Unit 1
Examining Distributions

What is Statistics?

Statistics is the set of methods for obtaining, organizing, summarizing, presenting and analyzing data.

Our **data** come from characteristics measured on **individuals** or **units**. These can be people, animals, places, things, etc.

The **population** is the totality of individuals about which we want information.

Sample & Population

A **sample** is a subset of the units in a population that we actually examine in order to gather information.

- **1000 voters** are asked which candidate they support in the upcoming election.
- **50 insomnia patients** are given a new treatment.
- **200 Canada geese** are tagged to study their migration patterns.

What is the **population** in each of these cases?

Practice Question

An MLA wants to know whether voters in her riding favour proposed legislation on new tax cuts. She asks her staff to mail out a questionnaire to a sample of 1000 voters in her riding. The population of interest is:

(A) the 1000 voters who receive the questionnaire.
(B) all voters in the MLA's riding.
(C) all voters in the province.
(D) all voters in the MLA's riding who favour the tax cuts.
(E) all voters who respond to the questionnaire.

Variables

A **variable** is a characteristic or property of an individual.

- **Time** until a light bulb burns out
- **Distance** traveled by a taxi driver in one day
- **Number of Heads** in five tosses of a quarter
- **Hair Colour**
- Your **Grade** in this course

Categorical Data

There are two broad classifications of data:

Categorical data represent values of **categorical variables** that place individuals into one of several groups.

- **Gender of a newborn baby**
- **Reason for taking this course**
- **Favourite television show**
- **Eye Colour**

Categorical Data

In some cases, there is a logical ordering to the values of a categorical variable:

- **Placing** in a hockey tournament (1st, 2nd, 3rd, etc.)
- **Service Rating** at a restaurant (Good, Fair, Poor)
- **Letter Grade** in a course (A+, A, B+, …, F)

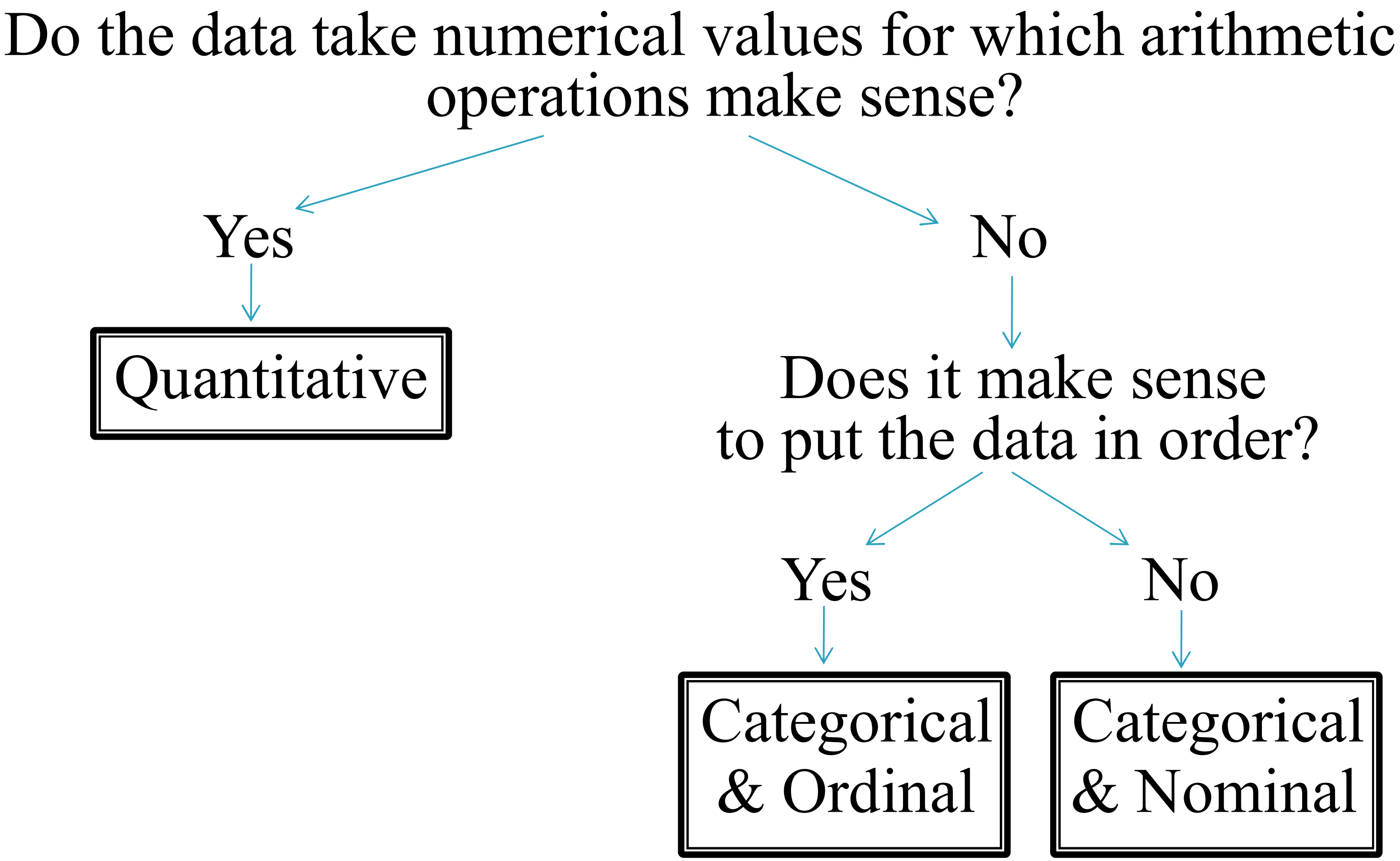
If ordering makes sense for the values of a categorical variable, it is called **categorical and ordinal**.
Otherwise, it is called **categorical and nominal** (like all variables on the previous slide).

Quantitative Data

Quantitative data represent values of **quantitative variables** for which arithmetic operations such as adding and averaging make sense:

- Final Exam **Score**
- **Height**
- **Volume** of air in a balloon
- **Sum** of the numbers shown on two rolled dice

Classifying Data/Variable Types



Do the data take numerical values for which arithmetic operations make sense?

Yes

Quantitative

No

Does it make sense to put the data in order?

Yes

Categorical & Ordinal

No

Categorical & Nominal

Example

State whether each of the following variables is

(i) Quantitative

(ii) Categorical and Nominal

(iii) Categorical and Ordinal

(a) Political Party a voter supports:
Liberal, Conservative, NDP, Green, etc.

(b) Speed of a passing vehicle

(c) Month of birth

Practice Question

A sample of elementary school students is asked to fill out a survey, which includes the following two questions:

(i) What is your **grade level**? (Kindergarten, Grade 1, 2, 3, 4, 5 or 6)
(ii) What is your **phone number**?

The variables (in **bold**) for these two questions are, respectively:

(A) (i) categorical & ordinal, (ii) categorical & ordinal
(B) (i) categorical & nominal, (ii) categorical & nominal
(C) (i) categorical & ordinal (ii) quantitative
(D) (i) categorical & nominal, (ii) categorical & ordinal
(E) (i) categorical & ordinal, (ii) categorical & nominal

Practice Question

Which of the following variables is **not** categorical and ordinal?

(A) Size of T-shirt purchased by a customer (S, M, L, XL, etc.)
(B) Condition of a patient (stable, serious, critical, etc.)
(C) Colour of an Olympic medal (gold, silver, bronze)
(D) Size of a farmer’s field (in acres)
(E) Date of first snowfall in Winnipeg next year

Data Distribution

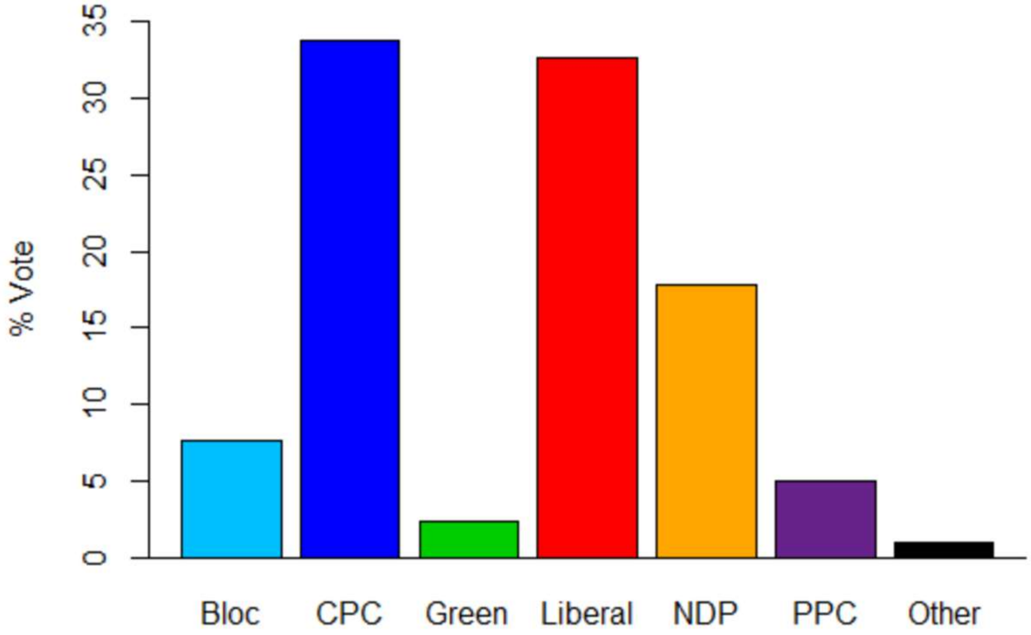
The **distribution** of a data set tells us what values a variable takes and how often it takes these values.

Note: In this sense, “value” doesn’t necessarily have to be quantitative. For example, “Blue” is a value of the variable Eye Colour.

There are several techniques we can use to display a data set in order to better study its features.

Bar Charts

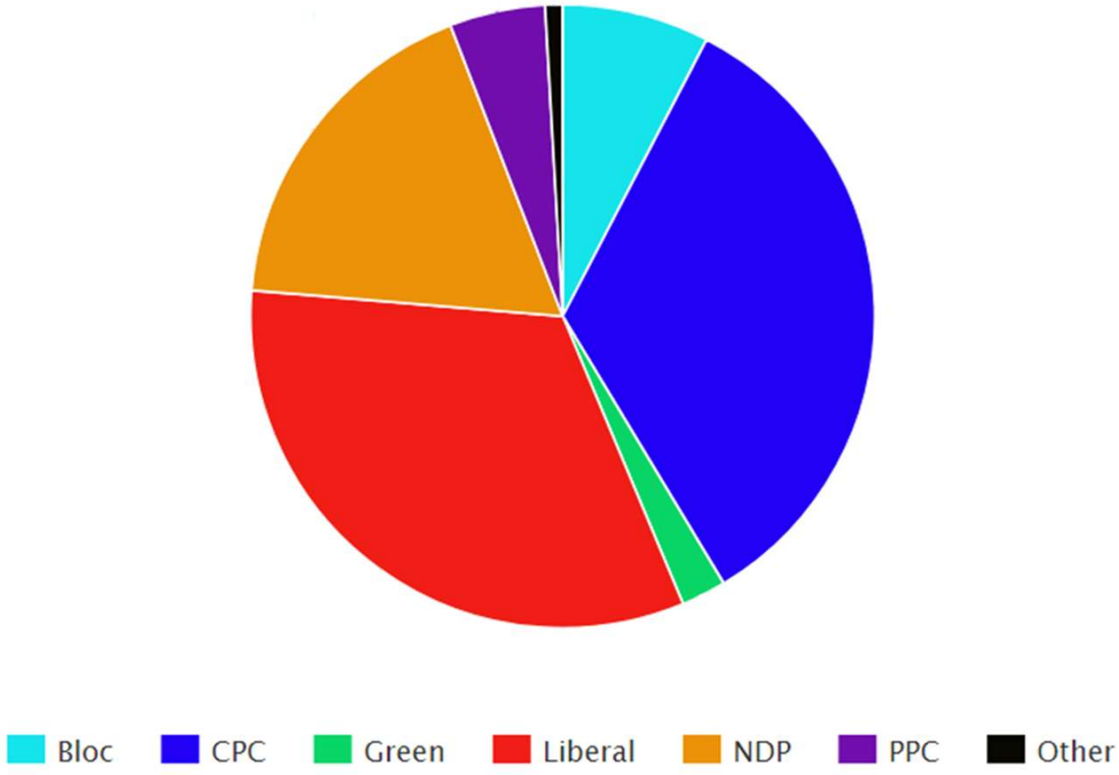
Bar charts display **categorical** variable values on one axis and frequencies on the other.



Note: There are spaces between bars so as not to imply continuity.

Pie Charts

Pie charts give us a visual representation of the relative frequency of the observed values for a categorical variable.



Example

Test scores for a class of 40 chemistry students are ordered and shown below:

31	37	40	44	49	50	51	53	56	56
62	64	67	67	68	68	69	70	71	72
73	73	74	75	77	78	78	81	82	84
84	87	89	89	92	92	94	95	96	98

Frequency Distribution

Consider the following **frequency distribution** for the data:

Score	Frequency
30 – 40	2
40 – 50	3
50 – 60	5
60 – 70	7
70 – 80	10
80 – 90	7
90 – 100	6

Frequency Distribution

A frequency distribution is a count of how many of our data values fall into various predetermined classes or intervals.

We choose the intervals ourselves. There is no correct choice of intervals, but we construct them in such a manner that we get a “nice” summary of our data. We usually use about 5 – 10 intervals.

Frequency Distribution

Our first interval must include the lowest data value (the **minimum**) and our last interval must include the highest data value (the **maximum**). All intervals should be of equal length.

Note that each interval includes the left endpoint but not the right.

i.e., 70 is included in the 70 – 80 interval.

Frequency Distribution

To avoid this problem, we could create the frequency distribution with non-overlapping intervals 30 – 39, 40 – 49, etc., but we would like the frequency distribution to reflect the **continuity** of the data.

For example, if a student got a score of 59.5, there would be no interval containing this value.

Types of Quantitative Variables

Note that there are in fact two types of quantitative variables.

A **continuous variable** can take any value within a given range. Examples of continuous variables include weight, age and distance.

A **discrete variable** can take only certain values. Examples of discrete variables include the number of children in a family, the number of days of rain in a month, and the highest denomination of bill in someone’s wallet.

Relative Frequency Distribution

A frequency distribution can be converted into a **relative frequency distribution** as follows:

Score	Frequency	Relative Frequency
30 – 40	2	2/40 = 0.050
40 – 50	3	3/40 = 0.075
50 – 60	5	5/40 = 0.125
60 – 70	7	7/40 = 0.175
70 – 80	10	10/40 = 0.250
80 – 90	7	7/40 = 0.175
90 – 100	6	6/40 = 0.150
	sum = 40	**sum = 1**

Relative Frequencies

By dividing the number of data values in each interval by the total number of data values (i.e., 40), we get the **relative frequency**, or **proportion** of individuals in the interval.

Proportions are values between 0 and 1, inclusive, and are **decimal representations of fractions**. Proportions convert to percentages when multiplied by 100. Note that the proportions for all intervals must add up to 1, since 100% of the data values are counted in this distribution.

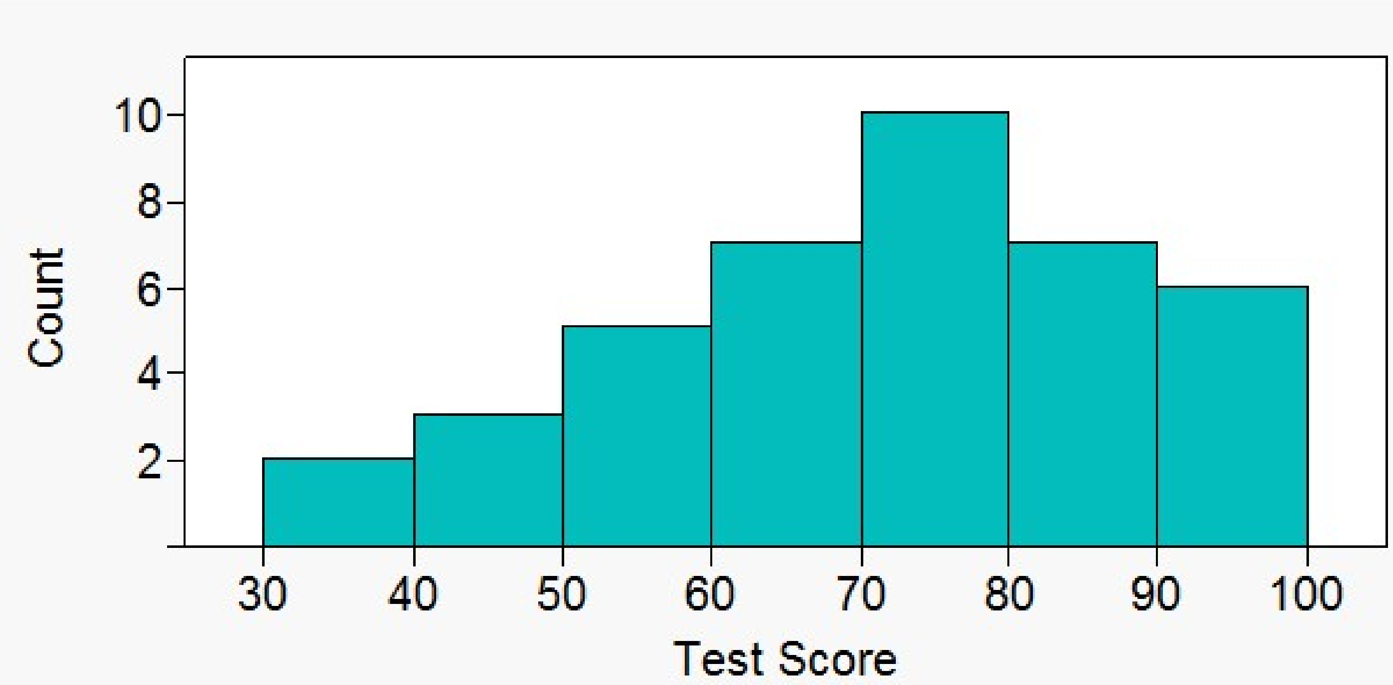
Histograms

If a frequency distribution is a representation of quantitative data, we can construct a more useful and more commonly used display known as a **histogram**.

Histograms are useful when we need to work with a large amount of data. They are graphical displays of the count (or proportion) of data values falling into each of several intervals.

Histograms

The following is a histogram for the test score data:



Histograms

A histogram is a form of bar graph (with no spaces between bars, so as to reflect the continuity of the data). The base of each rectangle represents the length of the interval. (All intervals should be of equal length.) The height represents the frequency (or relative frequency) of data values falling in the corresponding interval.

R Code

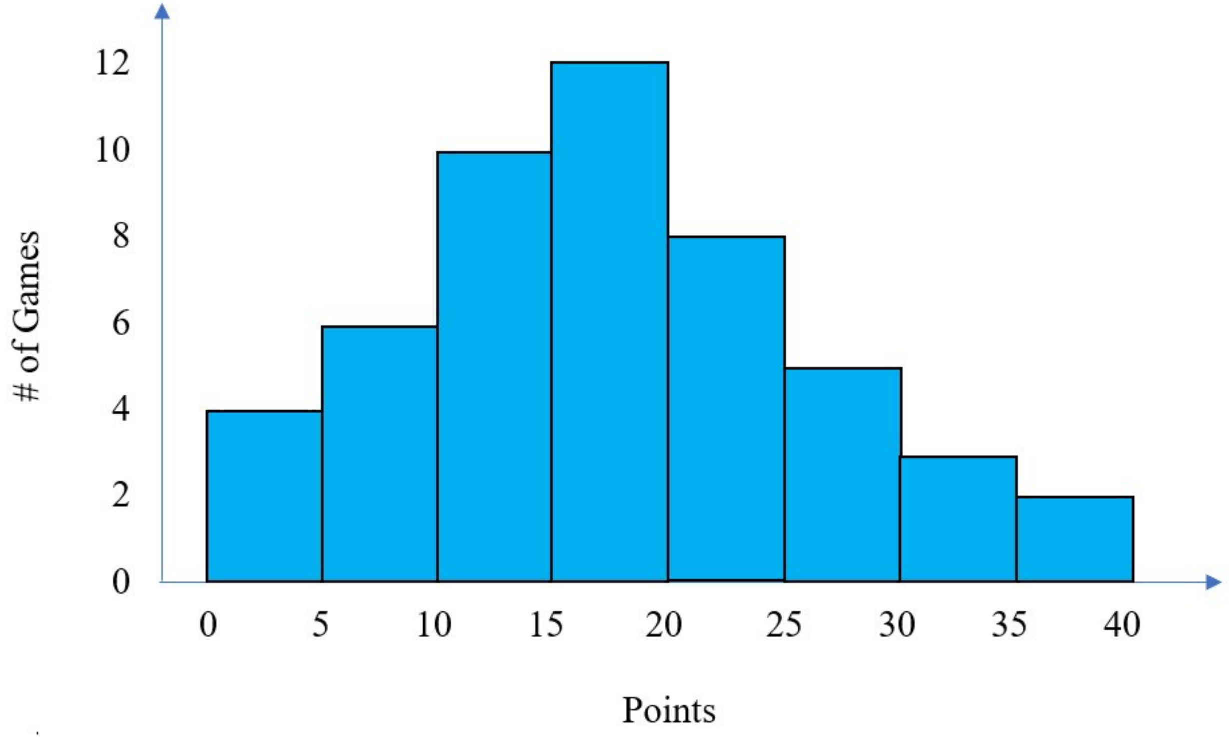
The histogram in the notes was not made with R, but we can make one in R with the following code:

```
> Score <- c(31, 37, 40, 44, 49, 50, 51, 53,
             56, 56, 62, 64, 67, 67, 68, 68,
             69, 70, 71, 72, 73, 73, 74, 75,
             77, 78, 78, 81, 82, 84, 84, 87,
             89, 89, 92, 92, 94, 95, 96, 98)

> hist(Score, col = "turquoise“, right = FALSE)
```

Practice Question

The points scored by a basketball player for a sample of 50 of his games are summarized below:



In what percentage of games did she score less than 15 points?

(A) 20% (B) 25% (C) 30% (D) 37.5% (E) 40%

Practice Question

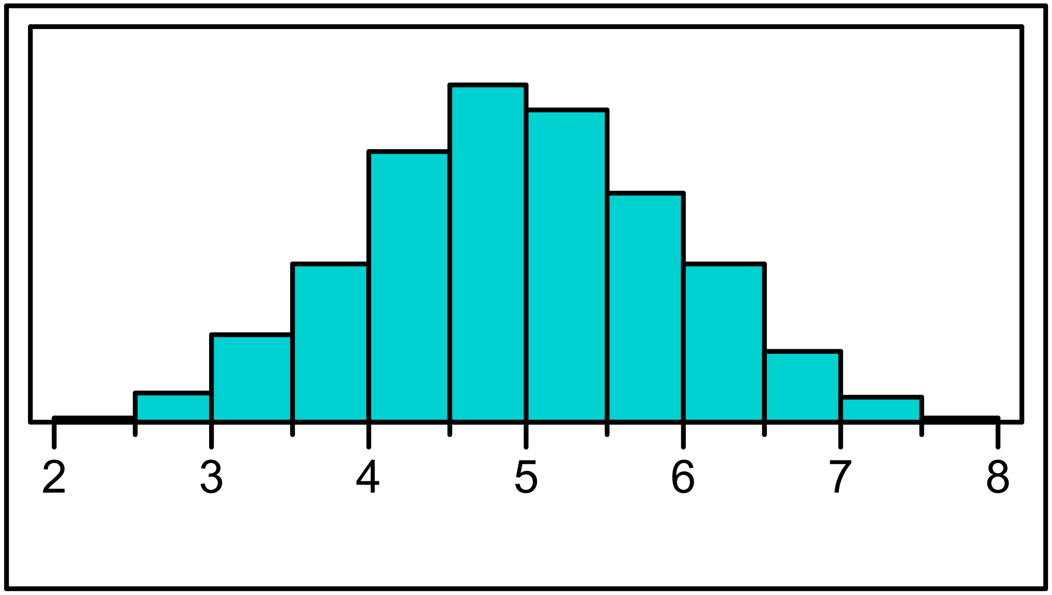
We would like to make a histogram (with vertical bars) of the annual salaries of all National Hockey League players who scored more than 20 goals last season. The horizontal and vertical axes represent, respectively:

(A) salary and number of players
(B) number of goals and number of players
(C) salary and number of goals
(D) number of goals and salary
(E) player and number of goals

Symmetry & Skewness

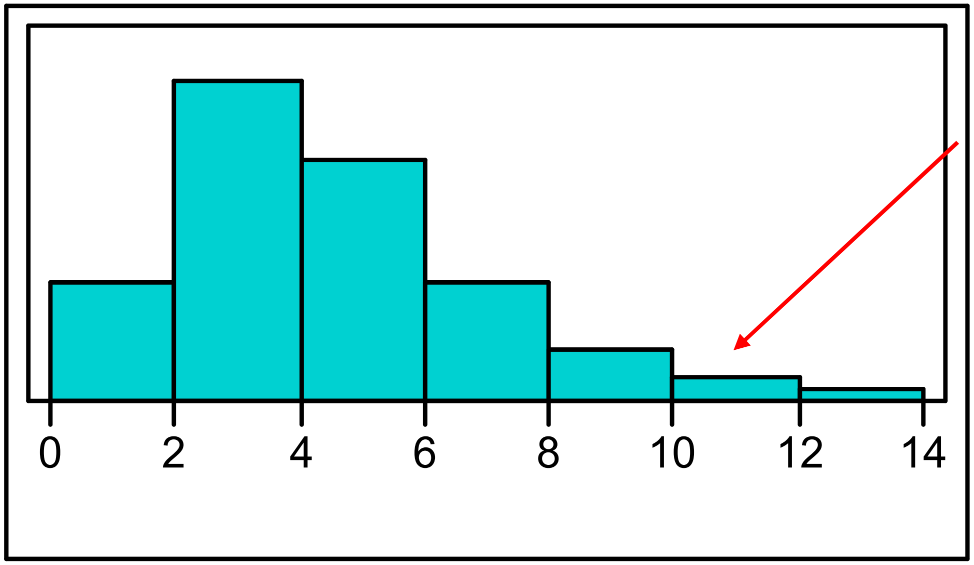
We can use a histogram to characterize the **shape** of a data distribution.

A distribution is said to be **symmetric** if its center divides it into two approximate mirror images.



Symmetry & Skewness

A distribution is said to be **skewed to the right** if the right side of the histogram (the larger half of the data values) extends much further out than the left side.

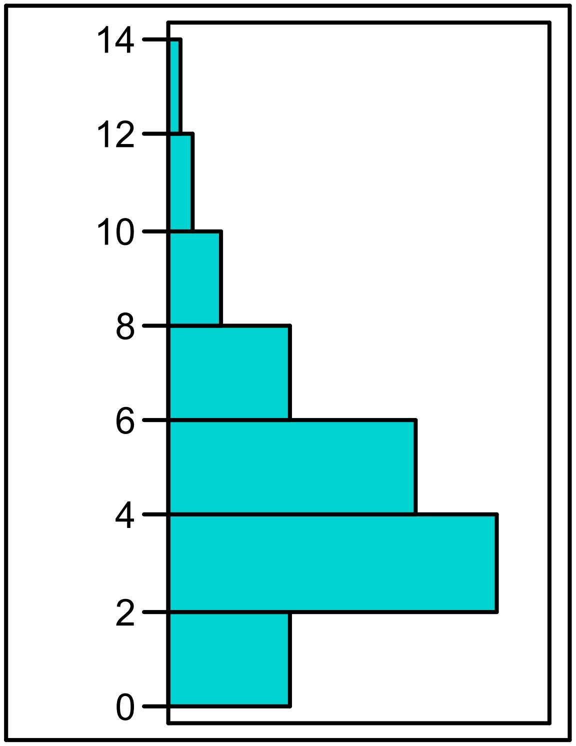


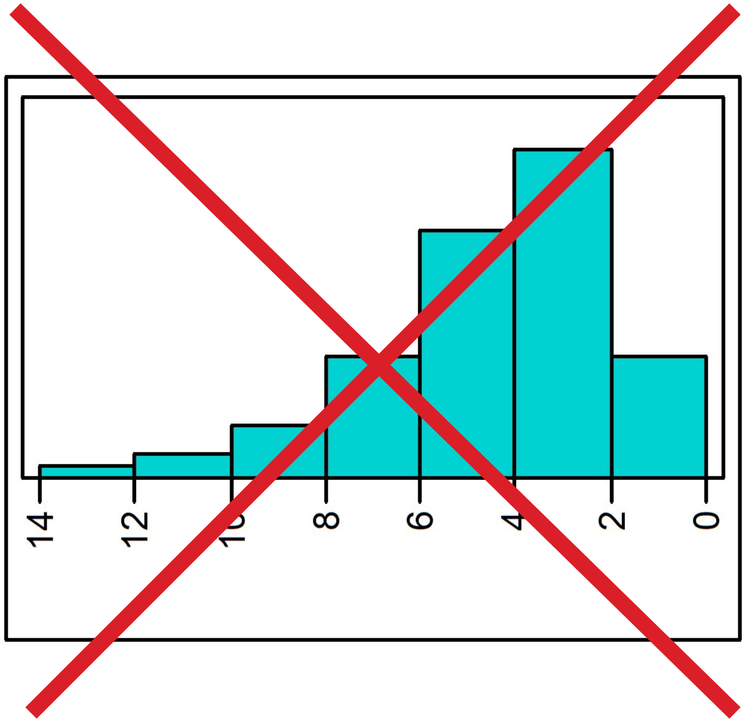
This is called The "tail" of the distribution.

The definition of **left skewness** follows analogously.

Symmetry & Skewness

If the histogram is displayed vertically, we must be careful interpreting the shape:





Practice Question

In a sample of 100 people, 25 have blue eyes, 35 have brown eyes, 20 have hazel eyes, 15 have green eyes and 5 have grey eyes. Which of the following statements about the distribution of eye colours of these people is true?

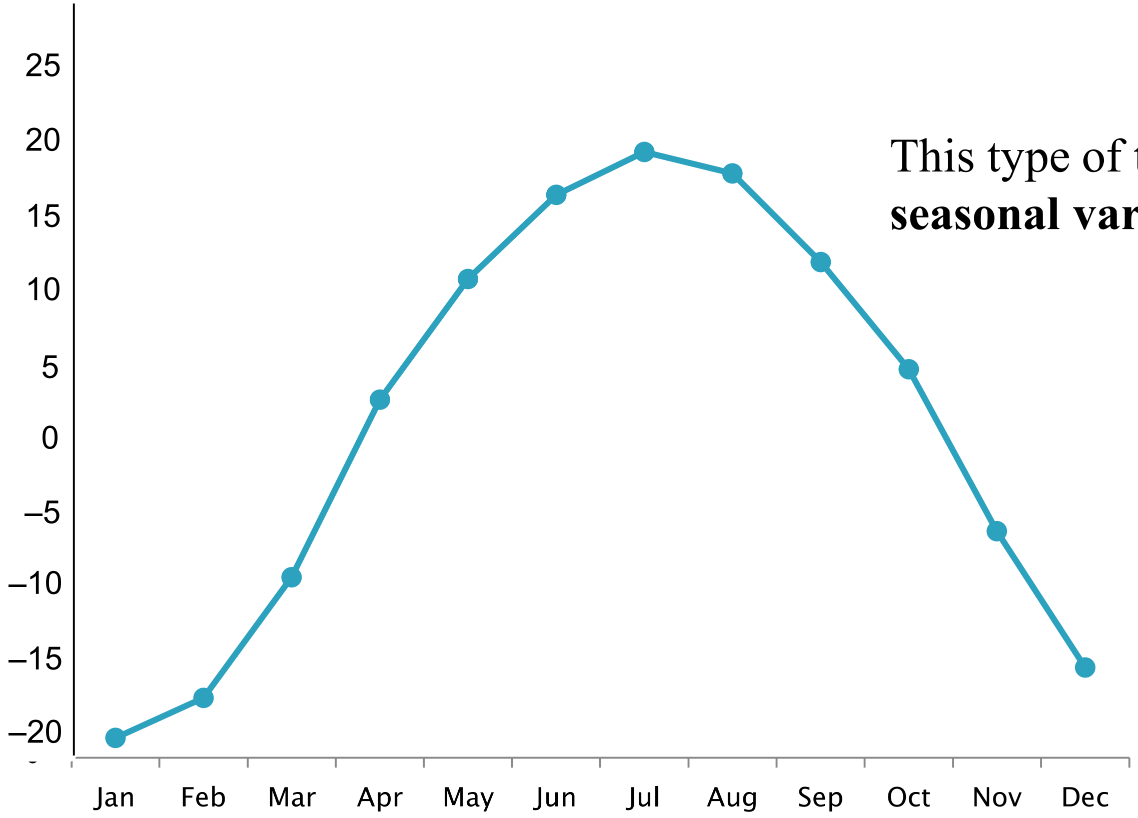
(A) A histogram would show the distribution is skewed to the left.
(B) A histogram would show the distribution is skewed to the right.
(C) A bar chart would show the distribution is skewed to the left.
(D) A bar chart would show the distribution is skewed to the right.
(E) It is meaningless to talk about the shape of this distribution.

Time Plots

Time plots are used for plotting **time series data**, which are values for some variable measured over time. Time is plotted on the *x*-axis, while variable values are plotted on the *y*-axis. Data values are represented by points, which are connected to better illustrate the pattern or **trend**.

Time Plots

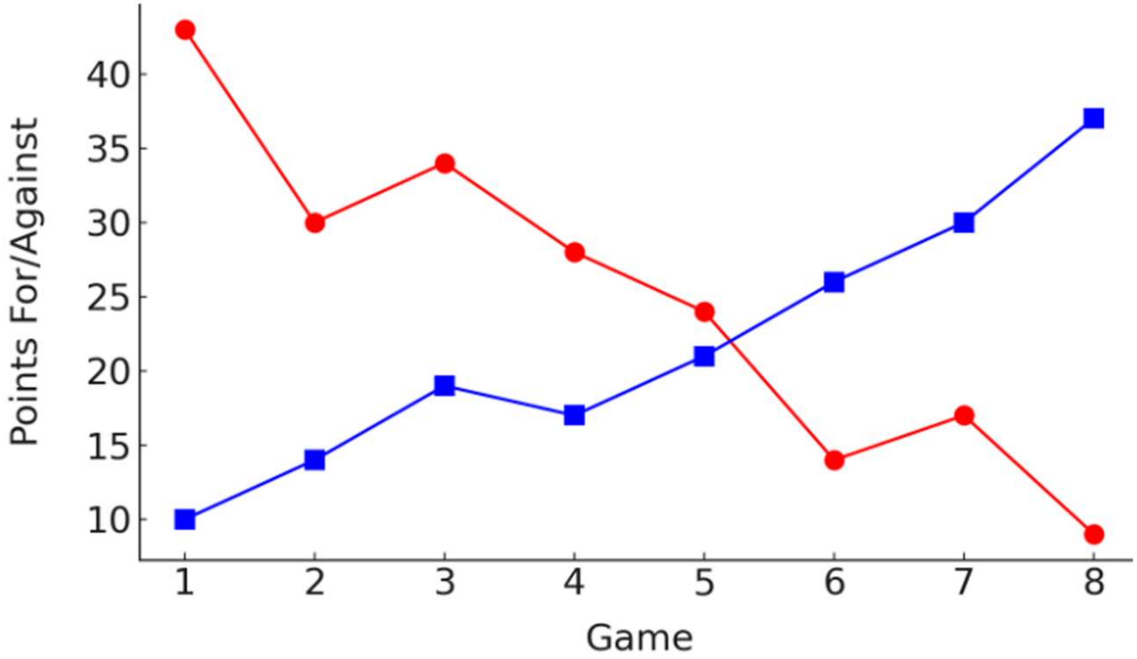
The timeplot below displays the average monthly temperature in Winnipeg over a one-year period.



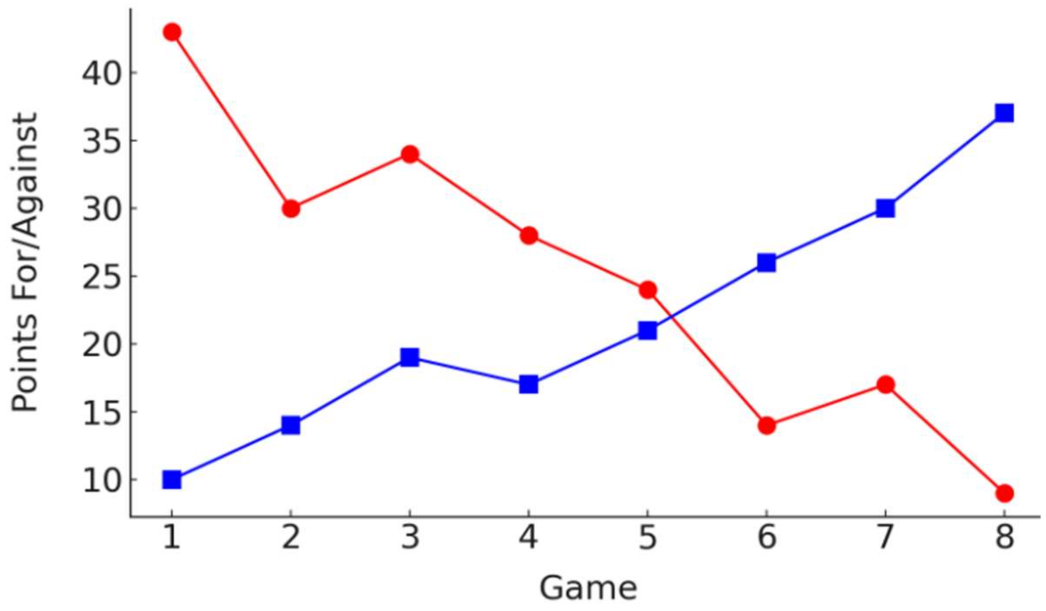
This type of trend is called **seasonal variation**.

Practice Question

A football team has played eight games so far this season. The timeplot below displays the number of points scored by the team (•) and the number of points their opponents scored against them (■) for each game:



Practice Question



Which of the following statements is/are true?

(I) There is an upward trend in the number of points the team has allowed per game.

(II) The team has won five games and lost three.

(III) The lines cross at one point, which means the team has tied one game.

(A) I only (B) II only (C) I and II only

(D) I and III only (E) I, II and III

Describing Distributions With Numbers

So far, we have seen some visual displays of data using bar charts, frequency distributions, histograms and time plots.

We will now examine some numerical summaries of a data set.

Two important features of a data set are its **location** and **variability**.

Measures of Location – Mode

Location is determined by where the **center** of our data falls.

We look at three measures of location:

The **mode** is the most frequently observed data value.

Example

A golfer records the distance (in yards) of a sample of 35 of his tee shots. The data are ordered and shown below:

168	177	178	184	188	189	193
194	196	199	199	201	203	206
206	207	207	208	212	215	217
217	217	219	222	223	227	229
233	236	240	241	244	247	254

Example

The mode for the golf data is **217**, as it is observed three times, more than any other value.

168	177	178	184	188	189	193
194	196	199	199	201	203	206
206	207	207	208	212	215	217
217	217	219	222	223	227	229
233	236	240	241	244	247	254

Note that it is possible for a data set to have more than one mode.

Measures of Location – Median

The **median** is the middle value in an ordered data set. Half of the data values are as small or smaller than the median and half of the values are as large or larger.

Measures of Location – Median

To find the median:

- Order the data from smallest to largest.
- Count n, the number of data values, and compute

$$\frac{n+1}{2}$$

- Count $\frac{n+1}{2}$ data values up from the lowest value.

Measures of Location – Median

If n is an odd number, the median is in position $\frac{n+1}{2}$.

If n is an even number, the median is the average of the two values on either side of position $\frac{n+1}{2}$.

Note: The median is **not equal** to $\frac{n+1}{2}$; it is the **data value** in that position.

Example

For the golf data, *n* = 35, so the median is in position (35 + 1)/2 = 18 of the ordered data.

168	177	178	184	188	189	193
194	196	199	199	201	203	206
206	207	207	**208**	212	215	217
217	217	219	222	223	227	229
233	236	240	241	244	247	254

The median is therefore equal to **208**.

R Code

```
> Distance <- c(168, 177, 178, 184, 188, 189, 193,
                194, 196, 199, 199, 201, 203, 206,
                206, 207, 207, 208, 212, 215, 217,
                217, 217, 219, 222, 223, 227, 229,
                233, 236, 240, 241, 244, 247, 254)

> median(Distance)
[1] 208
```

Example

For the test score data, *n* = 40, so the median is in position (40 + 1)/2 = 20.5 of the ordered data.

The median is the average of the data values in positions 20 and 21.

31	37	40	44	49	50	51	53	56	56
62	64	67	67	68	68	69	70	71	**72**
73	73	74	75	77	78	78	81	82	84
84	87	89	89	92	92	94	95	96	98

The median is therefore equal to (72 + 73)/2 = **72.5**.

Example

Consider the small data set

3, 8, 14, 15, 20

The median is 14, located in the third position.

Now consider the adjusted data set

3, 8, 14, 15, 200

Despite the extreme value, the median is **still** 14!

Outliers & The Median

Extreme values (also known as **outliers**) **do not affect the value of the median**.

For this reason, we say the median is **resistant** to the effect of outliers.

Practice Question

The following data are the magnitudes of a sample of earthquakes (measured on the Richter scale) around the world one year:

4.1 4.8 3.1 5.3 5.1 U 3.1 4.7 2.9 U 4.6 2.5

The equipment used is unable to measure magnitudes below 2.5, and so earthquakes with magnitudes below 2.5 are recorded as "U". The median magnitude of all earthquakes in this sample is:

(A) 3.1 (B) 4.65 (C) (U + 3.1)/2 (D) 4.95 (E) 3.6

Measures of Location – Mean

The third and most commonly used measure of center is the **mean**.

The mean, or **average**, is found by adding all of the *n* data values and then dividing by *n*.

Measures of Location – Mean

When our mean comes from a sample, we denote it as $\bar{x}$. The sample mean is calculated as follows:

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

The Greek symbol Σ (sigma) is used to indicate a summation. The formula tells us to add all the data values $x_1 + x_2 + \cdots + x_n$, and then divide by the sample size n.

Example

Consider again the small data set

3, 8, 14, 15, 20

The sample mean is equal to

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{3+8+14+15+20}{5} = \frac{60}{5} = 12$$

R Code

```
> Data <- c(3, 8, 14, 15, 20)

> mean(Data)
[1] 12
```

Example

Now consider again the adjusted data set

3, 8, 14, 15, 200

The sample mean is equal to

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{3+8+14+15+200}{5} = \frac{240}{5} = 48$$

Outliers & The Mean

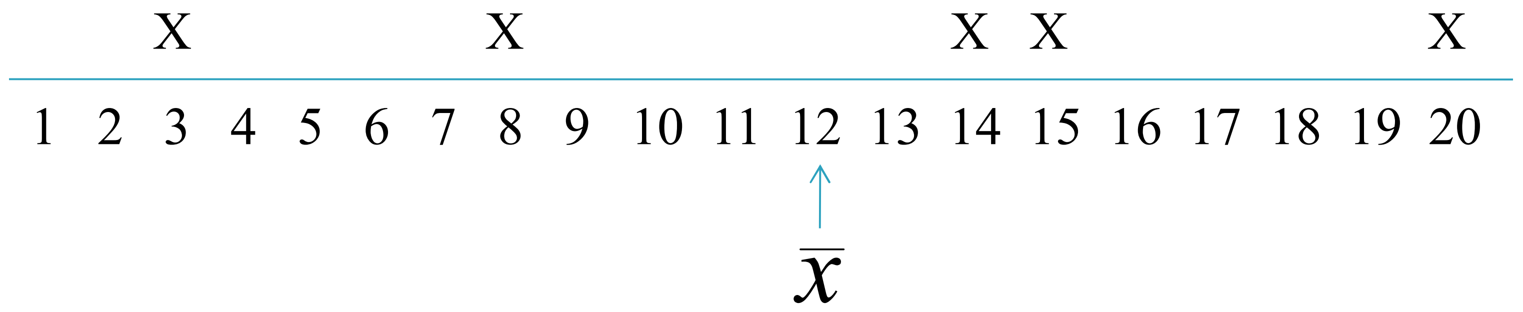
The value of the mean is strongly affected by the presence of the extreme value.

Therefore, the sample mean is **not** resistant to outliers.

The median measures the center of the data because it divides the data set into two halves of equal size. But how is the mean a measure of center?

Measures of Location – Mean

The mean is the “center of mass” or the “balance point” of the data:



This is where a teeter-totter would exactly balance if five people of equal mass were sitting in the positions of the five data points.

Example

The mean age of the six males in a class is 23.2 years. The mean age of the four females in the class is 21.7 years. What is the mean age for the whole class?

Example

Since there are more males than females, we can't simply take the average of the two means.

The mean age of the class is calculated as

$$\bar{x}_c = \frac{\sum_c x}{n_c} = \frac{\sum_m x + \sum_f x}{n_c}$$

Example

Now we must find the total ages of the males and the females separately:

$$\bar{x}_m = \frac{\sum_m x}{n_m} \quad \Rightarrow \quad \sum_m x = n_m \bar{x}_m = 6(23.2) = 139.2$$

Similarly,

$$\sum_f x = n_f \bar{x}_f = 4(21.7) = 86.8$$

Example

And so

$$\bar{x}_c = \frac{\sum_c x}{n_c} = \frac{\sum_m x + \sum_f x}{n_c} = \frac{139.2 + 86.8}{10} = \frac{226}{10} = 22.6$$

Practice Question

Now suppose that a student who is 26 years old drops the class, and two new students, who are both 21 years old, join the class. What is the new mean age of the class?

(A) 22.0 (B) 22.2 (C) 22.4 (D) 22.6 (E) 22.8

Example

There are ten dogs in an obedience training class. The average weight of the two bulldogs is 40 pounds. The average weight of the five golden retrievers is 51 pounds. The average weight of all ten dogs in the class is 38 pounds.

What is the average weight of the three pugs?

Example

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{\sum_{i=1}^{2} x_B + \sum_{i=1}^{5} x_G + \sum_{i=1}^{3} x_P}{10}$$

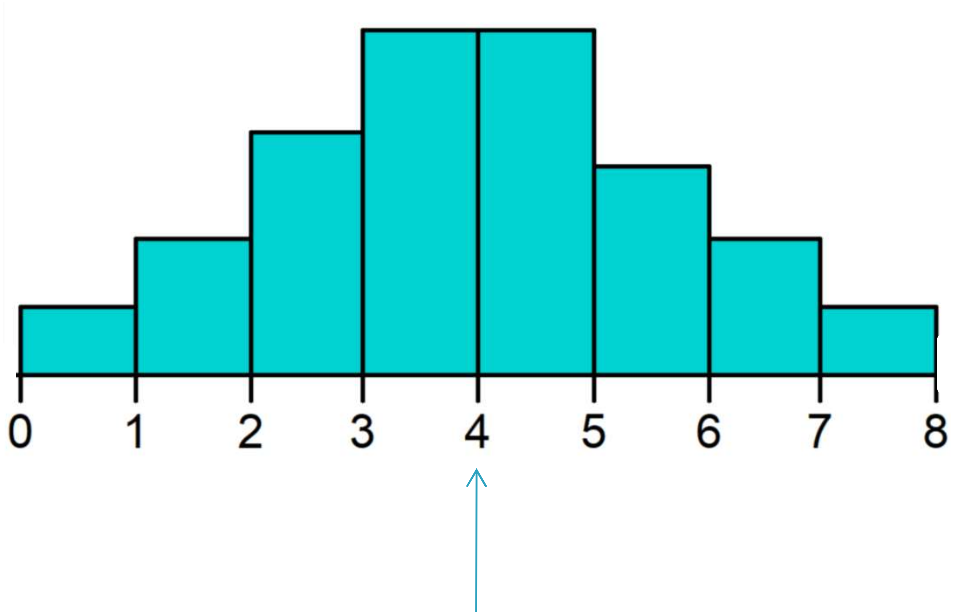
$$= \frac{n_B \bar{x}_B + n_G \bar{x}_G + n_P \bar{x}_P}{10} = \frac{2(40) + 5(51) + 3\bar{x}_P}{10} = 38$$

$$\Rightarrow \quad \frac{335 + 3\bar{x}_P}{10} = 38 \quad \Rightarrow \quad 335 + 3\bar{x}_P = 10(38) = 380$$

$$\Rightarrow \quad 3\bar{x}_P = 380 - 335 = 45 \quad \Rightarrow \quad \bar{x}_P = \frac{45}{3} = 15$$

Mean vs. Median

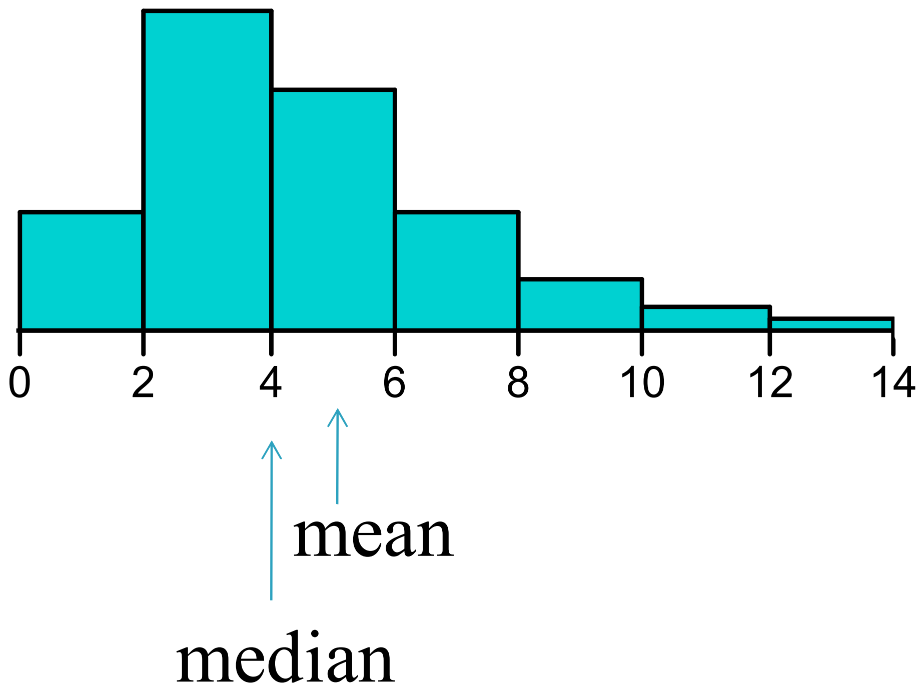
In a symmetric distribution, the mean and median are **equal**.



mean = median = 4

Mean vs. Median

In a skewed distribution, the mean is closer to the tail.



mean

median

Practice Question

The weights of 47 giant pumpkins entered into a competition at an agricultural fair are summarized in the following frequency distribution:

Weight	Number of Pumpkins
200-300	2
300-400	3
400-500	3
500-600	5
600-700	12
700-800	15
800-900	7

Practice Question

Which of the following statements is true?

(A) The distribution of weights is skewed to the right and so the median is greater than the mean.

(B) The distribution of weights is skewed to the left and so the median is greater than the mean.

(C) The distribution of weights is skewed to the right and so the mean is greater than the median.

(D) The distribution of weights is skewed to the left and so the mean is greater than the median.

(E) The distribution of weights is approximately symmetric and so the mean and median are approximately equal.

Weighted Mean

In some cases, when calculating the mean, some data values are given more weight than others. This may be due to the fact that some values are observed more frequently, or because some values are in some sense more important than others. In such cases, we calculate the **weighted mean**:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where w_i is the weight given to the i^{th} data value.

Example

A small restaurant has nine employees. The annual salary of the two chefs is $45,000. The annual salary of the six servers is $35,000. The annual salary of the restaurant manager is $60,000. What is the mean annual salary of all the restaurant's employees?

Example

We could calculate this using the formula for the regular sample mean. We would add the nine data values and divide by nine. It is simpler in this case to use the formula for the weighted mean:

$$\bar{x}_w = \frac{2(45{,}000) + 6(35{,}000) + 1(60{,}000)}{(2+6+1)}$$

$$= \frac{90{,}000 + 210{,}000 + 60{,}000}{9} = \frac{360{,}000}{9} = \$40{,}000$$

Example

A student's Grade Point Average is also calculated as a weighted mean. For example, consider one student who took the following courses one year and got the following grades:

Course	Credit Hours	Grade
STAT 1000	3	B+
PHED 3840	2	A+
PHIL 1200	6	C+
HIST 2050	3	B

Example

To calculate the student's GPA for the year, we can't just take the regular average, because it does not account for the fact that some courses (i.e., those worth more credit hours) count more than others towards a student's GPA. Instead, we take the weighted average, using credit hours as weights:

$$\text{GPA} = \bar{x}_w = \frac{3(3.5) + 2(4.5) + 6(2.5) + 3(3.0)}{(3 + 2 + 6 + 3)} = \frac{43.5}{14} = 3.107$$

Practice Question

There are two sections of a first-year math course. The table below gives the number of students in each section, as well as the midterm average for one of the sections:

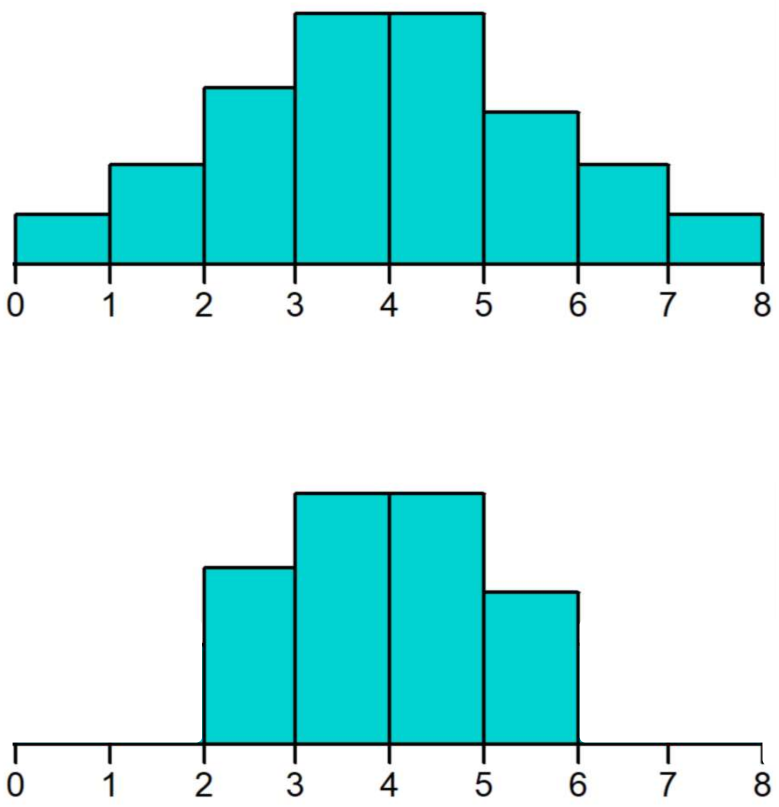
Section	# of students	midterm avg.
A01	140	67.0
A02	180	???

The midterm average of all students in the course was 64.3. What was the midterm average for A02?

(A) 61.6 (B) 61.8 (C) 62.0 (D) 62.2 (E) 62.4

Variability

How do the following two distributions differ?



Variability

They are both approximately symmetric, and so the mean and median are approximately equal, but the **variability** differs.

The variability of a data set refers to how spread out the data values are.

Variability is a very important concept in statistics. Just as we studied three measures of location, we will see several measures of variability.

Measures of Variability – Range

One simple measure of variability, or spread, is the **range** of the data

$$R = \text{maximum} - \text{minimum}$$

More than any descriptive measure we've seen so far, *R* is most heavily affected by extreme values (outliers).

Variability

In many cases, outliers arise from errors in measurement or data entry. In other cases, the outliers may be legitimate observations, but we may not be interested in including these extreme values in our numerical summary of the data. We would like a measure of spread that excludes outliers.

Measures of Variability – Interquartile Range

The **interquartile range** of a data set measures the length of an interval which covers the middle 50% of the ordered observations. It does not consider outliers (or in fact, any of the 50% of the observations furthest in position from the median).

Measures of Variability – Interquartile Range

The **first quartile *Q*1** of a data set is the ordered observation such that **at least 25%** of the data values are **as small or smaller** and **at least 75%** of the values are **as large or larger**.

The **third quartile *Q*3** of a data set is the ordered observation such that **at least 75%** of the data values are **as small or smaller** and **at least 25%** of the values are **as large or larger**.

Percentiles

In general, the ***p*th percentile** of a data set is the ordered observation such that **at least *p*%** of the data values are **as small or smaller** and **at least (100 – *p*)%** of the values are **as large or larger**.

As such, the first quartile is the 25th percentile and the third quartile is the 75th percentile.

Measures of Variability – Interquartile Range

To find $Q1$: Take the median of all data values lower in position than the median.

To find $Q3$: Take the median of all data values higher in position than the median.

Example

Consider again the golf data:

168	177	178	184	188	189	193
194	196	199	199	201	203	206
206	207	207	208	212	215	217
217	217	219	222	223	227	229
233	236	240	241	244	247	254

The minimum data value is **168** and the maximum is **254**. We previously calculated the median to be **208**.

Example

To find $Q1$, we take the median of all observations lower in position than the median. The median is in position 18 of the ordered data, so $Q1$ is the median of the first 17 ordered observations, i.e., position $(17 + 1)/2 = 9$. Therefore, $Q1 =$ **196**.

$Q3$ is the median of all observations higher in position than the median. We can either count nine positions up from the median, or nine positions down from the maximum. Therefore, $Q3 =$ **227**.

Five-Number Summary

The interquartile range is therefore

$$IQR = Q3 - Q1 = 227 - 196 = \mathbf{31}$$

We now have five numbers that describe this data distribution: the minimum, $Q1$, the median, $Q3$ and the maximum.

Together, this is known as the **five-number summary** and offers a good numerical description of the distribution of data values.

Example

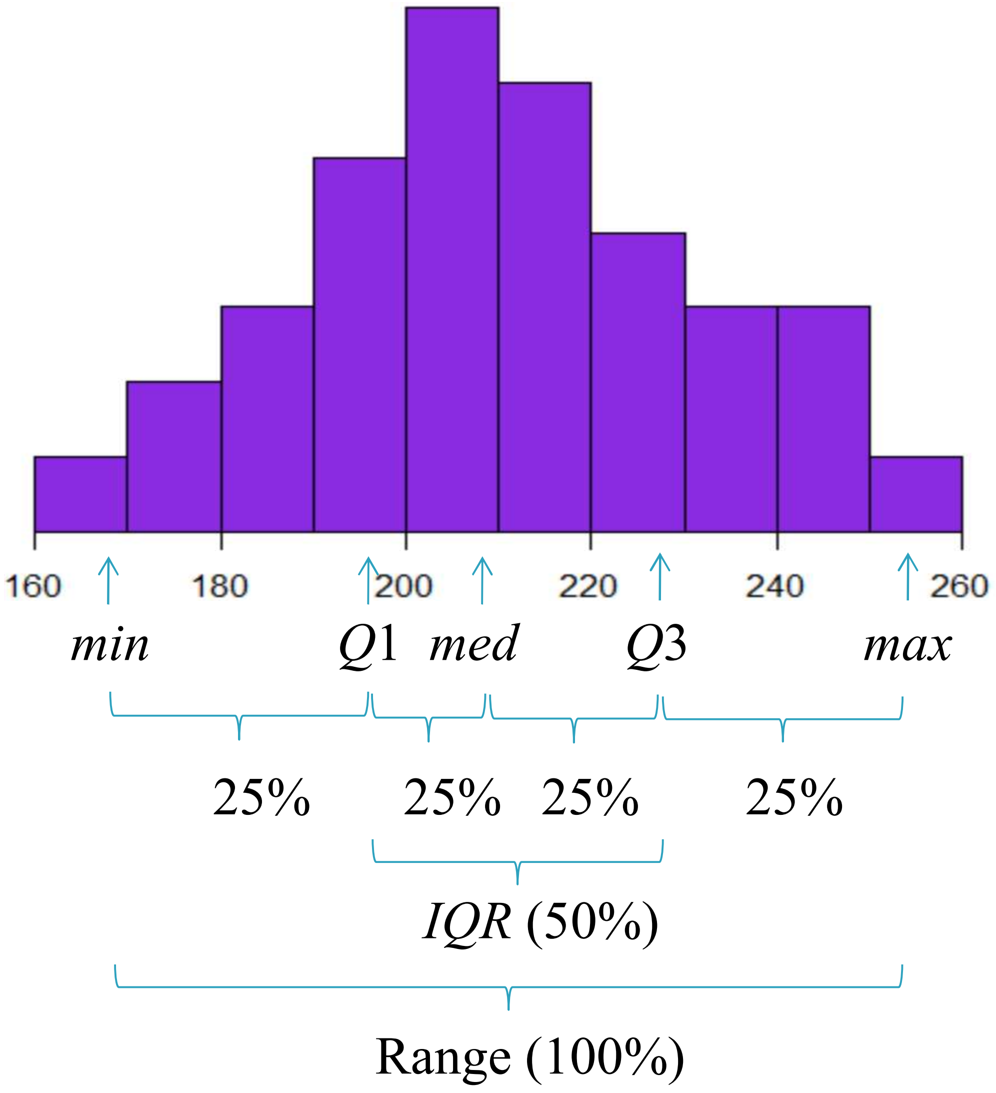
For the golf data, the five-number summary is

168 196 208 227 254

168	177	178	184	188	189	193
194	**196**	199	199	201	203	206
206	207	207	**208**	212	215	217
217	217	219	222	223	**227**	229
233	236	240	241	244	247	**254**

The five-number summary divides the data into four equal parts.

Five-Number Summary



min *Q*1 *med* *Q*3 *max*

25% 25% 25% 25%

IQR (50%)

Range (100%)

R Code

```
> fivenum(Distance)
[1] 168.0  197.5  208.0  225.0  254.0
```

Note that R uses a slightly different algorithm to calculate the quartiles than we do, so the values of $Q1$ and $Q3$ calculated by R will usually be slightly different than the values we calculate.

Example

Consider again the test score data:

31	37	40	44	49	50	51	53	56	56
62	64	67	67	68	68	69	70	71	72
73	73	74	75	77	78	78	81	82	84
84	87	89	89	92	92	94	95	96	98

The minimum data value is **31** and the maximum is **98**. We previously calculated the median to be **72.5**.

Example

There are 20 data values less than the median. *Q*1 is therefore in position (20 + 1)/2 = 10.5 of the ordered data, so we take the average of the 10th and 11th ordered observations. Thus, *Q*1 = (56 + 62)/2 = **59**.

Similarly, *Q*3 is the average of the 10th and 11th ordered values above the median (or equivalently, the average of the 10th and 11th values down from the maximum). Thus, *Q*3 = (84 + 84)/2 = **84**.

Example

The interquartile range is therefore

$$IQR = Q3 - Q1 = 84 - 59 = \mathbf{25}$$

and the five-number summary is

31 59 72.5 84 98

Practice Question

The ages of all 45 U.S. Presidents at the time of their (first) inauguration are ordered and shown below:

42 43 46 46 46 48 49 49 50 50
51 51 51 51 52 52 54 54 54 54
54 55 55 55 55 56 56 56 57 57
57 57 58 60 61 61 61 62 64 64
65 68 69 71 78

What is the interquartile range of inauguration ages?

(A) 8.0 (B) 8.5 (C) 9.0 (D) 9.5 (E) 10.0

Practice Question

A student's final exam score was equal to the 70th percentile in her class of 100 students. Therefore,

(A) approximately 30 students got 70% or less.

(B) approximately 70 students got 70% or less.

(C) approximately 30 students did as well or better than this student.

(D) approximately 70 students did as well or better than this student.

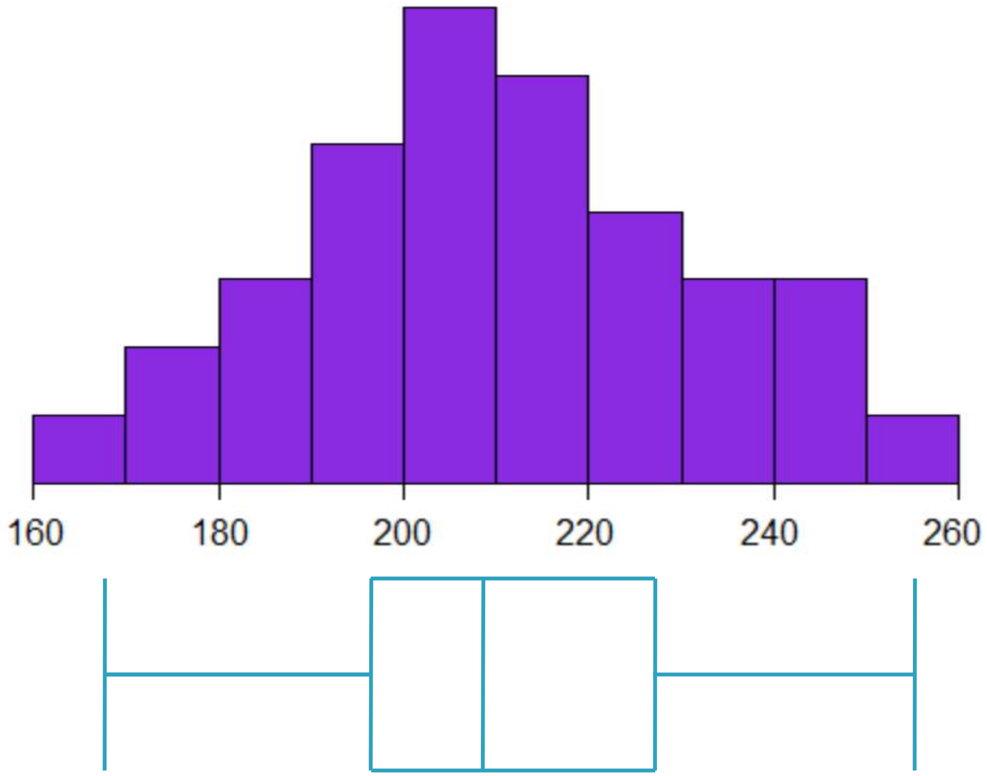
(E) the student received a score of 70% on her exam.

Five-Number Summary & Boxplots

The five-number summary describes the center, shape and spread of our data. We can use the five-number summary to get a “picture” of the data.

Quantile Boxplots

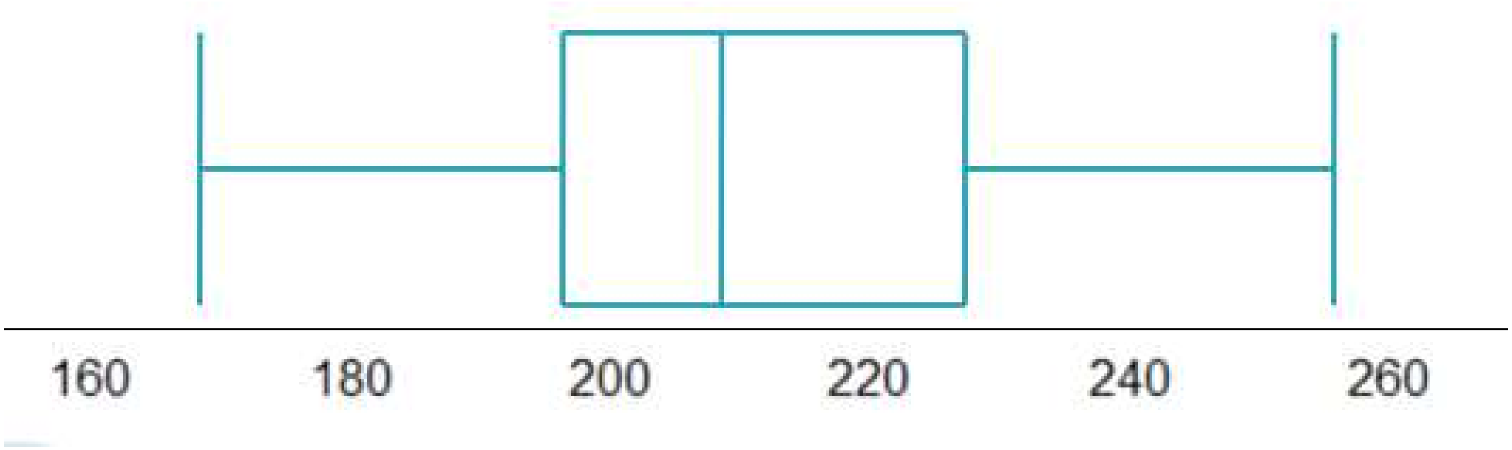
Below the histogram we see a boxplot for the golf shot data. This type of boxplot is called a **quantile boxplot**.



Quantile Boxplots

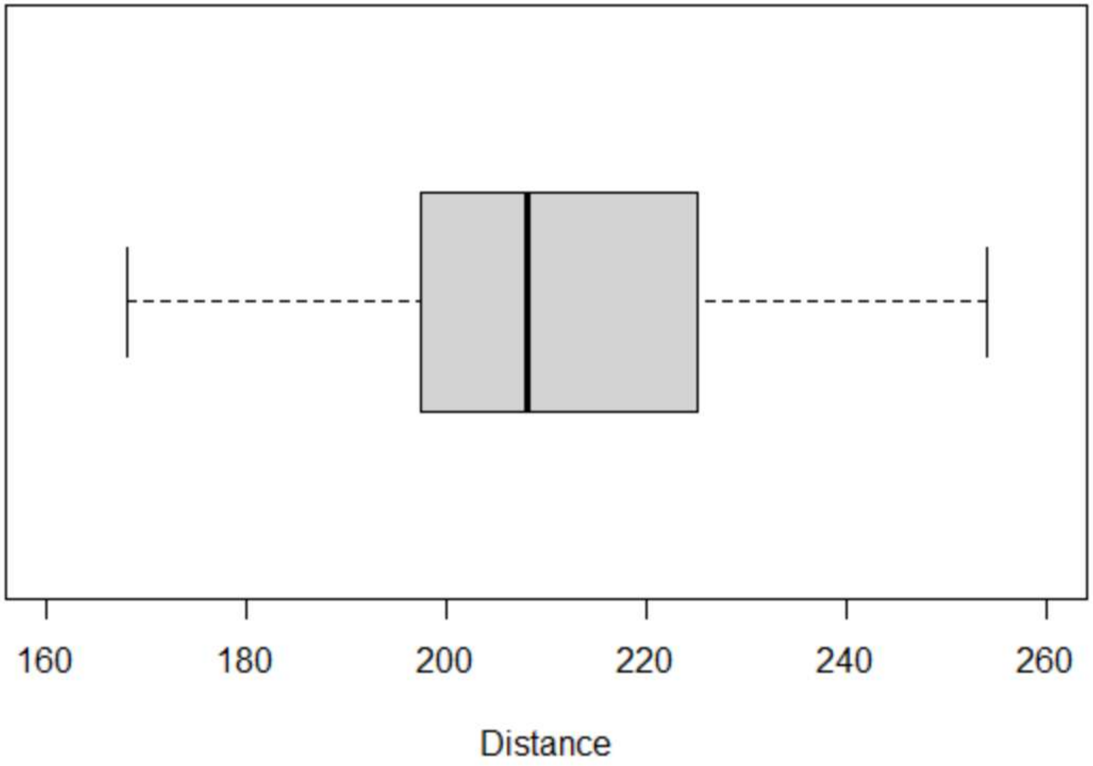
A quantile boxplot consists of:

- a line at the median
- a box that covers the IQR
- lines ("whiskers") that extend from the box out to the minimum and maximum



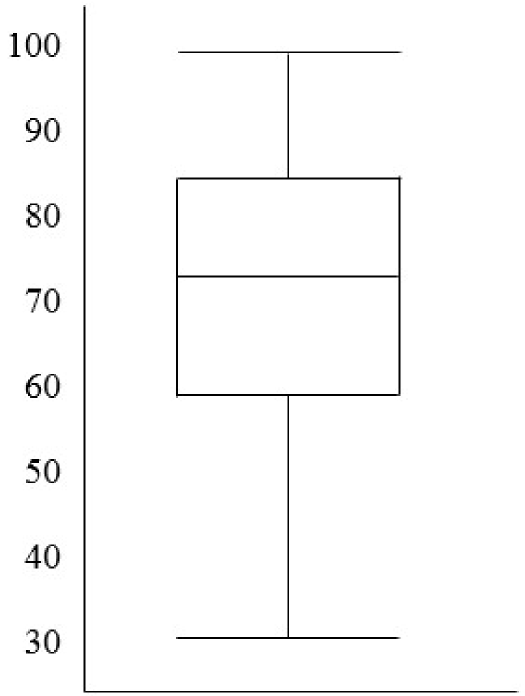
R Code

```
> boxplot(Distance, range = 0, horizontal = TRUE,
          ylim = c(160, 260), xlab = "Distance")
```



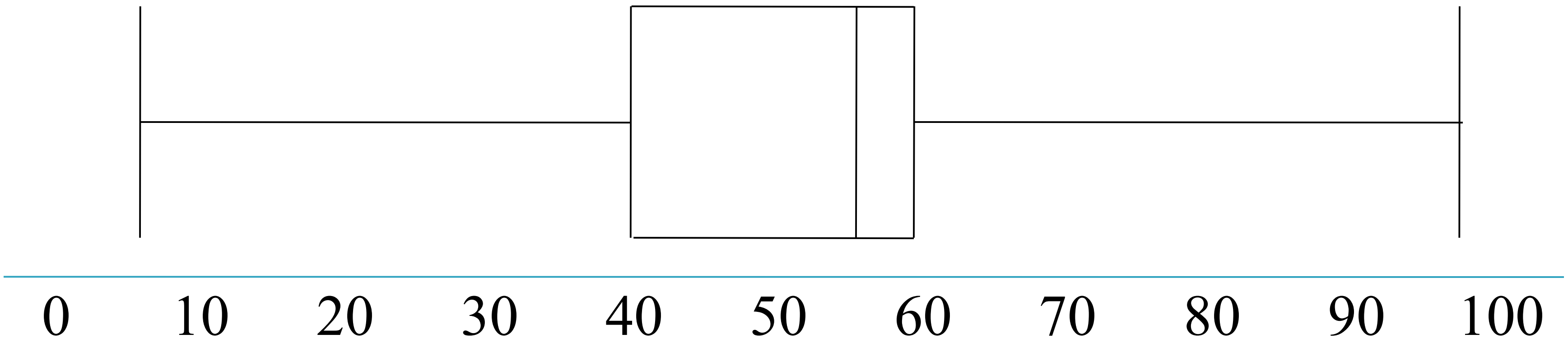
Quantile Boxplots

A boxplot can be displayed either horizontally or vertically. Below we see the quantile boxplot for the test score data in vertical orientation.



Practice Question

A quantile boxplot for the test scores of 120 Physics students is shown below:

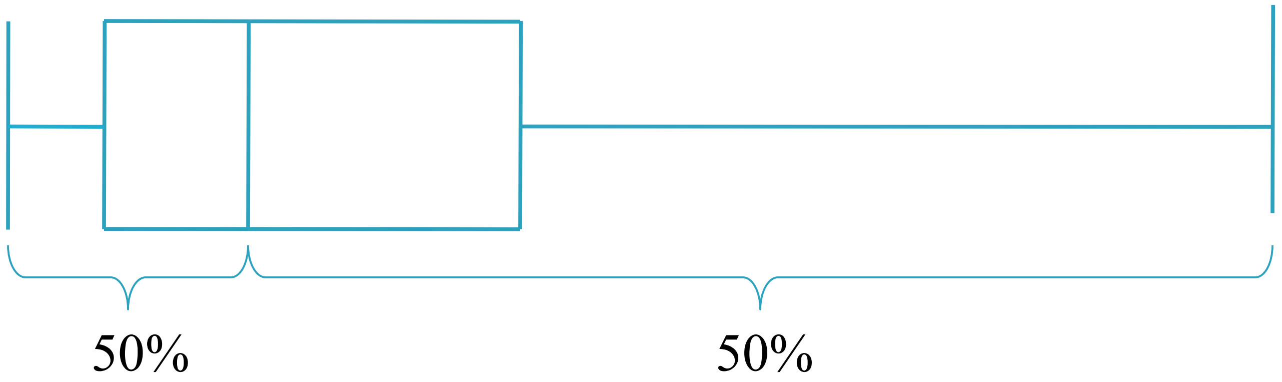


Approximately how many students scored above 40?

(A) 30 (B) 80 (C) 60 (D) 90 (E) 75

Boxplots – Shape

Like a histogram, a boxplot enables us to characterize the shape of a distribution:



Half of the data values fall within a relatively short interval on the left, while the other half covers a large interval on the right. This distribution is **skewed to the right**.

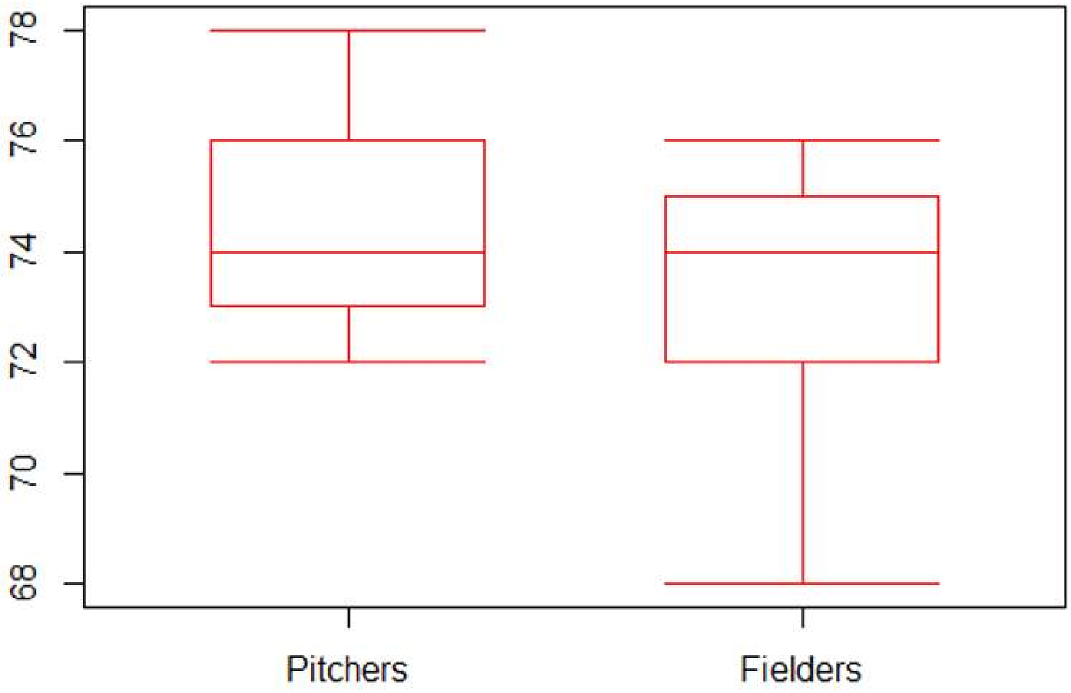
Side-by-Side Boxplots

We can compare two data distributions using **side-by-side boxplots**:

- used for plotting the same variable for samples representing different populations
- enable us to compare the distributions with respect to center, shape and spread
- must be constructed with a uniform scale to legitimize the comparison

Example

The side-by-side boxplots below compare the height distributions for Toronto Blue Jays pitchers and players in other fielding positions:.



R Code

```
> pitchers <- c(72, 75, 75, 73, 74, 72, 78, 76,
                76, 74, 74, 77, 73, 72)
> fielders <- c(68, 72, 72, 74, 72, 70, 74, 76,
                74, 75, 75, 76)
> position <- c("Pitchers", "Fielders")
> boxplot(pitchers, fielders, names = position,
          ylab = "Height")
```

Example

We see that the median heights for pitchers and fielders are equal.

The interquartile ranges for pitchers and fielders are equal, but the range for fielders is greater.

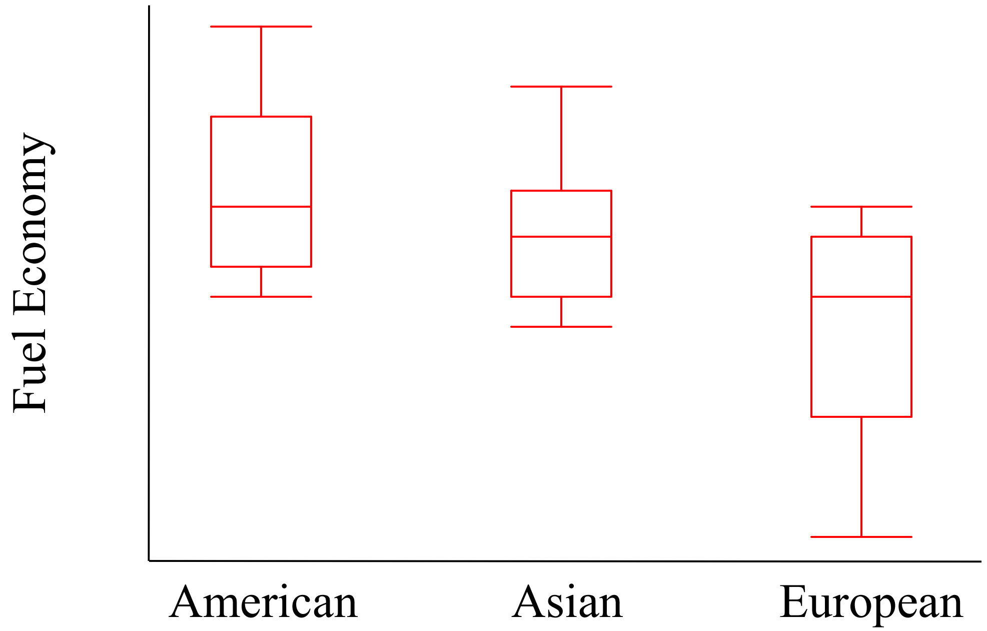
The distribution for pitchers is skewed to the right and the distribution for fielders is skewed to the left.

Although the medians are equal, we see that pitchers are more likely to be taller and fielders are more likely to be shorter.

Practice Question

Data on highway fuel economy (in miles per gallon) were obtained for several American, Asian and European automobiles produced in 2018. The data are displayed in the side-by-side quantile boxplots on the next page.

Practice Question



Which of the following statements is (are) true?

(I) At least 25% of Asian cars have fuel economies greater than the best European car.

(II) The distribution for American cars is the most variable.

(III) The distribution for European cars is skewed to the right.

(A) I only (B) II only (C) III only

(D) I and III only (E) I, II and III

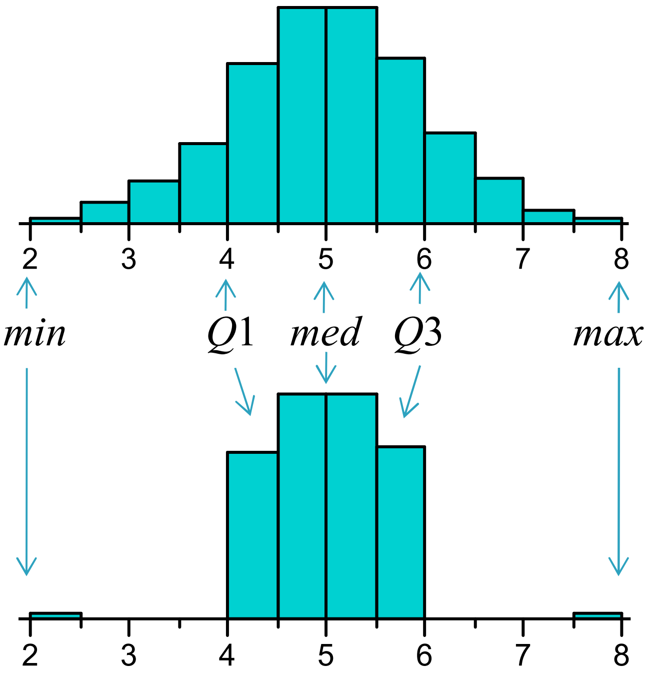
Boxplots & Outliers

As we have seen, extreme observations can affect our interpretation of the data if a numerical measure is not resistant to the effect of outliers.

We can have a similar problem with graphical tools. Since the lines in our quantile boxplot extend out to the minimum and maximum, we might not get an accurate picture of our data.

Boxplots & Variability

The two distributions shown below will have similar looking quantile boxplots, even through they are quite different in terms of variability.



Outlier Boxplots

We would like to construct a modified boxplot that takes these extreme observations into account.

An **outlier boxplot** is also based on quantiles.

Outlier Boxplots

In a quantile boxplot, the five lines correspond to the values in the five-number summary.

The middle three lines (the box) are the same for an outlier boxplot: $Q1$, median, $Q3$.

In this case, however, the lines coming out from the box (the whiskers) do not extend out to the minimum and maximum values.

Outlier Boxplots

Instead, we construct a “fence” with lower (*LF*) and upper (*UF*) values:

$$LF = Q1 - 1.5(IQR) = Q1 - 1.5(Q3 - Q1)$$

$$UF = Q3 + 1.5(IQR) = Q3 + 1.5(Q3 - Q1)$$

Outlier Boxplots

The line coming out from the left side of the box extends out to the lowest data value which is still greater than the lower fence.

Similarly, the line coming out from the right side of the box extends to the highest data value which is still less than the upper fence.

Outlier Boxplots

An outlier is considered to be any data point falling **outside the fences** (i.e., less than the lower fence or greater than the upper fence). As such, the lines extend out to the “new” minimum and maximum, after we remove the outliers from consideration.

The outliers are included in the boxplot as points along the axis.

Example

We record the number of cars caught speeding over the weekend by red-light cameras at 12 intersections. The data are ordered and shown below:

6	9	10	10	12	13
14	14	15	15	17	28

The five-number summary for these data is:

6 10 13.5 15 28

Example

We calculate the fences as follows:

$$
\begin{aligned}
LF &= Q1 - 1.5(IQR) = 10 - 1.5(15 - 10) \\
&= 10 - 7.5 = \mathbf{2.5}
\end{aligned}
$$

$$
\begin{aligned}
UF &= Q3 + 1.5(IQR) = 15 + 1.5(15 - 10) \\
&= 15 + 7.5 = \mathbf{22.5}
\end{aligned}
$$

Example

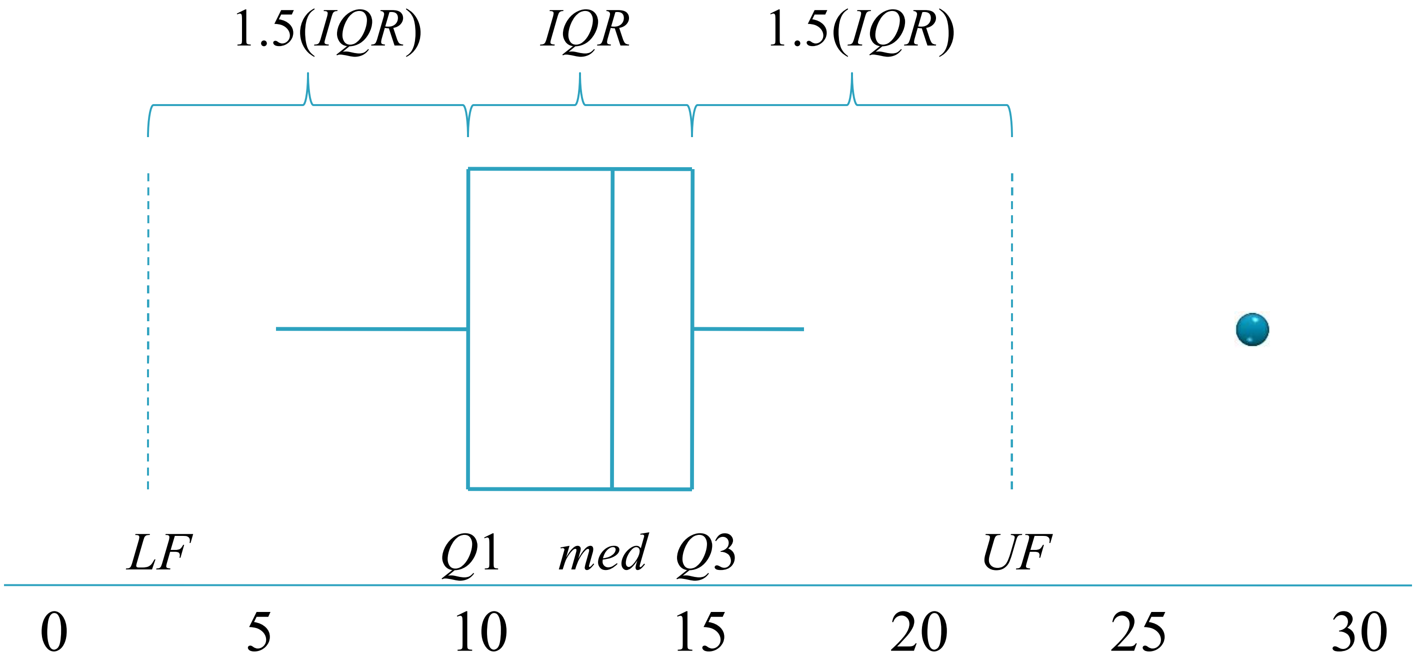
There are no data values less than LF, so there are no outliers on the left. The data value closest to but not less than LF is 6 (i.e., the "new minimum" is the same as the "old minimum").

There is one value greater than UF, $28 > 22.5$, so 28 is an outlier. The data value closest to but not greater than UF is 17 (i.e., the "new maximum").

Note that we draw the whiskers up to the "new" minimum and maximum, **not the fences**.

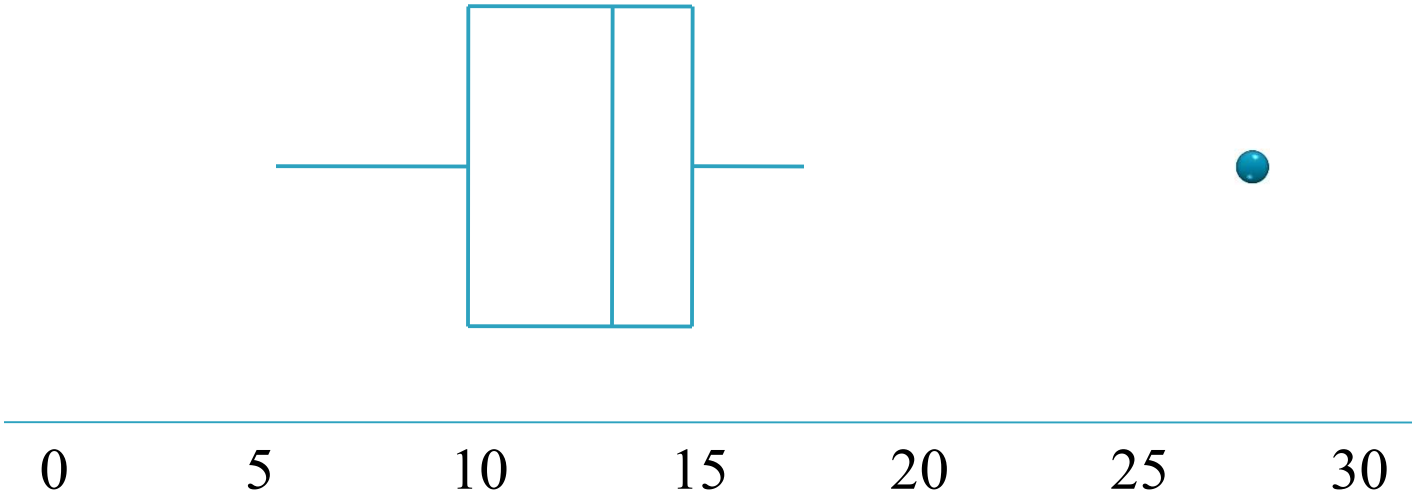
Example

The outlier boxplot for these data is shown below:



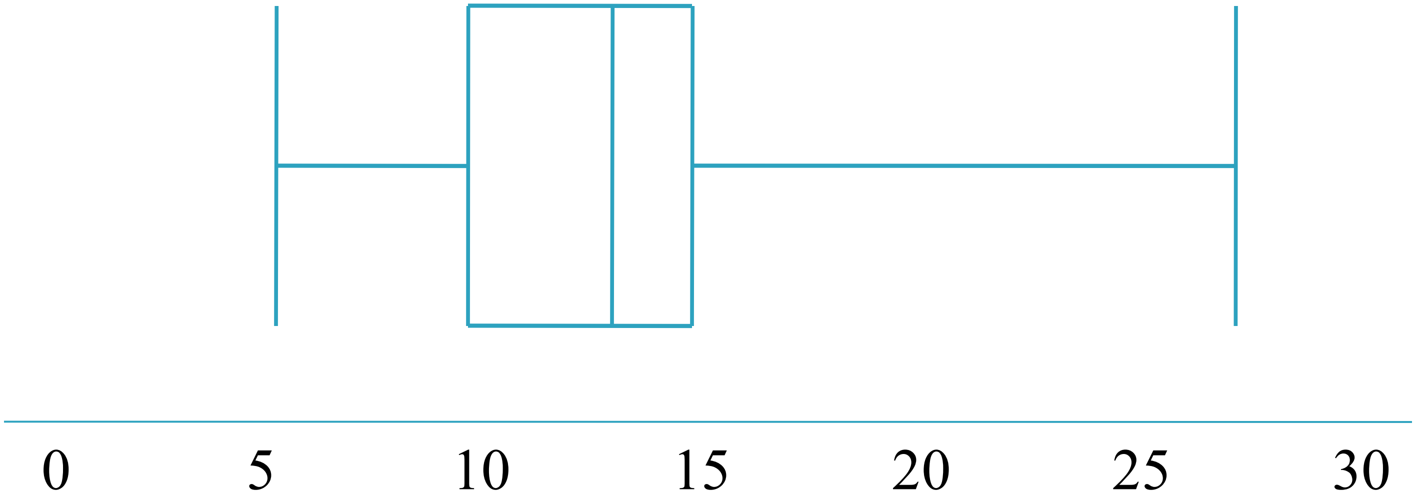
Example

Note that it is not necessary to include the labels or to draw the fences when constructing the modified boxplot.



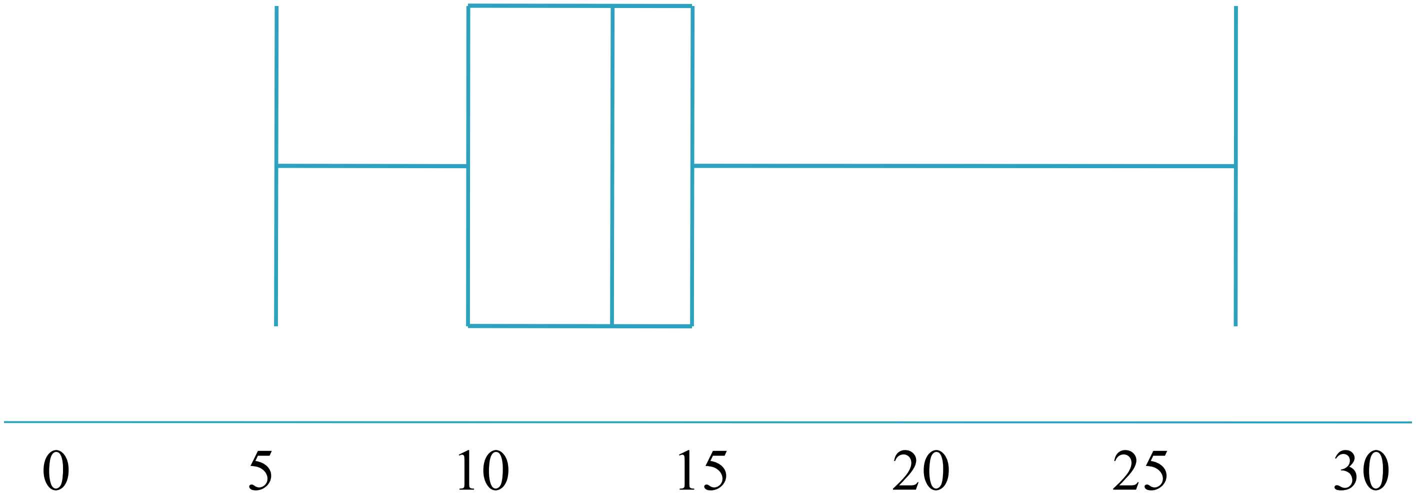
Example

Consider what would happen if we had instead constructed a quantile boxplot for these data:



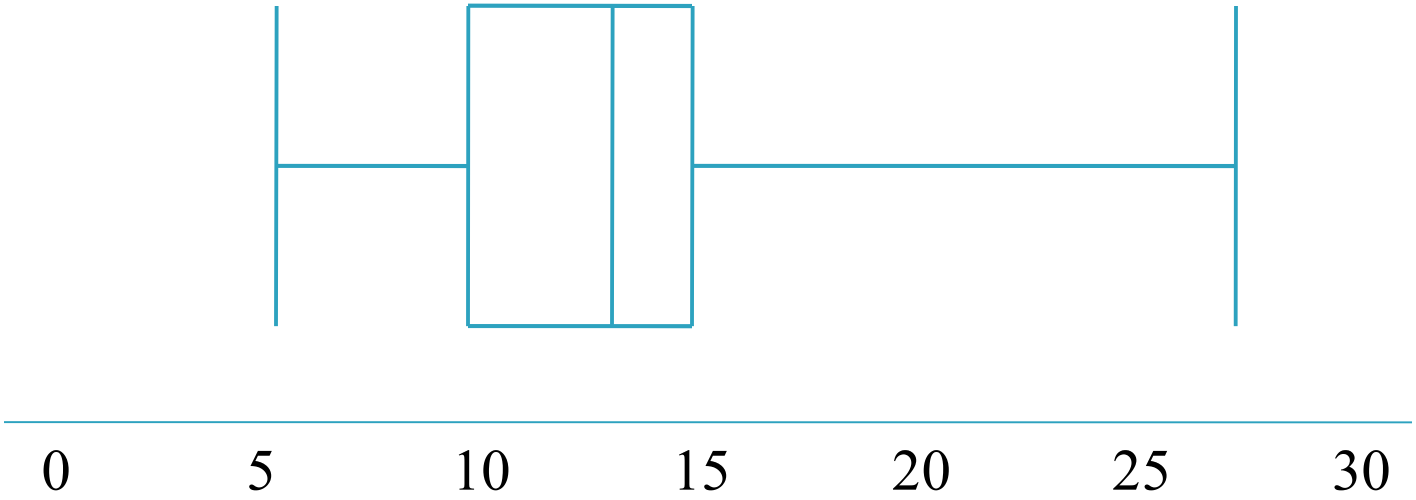
Example

Not only would we be given the impression that there is **high variability** in the data…



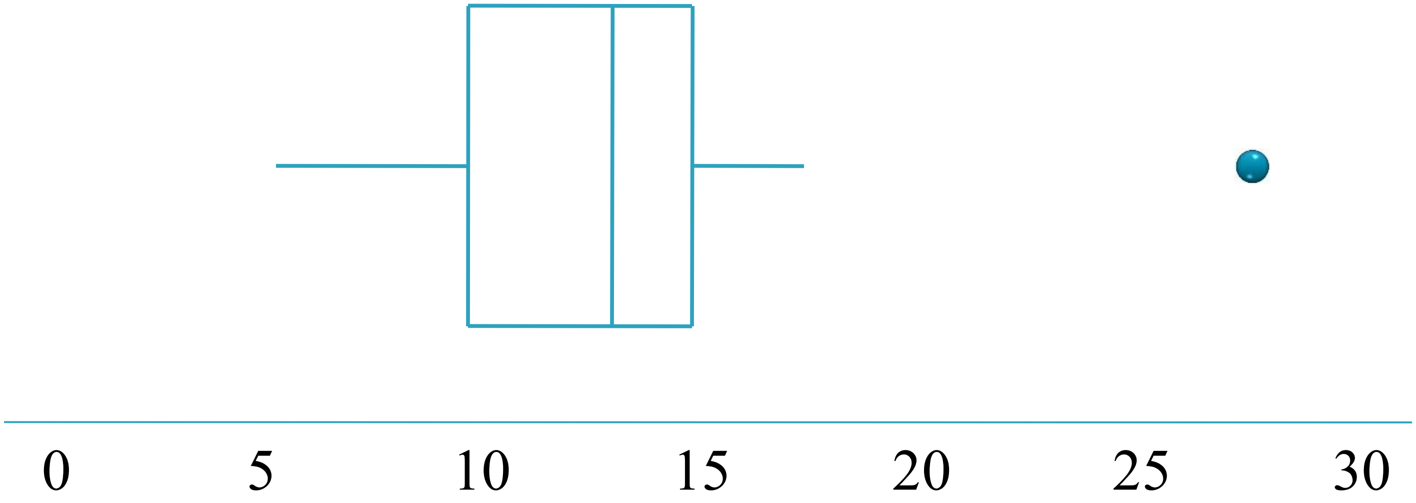
Example

but we may also conclude that the distribution is **skewed to the right**…



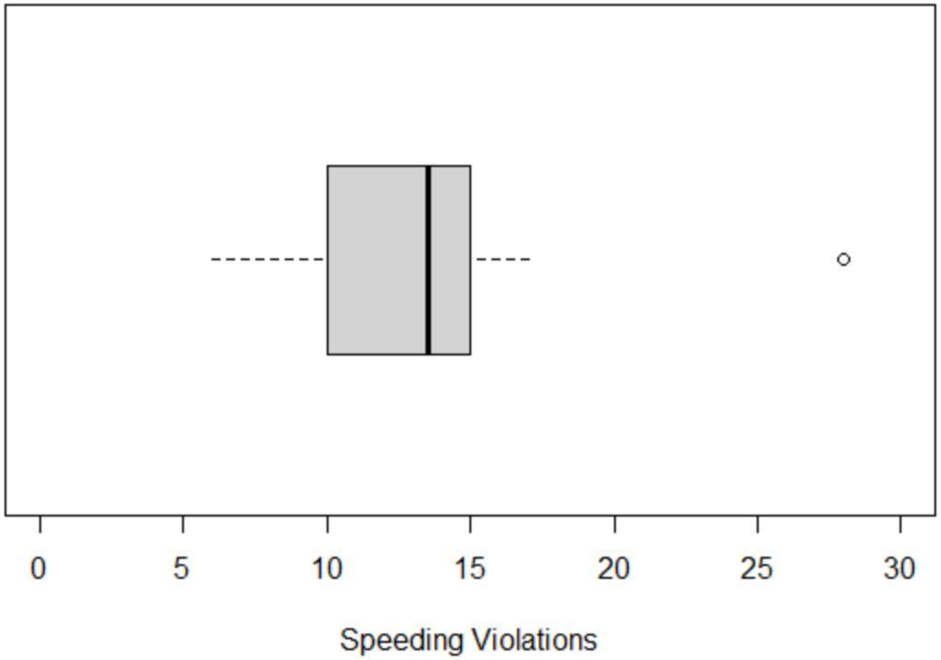
Example

when we see from the modified boxplot that it is in fact **skewed to the left**…



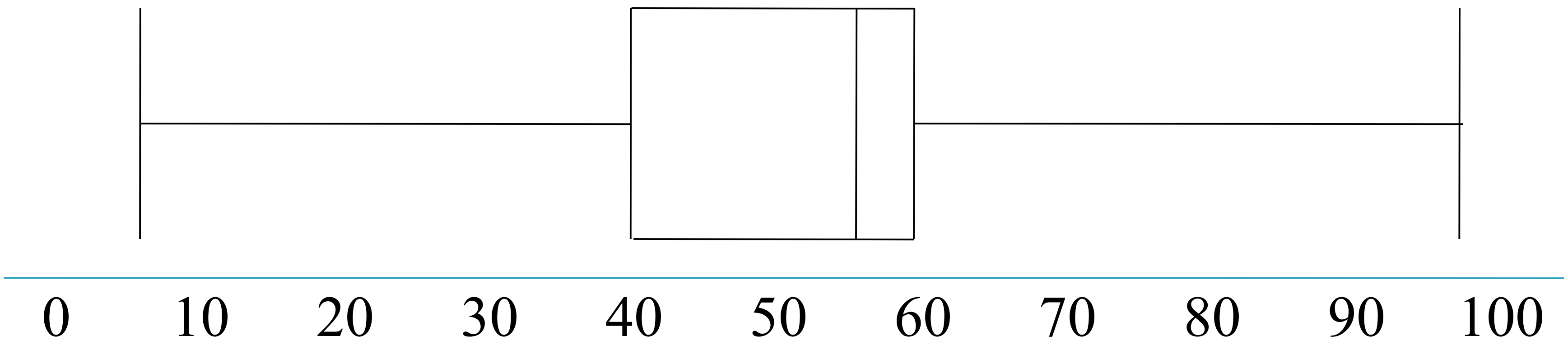
R Code

```
> speed <- c(6, 9, 10, 10, 12, 13, 14, 14, 15,
             15, 17, 28)
> boxplot(speed, horizontal = TRUE, ylim = c(0, 30),
          xlab = "Speeding Violations", staplelty = 0)
```



Practice Question

A basic quantile boxplot for the test scores of 120 Physics students is shown below:



If an outlier boxplot were to be constructed, what scores would be considered outliers?

(A) < 40 or > 60 (B) < 25 or > 85 (C) < 10 or > 90

(D) < 20 or > 80 (E) There are no outliers in this data set.

Practice Question

A bowler's scores for her last 30 games are ordered and are shown below:

139 142 144 149 156 166 171 178 179 179 181
182 183 185 185 189 190 190 190 191 193 195
200 202 207 212 220 227 228 235

The five-number summary for this data set is: 139 178 187 200 235

In constructing an outlier boxplot for these data, the whisker on the right would extend to which value?

(A) 227 (B) 235 (C) 220 (D) 233 (E) 228

Outlier Boxplots

Note that we only had a problem with the quantile boxplot when there were extreme values in our data. When there are no outliers, the quantile boxplot is **identical** to the outlier boxplot (since the “new” minimum and maximum are the same as the “old” minimum and maximum).

Variability

Back to the idea of variability…

The median gave us a measure of center, but we saw that the mean is a better measure of center in some cases, as it includes information about all data values.

Similarly, the range is a very crude description of the variability of a data set. Can we find another way to describe the spread numerically?

Sample Variance

The **variance** of a set of data values (also referred to as the **sample variance**) is defined as:

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

Loosely speaking, the variance can be thought of as the “average squared deviation” from the mean.

Sample Standard Deviation

The **standard deviation** of a data set is defined as the positive square root of the variance:

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

Sample Variance

To calculate the variance s^2 for a sample of size n:

1) Calculate the sample mean $\bar{x}$.
2) Calculate the n deviations $x_i - \bar{x}$.
3) Square the deviations $(x_i - \bar{x})^2$.
4) Add the squared deviations $\sum(x_i - \bar{x})^2$.
5) Divide by $n - 1$

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

Example

The number of goals scored by each of a sample of five NHL teams for the 2018/19 regular season are shown below:

Team	Goals
	247
TORONTO MAPLE LEAFS	286
	222
WASHINGTON capitals	278
	272

Example

We calculate 1) $\bar{x} = 261$

x_i	2) $x_i - \bar{x}$	3) $(x_i - \bar{x})^2$
247	– 14	196
286	25	625
222	– 39	1521
278	17	289
272	11	121
	sum = **0**	4) sum = **2752**

5) The sample variance is therefore equal to

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{2752}{4} = 688$$

Example

We could calculate the sample variance without a table as follows:

$$s^2 = \frac{(247-261)^2 + (286-261)^2 + (222-261)^2 + (278-261)^2 + (272-261)^2}{4}$$

$$= \frac{196 + 625 + 1521 + 289 + 121}{4} = \frac{2752}{4} = 688$$

To get the sample standard deviation, we simply take the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{688} = 26.23$$

R Code

```
> goals <- c(247, 286, 222, 278, 272)

> var(goals)
[1] 688

> sd(goals)
[1] 26.22975
```

Practice Question

The amounts (in $) spent on ice cream by a sample of four customers at Dairy Queen are shown below:

14 5 20 9

What is the value of the standard deviation for this sample?

(A) 5.61 (B) 6.48 (C) 7.75 (D) 31.50 (E) 42.00

Sum of Deviations from the Mean

Note that we calculated $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ for these data.

In fact, this is **always** the case, as shown below:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - n\bar{x}$$

$$= (x_1 + x_2 + \cdots + x_n) - n\left(\frac{x_1 + x_2 + \cdots x_n}{n}\right)$$

$$= (x_1 + x_2 + \cdots + x_n) - (x_1 + x_2 + \cdots + x_n) = 0$$

Practice Question

The mean and standard deviation of the midterm scores of a class of 170 students are calculated to be $\bar{x} = 56$ and $s = 10$, respectively. Based on the low marks, the professor concludes that the test was slightly too difficult, so she adds 4 marks to each student's score. What is the new standard deviation of scores for the test?

(A) 0 (B) 6 (C) 10 (D) 12 (E) 14

Units for Standard Deviation & Variance

What are the units for the standard deviation and variance?

The units for $\bar{x}$ are the same as the units for the individual x's.

For example, the mean hourly wage of a group of employees is still expressed in dollars.

Units for Standard Deviation & Variance

The units for variance are the **squared units** of the individual observations.

The sample variance of hourly wages for the group of employees is expressed in dollars^2.

Note: A variance of 10 dollars^2 does **not** mean that $s^2 = 100$. It means that $s^2 = 10$ and the units are $\$^2$.

Units for Standard Deviation & Variance

The units for the standard deviation are again the same as the units of the individual observations (in this case, dollars).

This is one reason why standard deviations are more commonly reported than variances – because they are easier to interpret in terms of units we understand.

Measures of Variability – Standard Deviation

We do, however, need to be careful when interpreting the meaning of the standard deviation.

Although the variance is loosely defined as the average squared deviation of an observation from the mean, the standard deviation is **not** the average deviation from the mean.

Variance: Nicer mathematical definition (less intuitive units)

Standard deviation: Nicer units (less intuitive mathematical definition)

Mean & Std. Dev. Vs. Five-Number Summary

We have seen two different numerical summaries for a given data set:

(1) the five-number summary

(2) the sample mean and standard deviation

Both provide us with descriptions of the center and variability of a data distribution.

Mean & Std. Dev. Vs. Five-Number Summary

So how do we know which one to report?

It depends on the **shape** of the data distribution!

Recall that the mean and median are equal for symmetric distributions. Therefore, if the distribution of data values is reasonably **symmetric** with **no outliers**, we should report the **mean and standard deviation**, since they contain more information about the sample than do the median, IQR and range.

Mean & Std. Dev. Vs. Five-Number Summary

However, for **skewed distributions**, or in the presence of **outliers**, the **five-number summary** should be reported, since the mean and standard deviation are strongly affected by extreme values.