Prospective, retrospective, and cross-sectional studies

Patrick Breheny

Study designs that can be analyzed with χ^2-tests

- One reason that χ^2-tests are so popular is that they can be used to analyze a wide variety of study designs
- In addition to controlled experiments, they are widely used in epidemiology, where investigators often conduct observational studies
- Broadly speaking, observational studies in epidemiology fall into three categories: *prospective studies*, *retrospective studies*, and *cross-sectional studies*

Study designs (cont'd)

- χ^2-tests and Fisher's exact test can be used to analyze all of these studies
- This is appealing as there is little risk of making of mistake: regardless of your study design, if it can be expressed as a 2x2 contingency table, a Fisher's exact test is always appropriate
- One caveat: there are a lot of tables in this world that have two rows and two columns; that does not necessarily make them contingency tables, so please don't try to perform χ^2-tests on them!

Prospective studies

- We have said that the double-blind, randomized controlled trial is the gold standard of biomedical research
- When this is not possible (or ethical), the prospective study (also called a *cohort study*) is the next best thing
- In a prospective study, investigators collect a sample, classify individuals in some way, and then wait to see if the individuals develop a condition
- The classification is usually based on exposure to a risk factor such as smoking or obesity

Risk factors for breast cancer

- For example, the CDC tracked 6,168 women in the hopes of finding risk factors that led to breast cancer
- One risk factor they looked at was the age at which the woman gave birth to her first child:

	Cancer	
	No	Yes
Before age 25	4475	65
25 or older	1597	31

Risk factors for breast cancer (cont'd)

- Performing a χ^2-test on the data, we obtain $p = .19$
- Thus, the evidence from this study is rather unconvincing as far as whether the risk of developing breast cancer depends on the age at which a woman gives birth to her first child
- In other words, although we observed that women who gave birth to their first child after age 25 were slightly more likely to develop breast cancer, the data is still consistent with the null hypothesis that the two groups have the same risk of developing breast cancer

Retrospective studies

- Not all researchers have the resources to follow thousands of people for decades to see if they develop a rare disease
- Instead, they often try the more feasible approach of collecting a sample of people with the condition of interest, a second sample of people without the condition of interest, and then ask them if they were exposed to a risk factor in the past
- For example, a much cheaper way to conduct the study of breast cancer risk factors would be to find 50 women with breast cancer, 50 women without breast cancer, and ask them when they had their first child
- This approach is called a *retrospective*, or *case-control* study

Fluoride poisoning in Alaska

- In 1992, an outbreak of illness occurred in an Alaskan community
- The CDC suspected fluoride poisoning from one of the town's water supplies

	Case	Control
Drank from supply	33	4
Didn't drink from supply	5	46

Fluoride poisoning in Alaska

- Testing whether this could be due to chance, the χ^2-test gives us $p = 6 \times 10^{-13}$
- The observed association was certainly not due to chance
- But the association still may be due to factors besides fluoride poisoning

Recall bias

- For example, people who got sick may think much harder about what they ate and drank than people who didn't
- This is called *recall bias*, and it is an important source of bias in retrospective studies
- How big is the problem? Depends on the study:
  - In the breast cancer example, it would not be much of a concern, since giving birth to a child is a major life event and a woman would know how old she was when it happened
  - On the other hand, if the risk factor was something like diet or exercise, recall bias would be a huge concern, as people are notoriously unreliable at recalling these things
- Furthermore, case-control studies are more prone to sampling bias than prospective studies

Electromagnetic field example

- For example, retrospective studies have been performed investigating links between childhood leukemia and exposure to electromagnetic fields (EMF)
- Families with low socioeconomic status are more likely to live near electromagnetic fields
- Families with low socioeconomic status are also less likely to participate in studies as controls
- Socioeconomic status does not affect the participation of cases, however (cases are usually eager to participate)
- This results in an observed association between EMF and leukemia potentially arising entirely due to bias

Cross-sectional studies

- The weakest type of observational study is the cross-sectional study
- In a cross-sectional study, the investigator simply gathers a single sample and cross-classifies them depending on whether they have the risk factor or not and whether they have the disease or not

Selection bias in cross-sectional studies

- Cross-sectional studies are the easiest design to carry out; however, because they don't pay any attention to time, they are weak at establishing cause and effect
- For example, suppose we obtain a cross-sectional sample of factory workers in the hopes of seeing whether they are more likely have asthma than non-factory workers
- This is a problematic design because workers who develop asthma from working in the factory are more likely to quit their job, and thus less likely to be included in our sample
- If we measure X and Y at the same time, then we don't know whether X caused Y, Y caused X, or whether both X and Y were caused by some third factor Z

Circulatory disease and respiratory disease

- As another example, one study surveyed 257 hospitalized individuals and determined whether each individual suffered from a disease of the respiratory system, a disease of the circulatory system, or both
- Their results:

	Respiratory	
Circulatory	Yes	No
Yes	7	29
No	13	208

- Could this association be due to chance?
- Not likely; $\chi^2 = 7.9$, so $p = 0.005$

Circulatory disease and respiratory disease (cont'd)

- Okay, so it's probably not due to chance
- But does that mean that you are more likely to get a respiratory disease if you have a circulatory disease?
- The same study surveyed nonhospitalized individuals as well:

	Respiratory	
Circulatory	Yes	No
Yes	15	142
No	189	2181

Circulatory disease and respiratory disease (cont'd)

- The evidence in favor of an association is now nonexistent: $p = 0.48$
- What's going on?
- Both samples were gathered carefully and are representative of their respective populations
- The problem, however, is that if a patient has both a circulatory disease and a respiratory disease, then he or she is much more likely to be hospitalized and therefore to be included in a sample of hospitalized patients

Summary

- In conclusion, prospective studies are the most trustworthy observational study, but like any observational study, they are subject to confounding
- Retrospective studies are often much more feasible, but potentially subject to recall bias and unrepresentative sampling
- Cross sectional studies provide a quick snapshot of an association, but need to be interpreted with care