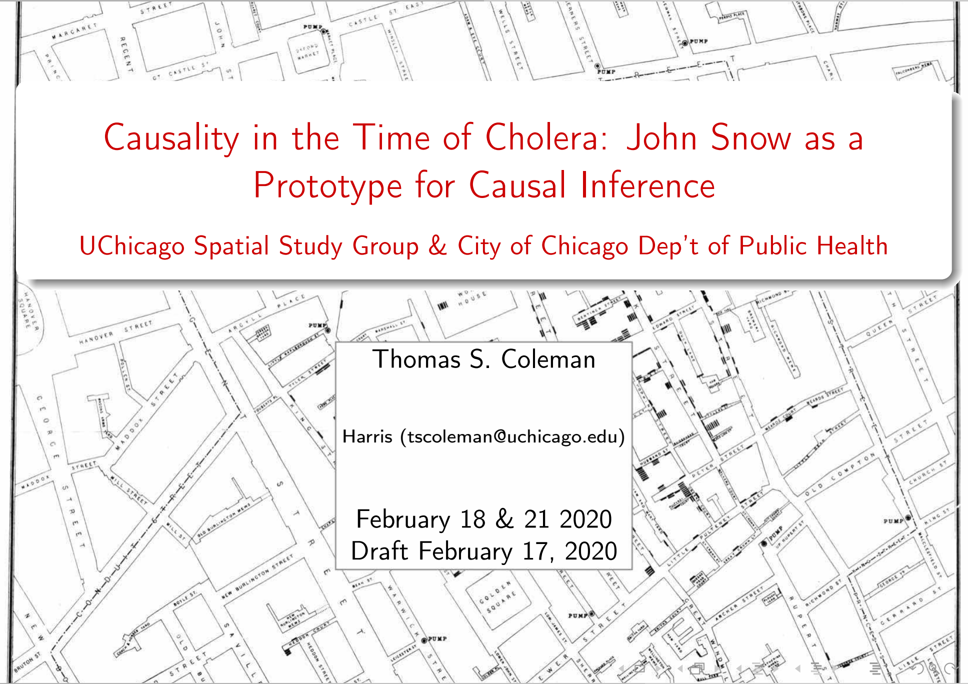
Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference

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

Thomas S. Coleman

Harris (tscoleman@uchicago.edu)

February 18 & 21 2020
Draft February 17, 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

Why John Snow and 1850s Cholera?

Three reasons:

1. **Rollicking Good Tale** – full of heroism, death, and statistics
2. **Causal Inference** – template for how to marshal evidence in support of a causal explanation
3. **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

Why John Snow and 1850s Cholera?

Three reasons:

1. **Rollicking Good Tale** – full of heroism, death, and statistics
2. **Causal Inference** – template for how to marshal evidence in support of a causal explanation
3. **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

Why John Snow and 1850s Cholera?

Three reasons:

1. **Rollicking Good Tale** – full of heroism, death, and statistics
2. **Causal Inference** – template for how to marshal evidence in support of a causal explanation
3. **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

Outline

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

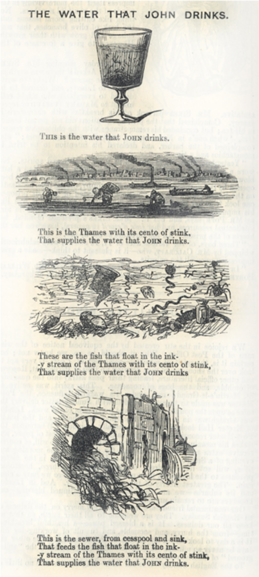
Cholera Loved Victorian London

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



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

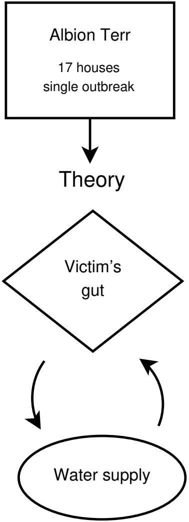
Implications for patterns of infection, across scales

- "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



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

Timeline


Cholera
Snow
Research
Data
Other
Modern pandemic
London 1st 6,526 deaths
London 2nd 14,137 deaths
London 3rd 10,738 deaths
London 4th last outbreak
Qualifies as doctor
Ether book
OMCC 1
OMCC 2 Vestry
Cholera & water
Dies
Albion Terrace
Broad St S London
Sewer connections
Great Stink
Bazalgette start
1817
1831-32
1838
1847
1849
1853-54
1855
1856
1858
1859
1866

I discuss Three Strands or "Blocks" of Evidence

1. Albion Terrace
   - 1849, Discovery of waterborne theory
   - single event, 17 houses
2. Broad Street Outbreak
   - Aug-Sep 1854, 700 deaths over roughly 2 weeks, 10 square blocks
3. South London "Grand Experiment"
   - Summer & Fall 1854, customers supplied by two water companies
   - large scale, 400k mixed (quasi-random) subjects

Data or Evidence Blocks

Albion Terr	Broad St	South London
17 houses single outbreak	~10 sq blocks 2wks, 700 deaths	summer/fall 1854 ~400k subjects mixed treated & untreated

Locations of Events & Data

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?"

1. Divide evidence into blocks or types of evidence
2. Assign to each block a *veritas* rating – quality of data
3. Develop groups of hypotheses
4. 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)
5. Organize evidence blocks by hypothesis into matrix
6. Choose hypothesis not contradicted
7. Strongest hypothesis checked

Theory, Data, Hypothesis Testing

Data or Evidence Blocks

Albion Terr
17 houses
single outbreak

Broad St
~10 sq blocks
2wks, 700 deaths

South London
summer/fall 1854
~400k subjects mixed
treated & untreated

Theory & Hypotheses

water & small intestine	miasma (airborne)	elevation, class, ...

Hypothesis or Testing Blocks

Albion Terr
- Narrative

Broad St
- Map
- Cases
- Contin

South London
- Diff-in-Diffs
- Mixing
- No sub-district pop
- With sub-district pop

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)

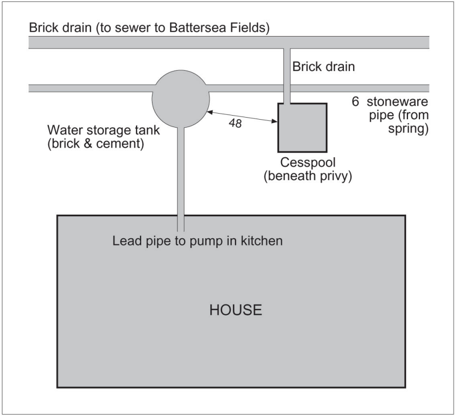
Provided Snow with final evidence that crystalized his theory

  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

Schematic of Cesspools & Water Tanks

17 houses sharing common water source



- 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

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

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

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*:

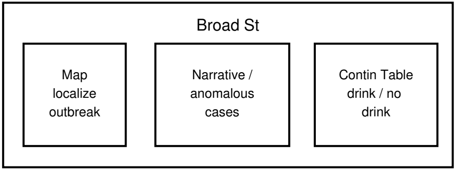
1. Placing the data in an appropriate context for assessing cause and effect
2. Making quantitative comparisons
3. Considering alternative explanations and contrary cases
4. Assessment of possible errors in the numbers reported in graphics

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

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

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

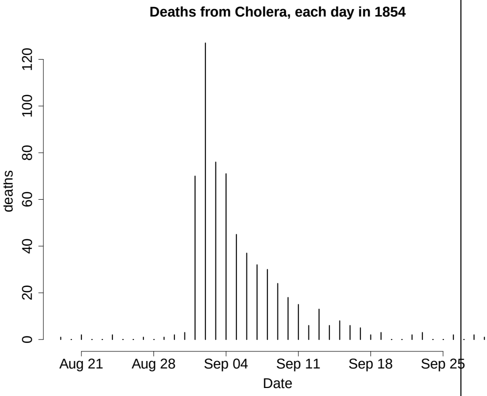
The raw data were a list of deaths by date – Virtually useless,

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)

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

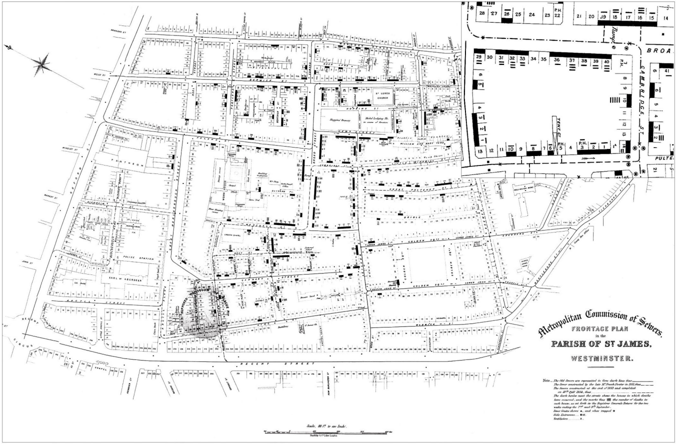
> *he knew ... that that kind of evidence, on its own, would not satisfy a miasmatist. The cluster could just as easily reflect some pocket of poisoned air that had settled over that part of Soho (Johnson p 140)*

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

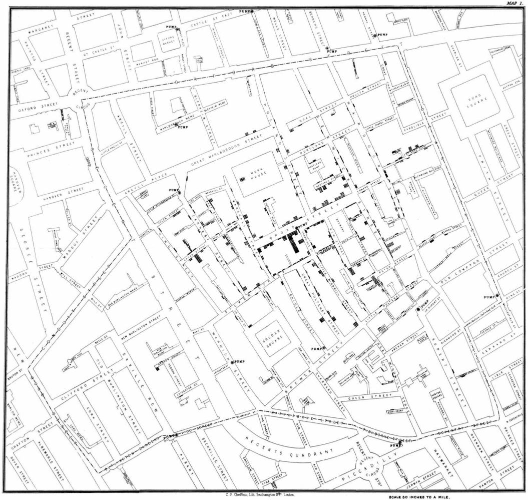
Cooper's Map Obscures: Too Much Detail



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): mappingQuantAnalysis

Tufte 3: Alternative Explanations & Contrary Cases

More Important than Map: Narratives & Anomalous Cases

Testing Competing Theories: "confronting the waterborne and alternative theories with evidence"

1. Those who should have died but escaped
   - Close to pump but did not die
   - Work House & Brewery (few-to-no deaths)
2. 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"
3. 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)

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

Anomalies to Test & Separate Theories

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

(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

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****.*

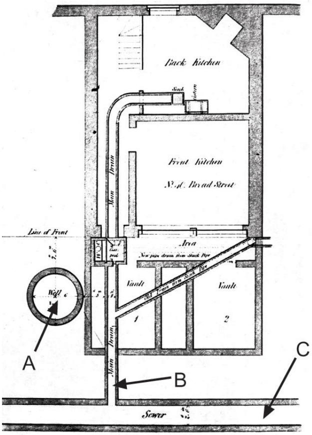
Wonderful example of Miasma auxiliary hypothesis to protect miasma core

- 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

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



Making Quantitative Comparisons

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

1. Compared to other pumps, Broad St stands out
   - All areas densely populated – problem with maps that reflect population
2. 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

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 drinkersdetail

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 χ^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

"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

South London Analysis – 2 Parts

Aggregate, Diff-in-Diffs

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

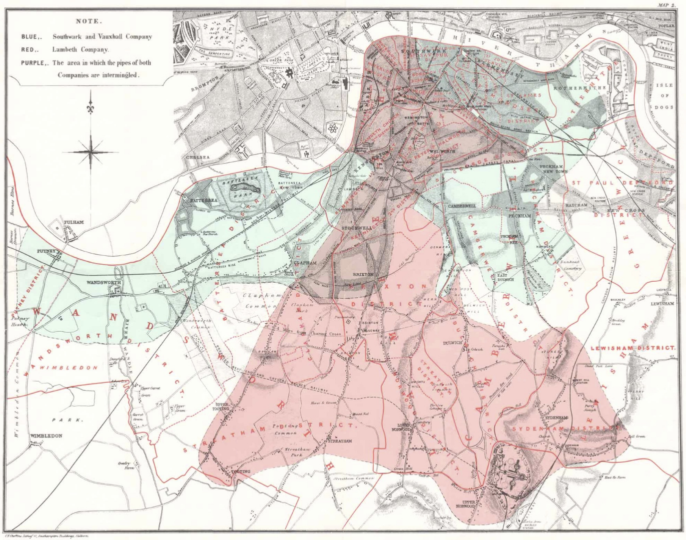
Mixed or quasi-random comparison

- 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."

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	
St. Mary, Newington .	143	92	Lambeth Company, and Southwark and Vauxhall Compy.
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 DiDdetails

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: DiDdetails

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)*

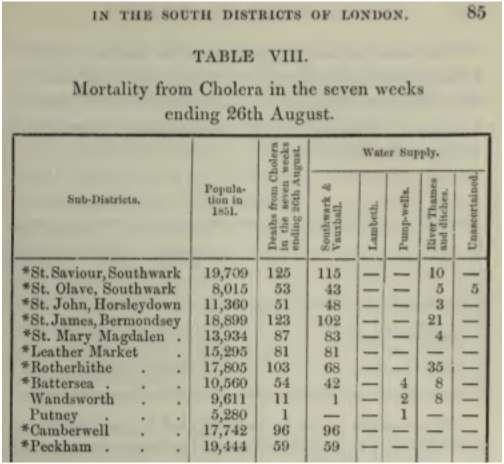
During August Snow visited every house with a death to identify supplier

- 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 "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	—	—	—	—

Snow's Comparison – Direct Control vs Treatment

Using Houses for all 32 sub-districts together

- Includes "first 12" Southwark-only sub-districts (& "last 4"), so not a clean comparison of "next 16" mixed sub-districts
- But – from diff-in-diffs – "first 12" & "next 16" differences small

Houses, Deaths, and Mortality per 10,000 Households, First Seven Weeks of 1854 Cholera Epidemic – Table IX p 86

Water Supplier	Number of houses	Deaths from Cholera	Deaths in each 10,000 houses
Southwark and Vauxhall	40,046	1,263	315.4
Lambeth Company	26,107	98	37.54
Reduction in mortality			–277.9
Naive *t*-ratio			–29.2

Note that this corrects a rounding error in the "Deaths in each 10,000 houses" for Lambeth in Snow's original table

Huge decrease – mortality lower by factor of 8

Naive t-ratio –29.2, but this is wrong. True closer to –11

- Still large, justifies Snow's claim for "the overwhelming influence of water"

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

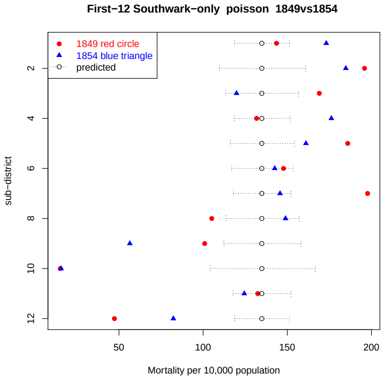
Exploit this variation to assess precision of our -57.0 estimate (-0.511 in logs)

- 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



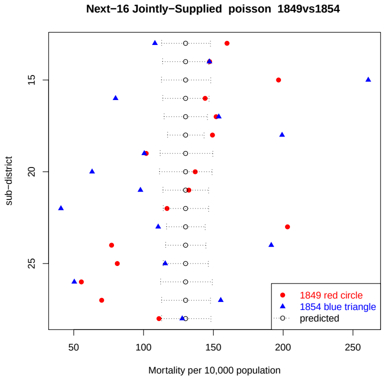


Figure: Mortality per 10,000, Poisson Count Model, Same Rate All Sub-Districts, Predicted (with 95% confidence bands) and Actual 1849 & 1854 (Adjusted for Time and Single Treatment Effect)

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

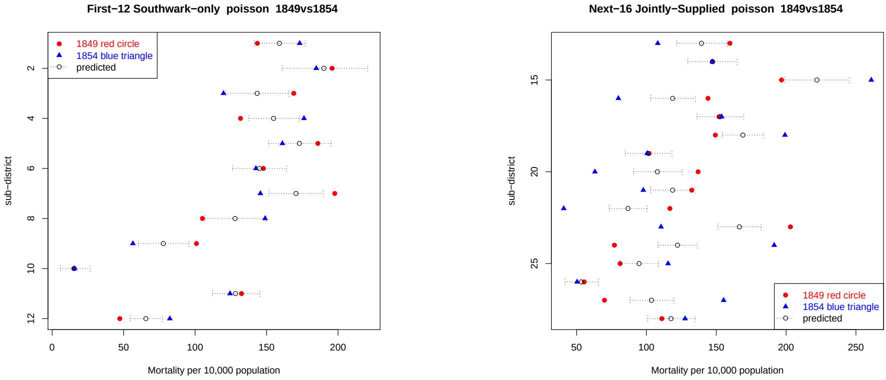
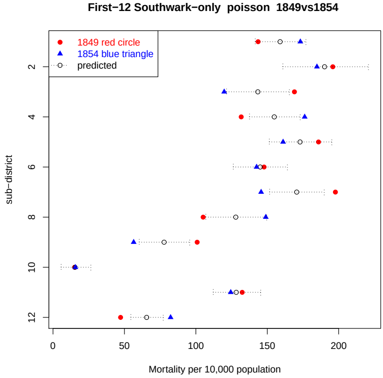


Figure: Mortality per 10,000, Poisson Count Model, Different Rates for Sub-Districts, Predicted (with 95% confidence bands) and Actual 1849 & 1854 (Adjusted for Time and Single Treatment Effect)

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)

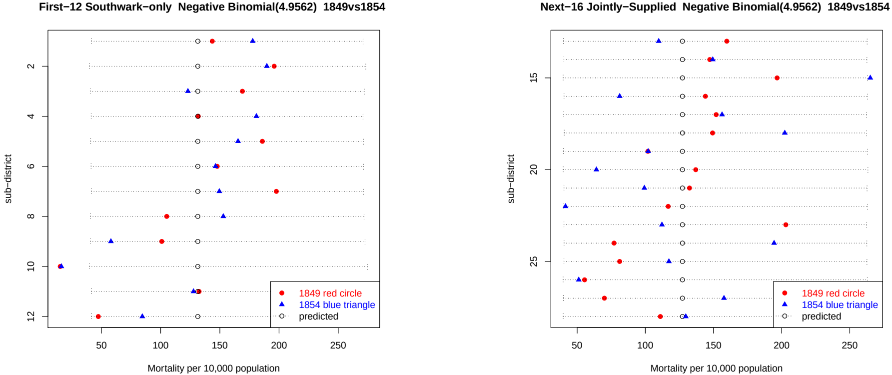
Variation *within* sub-districts harder – How can mortality *not* be Poisson?

- 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 Poisson	2 Poisson, sub-district Fixed Effects	3 Negative Binomial	4 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 [?] 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 Poisson	2 Poisson, sub-district Fixed Effects	3 Negative Binomial	4 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 [?] 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 Poisson	2 Poisson, District Fixed Effects	3 Negative Binomial	4 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

1. 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
2. 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

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

Conclusion: Theory, Data, Hypothesis Testing

Data or Evidence Blocks

Albion Terr
17 houses
single outbreak

Broad St
~10 sq blocks
2wks, 700 deaths

South London
summer/fall 1854
~400k subjects mixed
treated & untreated

Theory & Hypotheses

water & small intestine	miasma (airborne)	elevation, class, ...

Hypothesis or Testing Blocks

Albion Terr
- Narrative

Broad St
- Map
- Cases
- Contin

South London
- Diff-in-Diffs
- Mixing
- No sub-district pop
- With sub-district pop

Theory, Data, Hypothesis Testing

Table: 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

Still Much to Learn From John Snow

1. **Rollicking Good Tale** – full of heroism, death, and statistics
2. **Causal Inference**: template for how to marshal evidence in support of a causal explanation
3. **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

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
- mapQuantReturn

Major Innovation by Snow – Walking Neighborhood

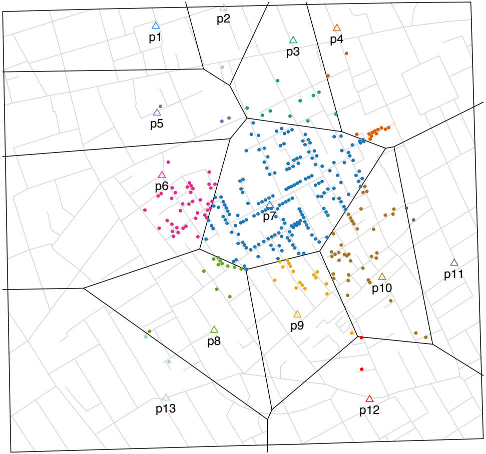
Detail showing the outline

- 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

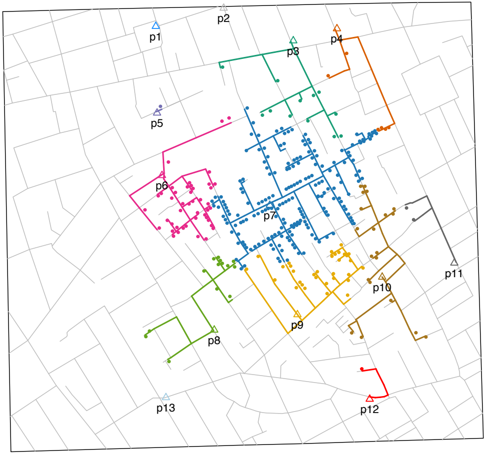
pump.id	Count	Percent	Expected	Pearson
1	0	0	19.5	19.5
2	1	0.31	6.2	4.4
3	10	3.12	14.0	1.1
4	13	4.05	30.4	10.0
5	3	0.93	26.5	20.8
6	39	12.15	39.9	0.0
7	182	56.7	27.2	881.0
8	12	3.74	22.1	4.6
9	17	5.3	15.5	0.1
10	38	11.84	19.0	19.0
11	2	0.62	24.6	20.8
12	2	0.62	29.7	25.8
13	2	0.62	46.4	42.5
Sum	321		Sum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

mapQuantReturn

Walking Neighborhoods

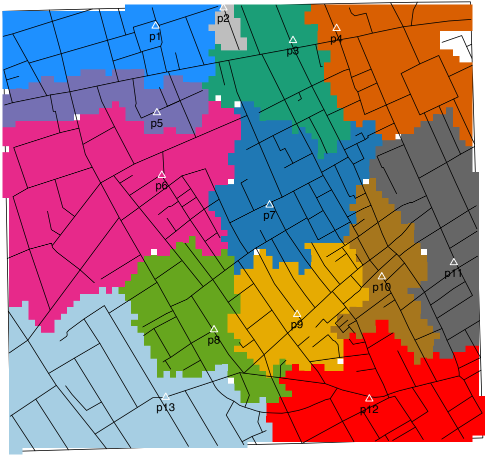
Even more fun – equal walking distance



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

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

mapQuantReturn

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

mapQuantReturn

Counts for Drinkers vs Non-Drinkers

Table: Count of Residents of Broad Street Categorized by Drinking and Illness

	Not ill	Ill, recovered	Ill, died	**TOTAL**
Did not drink from pump	279	7	13	**299**
Drank from pump	57	43	45	**145**
Probably drank from pump	–	2	10	**12**
Uncertain or Unknown	13	6	22	**41**
TOTAL	**349**	**58**	**90**	**497**

Counts of Broad Street residents collected by the Reverend Whitehead and reported in [?] p 128 ff. See text and footnotes for details on source for individual cells

drinkersreturn

Contingency Table Analysis for Drinkers vs Non-Drinkers

Table: Contingency Table Analysis for Drinking versus Illness

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**

Using cases for which drinking status (drinking from the pump versus not) could be determined. "Expected Counts" are expected if drinking and illness were independent (conditional on row and column sums). The Pearson chi-squared statistic is 154.7. Both the Pearson chi-squared and the Fisher exact test strongly reject the hypothesis that drinking and illness are independent (p-value far less than 0.0001). The Phi coefficient (a measure of association, the same as Cramér's V in this case) is +0.59, showing strong positive association between illness and drinking from the pump.

drinkersreturn

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: δ_{54} captures any difference between 1849 & 1854
- Region Effect: γ_J captures any difference between "First 12" versus "Next 16"
- Treatment Effect: β captures the effect of clean water
- **We care about the treatment effect β**
- Worry about region (γ_J) and time (δ_{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	μ	$\mu + \delta_{54}$	δ_{54}
"Next 16" Jointly Supplied	$\mu + \gamma_J$	$\mu+\gamma_J+\delta_{54}+\beta$	$\delta_{54} + \beta$
Diff Joint less Southwark	γ_J	$\gamma_J + \beta$	β

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	$I_{r=J} = 0$ $I_{t=54} = 0$	$I_{r=J} = 0$ $I_{t=54} = 1$	
"Next 16" Jointly Supplied	$I_{r=J} = 1$ $I_{t=54} = 0$	$I_{r=J} = 1$ $I_{t=54} = 1$	
Diff Joint less Southwark			

Get same table:

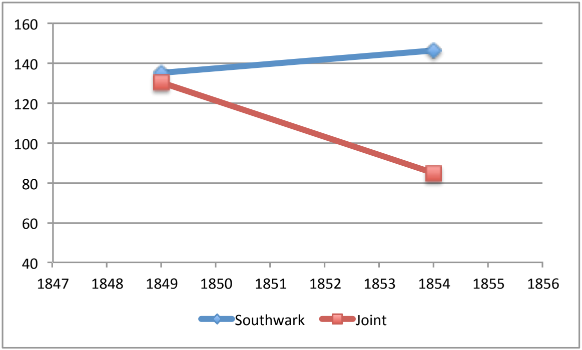
Region (Sub-Districts) Supplied by	1849 Deaths per 10,000	1854 Deaths per 10,000	Diff 1854 less 1849
"First 12" Southwark Only	μ	$\mu + \delta_{54}$	δ_{54}
"Next 16" Jointly Supplied	$\mu + \gamma_J$	$\mu + \gamma_J + \delta_{54} + \beta$	$\delta_{54} + \beta$
Diff Joint less Southwark	γ_J	$\gamma_J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

DiDreturn

Graphing the Treatment Effect

Comparing the "Southwark Only" vs "Joint" regions:

- They look very similar in 1849 – γ_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)) DiDreturn

Mortality Rates from Snow Table XII

	Sub-Districts	1849 per 10,000	1854 per 10,000	Water Supplier
1	St. Saviour, Southwark	144	188	SouthwarkVauxhall
2	St. Olave, Southwark	196	201	SouthwarkVauxhall
3	St. John, Horsleydown	169	130	SouthwarkVauxhall
4	St. James, Bermondsey	132	192	SouthwarkVauxhall
5	St. Mary Magdalen	186	175	SouthwarkVauxhall
6	Leather Market	148	155	SouthwarkVauxhall
7	Rotherhithe	198	158	SouthwarkVauxhall
8	Battersea	92	56	SouthwarkVauxhall
9	Wandsworth	115	178	SouthwarkVauxhall
10	Putney	15	17	SouthwarkVauxhall
11	Camberwell	132	135	SouthwarkVauxhall
12	Peckham	47	89	SouthwarkVauxhall
13	Christchurch, Southwark	160	71	Southwark&Lambeth
14	Kent Road	147	96	Southwark&Lambeth
15	Borough Road	197	170	Southwark&Lambeth
16	London Road	144	52	Southwark&Lambeth
17	Trinity, Newington	152	100	Southwark&Lambeth

Mortality Rates from Snow Table XII

	Sub-Districts	1849 per 10,000	1854 per 10,000	Water Supplier
18	St. Peter, Walworth	149	130	Southwark&Lambeth
19	St. Mary, Newington	102	66	Southwark&Lambeth
20	Waterloo Road (1st)	137	41	Southwark&Lambeth
21	Waterloo Road (2nd)	132	64	Southwark&Lambeth
22	Lambeth Church (1st)	117	27	Southwark&Lambeth
23	Lambeth Church (2nd)	203	72	Southwark&Lambeth
24	Kennington (1st)	77	125	Southwark&Lambeth
25	Kennington (2nd)	81	75	Southwark&Lambeth
26	Brixton	55	33	Southwark&Lambeth
27	Clapham	70	101	Southwark&Lambeth
28	St. George, Camberwell	111	83	Southwark&Lambeth
29	Norwood	5	25	Lambeth
30	Streatham	171	17	Lambeth
31	Dulwich	6	0	Lambeth
32	Sydenham	11	27	Lambeth
	First 12 sub-districts	135	147	first12
	Next 16 sub-districts	130	85	next16
	Last 4 sub-districts	85	19	last4
	TOTAL	130	104	