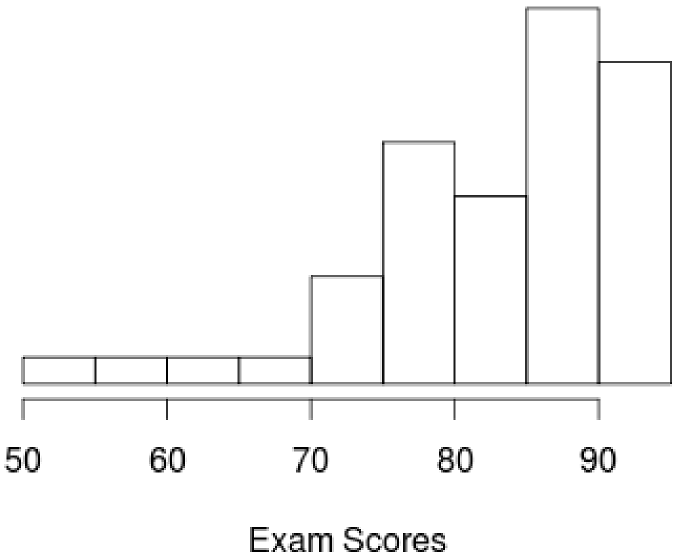
Section 2.3:

One Quantitative Variable – Measures of Spread

The exam scores shown in the histogram are



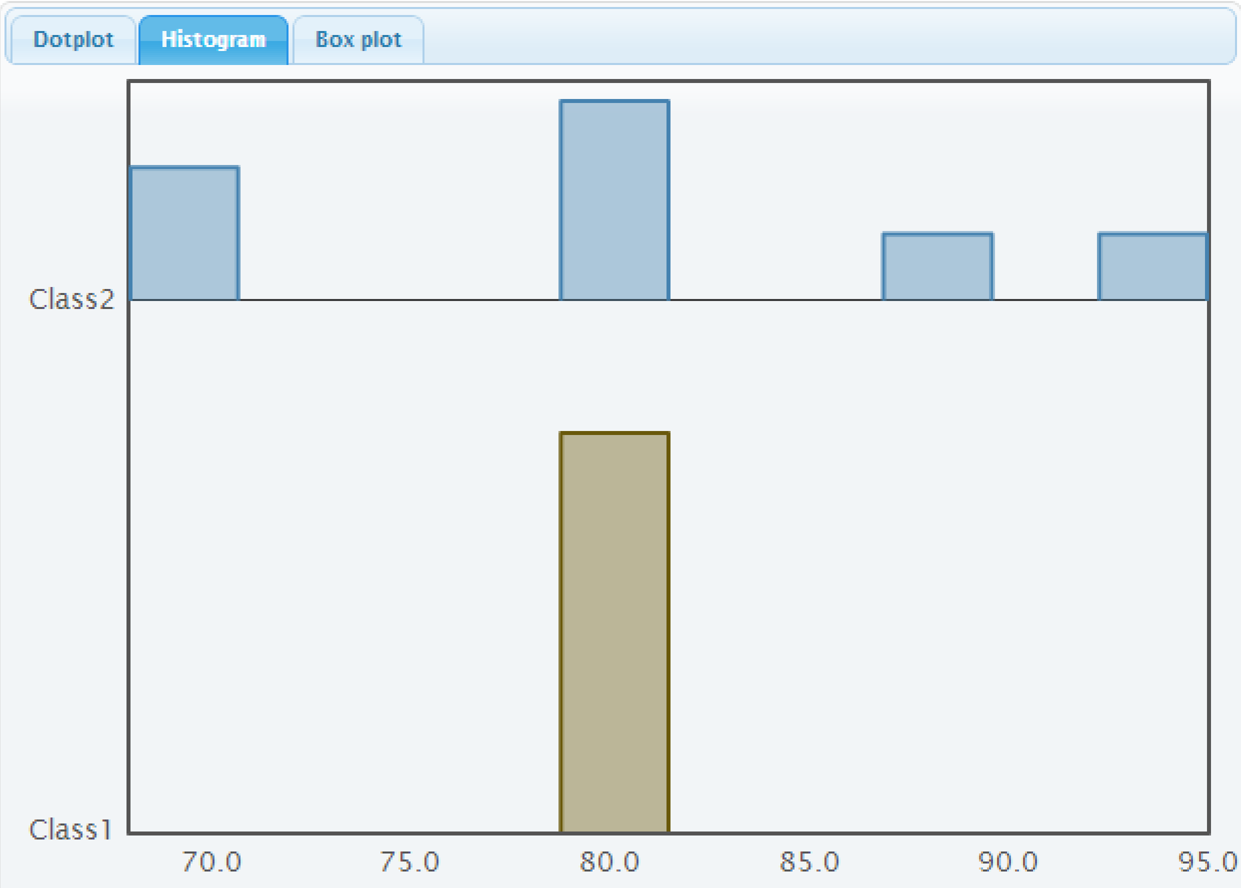
A. Symmetric
B. Left-skewed
C. Right-skewed

The Basics

Class 1	Class 2
80	68
80	70
80	80
80	80
80	80
80	87
80	95

- a) Consider the following two samples of quiz scores. What are their means? medians? modes?
- b) How are the samples different from each other?

The Basics



Summary Statistics

Statistics	Class1	Class2	Overall
Sample Size	6	7	13
Mean	80.0	80.0	80.0
Standard Deviation	0.000	9.3	6.6
Minimum	80	68	68
Q_1	80.00	75.00	80.00
Median	80.00	80.00	80.00
Q_3	80.00	83.50	80.00
Maximum	80	95	95

Histogram Controls Set Limits

Number of buckets: 7

lower bound	upper bound	frequency

Standard Deviation

The **standard deviation** for a quantitative variable measures the spread of the data

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

- **Sample** standard deviation: **s**
- **Population** standard deviation: **σ** (“sigma”)

Example: Our Only Time We Will Use the Formula

Sample Data: {1, 3, 3, 6, 7}

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

For our sample, $\bar{x} = 4$
and $n = 5$

$$= \sqrt{\frac{(x_1 - 4)^2 + (x_2 - 4)^2 + (x_3 - 4)^2 + (x_4 - 4)^2 + (x_5 - 4)^2}{5 - 1}}$$

$$= \sqrt{\frac{(1 - 4)^2 + (3 - 4)^2 + (3 - 4)^2 + (6 - 4)^2 + (7 - 4)^2}{4}}$$

Example: Our Only Time We Will Use the Formula

$$= \sqrt{\frac{(1-4)^2 + (3-4)^2 + (3-4)^2 + (6-4)^2 + (7-4)^2}{4}}$$

So What is a Standard Deviation?

$$x - \bar{x} = \text{ The "deviations" of each value from the mean}$$

$$\left(x - \bar{x}\right)^2 = \text{ Makes the deviations positive}$$

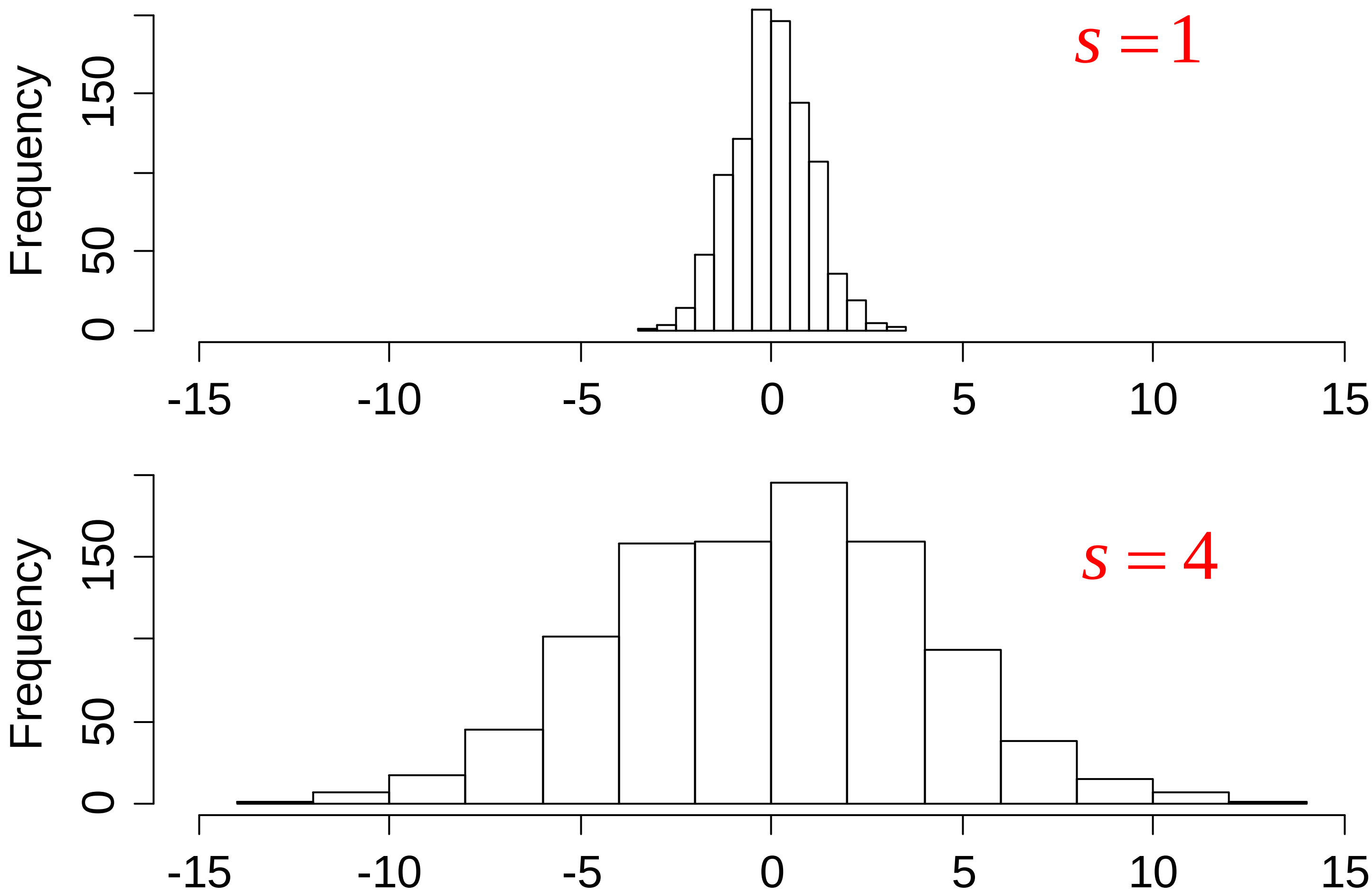
$$\frac{\Sigma\left(x - \bar{x}\right)^2}{n-1} = \text{ "Averages" the squared deviations}$$

$$\sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{n-1}} = \text{ Units are now the same as the original data}$$

So What is a Standard Deviation?

- The standard deviation gives a rough estimate of the typical distance of the data values from the mean
- The larger the standard deviation, the more variability there is in the data and the more spread out the data are.

Standard Deviation



Both of these distributions are ***bell-shaped***

Example 3: Arsenic in Toenails

- Arsenic is toxic to humans and people can be exposed to it through contaminated drinking water, food, dust, and soil. Scientists have devised an interesting new way to measure a person's level of arsenic poisoning: by examining toenail clippings.
- The table on the next slide gives toenail arsenic concentrations (in ppm) for 19 individuals with private wells in New Hampshire. (cont'd…)

Adapted from Karagas, M., et al., "Toenail Samples as an Indicator of Drinking Water Arsenic Exposure," *Cancer Epidemiology, Biomarkers and Prevention*, 1996; 5: 849-852.

Example 3: Arsenic in Toenails

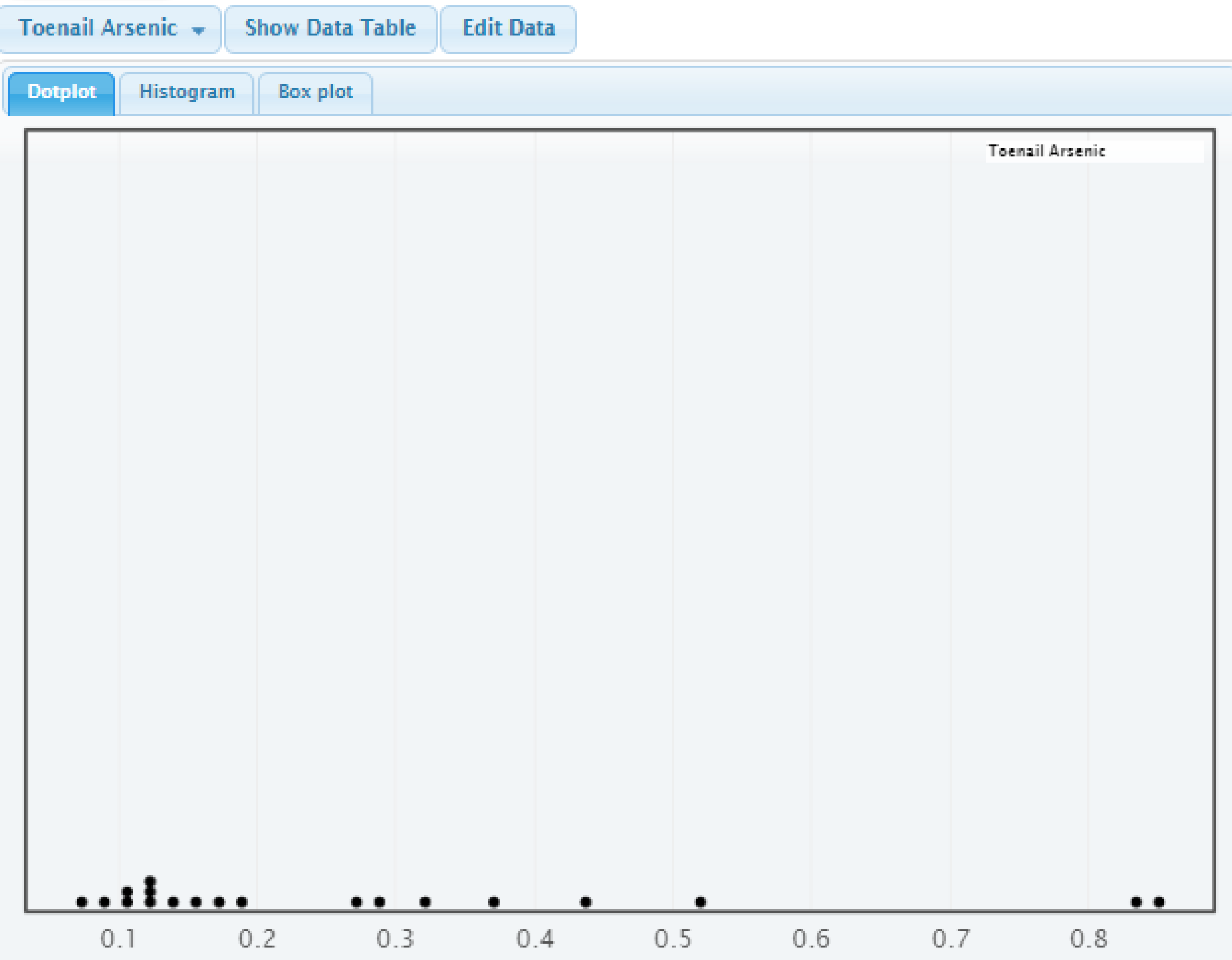
0.119	0.118	0.099	0.118	0.275	0.358
0.080	0.158	0.310	0.105	0.073	0.832
0.517	0.851	0.269	0.433	0.141	0.135
0.175					

a) Use technology to find the mean, median, max, and standard deviation.

Example 3: Arsenic in Toenails

[StatKey]

StatKey Descriptive Statistics for One Quantitative Variable



Summary Statistics

Statistic	Value
Sample Size	19
Mean	0.272
Standard Deviation	0.237
Minimum	0.073
Q_1	0.118
Median	0.158
Q_3	0.334
Maximum	0.851

Mean = 0.272

Median = 0.158

Max = 0.851

St. Dev = 0.237

95% Rule

If a distribution of data is approximately symmetric and bell-shaped, about 95% of the data should fall within two standard deviations of the mean.

- For a population, 95% of the data will be between

$$\mu - 2\sigma \text{ and } \mu + 2\sigma$$

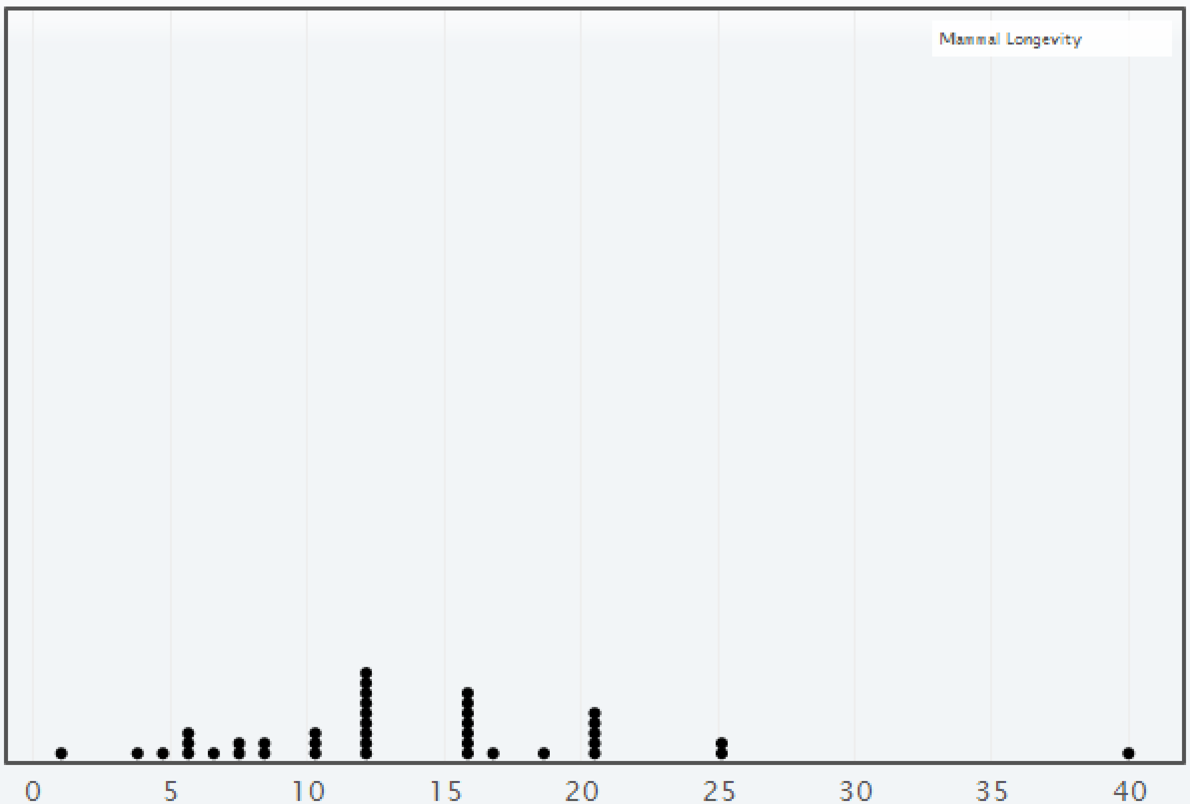
- StatKey

Example of 95% Rule

StatKey **Descriptive Statistics for One Quantitative Variable**

Mammal Longevity ▾ | Show Data Table | Edit Data

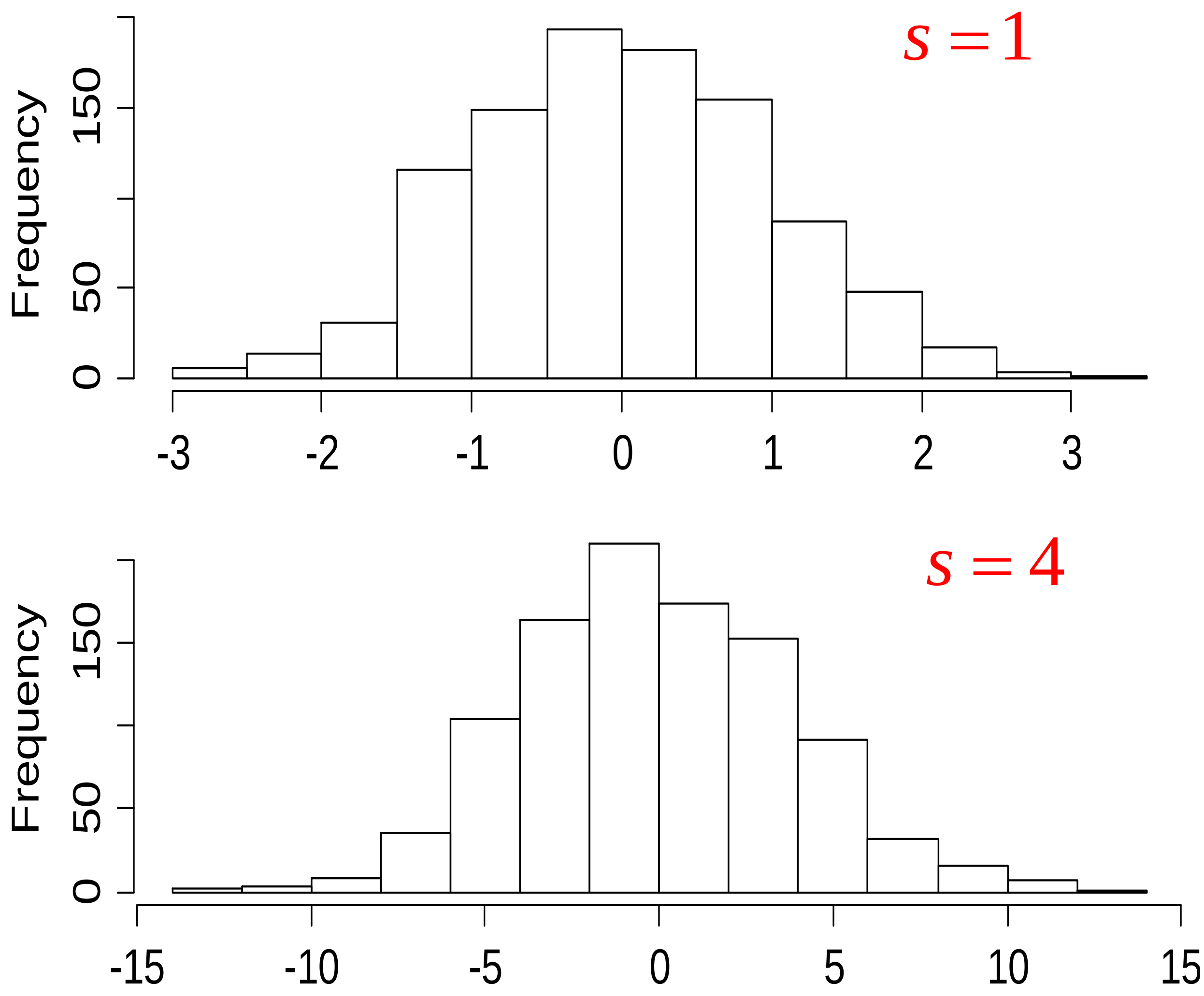
Dotplot | Histogram | Box plot



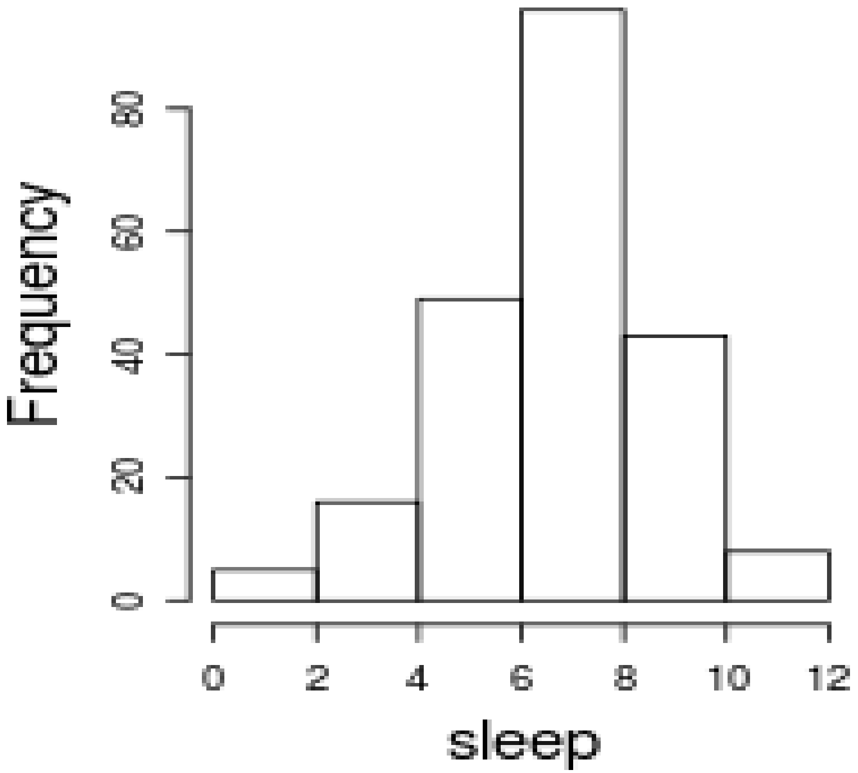
Summary Statistics

Statistic	Value
Sample Size	40
Mean	13.2
Standard Deviation	7.2
Minimum	1
Q_1	8.000
Median	12.000
Q_3	15.500
Maximum	40

The 95% Rule



$s = 1$
Frequency
0
50
150
-3
-2
-1
0
1
2
3
$s = 4$
Frequency
0
50
150
-15
-10
-5
0
5
10
15



The standard deviation for hours of sleep per night is closest to

1. ½
2. 1
3. 2
4. 4
5. I have no idea

Z-score

The ***z-score*** for a data value, x, is

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

- For a population, $\bar{x}$ is replaced with μ and s is replaced with σ
- Z-score values farther from 0 in either direction are more extreme

Z-score

- A z-score puts values on a common scale
- A z-score is the number of standard deviations a value falls from the mean
- 95% of all z-scores fall between what two values?

Which is better, an ACT score of 28 or a combined SAT score of 2100?

ACT: $\mu = 21, \sigma = 5$

SAT: $\mu = 1500, \sigma = 325$

Assume ACT and SAT scores have approximately bell-shaped distributions

1. ACT score of 28
2. SAT score of 2100
3. I don't know

Example 3: Arsenic in Toenails

0.119	0.118	0.099	0.118
0.275	0.358	0.080	0.158
0.310	0.105	0.073	0.832
0.517	0.851	0.269	0.433
0.141	0.135	0.175	

b) Compute the z-score for the largest concentration and interpret it.

Other Measures of Location

Maximum = largest data value

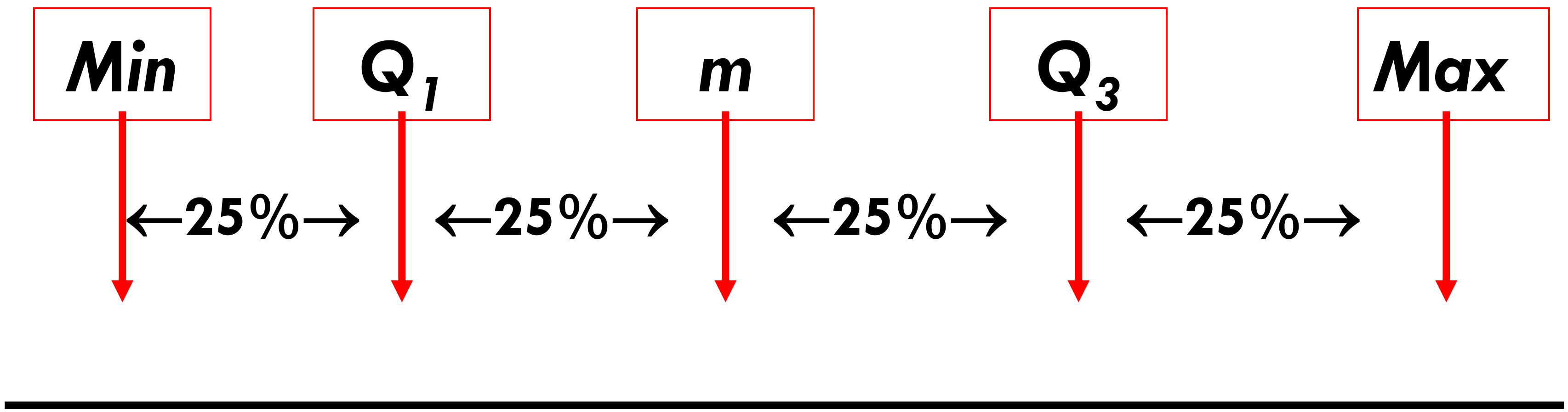
Minimum = smallest data value

Quartiles:

Q_1 = median of the values below m.

Q_3 = median of the values above m.

Five Number Summary



The five number summary of the number of hours that a sample of students reported studying is:

Min.	1st Qu.	Median	3rd Qu.	Max.
2.00	10.00	15.00	20.00	69.00

The distribution of number of hours spent studying each week is:

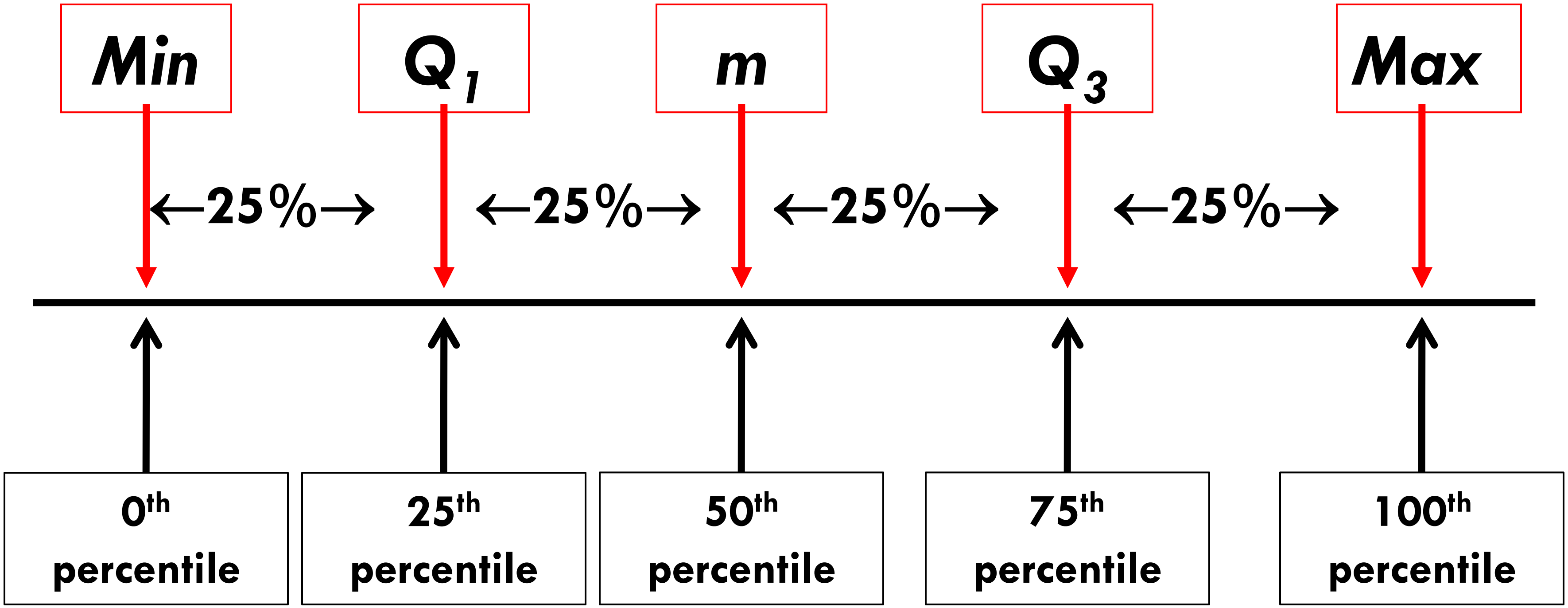
A. Symmetric
B. Right-skewed
C. Left-skewed
D. Impossible to tell

Percentile

The ***P^{th}* percentile** is the value which is greater than *P*% of the data

- We already used z-scores to determine whether an SAT score of 2100 or an ACT score of 28 is better
- We could also have used percentiles:
  - ACT score of 28: 91st percentile
  - SAT score of 2100: 97th percentile

Five Number Summary and Corresponding Percentiles



More Measures of Spread

Range = Max – Min

Interquartile Range (IQR) = $Q_3 - Q_1$

Is the range resistant to outliers?

1. Yes
2. No

Is the IQR resistant to outliers?

1. Yes
2. No

Comparing Statistics

- Measures of Center:
  - Mean (not resistant)
  - Median (resistant)
- Measures of Spread:
  - Standard deviation (not resistant)
  - IQR (resistant)
  - Range (not resistant)
- Most often, we use the mean and the standard deviation, because they are calculated based on all the data values, so use all the available information

Summary: One Quantitative Variable

- Summary Statistics
  - Center: mean, median
  - Spread: standard deviation, range, IQR
  - Measures of Location: z-scores, Percentiles, Quartiles

- Visualization
  - Dotplot
  - Histogram

- Other concepts
  - Shape: symmetric, skewed, bell-shaped
  - Resistance