Stem-and-Leaf Display: presents a graphical display of the data using the actual value of each observation. - good for small data sets

To construct a stem-and-leaf-display:

1) Order the data values from smallest to largest. (ascending order)
2) Divide each data value into two parts: the **stem** and **leaf**.
   - the stem usually consists of all digits except the final one.
   - data values may be rounded.
3) List the stems in a column, starting with the smallest stem at the top, and draw a vertical line to the right.
   - can "stretch" the stems by dividing each one into several lines.

     0-4, 5-9 or 0-1, 2-3, ..., 8-9
4) For each data value, record the leaf portion in the same row as the corresponding stem.
5) Provide a key to indicate your coding.

Example: Student Height in our Winter 2020 Stat 151 class:

Variable: height

Decimal point is 1 digit(s) to the right of the colon. → 150.0

Leaf unit = 1

15 : 012222234 → 150, 151, 152, ..., 154

15 : 555777777788889

16 : 000000000001222333333333344444 } 0-4

16 : 555555555555555556777788888889 9 } 5-9

17 : 0000000111122233333

17 : 555555555566777888888

18 : 0000002233333333334

18 : 5558

19 : 23

19 : 55 → 195

eg. leaf unit = 10 leaf unit = 0.1

1950 19.5

Example:

The prices of 22 textbooks in a bookstore were recorded in dollars:

60	93	94	86	55	60	86	82	76	57	60
94	89	60	62	72	30	68	65	75	84	81

Order from smallest to largest:

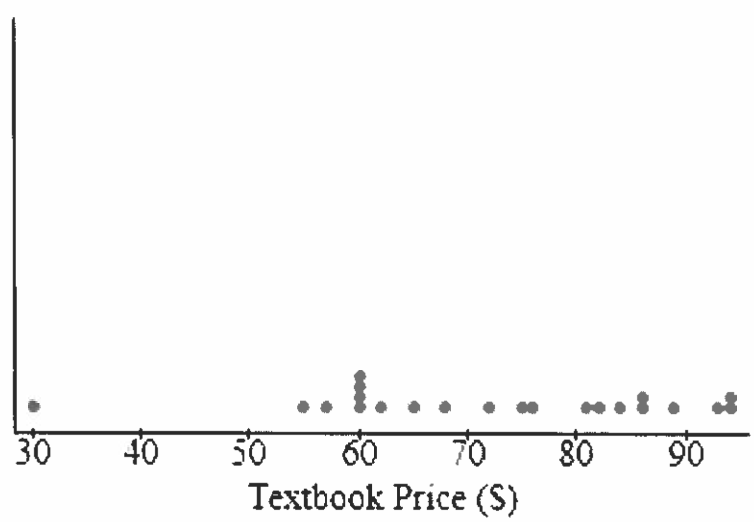
30	55	57	60	60	60	60	62	65	68	72
75	76	81	82	84	86	86	89	93	94	94

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

Textbook Price

(7|2 means $72)

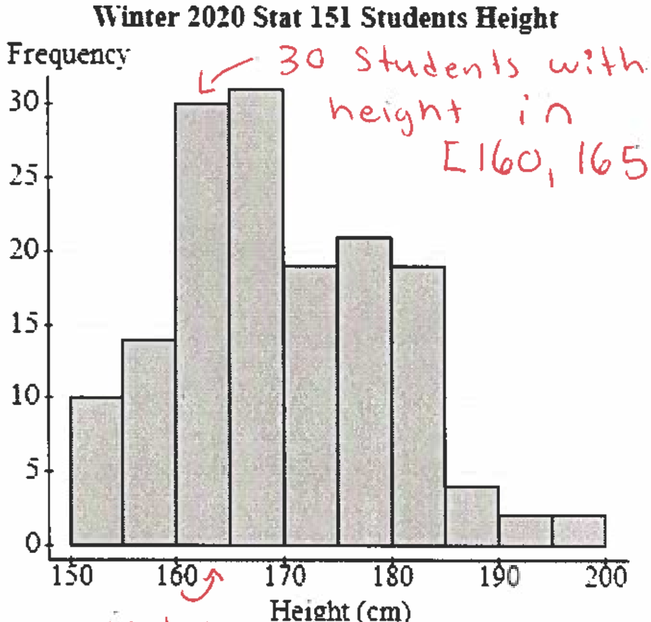
Textbook Prices at a Bookstore

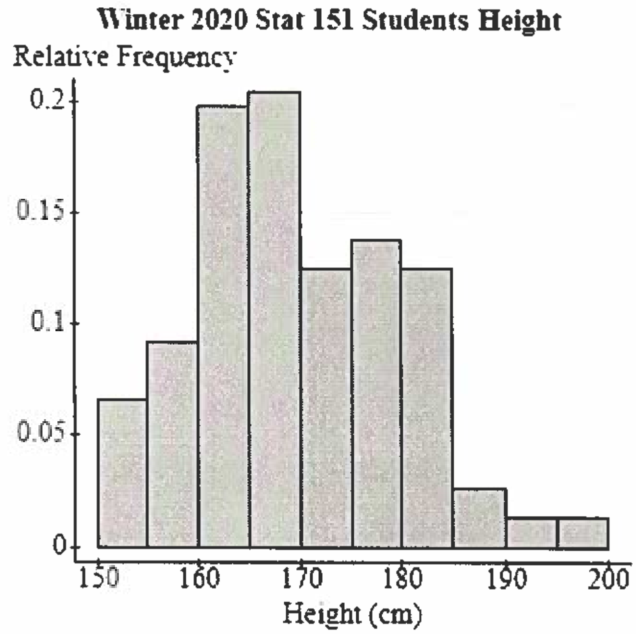


Histogram: a bargraph for a quantitative variable. The set of observations lies in some interval. This interval is partitioned into equal-width subintervals called **bins**. The height of each bar represent the number of observations that fall within the corresponding bin. frequency

- helps to visualize distribution
- good for large data sets
- doesn't show data values

-also use relative frequency





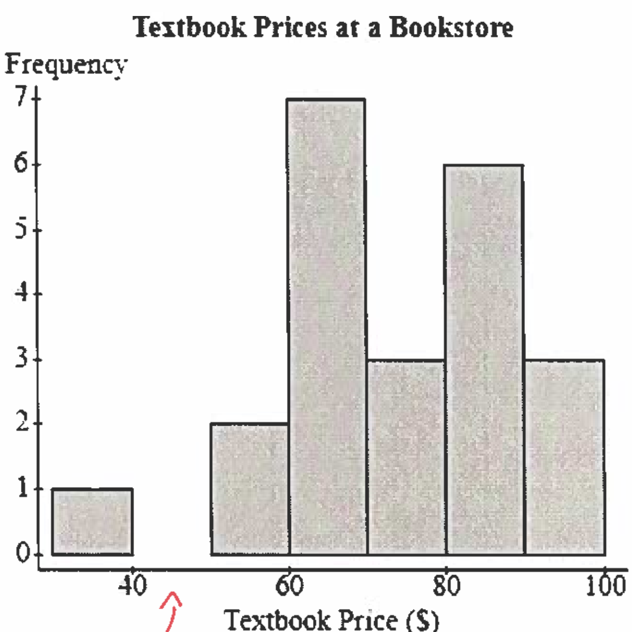
- Use between 5 and 30 bins (equal-width).
  - using too few may hide detail.
  - using too many may make unimportant features too prominent.
- Each bin is an interval of the form $[a, b)$ with "nice" boundaries.
- Compute the (relative) frequency of each bin. integer or 1 decimal place
- Mark the bins on a horizontal axis and the (relative) frequency on a vertical axis.
- Draw a bar above each bin such that the height of the bar is proportional to the (relative) frequency of the bin.

Example: Prices of Textbooks in \$: $n = 22$

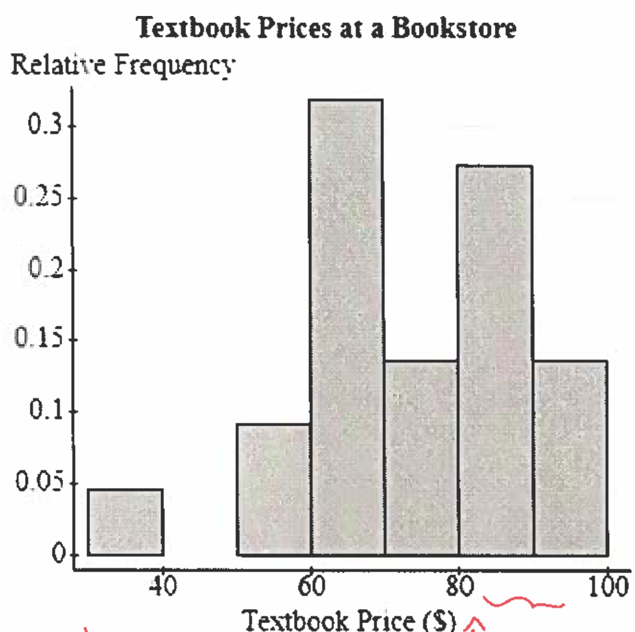
30 55 57 60 60 60 60 62 65 68 72
75 76 81 82 84 86 86 89 93 94 94

	Frequency	Relative Frequency
[30, 40)	1	0.05
[40, 50)	0	0
[50, 60)	2	0.09
[60, 70)	7	0.32
[70, 80)	3	0.14
[80, 90)	6	0.27
[90, 100)	3	0.14
Total	22	1

$\frac{1}{22} \simeq 0.05$



gap = no values



proportion of textbook prices below \$60 is about 14%

proportion of textbook prices at or above \$80 is about 41%

Shape

To describe the shape of a histogram / distribution, we will consider:

a) number of modes

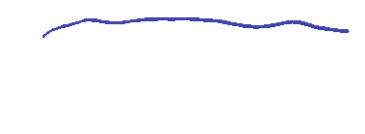
b) symmetry

c) unusual values / deviations from overall pattern: outliers and gaps

A **mode** is a hump or local high point in the shape of the distribution of a variable.

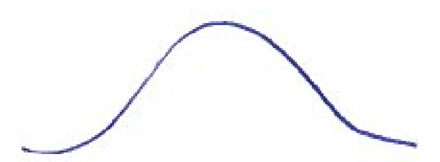
a) Number of Modes: a distribution / histogram is said to be

- **uniform**, if it appears flat, without any clear modes.



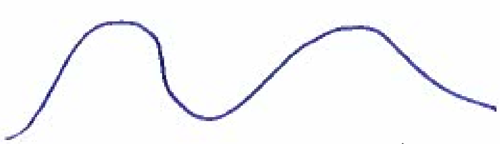
- flipping a coin

- **unimodal**, if it has one mode.



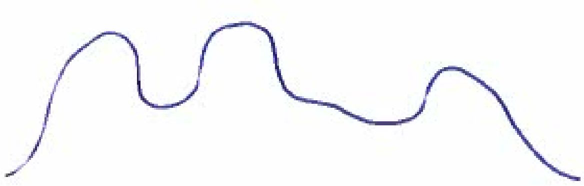
- birth weight

- **bimodal**, if it has two modes.



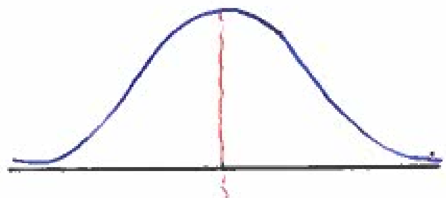
- may indicate a mixture of 2 distinct groups in data set
- hair length

- **multimodal**, if it has more than two modes.

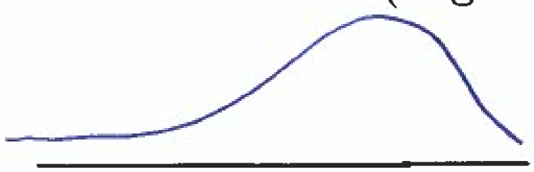


b) Symmetry: a distribution / histogram is said to be

- **symmetric** if the two halves on either side of the "centre" resemble mirror images of each other.

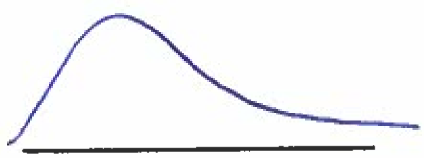


- **skewed** if it is not symmetric and one tail stretches out farther than the other.
  - **left-skewed (negatively skewed):** tail stretches to the left.



- ages of patients with heart disease

  - **right-skewed (positively skewed):** tail stretches to the right.



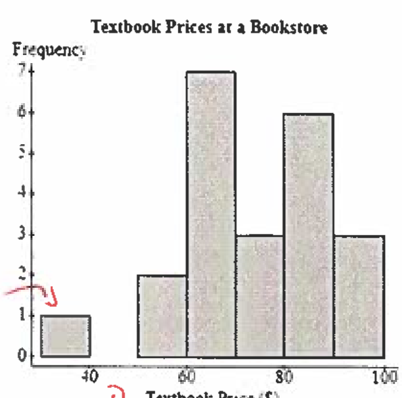
- income amounts

c) Deviations from Overall Pattern:

An **outlier** is a value that falls well above or well below the bulk of the data. It falls outside the overall pattern of the data.

- might be an error.
- might be an unusual value to investigate.

A **gap** is a region of the distribution / histogram where there are no values.



→ may indicate multiple modes corr. to distinct groups

potential outlier?

gap

Describing Quantitative Variables with Numbers

Centre

Measures of **centre** for data sets of quantitative variables:

a) the mean "typical" or "representative" value

b) the median

c) the mode - split data in "half"

Notation:

variable of interest: y

sample size / number of observations of the variable y: n

i^{th} observation of the variable y: y_i

y_1 1st observation

y_{15} 15th observation

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + y_3 + \cdots + y_n$$

Example: $y_1 = 4$, $y_2 = 1$, $y_3 = 5$, $y_4 = 2$ $n = 4$

$$\sum_{i=1}^{4} y_i = y_1 + y_2 + y_3 + y_4$$

$$= 4 + 1 + 5 + 2$$

$$= 12$$

a) **The Mean:** the mean of a set of observations of a quantitative variable is the sum of the observations divided by the number of observations.

- balancing point - Same units as data

Formula: If you have a sample of n observations $y_1, y_2, \ldots, y_n$, then the mean $\bar{y}$ of these values is:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$n = 7$

Example: Find the mean of the set of observations $\{1, 1, 2, 3, 5, 8, 13\}$.

$$\bar{y} = \frac{1+1+2+3+5+8+13}{7} = \frac{33}{7} \simeq 4.7$$

If $y_1, y_2, \ldots, y_k$ are the **distinct** observations of y and their respective frequencies are $f_1, f_2, \ldots f_k$, then the sample mean can be computed using the formula

$$\bar{y} = \frac{\sum_{i=1}^{k} f_i y_i}{n}$$

Note: $\sum_{i=1}^{k} f_i = n$

Example: Number of siblings of students in this class

# of Siblings	0	1	2	3	4	5	6	7	Total
Frequency	17	78	37	10	7	2	0	1	152

← right-skewed → good choice?

$$\bar{y} = \frac{17(0) + 78(1) + 37(2) + 10(3) + 7(4) + 2(5) + 0(6) + 1(7)}{152}$$

$$= \frac{78 + 74 + 30 + 28 + 10 + 7}{152} = \frac{227}{152} \simeq 1.49$$

b) The Median: the median of a set of observations of a quantitative variable is the middle value (midpoint) when the observations are ordered from smallest to largest.

(ascending order) — Same units as data

Note: one half of the data lie at or below the median and one half of the data lie at or above the median.

To compute the median:

1) Order the observations from smallest to largest.

2) Determine if n is odd or even:

 – if n is odd, the median is the value in position $\frac{n+1}{2}$.

 – if n is even, the median is the average of the values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

Example: # Siblings ex. median = 1

The median of $\{\underbrace{1, 1, 2}_{3}, 3, \underbrace{5, 8, 13}_{3}\}$ is 3

$n = 7$ $\frac{7+1}{2} = 4$

The median of $\{1, 1, 2, 3, 5, 8, 13, 21\}$ is 4

$n = 8$ $\frac{8}{2} = 4$ $\frac{3+5}{2} = \frac{8}{2} = 4$

c) The Mode: the mode of a set of observations of a quantitative variable is the value in the set that occur with the highest frequency.

or categorical

- if no value occurs more than once, then the data set has no mode.
- a data set may have more than one mode.

Siblings example. mode = 1

Example:

The mode of $\{\underbrace{1,1}_{2},\underbrace{2,2,2}_{3},\underbrace{3,3}_{2},\underbrace{4}_{1}\}$ is 2

The modes of $\{\underbrace{15}_{1},\underbrace{17,17}_{2},\underbrace{18}_{1},\underbrace{20}_{1},\underbrace{21,21}_{2},\underbrace{22}_{1}\}$ are 17 and 21

The set $\{\underbrace{100}_{1},\underbrace{102}_{1},\underbrace{104}_{1},\underbrace{110}_{1}\}$ has no mode.

Spread

Measures of **spread** for data sets of quantitative variables:

-variation

a) range

b) standard deviation / variance

c) interquartile range (IQR)

a) Range: the difference between the largest (maximum) observation and the smallest (minimum) observation.

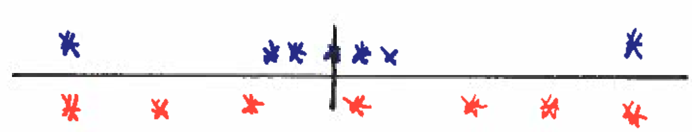
$$\text{Range} = \text{Max} - \text{Min}$$

← not an interval

Example: The range of $\{\underline{5}, 8, 8, 10, 15, 19, 25, \underline{31}\}$ is 26

$31 - 5 = 26$

1st Sample



2nd sample

b) Standard Deviation / Variance

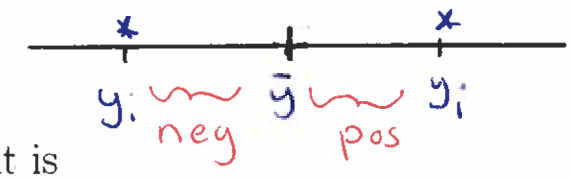
The **deviation** of an observation y_i from the sample mean $\bar{y}$ is

Sample size n

$$y_i - \bar{y}$$

n deviations: $y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}$

- If $y_i \geq \bar{y}$, then $y_i - \bar{y} \geq 0$.
- If $y_i \leq \bar{y}$, then $y_i - \bar{y} \leq 0$.
- The sum of all deviations is 0, that is



$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

(Sample)

The **variance** of a data set, denoted s^2, is the "average" of the squared deviations:

- in squared units
- want same units as data

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

← if use n, s^2 underestimates σ^2

(Sample)

The **standard deviation** of a data set, denoted s, is the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Example: $\{1, 3, 3, 4, 5, 8\}$ $n = 6$

$$\bar{y} = \frac{1+3+3+4+5+8}{6} = \frac{24}{6} = 4$$

$$s = \sqrt{\frac{(1-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (8-4)^2}{5}}$$

$$= \sqrt{\frac{9+1+1+0+1+16}{5}} = \sqrt{\frac{28}{5}} \approx 2.37$$

$\therefore$ $s \approx 2.37$

Note: Facts about standard deviation:

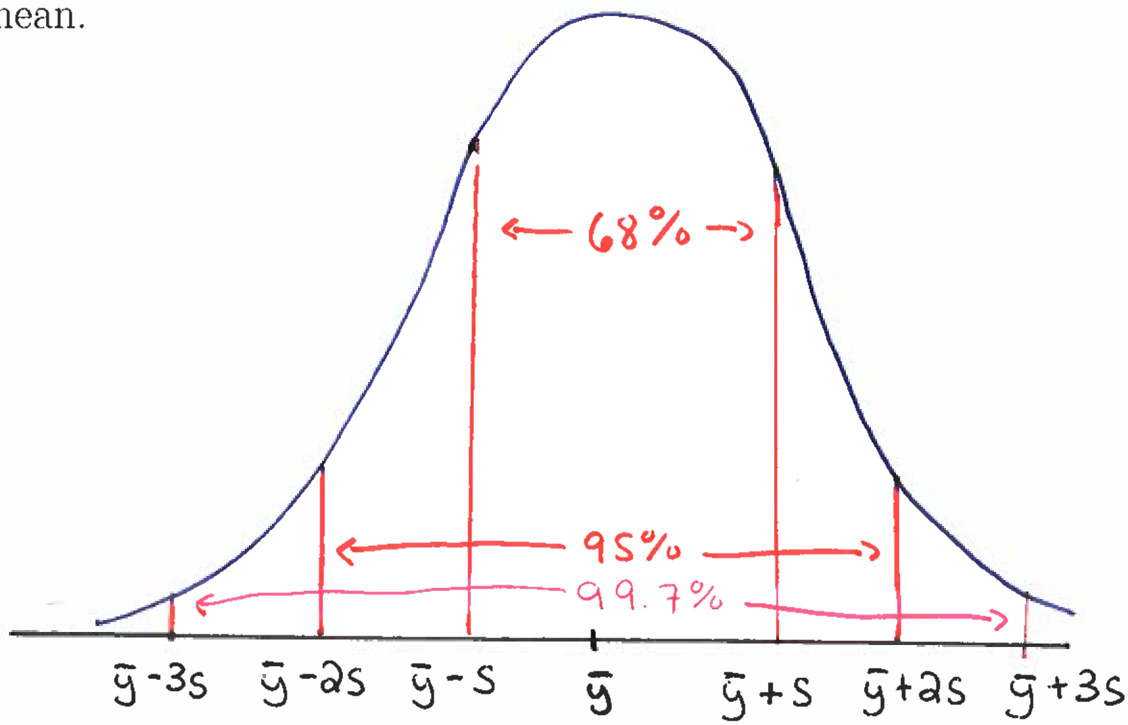
- It has the same units as the data.
- Both $\bar{y}$ and s are sample statistics, since they are computed from sample data (not population data).
- It represents the size of a "typical" deviation from the mean.
- It indicates how closely the observations in a data set are gathered around the mean.
- $s \geq 0$
- $s = 0$ if and only if all observations are the same (there is no spread/variation in the data).
- The larger the value of s, the more spread out the data.

- Symmetric
- Shaped like a bell

Empirical Rule

In a bell-shaped distribution, the mean $\bar{y}$ lies in the centre of the bell and approximately:

- 68% of the data fall within one standard deviation from the mean.
- 95% of the data fall within two standard deviations from the mean.
- almost all of the data fall within three standard deviations from the mean.



Note: The Empirical Rules gives a method of identifying outliers. If a data value lies more than three standard deviations from the mean, it can be considered as an outlier. ↳ above or below