Lecture Slides

Essentials of Statistics
5th Edition

and the Triola Statistics Series

by Mario F. Triola

Chapter 2
Summarizing and Graphing Data

2-1 Review and Preview

2-2 Frequency Distributions

2-3 Histograms

2-4 Graphs that Enlighten and Graphs that Deceive

Preview
Characteristics of Data

1. Center: A representative value that indicates where the middle of the data set is located.
2. Variation: A measure of the amount that the data values vary.
3. Distribution: The nature or shape of the spread of data over the range of values (such as bell-shaped, uniform, or skewed).
4. Outliers: Sample values that lie very far away from the vast majority of other sample values.
5. Time: Changing characteristics of the data over time.

11

Chapter 2
Summarizing and Graphing Data

2-1 Review and Preview

2-2 Frequency Distributions

2-3 Histograms

2-4 Graphs that Enlighten and Graphs that Deceive

Key Concept

When working with large data sets, it is often helpful to organize and summarize data by constructing a table called a frequency distribution.

Because computer software and calculators can generate frequency distributions, the details of constructing them are not as important as what they tell us about data sets.

Definition

- Frequency Distribution (or Frequency Table)

  shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number (frequency) of data values in each of them.

IQ Scores of Low Lead Group

Lower Class Limits

are the smallest numbers that can actually belong to different classes.

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Upper Class Limits

are the largest numbers that can actually belong to different classes.

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Class Boundaries

are the numbers used to separate classes, but without the gaps created by class limits.

49.5
69.5
89.5
109.5
129.5
149.5

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Class Midpoints

are the values in the middle of the classes and can be found by adding the lower class limit to the upper class limit and dividing the sum by 2.

59.5
79.5
99.5
119.5
139.5

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

IQ Scores of Low Lead Group

Class Width

is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

20
20
20
20
20

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1

Reasons for Constructing Frequency Distributions

1. Large data sets can be summarized.
2. We can analyze the nature of data.
3. We have a basis for constructing important graphs.

Constructing A Frequency Distribution

1. Determine the number of classes (should be between 5 and 20).
2. Calculate the class width (round *up*).

$$\text{class width} \approx \frac{\text{(maximum value)} - \text{(minimum value)}}{\text{number of classes}}$$

3. Starting point: Choose the minimum data value or a convenient value below it as the first lower class limit.
4. Using the first lower class limit and class width, proceed to list the other lower class limits.
5. List the lower class limits in a vertical column and proceed to enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to get the frequency.

Relative Frequency Distribution

includes the same class limits as a frequency distribution, but the frequency of a class is replaced with a relative frequencies (a proportion) or a percentage frequency (a percent)

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}}$$

$$\text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequencies}} \times 100\%$$

Relative Frequency Distribution

IQ Score	Frequency	Relative Frequency
50-69	2	2.6%
70-89	33	42.3%
90-109	35	44.9%
110-129	7	9.0%
130-149	1	1.3%

Cumulative Frequency Distribution

IQ Score	Frequency	Cumulative Frequency
50-69	2	2
70-89	33	35
90-109	35	70
110-129	7	77
130-149	1	78

Cumulative Frequencies

Critical Thinking: Using Frequency Distributions to Understand Data

In later chapters, there will be frequent reference to data with a normal distribution. One key characteristic of a normal distribution is that it has a "bell" shape.

- The frequencies start low, then increase to one or two high frequencies, and then decrease to a low frequency.
- The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

Gaps

- Gaps

The presence of gaps can show that we have data from two or more different populations.

However, the converse is not true, because data from different populations do not necessarily result in gaps.

Example

- The table on the next slide is a frequency distribution of randomly selected pennies.
- The weights of pennies (grams) are presented, and examination of the frequencies suggests we have two different populations.
- Pennies made before 1983 are 95% copper and 5% zinc.
- Pennies made after 1983 are 2.5% copper and 97.5% zinc.

Example (continued)

Weight (grams) of Penny	Frequency
2.40–2.49	18
2.50–2.59	19
2.60–2.69	0
2.70–2.79	0
2.80–2.89	0
2.90–2.99	2
3.00–3.09	25
3.10–3.19	8

The presence of gaps can suggest the data are from two or more different populations.

Chapter 2 Summarizing and Graphing Data

2-1 Review and Preview

2-2 Frequency Distributions

2-3 Histograms

2-4 Graphs that Enlighten and Graphs that Deceive

Key Concept

We use a visual tool called a histogram to analyze the shape of the distribution of the data.

Histogram

A graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data)

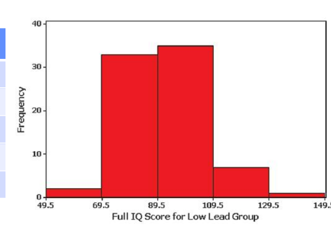
The horizontal scale represents the classes of quantitative data values and the vertical scale represents the frequencies.

The heights of the bars correspond to the frequency values.

Example

IQ scores from children with low levels of lead.

IQ Score	Frequency
50-69	2
70-89	33
90-109	35
110-129	7
130-149	1



Histogram

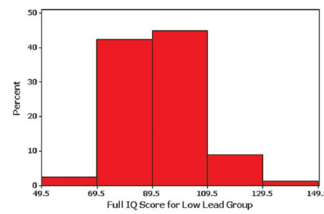
A histogram is basically a graph of a frequency distribution.

Histograms can usually be generated using technology.

Relative Frequency Histogram

has the same shape and horizontal scale as a histogram, but the vertical scale is marked with relative frequencies instead of actual frequencies

IQ Score	Relative Frequency
50-69	2.6%
70-89	42.3%
90-109	44.9%
110-129	9.0%
130-149	1.3%



Critical Thinking Interpreting Histograms

Objective is not simply to construct a histogram, but rather to *understand* something about the data.

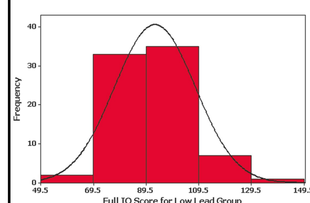
When graphed, a normal distribution has a "bell" shape. Characteristic of the bell shape are

(1) The frequencies increase to a maximum, and then decrease, and

(2) symmetry, with the left half of the graph roughly a mirror image of the right half.

The histogram on the next slide illustrates this.

Example – IQ Scores



- What is the shape of this distribution?
- What is the center?
- How much variation is in the data?
- Are there any outliers?

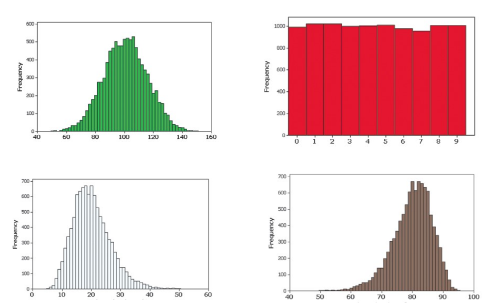
Skewness

A distribution of data is **skewed** if it is not symmetric and extends more to one side to the other.

Data **skewed to the right** (*positively skewed*) have a longer right tail.

Data **skewed to the left** (*negative skewed*) have a longer left tail.

Example – Discuss the Shape

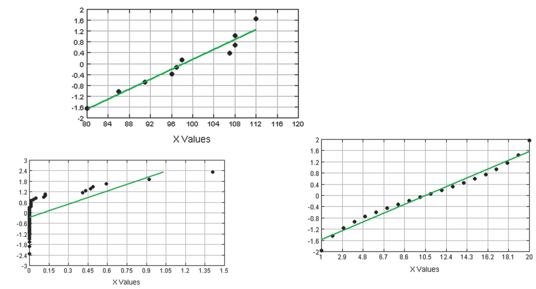


Assessing Normality with a Normal Quantile Plot

- Many methods we will use later in the text require that the sample data must be from a population with a normal distribution.

- A *normal quantile plot* can be interpreted on the following criteria:
  - Normal Distribution: Points are reasonably close to a straight line
  - Not a Normal Distribution: Points not reasonably close to a straight line or the points show some systemic pattern that is not straight

Assessing Normality with a Normal Quantile Plot



Chapter 2
Summarizing and Graphing Data

2-1 Review and Preview

2-2 Frequency Distributions

2-3 Histograms

2-4 Graphs that Enlighten and Graphs that Deceive

Key Concept

This section discusses other types of statistical graphs.

Our objective is to identify a suitable graph for representing the data set. The graph should be effective in revealing the important characteristics of the data.

Key Concept

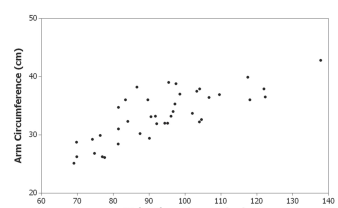
Some graphs are bad in the sense that they contain errors.

Some are bad because they are technically correct, but misleading.

It is important to develop the ability to recognize bad graphs and identify exactly how they are misleading.

Scatterplot (or Scatter Diagram)

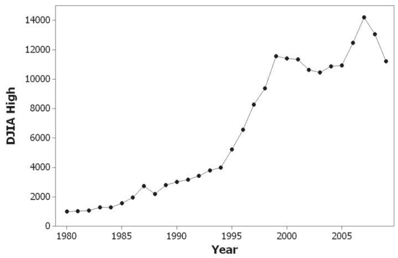
A plot of paired (*x*, *y*) quantitative data with a horizontal *x*-axis and a vertical *y*-axis. Used to determine whether there is a relationship between the two variables.



Randomly selected males – the pattern suggests there is a relationship.

Time-Series Graph

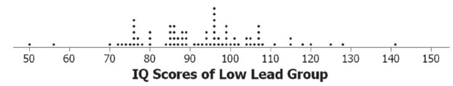
Data that have been collected at different points in time: *time-series data*



Yearly high values of the Dow Jones Industrial Average

Dotplot

Consists of a graph in which each data value is plotted as a point (or dot) along a scale of values. Dots representing equal values are stacked.



Stemplot (or Stem-and-Leaf Plot)

represents quantitative data by separating each value into two parts: the stem (such as the leftmost digit) and the leaf (such as the rightmost digit).

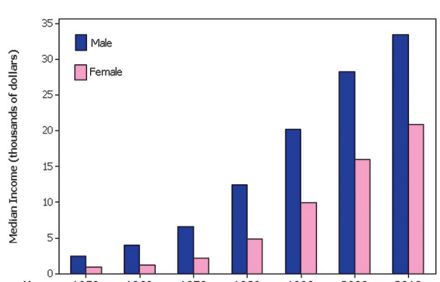
```
 5 | 06                      <- Lowest IQ scores are 50 and 56.
 6 |
 7 | 0234566666778
 8 | 000455556666677888999
 9 | 123444566666667789999
10 | 01124455677778
11 | 1558                    <- These IQ scores are 111, 115, 115, and 118.,
12 | 058
13 |
14 | 1
```


Bar Graph

Uses bars of equal width to show frequencies of categorical, or qualitative, data. Vertical scale represents frequencies or relative frequencies. Horizontal scale identifies the different categories of qualitative data.

A *multiple bar graph* has two or more sets of bars and is used to compare two or more data sets.

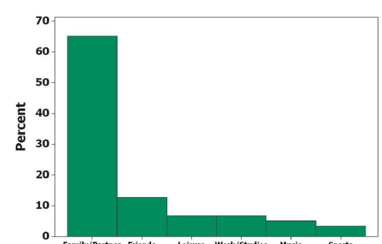
Multiple Bar Graph



Multiple Bar Graph: Median Income by Gender

Pareto Chart

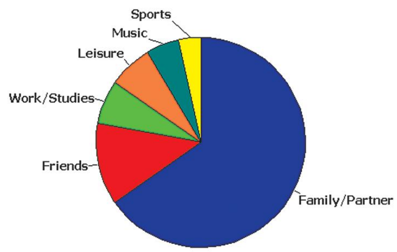
A bar graph for qualitative data, with the bars arranged in descending order according to frequencies



Pareto Chart: What Contributes Most to Happiness?

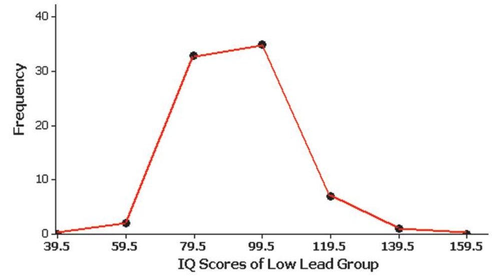
Pie Chart

A graph depicting qualitative data as slices of a circle, in which the size of each slice is proportional to frequency count



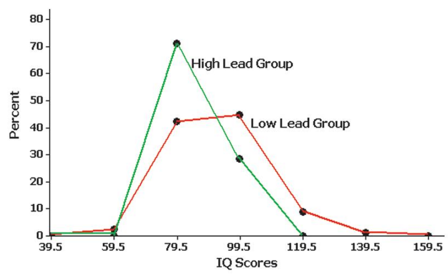
Frequency Polygon

uses line segments connected to points directly above class midpoint values.



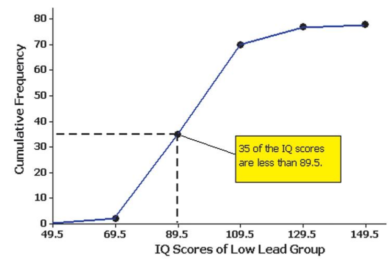
Relative Frequency Polygon

Uses relative frequencies (proportions or percentages) for the vertical scale.



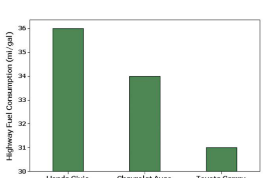
Ogive

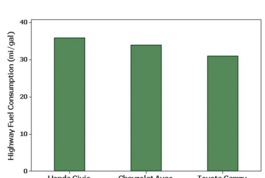
A line graph that depicts cumulative frequencies



Graphs That Deceive

Nonzero Axis: Graphs can be misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.





Pictographs

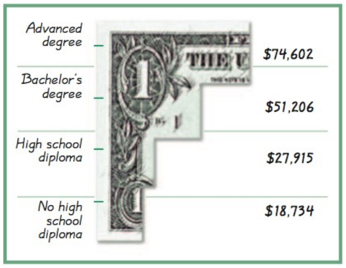
Drawings of objects. Three-dimensional objects - money bags, stacks of coins, army tanks (for army expenditures), people (for population sizes), barrels (for oil production), and houses (for home construction) are commonly used to depict data.

These drawings can create false impressions that distort the data.

If you double each side of a square, the area does not merely double; it increases by a factor of four; if you double each side of a cube, the volume does not merely double; it increases by a factor of eight.

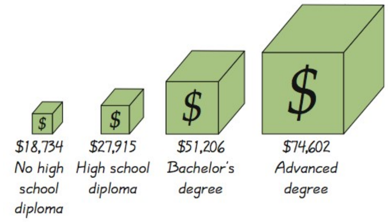
Pictographs using areas and volumes can therefore be very misleading.

Example – Income and Education



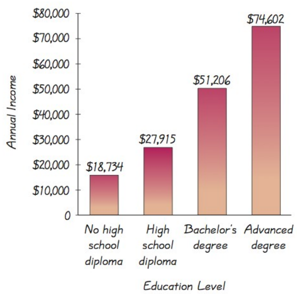
Bars have same width, too busy, too difficult to understand.

Example – Income and Education



Misleading. Depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.

Example – Income and Education



Fair, objective, unencumbered by distracting features.

Important Principles Suggested by Edward Tufte

For small data sets of 20 values or fewer, use a table instead of a graph.

A graph of data should make the viewer focus on the true nature of the data, not on other elements, such as eye-catching but distracting design features.

Do not distort data. Construct a graph to reveal the true nature of the data.

Almost all of the ink in a graph should be used for the data, not for the other design elements.