

Observational Studies and Experiments

The goal of a **comparative study** is to estimate the difference in some characteristic between groups, or to assess how one characteristic varies across the levels of another characteristic that is continuous.

Here are some examples of comparative studies:

Comparisons within populations A specific medical treatment is compared to giving no treatment, for a defined population of subjects; or a new treatment is compared to the current standard treatment.

Comparisons over time The number of US highway fatalities in 2000-2005 is compared to the number of US highway fatalities in 1995-1999.

Comparisons across populations The number of highway fatalities per kilometer driven in the US is compared to the number of highway fatalities per kilometer driven in Canada.

Comparisons across levels of a continuous factor Rates of home foreclosure are compared based on the amount of downpayment paid.

Non-comparative studies

A non-comparative study typically aims to estimate some quantity, e.g. the proportion of US children without health insurance – but the estimate is not immediately compared to anything.

Terminology

The characteristic being compared (i.e. successful treatment response, traffic fatalities, home foreclosure) is called the **outcome variable** (sometimes the **response variable** or **dependent variable**).

We will call the groups being compared the **treatment groups**, even though the “treatment” may be a natural phenomenon or social construct (e.g. when comparing the US to Canada).

We will call a continuous treatment variable such as the amount of downpayment made on a home the **treatment factor**.

Sometimes the treatment group or factor is called the **independent variable**.

Any additional variables measured at or before the time of treatment are called **covariates**. Any variables measured after treatment are **outcomes**.

Goal of a comparative study

Comparative studies are useful for identifying **associations**, e.g.:

- People with persistent anxiety who took medication A had a greater reduction in anxiety symptoms than people who took medication B.
- Countries with higher traffic density on their roads have higher traffic fatality rates per kilometer driven than countries with lower traffic density.
- The rate of home foreclosure among people who paid a higher downpayment on their home is lower than among people who paid a lower downpayment.

Only certain types of comparative studies can be used to identify factors that **cause** changes in the outcome.

Observational studies

A comparative study is **observational** if treatment group assignments are not imposed by researchers following a study design or protocol.

Here are some examples of observational studies:

- People who took either generic or brand name Tylenol are compared in terms of their pain relief.
- Counties in eastern Iowa that that received either high or low rainfall in a given year are compared in terms of their corn yields.
- Third grade students who followed either a phonics or a whole-language based reading curriculum are compared in terms of their standardized reading test scores.

Controlled Experiments

A controlled experiment is a type of comparative study that aims to isolate the effect of a single factor, by minimizing or eliminating the possible effects of all other factors.

One type of controlled experiment is a **randomized controlled experiment**, in which units are randomly assigned to treatment groups.

For example, in order to determine whether a new chemotherapeutic drug A has a greater effect than the current standard drug B, we could randomly assign patients to be treated by one or the other of the two drugs and compare the response rates.

We will discuss other forms of control later.

Causality

When analyzing a comparative study, many people want to draw **causal** conclusions. A causal conclusion makes a statement about what would be expected to happen under an intervention.

For example based on observed associations, the following interventions might be proposed:

- If anti-anxiety medication A appears to be more effective than anti-anxiety medication B, a national medical organization may call for doctors to preferentially prescribe medication A.
- If higher traffic density is associated with higher traffic fatalities, a national government might propose to build more roads.
- If smaller downpayments on home loans is associated with greater default rates, government regulators may encourage lenders to require greater downpayments.

Causality

For an intervention to perform as expected, it is necessary that the association used to justify the intervention be causal.

If anti-anxiety medication A causes a reduction in anxiety symptoms, or if traffic density causes traffic fatalities, or if small home downpayments cause loan defaults, these interventions will be successful at reducing the occurrence of something bad (anxiety, traffic fatalities, foreclosures).

However for many observed associations, the variation in outcomes is wholly or partly caused by different factor than the treatment called a **confounder**.

Confounders

A confounder (or lurking variable) is a variable that is associated with both treatment group status and the outcome variable.

Suppose we observe an association between treatment group status and outcomes. If confounders are present, we cannot state with any certainty that the treatment causes changes in the outcome, since it is also possible that any of the confounders causes the changes.

Here are possible confounding variables in the examples discussed earlier:

- Suppose anti-anxiety medication A produces a side effect that people with the most severe anxiety are least able to tolerate, leading some of them to switch to medication B.
- Suppose traffic density is negatively associated with enforcement of traffic laws.
- Suppose companies that allow low loan downpayments tend to operate in states with strong laws protecting assets from creditors.

Confounders

In each of the three examples, it may be the confounder rather than the treatment effect that causes changes in the outcome:

- Encouraging physicians to preferentially prescribe medication A may not lead to an overall reduction in the number of people with persistent anxiety (it may even make it higher if some people choose to be untreated after experiencing the side effects of medication A).
- Building more roads may not reduce traffic fatalities (it may even make things worse if it makes enforcing traffic laws more difficult or expensive).
- Requiring higher home downpayments may not reduce home loan defaults.

Randomization and balance

Randomization guarantees that potential confounders are statistically independent of treatment group assignment.

Balance means that a potential confounder has identical sample distributions within all treatment groups. For example, suppose patients with alcoholism are given zero, low, and high doses of a drug to reduce craving for alcohol. The treatment groups would be balanced for smoking status (a potential confounder) if the fraction of smokers is the same in each treatment group.

The study on the left is balanced, the study on the right is not:

	Dose		
	High	Low	Zero
Smokers	30	50	20
Non-smokers	30	50	20

	Dose		
	High	Low	Zero
Smokers	30	50	10
Non-smokers	50	50	20

Randomization and balance

Randomization ensures that potential confounders are balanced on average across treatment groups. However especially for small samples, confounding factors can fail to be balanced by chance.

If we assigned treatments at random, we might get something like this:

		Dose	
	High	Low	Zero
Smokers	33	54	19
Non-smokers	27	46	21

The benefit of randomization is that we achieve approximate balance with respect to unknown confounders.

Bias due to confounders

Suppose we aim to compare the mean level of an outcome variable Y between treated ($T = 1$) and untreated ($T = 0$) subjects. Suppose smoking status is a confounder that also influences the outcome ($S = 1$ is a smoker and $S = 0$ is a non-smoker).

Suppose an additive model holds, where

$$E(Y|T, S) = \alpha_T + \beta_S$$

and $\beta_0 = 0$ for identification (otherwise we could add a constant to the α 's and subtract the same constant from the β 's and get the same model).

For non-smokers, the treatment effect is

$$E(Y|T = 1, S = 0) - E(Y|T = 0, S = 0) = \alpha_1 - \alpha_0$$

Bias due to confounders (continued)

Smokers have the same treatment effect:

$$E(Y|T = 1, S = 1) - E(Y|T = 0, S = 1) = \alpha_1 + \beta_1 - (\alpha_0 + \beta_1) = \alpha_1 - \alpha_0.$$

If we don't control for smoking, the mean response among treated subjects is

$$\begin{aligned} E(Y|T = 1) &= E_S E(Y|T = 1) \\ &= P(S = 1|T = 1)E(Y|T = 1, S = 1) + P(S = 0|T = 1)E(Y|T = 1, S = 0) \\ &= p_1(\alpha_1 + \beta_1) + (1 - p_1)\alpha_1 \\ &= \alpha_1 + p_1\beta_1, \end{aligned}$$

where p_1 is the proportion of smokers in the treatment group.

Similarly, the mean response in the untreated subjects is

$$E(Y|T = 0) = \alpha_0 + p_0\beta_1,$$

where p_0 is the proportion of smokers in the control group.

The mean difference between treatment responses in a study that does not control for smoking status is therefore

$$E(Y|T = 1) - E(Y|T = 0) = \alpha_1 - \alpha_0 + \beta_1(p_1 - p_0).$$

If smoking rates are not the same in the two treatment groups ($p_0 \neq p_1$), the estimate of treatment response is biased.

Counterfactuals and control groups

The ideal comparison for a medical treatment is a **counterfactual**, meaning the outcomes for a set of treated people are compared to the outcomes the same people would have had, had they not been treated.

In this ideal experiment, there can be no confounders.

In practice a unit cannot be both treated and untreated. Thus to estimate the treatment effect we must compare treated units to untreated units that were most like the treated unit at baseline (and similarly for untreated units).

Known/unknown and measured/unmeasured confounders

Known confounders are potential confounding factors that can be identified, for example, a medical treatment may have different rates of success in women and men, or in people with different diets.

An **unknown confounder** is something that we have not recognized as being a potential confounding variable.

Measured confounders are potential confounding factors that we can measure either perfectly (e.g. gender) or partially (e.g. diet).

All unknown confounders are unmeasured, but a known confounder may be impossible to measure in practice, e.g. if we are looking at cancer incidence in older people it is very unlikely that we would have any useful information about their mothers' diets during pregnancy.

Balanced randomization

In small or moderate samples, confounders may be unbalanced by chance even if randomization is performed.

We can force known confounders to be balanced by using **balanced randomization**.

A simple example would be if gender was thought to be a confounder.

If we have 50 men and 50 women in our study, and we assign treatments at random (say by flipping a coin), we will on average have 25 treated men and 25 treated women, but these numbers will fluctuate by chance.

Under balanced randomization, we would randomly sample exactly 25 men (without replacement from the 50 men) and 25 women (without replacement from the 50 women) to be treated. This ensures exact balance for gender.

Balanced randomization is useful for exactly balancing known, measured confounders. For unknown and/or unmeasured confounders, we must rely on randomization to provide balance on average (but cannot expect to achieve exact balance).

Strategies for analyzing observational studies

Suppose we have data from an observational study in which known, measured confounders were not controlled.

Is there anything we can do to provide some sense of what we should expect from a better controlled study?

Common support

Suppose we have an outcome variable Y , a treatment variable T , and a potential confounder X .

The **support** of X for treated subjects is the range of X values that occur for treated subjects ($T = 1$). Similarly, the support of X for untreated subjects is the range of X values that occurs for untreated subjects ($T = 0$).

If the supports of X for $T = 0$ and $T = 1$ are identical, the study is perfectly balanced for X and we do not need to worry about X being a confounder.

But if the study did not control for X , it is unlikely that the supports are identical.

Common support

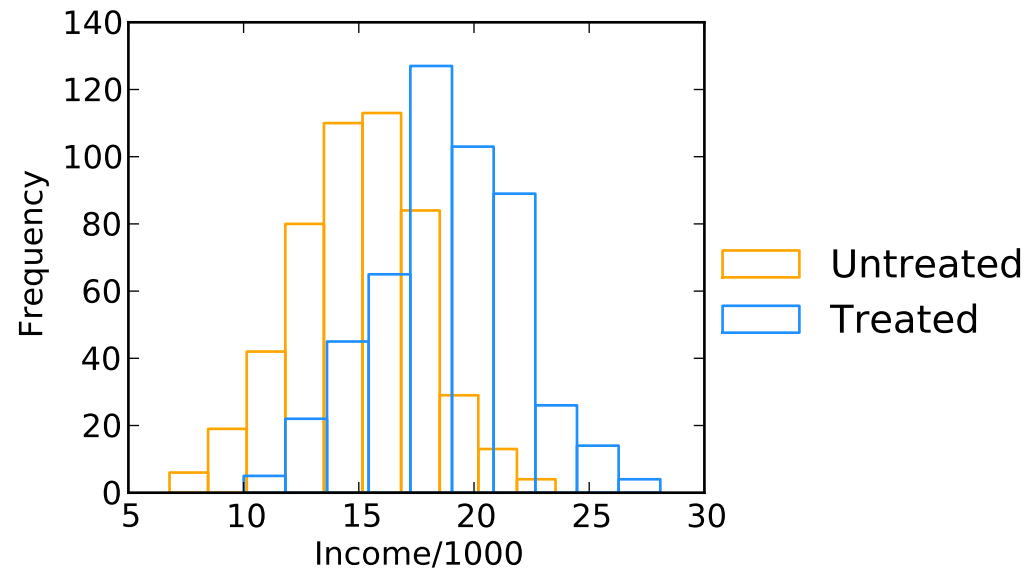
As an example, suppose we are interested in estimating the effect of a special program for high risk pre-schoolers. It is hoped that the program improves the childrens' reading scores in the third grade.

Suppose the program is made available to all children in a particular neighborhood. Some of the children are enrolled in the program and some are not by their parents' choice. Enrollment in the program is the treatment variable for this analysis.

Another major factor that is associated with the reading score is the parents' socioeconomic status (SES). For simplicity suppose this is measured by the net family income.

Common support

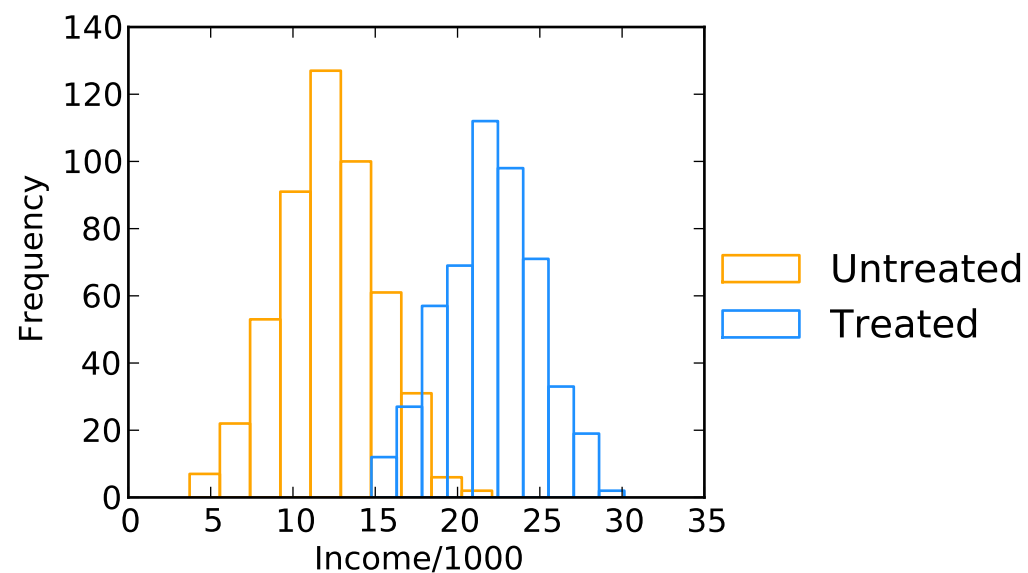
It is unlikely that the treatment and non-treatment groups will be perfectly balanced for parent income. For example, the two income distributions may look like this:



In this case, the common support is from around \$14,000 to \$21,000, and more than half the sample lies in the common support.

Common support

The following case is more extreme:



There is a small region of common support, but only around 10% of the sample lies in the common support.

Common support

If the common support is small, little can be done to reliably protect the analysis from strong biases.

In the first example, where there is a substantial common support, several options are available to us.

Stratification

Stratification can be used to protect against biases in observational studies when there is substantial common support for a potential confounder.

The details vary depending on the particular statistical analysis being used (e.g. Z-test, χ^2 test, ...), but there is a basic strategy that is followed.

Continuing with the previous example, suppose we intend to do a Z-test between childrens' test scores in the treated and untreated samples.

Instead of doing an overall test, we could divide the sample into 5 parts based on parent income. Within each income slice, the incomes are relatively similar, so the potential for confounding is minimized.

Stratification

Say we cut the income into 20% slices based on the marginal income distribution. Then we can estimate the treatment effect D_1, \dots, D_5 for the five slices separately. We can then use a weighted average

$$\bar{D} = w_1 D_1 + \dots + w_5 D_5,$$

where $w_1 + \dots + w_5 = 1$ to estimate the overall treatment effect with the confounding effect of income mostly removed.

We also want a confidence interval for the treatment effect, and a hypothesis test for it being nonzero. To do this, estimate the variances $V_i = \text{var}(D_i)$, and obtain

$$\text{var}(\bar{D}) = w_1^2 V_1 + \dots + w_5^2 V_5.$$

This allows us to form an overall Z-score

$$\bar{D} / \sqrt{\text{var}(\bar{D})}$$

to test whether the treatment effect is nonzero.

Stratification

The estimated treatment effect in the i^{th} slice is

$$D_i = \bar{Y}_T^{(i)} - \bar{Y}_U^{(i)}$$

and its variance is

$$V_i = \hat{\sigma}_T^{2(i)} / n_T^{(i)} + \hat{\sigma}_U^{2(i)} / n_U^{(i)},$$

where $\bar{Y}_T^{(i)}$, $\hat{\sigma}_T^{2(i)}$, and $n_T^{(i)}$ are the sample mean, sample variance, and sample size for the treated cases in the i^{th} slice, and $\bar{Y}_U^{(i)}$, $\hat{\sigma}_U^{2(i)}$, and $n_U^{(i)}$ are the sample mean, sample variance, and sample size for the untreated cases in the i^{th} slice.

Stratification

Here is what we get for a simulated data set with true treatment effect 0.4, and using $w_i = V_i^{-1} / \sum_j V_j^{-1}$.

n_T	n_U	$\bar{Y}_T^{(i)}$	$\bar{Y}_U^{(i)}$	D
8	112	1.14	1.28	-0.13
34	86	1.79	1.58	0.22
70	50	1.95	1.63	0.32
86	34	2.32	2.00	0.32
102	18	2.58	1.65	0.93

The overall estimated treatment effect is 0.38 with a Z-statistic of 3.9. The unstratified treatment effect estimate and Z-statistic are 0.71 and 8.1, respectively. Thus the treatment is clearly beneficial even after adjusting for parent income, but not nearly as strongly as it appeared to be when we did not adjust for the confounding effect of income.

Exercise: Using calculus and two strata for simplicity, show that the formula for w_i used here minimizes the variance of \bar{D} .

Matching

Suppose Y_i is the outcome for a treated unit (so $T_i = 1$). In an ideal comparative study, we would compare Y_i to the outcome for a unit that is identical to it in every way, except for being untreated.

In practice, we can only work with measured covariates that are potential confounders. Suppose we have two such covariates, X_1 and X_2 . We can assess the similarity between treated case i and untreated case j using

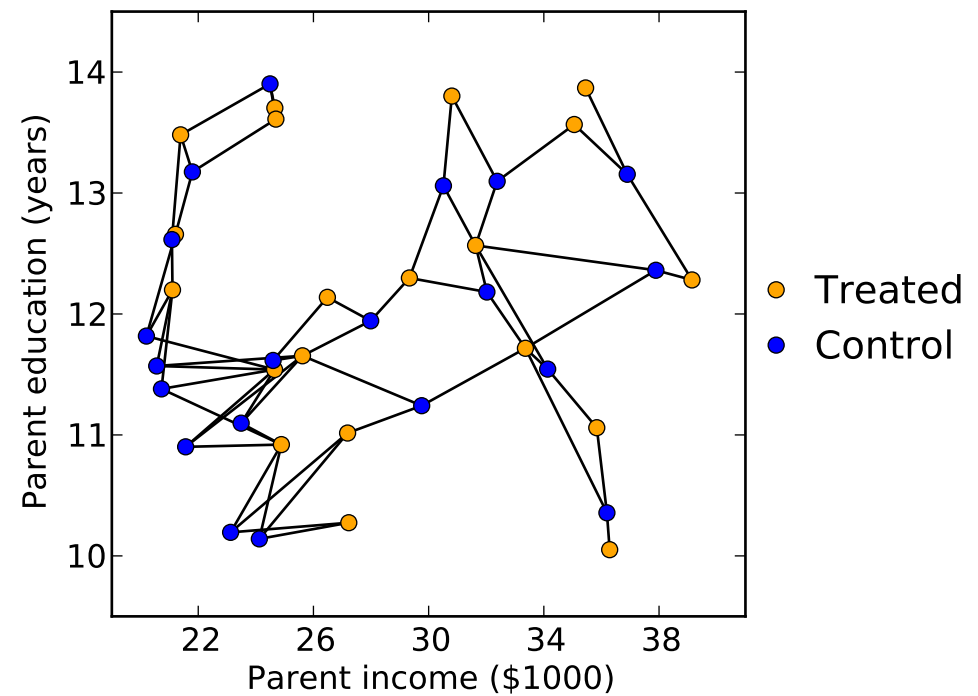
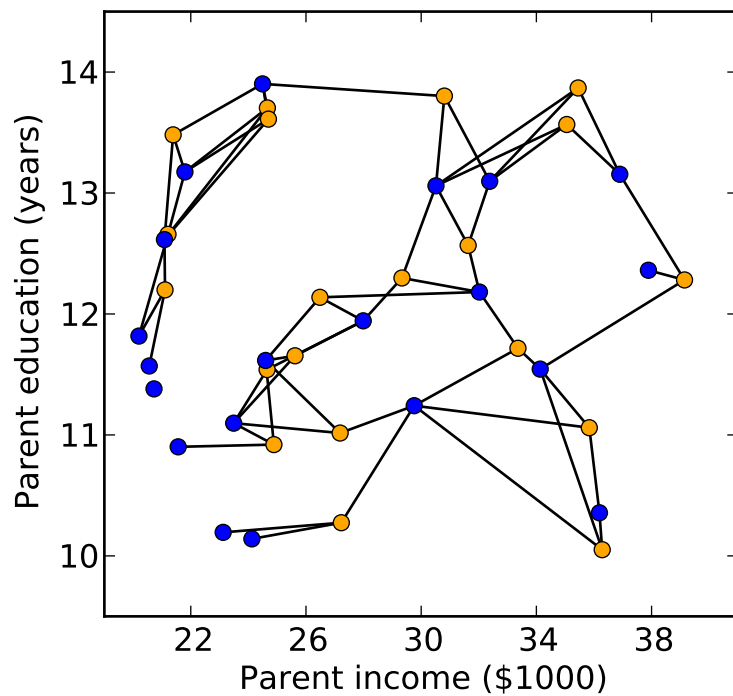
$$D_{ij} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}.$$

The matching estimate of the treatment effect is formed by selecting a number M of matches to use, and then forming the average outcome of the M untreated units that are closest to treated unit i . Denote this average by Y_i^* .

Similarly, for an untreated case j we can form the average outcome for the M closest treated cases, denoted by Y_j^* .

Matching

Matching each treated unit to the three closest control units (left), match each control unit to the three closest treated units (right):



Matching

The treatment effect can then be estimated using:

$$\hat{D} = \left(\sum_{T_i=1} Y_i - Y_i^* + \sum_{T_i=0} Y_i^* - Y_i \right) / (n_T + n_U).$$

It is possible to get an analytic formula for the variance of \hat{D} , but for this course we will use the bootstrap.

Covariate adjustment

Another common strategy for dealing with known, measured confounders is **covariate adjustment**. We will discuss this when we get into the section of the course on linear models.

Balance and representativeness

All research studies aim to generalize from the sample of analyzed units to the population from which the units were drawn.

In order to generalize we should study a representative sample – one that is reasonably similar to a random sample from the population of interest.

It is important to be aware that if extreme selection is used to create balance, it may be at the expense of representativeness.

For example, in the study of the effectiveness of the pre-school program, suppose our population of interest is 50% Hispanic. We can create a balanced sample that has no Hispanic children in either the treatment or control group, but we cannot be confident that our findings would generalize to a population that contains a substantial fraction of Hispanic children.

Sensitivity analysis

Earlier we looked at an example where the mean outcome is

$$E(Y|T, S) = \alpha_T + \beta_S,$$

where α_T ($T = 0, 1$) is the treatment effect and β_1 is the effect due to smoking, which acts as a confounder. We saw that if we don't control for smoking, the estimated treatment effect is expected to be

$$\alpha_1 - \alpha_0 + \beta_1(p_1 - p_0),$$

where p_1, p_0 are the proportions of smokers in the treated and untreated groups, respectively.

Suppose we carry out such an analysis and obtain a Z-score of 4, giving a two-sided p-value of 6.3×10^{-5} .

What would it take for confounding to reduce this Z-score to an uninteresting value (one that is below 2 in magnitude)?

Sensitivity analysis

Suppose we use a two-sample Z-test to compare the treated and untreated samples:

$$\sqrt{n} \frac{\bar{Y}_T - \bar{Y}_U}{\sqrt{\hat{\sigma}_T^2/q_T + \hat{\sigma}_U^2/q_U}}$$

where n is the total sample size, \bar{Y}_T, \bar{Y}_U are the sample mean outcomes for treated and untreated subjects, n_T, n_U are the numbers of treated and untreated subjects, and $q_T = n_T/n, q_U = n_U/n$ are the proportions of treated and untreated subjects in the sample.

For sensitivity analysis we need to simplify things a bit and focus on the key issues, so we'll ignore sampling variation in $\hat{\sigma}_T$ and $\hat{\sigma}_U$, and let

$$s^2 \equiv \hat{\sigma}_T^2/q_T + \hat{\sigma}_U^2/q_U,$$

and treat this as a fixed quantity.

Sensitivity analysis

If we further assume that $\sigma_T = \sigma_U = \sigma$, then

$$s^2 = \hat{\sigma}^2(1/q_T + 1/q_U) = \hat{\sigma}^2 f,$$

where $f = 1/q_T + 1/q_U$.

Thus the Z-score becomes

$$\sqrt{n}(\bar{Y}_T - \bar{Y}_U)/(\sigma\sqrt{f}).$$

Above we saw that $E(\bar{Y}_T - \bar{Y}_U) = \alpha_1 - \alpha_0 + \beta_1(p_1 - p_0)$, thus

$$\begin{aligned} \sqrt{n}(\alpha_1 - \alpha_0 + \beta_1(p_1 - p_0))/(\sigma\sqrt{f}) &= \sqrt{n}(\alpha_1 - \alpha_0)/(\sigma\sqrt{f}) + \sqrt{n}\beta_1(p_1 - p_0)/(\sigma\sqrt{f}) \\ &\approx 4. \end{aligned}$$

Sensitivity analysis

The Z-score based on the direct treatment effect is $\sqrt{n}(\alpha_1 - \alpha_0)/(\sigma\sqrt{f})$. For this to be greater than 2, we must have the Z-score contribution from confounding be less than 2:

$$\sqrt{n}\beta_1(p_1 - p_0)/(\sigma\sqrt{f}) < 2,$$

so

$$(\beta_1/\sigma)(p_1 - p_0) \approx 2\sqrt{f/n}.$$

The value of β_1/σ is the amount that smoking changes the expected response in SD units. We might have some idea about the plausible size of this and some of the other quantities. For example, suppose $n = 100$ and $f = 4$ (corresponding to equal numbers of treated and untreated subjects). Then

$$(\beta_1/\sigma)(p_1 - p_0) \approx 0.4.$$

Sensitivity analysis

Since $|p_1 - p_0| < 1$, we must have $\beta_1/\sigma > 0.4$ (assuming that the direction $\beta_1 \geq 0$ is known, which is typical). This is already a large effect, and if $|p_1 - p_0| \ll 1$ which is expected, the effect would have to be even larger.

Thus it seems unlikely that confounding by smoking could reduce our Z-score to an insignificant value in this example.

Imperfect experiments are somewhat like observational studies

Especially when working with human subjects, violations of an experimental protocol are common.

Treatment adherence Subjects assigned the treatment may inadvertently or deliberately deviate from the planned treatment regimen. If the treatment is difficult or has unpleasant side effects this is especially likely.

Drop out/loss to follow up Patients enrolled in a trial and randomized to either the treatment or control group may die, move away, or elect to stop participating in the study.

Missing/partial data Covariate or outcome data may be missing for some people because of errors made by research staff or equipment failures.

Measurement error in balancing variables If balanced randomization is being used, measurement errors in the balancing variables may create the appearance of balance when it is not in fact balanced. This could happen, for example, if balancing variables relating to socially disapproved behaviors like drug use are used.

Imperfect experiments

Subversion Subjects assigned to either the treatment or control group may actively subvert the study, e.g. by seeking out other treatments that may mask or amplify the effect of the treatment under study.

In any of these cases the experiment may need to be analyzed as if it were an observational study, using methods like matching or stratification to reduce the bias from violations of the experimental plan.

Natural experiments

Just as complex experiments can come to resemble observational studies, some observational studies resemble experiments.

Sometimes a natural event occurs that can arguably be treated as a randomized treatment assignment.

Here is one example:

Researchers studying autism have sought to relate childhood television viewing to the risk of developing autism.

A potentially relevant natural experiment was the introduction of cable television in the US in the 1970's. Much data shows that TV watching went up across most segments of society after cable TV was introduced. For logistical reasons, cable TV was introduced in different communities at different times. By comparing autism rates among people who lived as children in communities with cable to the rates among people who lived as children in communities without cable, it may be possible to estimate the contribution of TV viewing to autism risk.

Natural experiments

Here is another example of a natural experiment:

A fundamental question in biology is to identify the amount by which genetic factors influence a quantitative variable such as height.

In principal this can be done by comparing pairs of people who share some or all of their genes (i.e. siblings) to pairs of people who share a much lower fraction of their genes (i.e. unrelated people).

A potential confounder is environment – people who share genes tend to share a common environment (they have the same parents, live in the same home, etc.).

Nature provides a “natural experiment” – identical twins share a family environment and 100% of their genes, while same sex fraternal twins share a family environment and only 50% of their genes.

If the heights of identical twins are more similar than the heights of fraternal twins, a genetic source of variation is a likely explanation.

Natural experiments

Are these really as good as experiments?

No, in the autism example it is possible that the regions in which cable TV was introduced first are different in other important ways, such as having higher levels of environmental toxins.

For the genetics example, it is possible that parents inadvertently or deliberately create a more similar environment for identical twins than for fraternal twins. It seems unlikely that this could influence height, but for behavioral traits it is less clear that this possibility can be excluded.

However it is generally accepted that even when a natural assignment is an imperfect proxy for randomized assignment, it can provide very useful information about what the unconfounded treatment effect may be.