ISyE 3770, Spring 2024
Statistics and Applications

Descriptive Statistics

Instructor: Jie Wang
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

jwang3163@gatech.edu
Office: ISyE Main 445

Descriptive statistics

- An important aspect of dealing with data is organizing and summarizing the data in ways that facilitate its interpretation and subsequent analysis.
- This aspect of statistics is called descriptive statistics. It is a summary statistic that quantitatively describes or summarizes features of a collection of information.
- It is usually the first-attempt of data analytics.
- Sometimes they are sufficient for particular investigations.

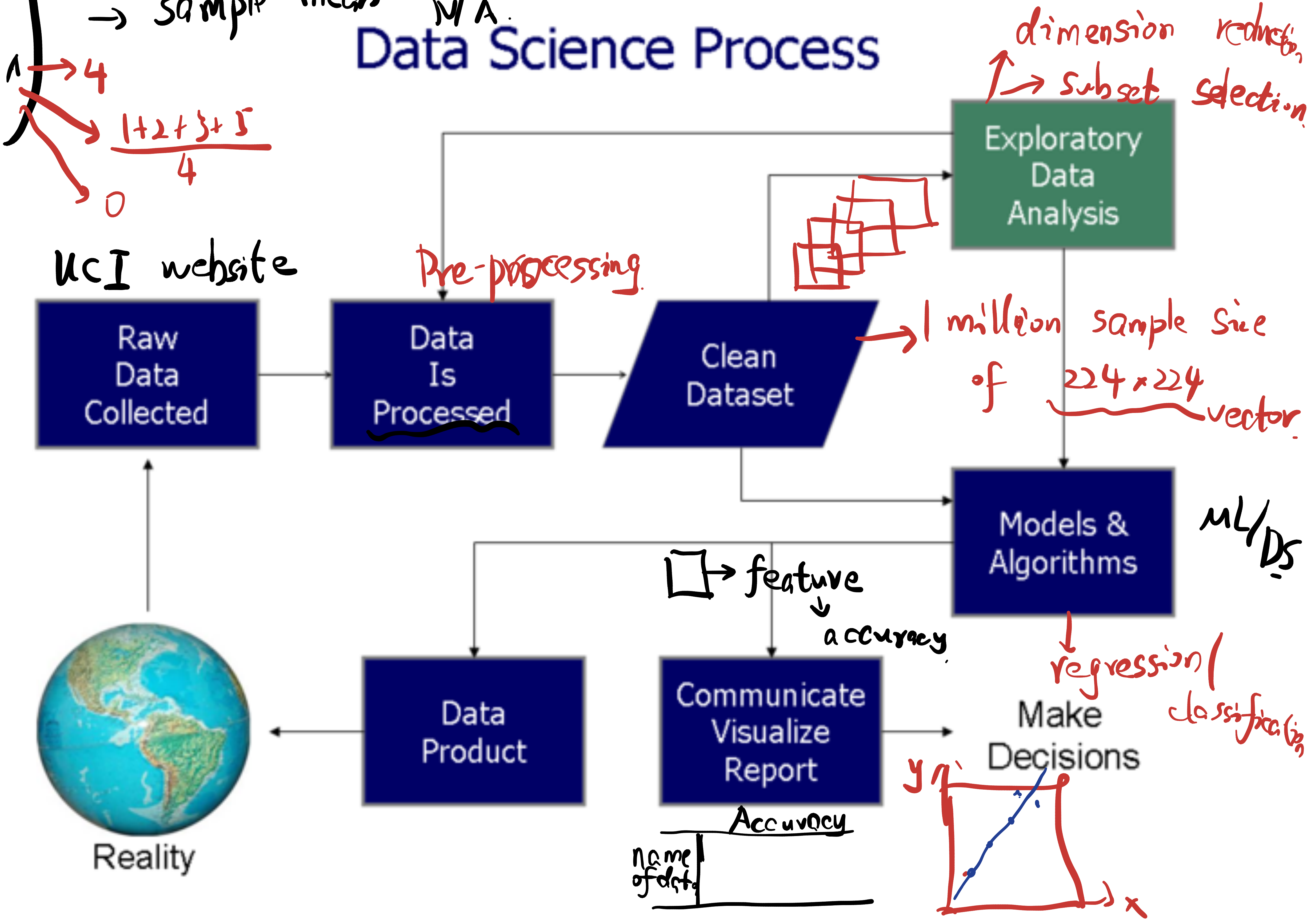
$$\begin{pmatrix}1\\2\\3\\N/A\\5\end{pmatrix} \to \text{sample mean} \quad N/A.$$

N/A → 4

N/A → $\frac{1+2+3+5}{4}$

N/A → 0

Data Science Process



Types of Summary Statistics

Numerical summaries

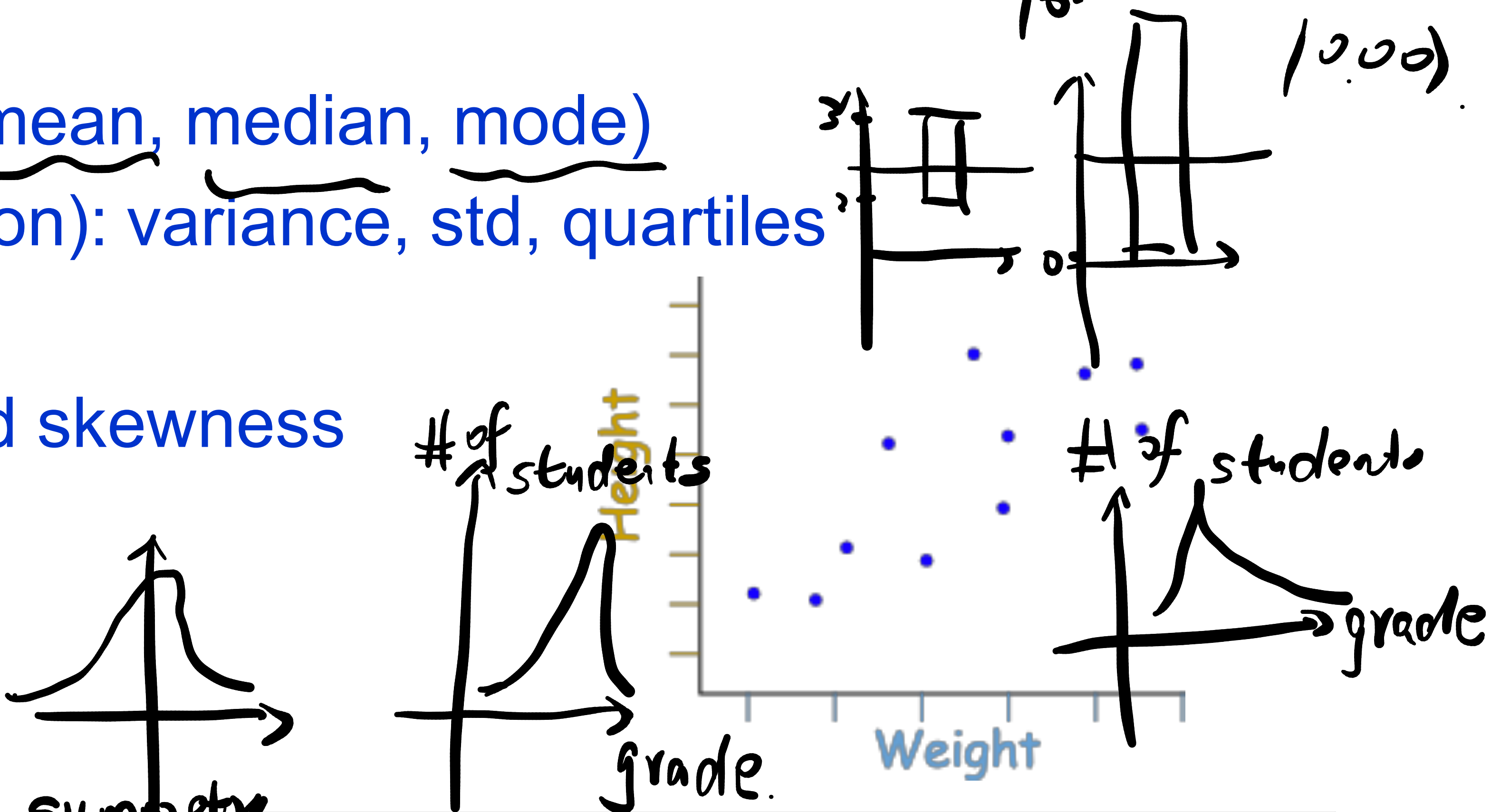
(1, 1, 6, 6, 6, 6, 1000, 1000)

Univariate

- Central tendency (mean, median, mode)
- Variability (dispersion): variance, std, quartiles
- Range: max, min
- Shape: kurtosis and skewness
- Probability plot

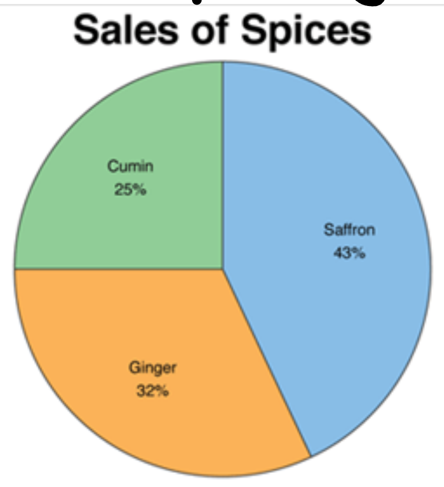
Bivariate

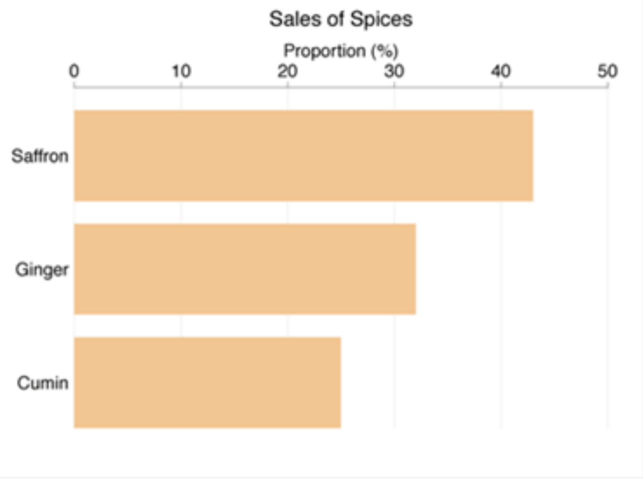
- Scatter plot.



Graphical

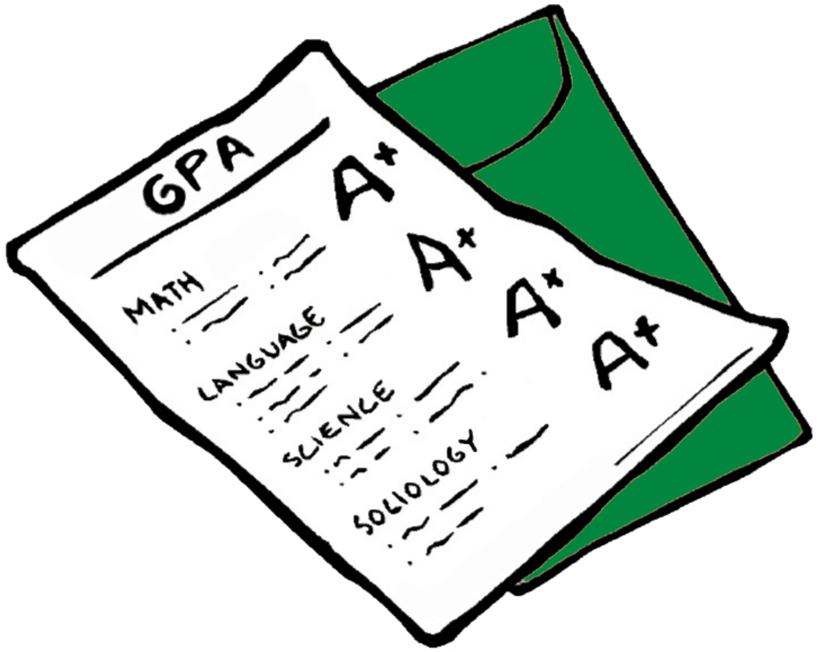
- Pie chart, bar chart
- Histogram, box plot
- Data visualization





Examples

- Shooting percentage in basketball is a descriptive statistic that summarizes the performance of a player or a team.
- GPA: grade point average. This single number describes the general performance of a student across the range of their course experiences



Data type

(20, 10, 2) ← Numerical

{"F", "M", "others"}

Categorical

Categorical or nominal data

Observe frequencies within several categories

Example: frequency of female and male attendance in this class

Numerical Data

Observed values are integer, real or complex numbers

Examples: IQ scores of GT students (integer values)

Lifetime of a computer chip (real values)

Class Activity

1. For each GT student, we record his/her blood type

A. numeric [✓ B. categorical]

2. For each GT student, we record his/her number of siblings.

[✓ A. numeric] B. categorical

3. For each GT student, we record his/her country of residence.

A. numeric [✓ B. categorical]

4. For each GT student, we record his/her height.

[✓ A. numeric] B. categorical

Sample Mean

If the n observations in a sample are denoted by $x_1, x_2, \dots, x_n$, the **sample mean** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (6\text{-}1)$$

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$. The sample mean is

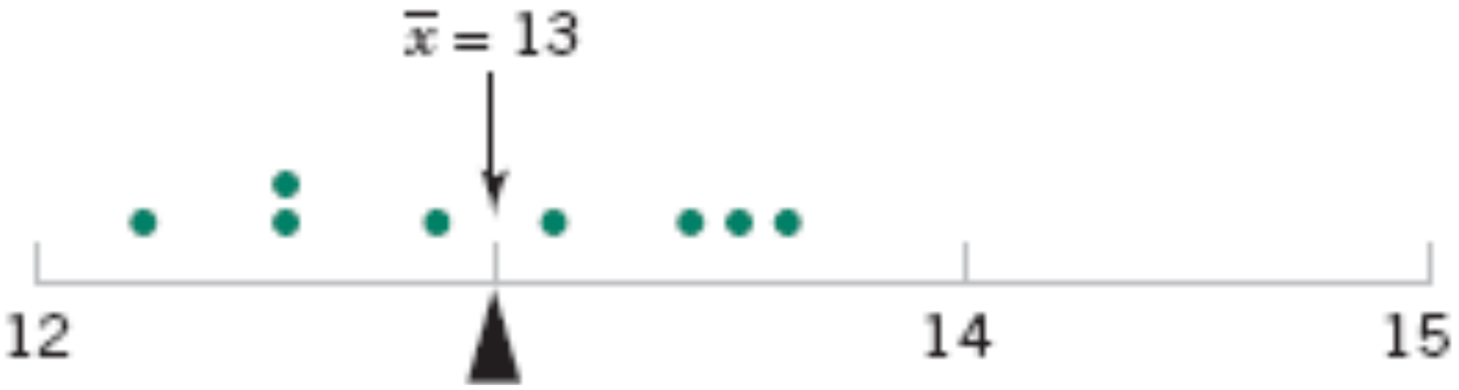
$$(1, 1, 1, 1000)$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \cdots + 13.1}{8}$$

$$= \frac{104}{8} = 13.0 \text{ pounds}$$

$$\bar{X} = \frac{1003}{4} \approx 225.$$

$$\gg 1$$



Sample Median

$\tilde{x}$ A value such that 50% of the data are at or above this value

x <- c(1, 2, 4, 6, 5)

x <- sort(x)

How to calculate:

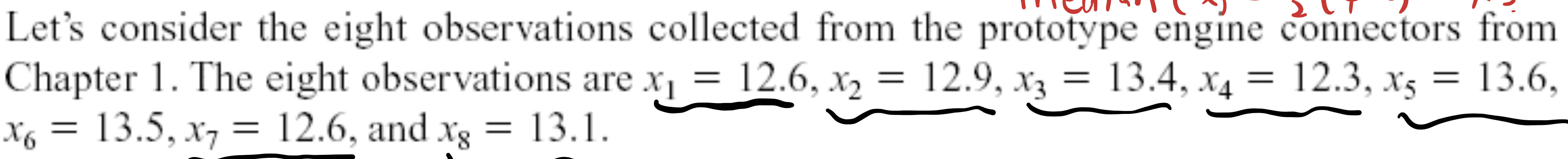
- Sort the data in ascending (or descending) order
- If n is an odd number, median is the $(n+1)/2$th number ;
- If n is an even number, median is the average of is the $n/2$th and $(n/2)+1$th numbers

x = c(1, 2, 4, 5, 6

median(x) = 4.

x = c(1, 2, 4, 5, 6, 7).

median(x) = $\frac{1}{2}(4+5)$ = 4.5



Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

12.3 12.6 12.6 $\boxed{12.9 \quad 13.1}$ 13.4. 13.5 13.6.

↓

median = 13.

Sample Mode

$\hat{x}$ Observation with the highest frequency

observations → (1, 1, 1, 1000)

Mode = 1.

observations ← (1, 2, 3, 4).

Mode is 1 or 2 or 3 or 4.

Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

observation ← (1, 1, 2, 2, 3, 4).

Mode = 12.6.

Mode ← 1 or 2.

Example 1:

Data = {3, 1, 1, 0, 5, 4, 13, 3}

1. Mean = ? 3.75
2. Median = ? 3
3. Mode = ? 1

Example 2:

Data = {3, 1, 1, 0, 5, 4, 3}

1. Mean = ? 2.42
2. Median = ? 3
3. Mode = ? 1

Estimating Data Variability

1. **Sample Range** = **[Largest item] –[smallest item]**

$(1, 5, 0, 4)$. Range = $5 - 0 = 5$.

Easy to calculate, but often misleading (due to outliers).

$(1, 5, 0, 4, 100)$ ⇒ 100

2. **Sample Variance** **measures deviation from the mean**

- Compute $\bar{x}$
- $\sum_i (x_i - \bar{x})^2$
- Divide $(n-1)$.

x_1: Sample variance undefined.

$x_1, \dots, x_n$: degree of freedom ⇒ variablity 's n-1

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n - 1}$$

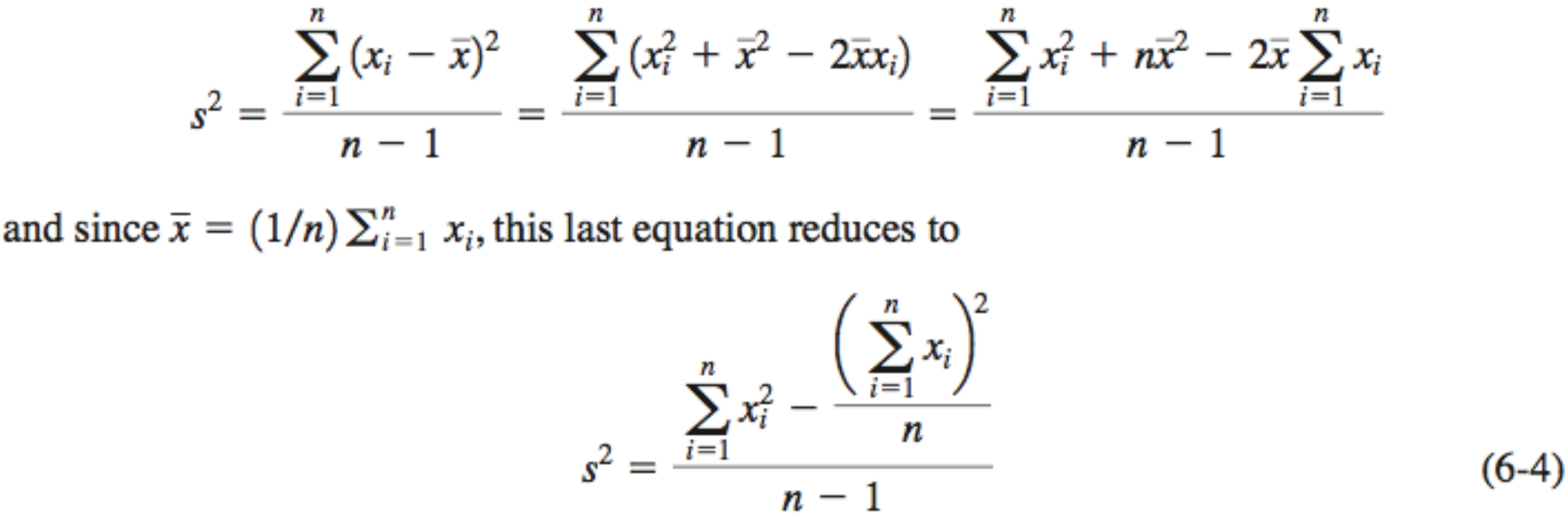
$$E[s^2] = E\left[\frac{\sum_i (x_i - \bar{x})^2}{n-1}\right] = \sigma^2,$$ assuming $X_1, \dots, X_n \sim X$

Pay attention to n-1 !

Standard deviation s

Most commonly used.

Derivation of the variance formula



$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2\bar{x}x_i)}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n}x_i}{n-1}$$

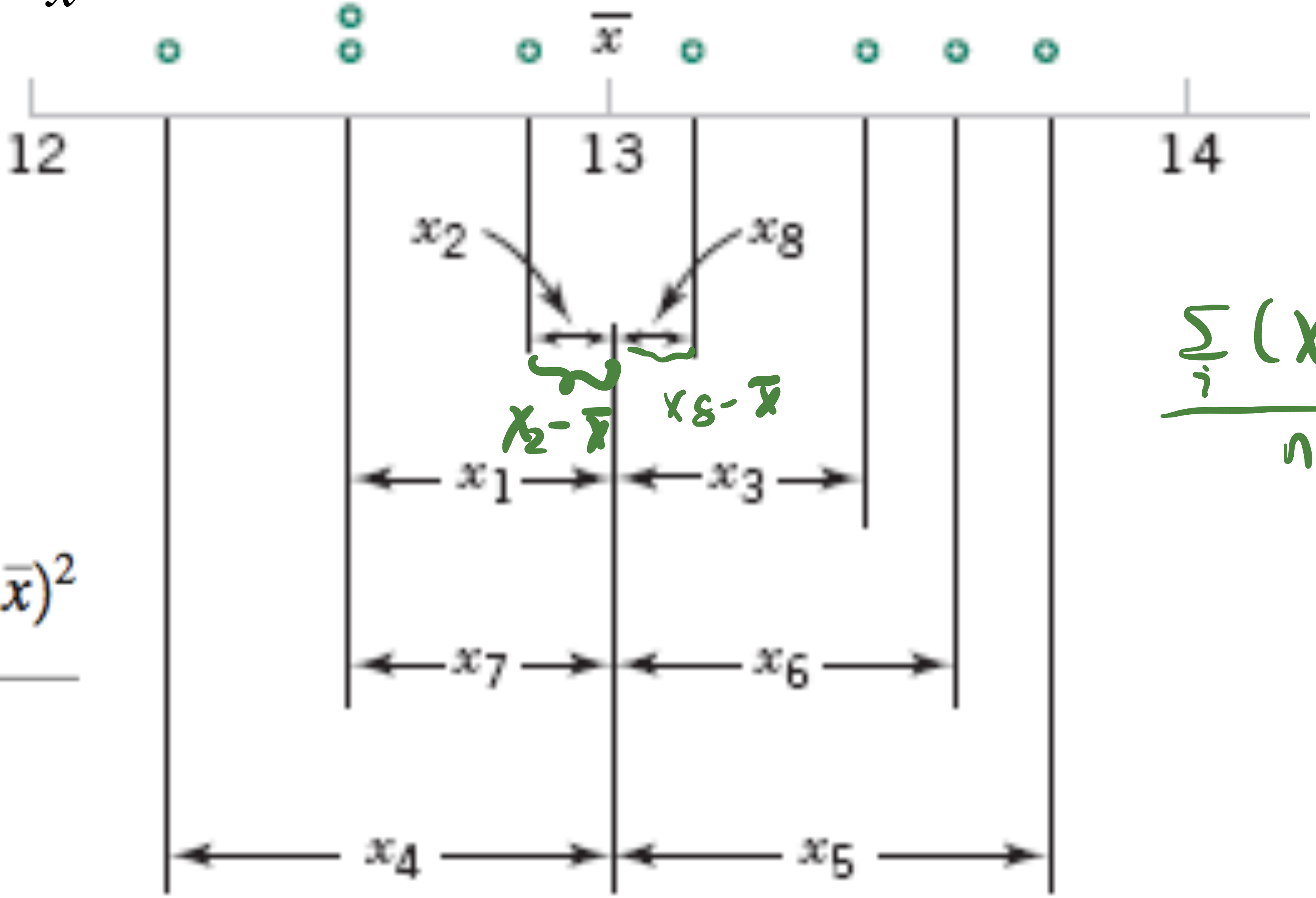
and since $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, this last equation reduces to

$$s^2 = \frac{\sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1} \qquad (6\text{-}4)$$

Sample Variance & Standard Deviation

How the sample variance measures variability through the deviations.

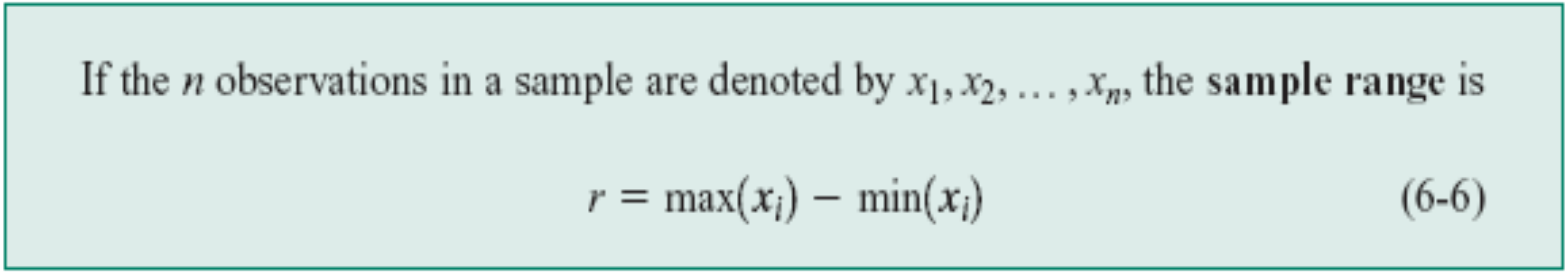
$$x_i - \overline{x}$$



$$\frac{\sum_{i} (x_i - \overline{x})^2}{n-1}$$

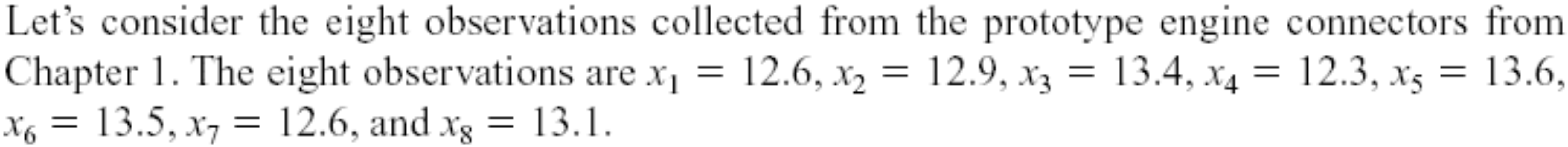
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Sample Range



If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the **sample range** is

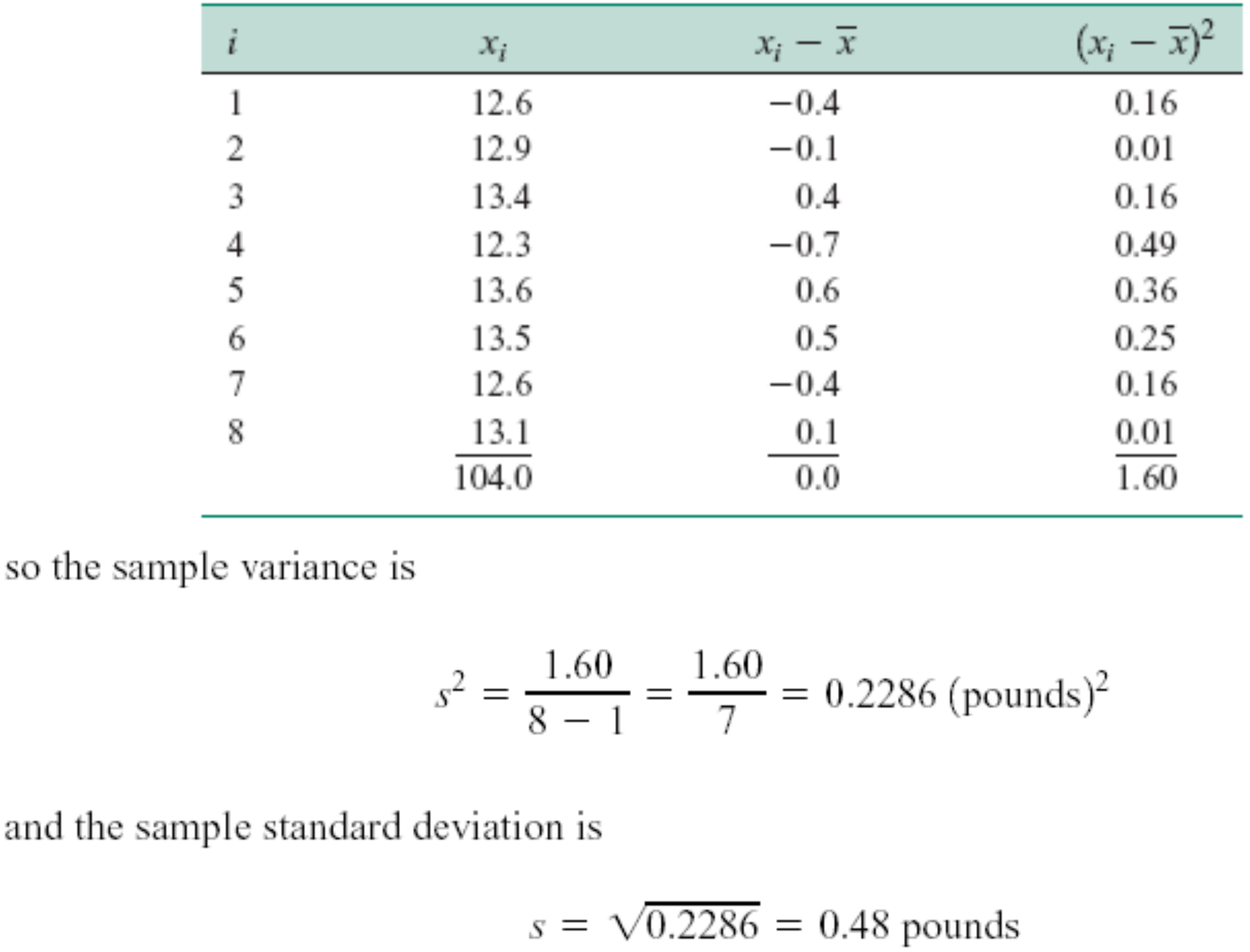
$$r = \max(x_i) - \min(x_i) \qquad (6\text{-}6)$$



Let's consider the eight observations collected from the prototype engine connectors from Chapter 1. The eight observations are $x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

$$r = x_{\max} - x_{\min} = 13.6 - 12.3 = 1.3$$

Example (pull-off force)



i	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	12.6	−0.4	0.16
2	12.9	−0.1	0.01
3	13.4	0.4	0.16
4	12.3	−0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	−0.4	0.16
8	13.1	0.1	0.01
	104.0	0.0	1.60

so the sample variance is

$$s^2 = \frac{1.60}{8 - 1} = \frac{1.60}{7} = 0.2286 \text{ (pounds)}^2$$

and the sample standard deviation is

$$s = \sqrt{0.2286} = 0.48 \text{ pounds}$$

Sample Quantiles

The (100p)-th ~~pth~~ quantile (x_p) is such that 100p% of the sample is smaller than x_p

$p = 0.25$. 25-th quantile is a number s.t. 25% of sample is smaller than it.

Hint : Imagine a sample of 100 ranked items. Think of the quantile as an item's rank.

Quartile refers to quarters:
25th quantile (Q1, lower-quartile, first-quartile),
50th quantile (Q2, second-quartile, median), and
75th percentile (Q3, upper-quartile, third-quartile)

3. Sample inter-quantile range (IQR) = Q3 – Q1

= best measure to capture data variability.

IQR is insensitive to extreme values.

Example:

Data = {3, 1, 1, 0, 5, 4, 13, 3}

1. Sample Range 13.

2. Sample Variance: 16.78

3. Sample IQR = Upper quartile - Lower quartile = 4.25 - 1 = 3.25

Data = {3, 1, 1, 0, 5, 4, 3}

1. Sample Range: 5.

2. Sample Variance: 3.28

3. IQR = 3.5 - 1 = 2.5.

Repeat the calculation above.

Numerical Summary of Data

Statistic: Any function of sampled observations is called a statistic

Central Tendency statistics

Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median $\tilde{x}$

A value such that 50% of the data are at or above this value.

Mode $\hat{x}$

Observation with the highest frequency

Variability statistics

Range $R = x_{\max} - x_{\min}$

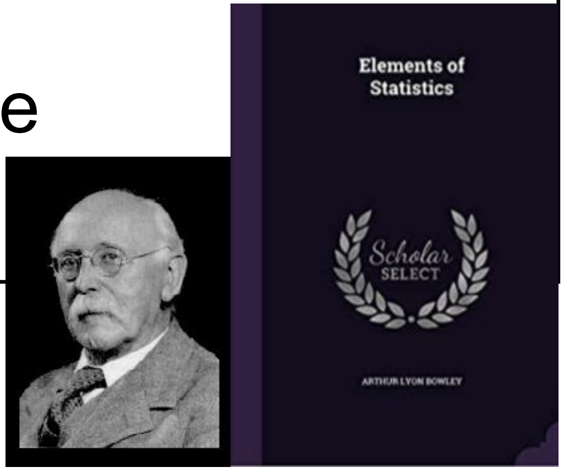
Variance

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard Deviation

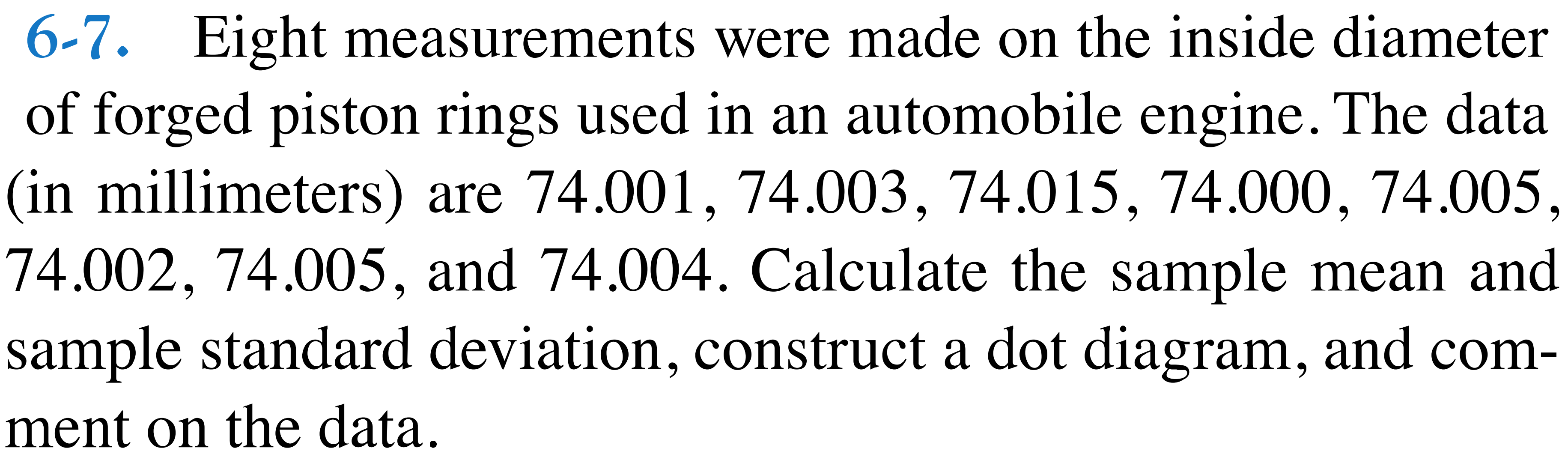
$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

IQR Q_3 - Q_1



Arthur Lyon Bowley

Data Excercise



6-7. Eight measurements were made on the inside diameter of forged piston rings used in an automobile engine. The data (in millimeters) are 74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005, and 74.004. Calculate the sample mean and sample standard deviation, construct a dot diagram, and comment on the data.

Central Tendency statistics

Mean 74.00437

Median 74.0035

Mode 74.005

Variability statistics

Range 0.015

Variance 2.169643e-05

Standard Deviation 0.004657943

IQR

74.005-74.00175 = 0.00325

Covariance between two variables

- **Sample covariance** : a sample estimate of covariance.
- $\frac{1}{n}(\sum_{i=1}^{n} x_i y_i) - \frac{1}{n^2}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) = \underbrace{\boxed{\frac{\sum_i x_i y_i}{n}}}_{\to E[XY]} - \underbrace{\boxed{\left(\frac{\sum_i x_i}{n}\right)\left(\frac{\sum_i y_i}{n}\right)}}_{E[X]\cdot E[Y]}$
- **Estimation of $cov(X, Y) = E(XY) - E(X)E(Y)$**

- each observation is a vector of dimension 2.

  (x_i, y_i), $i = 1, \dots, n$.

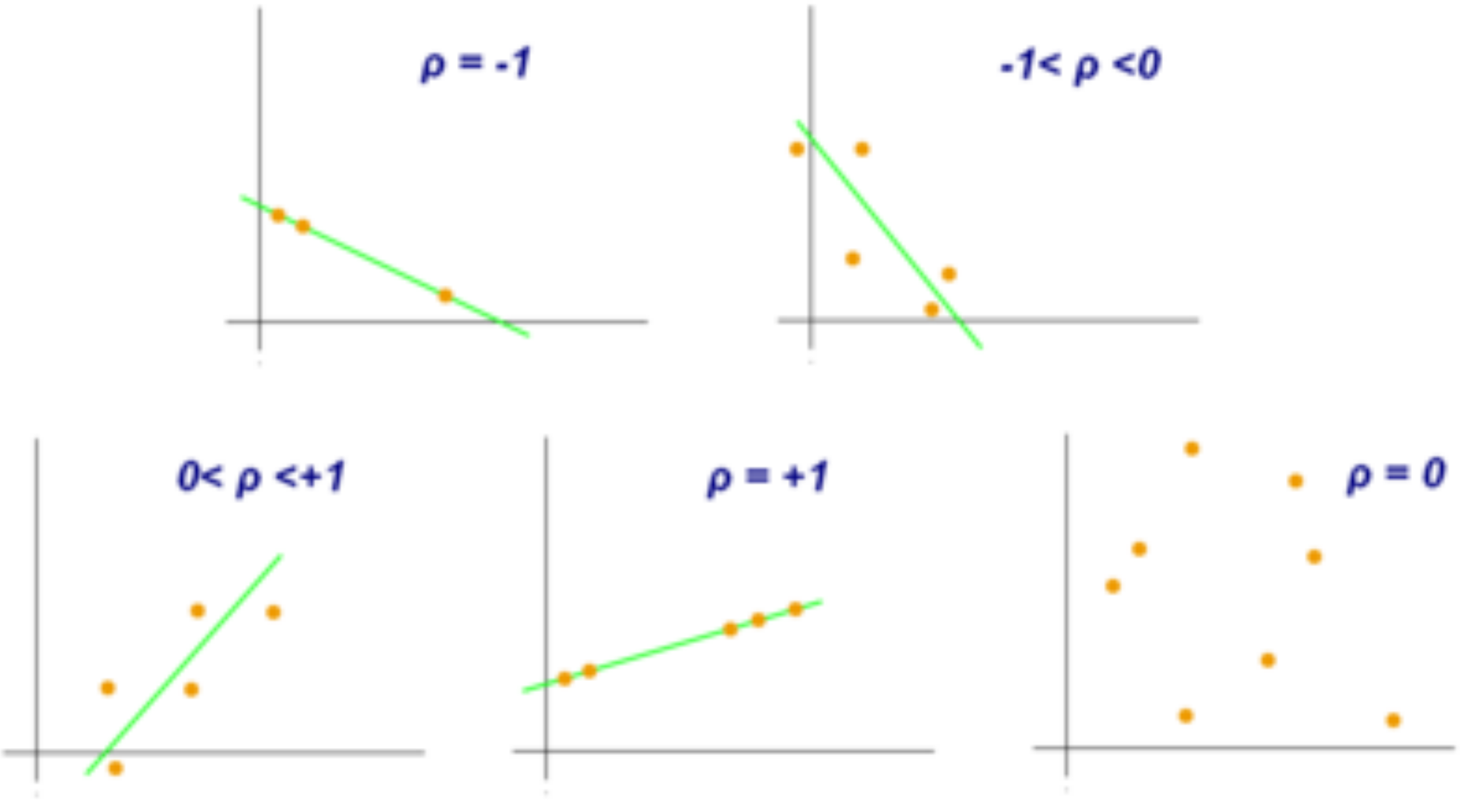
Pearson correlation coefficient

- **Describing the linear correlation between two variables**

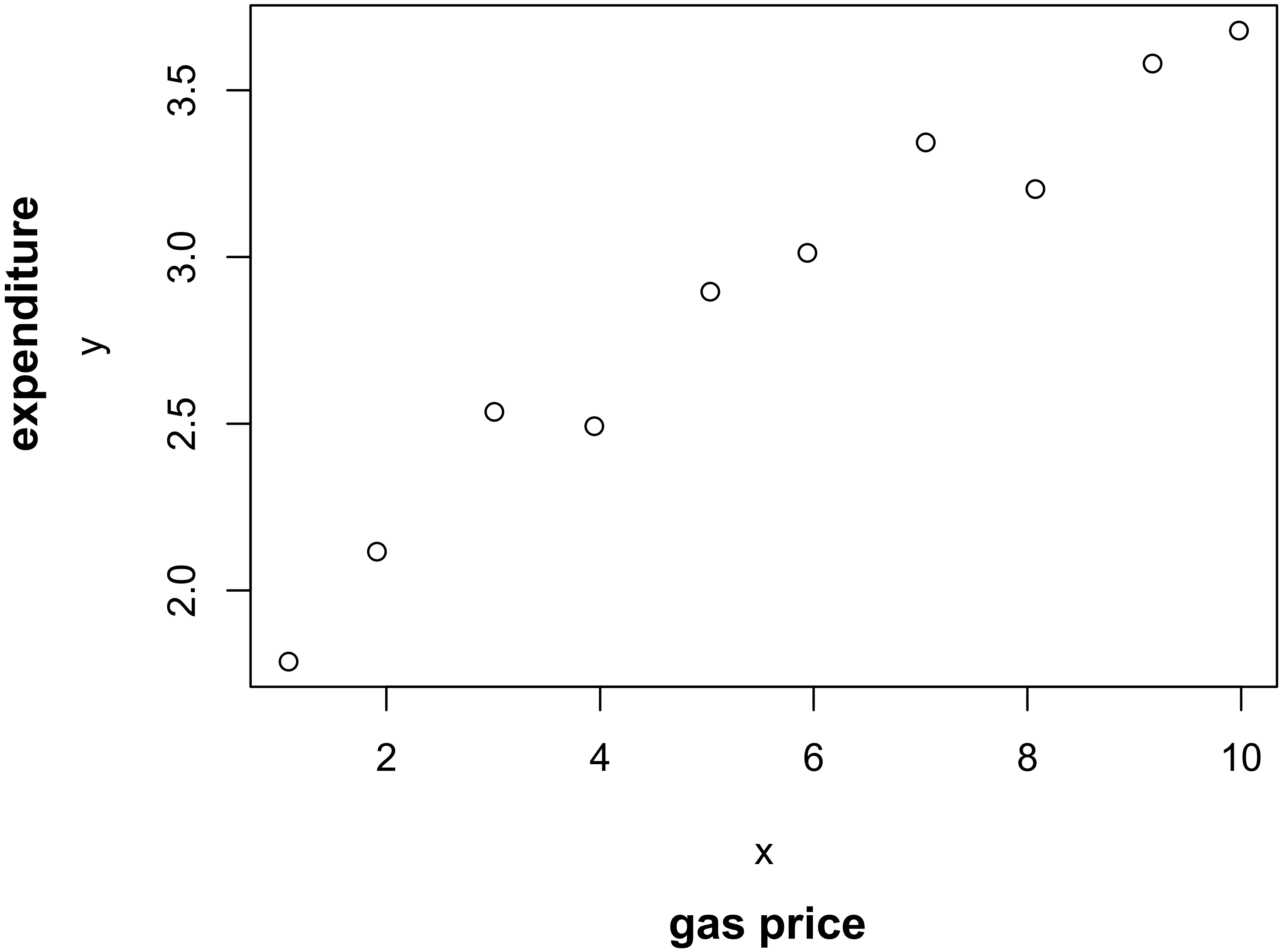
Cor(x,y)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$\in [-1, 1]$



Example

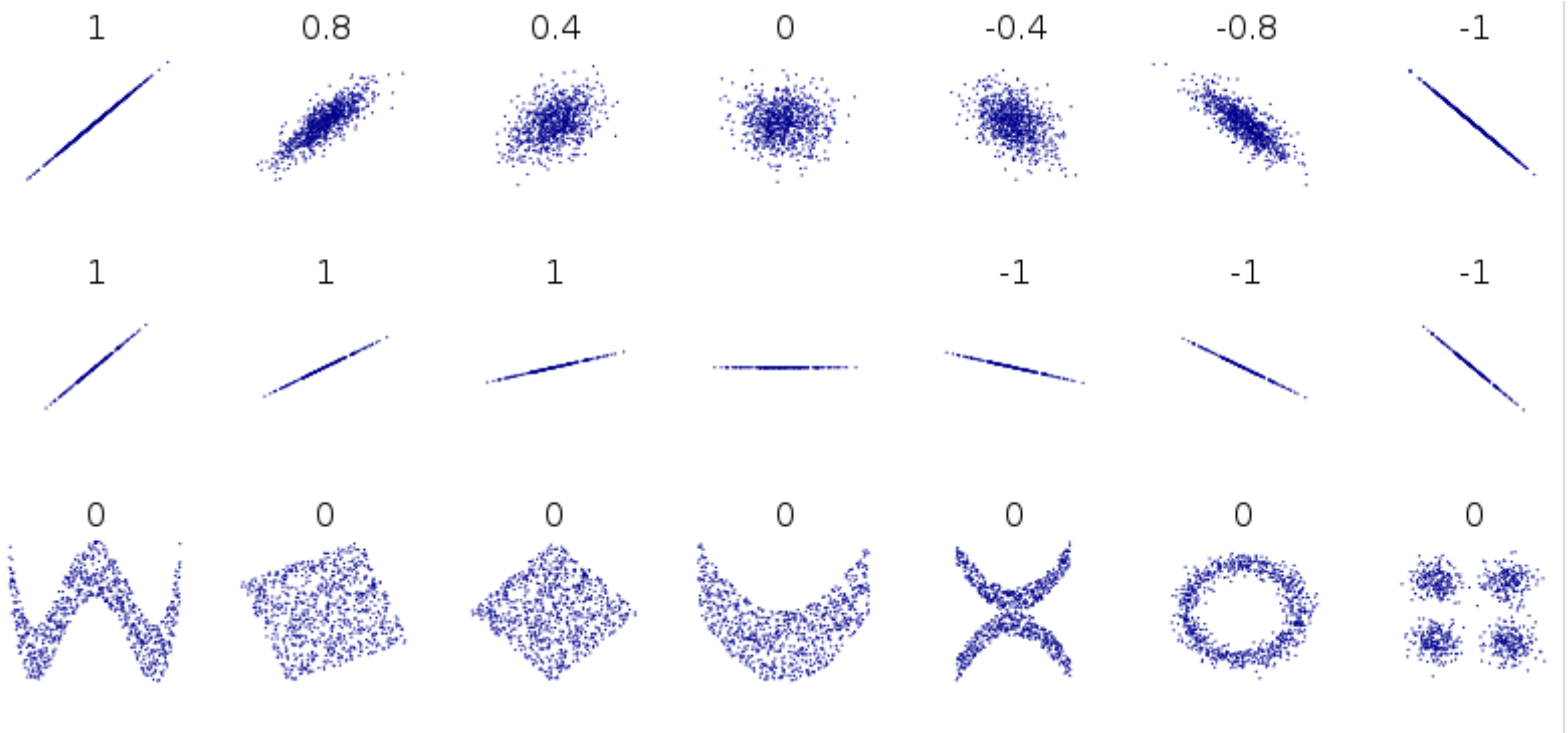


- **Data**

1.0840	1.7862
1.9112	2.1160
3.0100	2.5355
3.9455	2.4928
5.0304	2.8961
5.9400	3.0124
7.0490	3.3437
8.0739	3.2039
9.1712	3.5802
9.9806	3.6792

- **cov(x,y) = 1.865434**
- **cor(x,y) = 0.9774152**

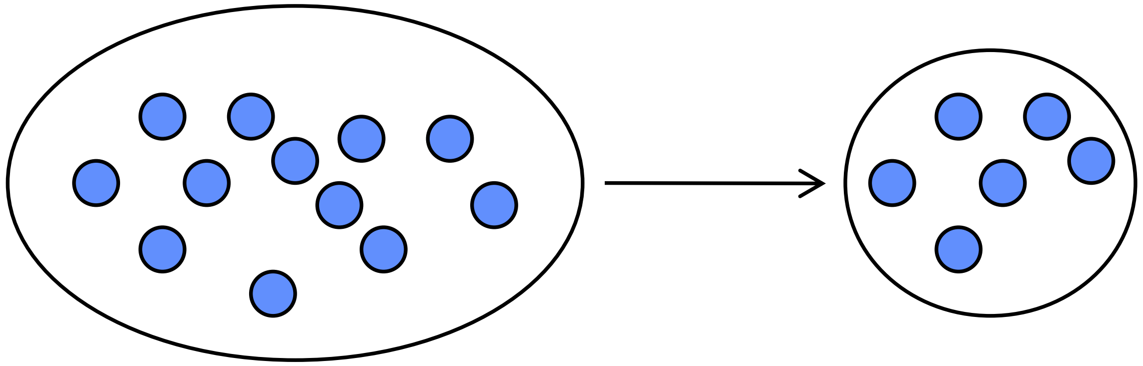
What correlation captures



Note that the correlation reflects the non-linearity and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom).

Population Mean

N: population size



n: sample size

Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

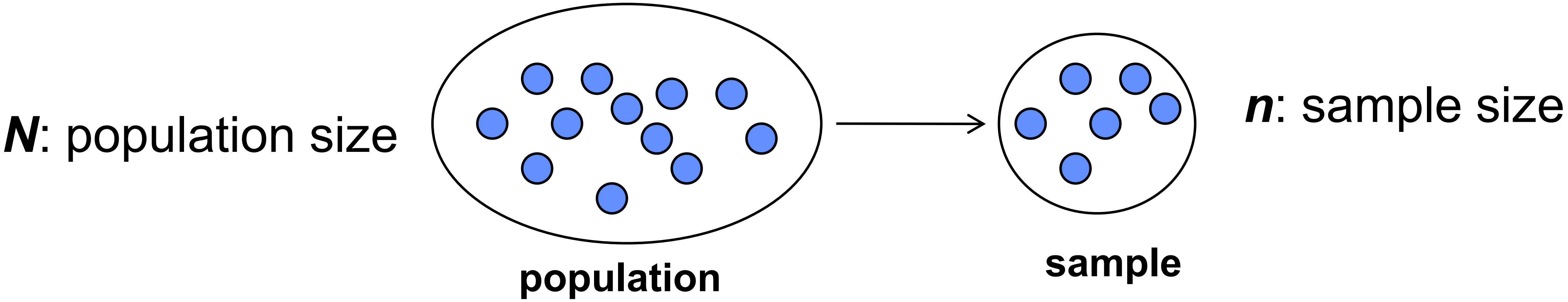
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample mean

The sample mean is a reasonable estimate of the population mean.

(More accurate when the sample size increases)

Population Variance



Population variance $$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ **Sample variance**

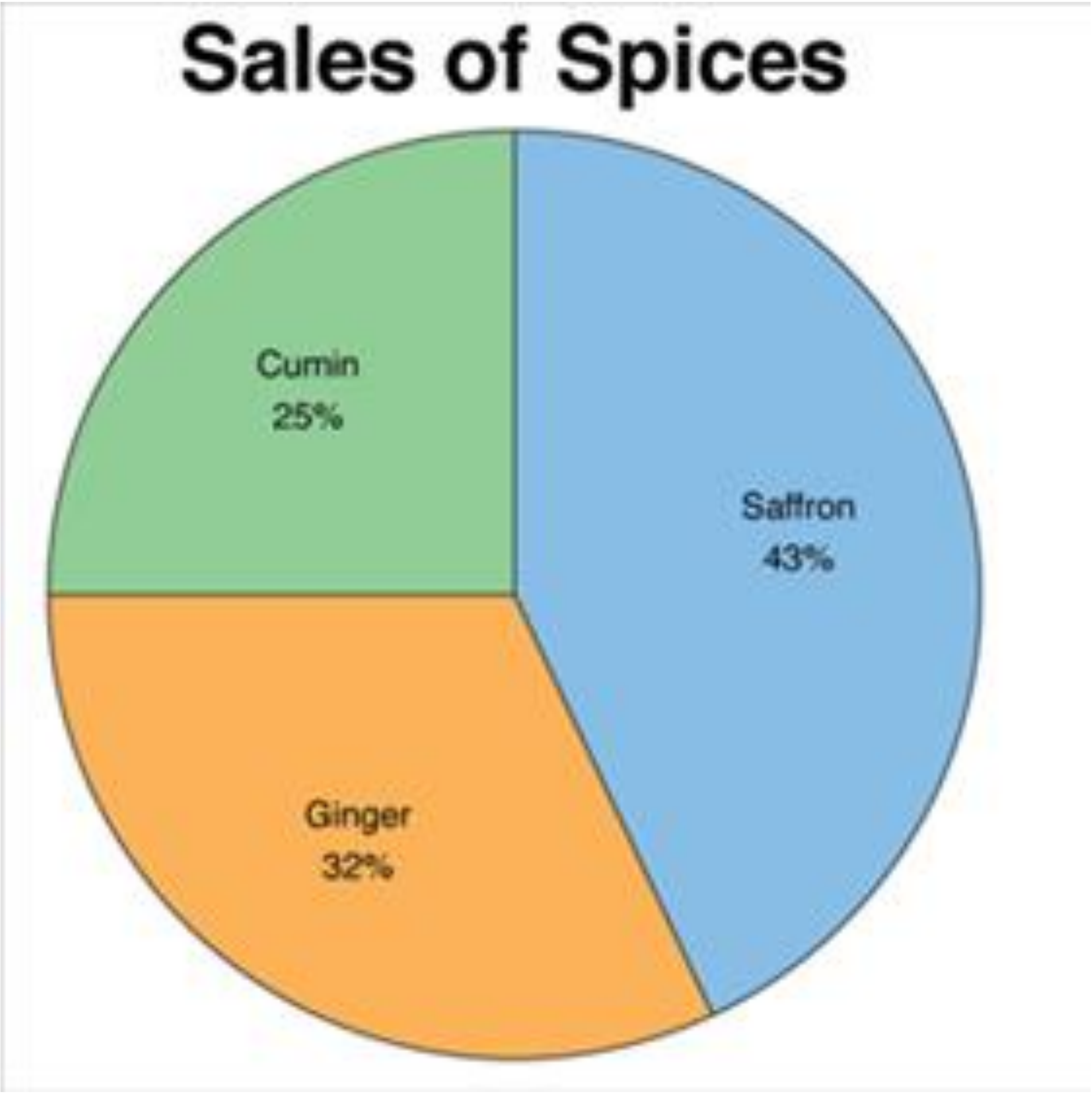
The sample variance is a reasonable estimate of the population variance.

(More accurate when the sample size increases)

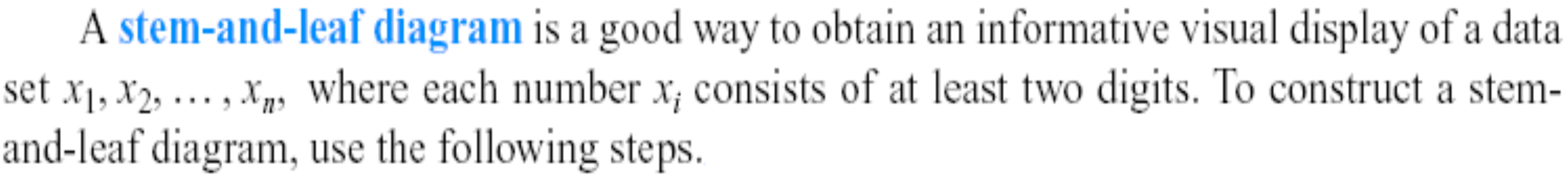
Graphical Descriptive Statistics

- Pie chart
- Stem-and-leaf diagram
- Frequency Table and Histogram
- Box plot
- Time-series plot

Example

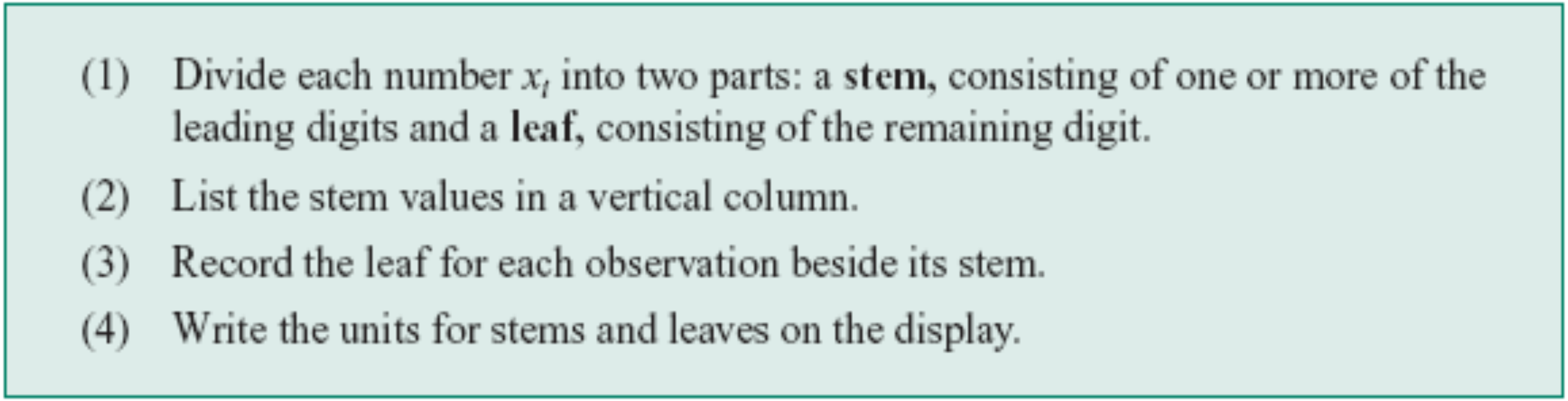


Stem-and-Leaf Diagrams



A **stem-and-leaf diagram** is a good way to obtain an informative visual display of a data set $x_1, x_2, \ldots, x_n$, where each number x_i consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

Steps for Constructing a Stem-and-Leaf Diagram



(1) Divide each number x_i into two parts: a **stem,** consisting of one or more of the leading digits and a **leaf,** consisting of the remaining digit.

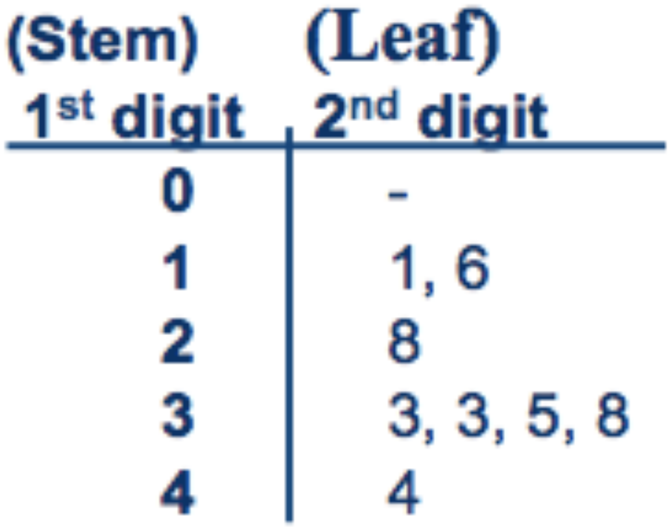
(2) List the stem values in a vertical column.

(3) Record the leaf for each observation beside its stem.

(4) Write the units for stems and leaves on the display.

Stem and Leaf Plots

Data = {33, 28, 16, 35, 11, 44, 33, 38}



(Stem) 1st digit	(Leaf) 2nd digit
0	-
1	1, 6
2	8
3	3, 3, 5, 8
4	4

Stem-and-Leaf Diagrams

Example 6-4

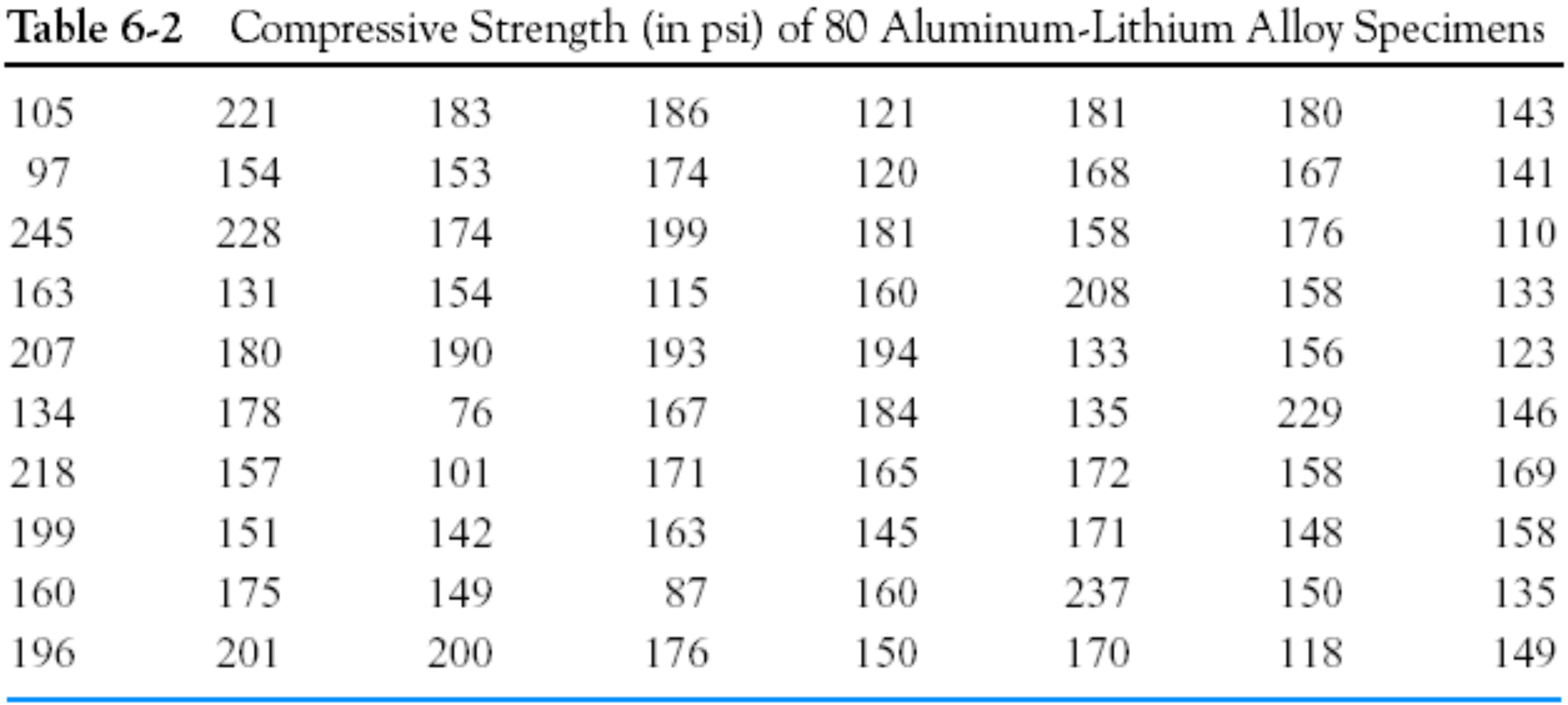
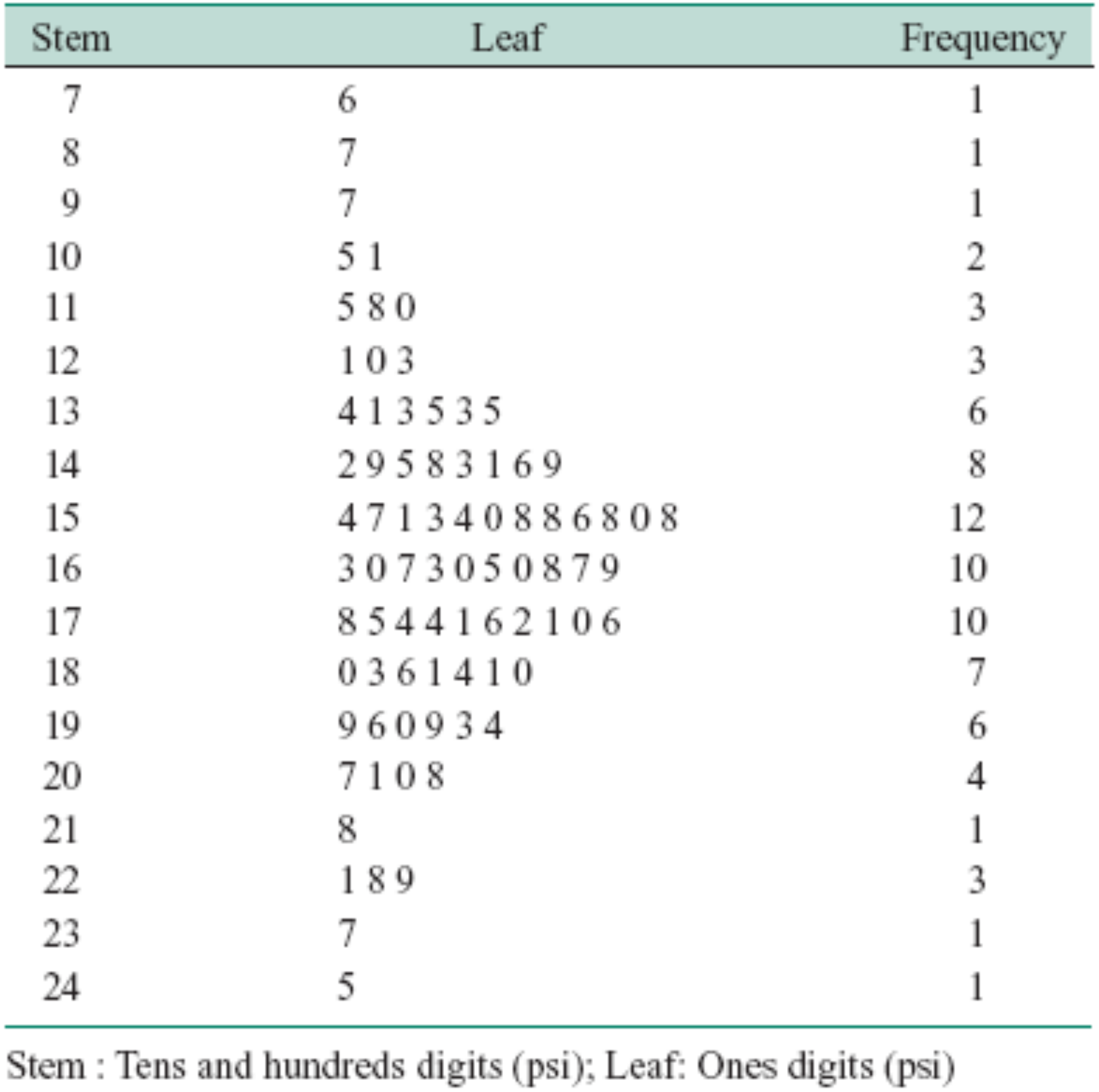


Table 6-2 Compressive Strength (in psi) of 80 Aluminum-Lithium Alloy Specimens

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Stem-and-Leaf Diagrams

Figure 6-4 Stem-and-leaf diagram for the compressive strength data in Table 6-2.



Stem	Leaf	Frequency
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Stem : Tens and hundreds digits (psi); Leaf: Ones digits (psi)

How to display data? Data Features

Shape of the data distribution: symmetric, skewed to the right or to the left.

Spread of the data: range, long or short tails

Outliers: extreme values that appear separate from the rest of the data

Modes: concentrations of the data – unimodal, bimodal, multimodal

Gaps: different subpopulations

Comparison of two or more datasets

Histograms

Divide observations into groups to construct frequency histogram. Often, this is the BEST way to communicate your findings from a set of data.

We can (generally) rely on computers to generate histograms

- **Keep bin widths equal**
- **Choose width that summarizes the data best**
- **Use Excel, R, Minitab, SAS, MATLAB, Python, etc**

Karl Pearson

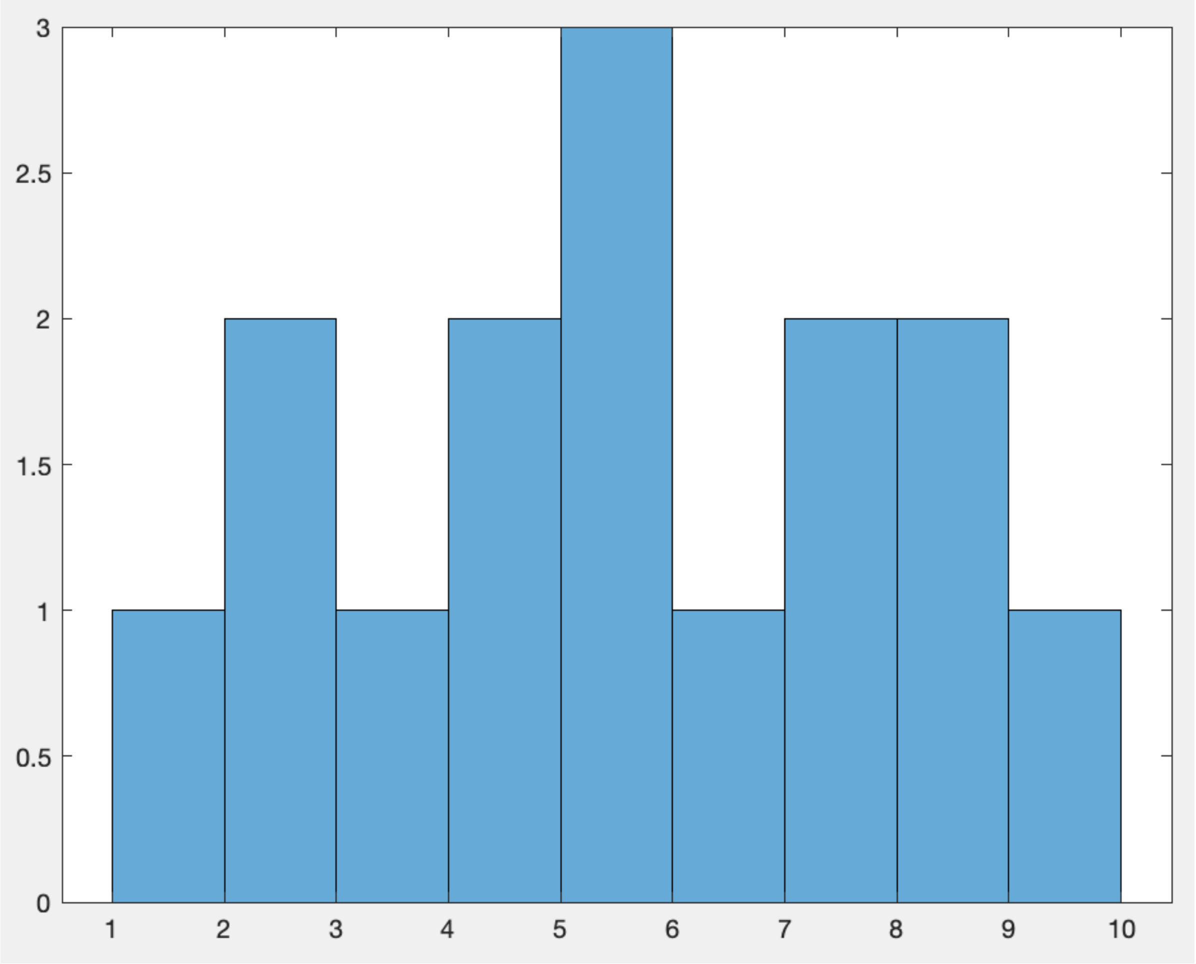
Frequency Table and Histogram

- To construct a frequency table
  1. Find the range of the data
     - start the lower limit for the first bin just slightly below the smallest data value
     - b_0=<min(x), b_m=max(x),
     - R = b_m - b_0
  2. Divide this range into a suitable number of equal intervals
     - m=4 ~ 20, or ?$\sqrt{n}$ (n is the total number of samples)
  3. Count the frequency of each interval
     - if $b_{i-1} <= x < b_i$

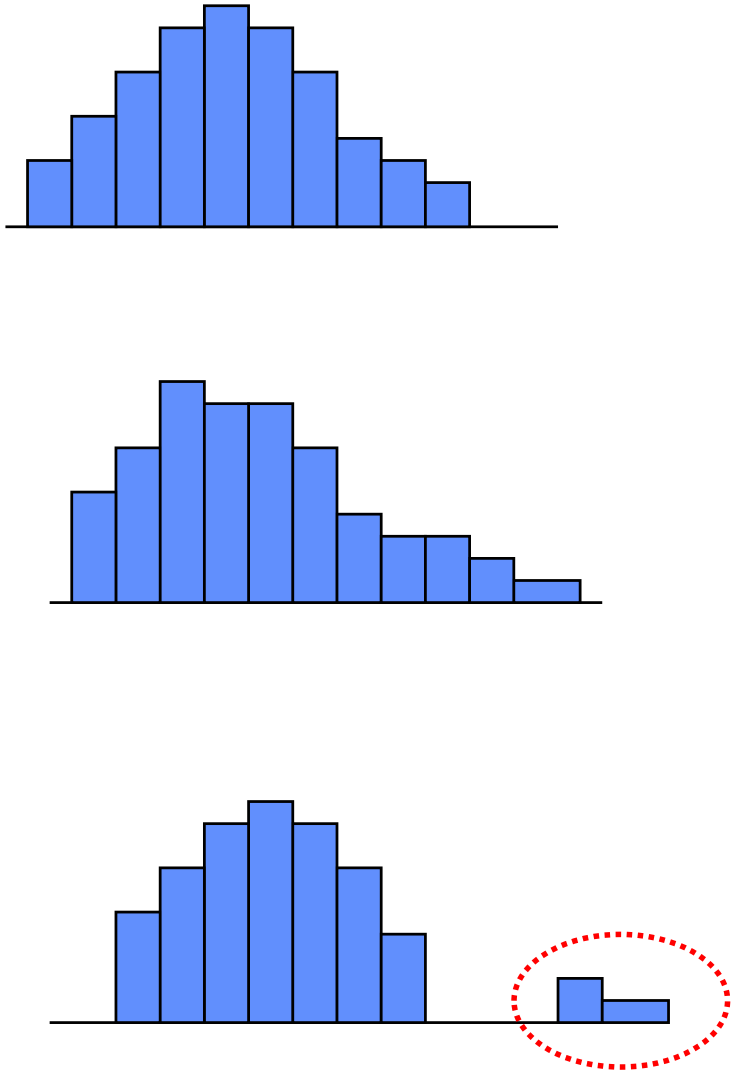
Example

- **Data:**

9, 5, 1, 4, 4, 7, 2, 5, 3, 8, 7, 6, 5, 8, 2



Histogram features



Shape of the data distribution: symmetric, skewed to the right or to the left.

Spread of the data: range, long or short tails

Outliers: extreme values that appear separate from the rest of the data

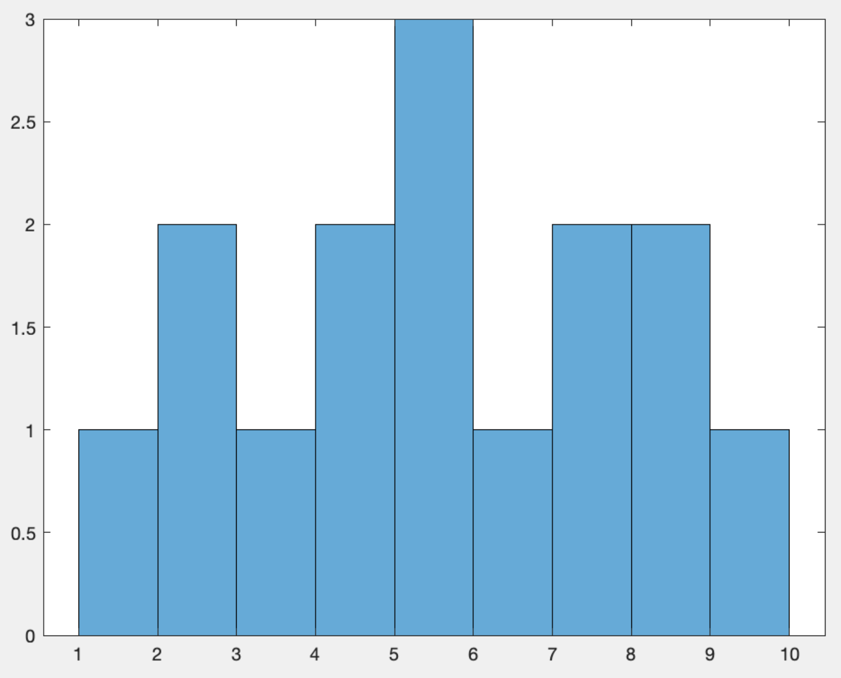
Modes: concentrations of the data – unimodal, bimodal

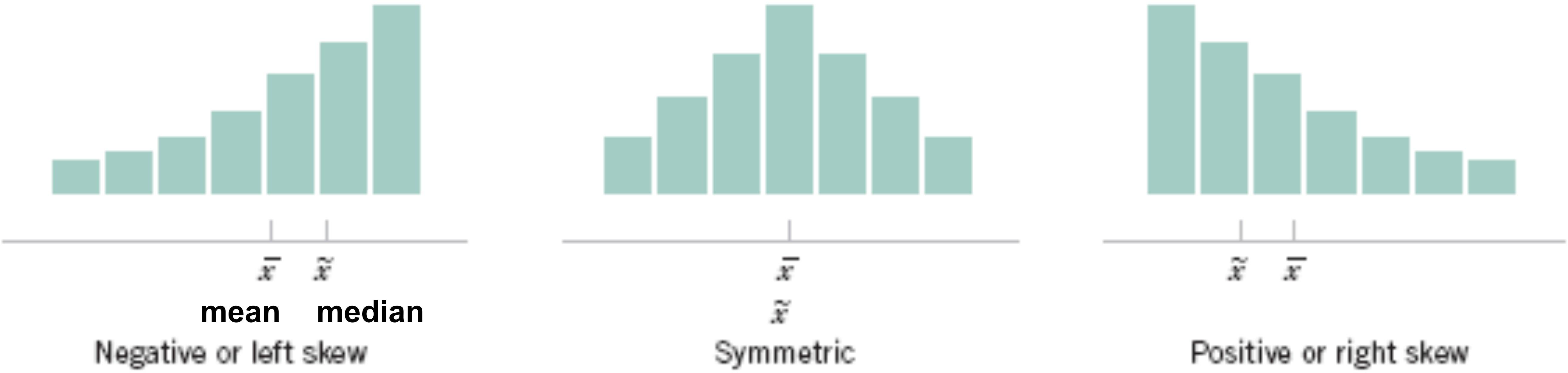
Gaps: different subpopulations

Interpretation based on Histogram

Three Properties of Sample Data

- **Shape:**
  - **roughly symmetric and unimodal**
- **The center tendency or location**
  - **the points tend to cluster near 5.**
- **Scatter or spread range**
  - **variability is relatively high** (min=1; max=9)





Relationship between population and samples

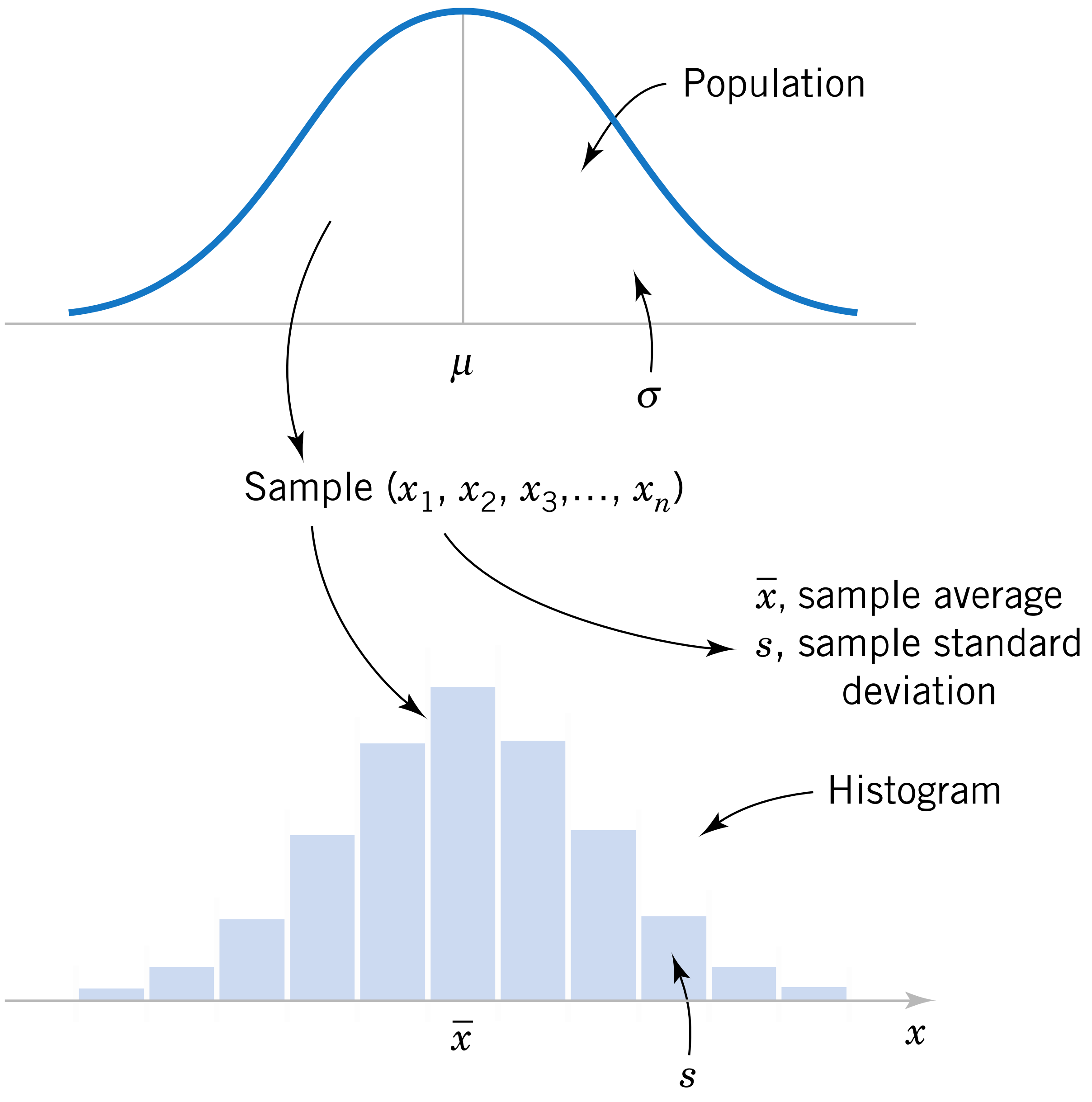
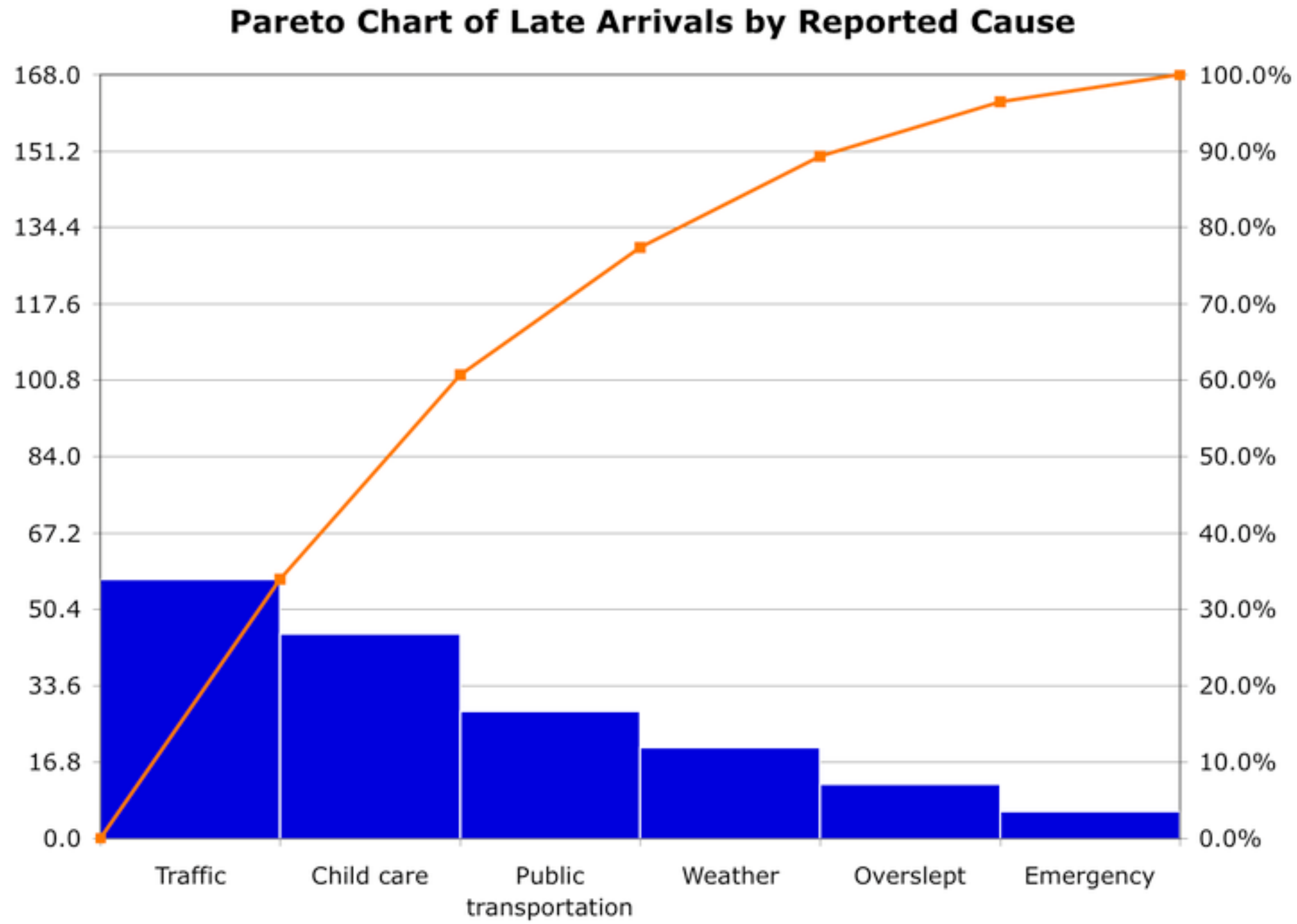

Population
μ
σ
Sample ($x_1, x_2, x_3, \ldots, x_n$)
$\overline{x}$, sample average
s, sample standard deviation
Histogram
$\overline{x}$
s
x

Figure 6-3 Relationship between a population and a sample.

Pareto chart

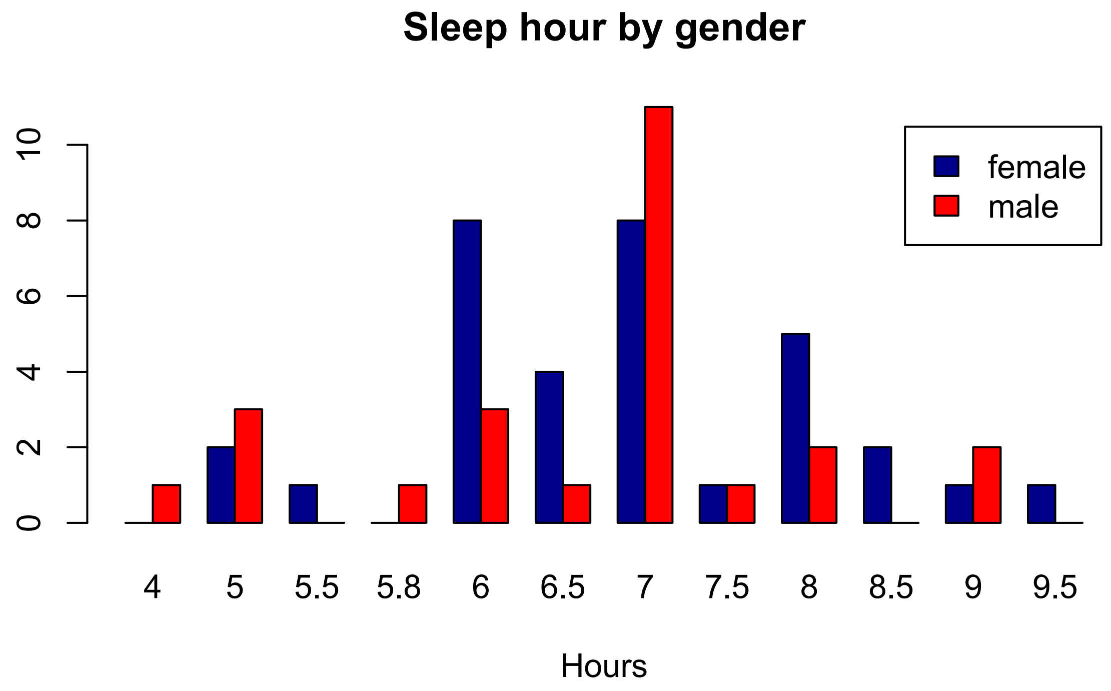
Highlight the most important among a (typically large) set of factors.

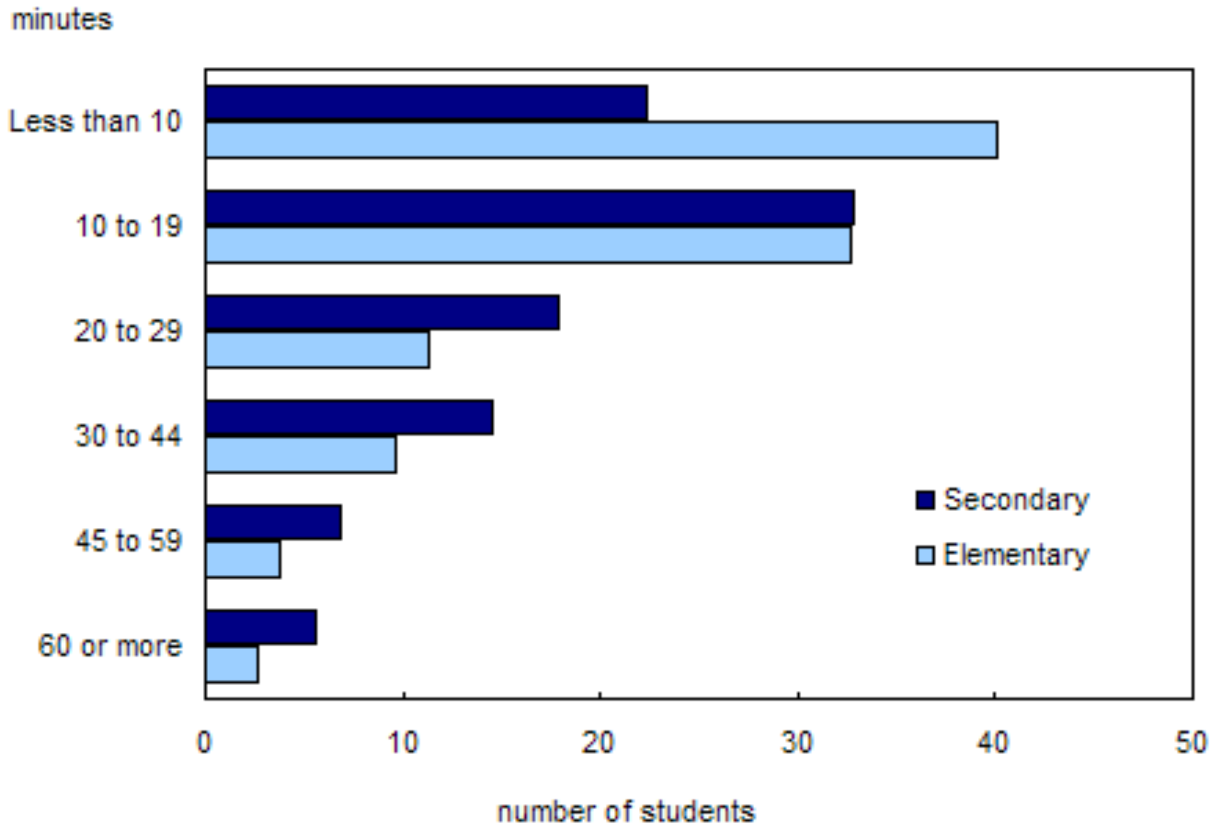


Simple example of a Pareto chart using hypothetical data showing the relative frequency of reasons for arriving late at work

Line Graphs: bar chart

Average time getting to school





Box Plots

- The box plot is a graphical display that simultaneously describes several important features of a data set, such as center, spread, departure from symmetry, and identification of observations that lie unusually far from the bulk of the data (outliers).

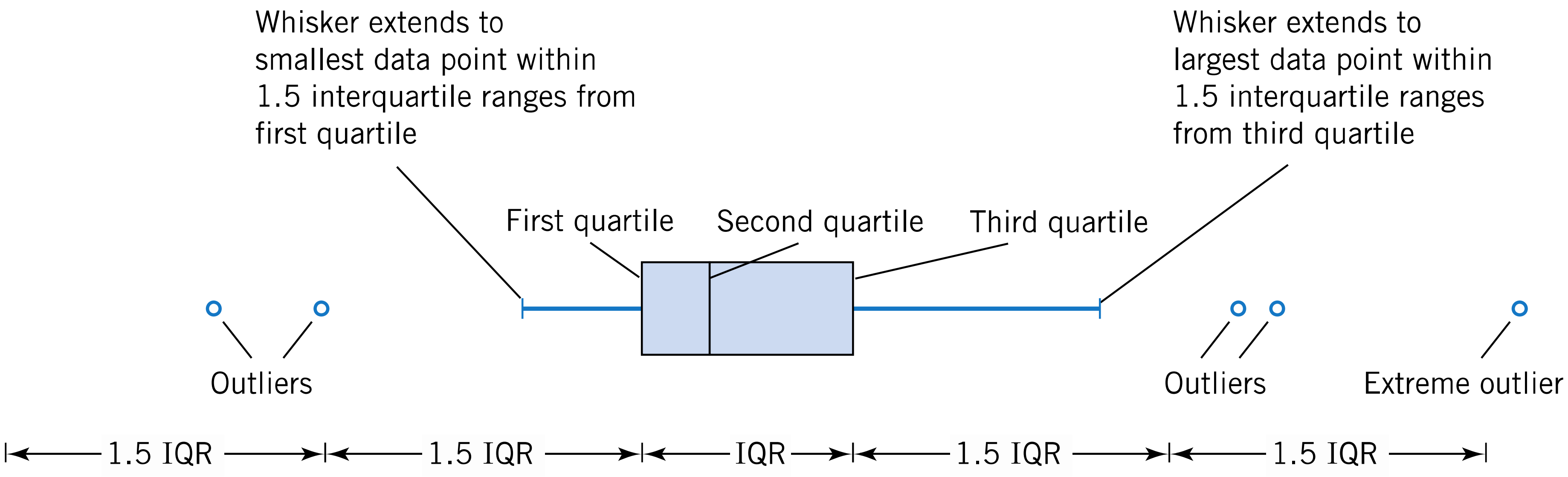
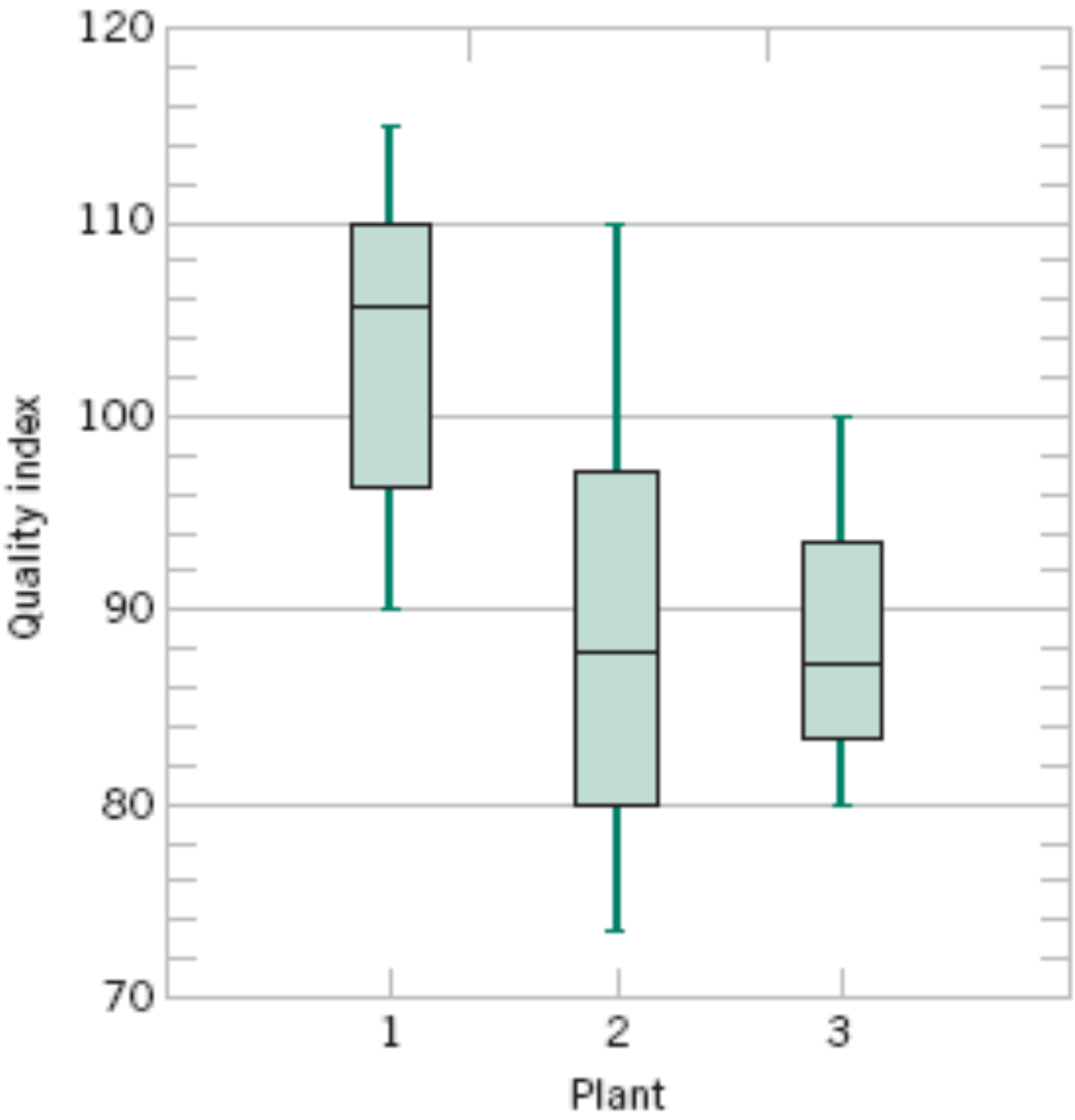

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

Figure 6-13 Description of a box plot.

Comparison using box plot

Box plots are useful in graphical comparison of datasets

Comparative box plots of a quality index at three plants.



Time Series Plot

- A time series or time sequence is a data set in which the observations are recorded in the order in which they occur.
- A time series plot is a graph in which the vertical axis denotes the observed value of the variable (say x) and the horizontal axis denotes the time (which could be minutes, days, years, etc.).
- When measurements are plotted as a time series, we often see patterns like trends, cycles, or other broad features of the data

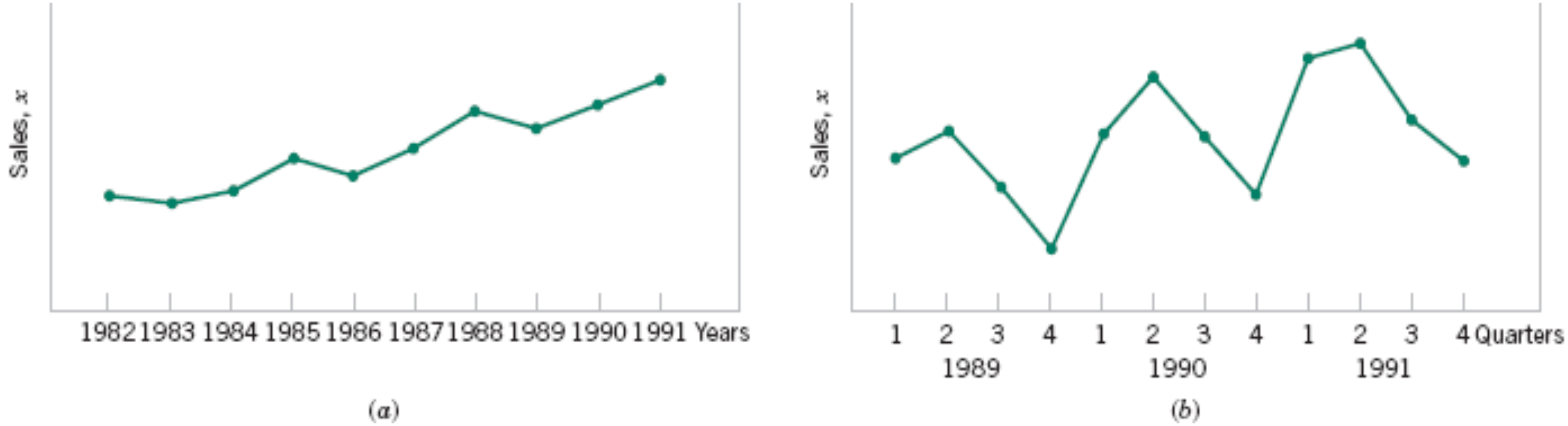
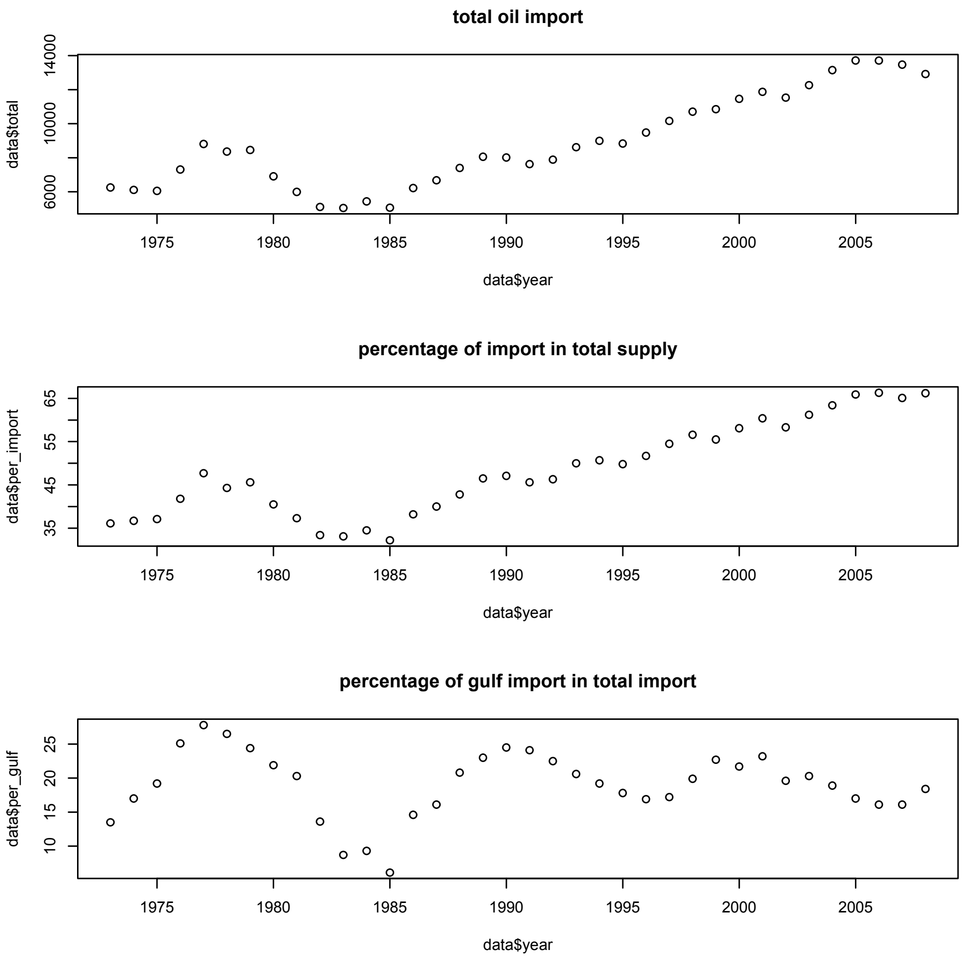


Figure 6-16 Company sales by year (a) and by quarter (b).

R example

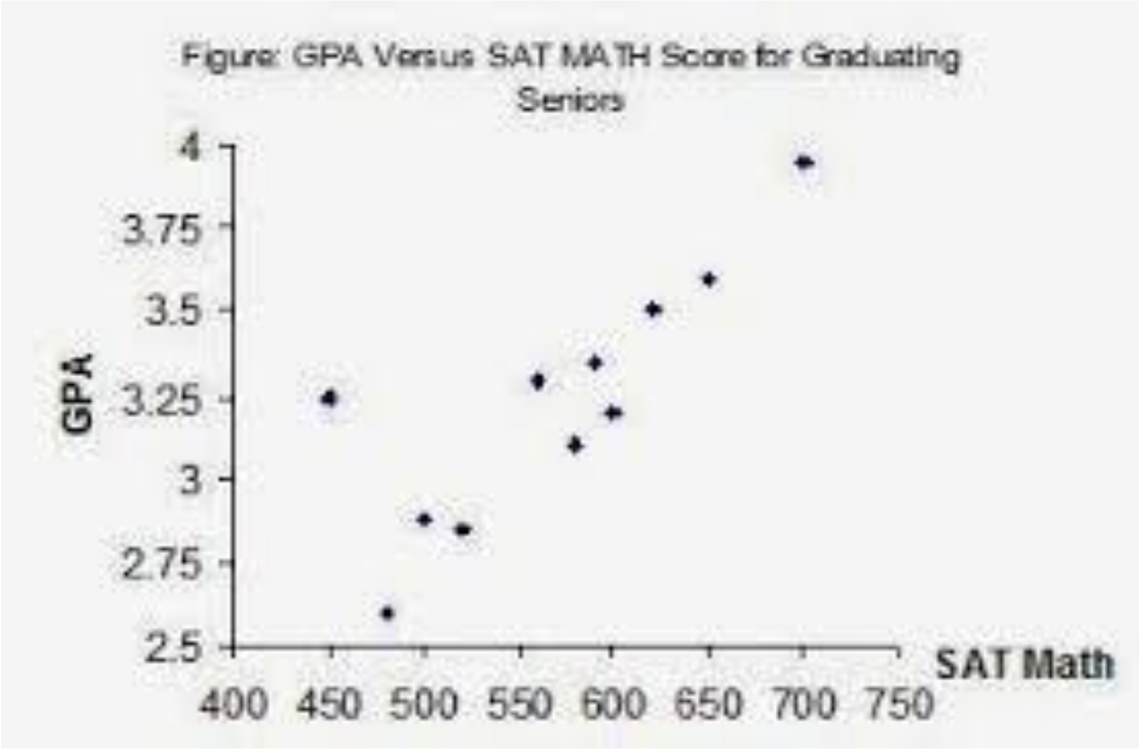
6-72. The following table shows U.S. petroleum imports, imports as a percentage of total, and Persian Gulf imports as a percentage of all imports by year since 1973 (source: U.S. Department of Energy Web site, http://www.eia.doe.gov/). Construct and interpret either a digidot plot or a separate stem-and-leaf and time series plot for each column of data.

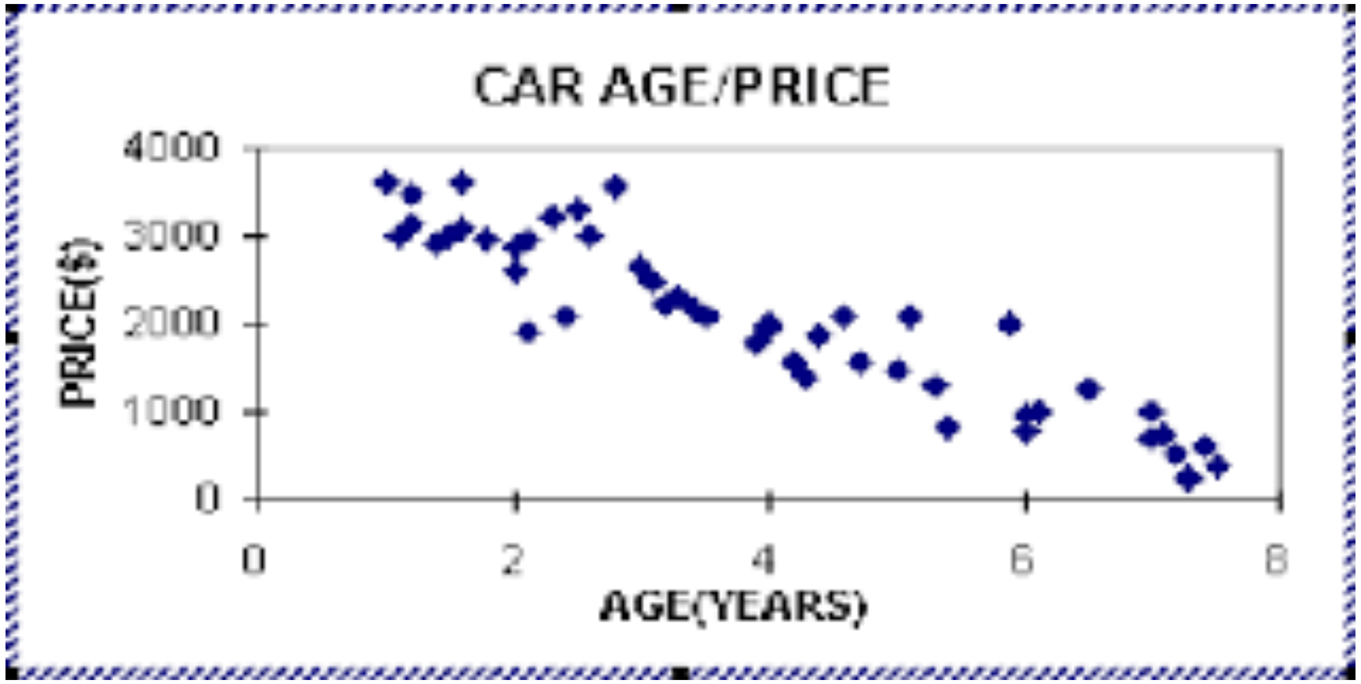


Year	Petroleum Imports (thousand barrels per day)	Total Petroleum Imports as Percent of Petroleum Products Supplied	Petroleum Imports from Persian Gulf as Percent of Total Petroleum Imports
1973	6256	36.1	13.5
1974	6112	36.7	17.0
1975	6055	37.1	19.2
1976	7313	41.8	25.1
1977	8807	47.7	27.8
1978	8363	44.3	26.5
1979	8456	45.6	24.4
1980	6909	40.5	21.9
1981	5996	37.3	20.3
1982	5113	33.4	13.6
1983	5051	33.1	8.7
1984	5437	34.5	9.3
1985	5067	32.2	6.1
1986	6224	38.2	14.6
1987	6678	40.0	16.1
1988	7402	42.8	20.8
1989	8061	46.5	23.0
1990	8018	47.1	24.5
1991	7627	45.6	24.1
1992	7888	46.3	22.5
1993	8620	50.0	20.6
1994	8996	50.7	19.2
1995	8835	49.8	17.8
1996	9478	51.7	16.9
1997	10,162	54.5	17.2
1998	10,708	56.6	19.9
1999	10,852	55.5	22.7
2000	11,459	58.1	21.7
2001	11,871	60.4	23.2
2002	11,530	58.3	19.6
2003	12,264	61.2	20.3
2004	13,145	63.4	18.9
2005	13,714	65.9	17.0
2006	13,707	66.3	16.1
2007	13,468	65.1	16.1
2008	12,915	66.2	18.4

Scatter Diagrams

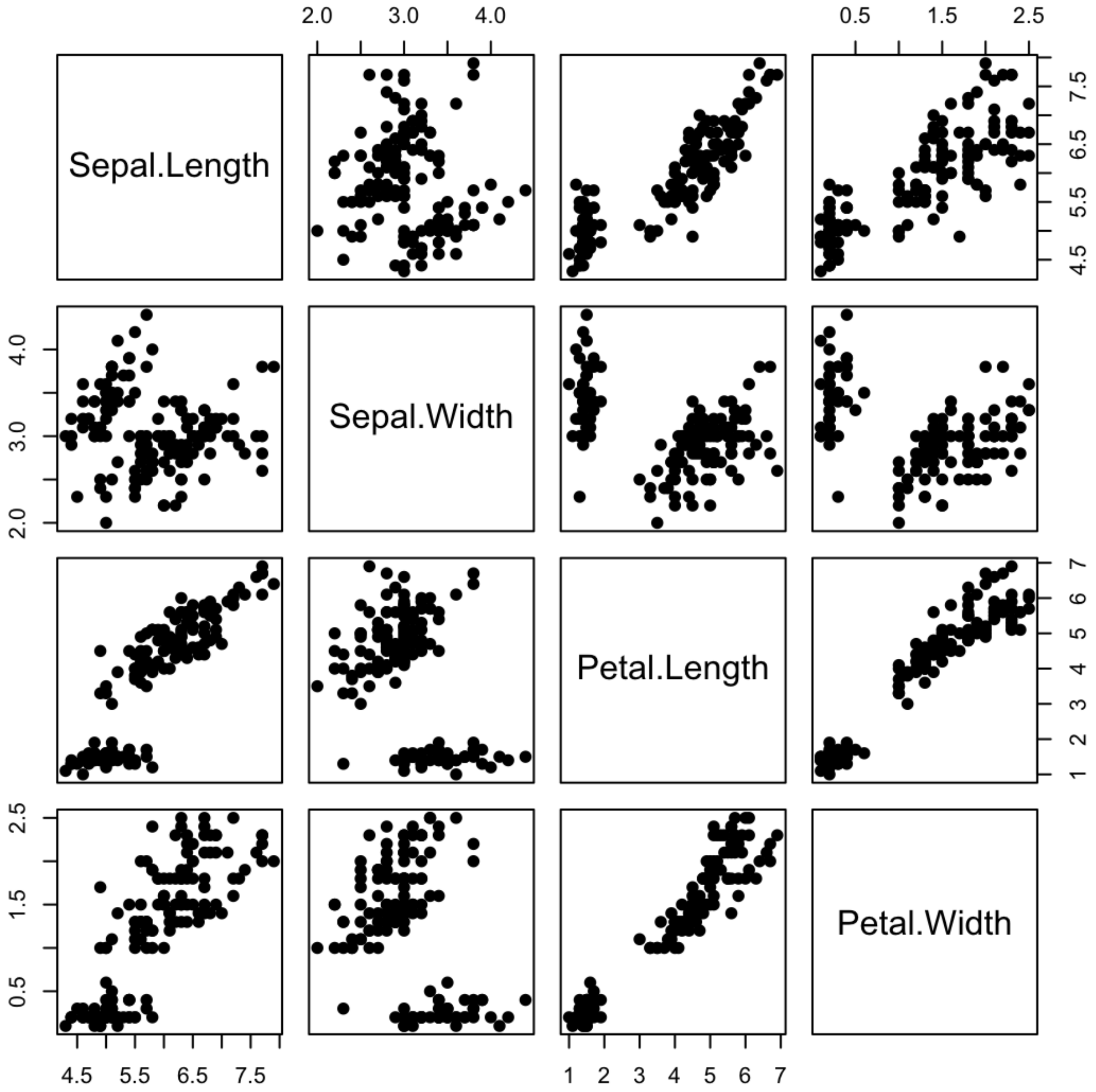
- In many problems, engineers and statisticians work with **multivariate** data
- Scatter plot can graphically display the potential relationship between two variables
- When the two variables are correlated, they should follow along a straight line





More than two variables

- When more than two variables are involved, can have a matrix of scattered diagrams



Example

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

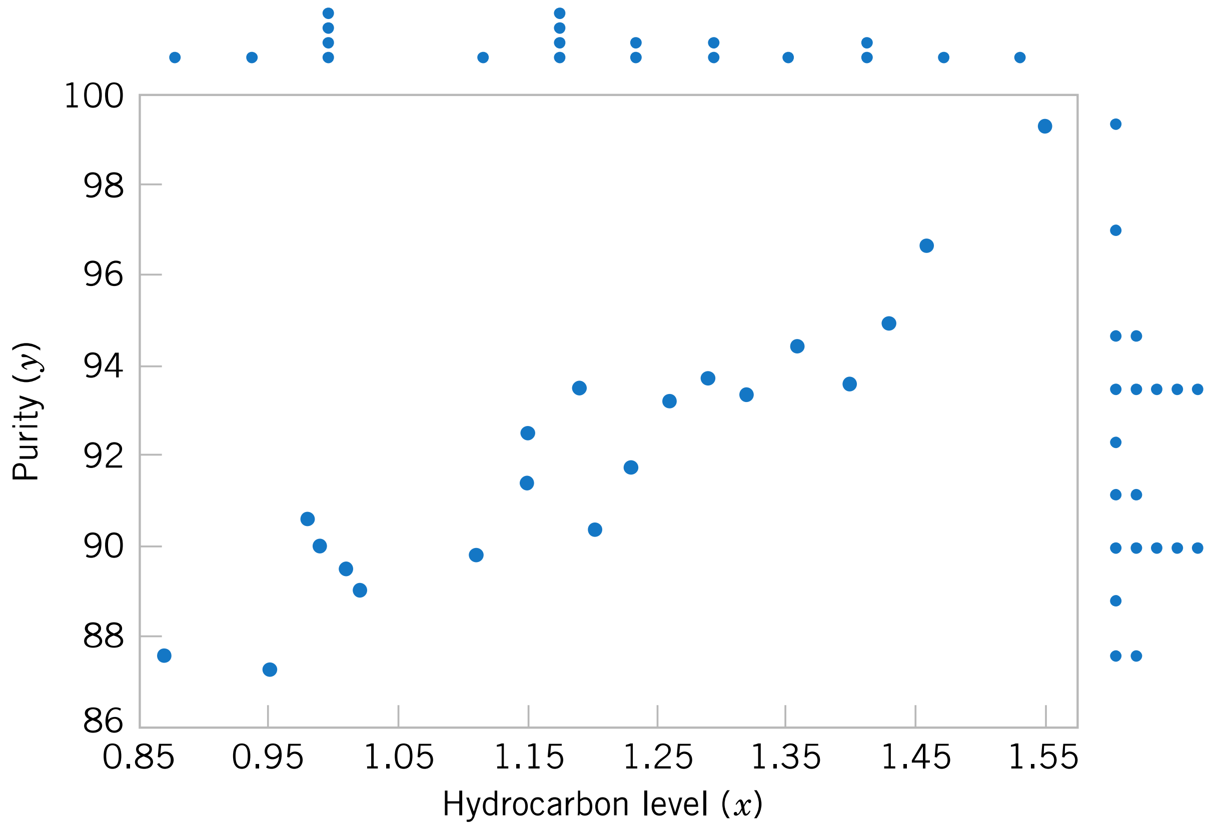
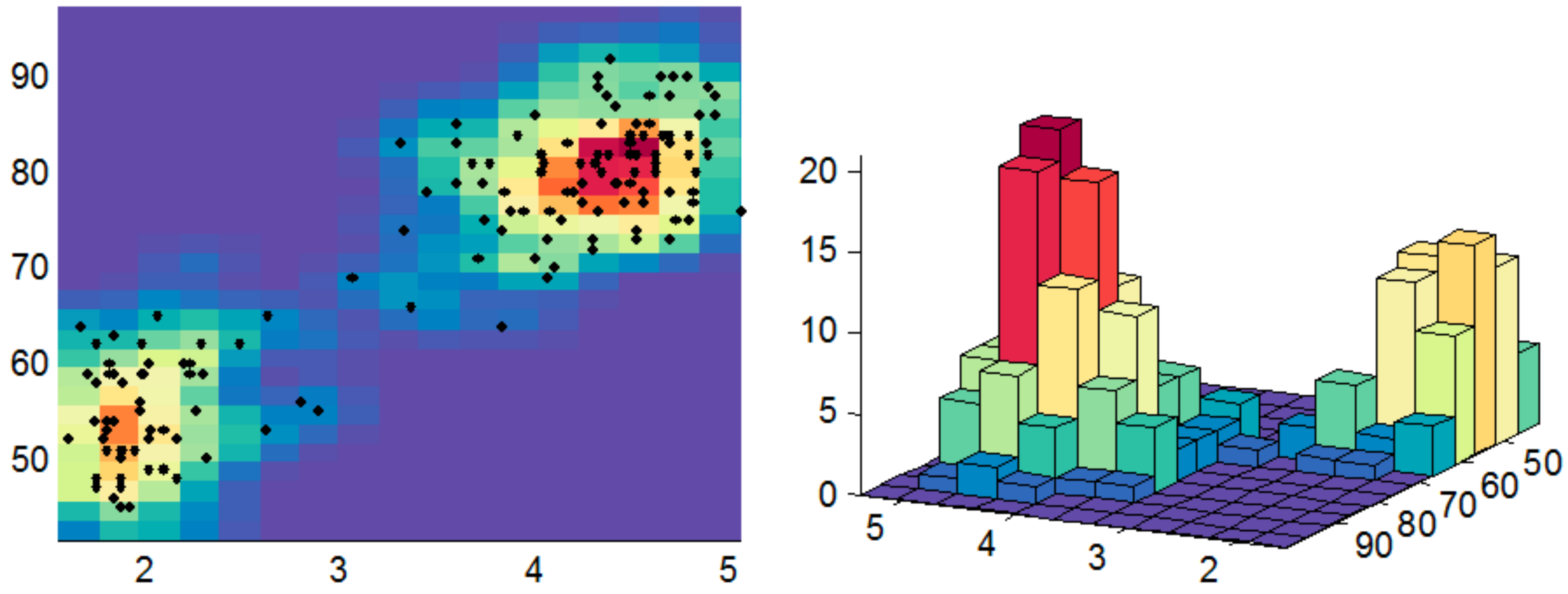
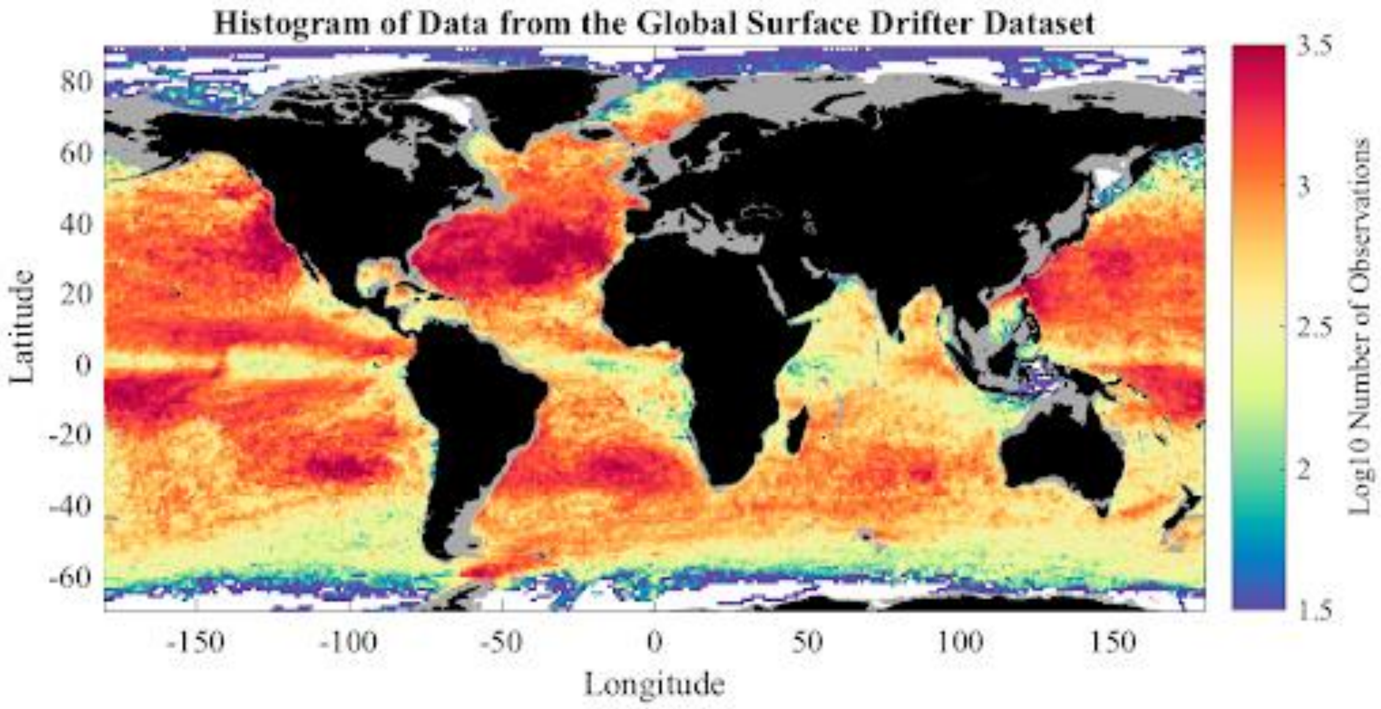


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

Two Dimensional Histogram




Histogram of Data from the Global Surface Drifter Dataset
80
60
40
20
0
-20
-40
-60
Latitude
-150
-100
-50
0
50
100
150
Longitude
3.5
3
2.5
2
1.5
Log10 Number of Observations

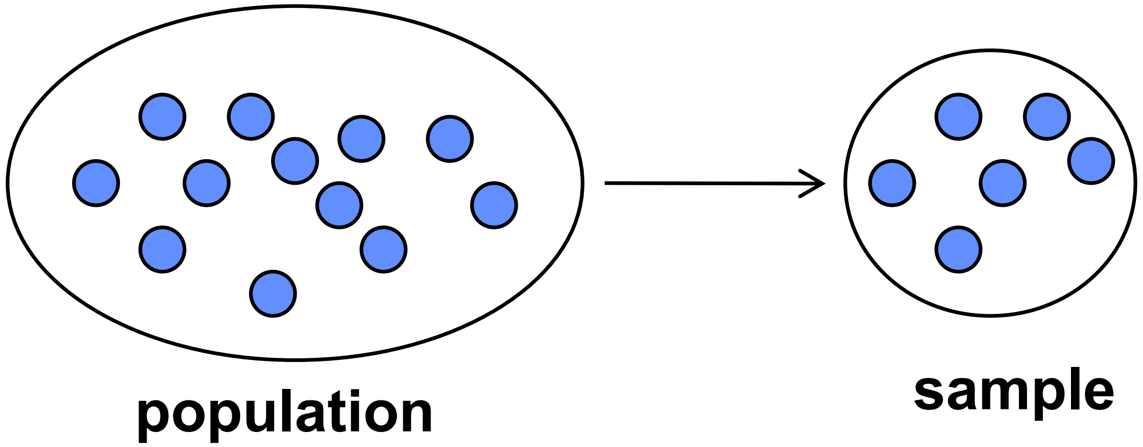
Descriptive Vs. Inferential Statistics

- Descriptive Statistics:

  A set of statistical techniques used to *organize*, *summarize*, *display*, and *describe* important features of data

- Inferential (a.k.a. inductive) Statistics:

  A set of statistical methods that uses ***sample*** information to draw conclusion about the ***population***


population
sample