Producing Data

Stephen B. Holt, Ph.D.

Rockefeller College of Public Affairs & Policy
University at Albany State University of New York

February 22, 2022

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
   - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
   - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
   - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
    - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
   - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
   - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
   - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
   - Pie and bar graphs - Depict the distribution of a categorical variable.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
   - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
   - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
   - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
   - Pie and bar graphs - Depict the distribution of a categorical variable.
   - Histogram - Depict the distribution of a quantitative variable.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
    - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
    - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
    - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
    - Pie and bar graphs - Depict the distribution of a categorical variable.
    - Histogram - Depict the distribution of a quantitative variable.
    - Scatterplot - Depict the relationship between two quantitative variables.

Basic Process

Most policy research involves deceptively simple steps:

1. Define the question you would like answered.
2. State hypotheses about the answer to the question.
3. Collect data that can answer the question.
4. Calculate measures to test hypotheses put forward about the relationship of interest.
   - Mean, Median - Measures of central tendency; describes value of X or Y in a typical case.
   - Quartiles, Standard deviation - Measures of spread; describes range of values in sample/population and measures deviations from the typical case.
   - Pearson's "r" coefficient - Measure of covariance; describes the strength and direction of the relationship between two variables.
5. Organize and report results.
   - Pie and bar graphs - Depict the distribution of a categorical variable.
   - Histogram - Depict the distribution of a quantitative variable.
   - Scatterplot - Depict the relationship between two quantitative variables.
   - Two-way table - Joint distribution of two categorical variables.

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

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

1. 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?)
2. Identify a theoretically important relationship (e.g., Does monopoly control of a service increase prices? Does denser development reduce driving?)
3. 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?)

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.
   - Many of the statistics we've learned to calculate to date are useful for observational studies.
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...highly technical
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 not generalizable outside of the lab.

Note that good research questions often imply what design should be used.

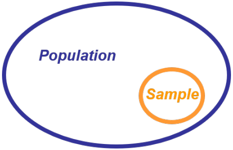
Collecting Data

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.
  - Used in observational studies and quasi-experiments.
- 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.



Designing a Sample

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. Examples: Some government surveys.
- Multiple stage random sample: select groups within a population in stages, resulting in a sample consisting of clusters of individuals. Examples: Many government studies.

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
- Multiple stage and stratified sampling can ensure representativeness with a smaller and less expensive sample; however, selection is not entirely random.
- 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, such as using a positive or negative lead to a question, or social desirability bias, such as lying about how much you study to look better for the researcher.

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**.
- Subjects, like all people, sometimes change their behavior when participating in a study, which could lead to **bias** known as the **placebo effect**.

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:

$$ATE = (\overline{Y_t} - \overline{Y_c}) \quad (1)$$

- Equation (1) 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.

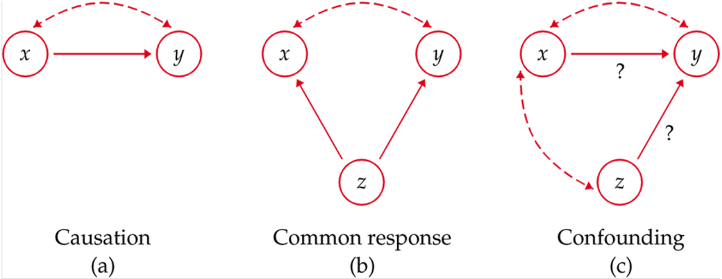
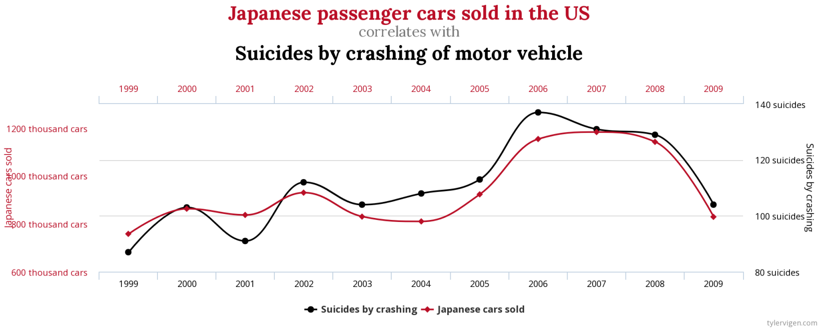


Figure 2.28
Introduction to the Practice of Statistics, Sixth Edition

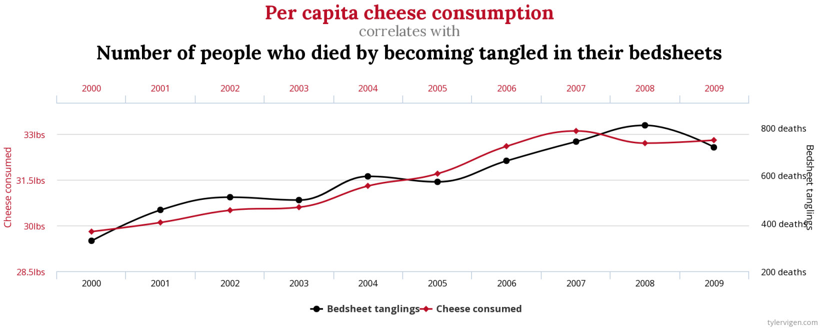
Spurious Correlations

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



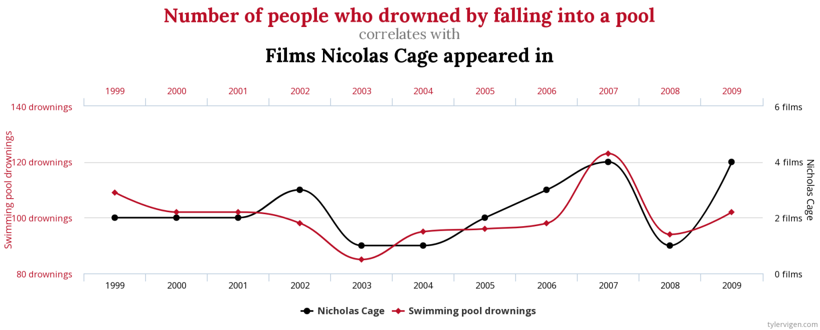
Credit: Tyler Vigen

Spurious Correlations



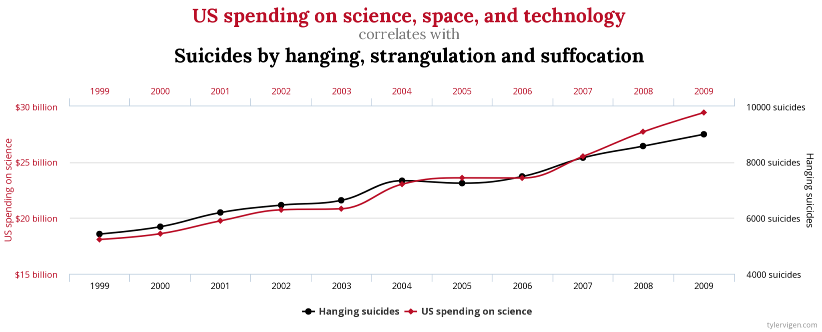
Credit: Tyler Vigen

Spurious Correlations



Credit: Tyler Vigen

Spurious Correlations



Credit: Tyler Vigen

So how do I randomize?

Randomizing comes in after you have a question, know your treatment, and have a sample. You follow these steps:

1. Use number to label experimental units
2. Use the table of random digits to select labels and assign units to two groups, treatment group and control group. Or in Excel, use function “=rand()”.
3. Only the treatment group will receive the treatment.
4. After the treatment, the changes of two groups are compared to determine the effect of the treatment.
5. Repeat the experiment with a different sample to see whether the results can be replicated.

Final Design Notes

- Observational studies are important and can tell us a lot about the world.
- BUT...there should be either a strong theoretical reason for a relationship or, even better, some strong experimental evidence to confirm it.
- Make sure that the design and sample you use match the question you are trying to answer.

Example

- RQ1: What classes do high school students take?

Example

- RQ1: What classes do high school students take?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools

Example

- RQ1: What classes do high school students take?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools
- Study design: Observational

Example

- RQ1: What classes do high school students take?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools
- Study design: Observational
- Statistics: Central tendency, spread, some joint distributions.

Example

- RQ2: Does taking calculus in high school affect college enrollment?

Example

- RQ2: Does taking calculus in high school affect college enrollment?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools OR administrative data OR voluntary participation from some high schools (if experiment)

Example

- RQ2: Does taking calculus in high school affect college enrollment?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools OR administrative data OR voluntary participation from some high schools (if experiment)
- Study design: Quasi-experimental or experimental.

Example

- RQ2: Does taking calculus in high school affect college enrollment?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools OR administrative data OR voluntary participation from some high schools (if experiment)
- Study design: Quasi-experimental or experimental.
- Treatment: Taking calculus

Example

- RQ2: Does taking calculus in high school affect college enrollment?
- Sample: Multiple stage sample - 1) Representative sample of high schools; 2) Random sample of students within schools OR administrative data OR voluntary participation from some high schools (if experiment)
- Study design: Quasi-experimental or experimental.
- Treatment: Taking calculus
- Statistics: Average Treatment Effect

Attendance