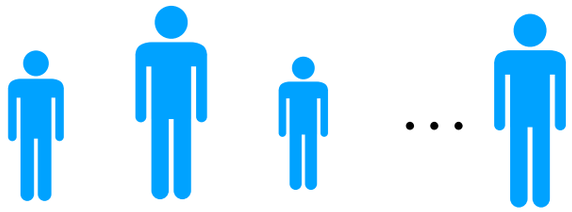
Single variable stats review

Sample statistics



h=height

$i \quad 1 \quad 2 \quad 3 \quad \dots \quad N$

$h_i \quad (h_1 \quad h_2 \quad h_3 \quad h_N) = \mathbf{h}^T$

matrix transpose --> rows become columns

ft **row vector** $1 \times N$

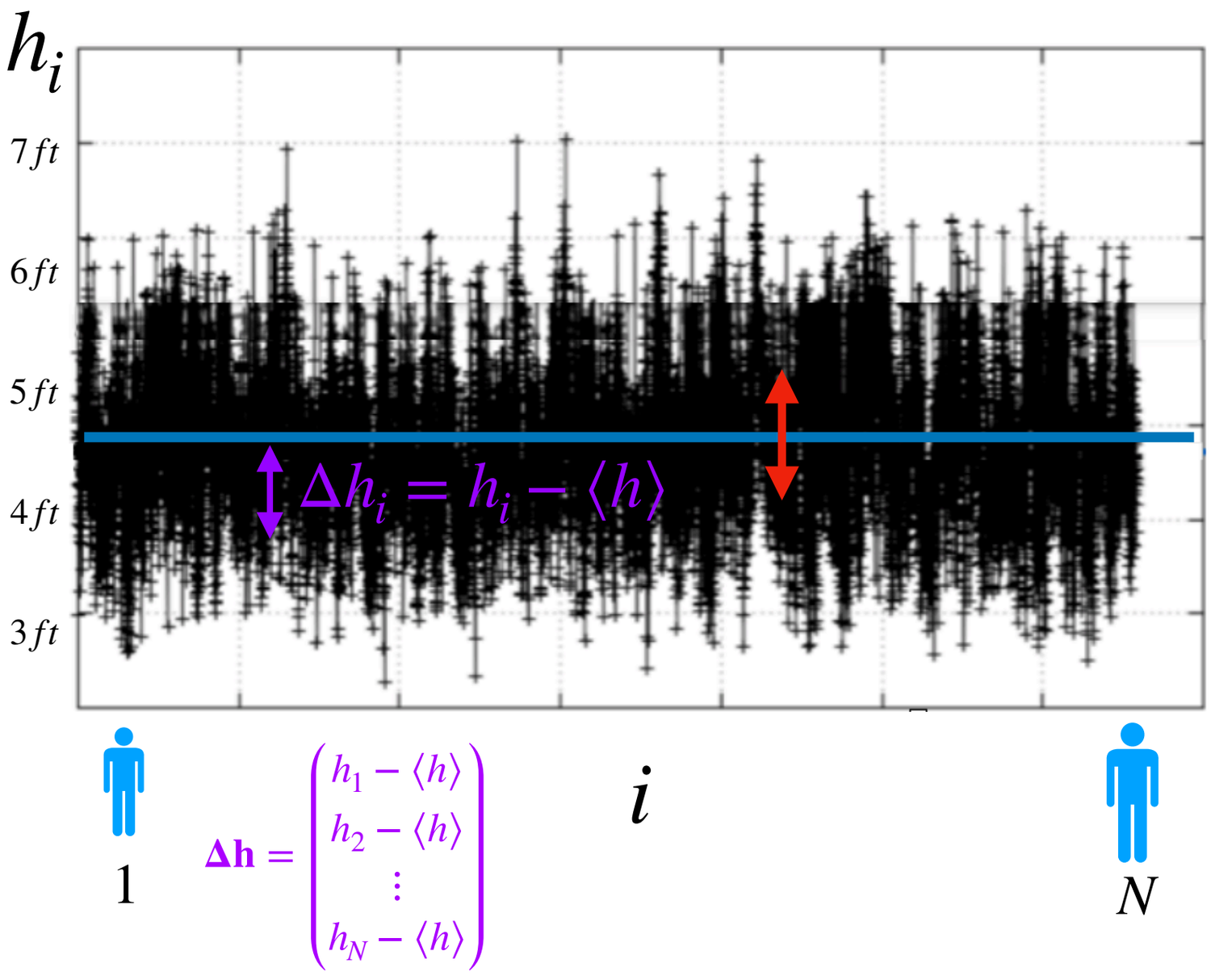
sample set

$$\mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_N \end{pmatrix}$$

sample: i=1

sample: i=2

column vector $N \times 1$



sample statistics

$\langle\rangle$ = **aveaging operator**

mean: ft $\quad \langle h \rangle = \frac{1}{N} \sum_{i=1}^{N} h_i$

measure of center

median (50%-50%)

mode (most common)

deviation: (ft) $\quad \Delta h_i = h_i - \langle h \rangle$

variance: ft^2 $\quad \langle \Delta h_i^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} (h_i - \langle h \rangle)^2$

mean square deviation

standard deviation ft $\quad \sqrt{\langle \Delta h_i^2 \rangle}$

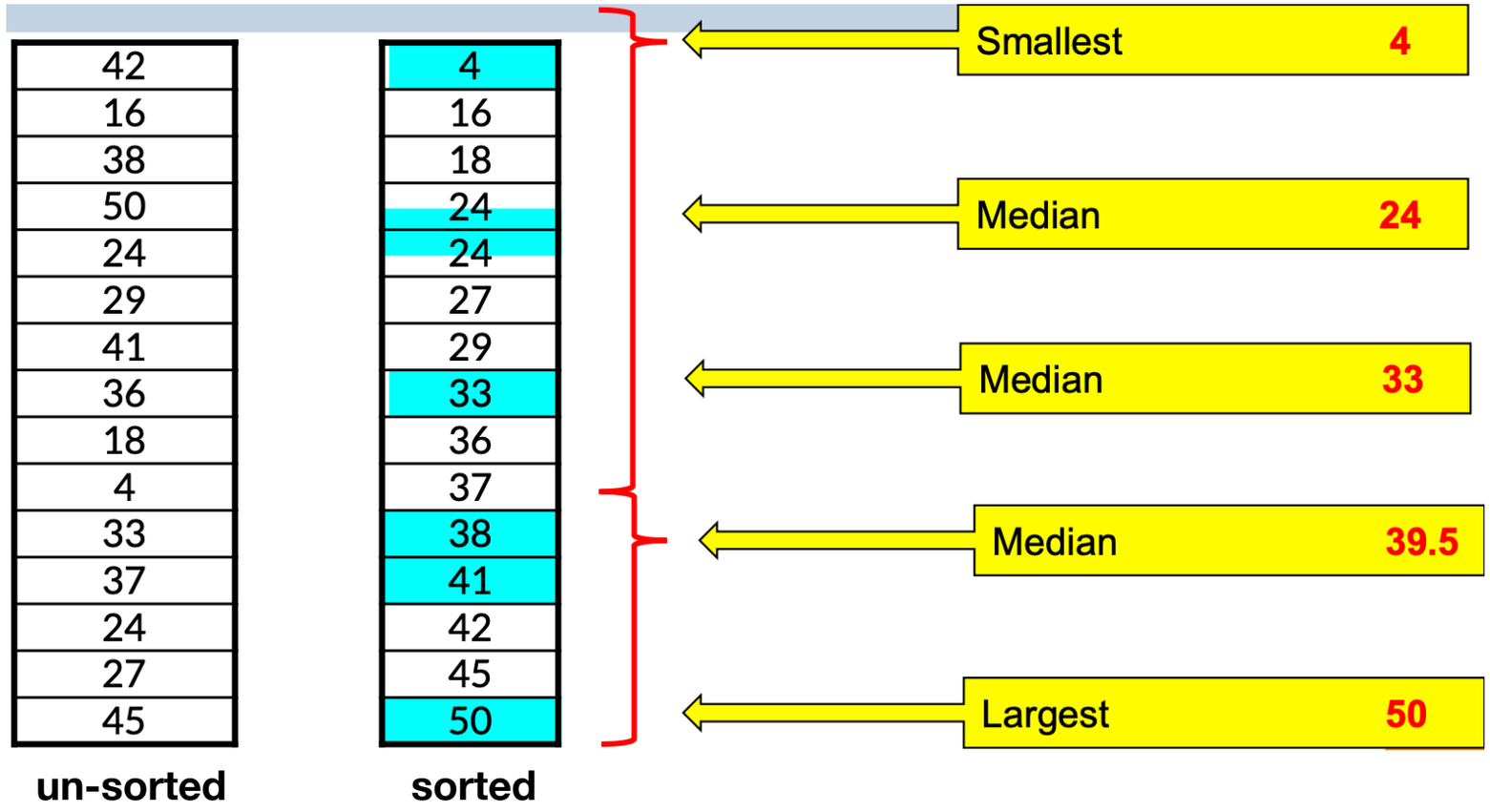
root mean square deviation

measure of width/spread

Five number summaries

- **Five-Number Summary: min, Q_1, Q_2, Q_3, Max**
- **Interquartile range (IQR):**

  IQR = Q3-Q1
- **Limits of the dataset:**
  - Lower limit = Q_1 - 1.5 x IQR
  - Upper limit = Q_3 + 1.5 x IQR



- **Outliers:** The objects below the lower limit and above the upper limit are potential outliers.
- **Percentile:** divides the data into hundredths (100 equal parts) P_1, P_2,…,P_{99}
- **Deciles:** divides the data into tenths (10 equal parts)
- **Quintiles:** divides the data into fifths (5 equal parts)
- **Quartiles:** divides the data into quarters (4 equal parts) Q_1, Q_2,Q_3

Source: Dr. Faisal Kamiran: https://github.com/FaizSaeed/Data-Science-Course/tree/master/Lectures

Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

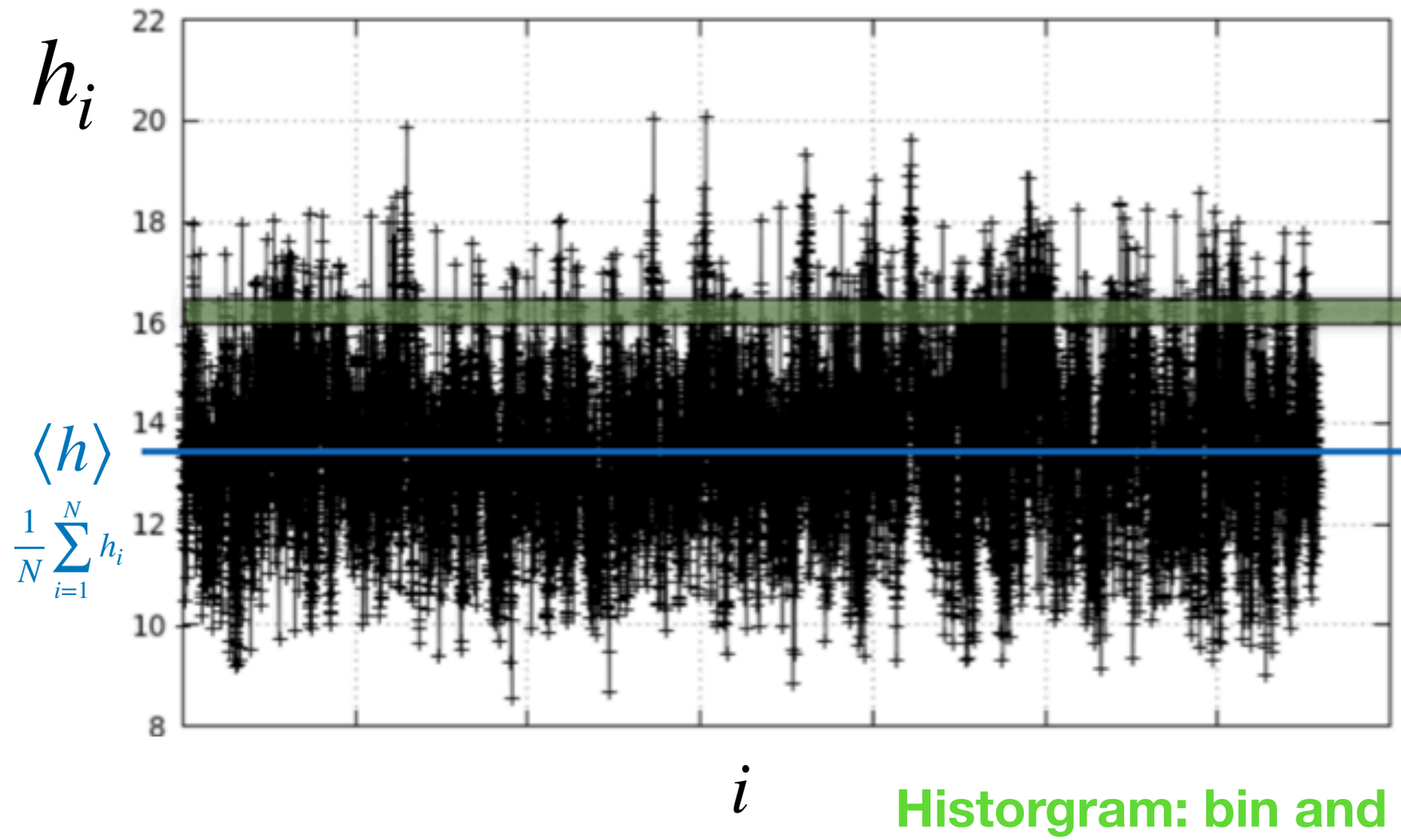
	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

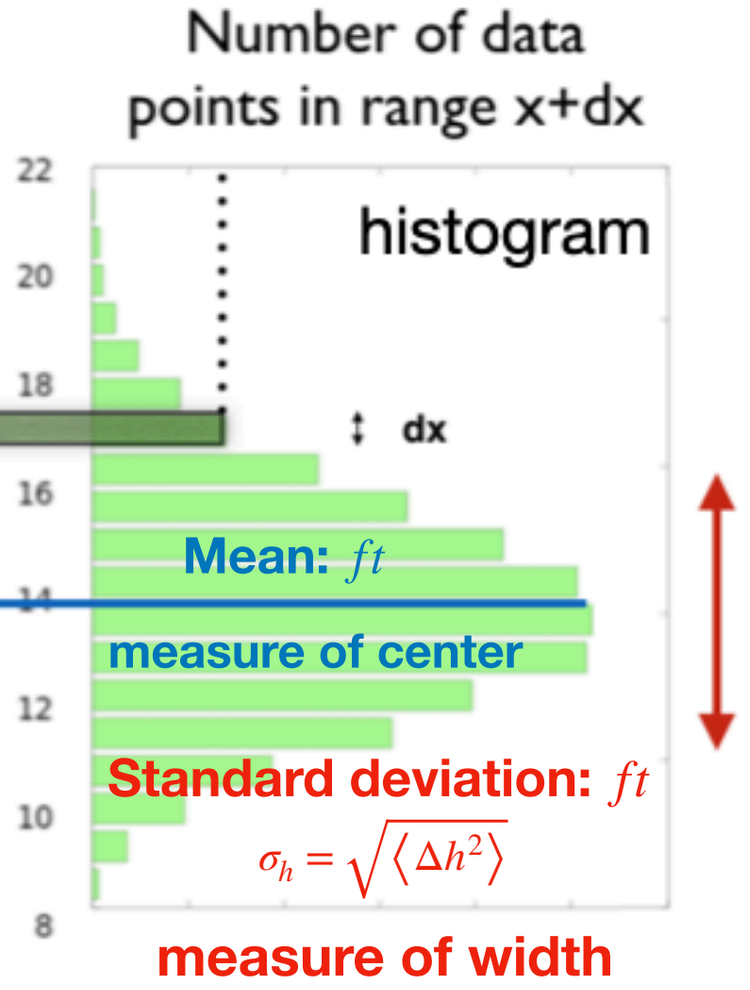
→

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

Source: Dr. Faisal Kamiran: https://github.com/FaizSaeed/Data-Science-Course/tree/master/Lectures

Historgram



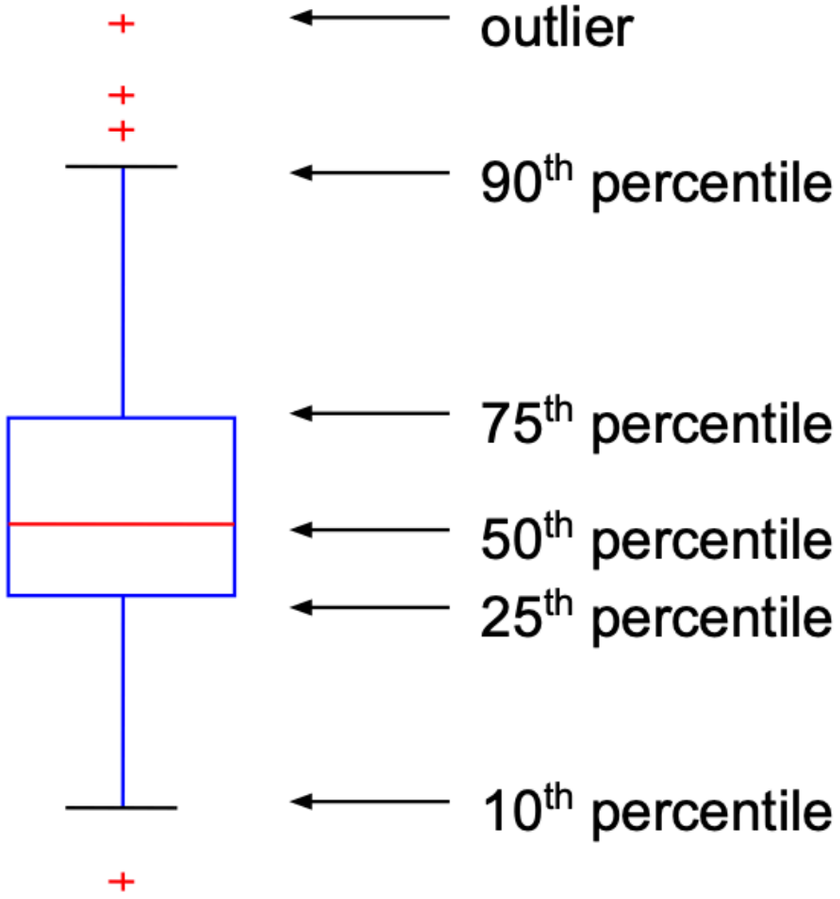


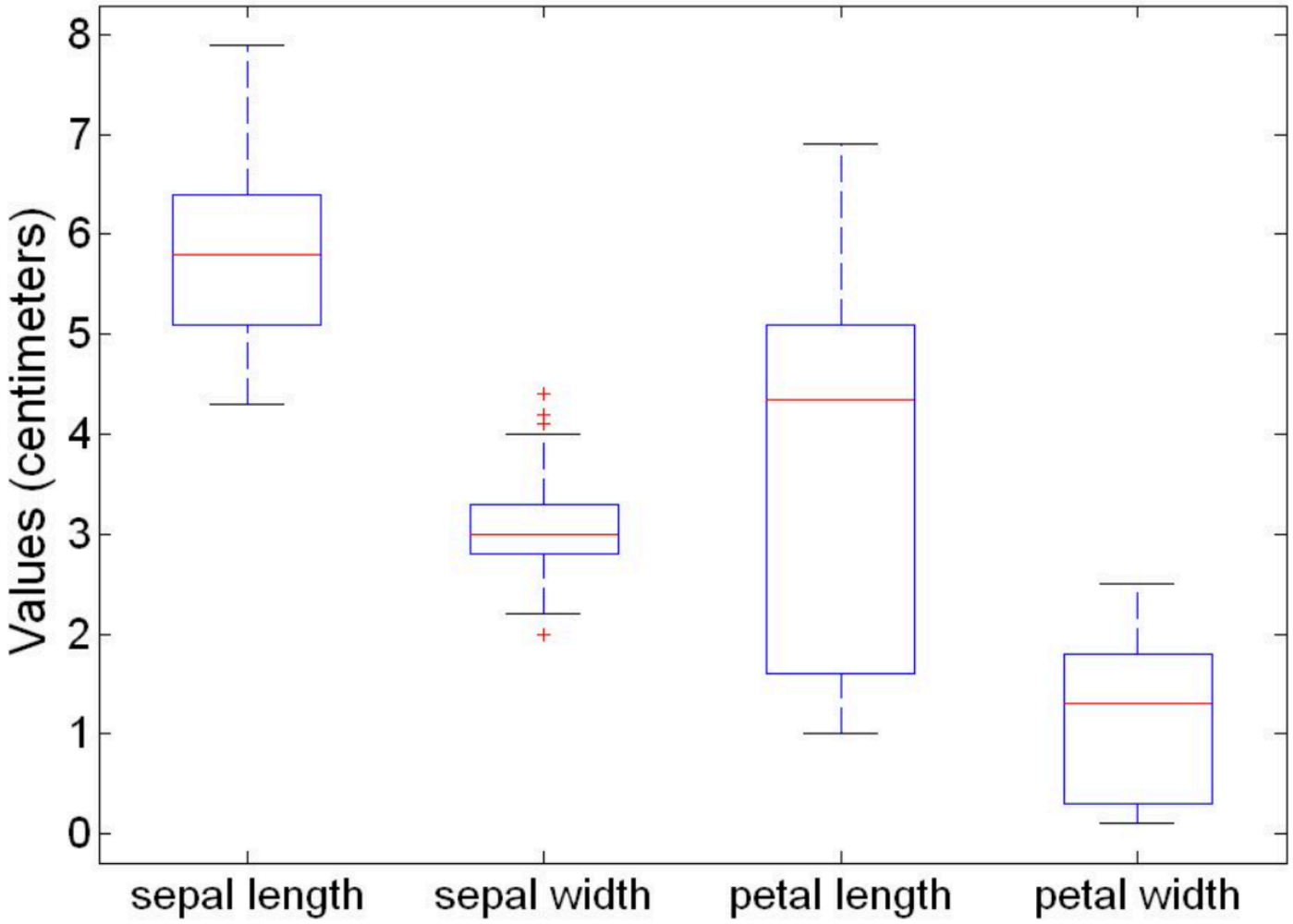
Historgram: bin and count samples in interval

- Usually shows the distribution of values of a single variable
- Divide the values into bins and show a bar plot of the number of objects in each bin
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins

Box plots

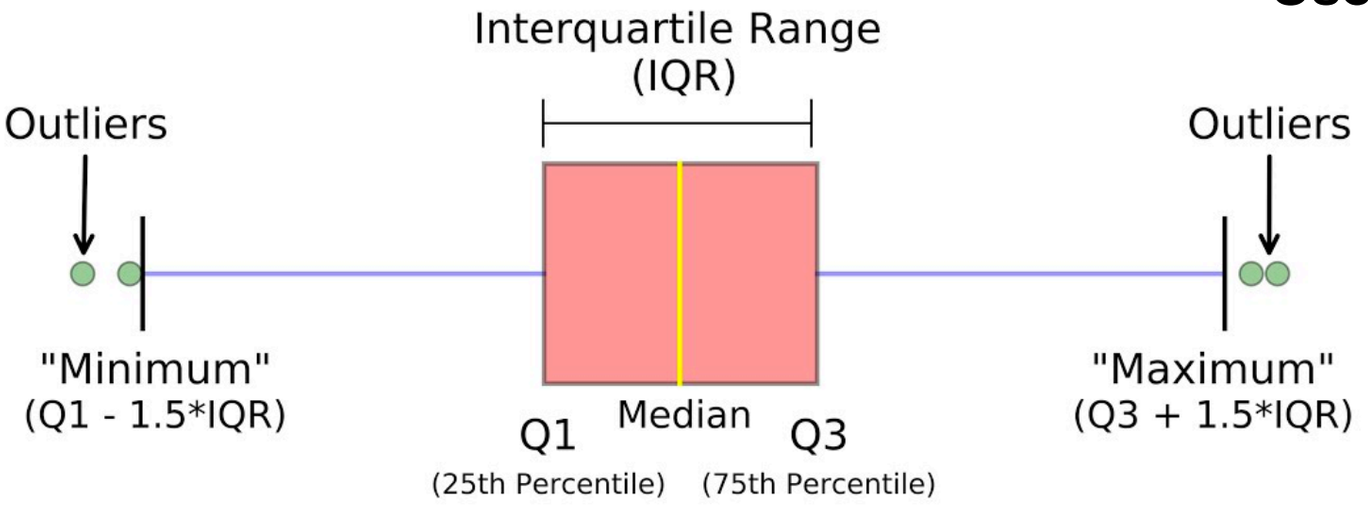
- Invented by J. Tukey
- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot





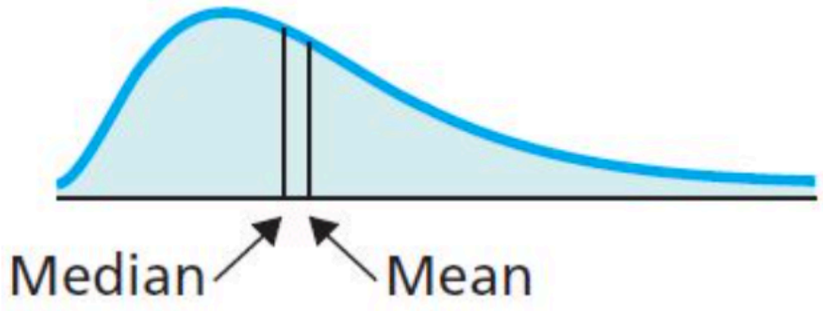
Useful for quick comparison of distributions

(Add code)

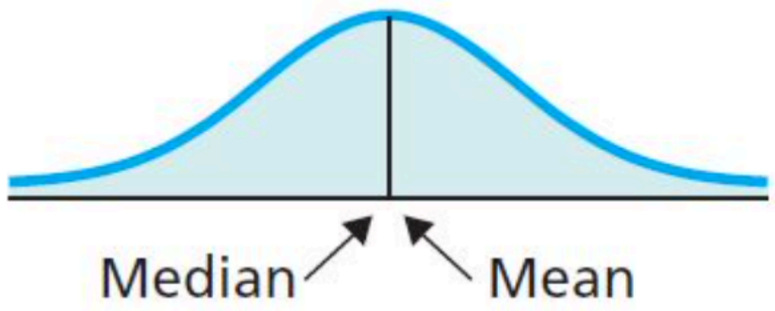


Source: Dr. Faisal Kamiran: https://github.com/FaizSaeed/Data-Science-Course/tree/master/Lectures

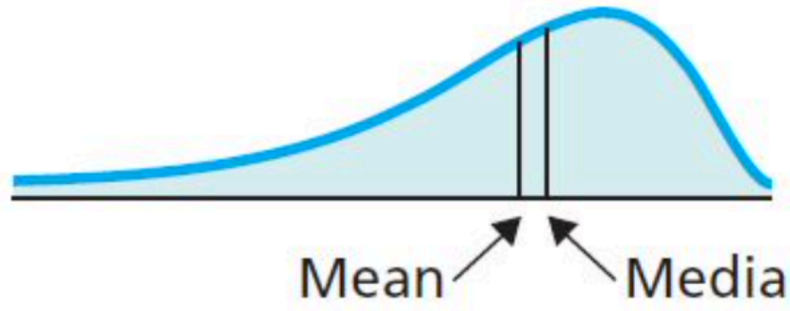
Skewness



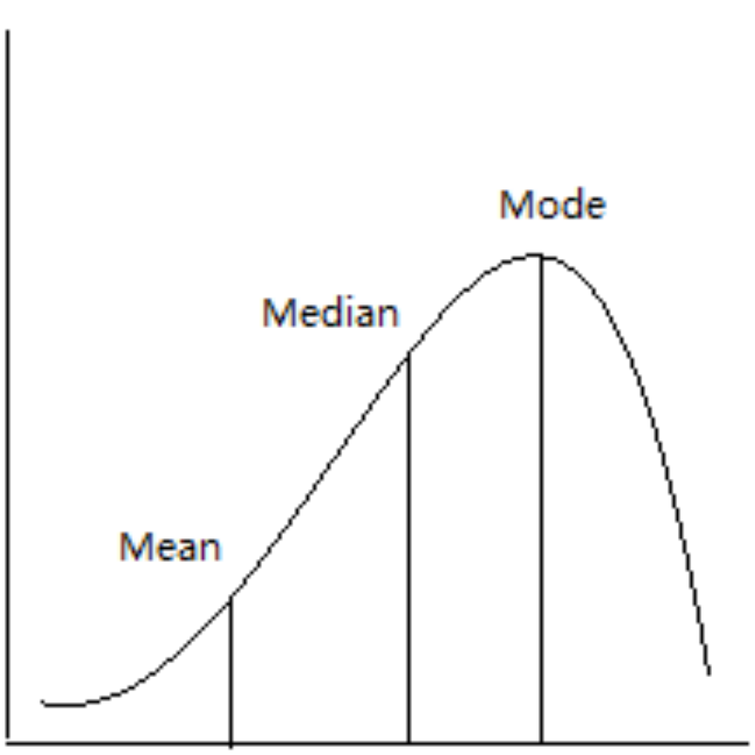
(a) Right skewed



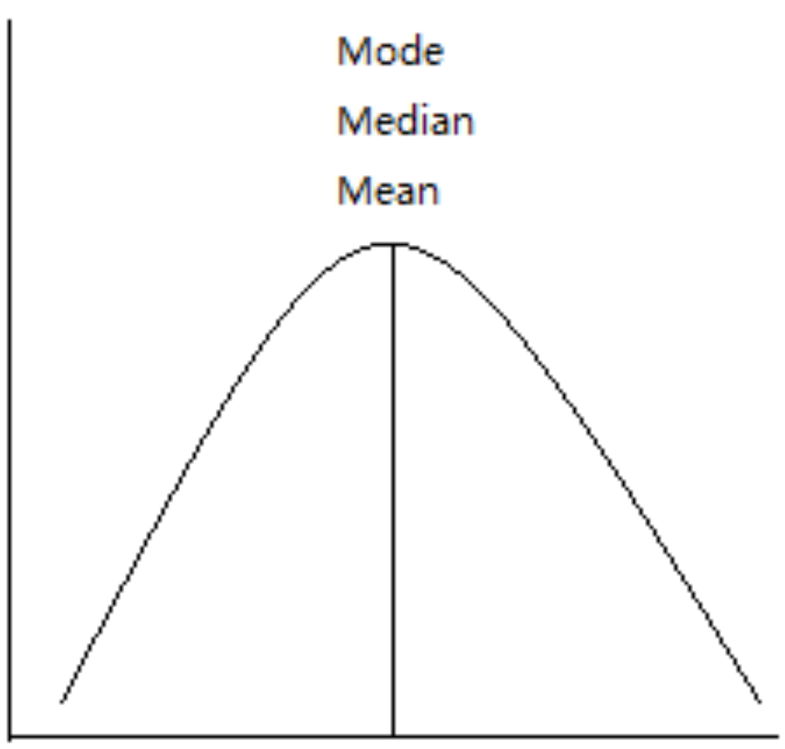
(b) Symmetric



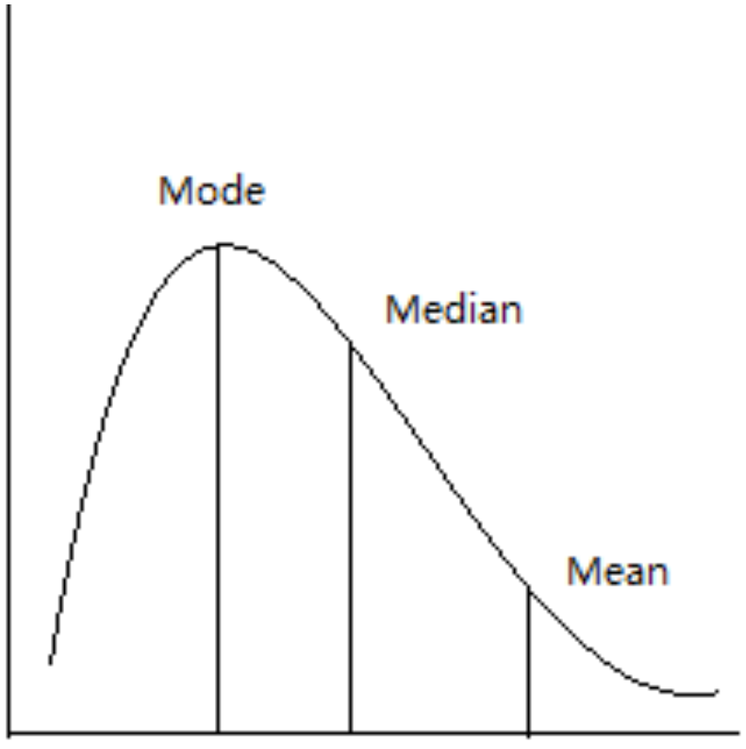
(c) Left skewed



Left skew



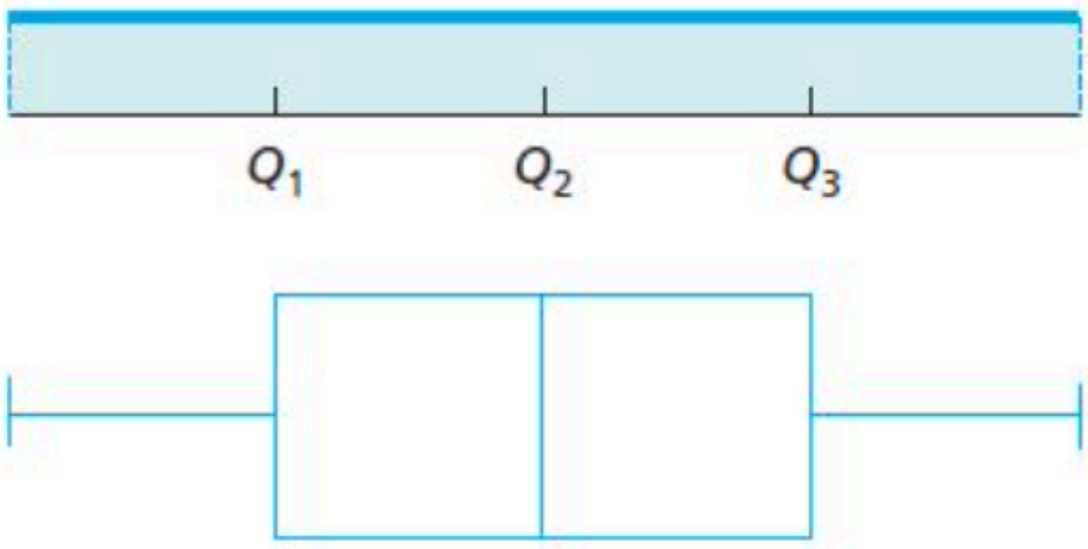
Normal Distribution



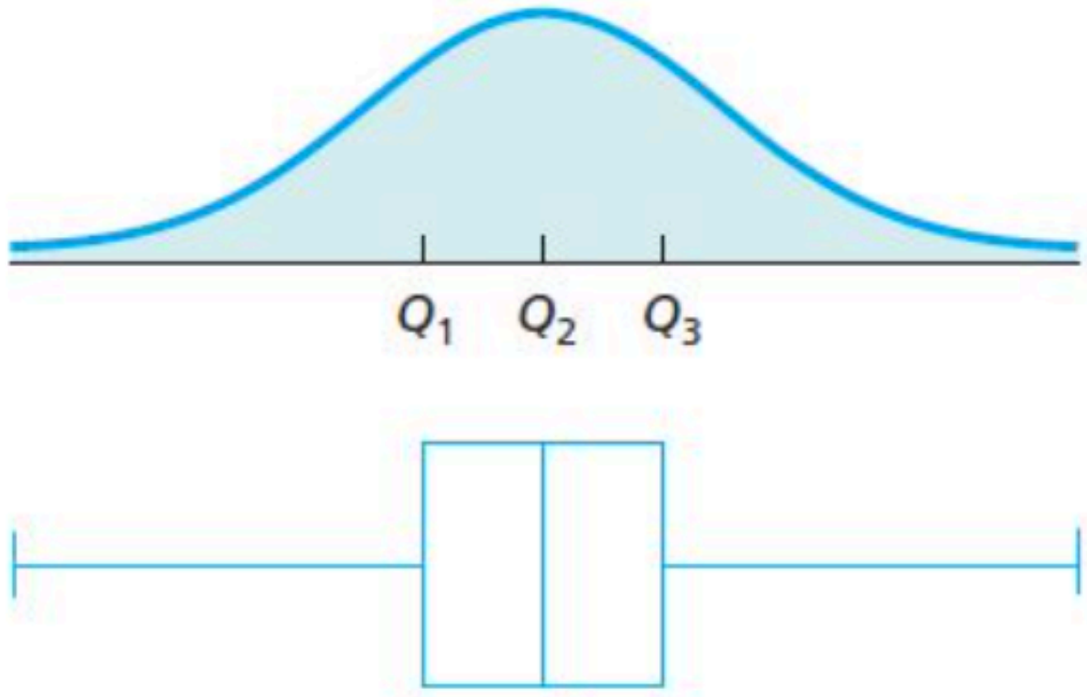
Right skew

Source: Dr. Faisal Kamiran: https://github.com/FaizSaeed/Data-Science-Course/tree/master/Lectures

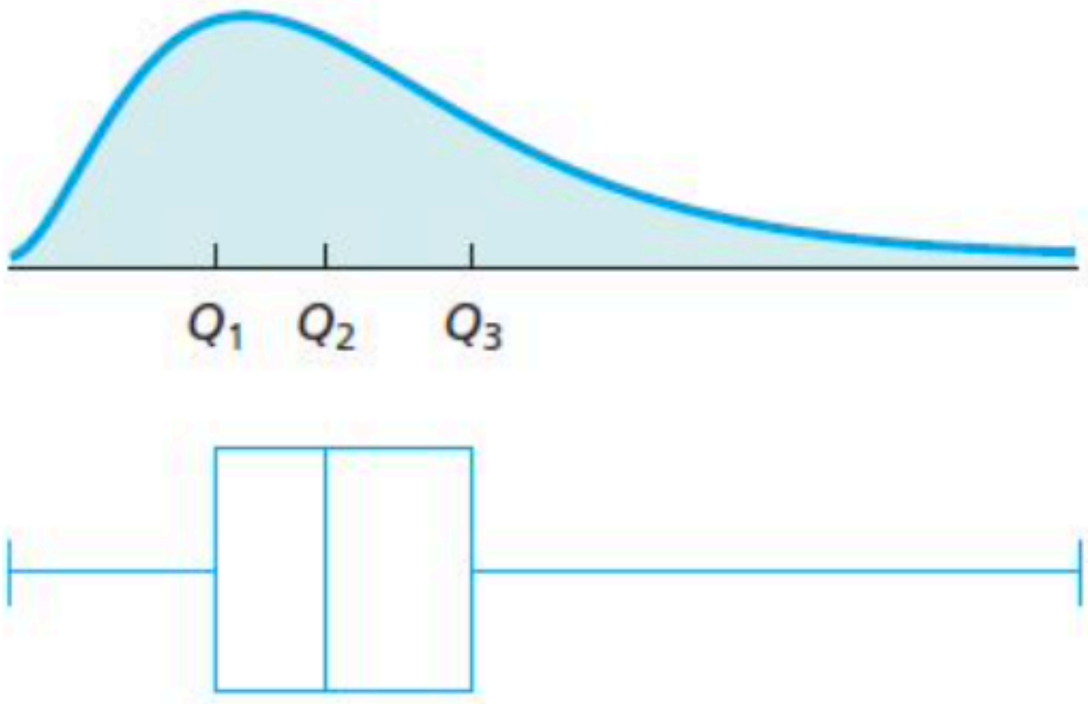
Distribution comparison



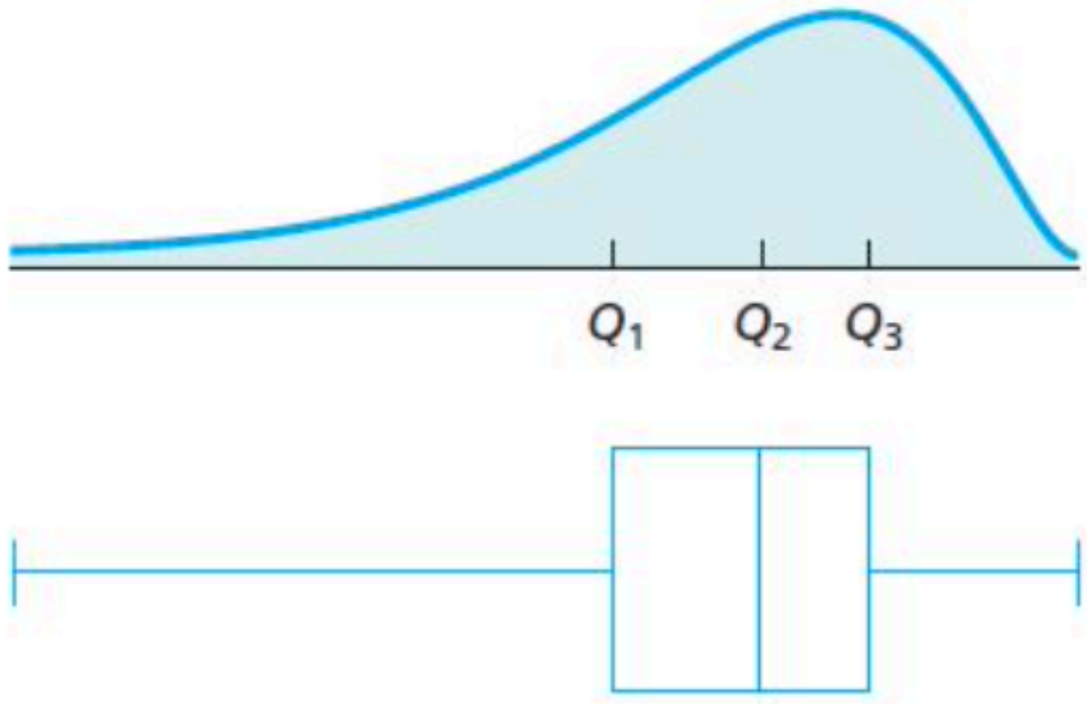
(a) Uniform



(b) Bell shaped



(c) Right skewed



(d) Left skewed

Source: Dr. Faisal Kamiran: https://github.com/FaizSaeed/Data-Science-Course/tree/master/Lectures