Methodological Tools for Public Policy

Introduction to Policy Research

Stephen B. Holt, Ph.D.

ROCKEFELLER COLLEGE
OF PUBLIC AFFAIRS & POLICY
UNIVERSITY AT ALBANY State University of New York

February 1, 2022

Introduction

- Research involves a variety of methodological approaches. Public policy typically clusters into two branches:
  - Qualitative research (focus groups, interviews - good for depth, **not covered here**)
  - Quantitative research (surveys, administrative data - good for breadth)
- Statistics in public policy
  - Describe systematically a body of information
  - Put intuitive ideas we have about a problem into empirical tests
  - Draw accurate inference about the world from a sample of data
  - Examine the influence of variables on some outcomes

Goals

- Learn statistical skills
  - Produce summary data from a sample
  - Analyze data for policy analysis
- Become a savvy consumer of statistical information
  - Think carefully about sources of data and how it affects conclusions drawn
  - Think about the methods used to analyze data
  - Consider the accuracy of how data are interpreted
  - Weigh whether conclusions drawn from data are accurate

Basic Process

Most policy research involves deceptively simple steps:

1. 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?).
2. 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).
3. 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).
4. 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).
5. Organize and report results.

Basic Terms

Statistical information is generally collected with three important dimensions: a unit of observation, individual observations, and variables.

- **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. Examples: inches, money, time, rankings
  - Categorical variables - captures a unit's grouping with other similar units. Examples: race and gender, species classifications, types of schools, employment status. Note: in many datasets, categorical variables are stored with numbers representing the different categories.

Example Dataset

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

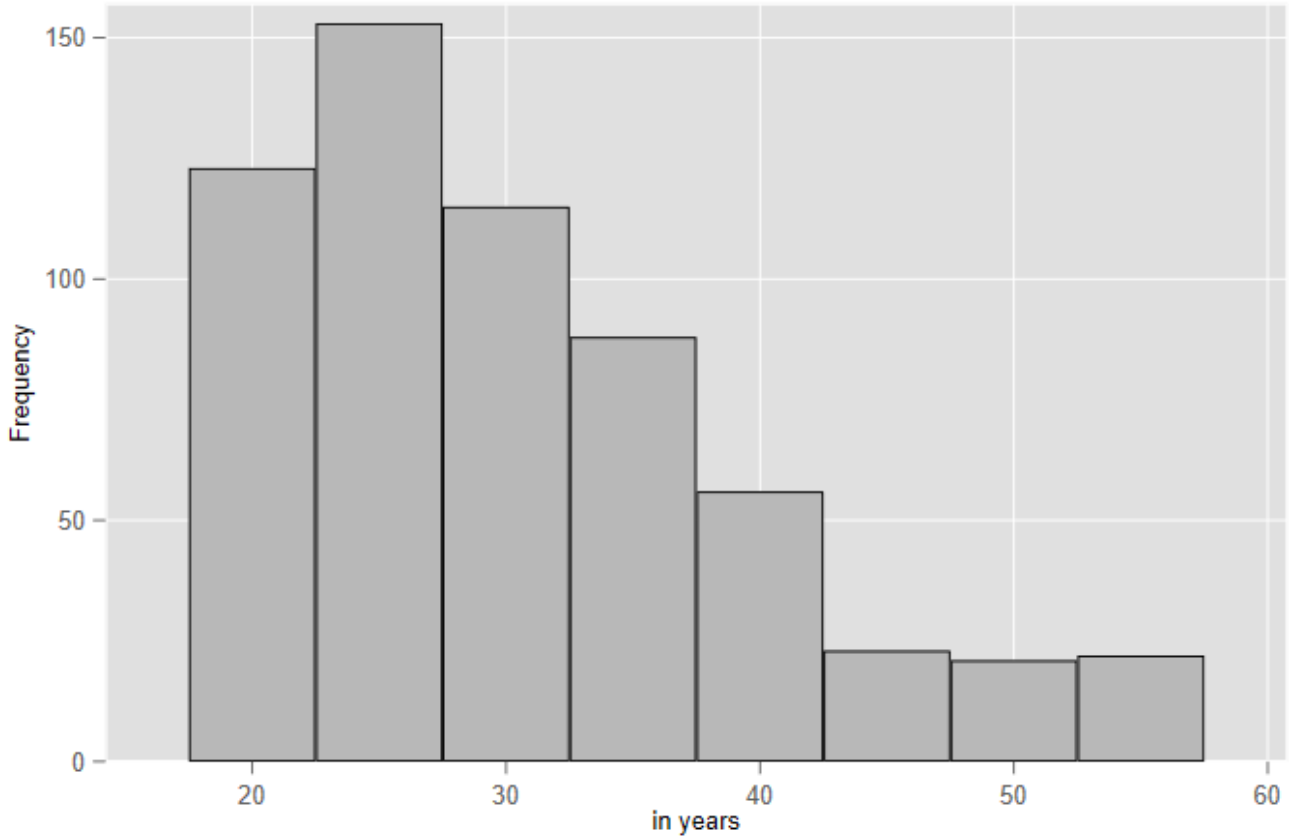
- Unit of observation: person
- Each row represents a person responding to a survey
- Each column is a different variable
- Categorical variables: male, relig (1 = anti-religious to 5 = very religious), occup, ratemarr (1 = very unhappy to 5 = very happy)
- Quantitative variables: age, yrsmarr, kids,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

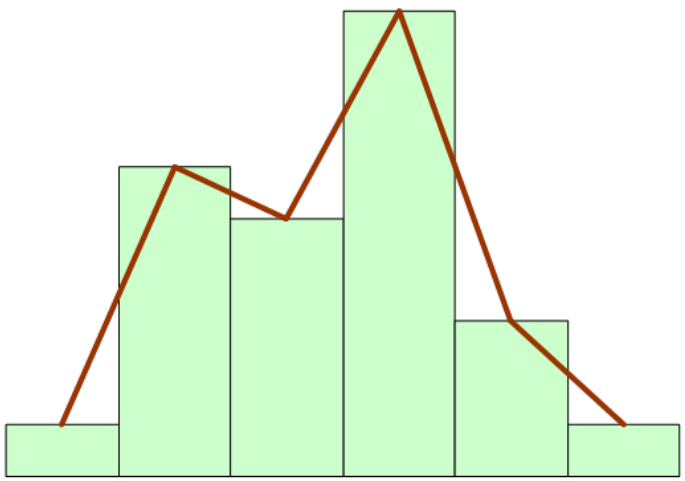
Histogram Example

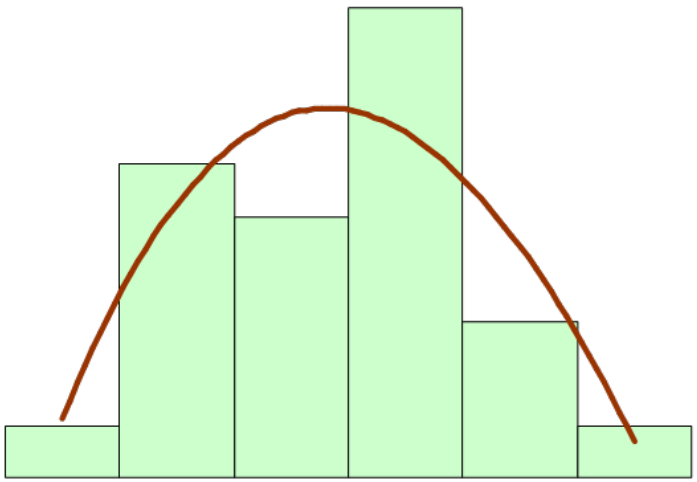
Below is a histogram of the age of our sample, ranging from 17.5 to 57, in 5 year intervals. The first bar suggests a little over 100 observations fall into the first interval, 17.5 to 22.5 years old. The last bar suggests there's fewer than 25 aged 52 to 57 in the sample.



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.

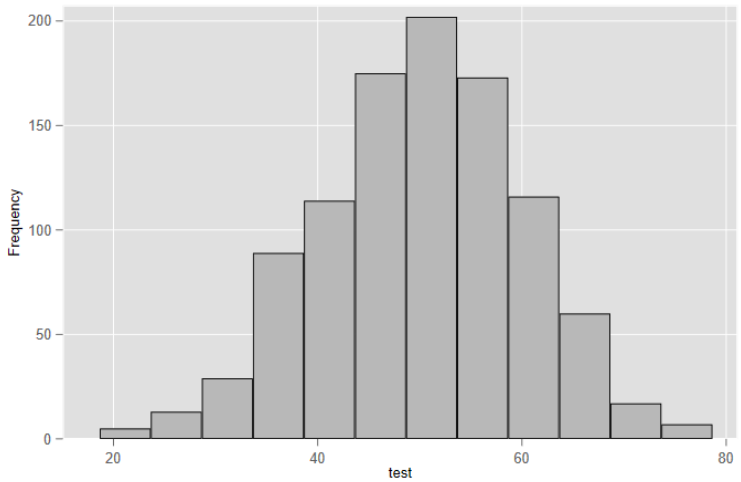




Histograms and Distributions

The shape of a distribution can tell us about how we might think about measures for our analysis. Common shapes that we will encounter in this class are symmetrical distributions, where the left and right sides of a histogram look about alike, and skewed, where one side of the distribution trails out farther than the other. Below is a distribution of test scores from a national sample of 12th grade students.

A normal, symmetrical distribution

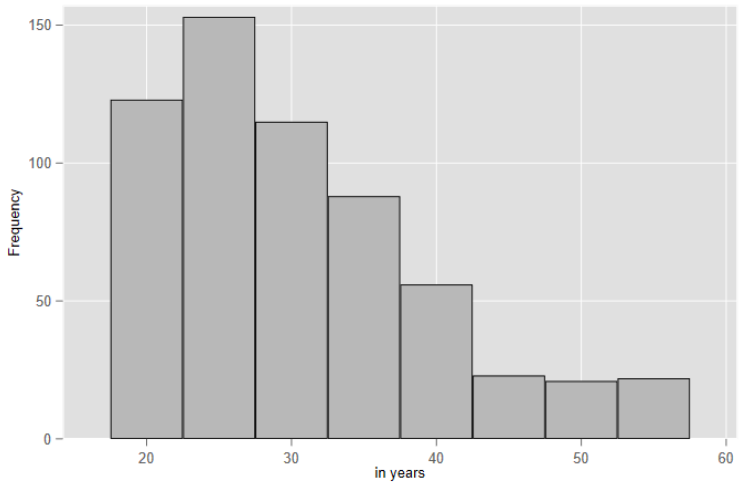


In normal and symmetrical distributions, like the one above, the mean and the median will be approximately equal.

Histograms and Distributions

A skewed distribution, such as the distribution of age from our example sample, has fewer bars on one side than the other. A distribution is skewed to the right if the side with larger values extends further than the smaller values. The distribution is skewed to the left if the smaller values are more represented than larger values.

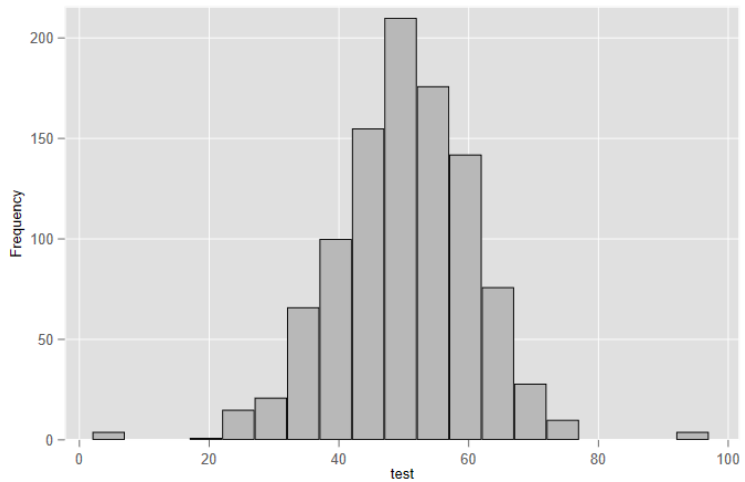
A right-skewed distribution



When there is a skew present, we know that the mean will be different from the median and will move in the direction of the skew.

Histograms and Distributions

Finally, histograms can help us spot potential outliers in our sample. Outliers reflect an important deviation from the center of a distribution because they lie outside the normal pattern of the variable in the population. If they are extreme enough, they can distort our understanding of a policy-related phenomenon. For instance, what if our sample of 12th graders had a test score distribution like the one below?



Measures of Center: Mean

- 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

Measures of Center: Mean

Expressed arithmetically:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \quad (1)$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (2)$$

$$\overline{x} = \frac{188.834}{25} = 7.553 \quad (3)$$

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 Center: Median

	yrsmarr
1	.75
2	.75
3	.75
4	.75
5	.75
6	1.5
7	1.5
8	1.5
9	1.5
10	1.5
11	4
12	6
13	7
14	7
15	7
16	7
17	10
18	10
19	10
20	10
21	10
22	15
23	15
24	15
25	15

- The **median**, by contrast, is the midpoint of a distribution-the value of a variable such that half of the observations are smaller and half are larger.
  1. Sort the observations by value of the variable, from smallest to largest. n = number of observations.
  2. $\Leftarrow$ If n is **odd**, the median is observation $(n+1)/2$ down the list. n = 25, $(25+1)/2 = 26/2 = 13$, median is 7.
  3. $\Rightarrow$ If n is **even**, the median is the mean of the middle two observations. n = 24, $24/2 = 12$, *Median* $= (6+7)/2 = 13/2 = 6.5$

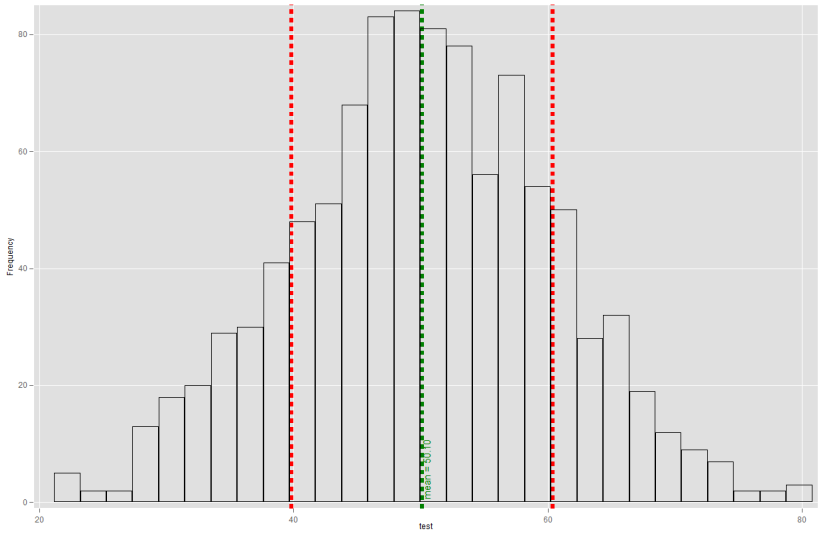
	yrsmarr
1	.75
2	.75
3	.75
4	.75
5	.75
6	1.5
7	1.5
8	1.5
9	1.5
10	1.5
11	4
12	6
13	7
14	7
15	7
16	7
17	10
18	10
19	10
20	10
21	10
22	15
23	15
24	15
25	15

Measures of Spread: Quartiles

	yrsmarr
1	.75
2	.75
3	.75
4	.75
5	.75
6	1.5
7	1.5
8	1.5
9	1.5
10	1.5
11	4
12	6
13	7
14	7
15	7
16	7
17	10
18	10
19	10
20	10
21	10
22	15
23	15
24	15
25	15

- Finally, in addition to a shape and center, every variable's distribution has a spread.
- The spread tells us how far from the center observations will fall in our sample. In short, measures of spread help us better understand how much our sample varies in values exists in our data on the variable of interest to us.
- Quartiles provide one measure to assess spread, and they can be calculated as the median of the bottom half of data and top half of the data, excluding the median.
- The first quartile is the mean of observation 6 and observation 7, which is the center of the bottom 12 observations. As it happens, this comes to 1.5 years, and means 25% of observations have been married for less than 1 and a half years.
- The top quartile is the mean of observations 19 and 20, which is the center of the top 12 observations, and works out to 10 years of marriage. 75% of the sample has been married for less than 10 years.

Measures of Spread: Standard Deviation



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 Calculation

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

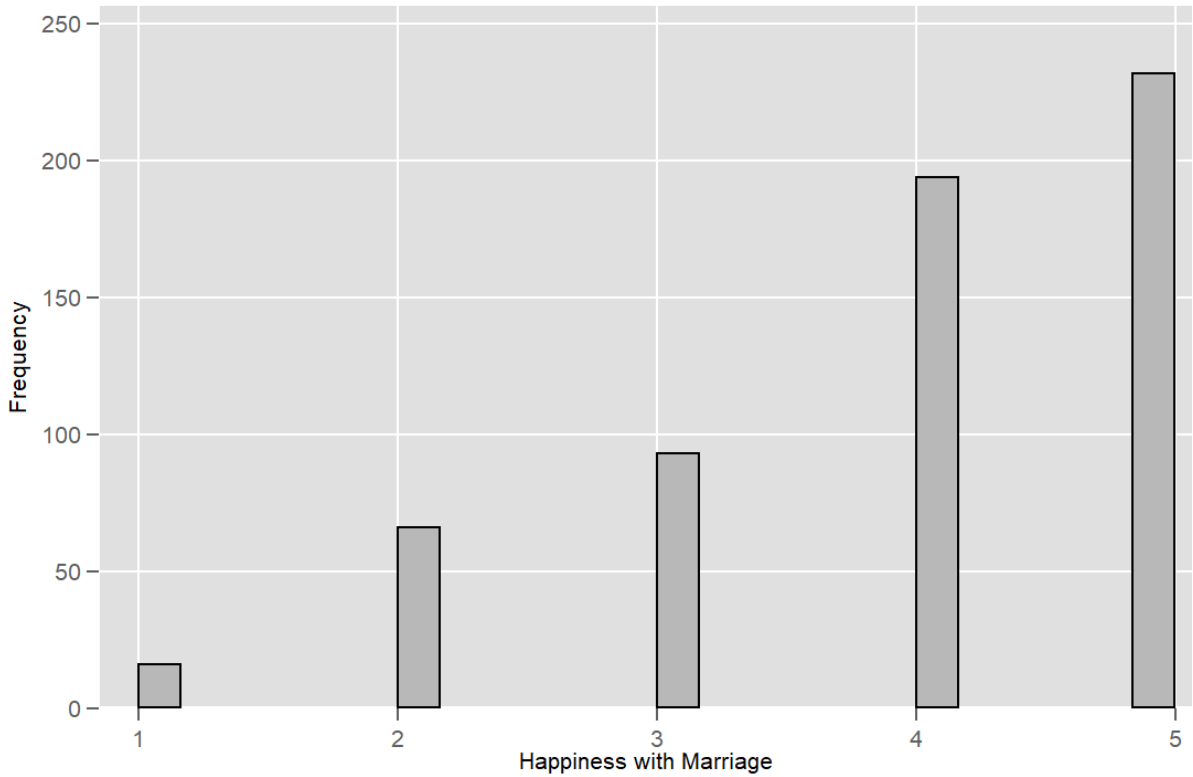
Mean: 8.1777
Sum of squared deviations from the mean: 716.2292
Degrees of freedom:
$df = n - 1 = 14$
s^2: $716.2292/14 = 51.1592$
s: $\sqrt{51.1592} = 7.1526$

i	yrsmarr (x)	$\bar{x}$	$(x - \bar{x})$	$(x - \bar{x})^2$
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:

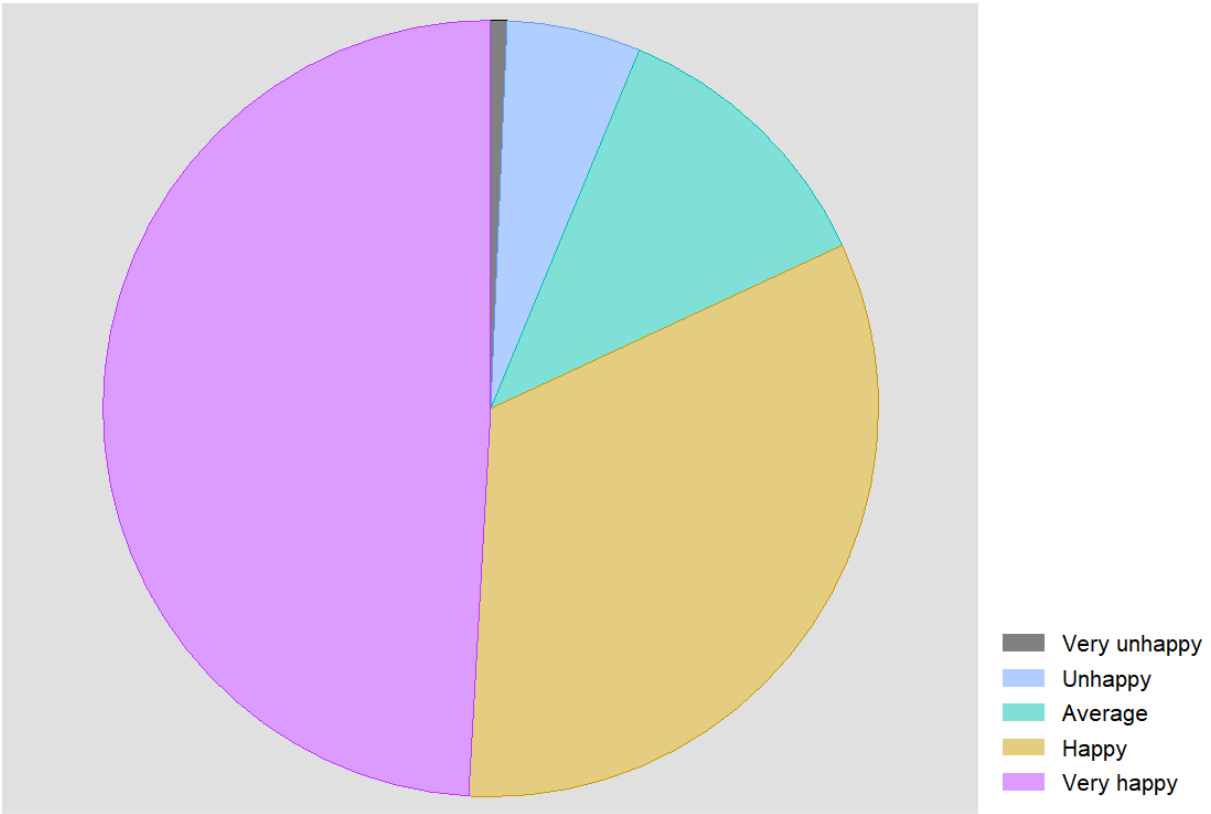
Respondents' happiness with their marriage



Visualizing Categorical Variables

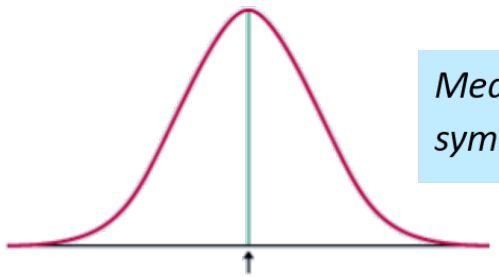
...or pie graphs, where slices represent each category's proportion of the whole sample:

Respondents' happiness with their marriage

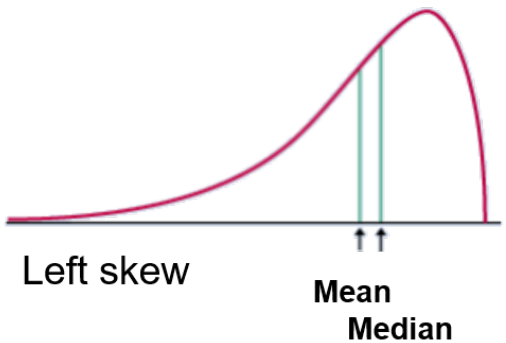


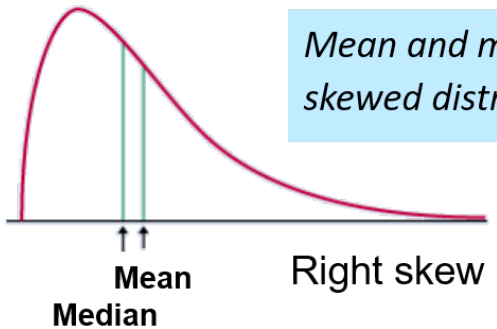
Visualizing Skew

Finally, now that we know how to calculate measures of center, here is a quick visual guide for the link between distribution shape and measures of center. Remember, the mean and median are only the same when the distribution is symmetrical, and the mean is influenced by skew and outliers while the median is not.



Mean and median for a symmetric distribution



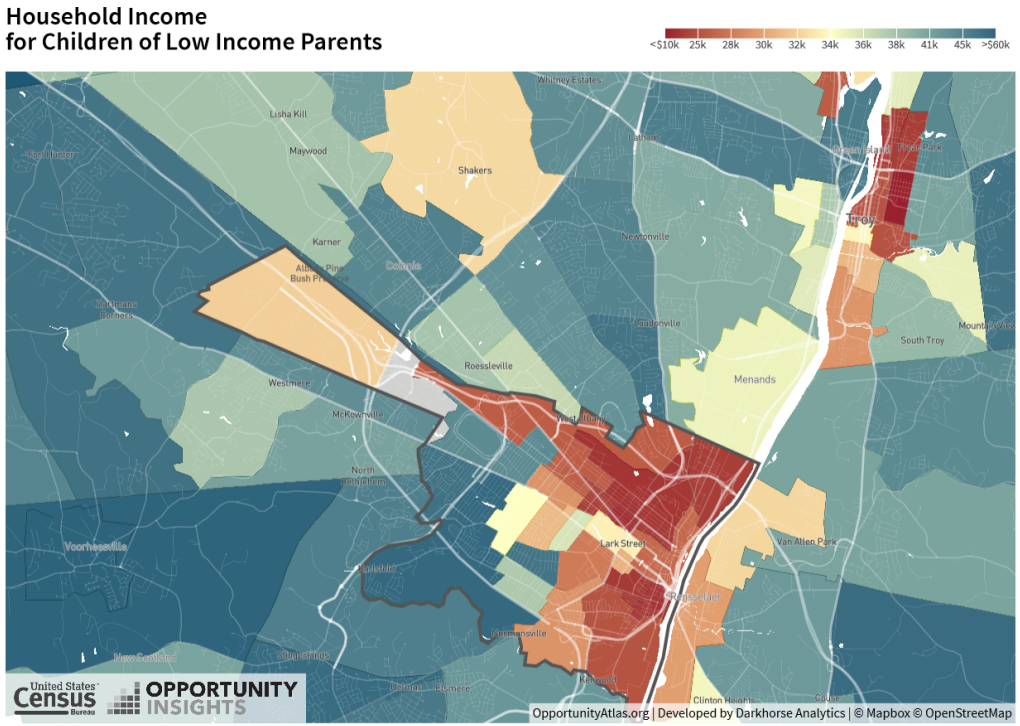


Mean and median for skewed distributions

Simple Averages Identifying Issues

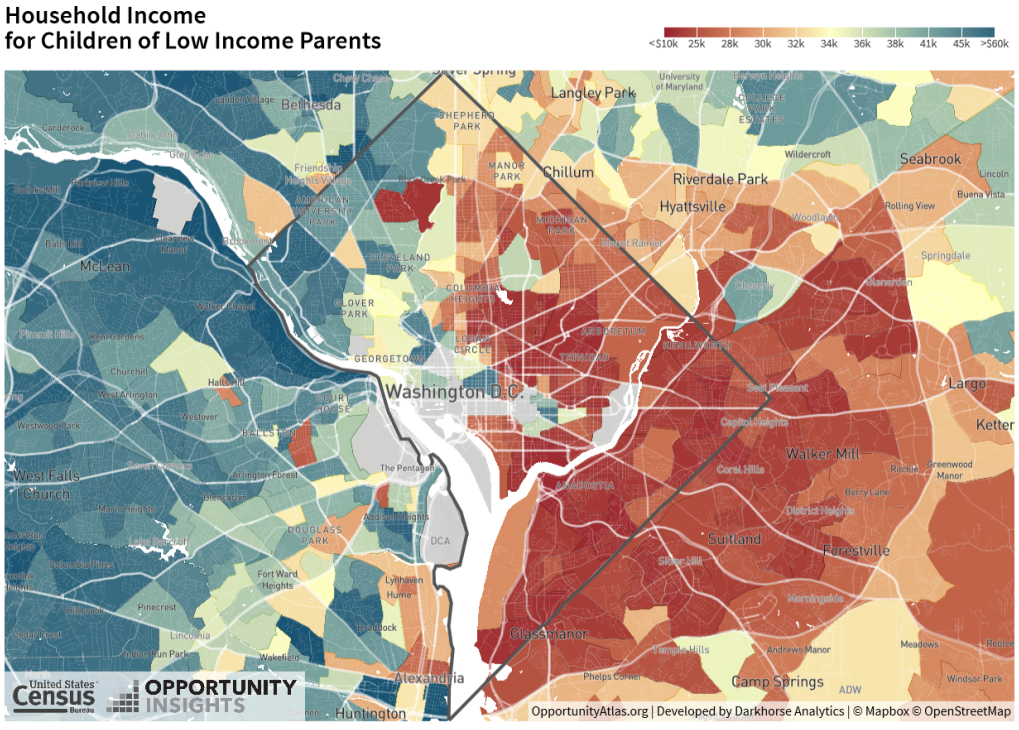
Example: Socioeconomic Mobility

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.



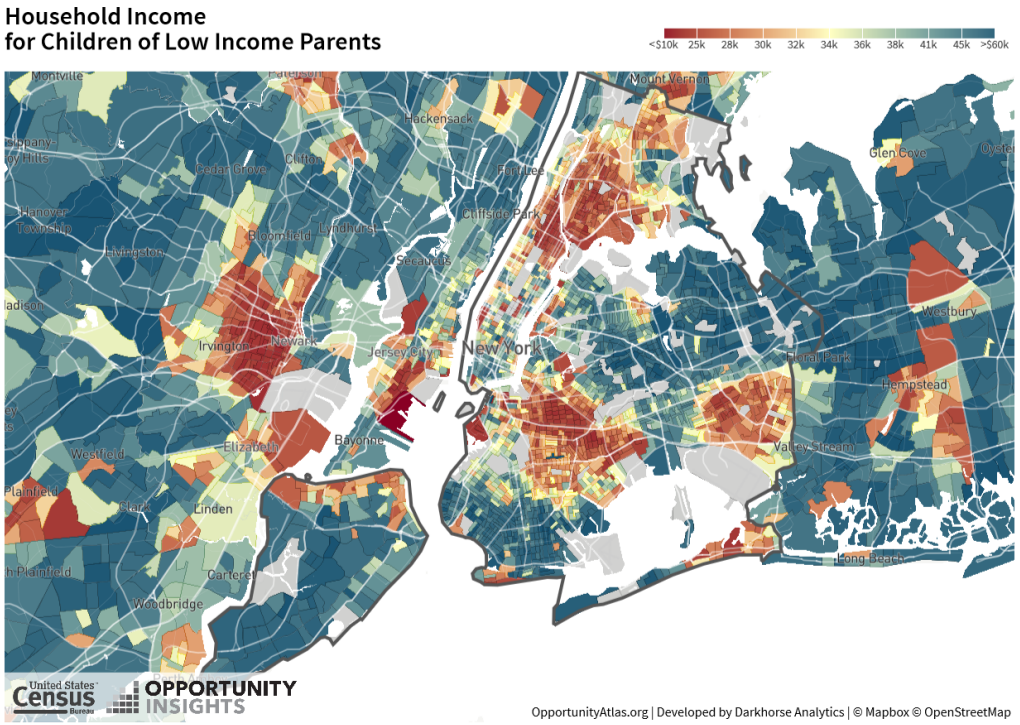
Simple Averages Identifying Issues

Example: Socioeconomic Mobility
Chetty et al. continued



Simple Averages Identifying Issues

Example: Socioeconomic Mobility
Chetty et al. continued

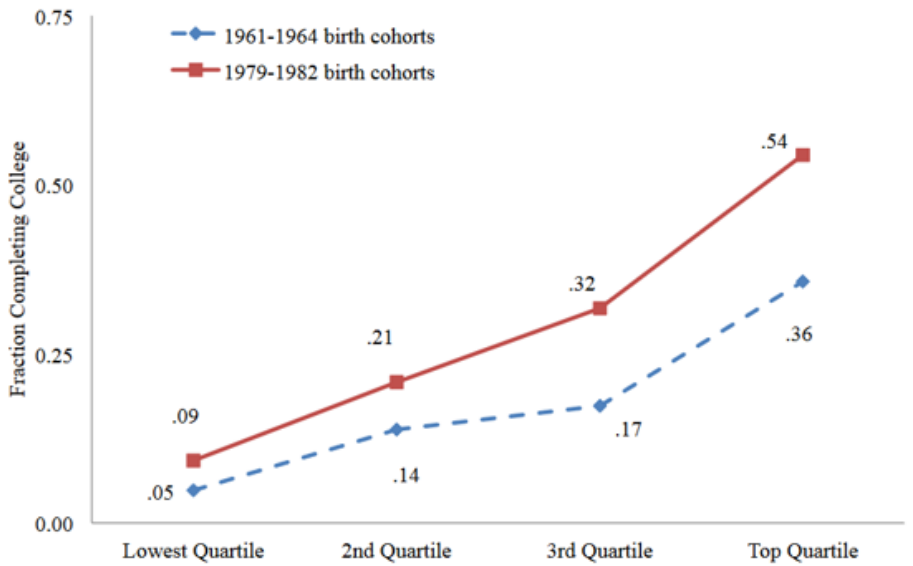


More at: Raj Chetty et al.'s Opportunity Atlas

Using Means and Spread

Using longitudinal data, Bailey and Dynarski look at how college completion rates have changed by income quartile over time.

Figure 3: Fraction of Students Completing College, by Income Quartile and Year of Birth



Source: Author's calculation based on data from the National Longitudinal Survey of Youth, 1979 and 1997 (U.S. Bureau of Labor Statistics, 2010a, 2010b).

Martha J. Bailey and Susan M. Dynarski. 2011. Gains and Gaps: Changing Inequality in U.S. College Entry and Completion. NBER Working Paper No. 17633.