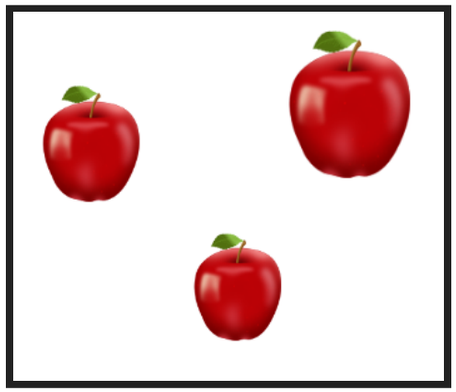
DESCRIBING DATA SETS

GONZALO G. PERAZA MUES

CASE 1: A COUPLE OF APPLES

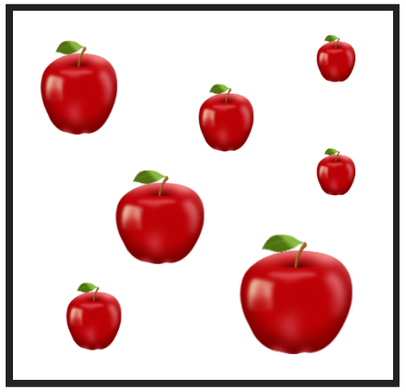


3 APPLES, 4 VARIABLES

num	Weight oz	Carbs	Acidity	Shape
1	4	20.0	medium	round
2	6	24.3	high	oval
3	7	25.0	medium	round

No big deal.

CASE 2: MORE APPLES

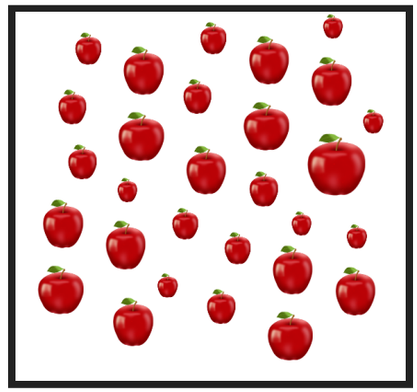


7 APPLES, 4 VARIABLES

num	Weight oz	Carbs	Acidity	Shape
1	4	20.0	medium	round
2	6	24.3	high	oval
3	7	25.0	medium	round
4	7	25.5	low	square
5	6	24.7	medium	round
6	8	26.1	low	round
7	6	25.2	high	square

Still OK.

CASE 3: LOTS OF APPLES



25 APPLES, 4 VARIABLES

num	Weight oz	Carbs	Acidity	Shape
1	4	20	medium	round
2	6	24.3	high	oval
3	7	25	medium	round
4	7	25.5	low	square
5	6	24.7	medium	round
6	8	26.1	low	round
7	6	25.2	high	square
8	9	28.4	high	oval
9	10	30	medium	square
10	8	27.5	medium	round
11	4	20	medium	round
12	6	24.3	high	oval
13	7	25	medium	round
14	8	26.1	low	round
15	6	25.2	high	square
16	4	20	medium	round
17	6	24.2	high	oval
18	7	25	medium	round
19	8	24	medium	square
20	9	28.5	high	oval
21	4	20	medium	round
22	6	24.3	high	oval
23	7	25	medium	round
24	9	29	high	round
25	8	26.3	low	oval

Oh oh.

TOO MUCH DATA.

MAIN IDEA

Make a large or complicated dataset more compact and easier to understand by organizing it in a table, chart, or few key values.

FREQUENCY TABLES

Number of days of sick leave taken by each of 50 workers in a company over 6 weeks

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

Values	Frequency	Value	Frequency
0		5	
1		6	
2		7	
3		8	
4		9	

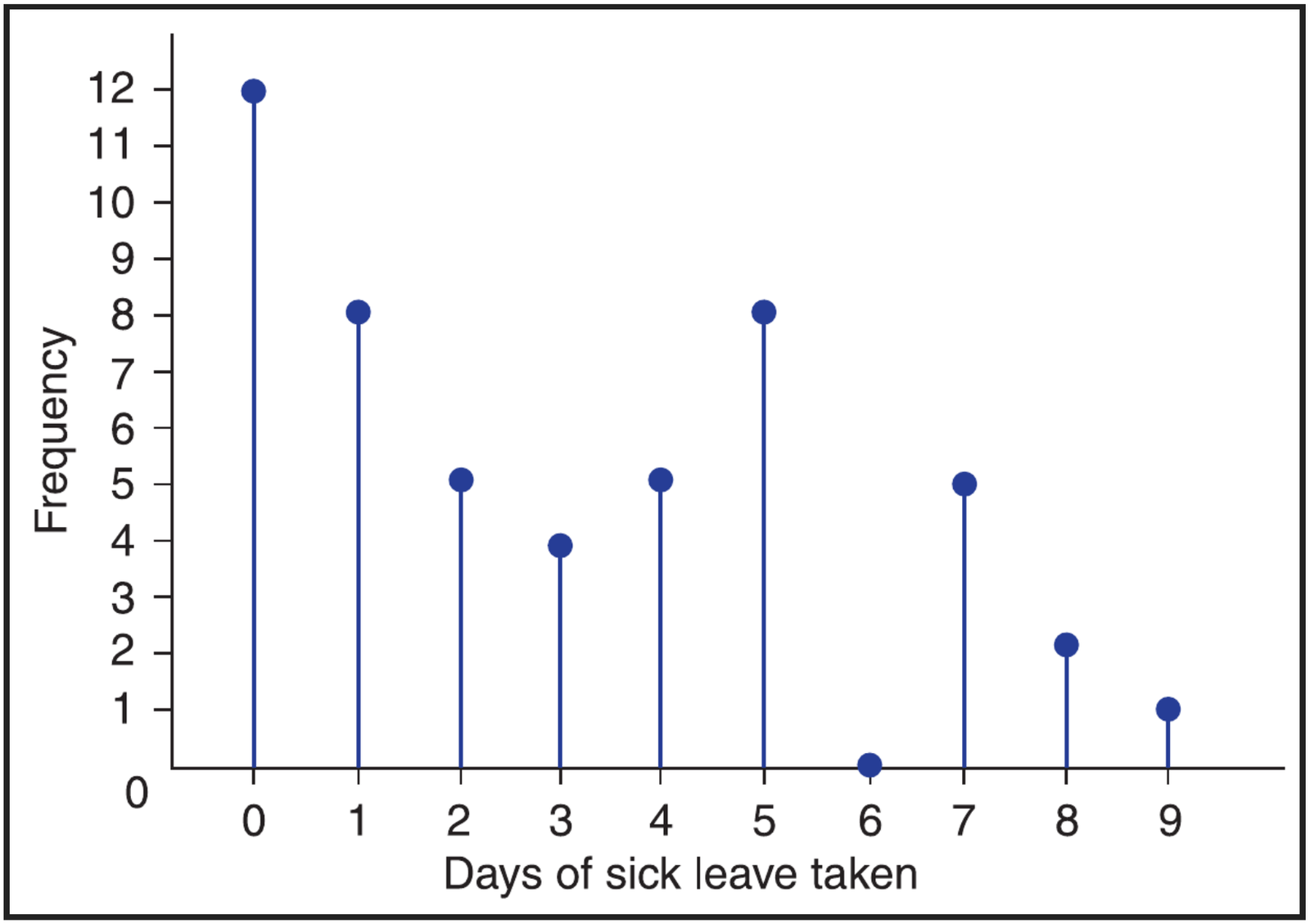
Values	Frequency	Value	Frequency
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

How many workers had at least 1 day of sick leave?

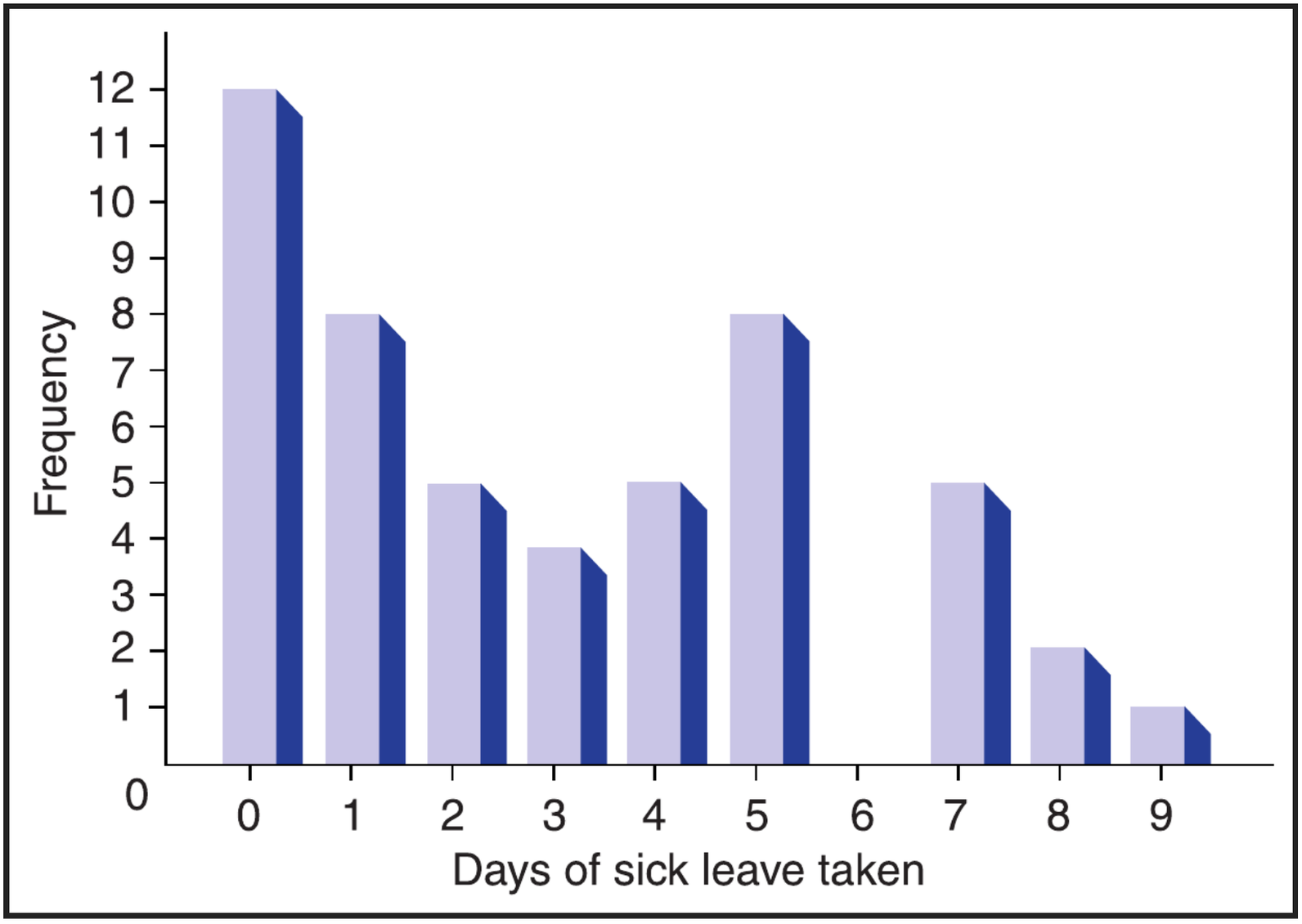
How many workers had between 3 and 5 days of sick leave?

How many workers had more than 5 days of sick leave?

LINE GRAPHS


Frequency
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
Days of sick leave taken

BAR GRAPHS


Frequency
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
Days of sick leave taken

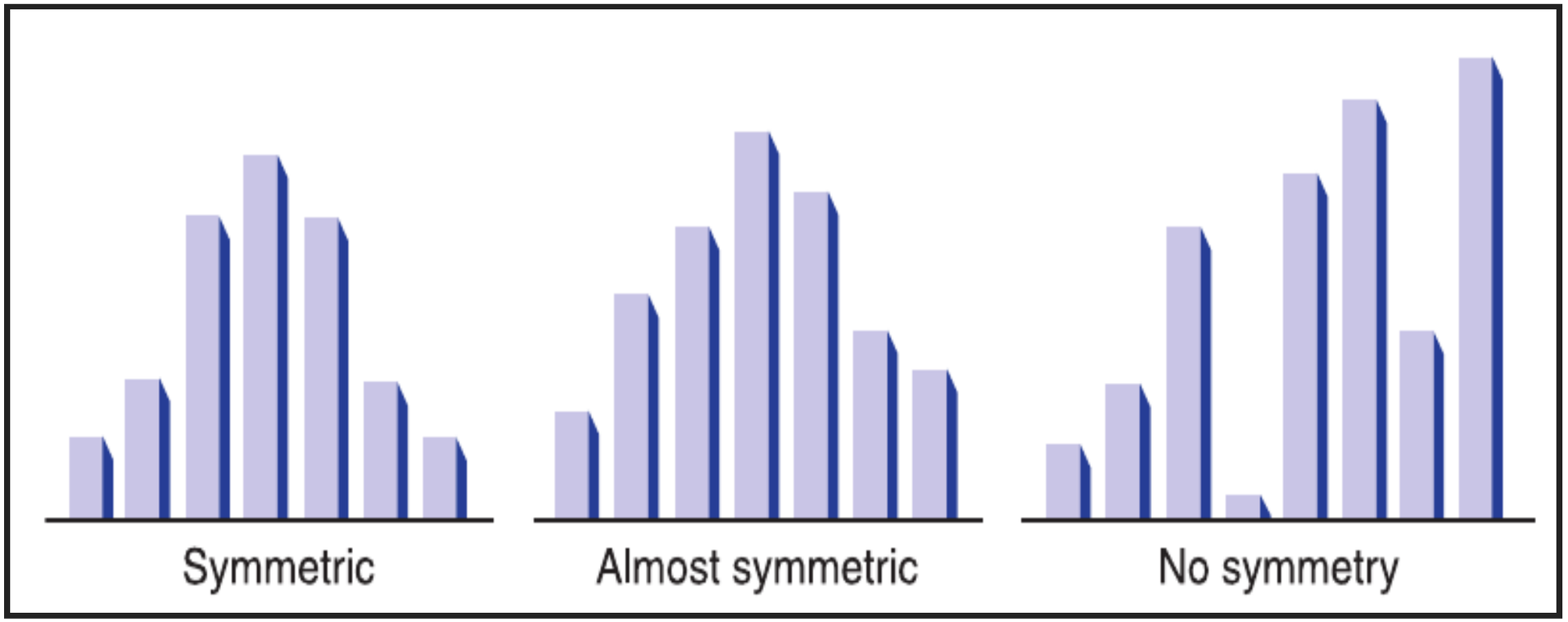
FREQUENCY POLYGON

SYMMETRY OF A DATA SET

Symmetric about x_0 if, for all c, frequencies:

$$f(x_0 - c) = f(x_0 + c)$$

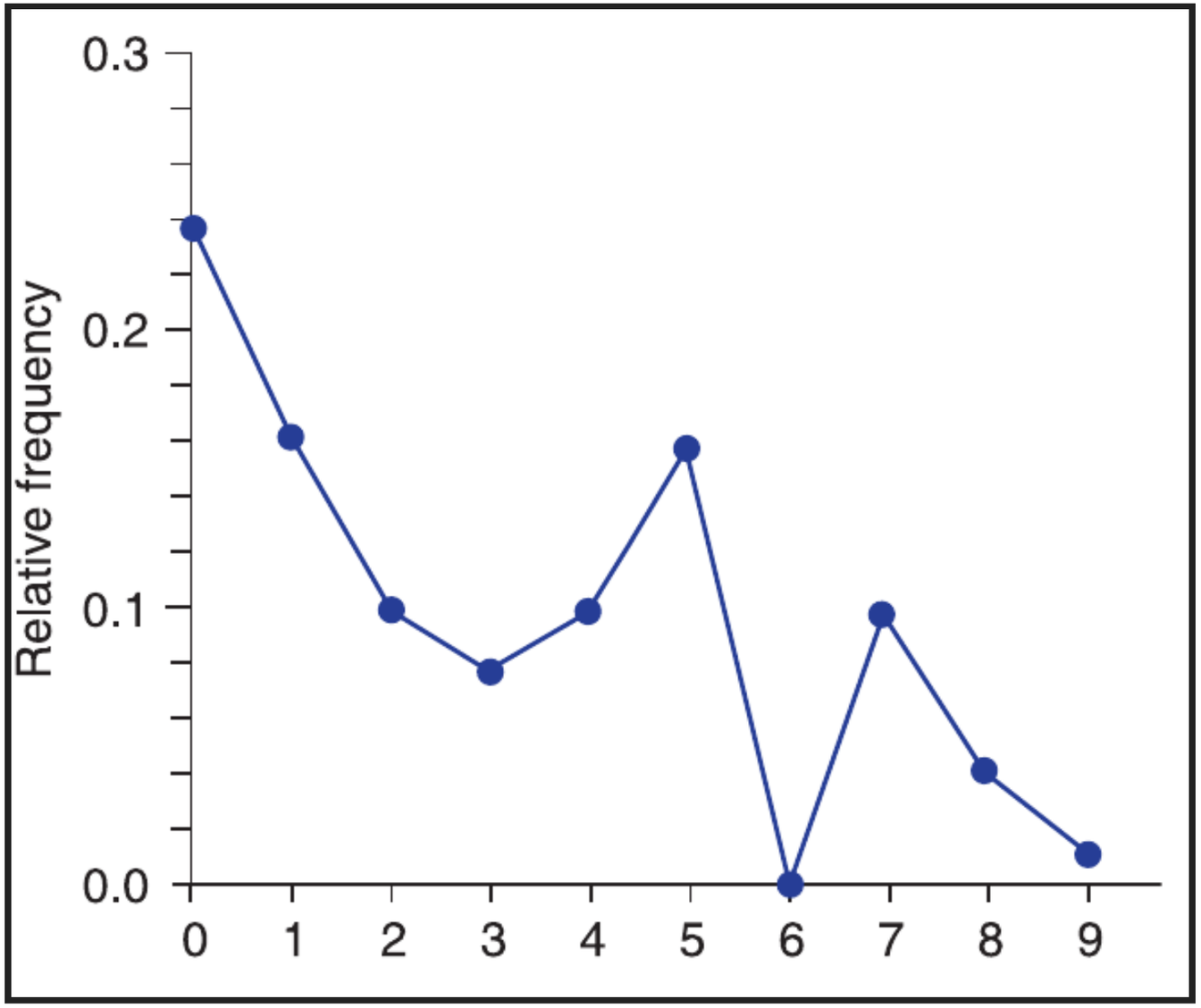
Values	Frequency	Value	Frequency
0	1	4	2
2	2	6	1
3	3	0	0


Symmetric
Almost symmetric
No symmetry

RELATIVE FREQUENCY GRAPHS

- Relative frequency $= \frac{f}{n}$
- How to construct a relative frequency table
  - Arrange the dataset in increasing order of values
  - Determine the distinct values and how often they occur
  - List these distinct values along side their frequency f and their relative frequency f/n

Value	Frequency	Relative frequency
0	12	
1	8	
2	5	
3	4	
4	5	
5	8	
6	0	
7	5	
8	2	
9	1	


0.3
0.2
0.1
0.0
Relative frequency
0
1
2
3
4
5
6
7
8
9

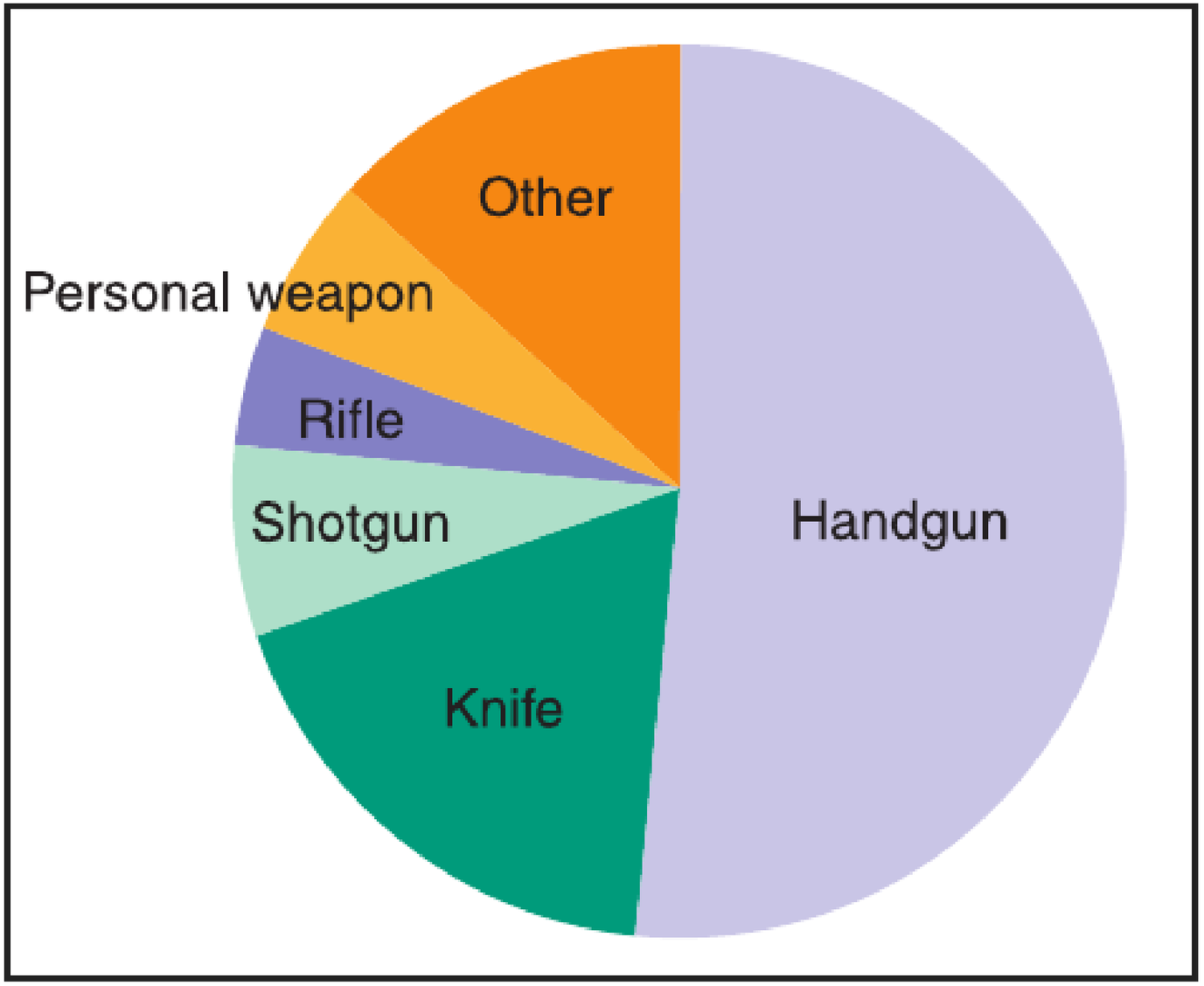
The Masters Golf Tournament Winners

Year	Winner	Score	Year	Winner	Score
1968	Bob Goalby	277	1987	Larry Mize	285
1969	George Archer	281	1988	Sandy Lyle	281
1970	Billy Casper	279	1989	Nick Faldo	283
1971	Charles Coody	279	1990	Nick Faldo	278
1972	Jack Nicklaus	286	1991	Ian Woosnam	277
1973	Tommy Aaron	283	1992	Fred Couples	275
1974	Gary Player	278	1993	Bernhard Langer	277
1975	Jack Nicklaus	276	1994	J.M. Olazabal	279
1976	Ray Floyd	271	1995	Ben Crenshaw	274
1977	Tom Watson	276	1996	Nick Faldo	276
1978	Gary Player	277	1997	Tiger Woods	270
1979	Fuzzy Zoeller	280	1998	Mark O'Meara	279
1980	Severiano Ballesteros	275	1999	J.M. Olazabal	280
1981	Tom Watson	280	2000	Vijay Singh	278
1982	Craig Stadler	284	2001	Tiger Woods	272
1983	Severiano Ballesteros	280	2002	Tiger Woods	276
1984	Ben Crenshaw	277	2003	Mike Weir	281
1985	Bernhard Langer	282	2004	Phil Mickelson	279
1986	Jack Nicklaus	279			

PIE CHARTS

- Plot relative frequencies when the data are not numeric
- Area of sector = relative frequency * area of circle
- Angle of sector = 360^o * relative frequency

Type of weapon	Percentage of murders
Handgun	52
Knife	18
Shotgun	7
Rifle	4
Personal weapon	6
Other	13


Other
Personal weapon
Rifle
Shotgun
Handgun
Knife

GROUPED DATA

- How to graph a large number of distinct values?

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

WE GROUP THE DATA.

CLASS INTERVALS

Trade off between:

- Choosing too few classes at a cost of loosing information
- Choose too many classes -> small frequencies -> no longer match the population

Trial and error.

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

CLASS BOUNDARIES

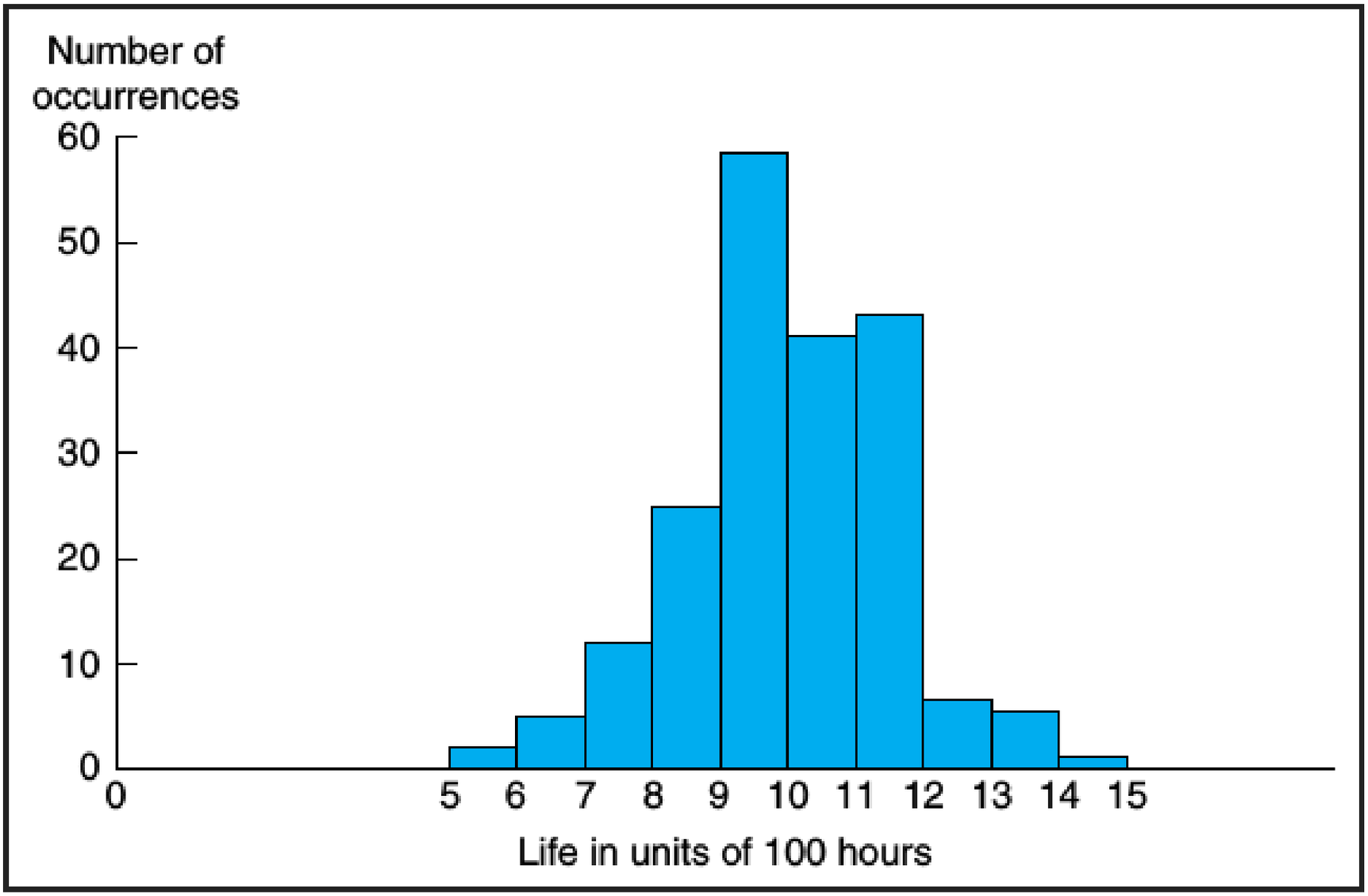
- Equal length

$$\Delta x = x_{i+1} - x_i$$

- Left end inclusion convention

$$\left[x_i, x_{i+1}\right)$$

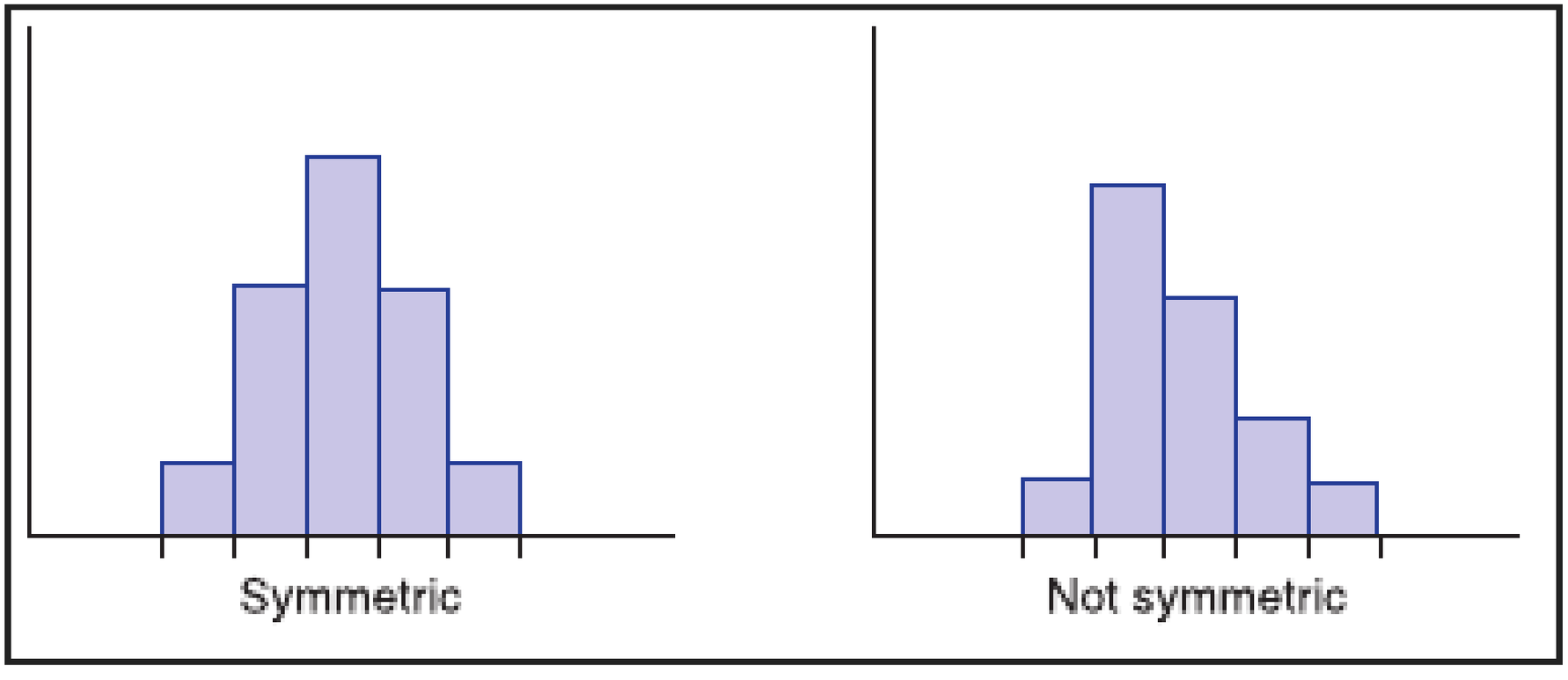
HISTOGRAMS

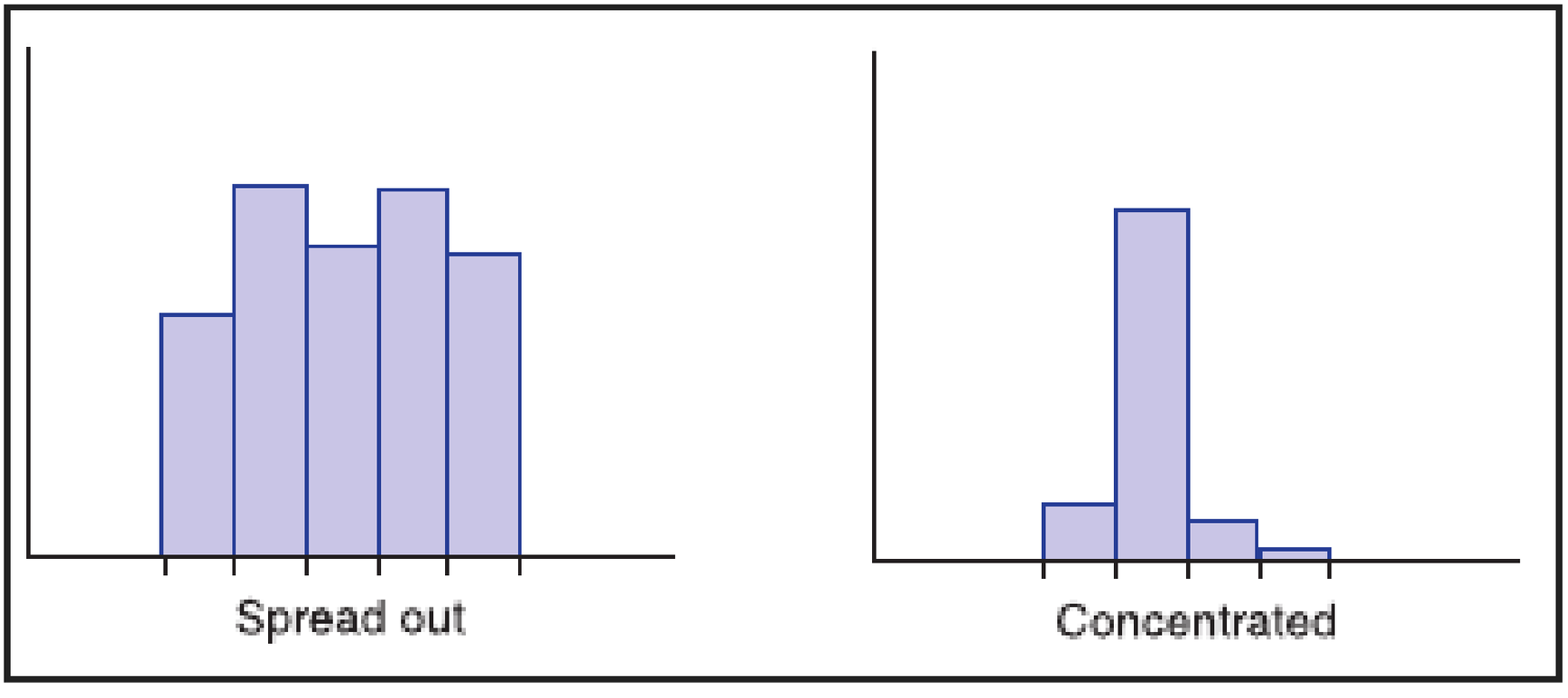

Number of occurrences
60
50
40
30
20
10
0
0
5
6
7
8
9
10
11
12
13
14
15
Life in units of 100 hours

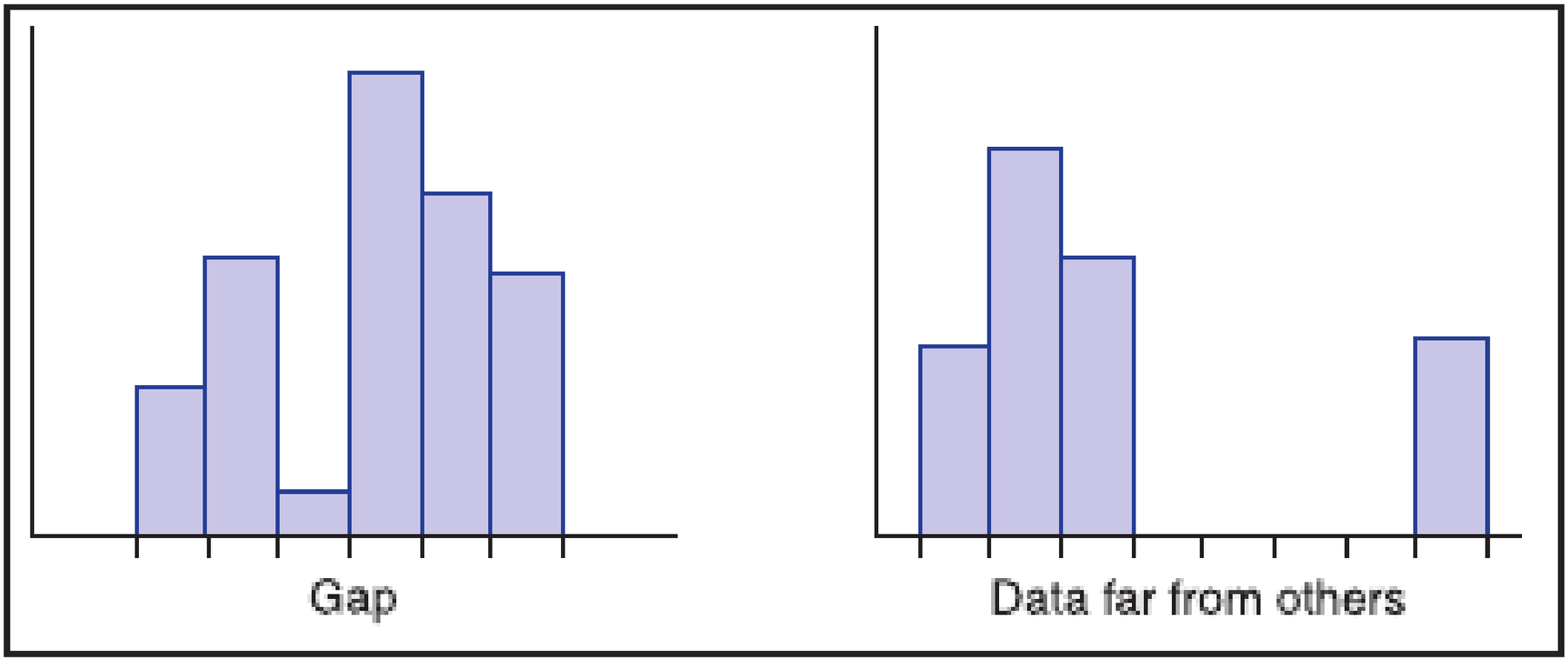
HOW DO WE MAKE THE PREVIOUS HISTOGRAM A RELATIVE FREQUENCY HISTOGRAM?

GRAPHICAL INFORMATION FROM A HISTOGRAM

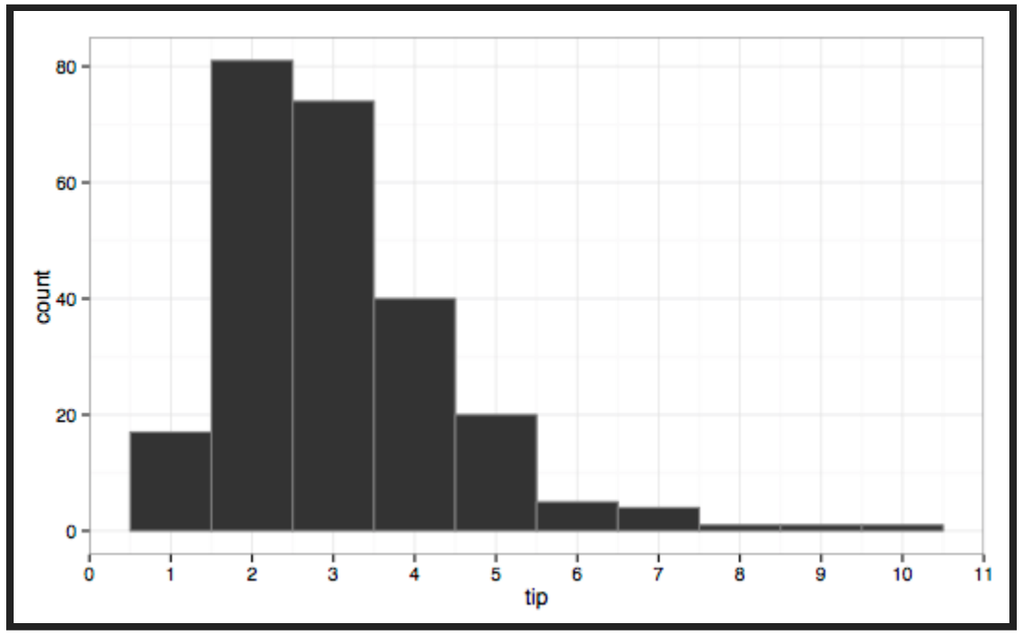
- How symmetric the data are.
- How spread the data are.
- Identifying intervals with high levels of concentration.
- Identifying gaps.
- Whether some data values are far apart from others.

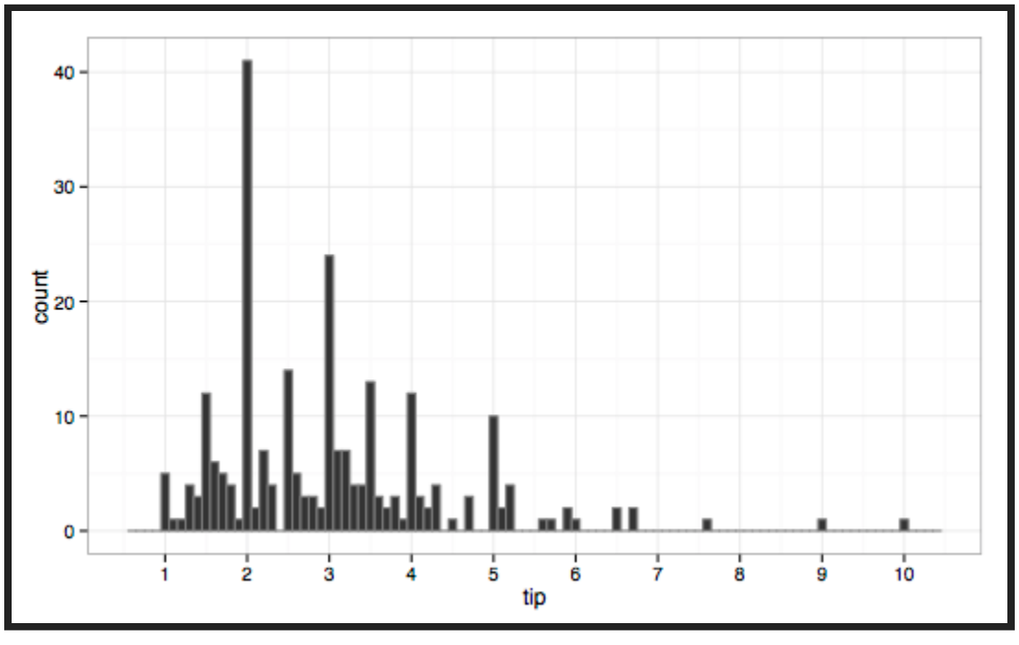

Symmetric
Not symmetric


Spread out
Concentrated


Gap
Data far from others

REMEMBER: INFORMATION IS BEING LOST.


80
60
40
20
0
count
0
1
2
3
4
5
6
7
8
9
10
11
tip


40
30
20
10
0
count
1
2
3
4
5
6
7
8
9
10
tip

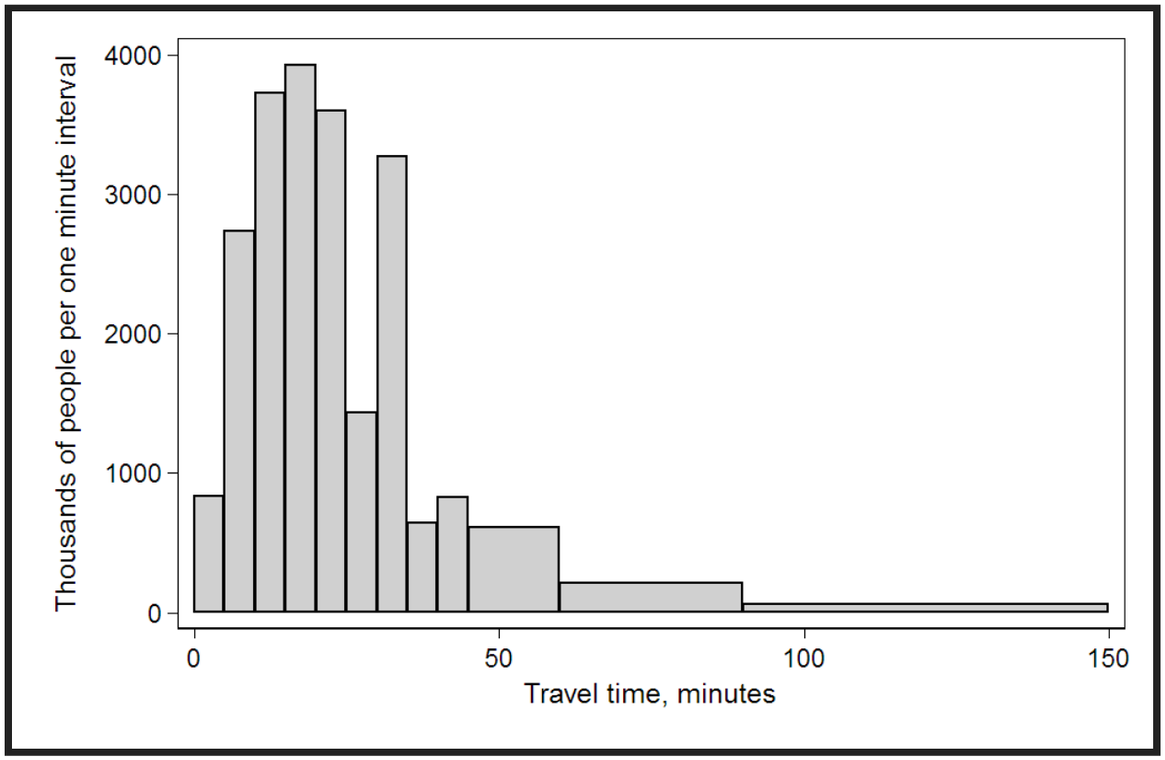
However, bins need not be of equal width; in that case, the erected rectangle is defined to have its area proportional to the frequency of cases in the bin. The vertical axis is then not the frequency but frequency density — the number of cases per unit of the variable on the horizontal axis.

TIME OCCUPIED BY TRAVEL TO WORK

Interval	Width	Quantity	Quantity/width
0	5	4180	836
5	5	13687	2737
10	5	18618	3723
15	5	19634	3926
20	5	17981	3596
25	5	7190	1438
30	5	16369	3273
35	5	3212	642
40	5	4122	824
45	15	9200	613
60	30	6461	215

Interval	Width	Quantity	Quantity/width
90	60	3435	57

This histogram shows the number of cases per unit interval as the height of each block, so that the area of each block is equal to the number of people in the survey who fall into its category. The area under the curve represents the total number of cases (124 million).



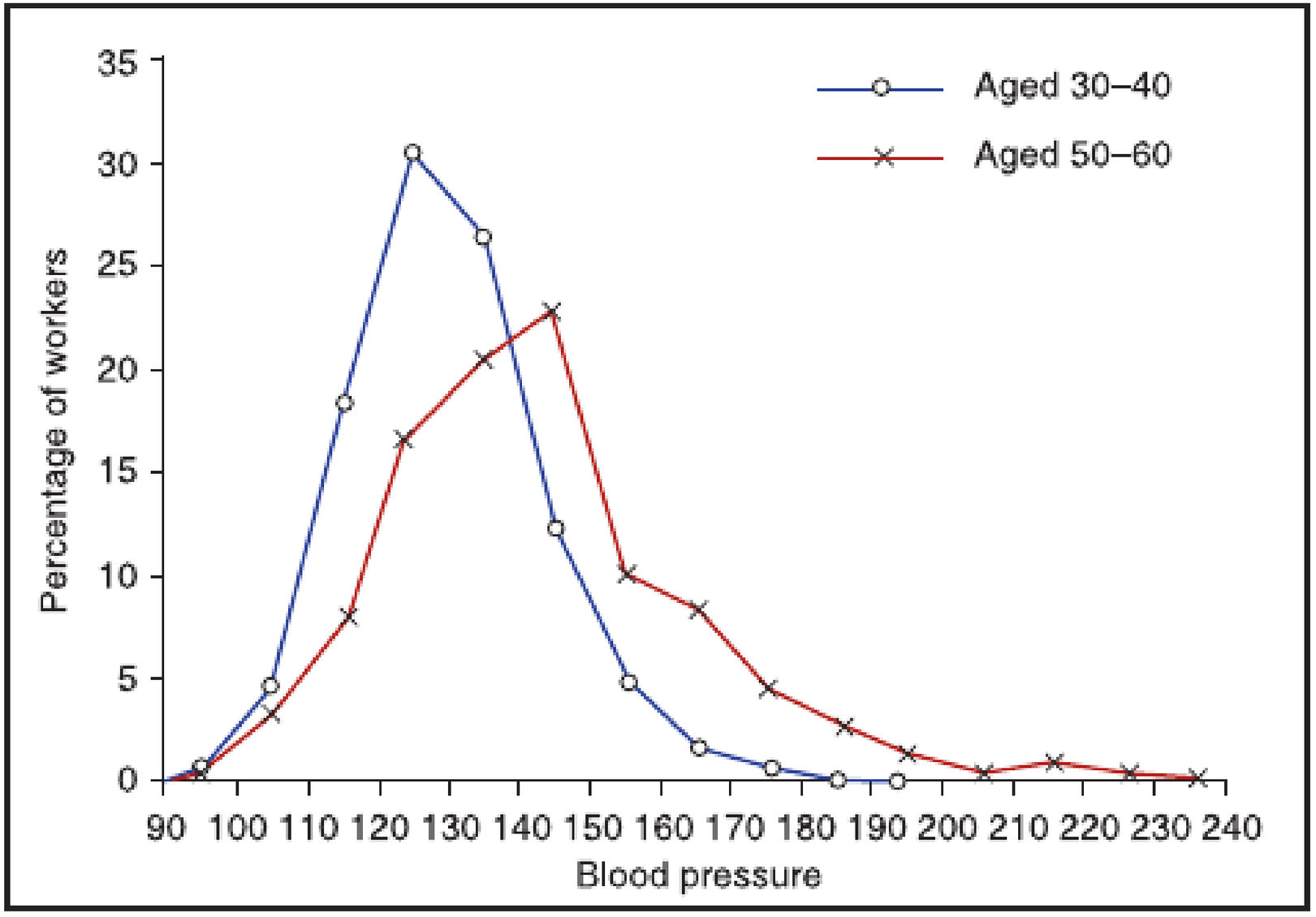
EXAMPLE

Table 2.9 Class Frequencies of Systolic Blood Pressure of Two Groups of Male Workers

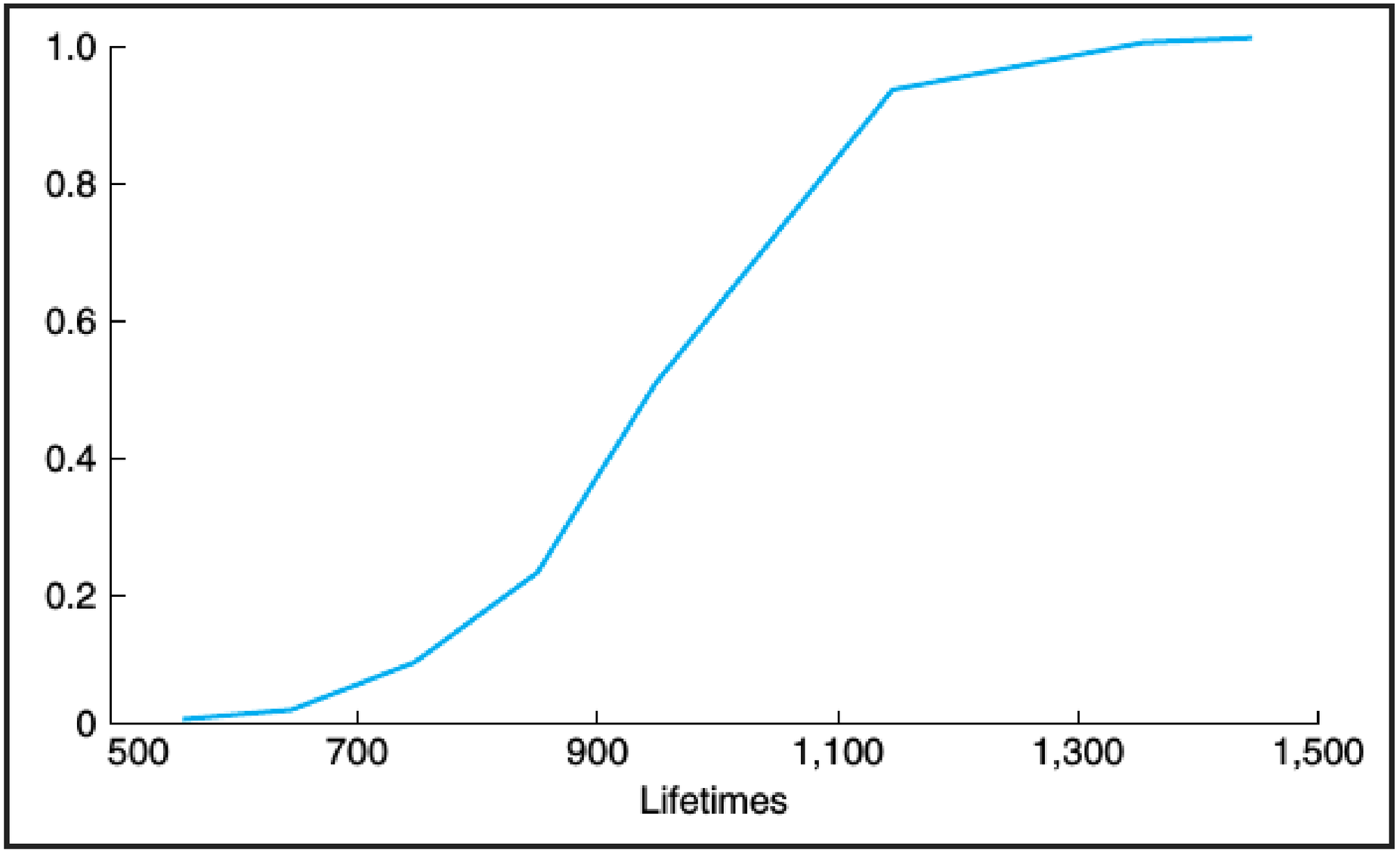
Blood pressure	Number of workers	
	Aged 30–40	Aged 50–60
Less than 90	3	1
90–100	17	2
100–110	118	23
110–120	460	57
120–130	768	122
130–140	675	149
140–150	312	167
150–160	120	73
160–170	45	62
170–180	18	35
180–190	3	20
190–200	1	9
200–210		3
210–220		5
220–230		2
230–240		1
Total	**2540**	**731**

Table 2.10 Relative Class Frequencies of Blood Pressures

Blood pressure	Percentage of workers	
	Aged 30–40	Aged 50–60
Less than 90	0.12	0.14
90–100	0.67	0.27
100–110	4.65	3.15
110–120	18.11	7.80
120–130	30.24	16.69
130–140	26.57	20.38
140–150	12.28	22.84
150–160	4.72	9.99
160–170	1.77	8.48
170–180	0.71	4.79
180–190	0.12	2.74
190–200	0.04	1.23
200–210		0.41
210–220		0.68
220–230		0.27
230–240		0.14
Total	**100.00**	**100.00**


Aged 30–40
Aged 50–60
Percentage of workers
35
30
25
20
15
10
5
0
90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240
Blood pressure

CUMULATIVE FREQUENCY PLOT (OGIVE)



STEM AND LEAF PLOT

84

Stem	Leaf
8	4

87, 84

Stem	Leaf
8	4, 7

The following data represent the proportion of public elementary school students that are classified as minority in each of 18 cities.

55.2, 47.8, 44.6, 64.2, 61.4, 36.6, 28.2, 57.4, 41.3,
44.6, 55.2, 39.6, 40.9, 52.2, 63.3, 34.5, 30.8, 45.3

2	8.2
3	0.8, 4.5, 6.6, 9.6
4	0.9, 1.3, 4.6, 4.6, 5.3, 7.8
5	2.2, 5.2, 5.2, 7.4
6	1.4, 3.3, 4.2