tufte highlights aspects of Snow's analysis

- A *good idea* – a theory.
- "A shrewd intelligence about evidence, a clear logic of data display and analysis"
- A *good method*

Tufte emphasize four components of *good method*:

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

II. Making quantitative comparisons

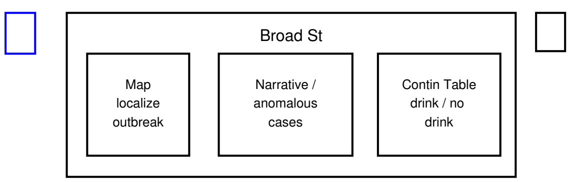
III. Considering alternative explanations and contrary cases

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

that I compress into three: Mapping; Cases & Anomalies; Quantitative & Statistics (with my contingency table contribution)

**Broad Street Pump Analysis – 3 Parts**
**Mapping**

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




- Icon: encapsulating and promoting waterborne theory

*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)

**Quantitative & Statistics** (also Whitehead, extending Snow)

- Statistical Tests of Clustering
- Contingency Testing – Drinkers vs Non-Drinkers and Survivorship Bias

## 4.1 Mapping & Tufte's Narrative

# Contents

**Snow's Data: Raw List → Time Series → Map**

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

The raw data were a list of deaths by date – Virtually useless,

So recast as time-series, which at least shows there was an epidemic

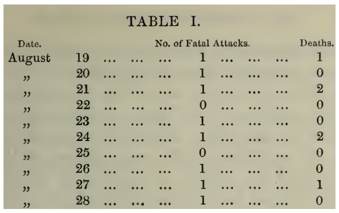
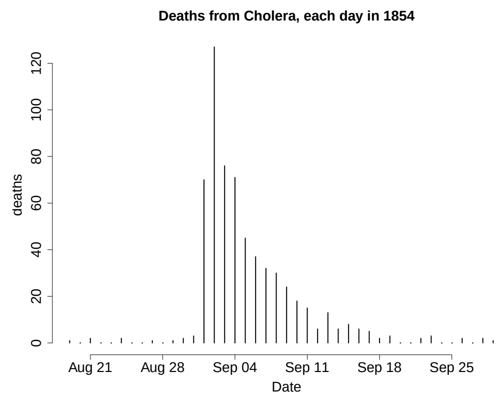


TABLE I.

| Date. | | No. of Fatal Attacks. | Deaths. |
|---|---|---|---|
| August | 19 | 1 | 1 |
| ,, | 20 | 1 | 0 |
| ,, | 21 | 1 | 2 |
| ,, | 22 | 0 | 0 |
| ,, | 23 | 1 | 0 |
| ,, | 24 | 1 | 2 |
| ,, | 25 | 0 | 0 |
| ,, | 26 | 1 | 0 |
| ,, | 27 | 1 | 1 |
| ,, | 28 | 1 | 0 |

Snow (1855) p 49




"descriptive narration is not causal explanation" (Tufte p 7)

Note for later: pump-handle off on Sep 8 – after peak

The analysis of Tufte ([Tufte(1997a), Tufte(1997b)])

"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's Maps – Analysis & Convincing Display**

Snow identified the pump just by walking the streets:

> *On proceeding to the spot, I found that nearly all of the deaths had taken place within a short distance of the pump (Snow p 39)*

But Snow needed more – a way to make it jump out to others

> *he knew ... 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 (Johnson p 140)*

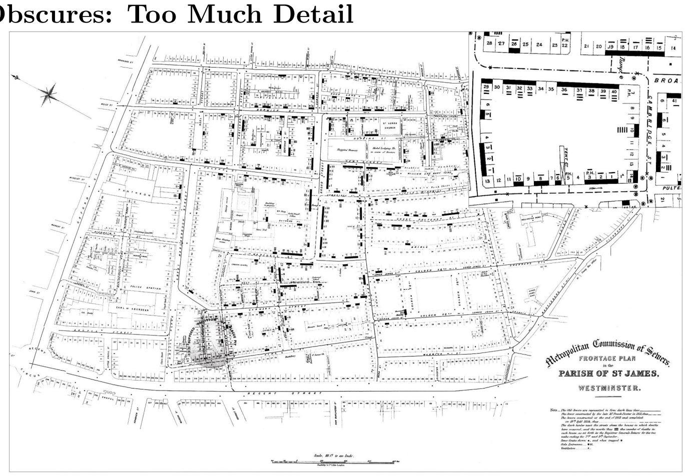
Snow was not the first to map the outbreak – Edmund Cooper, Metropolitan Commission of Sewers first

- Partly in response to concerns about Plague Pit, sewer line digging

Cooper's map was too busy, too much information

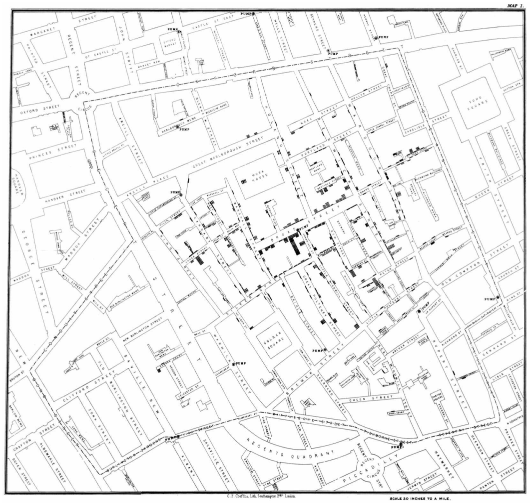
> *He could see at a glance that he'd be able to demonstrate that the outbreak was clustered around the pump, yet he knew from experience that that kind of evidence, on its own, would not satisfy a miasmatist. The cluster could just as easily reflect some pocket of poisoned air that had settled over 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*
>
> *Snow did not draw a map until December, 1854; the first spot map was produced in September of that year by Edmund Cooper, an engineer for the Metropolitan Commission of Sewers (see figure 3) [12]. Cooper's investigation resulted from public complaints linking the sewers to the cholera outbreak. Rumors held sway that sewer works had disturbed the soil of an ancient pit where bodies had been buried during the plague of 1665. Many feared that this process had freed or generated noxious gases that caused the cholera. Some alleged further that cholera deaths had been especially numerous in houses next to gully-holes, the openings through which sewer gases were vented to the surface. [Frerichs()], http://www.ph.ucla.edu/epi/snow/mapmyth/mapmyth3.html*

**Cooper's Map Obscures: Too Much Detail**




Cooper, from Vinten-Johansen at al Figure 12.4

**Snow was Masterful, Stripping Out Extraneous Detail**
Dot-Map demonstrates centrality of Broad Street pump

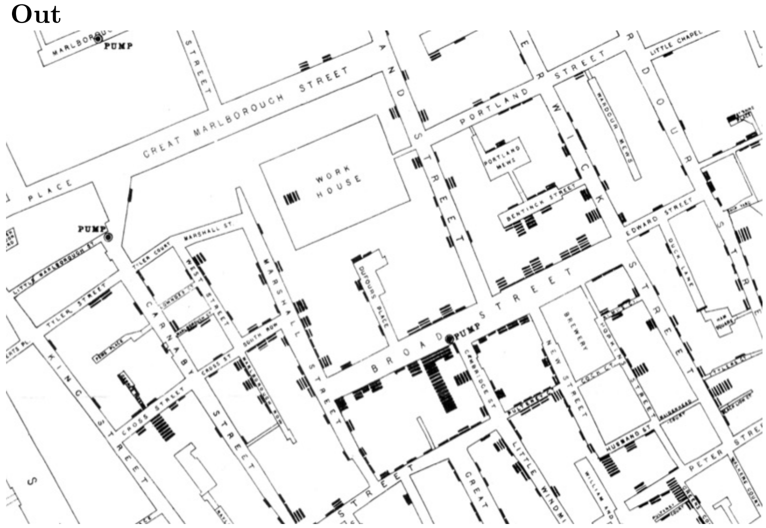


Snow's great contribution was to simplify & clarify – highlight the deaths and the pumps
Snow 1855

- Deaths & pumps only
- Deaths dark bars, pumps clearly marked

Clustering around pump jumps out

**Pump Jumps Out**




More mapping (quantitative analysis): mappingQuantAnalysismappingQuantAnalysis

### 4.2 Case Studies & Narrative: Tracking Individual Cases & Anomalies

## Contents

*There is a brewery in Broad Street, near to the pump, and on perceiving that no brewer's men were registered as having died of cholera, I called on Mr. Huggins, the proprietor. He informed me that there were above seventy workmen employed in the brewery, and that none of them had suffered from cholera—at least in severe form—only two having been indisposed, and that not seriously, at the time the disease prevailed. The men are* *allowed a certain quantity of malt liquor, and Mr. Huggins believes they do not drink water at all; and he is quite certain that the workmen never obtained water from the pump in the street. There is a deep well in the brewery, in addition to the New River water. ([Snow(1855)] p. 42)*

*The Workhouse in Poland Street is more than three-fourths surrounded by houses in which deaths from cholera occurred, yet out of five-hundred-thirty-five inmates only five died of cholera, the other deaths which took place being those of persons admitted after they were attacked. The workhouse has a pump-well on the premises, in addition to the supply from the Grand Junction Water Works, and the inmates never sent to Broad Street for water. If the mortality in the workhouse had been equal to that in the streets immediately surrounding it on three sides, upwards of one hundred persons would have died. ([Snow(1855)]p. 42)*

*Dr. Fraser also first called my attention to the following circumstances, which are perhaps the most conclusive of all in proving the connexion between the Broad Street pump and the outbreak of cholera. In the 'Weekly Return of Births and Deaths' of September 9th, the following death is recorded: 'At West End [Hampstead], on 2nd September, the widow of a percussion-cap maker, aged 59 years, diarrhea two hours, cholera epidemica sixteen hours.' I was informed by this lady's son that she had not been in the neighbourhood of Broad Street for many months. A cart went from Broad Street to West End every day, and it was the custom to take out a large bottle of the water from the pump in Broad Street, as she preferred it. The water was taken on Thursday, 31st August, and she drank of it in the evening, and also on Friday. She was seized with cholera on the evening of the latter day, and died on Saturday. . . . A niece, who was on a visit to this lady, also drank of the water; she returned to her residence, in a high and healthy part of Islington, was attacked with cholera, and died also. There was no cholera at the time, either at West End or in the neighbourhood where the niece died. ([Snow(1855)] p 44-45, quoted in [Tufte(1997a)] p 10)*

**Tufte 3: Alternative Explanations & Contrary Cases**
More Important than Map: Narratives & Anomalous Cases
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.

[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.

> 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)

**Imre Lakatos and "Protective Belt" of Auxiliary Hypotheses**
Scientific theories and the evidence to reject them are difficult things

- Evidence rarely or never speaks clearly and unambiguously – few "definitive experiments"
- Theories built on both "Core" & "Auxiliary" ("protective belt") hypotheses
- Evidence often rejects the (necessary) auxiliary hypotheses – core protected

We can only judge evidence in concert with judgement about theory

- Lakatos discusses Michelson Morley (speed-of-light) experiment
- Only in hindsight a "definitive" rejection of aether theory
- Many years' debate over "auxiliary" hypotheses of aether drag, ...

Snow's water-borne theory (and competitors) no different

- Must consider both core and auxiliary hypotheses
- Need to apply judgment to theory – data never speak unambiguously

**Anomalies to Test & Separate Theories**

- Water theory: evidence rejects neither core nor auxiliary
- Miasma: hard (but not impossible) to develop auxiliaries that protect core

(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

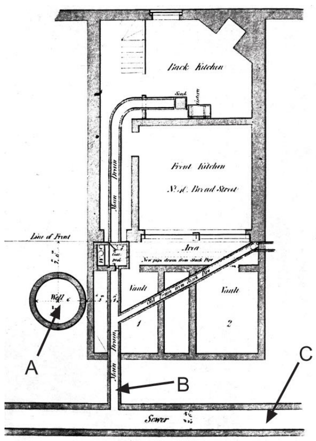
- Demonstrates that virtually any "core" can be protected by "auxiliary"
- An auxiliary we now recognize as foolish, cooked up to protect Miasma
- Miasma protected by auxiliary hypothesis allowing miasma to match drinking patterns

[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.*

**Additional Evidence & Analysis – Index Case**
Already compelling, Snow (& The Reverend Henry Whitehead, vicar of St Luke's church) did yet more

- Whitehead interviewed those who didn't die, to find out whether they drank from pump
  - If those who didn't die drank, evidence *against* water theory
  - Mortality: non-drinkers 1/10, drinkers 6/10
  - Trying to disprove theory & failing strngthens argument
- Whitehead identified *index case* at 40 Broad
  - Digging into pump showed leakage from 40 Broad into well




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

## Contents

**Making Quantitative Comparisons**
We see deaths clustered around Broad St pump – *But compared to what?*

I. Compared to other pumps, Broad St stands out
   - All areas densely populated – problem with maps that reflect population

II. Mortality among those who drank (6/10) vs those who did not (1/10)
   - Not in map – Whitehead's work for Vestry report

Comparison (1) helps identify Broad St, but not compare water vs miasma

- Could easily be miasma from pump

Comparison (2) helps disprove miasma

- Drinkers & non-drinkers would be equally at-risk from miasma
- Snow's theory and miasma predicted differently – miasma lost

*Whitehead began his assault on the pump-contamination theory by examining a crucial* *absence in Snow's original survey of the neighborhood. Snow had focused almost exclusively on the Soho residents who had perished in the outbreak, detecting that an overwhelming majority of them had consumed Broad Street water before falling ill. But Snow had not investigated the drinking patterns of the neighborhood residents who had* survived *the epidemic. If that group turned out to have drunk from the Broad Street pump at the same rate, then the whole basis for Snow's theory would dissolve. ... In the end, he [Whitehead] tracked down information on 497 residents of Broad Street, more than half the population that had lived there in the weeks before the outbreak. [Johnson(2007)] p. 173*

*Among the pump-water-drinking population, the rates of infection were along the lines that Snow had outline in his original survey: for every two Broad Street drinkers who were not affected, there were three who fell ill. That ratio seemed even more striking when you compared it to the infection rates among those who had not drunk from the well: only on in ten of that group had been seized with the cholera. [Johnson(2007)] p. 175*

**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

Contingency Table Analysis for Drinking versus Illness drinkersdetaildrinkersdetail

| Actual Counts | Not ill | Yes ill | **TOTAL** |
|---|---|---|---|
| No drink | 279 | 20 | **299** |
| Yes drink | 57 | 88 | **145** |
| **TOTAL** | **336** | **108** | **444** |

| Expected Counts | Not ill | Yes ill | **TOTAL** |
|---|---|---|---|
| No drink | 226.3 | 72.7 | **299** |
| Yes drink | 109.7 | 35.3 | **145** |
| **TOTAL** | **336** | **108** | **444** |

Fewer non-drinkers and more drinkers fall ill than expected if independent

- Statistical tests strongly reject independence (Pearson $\chi^2$ and Fisher exact $p$-value far lest than .01%)
- 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 drinking
  - 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

# 5 South London "Grand Experiment"

## Contents

**"Grand Experiment" – Water Supply Changes**
Two water companies served south London – Southwark & Vauxhall Co and Lambeth Co. – 486,936 customers, 300,000 **intimately mixed**

- In 1830s & 1840s companies competed for customers, often on same street

  > *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)*

1849 epidemic

- Both companies drew water from low in the Thames – near Vauxhall bridge

1852

- Lambeth Company moved source to Thames Ditton (upstream of London)
- In response to Act of Parliament, requiring move (by 1855)

1854 epidemic

- Southwark & Vauxhall Co supplied dirty water
- Lambeth Co supplied cleaner water

> *[Customers] were divided into two groups without their choice, and, in most cases, without their knowledge; one group being supplied water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity. (Snow 1855 p 75)*

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)

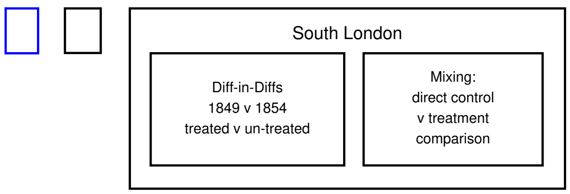
II. They served south London, part jointly, part separately (see [Snow(1855)]p. 68)

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)



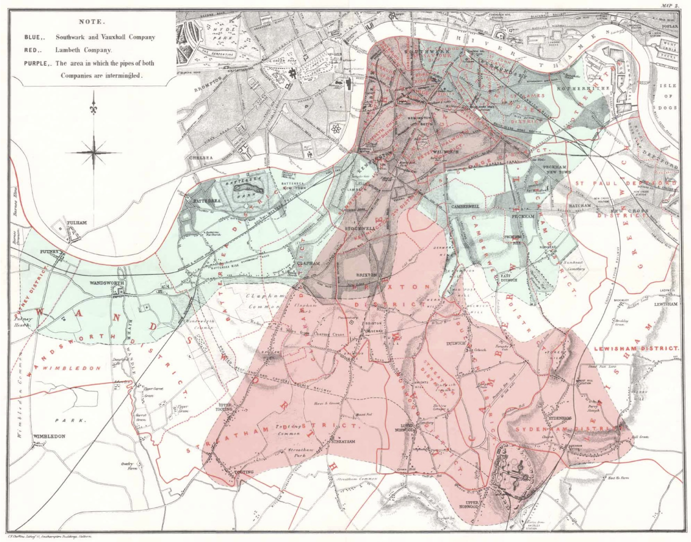

**Mixed** or quasi-random comparison

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

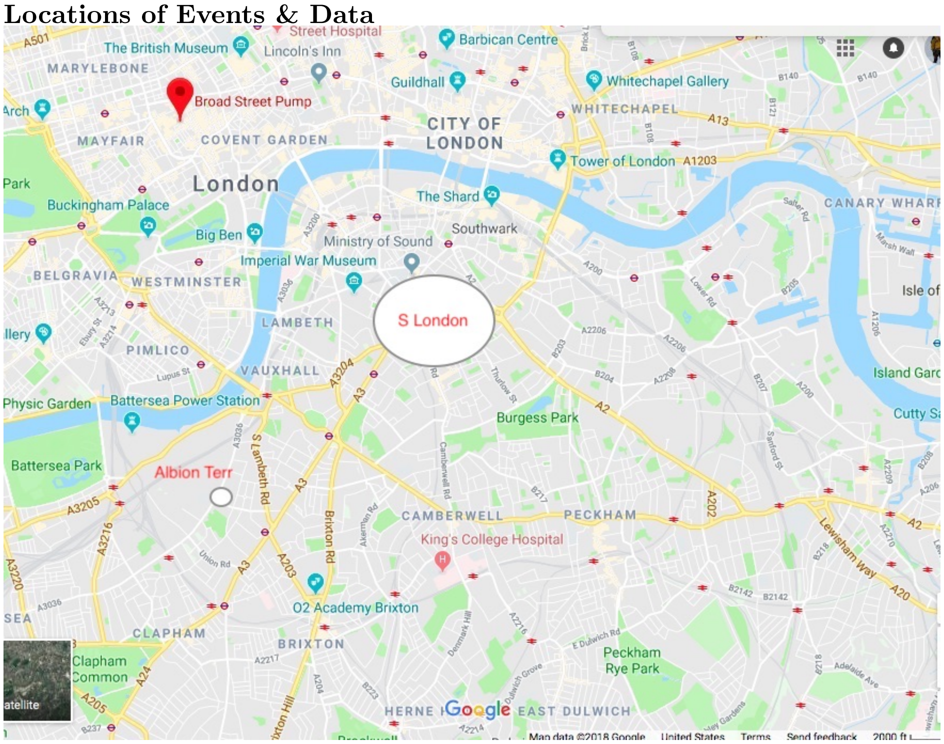
Registration Districts & Sub-Districts – Need to keep straight

- 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

## Locations of Events & Data




## Locations of Events & Data



**Learning From South London – Statistics & Methodology**
**Experimental Design & Control for Omitted Variables**
Early examples of two widely-used & valuable methodologies / designs

- Difference-in-differences: Exploit control vs treatment comparison
  - Use over-time comparison to control for confounding factors
  - Widely-used when experiment and randomization not possible
- Randomization & Mixing: Randomized Control Trial
  - Mixing by age, sex, class, income – controls for confounders

If clean vs dirty water shows big effect, hard to argue confounded by other factors

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

**Statistical Methodology – Careful Error Analysis**
Tempted to take large sample (400,000) as evidence of statistical significance

- Naive analysis (for DiD): $t$-ratio 11.7. Actually, closer to 2.0
- Using observed variation: what Stigler calls "intercomparison" (from Galton)

Extends Freedman (1991) idea to using statistical technique in concert with "good design, relevant data, and testing predictions against reality in a variety of settings."

### 5.1 Difference-in-Differences

# Contents

**Snow's "Before-vs-After" Comparison**

| | | | |
|---|---|---|---|
| Christchurch, Southwark | 256 | 113 | |
| Kent Road . . . | 267 | 174 | |
| Borough Road . . | 312 | 270 | |
| London Road . . | 257 | 93 | |
| Trinity, Newington . | 318 | 210 | |
| St. Peter, Walworth . | 446 | 388 | Lambeth Company, and Southwark and Vauxhall Compy. |
| St. Mary, Newington . | 143 | 92 | |
| Waterloo Road (1st) . | 193 | 58 | |
| Waterloo Road (2nd) . | 243 | 117 | |
| Lambeth Church (1st) . | 215 | 49 | |
| Lambeth Church (2nd) . | 544 | 193 | |
| Kennington (1st) . . | 187 | 303 | |
| Kennington (2nd) . | 153 | 142 | |
| Brixton . . . | 81 | 48 | |
| Clapham . . . | 114 | 165 | |
| St. George, Camberwell | 176 | 132 | |
| Norwood . . . | 2 | 10 | |
| Streatham . . . | 154 | 15 | Lambeth Company only. |
| Dulwich . . . | 1 | — | |
| Sydenham . . . | 5 | 12 | |
| First 12 sub-districts . | 2261 | 2458 | Southwk. & Vauxhall. |
| Next 16 sub-districts . | 3905 | 2547 | Both Companies. |
| Last 4 sub-districts . | 162 | 37 | Lambeth Company. |

Death statistics collected by government

- 1849 & 1854
- Snow copied, then summed up by sub-district
- Three regions, based on *water supplier*: Southwark&Vauxhall Co., Southwark Co. + Lambeth Co., Lambeth Co.

Exploit important fact:

- In 1852 (between 1849 & 1854) Lambeth changed to clean water – change in "treatment"

**Summarizing "Before-vs-After" Comparison**

> *[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)*

Population & Mortality (Counts), 1849 & 1854, Snow Table XII & Table VIII

| | 1851 Population | 1849 Deaths | 1854 Deaths |
|---|---|---|---|
| First 12 (Southwark & Vauxhall Water Company Only) | 167,654 | 2,261 | 2,458 |
| Next 16 (Joint Southwark & Vauxhall and Lambeth Companies) | 300,149 | 3,905 | 2,547 |
| TOTAL | 467,803 | 6,166 | 5,005 |

We can sharpen, considerably, tabulating as Diff-in-Diffs in rates (or log rates)

- Not sure why Snow didn't express as rates

**Better: Difference-in-Differences (1849 vs 1854)**

Mortality per 10,000 Persons, 1849 & 1854, Snow Table XII & Table VIII DiDdetailsDiDdetails

| Region or Sub-District Subtotals (Supplied by) | 1849 Before | 1854 After | Diff Before vs After |
|---|---|---|---|
| First 12 (Southwark & Vauxhall Co Only) – Dirty | 134.9 | 146.6 | *+11.8* |
| Next 16 (Joint Southwark & Vauxhall and Lambeth Cos) – Dirty / Clean | 130.1 | 84.9 | *–45.2* |
| Diff Water Supply Co.: Next 16 less First 12 | *-4.8* | *–61.8* | **–57.0** |

- Difference across regions to remove ("control for") regional differences
  - Diff in 1849 tells us "before treatment" difference: only -5
- Difference across time to remove ("control for") time differences
  - Diff for "First 12" shows pure time difference: +12
- Evidence that confounding factors not very important
- Difference the differences to produce *treatment effect*
  - Treatment effect = –57
  - Big reduction in mortality

Seems to support Snow's claim for "the overwhelming influence which the nature of the water supply exerted over the mortality" (1856 p248)

**Rules Out Most Everything Except Water**
Logic (mixing) and Data (1849) show "First 12" and "Next 16" similar

- Mixing: houses close and similar so miasma, elevation, weather, income, age, social class similar
- 1849: rates close when everyone gets dirty water

Rules out all those unobserved factors as causing differences in mortality rates

- If those factors similar should not cause differences
- 1849 shows no big differences in rates

Change water, now see difference

- 1854 different for "Next 16"

Doesn't "prove" water causes cholera, but hard to think of other explanations

**Naive Error Analysis for Difference-in-Differences – Wrong**
Like to think: sample of 467,864 overall ⇒ result is statistically significant

- Rates should be Binomial → Normal, so diff in column or row should have

$$SE(r1 - r2) = \sqrt{{}^{r1(1-r1)}/_{n1} + {}^{r2(1-r2)}/_{n2}}$$

Mortality per 10,000 Persons & Naive Error Analysis, 1849 & 1854

| | 1849 Deaths per 10,000 | 1854 Deaths per 10,000 | Diff 1854 less 1849 | Std Err of Diff | $t$-ratio |
|---|---|---|---|---|---|
| First 12 (Southwark & Vauxhall Water Company Only) | 134.9 | 146.6 | +11.8 | 4.07E-04 | 2.9 |
| Next 16 (Joint Southwark & Vauxhall and Lambeth Companies) | 130.1 | 84.9 | –45.2 | 2.66E-04 | –17.0 |
| Diff Water Supply Co.: Next 16 less First 12 | -4.8 | –61.8 | –57.0 | 4.86E-04 | –11.7 |
| Standard Error of Difference | 3.49E-04 | 3.38E-04 | 4.86E-04 | | |
| $t$-ratio | -1.4 | -18.3 | -11.7 | | |

But this is *wrong*: $t$-ratio of 11.7 is wrong, and actually closer to 2.0

- Variation across sub-districts & time imply rates & counts *not* Binomial

More detail on Difference-in-Differences: DiDdetailsDiDdetails

### 5.2 Mixing & Randomization

# Contents

**Mixing – Quasi-Randomized Control Trial**
Registrar-General recorded deaths weekly by sub-district – but not *water supplier*

- 16 sub-Districts (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

- The design provides close to random mixing
- Snow's data collection provided the needed data on deaths by supplier
- Randomization 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 Comparis**
Using Houses for all

- Includes "first
"next 16" mixe
- But – from dif

Houses, Deaths, and M

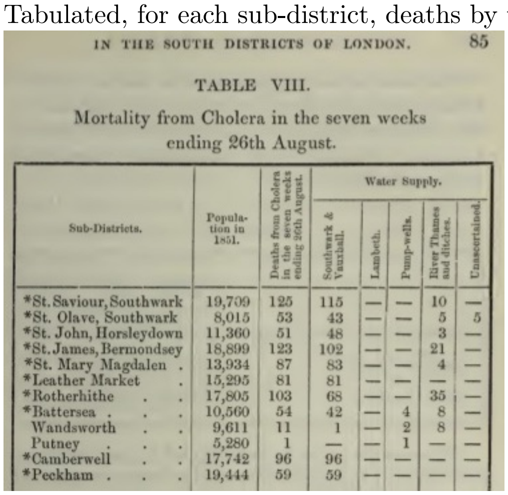
| Water |
|---|
| Southw |
| Lambe |
| Reduct |
| Naive |

Note that this co

Huge decrease – mo
Naive t-ratio –29.2,

- Still large, just

**Snow's "Shoe Leather" Work**
Tabulated, for each sub-district, deaths by water source



IN THE SOUTH DISTRICTS OF LONDON. 85

TABLE VIII.

Mortality from Cholera in the seven weeks ending 26th August.

| Sub-Districts. | Population in 1851. | Deaths from Cholera in the seven weeks ending 26th August. | Water Supply. Southwark & Vauxhall. | Lambeth. | Pump-wells. | River Thames and ditches. | Unascertained. |
|---|---|---|---|---|---|---|---|
| *St. Saviour, Southwark | 19,709 | 125 | 115 | — | — | 10 | — |
| *St. Olave, Southwark | 8,015 | 53 | 43 | — | — | 5 | 5 |
| *St. John, Horsleydown | 11,360 | 51 | 48 | — | — | 3 | — |
| *St. James, Bermondsey | 18,899 | 123 | 102 | — | — | 21 | — |
| *St. Mary Magdalen | 13,934 | 87 | 83 | — | — | 4 | — |
| *Leather Market | 15,295 | 81 | 81 | — | — | — | — |
| *Rotherhithe | 17,805 | 103 | 68 | — | — | 35 | — |
| *Battersea | 10,560 | 54 | 42 | — | 4 | 8 | — |
| Wandsworth | 9,611 | 11 | 1 | — | 2 | 8 | — |
| Putney | 5,280 | 1 | — | — | 1 | — | — |
| *Camberwell | 17,742 | 96 | 96 | — | — | — | — |
| *Peckham | 19,444 | 59 | 59 | — | — | — | — |

$SE(r1 - r2) = \sqrt{r1(1-r1)/n1 + r2(1-r2)/n2}$. Here r1=.03154, n1=40,046, r2=.00375, n2=26,107, so that SE=.000952 => t-ratio = -.02779/.000952 = -29.2

## 5.3 Detailed Error Analysis

# Contents

**Error Process / Statistical Model for Diff-in-Diffs**

Naive error analysis is wrong

Mortality per 10,000 Persons & Naive Error Analysis, 1849 & 1854

| | 1849 Deaths per 10,000 | 1854 Deaths per 10,000 | Diff 1854 less 1849 | Std Err of Diff | $t$-ratio |
|---|---|---|---|---|---|
| First 12 (Southwark & Vauxhall Water Company Only) | 134.9 | 146.6 | +11.8 | 4.07E-04 | 2.9 |
| Next 16 (Joint Southwark & Vauxhall and Lambeth Companies) | 130.1 | 84.9 | –45.2 | 2.66E-04 | –17.0 |
| Diff Water Supply Co.: Next 16 less First 12 | -4.8 | –61.8 | –57.0 | 4.86E-04 | –11.7 |
| Standard Error of Difference | 3.49E-04 | 3.38E-04 | 4.86E-04 | | |
| $t$-ratio | -1.4 | -18.3 | -11.7 | | |

Why? Large variation across and within sub-districts (mortality per 10,000)

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

| | Sub-Districts | 1849 | 1854 | Water Supplier |
|---|---|---|---|---|
| 1 | St. Saviour, Southwark | 144 | 188 | SouthwarkVauxhall |
| 8 | Battersea | 92 | 56 | SouthwarkVauxhall |

**Error Process / Statistical Model for Diff-in-Diffs**


| | Sub-Districts | 1849 | 1854 | Water Supplier |
|---|---|---|---|---|
| 1 | St. Saviour, Southwark | 144 | 188 | SouthwarkVauxhall |
| 8 | Battersea | 92 | 56 | SouthwarkVauxhall |


Exploit this variation to assess precision of our -57.0 estimate (-0.511 in logs)

- Stigler's "intercomparison" (from Galton)

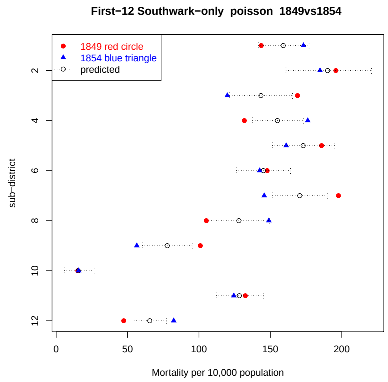
Need Statistical Model that maps our problem to usable mathematical framework

- 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
- Statistical Model 2: prob Poisson, but sub-districts vary – still not enough
- Statistical Model 3: random variation (mixture) in Poissons, across sub-districts & time
  - Poisson mixture, Gamma mixing ⇒ Negative Binomial Counts (deaths)

**Model 1: Poisson Same for All – Too Much Variation**

**Model 2: Poisson Varies by Sub-District – Still Too Much**

## Excess Variation (“Overdispersion”) Slightly Puzzling




Variation *across* sub-districts easy to understand

- Sub-districts characteristics (housing density, social class, ...) vary in ways that cause different mortality rates
- Easy to model: each sub-district has its own mean (fixed effect)

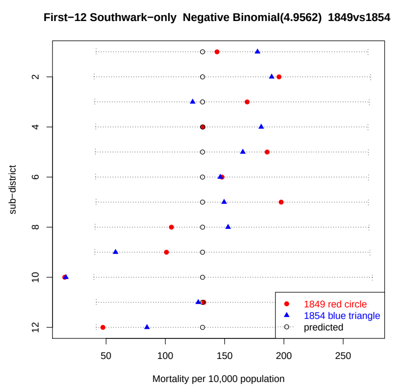
Variation *within* sub-districts harder – How can mortality *not* be Poisson?

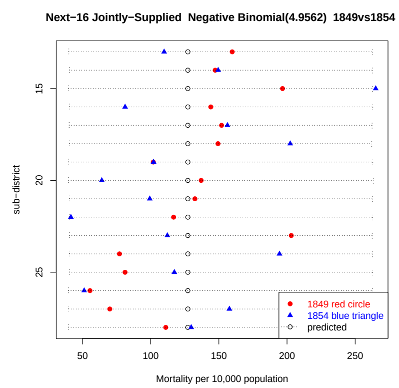
- Poisson good approx for mortality process
- Even if individuals different Poisson rates, sum of Poissons still Poisson
- Why does mortality vary in (seemingly) random manner?

Artificial example: tea drinkers (immune)

- Sub-districts vary in fraction of tea drinkers, and thus mortality
- But price of tea changes 1849-to-1854
- Sub-district changes appear random

## Model 3: Negative Binomial – Enough Variation




This statistical model “works” – consistent with data

## DiD Poisson Regressions – Inference (SEs) Wrong

| | 1<br>Poisson | 2<br>Poisson, sub-district Fixed Effects | 3<br>Negative Binomial | 4<br>Negative Binomial, 2 Lambeth Effects |
|---|---|---|---|---|
| **Single Treatment** | **-0.511** | **-0.511** | **-0.500** | **-0.338** |
| **standard error** | **0.039** | **0.039** | **0.246** | **0.248** |
| z-ratio (coeff/SE) | -13.20 | -13.20 | -2.03 | -1.36 |
| Robust z-ratio | -2.43 | -2.18 | -2.17 | -1.40 |
| **"More Lambeth" Treatment** | | | | **-1.132** |
| **standard error** | | | | **0.353** |
| z-ratio (coeff/SE) | | | | -3.20 |
| Robust z-ratio | | | | -3.84 |
| Joint region (single) control* | -0.036 | | -0.032 | -0.064 |
| Joint region (more Lambeth) control* | | | | 0.059 |
| Time control* | 0.084 | 0.084 | 0.057 | 0.057 |
| Residual Deviance | 1541.6 | 456.8 | 59.8 | 60.0 |
| p-value | 0.00% | 0.00% | 21.45% | 15.74% |
| theta (Gamma "size") | | | 4.96 | 5.57 |
| Pseudo-$R^2$ | 24.2% | 77.5% | 16.8% | 25.1% |

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, with population from Snow's Table VIII. Total 56 observations.

- Throw out Poisson & Poisson FE models – standard errors and inference wrong
- Estimates OK (-0.511 same as "by hand" in logs)

## DiD Negative Binomial – Single Treatment Marginal

| | 1<br>Poisson | 2<br>Poisson, sub-district Fixed Effects | 3<br>Negative Binomial | 4<br>Negative Binomial, 2 Lambeth Effects |
|---|---|---|---|---|
| **Single Treatment** | **-0.511** | **-0.511** | **-0.500** | **-0.338** |
| **standard error** | **0.039** | **0.039** | **0.246** | **0.248** |
| z-ratio (coeff/SE) | -13.20 | -13.20 | -2.03 | -1.36 |
| Robust z-ratio | -2.43 | -2.18 | -2.17 | -1.40 |
| **"More Lambeth" Treatment** | | | | **-1.132** |
| **standard error** | | | | **0.353** |
| z-ratio (coeff/SE) | | | | -3.20 |
| Robust z-ratio | | | | -3.84 |
| Joint region (single) control* | -0.036 | | -0.032 | -0.064 |
| Joint region (more Lambeth) control* | | | | 0.059 |
| Time control* | 0.084 | 0.084 | 0.057 | 0.057 |
| Residual Deviance | 1541.6 | 456.8 | 59.8 | 60.0 |
| p-value | 0.00% | 0.00% | 21.45% | 15.74% |
| theta (Gamma "size") | | | 4.96 | 5.57 |
| Pseudo-$R^2$ | 24.2% | 77.5% | 16.8% | 25.1% |

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, with population from Snow's Table VIII. Total 56 observations.

- Single Treatment Effect Only Marginally Significant
- Some sub-districts more Lambeth Co. customers – when split, get significance (-1.132 or factor of 3)

**Same for Quasi-Randomized: Poisson Doesn't Fit**

Poisson and Negative Binomial Regressions for Sub-District Mixing, Seven Weeks Ending 26th August

| | 1<br>Poisson | 2<br>Poisson, District Fixed Effects | 3<br>Negative Binomial | 4<br>Negative Binomial + Housing Density |
|---|---|---|---|---|
| **Lambeth (treatment) Effect** | **-2.101** | **-2.027** | **-2.099** | -2.097 |
| **standard error** | **0.104** | **0.107** | **0.194** | 0.177 |
| z-ratio (coeff/SE) | -20.15 | -18.93 | -10.84 | -11.86 |
| Robust z-ratio | -9.87 | -6.90 | -8.56 | -9.20 |
| Housing Density | | | | 0.215 |
| z-ratio (coeff/SE) | | | | 2.07 |
| Robust z-ratio | | | | 1.24 |
| Residual Deviance | 114.9 | 11.8 | 18.2 | 17.3 |
| p-value | 0.00% | 6.69% | 19.60% | 18.75% |
| theta (Gamma "size") | | | 12.08 | 16.42 |
| Pseudo-$R^2$ | 86.4% | 98.5% | 85.9% | 89.3% |

Data on deaths by District and by supplier (Southwark & Vauxhall Co versus Lambeth Co)

- Reject Poisson (see "Residual Deviance")
- Less data (no "across-time") so harder to decide on "Poisson FE" model 2, but probably no
- Negative Binomial: Treatment effect very large (-2.1 or factor of 8), even if include housing density

**Conclusion: Treatment Effect Survives, But not Simple**

I. The "Treatment Effect" of being a Lambeth Co. customer and getting clean water is statistically & substantively very significant

- But getting there is not easy
- Simple Binomial / Poisson assumption (standard for clinical trials) is rejected
- Need to broaden our thinking to random variation in mortality rates
- But – will be less important for small samples, where small-sample Poisson variation dominates

II. Some confidence that this result carries over to other regions, other periods

- DiD shows no large variation (in aggregate) over time
- Treatment effect survives observed variation across sub-districts (Stigler's *intercomparison*) so more likely to survive in other parts of London

Variance of Poisson is *count* so variance of rate is $^{count}/_{n^2} = {}^{rate}/_{n}$. Variance of Negative Binomial $count + {}^{count^2}/_{\theta}$ so the variance of the rate is $^{rate}/_{n} + {}^{rate^2}/_{\theta}$. For the sample sizes we are dealing with (say sub-district rates 0.02 and populations 15,000, $\theta = 5.0$) this gives

$$SD_{Poiss} = \sqrt{\frac{r}{n}} = \sqrt{\frac{.02}{15,000}} = .0012;\ SD_{NB} = \sqrt{\frac{r}{n} + \frac{r^2}{\theta}} = \sqrt{\frac{.02}{15,000} + \frac{.0004}{5}} = .0090$$

so that the Negative Binomial is 7.5-times larger.

For small sample sizes, say 200,

$$SD_{Poiss} = \sqrt{\frac{r}{n}} = \sqrt{\frac{.02}{200}} = .010;\ SD_{NB} = \sqrt{\frac{r}{n} + \frac{r^2}{\theta}} = \sqrt{\frac{.02}{15,000} + \frac{.0004}{5}} = .013$$

so that the Negative Binomial is 1.3-time larger.

**Supporting and Extending David Freedman's Comments**
This detailed analysis of Snow's work supports Freedman's (1991) comments about Snow:

> *Snow's work is ... a success story for scientific reasoning based on nonexperimental data*
>
> *statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings,*

But it modifies Freedman's skepticism about statistical arguments

> *I do not think that regression can carry much of the burden in a causal argument, [and] Arguments based on statistical significance of coefficients seem generally suspect.*

to a more nuanced view: Snow's work proves the importance of marrying good design with good statistical analysis

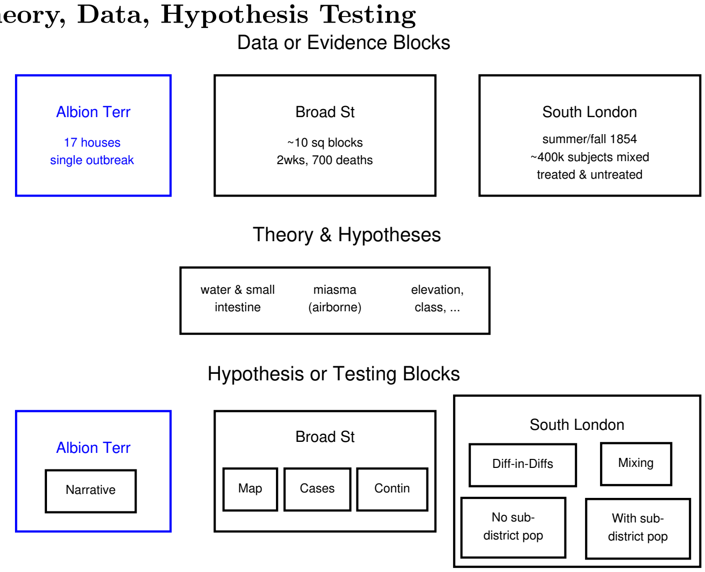
[Freedman(1999), Freedman(1991)] discusses Snow's work with particular focus on the methodology, arguing that "Snow's work is ... a success story for scientific reasoning based on nonexperimental data." ([Freedman(1991)] p 291) Freedman also argues that "statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings," ([Freedman(1991)] p 291) and that Snow's work provides a wonderful example of good design, relevant data, etc..

I have written this essay precisely to support and extend Freedman's assertion. Nonetheless I differ from Freedman on his somewhat negative view of the value of regression and other statistical tools ("I do not think that regression can carry much of the burden in a causal argument," and "Arguments based on statistical significance of coefficients seem generally suspect." ([Freedman(1991)] p 292)

# 6 Conclusion

# Contents

**Conclusion: Theory, Data, Hypothesis Testing**




**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

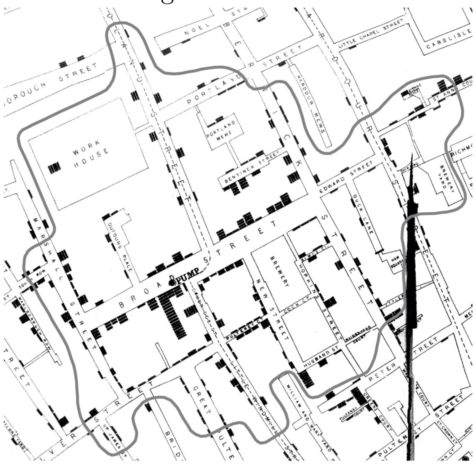
## A.1 Quantitative Analysis of Maps – Walking Neighborhoods

# Contents

**Major Innovation by Snow – Walking Neighborhood**
Snow's version for the 1855 Vestry Report adds in "walking neighborhood"

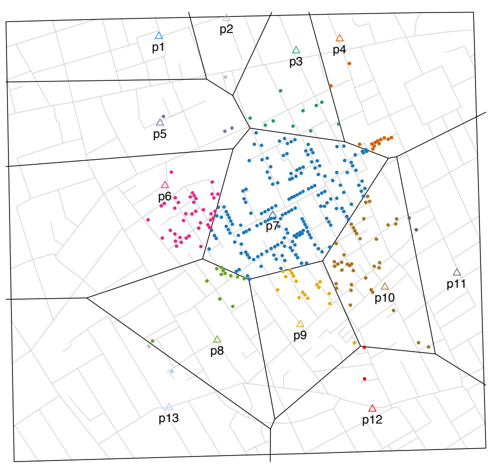
- Shows all deaths "equal walking distance" to Broad St pump
- Carries on Tufte's idea of "Quantiative Comparisons"
- Allows comparison of regions where pumps close or far
- Also, corrected pump position to 40 Broad St
- mapQuantReturnmapQuantReturn

**Major Innovation by Snow – Walking Neighborhood**
Detail showing the outline




- Shows all deaths "equal walking distance" to Broad St pump
- Neighborhood stretches out along streets
- Allows comparison of regions where pumps close or far

**Building on Snow's Neighborhoods - Voronoi**
Fun with R Package "cholera"
Start with "Voronoi Neighborhoods": Boundaries equidistant from pumps




- Examine how many deaths within Pump 7 region
- Versus other regions
- What about Pump 6? (Marlborough)
- Bad taste

**Formalize Testing for "Actual vs Predicted"**

Formal statistical testing of how many deaths in pump neighborhoods

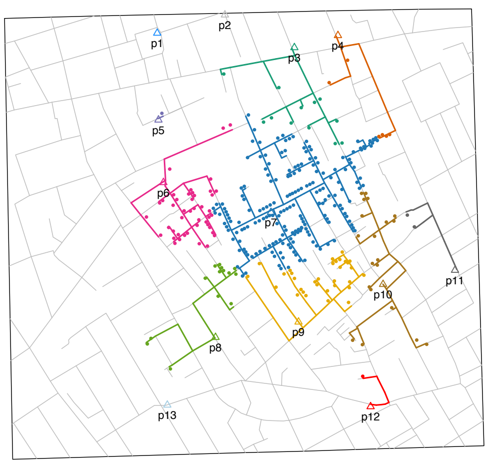
| pump.id | Count | Percent | Expected | Pearson |
|---|---|---|---|---|
| 1 | 0 | 0 | 19.5 | 19.5 |
| 2 | 1 | 0.31 | 6.2 | 4.4 |
| 3 | 10 | 3.12 | 14.0 | 1.1 |
| 4 | 13 | 4.05 | 30.4 | 10.0 |
| 5 | 3 | 0.93 | 26.5 | 20.8 |
| 6 | 39 | 12.15 | 39.9 | 0.0 |
| 7 | 182 | 56.7 | 27.2 | 881.0 |
| 8 | 12 | 3.74 | 22.1 | 4.6 |
| 9 | 17 | 5.3 | 15.5 | 0.1 |
| 10 | 38 | 11.84 | 19.0 | 19.0 |
| 11 | 2 | 0.62 | 24.6 | 20.8 |
| 12 | 2 | 0.62 | 29.7 | 25.8 |
| 13 | 2 | 0.62 | 46.4 | 42.5 |
| Sum | 321 | | Sum Sq | 1049.7 |

- "Expected" or "Predicted" is if deaths were even across the map
- "Pearson" is "Pearson's chi-squared statistic": $(act - exp)^2/exp$
- Large sum means actual is not random

mapQuantReturnmapQuantReturn

**Walking Neighborhoods**

Even more fun – equal walking distance

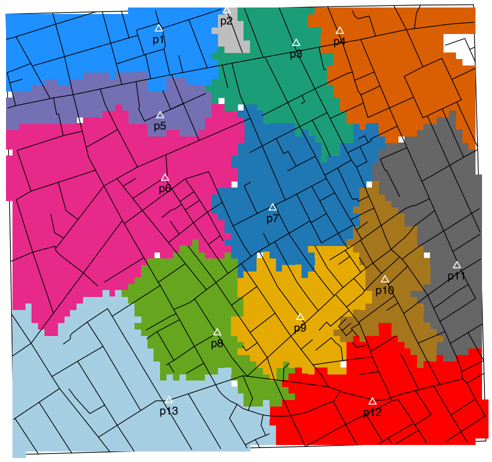



- Examine how many deaths within Pump 7 neighborhood
- Versus other regions
- What about Pump 6? (Marlborough)
- Bad taste

Number of deaths in each "pump neighborhood", calculated using the "neighborhoodWalking()" function of [Li()]

**Walking Neighborhoods**

Here are filled-in neighborhoods – put many cases on streets, figure out which pump is closest (by walking along street)




- Can use this to ask "how many deaths in a neighborhood?"
- Compare actual vs predicted

mapQuantReturnmapQuantReturn

plot(neighborhoodWalking(case.set = "expected"), type = "area.points")

**Formalize Testing for "Actual vs Predicted"**

Formal statistical testing of how many deaths in pump neighborhoods

| pump.id | Actual | Expected | Pearson |
|---|---|---|---|
| 1 - Market Place | 0 | 23.0 | 23.0 |
| 2 - Adam and Eve Court | 0 | 1.7 | 1.7 |
| 3 - Berners Street | 12 | 19.3 | 2.8 |
| 4 - Newman Street | 6 | 26.6 | 16.0 |
| 5 - Marlborough Mews | 1 | 13.8 | 11.9 |
| 6 - Little Marlborough Street | 44 | 55.8 | 2.5 |
| 7 - Broad Street | 189 | 27.6 | 942.4 |
| 8 - Warwick Street | 14 | 21.4 | 2.5 |
| 9 - Bridle Street | 32 | 19.9 | 7.4 |
| 10 - Rupert Street | 20 | 15.0 | 1.7 |
| 11 - Dean Street | 2 | 25.0 | 21.2 |
| 12 - Tichborne Street | 1 | 28.6 | 26.6 |
| 13 - Vigo Street | 0 | 43.2 | 43.2 |
| Sum | 321 | 321 | 1102.8 |

- "Expected" or "Predicted" is if deaths were even across the map
- "Pearson" is "Pearson's chi-squared statistic": $(act - exp)^2/exp$
- Large sum means actual is not random

mapQuantReturnmapQuantReturn

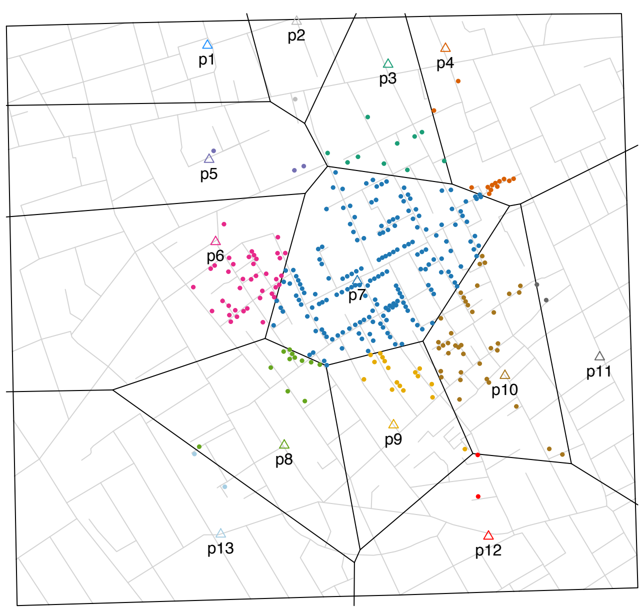
Analysis Using the "Voronoi Neighborhood" functions of [Li()]

| pump.id | Count | Percent | Expected | Pearson |
|---|---|---|---|---|
| 1 | 0 | 0 | 19.5 | 19.5 |
| 2 | 1 | 0.31 | 6.2 | 4.4 |
| 3 | 10 | 3.12 | 14.0 | 1.1 |
| 4 | 13 | 4.05 | 30.4 | 10.0 |
| 5 | 3 | 0.93 | 26.5 | 20.8 |
| 6 | 39 | 12.15 | 39.9 | 0.0 |
| 7 | 182 | 56.7 | 27.2 | 881.0 |
| 8 | 12 | 3.74 | 22.1 | 4.6 |
| 9 | 17 | 5.3 | 15.5 | 0.1 |
| 10 | 38 | 11.84 | 19.0 | 19.0 |
| 11 | 2 | 0.62 | 24.6 | 20.8 |
| 12 | 2 | 0.62 | 29.7 | 25.8 |
| 13 | 2 | 0.62 | 46.4 | 42.5 |
| Sum | 321 | | Sum Sq | 1049.7 |

Number of deaths in each "pump neighborhood", calculated using the "neighborhoodVoronoi()" function of [Li()]

The "Expected" counts are based on area of the Voronoi neighborhoods, assuming that deaths are uniformly distributed across the plane. The "Pearson" is (Count-Exp)^2/Exp and the sum of squares should be distributed as chi-squared. (A sum of 1,050 is very large so there is essentially zero probability that the observed count would be observed if deaths were uniformly distributed.)

Broad St pump = "pump 7" and Marlborough St = "pump 6"



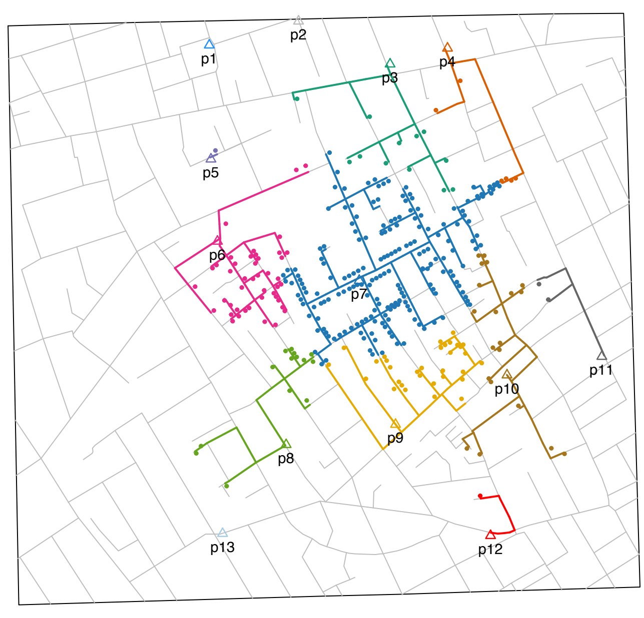

Analysis Using the "Walking Neighborhood" functions of [Li()]

| Pump Number | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of Deaths | 12 | 6 | 1 | 44 | 189 | 14 | 32 | 20 | 2 | 1 |
| Excluding Pump 6 | 12 | 6 | 6 | | 224 | 18 | 32 | 20 | 2 | 1 |

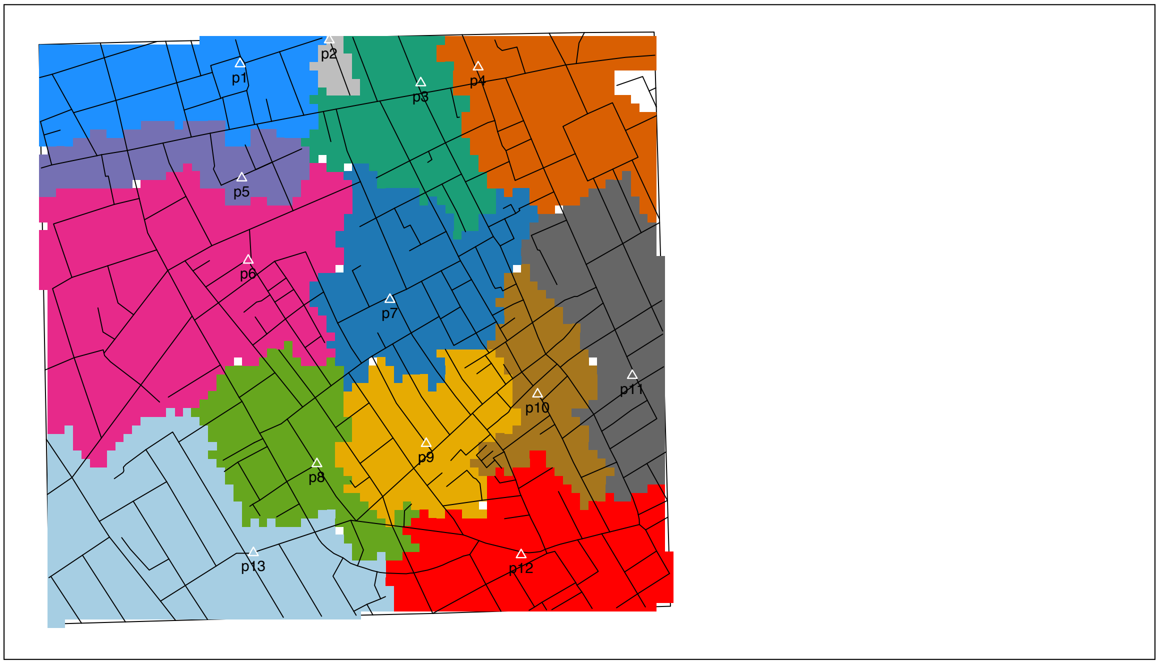
| pump.id | Walking Count | Expected (4972) | Expected (normalized) | Pearson | Voronoi Count |
|---|---|---|---|---|---|
| 1 | 0 | 356 | 23.0 | 23.0 | 0 |
| 2 | 0 | 27 | 1.7 | 1.7 | 1 |
| 3 | 12 | 299 | 19.3 | 2.8 | 10 |
| 4 | 6 | 412 | 26.6 | 16.0 | 13 |
| 5 | 1 | 214 | 13.8 | 11.9 | 3 |
| 6 | 44 | 865 | 55.8 | 2.5 | 39 |
| 7 | 189 | 428 | 27.6 | 942.4 | 182 |
| 8 | 14 | 331 | 21.4 | 2.5 | 12 |
| 9 | 32 | 308 | 19.9 | 7.4 | 17 |
| 10 | 20 | 232 | 15.0 | 1.7 | 38 |
| 11 | 2 | 388 | 25.0 | 21.2 | 2 |
| 12 | 1 | 443 | 28.6 | 26.6 | 2 |
| 13 | 0 | 669 | 43.2 | 43.2 | 2 |
| Sum | 321 | 4972 | 321 | 1102.8 | 321 |

Number of deaths in each “pump neighborhood”, calculated using the “neighborhoodWalking()” function of [Li()]

Broad St pump = “pump 7” and Marlborough St = “pump 6”




plot(neighborhoodWalking(case.set = "expected"), type = "area.points")




### 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 More Detail for Difference-in-Differences

# Contents

**Writing Table With Variables**

| Region (Sub-Districts) Supplied by | 1849 Deaths per 10,000 | 1854 Deaths per 10,000 | Diff in Time |
|---|---|---|---|
| "First 12" Southwark Only | 135 | 147 | *+12* |
| "Next 16" Jointly Supplied | 130 | 85 | *-45* |
| Diff Joint less Southwark | *-5* | *-62* | **-57** |

- Time Effect: $\delta_{54}$ captures any difference between 1849 & 1854
- Region Effect: $\gamma_J$ captures any difference between "First 12" versus "Next 16"
- Treatment Effect: $\beta$ captures the effect of clean water
- **We care about the treatment effect** $\beta$
- Worry about region ($\gamma_J$) and time ($\delta_{54}$) effects
- Control by differencing – across *region* and across *time* ("difference-in-differences")

| Region (Sub-Districts) Supplied by | 1849 Deaths per 10,000 | 1854 Deaths per 10,000 | Diff 1854 less 1849 |
|---|---|---|---|
| "First 12" Southwark Only | $\mu$ | $\mu + \delta_{54}$ | $\delta_{54}$ |
| "Next 16" Jointly Supplied | $\mu + \gamma_J$ | $\mu + \gamma_J + \delta_{54} + \beta$ | $\delta_{54} + \beta$ |
| Diff Joint less Southwark | $\gamma_J$ | $\gamma_J + \beta$ | $\beta$ |

**Write Difference-in-Differences as Equation**

$$R_{rt} = \mu + \gamma_J \cdot I_{r=J} + \delta_{54} \cdot I_{t=54} + \beta \cdot I_{r=J} \cdot I_{t=54}$$

With appropriately chosen Indicators:

| Region (Sub-Districts) Supplied by | 1849 Deaths per 10,000 | 1854 Deaths per 10,000 | Diff 1854 less 1849 |
|---|---|---|---|
| "First 12" Southwark Only | $I_{r=J} = 0$<br>$I_{t=54} = 0$ | $I_{r=J} = 0$<br>$I_{t=54} = 1$ | |
| "Next 16" Jointly Supplied | $I_{r=J} = 1$<br>$I_{t=54} = 0$ | $I_{r=J} = 1$<br>$I_{t=54} = 1$ | |
| Diff Joint less Southwark | | | |

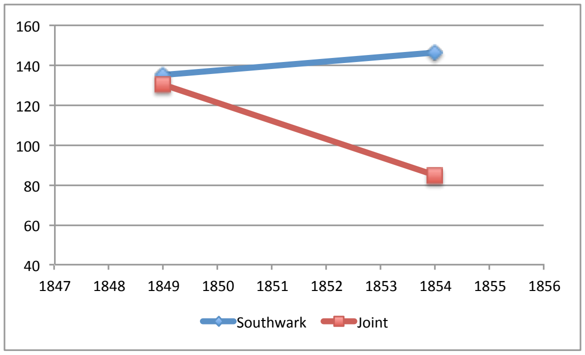
Get same table:

| Region (Sub-Districts) Supplied by | 1849 Deaths per 10,000 | 1854 Deaths per 10,000 | Diff 1854 less 1849 |
|---|---|---|---|
| "First 12" Southwark Only | $\mu$ | $\mu + \delta_{54}$ | $\delta_{54}$ |
| "Next 16" Jointly Supplied | $\mu + \gamma_J$ | $\mu + \gamma_J + \delta_{54} + \beta$ | $\delta_{54} + \beta$ |
| Diff Joint less Southwark | $\gamma_J$ | $\gamma_J + \beta$ | $\beta$ |

DiDreturnDiDreturn

**Graphing the Treatment Effect**
Comparing the "Southwark Only" vs "Joint" regions:

- They look very similar in 1849 – $\gamma_J$ small, looks like regions the same
- Useful – the regions look comparable. More confidence that the change in 1854 in the joint area is only due to water




**Calculating Treatment Effect in Logs: -0.51, 1.67x**
Usually want to compare *rates* in log (ratio) terms

- Rates cannot go negative
- Logs ensures we can't go negative

Equation becomes

$$\ln R_{rt} = \mu + \gamma_J \cdot I_{r=J} + \delta_{54} \cdot I_{t=54} + \beta \cdot I_{r=J} \cdot I_{t=54}$$

Table becomes

| Region or Sub-Districts – Supplied by | 1849 Death Rate (log) | 1854 Death Rate (log) | Diff 1854 less 1849 |
|---|---|---|---|
| First 12 – Southwark Only | $\ln(.0135) = -4.306$ | $\ln(.0147) = -4.223$ | 0.084 |
| Next 16 – Joint Southwark and Lambeth | $\ln(.0130) = -4.342$ | $\ln(.0085) = -4.769$ | -0.427 |
| Diff Joint less Southwark | -0.036 | -0.547 | **-0.511** |

–0.511 says (partially) clean water reduces death by 1.67x (exp(–0.511)) DiDreturnDiDreturn

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

## A.4 Raw Mortality Rates for 1849 & 1854

# Contents

**Mortality Rates from Snow Table XII**
SnowTableXIIreturnSnowTableXIIreturn

**Mortality Rates from Snow Table XII**
SnowTableXIIreturnSnowTableXIIreturn

## References


[Angrist and Pischke(2008)] Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, 1 edition edition, December 2008.

[Angrist and Pischke(2014)] Joshua D. Angrist and Jorn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton University Press, Princeton ; Oxford, with french flaps edition edition, December 2014. ISBN 978-0-691-15284-4.

[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`.

[Freedman(1999)] David Freedman. From association to causation: some remarks on the history of statistics. *Statistical Science*, 14(3):243–258, August 1999. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009212409. URL `https://projecteuclid.org/euclid.ss/1009212409`.

[Frerichs()] Ralph R. Frerichs. John Snow Site. URL `http://www.ph.ucla.edu/epi/snow.html`.

[Greene()] William Greene. *Econometric Analysis, 8th Edition.* ISBN 978-0-13-446136-6. URL `https://www.pearson.com/us/higher-education/program/Greene-Econometric-Analysis-8th-Edition/PGM3`

[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.

[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`.

[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, February 1997b. ISBN 978-1-930824-15-7. https://www.edwardtufte.com/tufte/books_visex.

[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.

Table 3: Theory & Hypotheses by Evidence Block

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

Table 6: 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 |