Basic Mathematics and Statistics

CHAPTER 12: NUMERICAL METHODS: MEASURE OF DISPERSION AND POSITION

Dr. Khaing S. Htun

12.1 Why study Dispersion

- Mean or median – only locate the center of the data
- **The spread of the data** *(spread of the distribution)*
- Clustered closely around arithmetic mean
- Large dispersion → mean is not reliable
- **To evaluate the reliability** of two or more averages

12.2 Measures of Dispersion for Ungrouped Data

Range

Range - simplest measure of dispersion

Range = Largest value – Smallest value

The major characteristics of the range are:

a. Only **two values are used** in its calculation.

b. It is **influenced by extreme values**.

c. It is **easy to compute** and to **understand**.

12.2 Measures of Dispersion for Ungrouped Data

Mean Absolute Deviation (MAD)

MAD - the arithmetic mean of the absolute values of the deviations from the arithmetic mean

1, 2, 5, 8, 9, 10 $\quad n=6$

$\bar{x} = \frac{2+4+5+8+9+10}{6} = \frac{38}{6} = 6.3$

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

where

x_i is the value of each observation.

$\bar{x}$ is the arithmetic mean of the data set

n is the number of observations in the sample.

The major characteristics of *MAD* are:

a. All values are used in the calculation.

b. It is **not unduly influenced by large or small values**.

c. It is **easy to compute** and to **understand**.

10 − 6.3

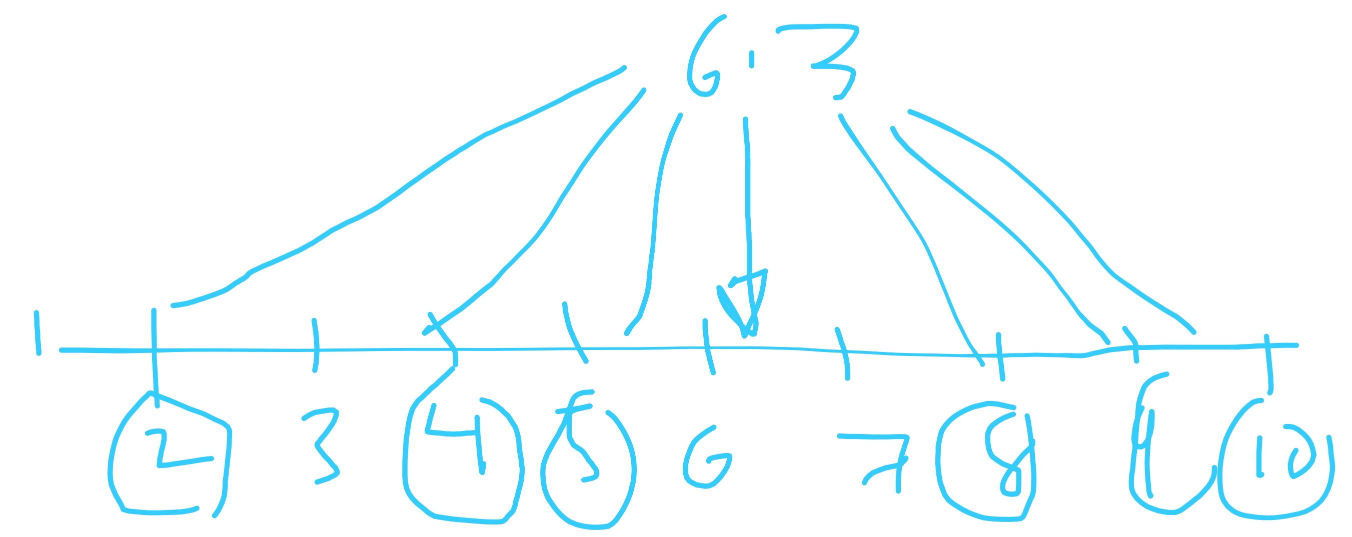
4

2

5

8

9



12.2 Measures of Dispersion for Ungrouped Data

Variance and Standard Deviation

Variance - the arithmetic mean of the squared deviations from the mean

The population variance:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}$$

The sample variance:

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

The major characteristics of the variance are:

a. **All values are used** in the calculation.

b. It is **not unduly influenced by extreme values**.

12.2 Measures of Dispersion for Ungrouped Data

Variance and Standard Deviation

Standard Deviation - the positive square root of the variance

The population standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

The sample standard deviation:

$$S = \sqrt{S^2}$$

12.2 Measures of Dispersion for Ungrouped Data

The Coefficient of Variation

COV

- a measure of relation dispersion
- useful for comparing distributions with different units

$$\text{Coefficient of Variation} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100$$

For population data:

$$CV = \frac{\sigma}{\mu} \times 100$$

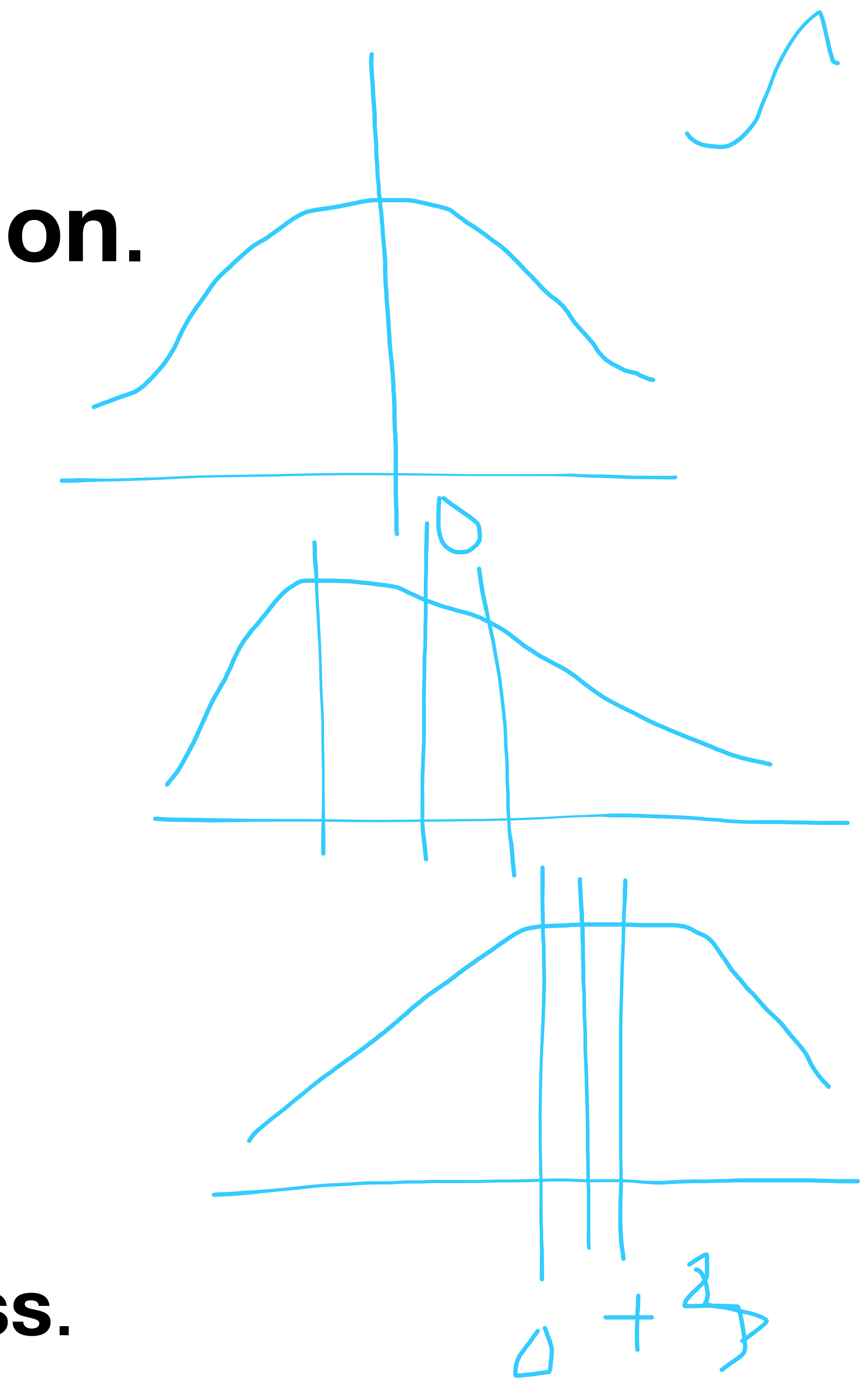
For sample data:

$$CV = \frac{S}{\bar{x}} \times 100$$

12.2 Measures of Dispersion for Ungrouped Data

The Coefficient of Skewness

The coefficient of skewness measures **the symmetry of a distribution.**



$$Sk = \frac{3(\bar{x} - median)}{s}$$

The major characteristics are:

a. It may range **from –3.00 up to 3.00**.

b. A value **near 0** means **the distribution is symmetric.**

c. A value **near 3** means **the distribution is positive skewness or right skewness.**

d. A value **near -3** means **the distribution is negative skewness or left skewness.**

12.3 Measures of Dispersion for Grouped Data

Range

Range - simplest measure of dispersion

Range = the upper limit of the largest class – the lowest limit of the smallest class

12.3 Measures of Dispersion for Grouped Data

Variance

Variance - the arithmetic mean of the squared deviations from the mean

The population variance:

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{\sum f_i M_i^2 - \frac{(\sum f_i M_i)^2}{N}}{N}$$

The sample variance:

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

$$S^2 = \frac{\sum f_i M_i^2 - \frac{(\sum f_i M_i)^2}{n}}{n-1}$$

where

M_i is the midpoint of each class.

f_i is the frequency in each class.

N is the total number of frequencies or the population size $N = \Sigma f$

n is the total number of frequencies or the sample size $\Sigma = f$

12.2 Measures of Dispersion for Grouped Data

Standard Deviation

Standard Deviation - the positive square root of the variance

The population standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

The sample standard deviation:

$$S = \sqrt{S^2}$$

12.4 Standard Score (or Z-score)

Z-score – making comparision between scores from two separate populations (from the same population as well) with different means and standard deviations

For population data:

$$z = \frac{x_i - \mu}{\sigma}$$

For sample data:

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

Standard scores (Z-score) is used with the **Empirical rule** and **Chebyshev's theorem** **to help identify unusual values within a given data set**.

12.4 Standard Score (or Z-score)

Chebyshev's Theorem

At least $\left(\mathbf{1}-\frac{\mathbf{1}}{\boldsymbol{z}^{\mathbf{2}}}\right)$ of the data values must be within z standard deviation of the mean.

where z is any value greater than 1

Some of the implications of this theorem, with z = 2, 3, and 4 standard deviations, follow.

- At least 0.75, or 75% of the data values must be within z = 2 standard deviations of the mean.
- At least 0.89, or 89% of the data values must be within z = 3 standard deviations of the mean.
- At least 0.94, or 94% of the data values must be within z = 4 standard deviations of the mean.

One of the advantages of Chebyshev's theorem is that it **applies to any data set** regardless of the shape of the distribution of the data.

12.4 Standard Score (or Z-score)

Chebyshev's Theorem

Assume that the midterm test scores for 100 students in a principles of statistics course had a mean of 70 and a standard deviation of 5. How many students had test scores between 58 and 82?

$$Z = \frac{x_i - \bar{x}}{s}$$

$z = \frac{58-70}{5} = -2.4$

below mean

$z = \frac{82-70}{5} = +2.4$

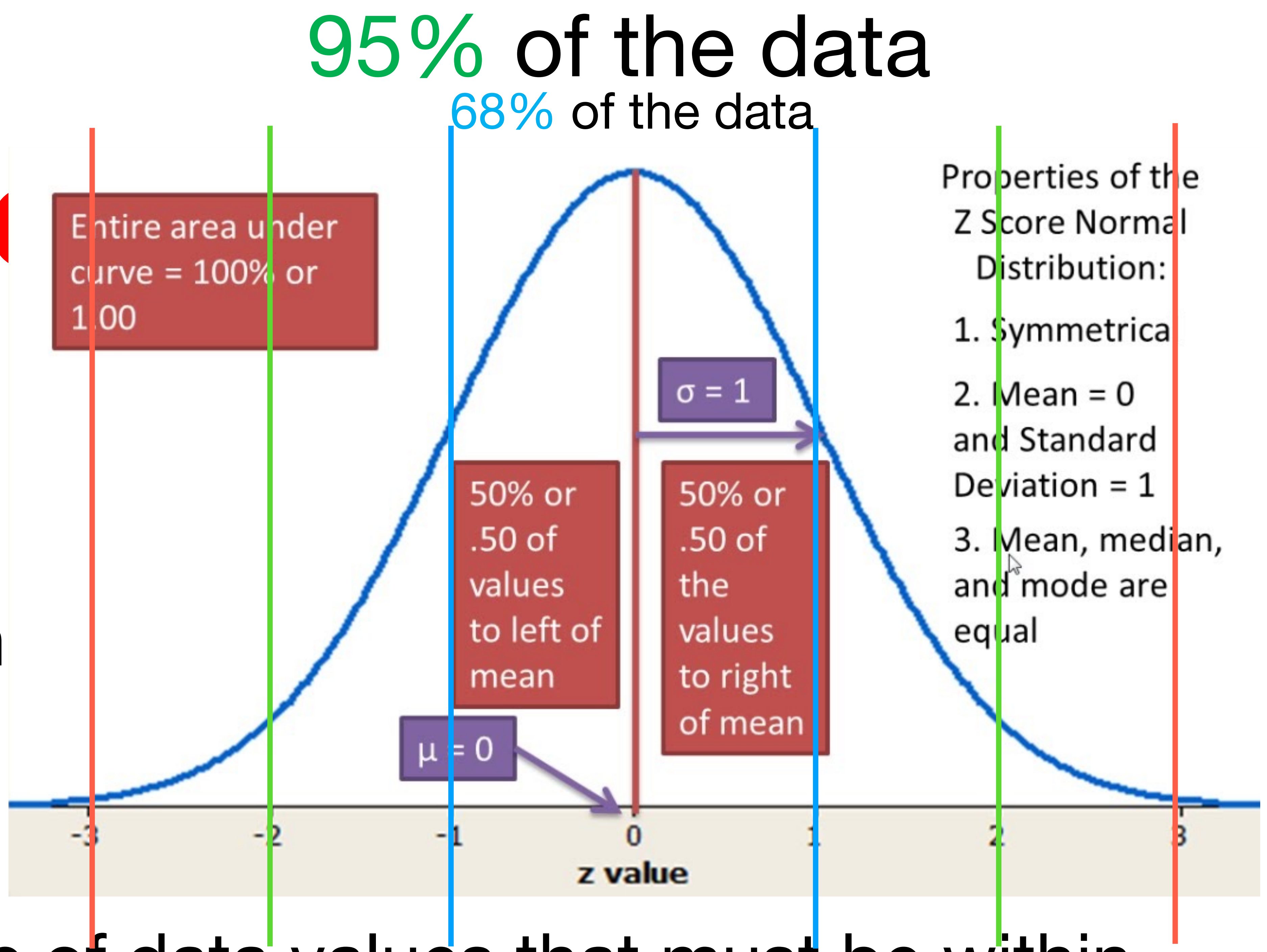
above mean

Chebyshev's Theorem→ $\left(1 - \frac{1}{z^2}\right) = \left(1 - \frac{1}{(2.4)^2}\right) = 0.826$

At least 82.6% of the students must have test scores between 58 and 82

12.4 Standard Score (or Z-s

Empirical Rule



Common shape - mound-shaped or bell-shaped distribution

The **empirical rule** can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.

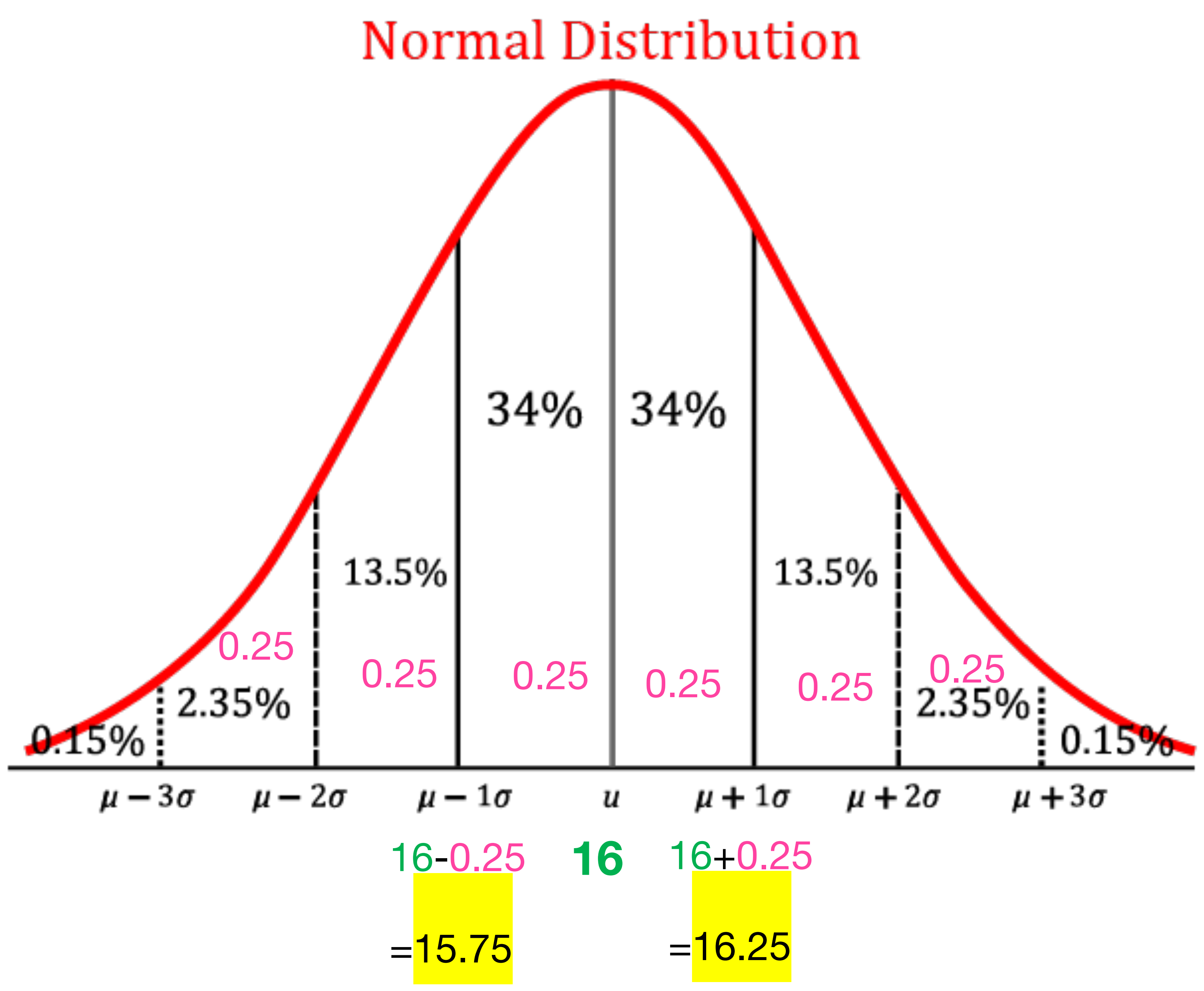
For data having a bell-shaped distribution:

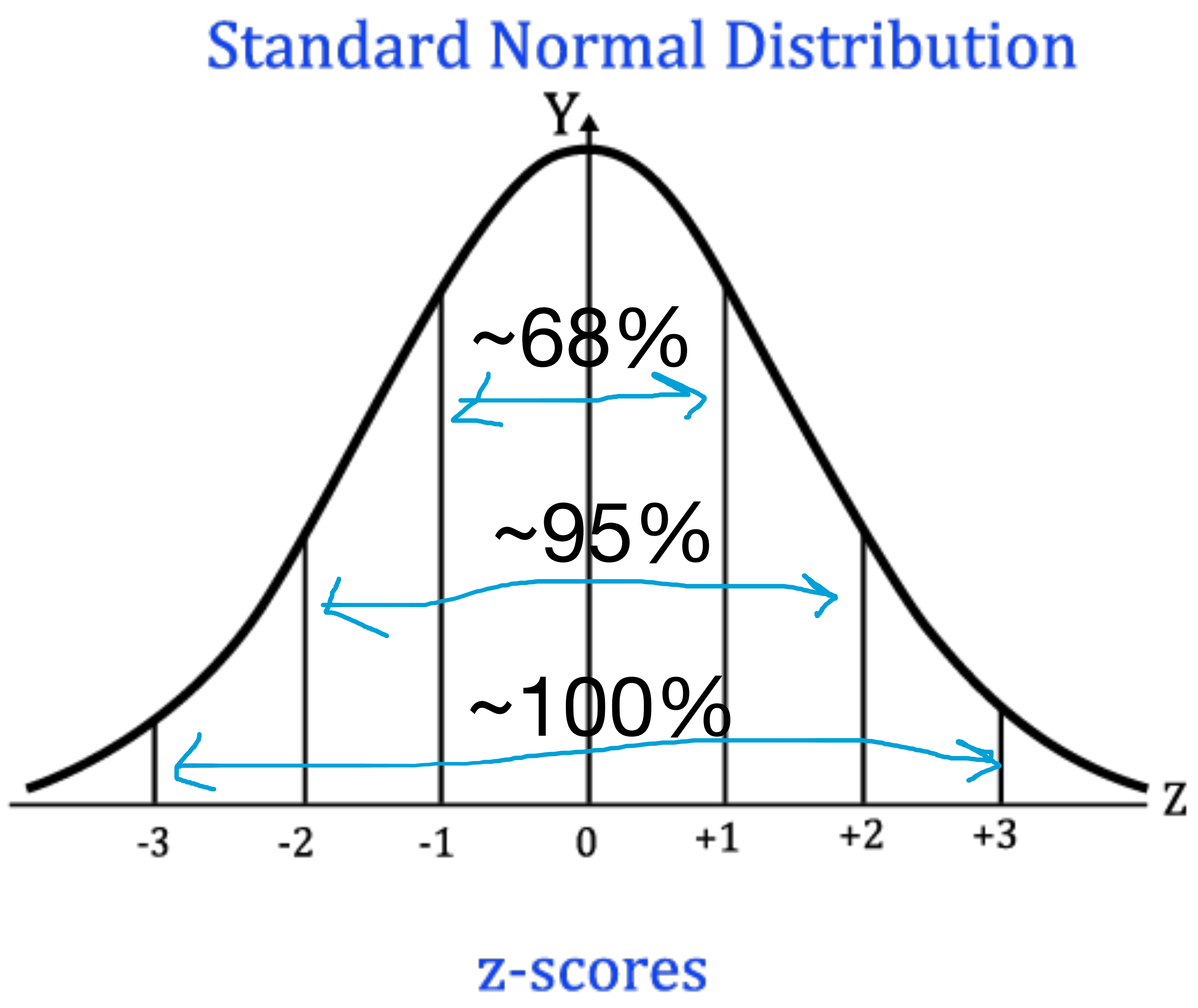
- Approximately 68% of the data values will be within one standard deviation of the mean.
- Approximately 95% of the data values will be within two standard deviations of the mean.
- Almost 100% of the data values will be within three standard deviations of the mean.

12.4 Standard Score (or Z-score)

Example

Liquid detergent cartons are filled automatically on a production line. Filling weights frequently have a bell-shaped distribution. If the mean filling weight is 16 ounces and the standard deviation is 0.25 ounces, we can use the empirical rule to draw the following conclusions.





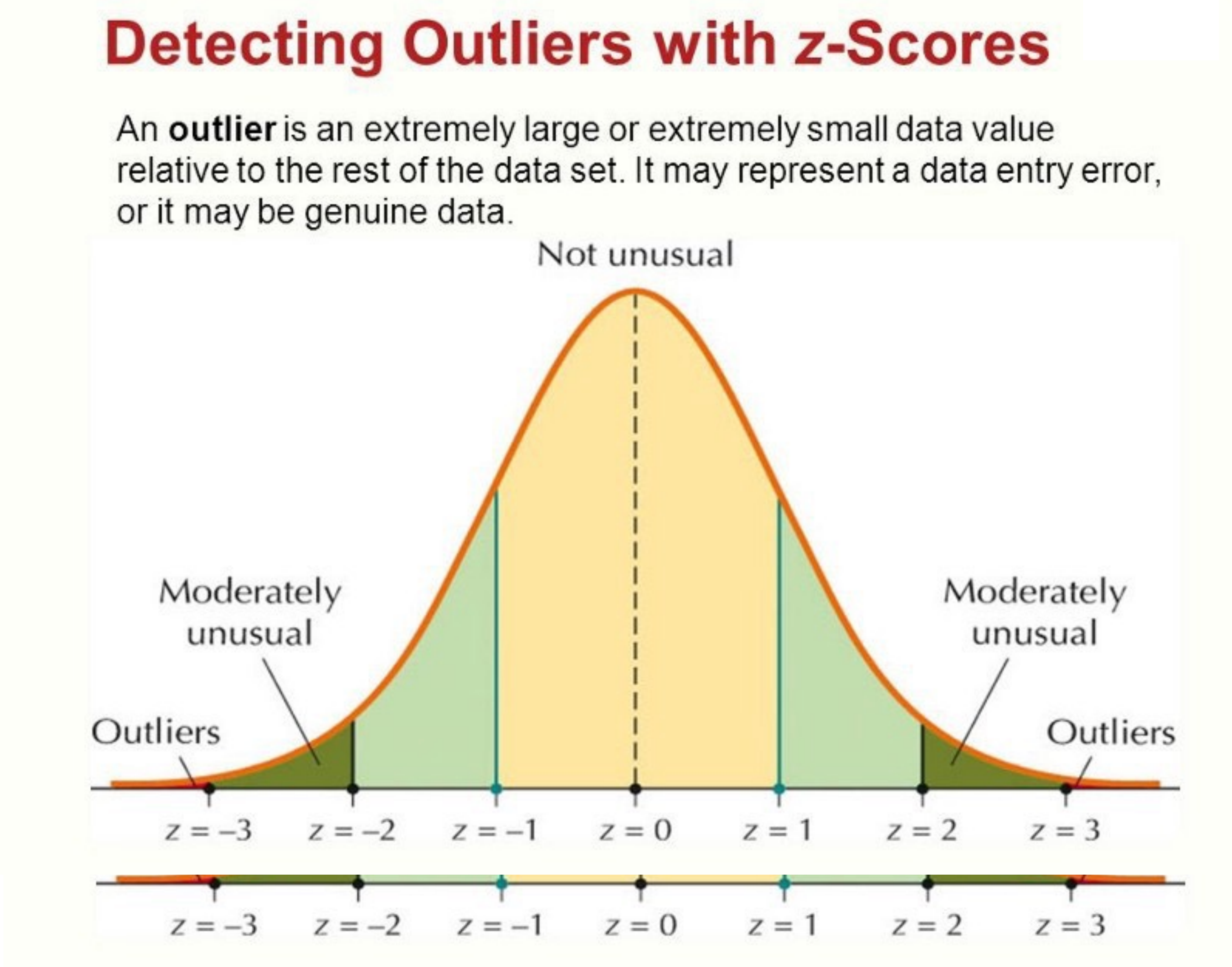
- Approximately 68% of the filled cartons will have weights between 15.75 and 16.25 ounces (that is, within one standard deviation of the mean).
- Approximately 95% of the filled cartons will have weights between 15.50 and 16.50 ounces (that is, within two standard deviations of the mean).
- Almost all filled cartons will have weights between 15.25 and 16.75 ounces (that is, within three standard deviations of the mean).

12.4 Standard Score (or Z-score)

Remark:

Z–scores can be used to identify **outliers,** a value that is inconsistent with the rest of data (sometimes it was called **extreme values**)

Outliers - any data value with at ***z*–score** less than **– 3** or greater than **+ 3**

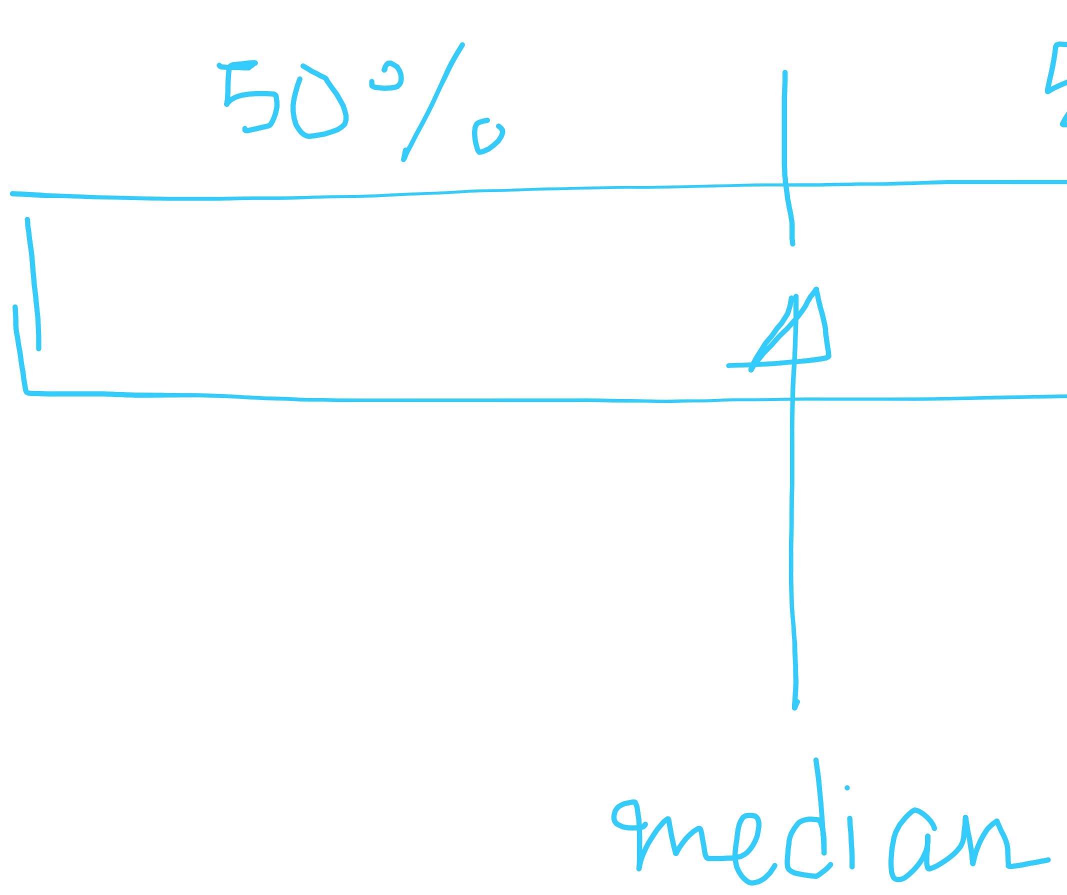


12.5 Measure of Position

Median

The median is the **midpoint in the data set** that has been **ranked in increasing order**.

It means that there is about 50% of the values in data set are smaller the median and about 50% are larger than the median.



1 2 3 4 5

2, 3, 6, 9, 10

$m = 6$

1 2 2.5 3 4

2 3 6 9

$m = \frac{3+6}{2} = 4.5$

12.5 Measure of Position

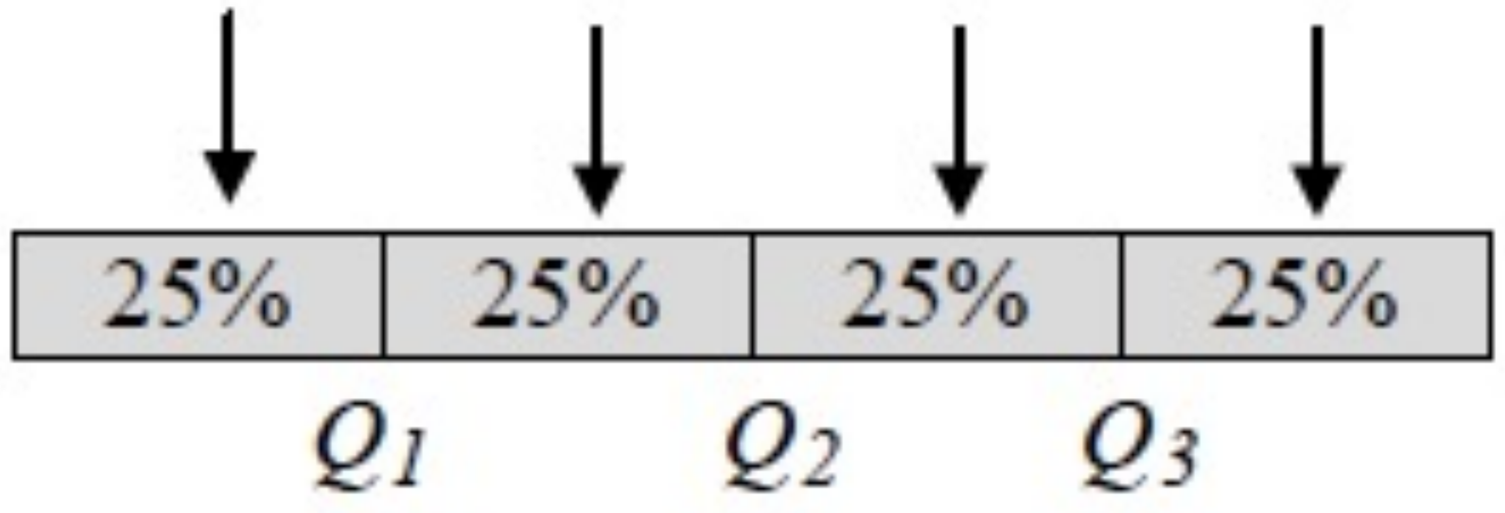
Quartiles

Quartiles is the measures that divide a ranked data set into **four** equal parts.

- The **first quartile** (Q_1) means that there is about 25% of the values in data set are smaller and about 75% are larger than the first quartile.

- The **second quartile** (Q_2) means that there is about 50% of the values in data set are smaller and about 50% are larger than the second quartile.

- The **third quartile** (Q_3) means that there is about 75% of the values in data set are smaller and about 25% are larger than the third quartile.



12.5 Measure of Position

Quartiles

The calculation of the quartiles:-

1. Rank the given data set in increasing order.
2. Find the position of $Q_q = \dfrac{(n+1)q}{4}$.

   That is position of $Q_1 = \dfrac{(n+1)}{4}$,

   position of $Q_2 = \dfrac{(n+1)2}{4} = \dfrac{(n+1)}{2}$,

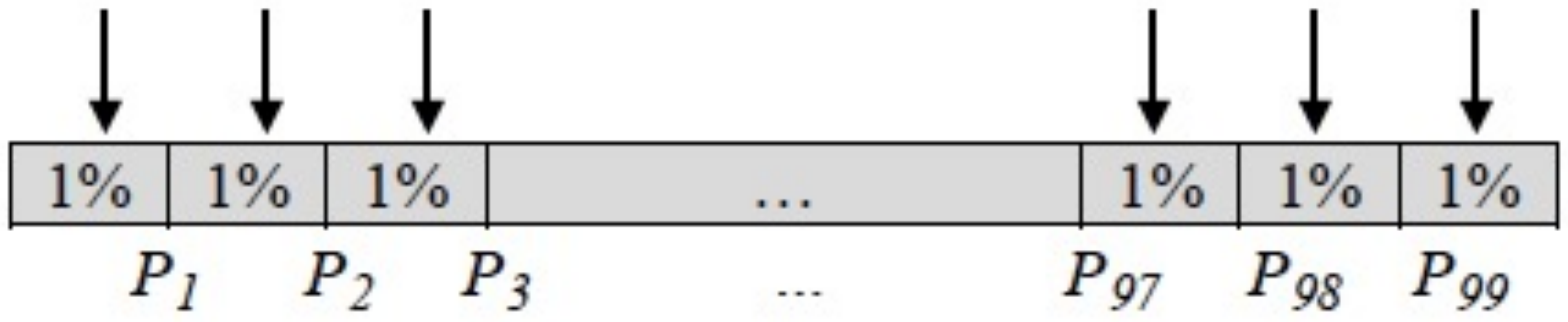
   position of $Q_3 = \dfrac{(n+1)3}{4}$.
3. Find the value of the quartile at the position Q_q.

12.5 Measure of Position

Percentiles

Percentiles is the measures that divide a ranked data set into **100** equal parts.

The ***p*****th percentile** is a value such that at least *p* percent (p%)of the observations are less than or equal to this value and at least (100-*p*) percent of the observations are greater than or equal to this value.



The calculation of the percentiles:-

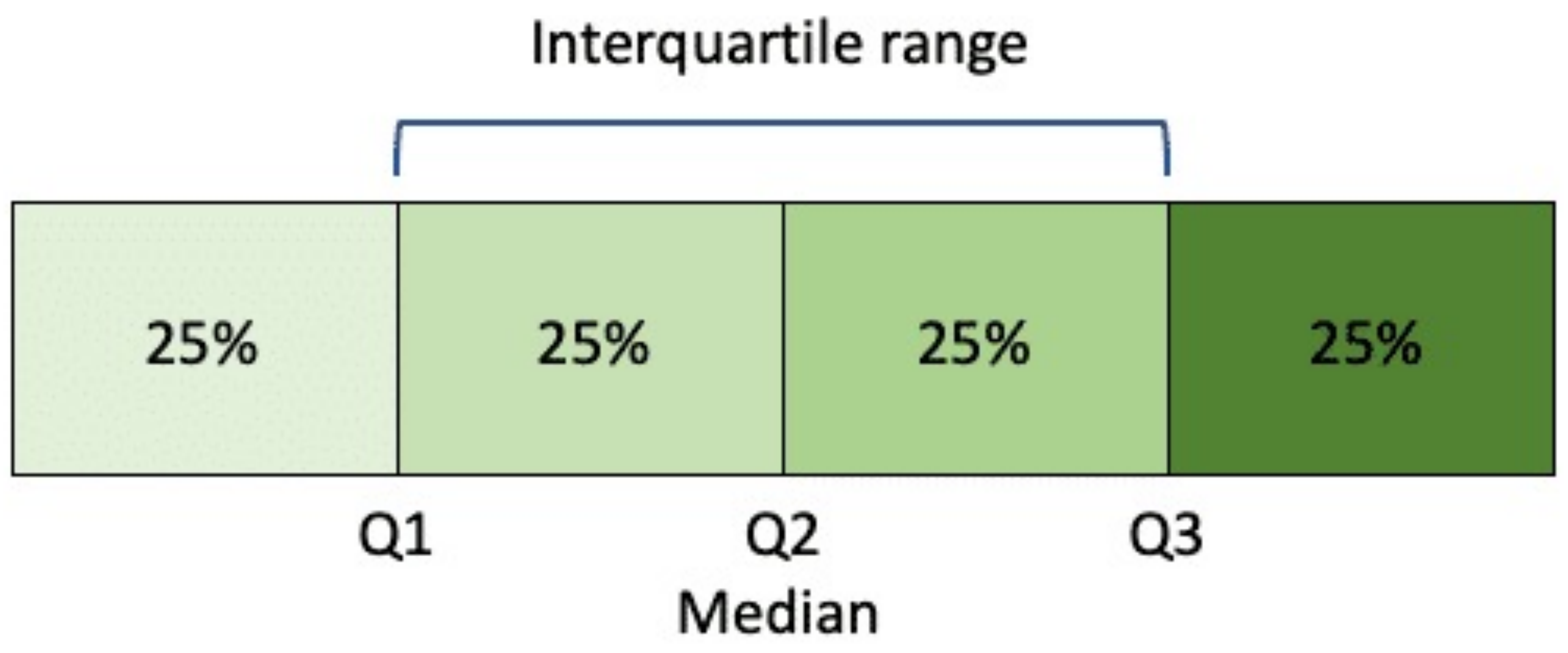
1. Rank the given data set in increasing order.
2. Find the position of $P_p = \frac{(n+1)p}{100}$.
3. Find the value of the quartile at the position P_p .

12.5 Measure of Position

Interquartile Range (IQR)

Interquartile Range (*IQR*) is the difference between the third and the first quartiles,

$$IQR = Q_3 - Q_1$$



12.5 Measure of Position

Example 12.7 The following are the scores of 12 students in a mathematics class (total scores = 100 marks). 75 80 68 53 99 58 76 73 85 88 91 79

a.) Compute the first quartile and the third quartile and interpret its value.

b.) Find the Interquartile range and interpret its value.

c.) Find the value of the 85^{th} percentile and interpret its value.

ranked data:	53	58	68	73	75	76	79	80	85	88	91	99
order :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

12.5 Measure of Position

a.) Compute the first quartile and the third quartile and interpret its value.

25% 69.25 25% 25% 25%

ranked data:	53	58	68	73	75	76	79	80	85	88	91	99
order :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

Q_1 Q_3

a)

the position of $Q_1 = \frac{(n+1)}{4} = \frac{(12+1)}{4} = 3.25$

the value of $Q_1 = 68 + 0.25(73-68) = 68 + 1.25 = 69.25$

It means that about 25% of the students in a mathematics class have the scores smaller than 69.25 marks and about 75% of the students of the students have the scores larger than 69.25 marks.

the position of $Q_3 = \frac{(n+1)3}{4} = \frac{(12+1)3}{4} = 9.75$

the value of $Q_3 = 85 + 0.75(88-85) = 85 + 2.25 = 87.25$

It means that about 75% of the students in a mathematics class have the scores smaller than 87.25 marks and about 25% of the students of the students have the scores larger than 87.25 marks.

12.5 Measure of Position

b.) Find the Interquartile range and interpret its value

$$IQR = Q_3 - Q_1 = 87.25 - 69.25 = 18$$

It means that the IQR of the mathematics scores is about 18 marks

91.4

85th

c.) Find the value of the 85th percentile and interpret its value. 85% 15%

ranked data:	53	58	68	73	75	76	79	80	85	88	91	99
order :	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

P_{85}

$$\text{the position of } P_{85} = \frac{(n+1)85}{100} = \frac{(12+1)85}{100} = 11.05$$

$$\text{the value of } P_{85} = 91 + 0.05(99 - 91) = 91 + 0.4 = 91.4$$

Thus, approximately 85% of the scores are less than 91.4 marks and 15% are greater than 91.4 marks.

12.6 A Box-Plot

A box plot is a graphic display of a set of data that shows the center, spread, and skewness of a data set

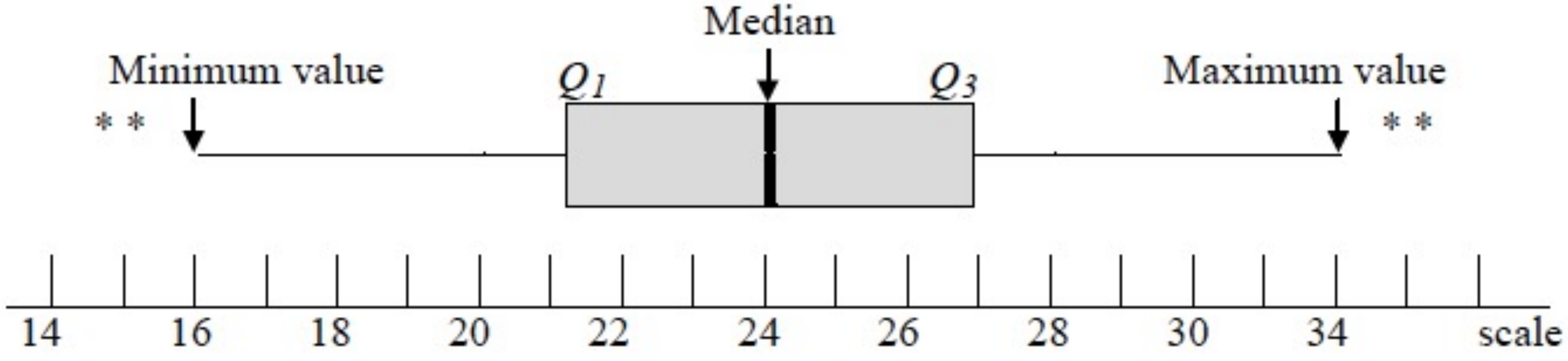


Figure 12.3 The box plot for a symmetric data set

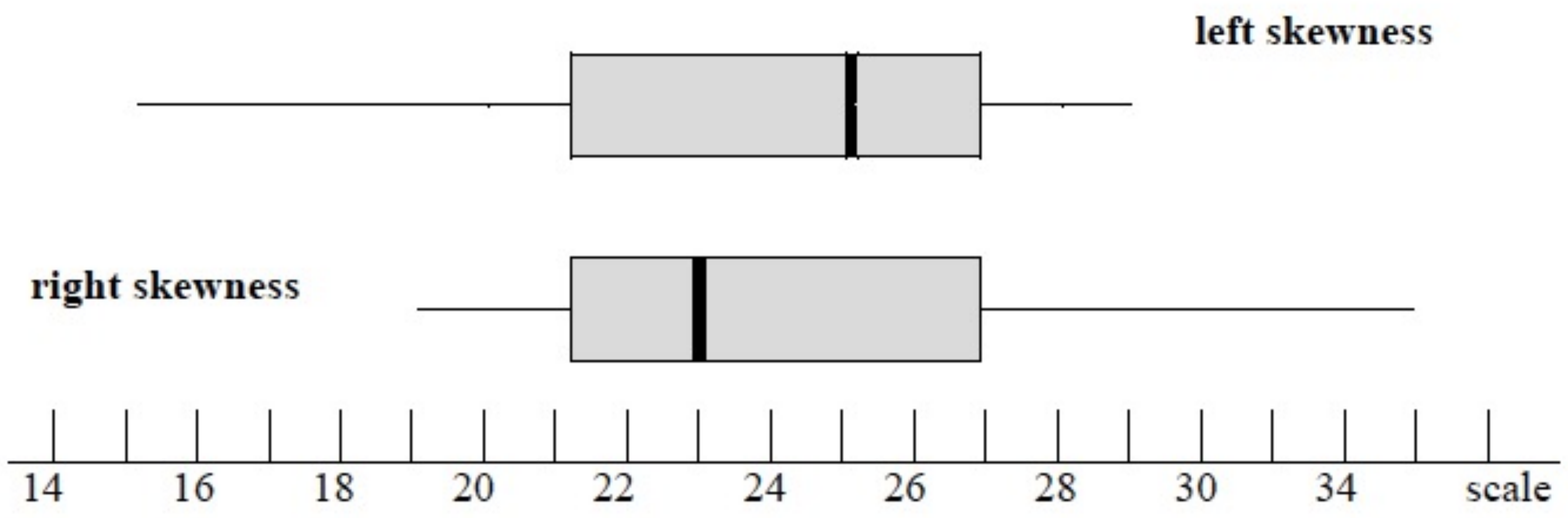


Figure 12.4 The box plot for skewed data set

12.6 A Box-Plot

- A box is drawn connecting the first and the third quartiles.

a) A line through the inside of the box shows the median.

b) The line segments, is called **whiskers**, from the third quartile to the largest value and from the first quartile to the smallest value shows the range of the largest 25% of the observations and the smallest 25%.

- A box plot is based on five statistics: **the smallest and the largest values, the first and the third quartiles, and the median**.

- It might have an asterisk(*) sign which indicates an outlier.

a) An **outlier** is a value that is inconsistent with the rest of data.

b) The standard definition of **an outlier** is a value that is more than $Q_3 + 1.5(Q_3 - Q_1)$ or less than $Q_1 - 1.5(Q_3 - Q_1)$.

42	70	64	47	66	55	85	10	24	45
16	40	81	15	35	38	79	35	36	23
31	38	52	16	81	69	73	38	48	25
31	62	47	63	84	17	40	36	44	17
64	75	53	31	60	12	61	43	30	33
212	239	240	218	222	249	265	224	257	271
266	234	239	219	255	260	253	261	249	230
246	263	235	229	218	238	254	249	250	263
229	221	253	227	270	257	261	238	240	239
237	220	226	239	258	259	230	262	255	226

Ungrouped (Raw) Data

$\boxed{10}$ sample data

every 8th score data

$N = 100$

mean ✓
median –
mode → 263
variance ✓
std. dev.

$\{10, 38, 16, 62, 17, 43, 249, 219, 263, 263\}$

$$\text{mean}, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{10+38+16+62+17+43+249+219+263+263}{10} = \frac{1180}{10} = 118$$

ranked data
order

$\{\underset{1}{10}, \underset{2}{16}, \underset{3}{17}, \underset{4}{38}, \underset{5}{43} \mid \underset{6}{62}, \underset{7}{219}, \underset{8}{249}, \underset{9}{263}, \underset{10}{263}\}$

midpoint

$$\text{median} = \frac{43+62}{2} = \frac{105}{2} = 52.5$$

$$s^2 = \frac{(10-118)^2 + (16-118)^2 + (17-118)^2 + \cdots}{9} = \frac{116842}{9} = 12982.44$$

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

$\{10, 16, 17, 38, 43 \mid 62, 219, 249, 263, 263\}$

$x_i^2 = 100, 256, 541, \cdots$

$$S = \sqrt{s^2} = \sqrt{12982.4444} = 113.94$$

{10, 16, 17, 38, 43 | 62, 219, 249, 263, 263}

{10, 16, 17, 38, 43 | 62, 219, 249, 263, 263}

20

5 → 20

±0.5

Class	class limit	Tally	Frequency	Relative freq (%)	Class boundaries	Midpoint	CF
1 A	10 — 60	~~||||~~ ~~||||~~	10	$\frac{10}{20} \times 100 = 50$	9.5 — 60.5	35	10
2 B	61 — 111	||	2	$\frac{2}{20} \times 100 = 10\%$	60.5 — 111.5	86	12
3 C	112 — 162	—	0	0	111.5 — 162.5	137	12
4 D	163 — 213	—	0	0	162.5 — 213.5	118	12
5 E	214 — 264	~~||||~~ |||	8	$\frac{8}{20} \times 100 = 40\%$	213.5 — 264.5	239	20
Total			20	100%			

$$\text{Class width} = \frac{\text{range}}{\text{no. of class}} = \frac{263-10}{5} = \frac{253}{5} = 50.6 \approx 51$$

$10+51=61+51$

$213+51$

$$\text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

Assignment

- Exercises 12