# Module 9: Frequency Distributions

This module includes descriptions of frequency distributions, frequency tables, histograms and frequency polygons. Also included are stem and leaf plots.

# Frequency Distribution

A frequency distribution is the organization of a data set into contiguous, mutually exclusive intervals so that the number or proportion of observations falling in each interval is apparent.

# Frequency Distribution Example

An example of a frequency distribution is the age distribution of one of the classes for an in-class version of this course. The class included n = 121 students, with the age distribution as shown on a following slide.

Two things are notable about this distribution. One has to do with the way we measure age, specified as age at last birthday, not age at nearest birthday.

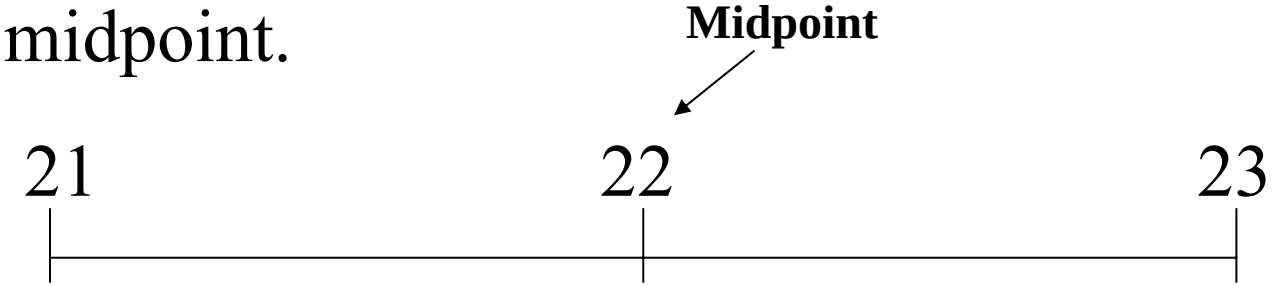
The other is the ages listed as 35+ years, with 21 students having this age. This is an *interval with indefinite length* and it will require special handling.

| Age | No. |
|---|---|
| 21 | 6 |
| 22 | 16 |
| 23 | 11 |
| 24 | 9 |
| 25 | 17 |
| 26 | 13 |
| 27 | 6 |
| 28 | 5 |
| 29 | 4 |
| 30 | 3 |
| 31 | 1 |
| 32 | 4 |
| 33 | 3 |
| 34 | 2 |
| 35+ | 21 |
| Total | 121 |

# Intervals for a Frequency Distribution

- Use not less than 6 intervals, generally 8-15
- Use equal width intervals if feasible and appropriate
- Avoid intervals with indefinite length, if possible
- Select interval width by dividing difference between smallest and largest observation by 10
- Take into account the manner measurements were made when setting up intervals

For the age distribution example, the ages range from 21 years to 35 years and older; to construct at least 6 intervals for this age range, each interval has to be at least 2 years wide. To create the intervals, we need to consider the midpoint of the intervals; The midpoint for the age interval for persons 21 years old is 21.5 years since they will not become 22 until their 22nd birthday. Thus, the age interval for persons 21 and 22 years old goes from 21 to 23, with 22 as the midpoint.




The intervals will be written as 21 – 23, 23 – 25, 25 – 27 etc. We will deal with the 35+ interval in a subsequent slide.

| Age | No. | Interval Midpoint |
|---|---|---|
| 21 | 6 | 22 |
| 22 | 16 | |
| 23 | 11 | 24 |
| 24 | 9 | |
| 25 | 17 | 26 |
| 26 | 13 | |
| 27 | 6 | 28 |
| 28 | 5 | |
| 29 | 4 | 30 |
| 30 | 3 | |
| 31 | 1 | 32 |
| 32 | 4 | |
| 33 | 3 | 34 |
| 34 | 2 | |
| 35+ | 21 | ? |
| Total | 121 | |

| | | Intervals | |
|---|---|---|---|
| Age | No. | Midpoint | No. |
| 21 | 6 | 22 | 22 |
| 22 | 16 | | |
| 23 | 11 | 24 | 20 |
| 24 | 9 | | |
| 25 | 17 | 26 | 30 |
| 26 | 13 | | |
| 27 | 6 | 28 | 11 |
| 28 | 5 | | |
| 29 | 4 | 30 | 7 |
| 30 | 3 | | |
| 31 | 1 | 32 | 5 |
| 32 | 4 | | |
| 33 | 3 | 34 | 5 |
| 34 | 2 | | |
| 35+ | 21 | ? | ? |
| Total | 121 | | |

| | | Intervals | | |
|---|---|---|---|---|
| Age | No. | Midpoint | No. | % |
| 21 | 6 | 22 | 22 | 18 |
| 22 | 16 | | | |
| 23 | 11 | 24 | 20 | 17 |
| 24 | 9 | | | |
| 25 | 17 | 26 | 30 | 25 |
| 26 | 13 | | | |
| 27 | 6 | 28 | 11 | 9 |
| 28 | 5 | | | |
| 29 | 4 | 30 | 7 | 6 |
| 30 | 3 | | | |
| 31 | 1 | 32 | 5 | 4 |
| 32 | 4 | | | |
| 33 | 3 | 34 | 5 | 4 |
| 34 | 2 | | | |
| 35+ | 21 | ? | ? | ? |
| Total | 121 | | | |

| | | Intervals | | | |
|---|---|---|---|---|---|
| Age | No. | Midpoint | No. | % | Cum % |
| 21 | 6 | 22 | 22 | 18 | 18 |
| 22 | 16 | | | | |
| 23 | 11 | 24 | 20 | 17 | 35 |
| 24 | 9 | | | | |
| 25 | 17 | 26 | 30 | 25 | 60 |
| 26 | 13 | | | | |
| 27 | 6 | 28 | 11 | 9 | 69 |
| 28 | 5 | | | | |
| 29 | 4 | 30 | 7 | 6 | 75 |
| 30 | 3 | | | | |
| 31 | 1 | 32 | 5 | 4 | 79 |
| 32 | 4 | | | | |
| 33 | 3 | 34 | 5 | 4 | 83 |
| 34 | 2 | | | | |
| 35+ | 21 | ? | ? | ? | ? |
| Total | 121 | | | | |

# Dealing with the Indefinite Interval

The interval with indefinite length, 35+, creates a problem for finding the midpoint since we don't know where the interval ends. Hence, we must make an arbitrary decision as to the upper age limit for the oldest person in the class. In this case, we made the arbitrary decision that the oldest student was 50 years old. Hence this interval begins at 35 and ends at 51 so that it is 16 years long and the midpoint is 35 + 8 = 43 years.

| | | Intervals | | | |
|---|---|---|---|---|---|
| Age | No. | Midpoint | No. | % | Cum % |
| 21 | 6 | 22 | 22 | 18 | 18 |
| 22 | 16 | | | | |
| 23 | 11 | 24 | 20 | 17 | 35 |
| 24 | 9 | | | | |
| 25 | 17 | 26 | 30 | 25 | 60 |
| 26 | 13 | | | | |
| 27 | 6 | 28 | 11 | 9 | 69 |
| 28 | 5 | | | | |
| 29 | 4 | 30 | 7 | 6 | 75 |
| 30 | 3 | | | | |
| 31 | 1 | 32 | 5 | 4 | 79 |
| 32 | 4 | | | | |
| 33 | 3 | 34 | 5 | 4 | 83 |
| 34 | 2 | | | | |
| 35+ | 21 | 43 | 21 | 17 | 100 |
| Total | 121 | | | | |

# Histogram

- Use: Graph of a frequency distribution
- For: Continuous variables

## Rules

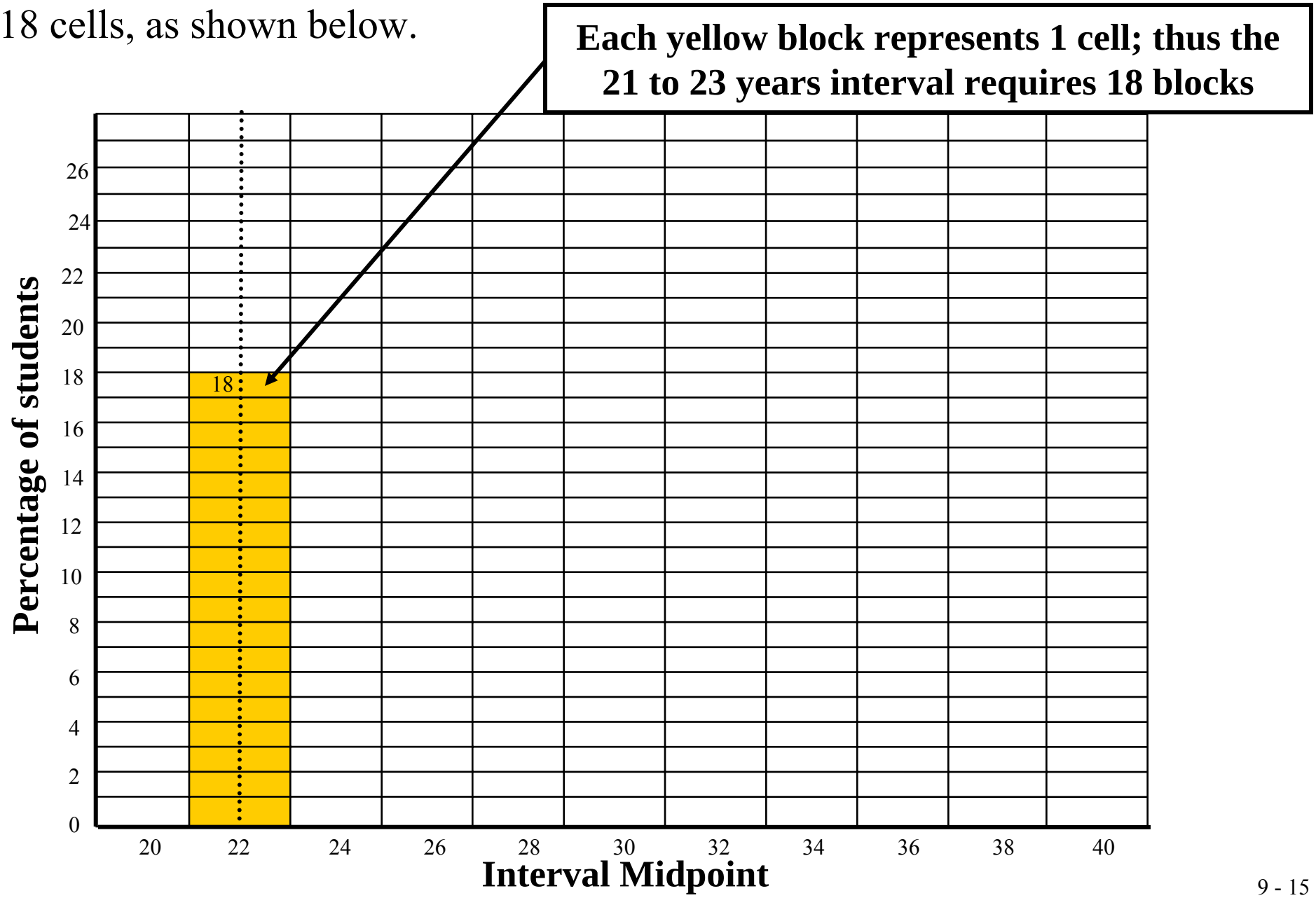
- No space between bars
- Equal areas must represent equal percentages or numbers
- Percentages often preferred to numbers for vertical axis

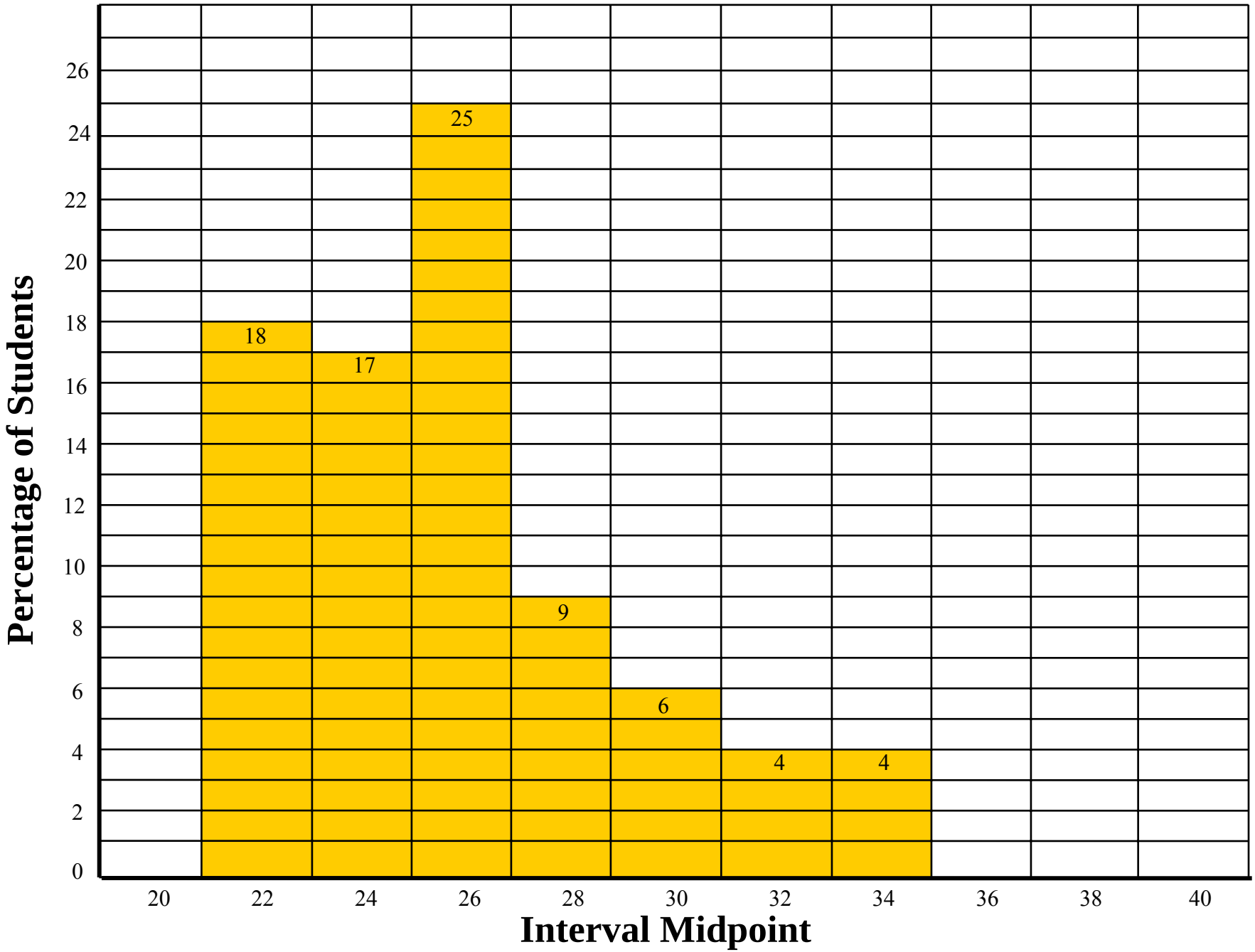
# From a Frequency Table to a Histogram

*Histo* comes from the Greek word for cell. In this case, the reference is to a cell like those in bee honeycombs, all of which have an equal area when viewed from the top. That is, a histogram is an equal cell or equal area graph, with each percentage represented by the same area. To prepare a histogram from a frequency table, we must first decide the shape and area for the cell that represents a single %. We can then stack the appropriate number of these cells on top of each other to represent the percentage of the frequency distribution located within the interval of interest.

For this example, we will make each cell be two years wide and one unit high so that each percentage point for a specific interval will look something like:

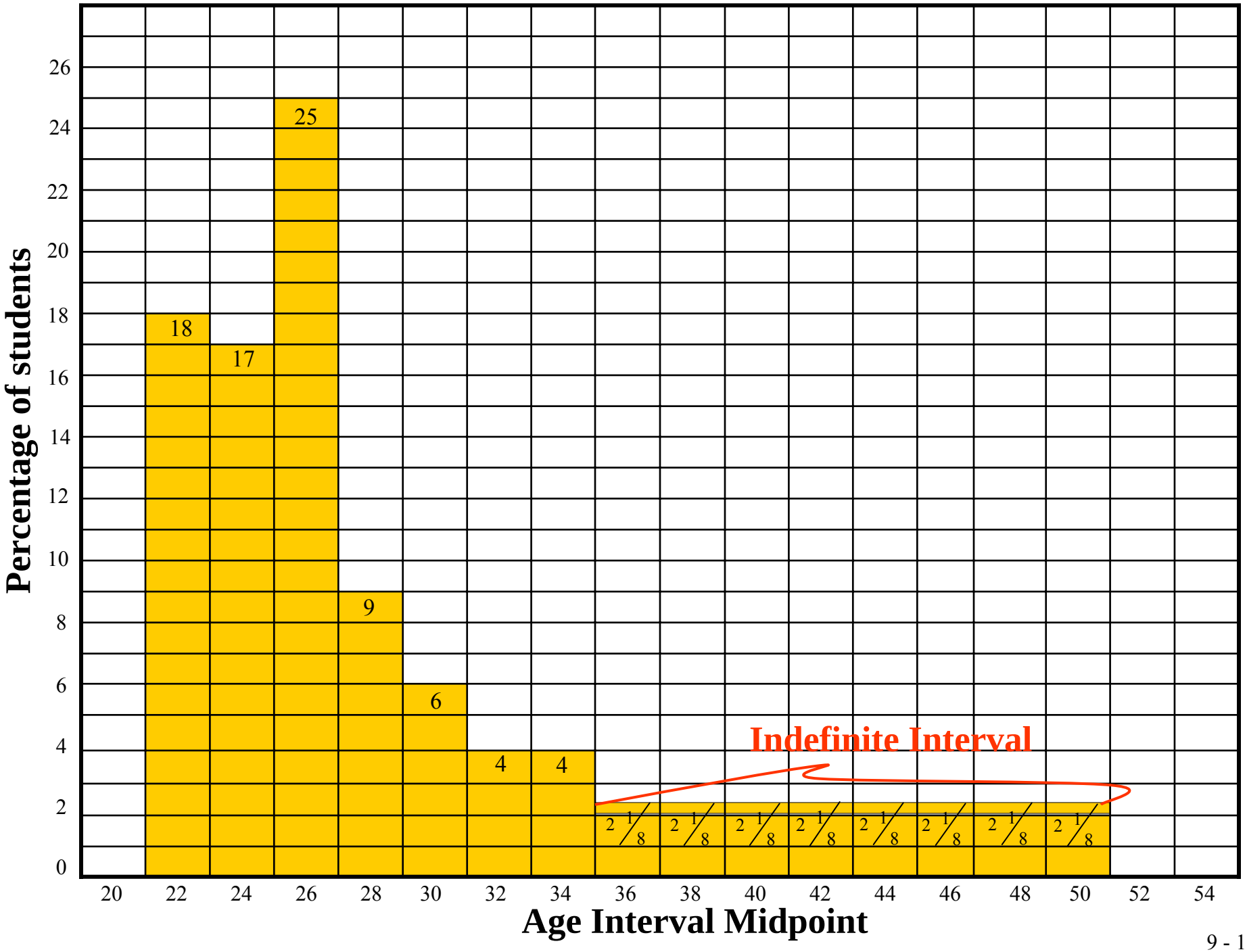
The 21 to 23 years interval of the age frequency distribution includes 18% of the observations, with a midpoint of 22; so for this interval, we need to draw 18 cells, as shown below.

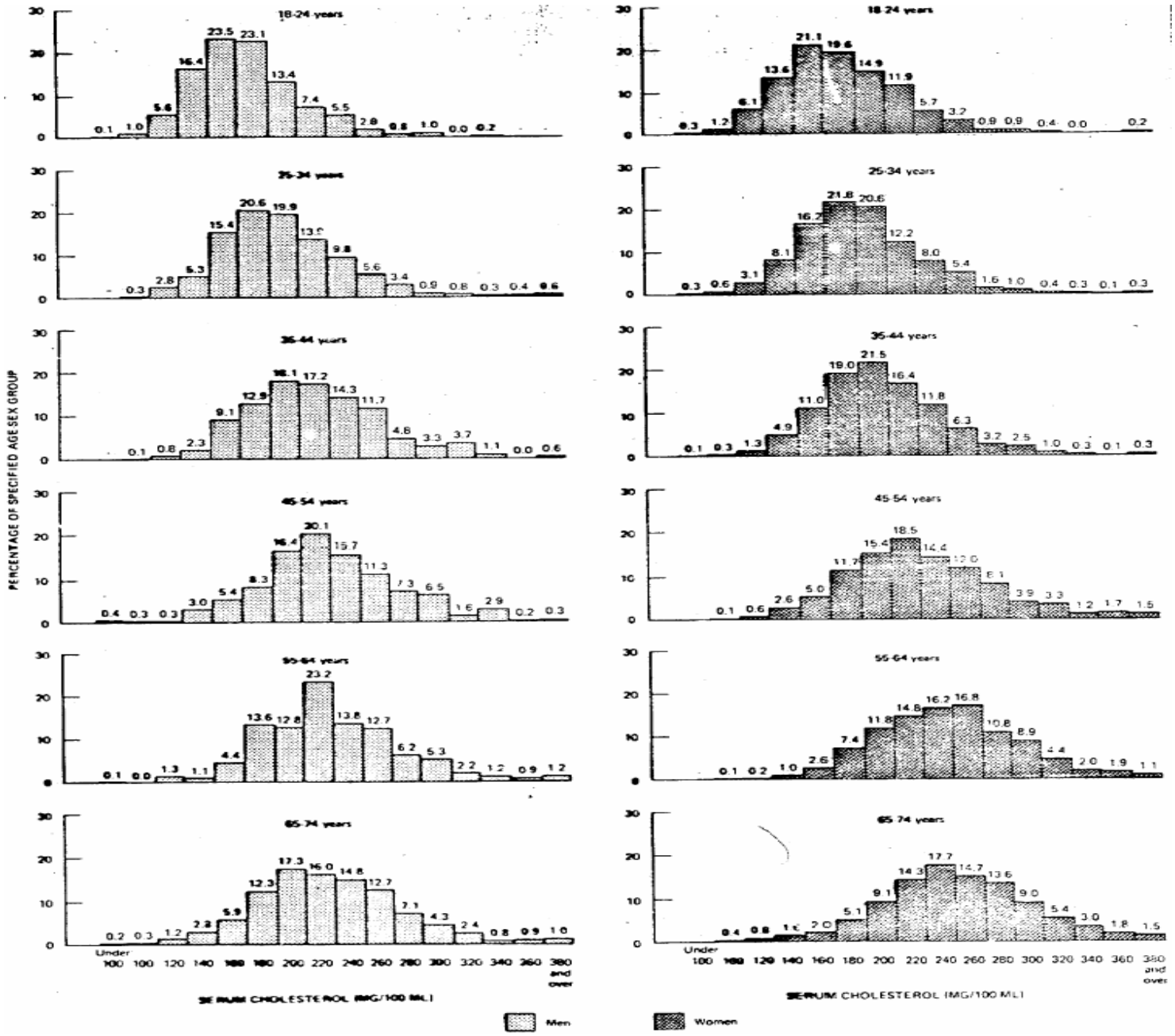



Percentage of Students
0
2
4
6
8
10
12
14
16
18
20
22
24
26
18
17
25
9
6
4
4
20
22
24
26
28
30
32
34
36
38
40
Interval Midpoint

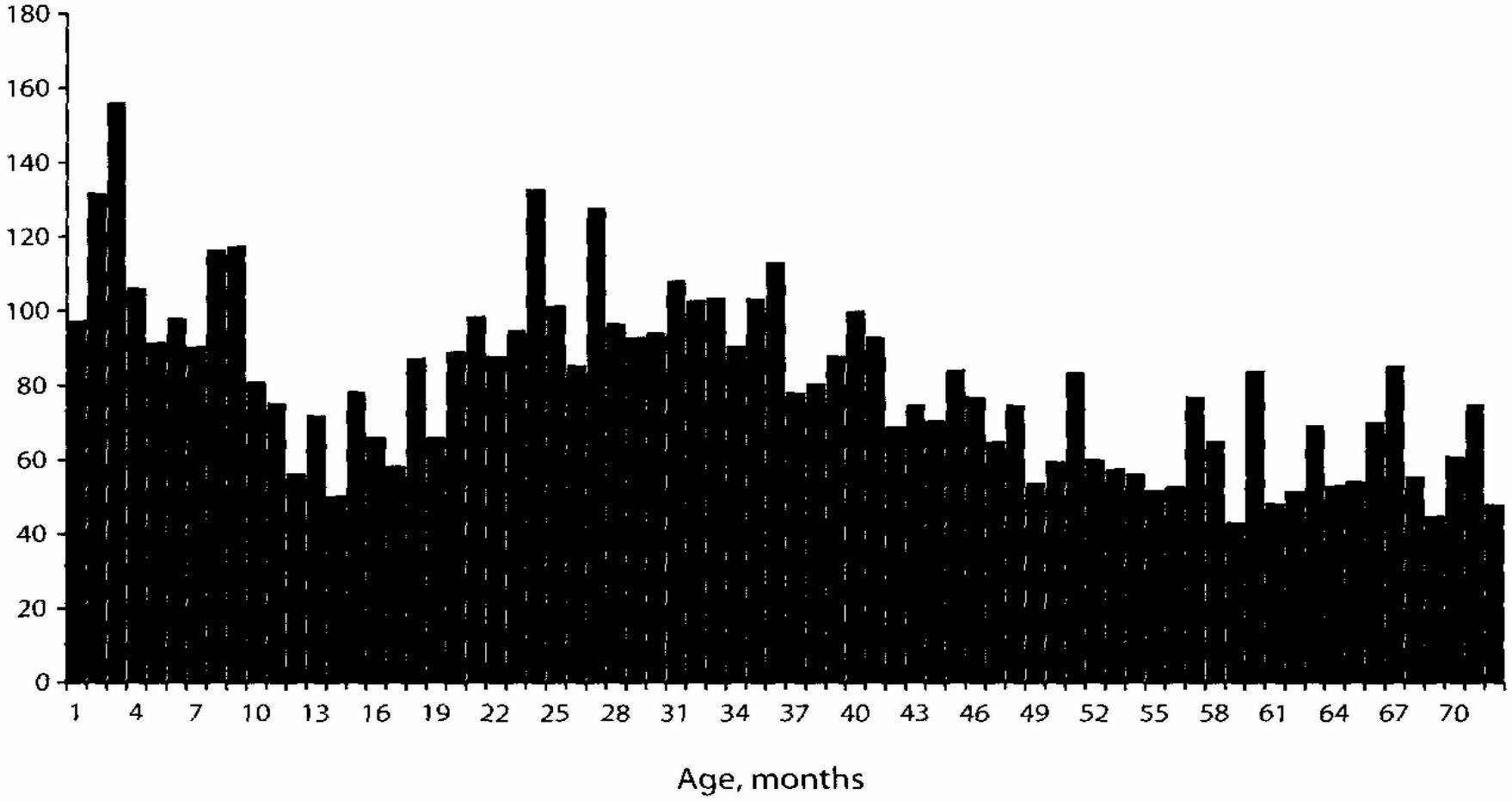
# Dealing with our Interval with Indefinite Length

- The ***area*** shown for an indefinite interval should be on same basis as it is for other intervals. Our indefinite interval contains 17% of the distribution so it needs to include the area equivalent to that of 17 cells, where each cell is 2 years wide by 1% tall.
- The ***width*** of the indefinite interval is 16 years so that it is 8 cells wide
- The ***height*** of the bar for this interval thus needs to be 17/8 or two and 1/8%.


Percentage of students
0
2
4
6
8
10
12
14
16
18
20
22
24
26
25
18
17
9
6
4
4
Indefinite Interval
2 1/8
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
Age Interval Midpoint


PERCENTAGE OF SPECIFIED AGE SEX GROUP
18-24 years
25-34 years
35-44 years
45-54 years
55-64 years
65-74 years
Under 100
100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 and over
SERUM CHOLESTEROL (MG/100 ML)
Men
Women

# *Source: Am J Public Health, April 2004;94:559*




*Source*. Kids' Inpatient Database, 1997.[6]

**FIGURE 1—Estimated number of femur fractures among children in the United States, by month of age.**

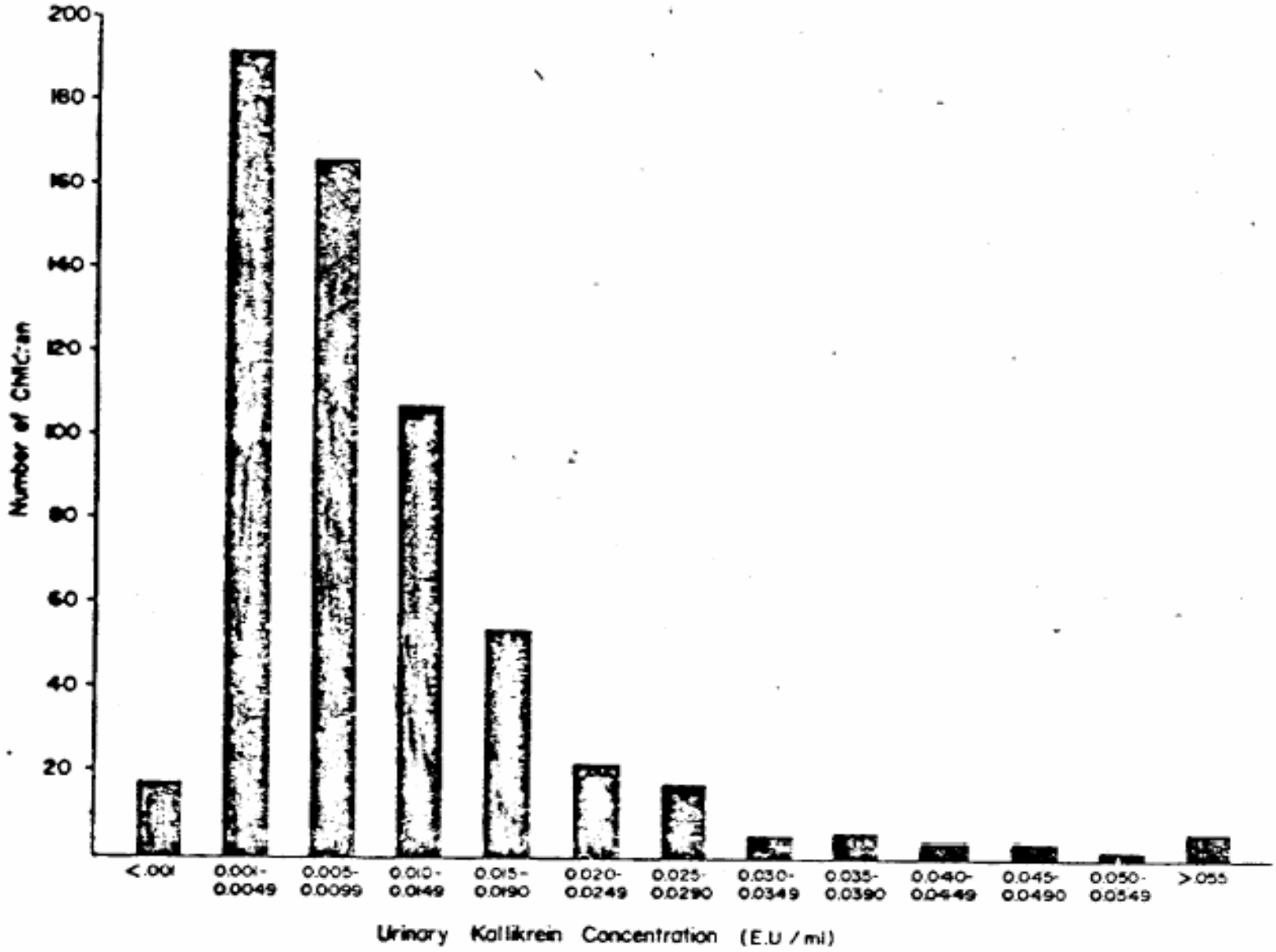



FIGURE 1. Distribution of urinary kallikrein concentration in α-N-p-tosyl-L arginine-[$^3$H]-methyl esterase units (EU/ml) in 601 children.

# Source: *Am J Public Health, June 2004;94:559*

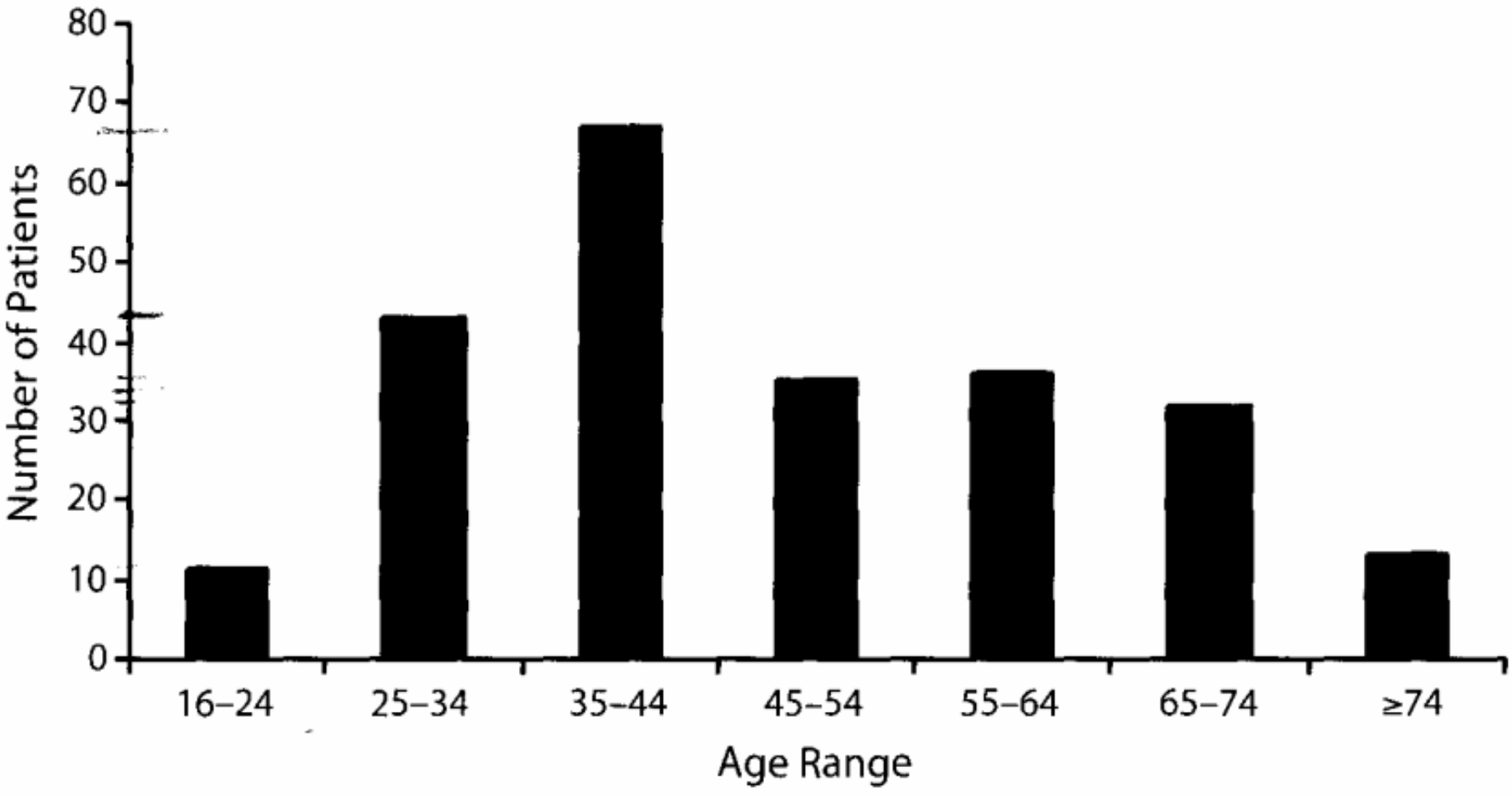
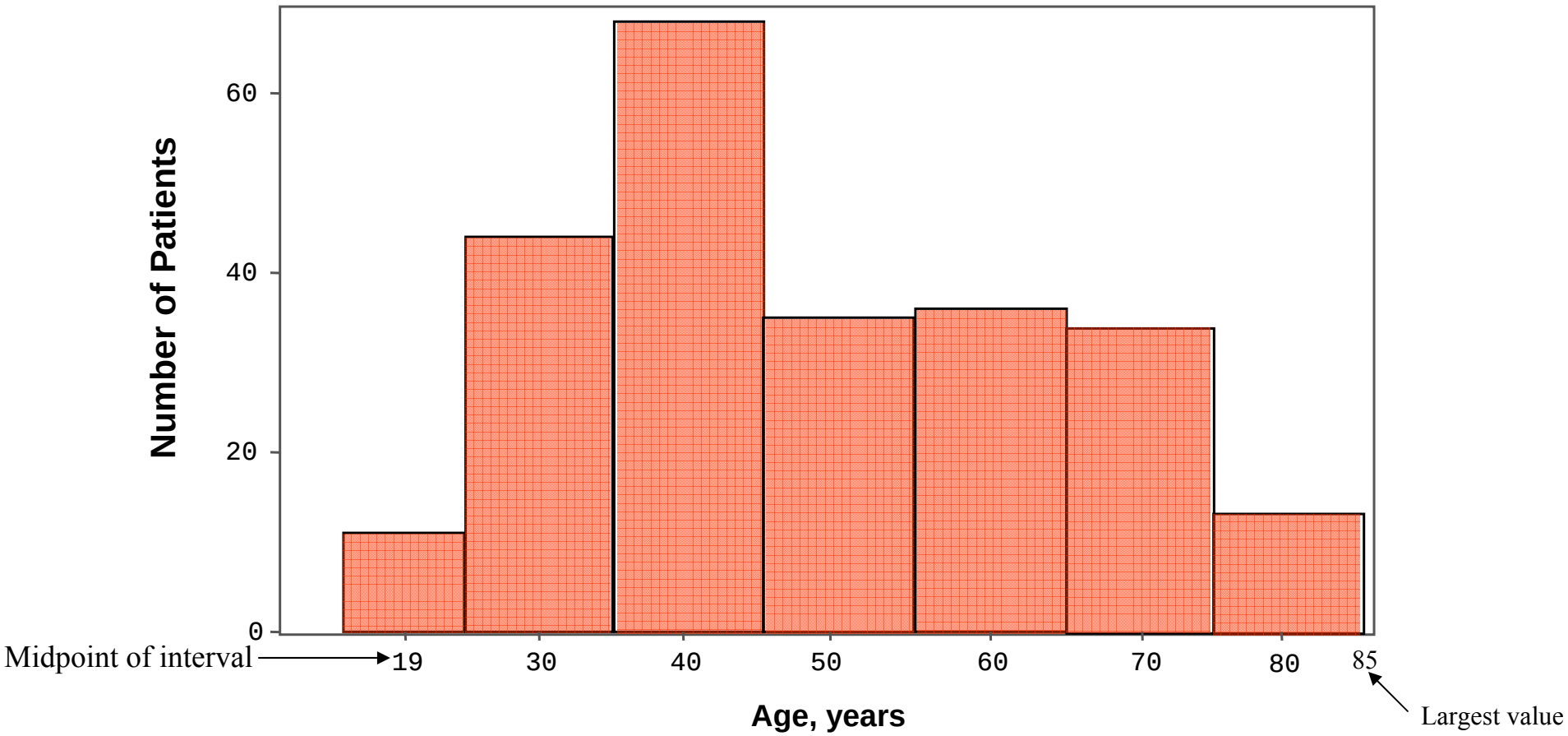


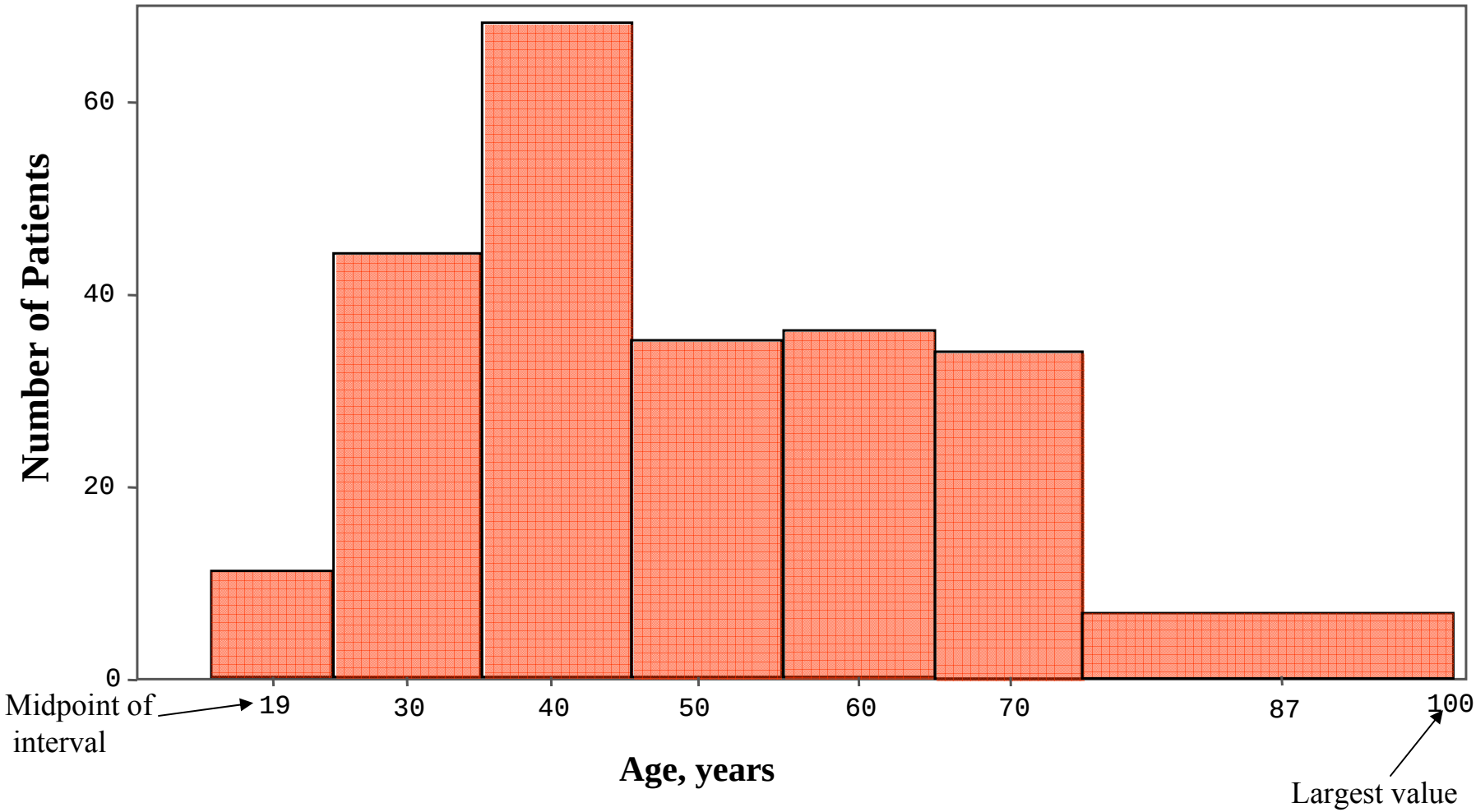

**FIGURE 1—Age distribution of survey participants.**

# Data Source: *Am J Public Health, June 2004;94:559*



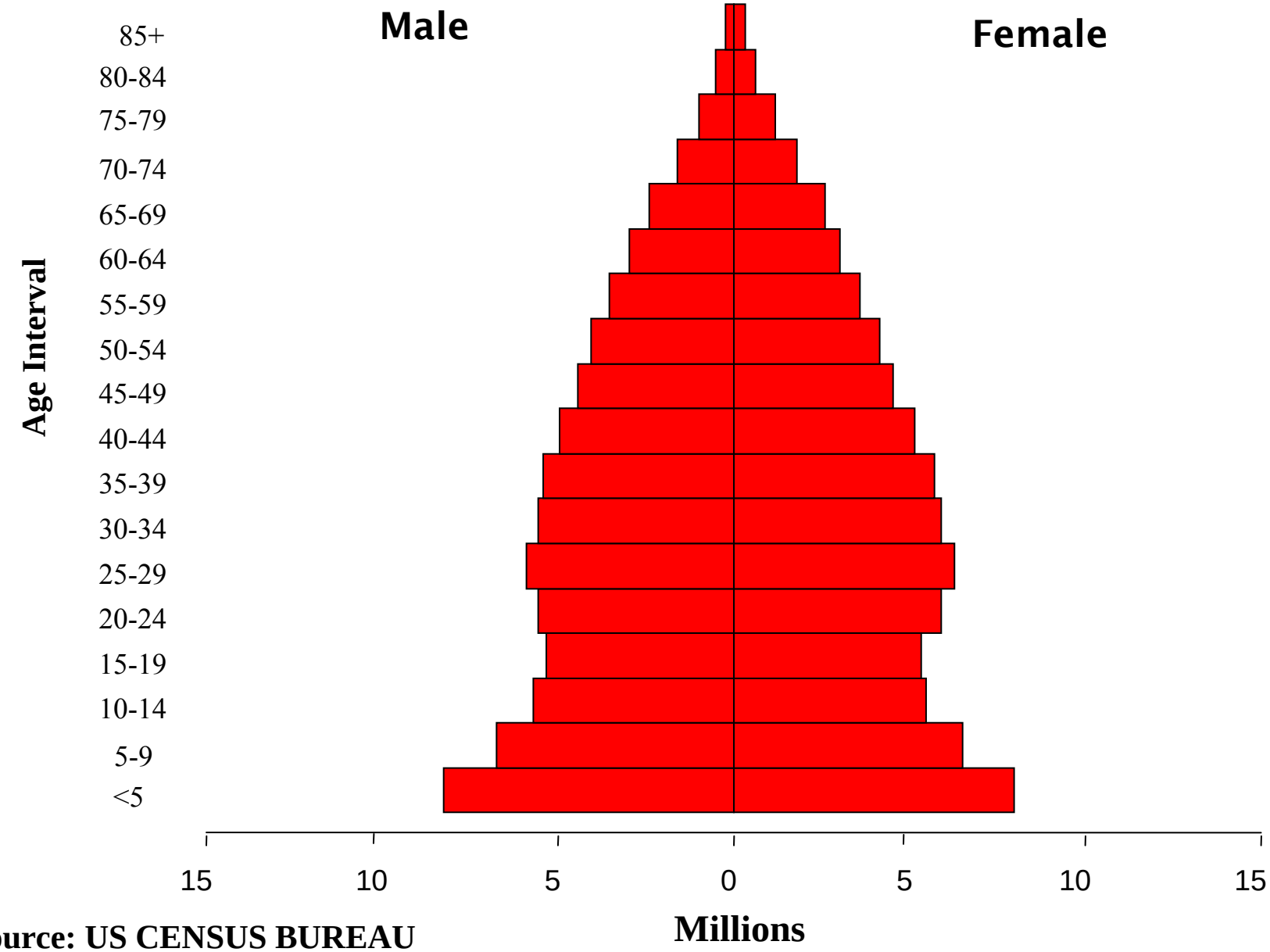

**A revised version of the histogram on the previous slide;** to deal with the indefinite interval for patients who are ≥ 74 years old, we made an arbitrary decision that the oldest patient was 85 years old.

# Data Source: *Am J Public Health, June 2004;94:559*



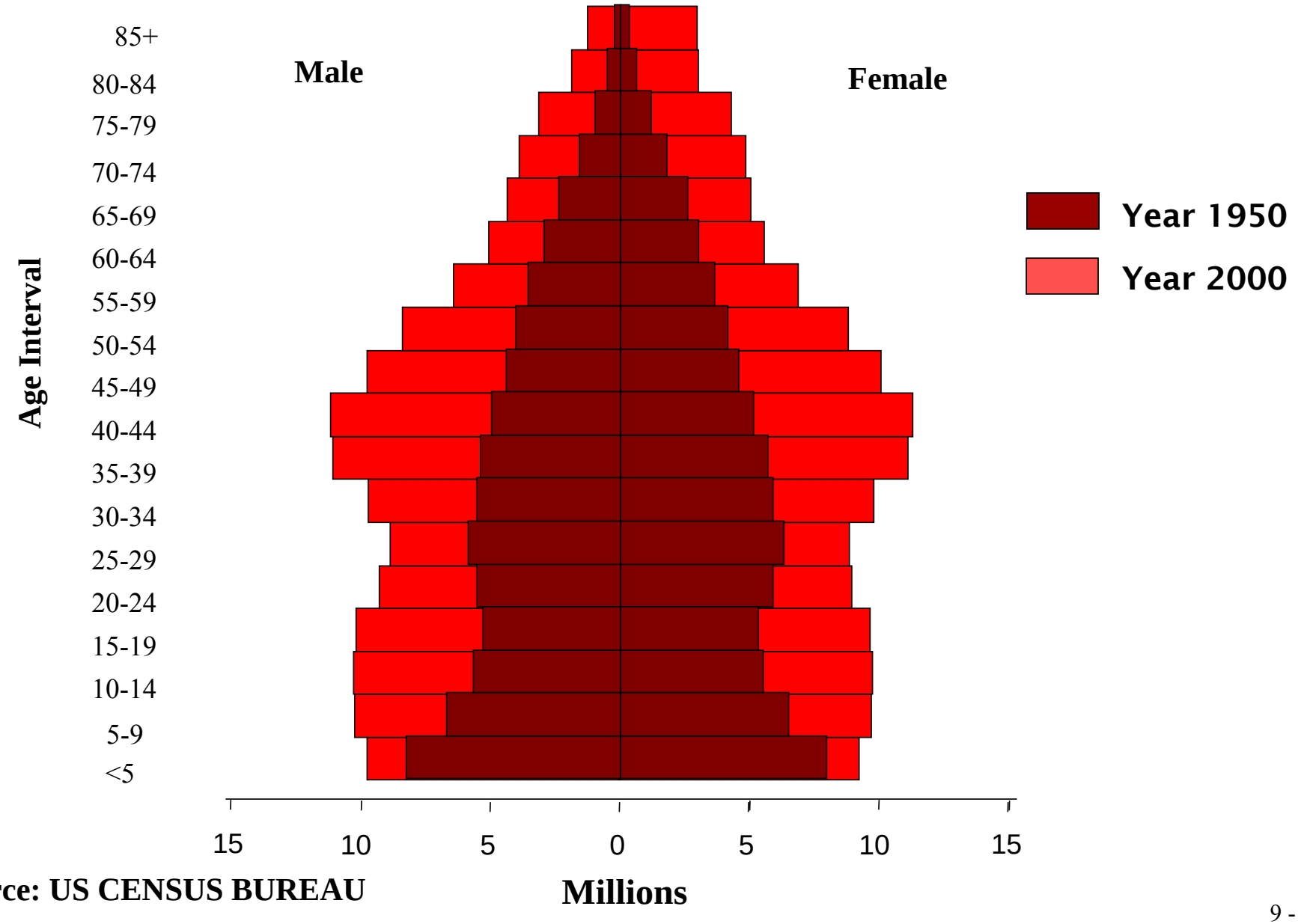

A histogram of data from *AJPH*, June 2004; 94:559 article with oldest patient assumed to be 100 years old.

# Age Distribution of the U.S. Population, by Sex: 1950



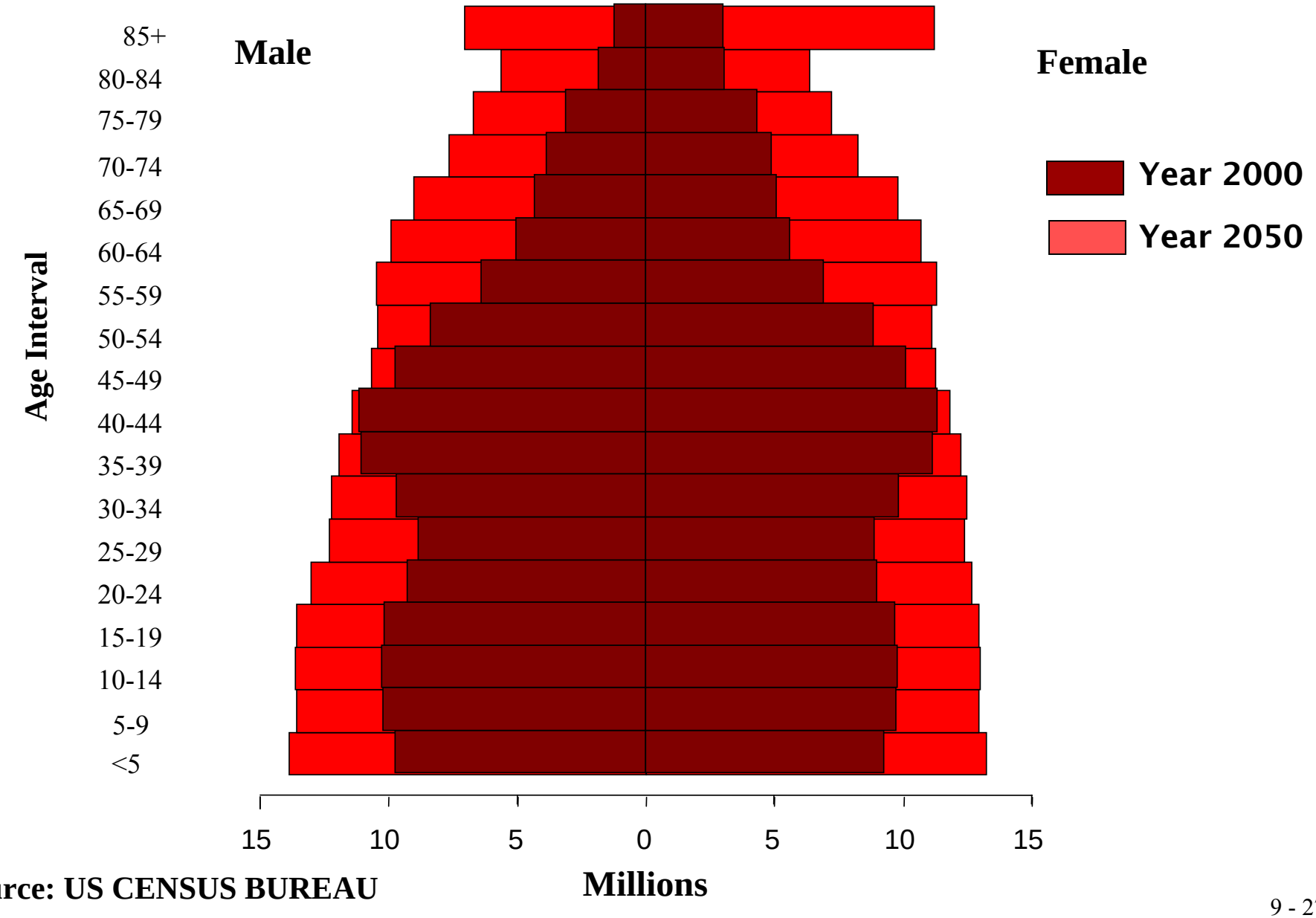

**Source: US CENSUS BUREAU**

# Age Distribution of the U.S. Population, by Sex: 2000



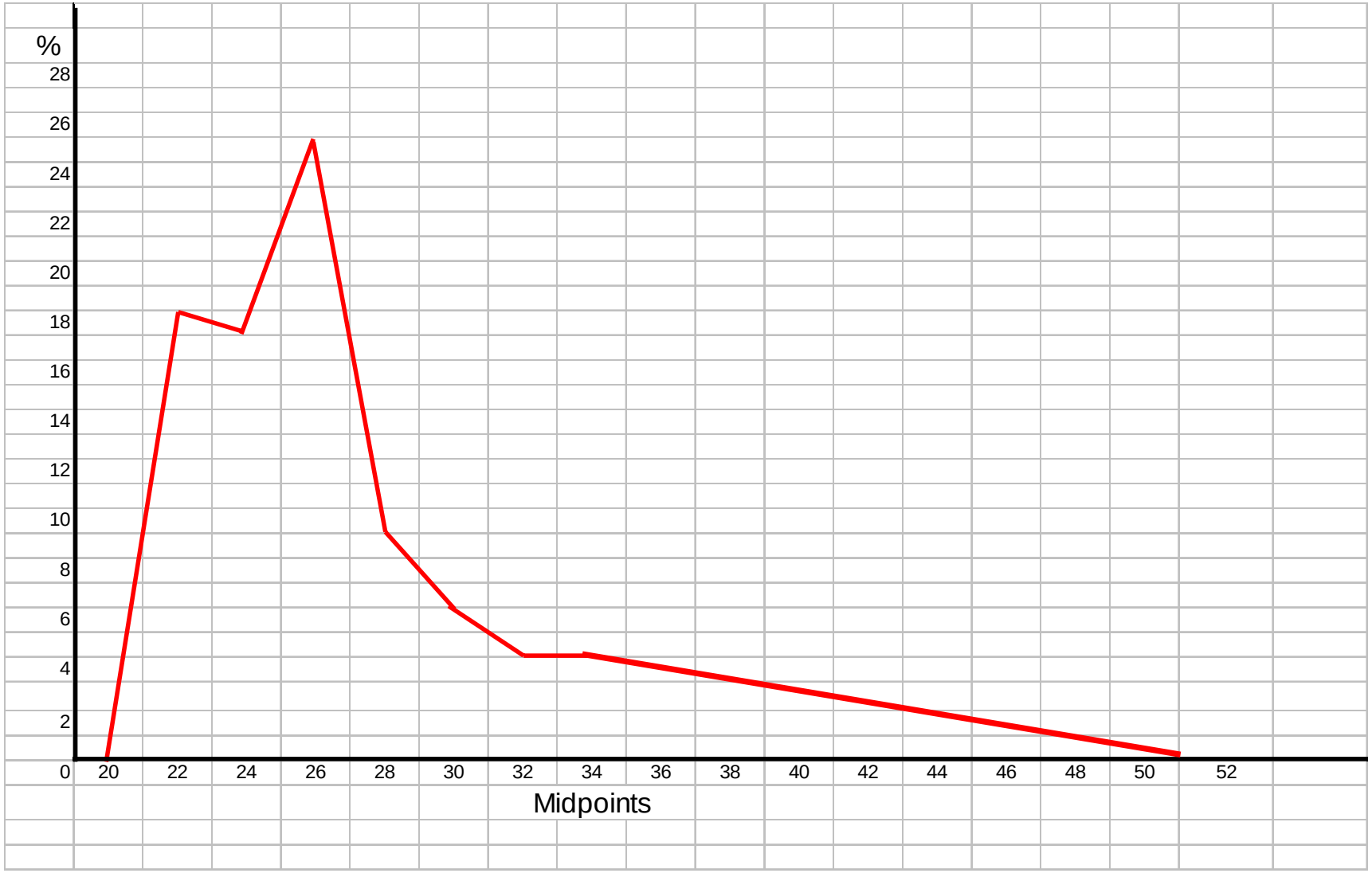

**Source: US CENSUS BUREAU**

# Age Distribution of the U.S. Population, by Sex: 2050




**Source: US CENSUS BUREAU**

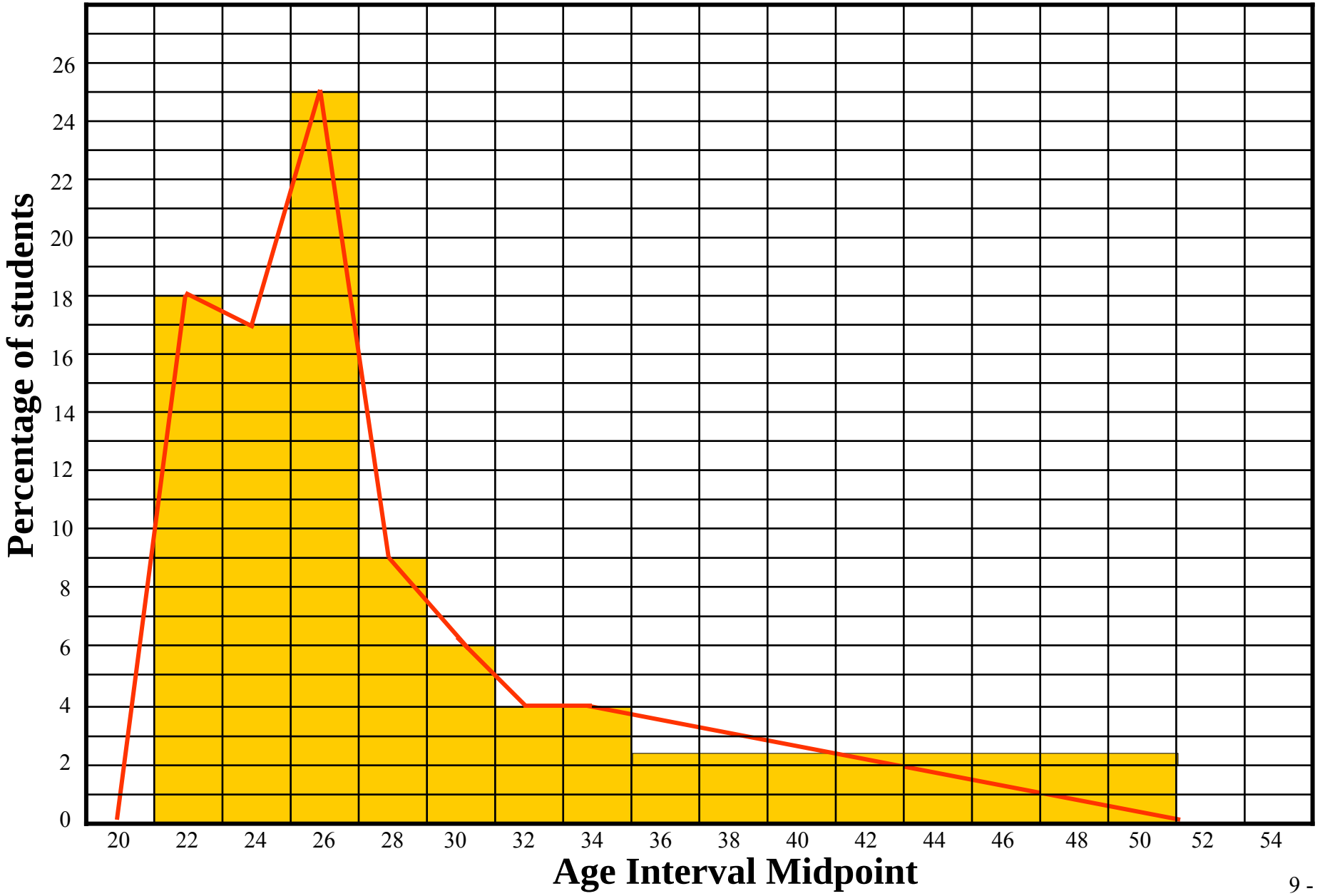
# Frequency Polygon



# Frequency Polygon

A frequency polygon is prepared by connecting the midpoints of the tops of the histogram bars with straight lines in such a manner that the area covered by the resulting figure includes 100% of the histogram area.

# Frequency Polygon for Age Example



# Source: *Am J Public Health, Aug. 2001;91:1209*

**TABLE 1—US Standard Million Population for 1940, 1970, and 2000**

| Age, y | 1940 | 1970 | 2000 |
|---|---|---|---|
| <1 | 15 343 | 17 150 | 13 818 |
| 1-4 | 64 718 | 67 265 | 55 317 |
| 5-14 | 170 355 | 200 511 | 145 565 |
| 15-24 | 181 677 | 174 405 | 138 646 |
| 25-34 | 162 066 | 122 567 | 135 573 |
| 35-44 | 139 237 | 113 616 | 162 613 |
| 45-54 | 117 811 | 114 265 | 134 834 |
| 55-64 | 80 294 | 91 481 | 87 247 |
| 65-74 | 48 426 | 61 192 | 66 037 |
| 75-84 | 17 303 | 30 112 | 44 842 |
| ≥85 | 2 770 | 7 436 | 15 508 |
| Total | 1 000 000 | 1 000 000 | 1 000 000 |

*Source.* Anderson and Rosenberg[6] and Devessa et al.[14]

# Source: *Am J Public Health, Sept. 1976;66:874*

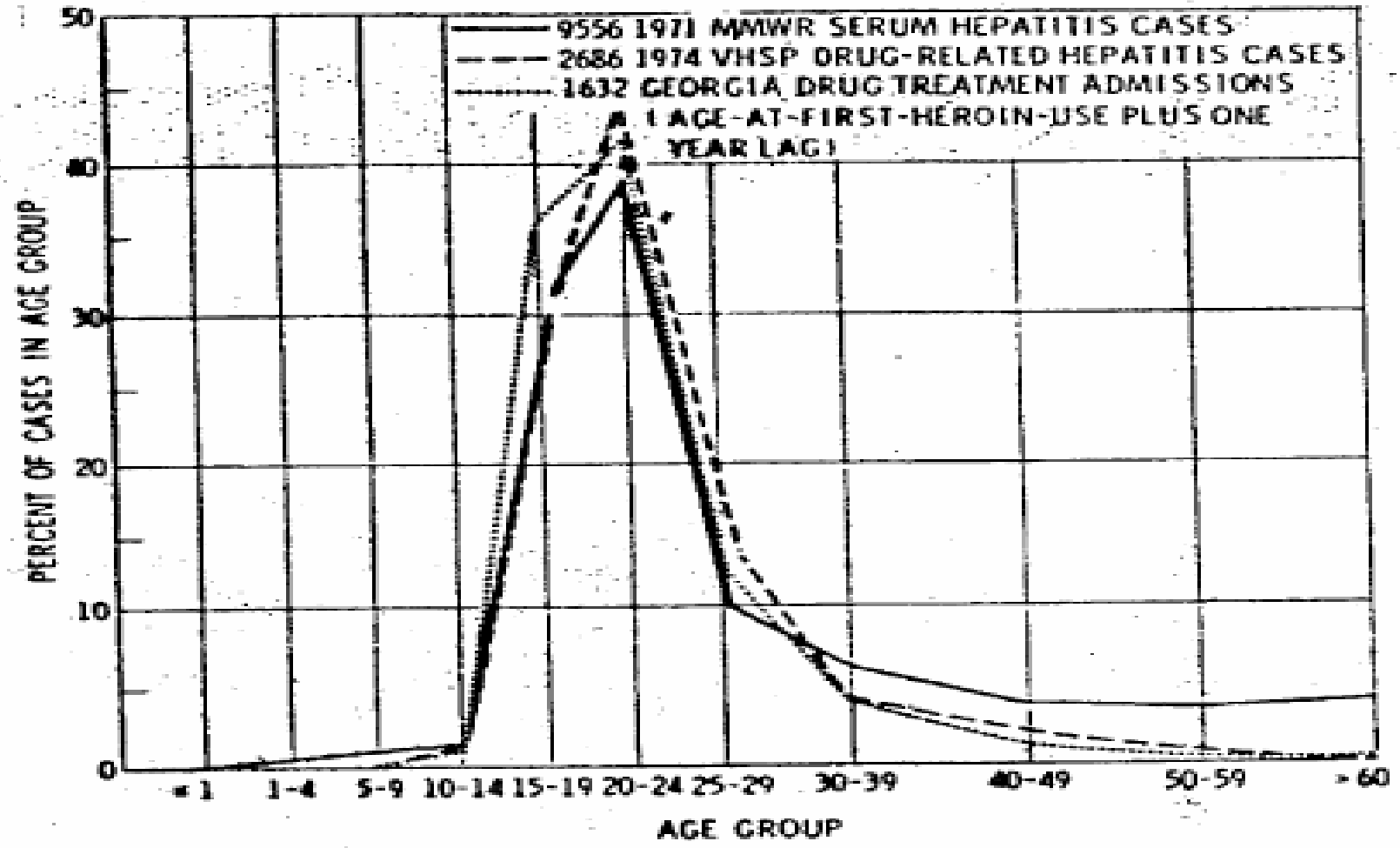



**FIGURE 1—Characteristic Age Distribution of I.V. Drug Use as Seen in Hepatitis and Drug Treatment Data**

# Source: *Am J Public Health, Sept. 1976;66:874*

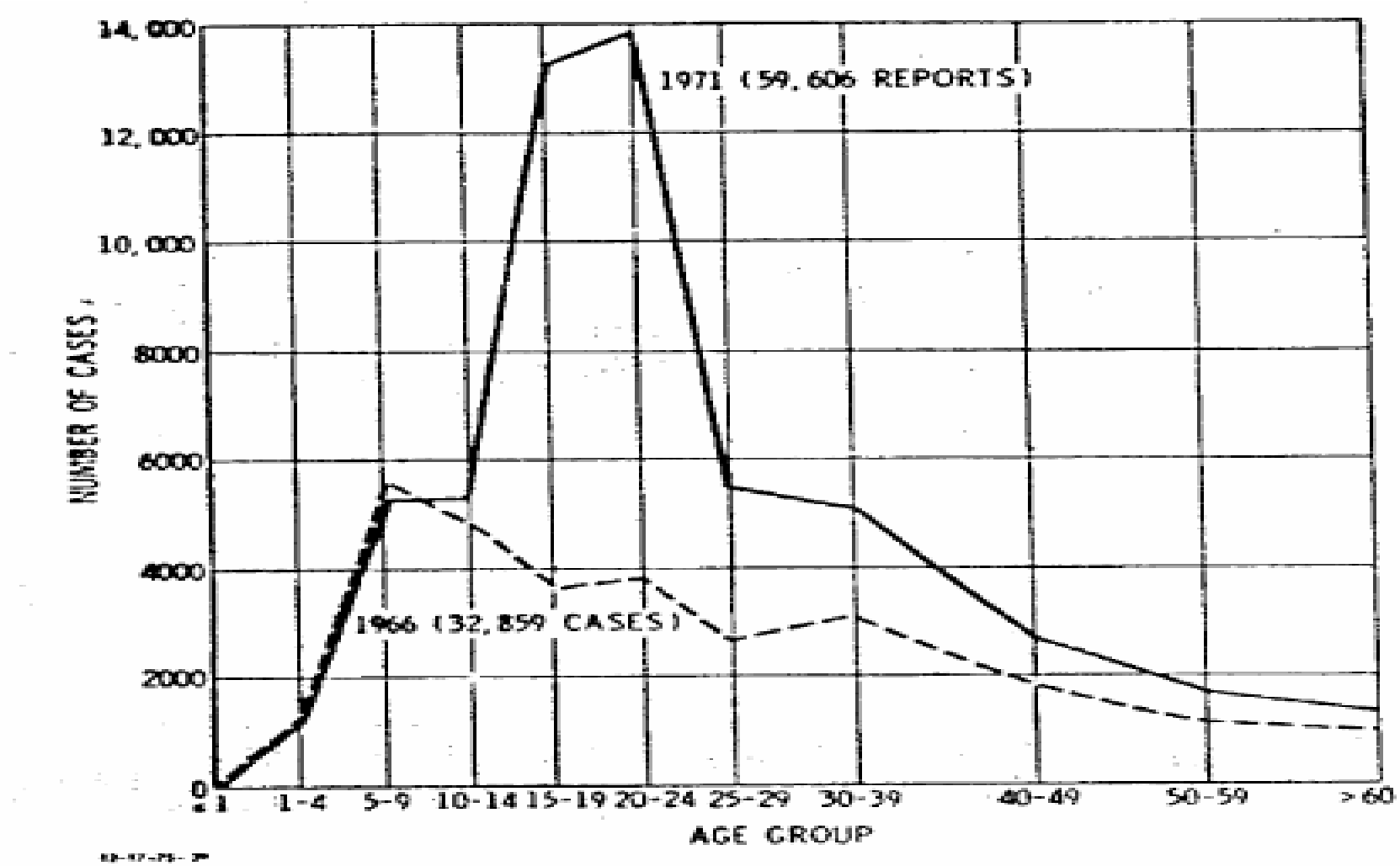



FIGURE 2—Change in the Age Distribution of United States Infectious MMWR Cases from 1966–1971

# Source: *Am J Public Health, Sept. 1976;66:874*

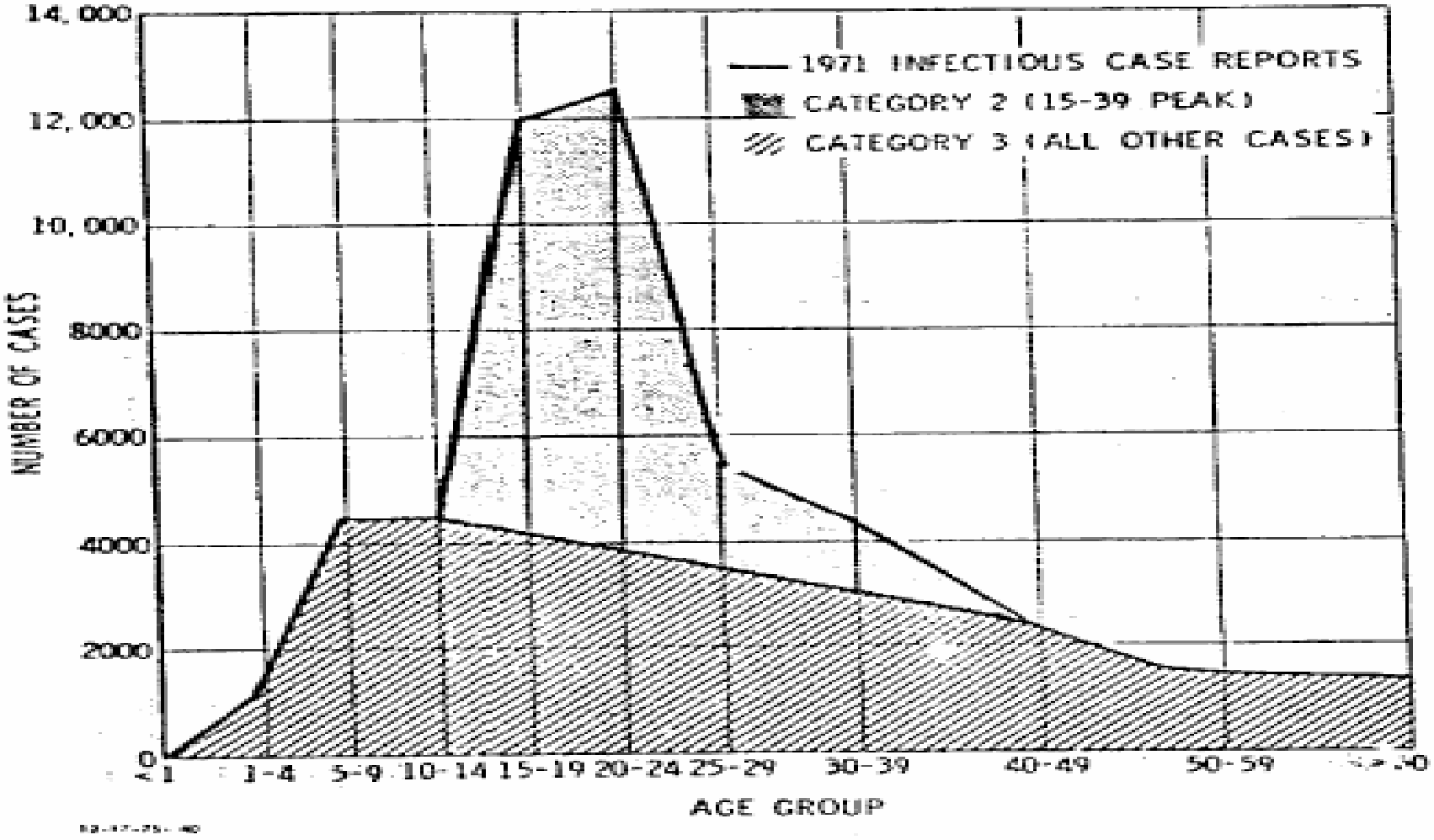
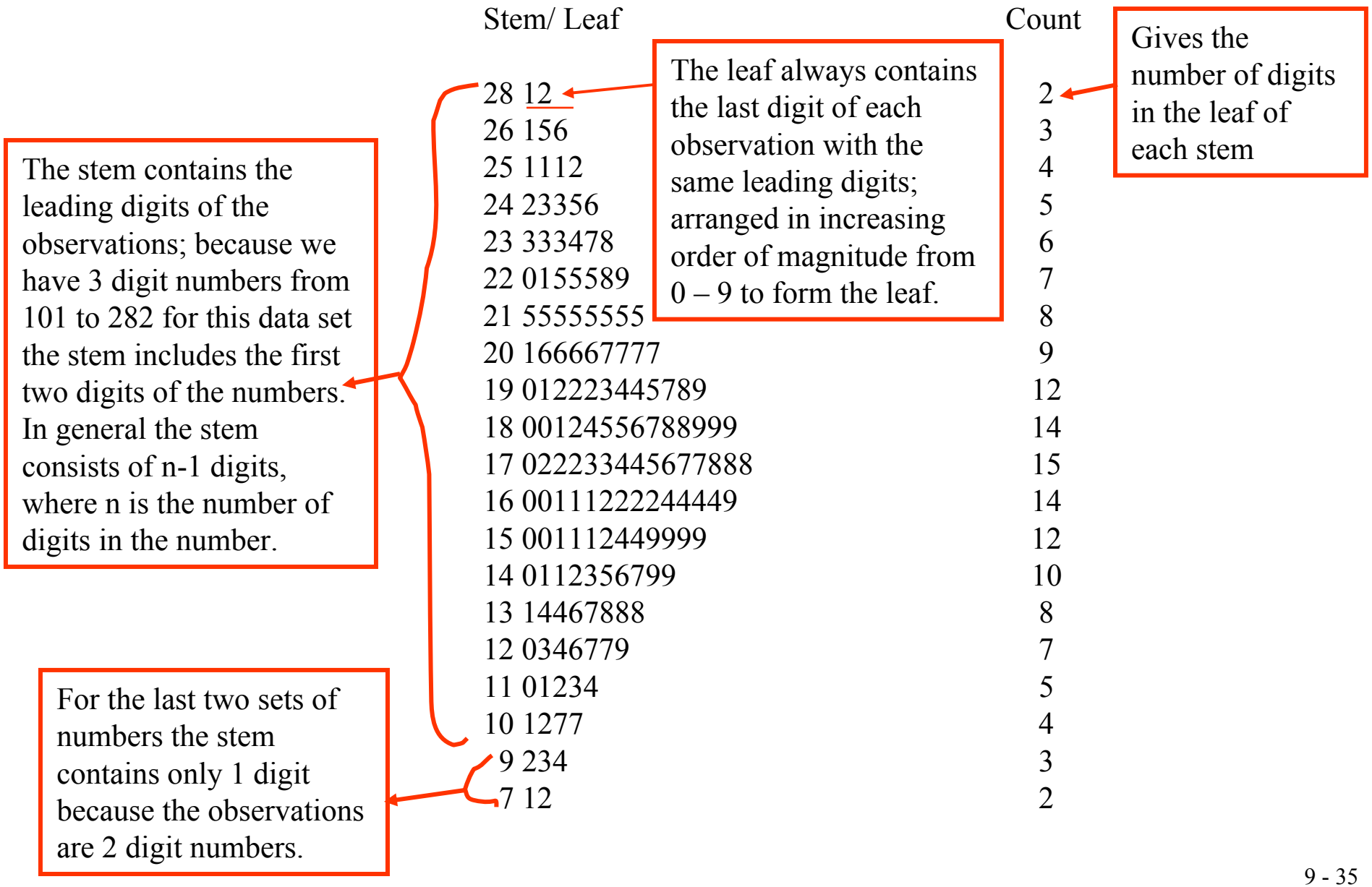



FIGURE 3—Separation of Infectious Hepatitis into Potentially-Drug-Related and Non-Drug-Related Categories

# Stem and Leaf Plot



# Stem and Leaf Plot

- A chart that displays a frequency distribution similar to a histogram.
- A stem and leaf plot shows
  - The spread of the data
  - The mode
  - Whether the distribution is skewed
  - Whether there are gaps in the data
  - Whether there are any unusual data points

# Stem and Leaf Plot

A stem and leaf plot typically consists of three columns, the first two being separated by a single blank space. The first column is the stem or leading digits. The second is the leaf which represents the values in the interval following the stem digits. The third column indicates the number of data values or count in the interval.

# Stem and Leaf Plot Example

**Question**: Display the information in the table below in a Stem and Leaf Plot.

Blood Cholesterol Values

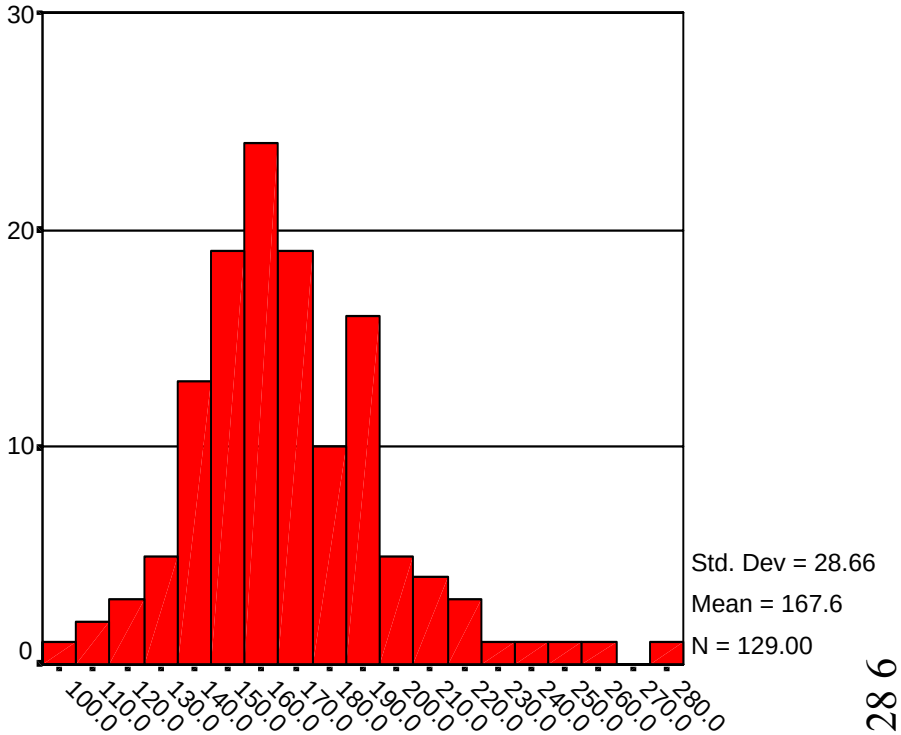
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 201 | **26** | 172 | **51** | 162 | **76** | 127 | **101** | 206 | **126** | 172 |
| **2** | 182 | **27** | 164 | **52** | 151 | **77** | 147 | **102** | 159 | **127** | 276 |
| **3** | 199 | **28** | 136 | **53** | 197 | **78** | 164 | **103** | 141 | **128** | 124 |
| **4** | 136 | **29** | 161 | **54** | 206 | **79** | 161 | **104** | 166 | **129** | 180 |
| **5** | 152 | **30** | 160 | **55** | 160 | **80** | 178 | **105** | 154 | | |
| **6** | 195 | **31** | 165 | **56** | 131 | **81** | 177 | **106** | 111 | | |
| **7** | 162 | **32** | 169 | **57** | 193 | **82** | 176 | **107** | 228 | | |
| **8** | 206 | **33** | 159 | **58** | 235 | **83** | 146 | **108** | 95 | | |
| **9** | 138 | **34** | 168 | **59** | 192 | **84** | 179 | **109** | 188 | | |
| **10** | 190 | **35** | 185 | **60** | 221 | **85** | 185 | **110** | 134 | | |
| **11** | 152 | **36** | 189 | **61** | 194 | **86** | 155 | **111** | 198 | | |
| **12** | 120 | **37** | 174 | **62** | 153 | **87** | 150 | **112** | 140 | | |
| **13** | 169 | **38** | 114 | **63** | 168 | **88** | 167 | **113** | 188 | | |
| **14** | 136 | **39** | 161 | **64** | 162 | **89** | 154 | **114** | 154 | | |
| **15** | 141 | **40** | 153 | **65** | 162 | **90** | 159 | **115** | 191 | | |
| **16** | 194 | **41** | 165 | **66** | 158 | **91** | 187 | **116** | 169 | | |
| **17** | 173 | **42** | 142 | **67** | 143 | **92** | 164 | **117** | 156 | | |
| **18** | 158 | **43** | 173 | **68** | 184 | **93** | 151 | **118** | 141 | | |
| **19** | 181 | **44** | 138 | **69** | 133 | **94** | 155 | **119** | 172 | | |
| **20** | 247 | **45** | 174 | **70** | 180 | **95** | 159 | **120** | 206 | | |
| **21** | 192 | **46** | 186 | **71** | 165 | **96** | 261 | **121** | 145 | | |
| **22** | 192 | **47** | 175 | **72** | 149 | **97** | 169 | **122** | 138 | | |
| **23** | 123 | **48** | 164 | **73** | 155 | **98** | 137 | **123** | 170 | | |
| **24** | 149 | **49** | 220 | **74** | 129 | **99** | 154 | **124** | 151 | | |
| **25** | 158 | **50** | 150 | **75** | 217 | **100** | 189 | **125** | 154 | | |

# Stem Leaf Plot for Blood Cholesterol Data

**SAS Output**

| Stem/Leaf | Count |
|---|---|
| 27 6 | 1 |
| 26 1 | 1 |
| 25 | 0 |
| 24 7 | 1 |
| 23 5 | 1 |
| 22 018 | 3 |
| 21 7 | 1 |
| 20 16666 | 5 |
| 19 012223445789 | 12 |
| 18 0012455678899 | 13 |
| 17 0222334456789 | 13 |
| 16 00111222244455567889999 | 24 |
| 15 001112233444455568889999 | 25 |
| 14 01112356799 | 11 |
| 13 1346667888 | 10 |
| 12 03479 | 5 |
| 11 14 | 2 |
| 10 | 0 |
| 9 5 | 1 |

# Histogram and Stem & Leaf plot for the Cholesterol Data




```
28 6
27
26 1
25
24 7
23 5
22 018
21 7
20 16666
19 012223445789
18 0012455678899
17 022233445678 9
16 00111222244445556788 9999
15 00111223344444555688 89999
14 011123567 99
13 1346667888
12 03479
11 14
10
9  5
```

Note: The stem & Leaf plot is plotted in decreasing order of magnitude, while the Histogram is plotted in increasing order, from smallest to the highest midpoint.