Study Design and Contingency Tables

Ryan Miller

Grinnell College
Statistics

Introduction

Benzene is a compound used in chemical manufacturing that is partially responsible for the way gasoline smells. It is suspected that benzene might cause leukemia, a type of cancer.

- How would you design a study to evaluate the hypothesis that benzene causes leukemia?
- Which is more important for establishing support for a cause-effect relationship, the study's design or the type of statistical analysis used?

Study Design and Causation

Study design is the primary tool used to establish causation, which requires:

- The exposure must precede the outcome
- The exposure must be associated with the outcome
- All alternative explanations for the association must be ruled out
  - The exposed and unexposed groups must be similar in every meaningful way
  - Measurement (and other) biases must be eliminated

An observed association can only be interpreted as evidence of a causal relationship if these conditions are met

Randomized Experiments

Randomized experiments are the study design best suited for establishing causation:

- Because the researchers assign subjects to the experimental conditions, the exposure will always precede the outcome
- Random assignment ensures that all other variables that might be associated with the outcome are similarly distributed in each group
- Tools like *blinding* (preventing participants and/or researchers from knowing treatment assignments) and *placebo* (fake treatments that are indistinguishable from the real treatment) can help prevent bias

Thus, observed differences are more plausibly attributed to the treatment than an alternative explanation.

Randomized Experiments and Benzene

- From a purely research perspective, what would the ideal randomized experiment studying the effects of benzene on leukemia look like?
- Can you name at least two barriers to such an experiment ever happening in the real world?
- What is a more ethical study design that could fulfill at least some of the elements needed to establish causation?

Prospective (Cohort) Designs

A prospective study follows individuals forward in time:

1. Subjects are first categorized according to their exposure status
2. Researchers then observe whether each subject eventually experiences the outcome

- Because exposure is recorded before the outcome occurs, prospective studies generally provide strong evidence that the exposure precedes the outcome
  - However, exposed and unexposed groups may differ in important ways, so confounding remains a concern

Prospective Study of Benzene

Suppose researchers identify a cohort of 100,000 workers and classify each worker according to whether they are regularly exposed to benzene. The researchers follow these workers for 20 years and record who develops leukemia.

- Benzene exposure is relatively uncommon
- Leukemia is also a rare disease
- As a result, very large samples and long follow-up periods may be required to observe enough cases for meaningful statistical analysis

Case-Control Designs

A retrospective study looks backward in time, selecting individuals using their outcome status:

1. Researchers recruit groups of individuals who have already experienced the outcome (cases) and those who have not (controls)
2. Past exposure for the members of each group are then determined

- Retrospective studies are less expensive and more efficient for studying rare outcomes
  - However, they are more vulnerable to bias (recall, sampling, and others) than prospective studies

Case-Control Study of Benzene

Instead of following 100,000 workers for decades, researchers could:

- Recruit 500 individuals who have already developed leukemia (cases)
- Recruit 500 similar individuals without leukemia (controls)
- Determine each person's prior exposure to benzene

Researchers then compare the prevalence of previous benzene exposure between cases and controls.

- This design focuses directly on individuals with the rare outcome
- It is much faster and less expensive than a prospective study
- However, accurate reconstruction of past exposures can be challenging

Cross-sectional Designs

A cross-sectional study observes a sample at a single point in time:

- Exposure and outcome are measured simultaneously
  - Thus, it generally cannot be determined whether the exposure preceded the outcome
  - Selection bias is also a major concern
- However, cross-sectional studies are inexpensive to conduct and can identify associations that might warrant further study using a better design

Cross-sectional Study of Benzene

Researchers survey workers at several chemical plants during a single year and record their current benzene exposure and current leukemia diagnosis status

- It is difficult to determine whether the exposure preceded the disease
  - Some workers might have been diagnosed with leukemia before starting their current job
- Individuals who become ill due to benzene may have already left these employers, creating selection bias

Study Design Hierarchy

Randomized Experiment

↓

Prospective (Cohort) Study

↓

Case-Control Study

↓

Cross-Sectional Study

Moving down the hierarchy generally means: Less ability to establish causation, but lower cost and faster completion

Effect Size for Rare Outcomes and Exposures

Consider a cohort study that follows 100,000 workers for 20 years:

	Leukemia	No Leukemia
Benzene Exposure	20	980
No Exposure	180	98,820

- Is there *statistical evidence* of an association? If so, how large is the effect?

Relative Risk

In our example, the absolute difference in risk was only:

$$\text{Risk Difference} = \frac{20}{1000} - \frac{180}{99000} = 0.018$$

Since leukemia is rare, the risk difference *understates the strength of association*. For rare outcomes it is better to use **relative risk**:

$$RR = \frac{\hat{p}_{\text{event|exposed}}}{\hat{p}_{\text{event|unexposed}}}$$

Using the benzene data:

$$RR = \frac{20/1000}{180/99000} = 11$$

Relative Risk and Case-Control Studies

Now suppose researchers conduct a case-control study:

	Leukemia	No Leukemia
Benzene Exposure	80	40
No Exposure	120	160

- Can we calculate the risk of leukemia among exposed workers? Among non-exposed workers?
- Suppose the researchers collect data on an additional 100 cases (using the same methodology), should this substantially change the relative risk?

Odds and Odds Ratios

- Because the researchers can determine the relative balance of cases and controls, conditional probabilities and relative risk cannot be estimated from a case control study
  - Thus, an alternative measure, the **odds ratio** is often used
- In statistics, *odds* are the ratio of how often an event occurs relative to how often it does not
  - Odds of 1 reflect an event with a 50% probability
  - Odds of 3 reflect an event with a 75% probability
- An odds ratio is simply a ratio of the odds calculated for separate groups (exposed and unexposed)

Odds and Odds Ratios

Using the benzene case-control study:

- The odds of leukemia among the exposed are $80/40 = 2$
- The odds among the unexposed are $120/160 = 0.75$

Thus, the odds ratio is:

$$OR = \frac{2}{0.75} = 2.67$$

- So, the odds of leukemia among exposed are 2.67 times higher than the odds among unexposed
- Note that an odds ratio of 1.0 reflects no association

Odds Ratios in R

We can use the `fisher.test()` function to compute odds ratios and the corresponding confidence interval estimates:

```
my_data = data.frame(cases = c(80, 120),
                     controls = c(40, 160))
fisher.test(my_data, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  my_data
## p-value = 1.844e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.667842 4.290508
## sample estimates:
## odds ratio
##   2.660011
```

Practice #1

A study surveyed 257 hospitalized individuals, classifying whether suffered from a circulatory disease, a respiratory disease, both, or neither. The results are displayed below:

	Respiratory Disease	No Respiratory Disease
Circulatory Disease	7	29
No Circulatory Disease	13	208

1. Conduct a statistical analysis to assess whether the observed association between presence of a circulatory disease and presence of a respiratory disease could be due to chance (sampling variability)
2. Describe the design of this study. Considering this design, do you believe this study supports causation? If not, what else might explain the relationship?

Practice #2

In 1980, the New England Journal of Medicine published a randomized controlled trial where subjects were assigned to receive clofibrate, a cholesterol lowering drug, or a placebo pill. Below are results for those who took the drug vs. those who didn't:

	Deaths	Survivors
Took Clofibrate	708	4012
Didn't Take Clofibrate	357	1071

1. Does this study provide statistical evidence of a difference between those who took the drug and those who didn't?
2. Describe the design of this study. Considering this design, do you believe the study supports causation? If not, what else might explain the relationship?

Practice #2 - Adherence to Randomization

The subjects were *not randomized* with respect to their *adherence*, so the groups shown in the previous table might differ in meaningful ways:

	Deaths (trt)	Percent	Deaths (ctrl)	Percent
Adhered	708	15%	1813	15%
Didn't Adhere	357	25%	882	28%
Total	1065	18%	2695	19%

If we'd like to make a causal conclusion, we have to compare the groups as they were randomized, which are not significantly different

Conclusion

For our next exam you are expected to be familiar with:

1. The importance of study design, and how various features of a study's design influence the conclusions that can be drawn from the sample data.
2. How to measure association in contingency tables (two-way frequency tables) using risk differences, relative risks, and odds ratios. This includes knowing how/why these measures are used, and how they are interpreted.