CHAPTER 2
Frequency Distributions and Graphs

Objectives

• Organize data using frequency distributions.
• Represent data in frequency distributions graphically using histograms, frequency polygons, and ogives.
• Represent data using Pareto charts, time series graphs, and pie graphs.
• Draw and interpret a stem and leaf plot.
• Draw and Interpret a scatter plot for a set of paired data.

Introduction

This chapter will show how to organize data and then construct appropriate graphs to represent the data in a concise, easy-to-understand form.

Section 2.1 Organizing Data

Basic Vocabulary

• When data are collected in original form, they are called raw data.

• A frequency distribution is the organization of raw data in table form, using classes and frequencies.

• The two most common distributions are categorical frequency distribution and the grouped frequency distribution.

Frequency Distributions

Categorical Frequency Distributions count how many times each distinct category has occurred and summarize the results in a table format

Example 1: Letter grades for Math 227 Spring 2005:

<table>
<thead>
<tr>
<th>C</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>F</th>
<th>B</th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>F</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>D</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>F</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

a) Construct a frequency distribution for the categorical data.

b) What percentage of the students pass the class with the grade C or better?
Frequency Distributions

Grouped Frequency Distributions—When the range of the data is large, the data must be grouped into classes that are more than one unit in width

Basic Vocabulary
The lower class limit represents the smallest value that can be included in the class.

The upper class limit represents the largest value that can be included in the class.

The class boundaries are used to separate the classes so that there are no gaps in the frequency distribution.

Class Boundaries Significant Figures

• Rule of Thumb: Class limits should have the same decimal place value as the data, but the class boundaries have one additional place value and end in a 5.

  e.g. data were whole numbers
  lower class boundary = lower class limit – 0.5
  upper class boundary = upper class limit + 0.5

  e.g. data were one decimal place
  lower class boundary = lower class limit – 0.05
  upper class boundary = upper class limit +0.05

Class Midpoints
The class midpoint (mark) is found by adding the lower and upper boundaries (or limits) and dividing by 2.

Class Width
The class width for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the the lower (or upper) class limit of the next class.
Class Width Class Rules
• There should be between 5 and 20 classes.
• The class width should be an odd number.
• The classes must be mutually exclusive.
• The classes must be continuous.
• The classes must be exhaustive.
• The classes must be equal width.

Class width as an odd number
The class width being an odd number is preferable since it ensures that the midpoint of each class has the sample place value as the data. If the class width is an even number, the midpoint is in tenths. For example, if the class width is 6 and the class limits are 6 and 11, the midpoint is:

Relative Frequency
Relative Frequency is the frequency of each class divided by the total number.

Cumulative Frequency
Cumulative Frequency is the sum of the frequencies accumulated up to the upper boundary of a class.

Procedure for constructing a grouped frequency distribution
1. Decide on the number of classes you want. (5 to 20 classes)
2. Calculate (round up) the class width
3. Choose a number for the lower limit of the first class
4. Use the lower limit of the first class and the class width to list the other lower class limits.
5. Enter the upper class limits.
6. Tally the frequency for each class
Example 1: Construct a grouped frequency table for the following data values.

44, 32, 35, 38, 35, 39, 42, 36, 36, 40, 51, 58
58, 62, 63, 72, 78, 81, 25, 84, 20

Tip: Consider reordering the data.

Frequency Distributions
An ungrouped frequency distribution is used for numerical data and when the range of data is small.

Example: The number of incoming telephone calls per day over the first 25 days of business:

4, 4, 1, 10, 12, 6, 4, 6, 9, 12, 12, 1, 1, 12, 10, 4, 6, 4, 8, 8, 9, 8, 4, 1

Construct an ungrouped frequency distribution

Types of Frequency Distributions (summary)

A categorical frequency distribution is used when the data is nominal.
• A grouped frequency distribution is used when the range is large and classes of several units in width are needed.

• An ungrouped frequency distribution is used for numerical data and when the range of data is small.

**Why Construct Frequency Distributions?**

• To organize the data in a meaningful, intelligible way.
• To enable the reader to make comparisons among different data sets.
• To facilitate computational procedures for measures of average and spread.
• To enable the reader to determine the nature or shape of the distribution.
• To enable the researcher to draw charts and graphs for the presentation of data.

**Section 2.2 Histogram, Frequency, Polygons, Ogives**

This chapter will show how to organize data and then construct appropriate graphs to represent the data in a concise, easy-to-understand form.

**The Role of Graphs**

• The purpose of graphs in statistics is to convey the data to the viewer in pictorial form.
• Graphs are useful in getting the audience’s attention in a publication or a presentation.

**Three Most Common Graphs**

• The histogram displays the data by using vertical bars of various heights to represent the frequencies.

• The frequency polygon displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes.

• The cumulative frequency or ogive represents the cumulative frequencies for the classes in a frequency distribution.

**Relative Frequency Graphs**

• A relative frequency graph is a graph that uses proportions instead of frequencies. Relative frequencies are used when the proportion of data values that fall into a given class is more important than the frequency.

**Example 1:**

The following data are the number of the English-language Sunday Newspaper per state in the United States as of February 1, 1996.

| 2 3 3 4 4 4 4 4 5 6 6 6 7 7 8 10 11 11 11 12 12 13 14 14 14 15 15 16 16 16 16 18 18 19 21 21 23 27 31 35 37 38 39 40 44 62 85 |
a) Using 1 as the starting value and a class width of 15, construct a grouped frequency distribution.

b) Construct a histogram for the grouped frequency distribution.
   (x-axis: class boundaries; y-axis: frequency)

c) Construct a frequency polygon.
   (x-axis: class midpoints(marks); y-axis: frequency)

d) Construct an ogive.
   (x-axis: class boundaries; y-axis: cumulative frequency)

e) Construct a (i) relative frequency histogram,
   (ii) relative frequency polygon,
   and (iii) relative cumulative frequency ogive.
### Distribution shapes

#### Section 2.3 Other Types of Graphs

A **Pareto chart** is used to represent a frequency distribution for categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

(x-axis: categorical variables; y-axis: frequencies, which are arranged in order from highest to lowest)

A **pie graph** is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

**Example 1:** Grade received for Math 227

| C | A | B | B | D | C | C | C | B | B | A | F | F |

a) Construct a pareto chart.

b) Construct a pie chart.
Other Types of Graphs (cont.)

• A time series graph represents data that occur over a specific period of time.

Example 1: The percentages of voters voting in the last 5 Presidential elections are shown here. Construct a time series graph.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>% of voters voting</td>
<td>74.63%</td>
<td>72.48%</td>
<td>78.01%</td>
<td>65.97%</td>
<td>67.50%</td>
</tr>
</tbody>
</table>

Stem-and-Leaf Plots

• A stem-and-leaf plot is a data plot that uses part of a data value as the stem and part of the data value as the leaf to form groups or classes.

• It has the advantage over grouped frequency distribution of retaining the actual data while showing them in graphic form. Digits of each data to the left of a vertical bar are called the stems.
Digits of each data to the right of the appropriate stem are called the leaves.

Example 1: The test scores on a 100-point test were recorded for 20 students

\[
\begin{align*}
61 & \quad 93 & \quad 91 & \quad 86 & \quad 55 & \quad 63 & \quad 86 & \quad 82 & \quad 76 & \quad 57 \\
94 & \quad 89 & \quad 67 & \quad 62 & \quad 72 & \quad 87 & \quad 68 & \quad 65 & \quad 75 & \quad 84
\end{align*}
\]

Construct an ordered stem-and-leaf plot

Example 2: Use the data in example 1 to construct a double stem and leaf plot. e.g. split each stem into two parts, with leaves 0 – 4 on one part and 5 – 9 on the other.

A stem-and-leaf plot portrays the shape of a distribution and restores the original data values. It is also useful for spotting outliers. Outliers are data values that are extremely large or extremely small in comparison to the norm.

2.4 Paired Data and Scatter Plots

- Many times researchers are interested in determining if a relationship between two variables exist.
- To do this, the researcher collects data consisting of two measures that are paired with another.
- The variable first mentioned is called the independent variable; the second variable is the dependent variable.

Scatter Plot – is a graph of order pairs values that is used to determine if a relationship exists between two variables.

Analyzing the Scatter Plot
• A **positive linear relationship** exists when the points fall approximately in an ascending straight line and both the \(x\) and \(y\) values increase at the same time.

• A **negative linear relationship** exists when the points fall approximately in a straight line descending from left to right.

• A **nonlinear relationship** exists when the points fall along a curve.

• **No relationship** exists when there is no discernable pattern of the points.

Examples of Scatter Plots and Relationships

Example 1: A researcher wishes to determine if there is a relationship between the number of days an employee missed a year and the person's age. Draw a scatter plot and comment on the nature of the relationship.

| Age, \(x\) | 22 | 30 | 25 | 35 | 65 | 50 | 27 | 53 | 42 | 58 |
| Days missed, \(y\) | 0  | 4  | 1  | 2  | 14 | 7  | 3  | 8  | 6  | 4  |
Summary of Graphs and Uses

- *Histograms*, *frequency polygons*, and *ogives* are used when the data are contained in a grouped frequency distribution.
- *Pareto charts* are used to show frequencies for nominal variables.
- *Time series graphs* are used to show a pattern or trend that occurs over time.
- *Pie graphs* are used to show the relationship between the parts and the whole.
- When data are collected in pairs, the relationship, if one exists, can be determined by looking at a *scatter plot*.

Conclusions

- Data can be organized in some meaningful way using frequency distributions. Once the frequency distribution is constructed, the representation of the data by graphs is a simple task.