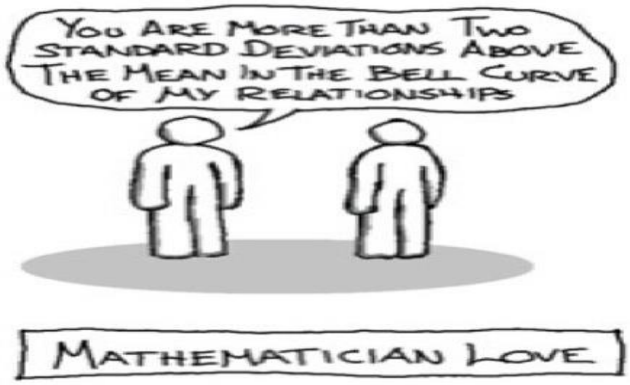
Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

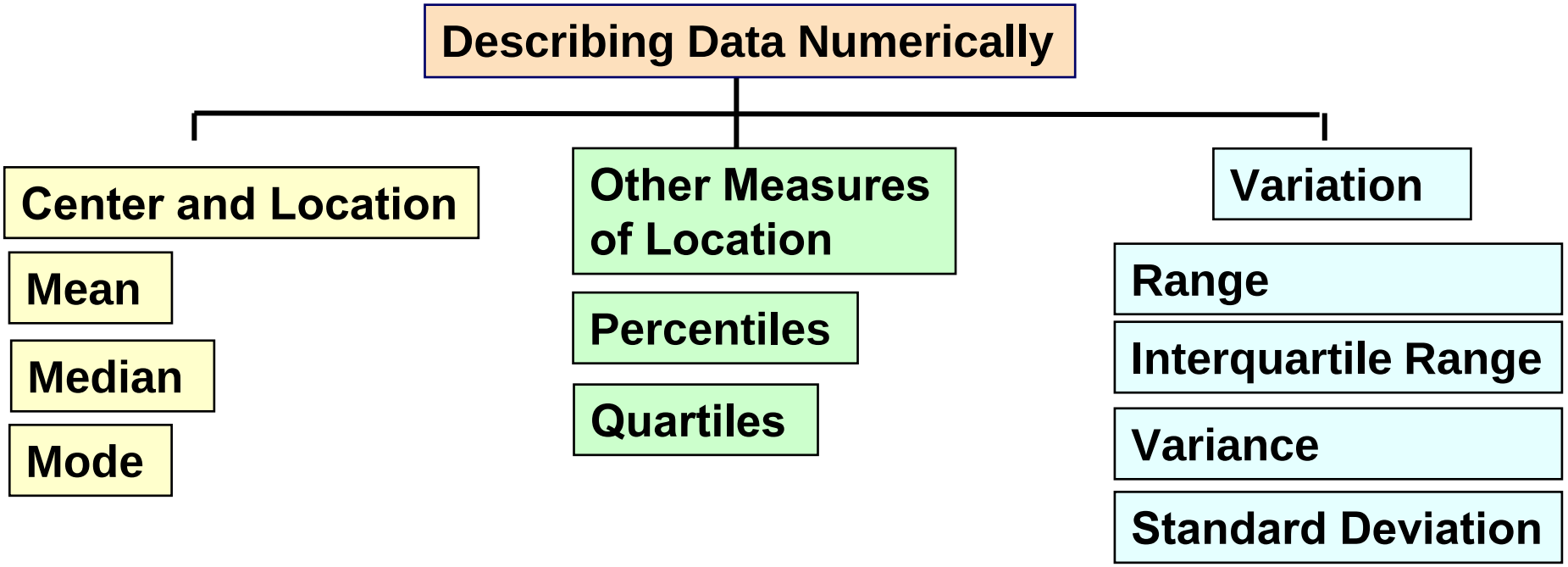
Chapter 3

Numerical Summary Measures



Measures of Central Tendency

- Tabular and graphical techniques provide useful summaries of data. However, these techniques are not suitable for statistical inference.
- Numerical summary measures are more precise, combine information in the data into a single number computed from a sample, and allow us to draw conclusions about the entire population used for inference.
- Measures of central tendency indicate where the majority of the data are centered, bunched, or clustered.



Notations

- x stands for a specific, fixed observation on a variable. Lowercase letters x, y, z are commonly used to represent observations.
- n is the number of observations in a data set, the sample size.
- If there are two data sets, we use m and n to denote their sample sizes. For more than two data sets, we use n_1, n_2, n_3, ...
- $x_1, x_2, x_3, \ldots, x_n$ refers to a set of fixed observations on a variable.
- The subscripts indicate the order in which the observations were selected, not the magnitudes of the observations
- Summation Notation $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$

  ✓ The left side is short-hand for the sum of n observations.

  ✓ Σ is the capital Greek letter sigma, and i is the index of summation with lower bound 1 and upper bound n.

Sample and Population Mean

The **sample** (**arithmetic**) **mean**, denoted $\overline{x}$, is the sum of the observations divided by the sample size. Written mathematically, it is

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

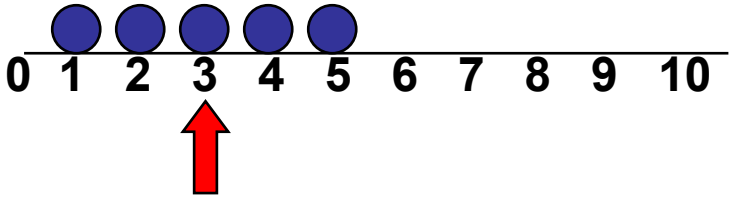
- Affected by extreme values (outliers)
- The **population mean** is denoted by the Greek letter mu, μ.
- Usually, μ is an unknown constant we would like to estimate. It describes the center of an entire population.
- The population mean is a fixed constant.
- The sample mean varies from sample to sample. It will not necessarily equal the population mean.

Median and Mode

- The **sample median**, denoted $\tilde{x}$, of the n observations $x_1, x_2, \ldots, x_n$ is the middle number when the observations are arranged in order from smallest to largest.
  1. If n is odd, the sample median is the single middle value
  2. If n is even, the sample median is the mean of the two middle values.
- The median is not affected by extreme values

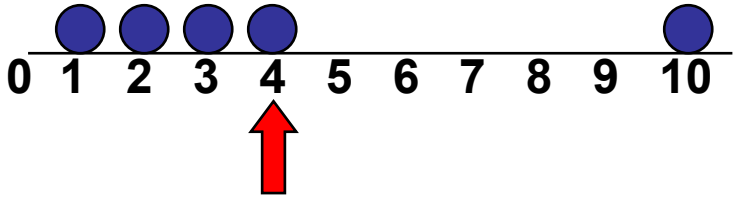
- The **mode**, denoted M, of a set of n observations $x_1, x_2, \ldots, x_n$ is the value that occurs most often.
- If all the observations occur with the same frequency, then the modes does not exist.
- If two or more observations occur with the same greatest frequency, then the mode is not unique: The distribution may be bimodal or multimodal.
- Used for either numerical or categorical data.
- Not affected by extreme values.

Examples



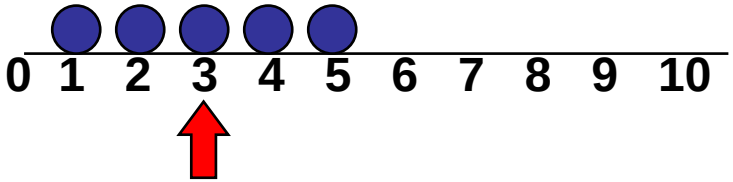
Mean = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

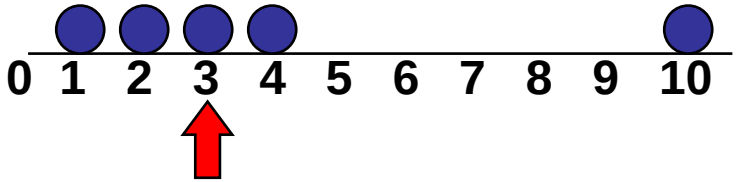


Mean = 4

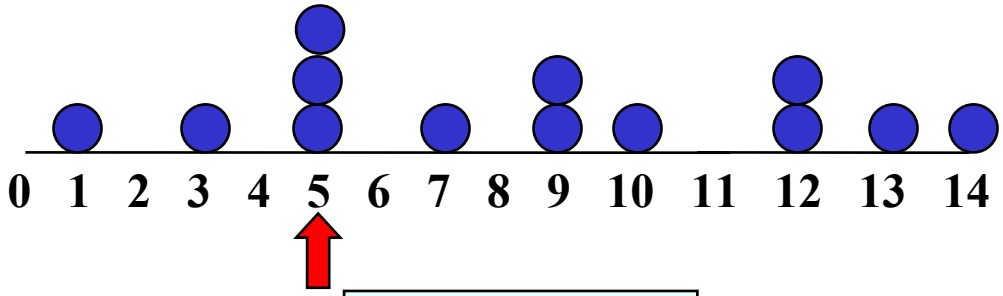
$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$



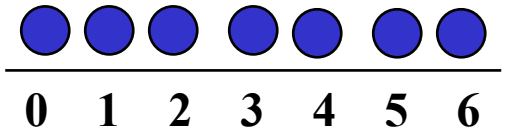
Median = 3



Median = 3



Mode = 5

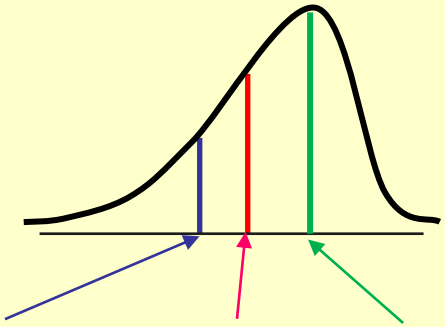


No Mode

Shape of Data Distribution

- Describes how data is distributed
- Symmetric or skewed
- The greater the difference between the mean and the median, the more skewed the distribution

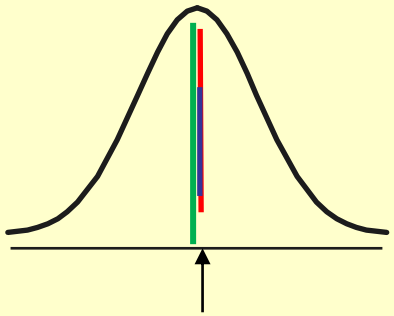
Left-Skewed (Negatively Skewed Distribution)



Mean < Median < Mode

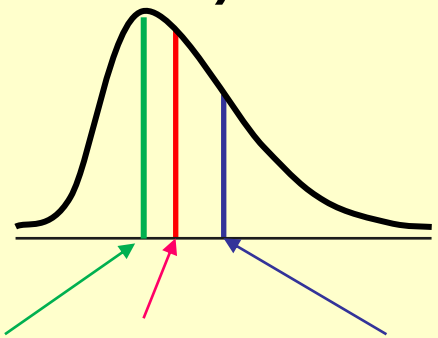
(Longer tail extends to left)

Symmetric



Mode ≈ Mean ≈ Median

Right-Skewed (Positively Skewed Distribution)



Mode < Median < Mean

(Longer tail extends to right)

Sample Proportion for Categorical Variable

For observations on a categorical variable with only two responses, the sample proportion of successes, denoted $\hat{p}$, is the relative frequency of occurrence of successes.

$$\hat{p} = \frac{\text{number of S's in the sample}}{\text{total number of responses}} = \frac{n(S)}{n}$$

- The **population proportion of successes** is denoted as p.
- A success is *not* necessarily a good thing. It just refers to the characteristic of interest to the researcher.
- The sample proportion can be thought of as a sample mean in disguise. Just change every success to a 1 and every failure to a 0.

$$\bar{x} = \frac{1}{n}(\text{a sum of 0's and 1's}) = \frac{n(\text{S})}{n} = \hat{p}$$

Example: Suppose the State Police recently established a checkpoint along a heavily traveled rural road. A success was recorded for a driver wearing a seat belt, and a failure recorded otherwise. The sample contains 30 observations and 25 successes given in the following table.

$$\hat{p} = \frac{n(S)}{n} = \frac{25}{30} = 0.8333$$

S	S	S	F	S	S	S	F	S	S	F	S	S	S	S
S	S	S	S	S	S	F	S	S	S	S	S	F	S	S

Approximately 83% of the drivers who were stopped at the checkpoint were wearing seat belts.

Which measure of location is the "best"?

- **Mean** is generally used, unless extreme values (outliers) exist
- Then **Median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers
- **Mode** is good for determining most likely to occur

  - **Mean:** ($3,000,000 / 5)
    = **$600,000**
  - **Median:** middle value of ranked data
    ME = **$300,000**
  - **Mode:** most frequent value
    Mode = **$100,000**

House Prices:

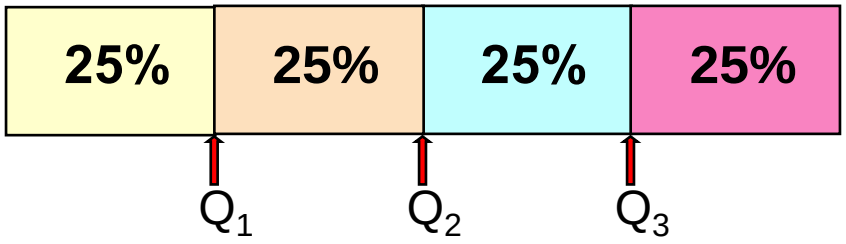
	$2,000,000
	500,000
	300,000
	100,000
	100,000
Sum	**3,000,000**

Other Location Measures

Quartiles

The quartiles divide/split the ordered data into four parts/groups.

- 1st quartile (Q_1) = 25th percentile
- 2nd quartile (Q_2) = 50th percentile, also, the median
- 3rd quartile (Q_3) = 75th percentile



Computing Quartiles: Suppose $x_1, x_2, \ldots, x_n$ is a set of observations.

1. Arrange the observations from smallest to largest.
2. To find Q_r, compute $d_r = (r/4)\,n$
   a) If d_r is a whole number, then p_r is the mean of the observations in positions d_r and $d_r + 1$.
   b) If d_r is not a whole number, round up to the next whole number (whatever the fraction) to find the position of Q_r.

Other Location Measures

Example1: Find the first quartile, Q_1

Sample Data in Ordered Array:	**11**	**12**	**13**	**16**	**16**	**17**	**18**	**21**	**22**

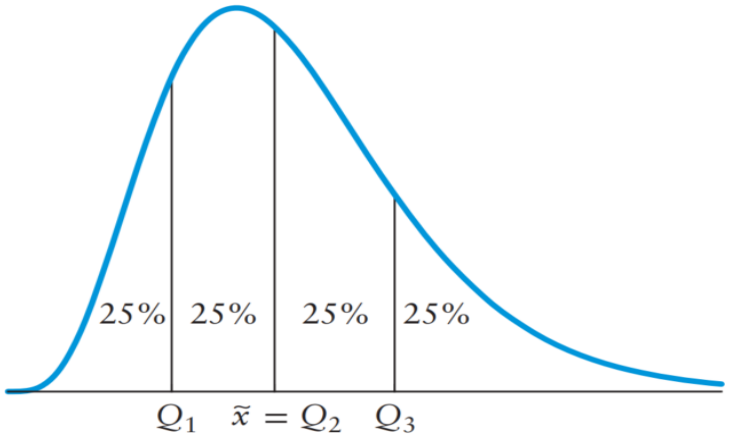
n = 9

$$d_1 = \frac{1}{4}(9) = 2.25$$

Round up to **3** since not an integer

So, round up and use the value in the 3rd position: **Q_1 = 13**

Interpretation: 25% of the data lie at or below 13



Other Location Measures

Example2: The following 10 observations represent the resting pulse rate for patients involved in an exercise study. 68 71 64 58 61 76 73 62 72 66
Find the first quartile, the third quartile, and the interquartile range.

Observation	58	61	62	64	66	68	71	72	73	76
Position	1	2	3	4	5	6	7	8	9	10

For Q_1: $d_1 = \frac{1}{4}(10) = 2.5$ Because d_1 is not a whole number, round up to 3

Q_1 is in the third position in the ordered list. So, $Q_1 = 62$

For Q_3: $d_3 = \frac{3}{4}(10) = 7.5$ Because d_1 is not a whole number, round up to 8

Q_3 is in the eighth position in the ordered list. So, $Q_3 = 72$

$$IQR = Q_3 - Q_1$$
$$= 72 - 62 = 10$$

Other Location Measures

Percentiles

Let $x_1, x_2, \ldots, x_n$ be a set of observations. The **percentiles** divide the ordered data set into 100 parts. For any integer r ($1 \le r \le 99$), the ***r*th percentile**, denoted p_r, is a value such that

- r% of the observations lie at or below p_r and
- (100 – r)% lie above p_r

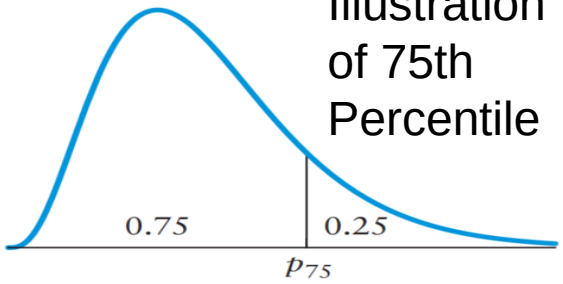


Illustration of 75th Percentile

- The 50th percentile is the median, $p_{50} = \tilde{x}$.
- The 25th percentile is the first quartile and the 75th percentile is the third quartile: $p_{25} = Q_1$, $p_{75} = Q_3$

Computing Percentiles: Suppose $x_1, x_2, \ldots, x_n$ is a set of observations.

1. Arrange the observations from smallest to largest.
2. To find p_r, compute $d_r = (r / 100)\, n$
   a) If d_r is a whole number, then p_r is the mean of the observations in positions d_r and $d_r + 1$.
   b) If d_r is not a whole number, round up to the next whole number (whatever the fraction) to find the position of p_r.

Other Location Measures

Example: A walk across the Brooklyn Bridge in New York City takes approximately 25–60 minutes. A random sample of people walking across the bridge was obtained, and their times are given in the table. Find the time at which it took 20% of the walkers to make it across the bridge.

$$d_r = \frac{r}{100}(n)$$

$$d_{20} = \frac{20}{100}(30) = 6$$

44	51	43	31	50	53	59	49	55	25
28	30	60	42	36	54	31	33	48	39
37	44	48	51	34	38	58	59	53	59

A portion of this ordered list is given in the following table

Observation	25	28	30	31	31	33	34	36	37	38
Position	1	2	3	4	5	6	7	8	9	10

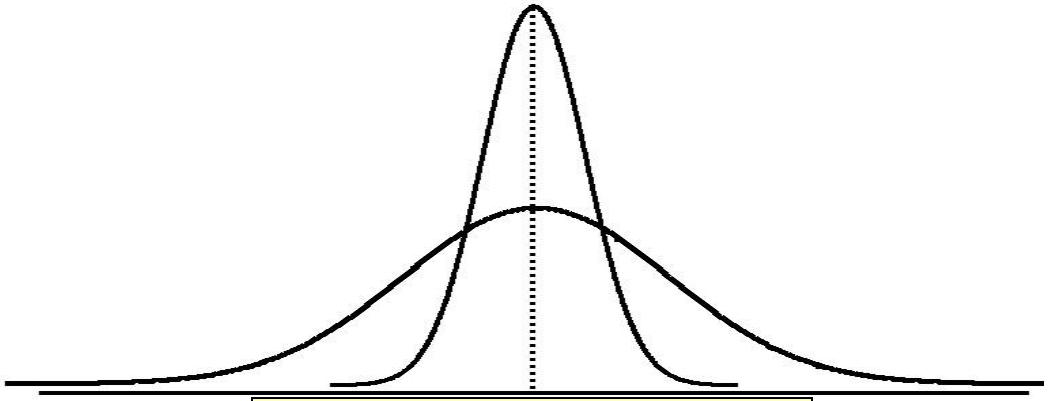
The 20th percentile, p_{20}, is the mean of the sixth and seventh observations.

$$p_{20} = \frac{33+34}{2} = 33.5$$

20% of the walkers made it across the bridge within 33.5 minutes, and 80% took longer than 33.5 minutes.

Measures of Variability

- Measures of central tendency alone are not sufficient to completely describe a sample.
- Two different data sets can have similar measures of central tendency, but different amounts of variability.
- Measures of variation give information on the spread or variability of the data values.
  - ✓ Smaller value of variation measure (less variation)
  - ✓ Larger value of variation measure (more variation)



Same center, different variation

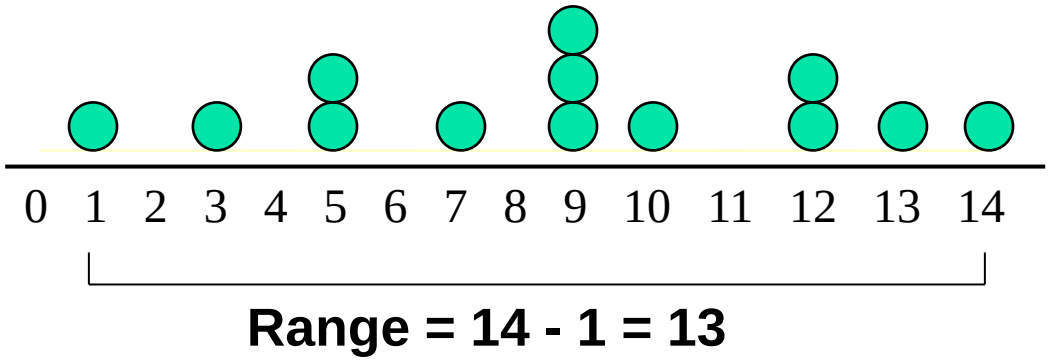
Range

The (**sample**) **range**, denoted R, of a set of n observations $x_1, x_2, \ldots, x_n$ is the largest observation minus the smallest observation. Written mathematically,

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

where $x_{\max}$ is the largest observation and $x_{\min}$ is the smallest observation.

Example:

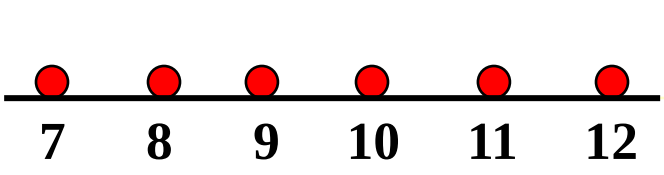


Range = 14 - 1 = 13

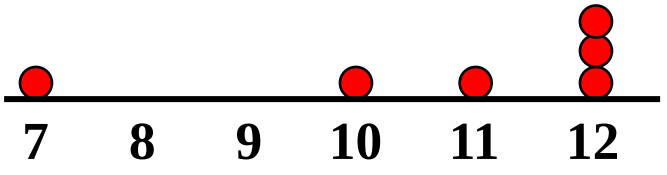
A data set with a small range has little variability.
A data set with a large range has lots of variability and is spread out.

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

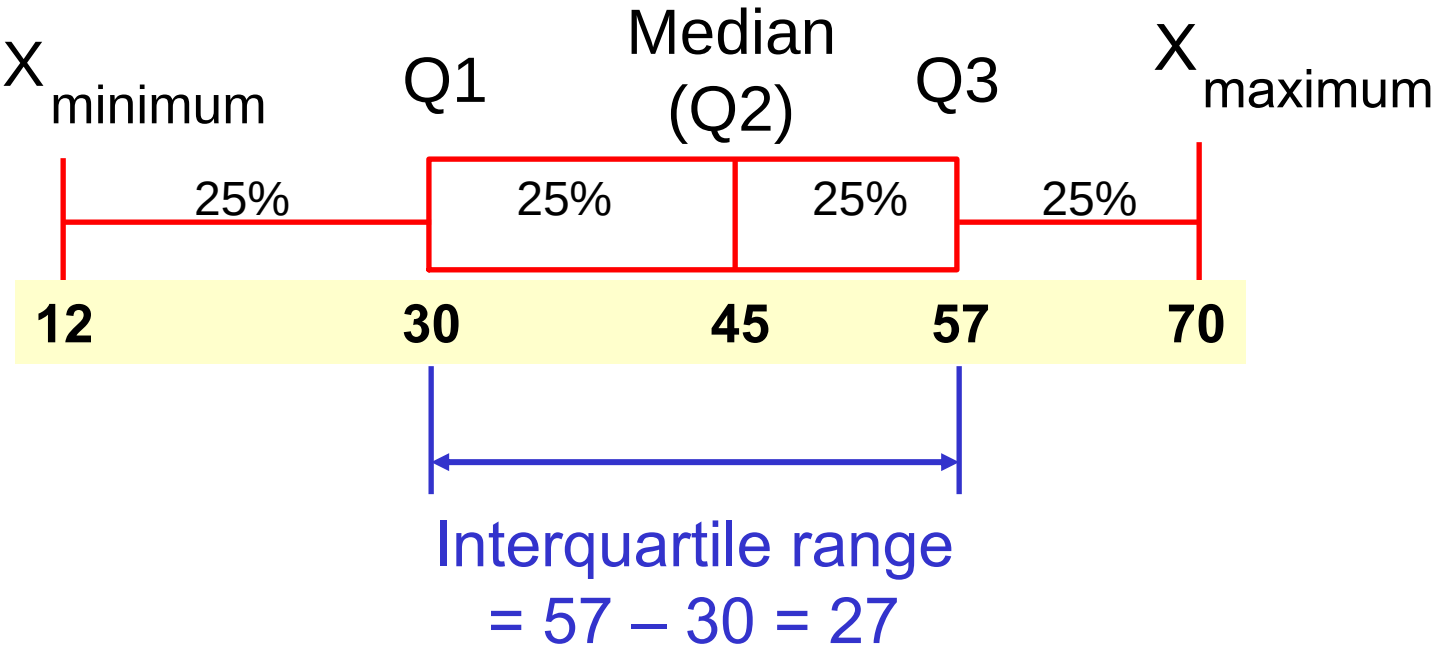
1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high-and low-valued observations and calculate the range from the remaining values.
- Interquartile range (IQR) = $Q_3 - Q_1$
- The IQR is the length of an interval that includes the middle half (middle 50%) of the data.

Example:



Variance and Standard Deviation

- The sum of the squared deviations about the mean divided by n – 1 (in squared units)

Population variance:

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

Sample variance:

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

- Most commonly used measure of variation is the SD.
- Shows variation about the mean.
- The SD has the same units as the original data.

Population standard deviation:

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

Sample standard deviation:

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

Computational Formula for Sample Variance

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

Example: The data below show how long (in minutes) seven patients (sample) have to wait to see their physician. To measure how the wait times vary from the expected wait-time (average), you are asked to calculate the standard deviation.

28, 29, 33, 32, 38, 28, 22

$$s^2 = \frac{1}{7-1}\left[6450 - \frac{1}{7}(210)^2\right] = 25$$

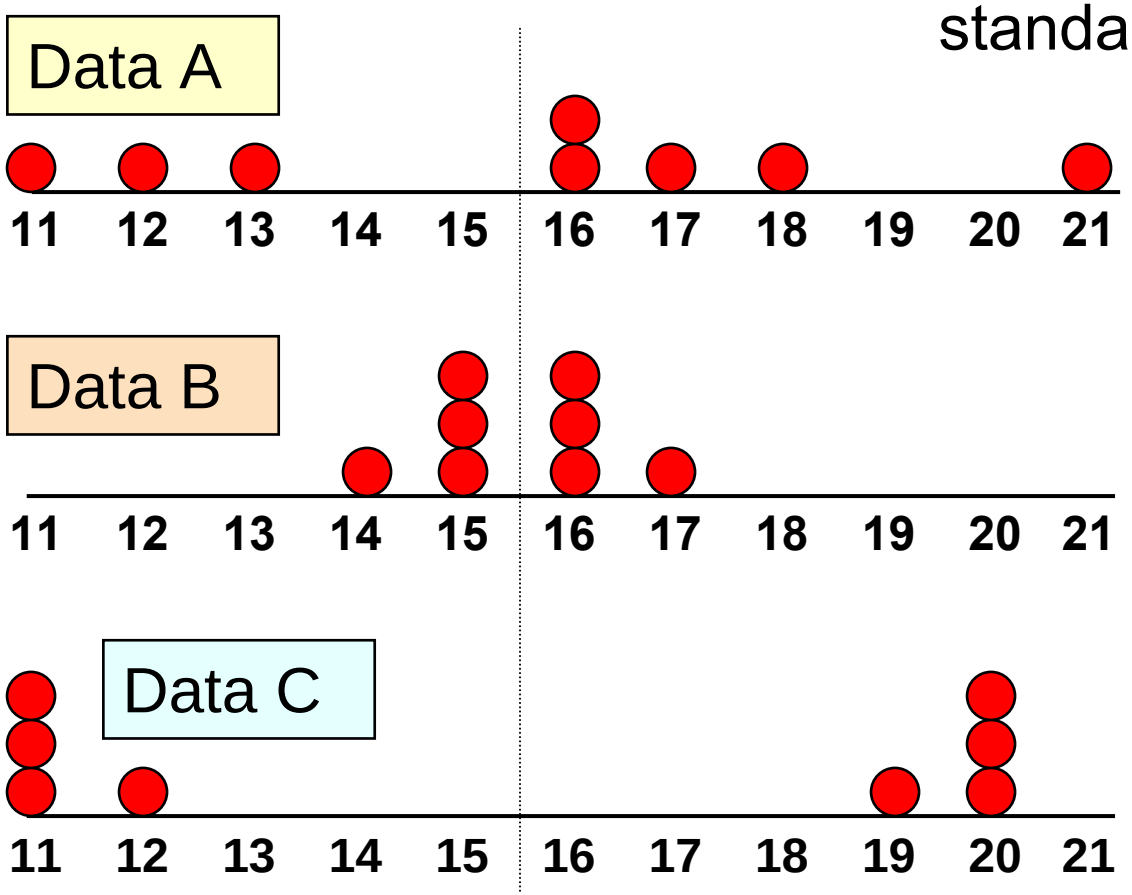
$$\bar{x} = \frac{\sum x}{n} = \frac{210}{7} = 30$$

$$s = \sqrt{s^2} = \sqrt{25} = 5$$

We expect the waiting time to vary ± 5 minutes from the expected or average wait time of 30 minutes. Thus, patients are expected to wait between 25 minutes and 35 minutes.

Comparing Standard Deviations

Same mean, but different standard deviations:



Mean = 15.5
s = 3.34

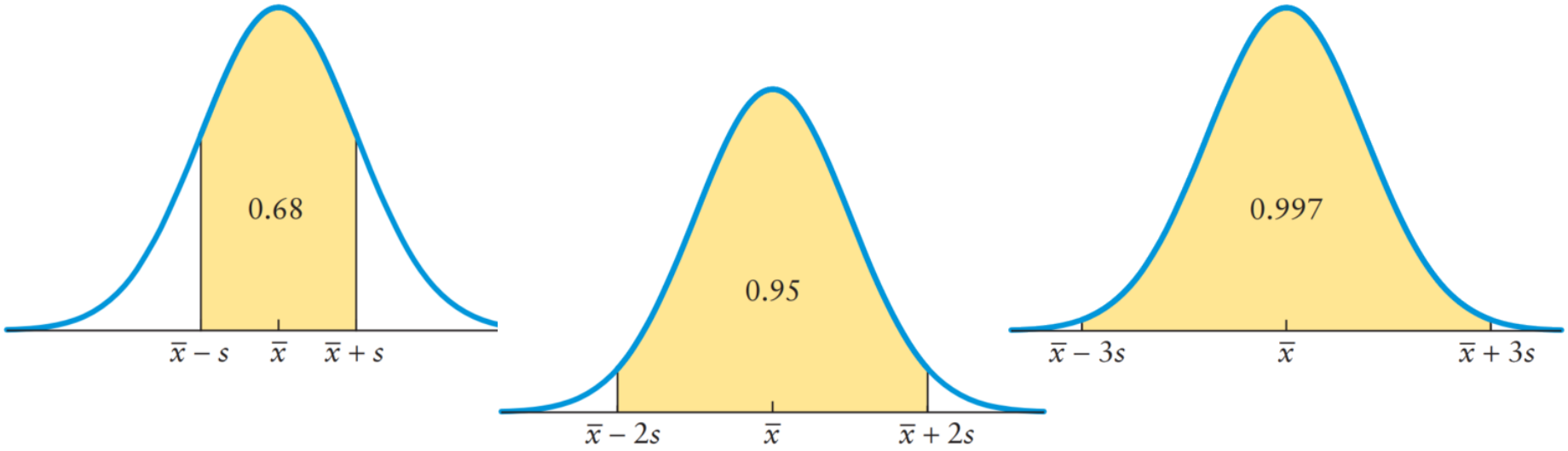
Mean = 15.5
s = .93

Mean = 15.5
s = 4.57

The Empirical Rule

If the shape of the distribution of a set of observations is approximately normal, then:

1. The proportion of observations within 1 standard deviation of the mean is approximately 0.68.
2. The proportion of observations within 2 standard deviations of the mean is approximately 0.95.
3. The proportion of observations within 3 standard deviations of the mean is approximately 0.997.



Example1: World Record Speeding Ticket

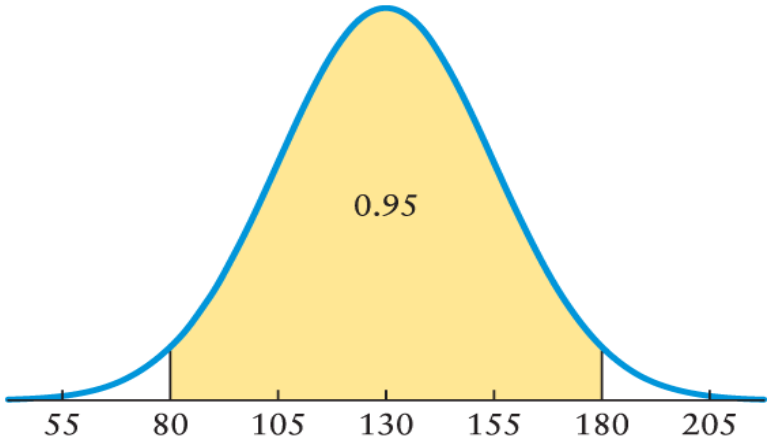
An ability-to-pay system led to the world's most expensive speeding ticket ($290,000) in 2010 to a Swedish man traveling 180 mph in a Mercedes. In a random sample of more traditional ticket fines in Alberta, Canada, in July 2018, suppose the shape of the distribution is approximately normal with $\bar{x} = 130$ and $s = 25$ (Canadian dollars). Approximately what proportion of observations is

a. between 80 and 180?
b. greater than 205 or less than 55?
c. greater than 205?
d. between 105 and 180?

$$\bar{x} \mp 2s$$

a. [130 – 2(25), 130 + 2(25)] = (80, 180)

Therefore, Approximately 95% of the observations lie within 2 standard deviations of the mean, in the interval (80, 180).



Example1: World Record Speeding Ticket

b. greater than 205 or less than 55?

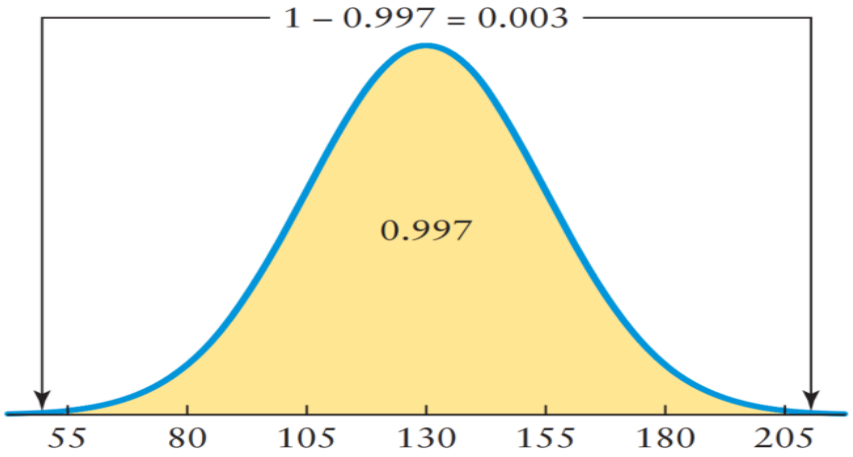
c. greater than 205?

d. between 105 and 180?

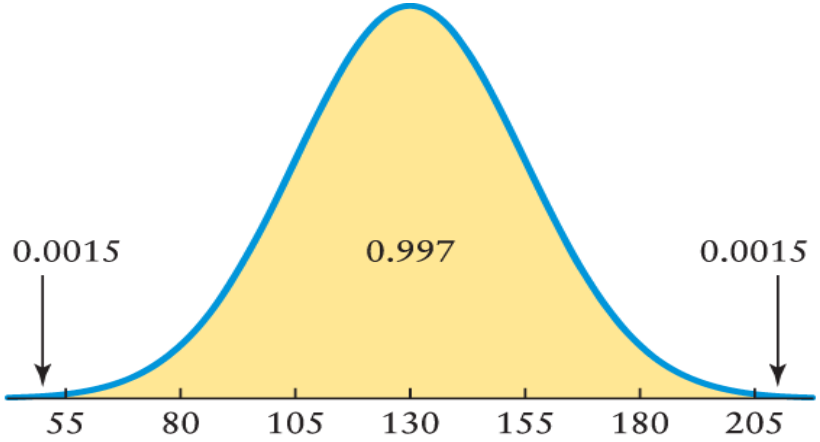
b. [130 – 3(25), 130 + 3(25)] = (55, 205)

$$\bar{x} \mp 3s$$

Therefore, 55 to 205 is a symmetric interval about the mean, 3 standard deviations in each direction. The Empirical Rule states that approximately 0.997 of the observations lie in this interval. So, 1 – 0.997 = 0.003 of the observations lie outside this interval.



c. The proportion outside the symmetric interval 3 standard deviations from the mean (55, 205) is evenly divided between the two tails. Therefore approximately 0.003/2 = 0.0015 (or 0.15%) of the observations are greater than 205.



Example1: World Record Speeding Ticket

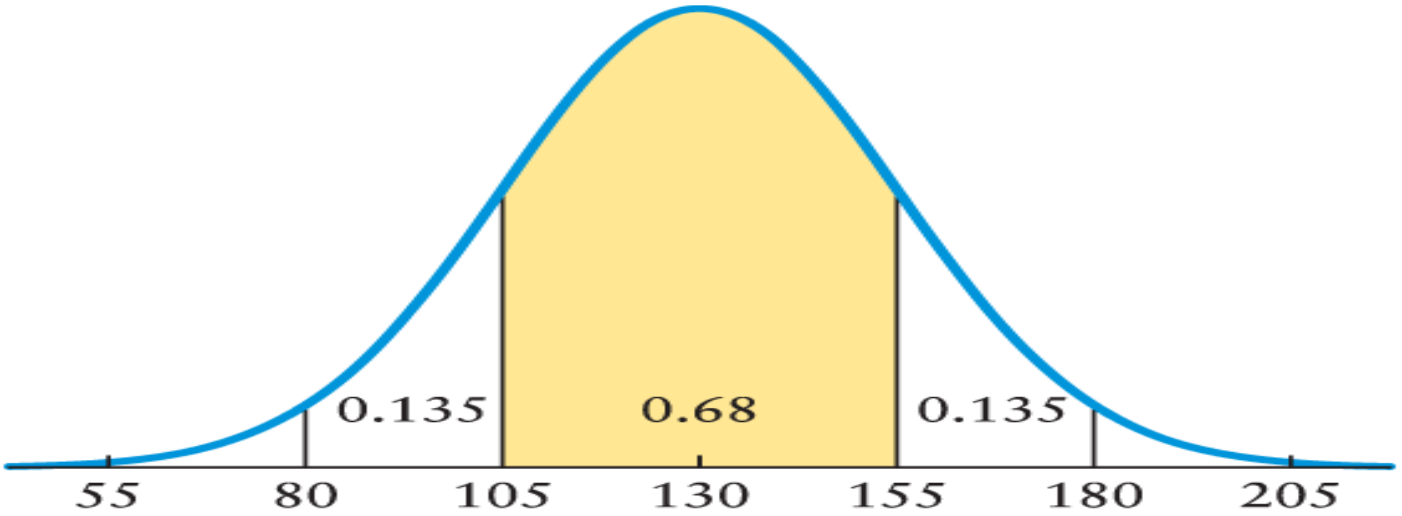
d. between 105 and 180? $\bar{x} \mp s$ → [130 – (25), 130 + (25)] = (105, 155)

Approximately 0.68 of the observations lie in the interval (105, 155) and approximately 0.95 of the observations lie in the interval (80, 180).

This means 0.95 – 0.68 = 0.27 of the observations lie in the intervals (80, 105) and (155, 180).

Because a normal distribution is symmetric, 0.27/2 = 0.135 of the observations lie between 155 and 180.

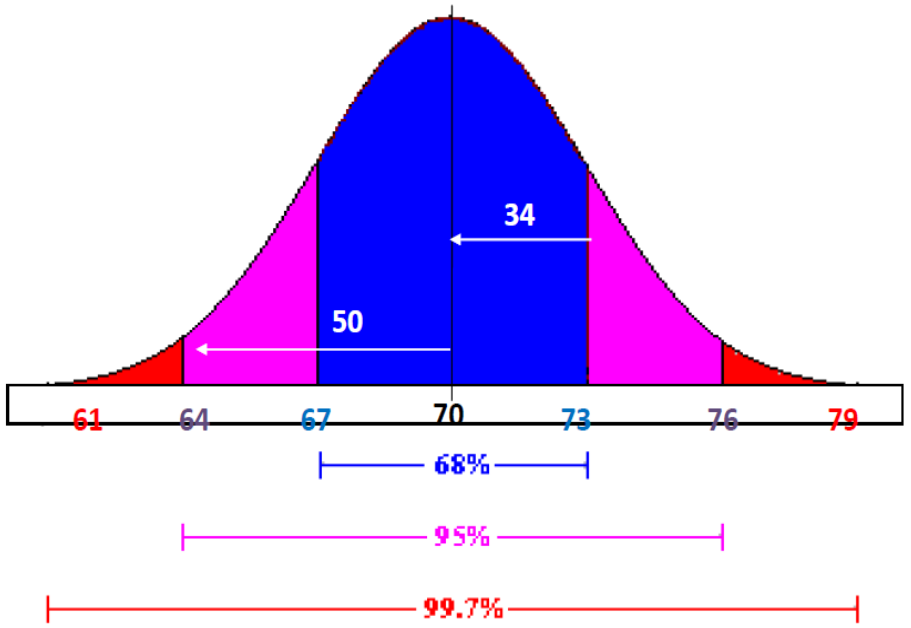
Therefore, a total of approximately 0.68 + 0.135 = 0.815 of the observations lie between 105 and 180.



More Examples

Example2: It is known that height of adult men in Canada follows a Normal distribution with mean 70 inches and standard deviation 3 inches. What percentage of adult men in Canada is shorter than 73 inches?

$\mu \pm \sigma = 70 \pm (3) = (67, 73)$



P(X 73) = 84%

Example3: If you know that the average of students' grades is 50 and the standard deviation is 10. According to the Empirical Rule, what are the two grades that 95% of grades is lied in between?

$P(? \le X \le ?) = 95\%$

$\mu \pm 2\sigma = 50 \pm 2(10)$

$P(30 \le X \le 70) = 95\%$

So, 95% of grades lie between 30 and 70

z-Score (Standardized Value)

Suppose $x_1, x_2, \ldots, x_n$ is a set of n observations with mean $\bar{x}$ and standard deviation s. The ***z*-score** corresponding to the ith observation x_i is given by

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

- z_i is a measure associated with x_i that indicates the distance from $\bar{x}$ in standard deviations
- Can be used to compare datasets
- z_i may be positive, negative, or zero.
  - A positive *z*-score indicates the observation is to the right of the mean.
  - A negative *z*-score indicates the observation is to the left of the mean.
- A *z*-score is a measure of relative standing; it indicates where an observation lies in relation to the rest of the data values.
- For any set of n observations, the sum of all *z*-scores is 0.

$$\sum z_i = 0$$

Standardized Value Examples

Example1: IQ scores in a sample have approximately a bell-shaped distribution with mean $\bar{x}$ = 100 and standard deviation s = 15. Find the standardized score (z-score) for a person with an IQ of 121.

$$z = \frac{x - \bar{x}}{s} = \frac{121 - 100}{15} = 1.4$$

Someone with an IQ of 121 is 1.4 standard deviations above the mean as we got positive z value.

Example2: Many colleges still require and strongly consider SAT or ACT scores in the admission process. These two admissions tests are scored on very different scales. Suppose the summary information for the math portion of each test is given in the table. One applicant to a college scored 670 on the SAT Math test and another scored 27 on the ACT Math test. Which score is better, in terms of statistics?

Test	Mean	Standard deviation
SAT Math	540	87
ACT Math	20.7	5.0

1st. applicant: $z = \frac{670 - 540}{87} \approx 1.49$

2nd. applicant: $z = \frac{27 - 20.7}{5} = 1.26$

The first applicant's score is better, because the test score is farther from the mean to the right, indicating, in this case, a better performance.

Five-Number Summary

The **five-number summary** for a set of observations $x_1, x_2, \ldots, x_n$ consists of the minimum value, the maximum value, the first and third quartiles, and the median.

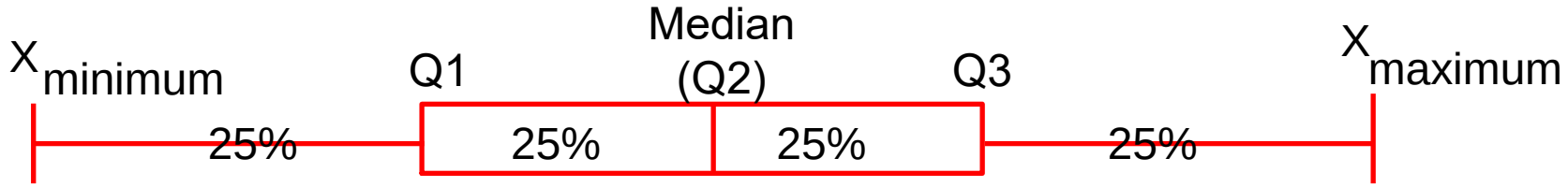
These numbers provide a glimpse into the symmetry, central tendency, and variability in a data set.

A **box plot**, or box-and-whisker plot, is a compact graphical summary that is constructed using the five-number summary for a set of observations.

Constructing a Standard Box Plot

Given a set of n observations: $x_1, x_2, \ldots, x_n$:

1. Find the five-number summary: $x_{\min}$, Q_1, $\tilde{x}$, Q_3, $x_{\max}$.
2. Draw a horizontal axis and sketch a box of any height that extends from Q_1 to Q_3.
3. Draw a vertical line in the box at the median.
4. Draw a horizontal line (whisker) from the left edge of the box to the minimum value (from Q_1 to $x_{\min}$) and a whisker from Q_3 to $x_{\max}$.



Five-Number Summary Examples

Example1: The data below show the number of volunteer hours completed by 15 students in a summer program. Summarize the data with five numbers and construct a box and whisker plot to describe the distribution.

12, 20, 20, 24, 25, 26, 27, 30, 32, 34, 36, 46, 48, 86, 128

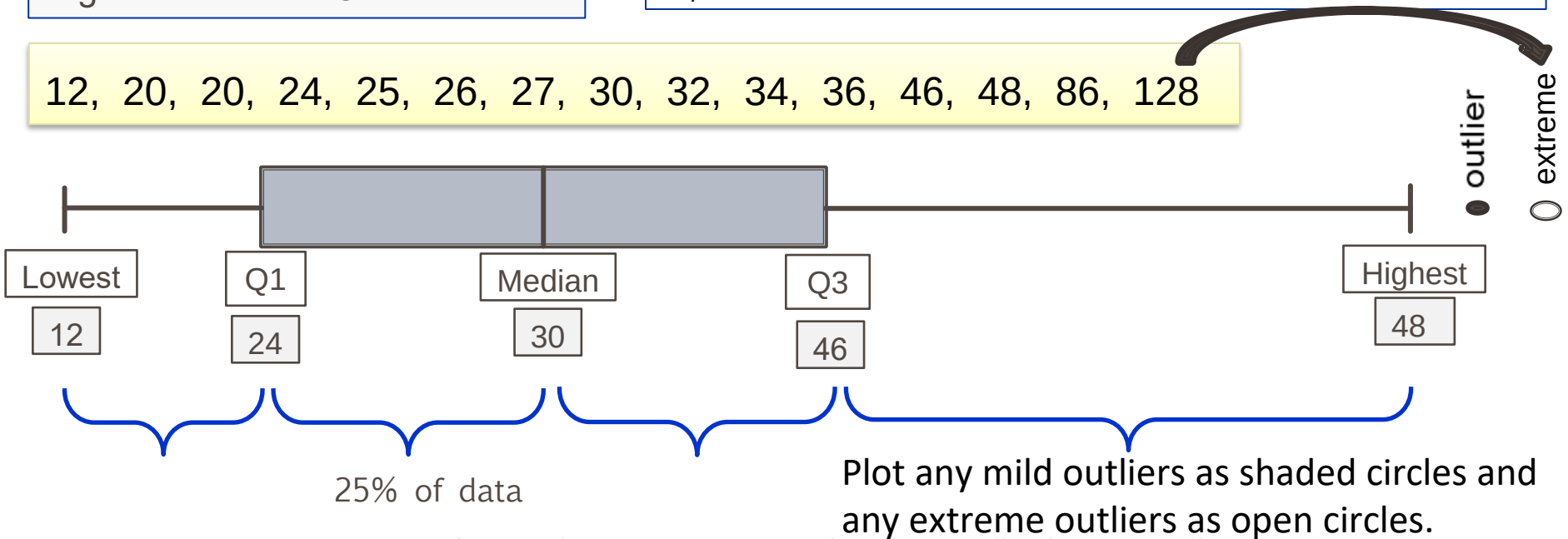
Lowest Value = 12
First Quartile = 24
Median Value = 30
Third Quartile = 46
Highest Value = 48

$d_r = (25/100)(15) = 3.75 \rightarrow 4^{th}$ ranked value

$d_r = (50/100)(15) = 7.5 \rightarrow 8^{th}$ ranked value

$d_r = (75/100)(15) = 11.25 \rightarrow 12^{th}$ ranked value

12, 20, 20, 24, 25, 26, 27, 30, 32, 34, 36, 46, 48, 86, 128



Plot any mild outliers as shaded circles and any extreme outliers as open circles.

Five-Number Summary Examples

A box-and-whisker plot provides a way to identify mild and extreme outliers in your data.

IQR = 46 – 24 = 22

Mild Outlier

Inner Lower Limit = Q_1 – 1.5 (IQR) ➔ 24 – 1.5 (22) = - 9
Inner Upper Limit = Q_3 + 1.5 (IQR) ➔ 46 + 1.5 (22) = 79

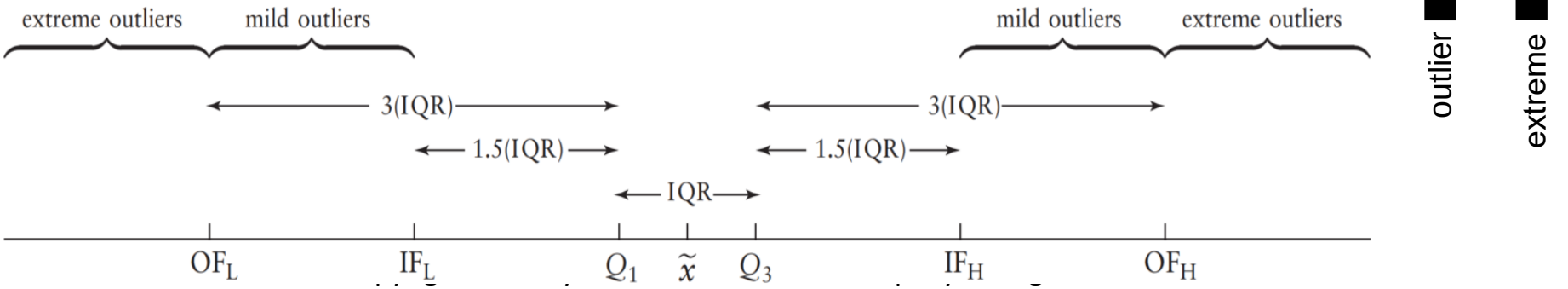
A value that is less than –9 or greater than 79 is a mild outlier

Extreme Outlier

Inner Lower Limit = Q_1 – 3.0 (IQR) ➔ 24 – 3.0 (22) = - 42
Inner Upper Limit = Q_3 + 3.0 (IQR) ➔ 46 + 3.0 (22) = 112

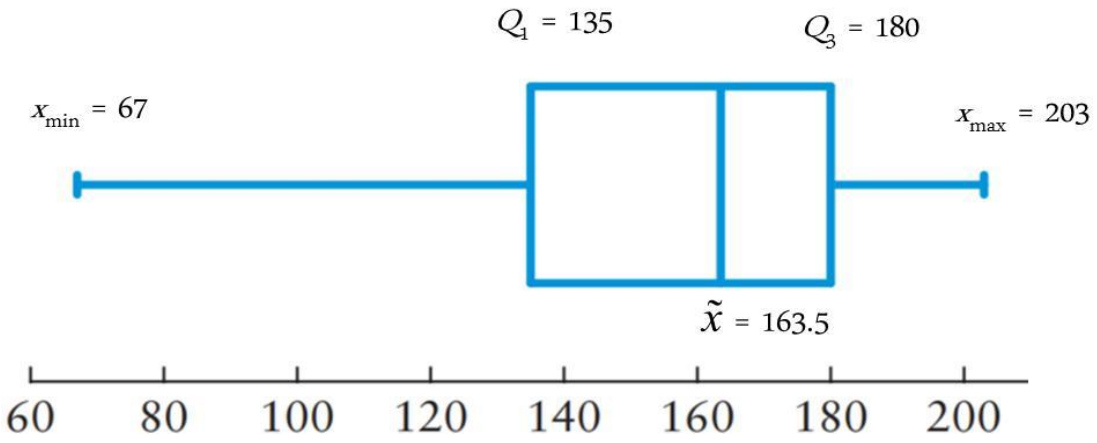
A value that is less than – 42 or greater than 112 is an extreme outlier

12, 20, 20, 24, 25, 26, 27, 30, 32, 34, 36, 46, 48, 86, 128



Five-Number Summary Examples

Example2: Systolic Blood Pressure



The box plot suggests that the data are negatively skewed or skewed to the left. The lower half of the data are much more spread out than the upper half are.

Example3: Sled Dog Trips

$Q_1 = 1.1$ $Q_3 = 3.7$

IQR = 3.7 – 1.1 = 2.6

$IF_L = 1.1 - 1.5(2.6) = -2.8$ $OF_L = 1.1 - 3(2.6) = -6.7$

$IF_H = 3.7 + 1.5(2.6) = 7.6$ $OF_H = 3.7 + 3(2.6) = 11.5$

Mild outliers: **7.9** and **9.5** are between 7.6 and 11.5

Extreme outlier: **11.9** is larger than 11.5.