2.3: A brief history of (bio)statistics

Introduction

Before discussing achievements and landmark moments in biostatistics, let's start with basic definitions.

Bioinformatics is loosely defined as a discipline of biology primarily concerned with work involving large data sets (e.g., databases), but a bioinformatician would not primarily be a statistician necessarily. Rather, a bioinformatician, in addition to having a foundation in statistical and mathematical training, would likely be fluent in at least one programming language and confident in the use and design of databases.

Biostatistics, then, refers to use of statistics in biology. Biostatistics encompasses application of statistical approaches to design, analyze, and interpret biological data collected through observation of use of experimentation. In turn, there are many broad disciplines or fields of specialty that trained biostatisticians may work.

Chance, the likelihood that a particular event will occur.

Data scientist, a very general label, is a person likely to work on "big data." Big data may be loosely and inconsistently identified as access to large detailed and unstructured data sets such as visits and behavior within websites of tens of millions of Internet "hits" to a web site like Amazon® or Google®. The data scientist would then be involved in extracting meaning from volumes of this data in a process called **data mining**. In the context of biology, web sites like ALFRED at Yale University that houses allele frequency information collected on human populations, the 1000 genome project, or any of the databases accessible at National Center for Biotechnology Information would constitute sources of big data for biological researchers.

Epidemiology refers to the statistics of patterns of and risk of disease in populations, particularly of humans and thus, an epidemiologist would also be considered to be a biostatistician. The statistics of epidemiology include all of the materials we will cover in this course, but perhaps if any particular analytical approach characterizes epidemiology, it would be **survival analysis** .

Event, an outcome to which a probability is assigned.

Likelihood, the probable chances of occurrence of an event that has already occurred.

Probability, the chance that an event will occur in the future.

Random. This is a good place to share a warning about vocabulary; statistics, like most of science, uses familiar words, but with refined and sometimes different meanings than our every day usage. Consider our everyday use of "random": "without definite aim, direction, rule, or method – subjects chosen at random" (Merriam-Webster online dictionary). In statistics, however, "random" refers to "an assignment of a numerical value to each possible outcome of an event" (Wikipedia). Thus, in statistics, random dictates a method of determining how likely a subject is to be included: If N represents the size of the population, then **random sampling** implies that each individual had $1/N$ chance of being selected. Thus, if $N = 100$, then each individual has a 1% $(1/100)$ chance of selection. This is quite different from Merriam-Webster's definition, in which no method is assumed. To a statistician, then, "random" as used in everyday conversation would imply **haphazard sampling** or **convenience sampling** from a population.

Statistics may be defined as the science of collecting, organizing, and interpreting data. Statistics is a branch of applied mathematics. Note that the word **statistic** is also used, but refers to a calculated quantity like the mean or standard deviation. A little confusing, but the context in which statistics or statistic is appropriate is usually not a major issue.

Some notes about history

The concepts of chance and probability, so crucial to **statistical reasoning**, were realized rather late in the history of mathematics. While people have been writing about applied and theoretical math for thousands of years, probability as a topic of interest by scholars seems to date only back to the late 17th century, beginning with letters written between Pierre de Fermat (1601-1665) and Blaise Pascal (1623 – 1662) and the substantial work on probability by Pierre-Simon Laplace (1749-1827). Often, research on probability developed under the watchful eyes of rich patrons more interested in gaming than to scientific applications. Work on permutations and combinations, essential for an understanding of probability, trace to India prior to Pascal's work (Raju 2011).

The history of statistics goes back further if you allow for the dual use of the term "statistics", both as a descriptor of the act of collecting data and as a systematic approach to the analysis of data. Prior to the 1700s, statistics was used in the sense of collection of data for use by the governments. It is not until the latter part of the 19th century that we see scholarship on statistical analytical techniques. Many of the statistical approaches we teach and use today were developed in the decades between 1880s and the 1930s.

For example, see the work by Francis Galton, Karl Pearson, R. A. Fisher, Sewell Wright, Jerzy Neyman, and Egon Pearson (Karl Pearson's son).

Since the 1950s, there has been an explosion of developments in statistics, particularly as related to power of computers. These include use of resampling, simulation, and Monte Carlo methods (Harris 2010). **Resampling** — the creation of new new samples based on a set of observed data — in particular is a key innovation in statistics. Its use led to a number of innovative ways to estimate the precision of an estimate (see Chapter 3.4 and Chapter 19). **Monte Carlo methods**, or MCM, which involves resampling from a probability distribution, is used to repeat (simulate) an experiment over and over again (Kroese et al 2014). Computers have so influenced statistics that some now define statistics as "…the study of algorithms for data analysis" (p. 175, Beran 2003). For more on the history of statistics, see Anderson (1992), Fienberg (1992), and Freedman 1999; for excellent, conversational books read Salsburg (2002) and McGrayne (2011). For influential women in early development of statisticians, see Anderson (1992).

Epidemiology

John Snow (1813-1858) is credited by some as the "Father of Epidemiology"(Ramsay 2006). During a London outbreak of cholera in 1953, Snow conducted work to establish cholera mortality with source and quality of drinking water. At the time, the prevailing explanation for cholera was that it was an airborne infection. Snow's map of cholera mortality in the Golden Square district of London in relation to a water pump on Broad Street is shown in Fig. 2.3.1. Snow's theory of contaminated water was not accepted as an explanation for cholera until after his death.

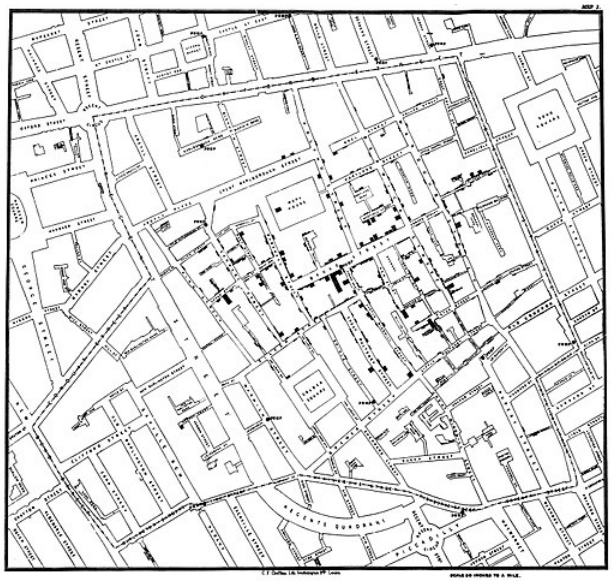

Figure 2.3.1: Original map by John Snow showing the clusters of cholera cases in the London epidemic of 1854, drawn and lithographed by Charles Cheffins. Image Public Domain, from Wikipedia

Snow's work and dataset can be viewed and thanks to Paul Lindman and others, the work can be expanded: for example, defining areas around pumps by walking distance (Fig. 2.3.2). The R package is `cholera` . Figure 2.6 shows a plot like Snow's annotated map.cholera.

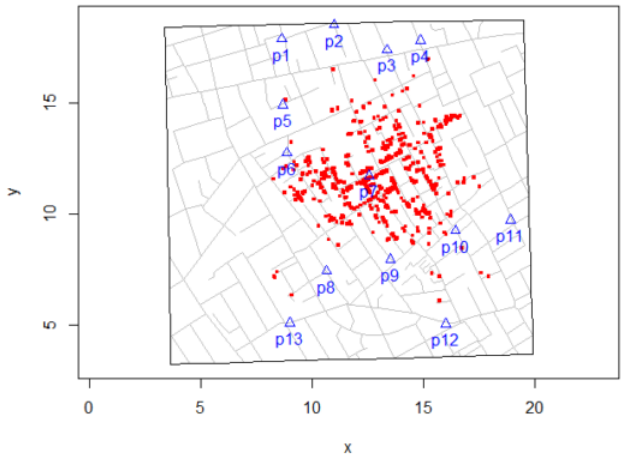

Figure 2.3.2: *Plot of Snow's London using R* `cholera` *package. Triangles marked with p1-p13 represent public water pumps. Red dots represent cholera cases.*

The R code to make the plot was

```
snowMap()
```

Snow's ideas about cholera were not accepted in his time and you should recognize that by itself, a cluster map supports both the airborne and waterborne theories. The `cholera` package contains additional data to help visualize the area, including setting regions by walking distance (Fig. 2.3.3).

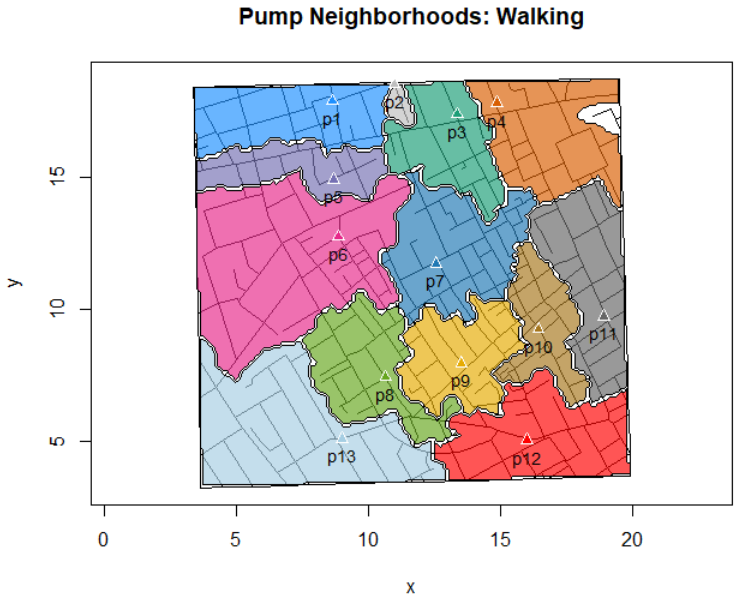

Figure 2.3.3: *Plot of Snow's London with walking areas drawn about the 13 water pumps. Created using R* `cholera` *package.*

R code to make the plot was

```
plot(neighborhoodWalking(case.set = "expected"), "area.polygons")
```

Epidemiology of cancer

This is an undeveloped section in my book. For now, please see Greenwald and Dunn (2009). Key landmarks in the history of epidemiology include

- Tobacco as a carcinogen
- Diet and cancer risk
- Obesity, exercise, and cancer risk
- Hormones and cancer risk
- Cancer risk and occupations: Ramazzini (1713), Pott (1775)

History of founders of statistics and eugenics

Many statistical methods in use today, including regression and analysis of variance methods, can trace their origins to the late 1800's and early 1900's (Kevles 1998). Many of these early statisticians developed statistical methods to further their interests in understanding differences between racial groups of humans. Sir Francis Galton, who developed regression and correlation concepts (the details and extensions of which were the works of Karl Pearson), also coined the term **Eugenics**, the "science" of improving humans through selective breeding. Sir R. A. Fisher, who invented analysis of variance and maximum likelihood techniques, and perhaps more importantly developed the concepts of sampling from populations, degrees of freedom, and his book *Statistical Methods for Research Workers*, is still relevant today.

Eugenics is still with us (click here to access the eugenics-watch website), but has been successfully and completely discredited on scientific grounds many times (click here for Eugenics Archive website). Do keep in mind that the times were different, but it is interesting nevertheless to learn a little about the murky history of statistics and the objectives of some of the very bright people responsible for many of the statistical analyses we use today (see Stephan J. Gould's "*The Mismeasure of Man*" at our Sullivan Library BF 431 G68 1981 or from Amazon.com; Gould, too, may be accused of some bias in his science — see NY Times article based on a PLOS Biology article). Here's an MIT web site with tremendous information about race in science).

Keep in mind also that statisticians were instrumental in showing why Eugenics was unscientific, at best. Here's a link to a non-peer reviewed article.

Questions

1. Find and copy definitions for "big data" and "data mining" from (a) one peer-reviewed, primary source (e.g., search Google Scholar), (b) one peer-reviewed, secondary source (e.g., search Google Scholar), and (c) Wikipedia. From these three sources, write your own definitions for big data processing and data mining.
