Math 227 Elementary Statistics

Bluman 5th edition

CHAPTER 3

Data Description

Objectives

- Summarize data using measures of central tendency, such as the mean, median, mode, and midrange.
- Describe data using the measures of variation, such as the range, variance, and standard deviation.
- Identify the position of a data value in a data set using various measures of position, such as percentiles, deciles, and quartiles.

Objectives (cont.)

- Use the techniques of exploratory data analysis, including boxplots and five-number summaries to discover various aspects of data.

Introduction

- Measures of average are called **measures of central tendency** and include the mean, median , mode, and midrange.

  *Loosely stated, the average means the center of the distribution or the most typical case.

Introduction

- Measures that determine the spread of the data values are called **measures of variation** or **measures of dispersion** and include the range, variance, and standard deviation.

  *Do the data values cluster around the mean, or are they spread more evenly throughout the distribution?

Introduction

- Measures of a specific data value’s relative position in comparison with other data values are called **measures of position** and include percentiles, deciles, and quartiles.

*Measures of position tells where a specific data value falls within the data set or its relative position in comparison with other data value?

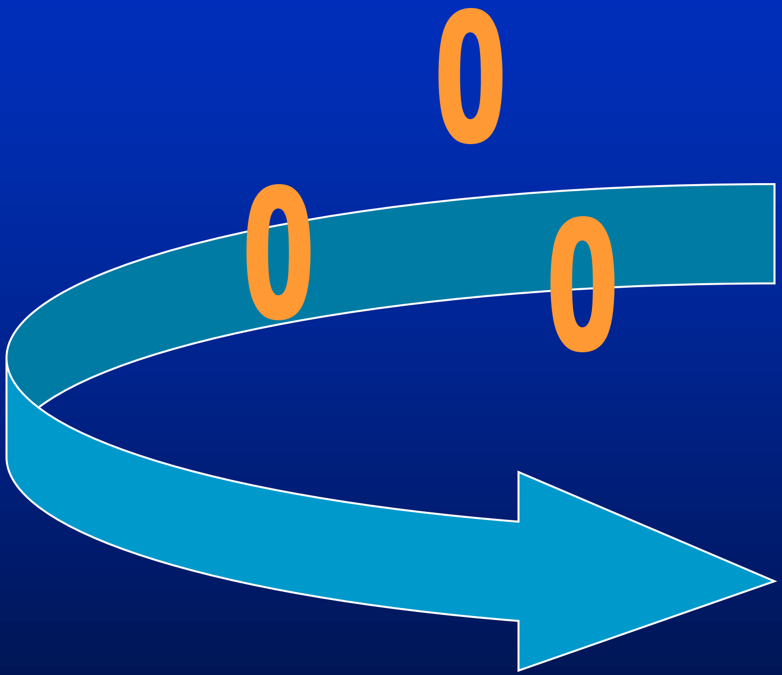
Section 3 - 1 Measures of Central Tendency

- A **statistic** is a characteristic or measure obtained by using the data values from a sample.

- A **parameter** is a characteristic or measure obtained by using all the data values for a specific population.

General Rounding Rule

- In statistics the basic rounding rule is that when computations are done in the calculation, rounding should not be done until the final answer is calculated.



I. Mean and Mode

The symbol for a population mean is μ (mu).

The symbol for a sample mean is $\bar{x}$ (read “x bar”).

The **mean** is the sum of the values, divided by the total number of values.

$$\left(\text{sample mean}\right) \rightarrow \bar{x} = \frac{\sum x}{n}$$

$$\left(\text{population mean}\right) \rightarrow \mu = \frac{\sum x}{N}$$

x is any data value from the data set.

n is the total number of data (*n* is called the sample size)

Rounding Rule for the Mean: The mean should be rounded to one more decimal place than occurs in the raw data.

Example 1: Find the mean of 24, 28, 36

$$\bar{x} = \frac{\sum x}{n} = \frac{24+28+36}{3}$$

$$\bar{x} = 29.3$$

Mode is the value that occurs most often in a data set. A data set can have more than one mode or no mode at all.

Example 2: Find the mode of 2.3 2.4 2.8 2.3 4.5 3.1

Mode = 2.3

Example 3: Find the mode of 3 4 7 8 11 13

There is no mode.

The mode is the only measure of central tendency that can be used in finding the most typical case when the data are categorical.

The procedure for finding **the mean for grouped data** uses the midpoints of the classes. The formula for finding the mean of grouped data is $\bar{x} = \frac{\sum f \cdot x_m}{n}$

The modal class is the class with the largest frequency.

Example 4:

Thirty automobiles were tested for fuel efficiency (in miles per gallon). Find the mean fuel efficiency and the modal class for the frequency distribution obtained from the thirty automobiles.

Class Boundaries	Frequency (f)	Midpoint (x_m)	$f \cdot x_m$
$7.5-12.5$	3	$\frac{7.5+12.5}{2}=10$	$3 \cdot 10 = 30$
$12.5-17.5$	5	$\frac{12.5+17.5}{2}=15$	$5 \cdot 15 = 75$
$17.5-22.5$	15	$\frac{17.5+22.5}{2}=20$	$15 \cdot 20 = 300$
$22.5-27.5$	5	$\frac{22.5+27.5}{2}=25$	$5 \cdot 25 = 125$
$27.5-32.5$	2	$\frac{27.5+32.5}{2}=30$	$2 \cdot 30 = 60$
	30		590

$$\bar{x} = \frac{\sum f \cdot x_m}{n}$$

$$= \frac{590}{30} \approx 19.7 mpg$$

Modal Class =

17.5 – 22.5

(highest frequency)

II. Median and Midrange

The **median** is the midpoint in a data set.

The symbol for a sample median is MD

1. Reorder the data from small to large.

2. Find the data that represents the middle position.

Example 1: Find the median

(a) 35, 48, 62, 32, 47

Reorder: 32, 35, 47, 48, 62

MD = 47

(b) 25.4, 26.8, 27.3, 27.5, 28.1, 26.4

Reorder: 25.4, 26.4, 26.8, 27.3, 27.5, 28.1

MD = (26.8 + 27.3) / 2 = 27.05

Example 2: Find the median

3,   5,   32,   6,   13,   11,   8,   19,   21,   6

Reorder: 3,   5,   6,   6,   8,   11,   13,   19,   21,   32

MD = (8 + 11) / 2 = 9.5

Midrange is the sum of the lowest and highest values in a data set, divided by 2.

Example 3: Find the midrange of 17, 16, 15, 13, 17, 12, 10

Reorder: 10, 12, 13, 15, 16, 17, 17

MR = (10 + 17) / 2 = 13.5

Example 4: The average undergraduate grade point average (GPA) for the top 9 ranked medical schools are listed below.

3.80 3.86 3.83 3.78 3.75 3.75 3.86 3.70 3.74

Find (a) the mean, (b) the median, (c) the mode, and (d) the midrange.

Reorder: 3.70 3.74 3.75 3.75 3.78 3.80 3.83 3.86 3.86

(a) Mean

(3.70 + 3.74 +..... + 3.86) / 9

≈ 3.786

(b) Median

5th data

MD = 3.78

(c) Mode

There are two modes: 3.75 and 3.86

(d) Midrange

$$MR = \frac{3.70+3.86}{2}$$

MR = 3.78

III. The Weighted Mean

Weighted Mean – Multiply each value by its corresponding weight and divide the sum of the products by the sum of the weights.

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots.. + w_n x_n}{w_1 + w_2 + w_3 + \ldots.. + w_n}$$

$$\bar{x} = \frac{\sum wx}{\sum w}$$

where w_1, w_2, ..., w_n are the weights and x_1, x_2, ..., x_n are the values.

Example 1:

An instructor gives four 1-hour exams and one final exam, which counts as two 1-hour exams. Find the student's overall average if she received 83, 65, 70, and 72 on the 1-hour exams and 78 on the final exam.

Scores (x)	Weights (w)	W · x
83	1	$83 \cdot 1 = 83$
65	1	$65 \cdot 1 = 65$
70	1	$70 \cdot 1 = 70$
72	1	$72 \cdot 1 = 72$
78	2	$78 \cdot 2 = 156$
	$\sum w = 6$	$\sum w \cdot x = 446$

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{446}{6} \approx 74.3$$

Example 2:

Grade distributions for a Math 227 class: In class-8%; tests-52%; computer exam-10%; and final exam-30%. A student had grades of 82, 75, 94, and 78 respectively on In class, tests, computer exam, and final exam. Find the student's final average.

	% (w)	Grades (x)	w · x
In class	8	82	$8 \cdot 82 = 656$
Tests	52	75	$52 \cdot 75 = 3,900$
Computer exam	10	94	$10 \cdot 94 = 940$
Final Exam	30	78	$30 \cdot 78 = 2,340$
$\sum$	$w = 100$		$\sum w \cdot x = 7,836$

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{7,836}{100} \approx 78.36$$

Properties of the Mean (pg 124)

- Uses all data values.
- Varies less than the median or mode
- Used in computing other statistics, such as the variance
- Unique, usually not one of the data values
- Cannot be used with open-ended classes
- Affected by extremely high or low values, called outliers

Properties of the Median (pg 124)

- Gives the midpoint
- Used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- Can be used for an open-ended distribution.
- Affected less than the mean by extremely high or extremely low values.

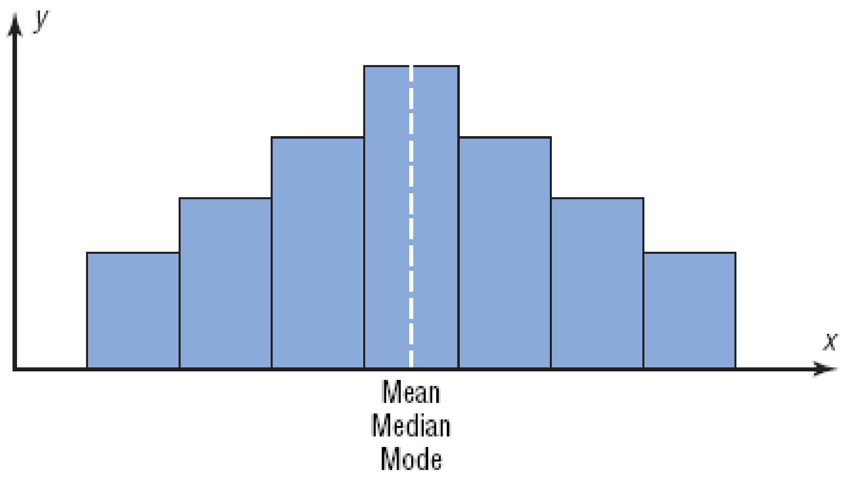
Properties of the Mode (pg 124)

- Used when the most typical case is desired
- Easiest average to compute
- Can be used with nominal data
- Not always unique or may not exist

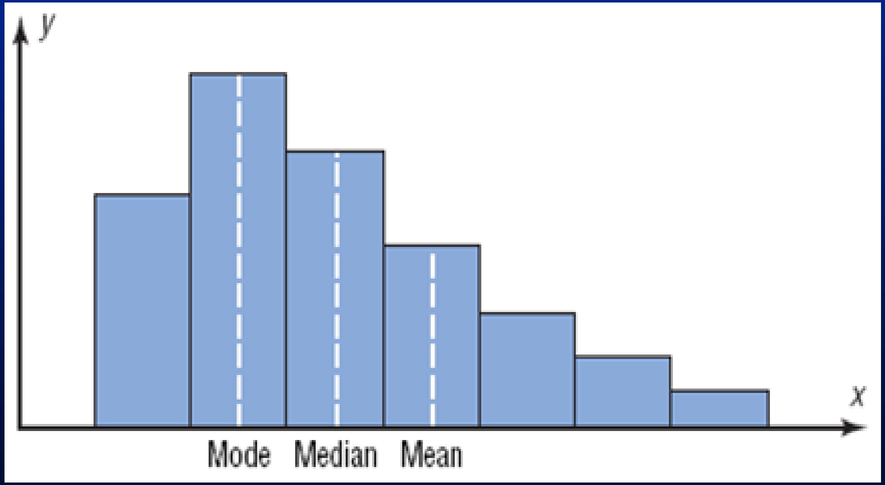
Properties of the Midrange (pg 124)

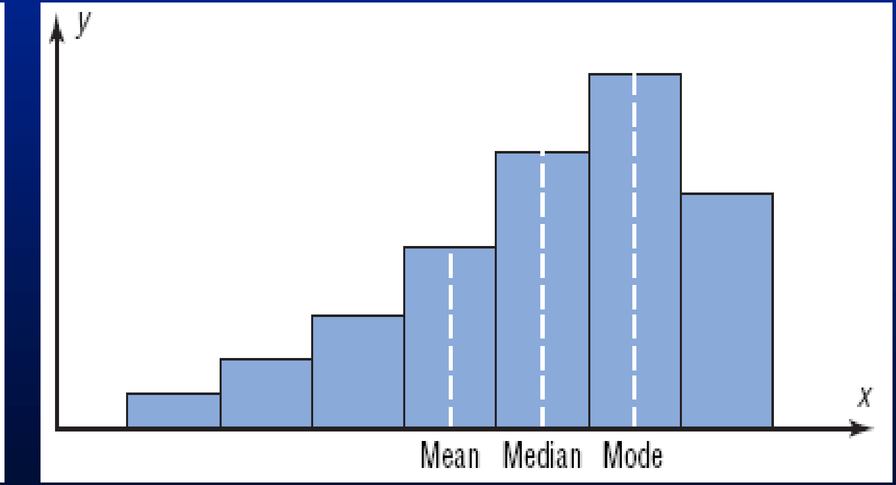
- Easy to compute.
- Gives the midpoint.
- Affected by extremely high or low values in a data set

Types of Distributions Figure 3-1



Symmetric





Positively skewed or right-skewed

Negatively skewed or left-skewed

Section 3.2 Measures of Variation

I. Range, sample variance, and sample standard deviation

Range is the highest value minus the lowest value.

R = highest value – lowest value

Example 1: Find the range of 32, 78, 54, 65, 89

R = Highest value – lowest value

R = 89 – 32 = 57

Example 3-18/19: Outdoor Paint

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean and the range of each group.

Example 3-18/19: Outdoor Paint

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Brand A: $\mu = \frac{\sum X}{N} = \frac{210}{6} = \boxed{35}$

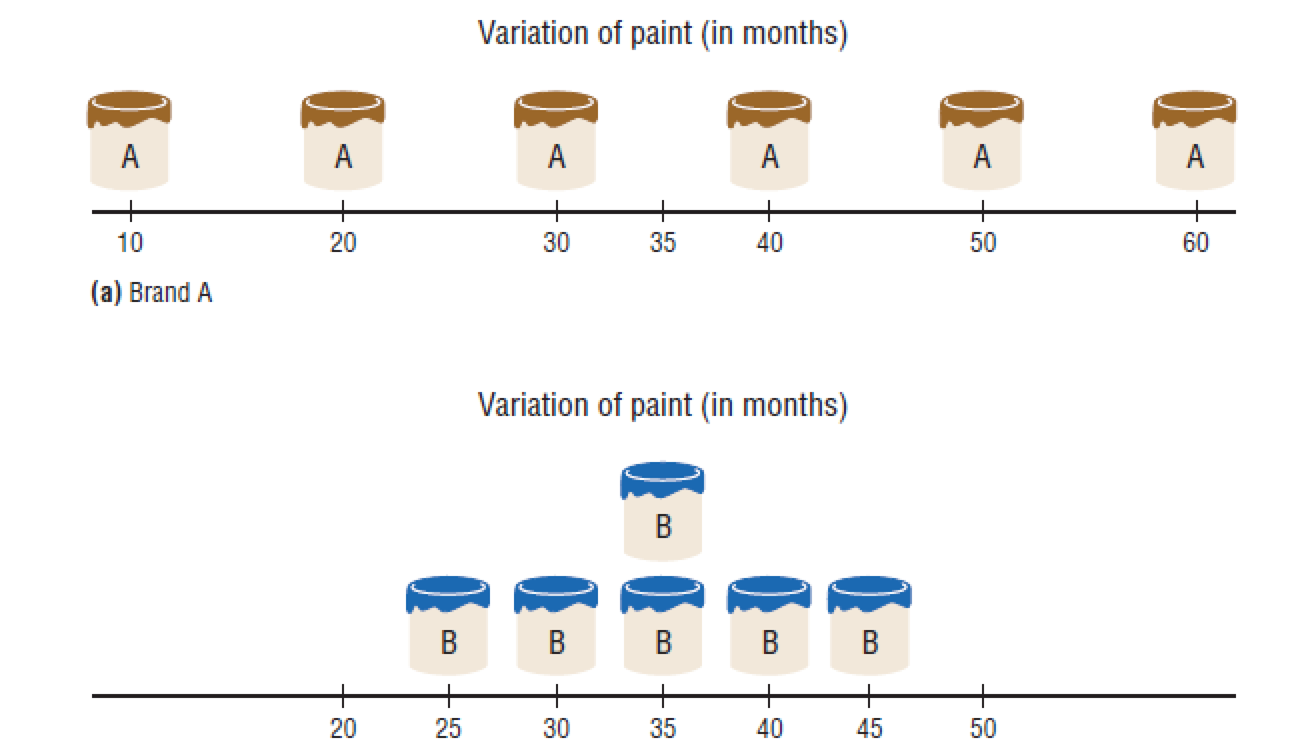
$R = 60 - 10 = \boxed{50}$

Brand B: $\mu = \frac{\sum X}{N} = \frac{210}{6} = \boxed{35}$

$R = 45 - 25 = \boxed{20}$

The average for both brands is the same, but the range for Brand A is much greater than the range for Brand B.

Which brand would you buy?



The above figure shows that brand B performs more consistently; it is less variable.

The measures of variance and standard deviation are used to determine the consistency of a variable.

Variance is the average of the square of the distance that each value is from the mean.

Measure the dispersion away from the mean

e.g. 5, 8, 11

$$\bar{x} = \frac{5+8+11}{3} = \frac{24}{3} = 8$$

Logically → sum up differences, then divide it by 3.

5 – 8 = -3

8 – 8 = 0 -3 + 0 + 3 = 0

11 – 8 = 3

Average of Difference = $\frac{-3+0+3}{3} = 0$

To avoid the cancellation, take the squared deviations.

Sum of squares = $(-3)^2 + 0^2 + 3^2 = 9 + 0 + 9 = 18$

Average of the sum of the squares (variance) =

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{18}{3-1} = \frac{18}{2} = 9$$

Standard deviation (Take the square root of the variance) =

$$s = \sqrt{\text{variance}}$$

$$s = \sqrt{9} = 3$$

Formulas for calculating variance and standard deviation

Definition Formulas

Variance of a sample

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

Standard Deviation of a sample

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$s = \sqrt{\text{variance}}$$

Computational Formulas

Variance of a sample

$$s^2 = \frac{\sum x^2 - \left[\left(\sum x \right)^2 / n \right]}{(n-1)}$$

Example 2: Use the definition formula to find the variance and standard deviation of 5, 8, 11

$$\bar{x} = \frac{5+8+11}{3} = \frac{24}{3} = 8$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
5	$5-8=-3$	$(-3)^2=9$
8	$8-8=0$	$0^2=0$
11	$11-8=3$	$3^2=9$
		$\sum(x-\bar{x})^2=18$

Sample variance : $$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{18}{3-1} = \frac{18}{2} = 9$$

Sample standard deviation : $$s = \sqrt{\text{variance}}$$

$$s = \sqrt{9} = 3$$

Example 3: Use the definition formula to find the standard deviation of 5.8, 4.6, 5.3, 3.8, 6.0

$$\bar{x} = \frac{5.8+4.6+5.3+3.8+6.0}{5} = \frac{25.5}{5} = 5.1$$

x	$x-\bar{x}$	$(x-\bar{x})^2$
5.8	$5.8-5.1=0.7$	$(0.7)^2=0.49$
4.6	$4.6-5.1=-0.5$	$(-0.5)^2=0.25$
5.3	$5.3-5.1=0.2$	$(0.2)^2=0.04$
3.8	$3.8-5.1=-1.3$	$(-1.3)^2=1.69$
6.0	$6.0-5.1=0.9$	$(0.9)^2=0.81$
		$\sum(x-\bar{x})^2=3.28$

Sample variance : $$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{3.28}{5-1} = \frac{3.28}{4} = 0.82$$

Sample standard deviation : $$s = \sqrt{\text{variance}} \qquad s = \sqrt{0.82} \approx 0.91$$

Example 4 : Use the computational formula to find the standard deviation of

5.8, 4.6, 5.3, 3.8, 6.0

x	x^2
5.8	33.64
4.6	21.16
5.3	28.09
3.8	14.44
6.0	36
$\sum x = 25.5$	$\sum x^2 = 133.33$

$$s^2 = \frac{\sum x^2 - [(\sum x)^2 / n]}{(n-1)}$$

$$= \frac{133.33 - (25.5)^2 / 5}{5-1}$$

$$= \frac{3.28}{4} = 0.82$$

$$s = \sqrt{0.82} = 0.91$$

Note : Both the mean and standard deviation are sensitive to extreme observations called the outliers. The standard deviation is used to describe variability when the mean is used as a measure of central tendency.

II. Variance and standard deviation for grouped data

The formula is similar to the computational formula of s^2 for a data set is

$$s^2 = \frac{\sum f \cdot x_m^2 - \left[\left(\sum f \cdot x_m\right)^2 / n\right]}{n-1}$$

Example 1:

These data represent the net worth (in millions of dollars) of 50 businesses in a large city. Find the variance and standard deviation.

Class Limit	Frequency f	Midpoint (x_m)	$f \cdot x_m$	$f \cdot x_m^2$
$10-20$	5	$\frac{10+20}{2}=15$	$5\cdot 15$ $=75$	$5\cdot 15^2$ $=1,125$
$21-31$	10	$\frac{21+31}{2}=26$	$10\cdot 26$ $=260$	$10\cdot 26^2$ $=6,760$
$32-42$	3	37	111	$4,107$
$43-53$	7	48	336	$16,128$
$54-64$	18	59	$1,062$	$62,658$
$65-75$	7	70	490	$34,300$
	$n=50$		$\sum=2,334$	$\sum=125,078$

Sample variance :

$$s^2 = \frac{\sum f \cdot x_m^2 - \left[\left(\sum f \cdot x_m \right)^2 / n \right]}{n-1}$$

$$= \frac{125{,}078 - \left[(2{,}334)^2 / 50 \right]}{50-1}$$

$$= \frac{125{,}078 - 108{,}951.12}{49} = 329.12$$

Sample standard deviation :

$$s = \sqrt{\text{variance}}$$

$$s = \sqrt{329.12} \approx 18.14$$

III. Coefficient of variation

The coefficient of variation is a measure of relative variability that expresses standard deviation as a percentage of the mean.

$$CVar = \frac{s}{\bar{x}} \cdot 100\%$$

When comparing the standard deviations of two different variables, the coefficient of variations are used.

Example 1:

The average score on an English final examination was 85, with a standard deviation of 5; the average score on a history final exam was 110, with a standard deviation of 8. Compare the variations of the two.

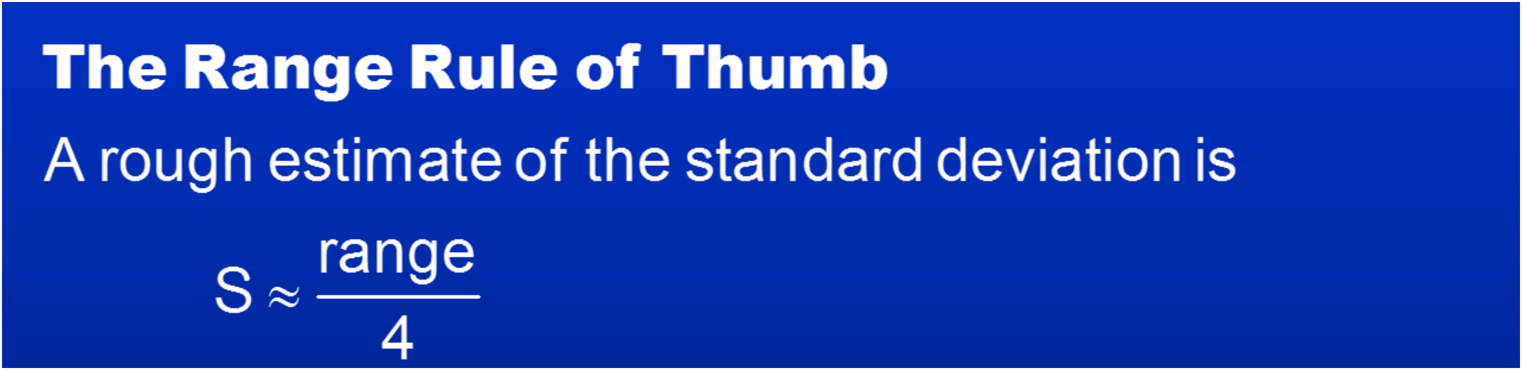
$$CVar = \frac{s}{\overline{x}} \cdot 100\% = \frac{5}{85} \cdot 100\% = 5.9\%$$

$$CVar = \frac{s}{\overline{x}} \cdot 100\% = \frac{8}{110} \cdot 100\% = 7.3\%$$

→ The average score on the history final exam was more variable than the average score on the English final exam.

IV. Range Rule of Thumb

The range can be used to approximate the standard deviation. This approximation is called the range rule of thumb.



The Range Rule of Thumb

A rough estimate of the standard deviation is

$$S \approx \frac{\text{range}}{4}$$

Example: Using the range rule of thumb, approximate the standard deviation for the data set 5, 8, 8, 9, 10, 12, and 13.

$$S \approx \frac{\text{range}}{4} = \frac{13-5}{4} = \frac{8}{4} = 2$$

Note: The range rule of thumb is only an *approximation* and should be used when the distribution of data is unimodal and roughly symmetric.

Measures of Variation: Range Rule of Thumb

Use $\bar{X}-2s$ to approximate the lowest value and $\bar{X}+2s$ to approximate the highest value in a data set.

$$\text{Example: } \bar{X}=10,\ Range=12$$

$$s\approx\frac{12}{4}=3 \qquad \begin{array}{l} LOW\approx 10-2(3)=\boxed{4} \\ HIGH\approx 10+2(3)=\boxed{16} \end{array}$$

V. Chebyshev’s Theorem and Empirical Rule

Chebyshev’s theorem (Any distribution shape)

The proportion of values from a data set that will fall within *k* standard deviation of the mean will be at least 1- 1 / k^2, where k is a number greater than 1.

Using Chebyshev's to find range

$$\text{Large}\,\# = \text{mean} + k \cdot s$$

$$\text{Small}\,\# = \text{mean} - k \cdot s$$

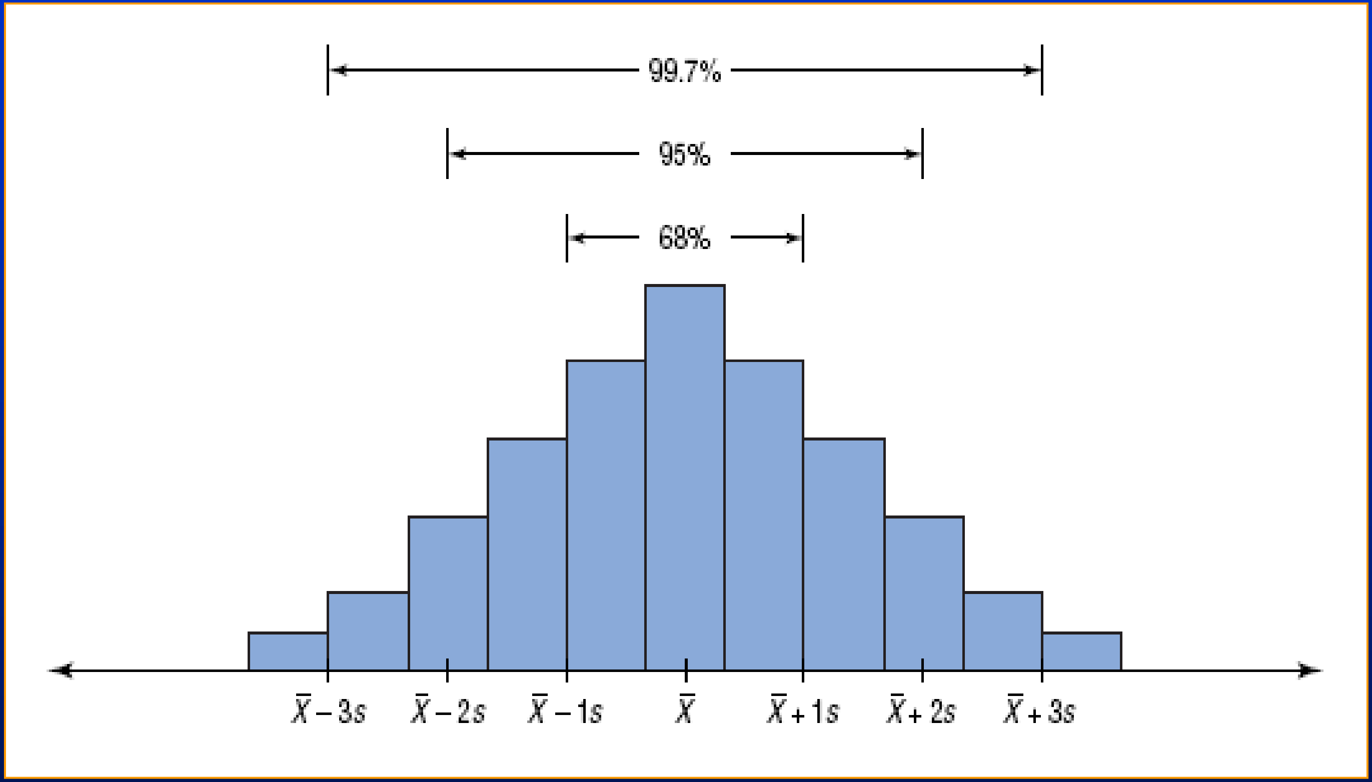
Empirical Rule (A bell-shaped distribution)

Approximately 68% of the data values will fall within 1 standard deviation of the mean.

Approximately 95% of the data values will fall within 2 standard deviations of the mean.

Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

The Empirical Rule



Example 1:

The average U.S. yearly per capita consumption of citrus fruit is 26.8 pounds. Suppose that the distribution of fruits amounts consumed is bell-shaped with a standard deviation equal to 4.2 pounds. What percentage of Americans would you expect to consume in the range of 18.4 pounds to 35.2 pounds of citrus fruit per year?

$\bar{x} = 26.8 \qquad s = 4.2 \qquad \% = ?$

$$k = \frac{\text{Large\#} - \text{Mean}}{\text{Stardard Deviation}} = \frac{35.2 - 26.8}{4.2} = 2$$

Since the data is a bell-shaped curve, Empirical Rule is used. According to the Empirical Rule, 95% of the data fall within 2 standard deviation.

Example 2:

Using the Chebyshev’s theorem, solve these problems for a distribution with a mean of 50 and a standard deviation of 5. At least what percentage of the values will fall between 40 and 60?

$\bar{x} = 50$ $s = 5$ Range = (40,60) % = ?

$$k = \frac{\text{Large\#} \ - \ \text{Mean}}{\text{Stardard Deviation}} = \frac{60-50}{5} = 2$$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2}$$

$$= 1 - \frac{1}{4} = 0.75$$

At least 75% of the values will fall between 40 and 60.

Example 3:

A sample of the labor costs per hour to assemble a certain product has a mean of \$2.60 and a standard deviation of \$0.15. Using Chebyshev’s theorem, find the range in which at least 88.89% of the data will lie.

$\bar{x} = 2.60$ $s = 0.15$ Range = ? % = 88.89

$$1 - \frac{1}{k^2} = 88.89\%$$

$$k = \sqrt{\frac{1}{1 - 0.8889}} = 3$$

$$1 - \frac{1}{k^2} = 0.8889$$

$$\frac{1}{k^2} = 1 - 0.8889$$

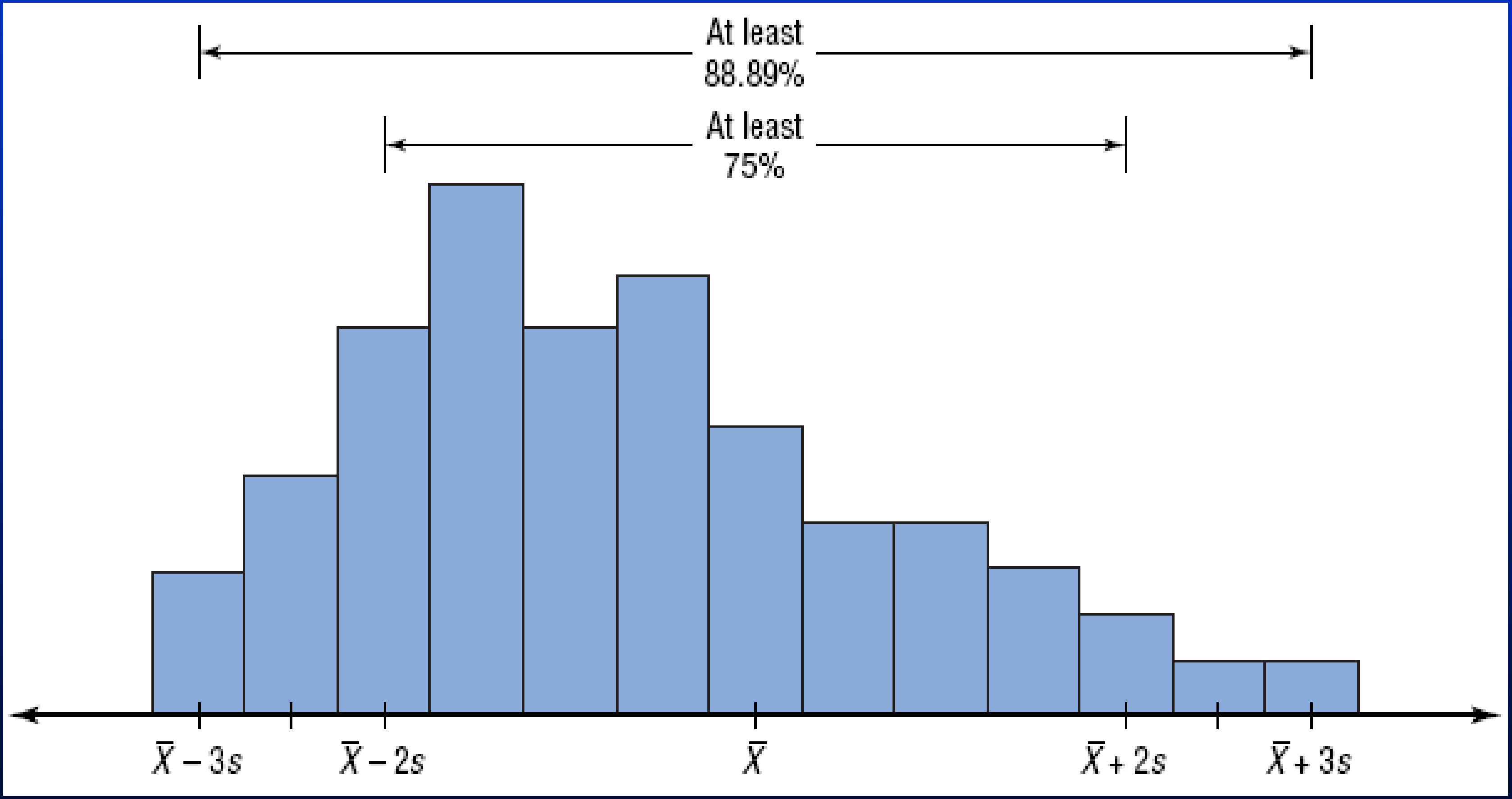
$$k^2 = \frac{1}{1 - 0.8889}$$

$$\bar{x} \pm k \cdot s$$

$$2.60 \pm 3 \cdot 0.15$$

$$= (2.15,\ 3.05)$$

Measures of Variation: Chebyshev's Theorem



Section 3.3 Measures of Position

I. z score

A z score represents the number of standard deviations that a data value lies above or below the mean.

$$z = \frac{x - \bar{x}}{s}$$

Example 1:

Which of these exam grades has a better relative position?

(a) A grade of 56 on a test with $\bar{x} = 48$ and s = 5.

(b) A grade of 220 on a test with $\bar{x} = 200$ and s = 10.

(a) $$z = \frac{x - \bar{x}}{s} = \frac{56 - 48}{5} = 1.6$$

(b) $$z = \frac{x - \bar{x}}{s} = \frac{220 - 200}{10} = 2$$

Part (b) has a better relative position.

Example 2:

Human body temperature have a mean of 98.20° and a standard deviation of 0.62 °. An emergency room patient is found to have a temperature of 101°. Convert 101° to a z score. Consider a data to be extremely unusual if its z score is less than -3.00 or greater than 3.00. Is that temperature unusually high? What does it suggest?

$$\bar{x} = 98.2 \qquad s = 0.62 \qquad x = 101$$

$$z = \frac{x - \bar{x}}{s}$$

$$z = \frac{101 - 98.2}{0.62}$$

$$\approx 4.52$$

Yes, the temperature is unusually high. It suggests that the patient has a fever.

II. Percentiles and Quartiles

Percentile Formula (Percentile Rank)

The Percentile corresponding to a given value *x* is computed by using the following formula:

$$Percentile = \frac{(number\ of\ \ values\ below\ x) + 0.5}{total\ \ number\ of\ \ values} \cdot 100\%$$

Example 1:

Find the percentile rank for each test score in the data set.

5, 15, 21, 16, 20, 12

Reorder: 5, 12, 15, 16, 20, 21 n = 6

For 5: $$\frac{0+0.5}{6}\cdot 100 \approx 8.33\% \quad or \quad 8th\ percentile$$

For 12: $$\frac{1+0.5}{6}\cdot 100 \approx 25\% \quad or \quad 25th\ percentile$$

For 15: $$\frac{2+0.5}{6}\cdot 100 \approx 41.7\% \quad or \quad 42th\ percentile$$

For 16: $$\frac{3+0.5}{6}\cdot 100 \approx 58.3\% \quad or \quad 58th\ percentile$$

For 20: $$\frac{4+0.5}{6}\cdot 100 \approx 75\% \quad or \quad 75th\ percentile$$

For 21: $$\frac{5+0.5}{6}\cdot 100 \approx 91.7\% \quad or \quad 92th\ percentile$$

Formula for finding a value corresponding to a given percentile (P_m)

P_m – is the number that separates the bottom *m*% of the data from the top (100 – m)% of that data.

e.g. If your test score represented 90th percentile means that 90% of the people who took the test scored lower than you and only 10% scored higher than you.

Finding the location of P_m :

Evaluate $(\frac{m}{100})n$

1. If $(\frac{m}{100})n$ is a whole number, then location of P_m is $\left[(\frac{m}{100})n+0.5\right]$.

   The percentile of P_m is halfway between the data value in position $(\frac{m}{100})n$ and the data value in the next position.

2. If $(\frac{m}{100})n$ is not a whole number, then location of P_m is the next higher whole number.

The percentile of P_m is the data value in this location.

Quartiles are defined as follows :

The first Quartile $Q_1 = P_{25}$

The second Quartile $Q_2 = P_{50}$

The third Quartile $Q_3 = P_{75}$

Example 2 : The number of home runs hit by the American League home rum leaders in the year 1959 – 1998. These ordered data are

22 32 32 32 32 33 36 36 37 39 39 39 40 40

40 40 41 42 42 43 43 44 44 44 44 45 45 46

46 48 49 49 49 49 50 51 52 56 56 61

Find the following :

(a) P_{77}

What is n ? $n = 40$

Location of P_{77}

What is 77% of 40? $30.8 \approx 31$

Count on reordered data until 31st data to get Answer. $P_{77} = 49$

(b) P_{42}

What is 42% of 40? 16.8 → 17th location

$P_{42} = 41$

(c) Q_1

$Q_1 = P_{25}$

What is 25% of 40?

$0.25 \cdot 40 = 10$ → change to location 10.5

$Q_1 = (39 + 39) / 2 = 39$

(d) Q_3

$Q_3 = P_{75}$

$0.75 \cdot 40 = 30$ → change to location 30.5

$Q_3 = (48 + 49) / 2 = 48.5$

III. The Interquartile Range

The interquartile range, or IQR

$IQR = Q_3 – Q_1$

The interquartile range is not influenced by extreme observations. If the median is used as a measure of central tendency, then the interquartile range should be used to describe variability.

Identifying Outliers (extremely high or low data value)

Any data value is smaller than $Q_1 – 1.5 \cdot IQR$ or larger than $Q_3 + 1.5 \cdot IQR$

is considered as an outlier.

The quick way to find Q_1 and Q_3:

Find the median of the data values that fall below Q_2 is Q_1.

Find the median of the data values that fall above Q_2 is Q_3.

Example 1 : Consider the following ranked data:

.09 .14 .25 .37 .55 .55 .56 .60 .77

.77 .86 .93 1.15 1.34 1.41 1.75 2.01 2.23 $n = 23$

3.69 3.90 4.50 4.88 7.79

(a) Find the interquartile range

$Q_1 = P_{25}$

$0.25 \cdot 23 = 5.75$

Position 6th

$Q_1 = 0.55$

$Q_3 = P_{75}$

$0.75 \cdot 23 = 17.25$

Position 18th

$Q_3 = 2.23$

$IQR = Q_3 – Q_1$

$= 2.23 – 0.55$

$IQR = 1.68$

(b) Is 7.79 an outlier?

$Q_1 \rightarrow \; Q_1 – 1.5 \cdot IQR = \; 0.55 – 1.5 \cdot 1.68 = -1.97$

$Q_3 \rightarrow \; Q_3 + 1.5 \cdot IQR = \; 2.23 + 1.5 \cdot 1.68 = 4.75$

(-1.97, 4.75)

Yes, 7.79 is an outlier since it falls outside the interval.

Example 2 : Check the following data set for outliers.

145 119 122 118 125 100

Reorder : 100 118 119 122 125 145 $n = 6$

Step 1: Find the interquartile range

$Q_1 = P_{25}$	$Q_3 = P_{75}$	$IQR = Q_3 – Q_1$
$0.25 \cdot 6 = 1.5$	$0.75 \cdot 6 = 4.5$	$= 125 – 118$
Position 2nd	Position 5th	$IQR = 7$
$Q_1 = 118$	$Q_3 = 125$	

Step 2 : Is there any outlier for the data set?

$Q_1 \rightarrow Q_1 – 1.5 \cdot IQR = 118 – 1.5 \cdot 7 = 107.5$

$Q_3 \rightarrow Q_3 + 1.5 \cdot IQR = 125 + 1.5 \cdot 7 = 135.5$

(107.5, 135.5)

Yes, 100 and 145 are outliers since they fall outside the interval.

Section 3.4 Explotory Data Ananalysis

I. Boxplot

The median and the interquartile range are used to describe the distribution using a graph called boxplot. From a boxplot, we can detect any skewness in the shape of the distribution and identify any outliers in the data set.

Find the 5-number summary consisting of the Low, Q_1, Q_2, Q_3, and High.

Construct a scale with values that include the Low and High.

Construct a box with two vertical sides called the hinges above Q_1 and Q_3 on the axis.

Also construct a vertical line in the box above Q_2.

Finally, connect the Low and High to the hinges using horizontal lines called the whiskers.

Example 1 :

Construct a boxplot for the number of calculators sold during a randomly selected week.

8, 12, 23, 5, 9, 15, 3

Reorder : 3, 5, 8, 9, 12, 15, 23 n = 7

Low = 3

$Q_1 = P_{25}$

$0.25 \cdot 7 = 1.75$

Position 2

$Q_1 = 5$

$Q_2 = P_{50}$

$0.50 \cdot 7 = 3.5$

Position 4

$Q_2 = 9$

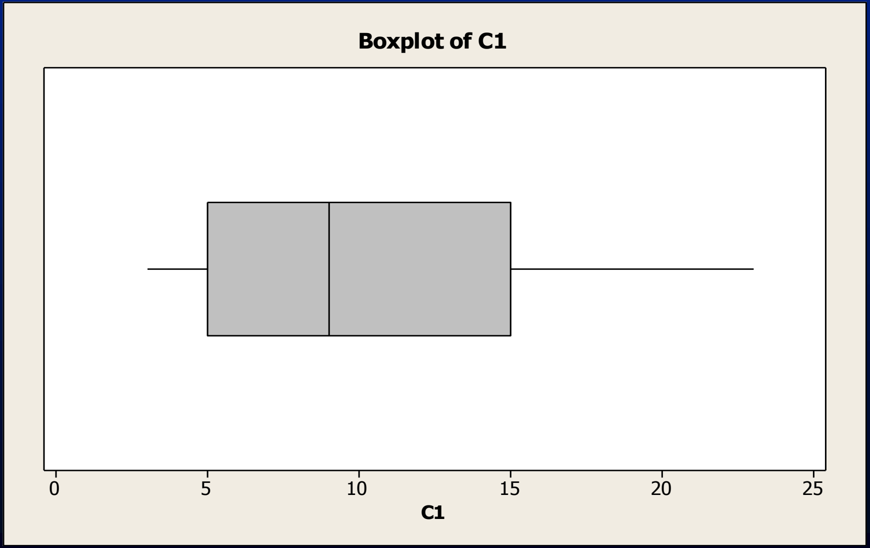
$Q_3 = P_{75}$

$0.75 \cdot 7 = 5.25$

Position 6

$Q_3 = 15$

High = 23



It is skewed to the right

Example 2 :

The following ranked data represent the number of English-language Sunday newspaper in each of the 50 states.

2 3 3 4 4 4 4 4 5 6 6 6 7

7 7 8 10 11 11 11 12 12 13 14 14 14

15 15 16 16 16 16 16 16 18 18 19 21 21

23 27 31 35 37 38 39 40 44 62 85

$n = 50$

Low = 2

$Q_1 = P_{25}$

$0.25 \cdot 50 = 12.5$

Position 13

$Q_1 = 7$

$Q_2 = P_{50}$

$0.50 \cdot 50 = 25$

Position 25.5

$Q_2 = (14 + 14) / 2$

$= 14$

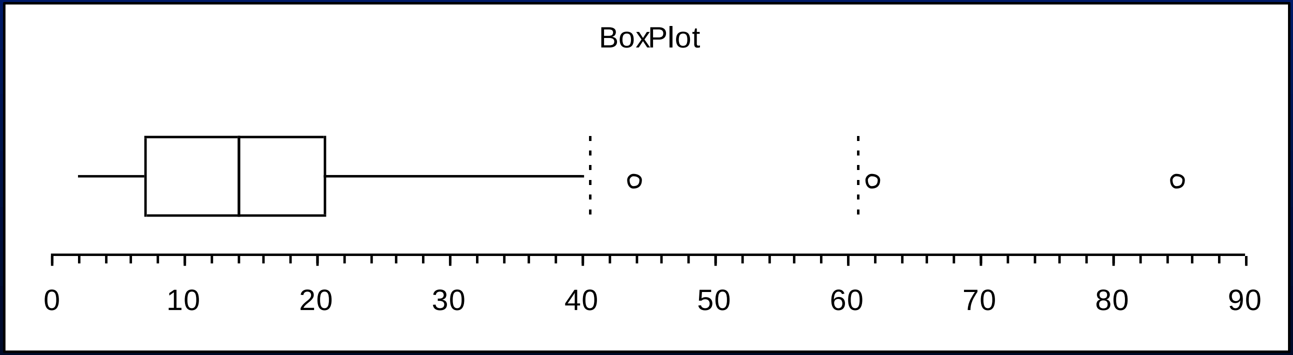
$Q_3 = P_{75}$

$0.75 \cdot 50 = 37.5$

Position 38

$Q_3 = 21$

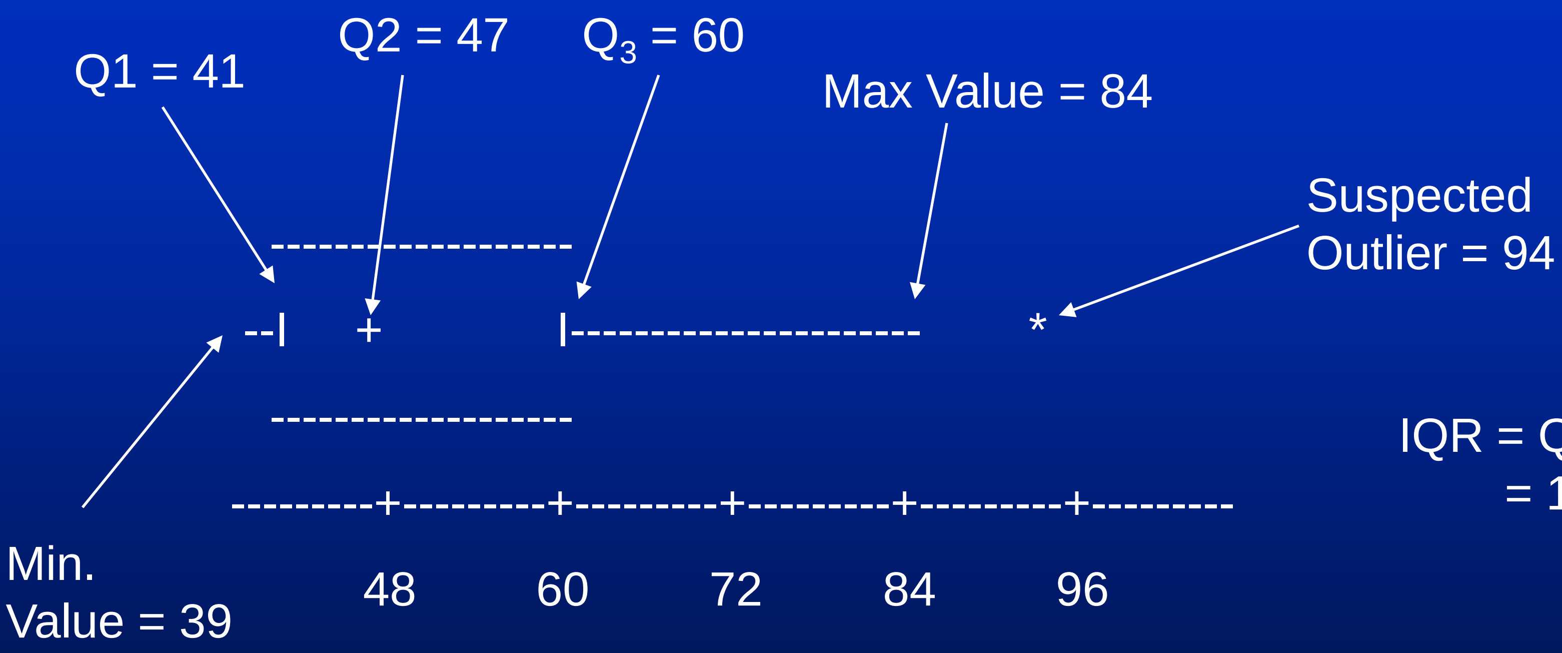
High = 85



It is skewed to the right.

Example 3 :

For the boxplot given below, (a) identify the maximum value, minimum value, first quartile, median, third quartile, and interquartile range; (b) comment on the shape of the distribution; (c) identify a suspected outlier



Q1 = 41

Q2 = 47

Q_3 = 60

Max Value = 84

Suspected Outlier = 94

```
    -------------------
  --I    +           I---------------------        *
    -------------------
---------+---------+---------+---------+---------+---------
        48        60        72        84        96
```

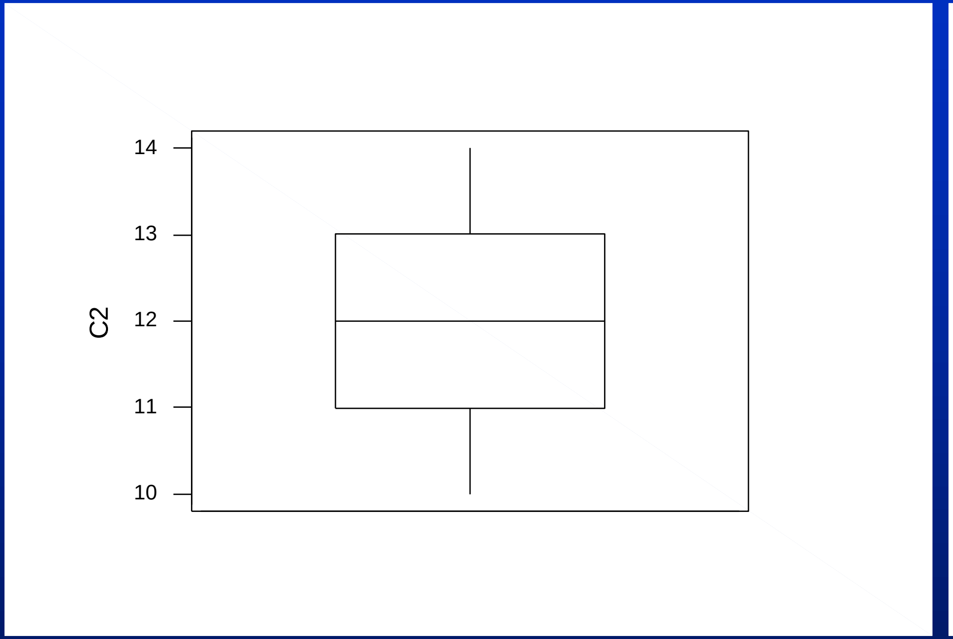
Min. Value = 39

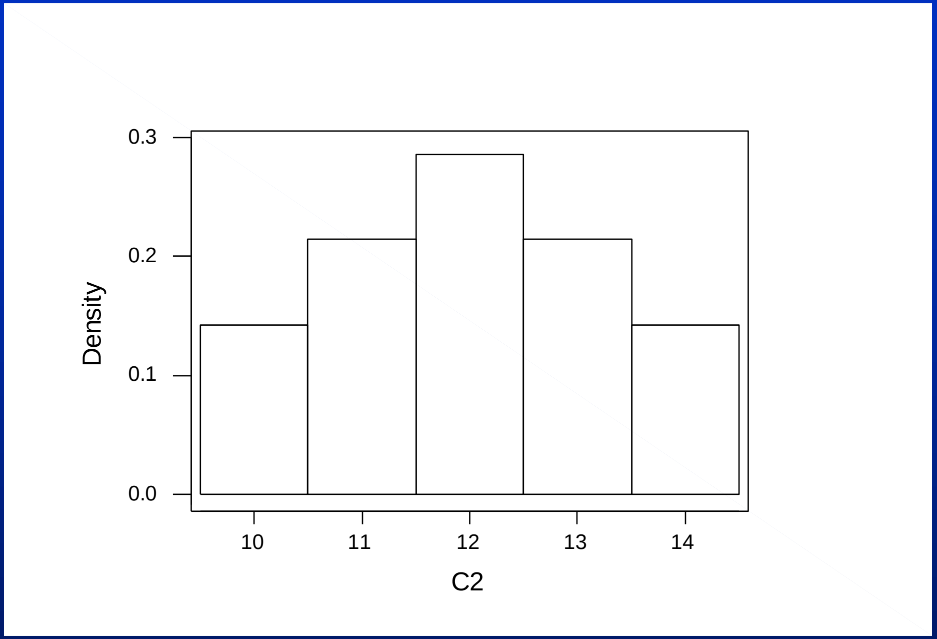
IQR = Q3 – Q1
= 19

The distribution is skewed to the right.

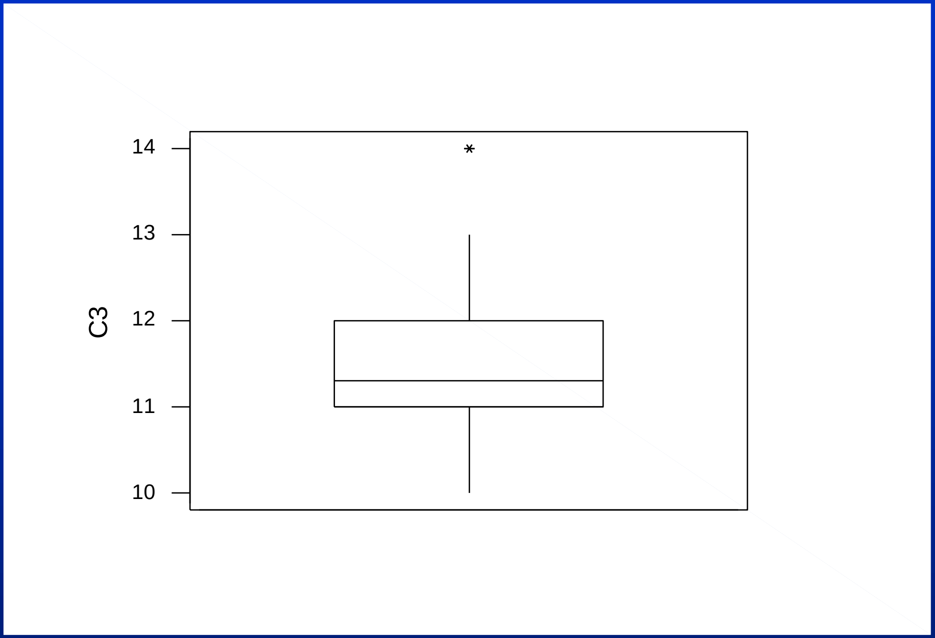
II. The distribution shape

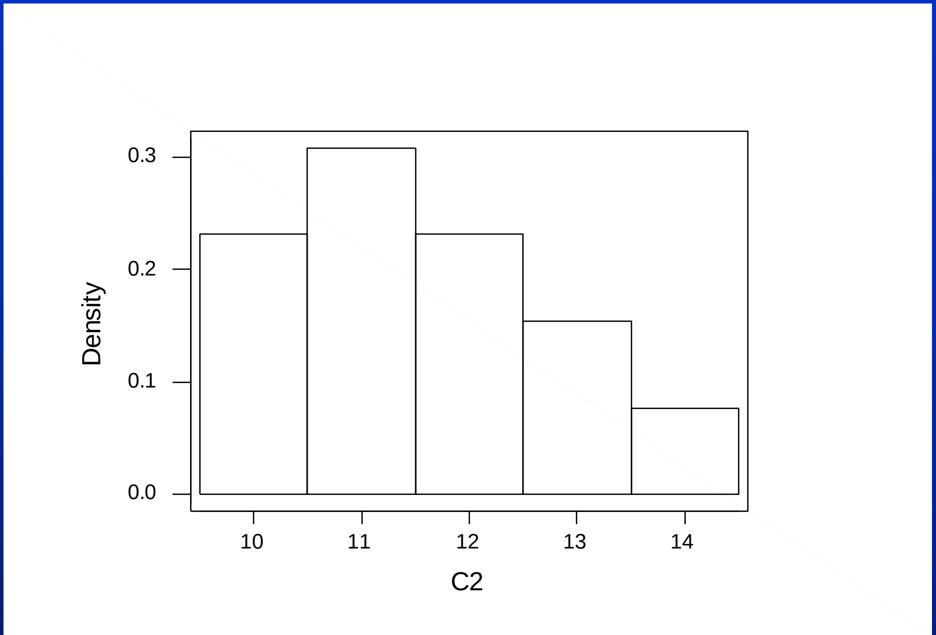
A bell-shaped distribution



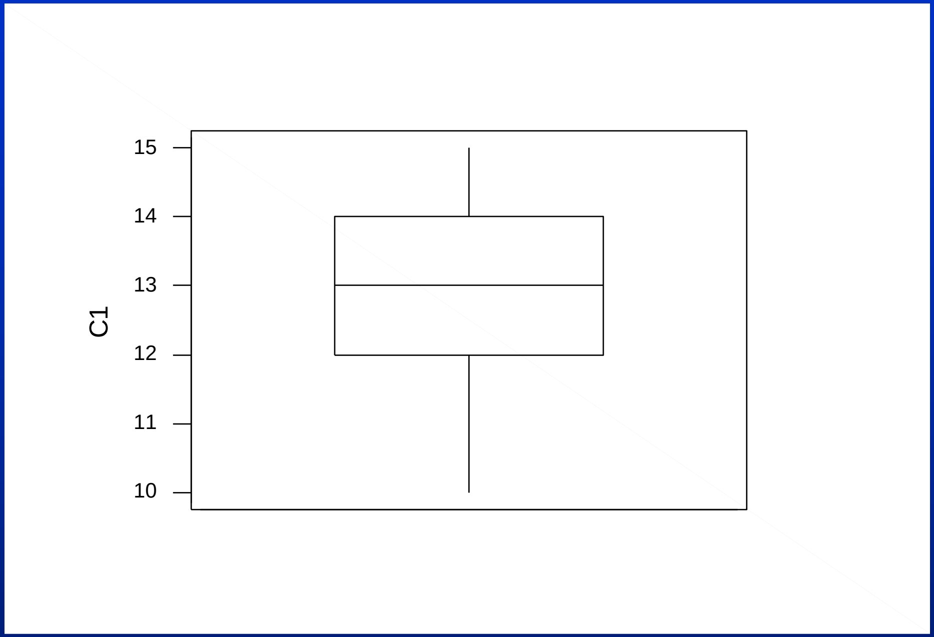


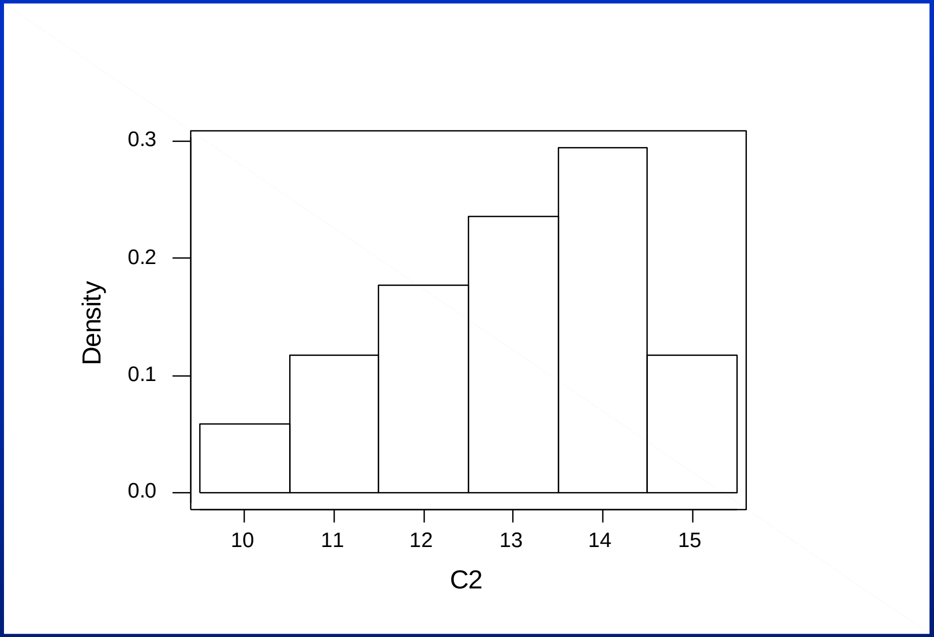
A skewed to the right distribution





A slightly skewed to the left distribution





Summary

- Some basic ways to summarize data include measures of central tendency, measures of variation or dispersion, and measures of position.
- The three most commonly used measures of central tendency are the mean, median, and mode. The midrange is also used to represent an average.

Summary (cont.)

- The three most commonly used measurements of variation are the range, variance, and standard deviation.
- The most common measures of position are percentiles, quartiles, and deciles.
- Data values are distributed according to Chebyshev’s theorem and in special cases, the empirical rule.