Statistics for the Biological, Environmental and Health Sciences

STAT 007

Introduction to Statistics

Chapter 1

Collecting Sample Data

Section 1-3

- In this section we will:
  - Describe methods for collecting data sets.
  - Describe methods for collecting samples.
  - Describe different types of data collection methods.

Sampling Errors

Definition

- *A **sampling error** (or random sampling error) occurs when the sample has been selected with a random method, but there is a discrepancy between a sample result and the true population result.*

  Comment: such an error results from chance sample fluctuations.

- *A **nonsampling error** is the result of human error, such as wrong data entries, computing errors, questions with biased wording, false data provided by respondents, forming biased conclusions, or applying statistical methods that are not appropriate for the circumstances.*

- *A **nonrandom sampling error** is the result of using a sampling method that is not random, such as using a convenience sample or a voluntary response sample.*

Practice

Look at the exercises at the end of Section 1-3 in page 31.

Specially, look at exercises:
1 to 36.

Exploring Data with Tables and Graphs

Chapter 2

Frequency Distributions for Organizing and Summarizing Data

Section 2-1

- In this section we will:
  - Discuss Frequency Distributions to summarize data.

- Information in data sets needs to be summarized somehow.
- In what follows we discuss *frequency distributions (or frequency tables)* for organizing and summarizing data from quantitative and categorical variables in a data set.
- Frequency distributions help us understand the nature of the distribution of the variables in a data set.
- Roughly, the distribution of a variable is the shape of the spread of the data over a range of values.

Frequency Distribution

Definition

*A **frequency distribution (or frequency table)** shows how data are partitioned among several categories (or classes) by listing the categories along with the number (frequency) of data values in each of them.*

*Terms used in a frequency distribution: **Lower class limit**, **Upper class limit**, **Class Boundary**, **Class Midpoint**, and **Class width**.*

Frequency Distribution

Example

Consider the following data of IQ scores:

96; 87; 101; 103; 127; 96; 88; 85; 97; 124

The following table corresponds to the frequency distribution of the IQ scores:

IQ	Frequency
85 - 93	3
94 - 102	4
103 - 111	1
112 - 120	0
121 - 129	2

Steps for constructing a frequency distribution:

a) Choose number of classes: between 5 and 20. Here 5.

b) Set the **class width**: $\frac{\text{maximum value} - \text{minimum value}}{\text{number of classes}} = \frac{127-85}{5} = 8.4$. Round to 9.

c) Choose the first **lower class limit**: either the minimum value or a convenient value. Here the minimum: 85.

d) Given the class width, get lower class limits for all classes by adding the class width to the previous lower class limit.
Here: 85; $85 + 9 = 94$; $94 + 9 = 103$; $103 + 9 = 112$; $112 + 9 = 121$.

e) Set all **upper class limits** as largest number for each class limit (without changing classes!).
Here: 93; 102; 111; 120; 129.

f) Find the **class boundary** for each class.
Here: 84.5; $\frac{93+94}{2} = 93.5$; $\frac{102+103}{2} = 102.5$; $\frac{111+112}{2} = 111.5$; $\frac{120+121}{2} = 120.5$; 129.5.

g) Assign each individual data to the class the belong and then count how many data values are in each class.

h) Find the **class midpoints**:
Here $\frac{85+93}{2} = 89$; $\frac{94+102}{2} = 98$; $\frac{103+111}{2} = 107$; $\frac{112+120}{2} = 116$; $\frac{121+129}{2} = 125$.

Variations of Frequency Distributions

- **Relative Frequency Distribution**: each class frequency is replaced by a relative frequency or a percentage frequency.
- **Cumulative Frequency Distribution**: each class frequency is replaced by the sum of the frequency of that class and all previous classes.

Example

Consider the following data of IQ scores:

96; 87; 101; 103; 127; 96; 88; 85; 97; 124

IQ	Frequency	Relative Frequency $\frac{\text{Frequency}}{\text{sample size}}$	Percentage Frequency Relative Frequency $\times$ 100%	Cumulative Frequency
85 - 93	3	0.3	30	3
94 - 102	4	0.4	40	7
103 - 111	1	0.1	10	8
112 - 120	0	0	0	8
121 - 129	2	0.2	20	10
	10	1	100	

Frequency Distributions for categorical variables

- For categorical frequency distributions, the classes correspond to the different categories (or labels) of the variable.

Example

The following table summarizes information regarding the highest seven sources of injuries resulting in a visit to a hospital emergency room in a recent year (based on data from the Centers for Disease Control and Prevention). The activity names are “Bicycling’, ’‘Football”, “Playground”, “Basketball”, “Soccer”, “Baseball”, and “All-terrain vehicle”.

TABLE 2-3 Annual ER Visits for Injuries from Sports and Recreation

Activity	Frequency
Bicycling	26,212
Football	25,376
Playground	16,706
Basketball	13,987
Soccer	10,436
Baseball	9,634
All-terrain vehicle	6,337

Practice

Look at the exercises at the end of Section 2-1 in page 48.

Specially, look at exercises:
1, 2, 3, 4, 5, 6, 7, 8, 13, 15, 19, 20, 23, 25, 26.
Skip the question regarding the normal distribution for exercises 13 and 15.

Histograms

Section 2-2

- A *histogram* is basically a graph of a frequency distribution.
- Definition: A **histogram** is a graph consisting of bars of *equal width* drawn adjacent to each other (unless there are gaps in the data). The horizontal scale represents classes of *quantitative* data values, and the vertical scale represents frequencies. The heights of the bars correspond to frequency values.
- A histogram is used understand characteristics of the data. We can learn about "CVDOT":
  - Show the location of ***C****enter* of the data.
  - Show the ***V****ariation* of the data.
  - Visually display the shape of the ***D****istribution* of the data.
  - Identify ***O****utliers*.
  - Show whether there is any change of the characteristics of the data over ***T****ime*.
- Frequency histogram and relative frequency histograms: the only difference bertween frequency and relative frequency histograms is the vertical scale. Frequency histograms show counts, relative frequency show percentages or proportions.

Frequency Histogram

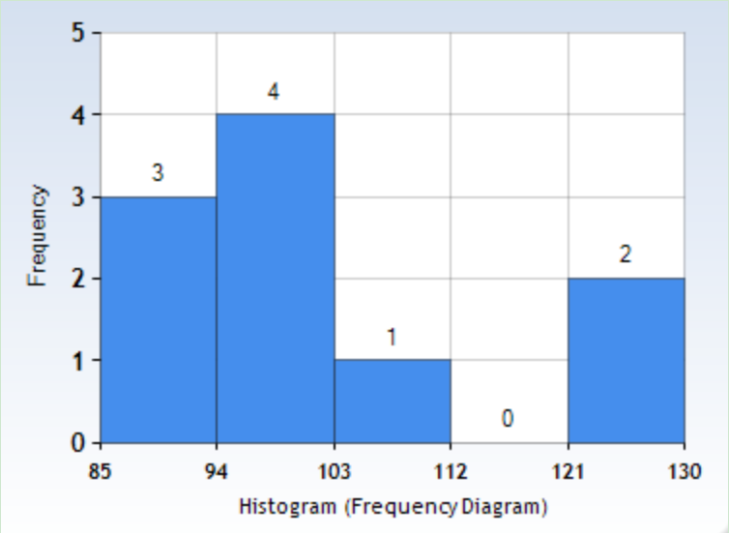
Example

Consider the following data of IQ scores:

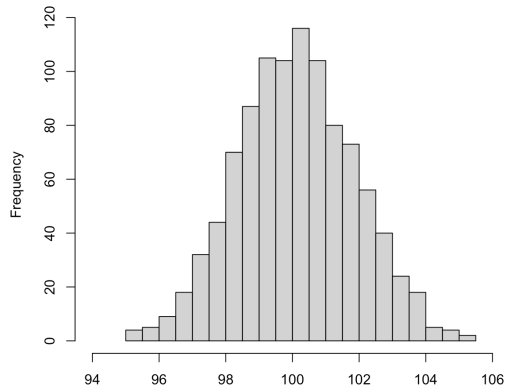
96; 87; 101; 103; 127; 96; 88; 85; 97; 124

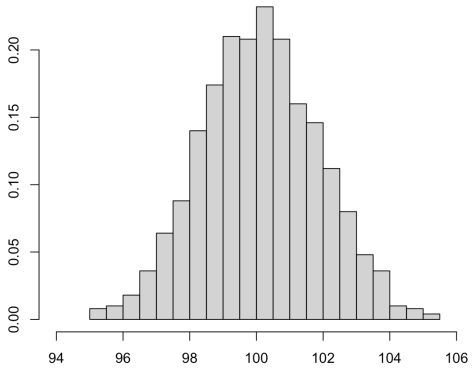
The following table corresponds to the frequency distribution of the IQ scores:

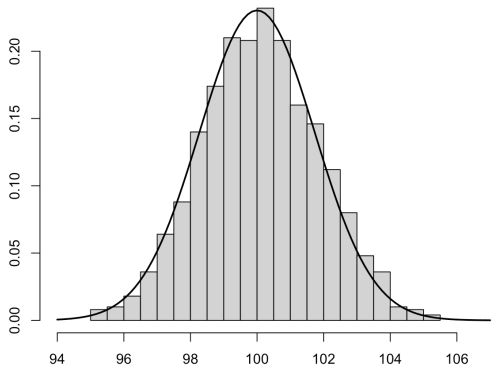
IQ	Frequency
85 - 93	3
94 - 102	4
103 - 111	1
112 - 120	0
121 - 129	2

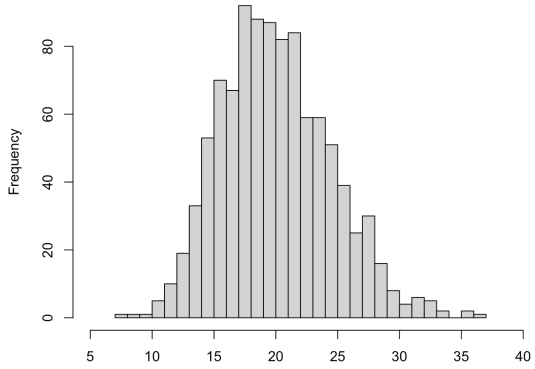


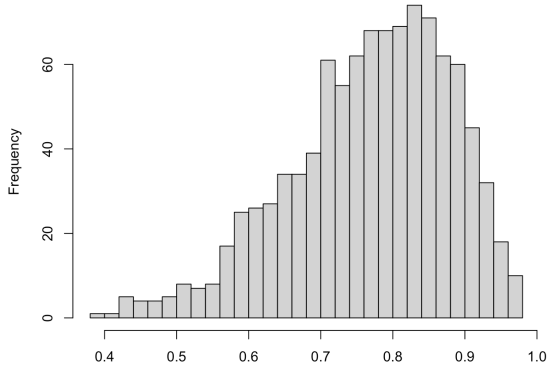
Histogram


Frequency
0
20
40
60
80
100
120
94
96
98
100
102
104
106


0.00
0.05
0.10
0.15
0.20
94
96
98
100
102
104
106


0.00
0.05
0.10
0.15
0.20
94
96
98
100
102
104
106


Frequency
0
20
40
60
80
5
10
15
20
25
30
35
40


Frequency
0
20
40
60
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Practice

Look at the exercises at the end of Section 2-2 in page 54.

Specially, look at exercises:
1, 2, 4, 5,6, 7, 8, 9, 10, 11, 12, 13, 14.

Graphs that Enlighten and Graphs that Deceive

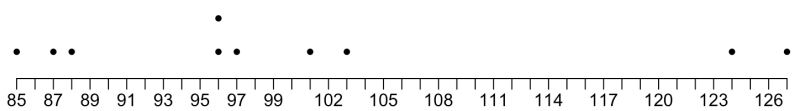
Section 2-3

- We will keep discussing plots for understanding data.
- We will group the types of graphs in three categories: graphs for quantitative data, graphs for categorical data, and graphs for measurements over time.
- We will discuss graphs that are deceptive because create misleading or wrong impressions about the data.
- Graphs should be constructed in a way that is fair and objective. The readers should be allowed to make their own judgments, instead of being manipulated by misleading graphs.

Graphs for Quantitative Data

Example

Consider the following data of IQ scores: `96; 87; 101; 103; 127; 96; 88; 85; 97; 124`



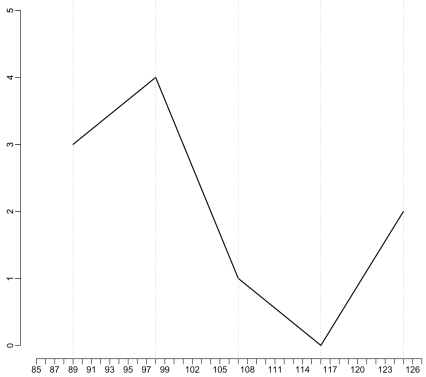
Dotplot Graph

- Displays shape of distribution of data
- Possible to recreate original data values

```
 8 | 578
 9 | 667
10 | 13
11 |
12 | 47
```

Stem-and-Leaf Graph

- Displays shape of distribution of data
- Retains original data values
- Sample data are sorted



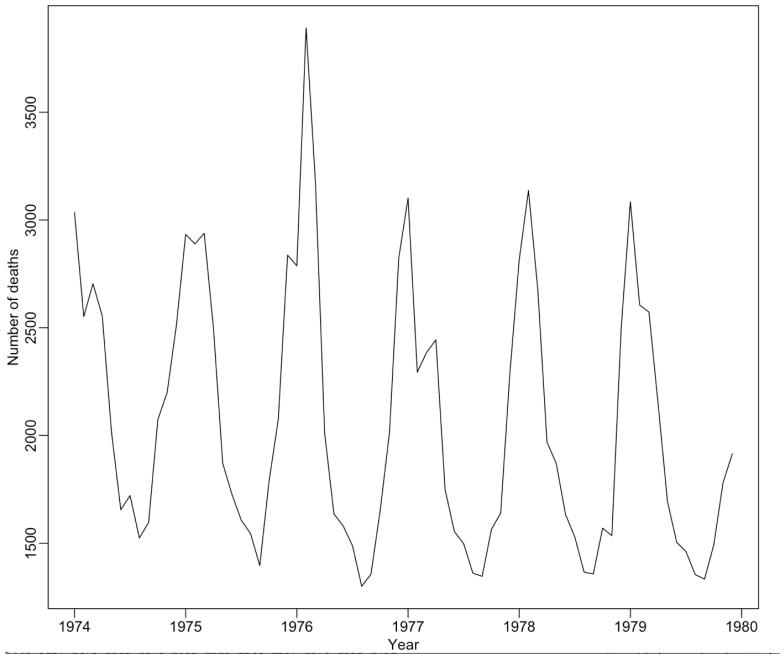
Frequency Polygon

- Similar to histograms, but instead of bars uses line segments joining class midpoints.
- Good for comparing multiple data sets.

Graphs for Measurements over Time

The following graph describes data from monthly deaths from bronchitis, emphysema and asthma in the UK, during 1974-1979.
The following is part of the data

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1974	3035	2552	2704	2554	2014	1655	1721	1524	1596	2074	2199	2512
1975	2933	2889	2938	2497	1870	1726	1607	1545	1396	1787	2076	2837



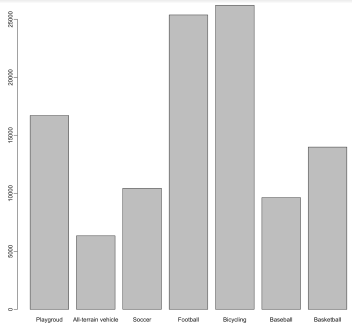
Time-Series Graph

- Reveals information about trends over time.

Graphs for Categorical Data

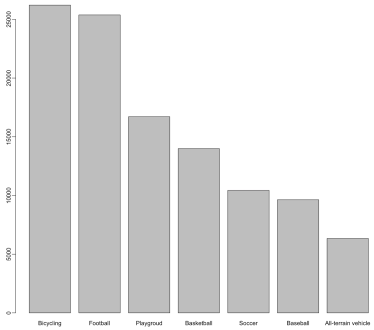
Example

Consider data regarding the highest seven sources of injuries resulting in a visit to a hospital ER from slide 10.



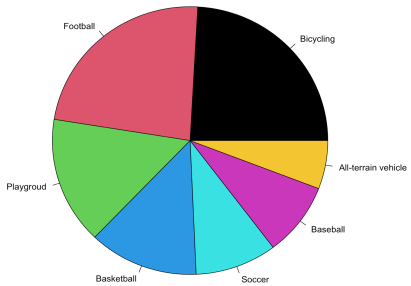
Bar Graph

- Displays distribution of categorical data.



Pareto Graph

- Displays distribution of categorical data.
- Draws attention to more important categories.



Pie Chart

- Displays distribution of categorical data in a commonly used format.

Graphs That Deceive

Example

USA Today survey: respondents were asked if they need caffeine to start up their brain for the day. Among 2,006 respondents, 74% said that they did need the caffeine.

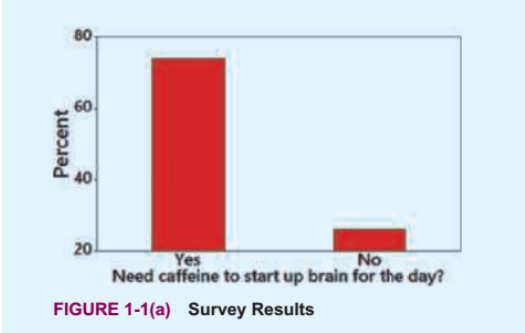
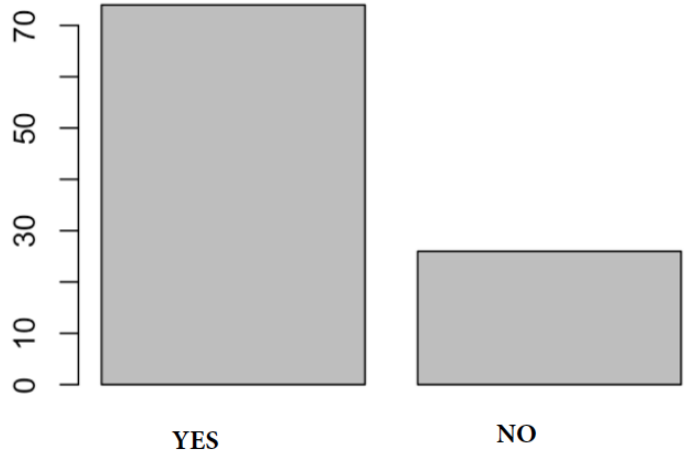
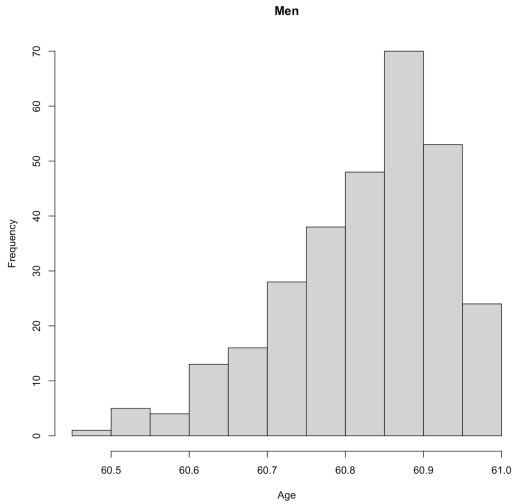
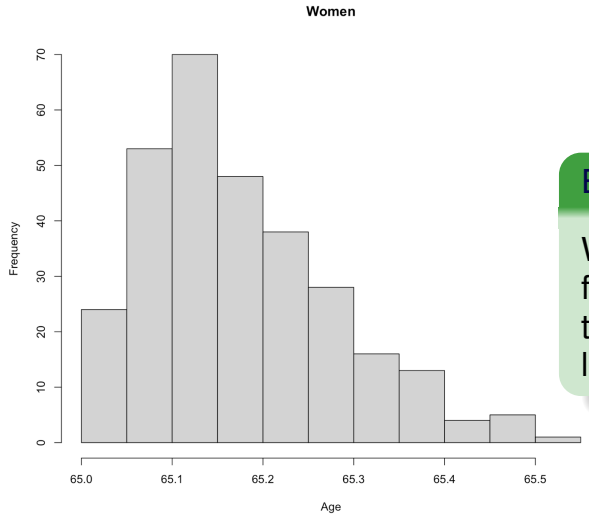


FIGURE 1-1(a) Survey Results







Example

What do you think about the following statement based on the previous graphs: “men retire later that women.”

Practice

Look at the exercises at the end of Section 2-3 in page 63.

Specially, look at exercises:
1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17.