CDS501: PRINCIPLES & PRACTICES OF DATA SCIENCE & ANALYTICS

Statistical Distribution

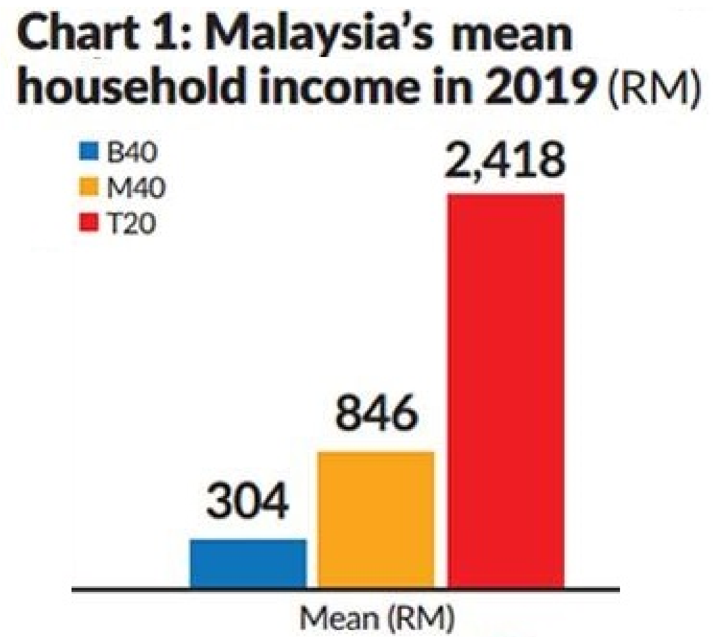
Outline

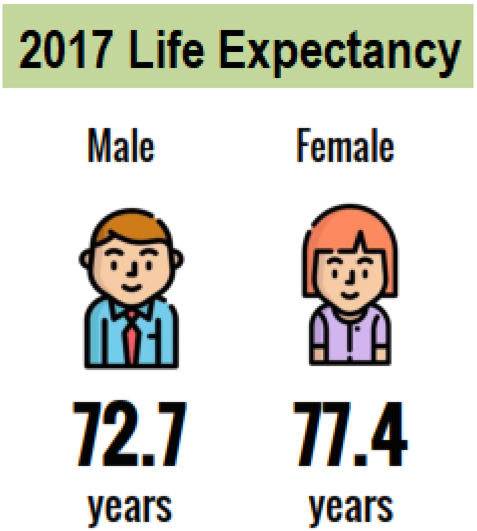
- Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Dispersion
  - Measures of Association
- Statistical Distributions
  - Basic concepts
  - Probability Density Function
  - Normal Distribution
  - Binomial Distribution
  - Multinomial Distribution

Descriptive Statistics

Measures of Central Tendency

Descriptive Statistics





A set of numbers that "describe" a data

Descriptive Statistics

- Measures of central tendency
  - Central aspect of the data
- Measures of dispersion
  - How spread-out the data is

Measures of Central Tendency

- Mean (average)
- Median
- Mode

Mean (Average)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{(5.2 + 2.8 + 1.2 + 6.1 + 4.8)}{5} = 4.02$$

Median

- The middle number that divides the ordered observations into two parts

1.2, 2.8, 4.8, 5.2, 6.1

Mean vs Median

Consider the following salaries:

RM 5500

RM 4800

RM 5900

RM 4900

RM 5200

RM 4500

RM 22000

Mean: RM7542.86

Median: RM5200.00

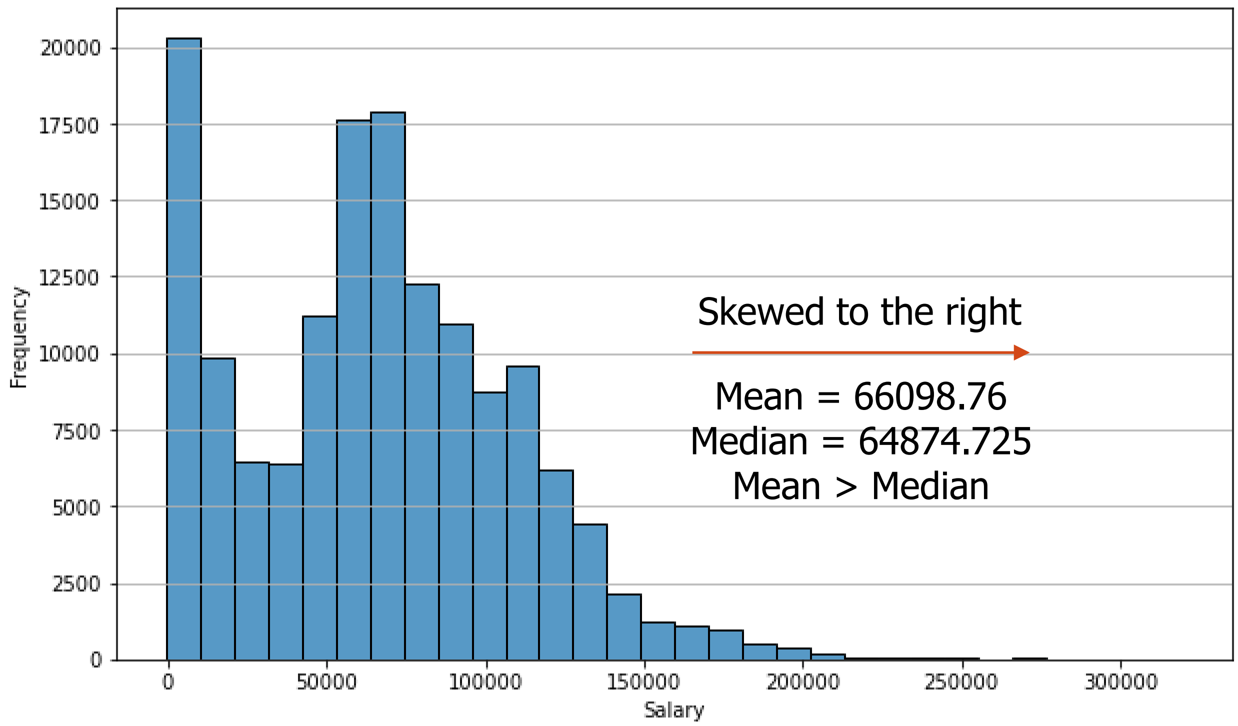
Mean and Median

- Mean is influenced by extreme observations
- Median is better the summary descriptor to use when there are extreme observations

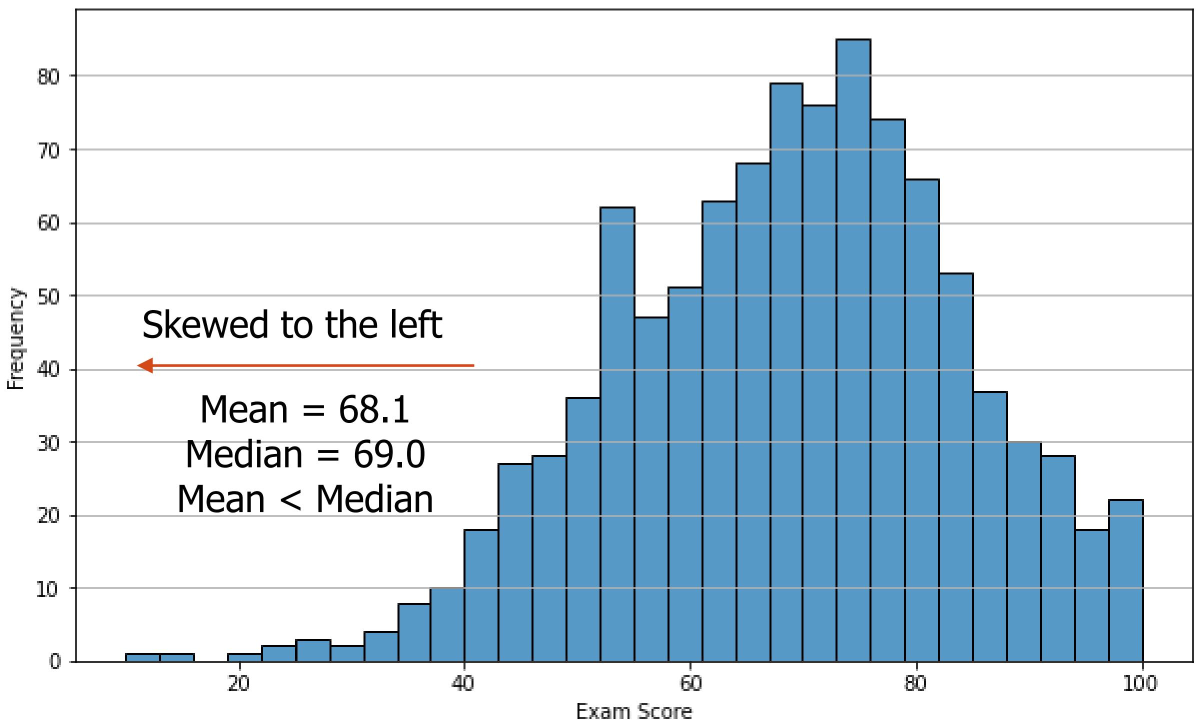
Mean and Median

- Mean and median relationship relates to the skewness of the data

Mean and Median



Mean and Median



Mode

- The most frequently occurring value in a set of data

Mode

- The most frequently occurring value in a set of data
- Not that relevant descriptive statistic when the data is continuous e.g. exchange rate USD to MYR

Date	USD to MYR
30/4/2020	4.326
29/4/2020	4.36
28/4/2020	4.3705
27/4/2020	4.354
24/4/2020	4.365
23/4/2020	4.365
22/4/2020	4.395
21/4/2020	4.3923
20/4/2020	4.388
17/4/2020	4.365
16/4/2020	4.374
15/4/2020	4.33
14/4/2020	4.323
13/4/2020	4.3225
10/4/2020	4.308
9/4/2020	4.341
8/4/2020	4.3398

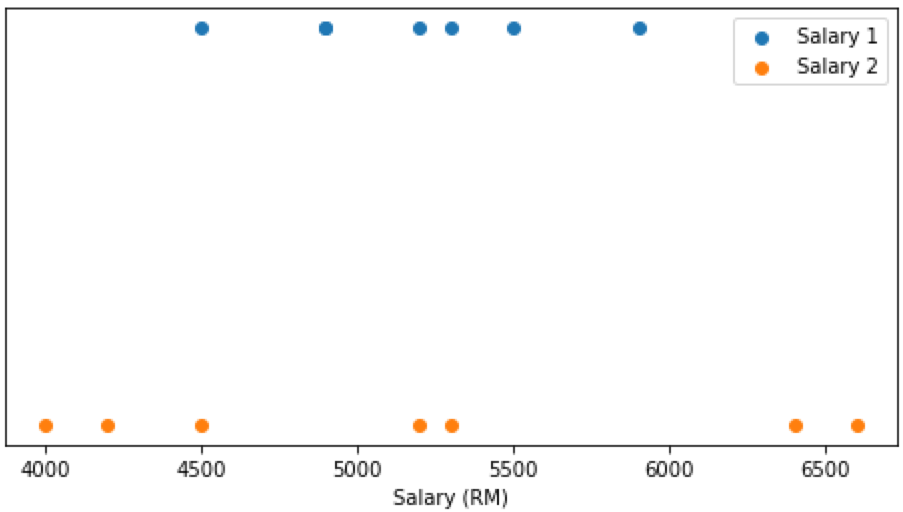
Descriptive Statistics

Measures of Dispersion

Measures of Dispersion

	Salaries 1 (RM)	Salaries 2 (RM)
	5500	6400
	4900	5300
	5900	4200
	4900	5200
	5200	4000
	4500	4500
	5300	6600
Mean:	5171.43	5171.43
Median:	5200.00	5200.00

Measures of Dispersion



Spread of Salaries 2 is greater than Salaries 1

Range

- Range of data

$$\text{Range} = \text{Max}(\boldsymbol{x}) - \text{Min}(\boldsymbol{x})$$

Range

- Range of Salaries 1

  = 5900 – 4500

  = 1400

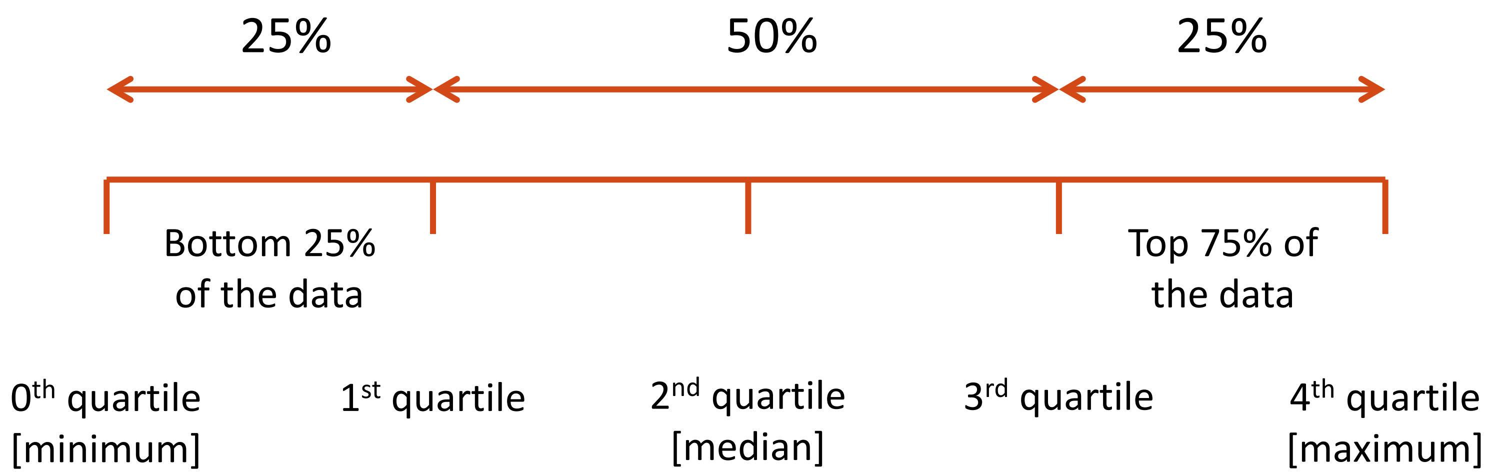
- Range of Salaries 2

  = 6600 – 4000

  = 2600

Interquartile Range

- IQR is the middle 50% of the data



$$IQR = Q3 - Q1$$

Interquartile Range

- IQR of Salaries 1

  4500, 4900, 4900, 5200, 5300, 5500, 5900

  Q1 = 4900 and Q3 = 5500

  IQR = 600

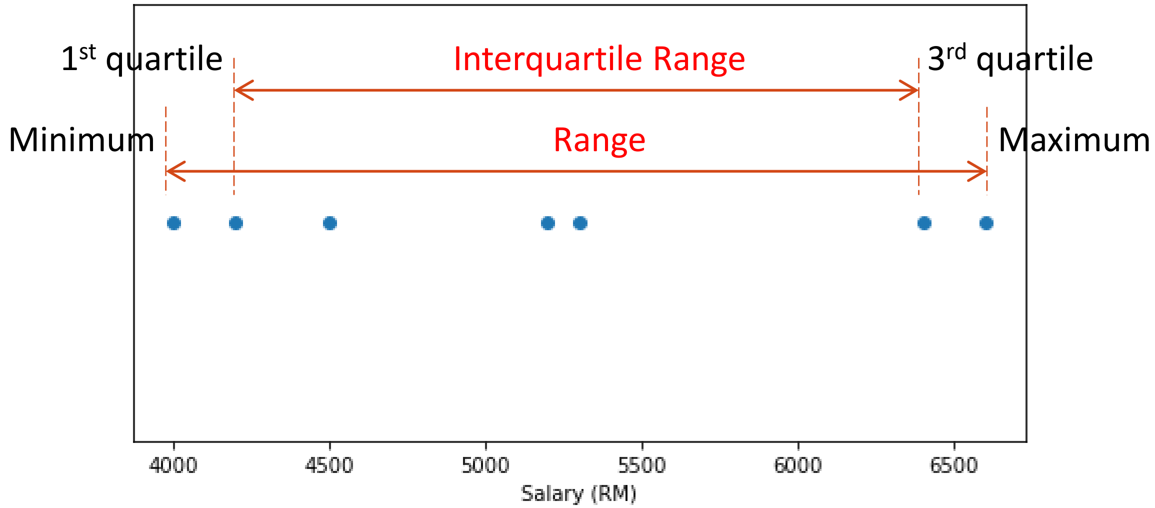
- IQR of Salaries 2

  4000, 4200, 4500, 5200, 5300, 6400, 6600

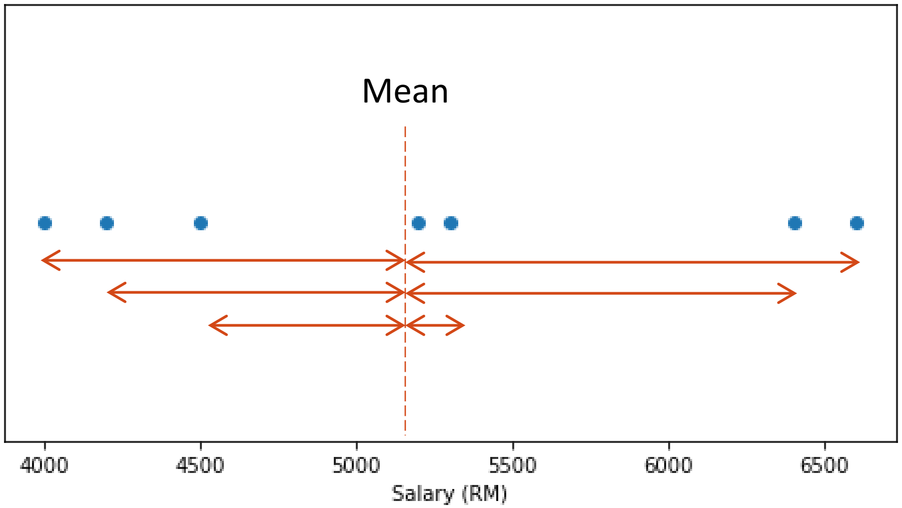
  Q1 = 4200 and Q3 = 6400

  IQR = 2200

Standard Deviation



Standard Deviation



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

Standard Deviation

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

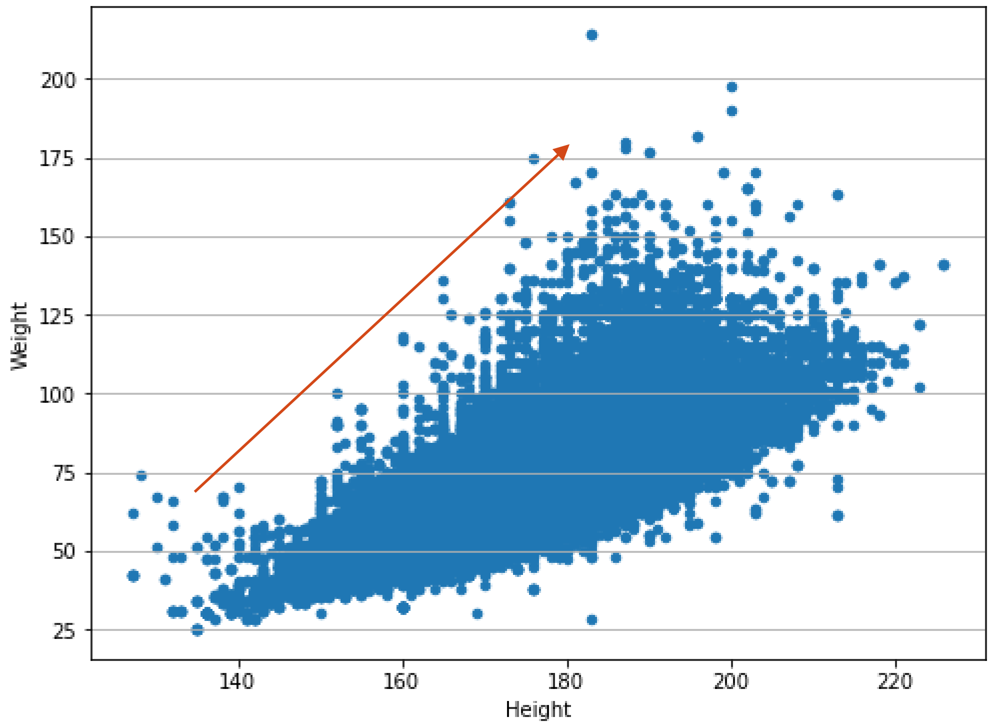
$$\text{variance} = \text{standard deviation}^2$$

Descriptive Statistics

Measures of Association

Covariance and Correlation

- How do two variables vary together



Covariance

$$Cov = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})$$

- $\bar{x}$: mean of variable x
- $\bar{y}$: mean of variable y
- The range of values: $-\infty$ to $+\infty$
- Positive value: positive relationship
- Negative value: negative relationship

Covariance does not describe the strength of the relationship

Correlation

$$Cor = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

- σ_x: standard deviation of variable x
- σ_y: standard deviation of variable y
- No affected by the units of measurement
- The range of values: -1 to $+1$
- Positive value: positive relationship
- Negative value: negative relationship
- $Cor > +0.5$ or $Cor < -0.5$ are considered strong

Statistical Distributions

Probability, Random Experiment, Random Variable and Probability Density Function

Probability

- A numerical measure of the frequency of occurrence of an event
- A scale from 0 to1
- Tossing a fair coin
- Rolling a dice

Random Experiment & Random Variable

- Random Experiment
  - Any situation where a process leads to more than one possible outcome
  - tossing a coin
  - rolling a dice
  - observing the number of goals in a football match
  - observing the number of phones sold by a shop in a year
  - observing the total sales of a shop in a day

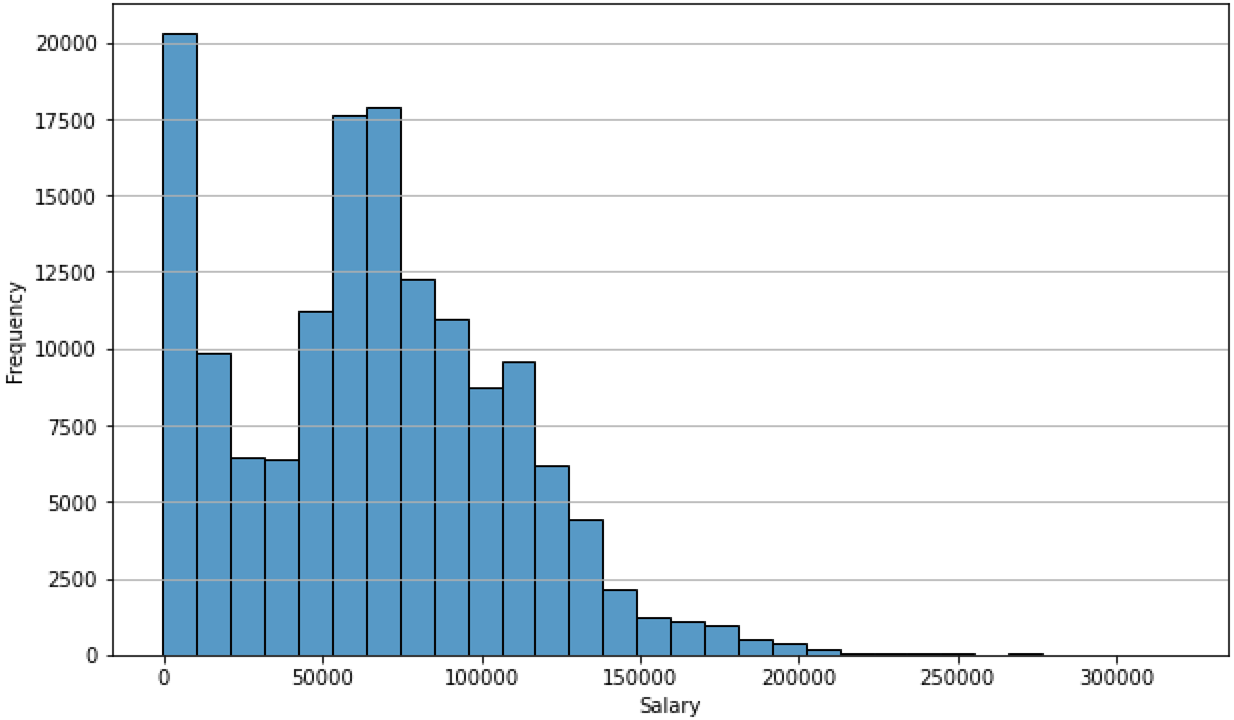
Random Experiment & Random Variable

- Random Variable
  - A variable that takes on values determined by the outcome of a random experiment

<u>Random Experiment</u>	<u>Random Variable</u>
Coin toss	outcome={head, tail}
Roll of a dice	outcome={1, 2, 3, 4, 5, 6}
Observe # phone sold	phone sold={0, 1, 2, 3, …}
Observe total sales	total sales=$\mathbb{R}_{\geq 0}$

Statistical Distribution

- Description of the frequency of all possible outcomes of a random variable



Random Experiment	Random Variable
Observing salary of workers	salary= $\mathbb{R}_{>0}$

Probability Density Function

- A rule that assigns probabilities to various possible values of a random variable can take when it is being approximated by a particular statistical distribution
- Probability Mass Function (for discrete data)

Probability Mass Function

- Consider a coin toss

Outcome of toss	Probability
Head	0.5
Tail	0.5
	1.0

Probability Mass Function

- Consider a coin toss

Outcome of toss	Probability
Head	0.5
Tail	0.5
	1.0

- Consider a roll of a dice

Outcome of roll	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
	1.0

The rule is known as Probability Mass Function (for discrete data)

Probability Mass Function

- What about more complex process such as the number of customers in a day
- Approximate the process using a statistical distribution and use the pmf of the distribution

Customers

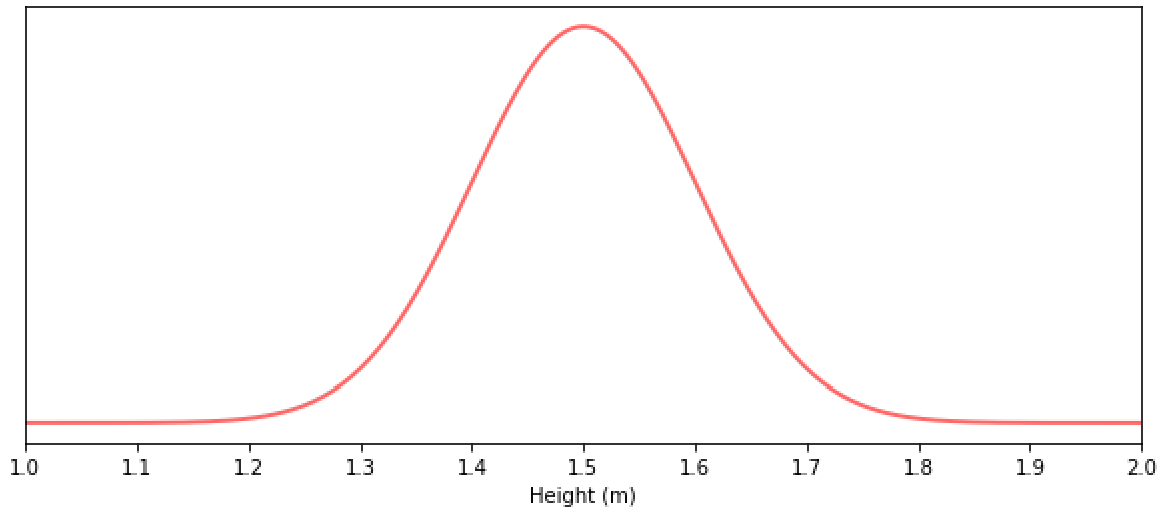
0 1 2 3 4 5 6

Zero probability between 3 and 4

Probability Density Function

- For continuous data
- The probability is not for a particular outcome but for ranges of outcomes
  - Probability of someone's height between 1.60m and 1.70m
  - Probability of someone's height less than 1.60m
  - Probability of someone's height greater than 1.70m

Probability Density Function

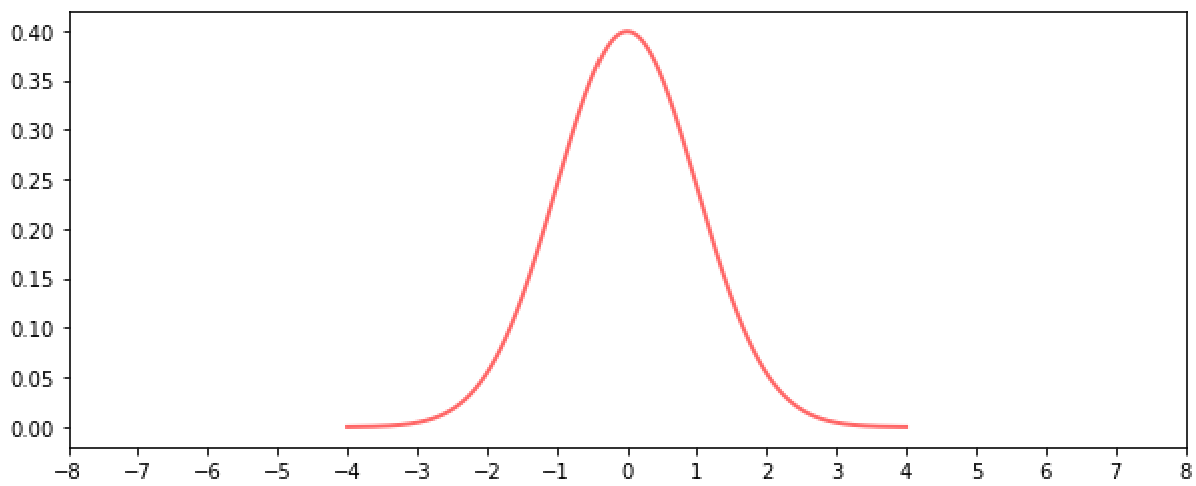


Statistical Distributions

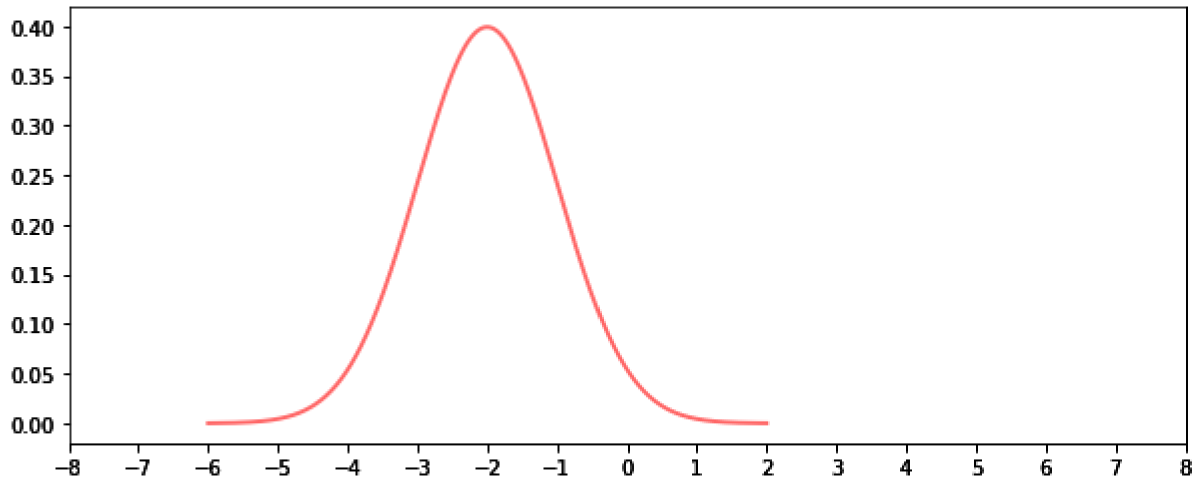
Normal Distribution

Normal Distribution

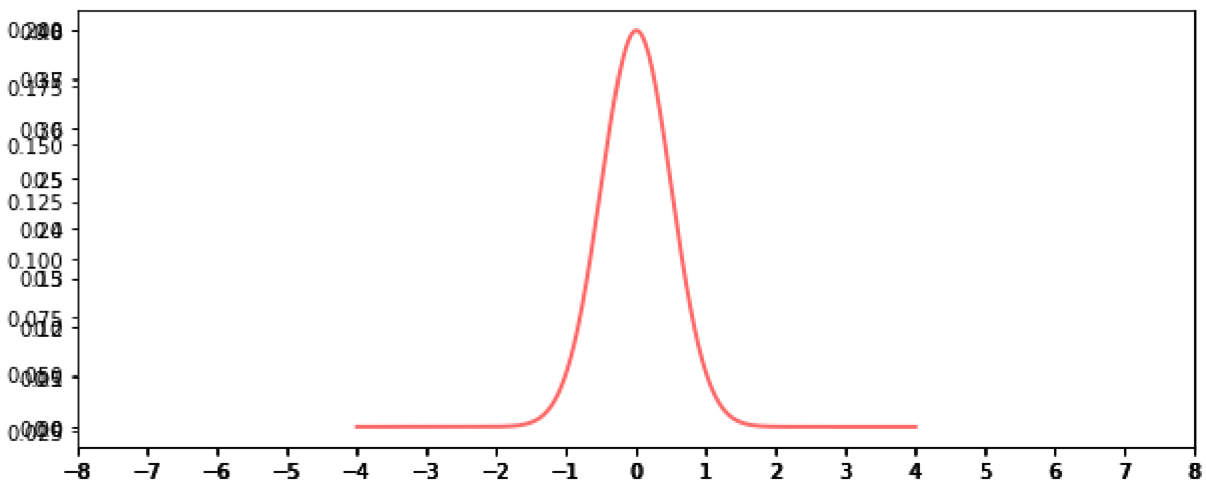
- Gaussian distribution or bell shape curve



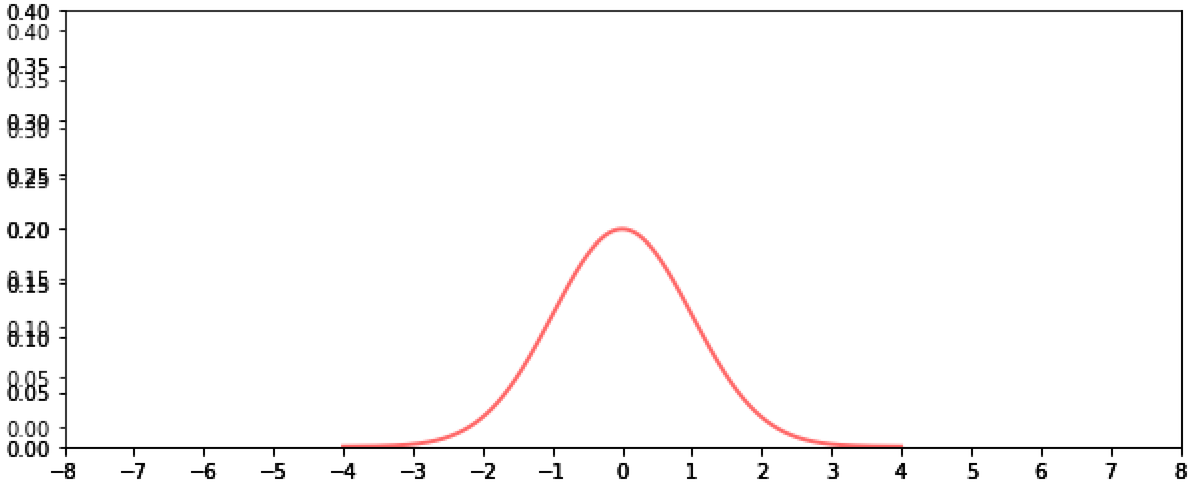
Normal Distribution



Normal Distribution

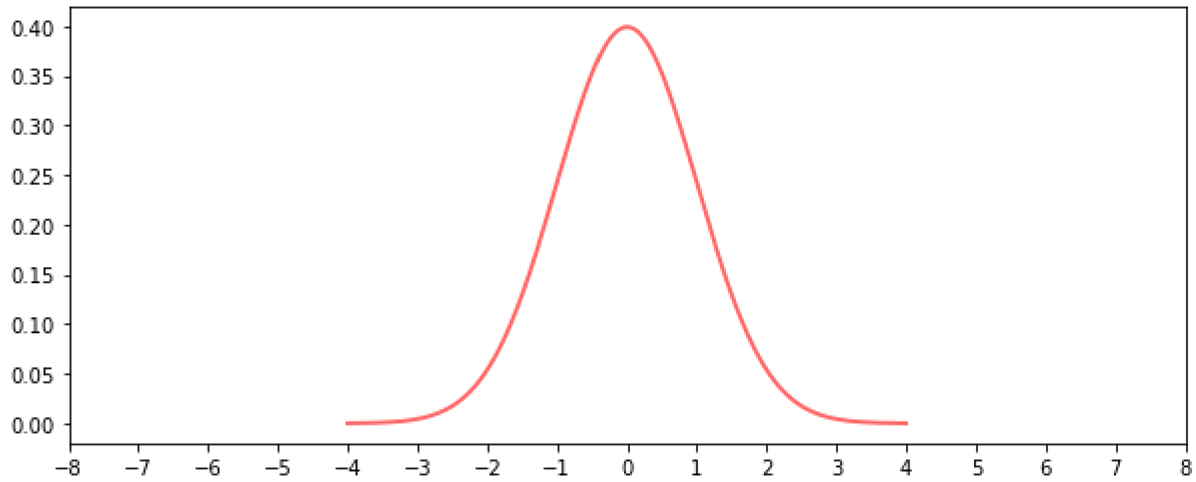


Normal Distribution



Normal Distribution

- Defined by two parameters
  - mean and standard deviation



Normal Distribution

$$\text{PDF} = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- μ is the mean
- σ is the standard deviation

Normal Distribution

- Assuming that number of customers that comes to a shop in a day can be approximated by a Normal distribution with the mean of 65 customers and standard deviation of 12 customers.
- What is the probability that on a particular day the number of customers is 50

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-65}{12}\right)^2}$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{50-65}{12}\right)^2}$$

$$0.033245 \times e^{-0.78125}$$

$$0.033245 \times 0.457833$$

0.015220

Statistical Distributions

Binomial Distribution

Binomial Distribution

- Bernoulli process is a situation where the random variable has only two mutually exclusive outcomes (success or failure)
- Coin toss → head/tail [1/2]
- Exam grade → pass/fail [1/2]
- Lucky draw → win/do not win [1/2]

Binomial Distribution

- Game of dice → win (roll a 6) / lose (otherwise)
- Probability of winning = 1/6 = 0.1667

- Probability of winning at least 4 times in 10 rolls?
- Probability of winning exactly 5 times in 10 rolls?

- Random Variable is number of times you win in 10 rolls

Binomial Distribution

- A probability distribution of x successful outcomes in n independent trials with the probability of success is p and the probability of failure is $1 - p$

$n = 10$

$p = 1/6$ (roll a 6)

$x = 5$ (winning exactly 5 times)

Binomial Distribution

- Probability mass function, P

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

$$\frac{10!}{5!(10-5)!} = 252$$

$$0.167^5 \cdot (1 - 0.167)^5 = 5.168e^{-5}$$

$$P(X = 5) = 252 \times 5.168e^{-5} = 0.013$$

Statistical Distributions

Multinomial Distribution

Multinomial Distribution

- Consider three-sided dice is tossed 10 times
- The probability of the three sides are $p_1 = 1/3, p_2 = 1/3$ and $p_3 = 1/3$
- What is the probability of getting five "1", three "2" and two "3"?

Multinomial Distribution

- n independent trials
- Each trial results in k mutually exclusive outcomes
- On a single trial, the probabilities of the k outcomes $p_1, p_2, \ldots, p_k$ where $\sum_{i=1}^{k} p_i = 1$

Multinomial Distribution

- Probability mass function, P

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1!\,x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

- n is the number of trials
- k is the number of outcomes
- x_i is the number of outcome x_i occurs
- p_i is the probability of outcome x_i occurs

$$\frac{10!}{5!3!2!} = 2520$$

$$0.333^5 \cdot 0.333^3 \cdot 0.333^2 = 1.693e^{-5}$$

$$P(X_1 = 5, X_2 = 3, X_3 = 2) = 2520 \times 1.693e^{-5} = 0.0426$$

End