# 2.2 Graphical Displays of Sample Data

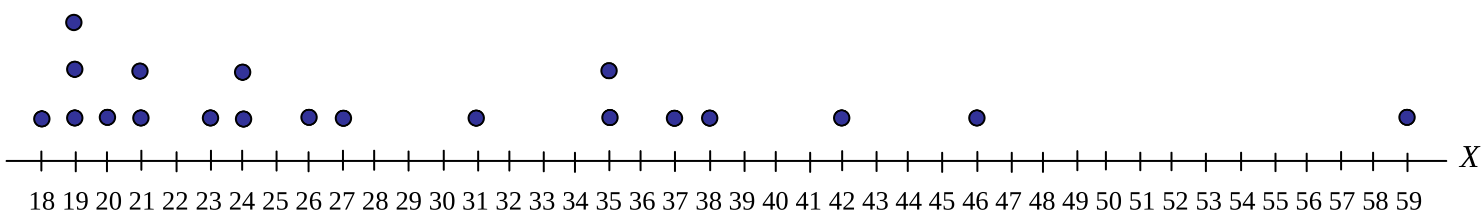
**Dotplots, Stem-and-Leaf Diagrams (Stemplots), Histograms**, Boxplots, Bar Charts, Pie Charts, Pareto Diagrams, …

Example: Random variable $X$ = "Age (years) of individuals at Memorial Union."

Consider the following *sorted* random sample of $n = 20$ ages:

{18, 19, 19, 19, 20, 21, 21, 23, 24, 24, 26, 27, 31, 35, 35, 37, 38, 42, 46, 59}

## ➢ Dotplot




*Comment*: Uses *all* of the values. Simple, but crude; does not summarize the data.

## ➢ Stemplot

Stem Leaves

| Tens | Ones |
|---|---|
| 1 | 8 9 9 9 |
| 2 | 0 1 1 3 4 4 6 7 |
| 3 | 1 5 5 7 8 |
| 4 | 2 6 |
| 5 | 9 |

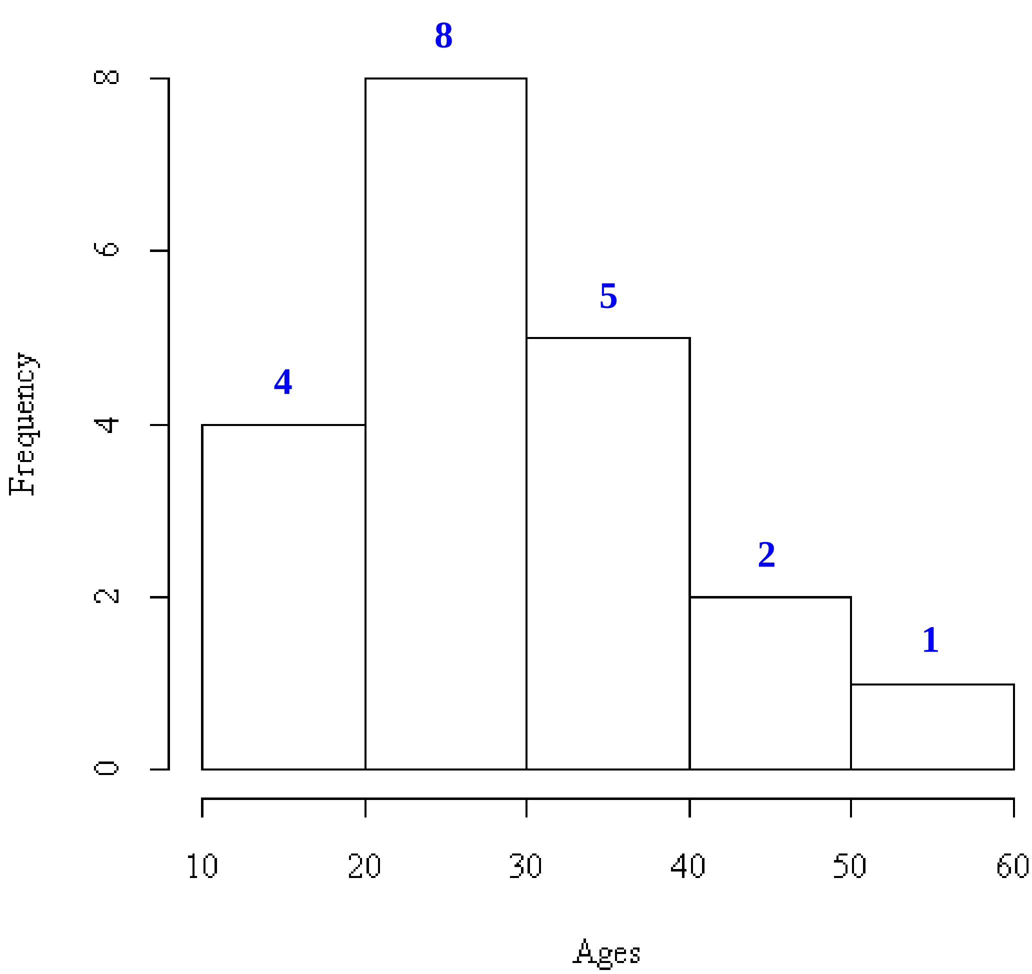
*Comment*: Uses all of the values more effectively. *Grouping summarizes the data better*.

## ➢ Histograms

| **Class Interval** | **Frequency** (# occurrences) |
|---|---|
| [10, 20) | **4** |
| [20, 30) | **8** |
| [30, 40) | **5** |
| [40, 50) | **2** |
| [50, 60) | **1** |

$n = 20$

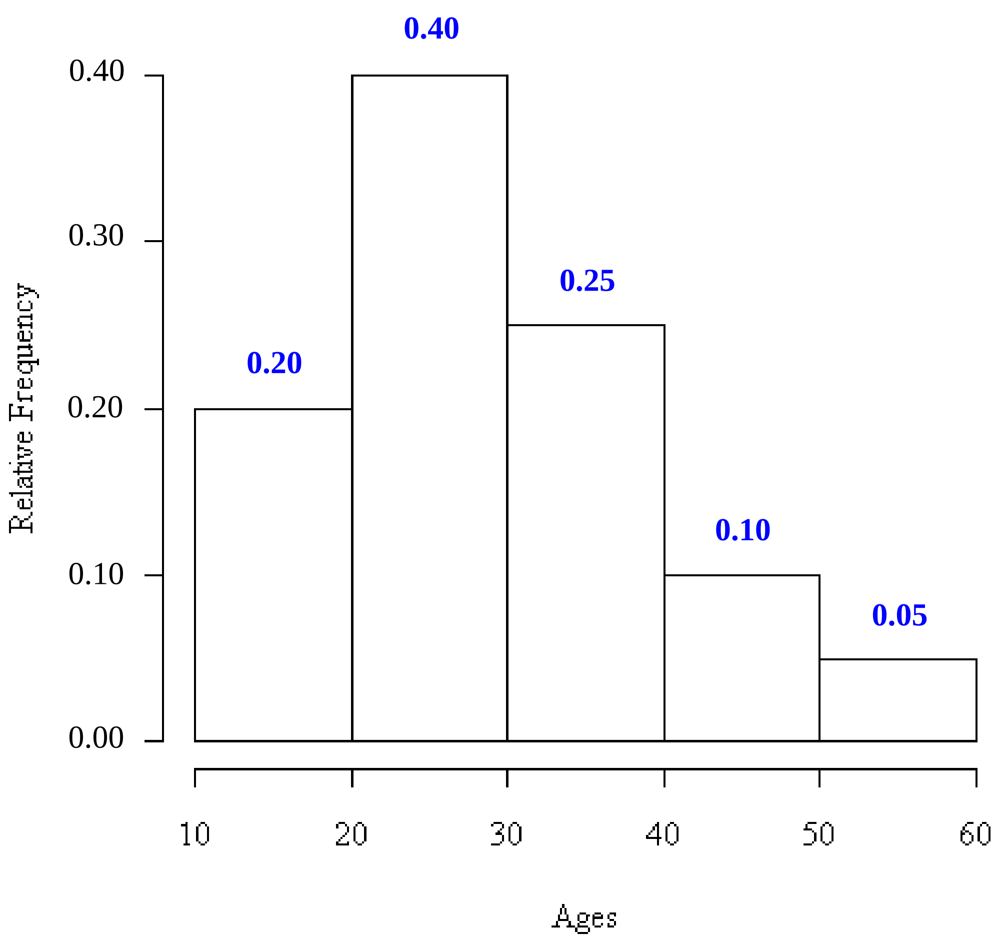
### Frequency Histogram



| **Class Interval** | **Absolute Frequency** (# occurrences) | **Relative Frequency** (Frequency ÷ $n$) |
|---|---|---|
| [10, 20) | 4 | $\frac{4}{20} = 0.20$ |
| [20, 30) | 8 | $\frac{8}{20} = 0.40$ |
| [30, 40) | 5 | $\frac{5}{20} = 0.25$ |
| [40, 50) | 2 | $\frac{2}{20} = 0.10$ |
| [50, 60) | 1 | $\frac{1}{20} = 0.05$ |
| | $n = 20$ | $\frac{20}{20} = 1.00$ |

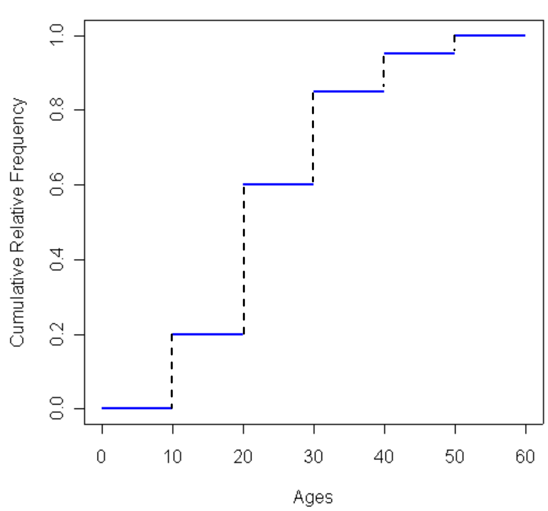
**Relative frequencies are *always* <u>between 0 and 1</u>, and their sum is *always* = 1 !**

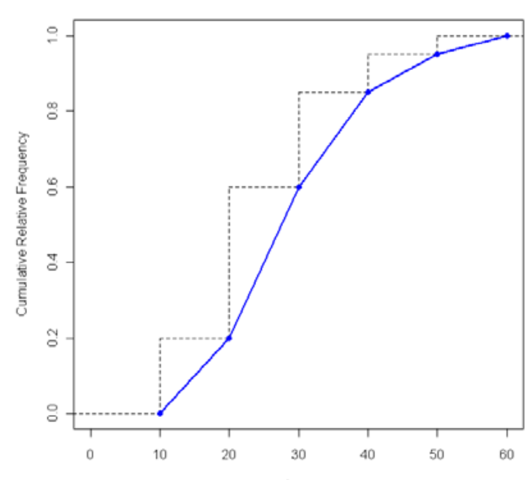
## Relative Frequency Histogram



| Class Interval | Absolute Frequency (# occurrences) | Relative Frequency (Frequency ÷ $n$) | Cumulative Relative Frequency |
|---|---|---|---|
| [0, 10) | 0 | 0.00 | **0.00** |
| [10, 20) | 4 | 0.20 | **0.20** = 0.00 + 0.20 |
| [20, 30) | 8 | 0.40 | **0.60** = 0.20 + 0.40 |
| [30, 40) | 5 | 0.25 | **0.85** = 0.60 + 0.25 |
| [40, 50) | 2 | 0.10 | **0.95** = 0.85 + 0.10 |
| [50, 60) | 1 | 0.05 | **1.00** = 0.95 + 0.05 |
| | $n$ = 20 | 1.00 | |

Often, it is of interest to determine the total relative frequency, *up to* a certain value. For example, we see here that 0.60 of the age data are *under* 30 years, 0.85 are *under* 40 years, etc. The resulting **cumulative distribution**, which always increases monotonically from 0 to 1, can be represented by the discontinuous "step function" or "staircase function" in the first graph below. By connecting the right endpoints of the steps, we obtain a *continuous* polygonal graph called the **ogive** (pronounced "o-jive"), shown in the second graph. This has the advantage of approximating the rate at which the cumulative distribution increases within the intervals. For example, suppose we wish to know the **median** age, i.e., the age that divides the values into equal halves, above and below. It is clear from the original data that **25** does this job, but if data are unavailable, we can still estimate it from the ogive. Imagine drawing a flat line from 0.5 on the vertical axis until it hits the graph, then straight down to the horizontal "Age" axis somewhere in the interval [20, 30); it is this value we seek. But the cumulative distribution up to 20 years is 0.2, and up to 30 years is 0.6… a rise of 0.4 in 10 years, or 0.04 per year, on average. To reach 0.5 from 0.2 – an increase of 0.3 – would thus require a ratio of 0.3 / 0.04 = 7.5 years from 20 years, or **27.5 years**. Medians and other **percentiles** will be addressed in the next section.



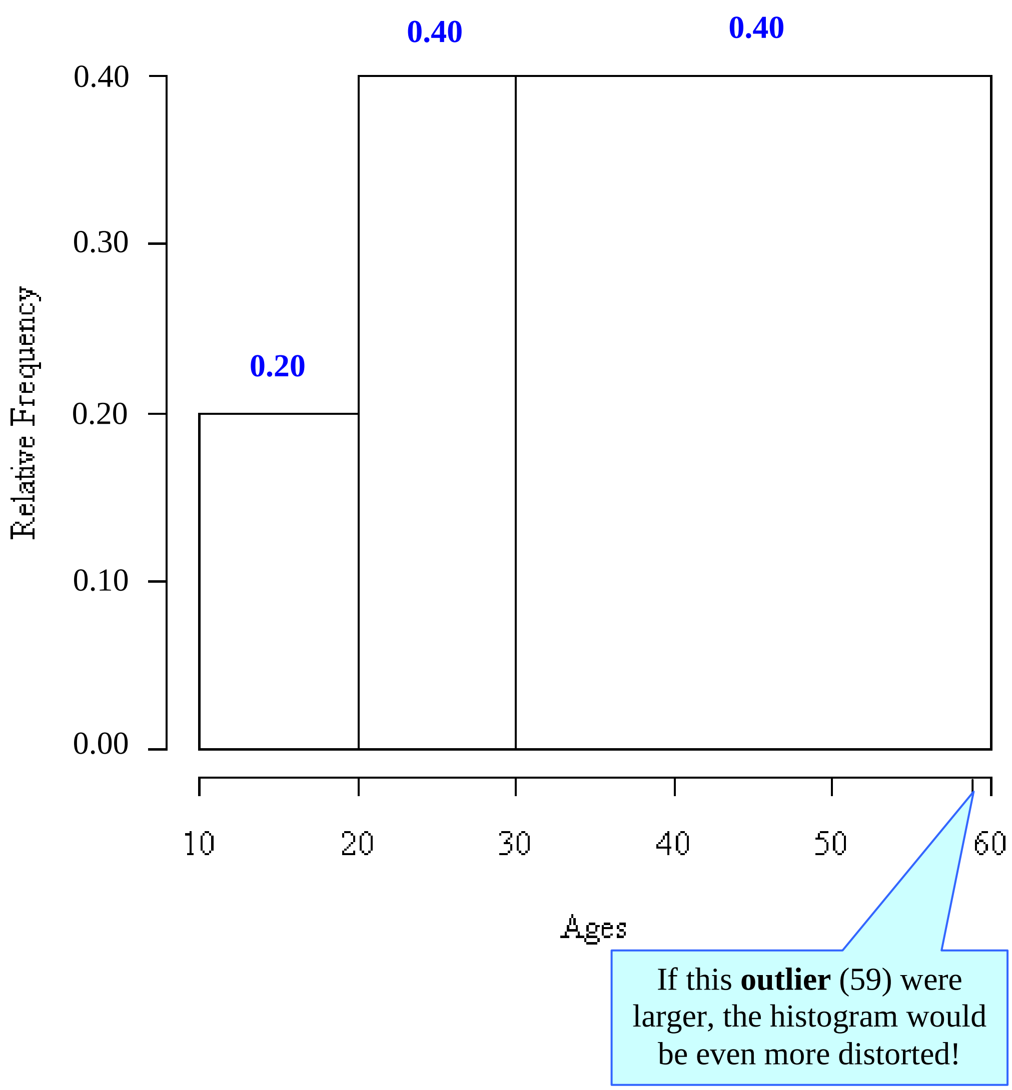



**Problem!** Suppose that all ages 30 and older are "lumped" into a single class interval:

{18, 19, 19, 19, 20, 21, 21, 23, 24, 24, 26, 27, **31, 35, 35, 37, 38, 42, 46, 59**}

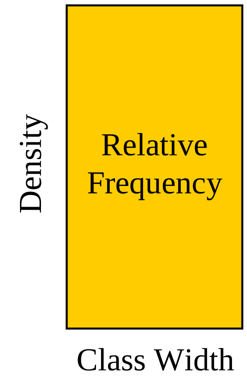
| Class Interval | Absolute Frequency (# occurrences) | Relative Frequency (Frequency ÷ $n$) |
|---|---|---|
| [10, 20) | 4 | $\frac{4}{20}$ = **0.20** |
| [20, 30) | 8 | $\frac{8}{20}$ = **0.40** |
| [30, 60) | 8 | $\frac{8}{20}$ = **0.40** |
| | $n$ = 20 | $\frac{20}{20}$ = 1.00 |

## Relative Frequency Histogram



***Remedy*:** Let… $\underbrace{\underline{\text{Area}}\text{ of each class rectangle}}_{\underbrace{\text{Height of rectangle}}_{\text{Density}} \times \text{ Class Width}} = \text{Relative Frequency}$

Therefore… $\textbf{Density} = \dfrac{\text{Relative Frequency}}{\text{Class Width}}$




| **Class Interval** | **Absolute Frequency** (# occurrences) | **Relative Frequency** (Frequency ÷ $n$) | **Density** (Rel Freq ÷ Class Width) |
|---|---|---|---|
| [10, 20); **width = 10** | 4 | $\frac{4}{20}$ = **0.20** | $\frac{0.20}{10}$ = 0.02 |
| [20, 30); **width = 10** | 8 | $\frac{8}{20}$ = **0.40** | $\frac{0.40}{10}$ = 0.04 |
| [30, 60); **width = 30** | 8 | $\frac{8}{20}$ = **0.40** | $\frac{0.40}{30}$ = 0.01333… |
| | $n$ = 20 | $\frac{20}{20}$ = 1.00 | |

## Density Histogram