More Relationships

Stephen B. Holt, Ph.D.

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

February 15, 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.

Categorical Variables

Categorical variables don't necessarily make sense in scatter plots. Observations stack into a limited number of values, and often those values stand-in for a different meaning than the number represented in the dataset (e.g., race or generation or education level).

Often, researchers are interested in the relationship between two categorical variables. For instance, have education levels changed across generations?

To answer this question, a researcher would use a two-way, or block, study design. A two-way design uses two categorical factors with several levels for both factors to answer the question. Here, generations are often defined using categories of ages (a proxy for birth cohorts) and education can be categorized by the highest degree a person has completed.

Two-way Tables

The researcher would descriptively answer the research question using a **two-way table**.
First factor, age grouping, defines the columns.
Second factor, education level, defines the rows.

	Age Group				
Education level	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
HS diploma	123462	116768	258297	428349	926876
Some college	94738	94191	181058	223464	593451
4-year college+	30534	146423	284465	309227	770649
Total	275728	384080	793209	1077709	2530726

Reading Two-Way Tables

- We call education the **row variable** and age group the **column variable**.
- Each combination of values for these two variables is called a cell.
- For each cell, we can compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions would be the **joint distribution** of the two variables.

	Age Group				
Education level	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
HS diploma	123462	116768	258297	428349	926876
Some college	94738	94191	181058	223464	593451
4-year college+	30534	146423	284465	309227	770649
Total	275728	384080	793209	1077709	2530726

Marginal Distributions

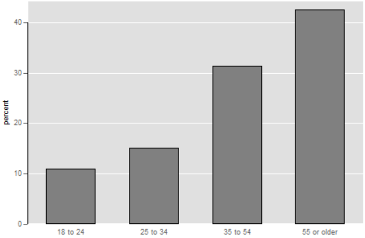
We can look at each categorical variable separately in a two-way table by studying the row totals and the column totals. They represent the **marginal distributions**, expressed in counts or percentages. (They are written as if in a margin.)

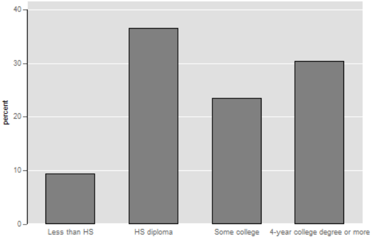
	Age Group				
Education level	18 to 24	25 to 34	35 to 54	55 or older	**Total**
Less than HS	26994	26698	69389	116669	**239750**
HS diploma	123462	116768	258297	428349	**926876**
Some college	94738	94191	181058	223464	**593451**
4-year college+	30534	146423	284465	309227	**770649**
Total	*275728*	*384080*	*793209*	*1077709*	2530726

Marginal Distributions

When we use bar graphs (or pie graphs) to show the distribution of a categorical variable, it captures the equivalent of the marginal distribution of that variable, and the marginal distribution is typically expressed in terms of percent of the total rather than a strict count of observations.

			Age Group		
Education level	18 to 24	25 to 34	35 to 54	55 or older	**Total**
Less than HS	26994	26698	69389	116669	**239750**
HS diploma	123462	116768	258297	428349	**926876**
Some college	94738	94191	181058	223464	**593451**
4-year college+	30534	146423	284465	309227	**770649**
Total	*275728*	*384080*	*793209*	*1077709*	2530726





Conditional Distribution

- In the table below, the 25 to 34 age group occupies the second column. To find the complete distribution of education in this age group, look only at that column. Compute each count as a percent of the column total.
- These percents should add up to 100% because all persons in this age group fall into one of the education categories. These four percents together are the conditional distribution of education, given the 25 to 34 age group.

			Age Group		
Education level	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
HS diploma	123462	116768	258297	428349	926876
Some college	94738	94191	181058	223464	593451
4-year college+	30534	146423	284465	309227	770649
Total	275728	384080	793209	1077709	2530726

Conditional Distributions

The percents within the table represent the conditional distributions. Comparing the conditional distributions allows you to describe the "relationship" between both categorical variables. $C.D. = \frac{cell}{columntotal}$

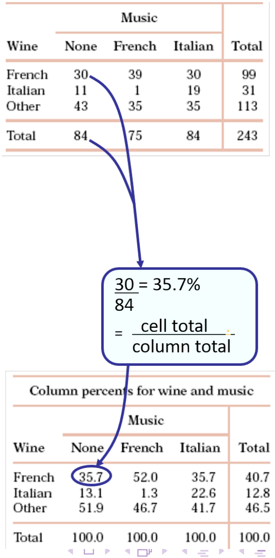
	Age Group				
Education	18 to 24	25 to 34	35 to 54	55 or older	Total
Less than HS	26994	26698	69389	116669	239750
	(9.79)	(6.95)	(8.75)	(10.83)	(9.47)
HS diploma	123462	116768	258297	428349	926876
	(44.78)	(30.40)	(32.56)	(39.75)	(36.62)
Some college	94738	94191	181058	223464	593451
	(34.36)	(24.52)	(22.83)	(20.74)	(23.45)
4-year college+	30534	146423	284465	309227	770649
	(11.07)	(38.12)	(35.86)	(28.69)	(30.45)
Total	275728	384080	793209	1077709	2530726
	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)

Example

	Level of Student				
Pet preferences	Freshmen	Sophomore	Junior	Senior	Total
Cat	0	3	3	2	8
	(0.00)	(75.00)	(37.50)	(33.33)	(40.00)
Dog	2	0	3	4	9
	(100.00)	(0.00)	(37.50)	(66.67)	(45.00)
Fish	0	1	0	0	1
	(0.00)	(25.00)	(0.00)	(0.00)	(5.00)
Other	0	0	1	0	1
	(0.00)	(0.00)	(12.50)	(0.00)	(5.00)
Reptile	0	0	1	0	1
	(0.00)	(0.00)	(12.50)	(0.00)	(5.00)
Total	2	4	8	6	20
	(100.00)	(100.00)	(100.00)	(100.00)	(100.00)
N	20				

Music and Wine Purchase Decisions

- What is the relationship between type of music played in supermarkets and type of wine purchased?
- We want to compare the conditional distributions of the response variable (wine purchased) for each value of the explanatory variable (music played). Therefore, we calculate column percents.
- Calculations: When no music was played, there were 84 bottles of wine sold. Of these, 30 were French wine. $30/84 = 0.357 \rightarrow 35.7\%$ of the wine sold was French when no music was played.



	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

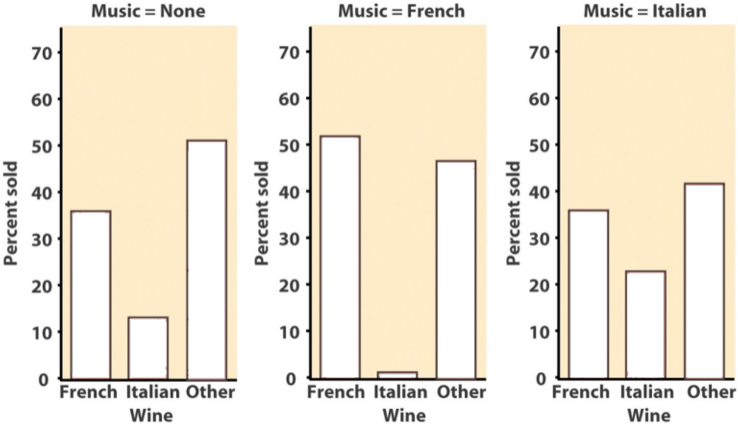
$\frac{30}{84} = 35.7\%$

$= \frac{\text{cell total}}{\text{column total}}$

Column percents for wine and music

	Music			
Wine	None	French	Italian	Total
French	35.7	52.0	35.7	40.7
Italian	13.1	1.3	22.6	12.8
Other	51.9	46.7	41.7	46.5
Total	100.0	100.0	100.0	100.0

Does background music affect wine purchases?



Caution with Association

- As we introduced last week, associations can be biased. This is true for categorical variables as well. Simpson's paradox provides one example of how relationships alone can be unintentionally misleading.
- **Simpson's Paradox**: An association or comparison that holds for all of several groups can reverse direction when the data are combined (aggregated) to form a single group.

	Day 1	Day 2	Total
Person A	63/90 (70%)	4/10 (40%)	67/100 (67%)
Person B	8/10 (80%)	45/90 (50%)	53/100 (53%)

Simpon's Paradox Examples

- Some analyses show men accepted to colleges at higher rates then women. However, each college accepts a higher share of women than men.
- A political party can receive more overall votes in a state and still lose the majority of individual districts in the state legislature.
- Generally, these incidents have to do with how much weight (i.e., the relative number of observations) a particular category has in an analysis.

Attendance