Measuring Association in Observational Studies

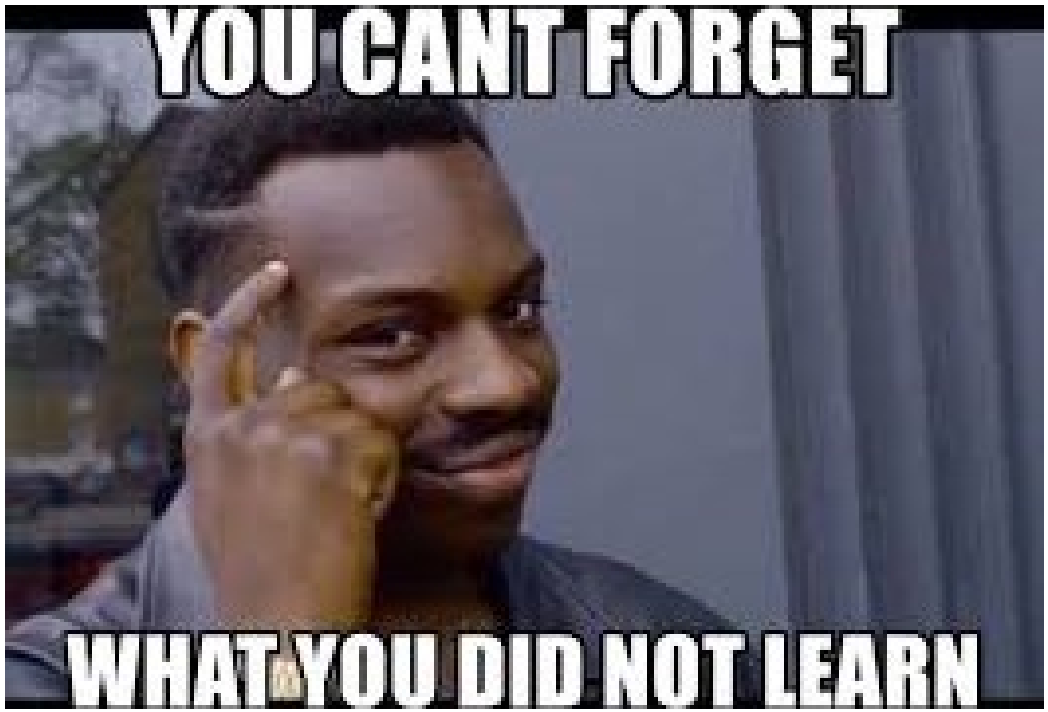
Javier E. Flores

April 8, 2019

Introduction

Way back when, we discussed two fundamental study designs: experimental and observational studies.

Recall that randomized controlled experiments are considered the "gold standard", but when they aren't feasible (due to ethics, cost, etc.), observational studies are often performed.



Introduction

The focus of this lecture will be on different kinds of observational studies, the analytic implications of each, and some additional measures of association.

The three types of observation studies we will discuss are:

- **Prospective Studies**
- **Retrospective Studies**
- **Cross-sectional Studies**

While we can apply some of our existing methods (namely χ^2 tests) for the analysis of data resulting from each type, there are some important considerations we should keep in mind for each design.

Prospective Studies

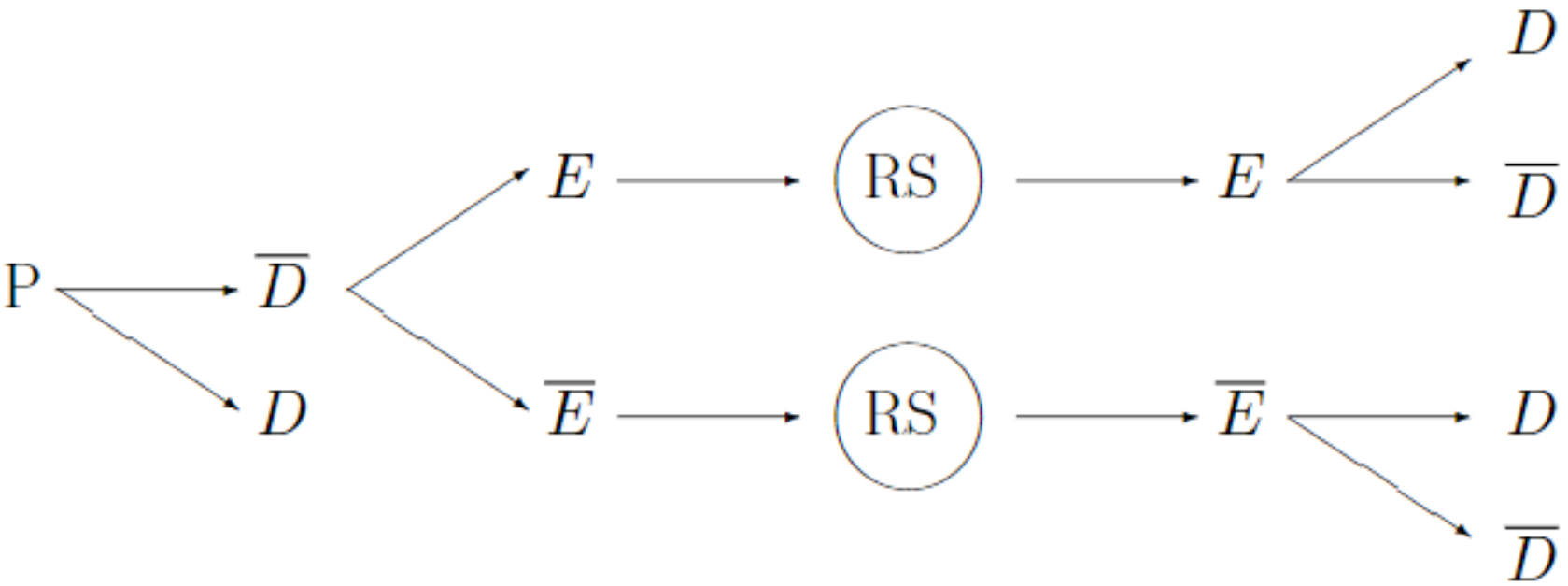
The first design we will discuss is a **prospective**, or **cohort**, study. This design is usually the next preferred when randomized controlled experiments are not an option.

In prospective studies, researchers obtain a representative sample of a population and follow the subjects forward in time.

Upon recruitment, individuals are classified into groups based on exposure to some risk factor.

As time progresses, researchers observe whether an outcome of interest manifests in each exposure group.

Prospective Study Flowchart



Prospective Study Example

In a study to determine the risk factors for breast cancer, CDC researchers recruited 6,168 women (all born in the 1960's) without breast cancer and followed them over time.

One potential risk factor of interest studied was the age at which each woman gave birth to their first child:

	Didn't Develop Cancer	**Developed Breast Cancer**
Child Before Age 25	4475	65
Child After Age 25	1157	31

1) With your groups, discuss whether you feel these data may be used to estimate the proportion of women in the population that develops breast cancer.
2) Determine whether these data provide evidence demonstrating that the age at which a women has their first child is a risk factor for breast cancer.

Solution

These data can be used to estimate the proportion of women in the population that develops breast cancer ($\frac{65+31}{65+31+4475+1157} = 1.6\%$). Since this was a prospective study, the data were obtained from a representative sample of cancer-free women.

Following this representative sample over time and tracking which develop breast cancer should allow us to draw inference on the the proportion that develop the disease.

To determine whether these data indicate first childbirth age as a risk factor, we can perform a χ^2 test for association.

In doing so, we obtain $\chi^2 = 7.8$ and a p-value of 0.005. This indicates that there is an association between the age of first child birth and breast cancer. However, since this prospective study is observational, this association may be attributed to one or more confounding factors.

Retrospective Studies

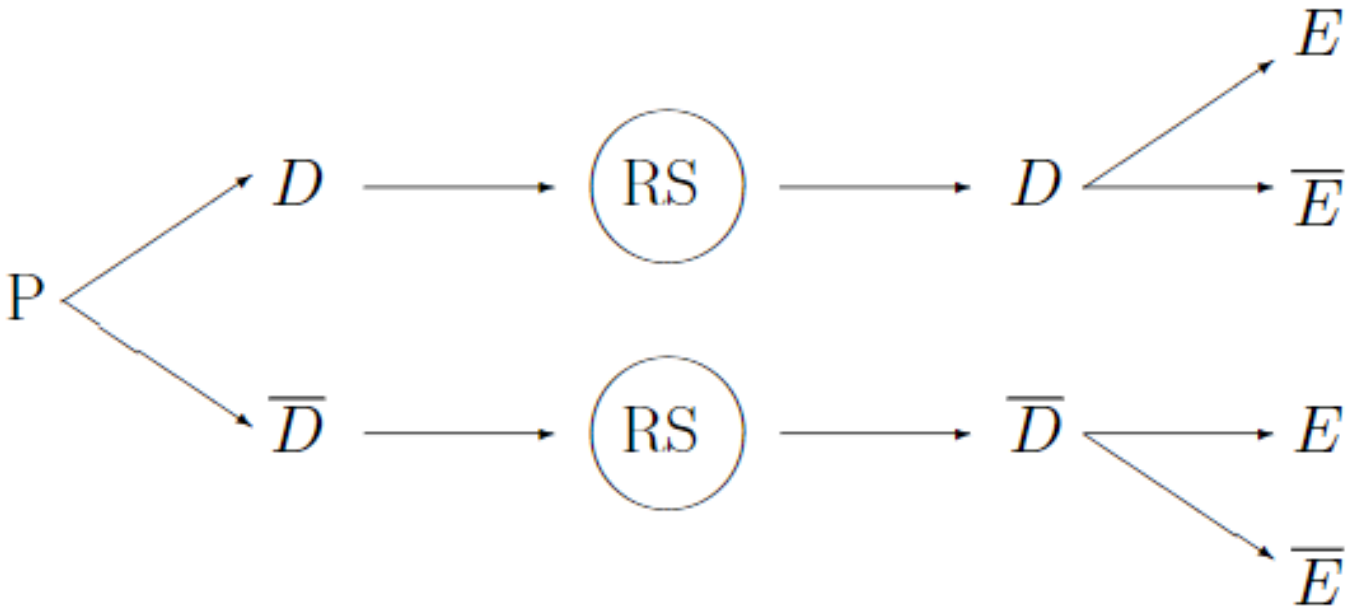
As you might imagine, tracking cohorts of individuals (especially when those cohorts are very large) over a long period of time comes at great cost financially and in terms of time invested.

For these reasons researchers sometimes consider performing **restrospective**, or **case-control**, studies.

In retrospective studies, researchers first randomly sample from two separate populations: individuals who have experience some outcome of interest, and those who have not.

After obtaining samples from these two populations, researchers look at each subject's history to determine whether or not they were exposed to some risk factor of interest.

Retrospective Study Flowchart



Retrospective Studies

While case-control studies are cheap and easy to conduct, they are much more prone to sampling biases relative to prospective studies.

In contrast to prospective studies, retrospective studies recruit from two separate populations - a "case" population and "control" population - which increases suspicions of confounding in the event that an association between some risk factor and the outcome is found.

Additionally, recall bias is often a concern. The ability of an individual to accurately recall important information necessary to assess exposure is often not very good.

Retrospective Study Example

In 1986, a retrospective study was conducted in order to determine whether a relationship between smoking and oral cancer was present.

Researchers sampled 304 individuals with oral cancer (cases) and 139 without (controls) and assessed their smoking frequency:

	Cases	**Controls**
$<$ 16 **cigarettes per day**	49	46
$\geq$ 16 **cigarettes per day**	255	93

1) Estimate the **prevalence** of oral cancer (i.e. the proportion in the population who have oral cancer). Do you believe this estimate is accurate? Why or why not?

2) Determine whether these data provide evidence of an association between smoking and oral cancer.

Solution

The estimated prevalence is $\frac{49+255}{49+255+46+93} = 68.6\%$. We should not trust this estimate or believe it accurately reflects the true prevalence.

By design, a fixed amount of cases and controls were sampled. For this reason, we cannot say that the proportion of cases in our total sample is representative of the population prevalence.

Despite the problems with estimating the prevalence, we can still rely on a χ^2 test to determine the presence of an association.

In doing so, we obtain $\chi^2 = 16.3$ and a p-value of nearly 0. This provides evidence of an association, but we still must be cautious of potential biases and confounding.

Cross-sectional Studies

The last observational study type we will discuss is the **cross-sectional** study.

In cross-sectional studies, researchers obtain a random sample of individuals and cross-classify them based upon who has been exposed to some risk factor and who has experienced some outcome.

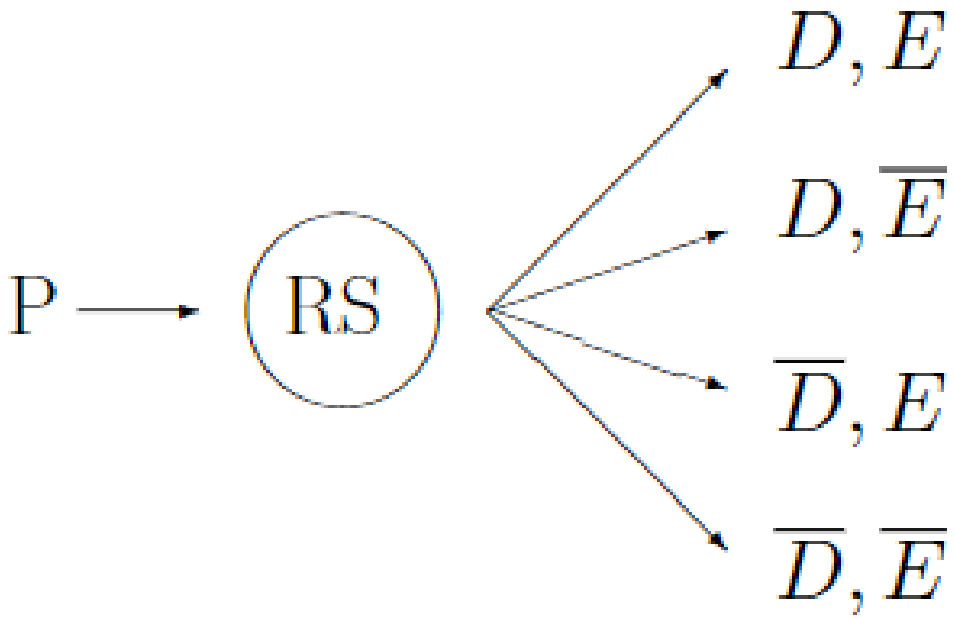
This sample is assessed at a single "cross-section" in time and not followed beyond this.

Cross-sectional studies are largely considered the weakest design given its high potential for confounding.

As an example, consider a study interested in determining whether working in a certain factory was associated with the development of asthma.

Workers who developed asthma are likely to quit their job at the factory and not be included in a cross-sectional sample.

Cross-sectional Study Flowchart



Measuring Association

For each of the types of observational studies discussed, if interest is in finding an association, performing a χ^2 test is a reliable option.

As we learned when first discussing χ^2 tests, this test provides a measure of evidence against the null hypothesis of independence (i.e. two variables are unrelated).

However this test is limited in that it tell us nothing about the *strength* of the association (i.e. effect size), only whether or not one exists.

One way we might think to characterize the effect size is through a difference in proportion, but this metric also has limitations in that it isn't necessarily applicable across all design types.

An additional limitation of the difference in proportion, or **risk difference**, is best illustrated through the following example...

Absolute and Relative Risks

Suppose that, over a ten year period, it was estimated that smokers have a 0.483% chance of developing lung cancer, while non-smokers were estimated to have a 0.045% chance of developing lung cancer over the same period.

Computing a risk difference, we obtain $0.00483 - 0.00045 = 0.4\%$ difference! Tiny!

Using this metric to characterize the impact of smoking on developing lung cancer would lead one to believe that smoking hardly makes a difference!

For this reason it is often preferred to find a **relative risk**, which is a ratio of prevalences.

This is particularly true in scenarios where each prevalence is extremely small. In this example, relative risk is 0.483/0.045=10.7. This suggests that smokers are 10.7 times as likely as non-smokers to develop lung cancer.

Absolute and Relative Risks

The relative risk does not always tell a substantially different story than the absolute risk. In the previous example, we saw this disparity primarily because of the magnitude of each proportion being compared.

An example of the contrary is the relationship between smoking and coronary artery disease (CAD).

Over a period of ten years, smokers are estimated to have a 2.947% chance of developing CAD while non-smokers are estimated to have a 1.695% chance.

Here, the risk difference is 1.25% while the relative risk is 1.7 - both of which are similar in magnitude.

Practice

A well-known case-control study published in 1969 examined the relationship between oral contraceptive use (OC) and the risk of blood clots. Data from this study is summarized in the table below:

	Clot	**No Clot**
No OC	42	145
OC	42	23

1) Find the absolute and relative risk of blood clotting between those who use oral contraceptives and those who don't.

2) Interpret these metrics and discuss their reliability.

Solution

The absolute risk is $\frac{42}{42+23} - \frac{42}{42+145} = 0.42$, or 42%. Individuals who use contraceptives are 42% more likely to develop a blood clot than those who don't.

The relative risk is $0.65/0.22 = 2.88$. Individuals who use contraceptives are 2.88 time more likely to develop a blood clot than those who don't.

Since these data are from a case-control study, the amount of cases and controls are fixed. Therefore, we can not rely on any kind of estimate involving a prevalence. This includes the risk difference and relative risk.

Bottom Line: Do not compute risk differences or relative risks for case-control (retrospective) studies.

Odds Ratio

Instead, compute the **odds ratio** when trying to quantify the strength of an association in a retrospective study.

Just as the relative risk is a ratio of two risks, the odds ratio is...you guessed it...the ratio of two **odds**!

As any gambler could tell you, the odds of an event is the number of times that an event occurs relative to the number of times that it doesn't.

Suppose the probability of an event occurring is a. Then,

$$\text{Odds} = \frac{a}{1-a}.$$

If the probability of an event is 75%, then Odds $= \frac{.75}{.25} = 3$. We may say that the odds of this event are "3 to 1".

Practice

	Clot	No Clot
No OC	42	145
OC	42	23

1) Compute the odds of developing a blood clot if you use oral contraceptives.

2) Compute the odds of developing a blood clot if you do not use oral contraceptives.

3) Use these odds to determine the odds ratio for the risk of blood clots given oral contraceptive use.

4) Now find the odds of using oral contraceptives if you've developed a blood clot, find these same odds if you haven't developed a blood clot, and find the odds ratio for oral contraceptive use given that you have a blood clot. How does this odds ratio compare to the odds ratio you found previously?

Solution

1) Odds of developing a blood clot if you use oral contraceptives: $(42/23) = 1.83$.

2) Odds of developing a blood clot if you do not use oral contraceptives: $(45/145) = 0.29$.

3) Odds ratio for the risk of blood clots given oral contraceptive use: $1.83/0.29 = 6.31$. Using OC increases the odds of blood clots by a factor of 6.31.

4a) Odds of using oral contraceptives if you have a blood clot: $(42/42) = 1$.

4b) Odds of using oral contraceptives if you do not have a blood clot: $(23/145) = 0.1586$.

4c) Odds ratio for the risk of blood clots given oral contraceptive use: $1/0.1586 = 6.31$.

The odds ratios are the same! This means that the odds ratio is symmetric!

Odds Ratios

A quicker way to compute the odds ratio is through the "cross-product" method. Suppose you have the following 2x2 table:

	Outcome	**No Outcome**
Exposure	a	b
No Exposure	c	d

The odds ratio for the risk of outcome given the exposure is:

$$\hat{OR} = \frac{ad}{bc}$$

In our previous example, the "exposure" row was the second row rather than the first. If we followed the above formula exactly, we would be computing the odds ratio for blood clots given *no* OC use: $\frac{42*23}{42*145} = 0.16$. The odds of blood clotting are decreased by 84% for those who do not use OC.

If we wanted to compute the odds ratio for the blood cloths given OC use, we could either swap the ordering of the rows and recompute the formula or just compute $1/0.16 = 6.31$.

Odds Ratios and Relative Risk

While odds ratios can be applied in any study design, they are often more difficult to interpret. Gaining a practical understanding of odds is much more difficult than understanding probabilities or risks.

For this reason, it is sometimes preferred to compute relative risks when possible.

When relative risks can not be computed, such as for case-control studies, there are instances in which computing the odds ratio will provide a good approximation to the relative risk.

If it is known beforehand that the overall prevalence of your outcome is very small, the odds ratio will be approximately equal to the relative risk.

Summary

Despite being somewhat difficult to interpret, odds ratios are a popular measure of association for categorical data.

In contrast to risk ratios (i.e. relative risks), odds ratios are symmetric and can be used across all types of observational studies.

Additionally, odds ratios derive popularity from their relationship with **logistic regression**, which is a regression method for modeling binary categorical outcomes.

While we won't discuss logistic regression in this class, it is widely considered to be part of the "bread and butter" for your average statistician.

Wrap-Up

Right now, you should...

- Be able to differentiate between and identify each observational study type.
- Know the various measures of association discussed, when they are appropriately used, and how they should be interpreted.

These notes are not covered by the textbook, but are still essential concepts and in theme with the material from the previous chapter. Please do not disregard these concepts as you will be expected to know and understand them.