Chapter 4: Gathering Data


Florence Nightingale

Florence Nightingale was born into an upper class family in England, but devoted her life to trying to improve conditions for the sick and injured. She served in the British military hospital in Constantinople from 1854 to 1856. The hospital cared for casualties of the Crimean War, and her famous rose diagrams impressed the public and authorities back in England with the fact that there were many more casualties due to sickness and injury than due to actual combat. On returning to England, she established the first professional school of nursing at Saint Thomas’ Hospital in London. “Her practices brought tremendous respect to the field of nursing, and she made great strides in the reform of hospital sanitation. Intensely charismatic and inspirational, Florence Nightingale was an internationally influential figure.” (Contagion, Harvard University Library)

Design of studies

*Interpretation of variables:*

- explanatory variables
- response variables

In some studies, the values of some of the variables, the explanatory variables, might be thought to affect the values of some other variables, the response variables.

If the values taken on by a variable are unaffected by the values taken on by another variable, then the variables are independent; otherwise, the variables are dependent. Dependent variables might exhibit a linear relationship. If an approximating line has a positive slope, then a linear relationship is positive. If an approximating line has a negative slope, then the linear relationship is negative. Dependent variables are said to be associated. Independent variables are not associated.

*Simpson’s paradox.* Imagine a large cloud of data points drooping slightly to the southeast. Ah, one thinks, a negative relationship. But then it is discovered that the large cloud is comprised of two sub-clouds, one on top of the other, and both rising to the northeast. Within those two groups, the association is positive. Aggregating data reversed the apparent trend.

*A hierarchy of evidence:*

- anecdote
- observational evidence (sample of convenience, retrospective study, volunteers)
- experiment (controlled, randomized, double-blind)

In an observational study, the researcher observes the values of the explanatory and response variables in the study sample.

In an experiment, the researcher randomly assigns the subjects in the study sample to the treatments and then observes the values of the response variables.

Sampling methods include taking simple random samples, stratified samples, and cluster samples.

Blocking during random assignment is analogous to stratified sampling during random sampling.
Outline for design of studies: research question, sample, variables, conclusions


Inference

What is required to support inference from a sample to a population? The basic requirement is that the sample must be representative of the population. A formal structure to achieve this is a simple random sample (each subset of size \( n \) of the population is equally likely to be chosen as the sample). Thus, studies performed with volunteers, retrospective studies, and samples of convenience do not support inference from a sample to a population. Conclusions from such studies apply only to the participants in the study.

Why does a random sample have this property of faithfully representing the population it is drawn from? The answer to that key question comes from probability in the form of two famous theorems (the Law of Large Numbers and the Central Limit Theorem, LLN and CLT). Furthermore, random sampling permits measurements of the extent to which the sample may not represent the population (margins of error), and guarantees that the larger the sample the more faithful the representation (\( \sigma/\sqrt{n} \)). We will be studying these concepts in the next section of the course on probability.

What is required to support inference of causality (this causes that)? The basic requirement is that members of the treatment groups should differ only in their treatments. The formal requirement to support this is an experiment with random assignment of the study subjects to the treatment groups. The gold standard in medicine is the controlled, randomized, double-blind experiment.

A famous saying in statistics: correlation does not imply causation. Lurking variables may be present. Think of the correlation of drowning deaths and ice cream sales along the famous Gold Coast near Brisbane, Australia. Both of them happen mostly in the summertime, and neither one causes the other.

Random sampling supports inference from the sample to the population.

Random assignment (to treatment groups) supports inference of causality (this causes that).

Correlation does not imply causation.

Outline for design of studies: research question, sample, variables, conclusions

Outline for inference: contingency table displaying random sample vs. random assignment, and the type of inference which is supported or not supported in each case (Çetinkaya-Rundel lecture slides, Stat 2, Spring 2013, Duke University)

Constructing a random sample

It is, conceptually at least, easy to construct a random sample. Obtain a list, called the sampling frame, of the population to be sampled and number each of the candidates. Suppose, for instance, that our population consists of 100 individuals consecutively numbered from 1 to 100. We wish to select a random sample of 5 individuals. In R, evaluate the following two lines of code:

```r
population <- 1:100
sample(population, 5)
# 29 79 41 86 91
```

The default for R’s `sample` is to sample without replacement, so there will not be any repeated numbers in the sample. The random number generator in R is designed in such a way that any two subsets of 5 numbers are equally likely to be selected, so the selected individuals form a simple random sample of the sampled population.
Literary Digest

The Literary Digest was a popular magazine back in the 1920’s and 1930’s, but it suddenly became much too well known for the poll that changed polling. It conducted a large scale national poll attempting to predict the outcome of the Landon vs. FDR presidential election in 1936. The poll indicated that Landon would win “in a landslide!” ... but actually, Landon won only Maine and Vermont. Roosevelt won every other state. The poll had failed because of selection bias and nonresponse bias, and the magazine folded soon thereafter.

Survey design

Prof. Dennis DeTurk, Dean of Arts and Sciences at the University of Pennsylvania, posts an interesting series of short articles on aspects of survey design covering the Truman vs. Dewey US presidential election of 1948, the US census of 2000 and the Salk polio vaccine field trials of 1954 in addition to his analysis of the poll that changed polling mentioned above.

Arthur Bradford Hill and Richard Doll

Medical statistician Austin Bradford Hill (1897-1991) is credited with a major milestone in medicine: the first randomized, controlled clinical trials. These experiments, conducted in the 1950’s, had to do with proving the efficacy of streptomycin in the treatment of tuberculosis, which was all the more poignant because he himself had been a victim of tuberculosis as a young man. He went on to collaborate with epidemiologist Richard Doll (1912-2005) in identifying smoking as an important risk-factor in lung cancer and other diseases. The evidence in that area was all observational, not experimental, and part of their contribution was to establish the case-control method for extracting useful medical information from that sort of evidence.

Two Principles of Inference

Random samples from the population at large and random assignment to treatment and control groups are two key constructs in statistics.

<table>
<thead>
<tr>
<th></th>
<th>Sample is representative of the population</th>
<th>Conclusions apply to the ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>random sample</td>
<td>YES</td>
<td>population</td>
</tr>
<tr>
<td>NOT a random sample</td>
<td>NO</td>
<td>sample</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Treatment group is essentially identical to control group except for treatment</th>
<th>Can conclude causality (this causes that)</th>
</tr>
</thead>
<tbody>
<tr>
<td>random assignment</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>NO random assignment</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>

Samples which fail to be random samples include samples of convenience, self-selected samples, and samples with undercoverage or high non-response rates.

Random assignment characterizes experiments. If random assignment does not occur, then the study is an observational study.

Note that the emphatic “YES” in these tables masks some important nuances. For instance, the accuracy of estimation improves with increasing sample size.
Exercises

We will attempt to solve some of the following exercises as a community project in class today. Finish these solutions as homework exercises, write them up carefully and clearly, and hand them in at the beginning of class next Friday.

*Exercises for Chapter 4:* 4.2 (blood pressure), 4.6 (Nurses’ Health Study, Women’s Health Initiative, Hodis, New England Journal of Medicine, 2003; Women’s Health Initiative, JAMA, 2002), 4.20 (polls), 4.31 (bias), 4.37 (smoking), 4.42 (vitamin C), 4.43 (blood pressure), 4.53 (caffeine), 4.54 (allergy), 4.55 (smoking)

Class work 4a – experiments

*Exercises from Chapter 4:* 4.2 (blood pressure), 4.6 (Nurses’ Health Study, Women’s Health Initiative, Hodis, New England Journal of Medicine, 2003; Women’s Health Initiative, JAMA, 2002), 4.20 (polls), 4.31 (bias), 4.37 (smoking)

Class work 4b – experiments

*Exercises from Chapter 4:* 4.42 (vitamin C), 4.43 (blood pressure), 4.53 (caffeine), 4.54 (allergy), 4.55 (smoking)