



Chapter 10

Quantitative Data Analysis

Commonly Used Terms

- **Univariate**
 - Examine 1 variable at a time
- **Bivariate**
 - Examine relationship between 2 variables
- **Multivariate**
 - Examine relationship between 3 or more variables
- **Distribution**
 - All of the values of a given variable for all cases under study
- **Discrete or categorical**
 - Values/attributes are separate and distinct from each other
 - Nominal and ordinal levels of measurement
- **Continuous**
 - Values/attributes are close together, on a continuum
 - Interval and ratio levels of measurement

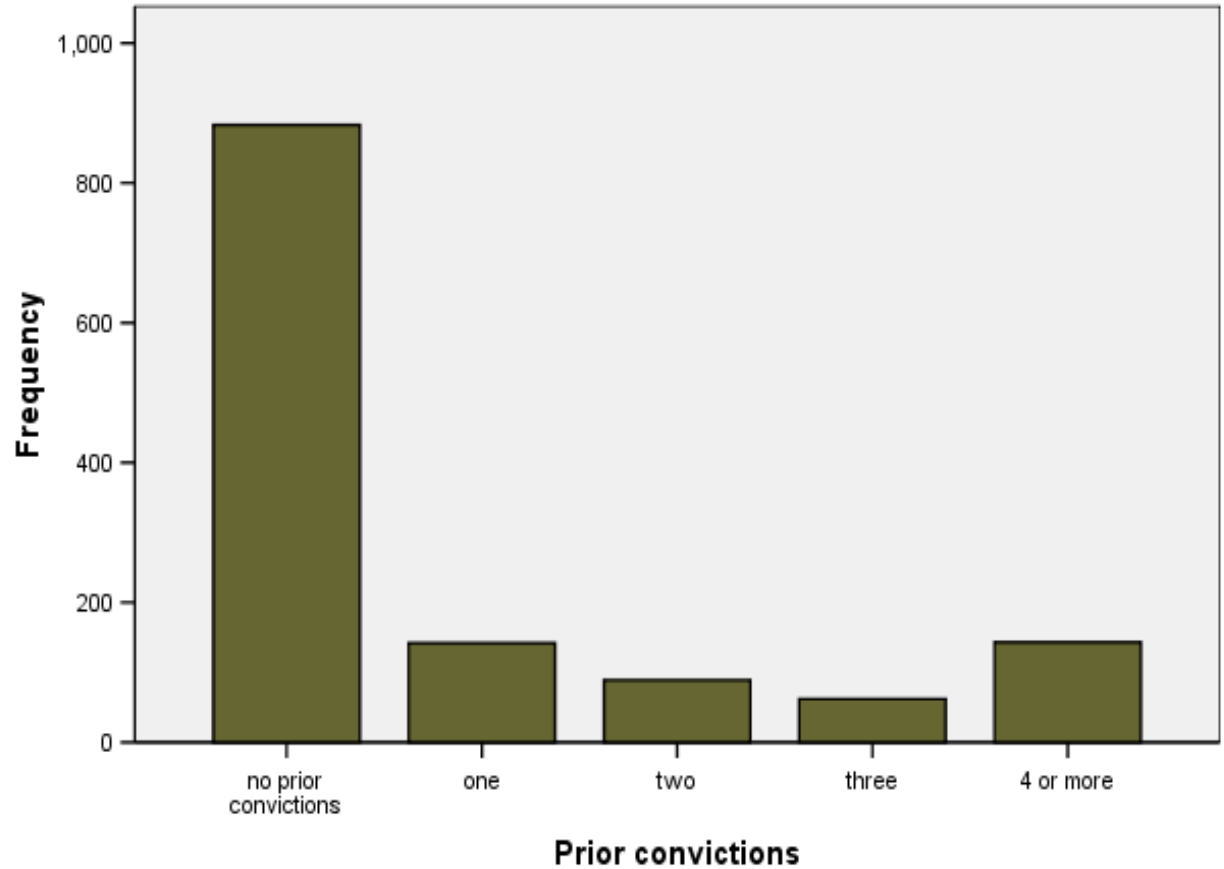
Displaying Univariate Distributions

- **Graphically**
 - Bar chart
 - Histogram
 - Frequency polygon Numerically
- **Frequency (or percentage) distributions**
 - Ungrouped – listing of all possible values of variable individually
 - Grouped – summarized listing of values collapsed into categories
- **Features of distribution shape**
 - Central tendency – most common value or value around which cases tend to center
 - Variability – how cases are spread out through the distribution or clustered
 - Skewness – extent to which cases are clustered more at one end of the distribution

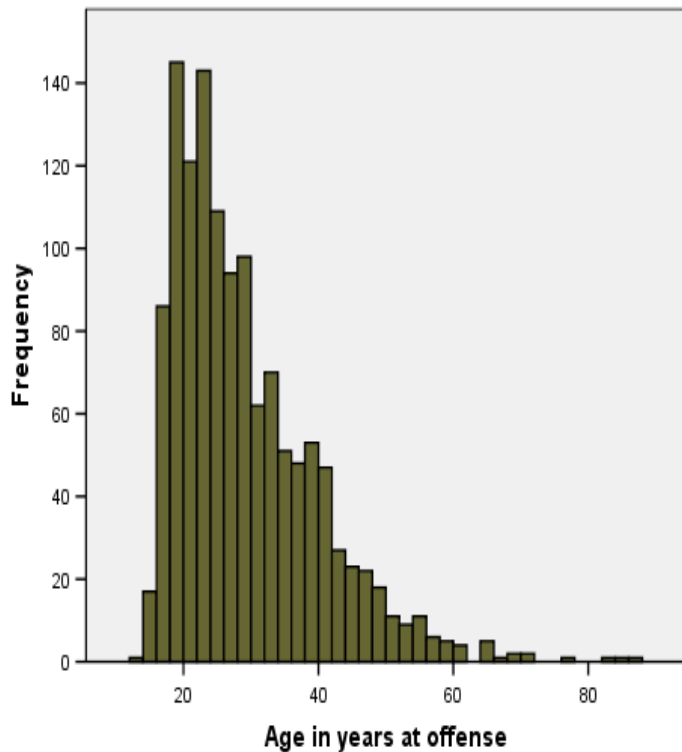
Bar Chart

*Appropriate for
all levels of
measurement*

Ordinal
data

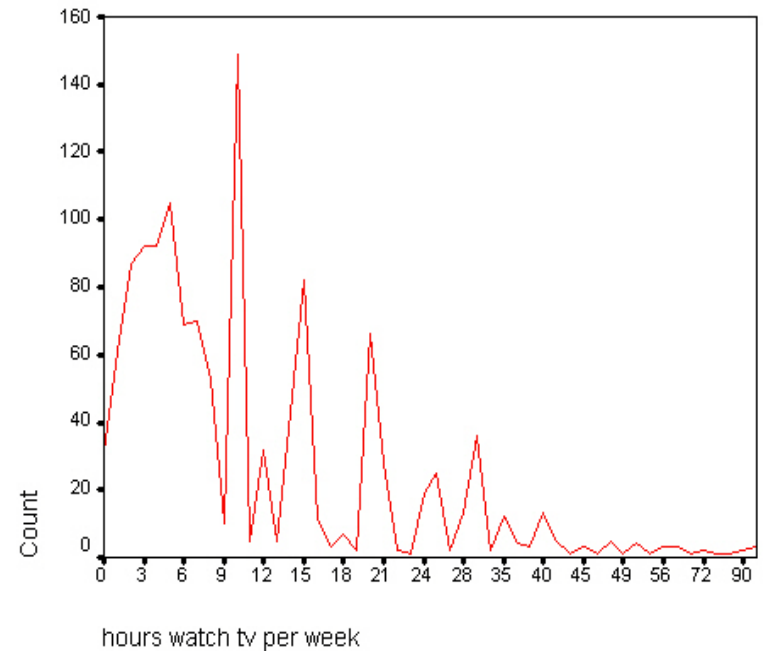


Histogram & Frequency Polygon



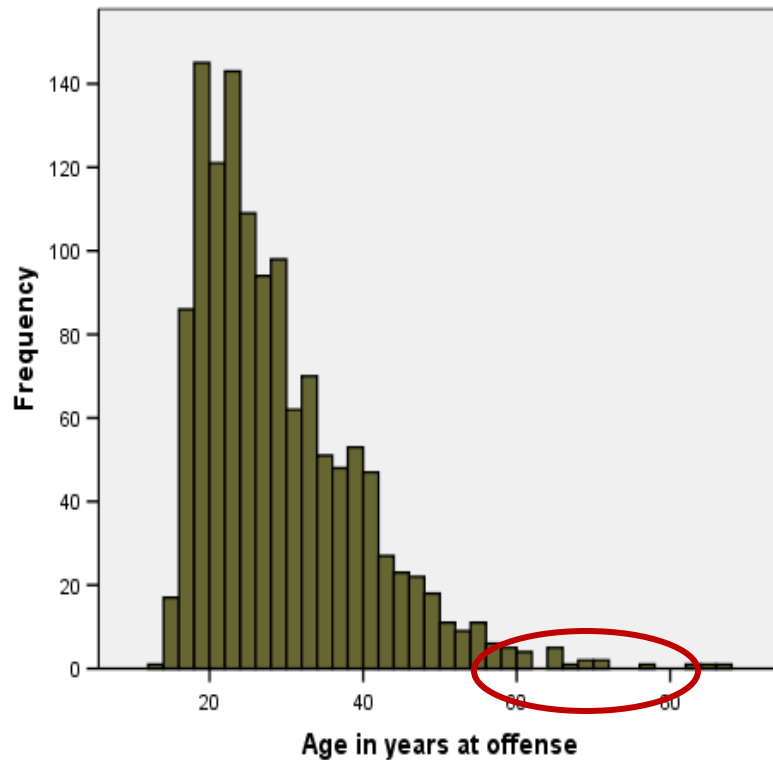
Histogram

Mean = 29.23
Std. Dev. = 10.634
N = 1,295



Frequency Polygon
(aka Line Graph)

Skewness



Mean = 29.23
Std. Dev. = 10.634
N = 1,295

- How values cluster
- Positive or negative
 - Positive
 - Values cluster to left
 - Right tail is longer
 - Negative
 - Values cluster to right
 - Left tail is longer

Frequency Distributions

| | Education Level | |
|--------------------------|-----------------|--------------|
| | Frequency | Percentage |
| Some high school | 169 | 34.1% |
| High school graduate | 163 | 32.9 |
| Some college | 103 | 20.8 |
| College graduate or more | 60 | 12.1 |
| Total | 495 | 99.9% |

Displaying Ungrouped Data

Many frequency distributions (and graphs) require grouping of some values after the data are collected. There are two reasons for grouping:

- There are more than 15–20 values to begin with, a number too large to be displayed in an easily readable table
- The distribution of the variable will be clearer or more meaningful if some of the values are combined

Ungrouped vs. Grouped Data

Age Percentage

| | |
|----|------|
| 18 | 1.5% |
| 19 | 1.6 |
| 20 | 1.5 |
| 21 | 1.6 |
| 22 | 1.5 |
| 23 | 2.0 |
| 24 | 2.0 |
| 25 | 1.9 |
| 26 | 2.2 |
| 27 | 1.6 |
| 28 | 2.6 |
| 29 | 1.5 |
| 30 | 1.9 |
| 31 | 1.7 |
| 32 | 1.9 |
| 33 | 2.0 |
| 34 | 2.3 |
| 35 | 2.2 |
| 36 | 2.1 |
| 37 | 1.5 |
| 38 | 1.9 |
| 39 | 1.7 |
| 40 | 2.5 |
| 41 | 2.0 |
| 42 | 1.9 |
| 43 | 2.3 |
| 44 | 2.4 |
| 45 | 2.2 |
| 46 | 1.9 |

...

Source: General Social Survey, 2004. Weighted.

Age Percentage

Group

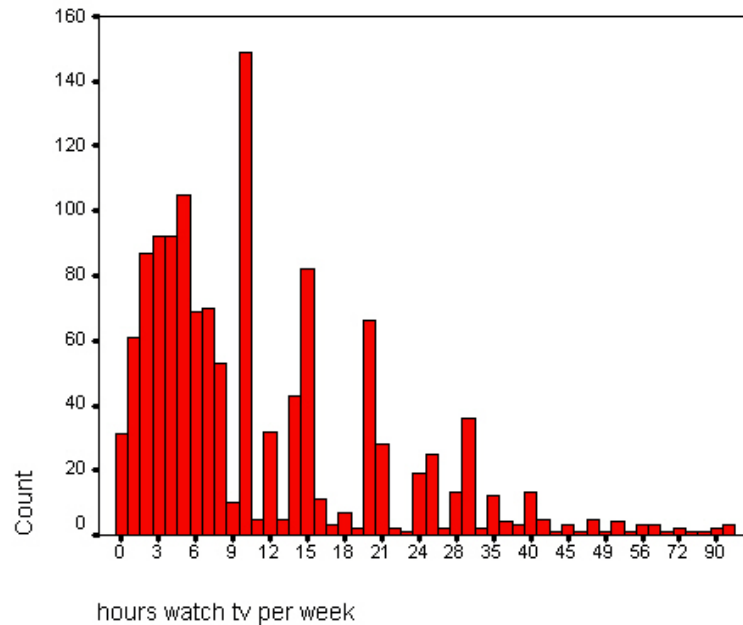
| | |
|-------|------|
| 18–19 | 3.1% |
| 20–29 | 18.4 |
| 30–39 | 19.3 |
| 40–49 | 21.9 |
| 50–59 | 18.0 |
| 60–69 | 11.3 |
| 70–79 | 5.2 |
| 80–89 | 42.8 |

100.0% (n=2801)

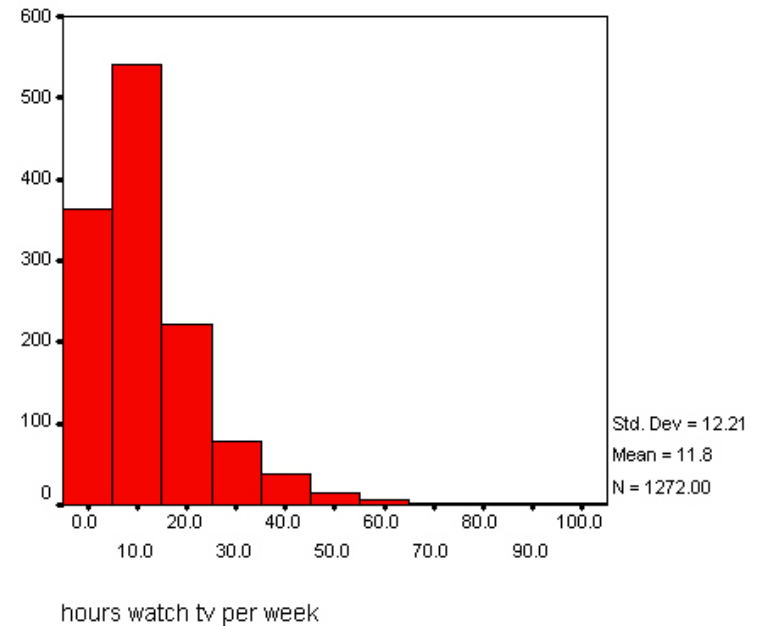
Guidelines for Grouping Values

- Categories should be...
 - logically defensible and preserve the distribution's shape
 - mutually exclusive and exhaustive
- Consider the level of measurement
 - Normally used for interval and ratio variables, but sometimes is appropriate for ordinal
 - Nominal variables can be grouped, but grouping is based on characteristics of the attributes, not on numerical values
- Intervals
 - First interval must contain the lowest value and the last interval must contain the highest value in the distribution
 - Each interval width, the number of values that fall within each interval, should be the same size.
 - Should be between 7 and 13 intervals

Grouped and Ungrouped Charts



Ungrouped



Grouped



Measures of Central Tendency

- Univariate
- How values of a variable cluster around the center of a distribution
- Three types
 - Mode
 - Median
 - Mean

Mode

- Value that occurs most often, most common value
- Any distribution may have more than one mode if there are two or more values that occur the same number of times
- Data may be grouped or ungrouped
- Valid measure for any level of measurement (nominal, ordinal, interval, ratio) – but not used often

2 8 9 0 3 45 7 8 3 4 6 5 8 8 3 2 4 6 9 44 27

2 8 9 0 3 45 7 8 7 4 7 5 8 8 3 7 4 6 9 44 27

bimodal distribution

Median

- Value in middle of distribution of values
- Half of the values are above the median and half the values are below the median
- Valid measure only for ordinal, interval, or ratio variables

Previous distribution, ordered from smallest value to largest value

0 2 2 3 3 3 4 4 5 6 **6** 7 8 8 8 8 9 9 27 44 45

$$\frac{N + 1}{2} = \frac{21 + 1}{2} = 11$$

N = number of elements in population/distribution

This formula gives the *POSITION* of the median – not its VALUE

▪ Here, median is 11th value from bottom (or top) of distribution, when ordered from smallest to largest (or largest to smallest) = 6

Median, continued

If distribution has an even number of values...

0 2 2 3 3 3 4 4 5 6 6 7 8 8 8 8 9 9 27 44 45 50

$$P1 = \frac{N}{2}$$

$$P2 = \frac{N+2}{2}$$

$$\text{Position of Median} = \frac{P1+P2}{2}$$

$$P1 = \frac{22}{2} = 11 \quad P2 = \frac{22+2}{2} = 12 \quad \text{Position of Median} = \frac{11+12}{2} = \frac{23}{2} = 11.5$$

$$VP1 = 6$$

$$VP2 = 7$$

$$\text{VALUE of Median} = \frac{VP1+VP2}{2} = \frac{6+7}{2} = 6.5$$

Value of P1

Note that in a distribution with an even number of values, the median may not be a value that actually appears in the distribution

The Mean

- Arith^metic mean, common average
- Requires interval or ratio data
- Sum of all values divided by number of values in distribution
- Sensitive to extreme values

In algebraic notation, the equation is:

$$\bar{x} = \frac{\sum_1^N x_i}{N}$$

\bar{x} = mean

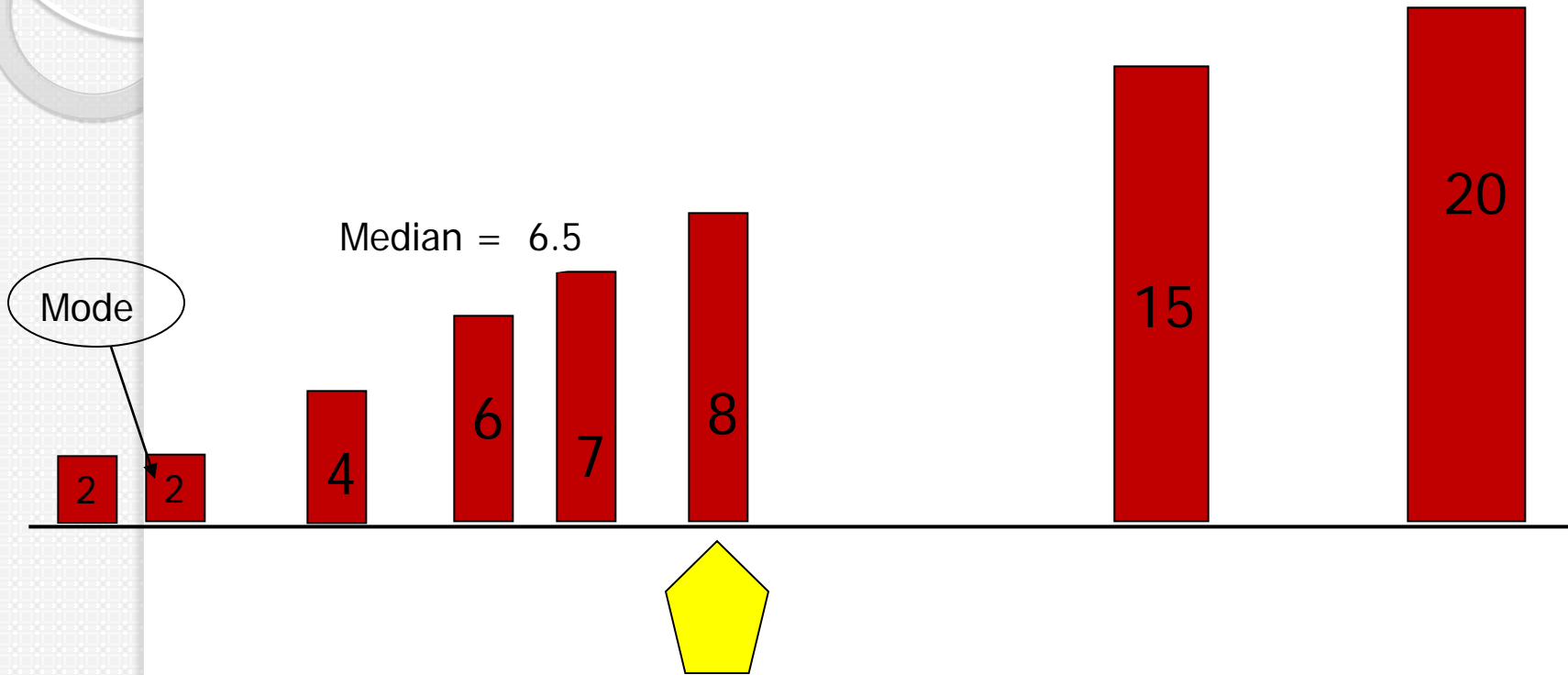
N = number of cases;

Σ = sum over all cases;

x_i = value of case i on variable x

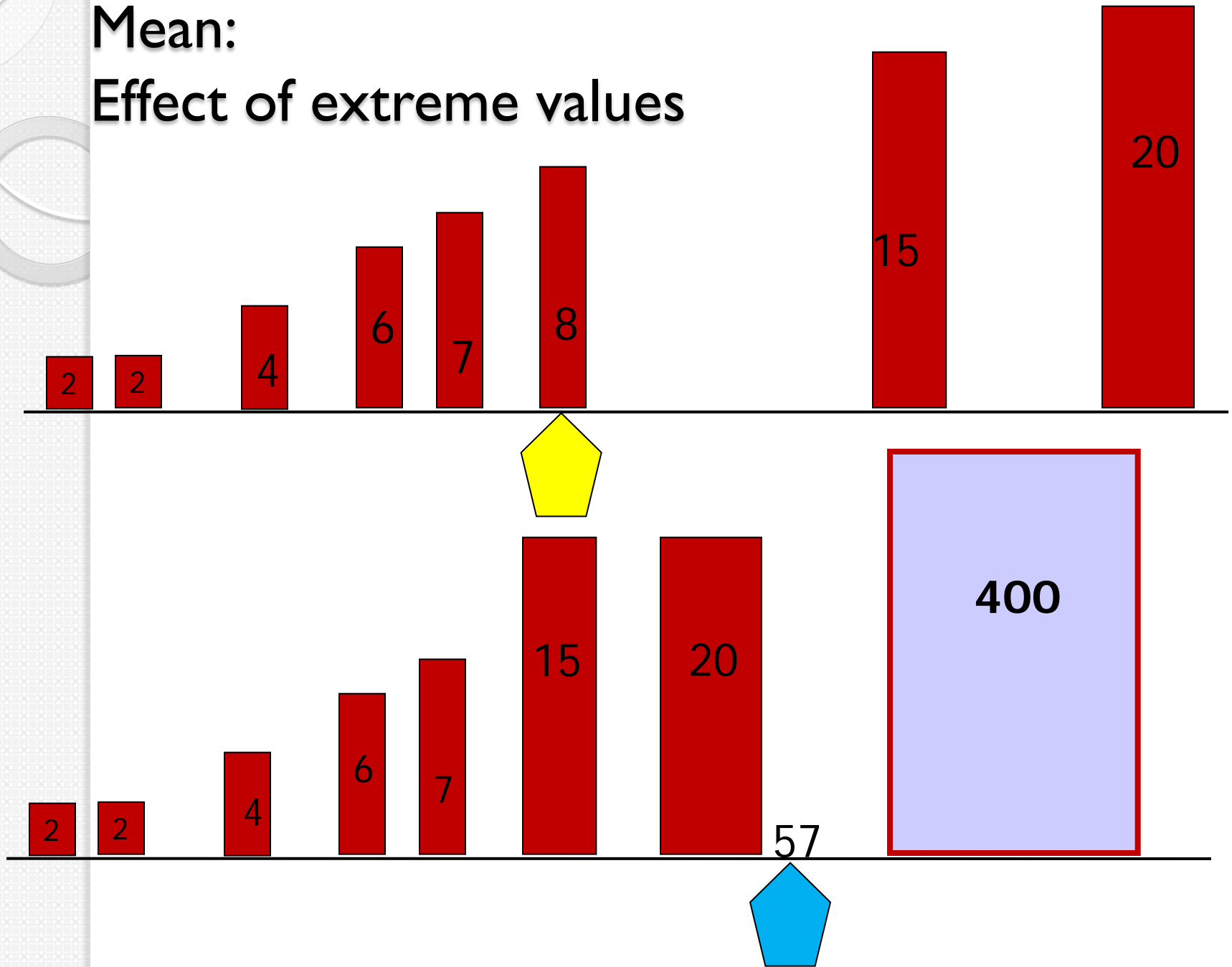
\sum_1^N = sum of all values from
1 (1st value) to
N (total number of elements)

Calculating the Mean



$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{2+2+4+6+7+8+15+20}{8} = \frac{64}{8} = 8$$

Mean: Effect of extreme values



Why calculate both the Median and the Mean?

- Each statistic tells something different about the distribution
- Mean
 - Reports the average across all scores
- Median
 - Tells what value is in middle of distribution, regardless of actual values of scores
- Reporting both tells something about whether there are extreme values



Measures of Variation (Dispersion)

Measures of variation capture how widely or densely spread the values are for the variable of interest.

- Range
- Interquartile Range
- Variance
- Standard Deviation

The Range

- The **range** is the simplest measure of variation
- Appropriate for ordinal, interval, and ratio levels of measurement
- Range of a distribution is calculated as:
 - $(\text{highest value} - \text{lowest value}) + 1$
- It often is important to report the range of a distribution, to identify the whole span (range) of possible values that might be encountered.

More about the Range

Say that you surveyed 11 college students and asked them how many times they have been arrested. Their answers looked like this:

However, since the range can be drastically altered by just one exceptionally high or low value (an **outlier**), it is not a good summary measure for most purposes.

| Number of times respondent was arrested |
|---|
| 0 |
| 2 |
| 2 |
| 3 |
| 4 |
| 4 |
| 5 |
| 20 |
| 2 |
| 1 |
| 6 |

Variance

- Requires Interval or Ratio data
- Average squared deviation of each case from the mean; take each case's distance from the mean, square that number, and take the average of all such numbers

$$\sigma^2 = \frac{\sum_1^N (x - \bar{x})^2}{N}$$

\bar{x}

= mean

N

= number of cases;

Σ

= sum over all cases;

x_i

= value of case i on variable x

\sum_1^N

= sum of all values from
1 (1st value) to
 N (total number of elements)

Calculating the Variance

| Number of times respondent was arrested | $(\bar{x}_i - \bar{x})$ | $(\bar{x}_i - \bar{x})^2$ |
|---|-------------------------|---------------------------|
| 0 | -4.45 | 19.8 |
| 2 | -2.45 | 6.0 |
| 2 | -2.45 | 6.0 |
| 3 | -1.45 | 2.1 |
| 4 | -.45 | .2 |
| 4 | -.45 | .2 |
| 5 | .55 | .3 |
| 20 | 15.55 | 241.8 |
| 2 | -2.45 | 6.0 |
| 1 | -3.45 | 11.9 |
| 6 | 1.55 | 2.4 |
| Total = 49 | 0 | 296.7 |

$$\sigma^2 = \frac{\sum_{i=1}^N (x - \bar{x})^2}{N}$$

$$\sigma^2 = \frac{296.7}{10}$$

$$\sigma^2 = 29.67$$

Characteristics of the Variance

- Takes into account the amount by which each case differs from the mean
- Affected by outliers, like the person who was arrested 20 times
- Mainly useful for computing the standard deviation

Standard Deviation

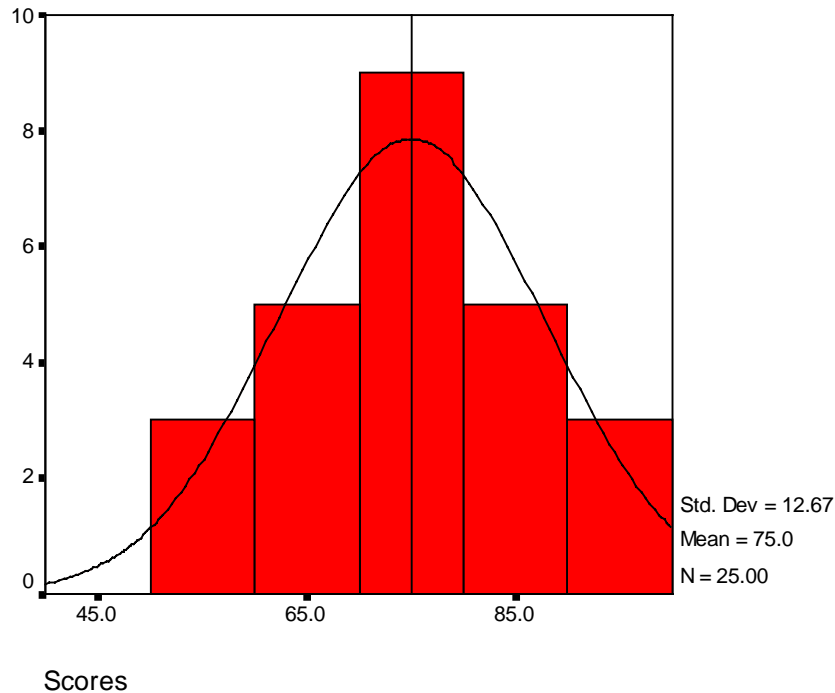
- Square root of variance (σ^2)

$$\sigma = \sqrt{\sigma^2}$$

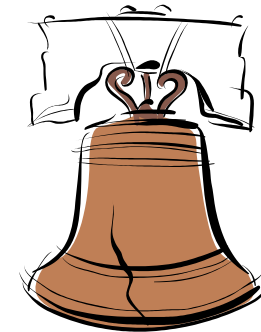
or

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}}$$

The Normal Distribution



Looks like a bell, with one “hump” in the middle, centered around the population mean, and the number of cases tapering off on both sides of the mean



Symmetric: If you folded it in half at its center (the population mean), the two halves would match perfectly

Required Level of Measurement

| | Nominal | Ordinal | Interval/Ratio |
|---------------------|---------|---------|----------------|
| Mode | X | X | X |
| Median | | X | X |
| Mean | | | X |
| Range | | | X |
| Interquartile Range | | | X |
| Variance | | | X |
| Standard Deviation | | | X |

Many researchers believe it is okay to use the mean, range, variance and standard deviation with variables measured at the ordinal level, too.



Summary Statistics

... describe particular features of a distribution and facilitate comparison among distributions.

The next step is to test for associations....

Crosstabulation

- Bivariate and multivariate comparisons are often presented in **crosstabulations** (“crosstabs”)
- A crosstabulation displays the distribution of one variable for each category of another variable
 - also called a **bivariate distribution**

Anatomy of a Table

| | | Columns | | | Row Marginal |
|------|---------------------|----------------|---------------------|-------------------------|--------------|
| | | Race/Ethnicity | | | Total |
| ROWS | Favor Death Penalty | Black | White/ Non-Hispanic | Asian/ Pacific Islander | Total |
| | Yes | 58 | 165 | 25 | 248 |
| | No | 68 | 92 | 9 | 169 |
| | Don't Know | 9 | 73 | 1 | 83 |
| | Total | 135 | 330 | 35 | 500 |

Column Marginal

Rules for Crosstabulating Variables

1. Make the independent variable the column variable and the dependent variable the row variable
2. Percentage the table column by column, on the column totals. The percentages in each column should add to 100 (or perhaps between 99 and 101, if there has been rounding error)
3. Compare the distributions of the dependent variable (the row variable) across each column

Crosstab Examples

Any crime for which or set of circumstances in which respondent favors the death penalty:

Frequency Table

| | Race/Ethnicity | | | |
|---------------|----------------|------------|-------------------------------|------------|
| | Black | White | Asian/ Pacific Islander | Total |
| Favor | | | | |
| Yes | 58 | 165 | 25 | 300 |
| No | 68 | 92 | 9 | 150 |
| Don't Know | 9 | 73 | 1 | 50 |
| Total | 135 | 330 | 35 | 500 |

Percentage Table

| | Race/Ethnicity | | |
|---------------|----------------|-------------|---------------------------------|
| | % Black | % White | % Asian/ Pacific Islander |
| Favor | | | |
| Yes | 43 | 50 | 72 |
| No | 50 | 28 | 25 |
| Don't Know | 7 | 22 | 3 |
| Total | 100% | 100% | 100% |

Describing Association

- Existence
 - Do the percentage distributions vary at all between categories of the independent variable?
- Strength
 - How much do the percentage distributions vary between categories of the independent variable?
- Direction
 - For quantitative variables, do values on the dependent variable tend to increase or decrease with an increase in value on the independent variable?
- Pattern
 - For quantitative variables, are changes in the percentage distribution of the dependent variable fairly regular (simply increasing or decreasing), or do they vary (perhaps increasing, then decreasing, or perhaps gradually increasing, then rapidly increasing)?