Chapter 5: Descriptive Statistics

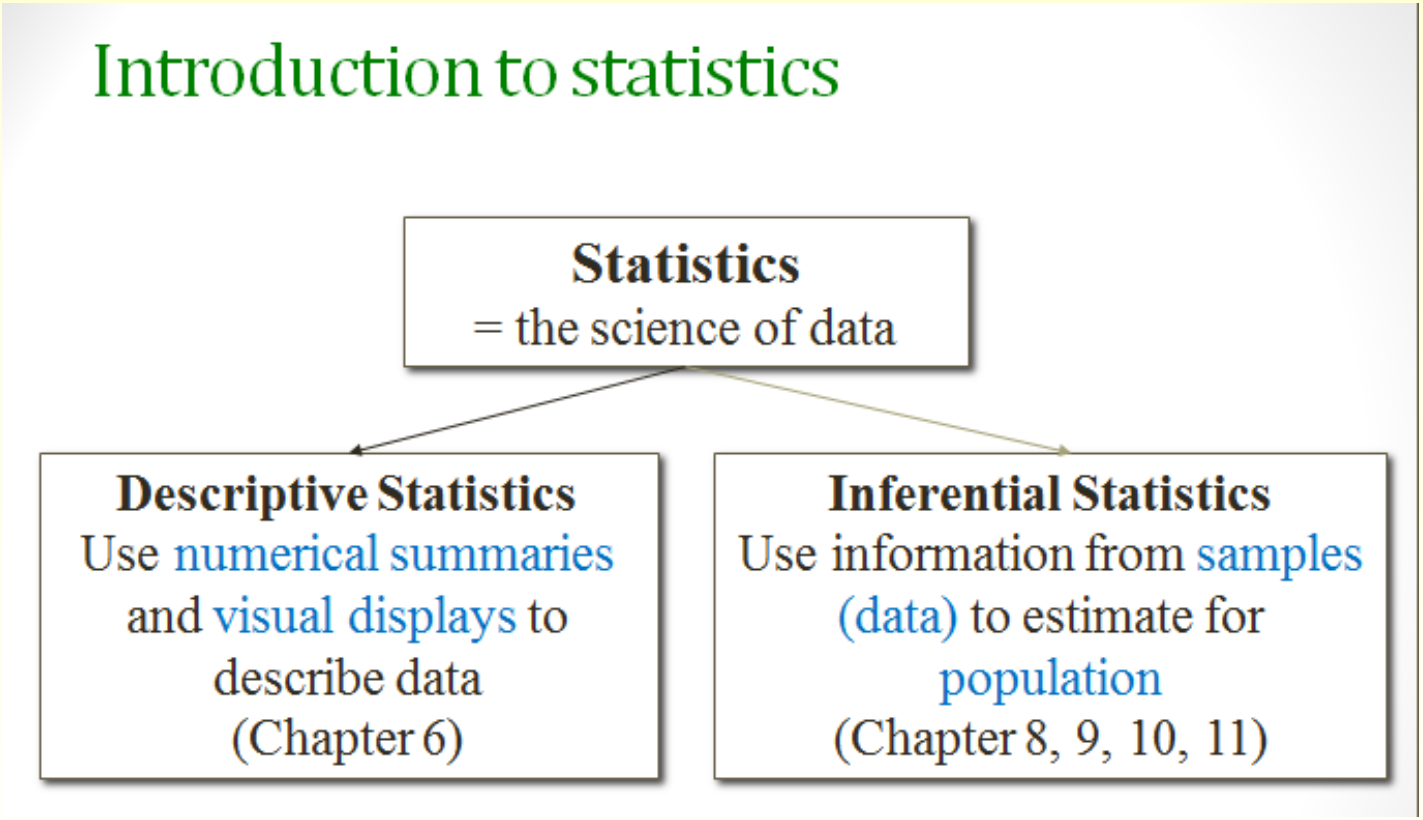
Course Name: PROBABILITY & STATISTICS

Lecturer: Duong Thi Hong

Hanoi, 2022

Content

Introduction



Content

1. Numerical Summaries of Data

Sample Mean

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

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: Let's consider the weight of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6 and 13.1. Find the sample mean.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{12.6 + 12.9 + ... + 13.1}{8} = 13$$

Question 1

During one recent year, U.S. consumers redeemed 6.79 billion manufacturers' coupons and saved themselves $2.52 billion. Calculate and interpret the mean savings per coupon.

Select one:

- a. The average savings was $0.26 per coupon.
- b. Half of all coupons were worth more than $0.37 in savings.
- c. The average savings was $0.37 per coupon.
- d. Half of all coupons were worth more than $0.26 in savings.

Question 2

The scores for a statistics test are as follows:

61	76	93	77	73	92	80	85	78	89
79	65	50	66	85	91	85	64	18	71

Compute the mean score.

Select one:

- a. 77.10
- b. 75
- c. 73.90
- d. 64.85

Question 3

Elaine gets quiz grades of 67, 64, and 87. She gets a 84 on her final exam. Find the mean grade if the quizzes each count for 15% and final exam counts for 55% of the final grade.

Select one:

- a. 75.5
- b. 78.3
- c. 72.1
- d. 78.9

Question 4

A data processing firm sampled 75 small businesses to find the number of days their computer systems were down during the previous three months. The distribution of responses is given below. Find the sample mean.

Days of down time	Frequency
0	16
1	10
2	14
3	20
4	8
5	7

Select one:

- a. 2.5
- b. 2.2
- c. 25
- d. 35

1. Numerical Summaries of Data

Although the sample mean is useful, it does not convey all of the information about a sample of data. The variability or scatter in the data may be described by the **sample variance** or the **sample standard deviation**.

Sample Variance

If $x_1, x_2, ..., x_n$ is a sample of n observations, the **sample variance** is

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

The **sample standard deviation**,s, is the positive square root of the sample variance.

Example 1

Table 6-1 Calculation of Terms for the Sample Variance and Sample Standard Deviation

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
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

Table 6-1 displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data. The sample variance is

$$s^2 = \frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{8-1} = 1.60/7 = 0.2286$$

The sample standard deviation is $s = \sqrt{s^2} = \sqrt{0.2286} = 0.48$ pounds

Question 1

Each year advertisers spend billions of dollars purchasing commercial time on network television. In the first 6 months of one year, advertisers spent $1.1 billion. In a recent article, the top 10 leading spenders and how much each spent (in million of dollars) were listed:

Company A: $73.7 Company F: $26.7
Company B: $63.9 Company G: $26.4
Company C: $57.9 Company H: $22.8
Company D: $57.1 Company I: $21.1
Company E: $32 Company J: $19.8

Calculate the sample variance.

Select one:

- a. 1987.406
- b. 2217.644
- c. 422.940
- d. 4003.428

Question 2

The mean $\overline{x}$ of a data set is 36.71, and the sample standard deviation **s** is 3.22. Find the interval representing measurements within one standard deviation of the mean.

Select one:

- a. (30.27, 43.15)
- b. (27.05, 46.37)
- c. (35.71, 37.71)
- d. (33.49, 39.93)

Question 3

The top speeds for a sample of five new automobiles are listed below. Calculate the standard deviation of the speeds.

105, 145, 190, 140, 175

Select one:

- a. 1092.50
- b. 24.02
- c. 33.05
- d. 576.96

Question 4

Find the variance for the given data. Round your answer to one more decimals than original data

1, 4, -5, -9, and 6

Select one:

- a. 39.2
- b. 39.3
- c. 31.4
- d. 39.4

Question 5

Find the variance for the given data. Round your answer to one more decimals than original data

1, 4, -5, -9, and 6

Select one:

- a. 39.2
- b. 39.3
- c. 31.4
- d. 39.4

Question 6

Find the variance for the given sample data

53 52 75 62 68 58 49 49

Select one:

- a. 67.9
- b. 8.2
- c. 9.5
- d. 89.6

Question 7

Find the standard deviation for the given sample data:

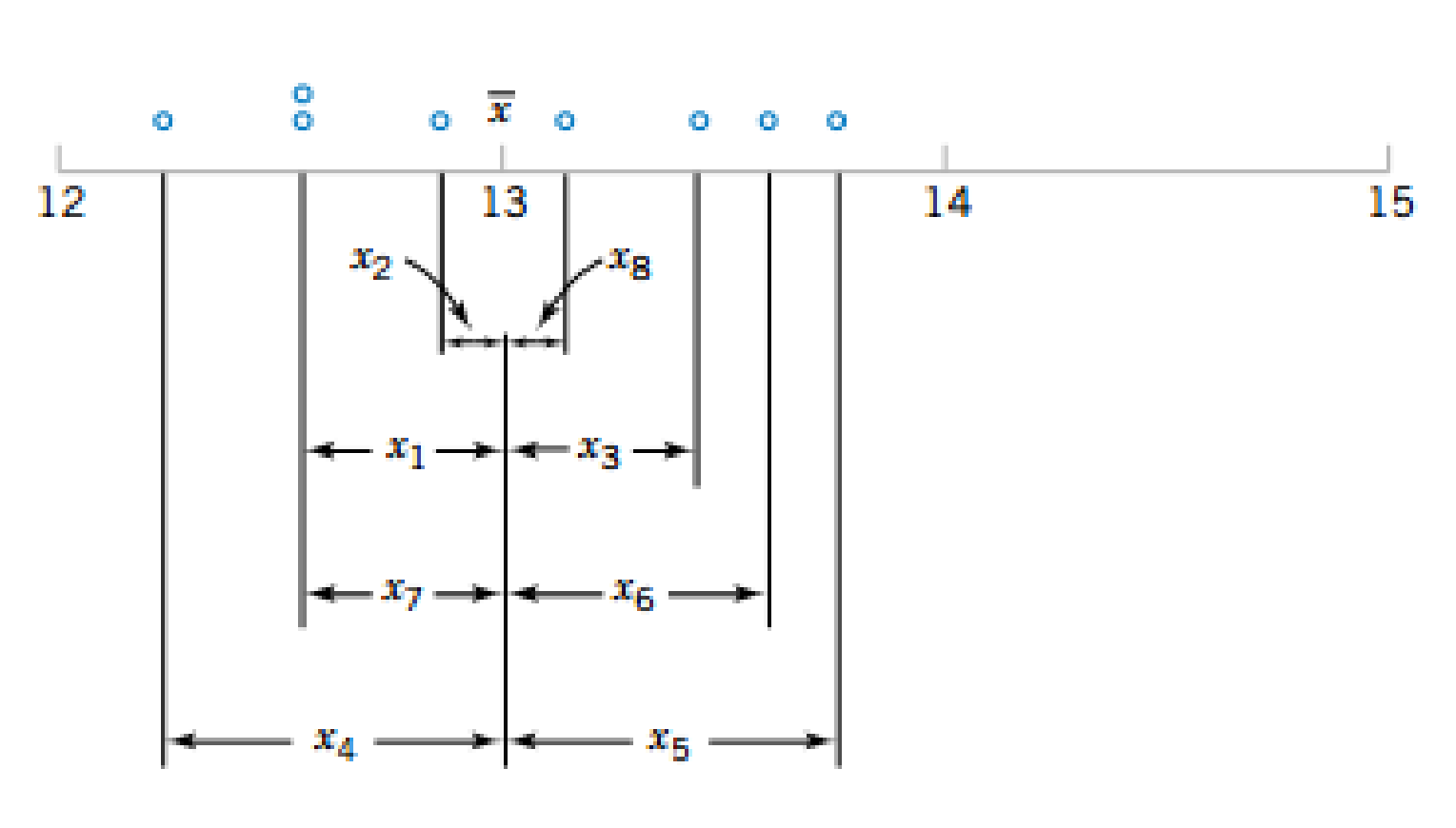
2 6 2 2 1 4 4 2 4 2 3 8 4 2 2 7 7 2 3 11

Select one:

- a. 2.7
- b. 2.4
- c. 2.5
- d. 2.6

Example 1

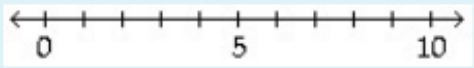
These data are plotted in Fig. 6-2.



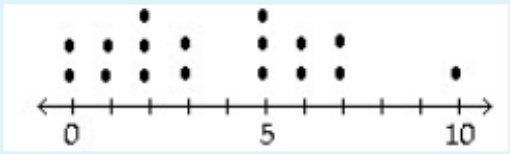
Question 1

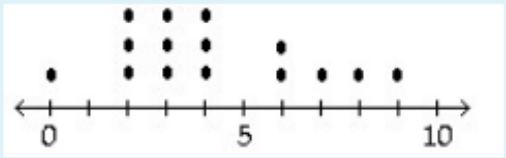
Attendance records at a school show the number of days each student was absent during the year. The days absent for each stude
0 2 3 4 2 3 4 6 7 2 3 4 6 9 8

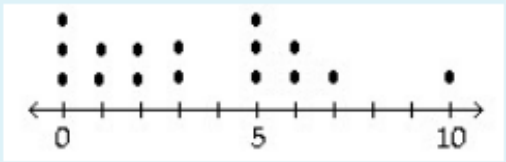
Construct the dot plot for the given data.

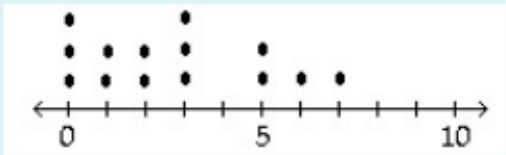


Select one:

a. 

b. 

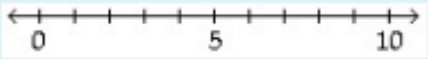
c. 

d. 

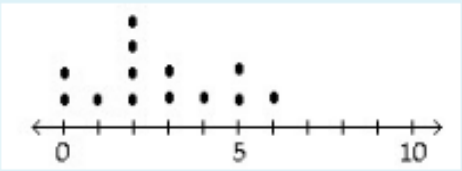
Question 2

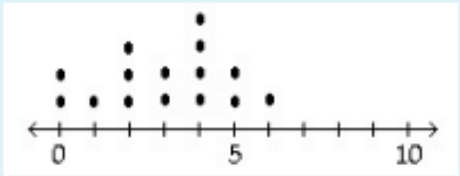
A manufacturer records the number of errors each work station makes during the week. The data are as follows.
6 3 2 3 5 2 0 2 5 4 2 0 1

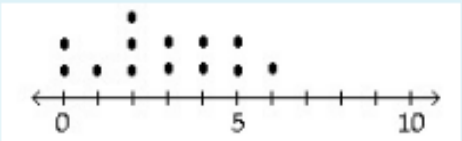
Construct the dot plot for the given data.

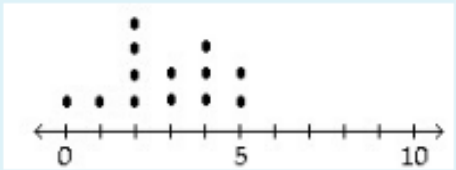


Select one:

a.


b.


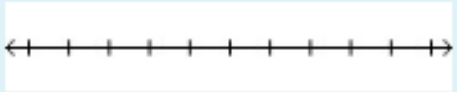
c.


d.


Question 3

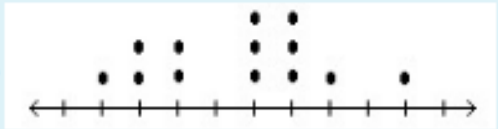
A store manager counts the number of customers who make a purchase in his store each day. The data are as follows.
10 11 8 14 7 10 10 11 8 7

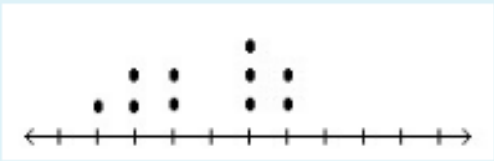
Construct the dot plot for the given data.

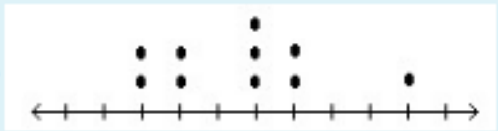


Select one:

a. 

b. 

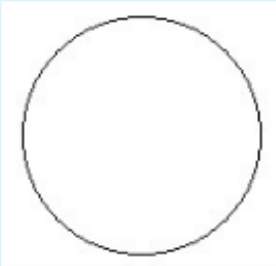
c. 

d. 

Question 4

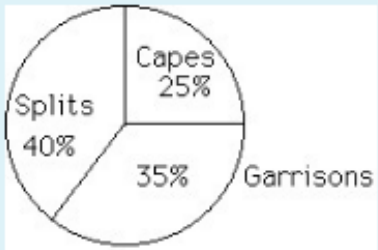
The following data give the distribution of the types of houses in a town containing 30,000 houses.
Capes: 7500, Garrisons: 10,500, Splits: 12,000

Construct a pie chart representing the given data set.

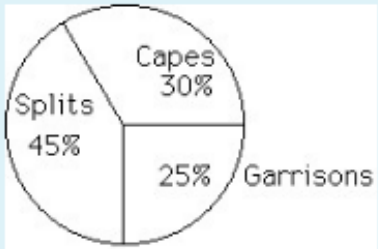


Select one:

a.



b.

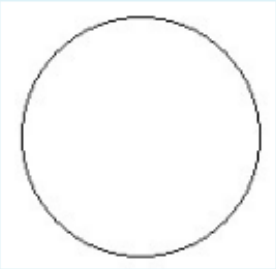


Question 5

After reviewing a movie, 800 people rated the movie as excellent, good, or fair. The following data give the rating distribution.

Excellent: 160, Good: 400, Fair: 240

Construct a pie chart representing the given data set.



Select one:

a.

b.

1. Numerical Summaries of Data

In addition to the sample variance and sample standard deviation, the **sample range**, or the difference between the largest and smallest observations, is a useful measure of variability.

Sample Range

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

$$r = \max(x_i) - \min(x_i)$$

Median

The **sample median** is a measure of central tendency that divides the data into two equal parts, half below the median and half above.

- If the number of observations is even, the median is halfway between the two central values.
- If the number of observations is odd, the median is the central value.

Example: The prices (in dollars) for a sample of round-trip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices:
872 432 397 427 388 782 397

1. Numerical Summaries of Data

Mode

The **sample mode** is the most frequently occurring data value.

Quartiles

When an ordered set of data is divided into four equal parts, the division points are called **quartiles**. The first or lower quartile, q_1 or Q_1, is a value that has approximately 25% of the observations below it and approximately 75%.

Interquartile Range (IQR)

An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.
- The interquartile range, $IQR = Q_3 - Q_1$

Question 1

Use the given sample data to find Q_1.
55, 52, 52, 52, 49, 74, 67, 55.

Select one:

- a. 67.0
- b. 6.0
- c. 52.0
- d. 55.0

Question 2

Find the mode(s) for the given data

6.8	5.9	5.7	6.5	5.4	6.8	5.9
4.9	4.8	6.5	6.8	4.9	6.5	4.7

Select one:

- a. 6.5
- b. none
- c. 6.8 and 6.5
- d. 6.8

Question 3

Find the mode(s) for the given dample data

98, 25, 98, 13, 25, 29, 56, 98

Select one:

- a. 42.5
- b. 25
- c. 55.3
- d. 98

Question 4

A store manager kept track of the number of newspapers sold each week over a seven-week period. The results are shown below:

95, 38, 221, 122, 258, 237, 233.

Find the median number of newspapers sold

Select one:

- a. 233
- b. 122
- c. 172
- d. 221

Question 5

Use the given sample data to find three quartiles:

15, 21, 3, 6, 10, 28, 36, 1

Select one:

- a. 4.5, 12.5, 24.5
- b. 5.25, 12.5, 24.5
- c. 4.5, 12.5, 22.75
- d. 5.25, 12.5, 22.75

Question 6

The distances traveled (in miles) to 7 different swim meets are given below:

71, 12, 18, 31, 46, 69, 85.

Find the median distance traveled.

Select one:

- a. 47 miles
- b. 31 miles
- c. 46 miles
- d. 69 miles

Question 7

Use the given sample data to find three quartiles:

5, 21, 13, 16, 11, 28, 36, 13, 22

Select one:

- a. 12, 16, 25
- b. 13, 16, 22
- c. None of the others
- d. 17, 11, 24.5

Question 8

Find the mode(s) for the given sample data

11, 13, 11, 23, 22, 24, 56, 22, 72, 15, 27

Select one:

- a. 11 and 22
- b. 24
- c. 11
- d. 15

Question 9

Health care issues are receiving much attention in both academic and political arenas. A sociologist recently conducted a survey of citizens over 60 years of age whose net worth is too high to qualify for Medicaid and have no private health insurance. The ages of 25 uninsured senior citizens were as follows:

60 61 62 63 64 65 66 68 68 69 70 73 73 74 75 76 76 81 81 82 86 87 89 90 92

Identify the first quartile of the ages of the uninsured senior citizens.

Select one:

- a. 65.5
- b. 81.5
- c. 73
- d. 75

Question 10

The data below represent the amount of grams of carbohydrates in a serving of breakfast cereal in a sample of 11 different servings.

11 15 23 29 19 22 21 20 15 25 17

What is the value of IQR?

Select one:

- a. 7
- b. 6
- c. 8
- d. 9

Question 11

Calculate the range of the following data set:

7, 8, 4, 1, 4, 15, 5, 8, 5

Select one:

- a. 14
- b. 15
- c. 16
- d. 1

Question 12

The manager of an electrical supply store measured the diameters of the rolls of wire in the inventory. The diameters of the rolls (in m) are listed below:

0.165 0.114 0.503 0.392 0.579 0.311. Find the range of data.

Select one:

- a. 0.503
- b. 0.114
- c. 0.146
- d. 0.465

Question 13

A sociologist recently conducted a survey of senior citizens who have net worths too high to qualify for Medicaid private health insurance. The ages of the 25 uninsured senior citizens were as follows:

69	74	67	77	87
75	62	90	66	91
70	93	77	63	82
64	69	82	71	74
61	88	76	65	83

Find the median of the observations.

Select one:

- a. 75
- b. 71
- c. 74
- d. 74.5

Question 14

The test scores of 32 students are listed below. Find Q_3.

32 37 41 44 46 48 53 55
56 57 59 63 65 66 68 69
70 71 74 74 75 77 78 79
80 82 83 86 89 92 95 99

Select one:

- a. 80
- b. 24
- c. 79
- d. 79.5

1. Numerical Summaries of Data

Exercise 1

1) Will the sample mean always correspond to one of the observations in the sample?
2) Will exactly half of the observations in a sample fall below the mean?
3) Will the sample mean always be the most frequently occurring data value in the sample?
4) Can the sample standard deviation be equal to zero? Give an example.

1. Numerical Summaries of Data

Exercise 1

1) Will the sample mean always correspond to one of the observations in the sample?
2) Will exactly half of the observations in a sample fall below the mean?
3) Will the sample mean always be the most frequently occurring data value in the sample?
4) Can the sample standard deviation be equal to zero? Give an example.

Exercise 2

Suppose that you add 10 to all of the observations in a sample. How does this change the sample mean? How does it change the sample standard deviation?

Content

2. Stem-and-Leaf Diagrams

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

Steps to Construct 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.

2. Stem-and-Leaf Diagrams

Example 2

To illustrate the construction of a stem-and-leaf diagram, consider the alloy compressive strength data in Table 6-2.

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

2. Stem-and-Leaf Diagrams

Example 2

We will select as stem values the numbers $7, 8, ..., 24$. The resulting stem-and-leaf diagram is presented in Fig. 6-4.

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

2. Stem-and-Leaf Diagrams

Example 2

The last column in the diagram is a frequency count of the number of leaves associated with each stem. Inspection of this display immediately reveals that:

- Most of the compressive strengths lie between 110 and 200 psi.
- A central value is somewhere between 150 and 160 psi.
- The strengths are distributed approximately symmetrically about the central value.

2. Stem-and-Leaf Diagrams

Example 3

Figure 6-5 illustrates the stem-and-leaf diagram for 25 observations on batch yields from a chemical process.

Stem	Leaf
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

(a)

Stem	Leaf
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

(b)

Stem	Leaf
6z	1
6t	3
6f	4 5 5
6s	6
6e	
7z	0 1 1
7t	3
7f	5
7s	7
7e	8 8 9
8z	1
8t	3
8f	4 4
8s	7
8e	8 8
9z	
9t	2 3
9f	5
9s	
9e	

(c)

2. Stem-and-Leaf Diagrams

Example 3

- In Fig. 6-5(a) we have used 6, 7, 8, and 9 as the stems: This results in too few stems, and the stem-and-leaf diagram does not provide much information about the data.
- In Fig. 6-5(b) we have divided each stem into two parts, resulting in a display that more adequately displays the data.
- Figure 6-5(c) illustrates a stem and-leaf display with each stem divided into five parts: There are too many stems in this plot, resulting in a display that does not tell us much about the shape of the data.

2. Stem-and-Leaf Diagrams

Exercise 1

1) When will the median of a sample be equal to the sample mean?
2) When will the median of a sample be equal to the mode?

2. Stem-and-Leaf Diagrams

Exercise 1

1) When will the median of a sample be equal to the sample mean?
2) When will the median of a sample be equal to the mode?

Exercise 2

Suppose that you add 10 to all of the observations in a sample. How does this change the sample mean? How does it change the sample standard deviation?

Question 1

Use the data to create a stemplot.

The following data show the number of laps run by each participant in a marathon.
46 65 55 43 51 48 57 30 43 49 32 56

Select one:

○ a.

3	0 2
4	3 3 6 8 9
5	1 5 6 7
6	5

○ b.

3	0 2
4	3 6 8 9
4	1 3 5 6 7
6	5

Question 2

Use the data to create a stemplot.

The midterm test scores for the seventh-period typing class are listed below.
85 77 93 91 74 65 68 97 88 59 74 83 85 72 63 79

Select one:

○ a.

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

○ b.

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

Question 3

Use the data to create a stemplot.

The attendance counts for this season's basketball games are listed below.
227 239 215 219
221 233 229 233
235 228 245 231

Select one:

- a.

21	5 9
22	1 7 8 9
23	1 3 3 5 9
24	5

- b.

21	5 7 9
22	1 8 9
23	1 3 3 5 9
24	5

Question 4

Find the origin data from the sterm-and-leaf plot

Stem	Leaves
8	5 8
9	1 8
10	5 5

Select one:

- a. 81, 85, 81, 98, 108, 105
- b. 85, 88, 91, 98, 105, 105
- c. 85, 81, 88, 91, 101, 105
- d. 85, 88, 91, 91, 105, 105

Question 5

Find the origin data from the sterm-and-leaf plot

Stem	Leaves
73	2 5 8
74	2 3 9
75	1 8

Select one:

- a. 73258, 74239, 7518
- b. 75, 78, 81, 76, 77, 83, 76, 83
- c. 732, 735, 738, 742, 743, 749, 751, 758
- d. 732, 735, 748, 742, 743, 749, 751, 768

Question 6

The attendace counts for this season's basketball games are listed below:

227 239 215 219 221 233 229 233 235 228 245 231

Use the data to creat a sterm plot.

Select one:

a.

21	5 7 8 9
22	1 8 9 0
23	2 3 5 9
26	5

b.

21	5 7 8 9
22	1 8 9 0
23	2 3 5 9
24	5

c.

21	5 7 9
22	1 8 9
23	1 3 3 5 9
24	5

d.

21	5 9
22	1 7 8 9
23	1 3 3 5 9
24	5

Question 7

A stem-and-leaf diagram for a set of examination scores is given below. Find sample median of these data.

Stem	Leaves
2	8
4	269
5	01389
6	246
7	05

Select one:

- a. 55.5
- b. 50.5
- c. 51
- d. None of the other choices is correct
- e. 53

Question 8

Construct the stem-and-leaf diagram for the below data.
16.9; 15.2; 17.5; 15.5; 16.8; 16.8; 17.1; 17.5; 15.3.

Select one:

○ a.

Stem	Leaf
15	253
16	789
17	225

○ b. None of the other choices is correct

○ c.

Stem	Leaf
15	235
16	889
17	155

○ d.

Stem	Leaf
15	235
16	89
17	255

Content

3. Frequency Distributions and Histograms

A frequency distribution is a more compact summary of data than a stem-and-leaf diagram. To construct a frequency distribution, we must divide the range of the data into intervals, which are usually called **class intervals**, **cells**,or **bins**.

A frequency distribution for the comprehensive strength data in Table 6-2 is shown in **Table 6-4**.

Table 6-4 Frequency Distribution for the Compressive Strength Data in Table 6-2

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

3. Frequency Distributions and Histograms

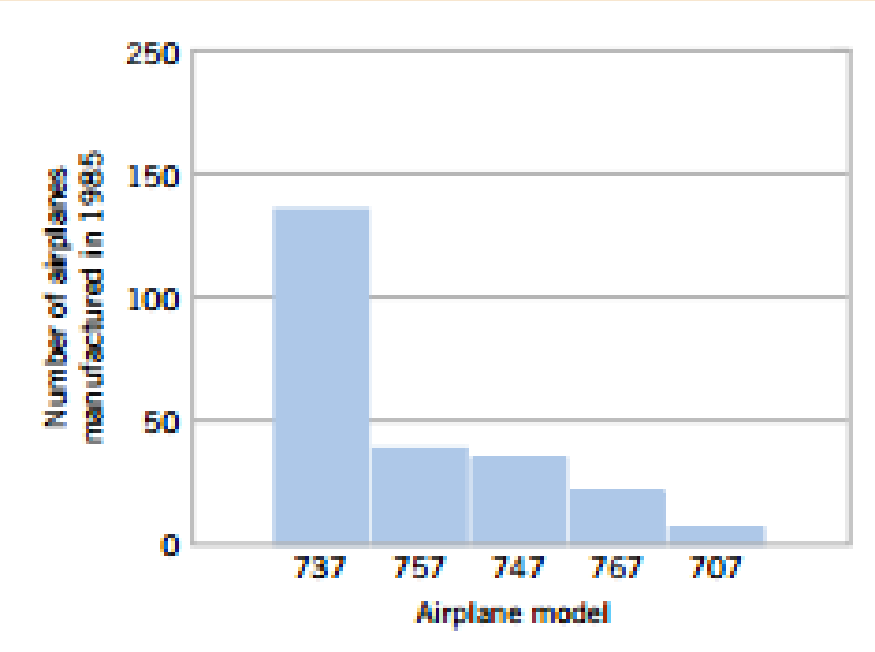
The **histogram** is a visual display of the frequency distribution. The steps for constructing a histogram follow

Constructing a Histogram (Equal Bin Widths)

1) Label the bin (class interval) boundaries on a horizontal scale.
2) Mark and label the vertical scale with the frequencies or the relative frequencies.
3) Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

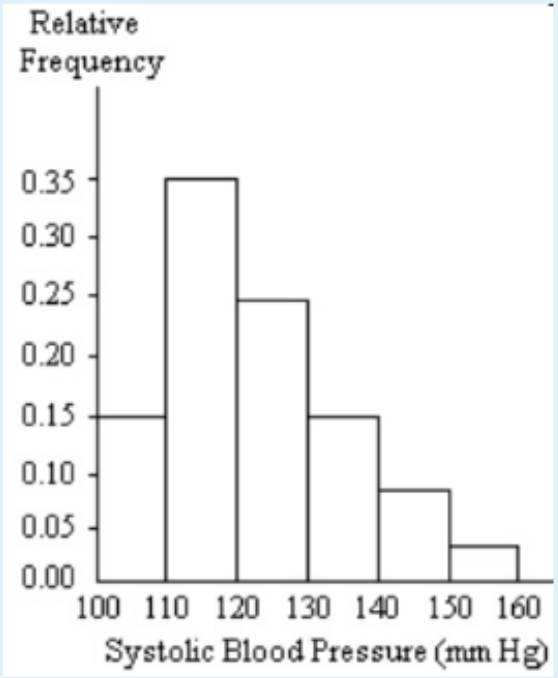
Example 4

Figure below presents the production of transport aircraft by the Boeing Company in 1985. Notice that the 737 was the most popular model, followed by the 757, 747, 767, and 707.



Question 1

A nurse measured the blood pressure of each person who visited her clinic. Following is a relative-frequence histogram for the systolic blood pressure readings for those people aged between 25 and 40. The blood pressure reading were given to the nearest whole number. Approximately what percentage of the people aged 25-40 had a systolic blood pressure reading between 110 and 119 inclusive?



Select one:

- a. 30%
- b. 3.5%
- c. 35%
- d. 0.35%

Question 2

Ten cartons of fragile ceramic castings were shipped on each of two air freight carries. On delivery at their destination the cartons were opened and inspected. The number of damaged items per carton were as follows:

17, 20, 1, 18, 5, 14, 18, 10, 6, 2.

Assume that you are finding the frequency distribution using groupings: 1-4 inclusively, 5-8 inclusively, 9-12 inclusively and so on.What is the frequency of the interval 5-8?

Select one:

- a. 2
- b. 1
- c. 4
- d. 0

Content

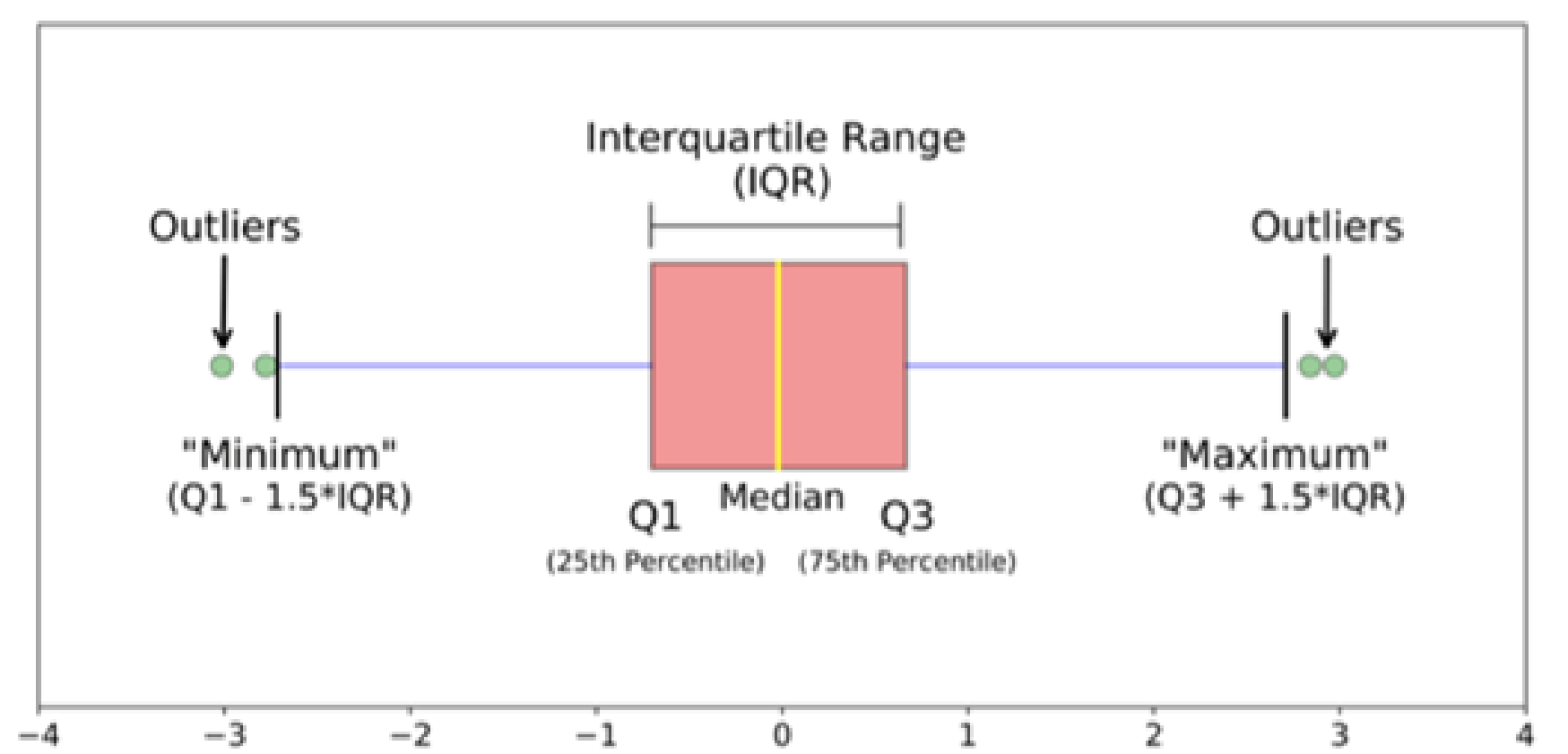
4. Box Plots

An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.
- The interquartile range, $IQR = Q_3 - Q_1$

4. Box Plots

A box-plot is a visual display that describes important features of data: three quartiles, the minimum/maximum values, and unusual observations (outliers).

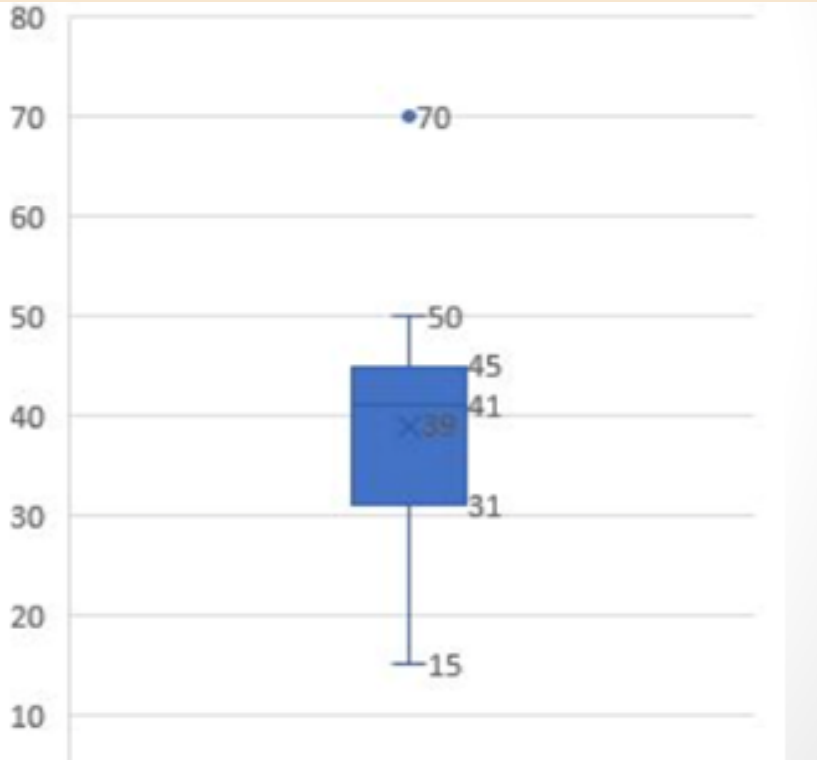


4. Box Plots

Example

Given a data of ages of 14 random adults from a village: 15, 20, 31, 31, 32, 40, 41, 41, 42, 43, 45, 45, 50, 70. Draw a box plot for this data.

Median	41
Q1	31
Q3	45
IQR	14
Outlier	70



Content

5. Times Sequence Plots

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.)

Years	Annual salary
2005	20614
2006	21175
2007	24766
2008	28100
2009	30189
2010	24618
2011	22006
2012	30912
2013	32523
2014	35285