

Social Research Methods: Qualitative and Quantitative Approaches

Eighth Edition



Chapter 12

Analysis of Quantitative Data

Learning Objectives

- 12.1** Organize raw data into appropriate formats for statistical analysis.
- 12.2** Analyze data using descriptive statistical methods.
- 12.3** Analyze data using bivariate statistical methods.
- 12.4** Analyze data relationships using multivariate techniques.
- 12.5** Analyze data using inferential statistical methods.

12.1 Dealing with Data (1 of 6)

Objective: Organize raw data into appropriate formats for statistical analysis.

- Coding Data
 - Data coding means systematically reorganizing raw data into a format that statistics software on computers can use.
 - Researchers create and consistently apply rules for transferring information from one form to another.
 - The **coding procedure** is a set of rules stating that you will assign certain numbers to variable attributes.

12.1 Dealing with Data (2 of 6)

- Coding Data

- A **codebook** is a document describing the coding procedure and the computer file location of data for variables in a specific format.
- Researchers should begin to think about a coding procedure and codebook before collecting any data.

Many researchers precode a questionnaire before collecting any data.

If precoding is not done, the first step after collecting data is to create a codebook.

12.1 Dealing with Data (3 of 6)

- Entering Data
 - In a grid, each row represents a respondent, participant, or case. In computer terminology, these are called **data records**.
 - Four ways to enter raw quantitative data into a computer:
 - Code sheet
 - Direct-method (including CATI)
 - Optical scan
 - Bar code or QR code

12.1 Dealing with Data (4 of 6)

- Cleaning Data
 - Errors made when coding or entering the data into a computer threaten the validity of the measures and cause misleading results.
 - Cleaning the data is the process of checking the accuracy of the coding of the data.
 - Many researchers code 10-15 percent of the data a second time. If there are no coding errors in the recorded sample, the researcher can proceed. If not, all the coding must be checked.

12.1 Dealing with Data (5 of 6)

Figure 12.1 Coded Data for Three Cases and Codebook

EXCERPT FROM SURVEY QUESTIONNAIRE

Respondent ID _____ Interviewer Name _____

Note the Respondent's Gender: _____ Male _____ Female

1. The first question is about the President of the United States. Do you Strongly Agree, Agree, Disagree, Strongly Disagree, or Have No Opinion About the following statement:

The President of the United States is doing a great job.

_____ Strongly Agree _____ Agree _____ Disagree _____ Strongly Disagree _____ No Opinion

2. How old are you? _____

EXCERPT OF CODED DATA

Column

0000000001111111112222222233333333444... etc. (tens)

123456789012345678901234567890123456789012... etc. (ones)

01 212736302 182738274 10239 18.82 3947461... etc.

02 213334821 124988154 21242 18.21 3984123... etc.

03 420123982 113727263 12345 17.36 1487645... etc.

etc.

Raw data for first three cases, columns 1 through 42.

EXCERPT FROM CODEBOOK

Column	Variable Name	Description
1–2	ID	Respondent identification number
3	BLANK	
4	Interviewer	Interviewer who collected the data: 1 = Susan 2 = Carlos 3 = Juan 4 = Sophia 5 = Clarence
5	Gender	Interviewer report of respondent's sex 1 = Male, 2 = Female
6	Pres.Job	The President of the United States is doing a great job. 1 = Strongly Agree 2 = Agree 3 = No Opinion 4 = Disagree 5 = Strongly Disagree Blank = Missing Information

12.1 Dealing with Data (6 of 6)

- Descriptive Statistics

- Two definitions of statistics

- A collection of numbers used in existing statistics research

- A branch of applied mathematics with techniques to manipulate and summarize numbers

- Descriptive statistics allow us to describe numerical data.

- We can categorize statistics by the number of variables involved.

12.2 Results with One Variable (1 of 4)

Objective: Analyze data using descriptive statistical methods.

- **Frequency distribution** displays the number of times each outcome in a sample occurred.
- Common Graphic Representations of Univariate Data
 - Histogram
 - Bar chart
 - Pie chart

12.2 Results with One Variable (2 of 4)

- Measures of Central Tendency

- **Mode**

- Easiest to use. It is the most common and frequently occurring number.

- **Median**

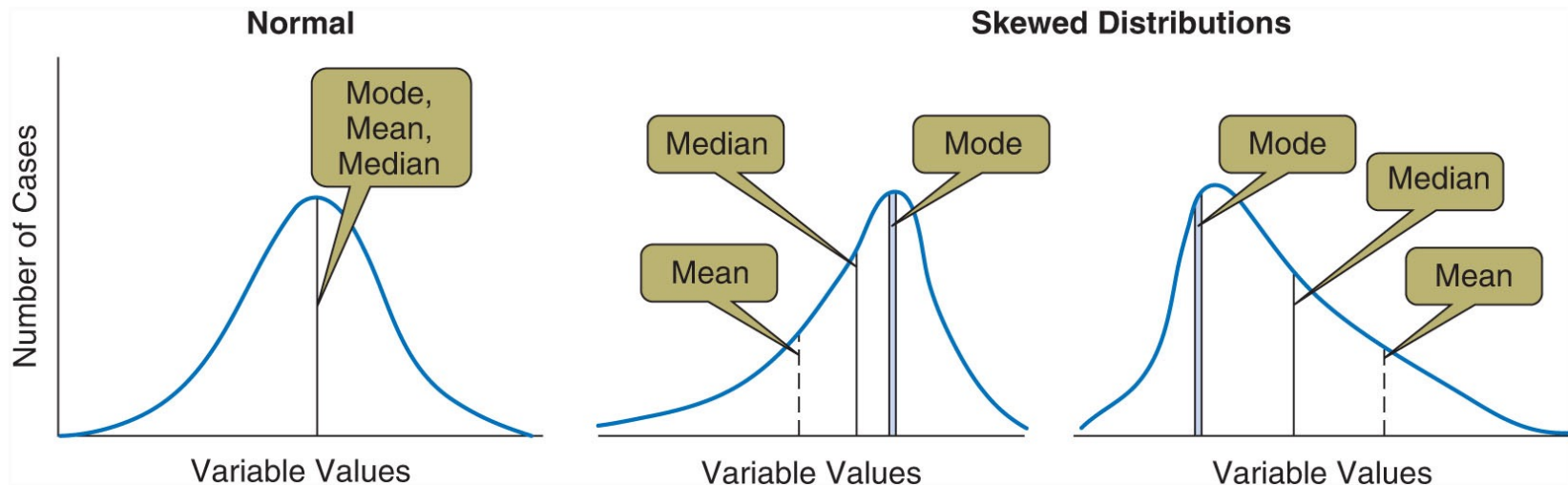
- The middle point. It is the point at which half of the cases are above it, and half of the cases are below it.

- **Mean**

- Also called arithmetic average

12.2 Results with One Variable (3 of 4)

Figure 12.3 Three Measures of Central Tendency in Normal and Skewed Distributions



12.2 Results with One Variable (4 of 4)

- Measures of Variation

- **Range**

- The largest and smallest scores

- **Percentiles**

- Describes the specific place within a distribution

- **Standard deviation**

- A measure of how spread out the numbers are

- It is of limited usefulness by itself.

- We use standard deviation and the mean to create **z-scores**.

12.3 Results with Two Variables (1 of 14)

Objective: Analyze data using bivariate statistical methods.

- A Bivariate Relationship
 - **Bivariate statistics** let us consider two variables together and describe the relationship between the variables.
 - Bivariate statistical analysis shows a **statistical relationship** between variables.

Covariation means that things are associated.

Statistical independence means there is no association between variables.

12.3 Results with Two Variables (2 of 14)

- Scattergram

- A **scattergram** (or scatterplot) is a graph on which you plot each case or observation.

Each axis represents the value of one variable.

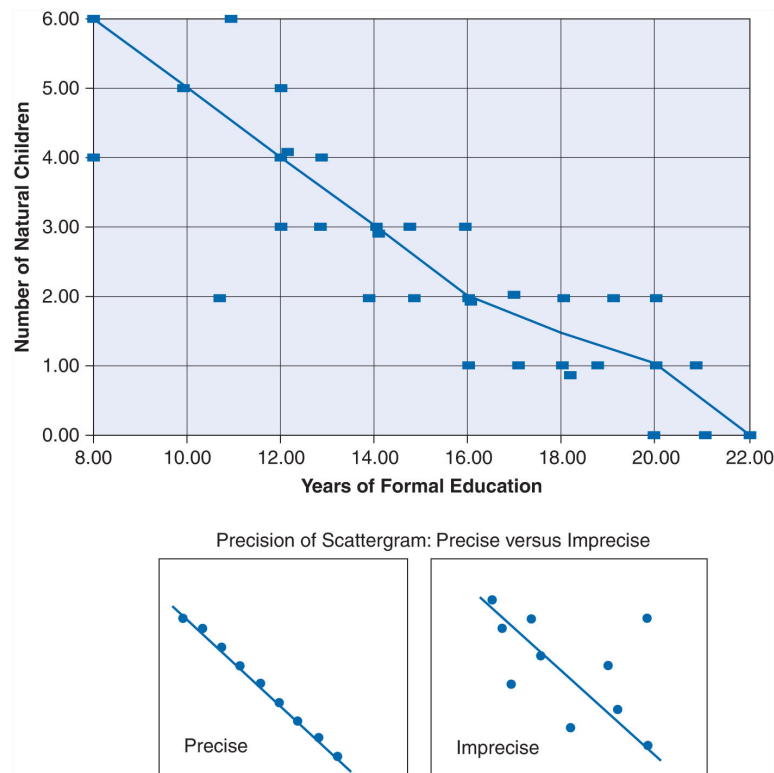
It is used for variables measured at the interval or ratio level.

- Constructing a Scattergram

- Form
- Direction
- Precision

12.3 Results with Two Variables (3 of 14)

Figure 12.5 Example of a Scattergram: Years of Education by Number of Natural Children for 33 Women



12.3 Results with Two Variables (4 of 14)

- Bivariate Tables
 - A bivariate **contingency table** presents the same information in a scattergram in a more condensed form.

Crosstabulation—tabulating two or more variables simultaneously

12.3 Results with Two Variables (5 of 14)

- Constructing a Percentaged Table

Example Box 12.2 Moving from Frequency Distributions to a Contingency Table*

1: Frequency Distributions of Two Variables

Health	Number of Cases
Excellent	228
Good	449
Fair	227
Poor	61
Total	965

Happiness Past Week	Number of Cases
None/Almost None of Time	32
Some of Time	207
Most of Time	479
All/Almost All of Time	247
Total	965

12.3 Results with Two Variables (6 of 14)

Example Box 12.2 [Continued]

2: Compound Frequency Distribution for Health and Week's Happiness

Health	Happiness Past Week	Number of Cases	
Excellent	None	2	228
Excellent	Some	29	
Excellent	Most	109	
Excellent	All	88	
Good	None	7	449
Good	Some	84	
Good	Most	239	
Good	All	119	
Fair	None	18	227
Fair	Some	65	
Fair	Most	111	
Fair	All	33	
Poor	None	5	61
Poor	Some	29	
Poor	Most	20	
Poor	All	7	

12.3 Results with Two Variables (7 of 14)

Example Box 12.2 [Continued]

3: Contingency Table of Health by Happiness Past Week (Count) (a) Happiness Past Week (b)

Happiness Past Week (b)

Health (b)	None	Some	Most	All	Total (c)
Excellent	2	29	109	88	228
Good	7(d)	84	239	119	449
Fair	18	65	111	33	227
Poor	5	29	20	7	61
Total (c)	32	207(e)	479	247	965

Don't Know/No Answer responses, Health = Happiness

12.3 Results with Two Variables (8 of 14)

Example Box 12.2 [Continued]

4: The Parts of the Contingency Table

5. Contingency Tables, Percentage calculated by Row or Column, Happiness in Past Week by Health

Column-Percentaged Table

Happiness Past Week

Health (b)	None	Some	Most	All	Total (c)
Excellent	6.3%	14.0%	22.8%	35.6%	23.6%
Good	21.9%	40.6%	49.9%	48.2%	449
Fair	56.3%	31.4%	23.2%	13.4%	23.5%
Poor	15.6%	14.0%	4.2%	2.8%	6.3%
Total (c)	100%	100%	100%	100%	100%
	(32)	(207)	(479)	(247)	(965)

12.3 Results with Two Variables (9 of 14)

Example Box 12.2 [Continued]

Row-Percentaged Table

Happiness Past Week

Health (b)	None	Some	Most	All	Total (c)	(N)
Excellent	0.9%	12.7%	47.8%	38.6%	100%	(228)
Good	1.6%	18.7%	53.2%	26.5%	100%	(449)
Fair	7.9%	28.6%	48.9%	14.5%	100%	(227)
Poor	8.2%	47.5%	32.8%	11.5%	100%	(61)
Total	3.3%	21.5%	49.6%	25.6%	100%	(965)

* Data are from the 2016 General Social Survey (GSS). A national sample of adult Americans.

Smith, Tom W., Peter Marsden, Michael Hout & Jibum Kim. 2016. **General Social Surveys, 1972–2016**, National Opinion Research Corporation. Chicago: NORC at the University of Chicago.

12.3 Results with Two Variables (10 of 14)

- Bivariate Tables Without Percentages

Table 12.1A Age by Schooling

Years of Schooling

Age	0–11	12	13–14	16+	Total
Under 30	5%	25	30	40	100
30–45	15	25	40	20	100
46–60	35	45	12	8	100
61+	45	35	15	5	100

Table 12.1B Age by Schooling

Years of Schooling

Age	0–11	12	13–14	16+	Total
61+	45%	35	15	5	100
46–60	35	45	12	8	100
30–45	15	25	40	20	100
Under 30	5	25	30	40	100

12.3 Results with Two Variables (11 of 14)

- Measures of Association
 - The measure of association is a single number that expresses the strength, and often the direction, of a relationship.
 - It condenses information about a bivariate relationship into a single number.
 - Many different measures of association
 - The correct one to use depends on the level of measurement of the data and specific research purposes.

12.3 Results with Two Variables (12 of 14)

Table 12.3 Six Measures of Association

- **Cramér's V** is used for nominal level data to indicate how strongly two categorical variables are associated. It ranges from 0 to 1.0 and can be used for any sized table.
- **Lambda** is used for nominal-level data. It is based on a reduction in errors based on the mode and ranges between 0 (independence) and 1.0 (perfect prediction or the strongest possible relationship).
- **Gamma** is used for ordinal-level data. It is based on comparing pairs of variable categories and seeing whether a case has the same rank on each. Gamma ranges from -1.0 to $+1.0$ with zero meaning no association.
- **Tau** is also used for ordinal-level data. It is based on a different approach than gamma and takes care of a few problems that can occur with gamma. Actually, there are several statistics named tau (it is a popular Greek letter), and the one here is Kendall's tau. Kendall's tau ranges from -1.0 to $+1.0$, with zero meaning no association.

12.3 Results with Two Variables (13 of 14)

Table 12.3 [Continued]

Rho is also called Pearson's product moment correlation coefficient (named after the famous statistician Karl Pearson and based on a product moment statistical procedure). It is the most commonly used measure of correlation, the correlation statistic people mean if they use the term correlation without identifying it further. It can be used only for data measured at the interval or ratio level. Rho is used for the mean and standard deviation of the variables and tells how far cases are from a relationship (or regression) line in a scatterplot. Rho ranges from -1.0 to $+1.0$ with zero meaning no association. If the value of rho is squared, sometimes

called *R*-squared R^2 it has a unique proportion reduction in error meaning. *R*-squared tells how the percentage in one variable (e.g., the dependent) is accounted for, or explained by, the other variable (e.g., the independent). Rho measures linear relationships only. It cannot measure nonlinear or curvilinear relationships. For example, a rho of zero can indicate either no relationship or a curvilinear relationship.

Chi-square has two different uses. It can be used as a measure of association in descriptive statistics like the others listed here or in inferential statistics. As a measure of association, chi-square can be used for nominal and ordinal data. It has an upper limit of infinity and a lower limit of zero, meaning no association.

12.3 Results with Two Variables (14 of 14)

Table 12.3 [Continued]

Summary Of Measures Of Association

Measure	Greek Symbol	Type of Data	High Association	Independence
Cramer's V	V	Nominal	1.0	0
Lambda	λ	Nominal	1.0	0
Gamma	γ	Ordinal	+1.0, -1.0	0
Tau (Kendall's)	τ	Ordinal	+1.0, -1.0	0
Rho	ρ	Interval, ratio	+1.0, -1.0	0
Chi-square	χ^2	Nominal, ordinal	Infinity	0

12.4 More than Two Variables (1 of 4)

Objective: Analyze data using multivariate techniques.

- Researchers can examine control variables with multivariate tables and statistics to decide whether a bivariate relationship is false.
- A **control variable** is a variable that represents an alternative explanation for a two-variable relationship.
 - The control variable takes into account the effects of variables other than the primary independent and dependent variables of the hypothesis.

12.4 More than Two Variables (2 of 4)

- The Elaboration Model of Percentaged Tables
 - Researchers can statistically introduce control variables in trivariate or three-variable tables.
 - A trivariate table consists of a separate bivariate table of the independent and dependent variables created for each category of the control variable. These are called **partials**.

12.4 More than Two Variables (3 of 4)

- The Elaboration Model of Percentaged Tables
 - The **elaboration paradigm** is a system for reading percentaged trivariate tables.
 - Five Patterns
 - Replication pattern
 - Specification pattern
 - Interpretation pattern
 - Explanation pattern
 - Suppressor variable pattern

12.4 More than Two Variables (4 of 4)

- Multiple Regression Analysis

- Requires interval- or ratio-level data
- Results tell the researcher two things:

The overall predictive power of the set of independent and control variables on the dependent variable

Multiple regression results give the direction and size of the effect of each variable on a dependent variable.

12.5 Inferential Statistics (1 of 3)

Objective: Analyze data using inferential statistical methods.

- The Purpose of Inferential Statistics
 - They rely on principles from probability sampling by which researchers use a random process to select cases from the entire population.
 - A precise way to talk about how confident researchers are when inferring results from the sample to the population

12.5 Inferential Statistics (2 of 3)

- Statistical Significance
 - **Statistical significance** indicates the probability of finding a relationship in the sample when there is none in the population.
 - The **level of statistical significance** is an easy way of talking about the likelihood that results are due to chance factors.

Usually 0.05, 0.01, or 0.001

12.5 Inferential Statistics (3 of 3)

- Type I and Type II Errors
- A Type I error occurs when the researcher says a relationship exists when none actually exists.
- A Type II error occurs when the researcher says a relationship does not exist when it actually does.

Conclusion

- Beginning researchers sometimes believe that they have done something wrong if their results do not support a hypothesis.
- There is nothing wrong with rejecting a hypothesis.
- Good research means guarding against possible errors or obstacles to true inferences from data to the social world.
- The only way to assign meaning to facts, charts, tables, or statistics is to use theory, insight, and understanding.
- The best way to protect yourself from being misled by statistics is to understand them.

Copyright



This work is protected by United States copyright laws and is provided solely for the use of instructors in teaching their courses and assessing student learning. Dissemination or sale of any part of this work (including on the World Wide Web) will destroy the integrity of the work and is not permitted. The work and materials from it should never be made available to students except by instructors using the accompanying text in their classes. All recipients of this work are expected to abide by these restrictions and to honor the intended pedagogical purposes and the needs of other instructors who rely on these materials.