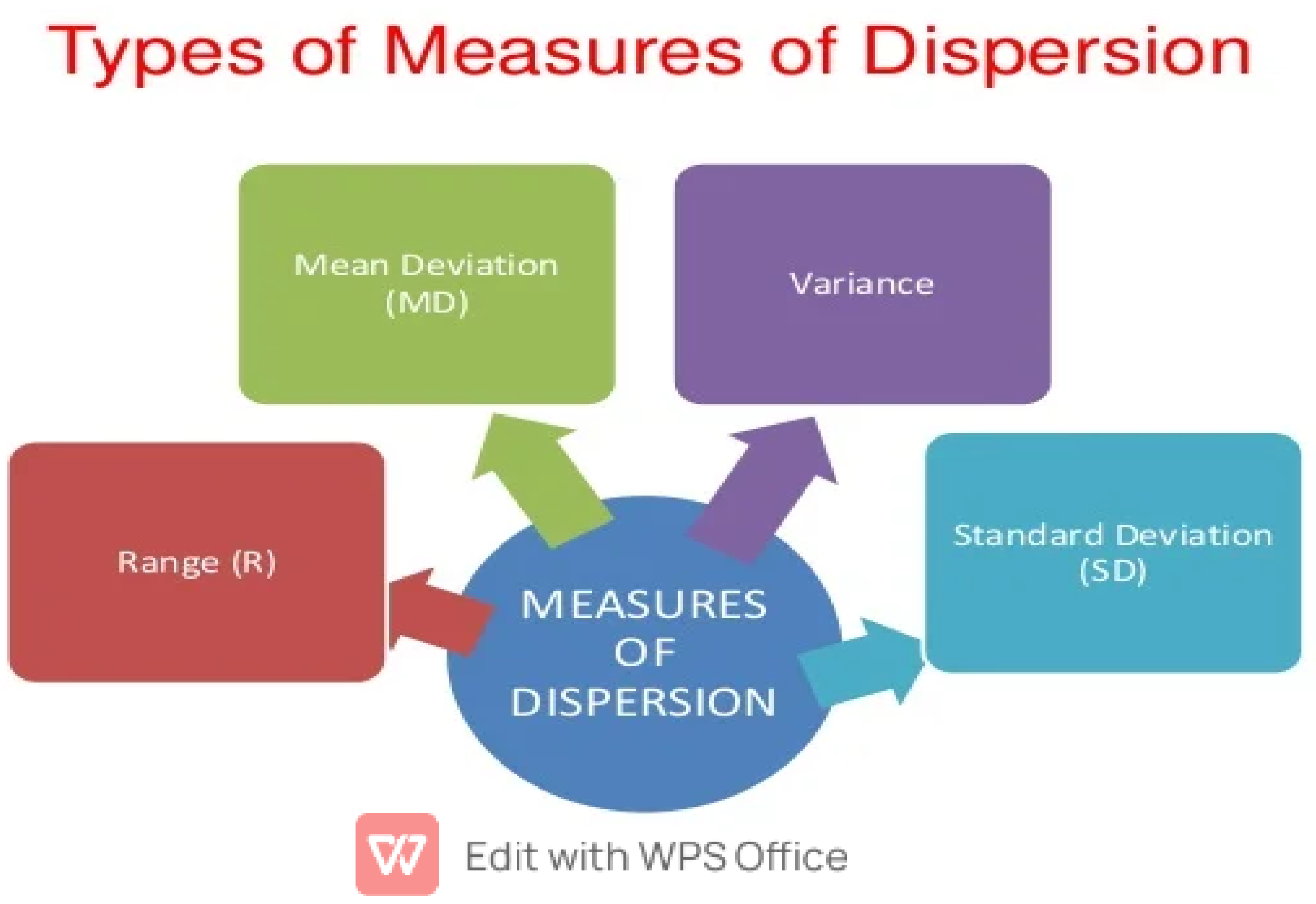
Day4

Mithun D J

Measures of dispersion
Normal Curve
Properties of a normal curve
Graphs

Measures of Dispersion

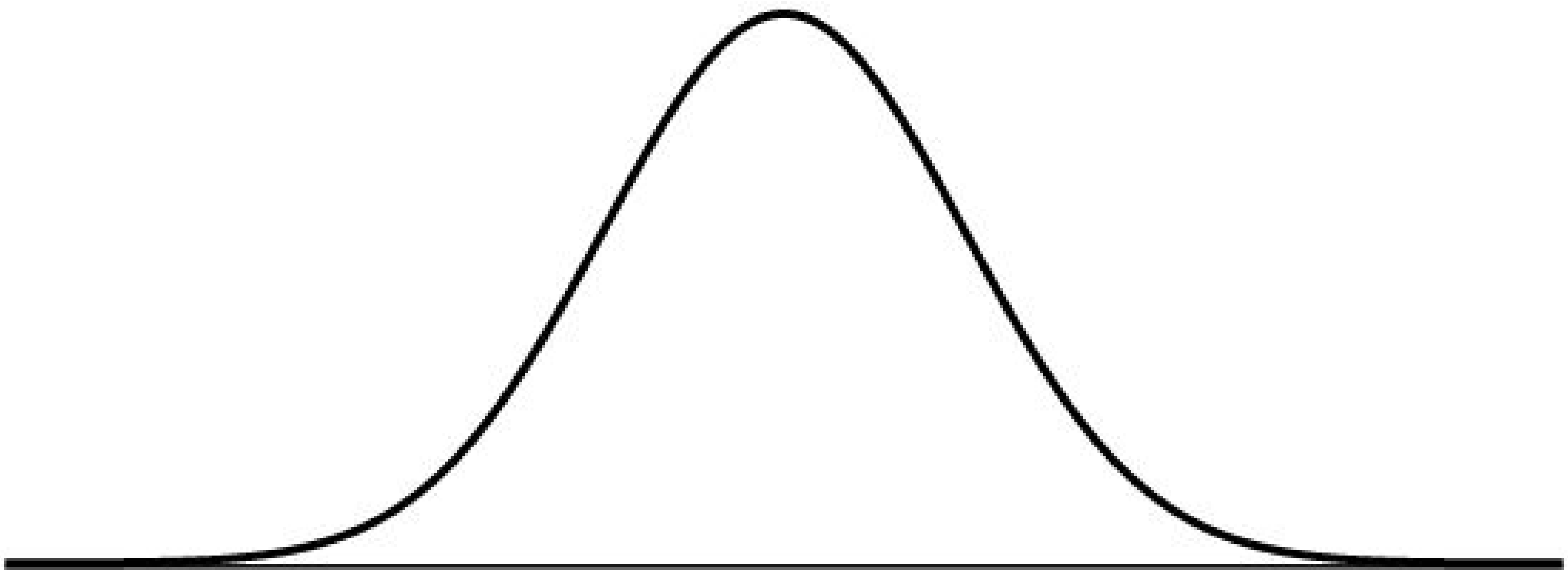


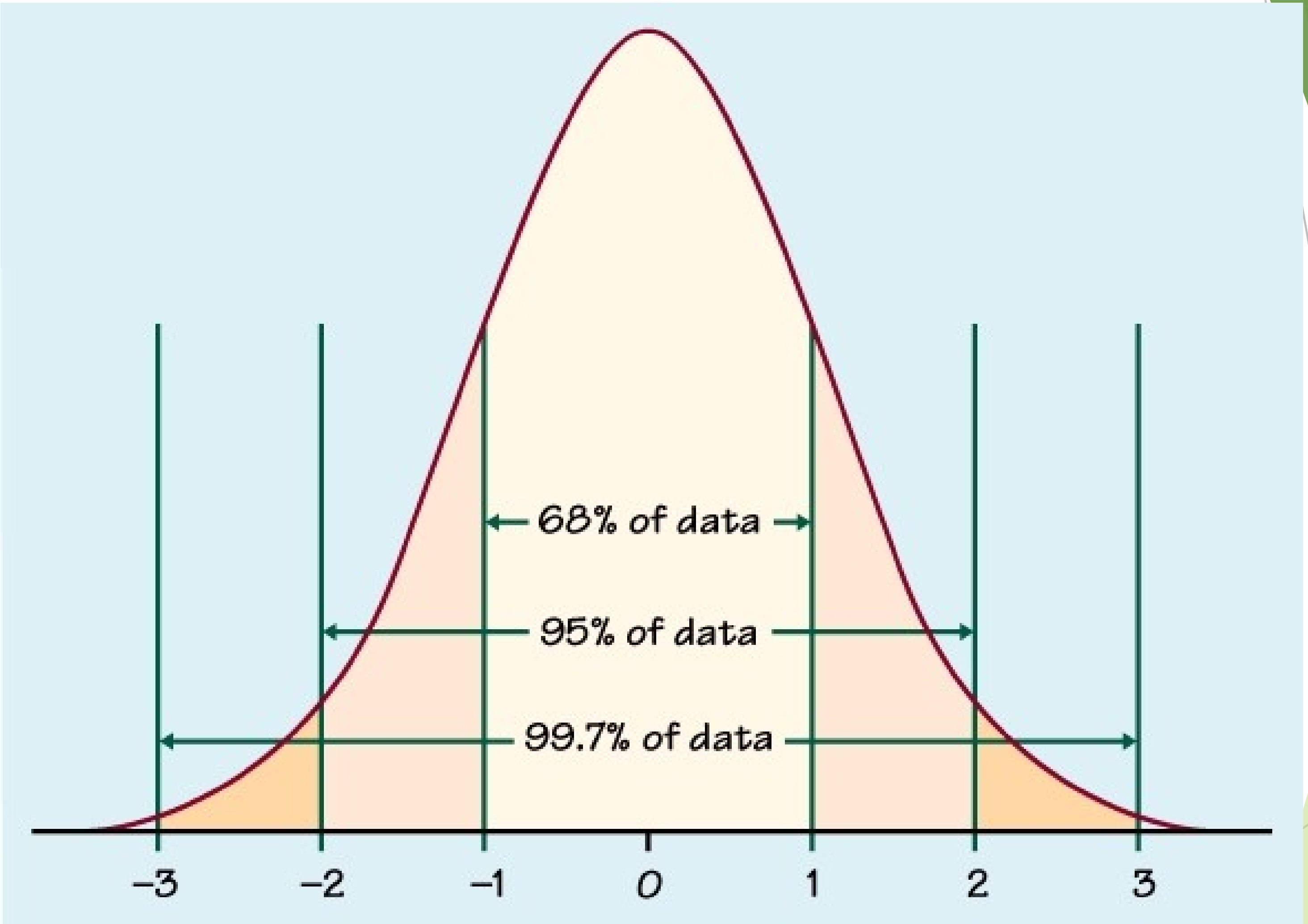
$$\text{Standard Deviation Formula} = \sqrt{\frac{\sum_{i-1}^{n} (x_i - \bar{x})^2}{N - 1}}$$

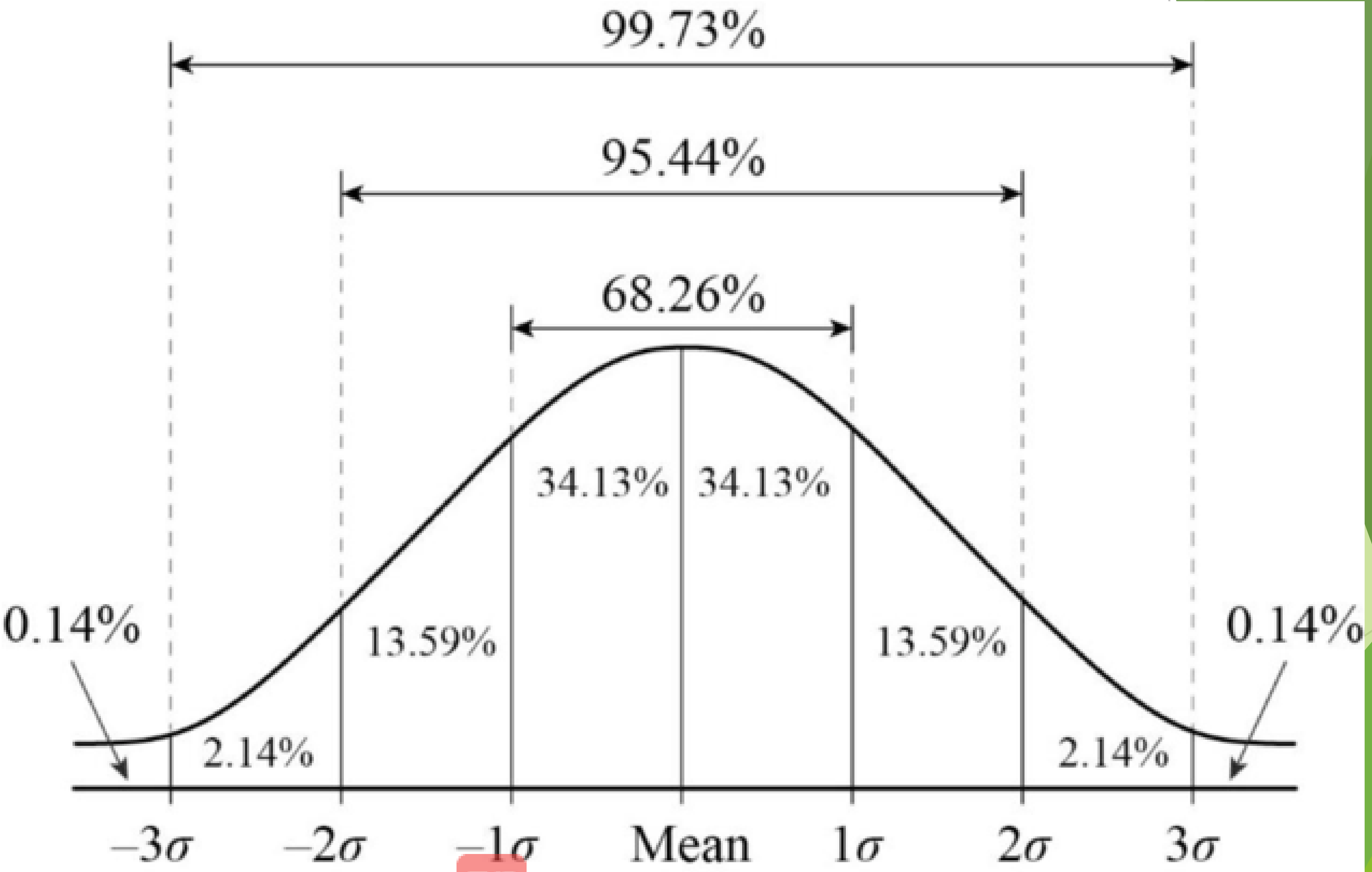
Normal Distribution

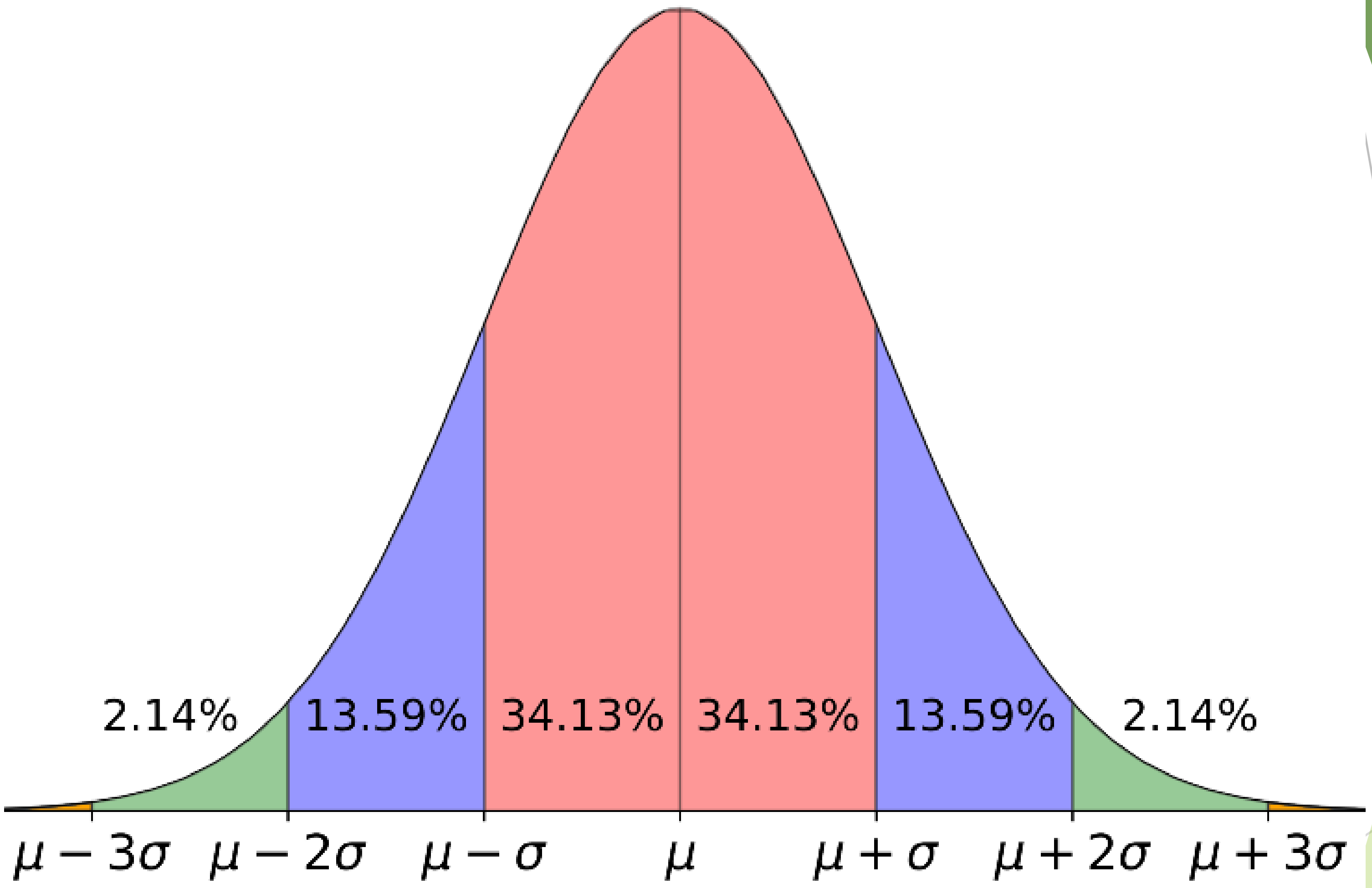
Bell shaped Curve
Mean Median Mode are same
68-95-99

Normal Curve

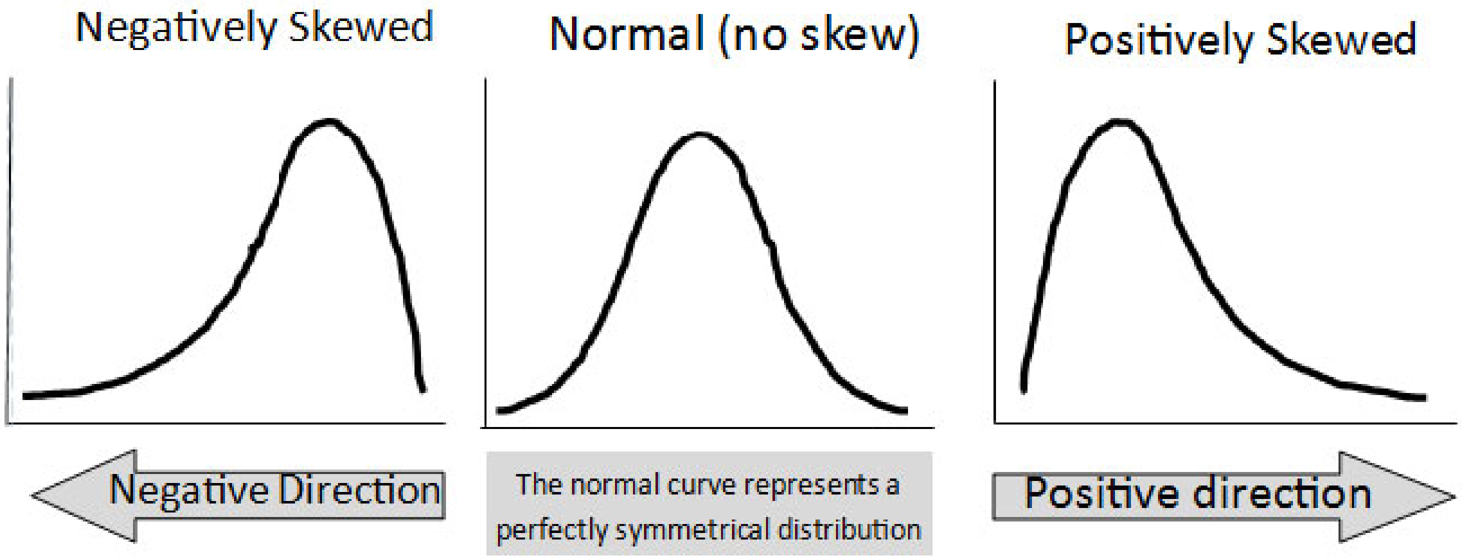


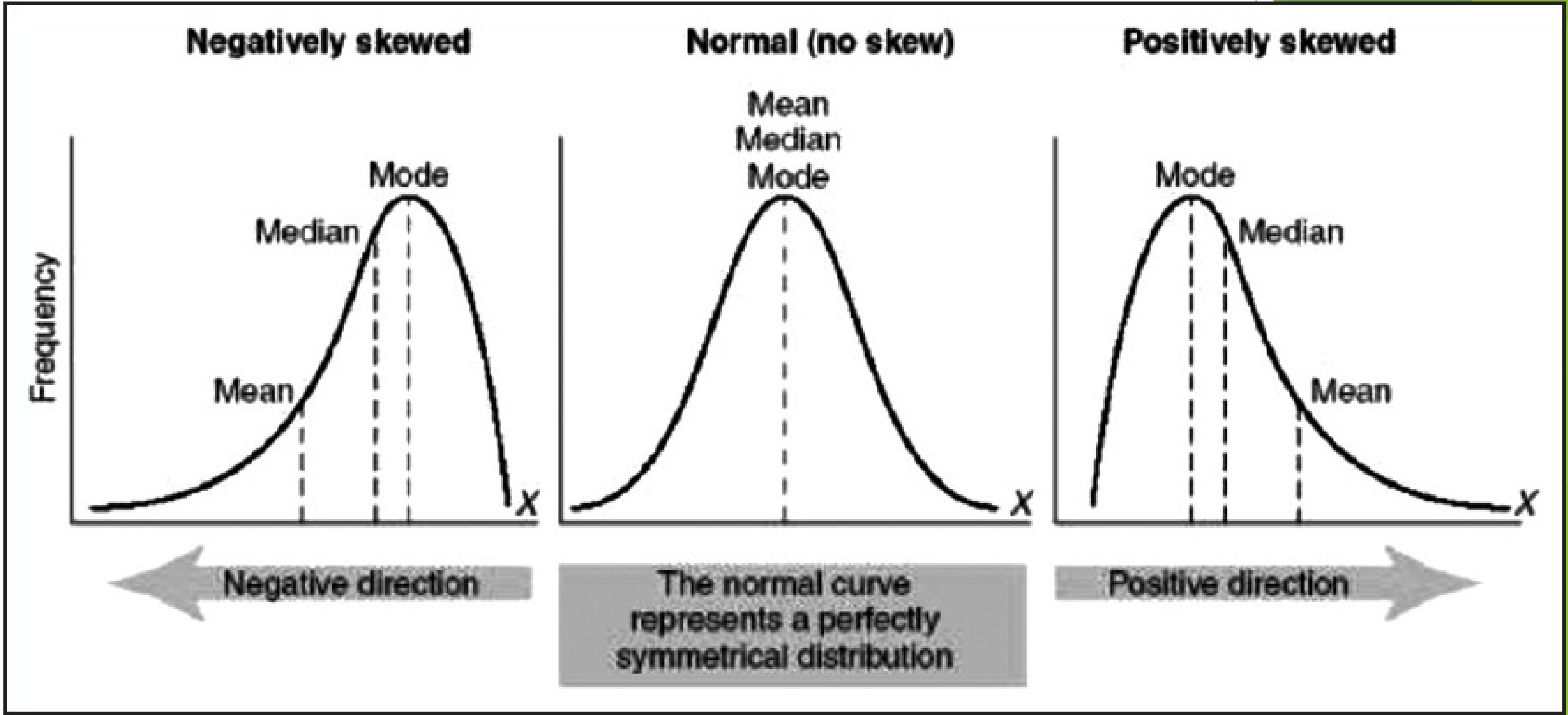

68% of data
95% of data
99.7% of data
–3
–2
–1
0
1
2
3


99.73%
95.44%
68.26%
34.13%
34.13%
0.14%
13.59%
13.59%
0.14%
2.14%
2.14%
–3σ
–2σ
–1σ
Mean
1σ
2σ
3σ

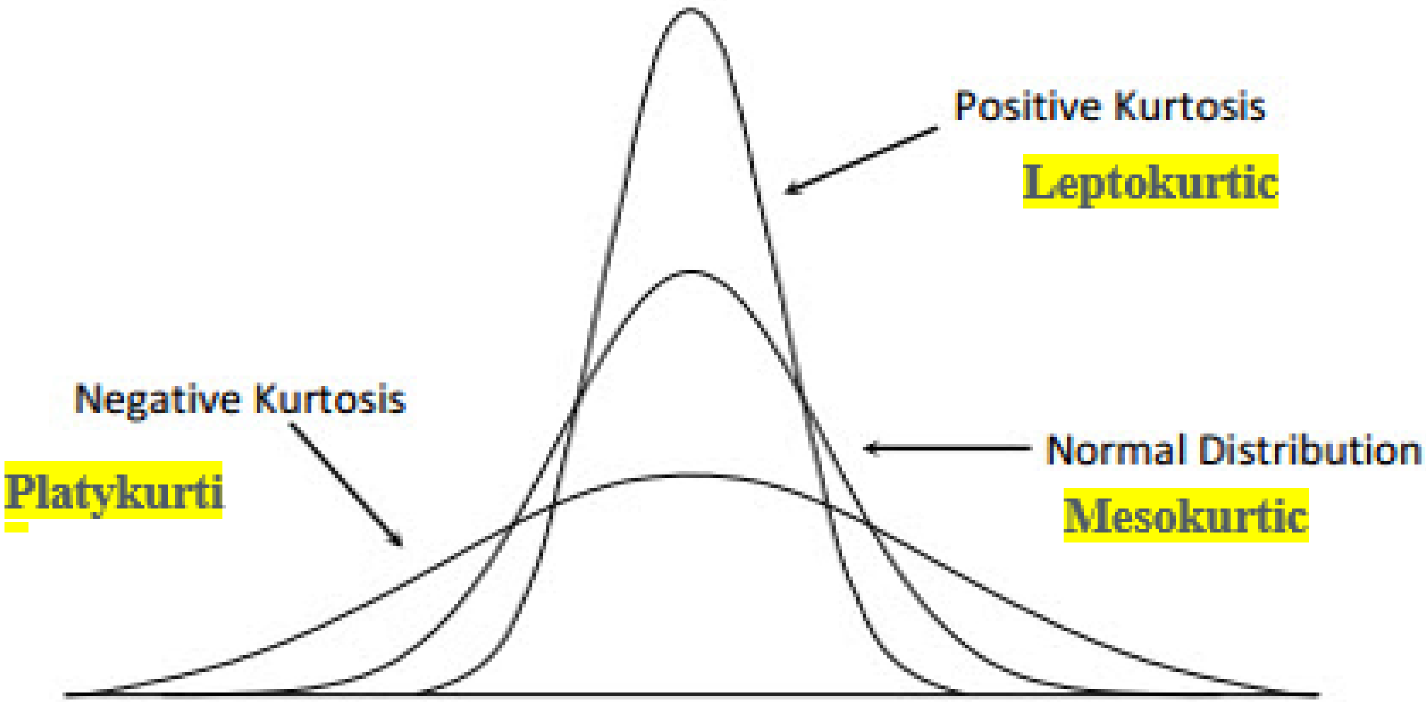

2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
μ − 3σ
μ − 2σ
μ − σ
μ
μ + σ
μ + 2σ
μ + 3σ

Skewness


Negatively Skewed
Normal (no skew)
Positively Skewed
Negative Direction
The normal curve represents a perfectly symmetrical distribution
Positive direction


Negatively skewed
Normal (no skew)
Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

Kurtosis


Positive Kurtosis
Leptokurtic
Negative Kurtosis
Platykurti
Normal Distribution
Mesokurtic

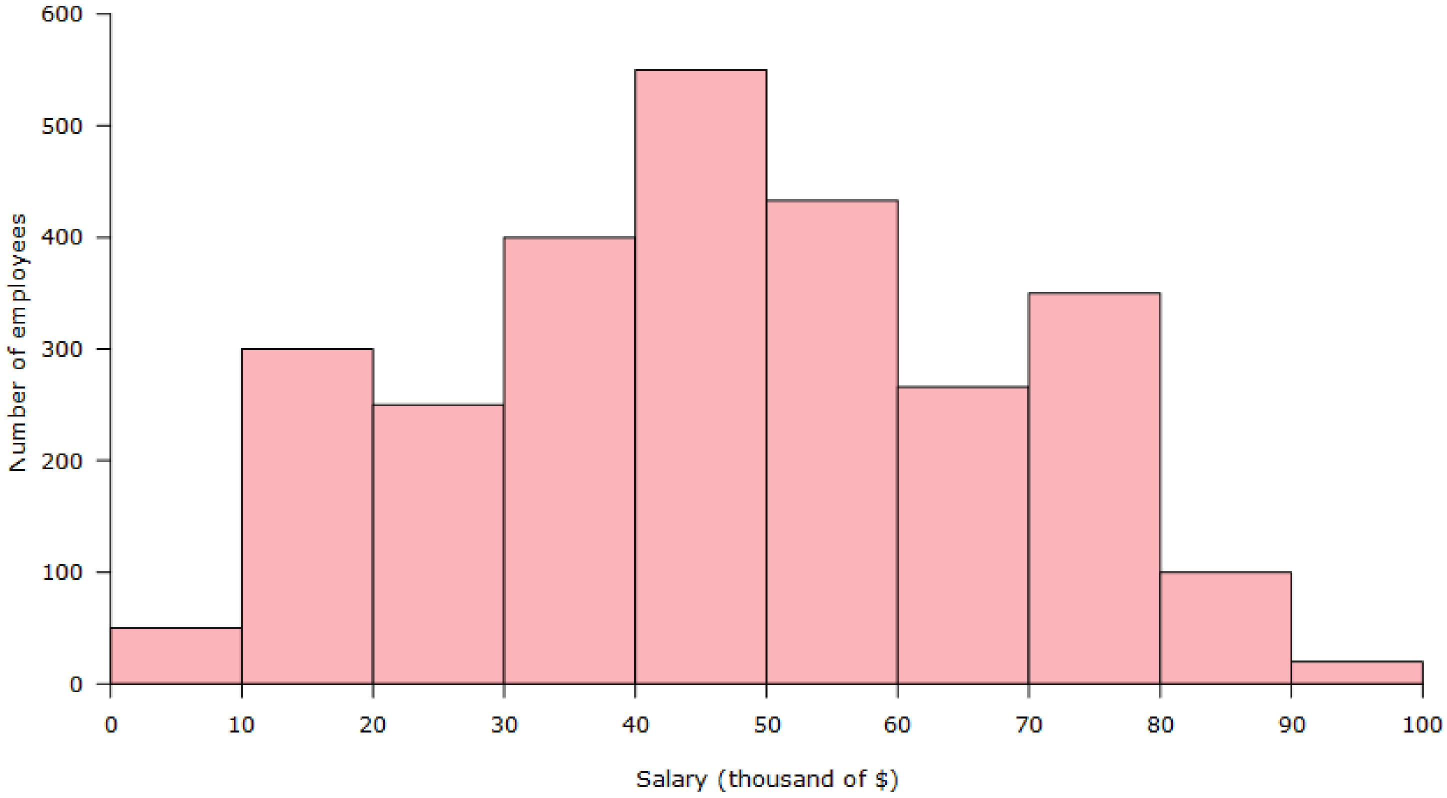
Reference Materials

1. Stats by Jim
2. Analytics Vidhya
3. Towards data science

Plots

- Histogram
- PP Plot
- QQ Plot
- Stem and Leaf Plot
- Box Plot

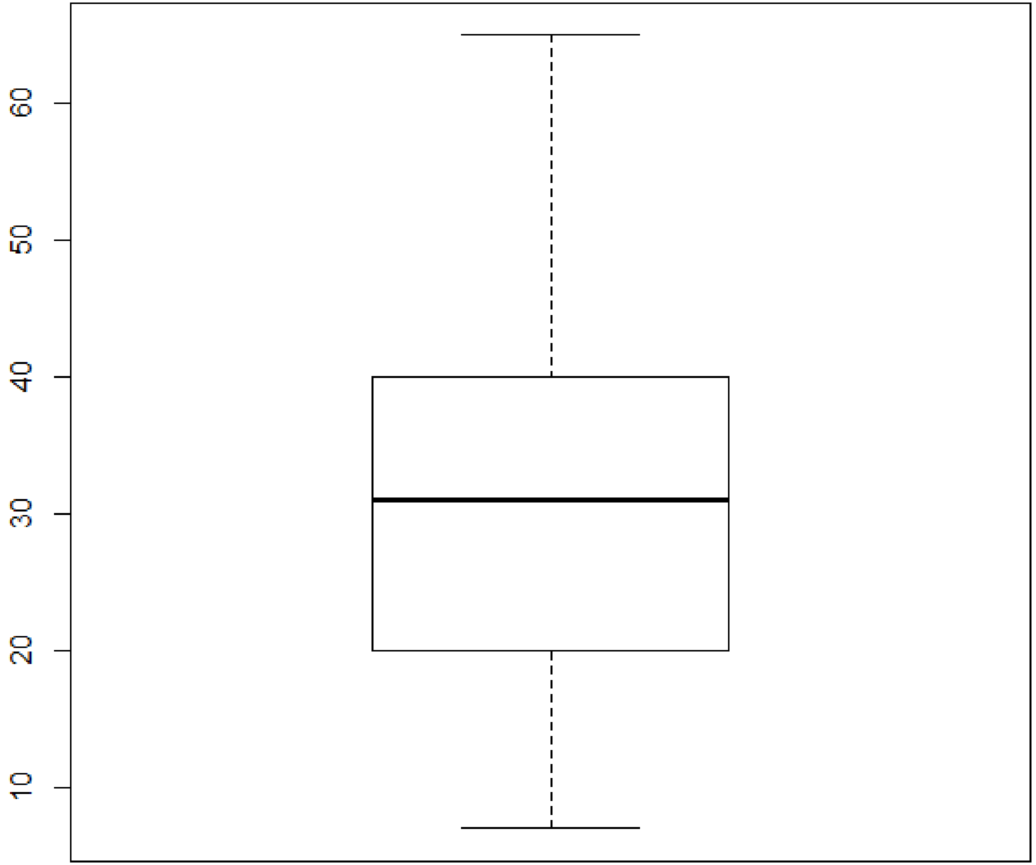
Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation



Steam and Leaf

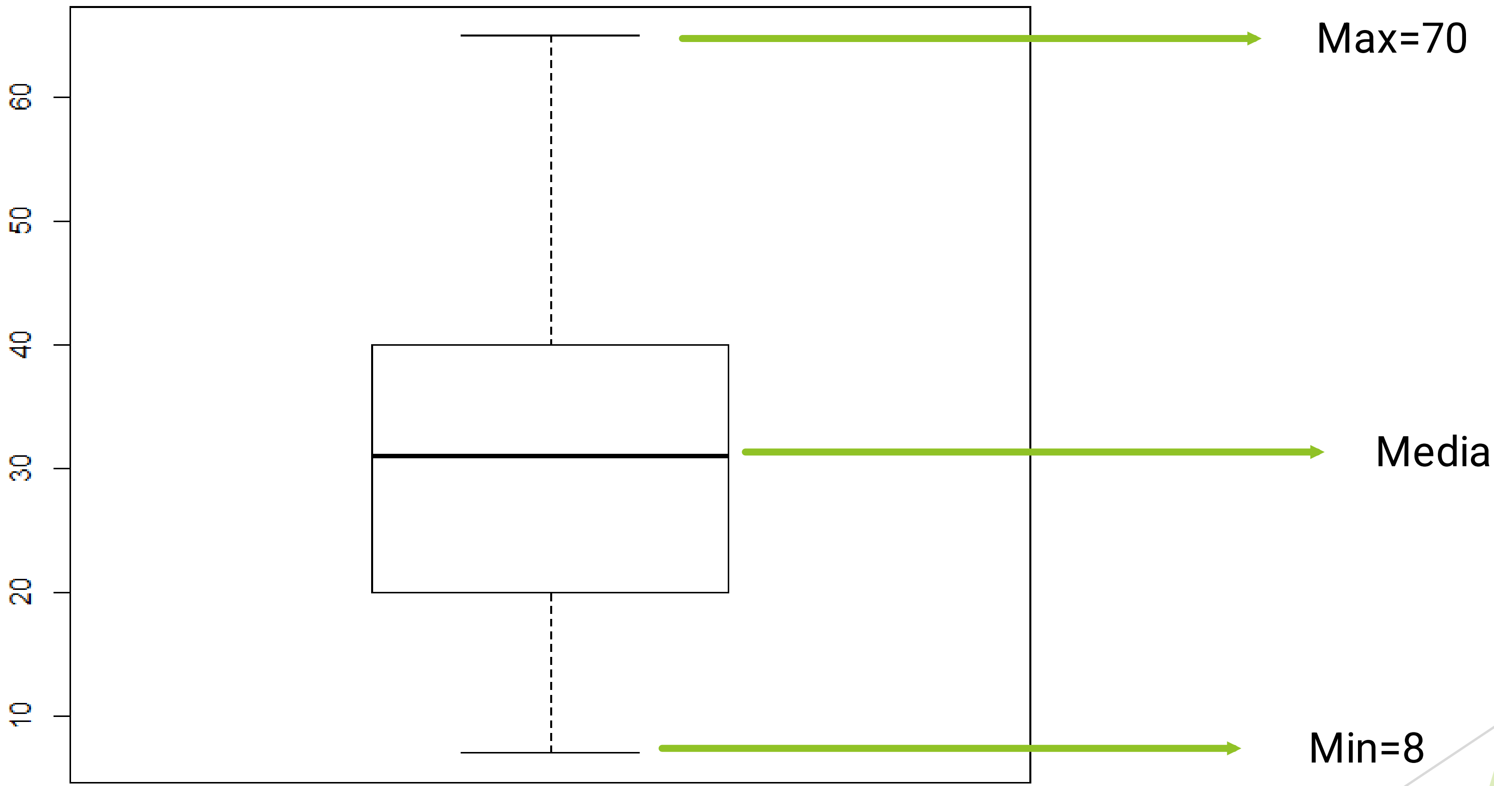
10, 15, 22, 25, 28, 23, 29, 31, 36, 45, 48

Stem	Leafs
1	0 5
2	2 5 8 3 9
3	1 6
4	5 8



60
50
40
30
20
10

Student marks: 8,9,10,15,20,25,30,35,40,45,50,55,60,70



Probability

Theoretical Probability Formula:

$$P(\text{event}) = \frac{\text{number of favorable outcomes}}{\text{number of possible outcomes}}$$

Experimental Probability Formula:

$$P(\text{event}) = \frac{\text{number of times desired outcome occurs}}{\text{number of possible outcomes}}$$

1) $0 \leq P(A) \leq 1$

2) $P(\bar{A}) = 1 - P(A)$

3) $P(\emptyset) = 0$

4) $If\ A \subseteq B,\ P(A) \leq P(B)$

5) Generally, for any two events

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

6) If the two events are mutually exclusive

- $P(A \cup B) = P(A) + P(B)$

7) If the two events are independent

- $P(A \cap B) = P(A) \times P(B)$

Terminologies

Experiment –

are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.

Outcome

is the result of a single trial. So, if it rains today, the outcome of today's trial from the experiment is "It rained"

Event

is one or more outcome from an experiment. "It rained" is one of the possible event for this experiment.

Probability

is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome "it rained" for tomorrow is 0.6

Probability distribution

Probability Distribution

- Indicates how different values of the random variable are distributed with some specified probability.
- a probability distribution is a set of all possible values that a random variable can take, along with the associated probability of each.

Random Variable

- To calculate the likelihood of occurrence of an event, we need to put a framework to express the outcome in numbers. We can do this by mapping the outcome of an experiment to numbers.

Let’s define X to be the outcome of a coin toss.

X = outcome of a coin toss

Possible Outcomes:

- 1 if heads
- 0 if tails

Random Variable and Discrete Probability Distribution

Toss two coins

outcomes :	H H	H T	T H	T T	Sample space
Random Variable X :	2	1	1	0	← x_i

Random function, X link outcomes to Real Numbers.

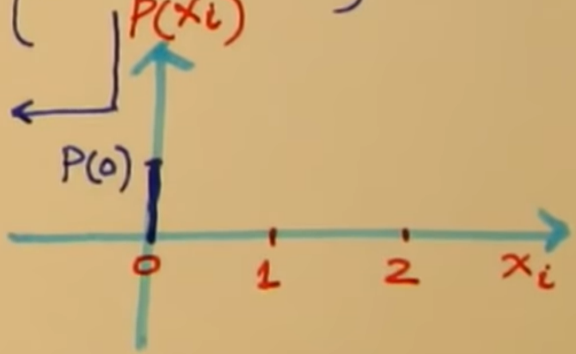
X : {Number of heads}

{Sample space} $\longrightarrow \boxed{f} \longrightarrow$ {Real Numbers.}

Probability Distribution

x_i	0	1	2
$P(x_i)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

- $\sum P(x_i) = \frac{1}{4} + \frac{2}{4} + \frac{1}{4} = 1$
- $P(x_i) \geq 0$



Data Types

- **Discrete Data**, as the name suggests, can take only specified values. For example, when you roll a die, the possible outcomes are 1, 2, 3, 4, 5 or 6 and not 1.5 or 2.45.

Example: The number of students in a class, number of workers in a company, etc.

- **Continuous Data** can take any value within a given range. The range may be finite or infinite. For example, A girl's weight or height, the length of the road. The weight of a girl can be any value from 54 kgs, or 54.5 kgs, or 54.5436kgs.

Examples: A person's height, Time, distance, etc.

Discrete distributions

- Uniform distribution
- Binomial distribution
- Bernoulli distribution
- Poisson distribution

Continuous distributions

- Normal distribution
- Standard Normal distribution
- Student's T distribution
- Chi-squared distribution

Important Terminologies

Terminologies

- **Probability Density Function (PDF)**

It is a statistical term that describes the probability distribution of a continuous random variable.

- **Probability Mass Function (PMF)**

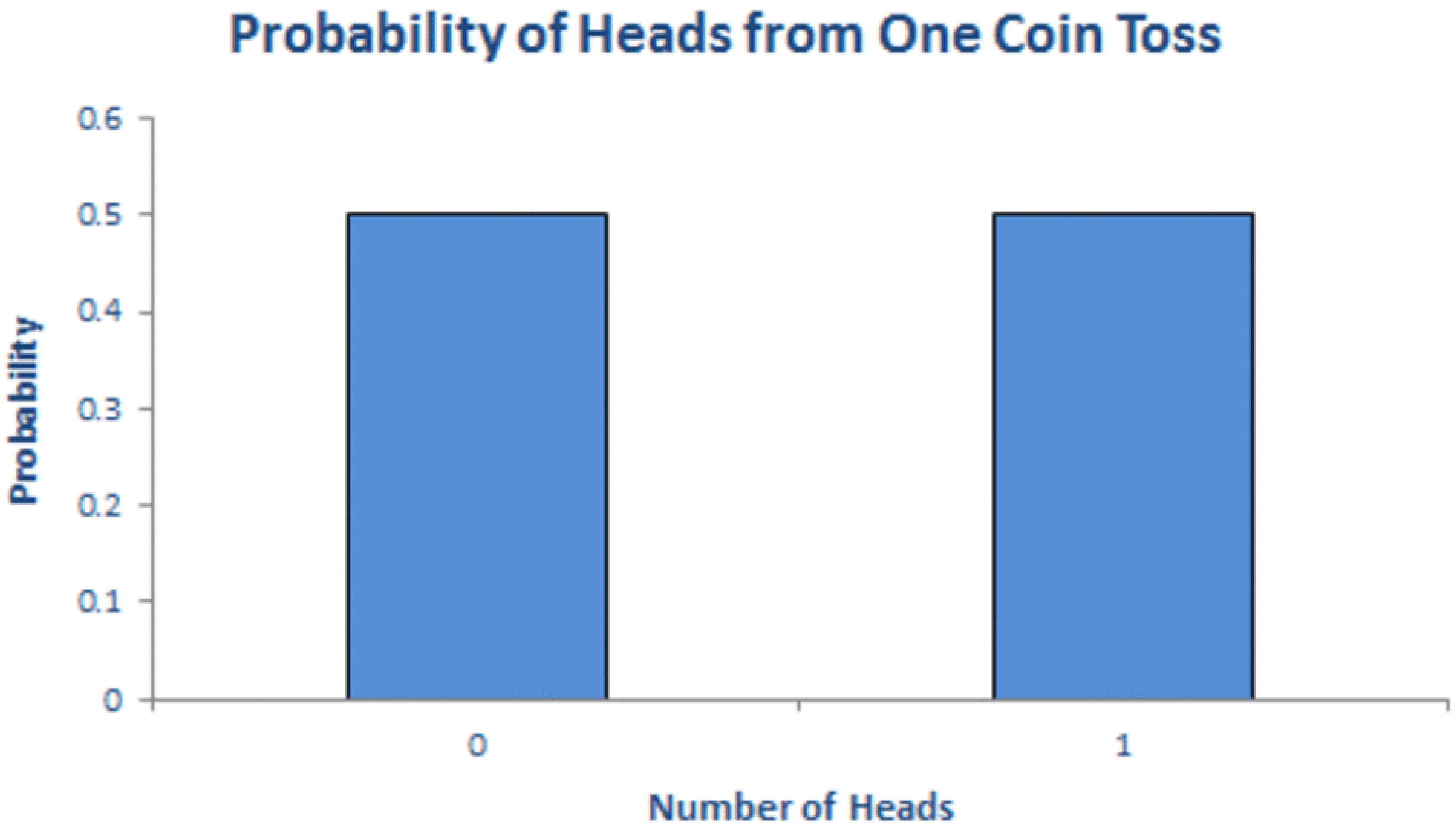
It is a statistical term that describes the probability distribution of a discrete random variable.

Bernoulli Distribution

Bernoulli Distribution

Background:

- All you cricket junkies out there! At the beginning of any cricket match, how do you decide who is going to bat or ball? A toss! It all depends on whether you win or lose the toss, right? Let's say if the toss results in a head, you win. Else, you lose. There's no midway.


Probability of Heads from One Coin Toss
0.6
0.5
0.4
0.3
0.2
0.1
0
Probability
0
1
Number of Heads

What is a Bernoulli Distribution?

- A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.
- Here, the occurrence of a head denotes success, and the occurrence of a tail denotes failure.
  Probability of getting a head = 0.5 = Probability of getting a tail since there are only two possible outcomes.

Examples of a Bernoulli Distribution

- Will you pass or fail a test?
- Will your favorite sports team win or lose their next match?
- Will you be accepted or rejected for that job you applied for?
- Will you roll a six in the opening round of your favorite board game?
- Will you win or lose the lottery?

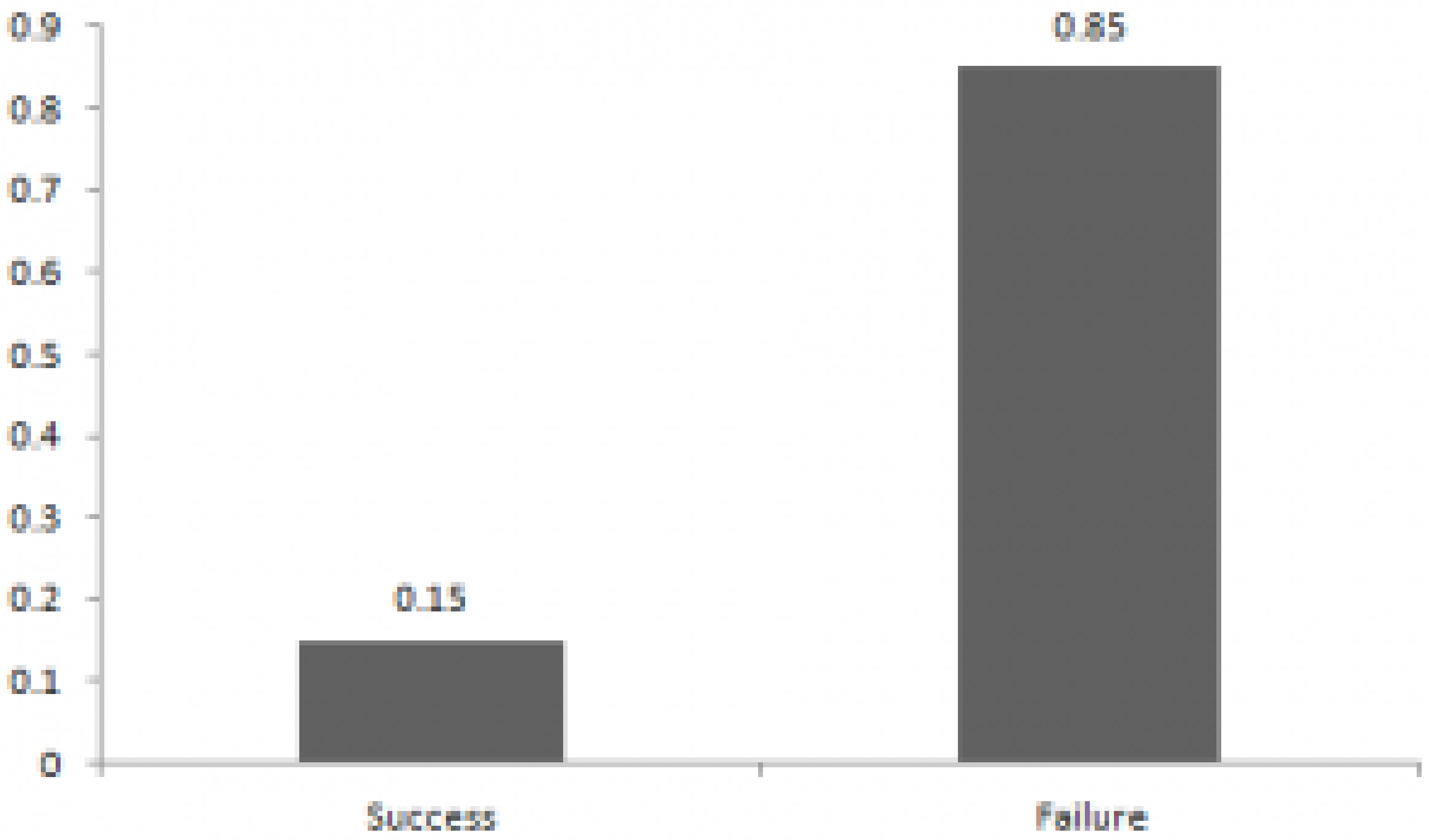
Properties of Bernoulli Distribution

PMF	$p^x(1-p)^{1-x}$
Mean	p
Variance	$pq = p(1-p)$

PMF of Bernoulli Distribution

$$P(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

The probabilities of success and failure need not be equally likely, like the result of a fight between you and Undertaker. He is pretty much certain to win. So in this case probability of your success is 0.15 while my failure is 0.85
Here, the probability of success(p) is not same as the probability of failure. So, the chart below shows the Bernoulli Distribution of our fight.


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.15
0.85
Success
Failure

Problems on Bernoulli Distribution

Binomial Distribution

Binomial Distribution

- **$X \sim B(n, p)$**
- X is a discrete random variable that follows Binomial distribution with parameters n, p.
- n is the no. of trials,
- p is the success probability for each trial.
- Binomial distribution is a discrete probability distribution of the number of successes in 'n' independent experiments sequence. The two outcomes of a Binomial trial could be Success/Failure, Pass/Fail/, Win/Lose, etc.
- Generally, the outcome **success** is denoted as **1,** and the probability associated with it is **p**.
- And **Failure** is denoted as **0**, and the probability associate with it is **q = 1-p**.

PMF of a Binomial Distribution

$$P(X) = {}_nC_x \, p^x q^{n-x}$$

- **P** is the probability of success on any trail.
- **q = 1- P** – the probability of failure
- **n** – the number of trails/experiments
- **x** – the number of successes, it can take the values 0, 1, 2, 3, . . . n.
- $_nC_x$ **= n!/x!(n-x)** and denotes the number of combinations of **n** elements taken **x** at a time.

Bernoulli Distribution	**Binomial Distribution**
Bernoulli distribution is used when we want to model the outcome of a single trial of an event.	If we want to model the outcome of multiple trials of an event, Binomial distribution is used.
It is represented as X ~~ Bernoulli (p). Here, p is the probability of success.	It is denoted as X ~~ Binomial (n, p). Where n is the number of trials.
Mean, E[X] = p Variance, Var[X] = p(1-p)	Mean, E[X] = np Variance, Var[X]= np(1-p)
Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then the Bernoulli distribution can be used to model the passing or failing in such an exam.	Example: Suppose the probability of passing an exam is 80% and failing is 20%. Then if we want to find the probability that a student will pass in exactly 4 out of 5 exams, we use the Binomial Distribution.

Poisson Distribution

Poisson Distribution

- Suppose you work at a call center, approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are

1. The **number of** emergency calls recorded at a hospital in a day.
2. The **number of** thefts reported in an area on a day.
3. The **number of** customers arriving at a salon in an hour.
4. The **number of** suicides reported in a particular city.
5. The **number of** printing errors at each page of the book.

Poisson distribution

- A distribution is called **Poisson distribution** when the following assumptions are valid:

Assumptions:

- Any successful event should not influence the outcome of another successful event.
- The probability of success over a short interval must equal the probability of success over a longer interval.
- The probability of success in an interval approaches zero as the interval becomes smaller.

Notation used in Poisson Distribution

- λ is the rate at which an event occurs,
- t is the length of a time interval,
- And X is the number of events in that time interval.

$$P(x) = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

Properties of Poisson Distribution

Probability Mass Function	$P(X) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$
Mean	$\mu = E(X) = \lambda$
Variance	$\sigma^2 = V(X) = \lambda$
Standard Deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\lambda}$

$0 \le x \le \infty$

λ is the average number of occurrences in an interval

Exponential Distribution

Exponential Distribution

- Let’s consider the call center example one more time. What about the interval of time between the calls ? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.
- Other examples are:

1. Length of time between metro arrivals,
2. Length of time between arrivals at a gas station
3. The life of an Air Conditioner

Exponential Distribution

- From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.
- A random variable X is said to have an **exponential distribution** with PDF:
  - $f(x) = \{ \lambda e^{-\lambda x}, \ x \geq 0$
- and parameter **λ>0** which is also called the rate.
- For Reliability analysis, λ is called the failure rate of a device at any time t, given that it has survived up to t.
- Mean and Variance of a random variable X following an exponential distribution:
  - Mean -> $E(X) = 1/\lambda$
  - Variance -> $Var(X) = (1/\lambda)^2$

Application

- Predict the time when an Earthquake might occur.
- Life Span of Electronic Gadgets.
- Predict the time duration between the arrival of two consecutive customers.
- Average Time a Call Centre Employee Spends With the Customer

Working Rule to solve problem for Exponential Distribution

- $P(X \leq x) = 1 - e^{-\lambda x}$

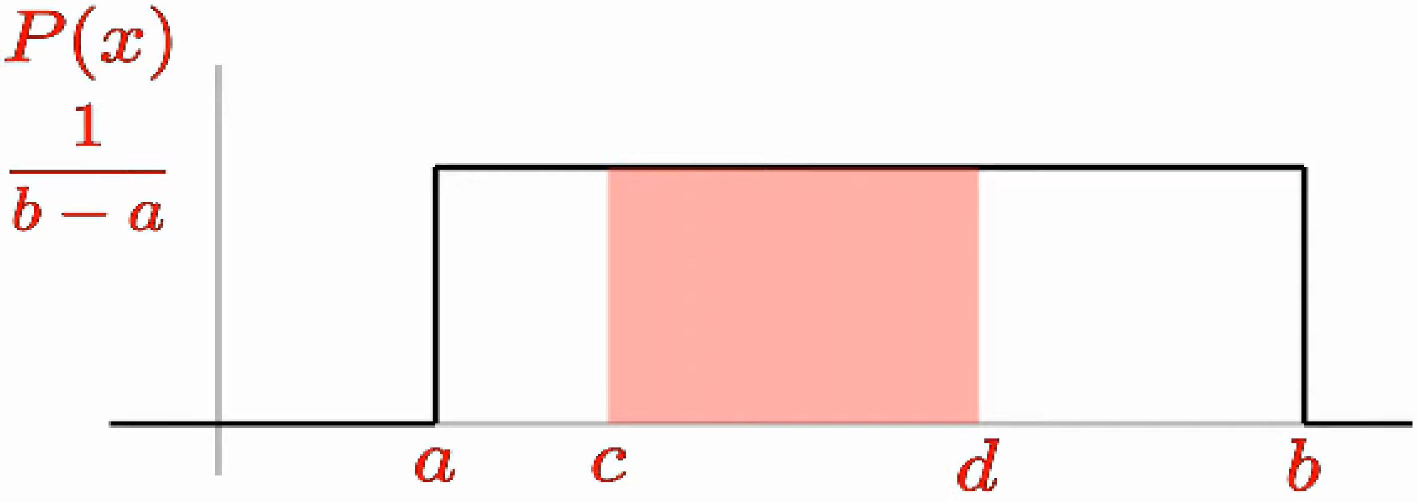
Uniform Distribution

Uniform Distribution

- A uniform distribution is characterized by a constant probability within a certain domain
- When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are **equally likely** and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.
- A variable X is said to be uniformly distributed if the density function is f(x)=1/b-a

Uniform Distribution

Uniform Distribution



$$Mean:\ \mu = \frac{a+b}{2}$$

$$S.D.:\ \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Probability

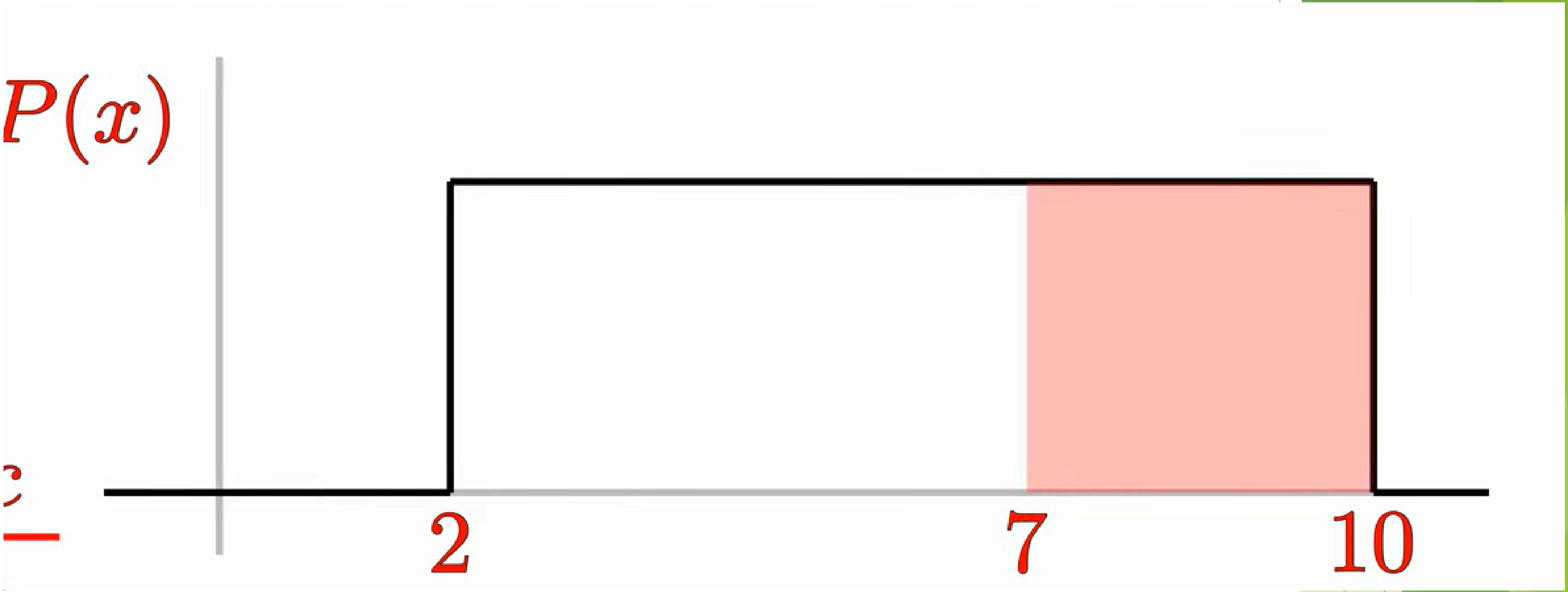
$$P(c \le X \le d) = \frac{d-c}{b-a}$$


f(x)
$\frac{1}{b-a}$
a
b
x

$$f(x) = \frac{1}{b-a}$$

$$for\ -\infty < a \le x \le b < \infty$$

Problems


$P(x)$
2
7
10

Problems

Normal Distribution

Normal Distribution

- **Normal distribution** represents the behavior of most of the situations in the universe (That is why it's called a "normal" distribution.
- The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application.
- Any distribution is known as Normal distribution if it has the following characteristics:

Properties

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line $x=\mu$.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

- Gaussian PDF:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standard Normal Distribution

- Mean=0
- Variance=1

Normal Distribution

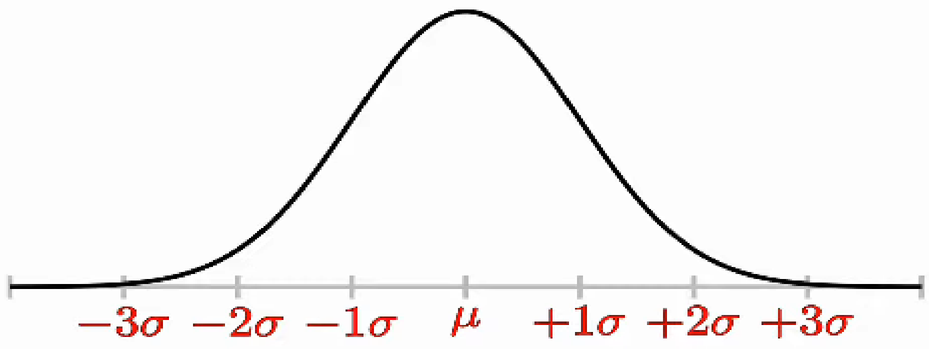
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty \le x \le \infty$$

$$= \frac{1}{1\times\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} \qquad -\infty \le x \le \infty$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} \qquad -\infty \le z \le \infty \qquad \left(\frac{x-0}{1} = z\right)$$

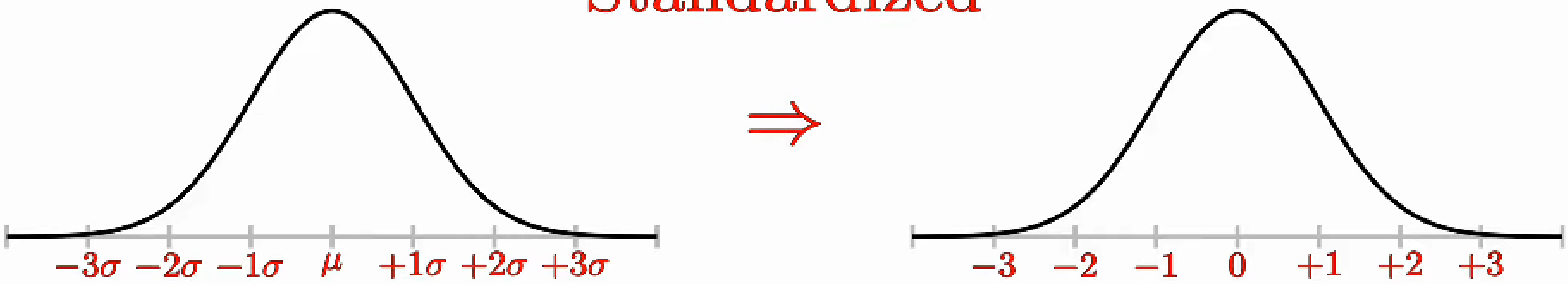
Standard Normal Distribution (Mean=0, Variance=1)

Normal Distribution



Normal Distribution

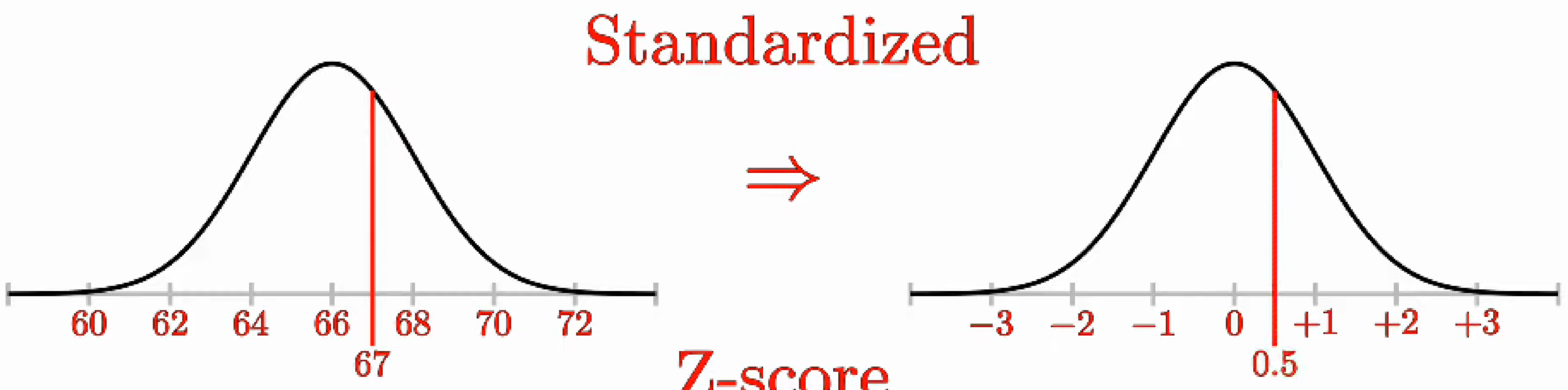
Standardized



<u>Z-score</u>

$$z = \frac{x - \mu}{\sigma}$$

Normal Distribution



Z-score

$$z = \frac{x - \mu}{\sigma}$$

Ex

Heights:

$\mu = 66$ in.
$\sigma = 2$ in.

$x = 67$ in. $\Rightarrow$ $z = \dfrac{67 - 66}{2} = \dfrac{1}{2} = 0.5$

Normal Distribution

Probability $\Rightarrow$ Use Chart

z	.00	.01	.02	.03	.04	.05	.06	.07
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418

Normal Distribution

Probability

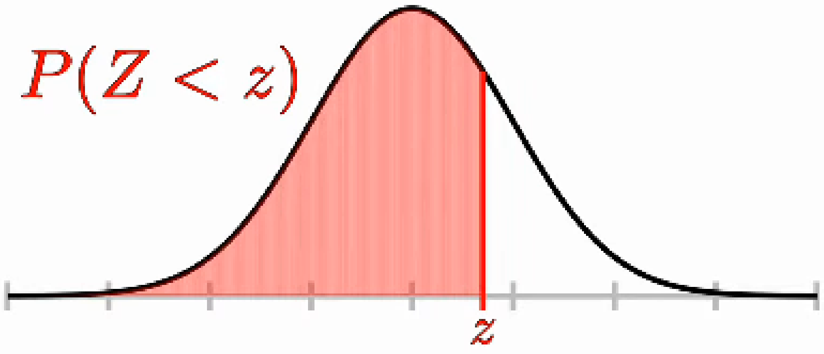
$$z = 0.63$$

z	.00	.01	.02	.03	.04	.05	.06	.07
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418

Normal Distribution

Probability



z	.00	.01	.02	.03	.04	.05	.06	.07
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418

Working Rule to solve problems

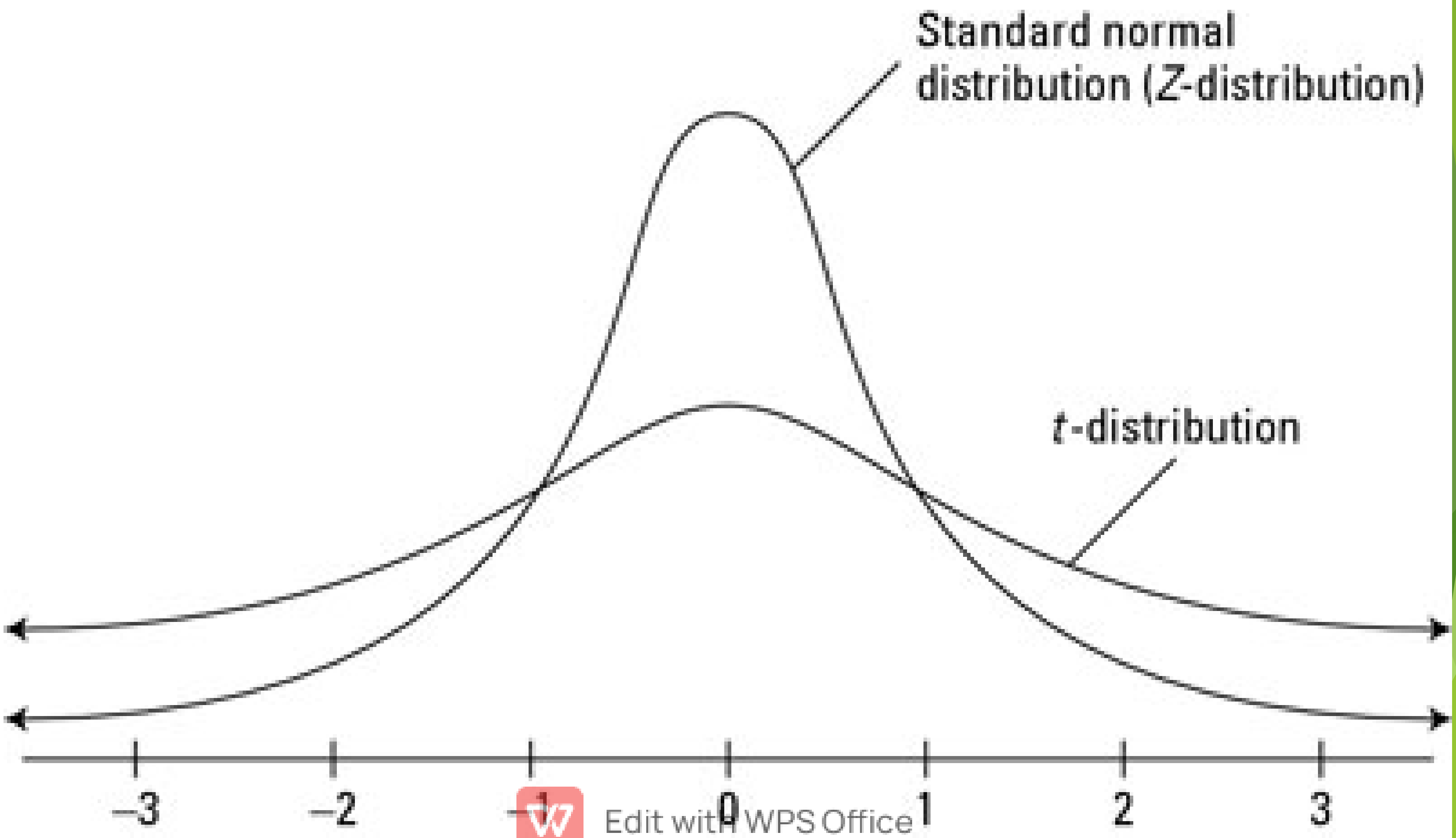
- **Step1:**
  - Convert x value to z score (Subtract mean and divide by std dev)
- **Step2:**
  - Look up the Standard normal table
- **Step3:**
  - Obtain the area corresponding to the z score from the table (Note: By default the Standard normal table gives the area to the left of z score).

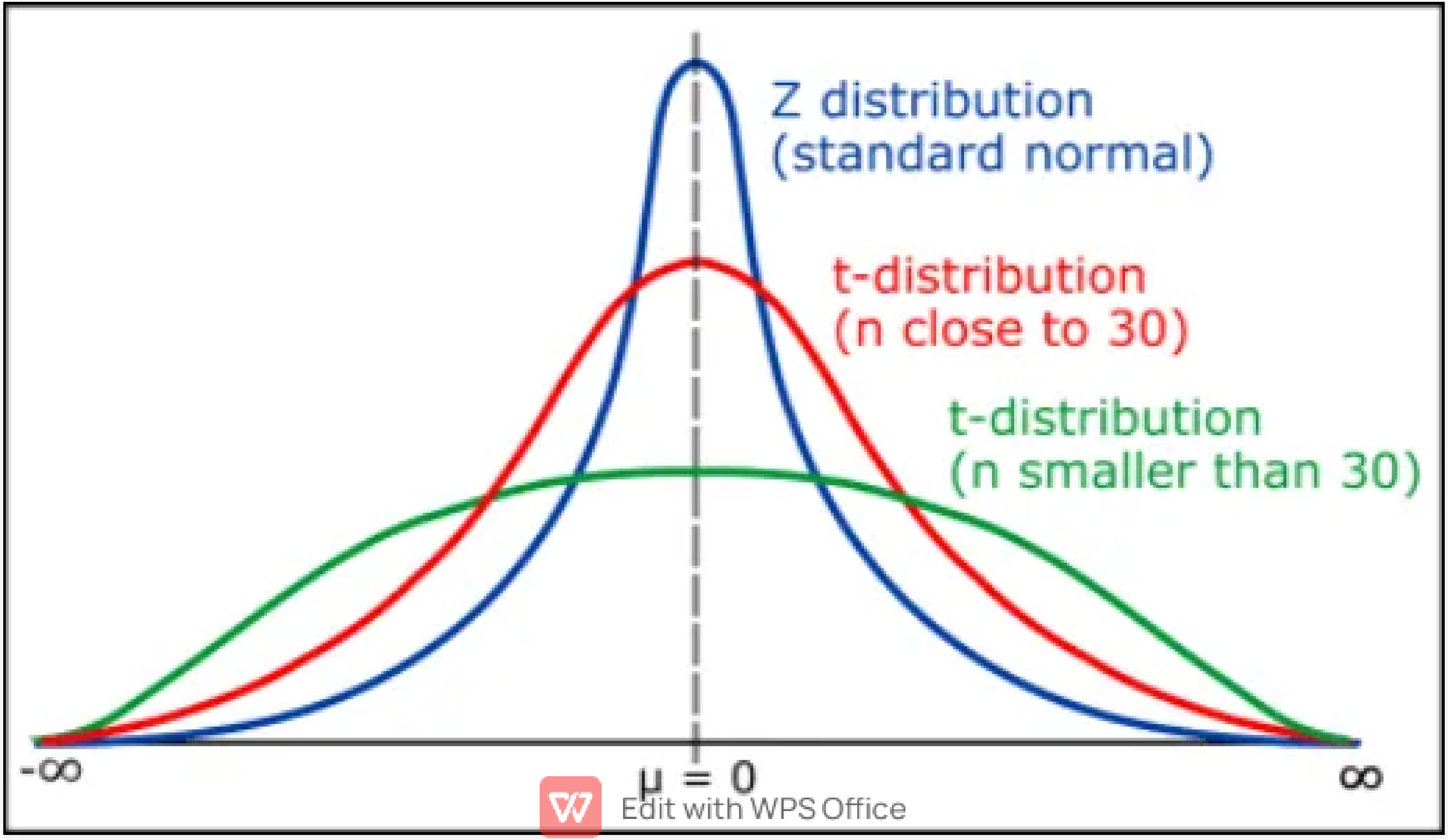
Problems

T Distribution

Introduction

- In probability and statistics, the normal distribution is a bell-shaped distribution whose **mean is μ and the standard deviation is σ**.
- The *t*-distribution is a type of normal distribution that is used for smaller sample sizes.
- Standard Deviation is unknown
- It is used to make presumptions about a mean when the standard deviation is not known to us.
- It is symmetrical, bell-shaped distribution, similar to the standard normal curve.


Standard normal distribution (Z-distribution)
t-distribution
−3
−2
−1
0
1
2
3


Z distribution
(standard normal)
t-distribution
(n close to 30)
t-distribution
(n smaller than 30)
-∞
μ = 0
∞

Properties

- It ranges from $-\infty$ to $+\infty$.
- It has a bell-shaped curve and symmetry similar to normal distribution.
- The shape of the t-distribution varies with the change in degrees of freedom.
- The variance of the t-distribution is always greater than '1' and is limited only to 3 or more degrees of freedom. It means this distribution has a higher dispersion than the standard normal distribution.

T distribution Formula

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

Where, $\bar{x}$ is the mean of first sample.

μ is the mean of second sample.

$\frac{s}{\sqrt{N}}$ = the estimate of the standard error of difference between the means.

Applications

- Testing for the hypothesis of the population mean
- Testing for the hypothesis of the difference between two means.

Chi squared Distribution

Introduction

- A chi-square distribution is a continuous probability distribution.
- The shape of a chi-square distribution depends on its degrees of freedom, k.
- **The mean of a chi-square distribution is equal to its degrees of freedom (k) and the variance is 2k**.
- Very few real-world observations follow a chi-square distribution. The main purpose of chi-square distributions is hypothesis testing, not describing real-world distributions.

Properties

- Imagine taking a random sample of a standard normal distribution (Z). If you squared all the values in the sample, you would have the chi-square distribution with $k = 1$.

  $X^2_1 = (Z)^2$

- Now imagine taking samples from two standard normal distributions (Z_1 and Z_2). If each time you sampled a pair of values, you squared them and added them together, you would have the chi-square distribution with $k = 2$.

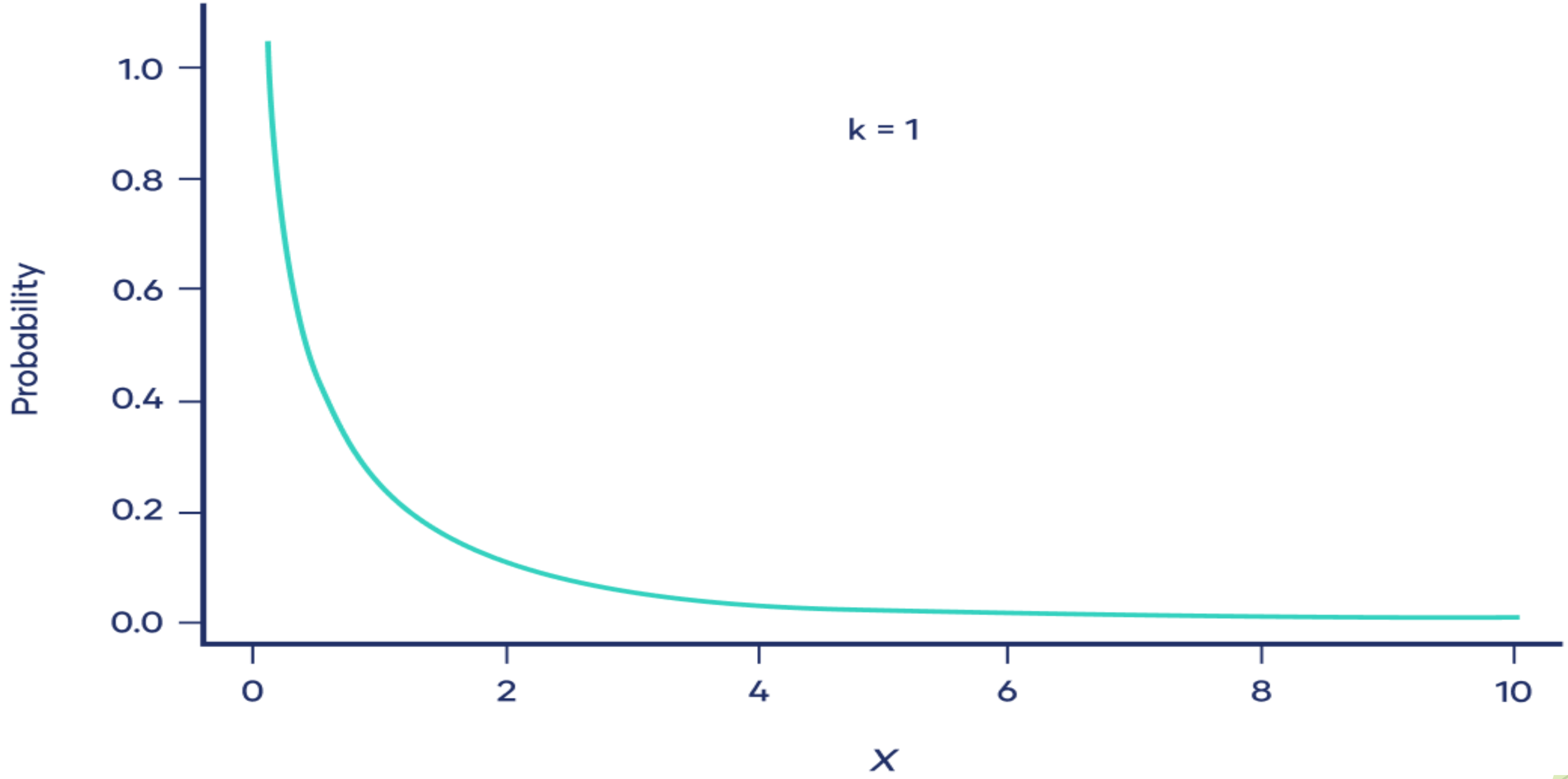
  $X^2_2 = (Z_1)^2 + (Z_2)^2$

- More generally, if you sample from k independent standard normal distributions and then square and sum the values, you'll produce a chi-square distribution with k degrees of freedom.

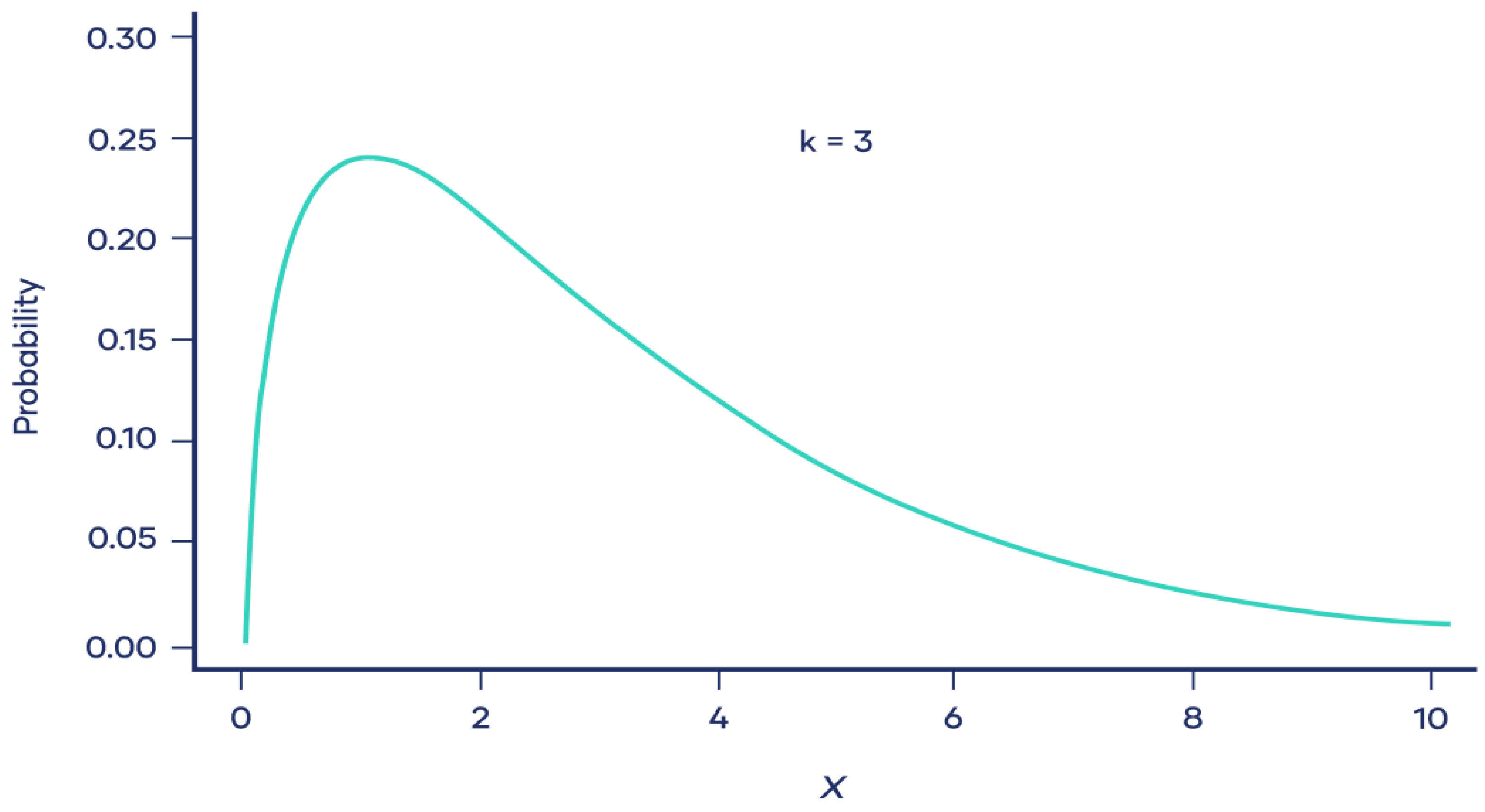
  $X^2_k = (Z_1)^2 + (Z_2)^2 + \ldots + (Z_k)^2$

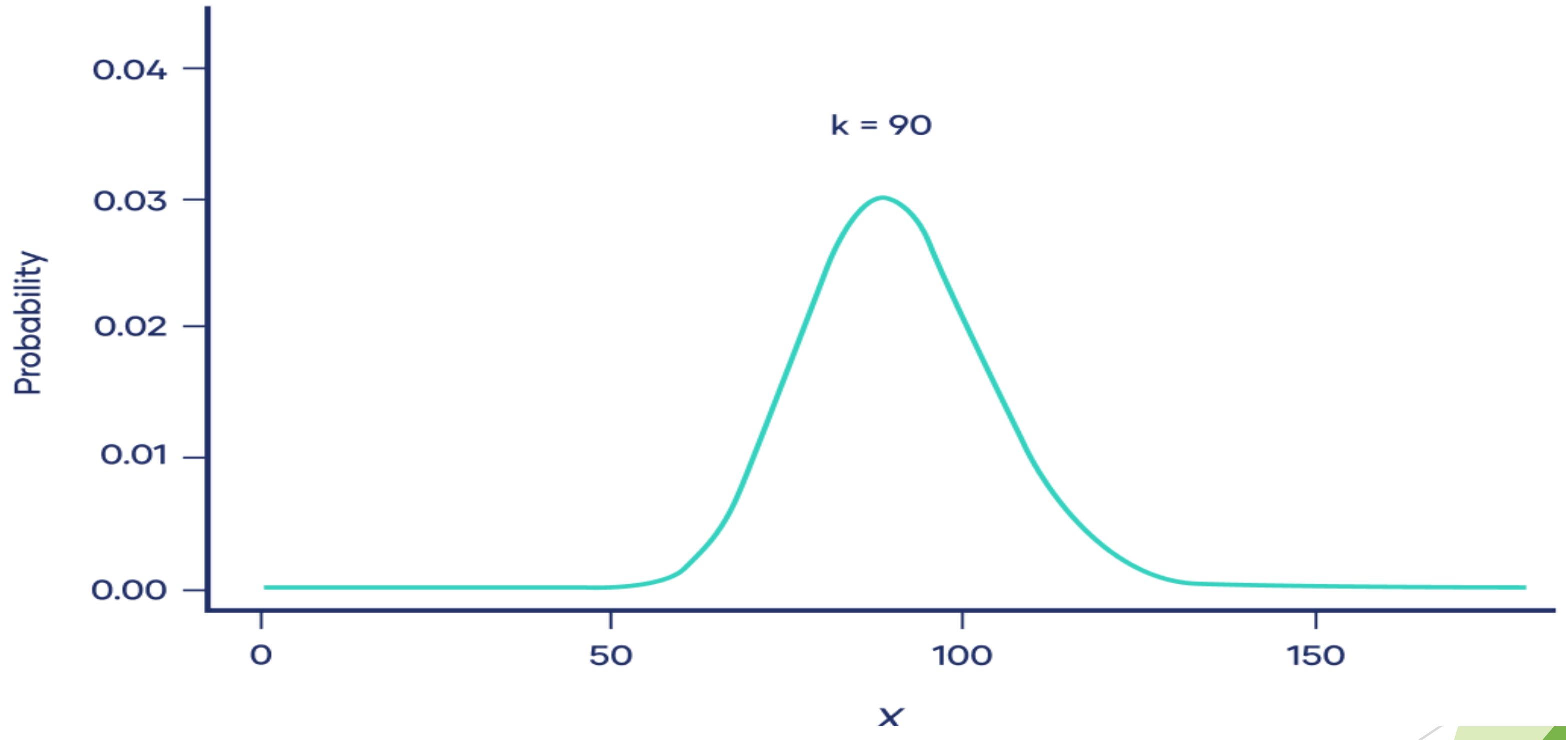
Formula

Formula	Explanation
$X^2 = \sum \frac{(O-E)^2}{E}$	Where • X^2 is the chi-square test statistic • $\sum$ is the summation operator (it means “take the sum of”) • O is the observed frequency • E is the expected frequency



1.0
0.8
0.6
0.4
0.2
0.0
Probability
k = 1
0
2
4
6
8
10
X


k = 3
Probability
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
x


k = 90
Probability
0.04
0.03
0.02
0.01
0.00
0
50
100
150
x

Addition Rule of Probability

Addition Rule

- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Multiplication Rule

Independent Events

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Dependent Events

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

Multinomial Distribution

Multinomial Distribution

- A **multinomial experiment** is a statistical experiment that has the following properties:
- The experiment consists of *n* repeated trials.
- Each trial has a discrete number of possible outcomes.
- On any given trial, the probability that a particular outcome will occur is constant.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Example

- Consider the following statistical experiment. You toss two dice three times, and record the outcome on each toss. This is a multinomial experiment because:
- The experiment consists of repeated trials. We toss the dice three times.
- Each trial can result in a discrete number of outcomes - 2 through 12.
- The probability of any outcome is constant; it does not change from one toss to the next.
- The trials are independent; that is, getting a particular outcome on one trial does not affect the outcome on other trials.

Note:

A Binomial experiment is a special case of a multinomial experiment. Here is the main difference. With a binomial experiment, each trial can result in two - and only two - possible outcomes. With a multinomial experiment, each trial can have two *or more* possible outcomes.

Multinomial

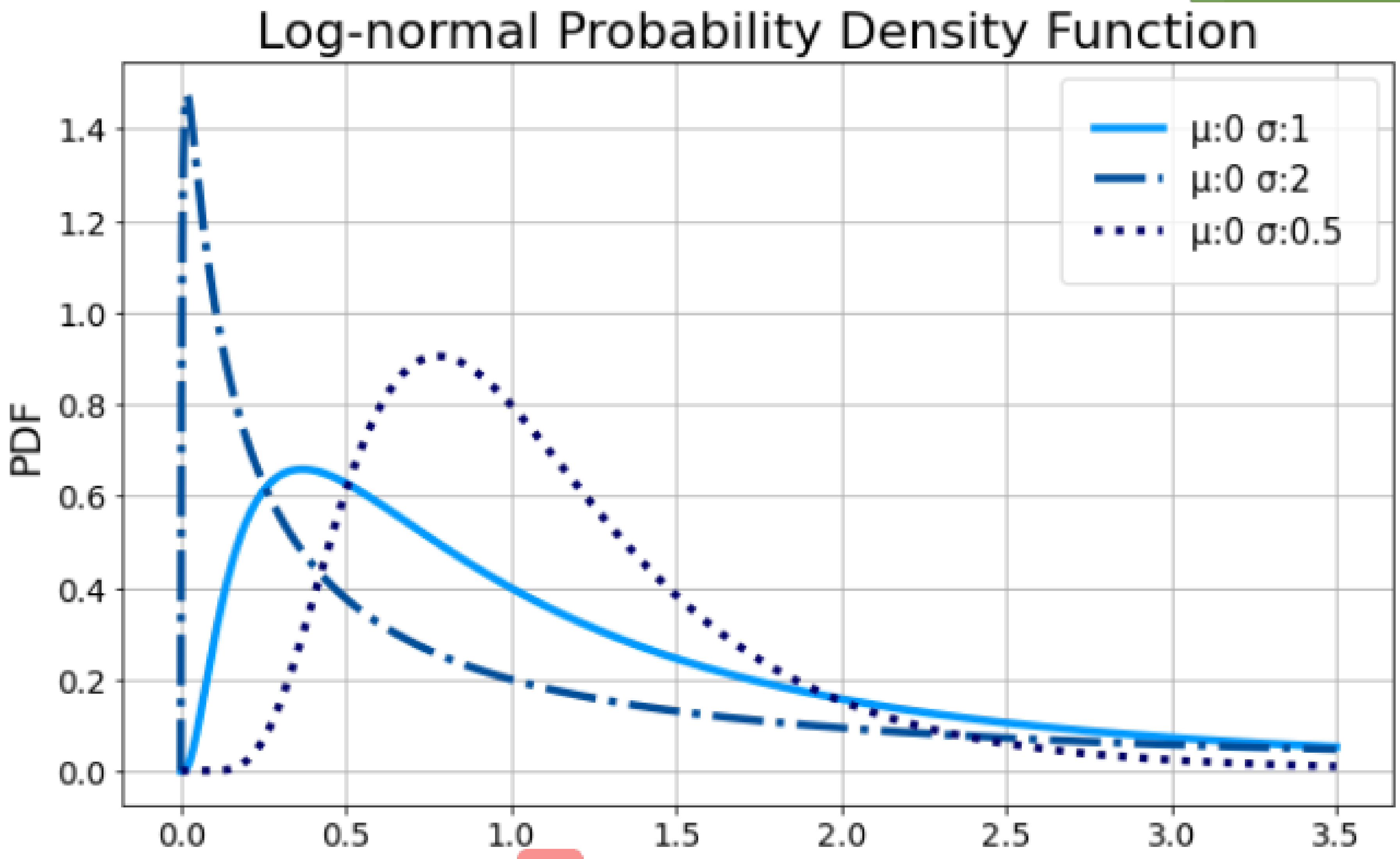
Parameters	$n > 0$ number of trials (integer) $p_1, \ldots, p_k$ event probabilities ($\Sigma p_i = 1$)
Support	$x_i \in \{0, \ldots, n\}, \quad i \in \{1, \ldots, k\}$ $\Sigma x_i = n$
pmf	$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$

Log Normal Distribution

Definition

- The log-normal distribution is a right skewed continuous probability distribution, meaning it has a long tail towards the right. It is used for modelling various natural phenomena such as income distributions, the length of chess games or the time to repair a maintainable system and more.
- The data points for our log-normal distribution are given by the *X* variable. When we log-transform that *X* variable *(Y=ln(X))* we get a *Y* variable which is normally distributed.

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



Log-normal Probability Density Function
PDF
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
μ:0 σ:1
μ:0 σ:2
μ:0 σ:0.5

Normal Distribution

$$\hat{\mu} = \frac{\sum_i x_i}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_i (x_i - \hat{\mu})^2}{n}$$

$$\hat{\mu} = \frac{\sum_k \ln x_k}{n} \text{ and } \hat{\sigma}^2 = \frac{\sum_k (\ln x_k - \hat{\mu})^2}{n}$$

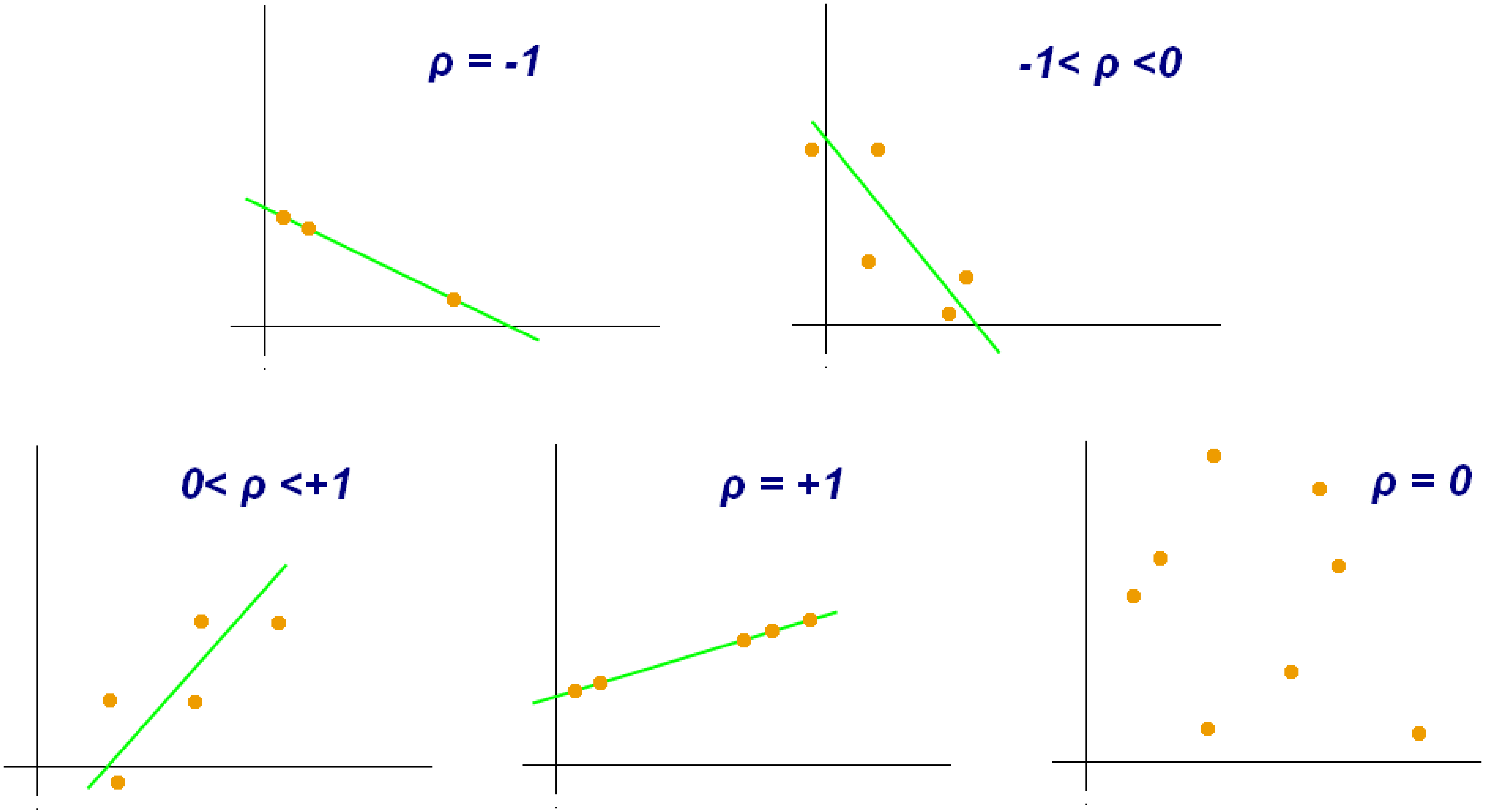
Correlation

Correlation

- The correlation coefficient is a measure of the association between two variables. It is used to find the relationship between data and a measure to check how strong it is. The formulas return a value between -1 and 1, where -1 shows negative correlation and +1 shows a positive correlation.
- The correlation coefficient value is positive when it shows that there is a correlation between the two values and the negative value shows the amount of diversity among the two values.

How to find correlation?

- Graphs
  - Scatterplot
- Mathematical Formula
  - Pearson’s Correlation
  - Spearman’s Correlation
  - Kendal’s Tau b


ρ = -1
-1< ρ <0
0< ρ <+1
ρ = +1
ρ = 0

Data Pre requisites

Data Pre requisites (SLOB)

- Scale
- Linear
- No Outlier
- Bivariate Normality

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Sl. No	Notation	Description
1	n	Quantity of Information
2	Σx	Total of the First Variable Value
3	Σy	Total of the Second Variable Value
4	Σxy	Sum of the Product of First & Second Value
5	Σx^2	Sum of the Squares of the First Value
6	Σy^2	Sum of the Squares of the Second Value

Box cox Transformation

WHAT IS A BOX COX TRANSFORMATION?

- A TRANSFORMATION OF NON-NORMAL DEPENDENT VARIABLES INTO A NORMAL SHAPE.
- RESULT: YOU CAN RUN A BROADER NUMBER OF TESTS.
- NAMED AFTER GEORGE BOX AND SIR DAVID ROXBEE COX (1964).

WHAT IS A BOX COX TRANSFORMATION?

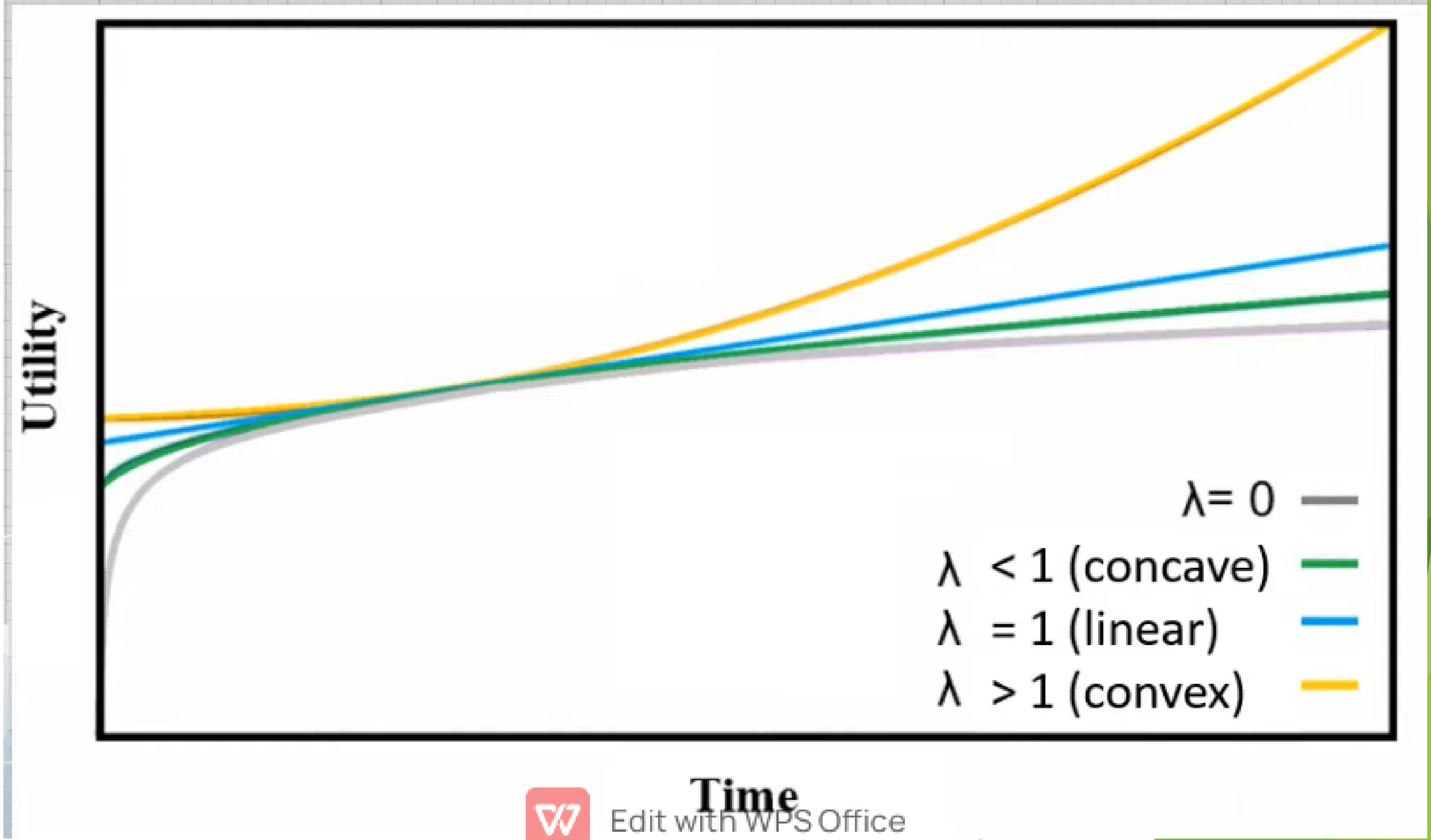
- LAMBDA (λ)VARIES FROM -5 TO 5.
- "OPTIMAL VALUE" = BEST APPROXIMATION OF A NORMAL DISTRIBUTION CURVE.

POSITIVE VALUES

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

NEGATIVE VALUES

$$y(\boldsymbol{\lambda}) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y+\lambda_2), & \text{if } \lambda_1 = 0. \end{cases}$$



Utility
λ= 0
λ < 1 (concave)
λ = 1 (linear)
λ > 1 (convex)
Time

COMMON BOX-COX TRANSFORMATIONS

Lambda value (λ)	Transformed data (Y')
-3	$Y^{-3} = 1/Y^3$
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\sqrt{(Y)})$
0	log(Y)**
0.5	$Y^{0.5} = \sqrt{(Y)}$
1	$Y^1 = Y$
2	Y^2
3	Y^3

- **LOG(Y) TRANSFORMS TO ZERO,
- OTHERWISE ALL DATA WOULD TRANSFORM TO $Y^0 = 1$.

CALCULATIONS

- TESTING ALL POSSIBLE VALUES BY HAND IS UNNECESSARILY LABOR INTENSIVE.
- MANY SOFTWARE PACKAGES INCLUDE AN OPTION, INCLUDING:
  - R: USE THE COMMAND BOXCOX(OBJECT, ...).
  - MINITAB: CLICK THE OPTIONS BOX AND THEN CLICK "BOX-COX TRANSFORMATIONS/OPTIMAL λ."

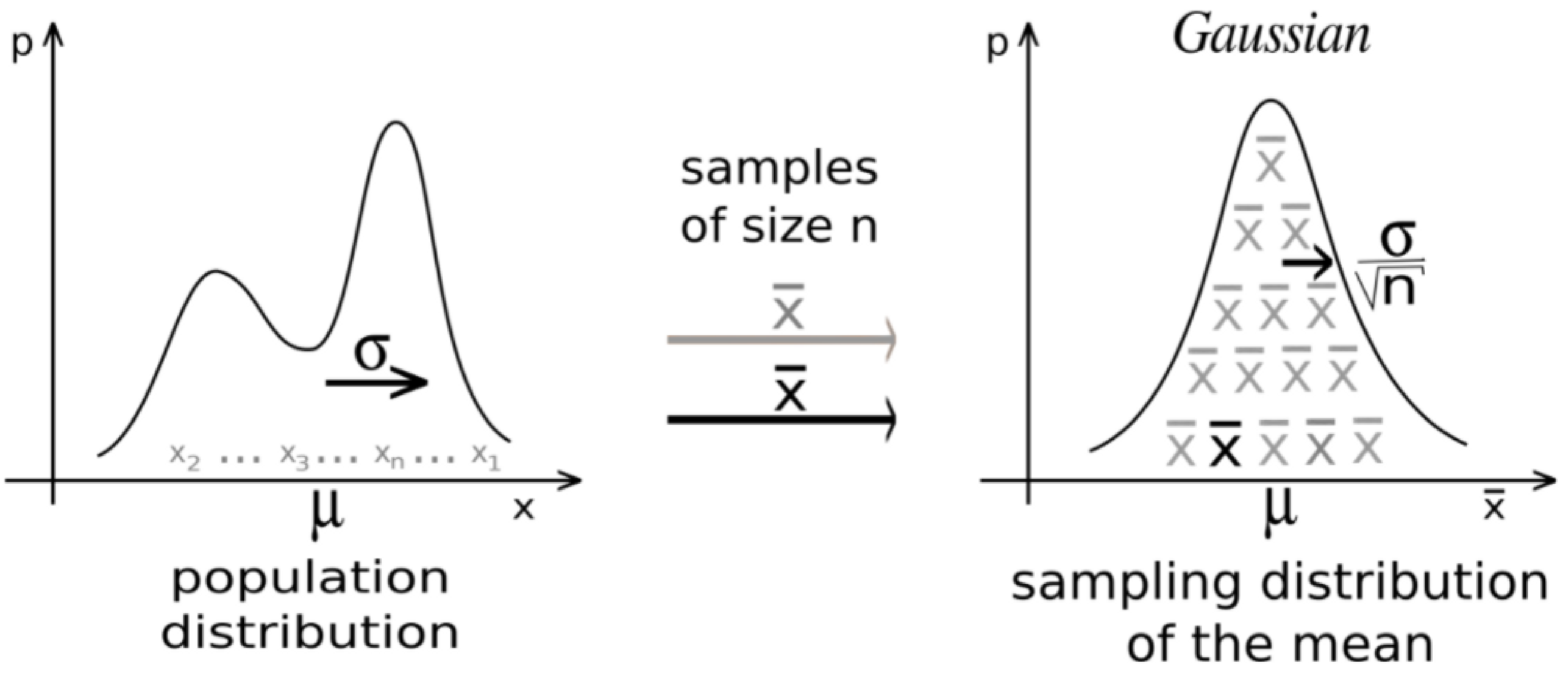
Central Limit Theorem

Definition

- The central limit theorem (CLT) states that **the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution**. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.

Definition2

- No matter what the distribution of the population is, the distribution of mean samples from the population will always be Normally distributed.



p
x₂ ... x₃ ... xₙ ... x₁
σ
μ
x
population distribution
samples of size n
x̄
x̄
p
Gaussian
σ / √n
μ
x̄
sampling distribution of the mean

Recap (chapter1)

- Business Intelligence Vs Data Analysis Vs Data Science
- Data Science Roles
- Different disciplines of Data Science

Chapter2

2.1 Sampling and Data
2.1.1 Definitions of Statistics, Probability, and Key Terms
2.1.2 Frequency, Frequency Tables, and Levels of Measurement
2.1.3 Concept of Correlation and Causal relationship between variables
2.1.4 Idea of Population Sample Vs. Population

Chapter3

2.2 Descriptive Statistics

2.2.1 Introduction to Mean Median and Mode

2.2.2 Introduction to Skewness and Kurtosis

2.2.3 Introduction to Data Spread and Range

2.2.4 Introduction Quartiles and Quartile Range

2.2.5 Measures of Dispersion

2.2.6 Introduction to Outliers

2.2.7 Stem-and-Leaf Graphs , Line Graphs, and Bar Graphs

2.2.8 Histograms, Frequency Polygons, and Time Series Graphs

2.2.9 Impact of Scaling and Shifting of origin

Chapter4

2.3 Probability Distribution
2.3.1 What is Frequency Distribution, Cumulative Frequency Distribution, Discrete and Continuous Probabilty Distributions
2.3.2 Different type of Probability Distribution
2.3.2.1 Bernoulli's Distribution
2.3.2.2 Binomial Distribution
2.3.2.3 Gaussian Distribution
2.3.2.4 Poisson Distribution
2.3.2.5 Exponential Distribution
2.3.2.6 Multinomial Distribution
2.3.2.7 T- Distribution
2.3.2.8 Uniform Distribution
2.3.2.9 Frequency Distribution
2.3.2.10 Log Normal Distribution
2.3.2.11 Boxcox Transform
2.3.3 Introduction to Z-scores

2.4 Probability Theory

2.4 Probability Theory

2.4.1 Introduction to Probability

2.4.2 Rule of Addition and Multiplication

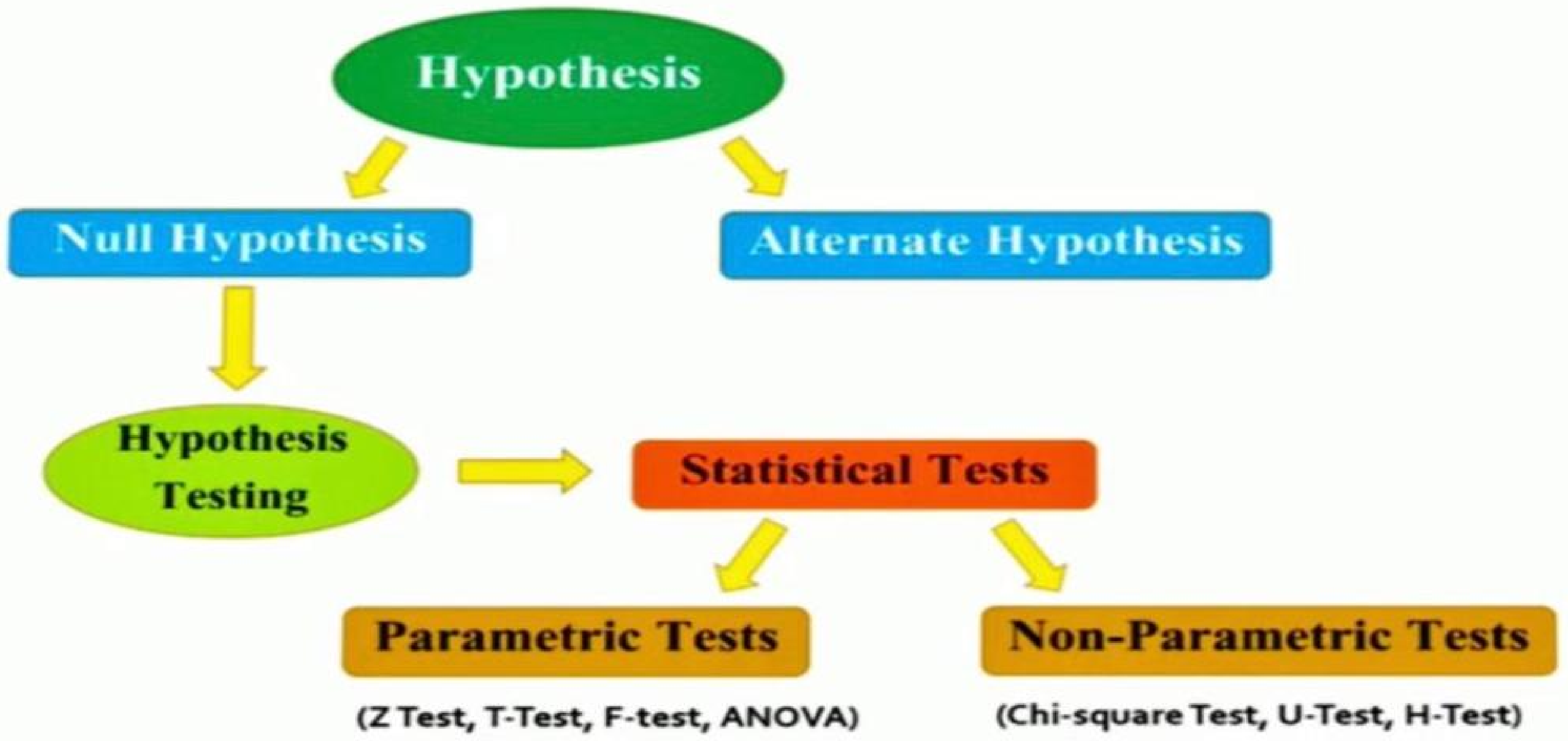
2.4.3 Conditional Probability and Bayes' Theorem

2.4.4 Central Limit Theorem

Statistical Inference

Hypothesis

- Suppose we have a huge amount of data. We take out a sample from the dataset and make some claims. Note that claims are not always valid, these are just assumptions or guesses, this type of claim or assumption is called **Hypothesis.**


Hypothesis
Null Hypothesis
Alternate Hypothesis
Hypothesis Testing
Statistical Tests
Parametric Tests
Non-Parametric Tests
(Z Test, T-Test, F-test, ANOVA)
(Chi-square Test, U-Test, H-Test)

Example

Example 1:
Let us take an example to understand it more clearly. According to laws food manufacturing companies should not put more than 2.5 ppm(particle per million) **Lead** in food. So, let us take a company XYZ and we claim that the average amount of lead in food that XYZ company manufactures contains is more than 2.5 ppm.

This is just a claim based on the limited amount of data and not valid for the whole population. Hypothesis testing helps us verifying a claim on statistic values.

Example 2:

- Let us take another example, suppose a person is charged for some trial where the jury has to decide whether the person is innocent or guilty.
- It can be converted to 2 hypotheses:
- **Hypothesis 0:** Defendant is innocent.
- **Hypothesis1** Defendant is guilty.
- These two opposing hypotheses are called the **null hypothesis** and **alternative hypothesis**.

Null hypothesis

Null Hypothesis

- The null hypothesis is a prevailing belief about the population. It states that there is no change or no difference in the situation.
- It assumes the status quo (the existing state of affairs) is true.
- In our example 2 defendant is a member of society, that is why he is considered innocent until proven guilty. So our null hypothesis claims the defendant is innocent just like he was before the charge.
- The Null hypothesis is represented as ***H0***
- Remember that the null hypothesis will always have these signs:

=   ≤   ≥

Alternative hypothesis

Alternative hypothesis

- In simple words, we can define the alternative hypothesis as the **opposite of the null hypothesis**
- Continuing the same example 2, our alternative hypothesis is that he is guilty.
- The Alternative hypothesis is represented as ***H1***
- Remember that the Alternative hypothesis will always have these signs:

!=   >   <

Important Points

Important Points

- H0 and H1 cannot be true at the same time.
- We only reject or not reject the null hypothesis, we never accept it. If **H1** is rejected it does not mean that **H0** has to be accepted there might be some other possibilities.

Situation1

Flipkart claimed that its total valuation in December 2016 was at least $14 billion. Here the claim contains a ≥ sign, so the null hypothesis is an original claim.

- **The hypothesis, in this case, can be formulated as:**
- *Total valuation ≥ $14 billion → Null Hypothesis*
- *Total valuation < $14 billion → Alternate Hypothesis*

Situation 2

- Flipkart claimed that its total valuation in December 2016 was greater than $14 billion. Here the claim contains > sign, so the null hypothesis is the complement of the original claim.
- **The hypothesis, in this case, can be formulated as:**
- *Total valuation ≤ $14 billion → Null Hypothesis*
- *Total valuation >$14 billion → Alternate Hypothesis*

Errors

Truth

Research	Hypothesis Testing	The Null Hypothesis Is True	The Alternative Hypothesis Is True
	The Null Hypothesis Is True	Accurate	Type II Error
	The Alternative Hypothesis Is True	Type I Error	Accurate

Type I Error

- A **type I error**, also known as an **error of the first kind**, occurs when the null hypothesis (H_0) is true, but is rejected.
- A type I error may be compared with a so called ***false positive.***
- A Type I error occurs when we believe a **falsehood**.
- The rate of the type I error is called the *size* of the test and denoted by **the Greek letter α (alpha).**
- It usually equals the **significance level of a test**.
- If type I error is fixed at 5 %, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true.

Type II Error

- **Type II error**, also known as an **error of the second kind**, occurs when the null hypothesis is false, but erroneously fails to be rejected.
- Type II error means accepting the hypothesis **which should have been rejected**.
- A type II error may be compared with a so-called ***False Negative.***
- A Type II error is committed when **we fail to believe a truth.**
- A type II error occurs when one rejects the alternative hypothesis (**fails to reject the null hypothesis**) when the **alternative hypothesis is true.**
- The rate of the type II error is denoted by the **Greek letter β (beta)** and related to the power of a test **(which equals 1-β)**.

Summary:

- Type 1 Error:
  - Reject null hypothesis when it is true.
- Type 2 Error:
  - Not Rejecting the null hypothesis when it is false.

Concepts

	H_0 is true	H_0 is false
Reject H_0	Type I error P(Type I error)=alpha	CORRECT
Accept H_0	CORRECT	Type II error P(Type II error)=beta

Making a Type I error or Type II error

Example

Ms. Rekha is a researcher and is testing her hypothesis that people in her town spend more money on coffee on Monday morning than they do on Tuesday morning. She doesn't know it, but her hypothesis is true: people actually *do* spend more money on coffee on Monday. She picks a random sample of people in her town and asks them how much money they spent on coffee each day. Say whether Rekha will make a Type I or Type II error.

	Monday	Tuesday
Average spend	$4.25	$5.45

Rekha's Null and Alternative Hypothesis are

- *H*0: People do not spend more on coffee on Monday than they do on Tuesday: $C_M \leq C_T$
- *Ha*: People spend more on coffee on Monday than they do on Tuesday: $C_M > C_T$

Interpretation

- In reality, her null hypothesis is true. But her sample data is showing that people spend more on Tuesday than they do on Monday. Which means she’s in danger of rejecting the null hypothesis when she shouldn’t since the null hypothesis is true.

Conclusion

	H_0 is true	H_0 is false
Reject H_0	**Type I error** **P(Type I error)=alpha**	CORRECT
Accept H_0	CORRECT	Type II error P(Type II error)=beta

Power of the Test

Power of the Test

- Sometimes we say that the **power** is the probability our test will reject the null hypothesis when it's false, which is a correct decision. Rejecting the null hypothesis when it's false is exactly what were hoping to do.

	H_0 is true	H_0 is false
Reject H_0	Type I error P(Type I error)=alpha	**CORRECT** **Power**
Accept H_0	CORRECT	Type II error P(Type II error)=beta

So the higher the power, the better off we are. Power is also equal to $1-\beta$.

Significance Level

Significance Level

α =10%:

Interpretation:

- When you decide on α, you're actually deciding how much you want to risk committing a Type I error. In other words, if you choose α=0.05, you're saying that, 5% of the time, or 1 out of 20 times, you'll reject the null hypothesis when the null hypothesis is actually true.

Lower Alpha

- Remember that we usually pick a confidence level of 90%, 95%, or 99%, and these correspond to α values of
- At 90% confidence, the alpha value is
  - alpha=1−90%=10%
- At 95% confidence, the alpha value is
  - Alpha=1−95%=5%
- At 99% confidence, the alpha value is alpha
  - Alpha=1−99%=1%

P value

What is p-value in hypothesis testing?

- The p value is **a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true**. P values are used in hypothesis testing to help decide whether to reject the null hypothesis.

P value

- **A p-value less than 0.05 (typically ≤ 0.05) is statistically significant**. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random).
- High p-values indicate that **your evidence is not strong enough to suggest an effect exists in the population**.

Introduction to Confidence Intervals

Confidence Interval

- In real life, we don't typically have access to the whole population. In these cases we can use the sample data that we do have to construct a **confidence interval** to estimate the population parameter with a stated level of confidence. This is one type of statistical inference.
- Confidence Interval
  - A range computed using sample statistics to estimate an unknown population parameter with a stated level of confidence

Example

- The teacher at a university want to estimate the average statistics anxiety score for all of their undergraduate students. It would be too time consuming and costly to give every undergraduate student at the university their statistics anxiety survey. Instead, they take a random sample of 50 undergraduate students at the university and administer their survey.
- Using the data collected from the sample, they construct a 95% confidence interval for the mean statistics anxiety score in the population of all university undergraduate students. They are using sample mean to estimate μ. If the 95% confidence interval for μ is 26 to 32, then we could say, “we are 95% confident that the mean statistics anxiety score of all undergraduate students at this university is between 26 and 32.” In other words, we are 95% confident that $26 \leq \mu \leq 32$. This may also be written as [26,32].

Margin of Error

- At the center of a confidence interval is the sample statistic, such as a sample mean or sample proportion. This is known as the **point estimate**. The width of the confidence interval is determined by the **margin of error.** The margin of error is the amount that is subtracted from and added to the point estimate to construct the confidence interval.
- Point Estimate
  - Sample statistic that serves as the best estimate for a population parameter.
- Margin of Error
  - Half of the width of a confidence interval; equal to the multiplier times the standard error

Confidence Interval

<u>**General Form of Confidence Interval**</u>

- Sample Statistic+/-Margin of Error.

<u>**General Form for 95% confidence Interval**</u>

- Sample Statistic+/- 2(standard Error)

Example: Proportion of Dog Owners

- At the beginning of the Spring 2017 semester a representative sample of STAT 200 students were surveyed and asked if they owned a dog. The sample proportion was 0.559. Bootstrapping methods, which we will learn later in this lesson, were used to compute a standard error of 0.022. Assume the bootstrap distribution is normally distributed. We can use this information to construct a 95% confidence interval for the proportion of all STAT 200 students who own a dog.
  - Sample Statistic+/- 2(standard Error)
  - $0.559 \pm 2(0.022)$
    0.559 ± 0.044
    $[0.515, 0.603]$
- Conclusion:
  - We are 95% confident that the proportion of all STAT 200 students in Spring 2017 that own a dog is between 0.515 and 0.603.

Example: Mean Height

- In a random sample of 525 Penn State World Campus students the mean height was 67.009 inches with a standard deviation of 4.462 inches. The standard error was computed to be 0.195. Construct a 95% confidence interval for the mean height of all Penn State World Campus students. Assume the bootstrap distribution is normally distributed.
  - sample statistic±2 (standard error)
    - 67.009 ± 2(0.195)
      67.009 ± 0.390
      [66.619, 67.399]
- I am 95% confident that the mean height of all Penn State World Campus students is between 66.619 inches and 67.399 inches.