Introduction to Data Science With Probability and Statistics

Lecture 3: Exploratory Data Analysis, Data Visualization and Wrangling

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

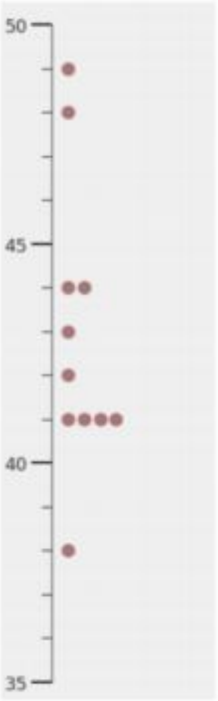
What will we learn today?

- Box-and-Whisker Plots
- Histograms
- Empirical Distribution Function

❑ *A Modern Introduction to Probability and Statistics, sections 15.1, 15.2, 16.4*

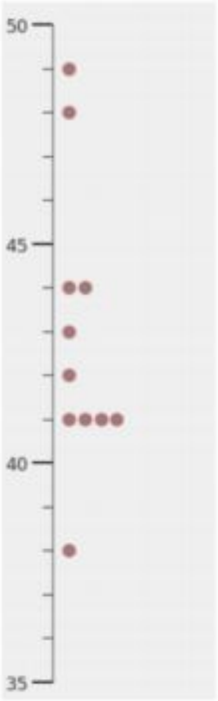
Quartiles

Example: Compute the quartiles and IQR of the data below.



Quartiles

Example: Compute the quartiles and IQR of the data below.



$$D \Rightarrow \overset{1}{38}, \overset{2}{41}, \underset{\uparrow Q_1}{\overset{3}{41}}, \overset{4}{41}, \overset{5}{41}, \underset{\uparrow Q_2}{\overset{6}{42}}, \overset{7}{43}, \overset{8}{44}, \underset{\uparrow Q_3}{\overset{9}{44}}, \overset{10}{48}, \overset{11}{49}$$

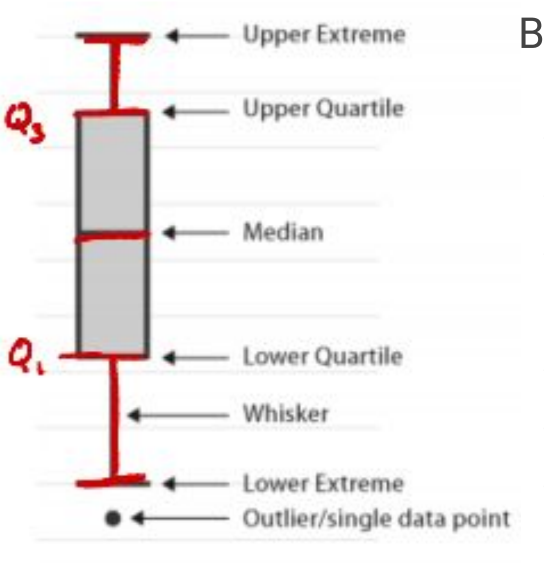
$Q_1 = 41$

$Q_2 = 42$

$Q_3 = 44$

$$IQR = Q_3 - Q_1$$
$$= 44 - 41 = \boxed{3}$$

Box-and-Whisker Plots (aka Boxplots)

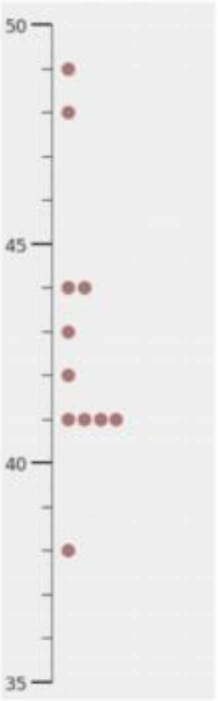


Box-and-Whisker plots are a convenient way to visualize data

- The **box** extends from Q_1 to Q_3
- The **median line** displays the median $\bar{x}$
- The **whiskers** extend to the farthest data point within 1.5 × IQR of each quartile
- The fliers or outliers are any points outside of the whiskers.
- The width of the box is unimportant
- Boxplots can be horizontally or vertically oriented

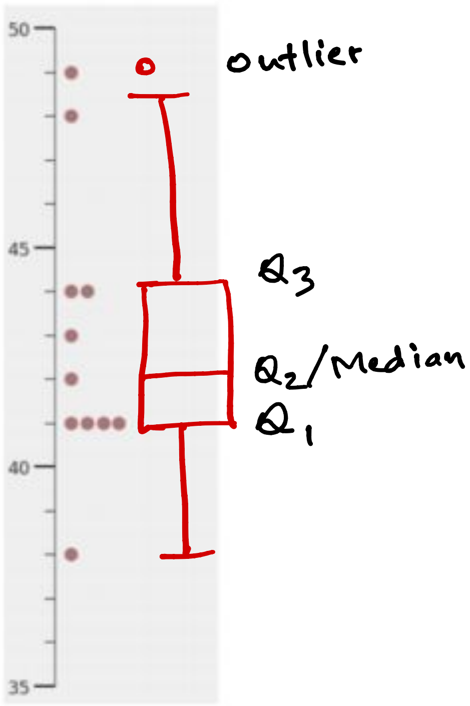
Box-and-Whisker Plots (aka Boxplots)

Example: Draw the box-and-whisker plot for the data on the left.



Box-and-Whisker Plots (aka Boxplots)

Example: Draw the box-and-whisker plot for the data on the left.



$Q_1 = 41$, $Q_3 = 44$, Med = 42

↑ Lower side of the box

↑ upper side of the box

↑ line inside the box

Limit ; $1.5 * IQR = 4.5$

max upper whisker $= 44 + 4.5 = \boxed{48.5}$

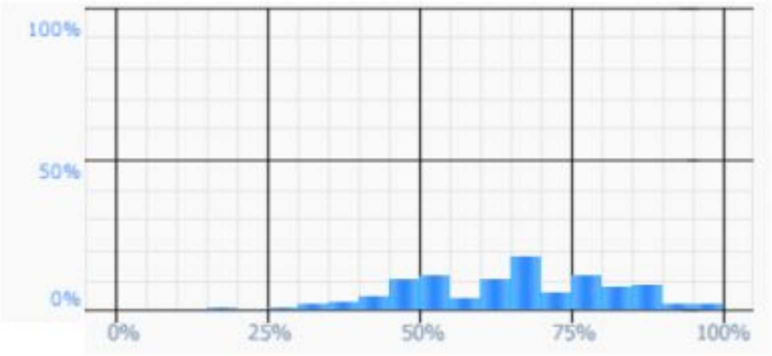
max lower whisker $= 41 - 4.5 = \boxed{36.5}$

Histograms

The **histogram** is a graphical representation of the distribution of numerical data

Construction:

- Lump or “bin” the observed values of the Variable of Interest (VOI)
    - Bins typically are consecutive, non-overlapping, and equal in width



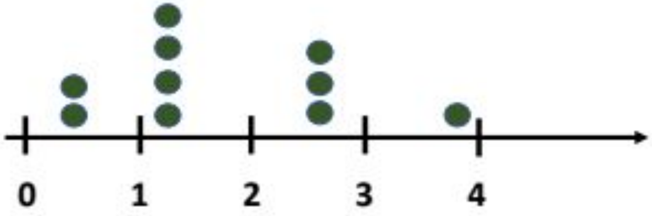
Example: Histogram of student grades on an exam.

Histograms

For a **frequency histogram**: count the number of data values that fall into a bin and draw a rectangle over that bin with height equal to the count.

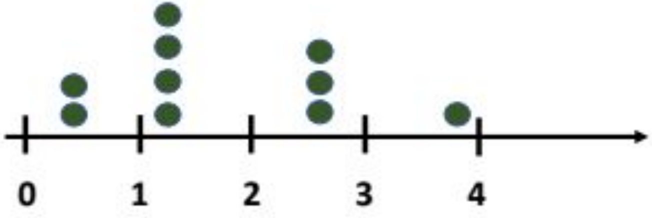
For a **density histogram**: count the number of data values that fall into a bin and adjust the height such that the sum of the area of all bins is equal to 1 (normalizing so that the sum of the heights = 1)

Example: Create a frequency histogram and a density histogram for the following data.



Histograms

Example: Create a frequency histogram and a density histogram for the following data.



Density histogram

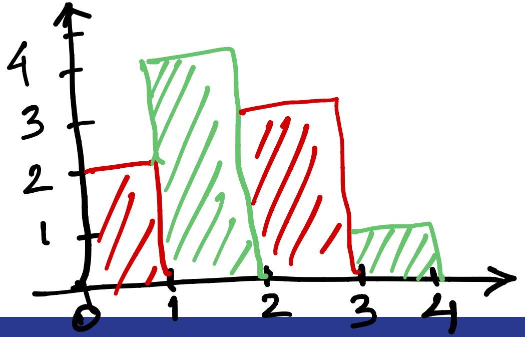
range	freq.	Density
0-1	2	2/10
1-2	4	4/10
2-3	3	3/10
3-4	1	1/10

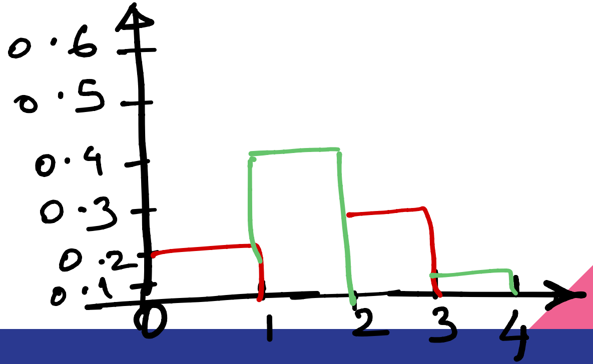
Total freq.

= 2+4+3+1

= 10

Frequency histogram

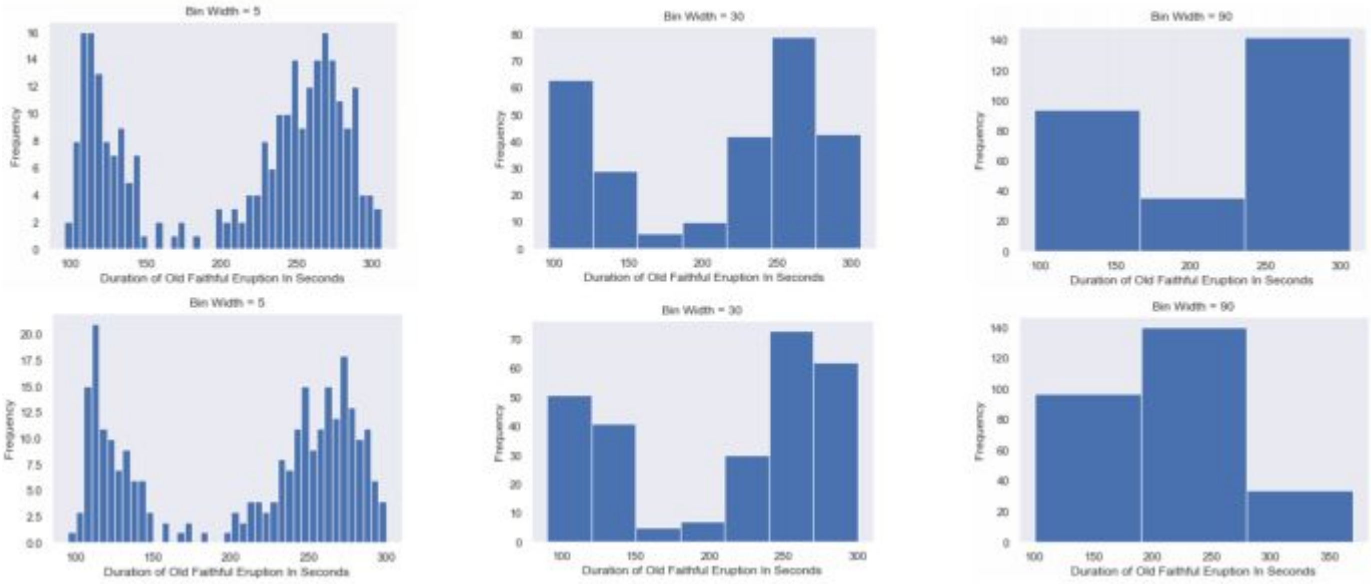




Histograms

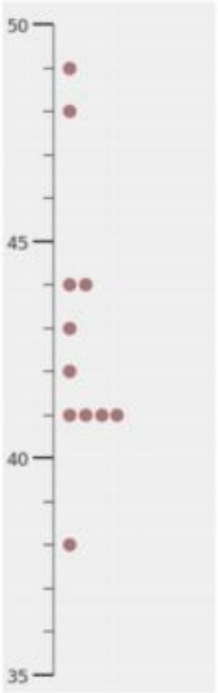
Note that choosing a different bin width can paint a very different picture of the data. Choosing the starting point for the bins also makes a difference.

Example: Old Faithful eruption duration data (from MIPS section 15.1 p. 208)



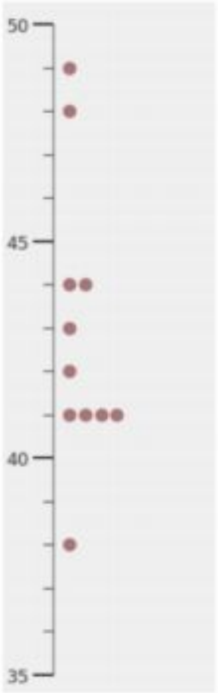
Histograms

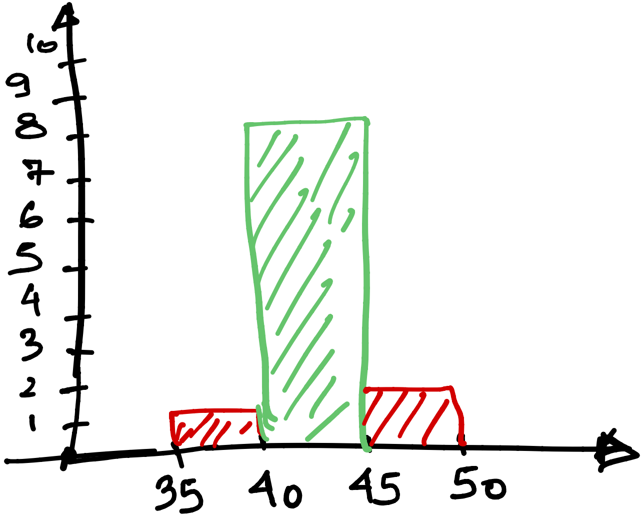
Example: Find the frequency histogram with bin width 5 of the data below, with left-most bin edge at 35.



Histograms

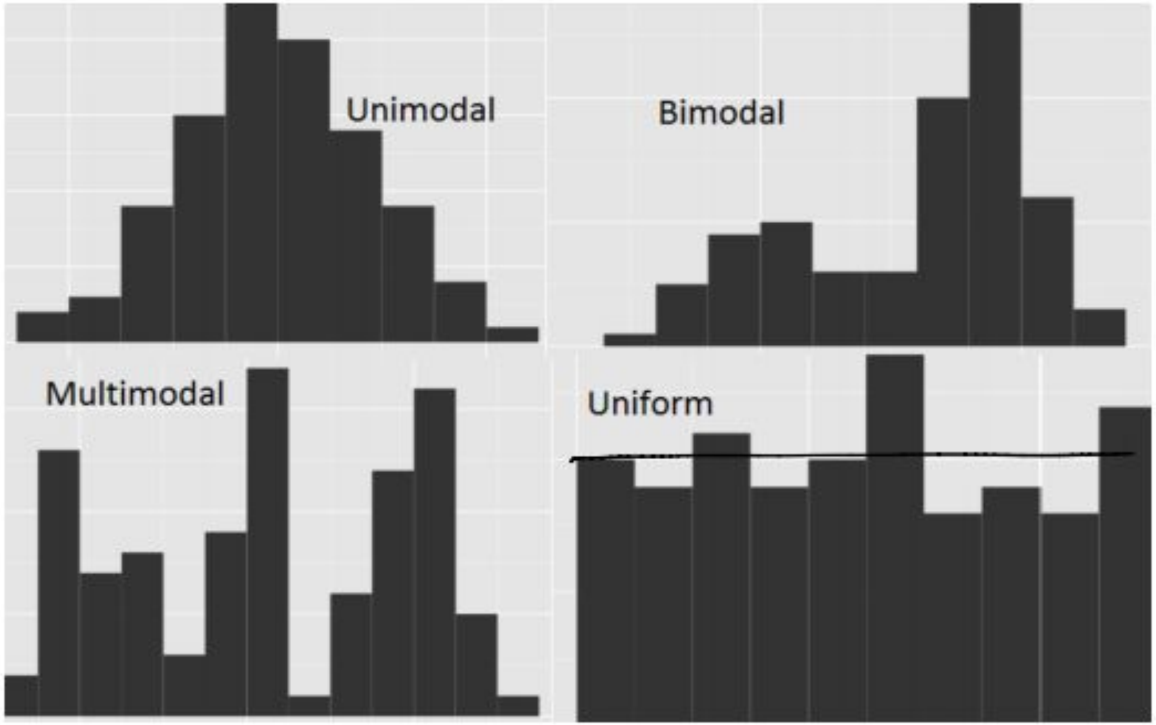
Example: Find the frequency histogram with bin width 5 of the data below, with left-most bin edge at 35.





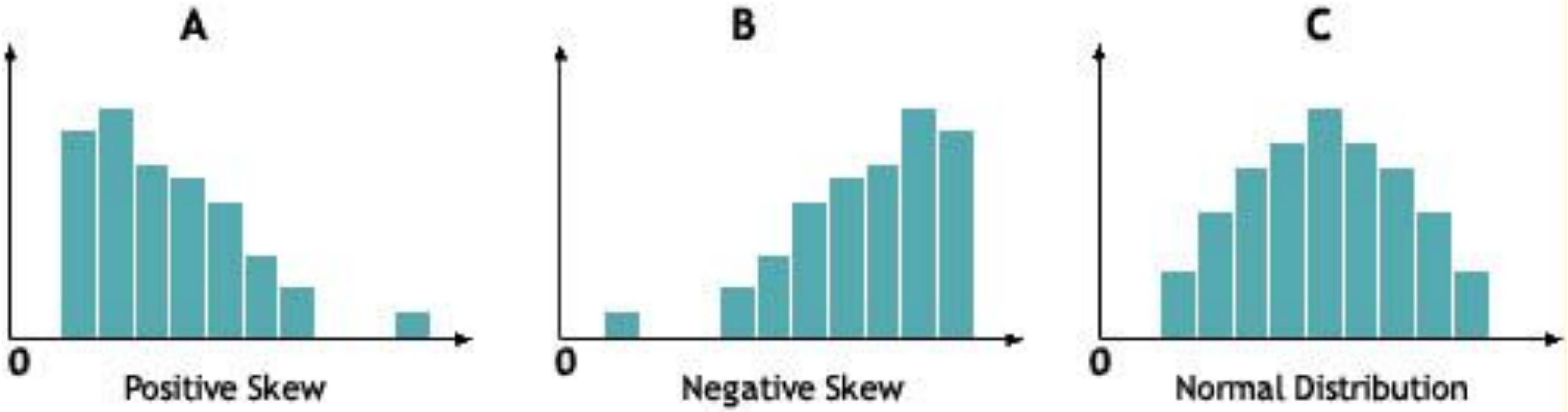
Histograms

Histograms come in a variety of shapes.



Histograms

Histograms come in a variety of shapes.



Empirical Distribution function

- This is used to plot a dataset in a cumulative manner.
- It is represented by F_n and it is defined such that $F_n(x)$ is the proportion of elements in the dataset that are less than or equal to x.

$$F_n(x) = \frac{number\ \ of\ \ elements\ \ in\ \ the\ \ dataset \leq x}{n}$$

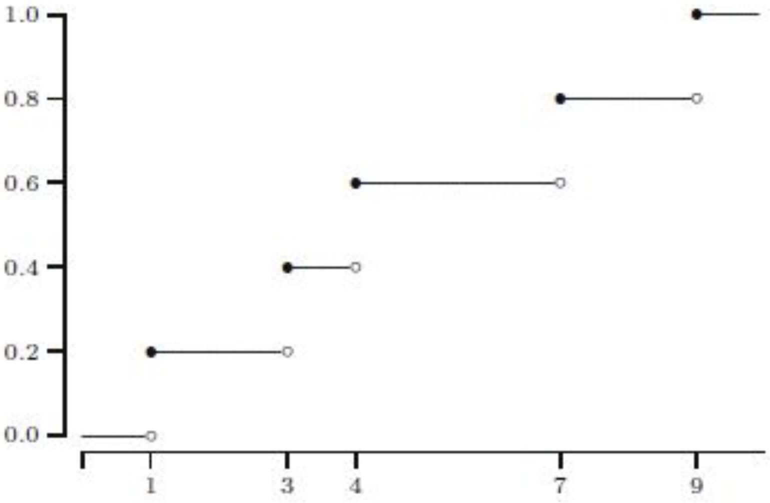
Empirical Distribution function - Properties

F_n satisfies the four properties of a distribution function:

1. It is continuous from the right
2. $F_n(x)$ -> 0 as x -> $-\infty$
3. $F_n(x)$ -> 1 as x -> ∞
4. F_n is non-decreasing

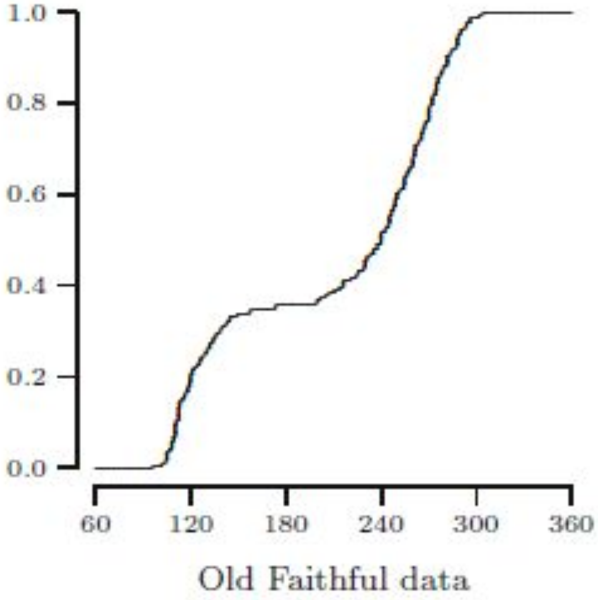
Empirical Distribution function - example

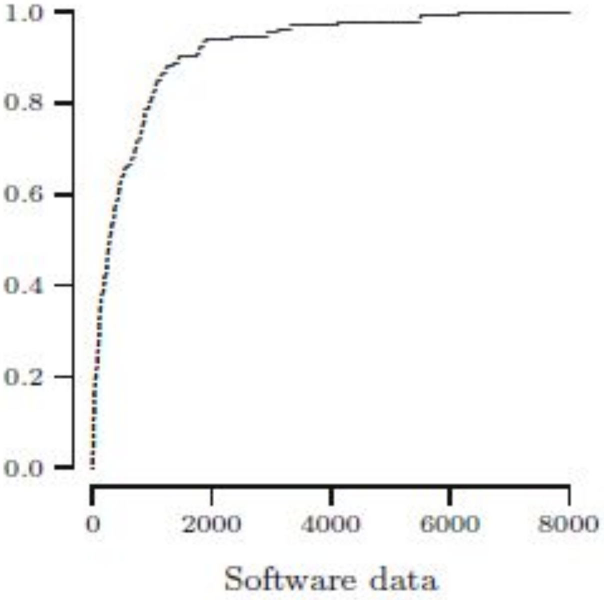
Given Data - 4   3   9   1   7



- ➢ $F_n(x) = 0 \quad \forall\ x < \min(\text{data})$
- ➢ $F_n(x) = 1 \quad \forall\ x > \max(\text{data})$

Empirical Distribution function - example plots





Empirical Distribution function - Question

Suppose that for a dataset consisting of 300 elements, the value of the empirical distribution function in the point 1.5 is equal to 0.7. How many elements in the dataset are strictly greater than 1.5?

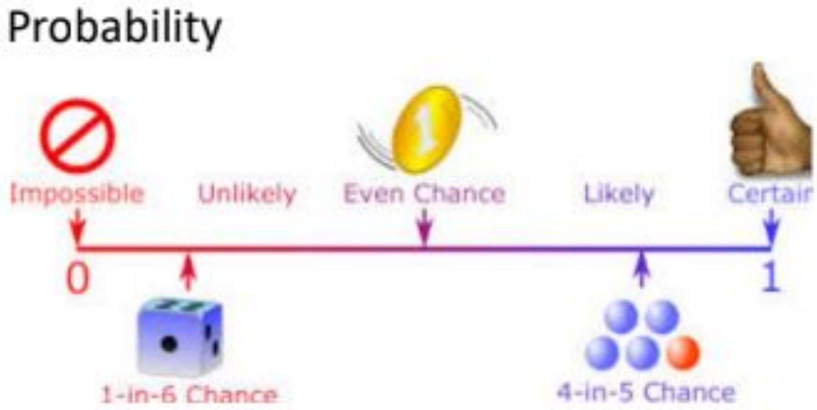
Empirical Distribution function - Question

Suppose that for a dataset consisting of 300 elements, the value of the empirical distribution function in the point 1.5 is equal to 0.7. How many elements in the dataset are strictly greater than 1.5?

- 70% of the datapoints are less than or equal to 1.5
- Thus, 30% of the remaining ones are strictly greater than 1.5

$\therefore$ 30% of 300 = $\underline{\underline{90}}$.

Next Time:


Probability
Impossible
Unlikely
Even Chance
Likely
Certair
0
1
1-in-6 Chance
4-in-5 Chance