Policy Research: The Art of Creating Convincing Evidence

Identification

Stephen B. Holt

2022-09-21

Basic Process

Most Policy Research

- Define the question you would like answered (e.g., do women have more affairs than men? or does the neighborhood in which you grow up influence your earnings as an adult?).
- State hypotheses about the answer to the question (e.g., based on previous research, men have more affairs than women or neighborhoods have no effect on adult earnings).
- Collect data that can answer the question (e.g., survey a sample of the population about the number of affairs they've had or use administrative data on both residences and earnings).
- Calculate measures to test hypotheses put forward about the relationship of interest (e.g., the average number of affairs men have relative to the average number of affairs women have, or the average adult earnings within a neighborhood).
- Organize and report results.

Defining the Question

How do we find and define topics for research questions?

- Policymakers (e.g., considering a new program, evaluating an existing program)
- Advocacy organizations or think tanks identify a problem
- Academic literature has found a problem, trend, or policy worth examining

Defining the Question

Research questions generally investigate a specific question on a specific topic to better inform the data needed to answer the question

- Describe a trend in policy specific behaviors (e.g., Are imprisoned persons more likely to re-offend? Are teachers leaving the profession at higher rates?)
- Identify a theoretically important relationship (e.g., Does monopoly control of a service increase prices? Does denser development reduce driving?)
- Identify a causal link between an intervention and outcome (e.g., Does higher funding levels for schools improve student learning? Does better street lighting reduce crime?)

Our Target

There is a parameter of interest in the population (either the typical value of a single variable or the link between two variables), something we rarely (if ever) observe directly

The parameter tells us something about people and society in a way that could be important for policy (explaining why it worked or didn't work as intended)

However, to get to identify that parameter, we must first collect data

Study Designs

Once we have a research question, we need to design our study. Generally three kinds:

1. Observational: Record data on individuals without attempting to influence the responses.
   - Good for describing a trend or theoretically important relationship.
2. Quasi-experimental: Use advanced statistical techniques to estimate effects of treatments on people in the real world.
   - Good for identifying a causal link between an intervention and an outcome.
   - BUT. . . assumption heavy
3. Experimental: Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.
   - Good for identifying a causal link between an intervention and an outcome.
   - BUT. . . expensive and potentially limited generalizability.

Collecting Data

We have a question, we have a design, now we need data! Most common ways to get data in policy research:

- Survey a sample or population of interest to the study.
- Observe and record information about a sample or population.
- Collect administrative data about a sample or population.

What is the difference between a sample and a population?

- Population refers to every individual in a given frame. Example: All humans, all residents in America, all public school students in California, all bees.
- Sample refers to the set of individuals we observe in our data. Example: A sample of public school students in California, a set of bees collected from different locations.

Samples

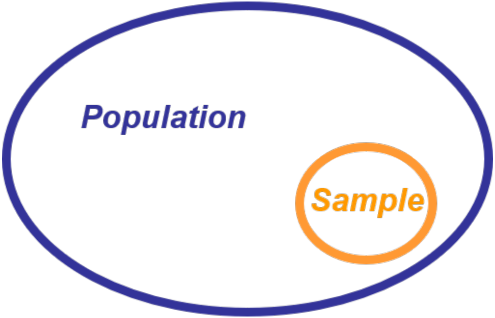


Figure 1: Sample

Sampling

Study results can be deeply influenced by decisions made when constructing the sample for analysis. Common sample designs you will see in research:

- Convenience sampling: Just ask whoever is around. Examples: Street polls, classroom polls, many marketing surveys.
- Voluntary Response Sampling: Individuals choose to be involved. Examples: Clinical trials, Internet polls
- Random sampling: Individuals are randomly selected. Each individual in the population has the same probability of being in the sample. Example: Public opinion polls.
- Stratified random sample: a series of random sampling performed on subgroups of a given population.
- Multiple stage random sample: select groups within a population in stages, resulting in a sample consisting of clusters of individuals.

Sampling Considerations

Two important factors for a good sample:

1. Sample is representative of the population of interest.
2. All eligible individuals have equal likelihood of selection into sample.

Many different designs try to optimize both representativeness and efficiency (lowest cost for needed statistical power). Considerations for samples:

- Random sampling might lead to **undercoverage**, an issue where a subpopulation is excluded or undercounted in the sample, and create an unrepresentative sample
- All samples can suffer from nonresponse, which occurs when people refuse to provide some or all information for the study despite being sampled, again yielding unrepresentative samples.
- Participants can be subject to question wording effects and other biases in questions

Sampling Considerations

Always consider the strength of a sample when analyzing it (or reading someone's analysis of it). Be skeptical of convenience and voluntary samples in observational studies!

Many observational studies want a realistic picture of a problem, trend, or relationship. For observational studies, in particular, it is important to ensure the sample is representative of the population of interest and not subject to nonrandom sample distortions.

Quasi-experiments and experiments use techniques that can compensate for some sample deficiencies (but should still avoid them).

Research Designs

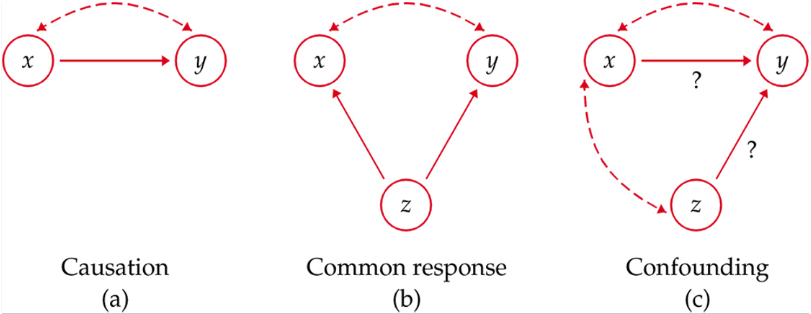
- Quasi-experiments and experiments distinguish themselves from observational studies by seeking to answer an additional question: what would the outcome have been if people had **not** been given a treatment? In other words, these designs try to estimate **potential outcomes**.
- **Treatment** in experimental designs refers to the factor, such as a policy change or program, that can be manipulated by researchers (or the larger world).
- To approximate the potential outcome that would have happened without the treatment, experimental designs also include a **control** group, or people not given the treatment.
- We refer to objects in an experiment **experimental units** or, if they are people, **subjects**.
- If something other than the treatment systematically effects one group and not another, we refer to this as **bias**.

Potential Outcomes and Experiments

- If our research question is "Does this treatment affect outcome Y?", we are interested in estimating the **average treatment effect**, which can be expressed as: $ATE = (\overline{Y_t} - \overline{Y_c})$
- This equation is just a fancy way of saying the difference between the average outcomes of people who received the treatment $(\overline{Y_t})$ and people who did not $(\overline{Y_c})$.
- We cannot observe what would happen to the people who got the treatment if they didn't receive it. . .
- BUT. . . if we randomly assign some people to get the treatment and others to not get it, we can construct a control group who, *on average*, only differs from the treatment group by not getting the treatment.
- Randomizing helps us ensure that we account for unobserved factors that might affect the outcome.
- The result: the difference in average outcomes between the two groups can only be attributed to the treatment!

Potential Outcomes and Experiments

More visually, some possible explanations for an observed association are below. The dashed lines show an association. The solid arrows show a cause-and-effect link. x is explanatory, y is response, and z is a lurking variable for which we lack data. Experiments control for Z because Z will be equal between treatment and control groups as a result of randomization.



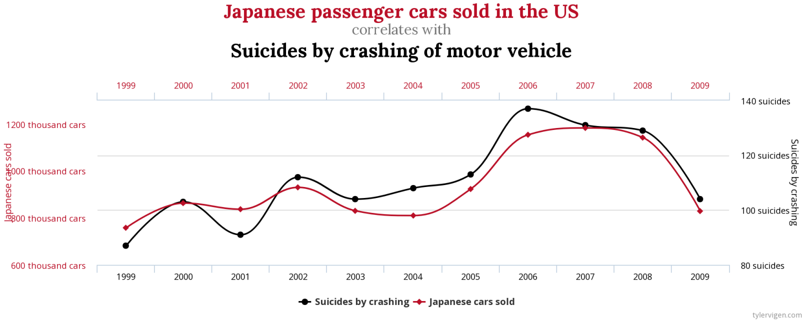
Causation
(a)

Common response
(b)

Confounding
(c)

Spurious Correlations

Put another way, sometimes X and Y can have the same pattern for no reason whatsoever (spurious correlations)!



Spurious Correlations

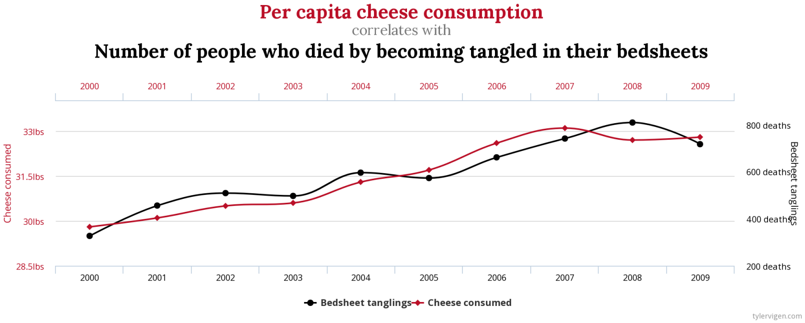


Figure 3: Cheese

Spurious Correlations

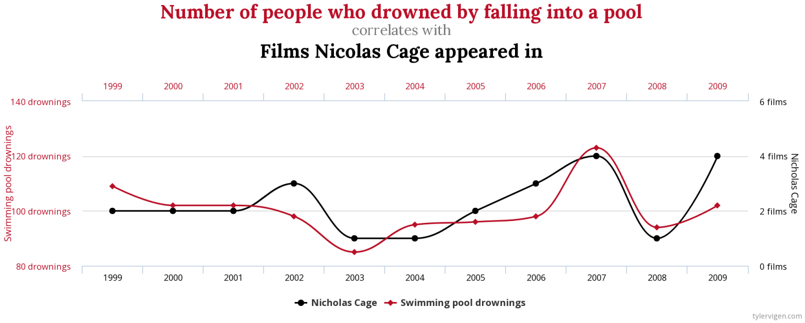


Figure 4: Cage

Spurious Correlations

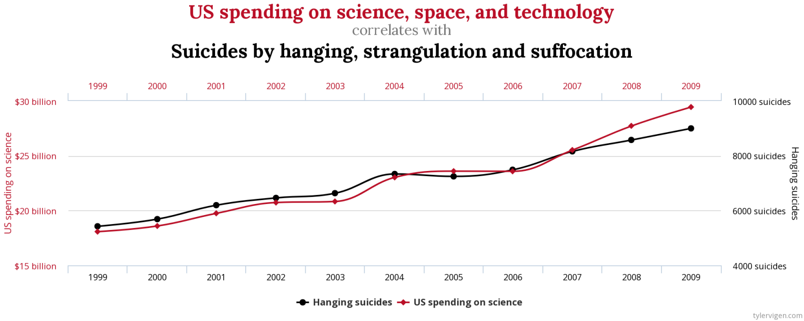


Figure 5: Hang

Paramters

The point here is theory and context of our research question provides insight about the parameter of interest that we hope to estimate with data.

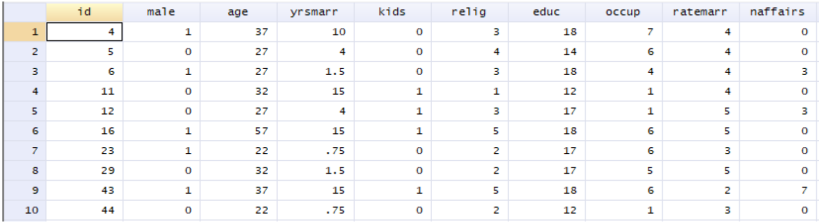
However, data must be collected and utilized with estimating that parameter correctly in mind. Sometimes correlations are the parameter all on their own! But you have to be clear about the parameter you need and the parameter you are estimating.

Data Vocabulary

- **Unit of Observation:** the unit about which information is being collected. Often denoted with a subscript in mathematical notation. Examples: schools, cities in the U.S., individual workers.
- **Observations:** constitutes an entry of all observed information collected about a unit in the data. In datasets, each row constitutes a single observation.
- **Variables:** a variable is a state, factor, or characteristic that is likely to change (*vary*) across observations or units of observation. In datasets, variables are stored in columns. There are a two major types of variables:
  - Quantitative variables - measured in numerical units.
  - Categorical variables - captures a unit's grouping with other similar units.

Data Example

Data on marital happiness and affairs



	id	male	age	yrsmarr	kids	relig	educ	occup	ratemarr	naffairs
1	4	1	37	10	0	3	18	7	4	0
2	5	0	27	4	0	4	14	6	4	0
3	6	1	27	1.5	0	3	18	4	4	3
4	11	0	32	15	1	1	12	1	4	0
5	12	0	27	4	1	3	17	1	5	3
6	16	1	57	15	1	5	18	6	5	0
7	23	1	22	.75	0	2	17	6	3	0
8	29	0	32	1.5	0	2	17	5	5	0
9	43	1	37	15	1	5	18	6	2	7
10	44	0	22	.75	0	2	12	1	3	0

- Unit of observation: person
- Each row represents a person responding to a survey
- Each column is a different variable
- Categorical variables: `male`, `relig` (1 = anti-religious to 5 = very religious), `occup`, `ratemarr` (1 = very unhappy to 5 = very happy)
- Quantitative variables: `age`, `yrsmarr`, `kids`, `educ`, `naffairs`

Properties of Data on Variables

- In a dataset, data for each variable will have a range of observed values and a frequency with which each value is observed.
- These two characteristics of data on a variable describe the **distribution** of the variable.
- The distribution can be visualized using a graph called a histogram.
- Creating a histogram divides the range of values into equally sized intervals, and shows the number of observations in each interval.

Histogram Example

Below is a histogram of the age of our sample, ranging from 17.5 to 57, in 5 year intervals. The first bar suggests a little over 100 observations fall into the first interval, 17.5 to 22.5 years old. The last bar suggests there's fewer than 25 aged 52 to 57 in the sample.

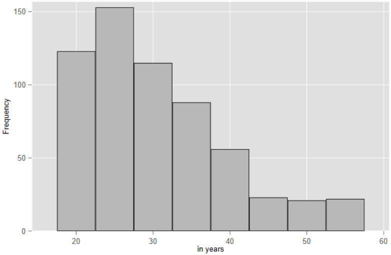


Figure 6: Age

Interpreting Histograms

The patterns depicted in histograms provide general information about a variable in a sample. These patterns can tell us about the **shape**, **center**, and **spread** of a variable's distribution in a sample. We look at patterns generally and think about the curves created by the bars rather than their precise connections.

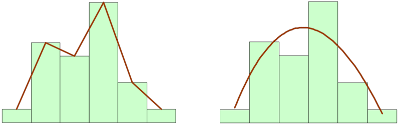


Figure 7: Examples

Measures of Center

- A starting point and building block of most analysis is a **measure of central tendency**. Simply put, measures of central tendency provide a way to assess the outcome of a typical case.
- The **mean** or **arithmetic average** is one of the most common measures of central tendency.
- The mean is calculated by summing the values of a variable and dividing by the number of observations.
- Sum of years married is 188.834. Divided by 25 people, the mean is 7.553.

respondent (i)	yrsmarr (x)
1	.417
2	.417
3	1.5
4	1.5
5	1.5
6	1.5
7	4
8	4
9	7
10	7
11	7
12	7
13	7
14	7
15	7
16	10
17	10
18	10
19	10
20	10
21	15
22	15
23	15
24	15
25	15

Means and their Ends

Expressed arithmetically:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \tag{1}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2}$$

$$\bar{x} = \frac{188.834}{25} = 7.553 \tag{3}$$

The average person in our sample has been married about 7 and a half years. If we knew nothing about a person in our sample, we would expect them to be married for about 7 and a half years judging by the sample average.

respondent (i)	yrsmarr (x)
1	.417
2	.417
3	1.5
4	1.5
5	1.5
6	1.5
7	4
8	4
9	7
10	7
11	7
12	7
13	7
14	7
15	7
16	10
17	10
18	10
19	10
20	10
21	15
22	15
23	15
24	15
25	15
n=25	$\sum = 188.834$

Measures of Spread - Standard Deviations

Distribution of student test scores.
Green: Mean
Red: ±1 standard deviation

- One of the most important measures of spread in statistics is the standard deviation, often denoted as s in mathematical notation.
- Similar to the mean, the standard deviation can be influenced by the skew in a distribution.
- Two steps in the calculation of s: calculate the variance (s^2) and take the square root.

$$s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2 \tag{4}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2} \tag{5}$$

Detailed Calculations

$$s = \sqrt{\frac{1}{df}\sum_{1}^{n}(x_i - \overline{x})^2} \qquad (6)$$

Mean: 8.1777

Sum of squared deviations from the mean: 716.2292

Degrees of freedom:

$df = n - 1 = 14$

s^2: $716.2292/14 = 51.1592$

s: $\sqrt{51.1592} = 7.1526$

i	yrsmarr (x)	$\overline{x}$	$(x - \overline{x})$	$(x - \overline{x})^2$
1	.75	8.177695	-7.427695	55.17066
2	.75	8.177695	-7.427695	55.17066
3	.75	8.177695	-7.427695	55.17066
4	1.5	8.177695	-6.677695	44.59161
5	1.5	8.177695	-6.677695	44.59161
6	1.5	8.177695	-6.677695	44.59161
7	1.5	8.177695	-6.677695	44.59161
8	15	8.177695	6.822305	46.54384
9	15	8.177695	6.822305	46.54384
10	15	8.177695	6.822305	46.54384
11	15	8.177695	6.822305	46.54384
12	15	8.177695	6.822305	46.54384
13	15	8.177695	6.822305	46.54384
14	15	8.177695	6.822305	46.54384
15	15	8.177695	6.822305	46.54384
	128.25	8.177695	5.585	716.2292

Visualizing Categorical Variables

Remember that in addition to quantitative data, there are some variables, such as ratings for how happy a person is in their marriage, that group people into categories. Categorical data can be presented in bar graphs, where bars represent each category:

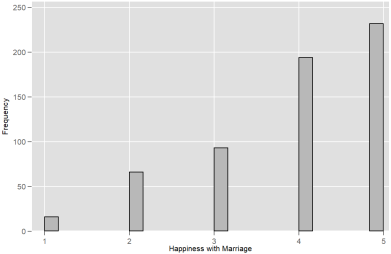


Figure 8: Happiness in Marriage

Visualizing Categorical Variables

. . . or pie graphs, where slices represent each category's proportion of the whole sample:

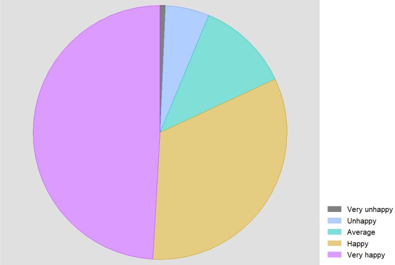


Figure 9: Happiness in Marriage

Simple Averages Identifying Issues

Raj Chetty and colleagues linked IRS data on earnings to address data to calculate average earnings in adulthood based on the neighborhood in which someone grew up. The results can tell us which neighborhoods promote upward mobility.

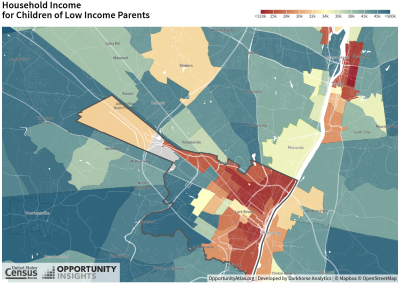
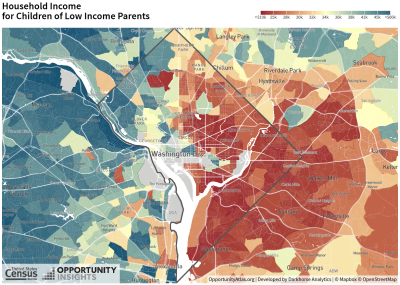

Household Income
for Children of Low Income Parents
<$10k 25k 28k 30k 32k 34k 36k 38k 41k 45k >$60k
Latham
Shakers
Karner
Roessleville
Westmere
Newtonville
Loudonville
Menands
Troy
South Troy
Lark Street
Van Allen Park
Voorheesville
United States Census Bureau
OPPORTUNITY INSIGHTS
OpportunityAtlas.org | Developed by Darkhorse Analytics | © Mapbox © OpenStreetMap

Figure 10: Mobility in Albany

Another Example



OpportunityAtlas.org | Developed by Darkhorse Analytics | © Mapbox © OpenStreetMap

Figure 11: Mobility in Albany

]