THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

The Difference-in-Difference Design: Snow’s South London Experiment

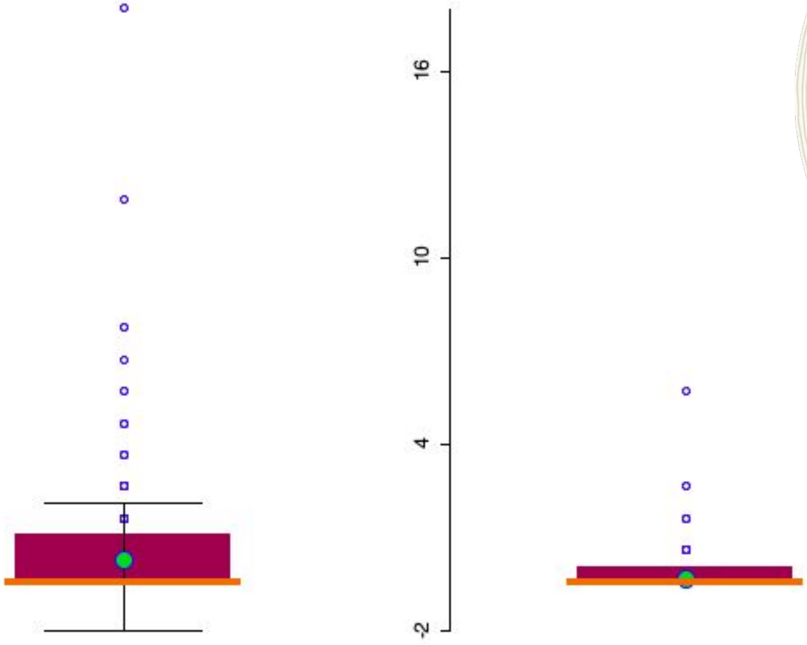
Julia Koschinsky, PhD
Executive Director and Senior Research Associate, Center for Spatial Data Science

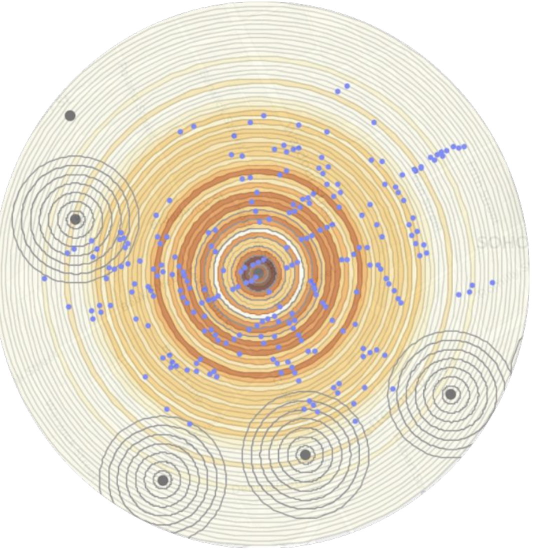
Cholera outbreaks in London in 1848–49 and 1853–54

Snow’s questions:

- Is there a causal relationship between ingesting contaminated water and getting sick from cholera?
- Is drinking contaminated water one of the causes or the dominant transmitter of cholera?

Comparing Two Areas in Neighborhood

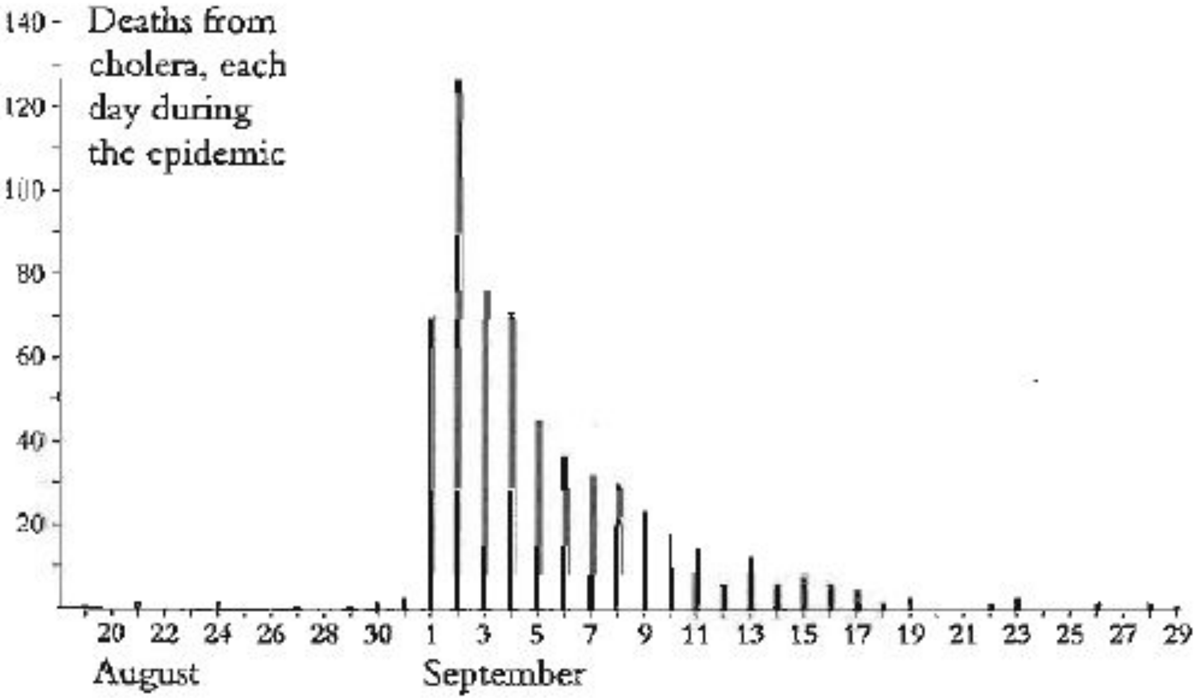




Deaths closer to Broad St pump

Deaths further away from Broad St pump

Ignoring time effects:
Just before pump was closed -
deaths near Broad St pump were already decreasing:
people were fleeing area and timing of epidemic curve

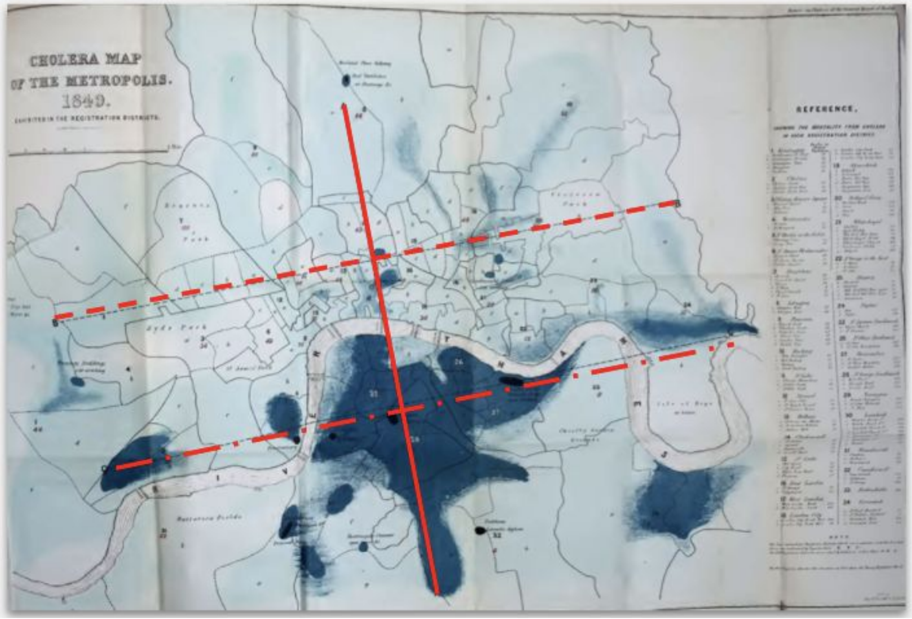


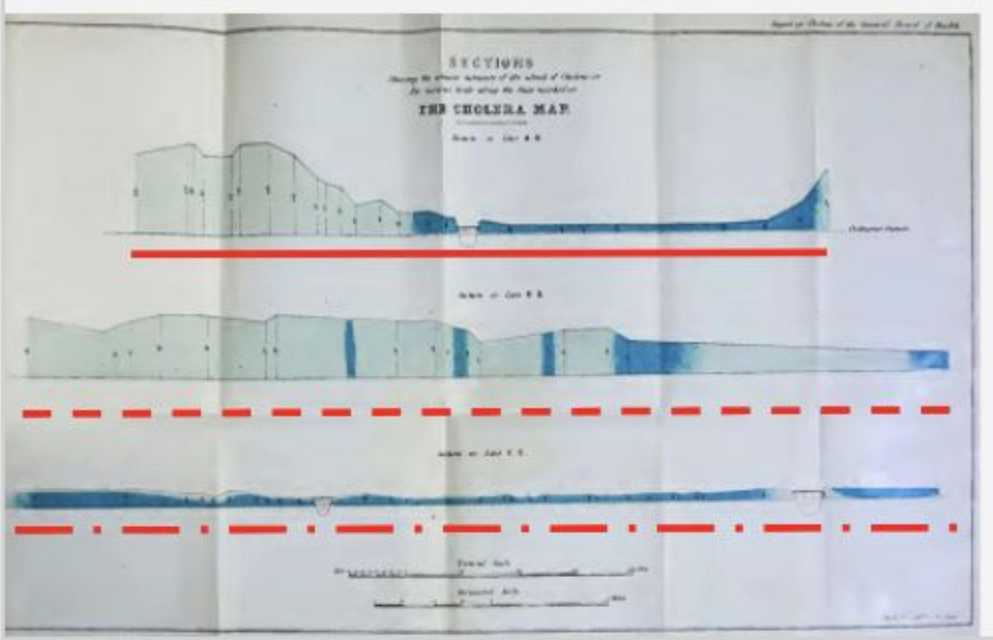
Comparing Two Areas in City

cholera in high and low elevation areas

Inspector for the General Board of Health Grainger (1849): cholera higher in low-elevation areas due to gases from sewage-contaminated parts of Thames settling there

Source: General Board of Health 1851

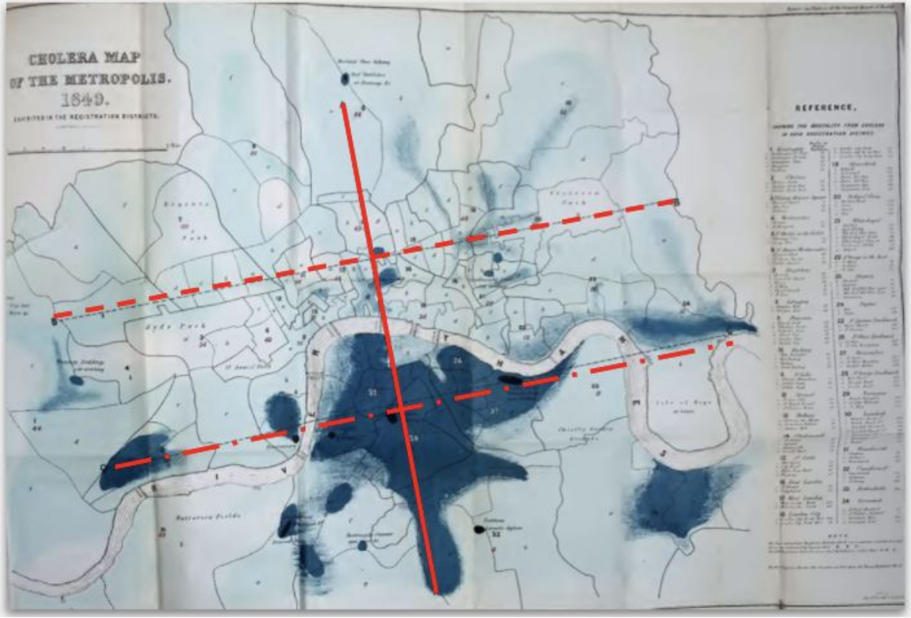


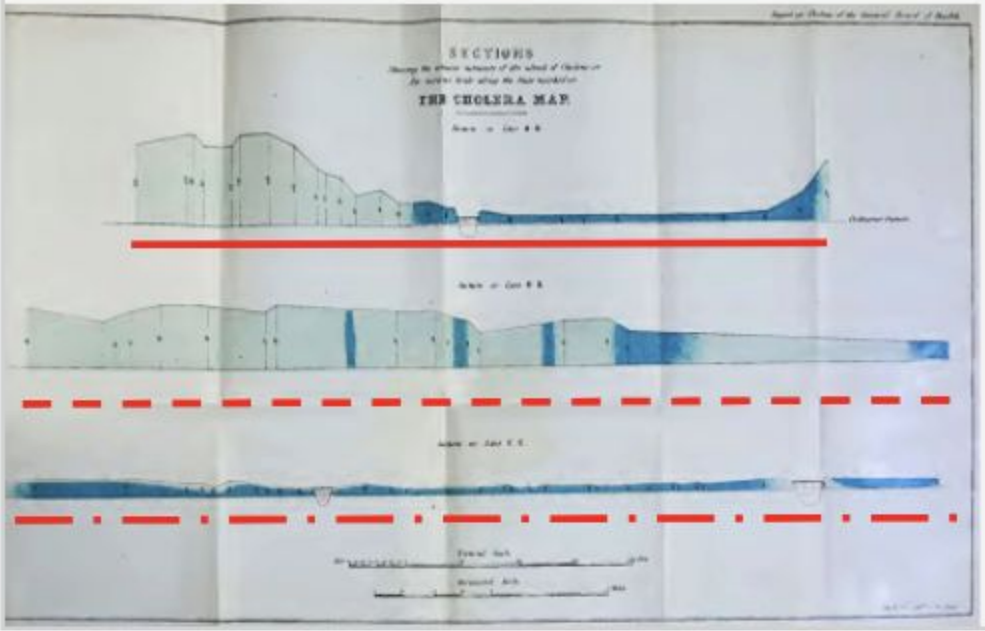


Problem with Comparing Two Areas

Correlation turned out to be spurious (correlation without causation)
We can't just compare cholera rates in low and high elevation because those areas might be different for other reasons.

Source: General Board of Health 1851





Cholera outbreaks in London in 1848–49 and 1853–54

Snow’s questions:

- Is there a causal relationship between ingesting contaminated water and getting sick from cholera?
- Is drinking contaminated water one of the causes or the dominant transmitter of cholera?

Snow realized the opportunity of a natural experiment: In 1852, the Lambeth water company moved its pump station out of the polluted part of the Thames. The Southwark water company staid until 1855. Water supplier assignment was basically random. Snow asked:

- Did cholera death rates vary for subdistricts getting their water from Lambeth or Southwark before and after 1852?

Difference-in-Difference (DID) Design

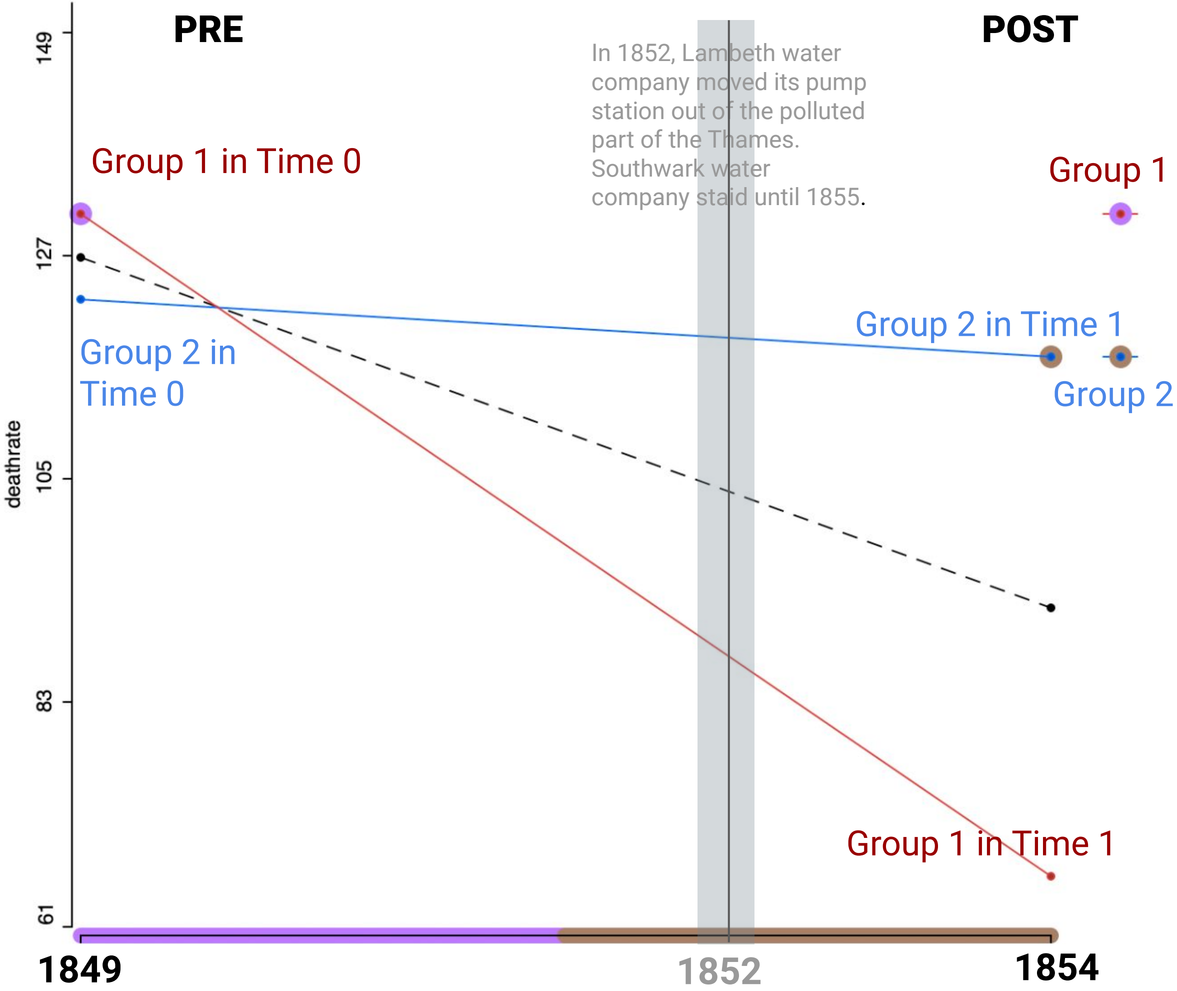
- Goal is to estimate potential causal relationships using observational data

- DID is useful when a treatment changes at different times for different units.

→ **treatment**: Lambeth moved pump in 1852, Southwark in 1855

→ **comparing two groups** (here subdistricts with different shares of contaminated water)

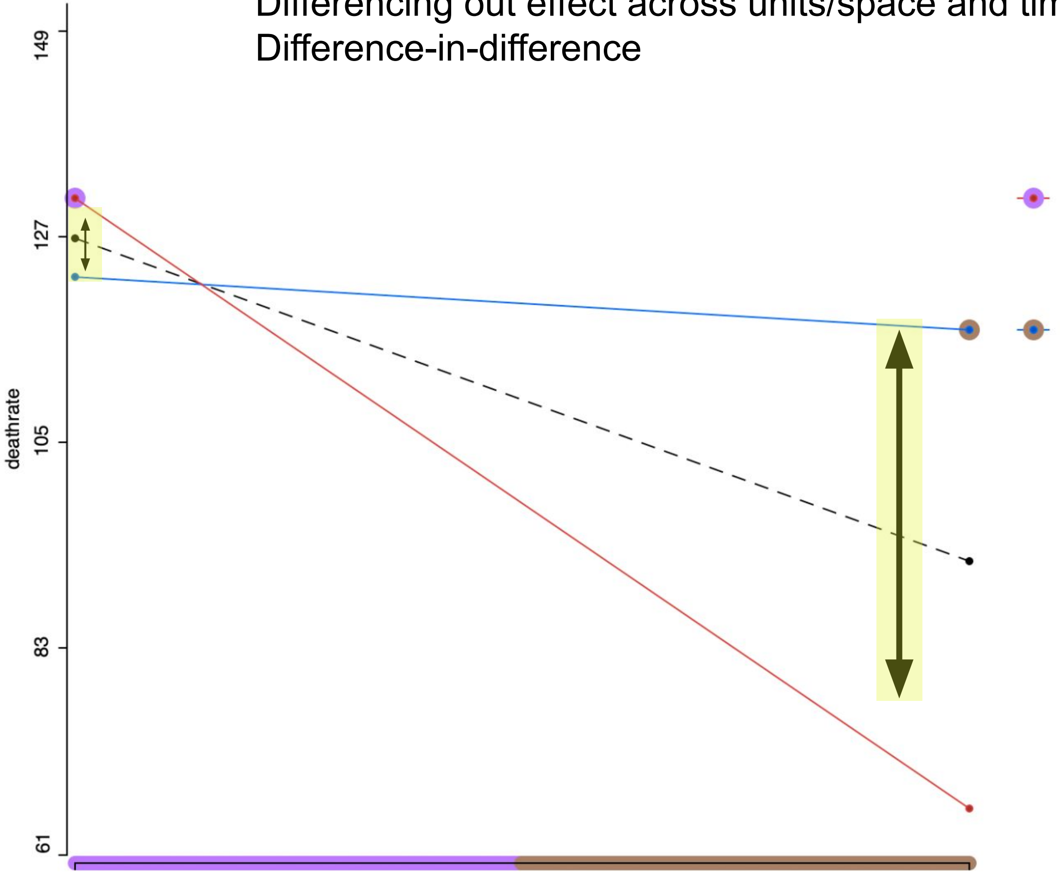
→ **comparing two times**: 1849 vs 1854 (pre and post treatment in 1852)



In 1852, Lambeth water company moved its pump station out of the polluted part of the Thames. Southwark water company staid until 1855.

Are these differences significant?

Differencing out effect across units/space and time:
Difference-in-difference

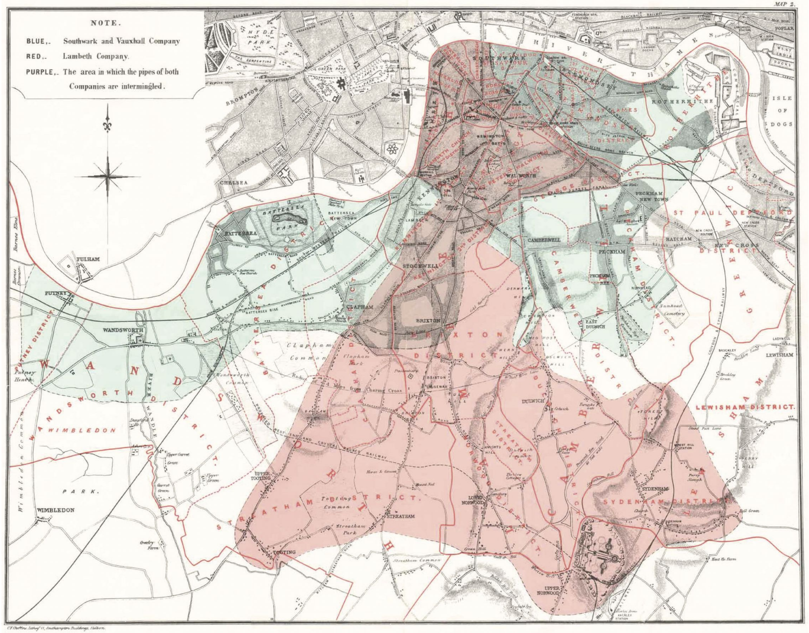


Map of Water Supply Areas

Southwark & Vauxhall,

Lambeth, and

subdistricts supplied by both



Source: Snow 1855

Mixed subdistricts: no difference in class, elevation, population density & air quality

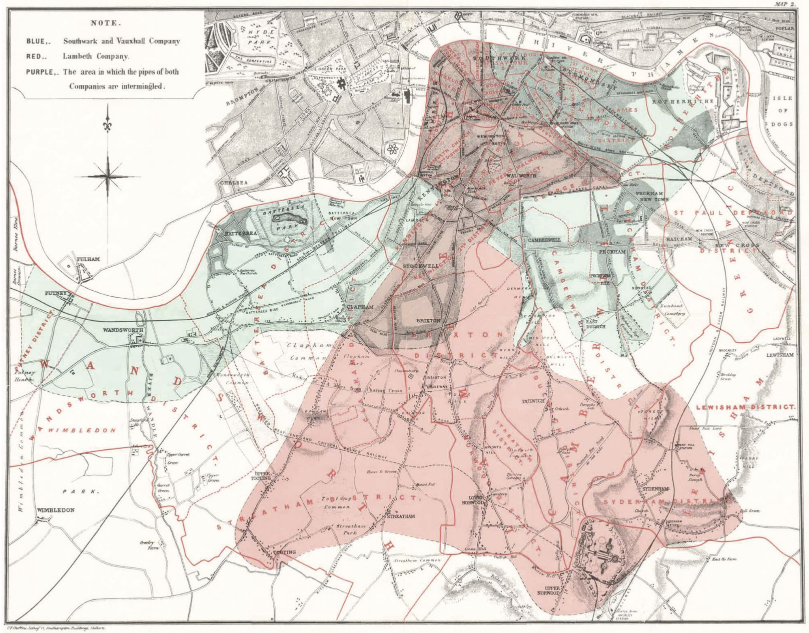
But difference in supplier of contaminated vs clean water

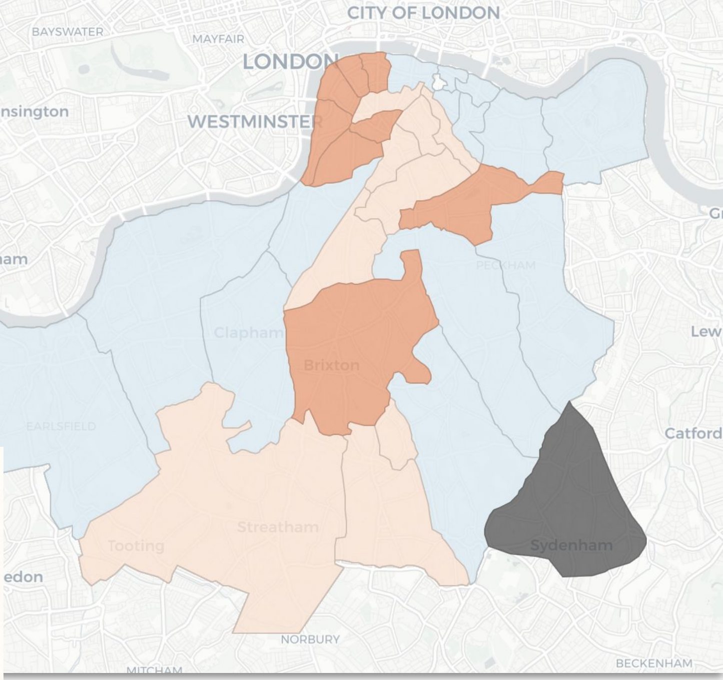
Map of Water Supply Areas

Southwark & Vauxhall,

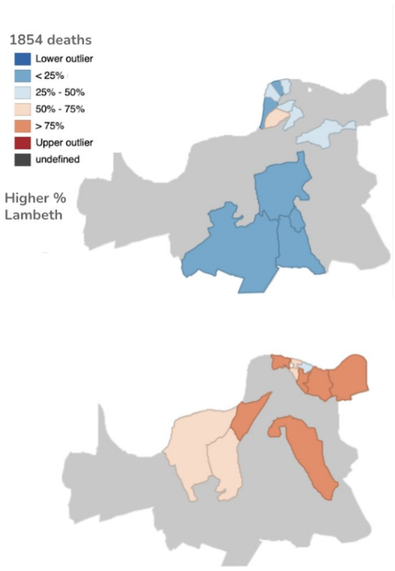
Lambeth, and

subdistricts supplied by both

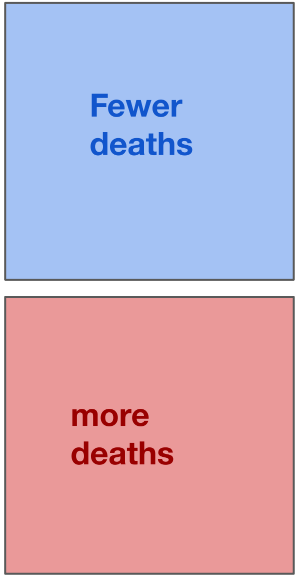




Percent of pop served by Lambeth

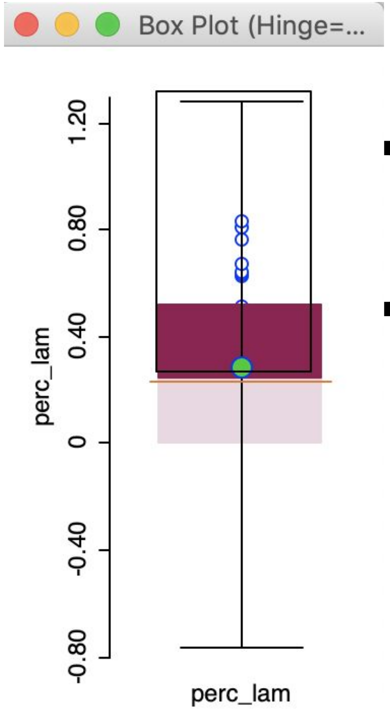


Expected Deaths:
Snow’s Theory



Fewer deaths

more deaths

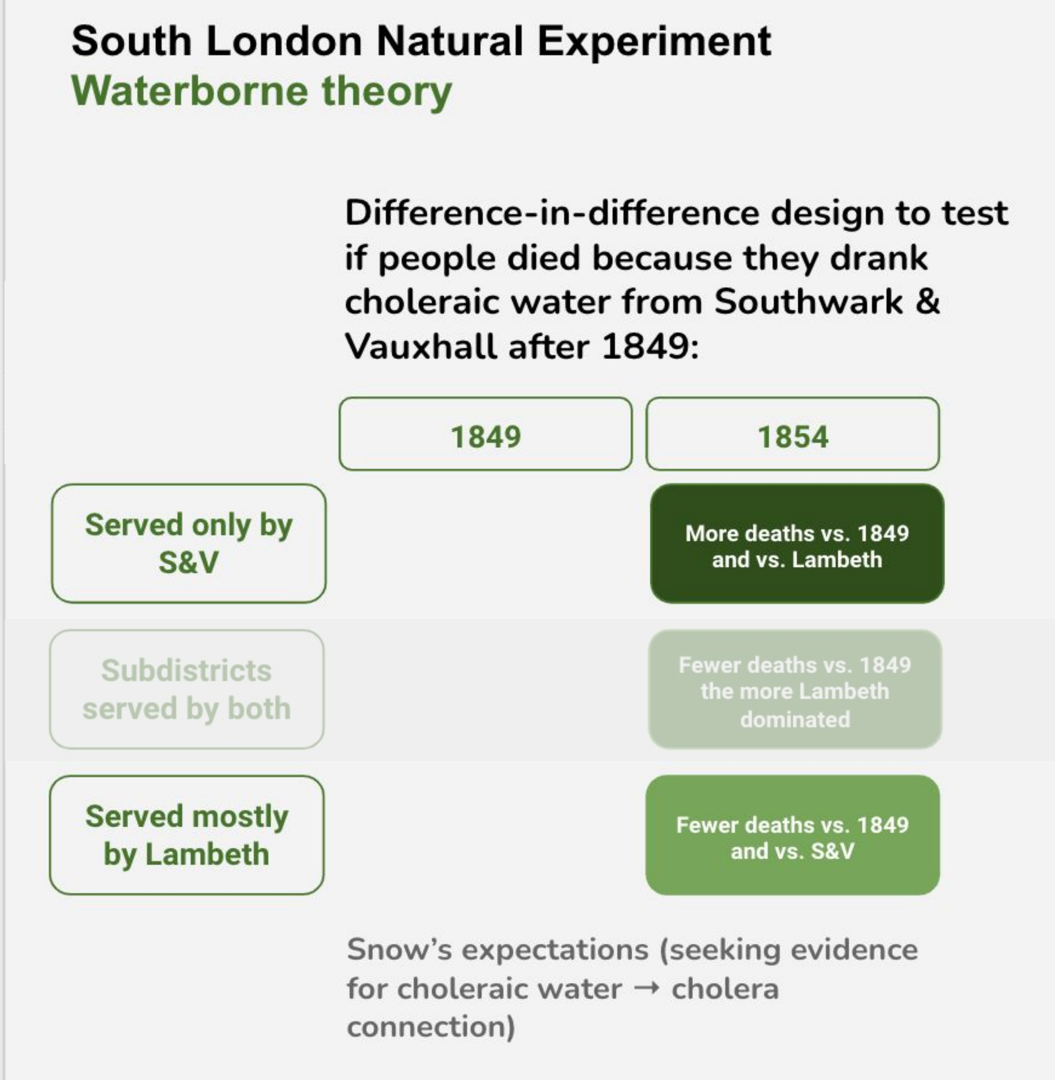


Subdistricts with an above-average share of Lambeth customers

Subdistricts with a below-average share of Lambeth customers

Adding a Time Comparison:

Pre and post 1852, when Lambeth stopped sourcing contaminated water but Southwark did not.



South London Natural Experiment

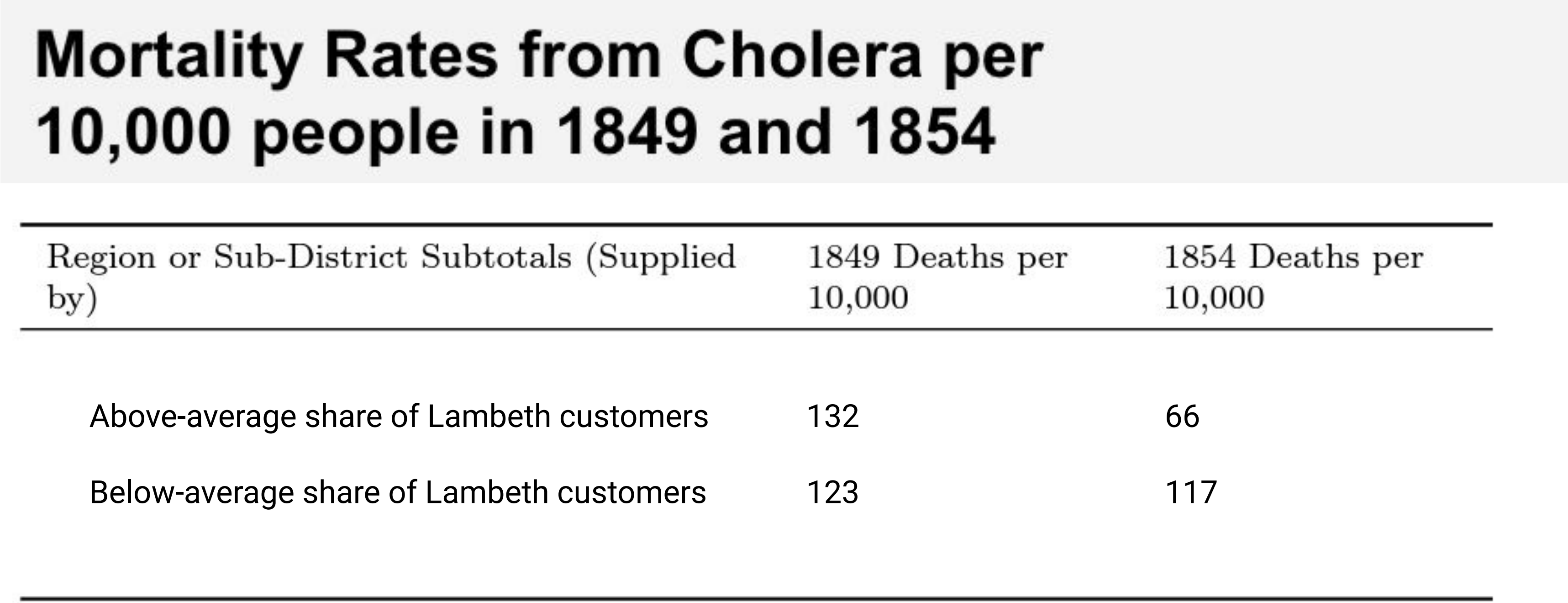
Waterborne theory

Difference-in-difference design to test if people died because they drank choleraic water from Southwark & Vauxhall after 1849:

	1849	1854
Served only by S&V		More deaths vs. 1849 and vs. Lambeth
Subdistricts served by both		Fewer deaths vs. 1849 the more Lambeth dominated
Served mostly by Lambeth		Fewer deaths vs. 1849 and vs. S&V

Snow's expectations (seeking evidence for choleraic water → cholera connection)

Did Cholera Death Rates Vary by Water Supplier? The South London Natural Experiment



Mortality Rates from Cholera per 10,000 people in 1849 and 1854

Region or Sub-District Subtotals (Supplied by)	1849 Deaths per 10,000	1854 Deaths per 10,000
Above-average share of Lambeth customers	132	66
Below-average share of Lambeth customers	123	117

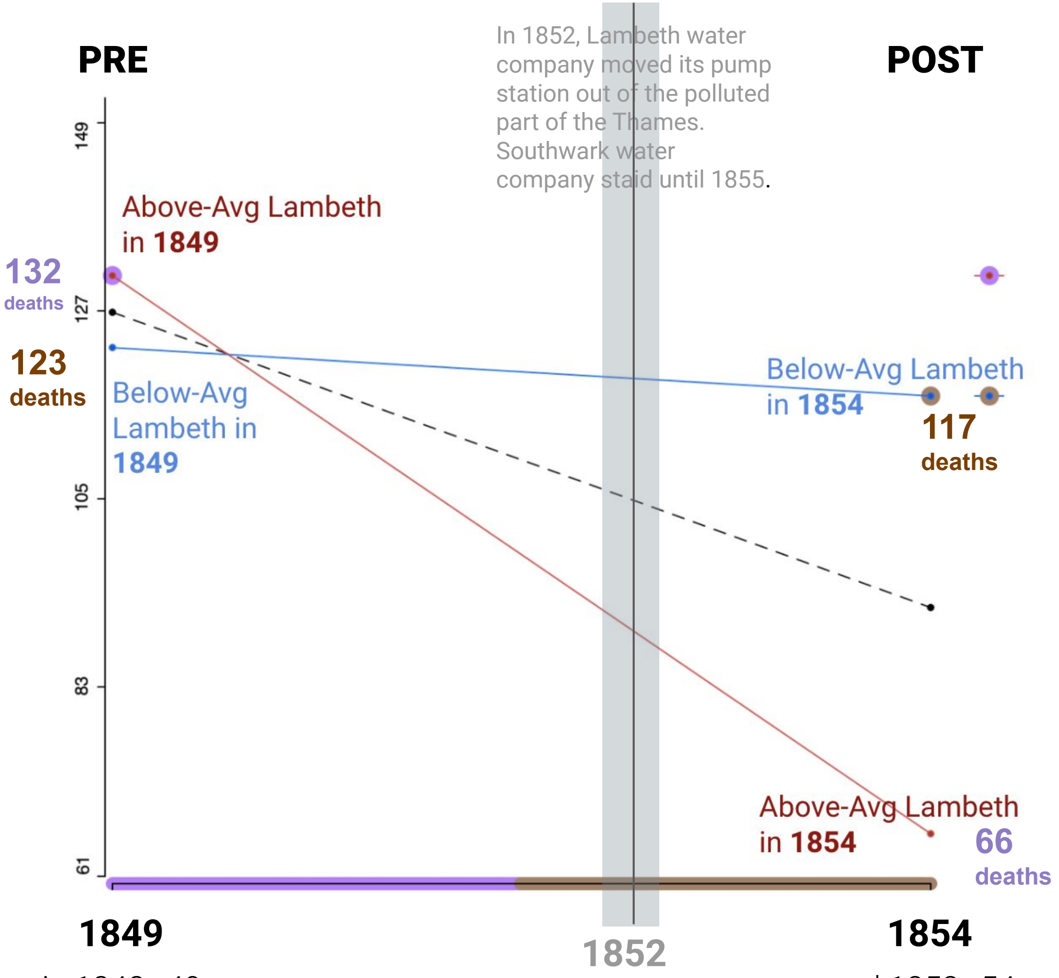
Snow compared differences across two dimensions:

Time

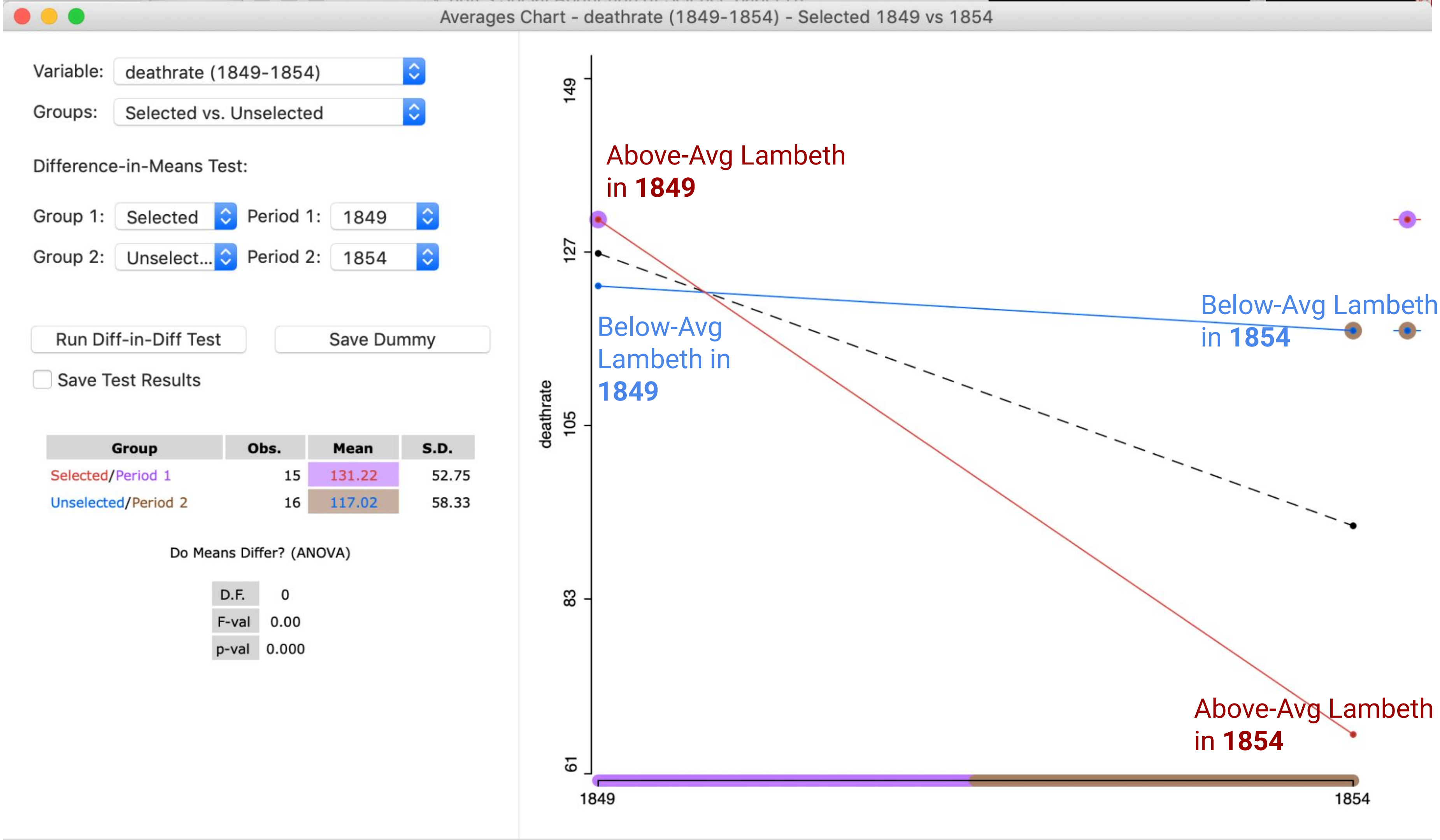
Deaths in 1854 vs 1849

Space

Subdistricts with larger vs smaller share of customers of Lambeth, which stopped pumping polluted water by 1854



cholera epidemics in 1848–49 and 1853–54



Averages Chart - deathrate (1849-1854) - Selected 1849 vs 1854

Variable: deathrate (1849-1854)

Groups: Selected vs. Unselected

Difference-in-Means Test:

Group 1: Selected Period 1: 1849

Group 2: Unselect... Period 2: 1854

□ Save Test Results

Group	Obs.	Mean	S.D.
Selected/Period 1	15	131.22	52.75
Unselected/Period 2	16	117.02	58.33

Do Means Differ? (ANOVA)

D.F.	0
F-val	0.00
p-val	0.000

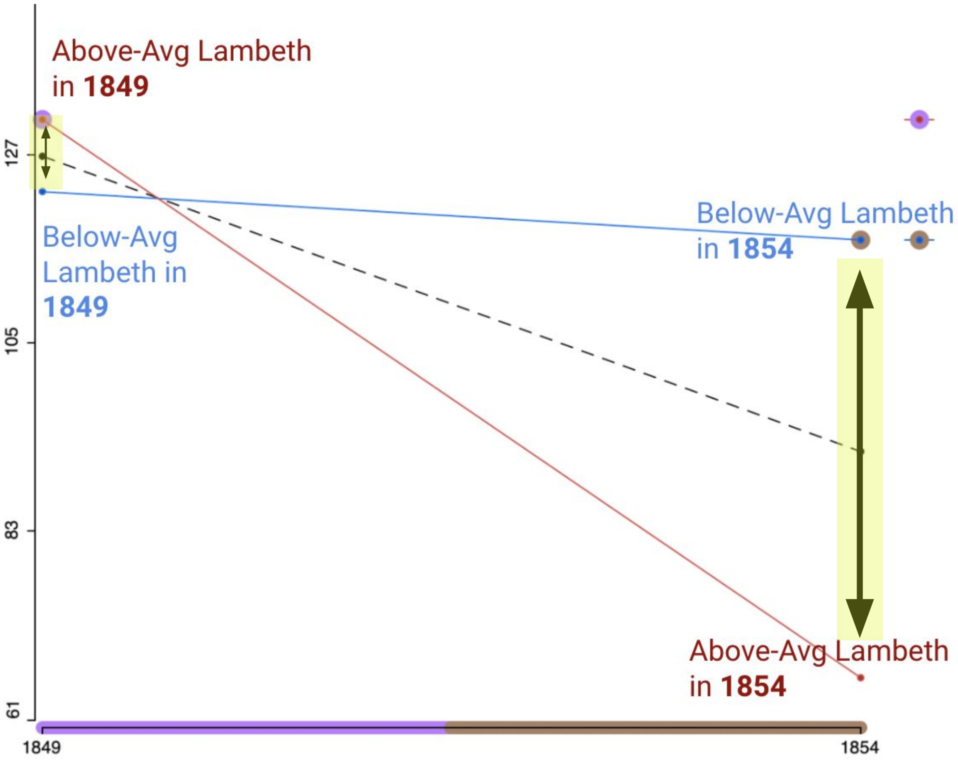
Above-Avg Lambeth in **1849**

Below-Avg Lambeth in **1849**

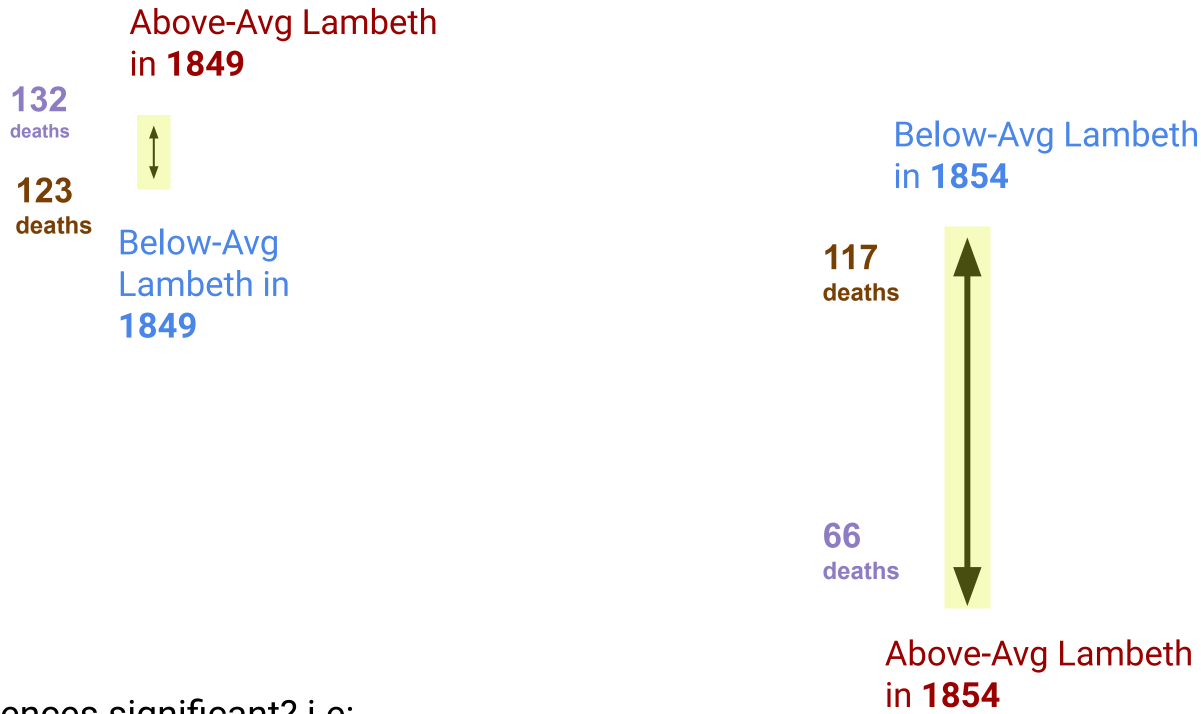
Below-Avg Lambeth in **1854**

Above-Avg Lambeth in **1854**

In **1849**, the death rate in areas with a larger share of Lambeth customers was **higher** than that in areas with lower shares of Lambeth customers.



In **1854**, the death rate in areas with a larger share of Lambeth customers was **lower** than that in areas with higher shares of Lambeth customers.



Above-Avg Lambeth
in **1849**

132
deaths

123
deaths

Below-Avg
Lambeth in
1849

Below-Avg Lambeth
in **1854**

117
deaths

66
deaths

Above-Avg Lambeth
in **1854**

Are these differences significant? i.e:

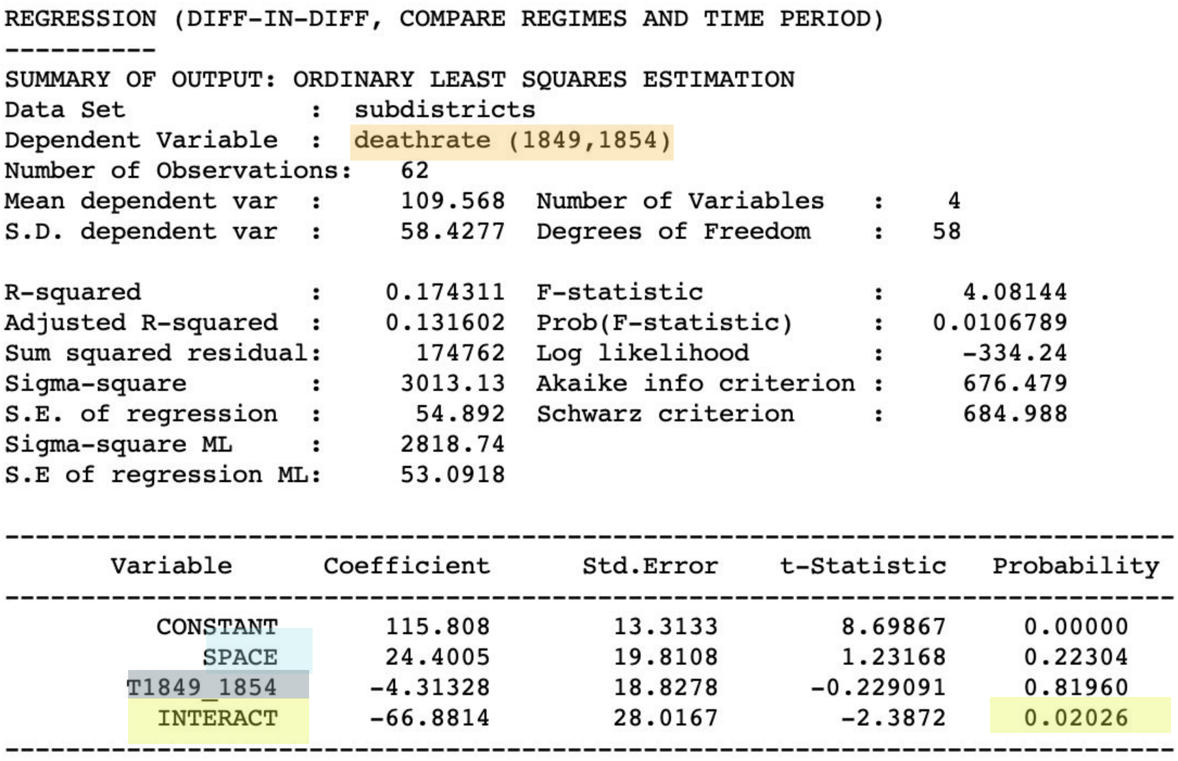
Are the lower death rates in Lambeth areas in 1854 statistically significant when we control for (difference out) the rates in non-Lambeth areas and rates in 1849?

We're differencing out constant differences between the two water supply areas. We're also differencing out any over-time trend that was the same for the two areas.

DID Regression

Deathrate = space + T1849_1854 + interact

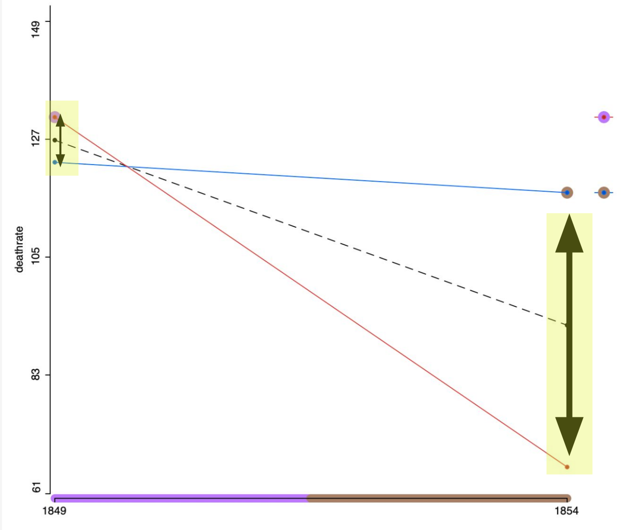
deaths/100,000 = above avg share of Lambeth = 1 + 1849 =1 + Lambeth: yes/no * 1949 vs 1854



```
REGRESSION (DIFF-IN-DIFF, COMPARE REGIMES AND TIME PERIOD)
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data Set             :  subdistricts
Dependent Variable   :  deathrate (1849,1854)
Number of Observations:   62
Mean dependent var   :     109.568  Number of Variables    :    4
S.D. dependent var   :     58.4277  Degrees of Freedom     :   58

R-squared            :    0.174311  F-statistic            :     4.08144
Adjusted R-squared   :    0.131602  Prob(F-statistic)      :   0.0106789
Sum squared residual:      174762  Log likelihood         :     -334.24
Sigma-square         :     3013.13  Akaike info criterion :     676.479
S.E. of regression   :      54.892  Schwarz criterion      :     684.988
Sigma-square ML      :     2818.74
S.E of regression ML:     53.0918

-----------------------------------------------------------------------------
       Variable      Coefficient       Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT       115.808         13.3133        8.69867       0.00000
             SPACE       24.4005         19.8108        1.23168       0.22304
        T1849_1854      -4.31328         18.8278      -0.229091       0.81960
          INTERACT      -66.8814         28.0167        -2.3872       0.02026
-----------------------------------------------------------------------------
```



Implications for Explaining How Cholera Was Transmitted

Stronger connection between choleraic water and contracting cholera than Broad St pump:
In the Broad Street case, living near the pump did not necessarily mean people drank its water. But in this experiment, it was highly likely that people who purchased choleraic water from Southwark & Vauxhall also drank it.

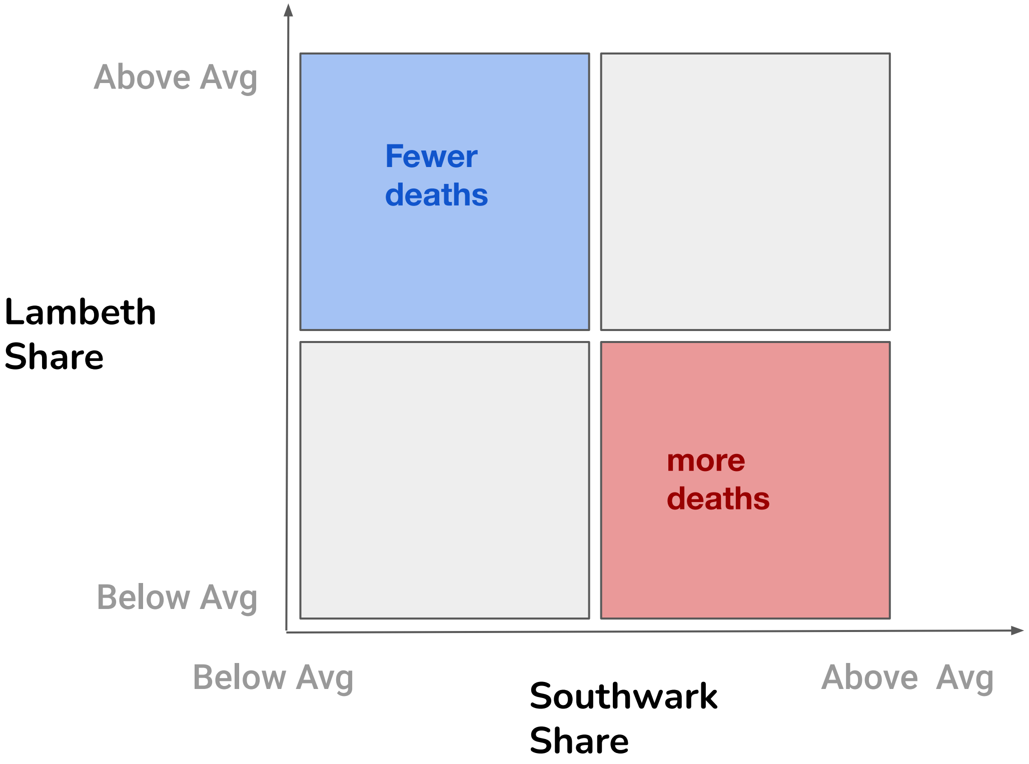
Low Elevation Correlation Spurious: Similar gas exposure in subdistricts with similar low elevations could not explain differences in cholera death rates in these areas – drinking sewage-contaminated water vs clean water better explanation.

Airborne theorists still generally held on to their beliefs despite the fact that Snow's findings strengthened the notion that impure water was the major predisposing cause of cholera.

Reference / Recommended Reading

de Mesquita, E. B., & Fowler, A. (2021). Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis. Princeton University Press. (Chapter 13: Difference-in-Difference Designs)

Expected Deaths According to Snow's Theory



Above Avg

Fewer deaths

Lambeth Share

more deaths

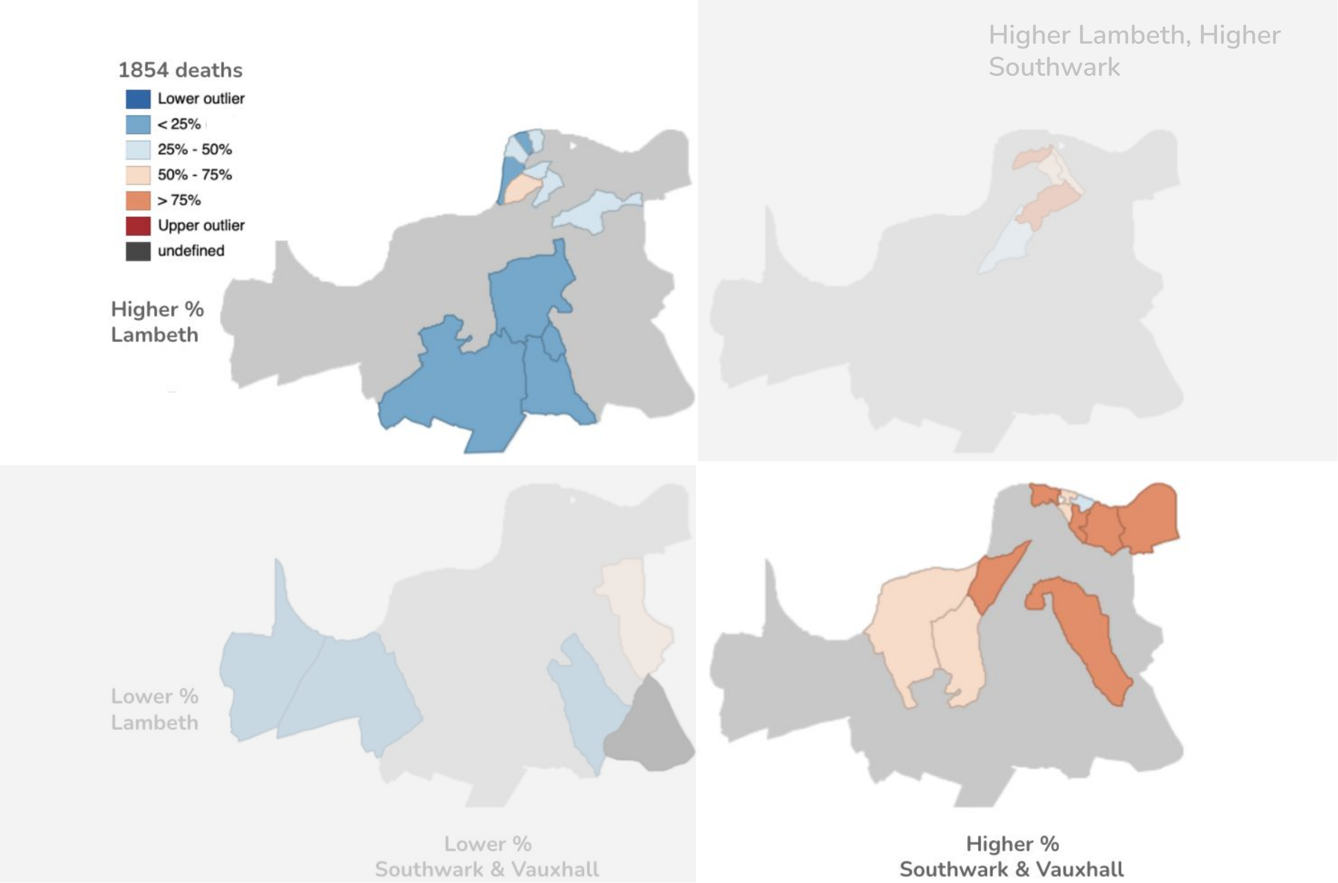
Below Avg

Below Avg

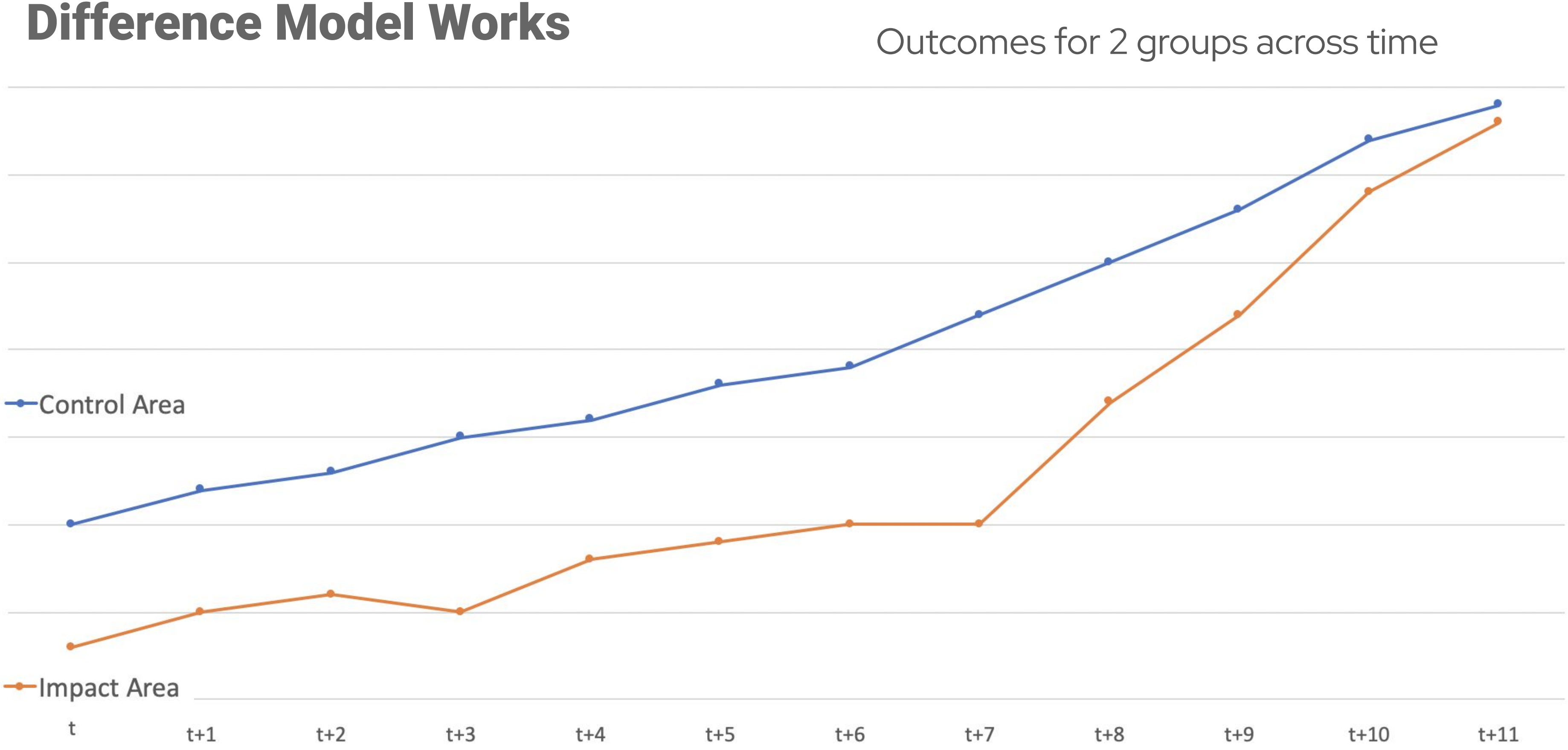
Southwark Share

Above Avg

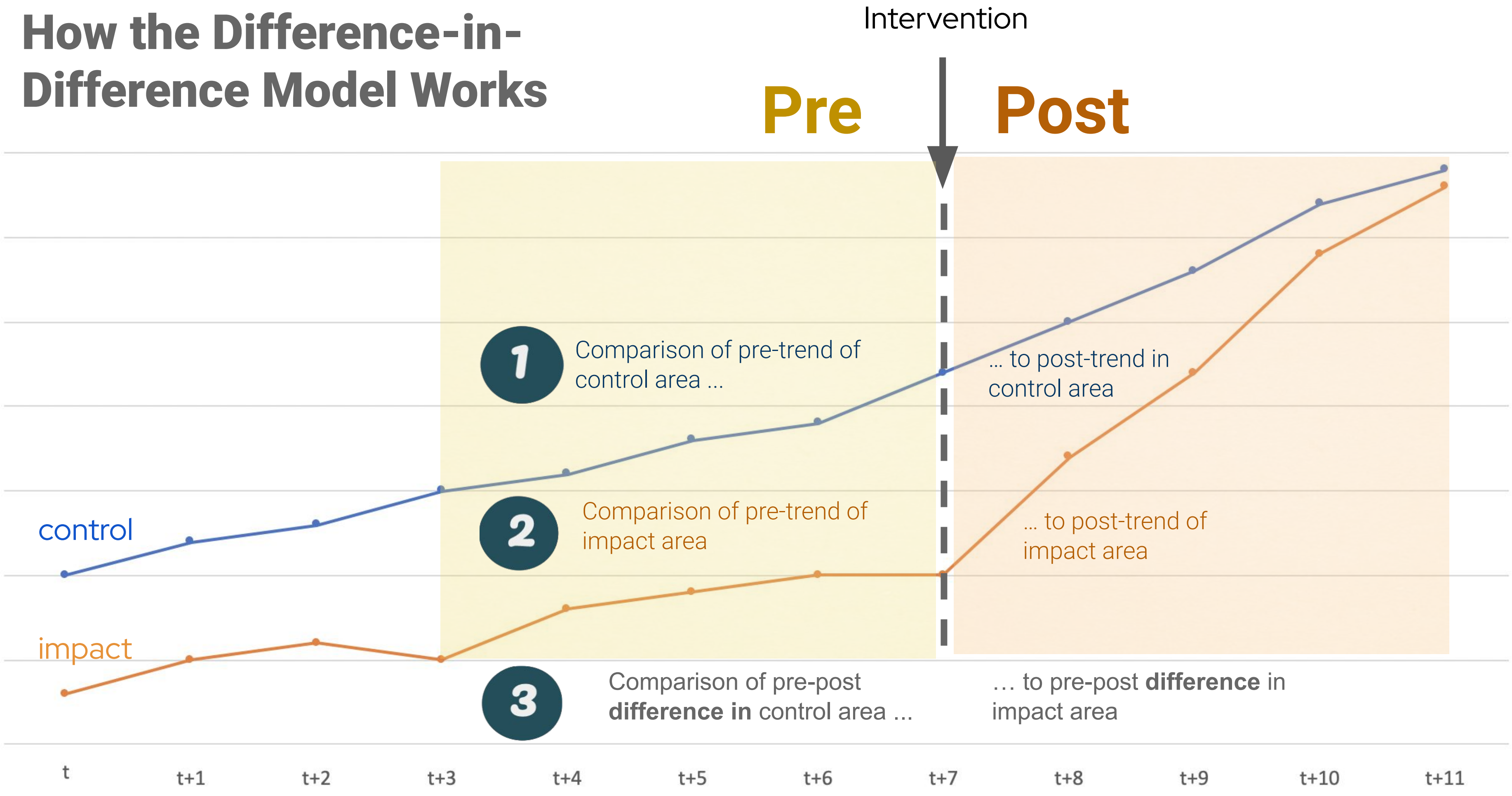
Creating Comparison Areas:
Subdistricts Above and Below Average Shares of Lambeth Customers



How the Difference-in-Difference Model Works



How the Difference-in-Difference Model Works



1 Comparison of pre-trend of control area ...

... to post-trend in control area

2 Comparison of pre-trend of impact area

... to post-trend of impact area

3 Comparison of pre-post **difference in** control area ...

… to pre-post **difference** in impact area