Lecture Notes for Chapter 3

Tony Baker
MATK123 Elementary Statistics Fall 2022 9-19-2022
Three Rivers Community College
Introductory Statistics 10th ed. by Weiss.

Key Topics:
3.1 Measures of Center
3.2 Measures of Variation
3.3 Chebyshev's Rule and the Empirical Rule
3.4 The Five-Number Summary; Boxplots
3.5 Descriptive Measures for Populations; Use of Samples

Descriptive measures that indicate where the center or most typical value of a data set lies are called **measures of central tendency** or, more simply, **measures of center**. Measures of center are often called *averages*.

In this section, we discuss the three most important measures of center: the *mean*, *median*, and *mode*. The mean and median apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative (categorical) data.

Definition 3.1: Mean of a Data Set

The **mean** of a data set is the sum of the observations divided by the number of observations.

Definition 3.2: Median of a Data Set

Arrange the data in increasing order.

- If the number of observations is odd, then the **median** is the observation exactly in the middle of the ordered list.
- If the number of observations is even, then the **median** is the mean of the two middle observations in the ordered list.

In both cases, if we let n denote the number of observations, then the median is at position $(n + 1)/2$ in the ordered list.

Definition 3.3: Mode of a Data Set

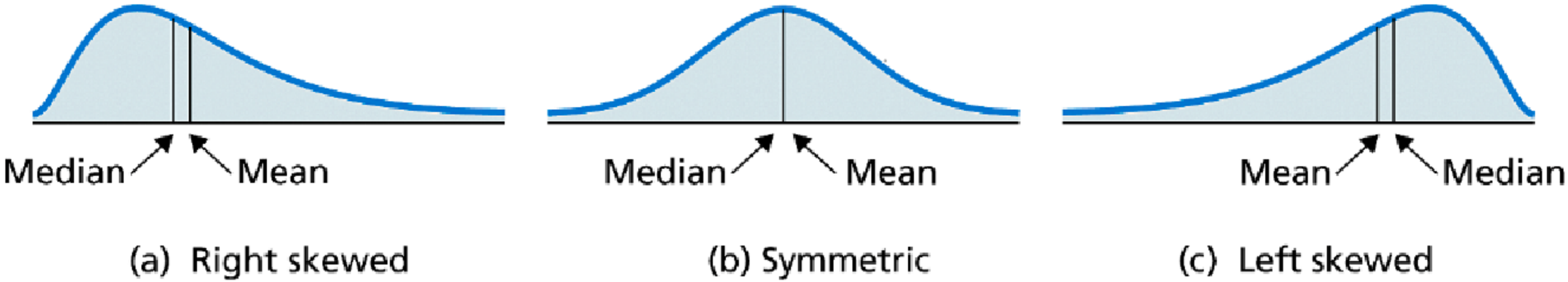
Find the frequency of each value in the data set.

- If no value occurs more than once, then the data set has *no mode*.
- Otherwise, any value that occurs with the greatest frequency is a **mode** of the data set.

Table 3.4 **Means, medians, and modes of salaries in Data Set I and Data Set II**

Measure of center	Definition	Data Set I	Data Set II
Mean	$\frac{\text{Sum of observations}}{\text{Number of observations}}$	$483.85	$474.00
Median	Middle value in ordered list	$400.00	$350.00
Mode	Most frequent value	$300.00	$300.00

Figure 3.1



Relative positions of the mean and median for archetypal (a) right-skewed, (b) symmetric, and (c) left-skewed distributions

Definition 3.5: Range of a Data Set

The **range** of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min},$$

where Max and Min denote the maximum and minimum observations, respectively.

Heights of Starting Players The heights, in inches, of the five starting players on Team I are 72, 73, 76, 76, and 78, as we saw in Fig. 3.2. Find the deviations from the mean.

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

$72, 73, 76, 76, 78 = \frac{375}{5}$

$\mu = 75$

$72 - 75 = -3^2 \quad 9$

$73 - 75 = -2^2 \quad 4$

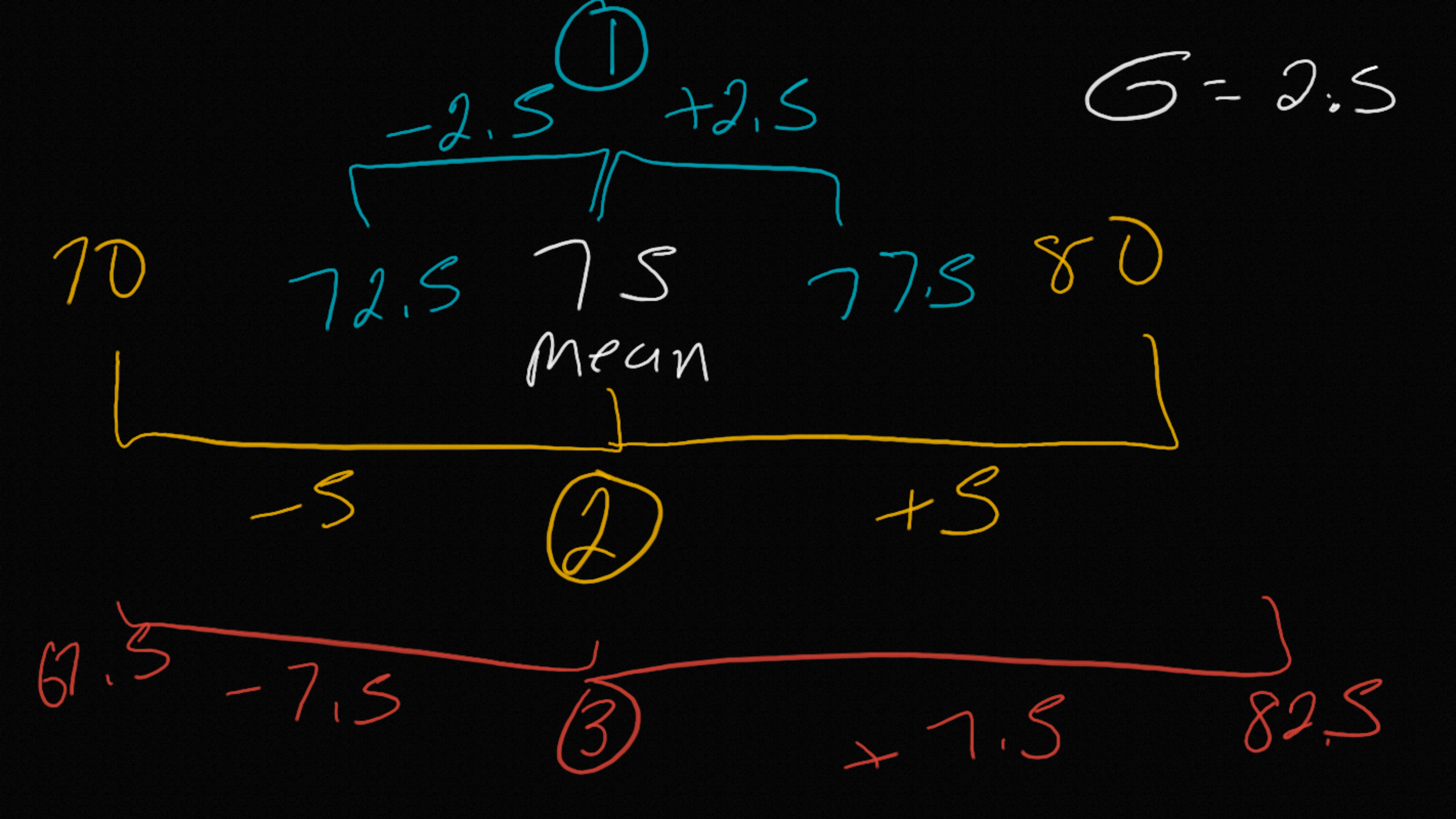
$76 - 75 = 1^2 \quad 1$

$76 - 75 = 1^2 \quad 1$

$78 - 75 = 3^2 \quad 9$

$\sigma = \sqrt{\frac{24}{4}}$

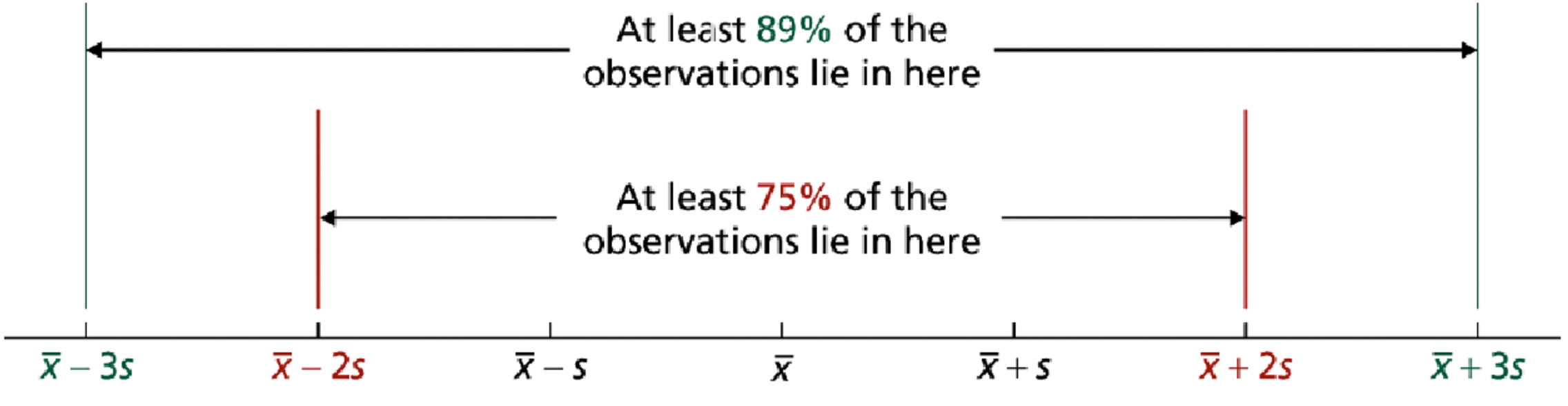
$\sigma = \sqrt{6} \approx 2.45$


1
−2.5
+2.5
σ = 2.5
70
72.5
75
77.5
80
Mean
−5
2
+5
67.5
−7.5
3
+7.5
82.5

Key Fact 3.3: Chebyshev's Rule

For any quantitative data set and any real number k greater than or equal to 1, at least $1 - 1/k^2$ of the observations lie within k standard deviations to either side of the mean, that is, between $\bar{x} - k \cdot s$ and $\bar{x} + k \cdot s$.

Figure 3.7



Forearm Length In 1903, K. Pearson and A. Lee published the paper "On the Laws of Inheritance in Man. I. Inheritance of Physical Characters" (*Biometrika*, Vol. 2, pp. 357–462). The article examined data on forearm length, in inches, for a sample of 140 men. The mean and standard deviation of the forearm lengths are 18.8 in. and 1.12 in., respectively.

a. Apply Chebyshev's rule with $k = 2$ to make pertinent statements about the forearm lengths of the men in the sample.

b. Repeat part (a) with $k = 3$.

$\mu = 18.8 \quad \sigma = 1.12$

15.44 16.56 17.68 18.8 19.92 21.04 22.16

$-1.12 \quad -1.12 \quad -1.12 \quad +1.12 \quad +1.12 \quad +1.12$

75%

$\geq 89\%$

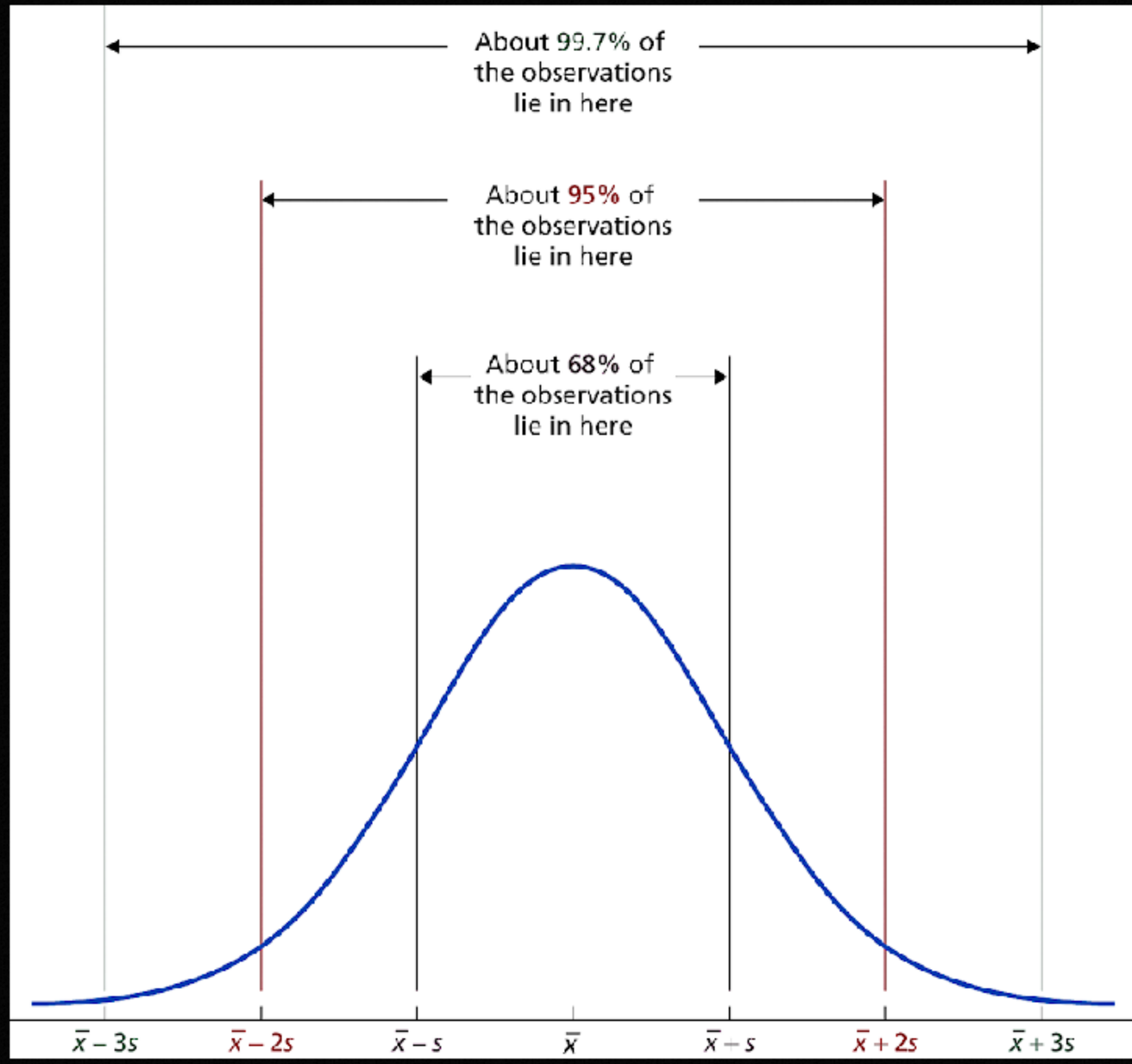
Key Fact 3.4: Empirical Rule

For any quantitative data set with roughly a bell-shaped distribution, the following properties hold.

Property 1: Approximately 68% of the observations lie within one standard deviation to either side of the mean, that is, between $\bar{x}$ s and $\bar{x}$ | s.

Property 2: Approximately 95% of the observations lie within two standard deviations to either side of the mean, that is, between $x - 2s$ and $x + 2s$.

Property 3: Approximately 99.7% of the observations lie within three standard deviations to either side of the mean, that is, between $\bar{x} - 3s$ and $\bar{x} + 3s$.



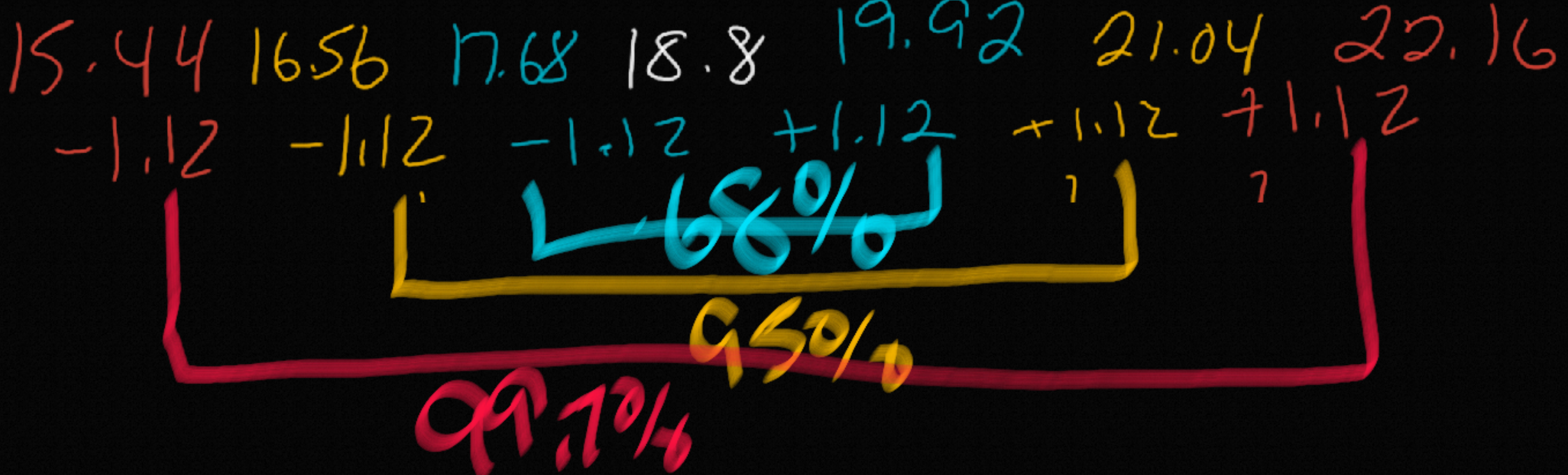
Forearm Length From Example 3.16 recall that the mean and standard deviation of forearm lengths for a sample of 140 men are 18.8 in. and 1.12 in., respectively.

In Example 3.16, we used Chebyshev's rule to make pertinent statements about the forearm lengths of the men in the sample. Presuming that the distribution of forearm lengths of the men in the sample is roughly bell shaped (which, in fact, it actually is), we can use the empirical rule to get more accurate estimates.

a. Apply Property 1 of the empirical rule to make pertinent statements about the forearm lengths of the men in the sample.

b. Repeat part (a) for Property 2 of the empirical rule.

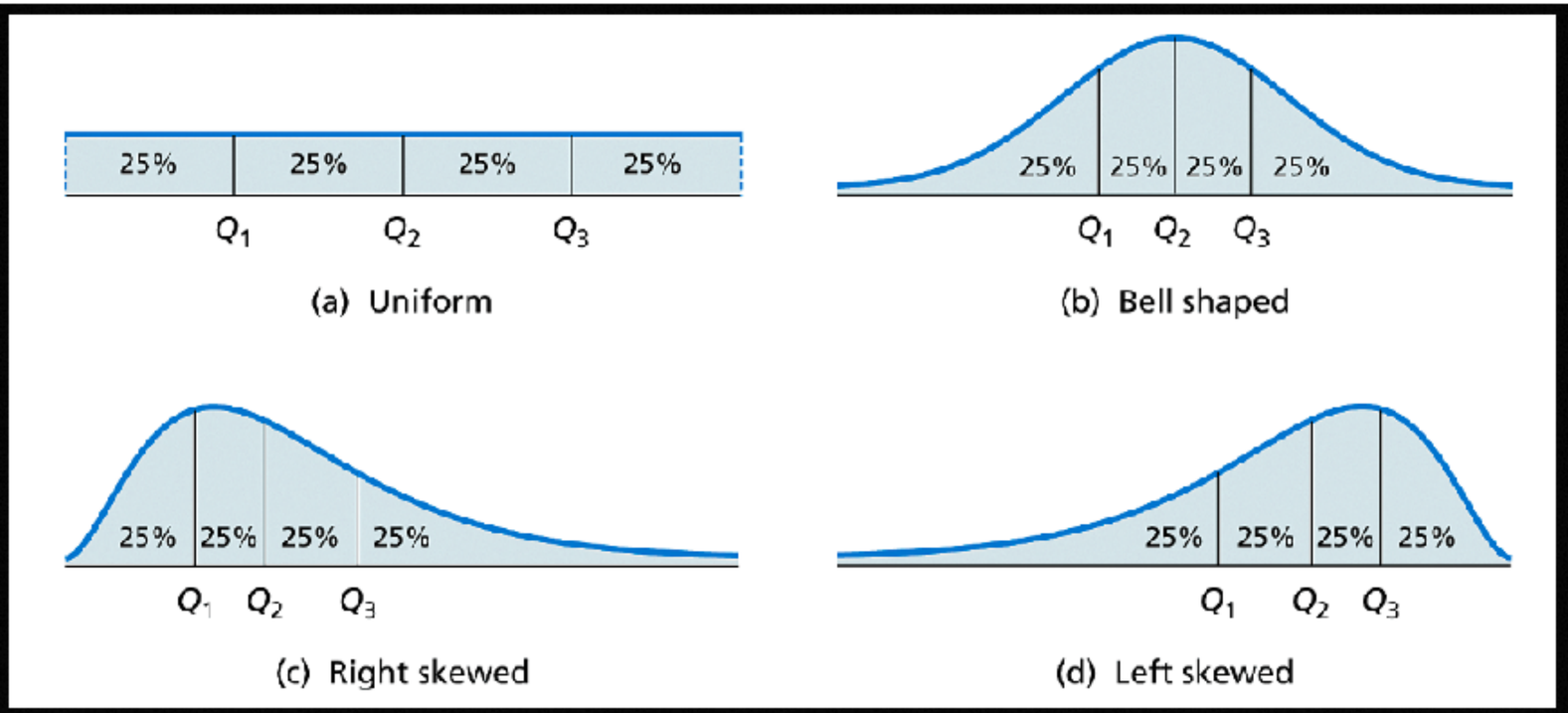
c. Repeat part (a) for Property 3 of the empirical rule.



Definition 3.7: Quartiles

First, arrange the data in increasing order. Next, determine the median. Then, divide the (ordered) data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

- The **first quartile** (Q_1) is the median of the bottom half of the data set.
- The **second quartile** (Q_2) is the median of the entire data set.
- The **third quartile** (Q_3) is the median of the top half of the data set.



~~25~~	41	~~27~~	32	43
66	35	31	~~15~~	~~5~~
34	~~26~~	32	38	~~16~~
~~30~~	38	~~30~~	~~20~~	~~21~~

5, 15, 16, 20, 21, 25, 26, 27, 30, 30

31, 32, 32, 34, 35, 38, 38, 41, 43, 66

5, 15, 16, 20, 21, 25, 26, 27, 30, 30

31, 32, 32, 34, 35, 38, 38, 41, 43, 66

$\text{Median} = \frac{30+31}{2} = 30.5$

$\text{1st Quartile} = \frac{21+25}{2} = 23$

$\text{3rd Quartile} = \frac{35+38}{2} = 36.5$

$\text{Min} = 5$ $\quad$ $\text{Max} = 66$

Definition 3.10: Lower and Upper Limits

The **lower limit** and **upper limit** of a data set are

$$\text{Lower limit} = Q_1 - 1.5 \cdot \text{IQR};$$
$$\text{Upper limit} = Q_3 + 1.5 \cdot \text{IQR}.$$

IQR =

$Q_3 - Q_1$

$IQR = 36.5 - 23 = 13.5$

$L.L. = 23 - 1.5(13.5) = 2.75$

$U.P. = 36.5 + 1.5(13.5) = 56.75$

To Construct a Boxplot

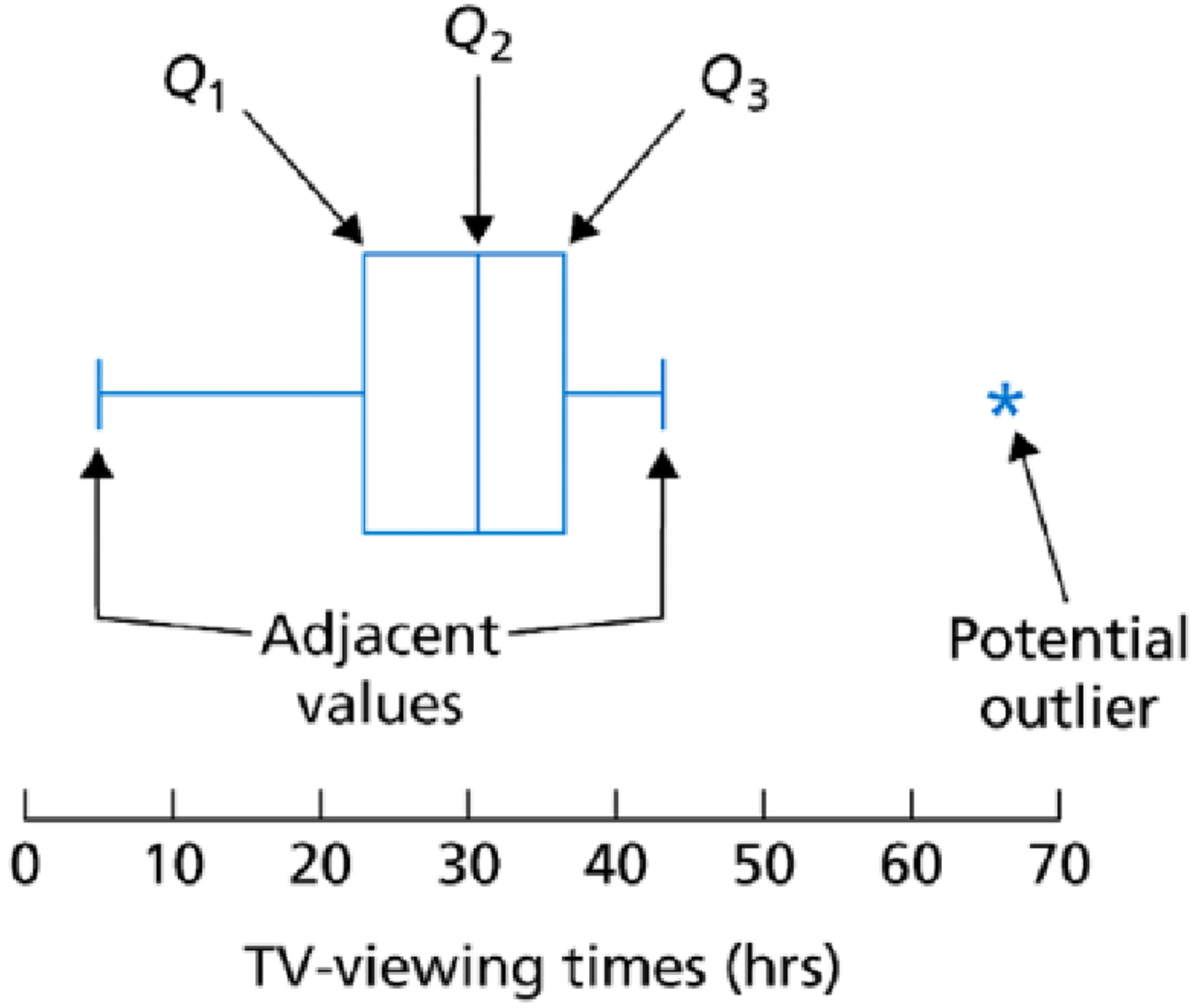
Step 1 Determine the quartiles.

Step 2 Determine potential outliers and the adjacent values.

Step 3 Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.

Step 4 Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.

Step 5 Plot each potential outlier with an asterisk.


Q_1
Q_2
Q_3
Adjacent values
Potential outlier
0
10
20
30
40
50
60
70
TV-viewing times (hrs)

(c)