A Note on Koch & Denike's Analysis of John Snow's 1856 "Cholera in the south district of London"

Thomas S. Coleman*

September 4, 2019

Abstract

Koch and Denike [2006] undertake a valuable exercise to examine Snow [1856] and Snow's analysis of the South London data. Their goal is a correction of a suspected methodological problem in Snow's argument. Unfortunately, they are mistaken in certain respects. First, the issue they identify is not a problem, but a result of mis-reading or mis-interpreting Snow's data and analysis. Second, I argue that the statistical test they propose for testing Snow's claim is not really appropriate, and is not applied appropriately. Finally (and presumably due to the mis-interpretation of Snow's data) Koch and Denike alter mortality data by moving deaths geographically. I argue that this invalidates their data, analysis, and conclusions.

Keywords: John Snow, cholera, causal inference, epidemiology, statistical methodology, history of science

JEL Classification: C18, N33, N93, B40, C52

Snow Project data and code: https://github.com/tscoleman/SnowCholera

Contents

*Harris School of Public Policy, University of Chicago, tscoleman@uchicago.edu

1 Introduction

John Snow's investigations of the cholera epidemics in London in 1849 and 1854 (primarily found in Snow [1849, 1855], Westminster and London School of Hygiene and Tropical Medicine [1855], Snow [1856]) are justly famous, covered in both popular books and specialist texts. (Johnson [2007], Tufte [1997], Freedman [1999, 1991], Hempel [2007], Vinten-Johansen et al. [2003], Rothman [2002], McLeod [2000] is a very partial list.)

Koch and Denike [2006] undertake a valuable exercise to examine Snow [1856] and Snow's analysis of the South London data. Their goal is the correction of Snow's analysis: "This paper describes a previously unacknowledged methodological and conceptual problem in Snow's 1856 argument. We review the context of the South London study, identify the problem and then correct it with an empirical Bayes estimation (EBE) approach."

Unfortunately they are mistaken. First, the "methodological and conceptual problem" they describe is a mis-reading or mis-interpretation of Snow's data and analysis. Specifically, Koch and Denike focus on a pool of 623 deaths for which the *water supply company* could not be determined (but for which the location was known). They say or imply that the *location* of these deaths is uncertain, when in fact the location was known but the *supplier* was not. Further, I argue Snow's method for assigning these deaths to *supplier* is reasonable and not a methodological problem.

The second issue is that the statistical test they propose for testing Snow's claim that "the calculated mortality bears a very close relation to the real mortality" (a paired t-test shown in Table 2) is neither appropriate nor properly applied. It is not appropriate because deaths are counts and the variance will vary across sub-districts, depending in a specific way on the sub-district population. It is not properly applied because they use counts (instead of rates). Actual versus predicted counts will differ when the underlying population-at-risk differs (as it does for Snow's Table VI) even when the underlying mortality *rate* is the same. If a paired t-test is properly applied using mortality rates as shown in Table 3, it would fail to reject Snow's claim.

The final and much the most important problem, apparently a result of mis-interpreting the source data, is that Koch and Denike alter the location for mortality data from the Registrar-General (used by Snow and others). Although not ill-intentioned this altering of the original mortality data does invalidate their data, analysis, and conclusions. [1]

[1] One additional issue to highlight is the large number of typos in the transcription of Snow's tables and in the paper overall. For example Koch and Denike [2006] Figure 4 (a transcription of Snow's 1856 Table VI) has errors in the 1851 population for Christchurch, St. Saviour, St. Mary Magdelen, and London Road, to name a few. Regarding the paper, page 280 states: "The results returned, $\gamma = 411/4177$ (equalling 0.09983) ...". The figure 4177 should read 4117 (the sum of 3,706 + 411, the figures from Snow's Table V for the assigned Southwark & Vauxhall and Lambeth deaths). The phrase should read "$\gamma = 411/4117$" which is indeed 0.09983. As far as I can determine the typos do not carry through to later calculations.

This paper lays out the problems with Koch and Denike [2006]. A companion paper (Coleman [2019b], also Coleman [2018]) re-examines and extends Snow's South London analysis using modern statistical tools. The conclusion is that the data strongly support Snow's contention for "the overwhelming influence which the nature of the water supply exerted over the mortality, overbearing every other circumstance which could be expected to affect the progress of the epidemic." (Snow [1856] p. 248).

2 Review of Snow [1856] Table VI

Snow wanted to show that difference in water supplier was predominant in causing mortality, overriding other factors such as crowding or proximity to the Thames. His approach in Snow [1856] (specifically Table VI) was to calculate a predicted mortality for each sub-district, compare this against the actual mortality by sub-district, and argue that "the calculated mortality bears a very close relation to the real mortality."

There are three issues with Snow's data and method. The first concerns data: the population estimates by water supplier have errors. Snow recognized and highlighted this problem (Snow [1856] pp. 245-246). For example if we look at the first row for Table 1 (Christchurch, Southwark) we see that the population for both suppliers together (16,149) exceeds the total 1851 census population (16,022). [2]

The second issue is that Snow predicts mortality for the two water company suppliers but compares with the overall mortality, for which there are three suppliers (the Southwark & Vauxhall Co., the Lambeth Co., and "Other"). Ignoring "Other" could be important because this is a large source for some sub-districts: St. Mary, Newington shows 40% supplied by "Other", and Norwood 73%.

The third issue is that Snow did not have the statistical tools to formalize his assertion that the actual and predicted "bear[s] a very close relation". Coleman [2019a] and Coleman [2018] apply modern statistical tools and shows that the data strongly support Snow's claim.

2.1 Snow's Table VI

Turning to Snow's analysis and predicted mortality, for each subdistrict Snow calculated:

$$\hat{R}_{subdis,both} = \frac{N_{subdis,Southwark} \cdot \hat{R}_{Southwark} + N_{subdis,Lambeth} \cdot \hat{R}_{Lambeth}}{N_{subdis,Southwark} + N_{subdis,Lambeth}} \tag{1}$$

$\hat{R}_{subdis,both}$ Predicted mortality rate for both suppliers (Southwark & Vauxhall Co and Lambeth Co) as a population-weighted average. Shown as the final column in Snow [1856] Table VI my Table 1

[2] Note that the errors are more substantive at the sub-district level than the District level. Most importantly, according to Snow, population for streets in which no death occurred was not reported at the sub-district level but was reported at the District level.

$N_{subdis,Southwark}$ Sub-district population estimated for Southwark & Vauxhall, from Snow [1856] Tables I and II (and also shown in Table VI)

$\hat{R}_{Southwark}$, $\hat{R}_{Lambeth}$ Mortality rate for supplier (overall for all sub-districts) calculated by Snow from the *overall* regional deaths, shown in the last row of Snow [1856] Table V. Snow calculated the supplier-specific rates after allocating the 623 "not ascertained" deaths to the two suppliers proportionally – see Snow [1856] p. 247. Snow calculated $\hat{R}_{Southwark} = 160$ and $\hat{R}_{Lambeth} = 27$

The $\hat{R}_{subdis,both}$ is a population-weighted average of the Southwark and Lambeth mortality rates. Note for future reference that this includes only the two water supply *companies* and excludes the third source of supply: "Other" meaning wells, ditches, or the Thames.

Snow showed the result of these calculations in his Table VI, reproduced here as Table 1. Snow wished to compare the calculated mortality with the actual and argued that they "bear a close relation" (Snow [1856] p. 248).

3 Koch & Denike Analysis

3.1 Testing with paired *t*-test

Koch & Denike apply a paired t-test to the differences in counts (deaths) by sub-district: "Basic statistical tests, available to us but not to Snow, suggest his conclusion [for the strong relation between actual versus predicted mortality] was less than convincing (Fig. 5)." (p. 275) For each sub-district they calculate the difference in deaths. The null hypothesis is that on average deaths are the same in each sub-district: the mean of the differences is zero.

There are two issues with using this test. First, as discussed in the appendix, the paired-sample t-test is not ideal because it ignores the differing variances across sub-districts. Second, when comparing actual versus predicted we should use mortality *rates* (per 10,000 population) rather than the raw deaths (counts). This is important because of the sometimes different populations-at-risk for actual versus calculated mortality. With different populations the counts will differ even when the underlying mortality rates are the same.[3]

Table 2 shows the test performed by Koch & Denike using deaths (counts). This test implies rejecting the equality of counts. But counts may differ either because the mortality rates differ or because the population-at-risk differs. The first we care about, the second we do not. Koch & Denike's test shown in Table 2 tests both jointly and would not be appropriate even *if* the paired t-test were an appropriate statistical approach.

[3]As pointed out above, Snow in Table VI is comparing the overall sub-district mortality (three suppliers) with a prediction based on only two suppliers (the Southwark & Vauxhall Company and the Lambeth Company). The populations served by all three may be larger than the population served by the two companies. For Putney the overall population is 5,280 while the combined Southwark-plus-Lambeth population is only 74: many houses were supplied by pump-wells or the Thames. For Kennington 1st the total population is 24,261 while the combined Southwark and Lambeth population is only 18,483. For Kennington 1st the observed death count (total population 24,261) is 305, giving an estimated rate of 125.7 per 10,000. If we apply the rate of 125.7 to the Southwark and Lambeth population (18,483) we would have an expected count of only 232.2. Comparing 305 versus 232.2 is a difference of 72.7 but this reflects only differing population size and not underlying mortality or infection.

Table 1: Snow [1856] Table VI: Mortality from Cholera in 1854, in Thirty-one Sub-Districts, as compared with Calculations founded on the Results shown in Table V

	1855 Seq	District	Sub-District	Pop 1851	Population Estimates by Supplier			Deaths 1854		Calculated Mortality			
					Southwark	Lambeth	Both Cos	Count	Rate	Southwark	Lambeth	Both	Rate
1	13	St. Saviour, S.	Christchurch, Southwark	16,022	2,915	13,234	16,149	113	*70.5*	*46.6*	*35.7*	*82.4*	*51.0*
2	1	St. Saviour, S.	St. Saviour, Southwark	19,709	16,337	898	17,235	378	*191.8*	*261.4*	*2.4*	*263.8*	*153.1*
3	2	St. Olave	St. Olave, Southwark	8,015	8,745	0	8,745	161	*200.9*	*139.9*	*0.0*	*139.9*	*160.0*
4	3	St. Olave	St. John, Horsleydown	11,360	9,360	0	9,360	152	*133.8*	*149.8*	*0.0*	*149.8*	*160.0*
5	4	Bermondsey	St. James, Bermondsey	18,899	23,173	693	23,866	362	*191.5*	*370.8*	*1.9*	*372.6*	*156.1*
6	5	Bermondsey	St. Mary Magdalen	13,934	17,258	0	17,258	247	*177.3*	*276.1*	*0.0*	*276.1*	*160.0*
7	6	Bermondsey	Leather Market	15,295	14,003	1,092	15,095	237	*155.0*	*224.0*	*2.9*	*227.0*	*150.4*
8	14	St. George, S.	Kent Road	18,126	12,630	3,997	16,627	177	*97.6*	*202.1*	*10.8*	*212.9*	*128.0*
9	15	St. George, S.	Borough Road	15,862	8,937	6,672	15,609	271	*170.8*	*143.0*	*18.0*	*161.0*	*103.1*
10	16	St. George, S.	London Road	17,836	2,872	11,497	14,369	95	*53.3*	*46.0*	*31.0*	*77.0*	*53.6*
11	17	Newington	Trinity, Newington	20,922	10,132	8,370	18,502	211	*100.9*	*162.1*	*22.6*	*184.7*	*99.8*
12	18	Newington	St. Peter, Walworth	29,861	14,274	10,724	24,998	391	*130.9*	*228.4*	*29.0*	*257.3*	*102.9*
13	19	Newington	St. Mary, Newington	14,033	2,983	5,484	8,467	92	*65.6*	*47.7*	*14.8*	*62.5*	*73.9*
14	20	Lambeth	Waterloo Road (1st)	14,088	3,548	11,939	15,487	59	*41.9*	*56.8*	*32.2*	*89.0*	*57.5*
15	21	Lambeth	Waterloo Road (2nd)	18,348	7,171	12,533	19,704	118	*64.3*	*114.7*	*33.8*	*148.6*	*75.4*
16	22	Lambeth	Lambeth Church (1st)	18,409	3,113	15,878	18,991	49	*26.6*	*49.8*	*42.9*	*92.7*	*48.8*
17	23	Lambeth	Lambeth Church (2nd)	26,784	7,868	16,023	23,891	195	*72.8*	*125.9*	*43.3*	*169.2*	*70.8*
18	24	Lambeth	Kennington (1st)	24,261	15,775	2,708	18,483	305	*125.7*	*252.4*	*7.3*	*259.7*	*140.5*
19	25	Lambeth	Kennington (2nd)	18,848	7,874	5,620	13,494	143	*75.9*	*126.0*	*15.2*	*141.2*	*104.6*
20	26	Lambeth	Brixton	14,610	1,922	9,356	11,278	48	*32.9*	*30.8*	*25.3*	*56.0*	*49.7*
21	29	Lambeth	Norwood	3,977	0	1,066	1,066	10	*25.1*	*0.0*	*2.9*	*2.9*	*27.0*
22	27	Wandsworth	Clapham	16,290	6,747	134	6,881	167	*102.5*	*108.0*	*0.4*	*108.3*	*157.4*
23	8	Wandsworth	Battersea	10,560	6,276	276	6,552	171	*161.9*	*100.4*	*0.7*	*101.2*	*154.4*
24	9	Wandsworth	Wandsworth	9,611	907	94	1,001	59	*61.4*	*14.5*	*0.3*	*14.8*	*147.5*
25	10	Wandsworth	Putney	5,280	74	0	74	9	*17.0*	*1.2*	*0.0*	*1.2*	*160.0*
26	30	Wandsworth	Streatham	9,023	0	3,244	3,244	15	*16.6*	*0.0*	*8.8*	*8.8*	*27.0*
27	31	Camberwell	Dulwich	1,632	0	25	25	0	*0.0*	*0.0*	*0.1*	*0.1*	*27.0*
28	11	Camberwell	Camberwell	17,742	9,139	639	9,778	242	*136.4*	*146.2*	*1.7*	*147.9*	*151.3*
29	12	Camberwell	Peckham	19,444	5,438	392	5,830	175	*90.0*	*87.0*	*1.1*	*88.1*	*151.1*
30	28	Camberwell	St. George, Camberwell	15,849	4,295	5,437	9,732	132	*83.3*	*68.7*	*14.7*	*83.4*	*85.7*
31	7	Rotherhithe	Rotherhithe	17,805	12,218	0	12,218	283	*158.9*	*195.5*	*0.0*	*195.5*	*160.0*
		Houses in streets with no death			28,929	23,338	52,267			*462.9*	*63.0*	*525.9*	*100.6*
		Not identified			2,712	165	2,877			*43.4*	*0.4*	*43.8*	*152.4*
		Totals		482,435	267,625	171,528	439,153			*4,282.0*	*463.1*	*4,745.1*	*108.1*
		Population per Registrar-General			266,516	173,748	440,264			*4,264.3*	*469.1*	*4,733.4*	*107.5*

This reproduces Snow [1856] Table VI with minor modifications. "Pop 1851" is the population estimate from the 1851 Census, reported in various tables in Snow [1855, 1856]. "Population Estimates by Supplier" are from Snow [1856] Table V using population estimates from Simon [1856]. Death counts for 1854 are from the Registrar-General and generally match Snow [1855] Table XII. (I assume changes reflect updates to the Registrar-General's counts subsequent to Snow's 1855 publication.) "Calculated Mortality: Southwark" and "Calculated Mortality: Lambeth" are the count of deaths based on mortality rates of 160 and 27 per 10,000 (Snow's calculation from his Table V) and the Southwark & Vauxhall and Lambeth sub-district populations – Equation 1. The final five columns (displayed *in italic*) are calculated based on the earlier columns, rounded to one decimal. Snow reported these numbers rounded to zero decimals, and there are a few minor differences due to errors in Snow's rounding. (See http://www.hilerun.org/econ/papers/snow/index.html for postings of Snow's original data.) The only substantive error is for the Christchurch sub-district: Snow reported a mortality rate of 57 per 10,000 where it should be 51. "1855 Seq" is the sequence number from Snow's 1855 tables (where sub-districts are sorted by "first 12" Southwark-only, then "next 16" jointly-supplied).

Table 3 shows the t-test on differences in *rates*, and here we find a t-ratio of 1.43 and we cannot reject the hypothesis that differences in mortality rates between actual and predicted are (on average) zero. Essentially, Koch & Denike apply a less-than-ideal test, and do not apply it appropriately.

	Mean	S. Dev	Std Err	t-test	Sig (2-tailed)
Difference in counts (deaths)	28.7609	47.0320	8.4472	3.405	0.002

Table 2: Paired t-test for death counts from Snow [1856] Table VI (my Table 1)

	Mean	S. Dev	Std Err	t-test	Sig (2-tailed)
Difference in mortality rates (per 10,000)	-10.1428	39.5349	7.1007	-1.4284	0.1635

Table 3: Paired t-test for mortality rates from Snow [1856] Table VI (my Table 1)

The paired t-test is not a good test for examining Snow's problem. A difference-in-differences (DiD) regressions is a more appropriate framework. Snow himself (Snow [1855] p. 89 and Table XII) performed a rudimentary DiD comparison on aggregate (not sub-district) observations. We can extend the DiD regression to sub-districts and, with the population data in Table VI, calculate a continuous DiD effect. This provides one approach for testing Snow's claim for a close connection between actual versus predicted. Coleman [2019a, 2018] do this.

A second approach for testing the effect of clean water is to directly compare the treated and untreated populations, as in a randomized control trial. Snow [1855] Table IX did this for the aggregate data but was limited to comparing *households* rather than individuals. Again, the population-by-supplier data in Snow [1856] provides the missing piece that allows us to extend and refine Snow's direct comparison. Coleman [2018] does this.

3.2 Re-Allocation of 623 Deaths

Turning from statistical testing to the data, the more serious issue with Koch and Denike [2006] is misunderstanding Snow's original data, which leads them to (unintentionally) alter and thus contaminate the original mortality data. The problem Koch & Denike set out to address is that for the 1854 cholera outbreak there were 623 deaths in the South London districts that were not assigned to water company *supplier* (Southwark & Vauxhall Company versus Lambeth Company).

Koch & Denike state "A problem in this approach was how Snow addressed data lacking spatial assignment: there were 623 houses in which cholera occurred that could not be assigned reflexively to any single district nor to either of the two water supplier areas." (p. 275) The first part of this statement is simply incorrect while the second part is slightly confusing. Regarding spatial assignment to district, the 623 deaths in houses were clearly assigned to Registration District and sub-district in the original reports from the Registrar-General. What could not be determined for these 623 was which *water supplier* supplied the house. The assignments to Registration District are clearly shown in Snow's Table V (Koch and Denike's Figure 3, my Table 4). There has never (to my knowledge) been any question about the reliability of the Registrar-General's assignment of

deaths to District or Sub-District – in contrast to assignment to water *supplier* within sub-district.[4][5] The second part of the statement is confusing because there never was assignment to water supplier *areas* but rather to water supply *companies* within areas that were (sometimes) jointly supplied by the two companies.[6]

Koch & Denike proceed to (incorrectly) re-allocate those 623 deaths across water supplier *and* Registration Districts (see their Figure 6, my Table 5). In doing so they move deaths across Districts, deaths that were reliably located – assigned to District by the Registrar-General. The re-assignment across Districts is neither necessary nor justified and corrupts the original data. The re-assignment introduces substantive errors in mortality rates for those districts with either below-average or above-average unassigned deaths. For example the mortality rate for St. Saviour, Southwark is increased from 137.4 to 156.8 (per 10,000), while Lambeth is reduced from 66.5 to 56.5. All of these changes in overall mortality at the Registration District level are arbitrary and counter to the observed data.[7]

The original problem in Snow's analysis of the 623 deaths not assigned to supplier can, I argue, be handled more simply by allocating at the District or sub-district level in proportion to the reported deaths for the two suppliers. Snow himself states that argument rather clearly:

[4]Throughout, Koch & Denike imply, incorrectly, that the problem with the 623 deaths is spatial location. On page 272 they state: "Snow attempted to extend the incomplete field using a then recently compiled but still incomplete inventory of deaths from cholera at registration district and sub-district levels through the simple expedient of allocating cases that *could not otherwise be spatially located* to water companies according to the best estimate then available" (emphasis added). In fact, all the 623 deaths were correctly *geographically* or spatially located within District and sub-district. (See, for example, Snow's 1856 Table V. Even more instructive is the appendix to Snow [1855] where Snow lists *all* deaths for the four weeks ending 5th August 1854. Snow had the address of each death and all deaths were assigned to sub-districts, as is obvious from perusing the list. Snow visited the households to ascertain the water source. In some cases Snow could not ascertain the water source; for example on p. 141 "At St. Thomas's Hospital, supposed from Red Cross Street, Southwark, on 31st July, a charwoman, aged 50, 'cholera'". In all cases, however, the location was known and recorded.)

[5]There are other instances where Koch & Denike's statements are not good representations of Snow's analysis and methods. Koch and Denike [2006] p. 273 state: "'All that was required,' he [Snow] wrote, 'was to learn the supply of water to each individual house where a fatal attack of cholera might occur' (Snow, 1855, p. 75). *Data permitting assignment of cholera deaths to either of the water suppliers was unavailable, however*" (emphasis added). This does not clearly represent the situation. Snow was stating the problem. But he then proposed *and implemented* the solution (see Snow [1855] p. 76 ff). As a result of Snow's efforts data the data was available: Snow himself collected the data through August 26 and convinced the Registrar-General to collect data following August 26. Another example is page 278 where Koch and Denike state "Required was some form of smoothing permitting a better appreciation of the reliability of risk of cholera in registration districts and thus a better assignment of the 623 unassigned cases (*561 from Southwark-Vauxhall, 62 from Lambeth Company*)" (emphasis added). This seems to be a mis-understanding (or poor explanation) of Snow's data. The fundamental problem is that we do not know how many of those 623 unassigned cases were supplied by the Southwark & Vauxhall Company and how many the Lambeth Company. The figures of 561 and 62 are in fact only Snow's estimates, his allocation of the unascertained cases between the Southwark company and the Lambeth company based on the reported (known) assignments of 3,706 and 411 (see Snow's Table V). The problem is not (as Koch and Denike seem to imply) the location or *geographic* assignment of deaths, but the assignment to water supplier for deaths that are well-identified as to location.

[6]In fairness, it is easy to confuse *areas* and *companies* because the names overlap. Southwark is the name of the general region and appears in the official names of three Districts (St. Saviour, Southwark; St. Olave, Southwark; and St. George, Southwark) as well as the in name of one of the water supply companies (the Southwark & Vauxhall Company). Lambeth is the name of one of the Districts as well as appearing in the names of two sub-districts (Lambeth Church, 1st; Lambeth Church, 2nd) and it is also the name of the other water supply company (the Lambeth Water Company). It is important to recognize, however, that the Southwark & Vauxhall Company and the Lambeth Water Company both supplied customers in many of the Districts and sub-districts. For example, in the Lambeth *District* both the Southwark & Vauxhall *Company* and the Lambeth *Company* supplied customers.

[7]I have focused on Koch & Denike's Figure 6 which shows results at the District level. Their Figure 7 re-assigns the 623 deaths at the sub-district level and some of those results alter mortality even more dramatically. Consider the District of Newington, comprised of three sub-districts with an overall population of 64,816. Koch and Denike's Figure 7 show re-allocated deaths from both companies (Southwark & Vauxhall Company and Lambeth Company) across the three Newington sub-districts totaling to 501.14. Adding the 2 pump-wells deaths gives a total of 503.14 versus the true total of 694 deaths. Dividing by the population gives a Newington District re-calculated mortality rate of 77.6 per 10,000, dramatically below the true rate of 107.1.

Table 4: Snow [1856] Table V – Reported Deaths and Mortality by Water Supply and Registration District for 1854 Outbreak

Registration District	Pop 1851	Southwark	Lambeth	Pump-well	Unasc	Total	Mortality per 10,000
1, St. Saviour, Southwark	35,731	406	72	10	3	491	137.4
2, St. Olave, Southwark	19,375	277	0	8	28	313	161.5
3, Bermondsey	48,128	821	0	25	0	846	175.8
4, St. George, Southwark	51,824	388	99	0	56	543	104.8
5, Newington	64,816	458	58	2	176	694	107.1
6, Lambeth	139,325	525	138	24	240	927	66.5
7, Wandsworth	50,764	268	7	106	40	421	82.9
8, Camberwell	54,667	352	33	115	49	549	100.4
9, Rotherhithe	17,805	207	0	46	30	283	158.9
10, Greenwich & sub-districts, Sydenham		4	4	2	1	11	–
TOTAL	482,435	3,706	411	338	623	5,078	105.3

The 1851 population for Camberwell is reported by Snow in his Table V as 54,607 when in fact it should be 54,667 (according to cross-checks by summing and comparing with his 1856 Table VI and 1855 Table VIII). This is one of the few errors or typos in Snow's publications that I have found.

> The instances in which the water supply was not specified, or not ascertained, in the returns made by the district registrars must evidently nearly all have been cases in which the house was supplied by one or other of the water companies, for, if the persons received no such supply, and obtained water from a pump well, canal, or ditch, there could be no difficulty in knowing the fact. Moreover, as the two water companies are guided by precisely the same regulations, the difficulty in ascertaining the supply is exactly the same with regard to one as the other; I, therefore, concluded that I could not be wrong in dividing the non-ascertained cases between the two companies in the same proportion as those which were ascertained, and I have done so at the foot of table V (Snow [1856] p. 247)

4 Appendix

4.1 Paired-Sample *t*-Test

A statistical test that captures some of Snow's original intent (although it is ultimately not an appropriate test) is a paired-sample t-test. For each of the 31 sub-districts in Table 1 take the difference between the actual and predicted mortality rates. This gives 31 observations (shown in Figure 1 below). If water quality (dirty Southwark & Vauxhall Co. versus clean Lambeth Co.) is indeed the most important factor in accounting for mortality we would expect no differences between the actual and predicted.

We would not expect the differences to be exactly zero for every one of the 31 sub-districts; there will be some random variation and possibly other factors than water. At a minimum we would require that the differences are zero *on average* and this is exactly what the paired-sample t-test tests: we ask whether the observed mean is close to zero, based on the variability we see in the data.

Table 5: Koch and Denike [2006] Figure 6 – Re-Calculated Deaths and Mortality by Water Supply and Registration District, Re-assigned *Across Districts* with Empirical Bayes Estimation Process

Registration District	Pop 1851	Southw	Lamb	Pump-well	Unasc	Total	Mortality per 10,000 K&D Calc	Original
1, St. Saviour, Southwark	35,731	*467.96*	*82.37*	*10*	*0*	*560*	*156.8*	137.4
2, St. Olave, Southwark	19,375	*317.63*	*1.28*	*8*	*0*	*327*	*168.7*	161.5
3, Bermondsey	48,128	*943.92*	*1.32*	*25*	*0*	*970*	*201.6*	175.8
4, St. George, Southwark	51,824	*447.88*	*112.82*	*0*	*0*	*561*	*108.2*	104.8
5, Newington	64,816	*527.36*	*66.72*	*2*	*0*	*596*	*92.0*	107.1
6, Lambeth	139,325	*605.61*	*157.72*	*24*	*0*	*787*	*56.5*	66.5
7, Wandsworth	50,764	*307.59*	*9.03*	*106*	*0*	*423*	*83.3*	82.9
8, Camberwell	54,667	*405.02*	*38.24*	*115*	*0*	*558*	*102.1*	100.4
9, Rotherhithe	17,805	*237.06*	*1.26*	*46*	*0*	*284*	*159.7*	158.9
10, Greenwich & sub-districts, Sydenham		*6.78*	*2.43*	*2*	*0*	*11*	–	–
TOTAL	482,375	4,267	473	338	0	5,078	105.3	105.3

The "Southwark" and "Lambeth" columns are from Koch and Denike [2006] Figure 6, "assignments using the empirical Bayes estimation process". Pump-well are from Snow [1856], the "Total" is the sum of the other columns. The 623 unassigned deaths from Snow [1856] Table V are (incorrectly) re-assigned across sub-district. The incorrectly re-assigned data reported by Koch & Denike are shown in *italic*.

The paired test is based on the assumption that each of the 31 differences are drawn from a normal distribution with constant (but unknown) variance. We calculate the ratio of the mean divided by standard error (estimated standard deviation normalized by the number of sub-districts). The ratio will be Student-t–distributed. Table 3 shows the results for such a t-test.

Table 3 shows that the mean of differences between actual and predicted is zero (more technically we cannot reject that the mean is zero): the t-ratio is small (-1.4) and statistically insignificant. This supports Snow's contention of a close relation.

It is important that in comparing actual versus predicted we use mortality *rates* per 10,000 population rather than the raw deaths (counts). This is necessary because of the sometimes very different populations between actual versus calculated. With different populations the counts will differ even when the underlying mortality rates are the same.

This testing framework, however, is not ideal. First it gives us no indication of *how much* of the variation in the original mortality rates is accounted for by the population-weighted predictions. Second, the test is simply not appropriate for mortality rates because the assumption of constant variance normality is not appropriate. We now turn to these issues.

4.2 Population-Dependent Variance – Graphical Examination of Error Process

In the next section we discuss count regression, which provides an appropriate statistical framework for analyzing Snow's mortality data. Before turning to the regression framework, however,

it is valuable to examine Snow's data graphically. Although a paired t-test is not a good testing framework, pursuing the idea reveals and highlights an important issue regarding the error process for the mortality rates examined here. For a paired t-test we would calculate the difference in mortality

rates for each sub-district. The left panel of Figure 1 shows these differences as solid circles. We would use these observations to estimate the underlying variance, the assumption for the paired t-test being that the differences are drawn from a normal distribution with a common variance. Figure 1 shows error bars under this normality assumption and using the variance calculated from the observed data.

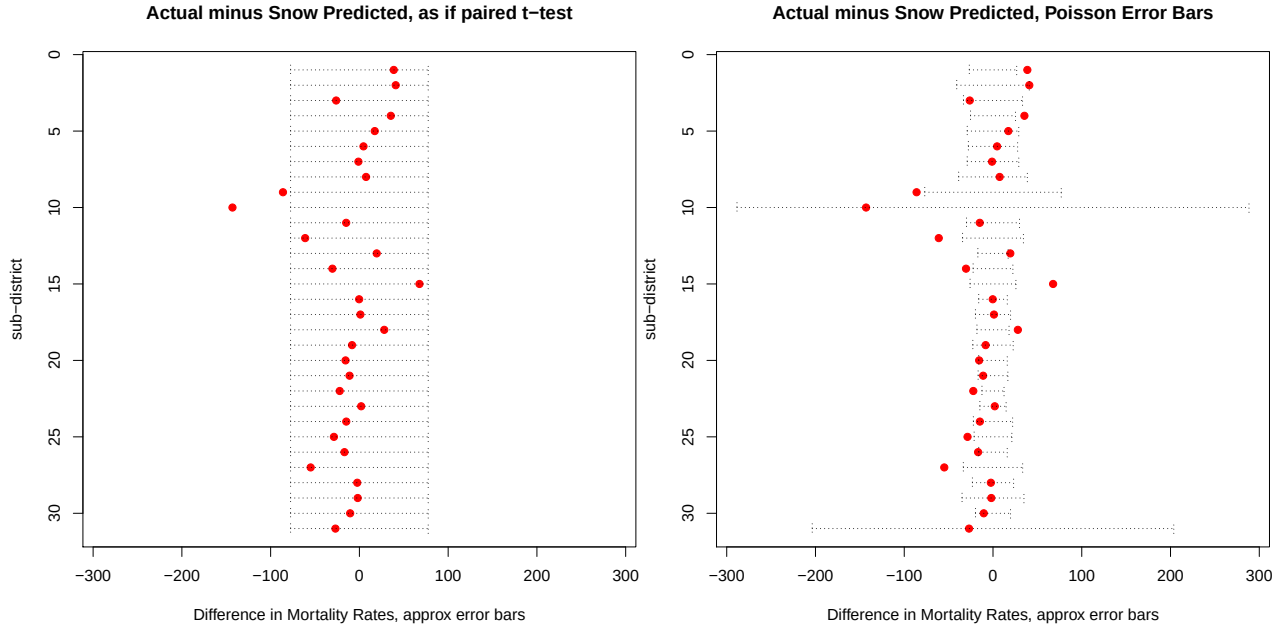

Figure 1: Difference in Mortality Rates (per 10,000) Between Actual and Snow's Predicted for 1854 for the 31 Sub-districts shown in Snow [1856] Table VI (my Table 1)

The red circles are the actual sub-district mortality rate minus Snow's predicted rate based on Southwark and Lambeth populations and Snow's calculated rates (160 per 10,000 for Southwark & Vauxhall and 27 for Lambeth – Equation 1). For consistency with other figures, sub-districts are sorted as in Snow [1855], the "1855 Sequence no." shown in Table 1. The left panel shows approximate 95% error bars *if* the difference in rates were normally distributed random variables with the same variance. The right panel shows approximate 95% error bars assuming the counts are generated by an underlying Bernoulli process with different rates and variances for each sub-district and for the calculated versus actual mortality rates: $SE(r1 - r2) = \sqrt{r1(1-r1)/n1 + r2(1-r2)/n2}$.

One important reason the paired t-test is not appropriate is that the assumption of common-variance normality is not true. Each sub-district mortality rate is itself an estimated mean (the sum of Bernoulli binary events – death or non-death – divided by the population – sample size) and each mean has its own standard error. For a large population the rates will go to normal by the central limit theorem. The difference between the two means will also be normal (since the difference of normals is normal). This forms the basis for the standard t-test for comparing rates in clinical trials. We will want to allow different variances if the samples sizes are different. (See the Appendix of Coleman [2018] and BMJ, Jakobsen et al. [2015].) The resulting formula for the variance of the difference in rates is $SE(r1 - r2) = \sqrt{r1(1-r1)/n1 + r2(1-r2)/n2}$. When one or both sample sizes are small the standard error will be large.

The right panel of Figure 1 shows the standard error for the differences based on the counts being generated by a Bernoulli process. (Or, as an approximation, a Poisson process.) Some of the error

bars are very wide: Putney (sub-district 10 or "1855 seq no" 10) is particularly wide because there are only 74 people supplied by the Southwark & Vauxhall and the Lambeth Companies combined. The assumption that all sub-districts have the same variance, embedded in the left panel of Figure and the paired t-test of Table 3, is clearly far from true.

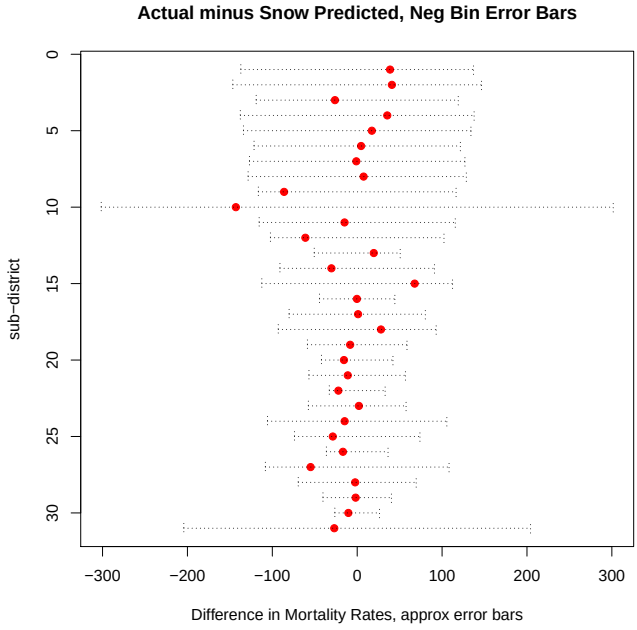

Figure 2: Difference in Mortality Rates (per 10,000) Between Actual and Snow's Predicted for 1854 for the 31 Sub-districts shown in Snow [1856] Table VI (my Table 1)

The red circles are the actual sub-district mortality rate minus Snow's predicted rate based on Southwark and Lambeth populations and Snow's calculated rates (160 per 10,000 for Southwark & Vauxhall and 27 for Lambeth – Equation 1). For consistency with other figures, sub-districts are sorted as in Snow [1855], the "1855 Sequence no." shown in Table 1. The approximate 95% error bars assume the counts are generated by an underlying Negative Binomial process (Poisson mixture process). This implies different rates and variances for each sub-district and for the calculated versus actual mortality rates:$SE(r1-r2) = \sqrt{\left(r1/n1 + r1^2/\theta\right) + \left(r2/n1 + r2^2/\theta\right)}$, with the Gamma mixing parameter $\theta = 12.8$.

But the Poisson error bars are still not the complete story. We want to incorporate both the variation *across* sub-districts (the error bars in the left panel of Figure 1) and *within* sub-districts (the right panel of Figure 1). As discussed in Coleman [2018] we need to use the statistical framework of counts, Poisson and Negative Binomial regression.[8] Figure 2 previews what we see when we incorporate both the within- and across-sub-district variation.

The left panel of Figure 1 assumes that each sub-district rate (or each difference in rates) is drawn from a random distribution (normal in this case) with common variance. The right panel of Figure 1 assumes that each sub-district rate is a fixed number, which differs across sub-districts. The difference is essentially assuming random versus fixed effects. Combining the across and within sub-district variation to give a Negative Binomial as mixture of Poissons assumes random effects.

[8]In the framework of Poisson count regressions (which is the appropriate mathematical extension of the Bernoulli assumption) we incorporate random sub-district rates by allowing the underlying sub-district rates to themselves be random variables, something very much like the left panel of Figure 1. We mix the rates that generate the Poisson distributions with an underlying distribution, often using a Gamma which then produces a Negative Binomial count distribution. This is discussed in the Appendix of Coleman [2018].

This raises a subtle but critically important issue. There are no data in Snow's Table VI (my Table 1) which could distinguish between the two hypotheses of fixed versus random effects. For the difference-in-differences analysis of Coleman [2019a]we do have some data (within sub-district comparisons of 1849 versus 1854) but the issue is still delicate. We know that sub-districts differ and so we should expect them to have somewhat different intrinsic rates, but there is also some evidence that rates within sub-district vary. The choice between random versus fixed effects does not have a major impact on estimates of parameters, but it has a dramatic impact on the standard errors and thus the confidence we assign to our estimates.

References

BMJ. 13. Study design and choosing a statistical test | The BMJ. URL https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design

Thomas S. Coleman. Causality in the Time of Cholera: John Snow as a Prototype for Causal Inference. SSRN Scholarly Paper ID 3262234, Social Science Research Network, Rochester, NY, October 2018. URL https://papers.ssrn.com/abstract=3262234.

Thomas S. Coleman. John Snow, Cholera, and the Birth of Difference-in-Differences Regression. SSRN Scholarly Paper, Social Science Research Network, Rochester, NY, July 2019a.

Thomas S. Coleman. Revisiting John Snow's 1856 "Cholera in the south district of London". SSRN Scholarly Paper, Social Science Research Network, Rochester, NY, July 2019b.

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.

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.

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.

JC Jakobsen, M Tamborrino, P Winkel, N Haase, A Perner, J Wetterslev, and C Gluud. Count Data Analysis in Randomised Clinical Trials. *Journal of Biometrics & Biostatistics*, 06(01), 2015. ISSN 21556180. doi: 10.4172/2155-6180.1000227. URL https://www.omicsonline.org/open-access/count-data-analysis-in-randomised-clinical-trials-2155-618

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.

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

K. S. McLeod. Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine (1982)*, 50(7-8):923–935, April 2000. ISSN 0277-9536.

Kenneth J. Rothman. *Epidemiology: an introduction*. Oxford University Press, New York, N.Y., 2002. ISBN 978-0-19-513553-4.

John Simon, editor. *Report on the last two cholera-epidemics of London: as affected by the consumption of impure water addressed to the Rt. Hon. The President of the General Board of Health, by the Medical Officer of the Board*. Printed by Eyre and Spottiswoode, for HMSO, London, 1856. URL https://collections-nlm-nih-gov.proxy.uchicago.edu/catalog/nlm:nlmuid-0260772-bk. OCLC: 14531255.

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

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

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

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

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

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