CHAPTER 4: CAUSAL INFERENCE FOR POLICY ANALYSIS

Study design and policy effects

There are a number of approaches for estimating the impact of policy changes on behavior.

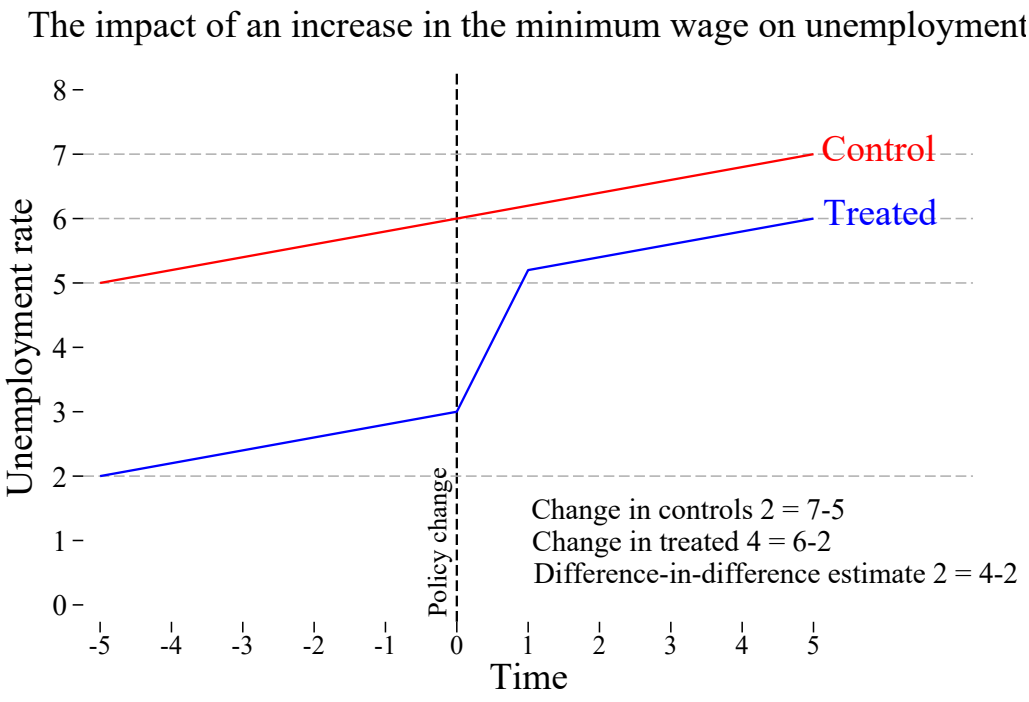
- Randomized trials in the real-world.
- Randomized trials in laboratory settings.
- Surveys (Ask people how their behavior would change).
- Cross-sectional studies.
- Pre-post comparisons.
- Pre-post with concurrent controls or "difference-in-difference" studies.
- Regression discontinuity designs.
- Natural experiments or instrumental variables studies.

Many studies that estimate the impact of policies use a difference-in-difference design. Suppose some states implemented a policy (for example, an increase in the minimum wage) and others did not. Individuals in states that raised the minimum wage are the treatment group and individuals in other states are the control group. The outcome is the likelihood of being employed. Suppose y is the outcome variable, like the unemployment rate. The difference-in-difference estimator is

$$(y^{T,POST} - y^{T,PRE}) - (y^{C,POST} - y^{C,PRE})$$

where T = treated and C = control. PRE = outcome before the policy change, POST = outcome after the policy change.

By comparing changes in outcomes rather outcomes at a point in time, the design removes bias due to time-invariant (i.e. unchanging) differences between the states. A picture is useful for illustrating this concept.



States in the treatment group may have raised the minimum age at different points in time. We can transform time so that time = 0 corresponds to the date when the wage increase went into effect. Time 0 will correspond to different calendar years in different states.

The *cross-sectional*, post-period difference between treatment and control states is -1 = (6 – 7). It implies that unemployment is *lower* in states that increased the minimum wage.

The cross-sectional comparison is biased by the fact that the states start out with different unemployment levels. The pattern makes sense: states that have a higher unemployment rate to begin with may be less willing to take the risk of increasing the minimum wage.

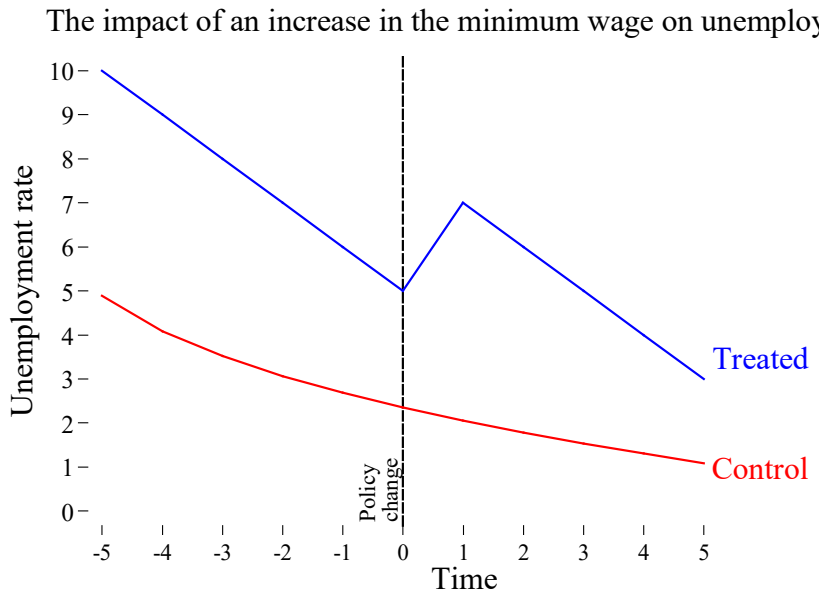
The *pre-post* difference in the treated group is 4 = (6 – 2). (Unemployment increased.)

The pre-post difference is biased by the fact that unemployment rates were trending up before the increase in the minimum wage. There was a "secular" trend.

The difference-in-difference estimate is 2 = (6 – 2) – (7 – 5). (Unemployment increased.)

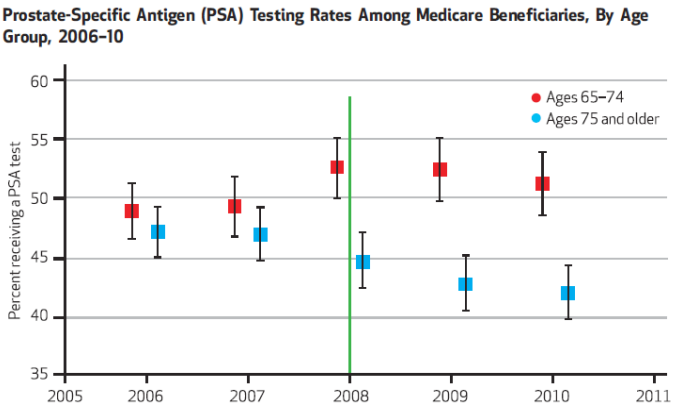
The difference-in-difference estimate adjusts for underlying differences in the unemployment rate between treatment and control states. It also adjusts for secular trends that are common to the treatment and control groups.

Here is a case where a difference-in-difference study probably would not yield the right outcome.



The pre-trends look similar, but the post-trends do not. The policy change may have had an impact, but unemployment in control states is close to 0. It is unusual for employment rates to be below 2%. This boundary effect may be why the rate of decline in unemployment rates in control states was lower.

Here is an example from a real study. [1] In 2008 the United States Preventive Services Task Force recommended against routine prostate cancer screening for men 75 years and older. In this study, men ages 65 to 74 are the control group. It appears that the recommendation had an impact. Screening rates decreased among men 75 years and older and were

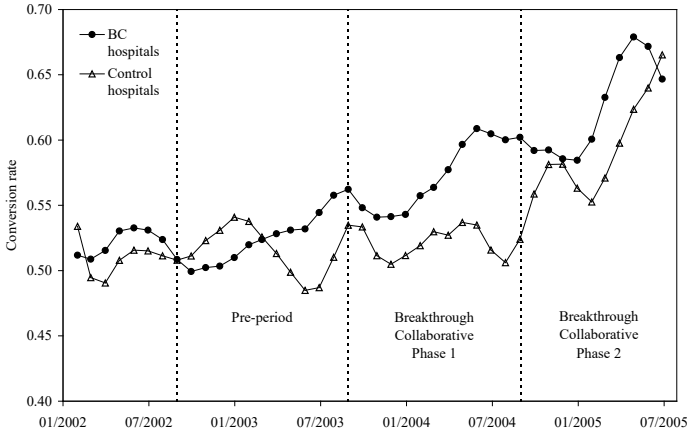


[1] Howard DH, Tangka F, Ekwueme D, Guy G, Lipscomb J. Prostate cancer screening in men ages 75 and older fell by 8 percentage points after Task Force recommendation. *Health Affairs* 2013;32(3):596-602.

basically unchanged among men ages 65 to 74.

Here is another example.[2] In 2003 the Department of Health and Human Services began a campaign (the "Breakthrough Collaborative on Organ Donation") to increase organ donation rates in selected hospitals. In 2003 the Department expanded to program to all hospitals. It looks like the Collaborative had an effect on donation rates, but rates were trending upward prior to the Collaborative.

Figure 1: Smoothed trend in the coversion rate among hospitals participating in the first phase of the Breakthrough Collaborative and control hospitals



BC: Breakthrough Collaborative. NOTE: The conversion rate is smoothed using the ksm command in Stata Verson 7.0 (Stata Corporation, College Station, TX) with a bandwidth of 0.15 and the "lowess" option.

Yet another example: Between 1997 and 2002 15 children were diagnosed with leukemia in Clark County Nevada, making Clark County one of a number of "cancer clusters" around the country. Davis (2004) studied the impact of the discovery of the cluster on home prices. He used Lyon County, which borders Clark County and has similar income and housing price levels, as a control. Although the first case was diagnosed in 1997, it was not until 2000 that there was a steep uptick in cases and newspapers began running stories on the cluster. The graph shows that trends in Churchill and Lyon Counties were similar before 2000. After 2000, they diverged. Estimates in the table to the left show that relative to prices in Lyon County, home prices in Churchill County declined by about 7.7 percent.

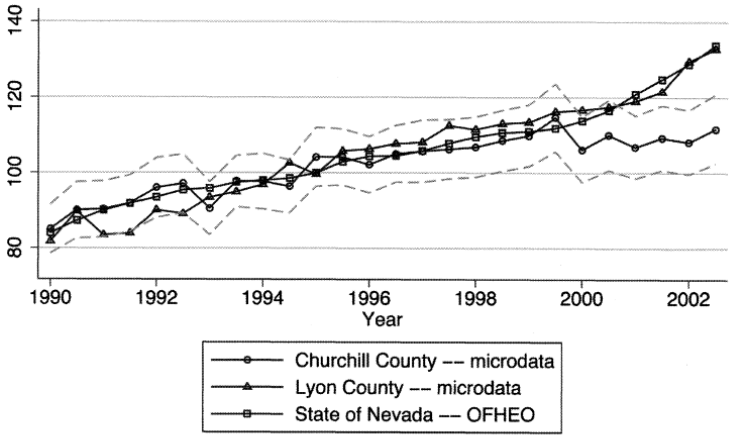


FIGURE 2. ASSESSING THE VALIDITY OF THE COUNTERFACTUAL: HOUSE-PRICE INDICES FOR SINGLE-FAMILY RESIDENCES, 1990–2002

[2] Howard DH, L Siminoff, V McBride, M Lin. Does quality improvement work? Evaluation of the Organ Donation Breakthrough Collaborative. *Health Services Research* 2007;42(6):2160-2173.

Regression analysis versus study design

A regression model is not a study design. A study design refers to how the effect is estimated. A regression model is an approach for implementing a study design. Consider the following model where the coefficient (β^1) on an indicator variable for treatment, $Treat^i$, is of interest.

$$y_{it} = \beta^0 + \beta^1 Treat_i + \beta^2 Age_i + \beta^3 Male_i + \varepsilon_{it}$$

You could use this model for a non-randomized cross-sectional analysis or a randomized trial. The model itself tells you nothing about whether treatment was randomly assigned or not.

A regression model for a difference-in-difference study is:

$$y_{it} = \beta^0 + \beta^1 Treat_i + \beta^2 Post_t + \beta^3 Treat_i \times Post_t + \beta^4 Age_i + \beta^5 Male_i + \varepsilon_{it}.$$

The coefficient on the interaction of the treatment indicator and post-period indicator, β^3, is of interest.

Different regression models (for example, logistic, generalized linear model) are designed to handle different types of data. You would use a logistic model to estimate effects when the outcome is dichotomous (0/1), regardless of whether the underlying study design was a randomized trial, pre-post analysis, cross-sectional, or a difference-in-difference study.

Interpreting effects from randomized trials with non-compliance

Randomized trials are usually considered the gold standard for estimating the effects of policies and medical treatments because they produce unbiased estimates. However, results may not be externally generalizable, especially when there are high rates of non-compliance.

Prostate specific antigen (PSA) testing is widely used in the US to screen men for prostate cancer. However, use of PSA screening is controversial. Many men are treated for prostate cancers detected via PSA screening that would never have become clinically apparent in the absence of screening.

The CAP Randomized Trial randomly assigned primary care practices in Britain to an intervention to increase PSA screening (patients received an invitation to a PSA testing clinic) or usual care.[3] Of the 189,386 men in the intervention group, only 36% actually had

[3] Martin RM, Donovan JL, Turner EL, Metcalfe C, Young GJ, Walsh EI, Lane JA, Noble S, Oliver SE, Evans S, Sterne JAC, Holding P, Ben-Shlomo Y, Brindle P, Williams NJ, Hill EM, Ng SY, Toole J, Tazewell MK, Hughes LJ, Davies CF, Thorn JC, Down E, Davey Smith G, Neal DE, Hamdy FC, . Effect of a Low-Intensity PSA-Based Screening Intervention on Prostate Cancer Mortality. The CAP Randomized Clinical Trial. *Journal of the American Medical Association* 2018;319(9):883-895.

blood drawn for a PSA test. The authors estimate that about 15% to 20% of the men in the control group had a PSA test.

The investigators compared men randomized to the treatment arm to men randomized to the control arm, regardless of whether they had a PSA test or not. This approach produces an "intent to treat" estimate. It will systematically underestimate the effect of the treatment. It estimates the effect of being *randomized* to the treatment arm, not the impact of the treatment itself. The CAP trial concluded that PSA screening did not reduce death from prostate cancer with a 10 year follow-up.

An alternative to the intent-to-treat estimate is the "per protocol" estimate. In the case of the CAP trial, a per protocol estimate would compare 1) men in the treatment arm who were screened to 2) men in the control arm who were not screened. (In the case of perfect compliance, the intent-to-treat and per protocol estimates would be the same.) The per protocol estimates the impact of being screened, as opposed to being randomized to the screening arm, but the estimate is biased. Men in the treatment arm who were screened probably differed from those who were not. Perhaps they were more likely to have a family history of cancer. Ditto for screened and unscreened men in the control arm. Focusing on only a non-randomly selected subset of participants in the treatment and control arm eliminates the benefits of randomization.

There is a statistical technique for trying to estimate the impact of an intervention in the face of non-compliance to treatment assignment (for example, some people in the control arm receive the treatment), but not all randomized trials report these adjusted estimates.

Interpreting and applying policy effect estimates

Studies describe effect estimates using many different measures: elasticities, risk ratios, odds ratios, etc. It is important to know how to interpret and apply these quantities.

If you cannot explain something in simple terms, you don't understand it.
–Richard Feynman

Estimates of the impact of a variable on a continuous outcome, like dollars, are usually stated in terms of the original scale or percent changes. Sometimes they are stated in terms of elasticities. If the price elasticity of smoking is -0.4, a 10% increase in price leads to a 4% decline in smoking.

Things get trickier when the outcome is binary. Consider an intervention that reduces the proportion of people who smoke from 20% to 15%.

The intervention reduces the smoking rate by 5 *percentage points*. When describing an effect in terms of percentage points, you should always use the verbiage "percentage points" and not "%".

Equivalently, you might say that the intervention led to a 25 *percent* (or %) decline in smoking rates (= 5 ÷ 20%).

It would be incorrect to say that the intervention caused smoking rates to decline by 5% (which would imply that smoking rates declined from 20% to 19%). It would be correct to say that the intervention caused rates to fall by 5 percentage points, or 25%.

The *relative risk* of smoking in the intervention group compared to the non-intervention group is 0.75 (= 15% ÷ 20%). Without knowing the baseline probability (20% in this case), you cannot convert a relative risk into a percentage point change.

The *odds ratio* would be

$$\frac{\frac{0.15}{1-0.15}}{\frac{0.2}{1-0.2}} = 0.71$$

It's close to the risk ratio, but not the same. A formula for converting odds ratios to risk ratios is

$$\frac{OR}{(1-P_o)+P_o \times OR} = RR$$

$$\frac{.71}{(1-.2)+.2\times.71} = 0.75$$

To apply this formula, you need to make an assumption about the baseline level of risk P_o. If the baseline risk is less than 30%, then the odds ratio will be close to the risk ratio.

Applying absolute or relative effects

Policy effect assumptions can be applied as absolute or relative effects. The absolute decline in smoking rates associated with the hypothetical intervention mentioned earlier is 5 percentage points. The relative decline is 25% (= 5 ÷ 20%).

Absolute versus relative effects

	Absolute decline (5 percentage points)		Relative decline (25%)		
Baseline	Change (percentage points)	New level	Change (percentage points)	New level	Difference
10%	5	5%	3	8%	3%
20%	5	15%	5	15%	0%
30%	5	25%	8	23%	-3%
40%	5	35%	10	30%	-5%
50%	5	45%	13	38%	-8%

Suppose you want to use these results to estimate the impact of expanding the intervention to a different population. Unless the baseline rate of smoking happens to be 20%, the same as in the original study, your prediction will depend on whether you apply a relative or absolute effect (see table).

Suppose the baseline smoking rate is 10%. Using the effect stated as an absolute decline, you would predict that smoking rates fall to 5%. Using the effect stated as a relative decline, you would predict that that rates fall to 8%. Which is correct? Maybe the smokers in this population are more committed, more addicted smokers than in the population included in the study (where the baseline rate was 20%). Maybe they are more resistant to efforts to get them to quit. In that case, you might want to use the relative effect, which yields a more conservative prediction. But it is a judgement call.

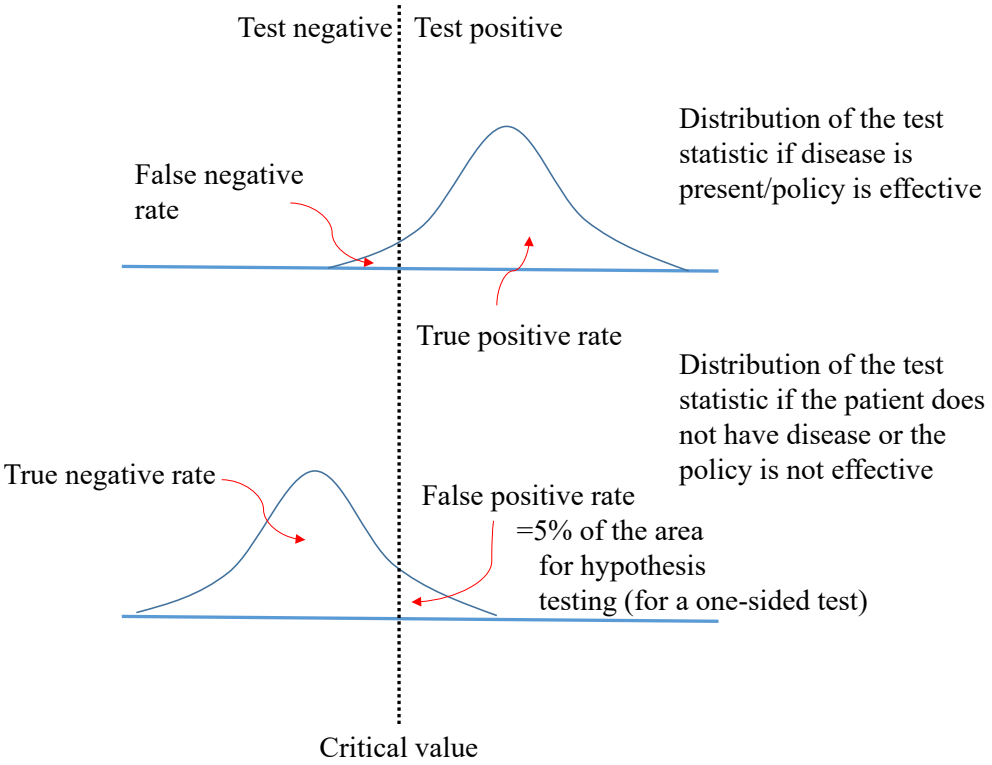
Statistical significance

In the academic literature, measures of policy effects are almost always reported alongside confidence intervals, t-statistics, p-values, etc. How should significance levels affect our interpretation of policy effects? It helps to think about the analogy between hypothesis testing in statistics and diagnostic testing in medicine.

In medicine, the true positive rate is the probability that a test is positive for a patient who is diseased. The true negative rate is the probability that a test is negative for a patient who does not have a disease. The people who design diagnostic tests (or the physician who interprets them) can control the true positive rate and true negative rate by adjusting the threshold of the test to balance the harms of failing to detect disease in a patient who has it (a false negative) and treating disease in a patient who does not (a false positive). There is a tradeoff: moving the threshold to increase the true positive rate will reduce the true negative rate.

In statistical hypothesis testing the convention is to set the threshold for a positive test such that the true negative rate is 95% and the false positive rate is 5%.

Suppose we want to analyze whether a policy works by performing a t-test comparing an outcome between subjects exposed to the policy and control subjects. Due to sampling variability, the observed value of the test statistic will be different every time we perform the analysis. The top panel of the Figure depicts the distribution of the test statistic if the policy works. The bottom panel depicts the distribution of the test statistic if the policy does not work (i.e. the null hypothesis). The distributions overlap. Unless the statistic is far to the right or left, we cannot tell which state of the world we are in based on the test statistic. We have to make an educated guess. If the test statistic is above the critical value, we conclude that we are probably in the state of the world where the policy is effective. Of course in hypothesis testing, there is a 5% chance that even if the policy doesn't work, the test statistic will be above the critical value (a false positive result).



The p-value of a test refers to the area to the right of the test statistic under the distribution of the test statistic in the state of the world where the policy does not work. In hypothesis testing, we set the critical value such that the false positive rate is 5%.

Diagnosis based on a lab test in medicine works much the same way, but clinicians set the critical value to balance the harms and benefits of detection, not based on convention. The false positive rate may be higher or lower than 5%. For example, if treatment is inexpensive and does not have side effects, then a false positive is not a bad outcome. A test maker might set the cutoff so that the false positive rate is higher than 5%.

Academic researchers adhere to the 5% convention (there is a little wiggle room). Policy analysts are not. In some cases the 5% convention maybe overly conservative, leading policymakers to reject a policy that, even taking the uncertainty into account, would pass a cost-benefit test. That said, it would be unusual to see a policy analysis of a policy that failed to produce a statistically significant improvement in its primary outcome. You might see analysts incorporate secondary endpoints where the effect is not statistically significant. For example, a policy that increases high school graduation rates might also reduce crime. A policy analyst may consider both effects, even if the impact on crime is not significant by conventional measures.