Describing Data

Department of Mathematics

Salt Lake Community College

(Slides by Adam Wilson)

Definition

A **frequency table** is a table with two columns. One column lists the categories, and the other the frequencies with which the items in the categories occur. (i.e. how many items fit into each category.)

Definition

A **frequency table** is a table with two columns. One column lists the categories, and the other the frequencies with which the items in the categories occur. (i.e. how many items fit into each category.)

Example 1

An insurance company determines vehicle insurance premiums based on known risk factors. If a person is considered a higher risk, their premiums will be higher. One potential factor is the color of your car. The insurance company believes that people with some color cars are more likely to get in accidents.

To research this, they examined police reports for recent total-loss collisions. The data is summarized in the frequency table.

Color	Frequency
Blue	25
Green	52
Red	41
White	36
Black	39
Grey	23

Definition

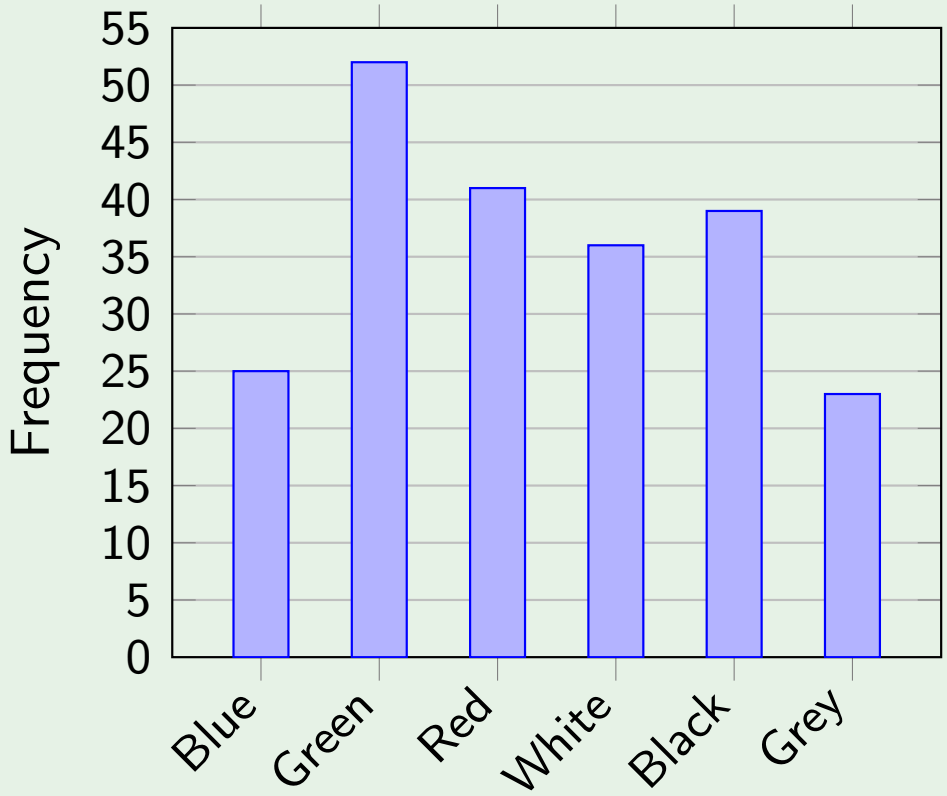
A **bar graph** is a graph that displays a bar for each category with the length of each bar indicating the frequency of that category.

Definition

A **bar graph** is a graph that displays a bar for each category with the length of each bar indicating the frequency of that category.

Example 2

Using the car data from Example 1, we have the following bar graph.



Definition

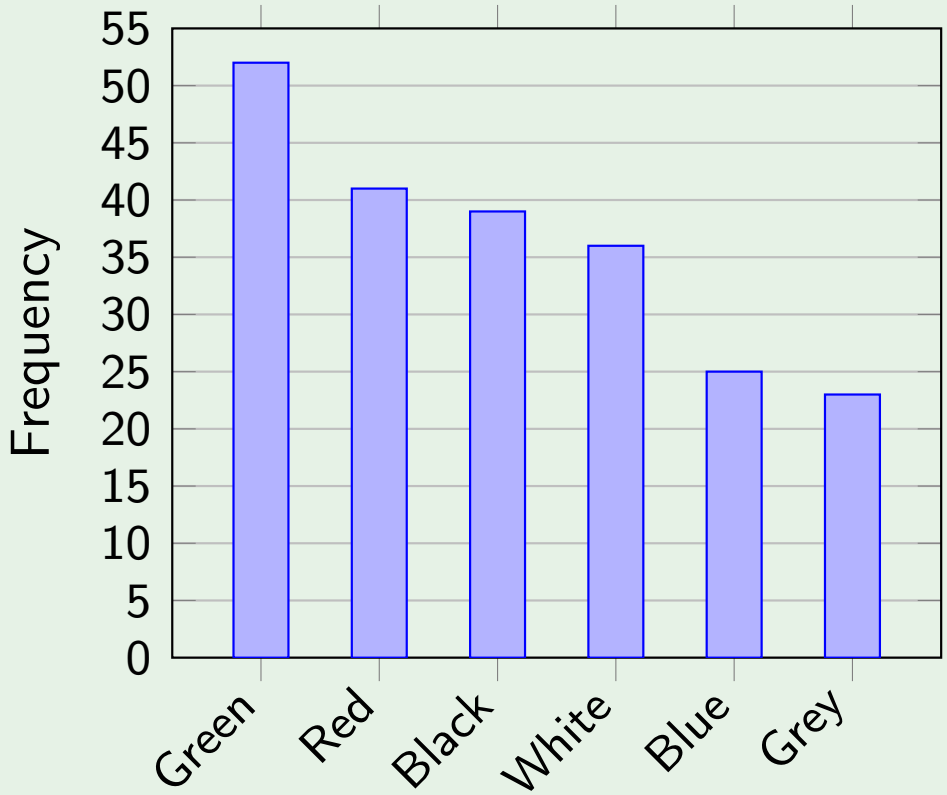
A **Pareto chart** is a bar graph ordered from highest to lowest frequency.

Definition

A **Pareto chart** is a bar graph ordered from highest to lowest frequency.

Example 3

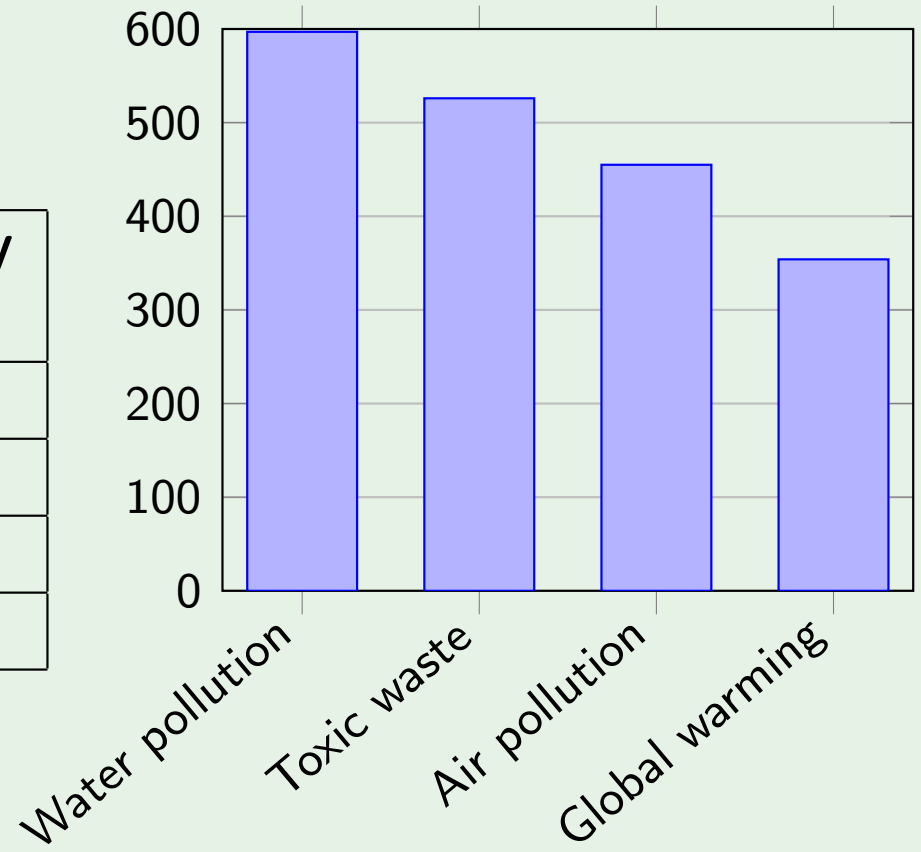
Using the car data from Example 1, we have the following Pareto chart.



Example 4

In a survey[1], adults were asked whether they personally worried about a variety of environmental concerns. The numbers (out of 1012 surveyed) who indicated that they worried "a great deal" about some selected concerns are summarized below.

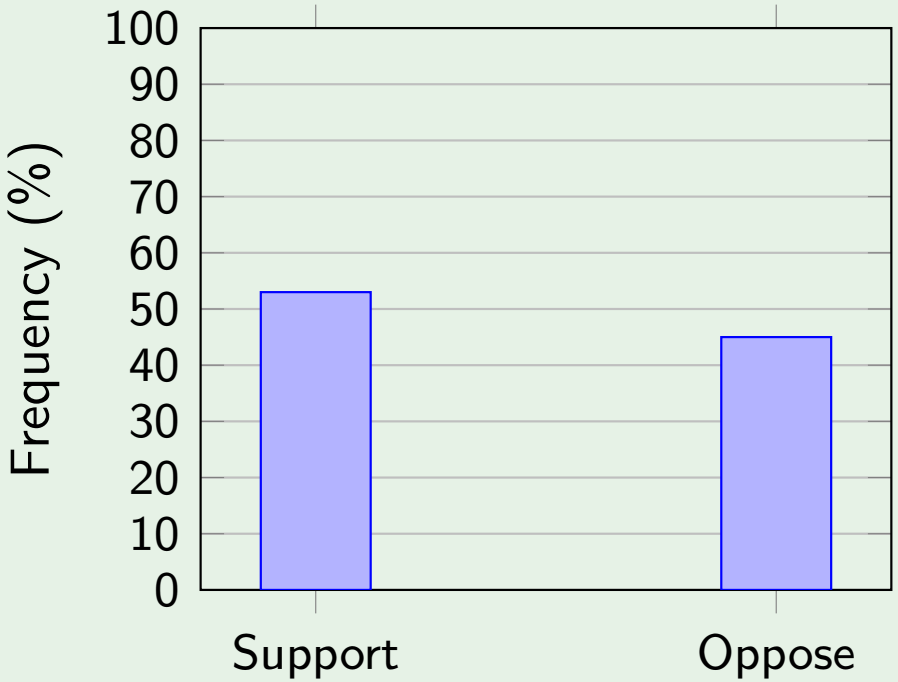
Environmental Issue	**Frequency**
Water pollution	597
Toxic waste	526
Air pollution	455
Global warming	354

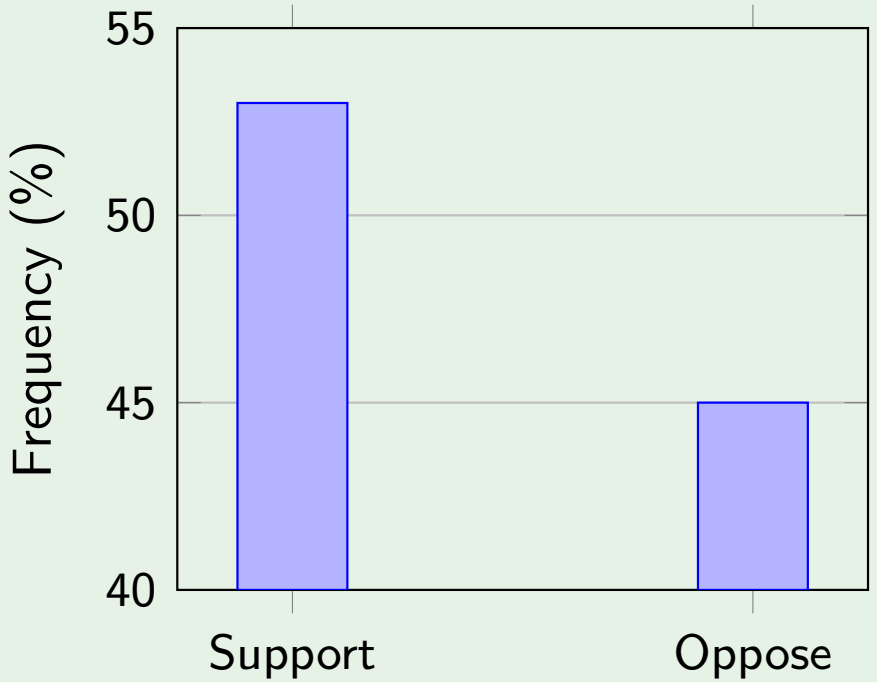


[1] Gallup Poll. March 5-8, 2009. http://www.pollingreport.com/enviro.htm

Example 5

Compare the two graphs below showing support for same-sex marriage rights from a poll taken in May 2013[2].





The difference in the vertical scale on the first graph suggests a different story than the true differences in percentages; the second graph makes it look like more than twice as many people support marriage rights as oppose it.

[2]Gallup Poll. May 2-7, 2013, from http://www.pollingreport.com/civil.htm

Definition

A **histogram** a bar graph, where the horizontal axis is a number line.

Definition

A **histogram** a bar graph, where the horizontal axis is a number line.

Definition

Class intervals are groupings of the data. In general, we define class intervals so that:

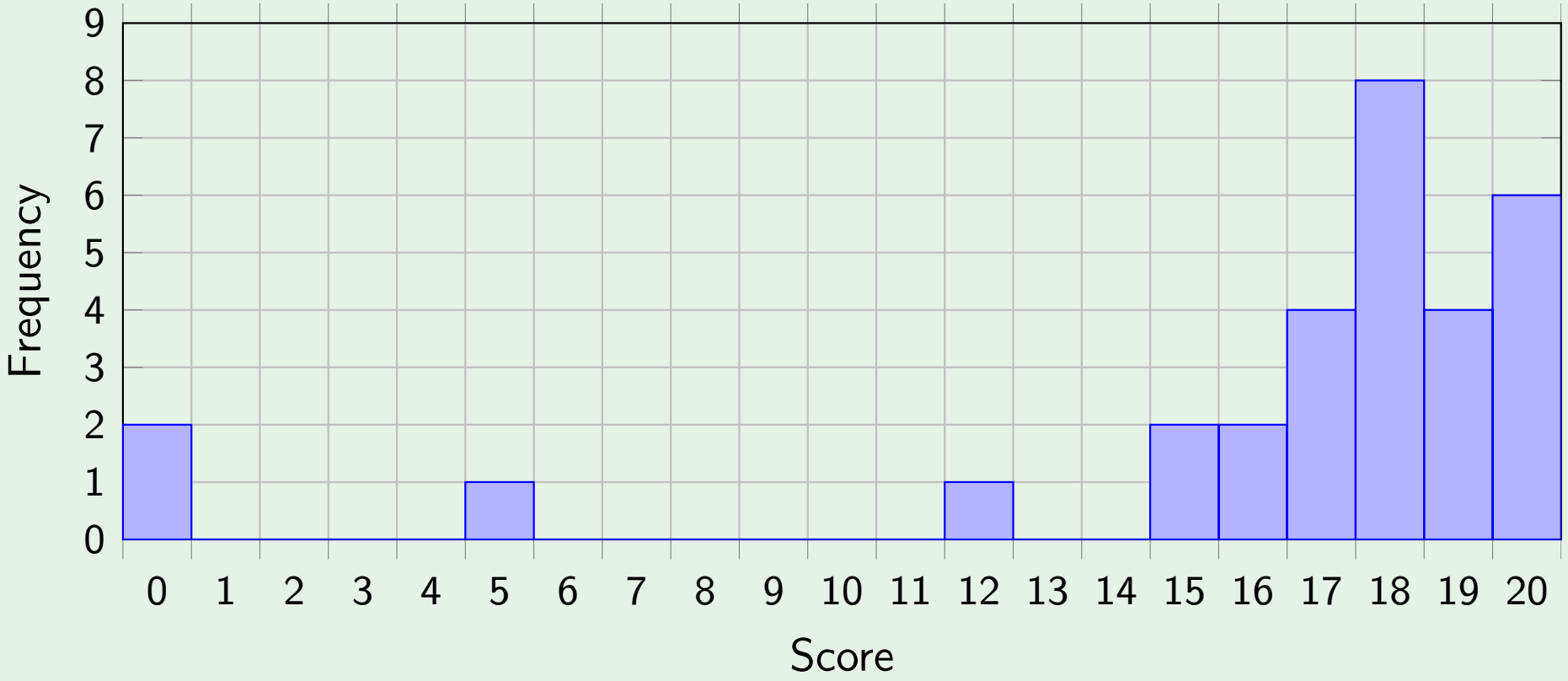
- Each interval is equal size. For example, if the first class contains values from 120-129, then the second class should include 130-139.
- We have somewhere between 5 and 20 classes, typically, depending upon the number of data we are working with.

Example 6

A teacher records scores on a 20-point quiz for the 30 students in his class.

19	20	18	18	17	18	19	17	20	18	20	16	20	15	17
12	18	19	18	19	17	20	18	16	15	18	20	5	0	0

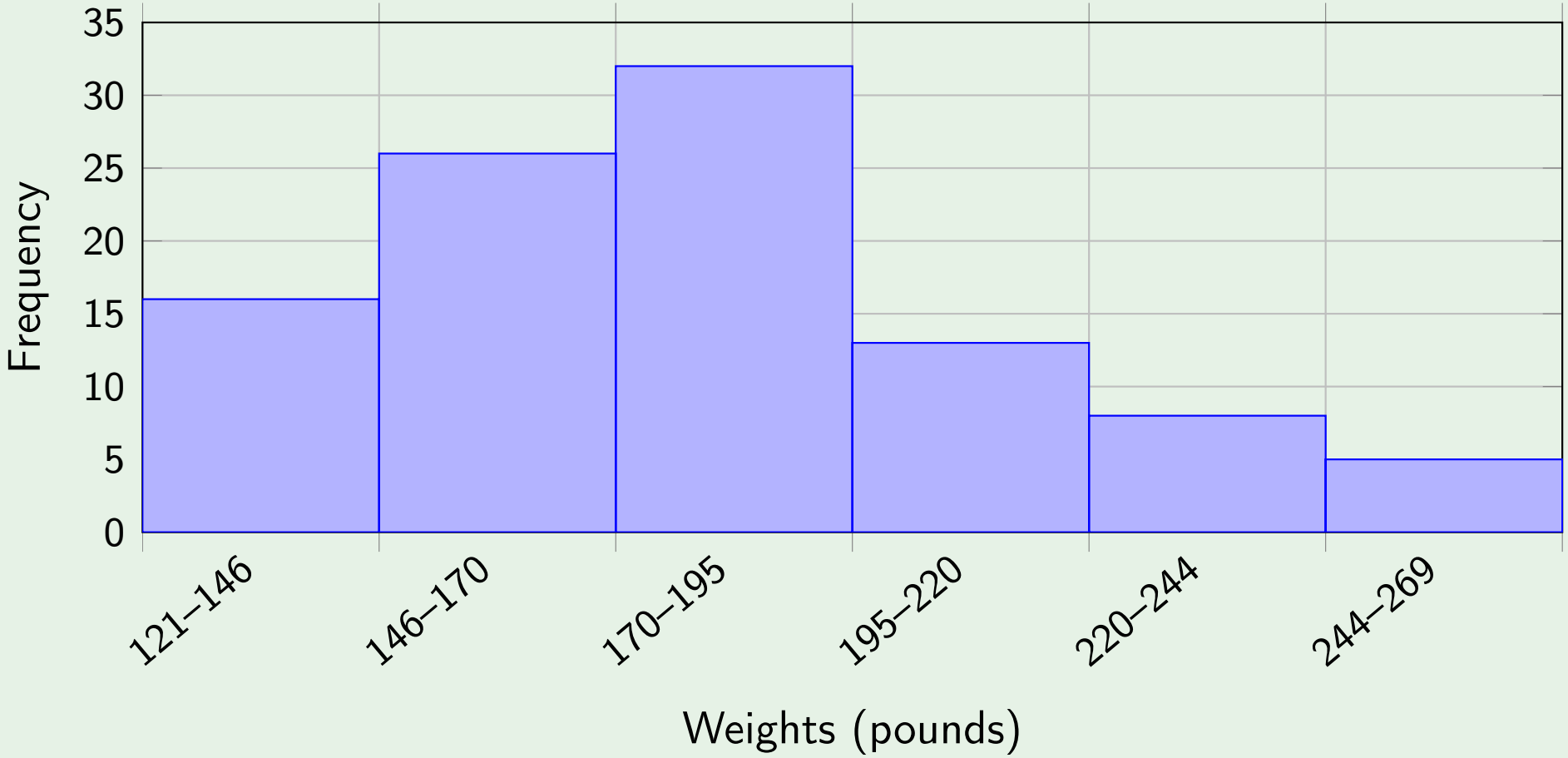
These scores could be summarized in the following histogram.



Example 7

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142.

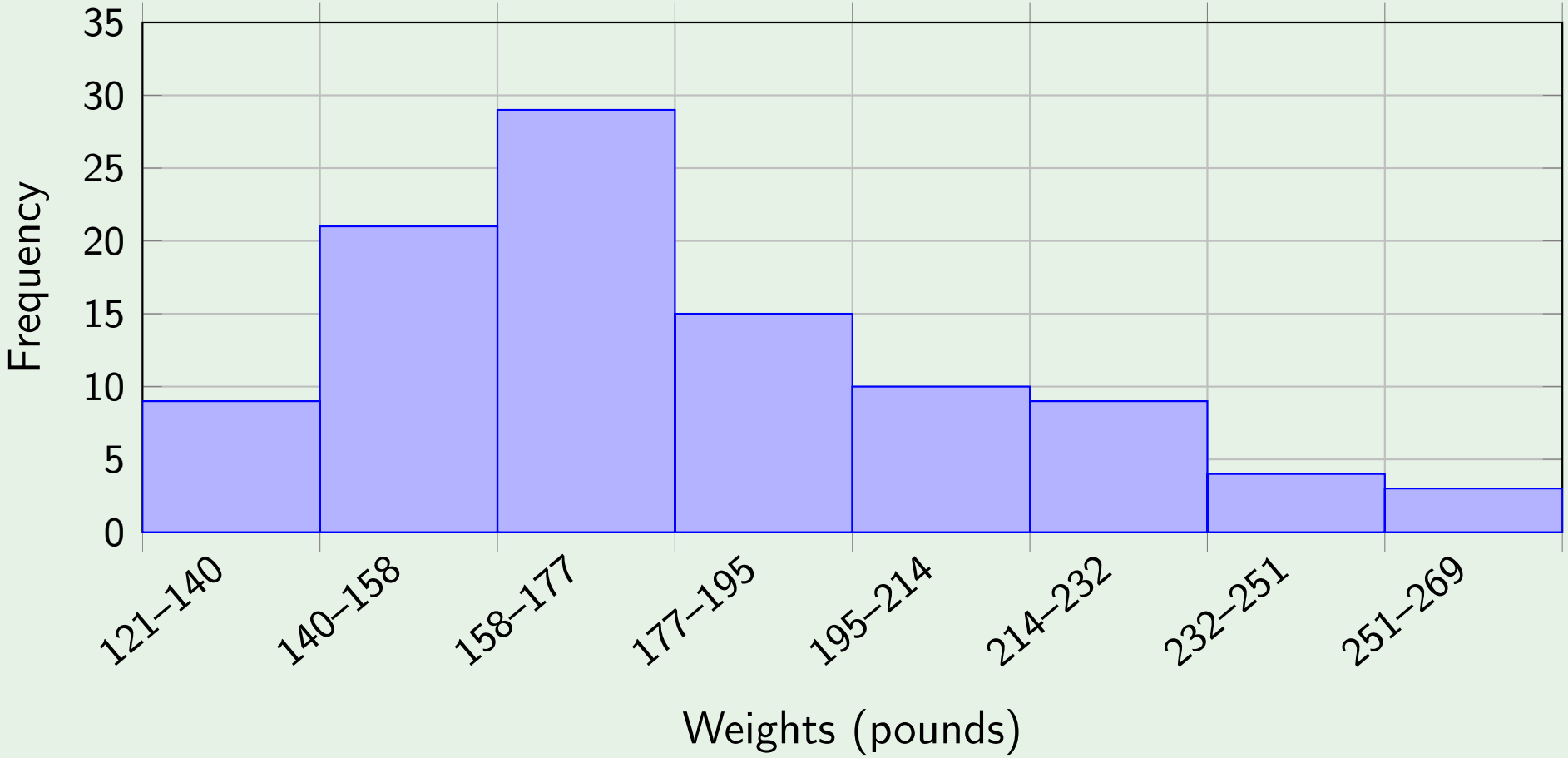
We could create 6 intervals with a width of around 20.



Example 7

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142.

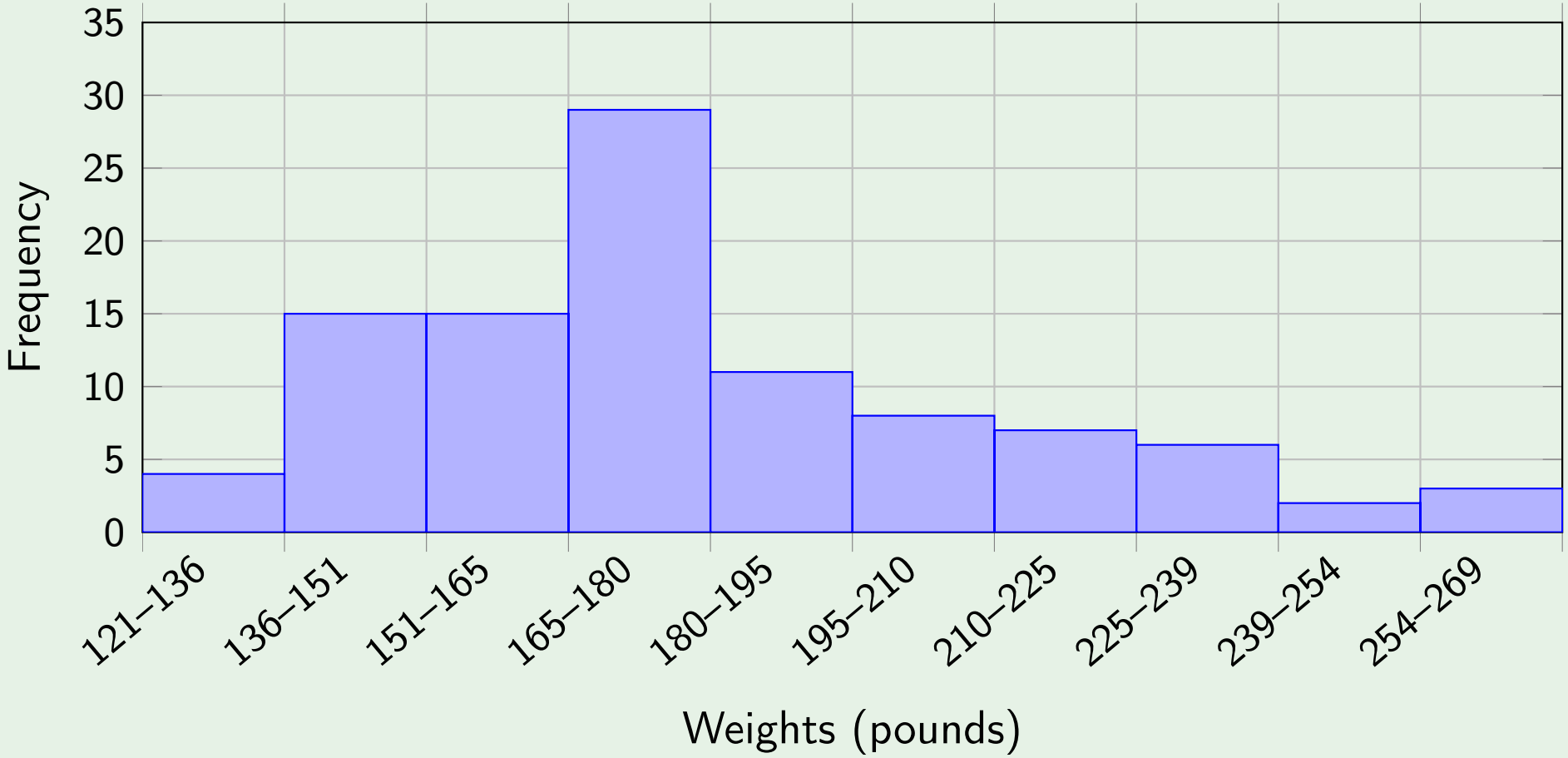
We could create 8 intervals with a width of around 18.



Example 7

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142.

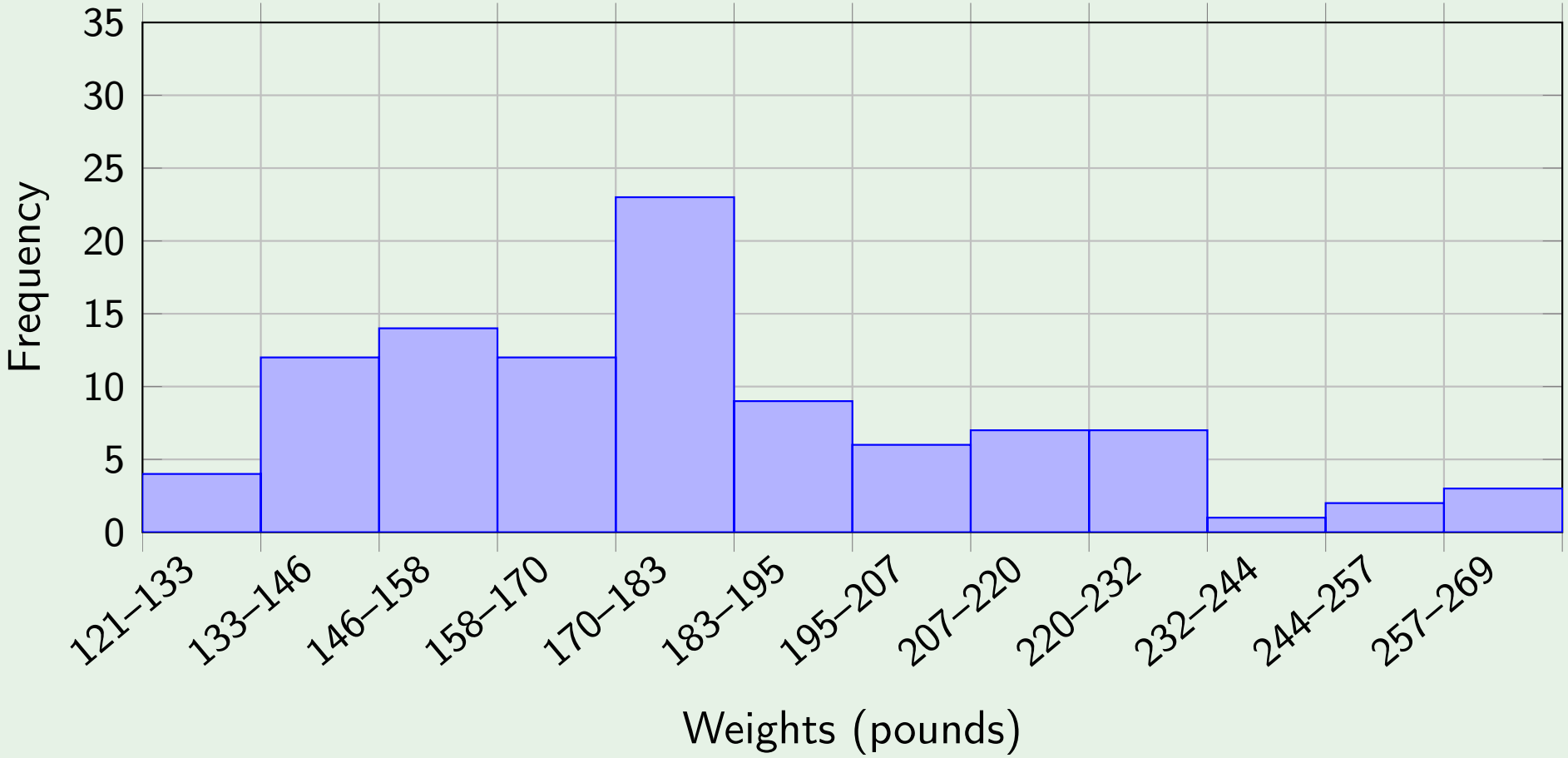
We could create 10 intervals with a width of around 14.



Example 7

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142.

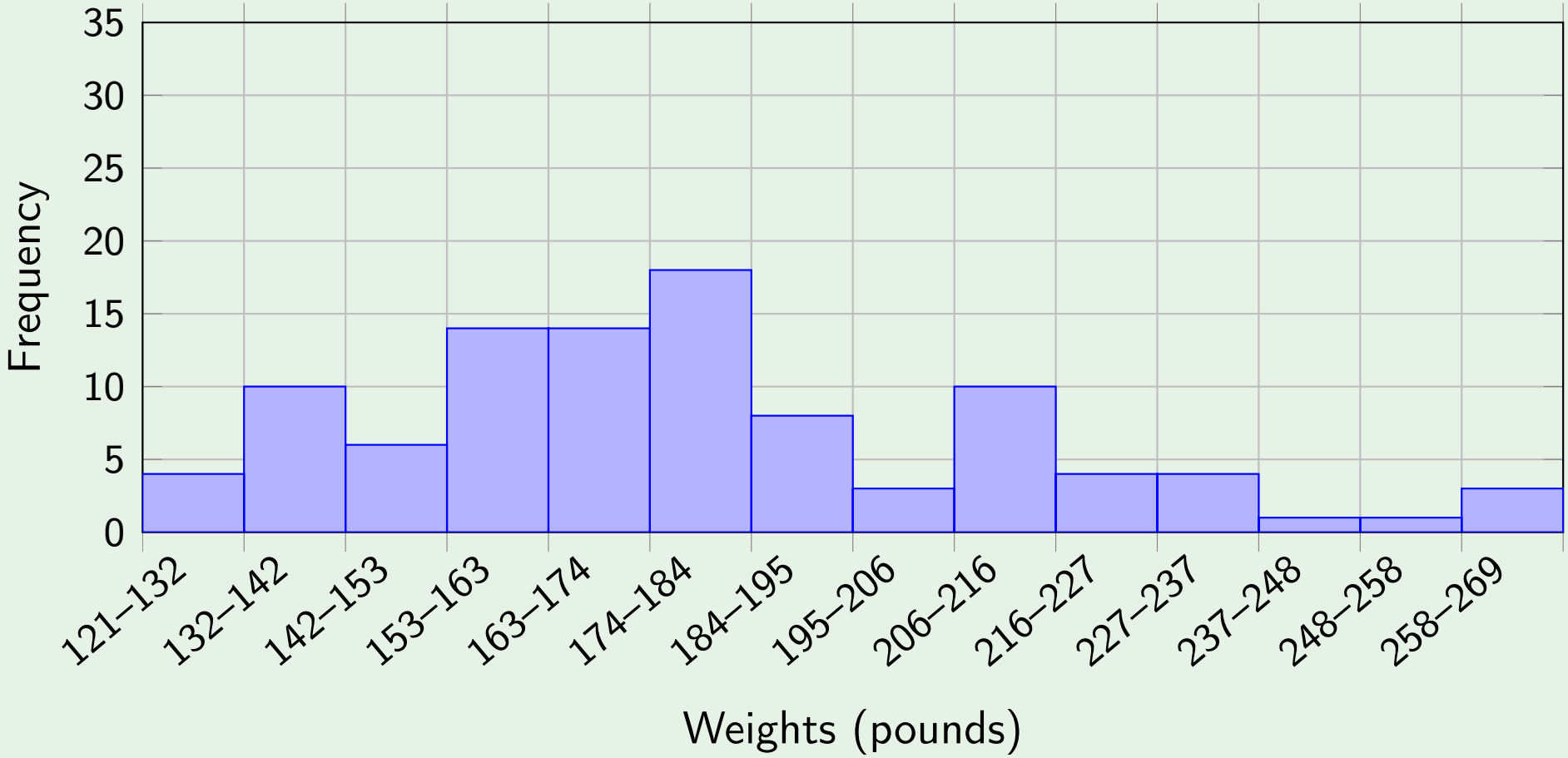
We could create 10 intervals with a width of around 12.



Example 7

Suppose that we have collected weights from 100 male subjects as part of a nutrition study. For our weight data, we have values ranging from a low of 121 pounds to a high of 263 pounds, giving a total span of 263-121 = 142.

We could create 14 intervals with a width of around 10.



Definition

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

Definition

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

Note

It is not uncommon to see the word "average" used instead of "mean".

Definition

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

Note

It is not uncommon to see the word "average" used instead of "mean".

Example 8

Marci's exam scores for her last math class were: 79, 86, 82, 94.
What is the mean of these values?

Definition

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

Note

It is not uncommon to see the word "average" used instead of "mean".

Example 8

Marci's exam scores for her last math class were: 79, 86, 82, 94.
What is the mean of these values?

$$\frac{79 + 86 + 82 + 94}{4} = \frac{341}{4} = 82.25 \approx 85.3$$

Typically we round means to one more decimal than the original data.

Definition

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

Note

It is not uncommon to see the word "average" used instead of "mean".

Example 8

Marci's exam scores for her last math class were: 79, 86, 82, 94.
What is the mean of these values?

$$\frac{79 + 86 + 82 + 94}{4} = \frac{341}{4} = 82.25 \approx 85.3$$

Typically we round means to one more decimal than the original data.

Example 9

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest $5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the mean of this data?

Example 9

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the mean of this data?

$$\begin{aligned}
&\frac{\overbrace{15+\cdots+15}^{6 \text{ terms}}+\overbrace{20+\cdots+20}^{8 \text{ terms}}+\overbrace{25+\cdots+25}^{11 \text{ terms}}+\cdots+\overbrace{50+\cdots+25}^{7 \text{ terms}}}{100} \\
=&\frac{15\cdot 6+20\cdot 8+25\cdot 11+30\cdot 17+35\cdot 19+40\cdot 20+45\cdot 12+50\cdot 7}{100} \\
=&\frac{3390}{100}=33.9
\end{aligned}$$

Definition

The **median** of a set of data is the value in the middle when the data is in numerical order.

Definition

The **median** of a set of data is the value in the middle when the data is in numerical order.

Note

To find the median, begin by listing the data in order from smallest to largest, or largest to smallest.

If the number of data values, N, is odd, then the median is the middle data value. This value can be found by rounding $\frac{N}{2}$ up to the next whole number.

If the number of data values is even, there is no one middle value, so we find the mean of the two middle values (values $\frac{N}{2}$ and $\frac{N}{2}+1$)

Example 10

Find the media of these quiz scores: 8 6 4 1 7 1 1

Example 10

Find the media of these quiz scores: 8 6 4 1 7 1 1

List the data in order: 1 1 1 4 6 7 8
$\Uparrow$

Since we have an odd number of data, we see the median is 4.

Example 10

Find the media of these quiz scores: 8 6 4 1 7 1 1

List the data in order: 1 1 1 4 6 7 8
$\Uparrow$

Since we have an odd number of data, we see the median is 4.

Example 11

Find the median of these quiz scores: 5 9 8 6 4 8 2 5 7 7

Example 10

Find the media of these quiz scores: 8 6 4 1 7 1 1

List the data in order: 1 1 1 4 6 7 8
$\Uparrow$

Since we have an odd number of data, we see the median is 4.

Example 11

Find the median of these quiz scores: 5 9 8 6 4 8 2 5 7 7

List the data in order: 2 4 5 5 6 7 7 8 8 9
$\Uparrow$

Since we have an even number of data, there is no middle item. In this case, we say the median is the mean of the two middle numbers, 6 and 7, which is 6.5.

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest $5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

There are 6 data values of 15. $\Rightarrow$ Values 1 to 6 are 15.

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

There are 6 data values of 15. $\Rightarrow$ Values 1 to 6 are 15.
The next 8 data values are 20. $\Rightarrow$ Values 7 to (6+8)=14 are 20.

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

There are 6 data values of 15. $\Rightarrow$ Values 1 to 6 are 15.
The next 8 data values are 20. $\Rightarrow$ Values 7 to (6+8)=14 are 20.
The next 11 data values are 25. $\Rightarrow$ Values 15 to (14+11)=25 are 25.

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

There are 6 data values of 15. $\Rightarrow$ Values 1 to 6 are 15.
The next 8 data values are 20. $\Rightarrow$ Values 7 to (6+8)=14 are 20.
The next 11 data values are 25. $\Rightarrow$ Values 15 to (14+11)=25 are 25.
The next 17 data values are 30. $\Rightarrow$ Values 26 to (25+17)=42 are 30.

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

There are 6 data values of 15. $\Rightarrow$ Values 1 to 6 are 15.
The next 8 data values are 20. $\Rightarrow$ Values 7 to (6+8)=14 are 20.
The next 11 data values are 25. $\Rightarrow$ Values 15 to (14+11)=25 are 25.
The next 17 data values are 30. $\Rightarrow$ Values 26 to (25+17)=42 are 30.
The next 19 data values are 35. $\Rightarrow$ Values 43 to (42+19)=61 are 35.

Example 12

The one hundred families in a particular neighborhood are asked their annual household income, to the nearest \$5 thousand dollars.

Income	Frequency	Income	Frequency
15	6	35	19
20	8	40	20
25	11	45	12
30	17	50	7

What is the median of this data?

We have 100 items, this means the median will be the mean of the 50^{th} and 51^{st} data values. So, we need to start counting up from the bottom:

There are 6 data values of 15. $\Rightarrow$ Values 1 to 6 are 15.
The next 8 data values are 20. $\Rightarrow$ Values 7 to (6+8)=14 are 20.
The next 11 data values are 25. $\Rightarrow$ Values 15 to (14+11)=25 are 25.
The next 17 data values are 30. $\Rightarrow$ Values 26 to (25+17)=42 are 30.
The next 19 data values are 35. $\Rightarrow$ Values 43 to (42+19)=61 are 35.

So, the median is $\frac{35+35}{2} = 35$, \$35 thousand dollars.

Definition

The **mode** is the element of the data set that occurs most frequently.

Definition

The **mode** is the element of the data set that occurs most frequently.

Note

The mode is useless with quantitative data. It is most commonly used for categorical data, for which median and mean cannot be computed.

Definition

The **mode** is the element of the data set that occurs most frequently.

Note

The mode is useless with quantitative data. It is most commonly used for categorical data, for which median and mean cannot be computed.

Example 13

In Example 1 we collected the data:

Color	Frequency	Color	Frequency
Blue	3	White	3
Green	5	Black	2
Red	4	Grey	3

For this data, Green is the mode.

Definition

The **mode** is the element of the data set that occurs most frequently.

Note

The mode is useless with quantitative data. It is most commonly used for categorical data, for which median and mean cannot be computed.

Example 13

In Example 1 we collected the data:

Color	Frequency	Color	Frequency
Blue	3	White	3
Green	5	Black	2
Red	4	Grey	3

For this data, Green is the mode.

Note

It is possible for a data set to have more than one mode.