Two Historic Case Studies
John Snow: Data Scientist
London in the 1850’s

- London in the 1850’s was wealthy, but many citizens lived in extreme poverty
- Disease was rampant, especially cholera
- Causality between unsanitary food and water and diseases was not yet accepted
- **Miasma theory**: bad air was generally believed to be the cause of cholera
Cholera in London in the mid 1800’s

- The first cases of cholera in London were reported in 1831
- Cholera killed hundreds of Londoners in one week in 1854
  “Within 250 yards of the spot where Cambridge Street joins Broad Street, there were upwards of 500 fatal attacks of cholera in 10 days,” Dr. Snow wrote.
- Patients died within a day or two of contracting the disease
- One wave of cholera could kill tens of thousands
John Snow

- John Snow, an obstetrician/anaesthesiologist, was skeptical of the miasma theory. If the disease was airborne, wouldn’t people living in the same area be breathing in the same toxins?
- But entire households were dying of cholera, while their neighbors were perfectly fine.
Snow’s Observations

- Symptoms included vomiting and diarrhea
- This evidence led Snow to believe the disease was food/drink borne as opposed to airborne
- He hypothesized that some drinking water was contaminated
Plotting Cholera

- August, 1854
- Mapped out deaths
- Black bars show deaths
- Black circles (inside red circles) show water pumps

(image drawn and lithographed by Charles Cheffins)
Snow’s Hypothesis

- Snow observed that deaths were clustered around Broad Street Pump
- Other Evidence
  - Lion Brewery
  - Rupert Street Pump
  - Scattered houses, school kids, prison

(image drawn and lithographed by Charles Cheffins)
Further Evidence

- Two women died of Cholera in Hampstead which was not near Soho
- Snow learned that the two women used to live on Broad Street
- They liked the taste of the water from the Broad Street pump so they had it delivered to them everyday
Could Snow prove that water from the Broad Street Pump was causing cholera?

- Snow suspected that the water supply was a clue to understanding cholera outbreaks
- But he could not prove this hypothesis
- He applied a statistical method, called the method of comparison
- But to be clear, statistics can never be used to prove anything; they can only be used to argue that an outcome is unlikely
Method of Comparison

- Individuals comprise the two test groups
  - Treatment
  - Control
- **Treatment**: a process that the “treatment” group undergoes
- **Control**: a group similar to the treatment group in every way, except that it does not undergo the treatment
- **Outcome**: the results observed in the two groups
Method of Comparison

- Question: does the treatment have an effect on the outcome?
- If the outcome between the treatment and control differs significantly, we can conclude that there is an association.
- But does the treatment cause the outcome to occur?
Confounding Factors

- **Confounding factor**: an underlying difference between the treatment and control groups that is *not* the treatment

- **TVs and SAT scores**
  - More TVs in students’ homes correlates with higher SAT scores
  - Does watching more TV *cause* students to do well on the SATs?

- **Coffee and Lung Cancer**
  - People who have lung cancer also tend to drink coffee
  - Does drinking coffee *cause* lung cancer?
Snow’s “Grand Experiment”

- Lambeth water company vs. Southwark and Vauxhall (S&V)
  - S&V homes were the treatment group
  - Lambeth houses were the control group

- Snow argued there were no confounding factors, because there were no substantive differences between the populations to whom the two companies provided water

“Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies … there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded …”
### Data collected by John Snow

<table>
<thead>
<tr>
<th>Supply Area</th>
<th># of Houses</th>
<th>Cholera Deaths</th>
<th>Deaths/10,000 Houses</th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;V</td>
<td>40,046</td>
<td>1,263</td>
<td>315</td>
</tr>
<tr>
<td>Lambeth</td>
<td>26,107</td>
<td>98</td>
<td>37</td>
</tr>
<tr>
<td>Rest of London</td>
<td>256,423</td>
<td>1,422</td>
<td>59</td>
</tr>
</tbody>
</table>
If we assume that the water is not different between Lambeth and the rest of London, then the observed data might just be statistical noise.

If we assume that the water is not different between S&V and the rest of London, then it is very unlikely we’d see data so extreme!
Causality

Snow observed that the only difference between the area where people were getting sick and the area where they weren’t was that “one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients, the other group having water quite free from impurity.”

Snow convinced the authorities to remove the handle of the Broad Street pump. It was later determined that there was a leaking cesspool a few feet away from the Broad Street pump well, and sewage from the households of cholera victims was seeping into the well water.
Success!

In 1866 there was another outbreak of Cholera in the Limehouse district of London.
In 1854 (at about the same time Snow was drawing these maps in London), an Italian scientist named Filippo Pacini discovered a bacterium called *Vibrio cholerae* that enters the small intestine and causes cholera. But because of the popularity of the miasma theory, Pacini’s discovery received little to no immediate attention. In 1883, a German scientist named Robert Koch discovered *Vibrio cholerae* again, and the root cause of cholera was finally uncovered.
An Early Example of the Power of Data Visualization

- Snow’s data collection and data visualization saved many lives
- Snow pioneered the use of data visualization to map disease
- Nowadays, it is standard to use disease maps to track epidemics
- To this day, scientists at the Center for Disease Control (CDC) still use the phrase “Where is the handle to this pump?” when trying to uncover the cause of an epidemic
Florence Nightingale: Data Scientist
Nurse or Data Scientist?

While Florence Nightingale (1820-1910) is best known as the founder of modern nursing and worldwide health-reform, in fact, she was at heart a statistician, and a pioneer in data visualization.
The Crimean War

- Nightingale worked as a nurse during the Crimean War.
- She found the hospital to be highly unsanitary.
- Many soldiers were dying, but why?
- From battle wounds, or other causes, such as poor hygiene?
- Nightingale decided to collect data in attempt to settle this question.
- Her approach worked. By the time Nightingale left Crimea, the mortality rate in hospitals had dropped from 42% to 2%.
The Power of Visualization

Afraid Queen Victoria and Parliament would not read or understand her statistical report, Florence Nightingale created a graphical representation to convey her findings.
Polar Area Graph (Coxcomb)

- Each wedge represents a month from April 1854 to March 1856
- **Blue**: death by disease
- **Red**: death by wounds
- **Black**: death by other causes
Saving Lives

- Deaths are greater from April 1854 to March 1855 than the following year
- Deaths by disease are much greater
- A sanitary commission arrived from Britain in March 1855 to improve the sanitation of the hospitals
Extras
Map of a later cholera outbreak in London, in 1866
Reproduction of Nightingale’s Graph
The Statistical Inference Process, via Example

● Start by choosing a population to study, and some features of that population: e.g., citizens of London in the 1850’s

● Establish a baseline for that population: what is the probability of someone living in London then contracting cholera?

● Sample the population: Snow studied two samples
  ○ One, the control, provided a baseline distribution (.37% chance)
  ○ The treatment provided a second distribution (3.15% chance)

● Snow argued that it was extremely unlikely that the treatment sample was drawn from the same distribution/population as the control

● Statistics can never be used to PROVE anything; but they can be used to argue that an outcome is likely or unlikely
Population

- A population is a set of individuals under study, chosen by the statistician, or nowadays, the data scientist.
- When the entirety of the population is observable, we can compute descriptive statistics like population mean, population variance, histograms, etc.
- But typically, the entirety of the population is unobservable.
Sample

- A **sample** is an observed subset of the population
- A statistician’s goal is to study the sample, and from it, draw inferences about the entirety of the population
- **Polling** is a typical application of statistical inference:
  - E.g., ask a few Brown students about their favorite *anything*; then infer the favorite *anything* of most Brown students.