2.1 Frequency Distributions of Nominal Data

Responses of Young Boys to Removal of Toy

Response of Child	*f*
Cry	25
Express Anger	15
Withdraw	5
Ply with another toy	5
	N=50

Characteristics of a frequency distribution of nominal data:

- Title
- Consists of two columns:
  - Left column: characteristics (e.g., Response of Child)
  - Right column: frequency (*f*)

2.1 Comparing Distributions

Comparisons clarify results, add information, and allow for comparisons

Response to Removal of Toy by Gender of Child

	Gender of Child	
Response of Child	*Male*	*Female*
Cry	25	28
Express Anger	15	3
Withdraw	5	4
Play with another toy	5	15
Total	50	50

2.2 Proportions and Percentages

Allows for a comparison of groups of different sizes

Proportion – number of cases compared to the total size of distribution

$$P = \frac{f}{N}$$

Percentage – the frequency of occurrence of a category per 100 cases

$$\% = (100)\frac{f}{N}$$

2.2 Ratio and Rates

Ratio – compares the frequency of one category to another

$$\text{Ratio} = \frac{f_1}{f_2}$$

Rate – compares between actual and potential cases

$$\text{Rate} = (1{,}000)\left(\frac{f \text{ actual cases}}{f \text{ potential cases}}\right)$$

2.3

TABLE 2.4 ***The Distribution of Marital Status Shown Three Ways***

Marital Status	*f*	Marital Status	*f*	Marital Status	*f*
Married	30	Single	20	Previously married	10
Single	20	Previously married	10	Married	30
Previously married	10	Married	30	Single	20
Total	60	Total	60	Total	60

Table 2.4

2.3

TABLE 2.5 ***A Frequency Distribution of Attitudes toward a Proposed Tuition Hike on a College Campus: Incorrect and Correct Presentations***

Attitude toward a Tuition Hike	*f*
Slightly favorable	2
Somewhat unfavorable	21
Strongly favorable	0
Slightly unfavorable	4
Strongly unfavorable	10
Somewhat favorable	1
Total	38

INCORRECT

Attitude toward a Tuition Hike	*f*
Strongly favorable	0
Somewhat favorable	1
Slightly favorable	2
Slightly unfavorable	4
Somewhat unfavorable	21
Strongly unfavorable	10
Total	38

CORRECT

Table 2.5

2.3 Grouped Frequency Distribution of Interval Data

Used to clarify the presentation of interval-level scores spread over a wide range

Class Intervals

- Smaller categories or groups containing more than one score
- Class interval size determined by the number of score values it contains

2.3

TABLE 2.7 ***Grouped Frequency Distribution of Final-Examination Grades for 71 Students***

Class Interval	*f*	%
95–99	3	4.23
90–94	2	2.82
85–89	4	5.63
80–84	7	9.86
75–79	12	16.90
70–74	17	23.94
65–69	12	16.90
60–64	5	7.04
55–59	5	7.04
50–54	4	5.63
Total	71	100[a]

[a] The percentages as they appear add to only 99.99%. We write the sum as 100% instead, because we know that .01% was lost in rounding.

Table 2.7

2.3 Class Limits and the Midpoint

Class Limits

- The point halfway between adjacent intervals
- Upper and lower limits
  - Distance from upper and lower limit determines the size of class interval

$$i = U - L$$

$i =$ size of a class interval

$U =$ upper limit of a class interval

$L =$ lower limit of a class interval

The Midpoint

- The middlemost score value in a class interval
  - The sum of the lowest and highest value in a class interval divided by two

$$m = \frac{\text{lowest score value} + \text{highest score value}}{2}$$

2.3 Cumulative Distributions

Cumulative Frequencies

- Total number of cases having a given score or a score that is lower
  - Shown as *cf*
  - Obtained by the sum of frequencies in that category plus all lower categories' frequencies

Cumulative Percentage

- Percentage of cases having a given score or a score that is lower

$$c\% = (100)\frac{cf}{N}$$

2.3

TABLE 2.7 ***Grouped Frequency Distribution of Final-Examination Grades for 71 Students***

Class Interval	*f*	%
95–99	3	4.23
90–94	2	2.82
85–89	4	5.63
80–84	7	9.86
75–79	12	16.90
70–74	17	23.94
65–69	12	16.90
60–64	5	7.04
55–59	5	7.04
50–54	4	5.63
Total	71	100[a]

[a] The percentages as they appear add to only 99.99%. We write the sum as 100% instead, because we know that .01% was lost in rounding.

Table 2.7

2.3 Percentiles

The percentage of cases falling at or below a given score

- Quartiles – points that divide a distribution into quarters
- Median – the point that divides a distribution in two, half above it and half below it

Unequal Class Sizes

TABLE 12 ***Frequency Distribution of Family Income Data***

Income Category	*f* (families in 1,000s)	%
$100,000 and over	8,391	11.8
$75,000–$99,999	7,826	11.0
$50,000–$74,999	15,112	21.3
$35,000–$49,999	12,357	17.4
$25,000–$34,999	9,079	12.8
$15,000–$24,999	9,250	13.0
$10,000–$14,999	4,054	5.7
$5,000–$9,999	2,887	4.1
Less than $5,000	1,929	2.7
	$N = 70,885$	100.0

TABLE 13 ***Frequency Distribution of Family Income Data (with Midpoints)***

Income Category	*m*	*f*	%
$100,000 and over	$125,000	8,391	11.8
$75,000–$99,999	$87,500	7,826	11.0
$50,000–$74,999	$62,500	15,112	21.3
$35,000–$49,999	$42,500	12,357	17.4
$25,000–$34,999	$30,000	9,079	12.8
$15,000–$24,999	$20,000	9,250	13.0
$10,000–$14,999	$12,500	4,054	5.7
$5,000–$9,999	$7,500	2,887	4.1
Less than $5,000	$2,500	1,929	2.7
		$N = 70,885$	100.0

To calculate the midpoint of the highest income category, invent an upper limit in accordance with the previous limits

2.4 Cross Tabulations

A cross-tabulation is a table that presents the distribution (frequencies and percents) of one variable (usually the dependent variable) across the categories of one or more additional variables (usually the independent variable).

2.4

TABLE 2.17 ***Cross-Tabulation of Victim–Offender Relationship by Victim Sex in U.S. Homicides—2010***

	Victim Sex		
Relationship	*Male*	*Female*	**Total**
Intimate	496	1,473	1,969
Family	1,309	606	1,915
Acquaintance	6,273	907	7,180
Stranger	3,334	349	3,683
Total	11,412	3,335	14,747

Table 2.17

2.4 Percents within Cross-Tabulations

Total Percents: $total\% = (100)\frac{f}{N_{total}}$

Row Percents: $row\% = (100)\frac{f}{N_{row}}$

Column Percents: $column\% = (100)\frac{f}{N_{column}}$

The choice comes down to which is more relevant to the purpose of the analysis

- If the independent variable is on the rows, use row percents
- If the independent variable is on the columns, use column percents
- If the independent variable is unclear, use whichever percent is most meaningful

2.4

TABLE 2.18 ***Cross-Tabulation of Victim–Offender Relationship by Victim Sex (with Total Percents) in U.S. Homicides—2010***

	Victim Sex		
Relationship	*Male*	*Female*	**Total**
Intimate	496 3.4%	1,473 10.0%	1,969 13.4%
Family	1,309 8.9%	606 4.1%	1,915 13.0%
Acquaintance	6,273 42.5%	907 6.2%	7,180 48.7%
Stranger	3,334 22.6%	349 2.4%	3,683 25.0%
Total	11,412 77.4%	3,335 22.6%	14,747 100.0%

Row marginal (row totals)

Total sample size

Column marginal (column totals)

Table 2.18

2.5 Graphic Presentations

Columns of numbers have been known to evoke fear, anxiety and boredom.

People are usually more interested in charts.

Pie Charts

Common graphic presentation types:

- **Pie charts**
- **Histograms**
- **Frequency polygons**
- **Line charts**

2.5

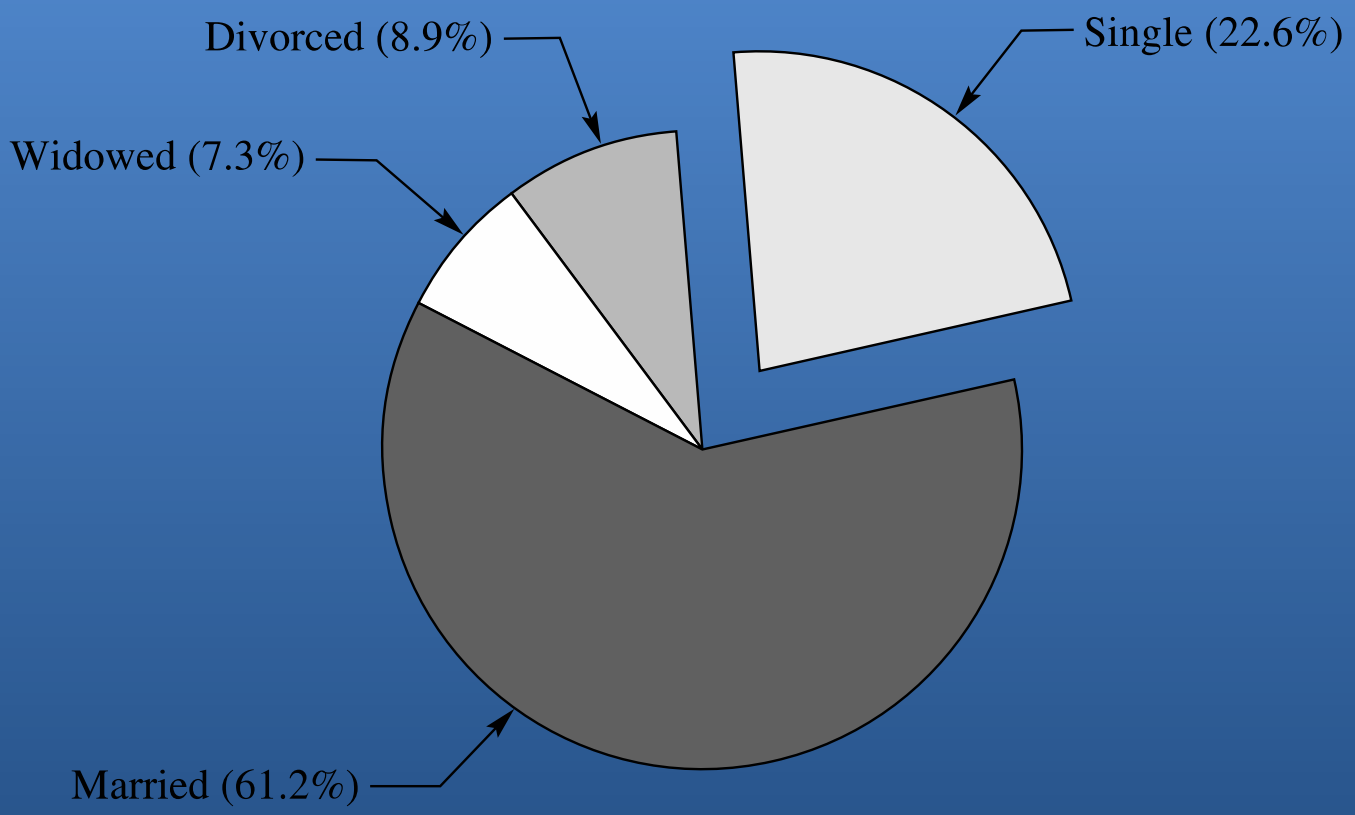


Figure 2.4

- **Useful for showing the differences in frequencies or percentages among categories of a nominal-level variable. (marital status)**

- **Not advisable to use a pie chart for data that are classified in ordered categories (level of education)**

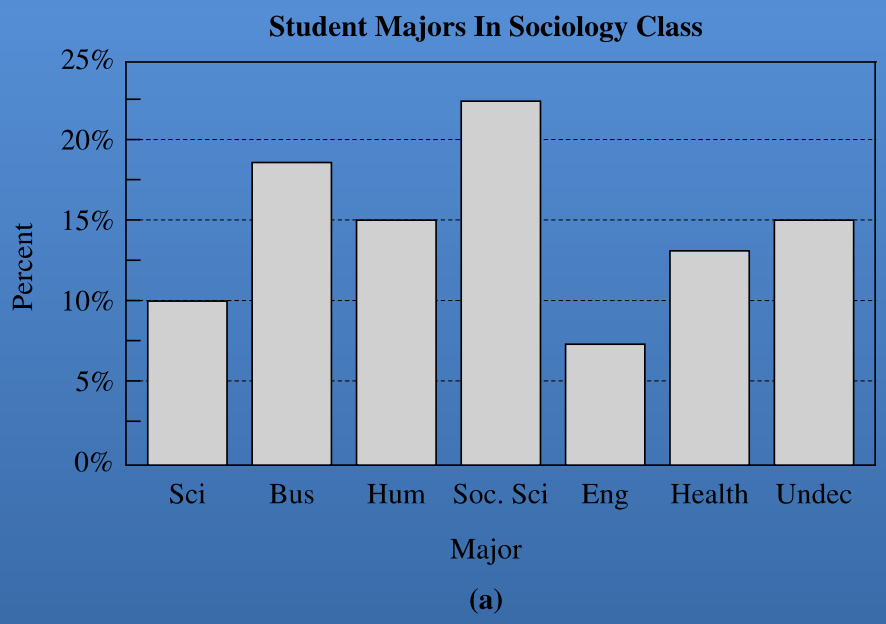
Bar graphs or histograms

- **The bar graphs or histograms can accommodate any number of categories at any level of measurement**

Bar graphs→nominal , discrete

Histograms→interval, continious

2.5


Student Majors In Sociology Class
25%
20%
15%
10%
5%
0%
Percent
Sci
Bus
Hum
Soc. Sci
Eng
Health
Undec
Major
(a)

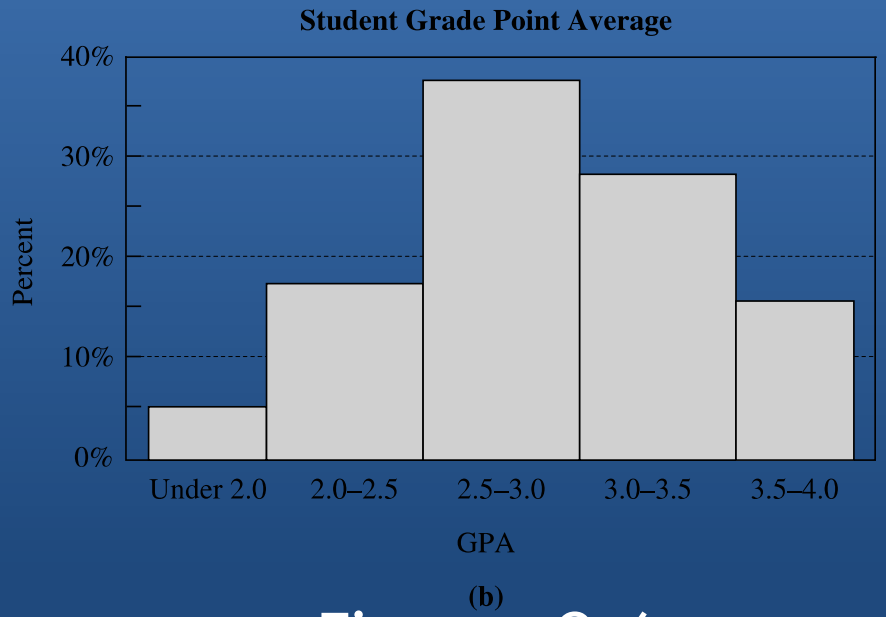

Student Grade Point Average
40%
30%
20%
10%
0%
Percent
Under 2.0
2.0–2.5
2.5–3.0
3.0–3.5
3.5–4.0
GPA
(b)

Figure 2.6

Frequency Polygons

- **Particularly useful for depicting ordinal and interval data.**
- **Frequencies are indicated by a series of points placed over the score values**
- **Adjacent points are connected with a straight line.**
- **The height of each point indicates frequency of occurrence.**

2.5

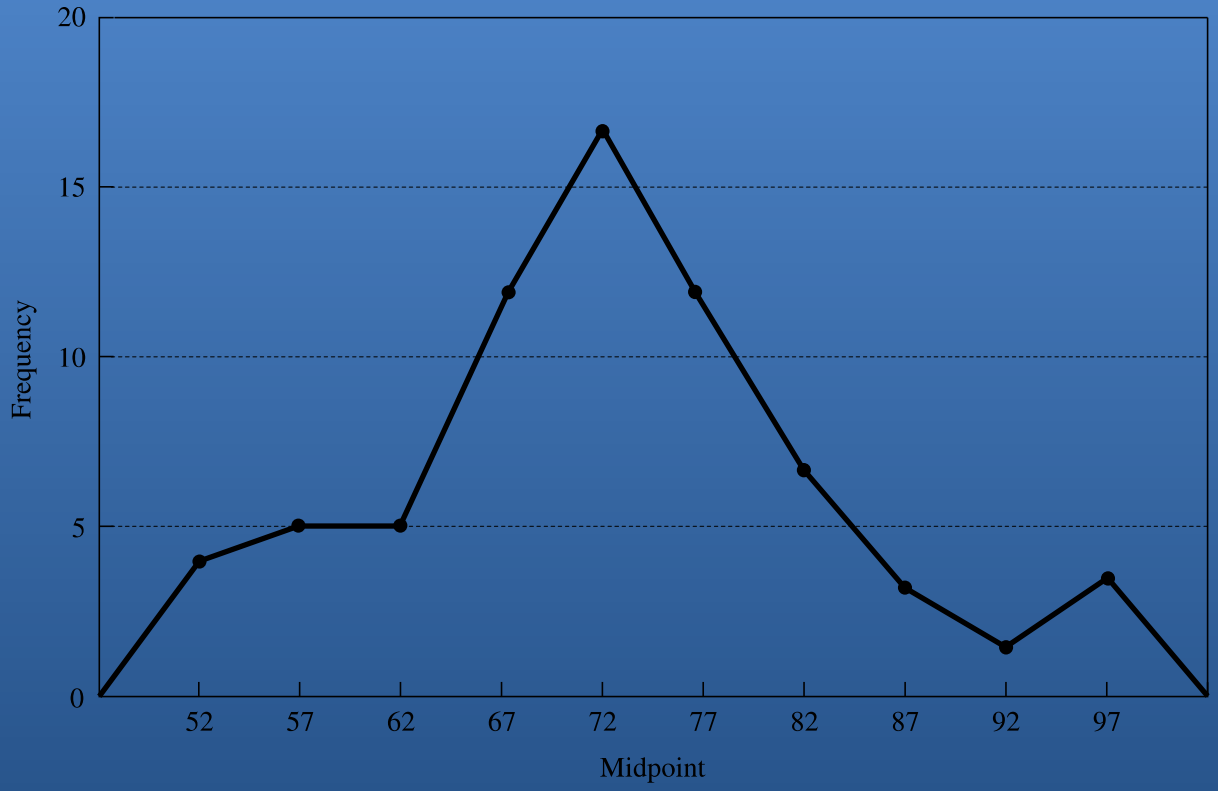


Figure 2.9

Line Charts

- **Trend data are customarily depicted with line charts.**

- **Increase / Decrease over the years**

2.5

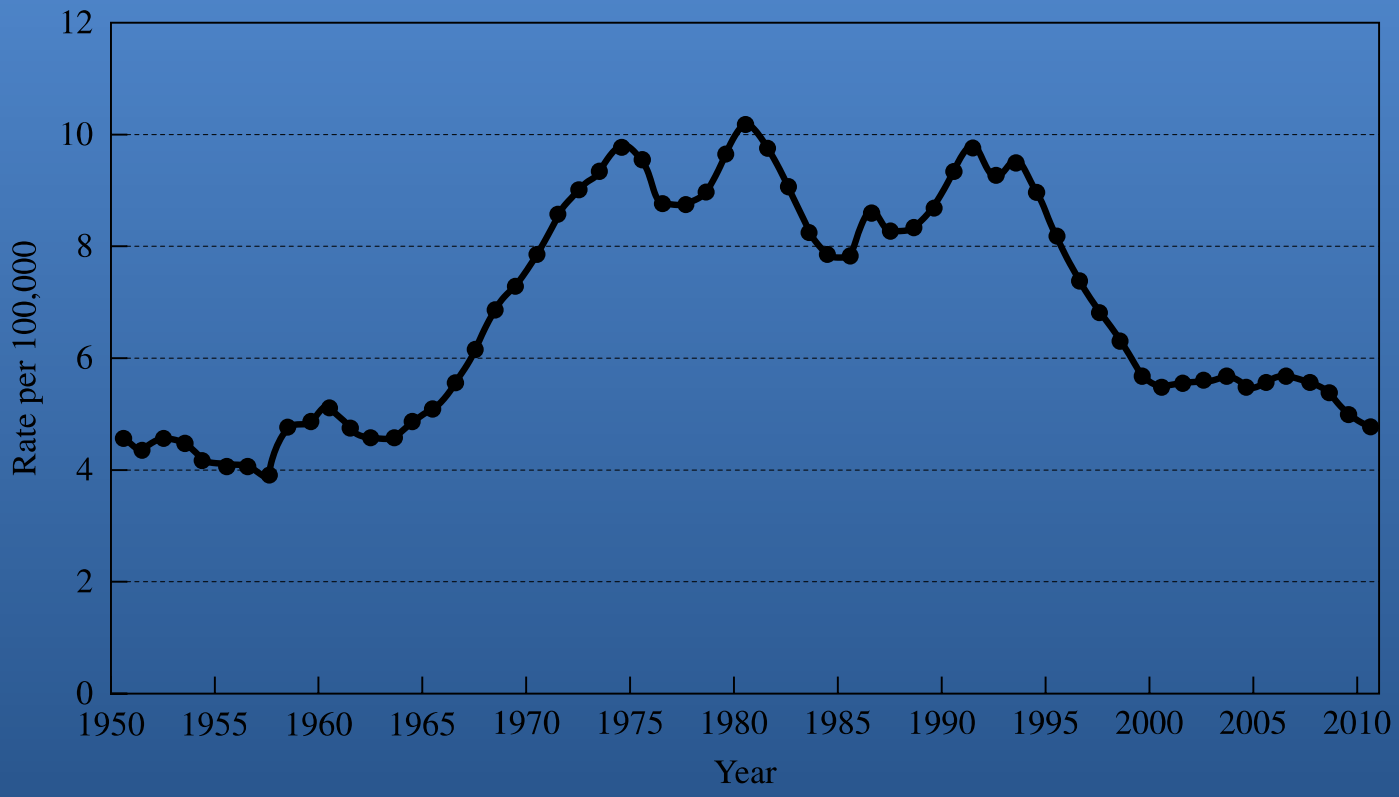


Figure 2.14

2.5

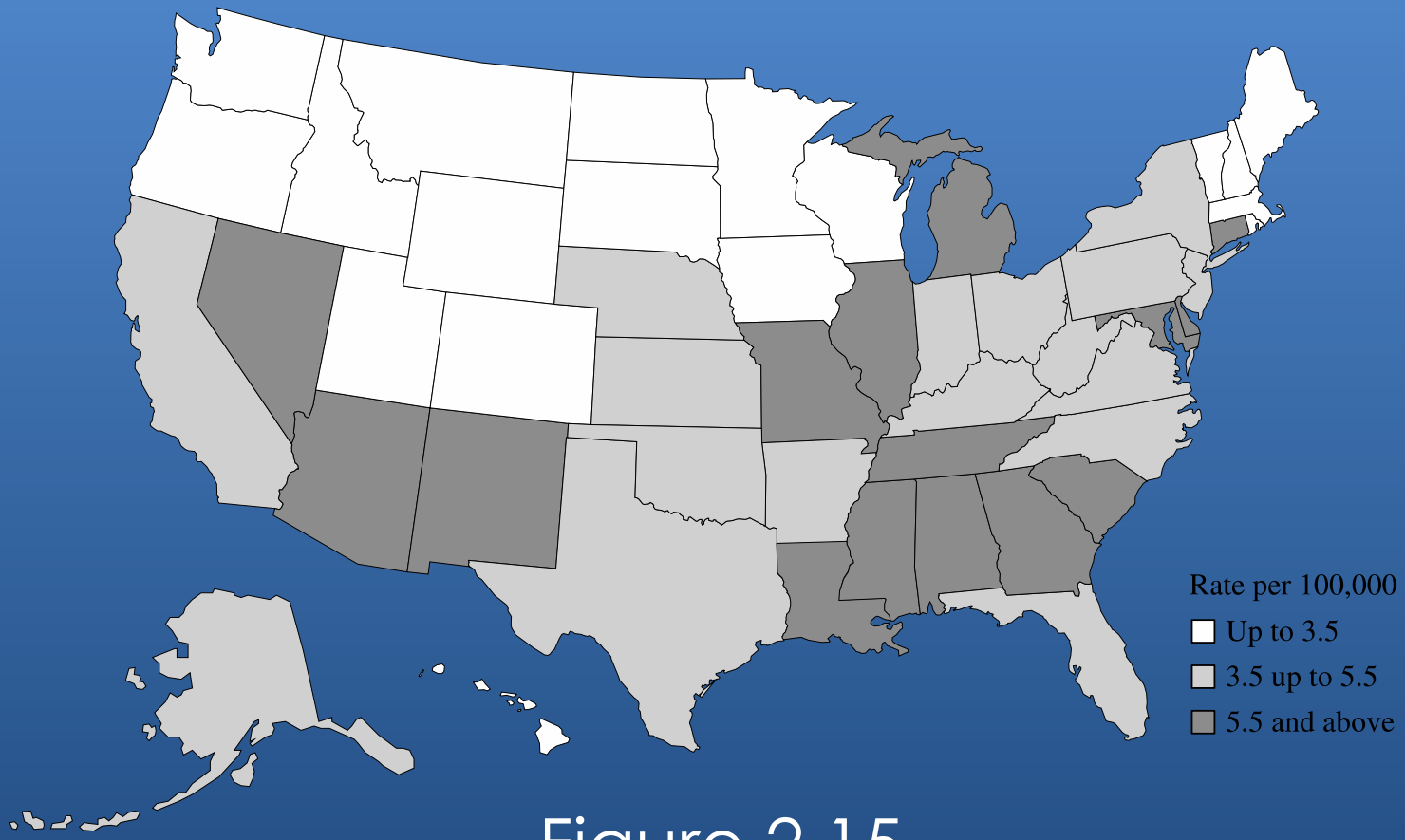


Figure 2.15

3.1 Introduction

Measures of Central Tendency

Mode

Median

Mean

3.1 The Mode

The most frequently occurring value in a distribution

- Example: 20, 21, 30, 20, 22, 20, 21, 20
  - Mode = 20

- Sometimes there is more than one mode
  - Example: 96, 91, 96, 90, 93, 90, 96, 90
  - Mode = 90 ***and*** 96

    This is a bimodal distribution

- The mode is the only measure of central tendency appropriate for nominal-level variables

3.1 The Mode

The mode is not the frequency of the most frequent score but the value

1,2,3,1,1,6,5,4,1,4,4,3→1 has the highest frequency (f=4)

Mode is 1 not 4

3.1

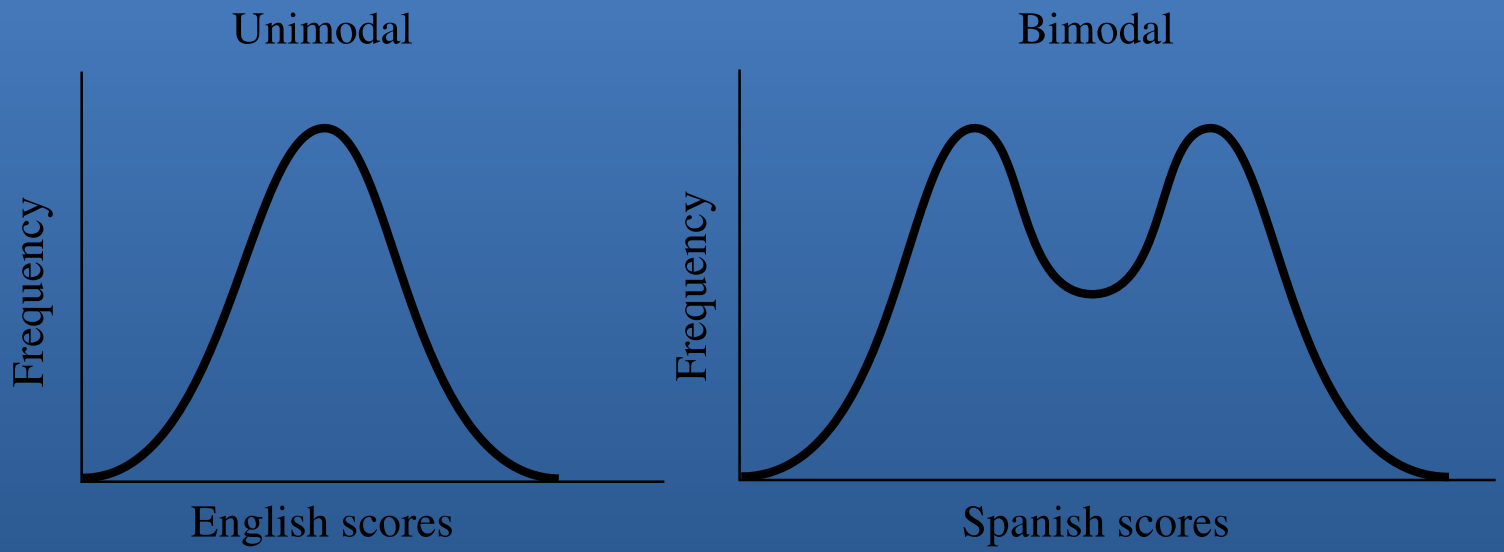

Unimodal
Frequency
English scores
Bimodal
Frequency
Spanish scores

Figure 3.1

3.1 The Median

The middlemost case in a distribution

- Appropriate for ordinal or interval level data

$$\text{Position of Median} = \frac{N+1}{2}$$

- How to find the median:
  - Cases must be ordered
  - If there are an odd number of cases, there will be a single middlemost case
  - If there are an even number of cases, there will be two middlemost cases
    - The halfway point between these two cases should be used as the median

3.1 The Median: Example 1

What is the median of the following distribution:

1, 5, 2, 9, 13, 11, 4

Step 1: Sort distribution from lowest to highest

1, 2, 4, 5, 9, 11, 13

Step 2: Locate the position of the median

$$\text{Position of Median} = \frac{N+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4$$

Step 3: Locate the median

1, 2, 4, 5, 9, 11, 13

3.1 The Median: Example 2

What is the median of the following distribution:

4, 3, 1, 1, 6, 2, 2, 4

Step 1: Sort distribution from lowest to highest

1, 1, 2, 2, 3, 4, 4, 6

Step 2: Locate the position of the median

$$\text{Position of Median} = \frac{N+1}{2} = \frac{8+1}{2} = \frac{9}{2} = 4.5$$

Step 3: Locate the median

1, 1, 2, 2, 3, 4, 4, 6

Step 4: Take the halfway point between the two cases

Median = 2.5

3.1 The Mean

The "center of gravity" of a distribution

- Appropriate for interval/level data

$$\bar{X} = \frac{\sum X}{N}$$

$\bar{X}$ = mean

$\sum$ = sum

X = raw scores in a set of scores

N = total number of scores in a set

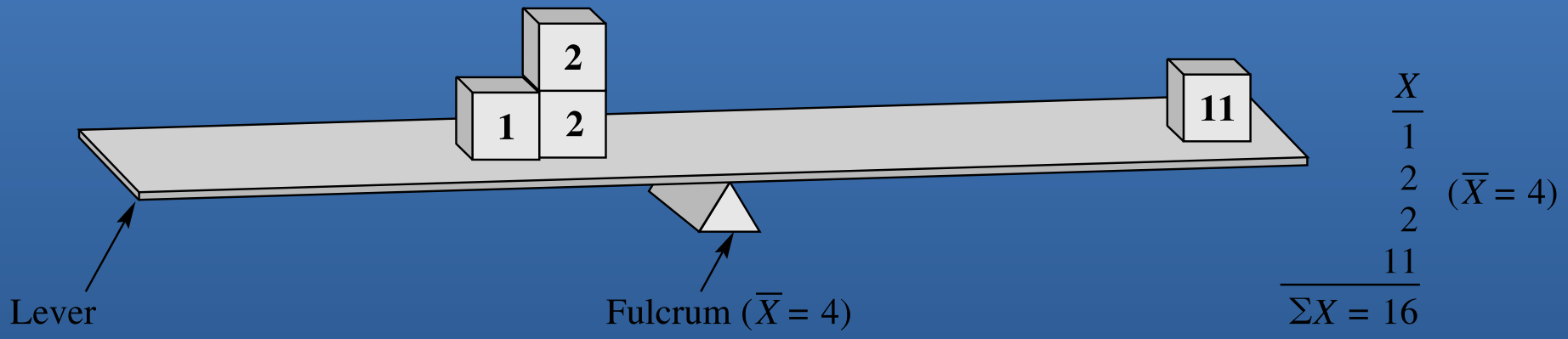

2
1
2
11
Lever
Fulcrum ($\overline{X} = 4$)
X
1
2
2
11
$\Sigma X = 16$
($\overline{X} = 4$)

Figure 3.2

3.1 The Mean: Example

What is the mean of the following distribution:

4, 8, 11, 2

$$\bar{X} = \frac{\sum X}{N}$$

$$\bar{X} = \frac{(4+8+11+2)}{4}$$

$$\bar{X} = \frac{25}{4}$$

$$\bar{X} = 6.25 \text{ years}$$

3.2 Deviations

The distance and direction of any raw score from the mean

$$\text{Deviation} = X - \bar{X}$$

- The sum of the deviations that fall above the mean is equal in absolute value to the sum of the deviations that fall below the mean.

3.4 Obtaining the Mode, Median, and Mean from a Simple Frequency Distribution

	X	f	cf	fX
	31	1	25	31
	30	1	24	30
	29	1	23	29
	28	0	22	0
	27	2	22	54
	26	3	20	78
	25	1	17	25
	24	1	16	24
$\overline{X}$ →	23	2	15	46
Mdn →	22	2	13	44
	21	2	11	42
	20	3	9	60
Mo →	19	4	6	76
	18	2	2	36

$$\text{Position of the Mdn} = \frac{N+1}{2} = \frac{25+1}{2} = \frac{26}{2} = 13$$

$$\overline{X} = \frac{\sum fX}{N} = \frac{575}{25} = 23$$

3.5 Comparing the Mode, Median, and Mean

Three factors in choosing a measure of central tendency

- Level of Measurement
- Shape of Distribution
- Research Objective

3.5 Level of Measurement

Level of Measurement	Mode	Median	Mean
Nominal	Yes	No	No
Ordinal	Yes	Yes	No
Interval	Yes	Yes	Yes

Suppose that a volunteer canvasses houses in her neighborhood collecting for a local charity. She receives the following donations (in dollars):

5 10 25 15 18 2 5

Step 1 Arrange the scores from highest to lowest.

25
18
15
10
5
5
2

Step 2 Find the most frequent score.

$$\text{Mo} = \$5$$

Step 3 Find the middlemost score. Because there are seven scores (an odd number), the fourth from either end is the median.

Step 4 Determine the sum of the scores.

$$\begin{array}{r} 25 \\ 18 \\ 15 \\ 10 \\ 5 \\ 5 \\ 2 \\ \hline \Sigma X = \$80 \end{array}$$

Step 5 Determine the mean by dividing the sum by the number of scores.

$$\overline{X} = \frac{\Sigma X}{N} = \frac{\$80}{7} = \$11.43$$

Thus, the mode, median, and mean provide very different pictures of the average level of charitable giving in the neighborhood. The mode suggests that the donations were typically small, whereas the median and the mean suggest greater generosity overall.

3.5 Shape of the Distribution

Symmetrical Distributions

- The mode, median, and mean have identical values

Skewed Distributions

- The mode is the peak of the curve
- The mean is closer to the tail
- The median falls between the two

Bimodal Distributions

- Both modes should be used to describe the data

3.5

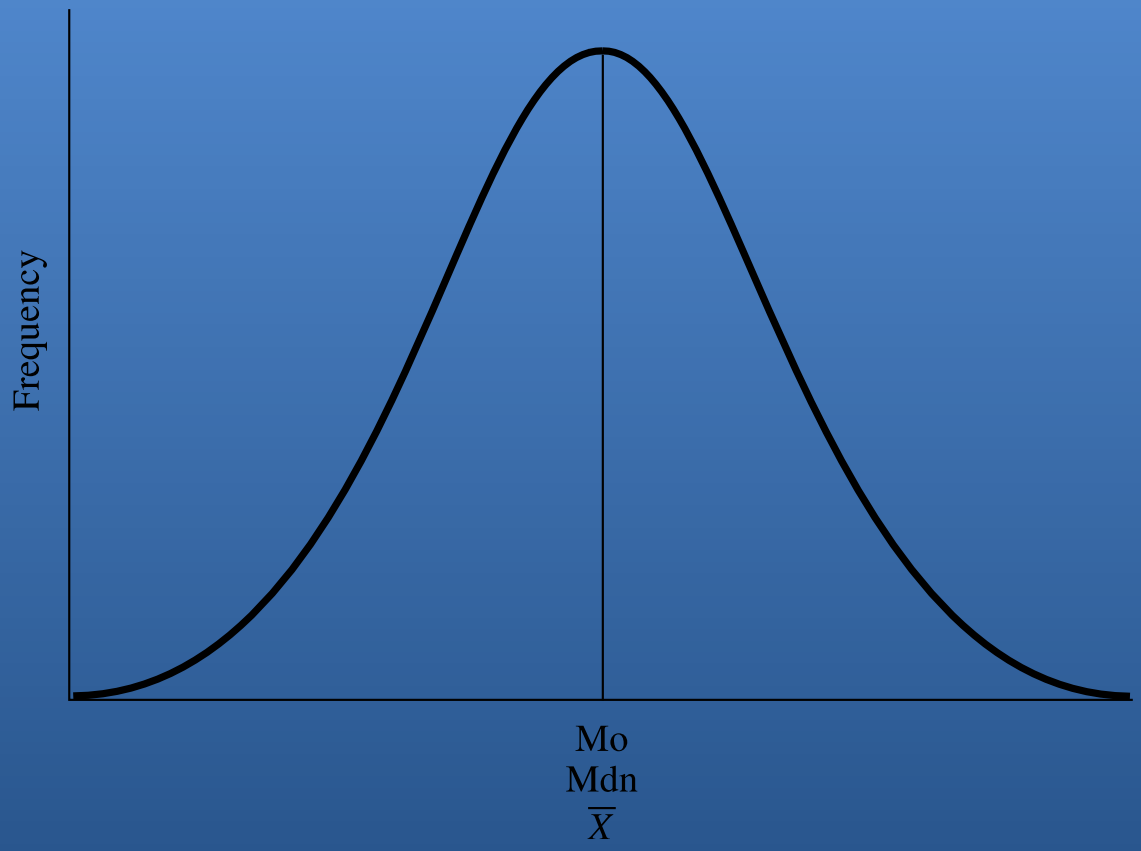


Figure 3.3

3.5

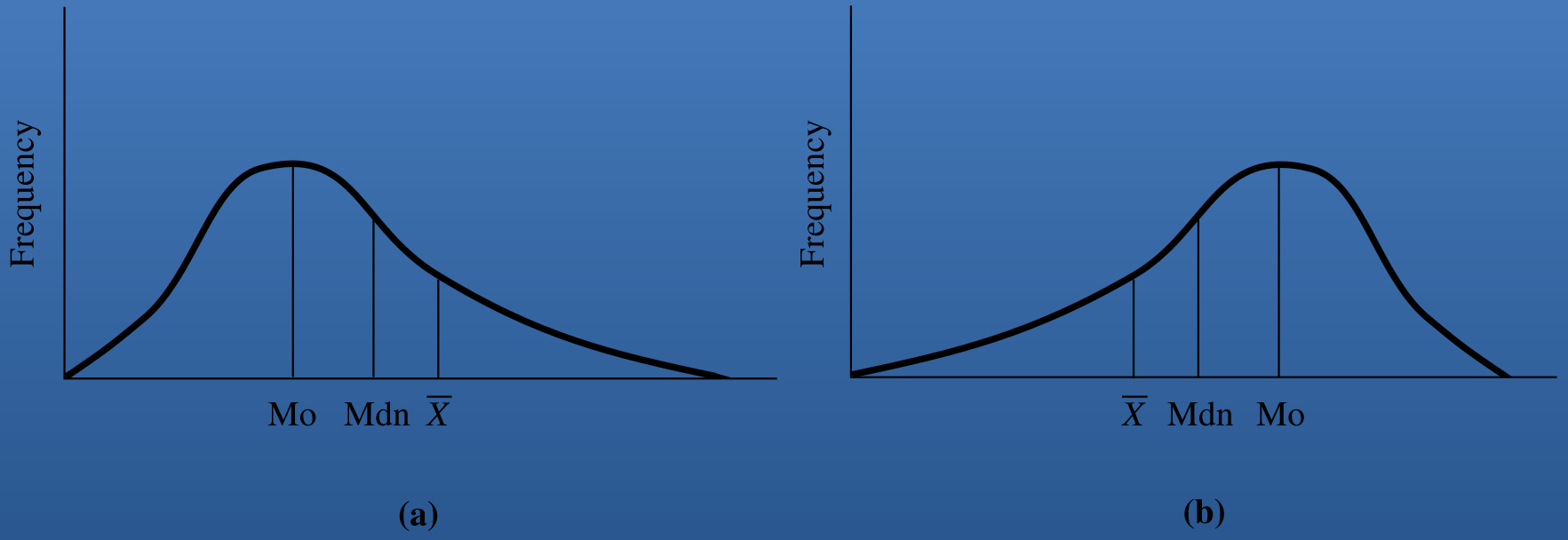


Figure 3.4

3.5 Research Objective

Fast and Simple Research → Mode

Skewed Distribution → Median

Advanced Statistics Analysis → Mean

There are three measures of central tendency: the mode, the median, and the mean

How individual scores compare to the mean of the distribution can be examined by calculating deviations

The mean of means, or the weighted mean, can be calculated for multiple groups

The mode, the median, and the mean can also be calculated when data are presented in a simple frequency distribution

Choosing which measure of central tendency to report is influenced by the level of measurement of the data, the shape of the distribution, as well as the research objective

Are male students more inclined to have tried marijuana?

Marijuana Use	Gender of Respondent: *Male*	*Female*
Number who have tried it	35	15
Number who have not tried it	65	85
Total	100	100

Marijuana Use	Gender of Respondent: *Male*	*Female*
Number who have tried it	30	20
Number who have not tried it	70	80
Total	100	100

Marijuana Use	Gender of Respondent: *Male*	*Female*
Number who have tried it	26	24
Number who have not tried it	74	76
Total	100	100

Which difference is significant enough to answer this question?

P Value

P value (Probability) → The probability, under the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed.

The probability of observing a difference even if our hypothesis is nor true.

Between 0-1 but never 0 or 1(0% or 100%).

0,05→5 %

0.25→25 %

	Cured	Not Cured	Total
Treatment A	56	15	71
Treatment B	67	30	97
Total	123	45	168
	p=0.565		

	Cured	Not Cured	Total
Treatment A	56	15	71
Treatment B	67	21	97
Total	123	45	168
	p=0.1565		

	Cured	Not Cured	Total
Treatment A	112	30	142
Treatment B	134	60	194
Total	246	90	336
	p=0.0451		

Obtaining a p value lower than a preset threshold indicates the 'correctness' of the hypothesis.

This threshold is usually considered to be 0.05

0,10 – 0,001

There is a significant difference between two treatments According to the results in the third table.

Are older men more likely to commit suicide?

- p=0.000043
- p=0.034
- p=0.087