2. ORGANIZING AND GRAPHING DATA

Frequency distribution tables, graphs, plots

Raw data

- Raw data – unprocessed data.
- Raw data is not organized, often recorded in sequence in which they are collected.
- They must be organized or processed into a manageable form.

Raw data

- Example of quantitative raw data:

Table 2.1 Ages of 50 Students

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

- Example of qualitative raw data:

Table 2.2 Status of 50 Students

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

Qualitative data

Example (donuts)

EXAMPLE 2–1 What Variety of Donuts Is Your Favorite?

A sample of 30 persons who often consume donuts were asked what variety of donuts is their favorite. The responses from these 30 persons are as follows:

glazed	filled	other	plain	glazed	other
frosted	filled	filled	glazed	other	frosted
glazed	plain	other	glazed	glazed	filled
frosted	plain	other	other	frosted	filled
filled	other	frosted	glazed	glazed	filled

- Categorical variables (glazed, filled, frosted, plain, other).
- Total 30 observations.

Frequency distribution

- Frequency is how often a category occurs.

- Relative frequency is the proportion (or percent) of times that the category occurs and is found using the formula:

$$\text{relative frequency} = \frac{\text{frequency of a category}}{\text{sum of all frequencies}}$$

- Percentage:

$$\text{percentage} = \text{relative frequency} \times 100\%$$

EXAMPLE 2–1 What Variety of Donuts Is Your Favorite?

A sample of 30 persons who often consume donuts were asked what variety of donuts is their favorite. The responses from these 30 persons are as follows:

glazed	filled	other	plain	glazed	other
frosted	filled	filled	glazed	other	frosted
glazed	plain	other	glazed	glazed	filled
frosted	plain	other	other	frosted	filled
filled	other	frosted	glazed	glazed	filled

- Frequency distribution table:

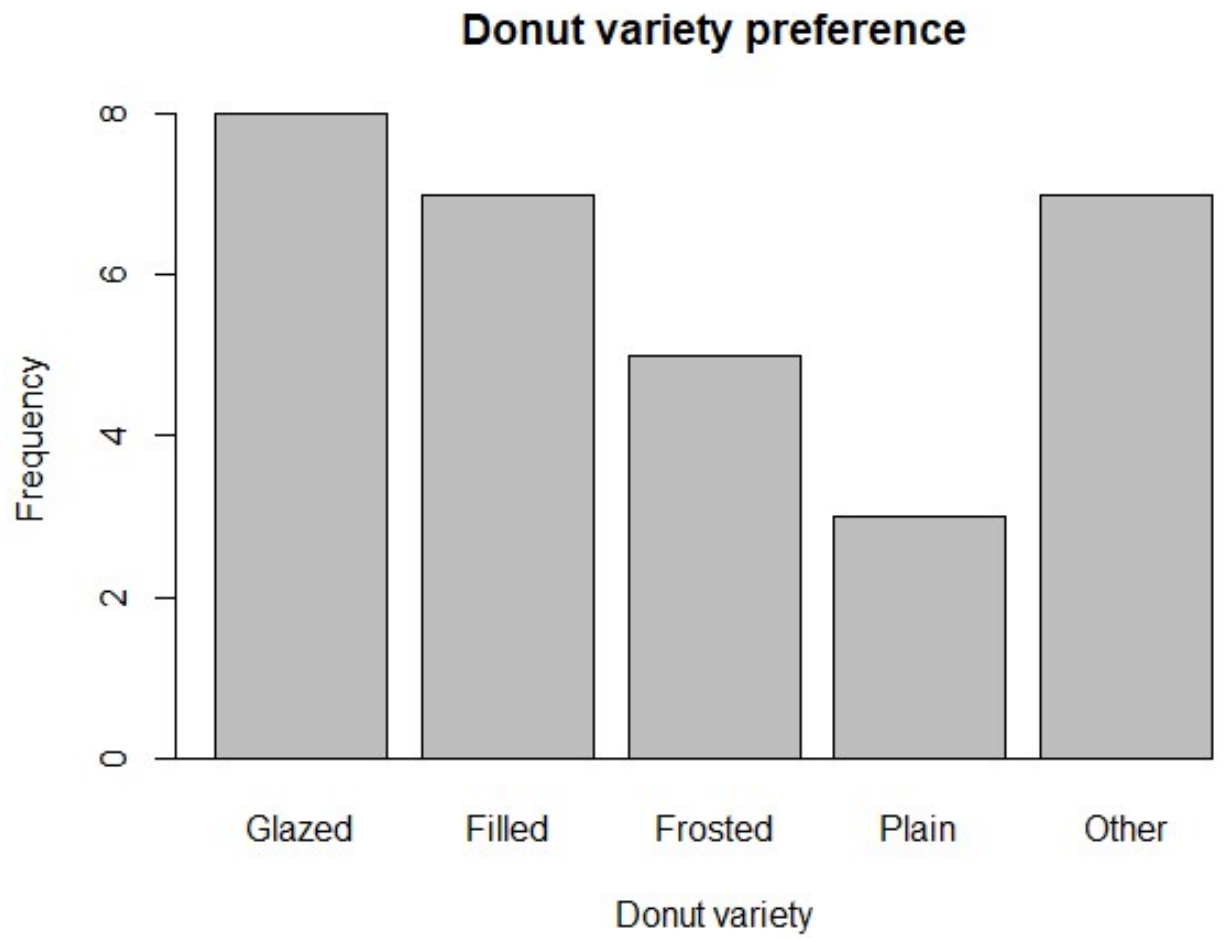
Donut Variety	Frequency	Relative frequency	Percentage (%)
Glazed	8	$8/30 = 0.267$	26.7
Filled	7	$7/30 = 0.233$	23.3
Frosted	5	$5/30 = 0.167$	16.7
Plain	3	$3/30 = 0.100$	10.0
Other	7	$7/30 = 0.233$	23.3
Total	30	1	100%

Bar graph

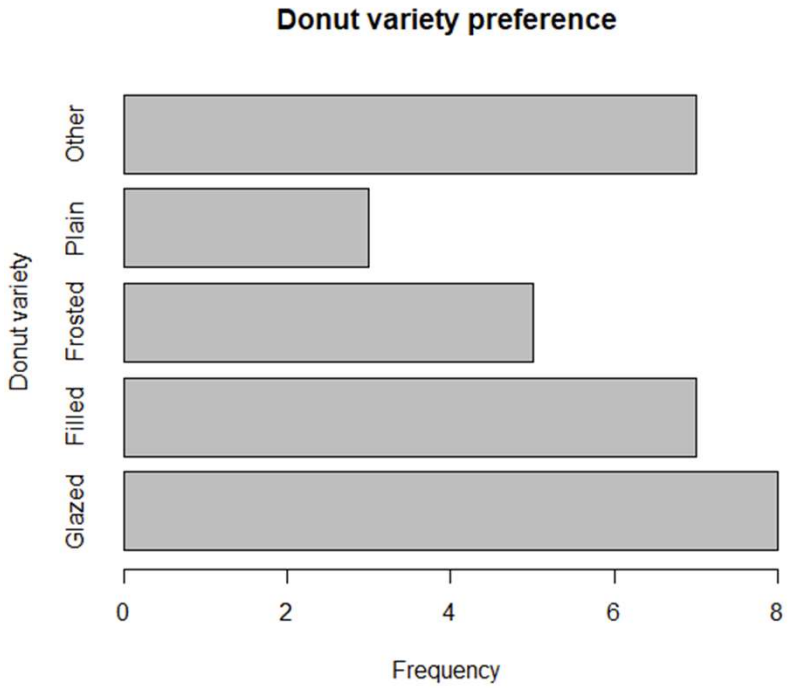
- We can use bar graph to visually inspect data.

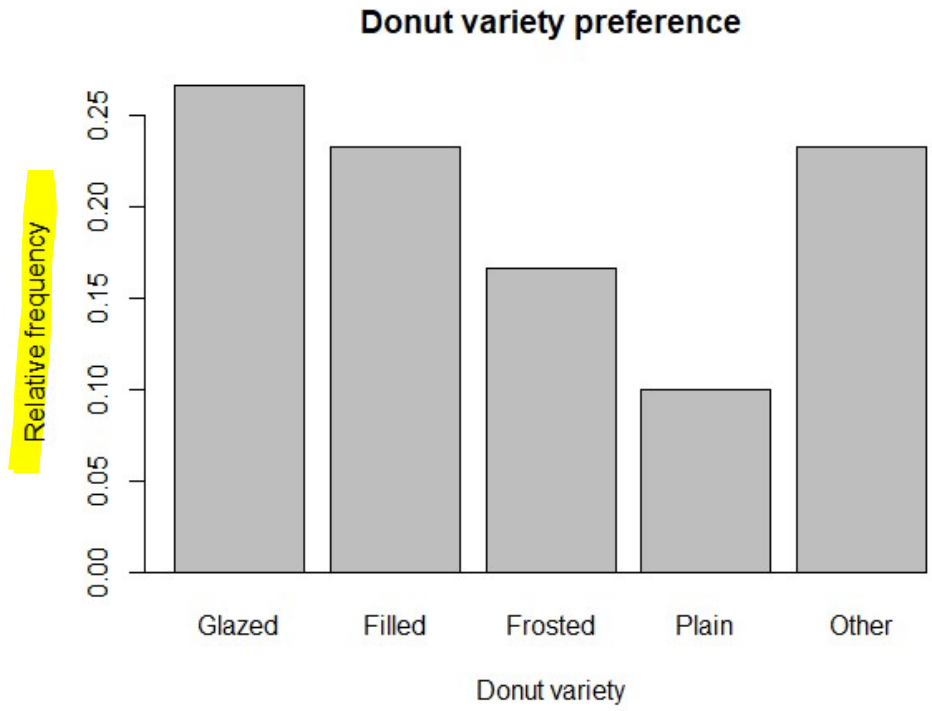
- Steps:
  1. Draw the x and y axes.
  2. Mark the various categories on the x-axis.
  3. Mark the frequencies on the y-axis.
  4. Draw one bar for each category such that the height of the bar represents the frequency of the corresponding category.
  5. Write down the correct label and title.

Example


Donut variety preference
Frequency
0
2
4
6
8
Glazed
Filled
Frosted
Plain
Other
Donut variety

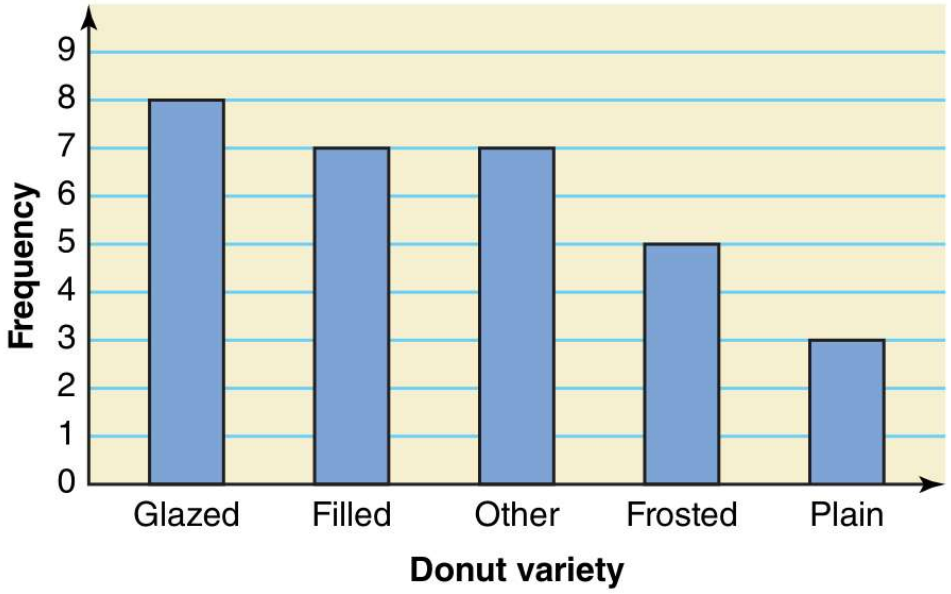
Bar graph (alternative forms)

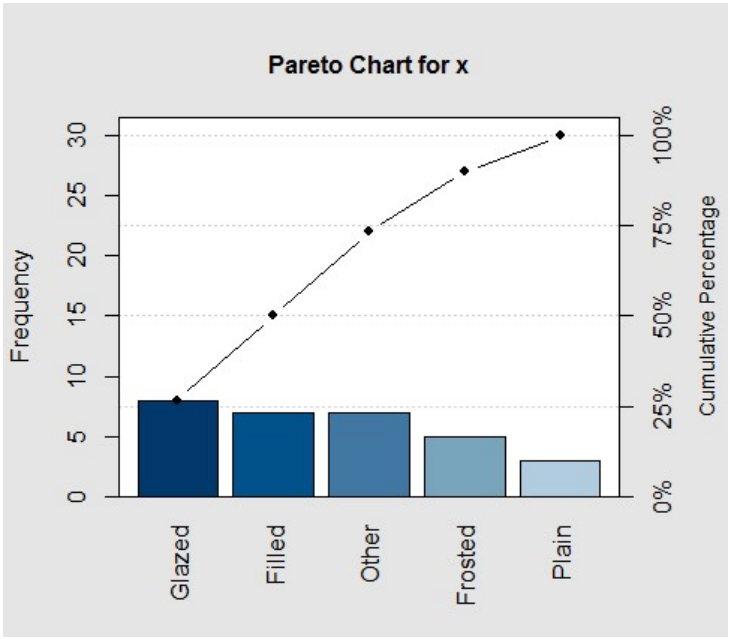




Pareto chart

- A Pareto chart is a bar graph with bars arranged by their heights in descending order.





Pie chart

- A pie chart is a circle divided into sectors. Each sector represents a category of data.

- The angle of each sector is proportional to the frequency of the category.

$$\text{Angle of a sector} = \frac{\text{frequency of a category}}{\text{sum of all frequency}} \times 360°$$

Pie chart

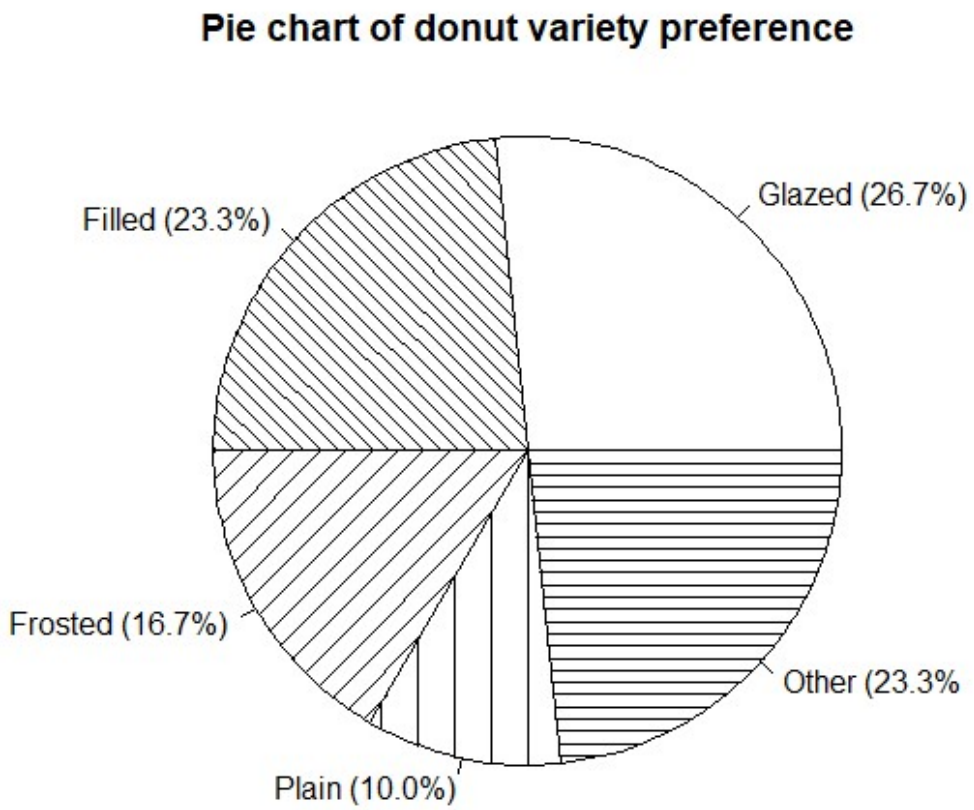
Steps:

1. Calculate the angle for each category.
2. Draw a circle to represent all the data.
3. Divide the circle into portions according to the angle of each category.
4. Write down the correct label and title.

Example

Donut Variety	Frequency	Percentage (%)	Angle (°)
Glazed	8	26.7	$8/30 \times 360 = 96$
Filled	7	23.3	$7/30 \times 360 = 84$
Frosted	5	16.7	$5/30 \times 360 = 60$
Plain	3	10.0	$3/30 \times 360 = 36$
Other	7	23.3	$7/30 \times 360 = 84$
Total	30	100%	360°

Example


Pie chart of donut variety preference
Glazed (26.7%)
Filled (23.3%)
Frosted (16.7%)
Other (23.3%
Plain (10.0%)

Exercise

2.3 The following data give the results of a sample survey. The letters A, B, and C represent the three categories.

A	B	B	A	C	B	C	C	C	A
C	B	C	A	C	C	B	C	C	A
A	B	C	C	B	C	B	A	C	A

a. Prepare a frequency distribution table.
b. Calculate the relative frequencies and percentages for all categories.
c. What percentage of the elements in this sample belong to category B?
d. What percentage of the elements in this sample belong to category A or C?
e. Draw a bar graph for the frequency distribution.

a)

Category	Frequency	Rel. Freq.	Percentages (%)
A	8	$8/30 = 4/15$	$4/15 \times 100\% = 26.7$
B	8	$8/30 = 4/15$	26.7
C	14	$14/30 = 7/15$	46.7
Total	30	1	100

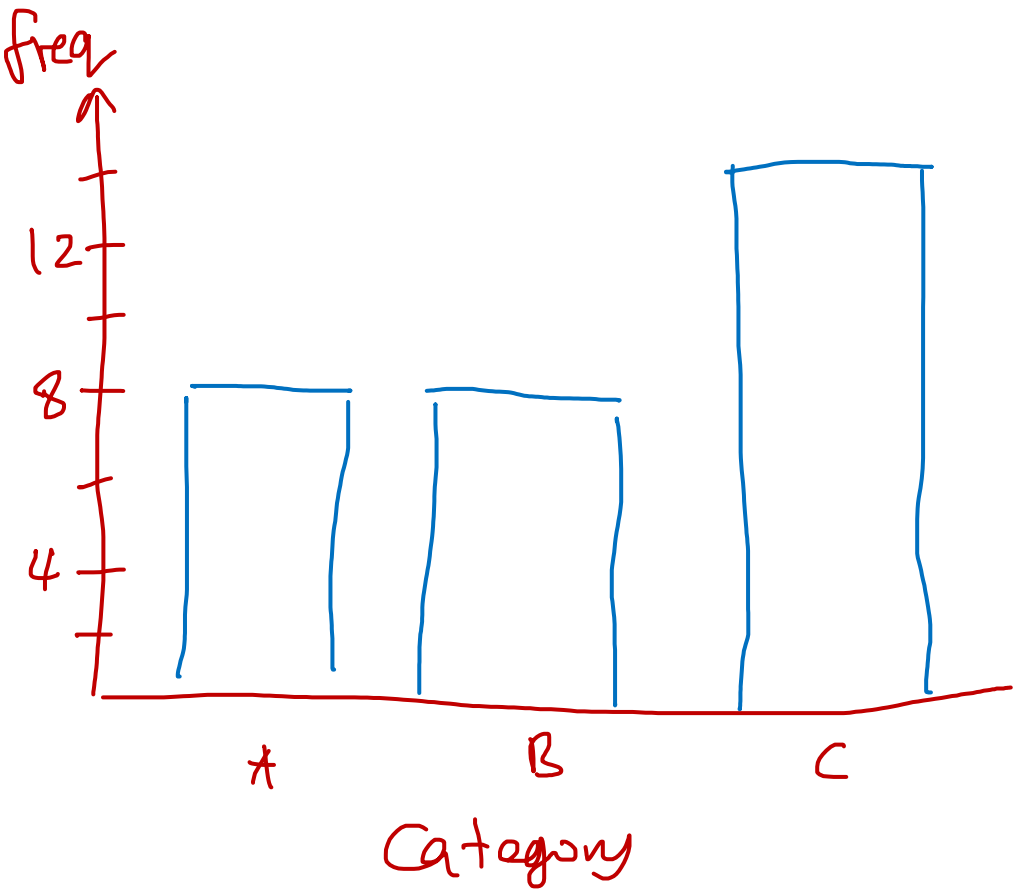
b)

c) 26.7%

d) 26.7 + 46.7 (A) (C)

= 73.3%

e)



Exercise

2.6 Thirty adults were asked which of the following conveniences they would find most difficult to do without: television (T), refrigerator (R), air conditioning (A), public transportation (P), or microwave (M). Their responses are listed below.

R	A	R	P	P	T	R	M	P	A
A	R	R	T	P	P	T	R	A	A
R	P	A	T	R	P	R	A	P	R

a. Prepare a frequency distribution table.

b. Calculate the relative frequencies and percentages for all categories.

c. What percentage of these adults named refrigerator or air conditioning as the convenience that they would find most difficult to do without?

d. Draw a bar graph for the relative frequency distribution.

a)

b)

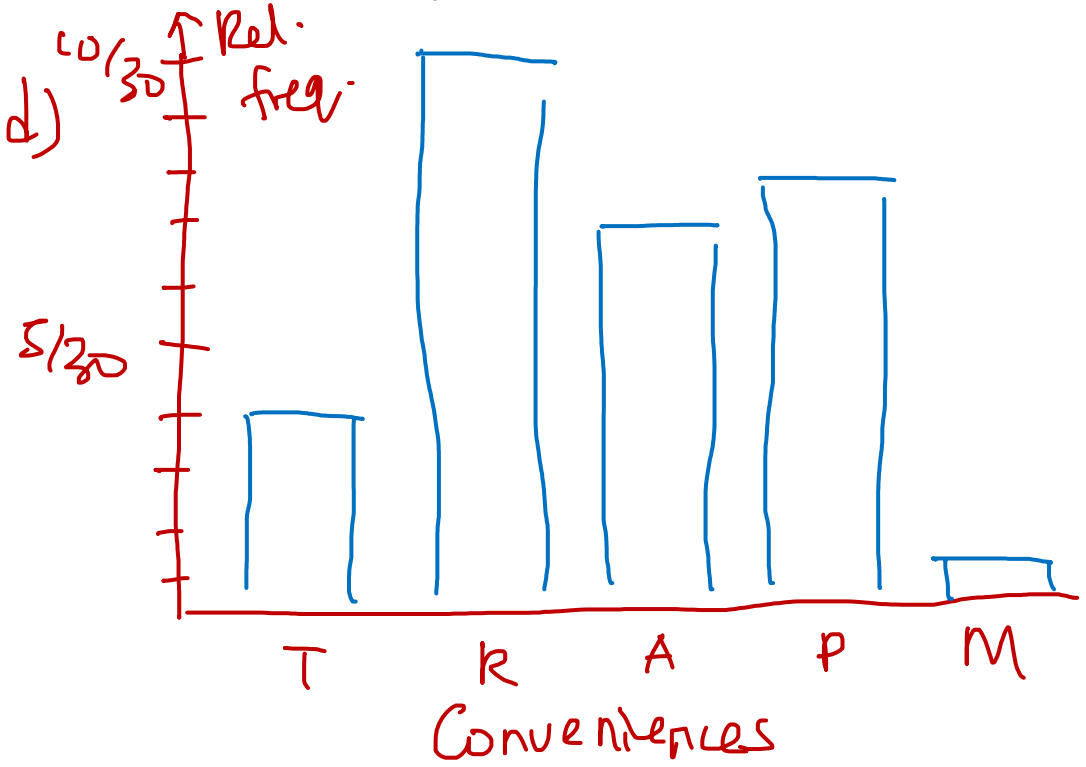
Conveniences	Freq.	Rel. Freq.	Percentage (%)
T	4	$4/30 = 2/15$	$2/15 \times 100\% = 13.33$
R	10	$1/3$	33.33
A	7	$7/30$	23.33
P	8	$4/15$	26.67
M	1	$1/30$	3.33
Total	30	1	100

c) (R) (A)

$33.33 + 23.33$

$= 56.67\%$

d)



Quantitative data

Organizing quantitative data

- The first step in summarizing quantitative data is to determine whether the data are discrete or continuous.
- If the data is discrete with few different values, we can use similar techniques as the categorical data.

Table 2.12 Frequency Distribution of the Number of Vehicles Owned

Vehicles Owned	Number of Households (f)
0	2
1	18
2	11
3	4
4	3
5	2
	$\Sigma f = 40$

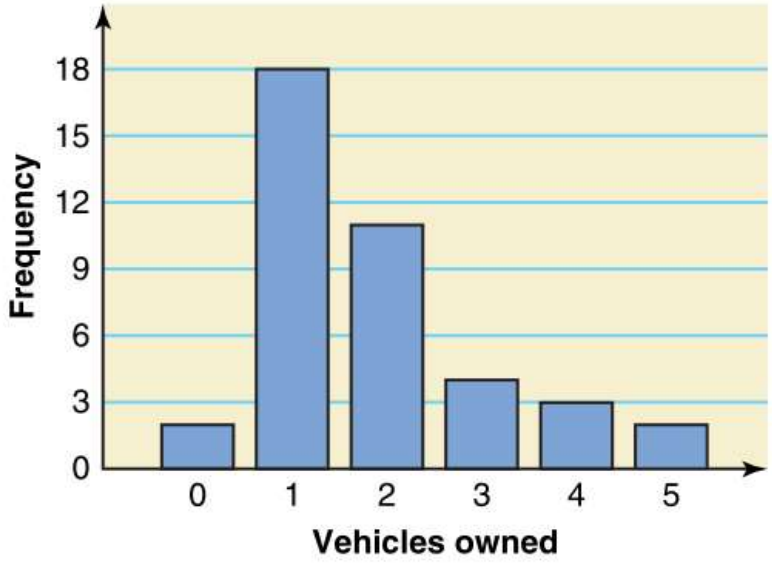


Figure 2.8 Bar graph for Table 2.12.

Organizing quantitative data

- But in the case where data is discrete with many different values, or if the data is continuous, they must first be grouped.

- We group the observations into classes.

Example

Table 2.6 Weekly Earnings of 100 Employees of a Company

Weekly Earnings (dollars)	Number of Employees f
801 to 1000	4
1001 to 1200	11
1201 to 1400	39
1401 to 1600	24
1601 to 1800	16
1801 to 2000	6

Variable → Weekly Earnings (dollars)

Number of Employees f ← Frequency column

Third class → 1201 to 1400

39 ← Frequency of the third class

Lower limit of the sixth class → 1801

2000 ← Upper limit of the sixth class

- First column – classes
- Second column – frequency of each classes

Frequency distribution table (grouped data)

- Class – categories in which data are grouped.
- Class frequency – how often data point falls inside class intervals.
- Lower limit/upper limit of a class – the lowest/highest value of data possible in the class
- Class width or class size

$$\text{class width} = \text{lower limit of next class} - \text{lower limit of current class}$$

- Class midpoint

$$\text{class midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

Frequency distribution table (grouped data)

- Lower boundary of a class – the boundary between current class and the previous class.

$$\text{lower boundary} = (\text{upper limit of previous class} + \text{lower limit of current class}) \div 2$$

- Upper boundary of a class – the boundary between current class and the next class.

$$\text{upper boundary} = (\text{upper limit of current class} + \text{lower limit of next class}) \div 2$$

Table 2.6 Weekly Earnings of 100 Employees of a Company

Weekly Earnings (dollars)	Number of Employees f
801 to 1000	4
1001 to 1200	11
1201 to 1400	39
1401 to 1600	24
1601 to 1800	16
1801 to 2000	6

Variable → Weekly Earnings (dollars)

Number of Employees f ← Frequency column

Third class → 1201 to 1400

39 ← Frequency of the third class

Lower limit of the sixth class → 1801

2000 ← Upper limit of the sixth class

- For the third class (1201 to 1400)
  - Lower limit = 1201
  - Upper limit = 1400
  - Class width = 1401 - 1201 = 200
  - Class midpoint = (1201+1400)/2 = 1300.5
  - Lower boundary = (1200+1201)/2 = 1200.5
  - Upper boundary = (1400+1401)/2 = 1400.5

Frequency distribution table (grouped data)

Universiti Kebangsaan Malaysia
The National University of Malaysia

Constructing frequency distribution table for grouped data:

1. Decide on the number of classes to be used. Normally it is around five to ten.
2. Calculate the approximated class width:

$$\text{approximate class width} = \frac{\text{largest value} - \text{smallest value}}{\text{number of classes}}$$

3. Round this number to a convenient number and use it as the class width.
4. Use any convenient number that is equal to or less than the smallest value in the data set as the lower limit of the first class.
5. Write down all the classes and find their frequencies. Sometimes the number of classes may not be as decided in step 1.

EXAMPLE 2–3 Values of Baseball Teams, 2015

The following table gives the value (in million dollars) of each of the 30 baseball teams as estimated by *Forbes* magazine (*source: Forbes* Magazine, April 13, 2015). Construct a frequency distribution table.

Values of Baseball Teams, 2015

Team	Value (millions of dollars)	Team	Value (millions of dollars)
Arizona Diamondbacks	840	Milwaukee Brewers	875
Atlanta Braves	1150	Minnesota Twins	895
Baltimore Orioles	1000	New York Mets	1350
Boston Red Sox	2100	New York Yankees	3200
Chicago Cubs	1800	Oakland Athletics	725
Chicago White Sox	975	Philadelphia Phillies	1250
Cincinnati Reds	885	Pittsburgh Pirates	900
Cleveland Indians	825	San Diego Padres	890
Colorado Rockies	855	San Francisco Giants	2000
Detroit Tigers	1125	Seattle Mariners	1100
Houston Astros	800	St. Louis Cardinals	1400
Kansas City Royals	700	Tampa Bay Rays	605
Los Angeles Angels of Anaheim	1300	Texas Rangers	1220
Los Angeles Dodgers	2400	Toronto Blue Jays	870
Miami Marlins	650	Washington Nationals	1280

Example

- From the data,
  - Smallest value $= 605$
  - Largest value $= 3200$
- Suppose we decide to use six classes with equal width.
- Approximate class width:

$$\text{approx class width} = \frac{3200 - 605}{6} = 432.5$$

- We round this approximate width to a convenient number, say 450.
- We take 601 as the lower limit of the first class because it is convenient.

Example

Value of team (in million $)	Frequency	Relative frequency
601 – 1050		
1051 – 1500		
1501 – 1950		
1951 – 2400		
2401 – 2850		
2851 – 3300		
Total	30	1

Example

Value of team (in million $)	Frequency	Relative frequency
601 – 1050	16	0.533
1051 – 1500	9	0.300
1501 – 1950	1	0.033
1951 – 2400	3	0.100
2401 – 2850	0	0.000
2851 – 3300	1	0.033
Total	30	1

Exercise

EXAMPLE 2–3

The following data give the total number of iPods® sold by a mail order company on each of 30 days. Construct a frequency distribution table.

8	25	11	15	29	22	10	5	17	21
22	13	26	16	18	12	9	26	20	16
23	14	19	23	20	16	27	16	21	14

- Use five classes with equal width.

5 classes

smallest = 5

largest = 29

approx. width $= \frac{29 - 5}{5}$

$= 4.8$

width = 5

	Number sold	Freq.	Rel-freq.
|||	5 – 9	3	$3/30 = 1/10$
|||||	10 – 14	6	$1/5$
||||| |||	15 – 19	8	$4/15$
||||| |||	20 – 24	8	$4/15$
|||||	25 – 29	5	$1/6$
	Total	30	1

Histograms

Histogram

- Useful to display the distribution of quantitative data

- Steps
  1. Prepare the frequency table.
  2. Mark classes based on the class boundaries on the x-axis and frequencies on the y-axis.
  3. Draw a rectangle for each class so that its height represents the frequency of that class.
  4. Write down the correct label and title.

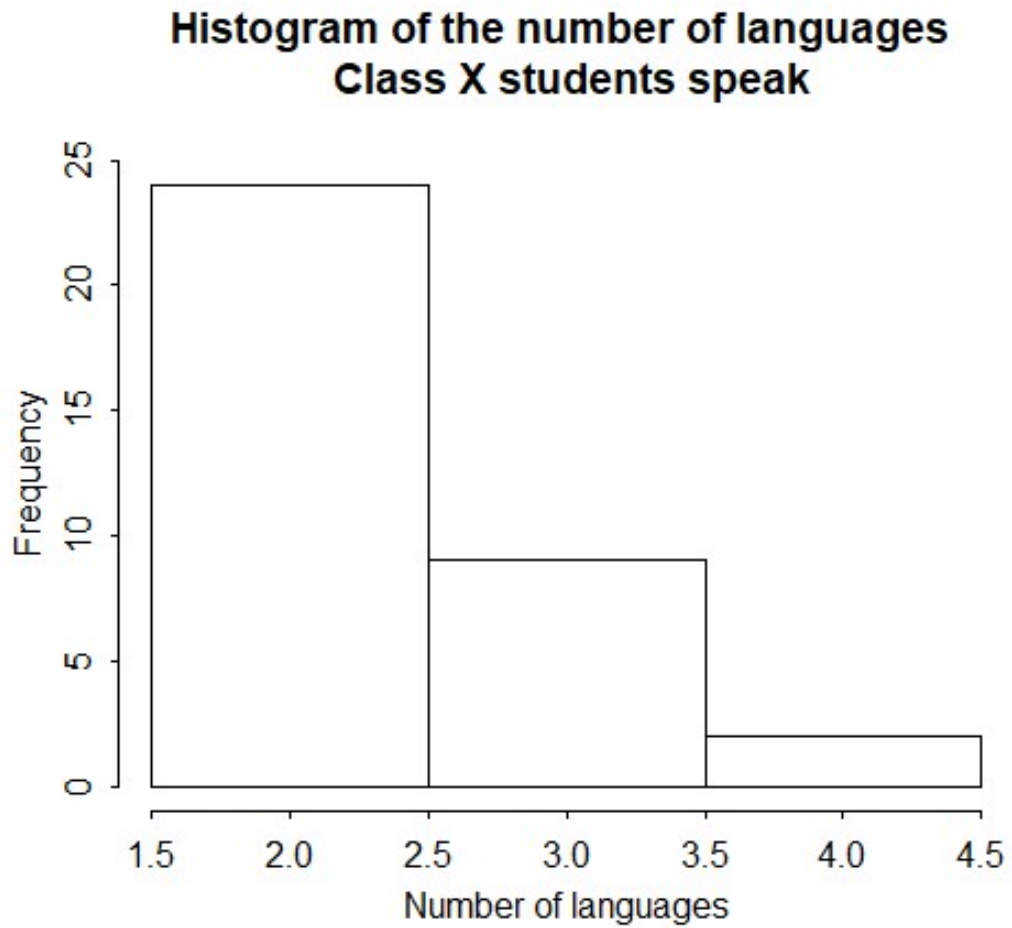
- Sometimes relative frequency or percentage is used for the y-axis.

Example (discrete data)

- Number of languages Class X students speak:

Language	Class boundary	Frequency
2	1.5 – 2.5	24
3	2.5 – 3.5	9
4	3.5 – 4.5	2

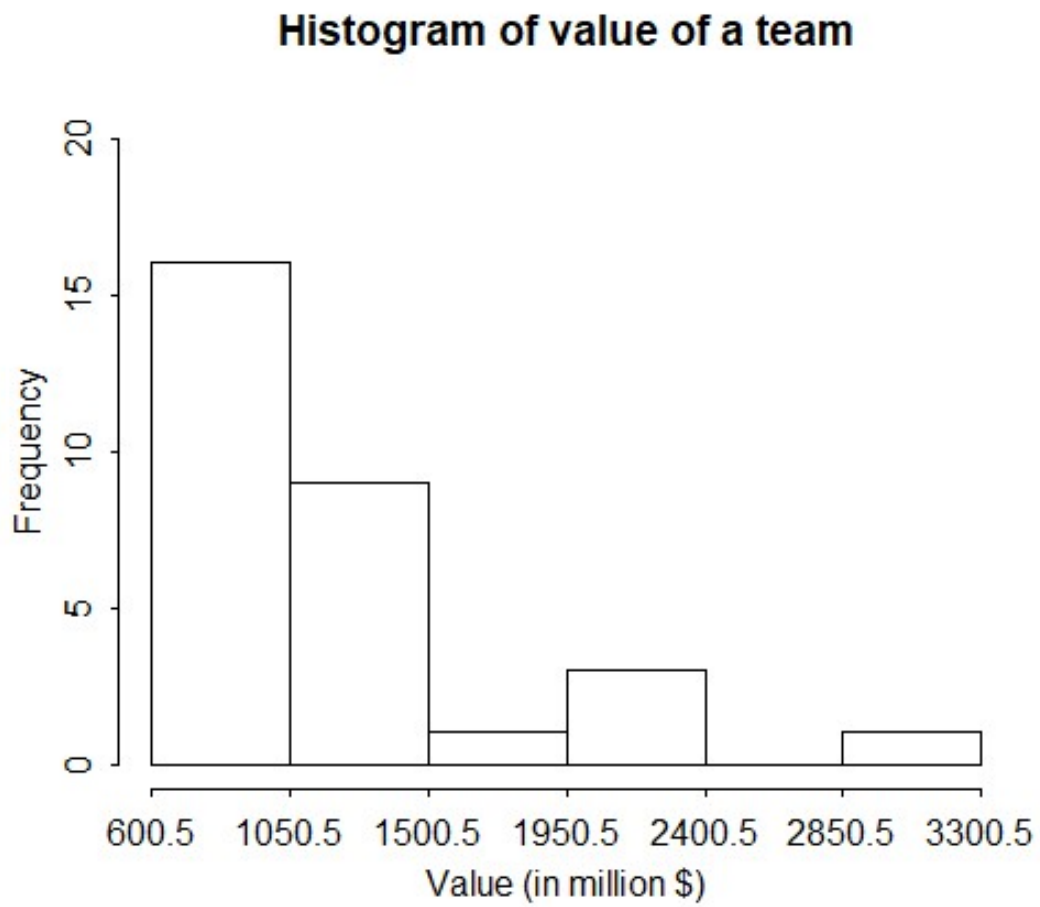
Example (discrete data)



Example (continuous data)

Value of team (in million $)	Frequency	Class boundary
601 – 1050	16	600.5 – 1050.5
1051 – 1500	9	1050.5 – 1500.5
1501 – 1950	1	1500.5 – 1950.5
1951 – 2400	3	1950.5 – 2400.5
2401 – 2850	0	2400.5 – 2850.5
2851 – 3300	1	2850.5 – 3300.5

Example (continuous data)

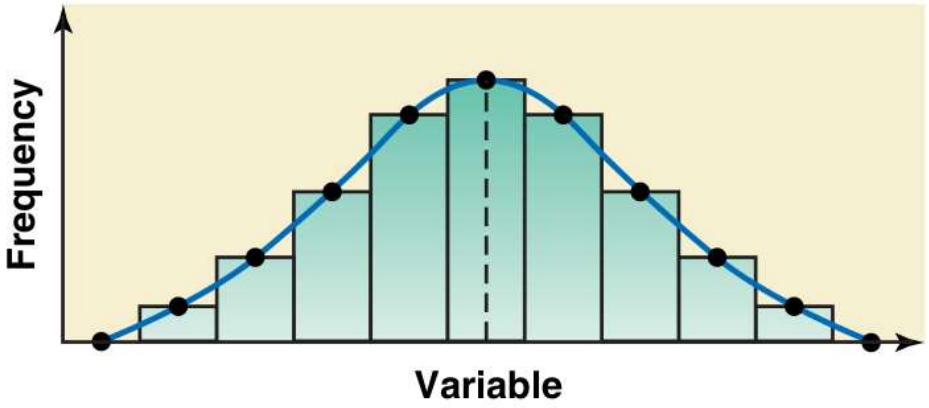


Shapes of histograms

The shape of histograms gives an idea on the distribution of data.

Some common shapes:

- Symmetric – identical on both sides of its central point



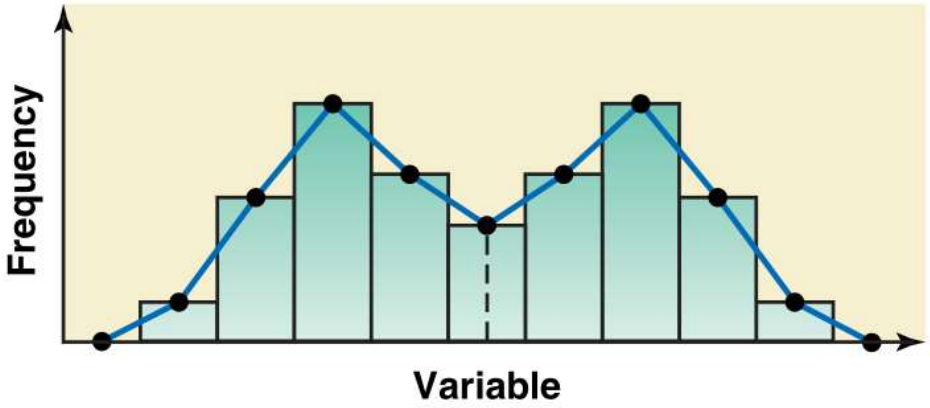


Figure 2.9 Symmetric histograms.

Shapes of histograms

- Skewed – nonsymmetric
  - Skewed-to-the-right – longer tail on the right side
  - Skewed-to-the-left – longer tail on the left side

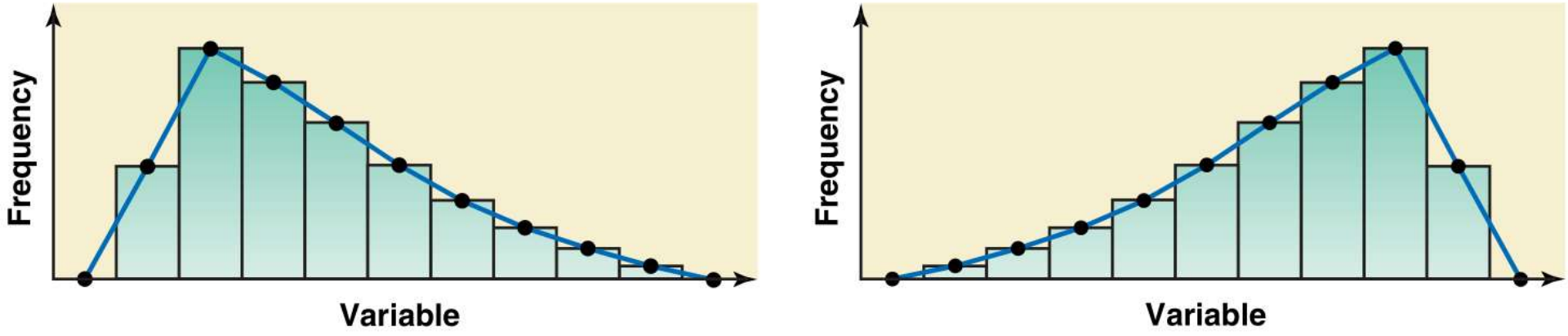
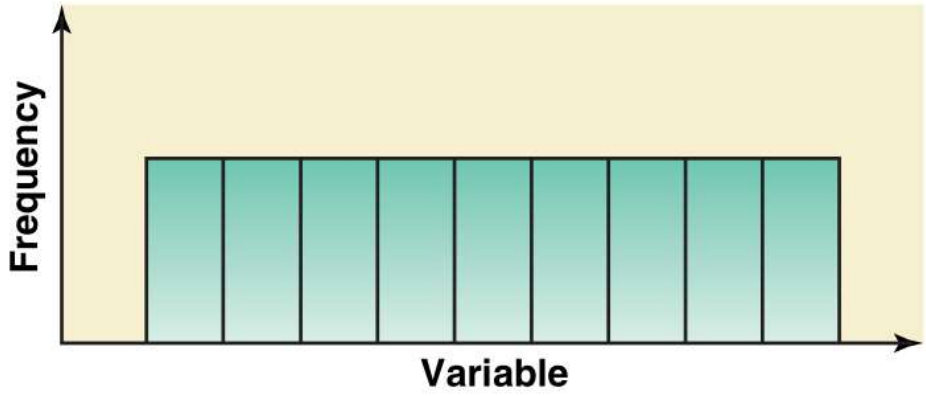


Figure 2.10 (*a*) A histogram skewed to the right. (*b*) A histogram skewed to the left.

Shapes of histograms

Uniform or rectangular – same frequency for each class.

Figure 2.11 A histogram with uniform distribution.



Exercise

2.20 The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45	52	48	41	56	46	44	42	48	53	51	53	51
48	46	43	52	50	54	47	44	47	50	49	52	

a. Make the frequency distribution table for these data.
b. Calculate the relative frequencies for all classes.
c. Construct a histogram for the relative frequency distribution.

6 classes

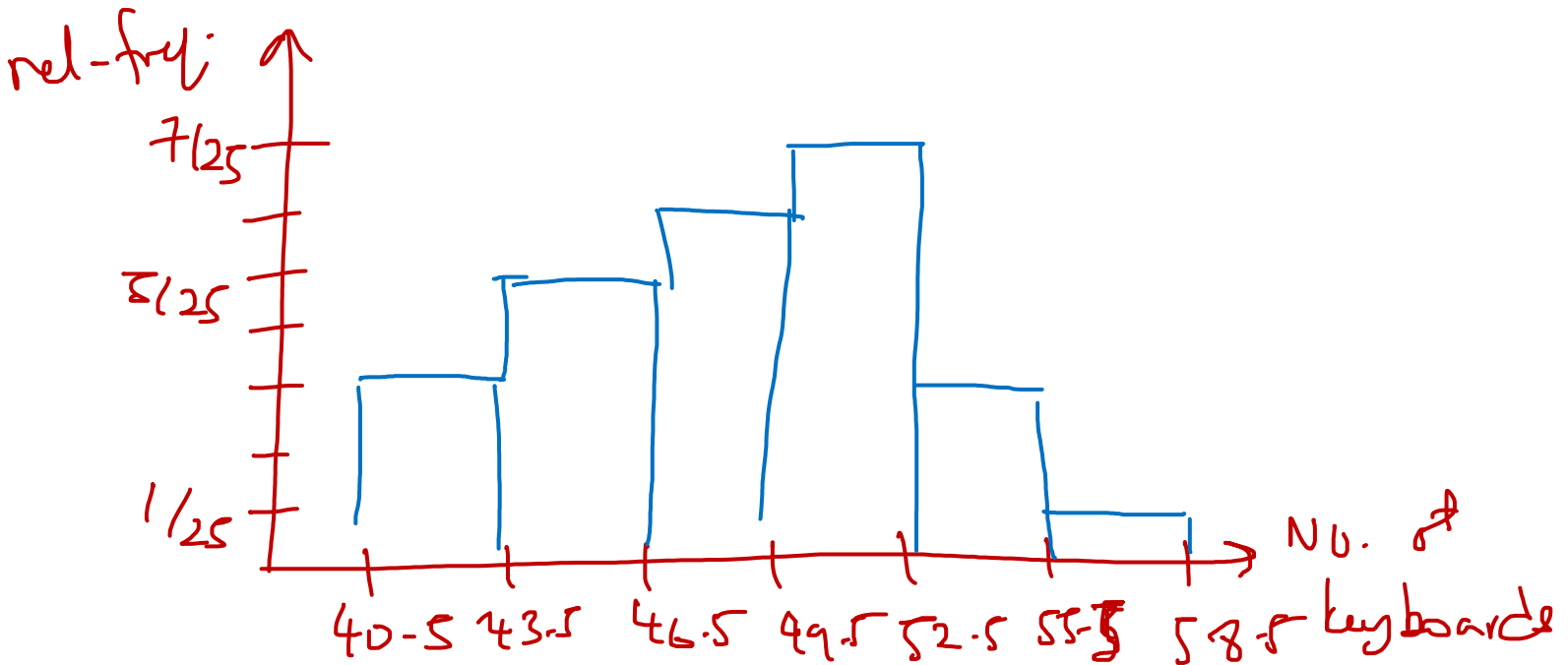
approx. width $= \frac{56 - 41}{6} = 2.5$

smallest = 41

largest = 56

width = 3

	No. of keyboards	Freq.	Rel. freq.
III	41 – 43	3	3/25
IIII	44 – 46	5	5/25 = 1/5
IIII I	47 – 49	6	6/25
IIII II	50 – 52	7	7/25
III	53 – 55	3	3/25
I	56 – 58	1	1/25
	Total	25	1



Other graphs and plots

Frequency polygon

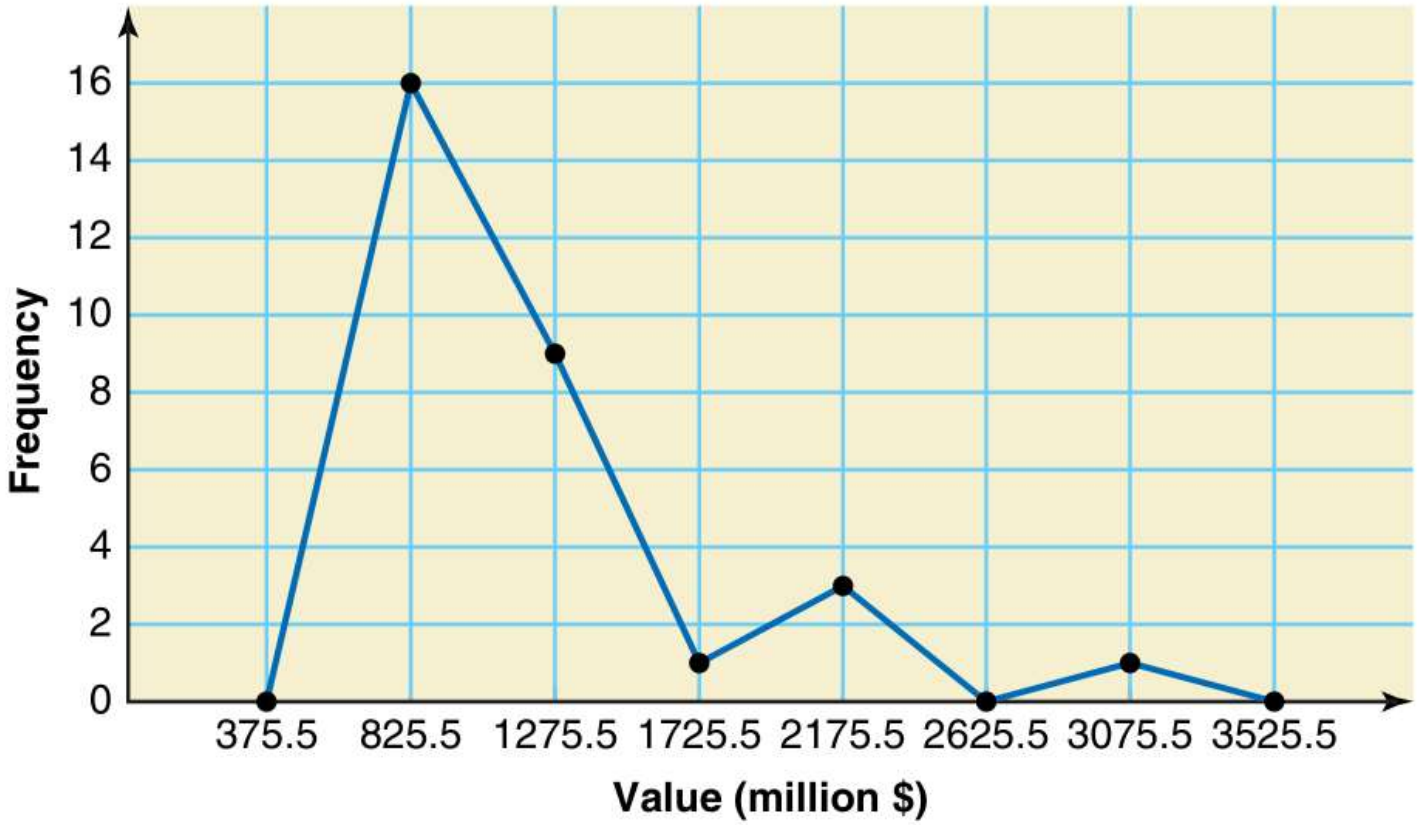
- Similar to the histogram, but uses class midpoints and lines to connect the frequency for the classes

- Steps:
  1. Prepare the frequency table, including class midpoints.
  2. Mark classes based on the class midpoints on the x-axis and frequencies on the y-axis.
  3. Mark a dot above the midpoint of each class at height equal to the frequency of that class.
  4. Add two more classes, one at each end, and mark their midpoints. These classes have zero frequencies
  5. Draw lines connecting the adjacent dots.
  6. Write down the correct label and title.

Frequency polygon

Value of team (in million $)	Frequency	Midpoint
601 – 1050	16	825.5
1051 – 1500	9	1275.5
1501 – 1950	1	1725.5
1951 – 2400	3	2175.5
2401 – 2850	0	2625.5
2851 – 3300	1	3075.5
Total	30	

Frequency polygon



Stem-and-leaf plot

- Another technique used to display data.
- Each observation/value is divided into two portions – a stem and a leaf.
- The first digit will be the stem, while the rest will be the leaf.

- Steps:
  - Arrange the data in increasing order.
  - Separate the data according to the first digit.
  - Construct the stem-and-leaf plot.

Example

EXAMPLE 2–8 Scores of Students on a Statistics Test

The following are the scores of 30 college students on a statistics test.

75	52	80	96	65	79	71	87	93	95
69	72	81	61	76	86	79	68	50	92
83	84	77	64	71	87	72	92	57	98

Construct a stem-and-leaf display.

Example

- Unordered stem-and-leaf:

5	2 0 7
6	5 9 1 8 4
7	5 9 1 2 6 9 7 1 2
8	0 7 1 6 3 4 7
9	6 3 5 2 2 8

Key: 5 | 2 = 52

Example

- Ordered stem-and-leaf:

5	0 2 7
6	1 4 5 8 9
7	1 1 2 2 5 6 7 9 9
8	0 1 3 4 6 7 7
9	2 2 3 5 6 8

Key: 5 | 2 = 52

Summary

- Frequency distribution table:
  - Qualitative data
  - Quantitative data
    - Grouped frequency distribution
- Graphs and plots:
  - Bar plot
  - Pie chart
  - Histogram
  - Frequency polygon
  - Stem-and-leaf plot