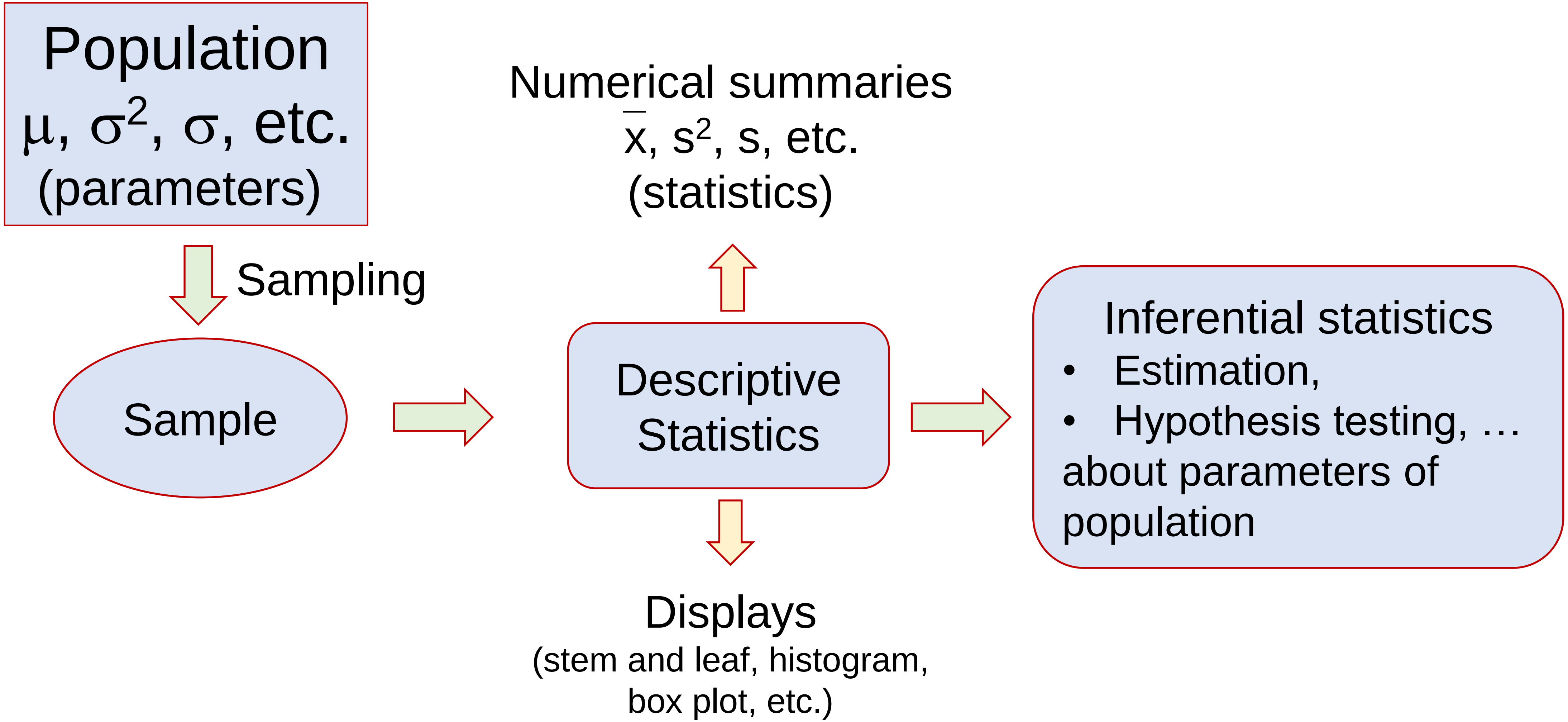
6 Descriptive Statistics

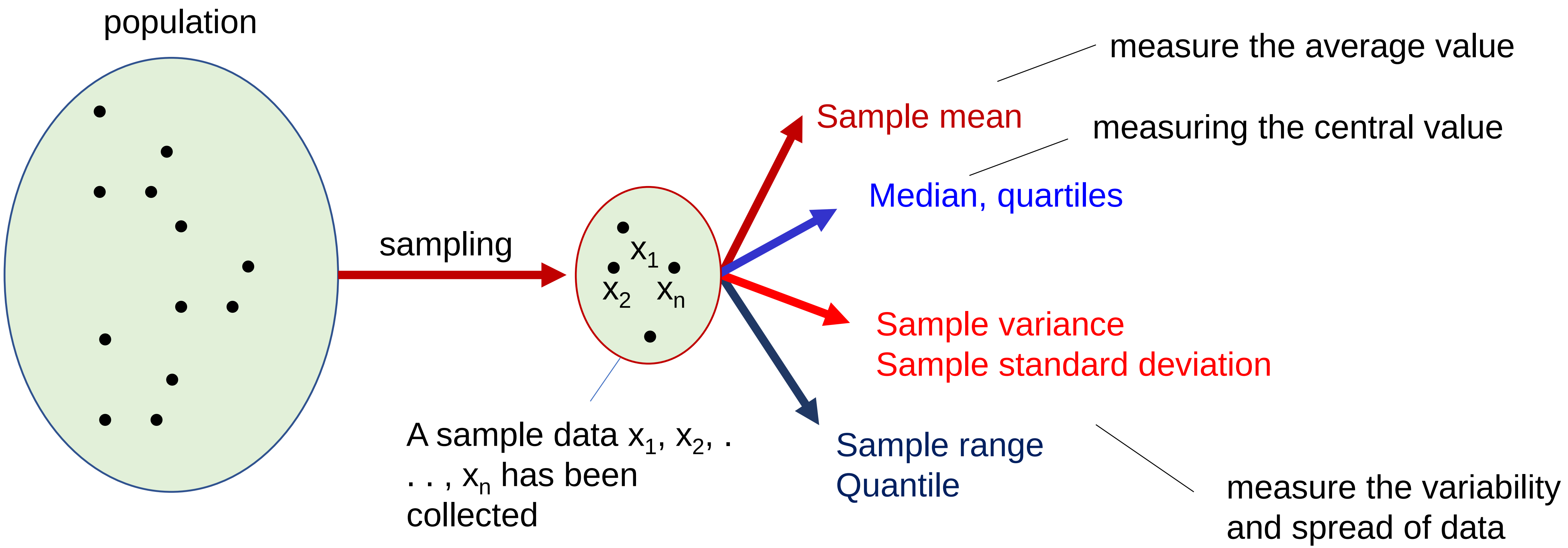
LO

- Find numeric summaries of data; describe data using stem-and-leaf diagrams, frequency distributions, histograms, time series plots.
- Determine quartiles and construct the box plot for a data; identify outliers.

Introduction



Simple descriptive statistics



Sample Mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

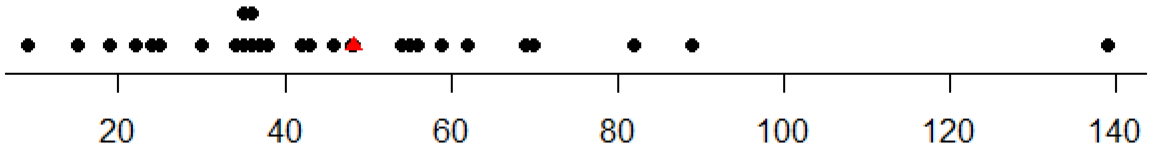
Sample Mean

$\bar{x}$ is a reasonable estimate of population mean μ

the location or central tendency in the data

does not convey all of the information about a sample of data

can be thought of as a "balance point" in a dot diagram



Ex. To evaluate effectiveness of a processor for a certain type of tasks, CPU time for n = 30 randomly chosen jobs (in seconds) were recorded:
70 36 43 69 82 48 34 62 35 15 59 139 46 37 42 30 55 56 36 82 38 89 54 25 35 24 22 9 56 19.
Then, the sample mean is $\bar{x} = 48.2333$

Ex

(Ex.6-38) The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data follow. Calculate the *sample mean*.

62 64 61 67 65 68 61 65 60 65 64 63 59 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 69 65 67 67 65 63 64 67 65

---

n = 37

The sample mean is $\bar{x} = (\sum x_i)/n = 2411/37 = 65.16$

Sample variance

- The *Sample variance* is given by

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}$$

- Ex. (CPU time, x_i)

70 36 43 69 82 48 34 62 35 15 59 139 46 37 42 30 55 56 36 82 38 89 54 25 35 24 22 9 56 19.

We have $n = 30$ and the sample mean is

$\bar{x} = 48.2333$,

➔ The *sample variance* $s^2 = 703.1476$ (sec^2)

➔ The *sample standard deviation* is

$s = 26.52$ sec

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	70	21.77	473.93
2	36	-12.23	149.57
3	43	-5.23	27.35
4	69	20.77	431.39
5	82	33.77	1140.41
6	48	-0.23	0.05
		⋮	
25	35	-13.23	175.03
26	24	-24.23	587.09
27	22	-26.23	688.01
28	9	-39.23	1538.99
29	56	7.77	60.37
30	19	-29.23	854.39
	Σx_i	$\Sigma(x_i - \bar{x})=0$	$\Sigma(x_i - \bar{x})^2$

Sample Variance - Note

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2\bar{x}x_i)}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i}{n-1}$$

and since $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, this last equation reduces to

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

Exercise

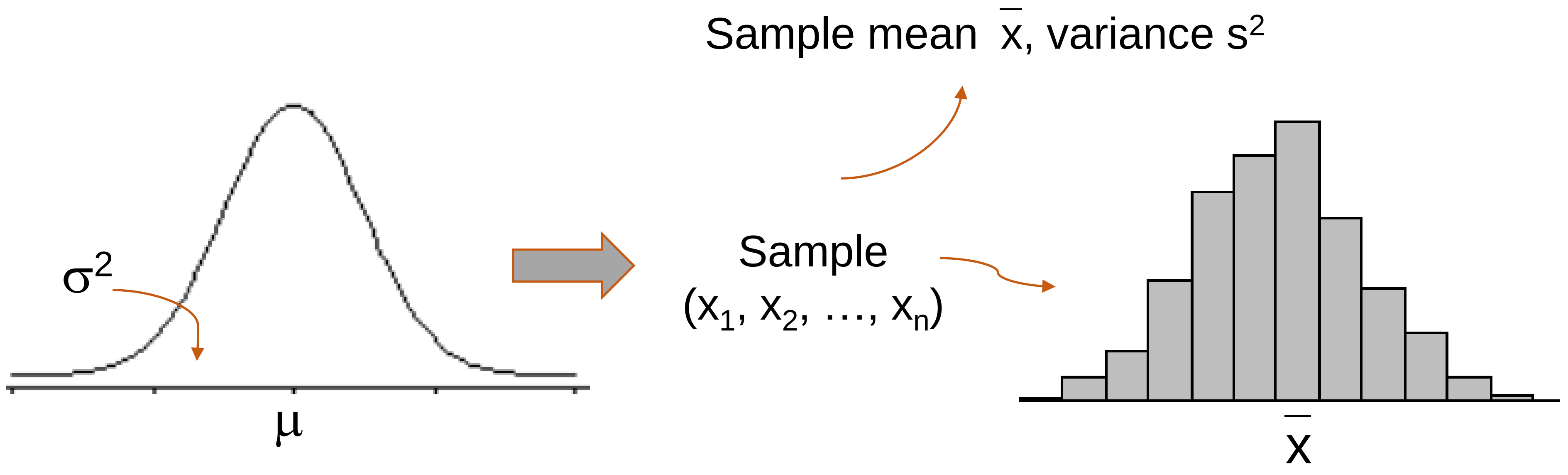
(Ex.6-38) The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data follow. Calculate the *sample variance* and standard deviation.

62 64 61 67 65 68 61 65 60 65 64 63 59 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 69 65 67 67 65 63 64 67 65

The sample variance is $s^2 = 6.47$

The standard deviation is $s = 2.54$

Relationship between a population and a sample



Probability distribution of Population

Histogram of sample

Sample Variance vs Population Variance

The sample variance

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

The population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Numerical summaries – Ex

A network provider investigates the load of its network. The number of concurrent users is recorded at ten locations (thousands of people),

17.2, 22.1, 18.5, 17.2, 18.6, 14.8, 21.7, 15.8, 16.3, 22.8

Compute the sample mean, variance, and standard deviation of the number of concurrent users.

\---

Sample mean: $\bar{x} = 18.5$ (thousands of people)

Sample variance $s^2 = 7.88$

Sample standard deviation: $s = 2.81$ (thousands of people)

Sample median - Example

Ex. (Median CPU time).

Since n = 30, (n + 1)/2 = 15.5 ➔ consider the 15-th smallest and 16-th smallest observations.

➔ (42 + 43)/2 = 42.5 the *sample median*

Sort the data:

1st ↓ … 15th ↓ 16th ↓ … 30th ↓

9 15 19 22 24 25 30 34 35 35 36 36 37 38 42 43 46 48 54 55 56 56 59 62 69 70 82 82 89 139

Exercise

(Ex.6-38) The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The data follow. Calculate the *sample median* of height.

62 64 61 67 65 68 61 65 60 65 64 63 59 68 64 66 68 69 65 67 62 66 68 67 66 65 69 65 69 65 67 67 65 63 64 67 65

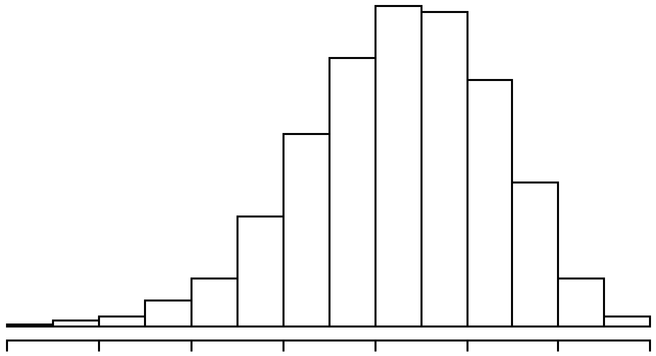
Sort the data

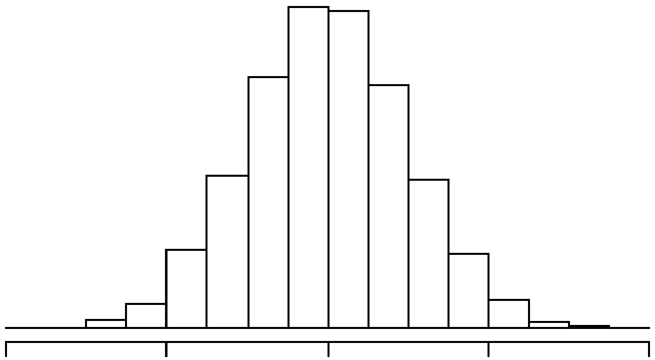
59 60 61 61 62 62 63 63 64 64 64 64 65 65 65 65 65 65 65 65 65 66 66 66 67 67 67 67 67 67 68 68 68 68 69 69 69

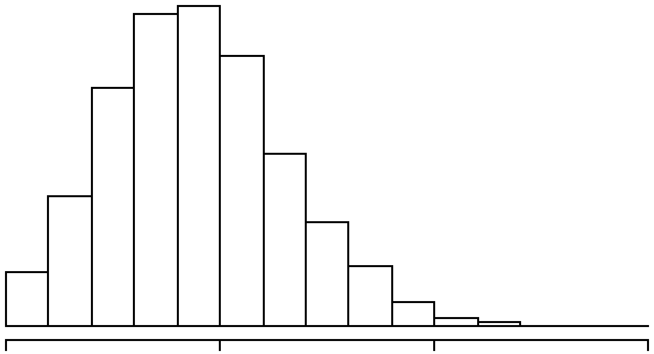
n = 37 is odd and (n + 1)/2 = 19

➔ The 19^{th} smallest observation = 65 is the *sample median*

Mean vs medium







Mean < Medium
Left-skewed
Negative skew

Medium = Mean
Symmetric

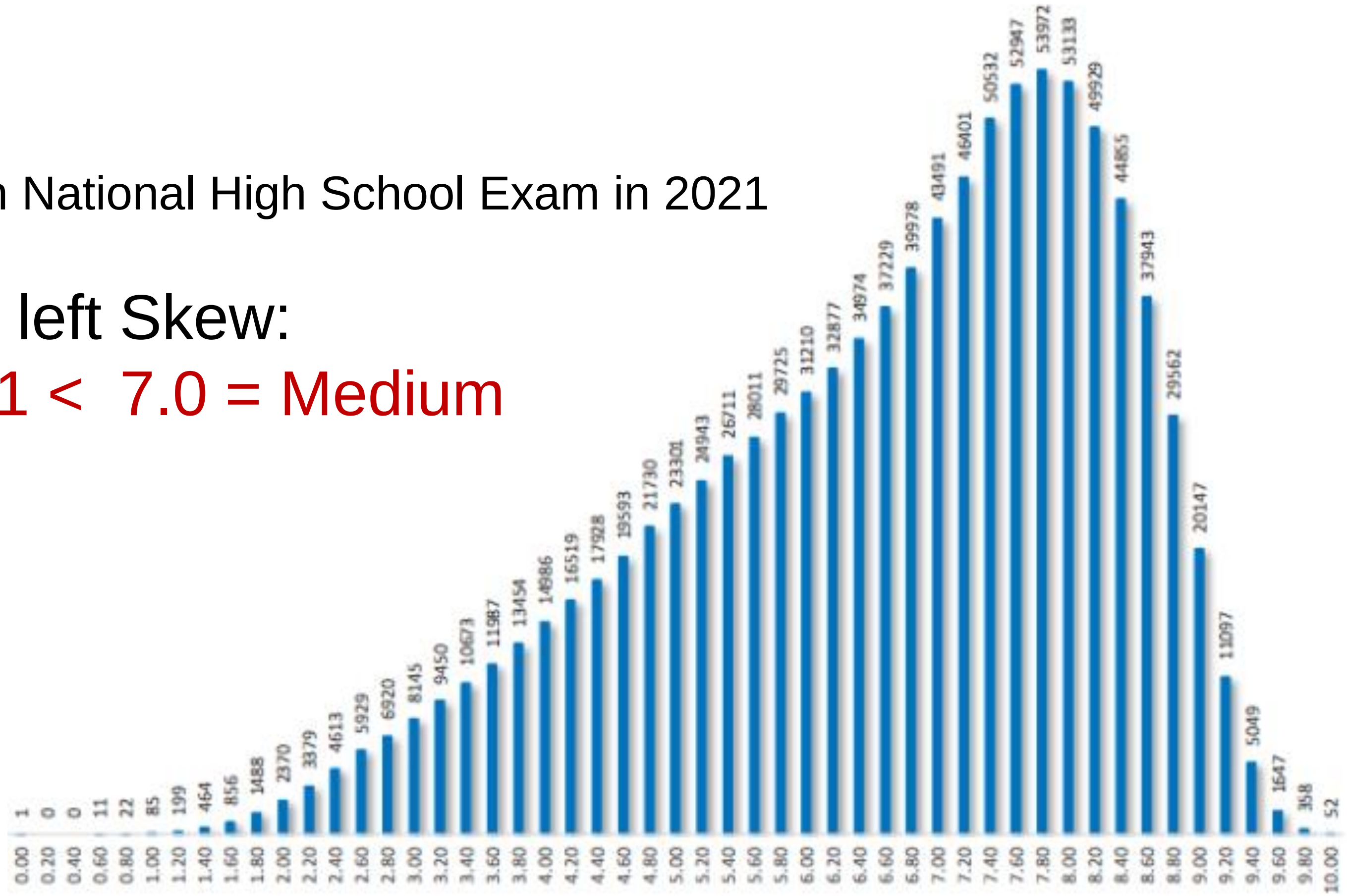
Medium < Mean
Right-skewed
Positive skew

Mean vs Medium – Ex

Math Scores on National High School Exam in 2021

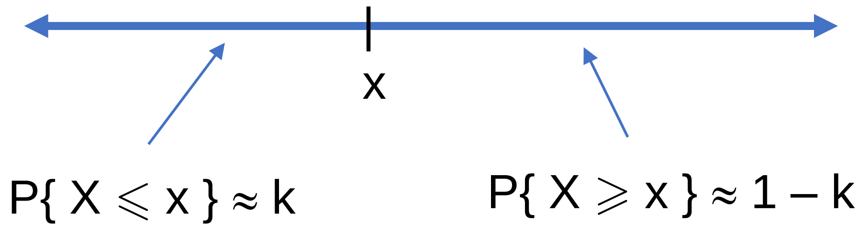
Negative or left Skew:

Mean = 6.61 < 7.0 = Medium



Quantiles, quartiles, percentiles

- A 100kth percentile is a value x such that
  - $P\{X \leqslant x\} \approx k$
  - $P\{X \geqslant x\} \approx 1 - k$



- Special cases of 100kth percentile
  - k = 0.25: first quartile, 25th percentile = q_1
  - k = 0.5: second quartile, 50th percentile = q_2 = median
  - k = 0.75: third quartile, 75th percentile = q_3
- The interquartile range, IQR = $q_3 - q_1$

Quantiles, quartiles, percentiles - Ex

```
> sort(CPUtime)
 [1]   9  15  19  22  24  25  30  34  35  35  36  36  37  38  42  43  46
[18]  48  54  55  56  56  59  62  69  70  82  82  89 139
```

1st quartile: $(n+1)/4 = 31/4 = 7.75$ ➔ between 7th and 8th observations: 33.5

3rd quartile: $3(n+1)/4 = 23.25$ ➔ between 23rd and 24th observations: 59.5

➔ $q_1 = 33.5$ and $q_3 = 59.5$

➔ $IQR = q_3 - q_1 = 59.5 - 33.5 = 26$

Quantiles – Note

```
> sort(CPUtime)
 [1]   9  15  19  22  24  25  30  34  35  35  36  36  37  38  42  43  46
[18]  48  54  55  56  56  59  62  69  70  82  82  89 139
```

```
> quantile(CPUtime, probs = c(0, 0.1,0.25, 0.3, 0.5,0.75,0.95, 1), type=1)
  0%    10%    25%    30%    50%    75%    95%    100%
   9     19     34     35     42     59     89     139
```

```
> quantile(CPUtime, probs = c(0, 0.1,0.25, 0.3, 0.5,0.75,0.95, 1), type=7)
   0%  10%    25%    30%    50%    75%    95%   100%
 9.00  21.70  34.25  35.00  42.50  58.25  85.85 139.00
```

Exercise

(Ex.6-38) The female students in an undergraduate engineering core course at ASU self-reported their heights to the nearest inch. The sorted data follow. Calculate the quartiles q_1, q_2, q_3.

59 60 61 61 62 62 63 63 64 64 64 64 65 65 65 65 65 65 65 65 65 66 66 66 67 67 67 67 67 67 68 68 68 68 69 69 69

---

n = 37, (n+1)/4 = 9.5, 2(n+1)/4 = 19, 3(n+1)/4 = 28.5

- $q_1 = 64$
- $q_3 = 67$

Sample range

- Sample range $r = \max(x_i) - \min(x_i)$

> sort(CPUtime)

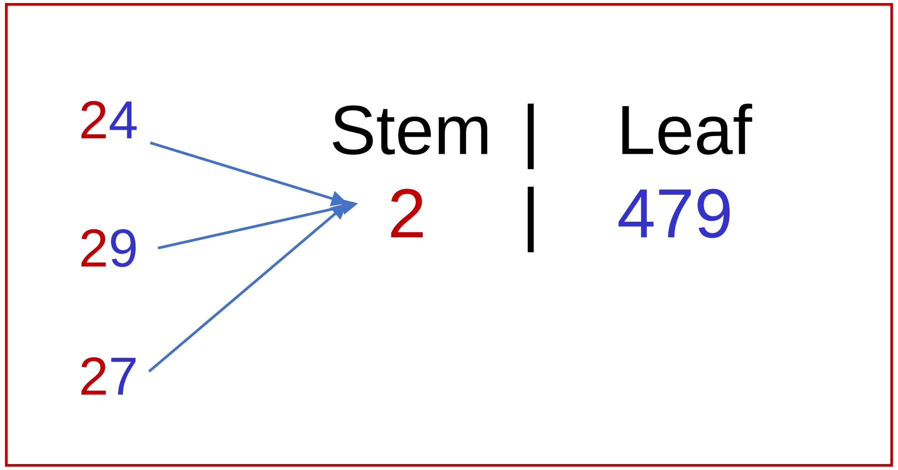
[1]   9 15 19 22 24 25 30 34 35 35 36 36 37 38 42 43 46

[18] 48 54 55 56 56 59 62 69 70 82 82 89 139

➔ $r = \max(x_i) - \min(x_i) = 139 - 9 = 130$

Stem-and-leaf diagrams

- Each number in the data point $x_1, x_2, \ldots, x_n$ consists of at least two digits
- Steps to construct stem-and leaf diagram
  - (1) divide each x_i into two parts: a stem: one or more of leading digits; and a leaf: the remaining digit
  - (2) List the stem values in a vertical
  - (3) Record the leaf beside its stem
  - (4) Write the units for stems and leaves



Stem-and-leaf diagrams - Ex

(CPU time) 70 36 43 69 82 48 34 62 35 15 59 139 46 37 42 30 55 56 36 82 38 89 54 25 35 24 22 9 56 19

Stem: tens and hundreds
Leaf: Ones digits

Stem	Leaf
0	9
1	59
2	245
3	04556678
4	2368
5	45669
6	29
7	0
8	229
9	
10	
11	
12	
13	9

Stem	Leaf	Data values
2	2, 4, 5	22, 24, 25

Stem-and-leaf diagrams – Ex

```
 0 | 9
 1 | 59
 2 | 245
 3 | 04556678
 4 | 2368
 5 | 45669
 6 | 29
 7 | 0
 8 | 229
 9 |
10 |
11 |
12 |
13 | 9
```

14 stems

```
 0 | 959
 2 | 24504556678
 4 | 236845669
 6 | 290
 8 | 229
10 |
12 | 9
```

7 stems

(CPU time)
9 15 19 22 24 25 30 34 35
35 36 36 37 38 42 43 46
48 54 55 56 56 59 62 69 70
82 82 89 139

If we use too many stems in a plot, the resulting in a display that may not tell us much about the shape of the data

Stem-and-leaf diagrams - Ex

An article in Technometrics (1977, Vol. 19, p. 425) presented the following data on the *motor fuel octane ratings of several blends of gasoline*:

```
 82 | 4
 84 | 333
 86 | 777456789
 88 | 23334556679023367889 9
 90 | 011134445678900011122566 88
 92 | 2223677702334 7
 94 | 2247
 96 | 15
 98 | 8
100 | 3
```

an informative visual display

The decimal point is at |

(n = 82 observations)

83.4	84.3	84.3	85.3	86.7	86.7	86.7	87.4	87.5
87.6	87.7	87.8	87.9	88.2	88.3	88.3	88.3	88.4
88.5	88.5	88.6	88.6	88.7	88.9	89.0	89.2	89.3
89.3	89.6	89.7	89.8	89.8	89.9	89.9	90.0	90.1
90.1	90.1	90.3	90.4	90.4	90.4	90.5	90.6	90.7
90.8	90.9	91.0	91.0	91.0	91.1	91.1	91.1	91.2
91.2	91.5	91.6	91.6	91.8	91.8	92.2	92.2	92.2
92.3	92.6	92.7	92.7	92.7	93.0	93.2	93.3	93.3
93.4	93.7	94.2	94.2	94.4	94.7	96.1	96.5	98.8
100.3								

Frequency Distributions and Histograms

- **Frequency Distributions**
  - More compact than a stem-and-leaf diagram
  - The range of the data is divided into intervals (**class intervals, cells, bins)**
  - A ***frequency histogram*** consists of columns, one for each bin, whose height is determined by the *number* of observations in the bin.
  - A ***relative frequency histogram*** has the same shape but a different vertical scale. Its column heights represent the *proportion* of all data that appeared in each bin.

Histograms

- The histogram is a visual display of the frequency distribution.
- Histograms have a shape similar to the pmf or pdf of data, especially in large samples.
- Histograms are stable and reliable for *large data sets*, preferably of size 75 to 100 or more.
- **Constructing a Histogram (Equal Bin Widths)**
  - (1) Label the bin boundaries on a horizontal scale
  - (2) Mark and label the vertical scale with the (relative) frequencies
  - (3) Above each bin, draw a rectangle where height is equal to the (relative) frequency corresponding to that bin

Histogram - Ex

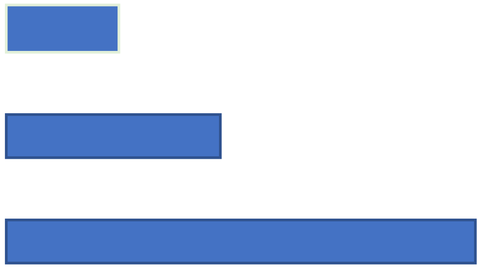
Ex. (CPU time) 70 36 43 69 82 48 34 62 35 15 59 139 46 37 42 30 55 56 36 82 38 89 54 25 35 24 22 9 56 19

Choosing intervals [0,10), [14,20), [20, 30), . . . as *bins*, we count

1 observation in [0, 10)

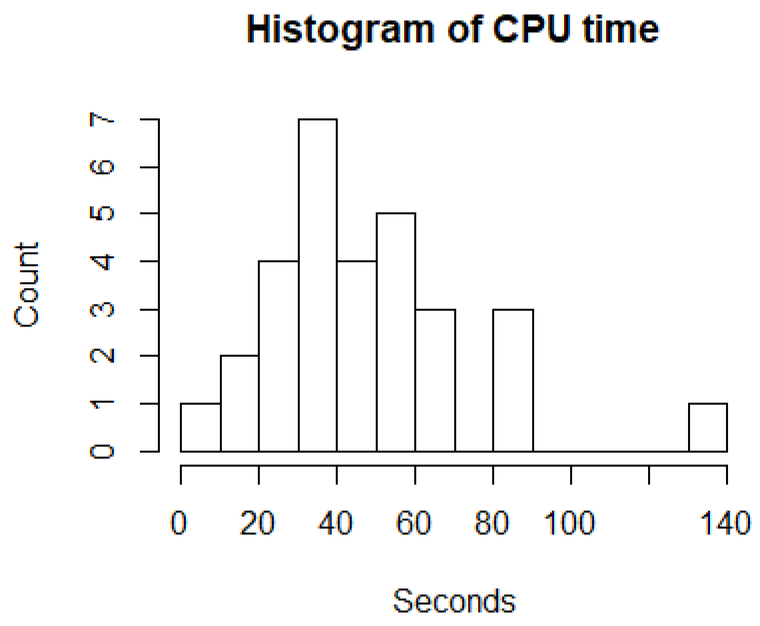
2 observations in [10, 20)

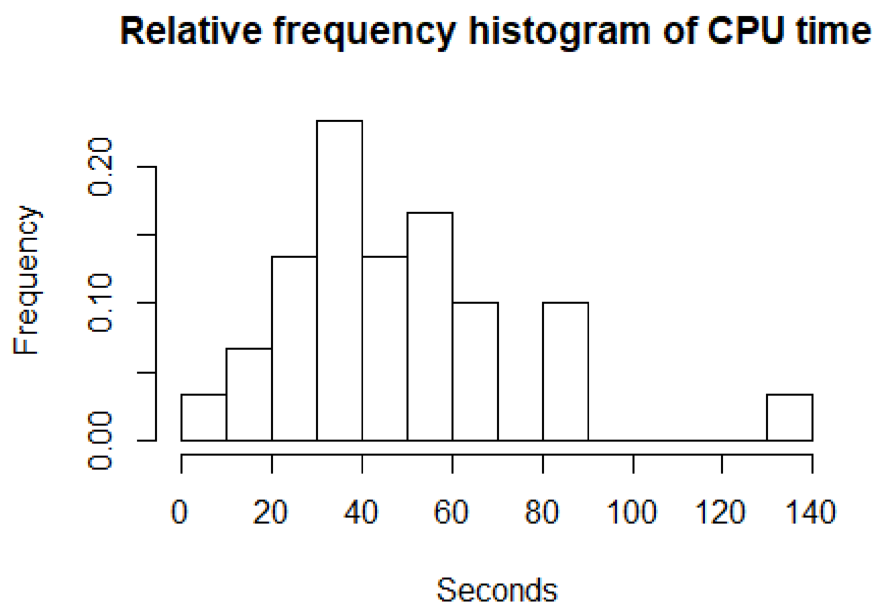
4 observations in [20, 30)



. . . .

Ex. Histogram of CPU time data

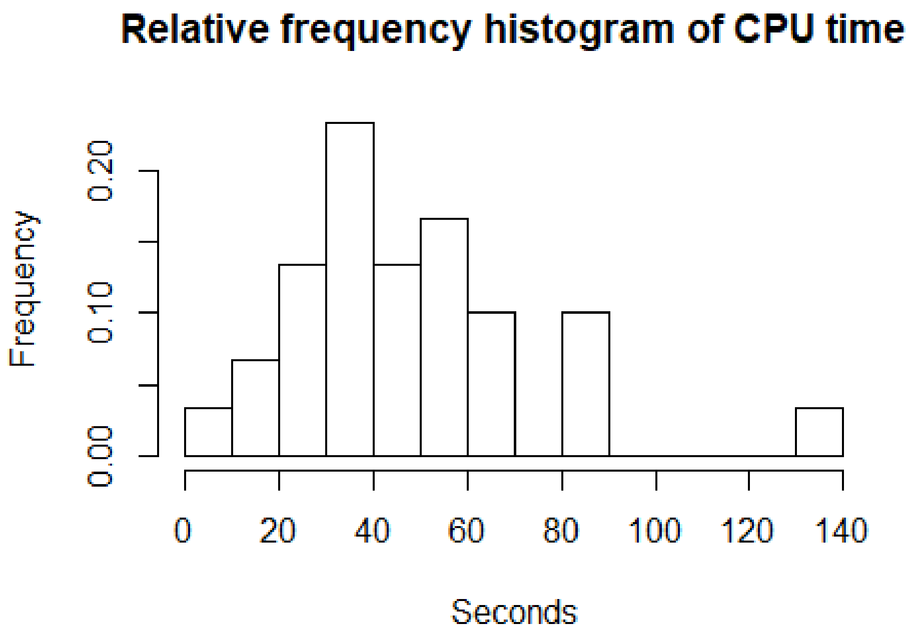


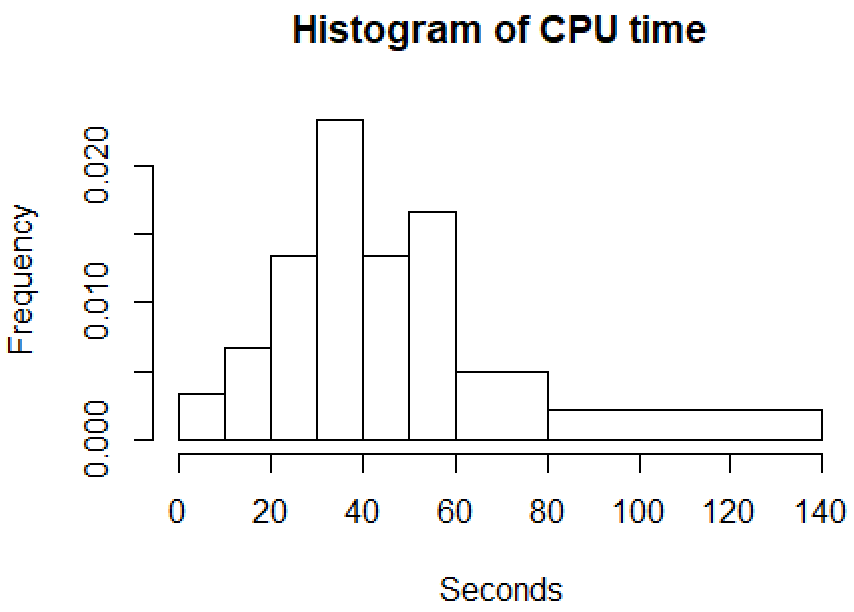


Some information can be drawn from the histograms:
- Distribution of CPU times is not symmetric; it is right-skewed as we see 5 columns to the right of the highest column and only 3 columns to the left.
- The time of 139 seconds stands alone suggesting that it is in fact an *outlier*.

Histogram

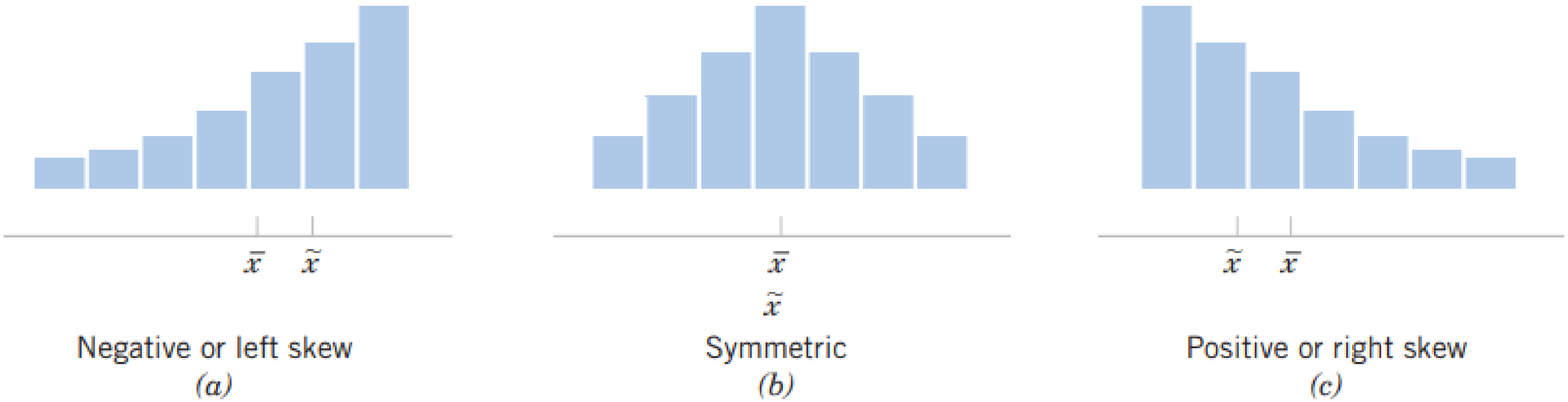
- With unequal bins, rectangular height = bin frequency/bin width





Unequal bins

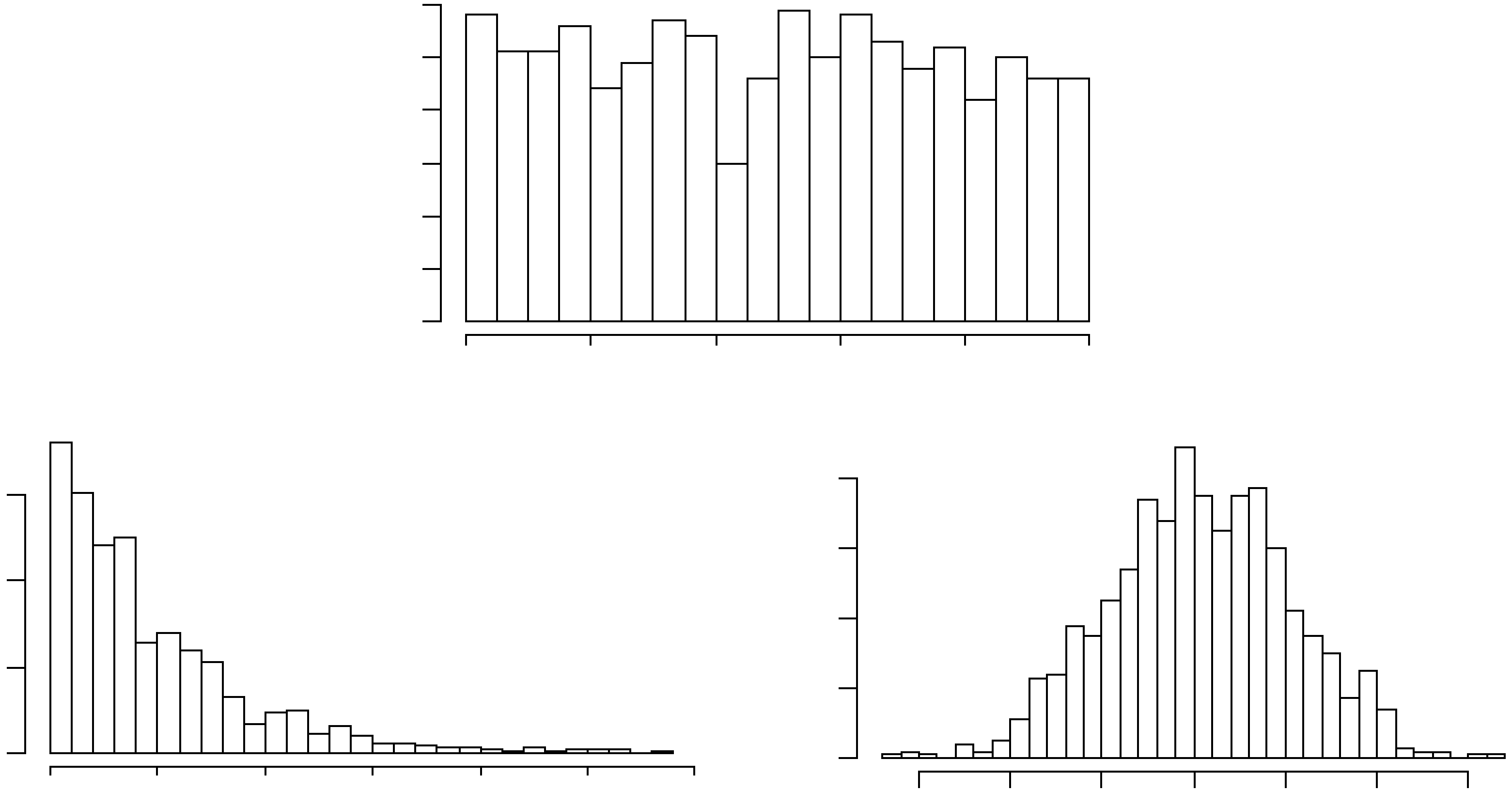
Histograms for symmetric and skewed distributions



Usually, we find that

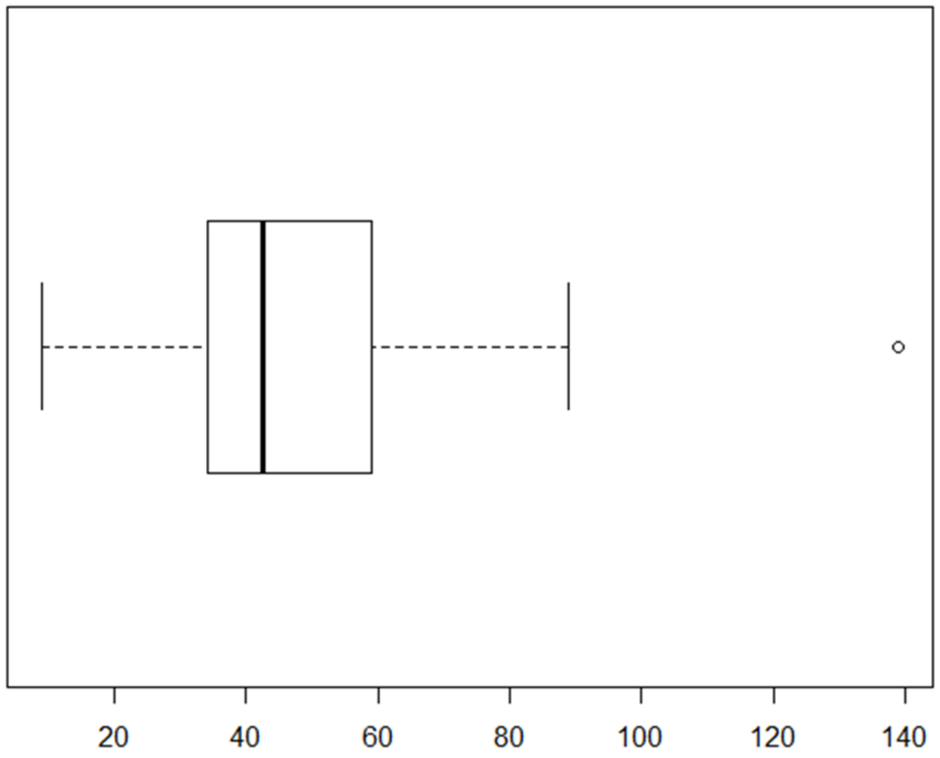
- if mode < median < mean, the distribution is right-skewed
- if mode > median > mean, the distribution is left-skewed

Histograms of various samples



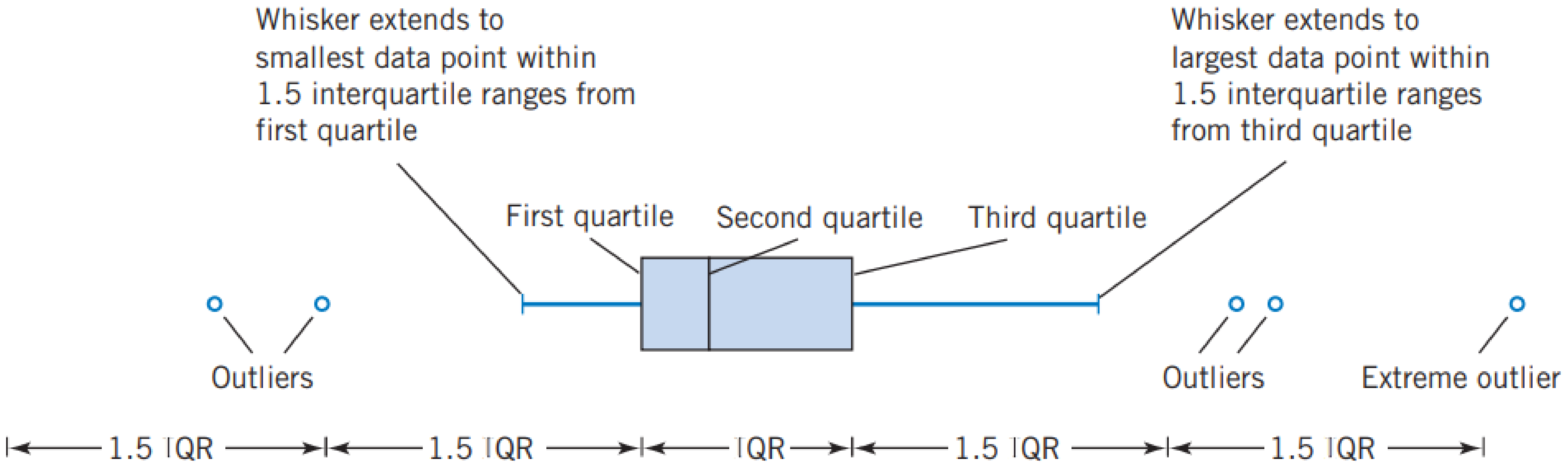
Box plot

- The box plot is a graphical display that simultaneously describes several important features of a data set:
  - Center
  - Spread
  - Symmetry
  - Outlier
  - Extreme outlier



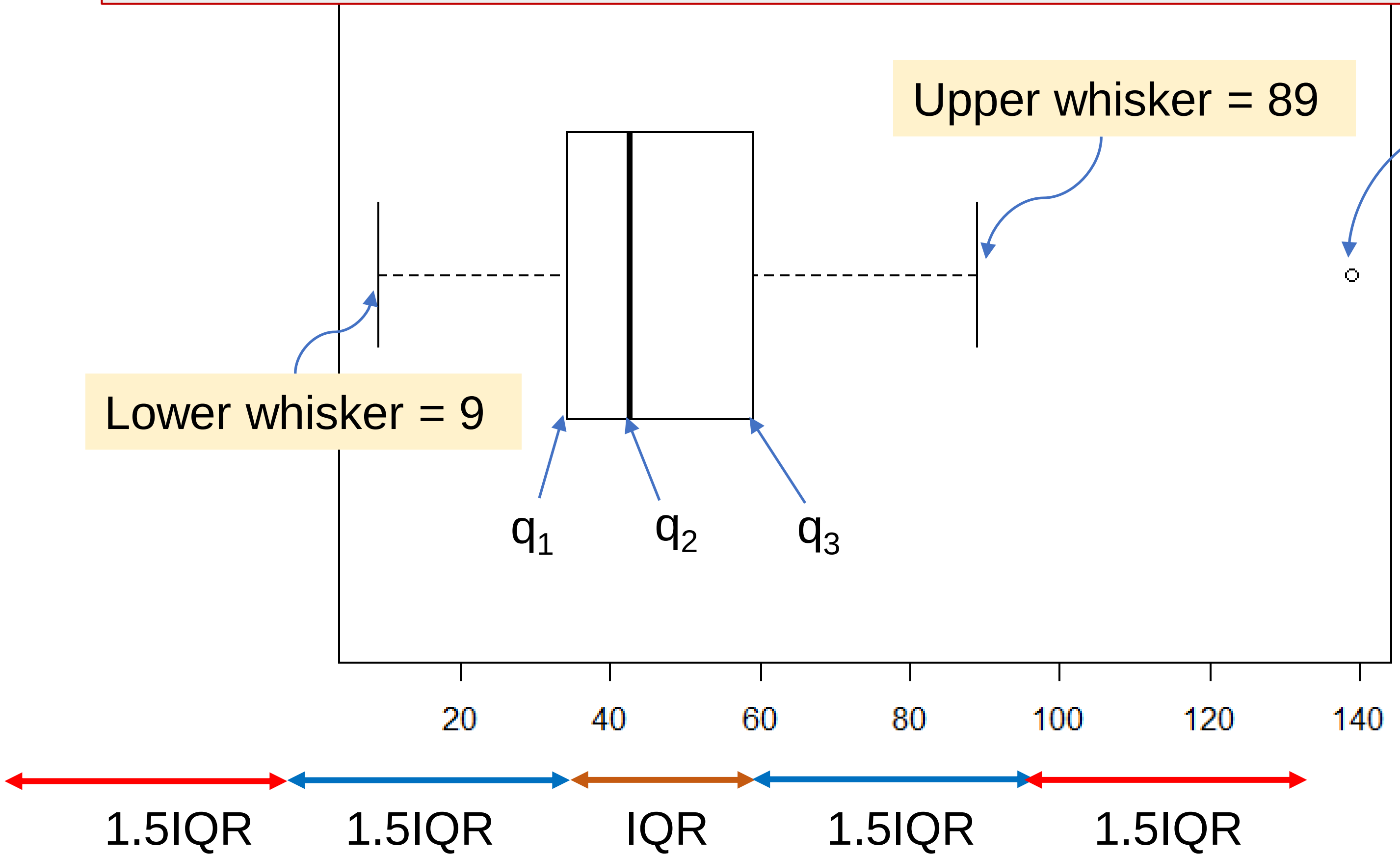
Description of a box plot

Five-point summary = $(\min(x_i), q_1, q_2, q_3, \max(x_i))$



Ex – CPU time Box plot

```
> sort(CPUtime)
[1]   9  15  19  22  24  25  30  34  35  35  36  36  37  38  42  43  46
[18]  48  54  55  56  56  59  62  69  70  82  82  89 139
```



IQR = $q_3 - q_1$ = 59.5 – 33.5 = 26

1.5IQR = 39

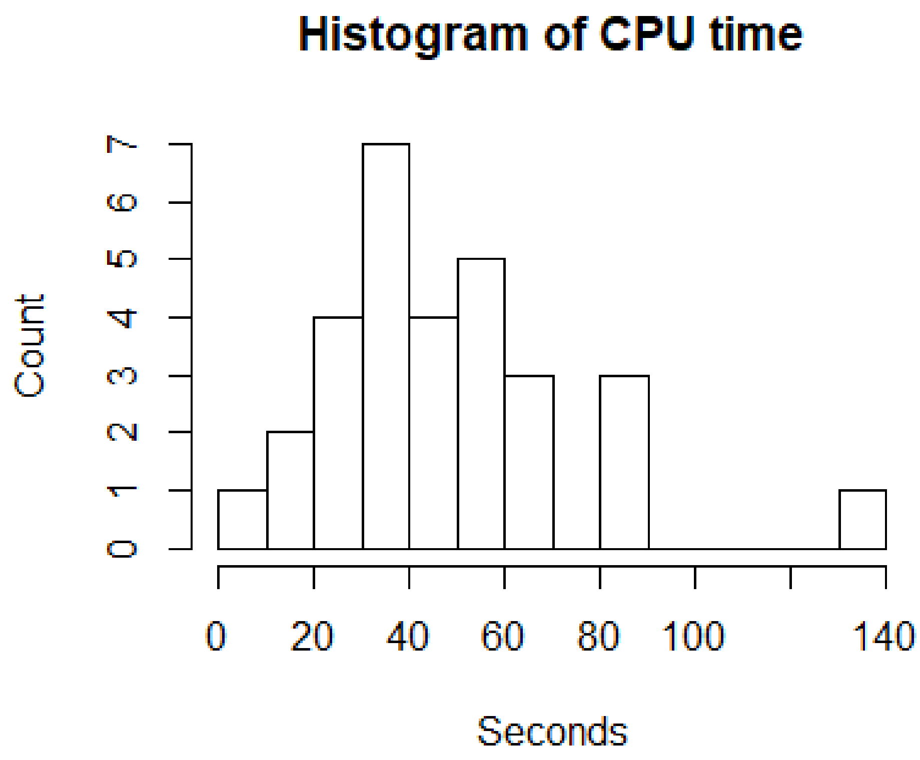
q_3 + 1.5IQR = 59.5 + 39 = 98.5

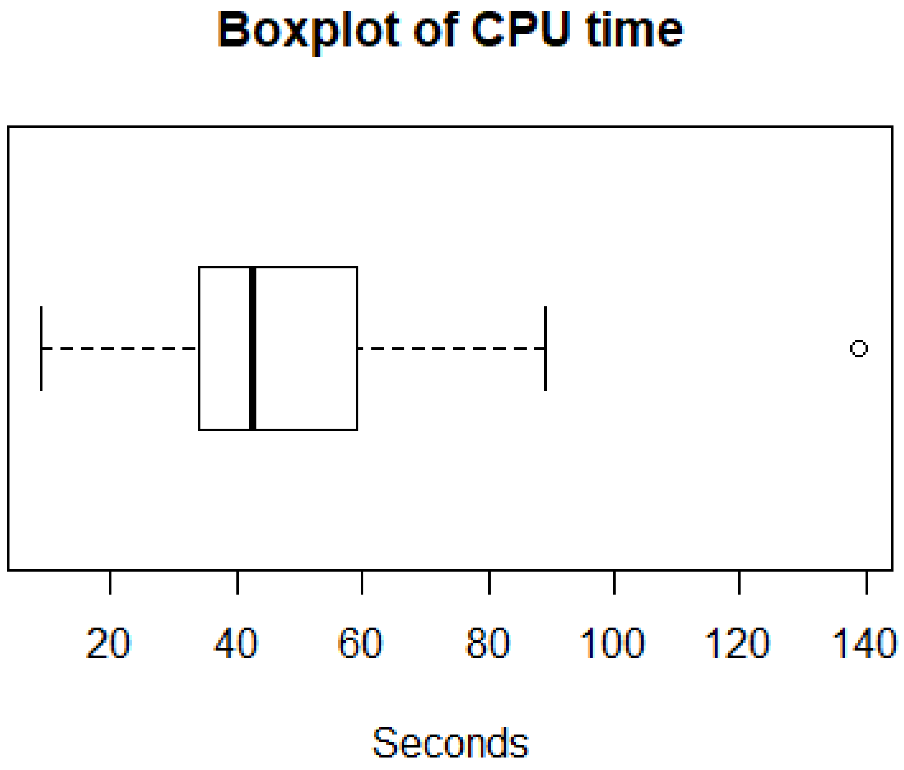
➔ Upper whisker is 89

98.5 + 1.5IQR = 137.5

➔ 139 is the *extreme outlier*

CPU time Ex – Histogram vs Boxplot

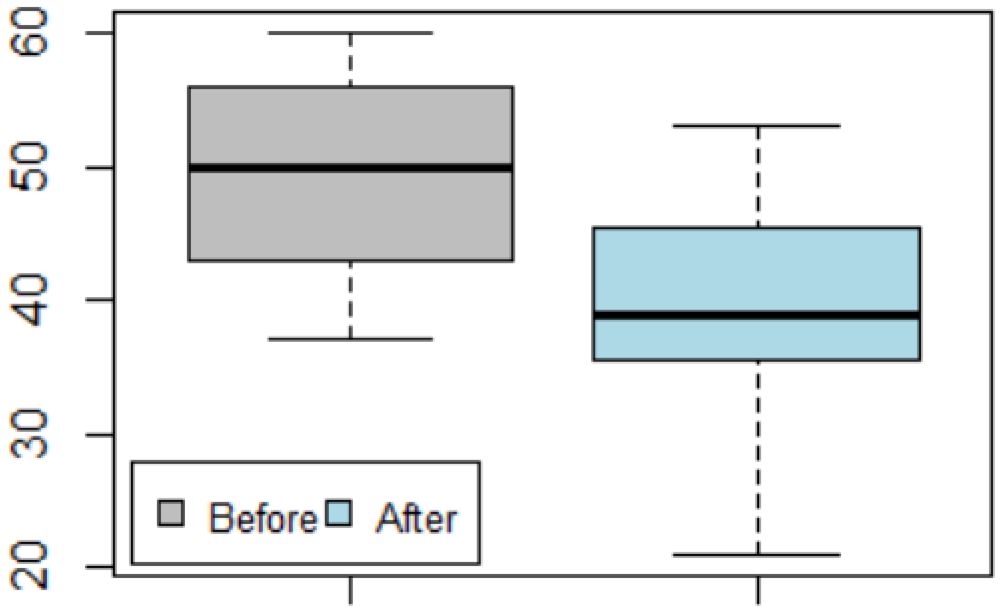




Comparative box plots

Ex. The numbers of blocked intrusion attempts on each day during the first two weeks of the month were 56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58. After the change of firewall settings, the numbers of blocked intrusions during the next 20 days were 53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37, 36, 39, 45.

- Parallel boxplots

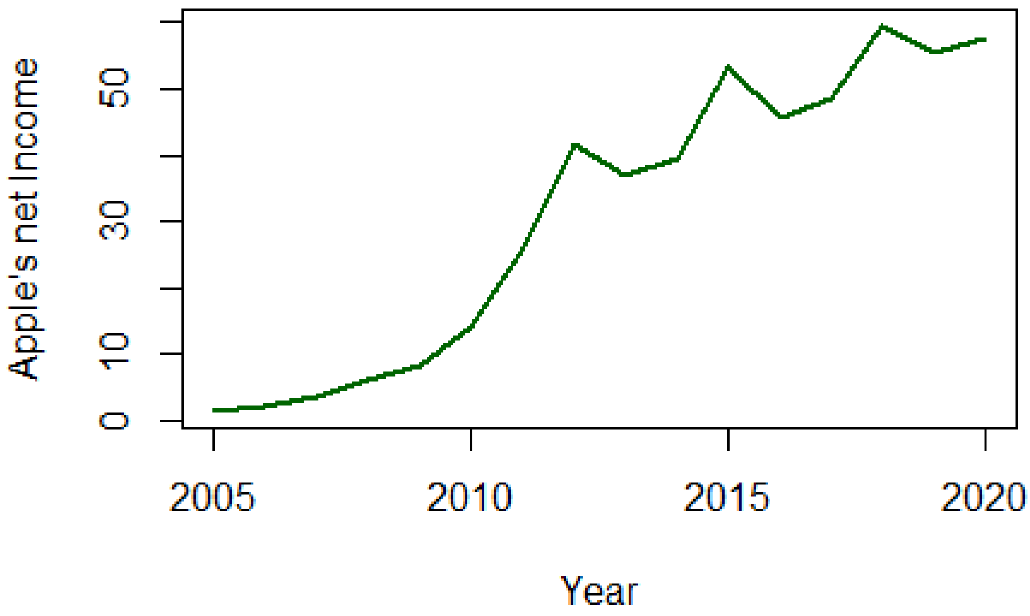


Time Sequence Plots

- A **time series** or **time sequence** is a data set in which the observations are recorded in the order in which they occur.
- A **time series plot**
  - the vertical axis denotes the observed value
  - the horizontal axis denotes the time
- In a time series plot, we often see
  - trends,
  - cycles,
  - or other broad features of the data

Ex- Apple's net Income (2005 - 2020)

Year	Income (B)
2005	1.33
2006	1.99
2007	3.5
2008	6.12
2009	8.24
2010	14.01
2011	25.92
2012	41.73
2013	37.04
2014	39.51
2015	53.39
2016	45.69
2017	48.35
2018	59.53
2019	55.26
2020	57.41



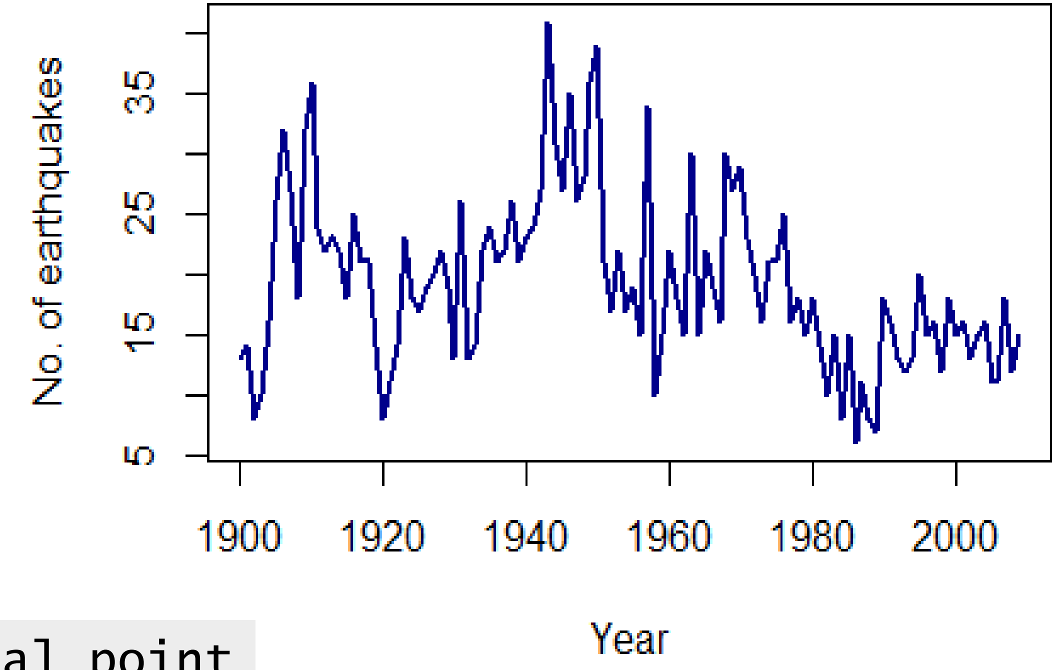
A time series plot of the annual income of Apple (2005-2020). The general impression from this display is that incomes show an *upward trend*.

The number of earthquakes per year of magnitude 7.0 and higher (1900-2009)

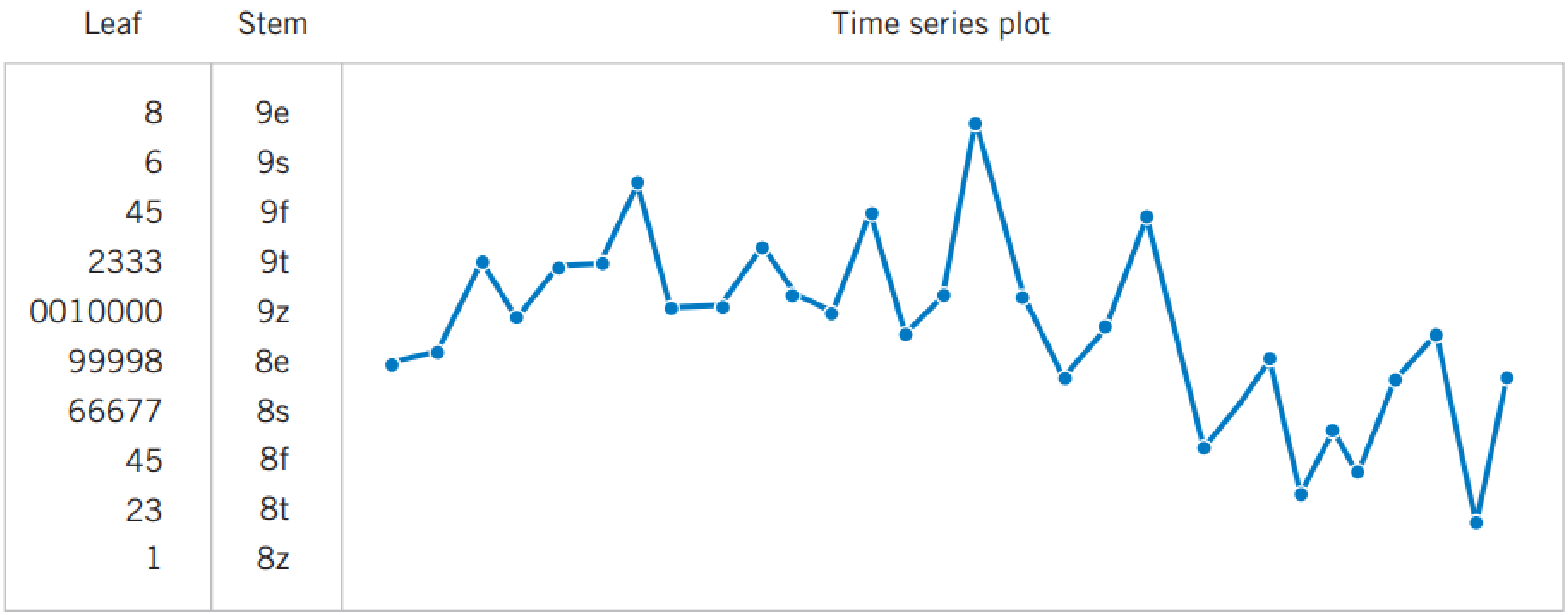
13, 14, 8, 10, 16, 26, 32, 27,
18, 32, 36, 24, 22, 23, 22,
18, 25, 21, 21, 14, 8, 11, 14,
23, 18, 17, 19, 20, 22, 19,
13, 26, 13, 14, 22, 24, 21,
22, 26, 21, 23, 24, 27, 41,
31, 27, 35, 26, 28, 36, 39,
21, 17, 22, 17, 19, 15, 34,
10, 15, 22, 18, 15, 30, 15,
22, 19, 16, 30, 27, 29, 23,
20, 16, 21, 21, 25, 16, 18,
15, 18, 14, 10, 15, 8, 15, 6,
11, 8, 7, 18, 16, 13, 12, 13,
20, 15, 16, 12, 18, 15, 16,
13, 15, 16, 11, 11, 18, 12, 15

```
 6 | 00
 8 | 0000
10 | 0000000
12 | 000000000
14 | 0000000000000000
16 | 00000000000
18 | 0000000000000
20 | 0000000000
22 | 000000000000
24 | 00000
26 | 00000000
28 | 00
30 | 000
32 | 00
34 | 00
36 | 00
38 | 0
40 | 0
```

```
The decimal point
is at the |
```



The digidot plot



Summary

- Summaries of data
  - Central
    - Mean
    - Median
    - Mode
  - Spread of data
    - Variance
    - Standard deviation
    - Range
    - Quantiles, quartiles
- Graphics
  - Stem-and-leaf diagrams,
  - Histograms
  - Box plot
  - Time series plots.

Exercises

$(13x + 17) \bmod 87$

$(11x + 21) \bmod 87$

$(9x + 29) \bmod 87$