HNDIT3072
Statistics for IT

2. Organizing and Summarizing Data

Course Objectives

After completing this module, students should be able to

1. Recognize different types of data
2. Describe data presented as a list
3. Describe discrete data presented in a table
4. Describe continuous data presented in a grouped frequency table

Raw Data

- Raw data are collected data that have not been organized numerically.

Example:

  - the set of heights of 100 male students obtained from an alphabetical listing of university records.

  - Marks sheet of students in a class in ascending order

Arrays

- An array is an arrangement of raw numerical data in ascending or descending order of magnitude.

Example:

Arrange the numbers 17, 45, 38, 27, 6, 48, 11, 57, 34, and 22 in an array.

SOLUTION

(a) In ascending order of magnitude, the array is: **6, 11, 17, 22, 27, 34, 38, 45, 48, 57.**

(b) In descending order of magnitude, the array is: **57, 48, 45, 38, 34, 27, 22, 17, 11, 6.**

Range

- The difference between the largest and smallest numbers is called the range of the data.

- Example:
  - if the largest height of 100 male students is 74 inches (in) and the smallest height is 60 in, the range is 74 - 60 = 14 in.

Data

- The first step in summarizing quantitative data is to determine whether the data is **discrete** or **continuous**.

  – If the data is **discrete**, the categories of data will be the observations (as in qualitative data),

  – if the data is **continuous**, the categories of data (called *classes*) must be created using intervals of numbers.

Data

Qualitative

Quantitative

Discrete Data

- 5 kids
- 96 workers
- 3 laptops

Continuous Data

- 3.25 kg
- 1.32 miles
- 7.25 inches

Frequency Table

- Table that displays the frequency of various outcomes or categories in a dataset.
- It is a way to summarize and organize data to understand the distribution or pattern of values within a dataset.
- Frequency tables are commonly used in data analysis and can be created for both categorical and numerical data.

Categorical Data

Eye Color	Frequency
Green	40
Blue	25
Hazel	35

Numerical Data

Height	Frequency
139	1
145	2
150	2
136	1
152	1
144	1
138	2

Frequency Distribution of Qualitative Data

- Political Party Affiliations Professor Weiss asked his introductory statistics students to state their political party affiliations as Democratic, Republican, or Other. The responses of the 40 students in the class are given in Table below. **Determine the frequency distribution of these data.**

Democratic	Other	Democratic	Other	Democratic
Republican	Republican	Other	Other	Republican
Republican	Republican	Republican	Democratic	Republican
Republican	Democratic	Democratic	Other	Republican
Democratic	Democratic	Republican	Democratic	Democratic
Republican	Republican	Other	Other	Democratic
Republican	Democratic	Republican	Other	Other
Republican	Republican	Republican	Democratic	Republican

Frequency Distribution of Qualitative Data

Democratic	Other	Democratic	Other	Democratic
Republican	Republican	Other	Other	Republican
Republican	Republican	Republican	Democratic	Republican
Republican	Democratic	Democratic	Other	Republican
Democratic	Democratic	Republican	Democratic	Democratic
Republican	Republican	Other	Other	Democratic
Republican	Democratic	Republican	Other	Other
Republican	Republican	Republican	Democratic	Republican

Party	Tally	Frequency
Democratic	𝍸 𝍸 | | |	13
Republican	𝍸 𝍸 𝍸 | | |	18
Other	𝍸 | | | |	9
		40

Relative Frequency Table

- A **relative frequency** is the fraction of times an answer occurs.
- To find the **relative frequencies**, divide each **frequency** by the total number of students in the sample

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{total number of frequencies}}$$

- A relative frequency is just a percentage expressed as a decimal.

Frequency Distribution of Qualitative Data

- Calculate the relative frequency of following .

Party	Tally	Frequency
Democratic	𝍸 𝍸 | | |	13
Republican	𝍸 𝍸 𝍸 | | |	18
Other	𝍸 | | | |	9
		40

Party	Relative frequency	
Democratic	0.325	← *13/40*
Republican	0.450	← *18/40*
Other	0.225	← *9/40*
	1.000	

Interpretation, we see that 32.5% of the students in Professor Weiss's introductory statistics class are Democrats, 45.0% are Republicans, and 22.5% are Other.

Frequency Distribution of Qualitative Data

- A frequency table is constructed by arranging collected data values in ascending order of magnitude with their corresponding frequencies

- Several methods can be used to group quantitative data into classes.

- Three of the most common methods:
  - Single-value grouping Method → **Non-grouped frequency**
  - Limit grouping Method
  - Cutpoint grouping Method

  → **Grouped frequency**

Frequency Distribution of Qualitative Data

- **Ungrouped Frequency Distribution:**

  (Single-value grouping Method)

  - class represents a single possible value
  - Such classes are called single-value classes
  - this method of grouping quantitative data is called single-value grouping
  - use the distinct values of the observations as the classes
  - particularly suitable for discrete data in which there are only a small number of distinct values.

Frequency Table

(Ungroup frequency distribution/ single-value grouping method)

Example 01:

- **The following data represent the number of available cars in a household based on a random sample of 50 households.**

3	0	1	2	1	1	1	2	0	2
4	2	2	2	1	2	2	0	2	4
1	1	3	2	4	1	2	1	2	2
3	3	2	1	2	2	0	3	2	2
2	3	2	1	2	2	1	1	3	5

- **Construct a frequency and relative frequency distribution.**

Frequency and Relative Frequency Table

Number of Cars	Tally	Frequency	Relative Frequency
0	𝍸	4	4/50 = 0.08
1	𝍸 𝍸 𝍸	13	13/50 = 0.26
2	𝍸 𝍸 𝍸 𝍸	22	0.44
3	𝍸	7	0.14
4	𝍸	3	0.06
5	|	1	0.02

3	0	1	2	1	1	1	2	0	2
4	2	2	2	1	2	2	0	2	4
1	1	3	2	4	1	2	1	2	2
3	3	2	1	2	2	0	3	2	2
2	3	2	1	2	2	1	1	3	5

Grouped Frequency Table- Terms

- Class Interval.
- Class Limits.
- Class Boundaries.
- Class Width.
- Class Marks.

Grouped Frequency Table- Terms

The following data represents the height of 100 student who are following HNDIT course .

Height (in)	No of Student
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
Total	100

- Find out **Class interval, Class limits, Class Boundaries, Class width, Class Marks.**

Grouped Frequency Table- Terms

Height (in)	No of Student
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8

- Class Interval:
  - A symbol defining a class, such as 60–62 in the given table, is called a class interval.
- Class Limits:
  - The end numbers, 60 and 62, are called class limits; the smaller number (60) is the **lower class limit**, and the larger number (62) is the **upper class limit**.
    - **lower class limit :**The smallest value that could go in a class.
    - **upper class limit**: The largest value that could go in a class
- Open Class Intervals:
  - A class interval that, at least theoretically, has either no upper class limit or no lower class limit indicated is called an open class interval.
    - For example, referring to age groups of individuals, the class interval ‘‘74 years and over **(74 <)**’’ is an open class interval.

Grouped Frequency Table- Terms

Height (in)	No of Student
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8

- Class Boundaries:
  - Subtract the first upper class limit from the second lower class limit
  - Divide the difference by 2
  - Subtract this value from all of the lower class limits and add the value to all the upper class limit.

Ex:

(63-62)/2 = **0.5**

  - If heights are recorded to the nearest inch, the class interval 60–62 theoretically includes all measurements from 59.5 in to 62.5 in.
  - These numbers, 59.5 and 62.5, are called **class boundaries**, the smaller number (59.5) is the **lower class boundary**, and the larger number (62.5) is the **upper class boundary**.

- Class Boundaries:

(63-62)/2 = **0.5**

Height (in)	Class Boundaries	No of Student
60-62	59.5 - 62.5	5
63-65	62.5 - 65.5	18
66-68	65.5 - 6.5	42
69-71	68.5 - 71.5	27
72-74	71.5 - 74.5	8
		Total 100

Grouped Frequency Table- Terms

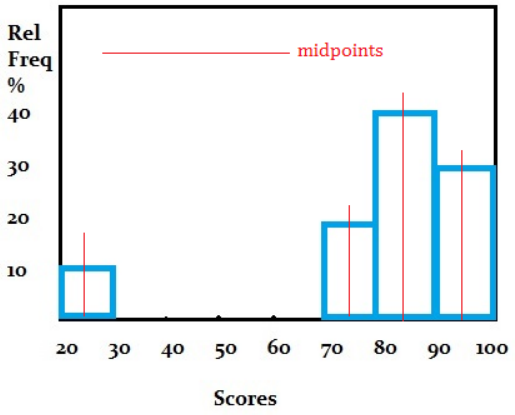
- The Size, or Width, of a Class Interval:
  - The size, or width, of a class interval is the **difference between the lower and upper class boundaries** and is also referred to as **the class width, class size, or class length.**
  - If all class intervals of a frequency distribution have equal widths, this common width is denoted by **c**. In such case c is equal to the difference between two successive lower class limits or two successive upper class limits.

Grouped Frequency Table- Terms

- The Class Mark (Midpoint):
  - The average of the two class limits of a class
  - A specific point in the center of the bins (categories) in a frequency distribution table
  - The class mark is obtained by adding the lower and upper class limits and dividing by 2.

Eg.

class mark of the interval (60 – 62) = $\frac{60+62}{2}$ = 61



Days to maturity	Tally	Frequency	Relative frequency
30–39	|||	3	0.075
40–49	|	1	0.025
50–59	𝍸 |||	8	0.200
60–69	𝍸 𝍸	10	0.250
70–79	𝍸 ||	7	0.175
80–89	𝍸 ||	7	0.175
90–99	||||	4	0.100
		40	1.000

- For instance, consider the class 50–59 in above table.
- The lower limit is 50, the upper limit is 59, **the width is 60 – 50 = 10, and the mark is (50 + 59)/2 = 54.5.**

General rules for forming Group frequency distributions

1. Determine the largest and smallest numbers in the raw data and thus find the range.

2. Divide the range into a convenient number of class intervals having the same size. The number of class intervals is usually between 5 and 20, depending on the data.

3. Determine the number of observations falling into each class interval; that is, find the class frequencies.

Frequency Distribution of Qualitative Data

- **Group Frequency Distribution:**
  - **Limit grouping Method**
  - **Cutpoint grouping Method**

Frequency Distribution of Qualitative Data

- **Limit grouping:**
  - use class limits
  - each class consists of a range of values
  - **particularly useful when the data are expressed as whole numbers and there are too many distinct values to employ single-value grouping.**

Frequency table

Example:

- These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

- Construct a (**grouped) frequency distribution** for the data using 7 classes.

Answer

- Range = Highest Value – Lowest Value = H – L
- Range = 134 – 100 = 34
- Width = $\frac{Range\ (R)}{Number\ of\ classes} = \frac{34}{7} = 4.9 \approx 5$
- **Round the answer up to the nearest whole number if there is a remainder:**
- Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

  105 - 1 = 104
- The first class is 100–104, the second class is 105–109, etc.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

frequency distribution

Class limits	Tally	Frequency
100–104	//	2
105–109	𝍸 ///	8
110–114	𝍸 𝍸 𝍸 ///	18
115–119	𝍸 𝍸 ///	13
120–124	𝍸 //	7
125–129	/	1
130–134	/	1
	$n = \Sigma f =$	50

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Frequency Distribution of Qualitative Data

- **Cutpoint Grouping :**
  - A third way to group quantitative data is to use class cutpoints.
  - As with limit grouping, each class consists of a range of values
  - The smallest value that could go in a class is called the lower cutpoint of the class, and the smallest value that could go in the next higher class is called the upper cutpoint of the class.
  - class is the same as its lower limit and that the upper cutpoint of a class is the same as the lower limit of the next higher class.
  - class is the same as its lower limit and that the upper cutpoint of a class is the same as the lower limit of the next higher class.
  - **particularly useful when the data are continuous and are expressed with decimals**

Frequency Distribution

- Weights shown in Table below, given to the nearest tenth of a pound, were obtained from a sample of 18- to 24-year-old males. Use cutpoint grouping to organize these data into frequency and relative-frequency distributions. Use a class width of 20 and a first cutpoint of 120.

129.2	185.3	218.1	182.5	142.8
155.2	170.0	151.3	187.5	145.6
167.3	161.0	178.7	165.0	172.5
191.1	150.7	187.0	173.7	178.2
161.7	170.1	165.8	214.6	136.7
278.8	175.6	188.7	132.1	158.5
146.4	209.1	175.4	182.0	173.6
149.9	158.6			

Frequency Distribution

Weight (lb)	Frequency	Relative frequency
120–under 140	3	0.081
140–under 160	9	0.243
160–under 180	14	0.378
180–under 200	7	0.189
200–under 220	3	0.081
220–under 240	0	0.000
240–under 260	0	0.000
260–under 280	1	0.027
	37	0.999

Note:

Although relative frequencies must always sum to 1, their sum in Table above is given as 0.999. This discrepancy occurs because each relative frequency is rounded to three decimal places, and, in this case, the resulting sum differs from 1 by a little. Such a discrepancy is called rounding error or roundoff error.

Choosing the Grouping Method

- There are three methods for grouping quantitative data: single-value grouping, limit grouping, and cutpoint grouping.
- The following table provides guidelines for deciding which grouping method should be used.

Grouping method	When to use
Single-value grouping	Use with discrete data in which there are only a small number of distinct values.
Limit grouping	Use when the data are expressed as whole numbers and there are too many distinct values to employ single-value grouping.
Cutpoint grouping	Use when the data are continuous and are expressed with decimals.

Histogram

- Another method for organizing and summarizing data is to draw a picture of some kind.

- Three common methods for graphically displaying quantitative data are histograms, dotplots, and stem-and-leaf diagrams.

- A histogram of quantitative data is the direct analogue of a bar chart of qualitative data, where we use the classes of the quantitative data in place of the distinct values of the qualitative data.

- Position the bars in a histogram so that they touch each other
- Frequencies, relative frequencies, or percents can be used to label a histogram.

Histogram

- A histogram displays the classes of the quantitative data on a horizontal axis and the frequencies (relative frequencies, percents) of those classes on a vertical axis.

- The frequency (relative frequency, percent) of each class is represented by a vertical bar whose height is equal to the frequency (relative frequency, percent) of that class.

- The bars should be positioned so that they touch each other
  - For **single-value grouping**, we use the distinct values of the observations to label the bars, with each such value centered under its bar.
  - For **limit grouping or cutpoint grouping**, we use the lower class limits (or, equivalently, lower class cutpoints) to label the bars. Note: Some statisticians and technologies use class marks or class midpoints centered under the bars.

Construct a Histogram - Steps

1. Obtain a frequency (relative-frequency, percent) distribution of the data.

2. Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies (relative frequencies, percents).

3. For each class, construct a vertical bar whose height equals the frequency (relative frequency, percent) of that class.

4. Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with “Frequency” (“Relative frequency,” “Percent”)

Construct Histogram for each

Number of TVs	Frequency	Relative frequency
0	1	0.02
1	16	0.32
2	14	0.28
3	12	0.24
4	3	0.06
5	2	0.04
6	2	0.04

(a) Single-value grouping

Days to maturity	Frequency	Relative frequency
30–39	3	0.075
40–49	1	0.025
50–59	8	0.200
60–69	10	0.250
70–79	7	0.175
80–89	7	0.175
90–99	4	0.100

(b) Limit grouping

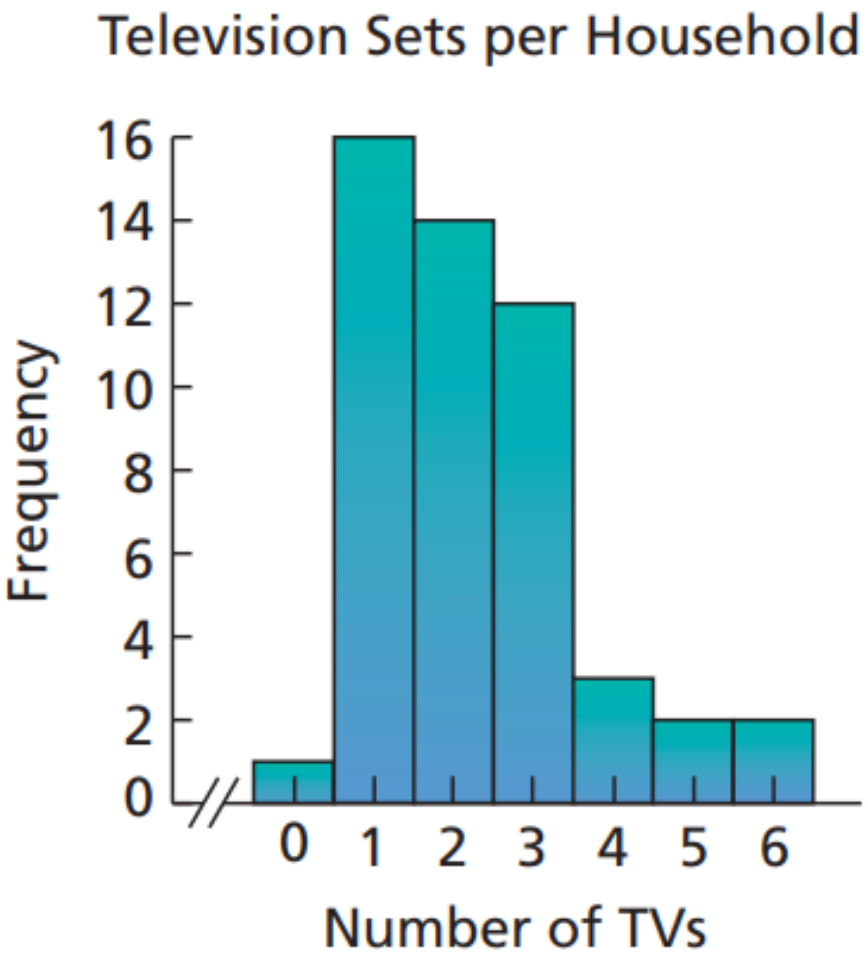
Weight (lb)	Frequency	Relative frequency
120–under 140	3	0.081
140–under 160	9	0.243
160–under 180	14	0.378
180–under 200	7	0.189
200–under 220	3	0.081
220–under 240	0	0.000
240–under 260	0	0.000
260–under 280	1	0.027

(c) Cutpoint grouping

Histogram for ungrouped frequencies

Number of TVs	Frequency	Relative frequency
0	1	0.02
1	16	0.32
2	14	0.28
3	12	0.24
4	3	0.06
5	2	0.04
6	2	0.04

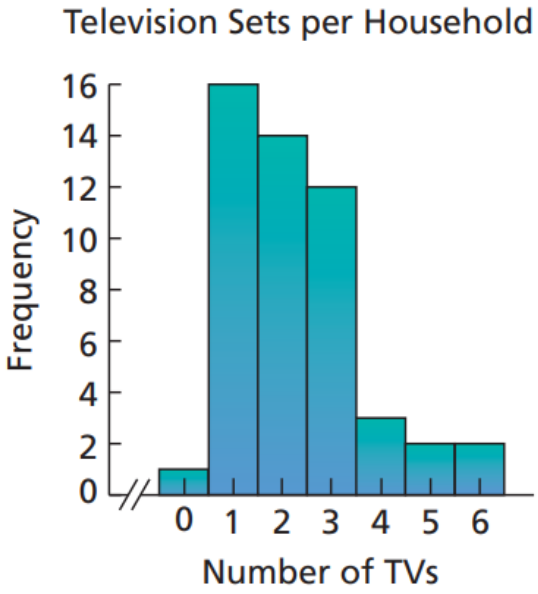
(a) Single-value grouping

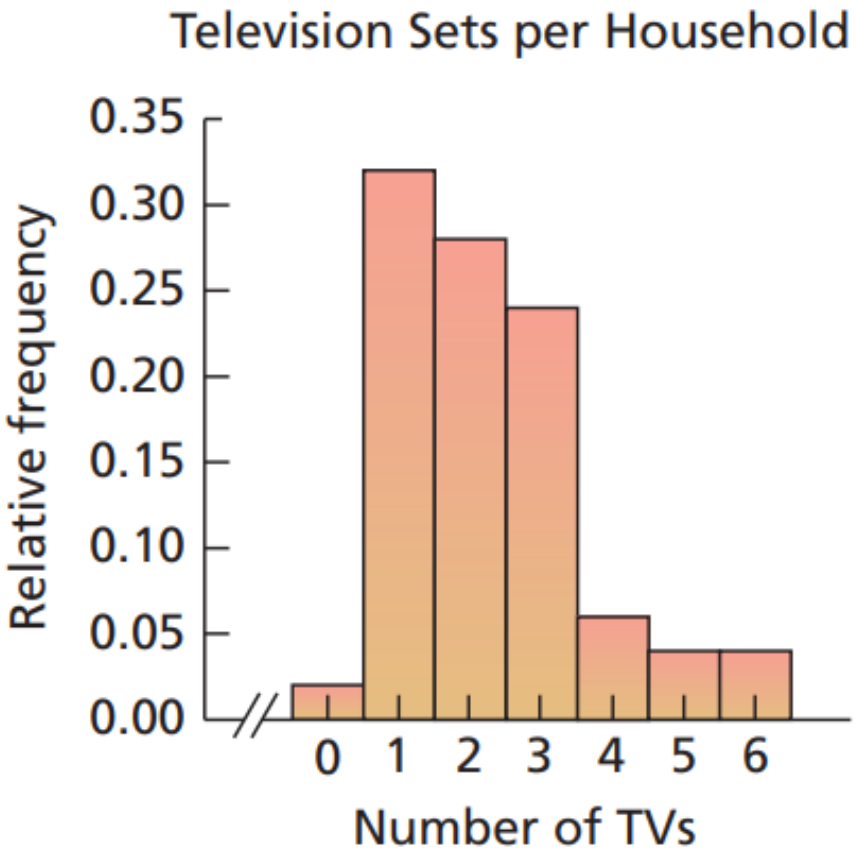


Histogram

Number of TVs	Frequency	Relative frequency
0	1	0.02
1	16	0.32
2	14	0.28
3	12	0.24
4	3	0.06
5	2	0.04
6	2	0.04

(a) Single-value grouping





Activity

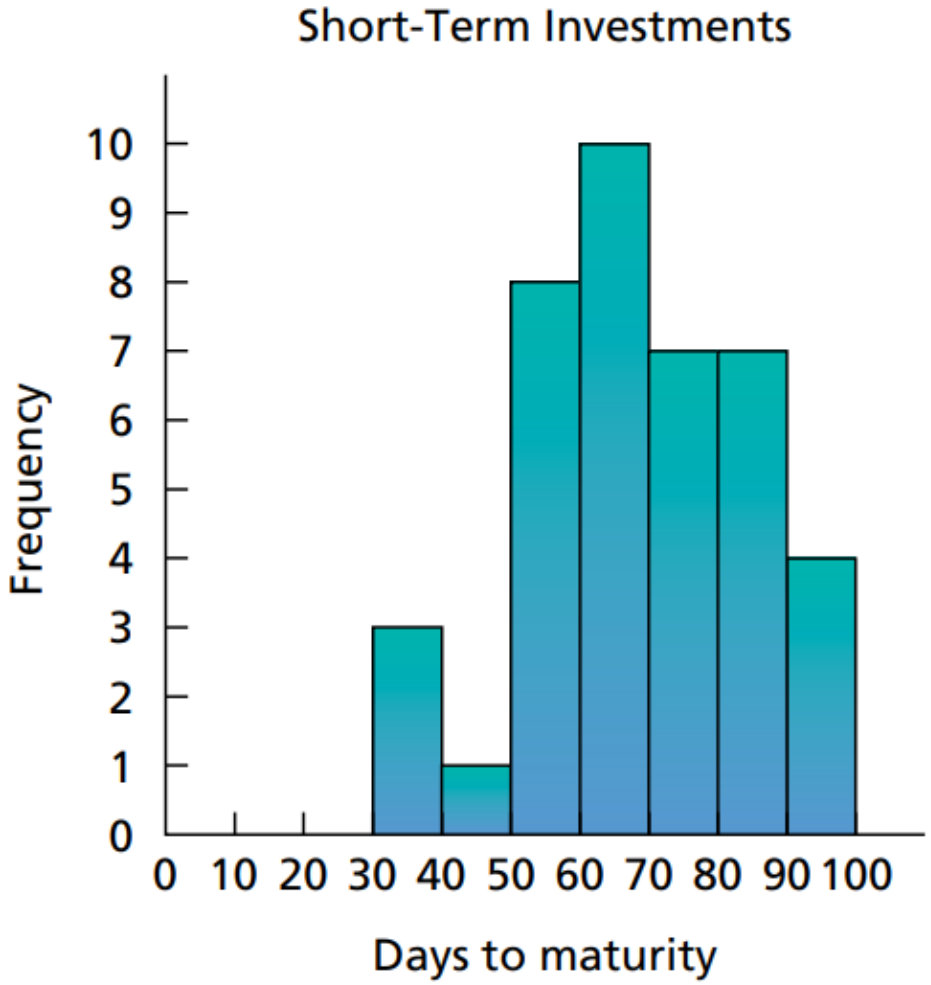
- Construct the Frequency Histogram

Number of Cars	*Tally*	*Frequency*	*Relative Frequency*
0	|||	4	4/50 = 0.08
1	|||| |||| |||	13	13/50 = 0.26
2	|||| |||| |||| |||| ||	22	0.44
3	|||| ||	7	0.14
4	|||	3	0.06
5	|	1	0.02

Histogram for Grouped frequencies

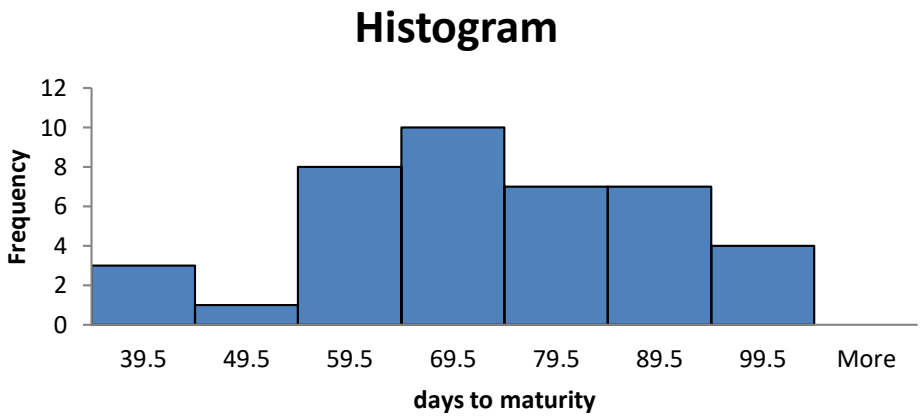
Days to maturity	Frequency	Relative frequency
30–39	3	0.075
40–49	1	0.025
50–59	8	0.200
60–69	10	0.250
70–79	7	0.175
80–89	7	0.175
90–99	4	0.100

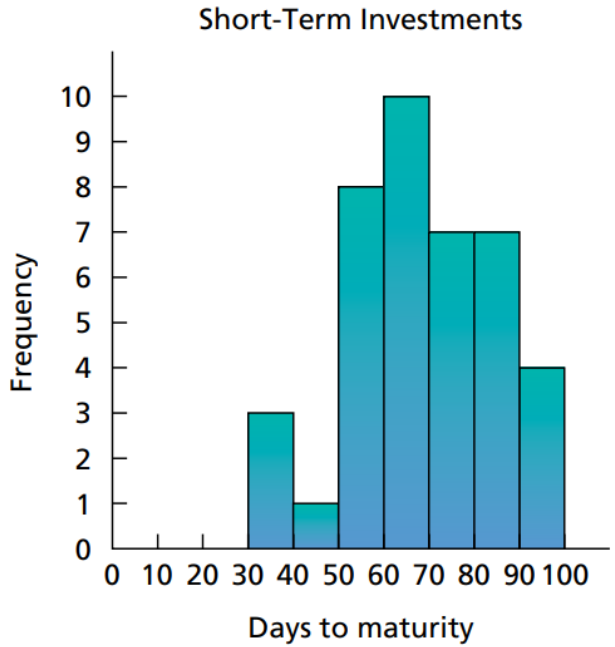
(b) Limit grouping



Histogram for Grouped frequencies

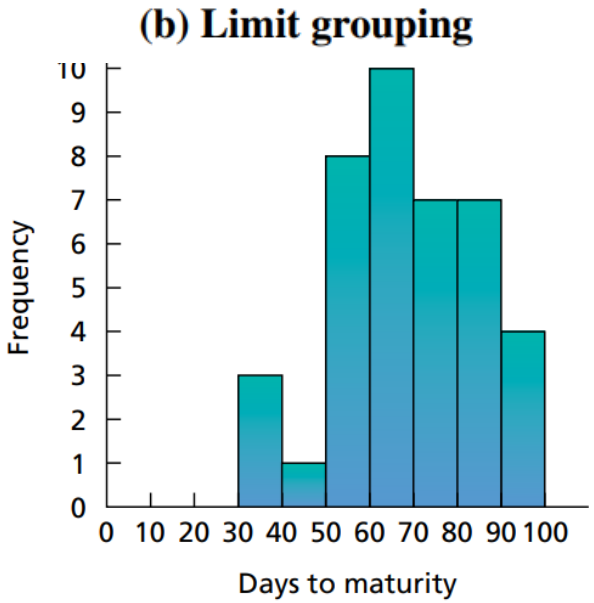
class limits	Class boundaries	Frequency
30-39	29.5-39.5	3
40-49	39.5-49.5	1
50-59	49.5-59.5	8
60-69	59.5-69.5	10
70-79	69.5-79.5	7
80-89	79.5-89.5	7
90-99	89.5-99.5	4

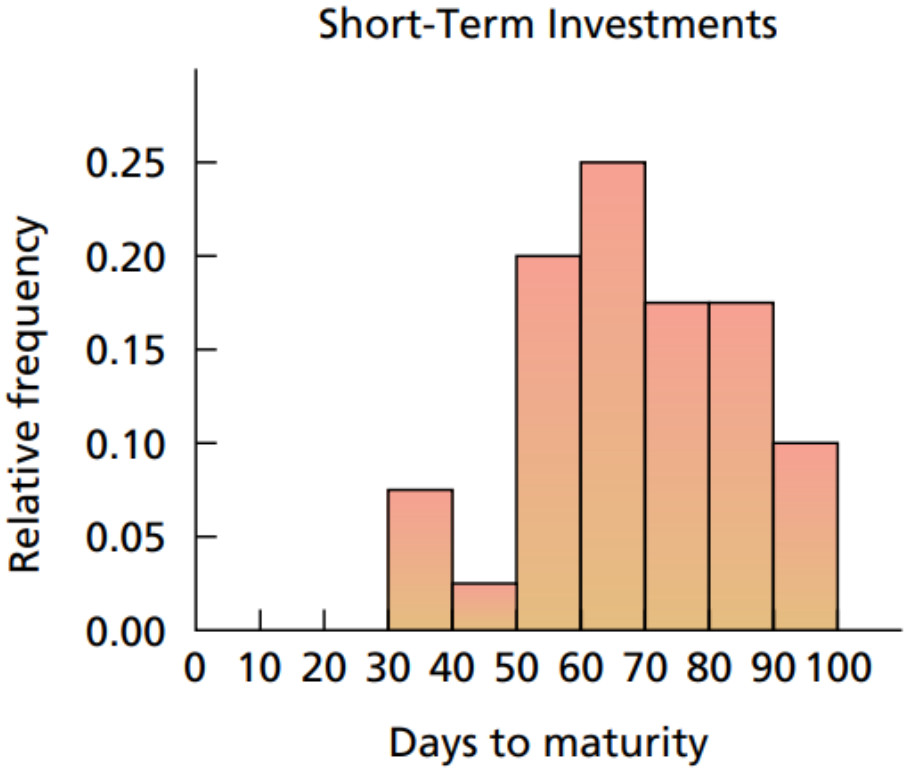




Histogram for Grouped frequencies

Days to maturity	Frequency	Relative frequency
30–39	3	0.075
40–49	1	0.025
50–59	8	0.200
60–69	10	0.250
70–79	7	0.175
80–89	7	0.175
90–99	4	0.100





Continuous data - Grouped frequency table

Example:

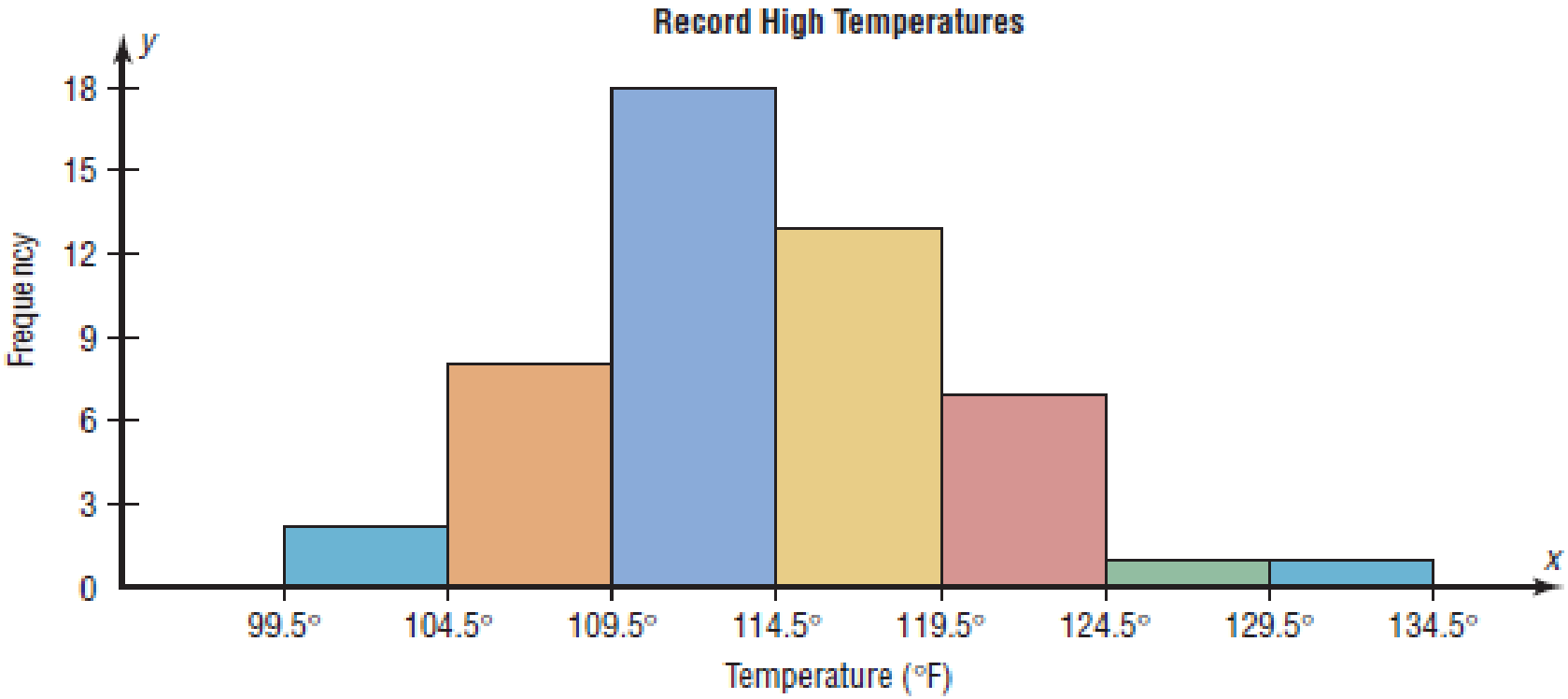
- These data represent the record high temperatures in degrees Fahrenheit (F) for each of the 50 states.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

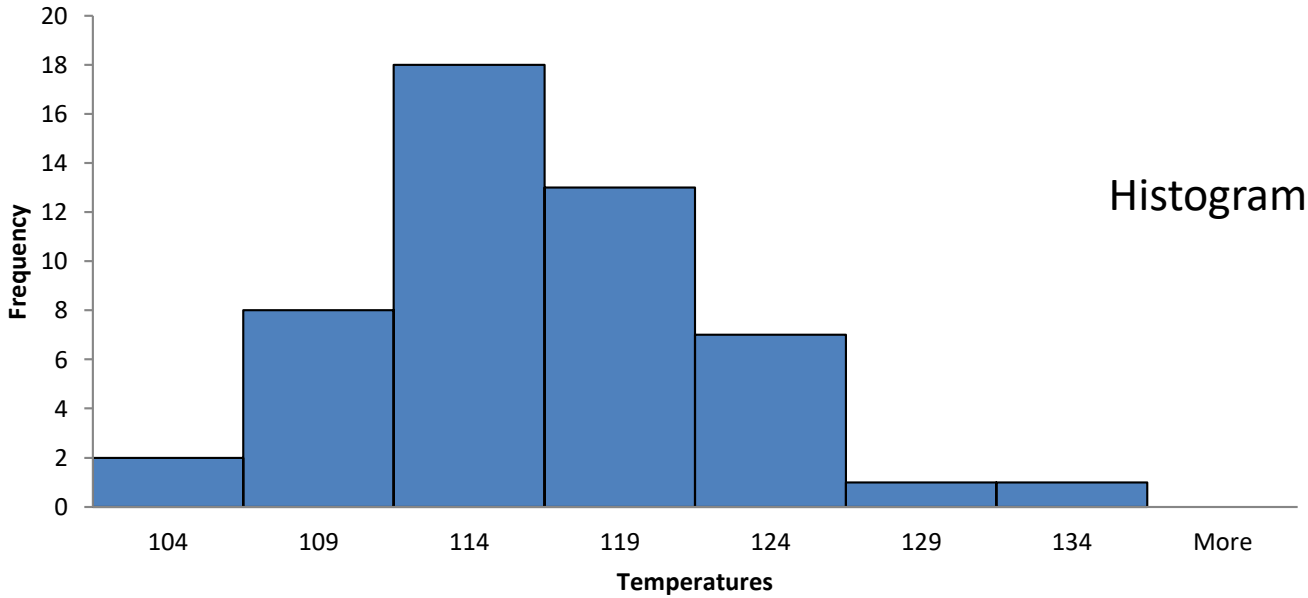
- Construct a **grouped frequency distribution** for the data using 7 classes.
- Construct a **Histogram** to represent the data shown for the record high temperatures for each of the 50 states.
- construct a **Frequency Polygon.**

Histogram for Grouped frequencies

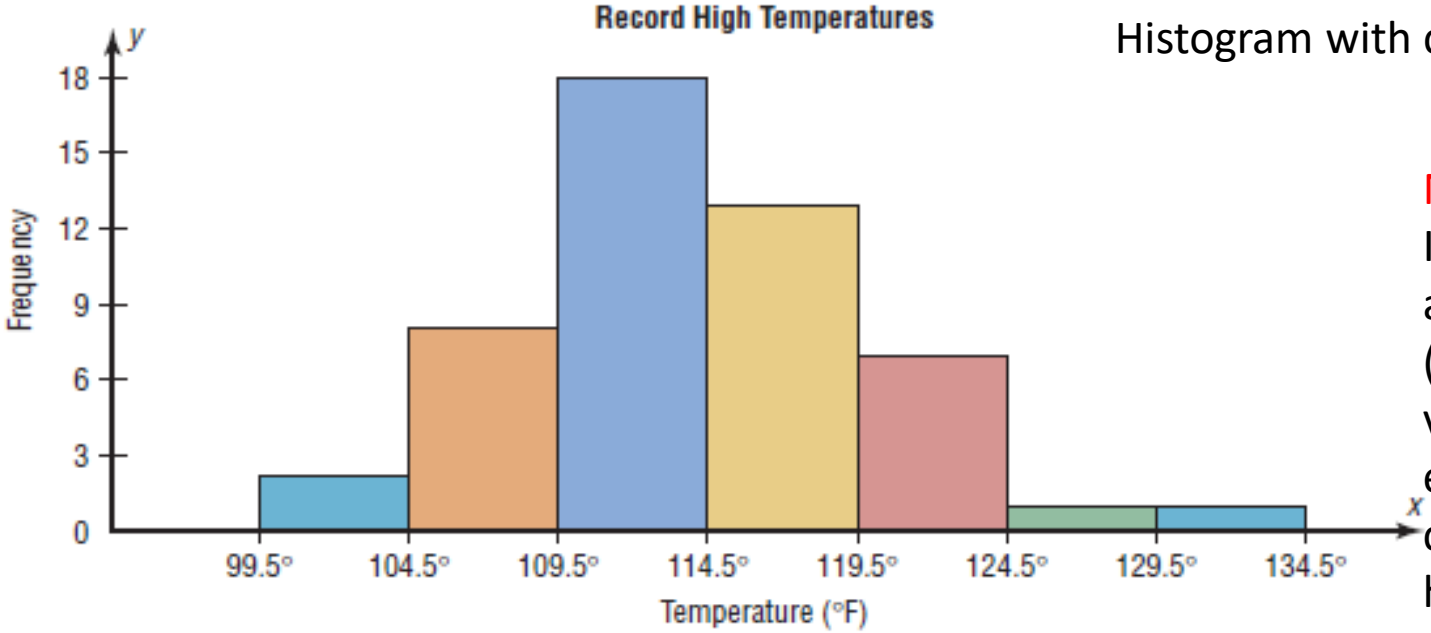
Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	𝍸 ///	8
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18
115–119	114.5–119.5	𝍸 𝍸 ///	13
120–124	119.5–124.5	𝍸 //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
		$n = \Sigma f =$	50



Histogram



Histogram with lower-class limits



Histogram with class boundaries

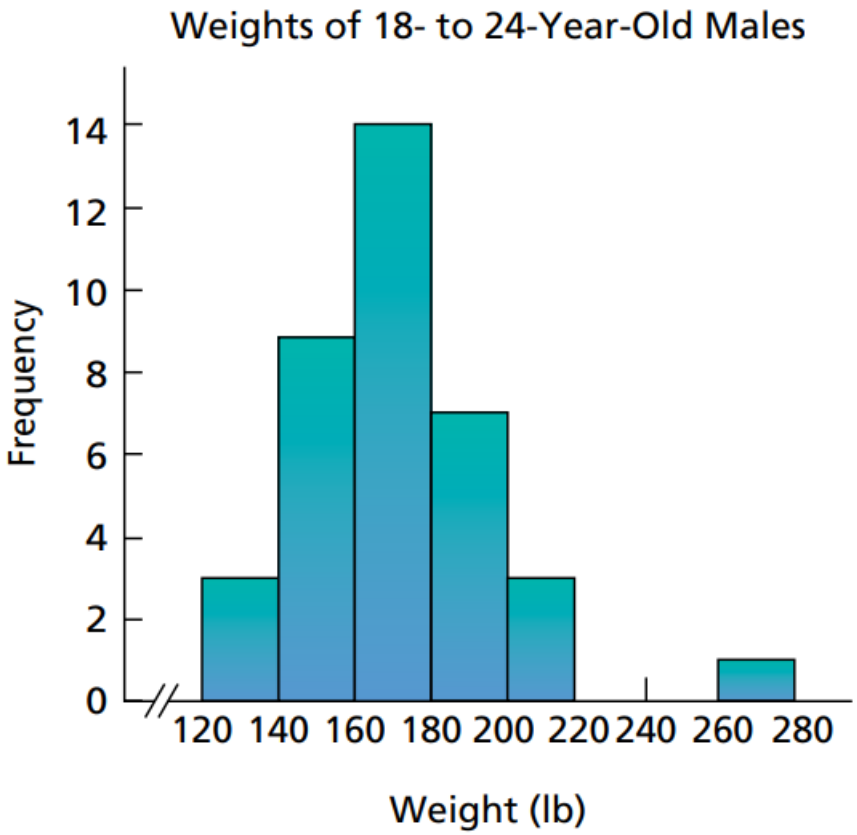
Note:

If the frequencies are discrete values (not decimal values) you can even use the lower class limit to draw a histogram.

Histogram for Grouped frequencies

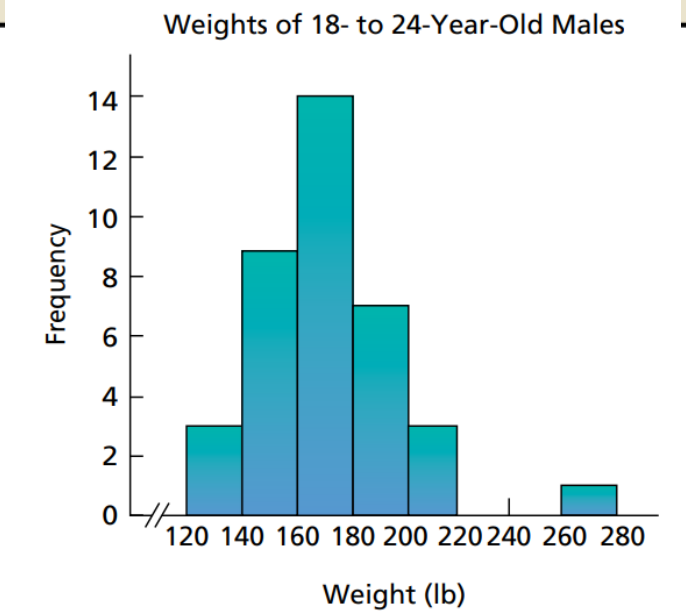
Weight (lb)	Frequency	Relative frequency
120–under 140	3	0.081
140–under 160	9	0.243
160–under 180	14	0.378
180–under 200	7	0.189
200–under 220	3	0.081
220–under 240	0	0.000
240–under 260	0	0.000
260–under 280	1	0.027

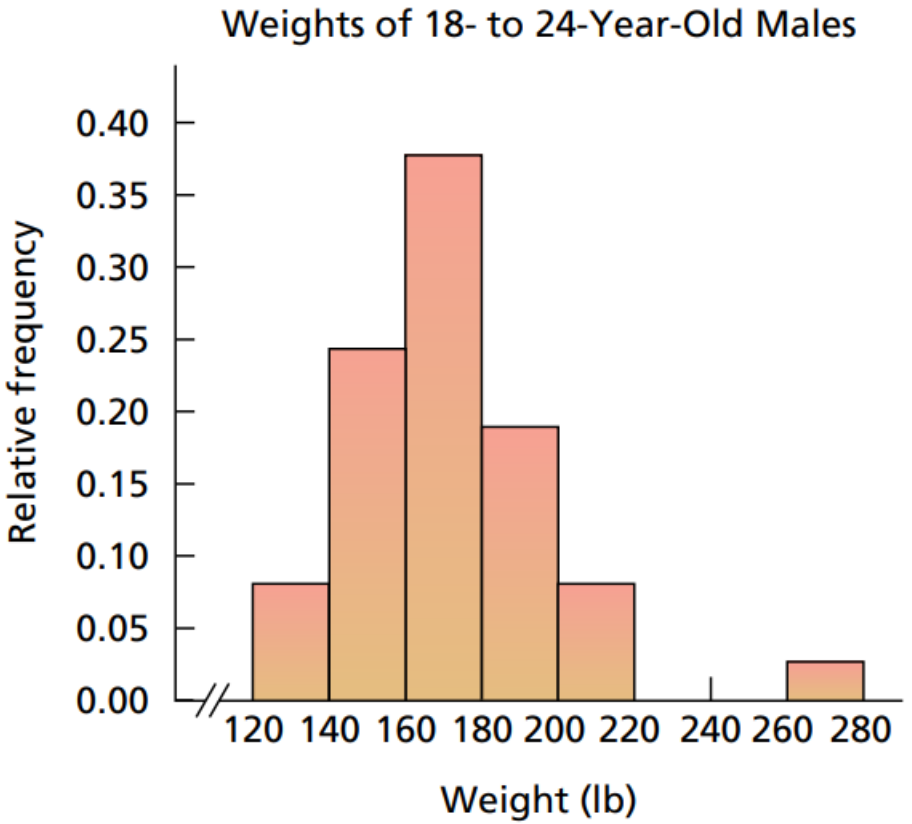
(c) Cutpoint grouping



Histogram

Weight (lb)	Frequency	Relative frequency
120–under 140	3	0.081
140–under 160	9	0.243
160–under 180	14	0.378
180–under 200	7	0.189
200–under 220	3	0.081
220–under 240	0	0.000
240–under 260	0	0.000
260–under 280	1	0.027





The Frequency Polygon

- Another way to represent the same data set is by using a **Frequency Polygon**.

- The frequency polygon is a graph that displays the data by using lines that connect points plotted for the frequencies at the **midpoints** of the classes.

- The frequencies are represented by the heights of the points.

Construct a Frequency Polygon

- Find the midpoints of each class.

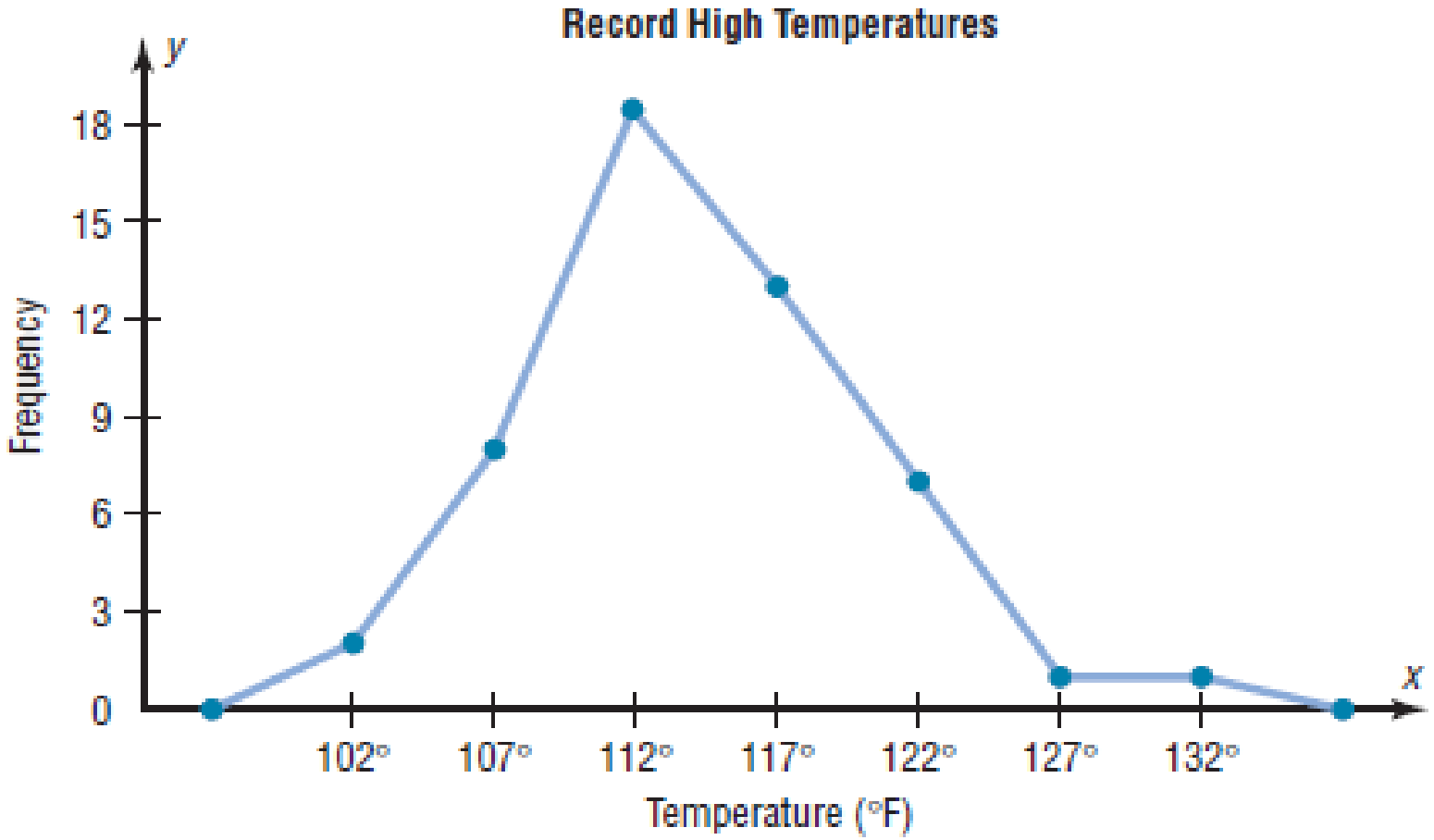
$$\frac{99.5+104.5}{2} = 102$$

$$\frac{104.5+109.5}{2} = 107$$

And so on. The midpoints are

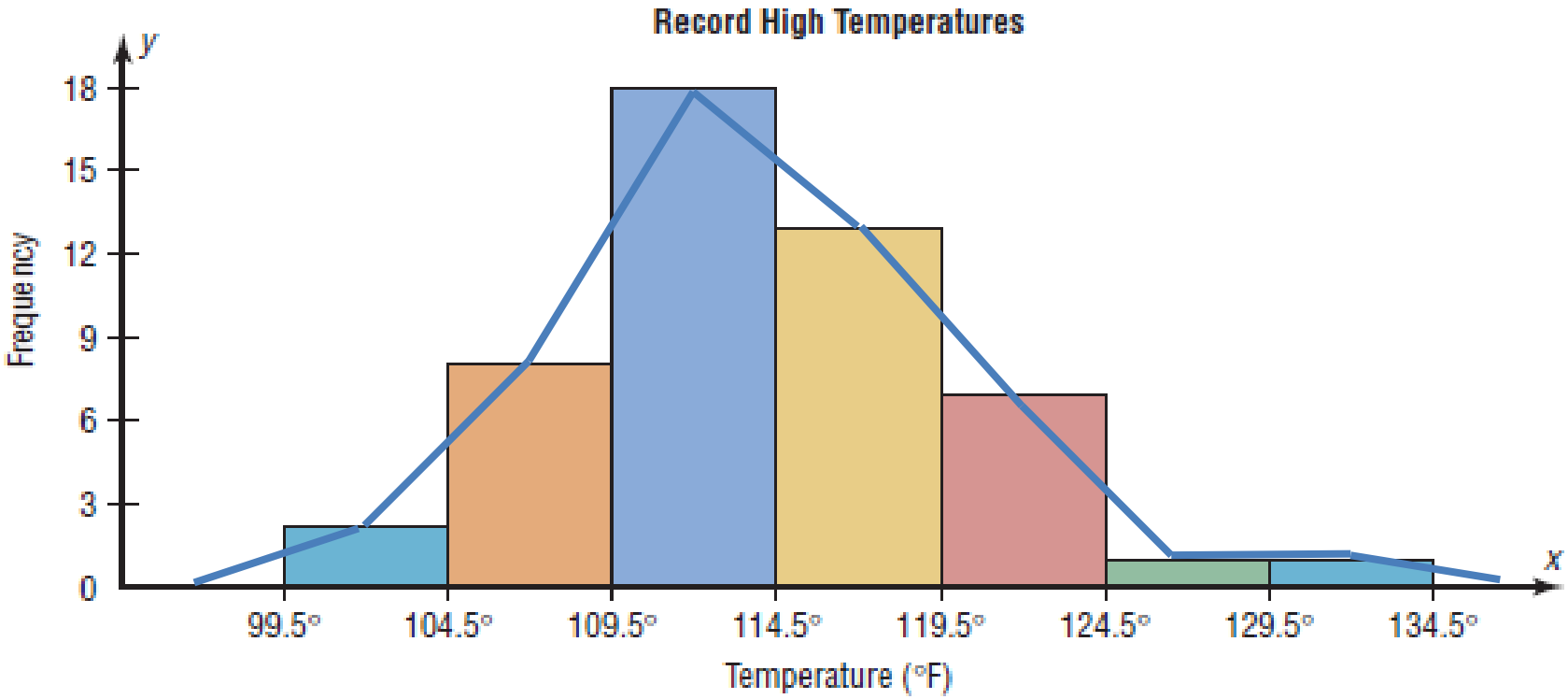
Class Boundaries	Mid Points	Frequency
99.5-104.5	102	2
104.5-109.5	107	8
109.5-114.5	112	18
114.5-119.5	117	13
119.5-124.5	122	7
124.5-129.5	127	1
129.5-134.5	132	1

Construct a Frequency Polygon



Frequency Polygon

Class limits	Class boundaries	Tally	Frequency
100–104	99.5–104.5	//	2
105–109	104.5–109.5	𝍸 ///	8
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18
115–119	114.5–119.5	𝍸 𝍸 ///	13
120–124	119.5–124.5	𝍸 //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
		$n = \Sigma f =$	50



Frequency distribution VS Relative Frequency distribution

- Relative-frequency distributions are better than frequency distributions for comparing two data sets. Because relative frequencies always fall between 0 and 1, they provide a standard for comparison.

Histogram VS Bar chart

Bar chat

- Typically used for categorical data, where each bar represents a category.
- Bars are spaced apart with gaps between them.
- Both axes (X and Y) have distinct categories or labels.
- Each bar stands for a separate category, and the length or height of the bar represents a value associated with that category.

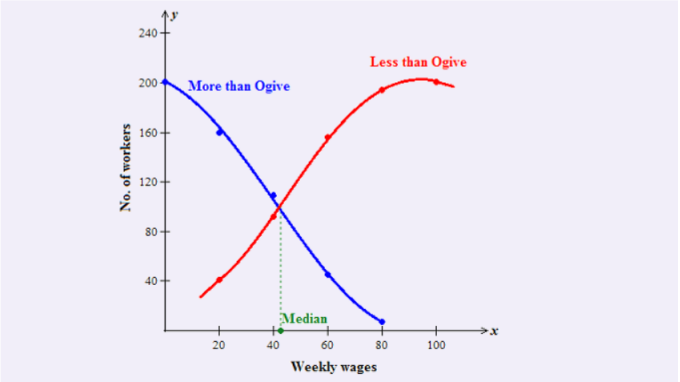
Histogram

- Used for continuous data, where the data is divided into intervals or bins.
- Bars are adjacent and touch each other.
- The X-axis represents the range of values (intervals or bins), and the Y-axis represents the frequency or count of data points.
- Bars represent ranges or intervals of continuous data, and the area of each bar represents the frequency or count of data points within that range.

In summary, while both bar charts and histograms use bars to represent data, the key distinction lies in the type of data they are designed to display. Bar charts are suitable for categorical data, while histograms are used for visualizing the distribution of continuous data.

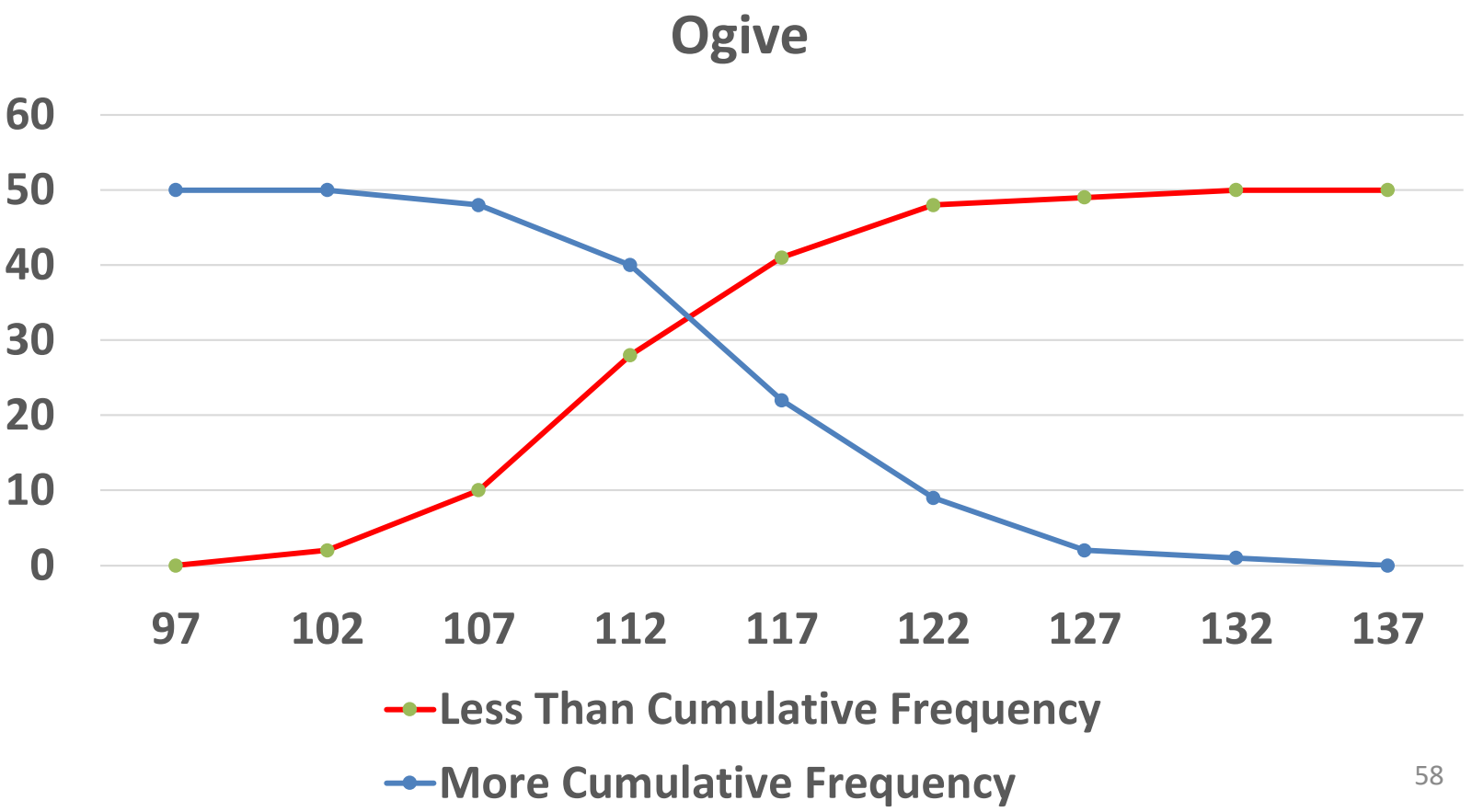
The Ogive

- The **Ogive** is a graph that represents the cumulative frequencies for the classes in a frequency distribution.
- There are two types: "**less than**" ogive and "more **than**" ogive.
  - **Less than Ogive:** This graph shows the cumulative frequency of values less than the upper class boundary.
  - **More than Ogive:** This graph shows the cumulative frequency of values greater than the lower class boundary.



Class Boundaries	Mid Points	Frequency	Less Than Cumulative Frequency	More Cumulative Frequency
	97		0	50
99.5-104.5	102	2	2	50
104.5-109.5	107	8	10	48
109.5-114.5	112	18	28	40
114.5-119.5	117	13	41	22
119.5-124.5	122	7	48	9
124.5-129.5	127	1	49	2
129.5-134.5	132	1	50	1
	137		50	0

Class Boundaries	Mid Points	Frequency	Less Than Cumulative Frequency	More Cumulative Frequency
	97		0	50
99.5-104.5	102	2	2	50
104.5-109.5	107	8	10	48
109.5-114.5	112	18	28	40
114.5-119.5	117	13	41	22
119.5-124.5	122	7	48	9
124.5-129.5	127	1	49	2
129.5-134.5	132	1	50	1
	137		50	0



Activity

1. Sample of 30 persons is showed their ages as follows;

20	18	25	68	23	25	16	22	29	37
35	49	42	65	37	42	63	65	49	42
53	48	65	72	69	57	48	39	58	67

a) Construct a frequency distribution for this data by selecting class width as 10. set the class range as 10-19.

b) Draw the Less than Ogive and More than Ogive graphs.

END