Theory Choice and Overdetermination of Evidence

The Case of The 1854 London Cholera Outbreak

James Michelson
Carnegie Mellon University, Department of Philosophy
jamesmic@andrew.cmu.edu

Abstract

Sed fringilla tempus hendrerit. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Etiam ut elit sit amet metus lobortis consequat sit amet in libero. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus vel sem magna. Nunc at convallis urna. isus ante. Pellentesque condimentum dui. Etiam sagittis purus non tellus tempor volutpat. Donec et dui non massa tristique adipiscing. Quisque vestibulum eros eu. Phasellus imperdiet, tortor vitae congue bibendum, felis enim sagittis lorem, et volutpat ante orci sagittis mi. Morbi rutrum laoreet semper. Morbi accumsan enim nec tortor consectetur non commodo nisi sollicitudin. Proin sollicitudin. Pellentesque eget orci eros. Fusce ultricies, tellus et pellentesque fringilla, ante massa luctus libero, quis tristique purus urna nec nibh.

Introduction

The 1984 Soho, London Cholera outbreak is widely considered to demarcate a "historical turning point" [p.162][2] in the development of modern epidemiology. John Snow's championing of the germ theory over the miasma theory of disease has been a fruitful case study for philosophers of science [5]. What has been less closely considered; however, is the degree to which the infamous Broad Street Pump handle removal was actually a straightforward application of probabilistic reasoning. Using a simple, Bayesian statistical model I demonstrate that if one starts from the assumption that the germ theory of disease is correct, one quickly arrives at the conclusion the Broad Street Pump is the origin of the cholera outbreak.

Our goal is *not* to propose an Bayesian model as realistic explanation of a scientist's actual thought process nearly two hundred years ago. Instead, I view this contribution as offering a path to more accurately reconstructing historical events from quantitative data. Given the rich historical record from John Snow's contemporaries we validate our conclusions from a statistical model with historical sources, showing that our simple model recovers many key features of a well-studied historical event.

Data

I use of the 1854 London Cholera outbreak data from `cholera` R package [3], which was compiled from numerous, disparate original sources. It provides both geographical coordinates for cholera fatalities, local landmarks, as well as supplementary time-series data. Notably, however, there is no unified view of the distributions of cholera attacks in both space and time: the location of of each individual fatality is recorded but not its associated date. The time-series begins on the 19th of August, 1854 and ends on the 30th of September 1854, a period of 43 days. There are 578 recorded fatalities in the Soho neighborhood, in the vicinity of water pumps.

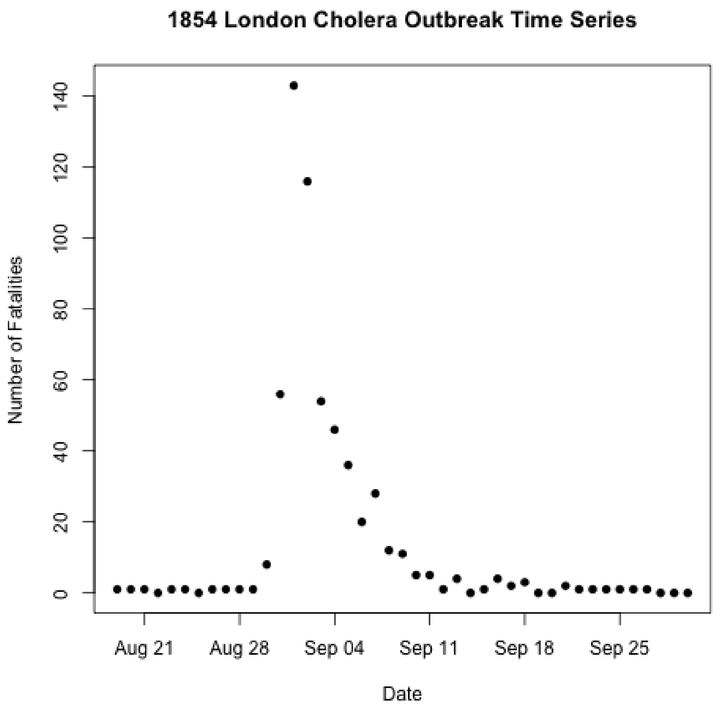

Figure 1: Time Series of Cholera Outbreak Fatalities

Model

I model John Snow's beliefs in the source of the Cholera outbreak over the $k = 13$ water pumps in Soho. The likelihood $f_{X|p_1,\ldots,p_k}(X|p_1,\ldots,p_k)$ is multinomial distribution $(x_1,\ldots,x_k) \sim \text{Multinomial}(n,(p_1,\ldots,p_k))$ (where $\sum_{i=1}^{k} p_i = 1$.) I use a Dirichlet distribution as a conjugate prior with $\alpha = (\alpha_1,\ldots,\alpha_k)$. I can obtain a uniform density by using the prior $\alpha_j = 1$ for all j [1, p.69]. The resulting posterior distrubtion is $\theta_j = \alpha_j + x_j$.

On each of the 43 days in our time series, I randomly sample the number of geographically-located fatalities based on the number of fatalities recorded in our time-series. This circumvents the lack of a single unified spatio-temporal view of the outbreak. Each day, these fatalities are tallied to the closest nearby pump based on their 'taxicab' or 'Manhattan' distance. This daily tally is the added to our prior, which is the previous day's tally. This posterior is continually updated over the time-series. The result randomly shuffles all 578 fatalities in time, recovers the same final result by conclusion of the time series.

Results

Starting with a uniform prior probability of each pump being the source of the outbreak, this simple Bayesian model clearly shows that the Broad Street Pump was more likely to be the source of the outbreak **before** the spike in deaths on September 1st 1854.

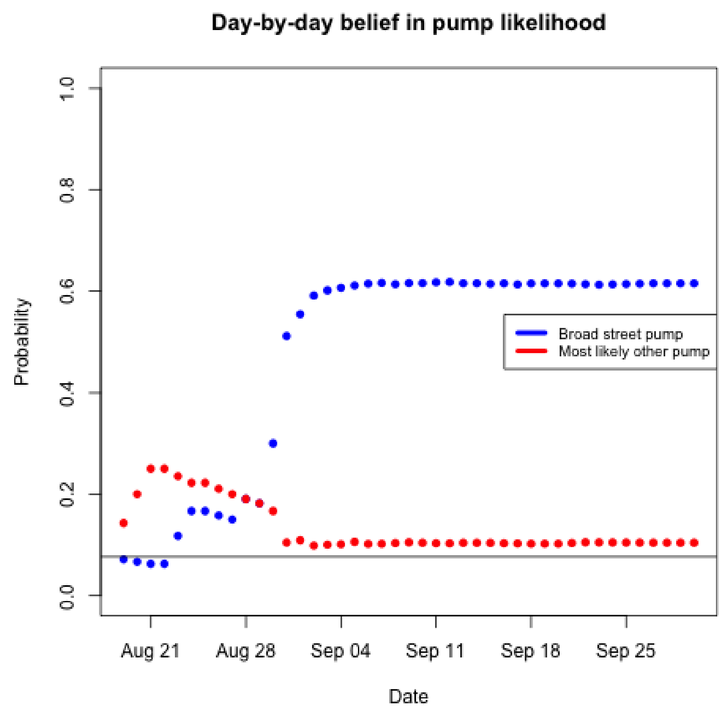

Figure 2: Day-by-day change in belief about the source of the outbreak

The convergence to $\tilde{6}0\%$ reflects the fact that this proportion of fatalities occurred in the neighborhood of the broad street pump[1]. Note, however, that by August 31st 1854 the probability the Broad Street Pump was the source of the outbreak was greater than all the other pumps combined. **these points should be violin plots!**

Discussion

A clear limitation of this study, which must frame the analysis of our results, is that I have implicitly assumed that John Snow learns of each fatality as soon as it occurs. This simplifying assumption is not as damaging as it might initially seem. John Snow extensively canvassed Soho for information about the Cholera Outbreak as it ravaged the neighborhood (see [2]) and so likely there was an information lag measured only in days. This is important to note if one juxtaposes the historical record alongside the time-series as it is presented here. In John Snow's own words,

> As soon as I became acquainted with the situation and extent of this irruption of cholera, I suspected some contamination of the water of the much-frequented street-pump in Broad Street, near the end of Cambridge Street; but on examining the water, on the evening of the 3rd September, I found so little impurity in it of an organic nature, that I hesitated to come to a conclusion. [4, §2]

This first-person account confirms the plausibility of my result: the broad street pump was a clear front-runner among sources of the outbreak. Indeed, the model shows that the removal of the Broad Street Pump handle on the 12th of September comes long after John Snow identified the pump as the probable source. This accords with the historical accounts which have documented the sway that the miasma theory of disease held over public officials at the time (see [2]). Thus, officials may have required much longer to accede the requests of a scientist making his case on the basis of an entirely different theory of disease.

Conclusion

This work shows how quantitative methods can be used to reach similar conclusions to those informed by careful historical scholarship. The approach here bears directly on other episodes in the history of science without the same depth of primary historical sources. In their absence, simple models like these can be used to ascertain whether hypotheses are plausible candidates for further study. Future work could investigate whether this model recovers the five pumps that John Snow sampled water from over the course his investigations [2, p.76].

References

[1] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[2] Steven Johnson. *Ghost Map*. Riverhead Books, 2nd edition, 2006.

[3] Peter Li. *cholera: R Package for Analyzing John Snow's 1854 Cholera Map*, 2019. R package version 0.7.0, available at `https://github.com/lindbrook/cholera`,.

[4] John Snow. *On the Mode of Communication of Cholera*. London: John Churchill, New Burlington Street, England, 1855.

[5] Dana Tulodziecki. A case study in explanatory power: John snow's conclusions about the pathology and transmission of cholera. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42(3):306 – 316, 2011.

Acknowledgements

I would like to thank Scott Weingart and David Danks.

[1] This result is robust to alternate distance specifications like Euclidean Distance