02-Probability
Notes and Examples

Topics

- Uncertainty and Probability
- What is Probability
- Terminology – Random Experiment, Sample Space, Event
- Types of Events
- Rules for Probability
- Marginal, Joint and Conditional Probability
- Bayes' Theorem
- Probability Distributions
  - Bernoulli, Binomial, Geometric (discrete)
  - Uniform, Exponential, Normal (continuous)
- Normal Distribution - Properties
- The Central Limit Theorem

Uncertainty

- Will it will rain today?
- Will India win the T20 world cup?
- Will Modi get elected in 2024?

Probability

- Helps in measuring Uncertainty
- Evaluating the chances of an event happening
- Probability Distribution helps in analyzing the future outcome
- We use the concept of probability to classify the data in Machine Learning
- Logistic Regression and Naïve Bayes are based on concept of Probability

What is Probability

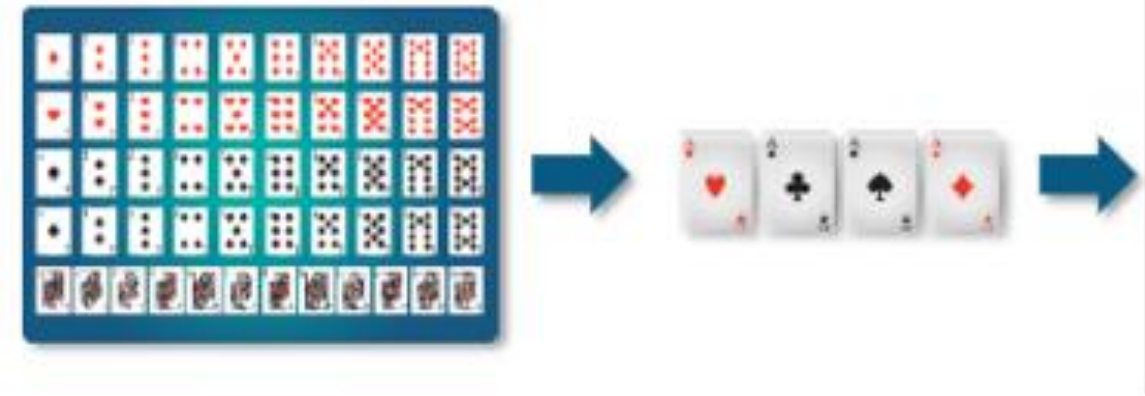
- Probability is the measure of how likely something will occur
- It is the ratio of desired outcomes to total possible outcomes:

  (# desired) / (# total)

- The sum of individual probabilities of all outcomes always equals to 1

Probability - example

- Let's assume that we have a deck of 52 cards
- The probability of drawing any particular card from the deck can be calculated
- Example – Probability of drawing any Ace card can be calculated as-



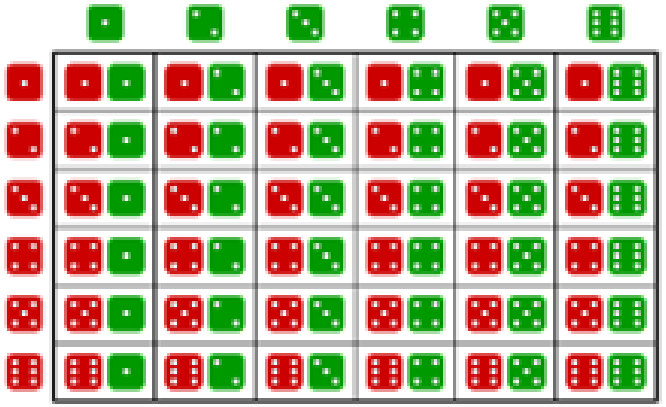
- Number of cards in deck = 52
- Number of aces = 4
- Ace_probability = 4/52 = 0.08

Terminology of Probability

Random Experiment

An experiment or a process for which the outcome cannot be predicted with certainty

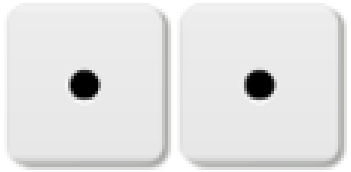
For example: Getting a head/tail in a toss.



Sample Space

Set of all possible outcomes of a random experiment in the sample space

(S) **For example:** Rolling of 2 dice has 36 different possible outcome



Event

A set of outcomes of an experiment, it is a subset of S

For example: Getting two 1 in rolling 2 dice

Disjoint and non-Disjoint events

Disjoint events -> Mutually Exclusive

Example: Rolling a Die:

- Events: Getting an odd number, Getting an even number
- These are mutually exclusive (Disjoint events)

- Events: Getting an odd number, Getting a number less than 3
- These are not mutually exclusive (non-Disjoint events)

Dependent and Independent Events

Two events are **independent** if the output of one is not affected by the output of the other

- **Example** – Toss of a coin or throw of a dice

Two events are **dependent** if the output of one is affected by the output of the other

- **Example** – Drawing two cards from a deck of card without replacing the first one

- The number obtained by each throw of the dice is independent of any previous outcome
- Thus, the event of obtaining any number, say 2, is an independent event

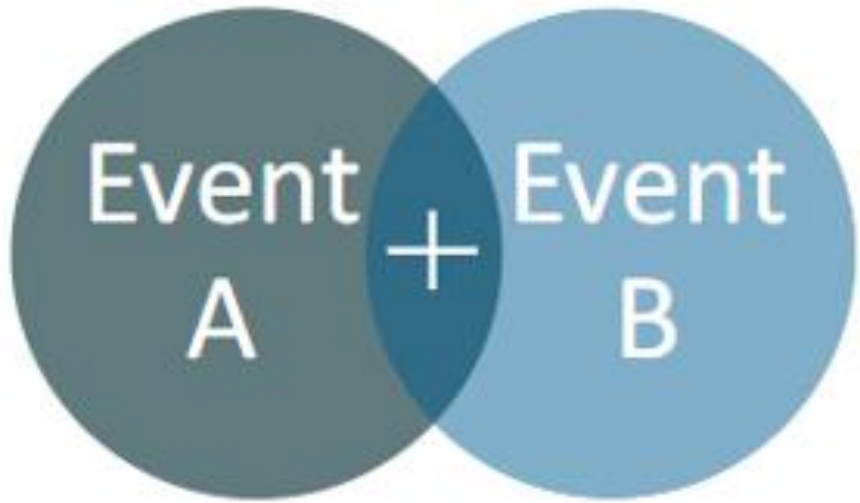
- At the time of drawing the second card, the sample space has reduced from 52 to 51
- Also, the probability of obtaining a desired card depends upon the first card that is drawn

Rules of Probability

- The probability of any event is between 0 and 1
  - **For any event A, $0 \leq P(A) \leq 1$**
- The sum of the probabilities of all possible outcomes is 1
  - $\sum P(E_i)=1$
- The Complement Rule
  - **For any event A, $P(A^C) = 1 - P(A)$**
- The Addition Rule
  - **$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$** **- non Disjoint events**
  - **$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B)$** **- Disjoint (mutually exclusive) events**
- The Product Rule
  - **$P(A \text{ and } B) = P(A \cap B) = P(A) * P(B \mid A)$** **- for Dependent events**
  - **$P(A \text{ and } B) = P(A \cap B) = P(A) * P(B)$** **- for Independent events**

The General Addition Rule

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



NOTE: When A and B are disjoint, **P(A and B) = 0**, Hence **P(A or B) = P(A) + P(B)**

Use Case - Scenario

Edureka has conducted a survey about its Statistics course

01

The most recent phase of the survey that polled 100,000 participants, estimates that 80% of the population agrees with the statement "The duration of the courses conducted by Edureka is just right"

02

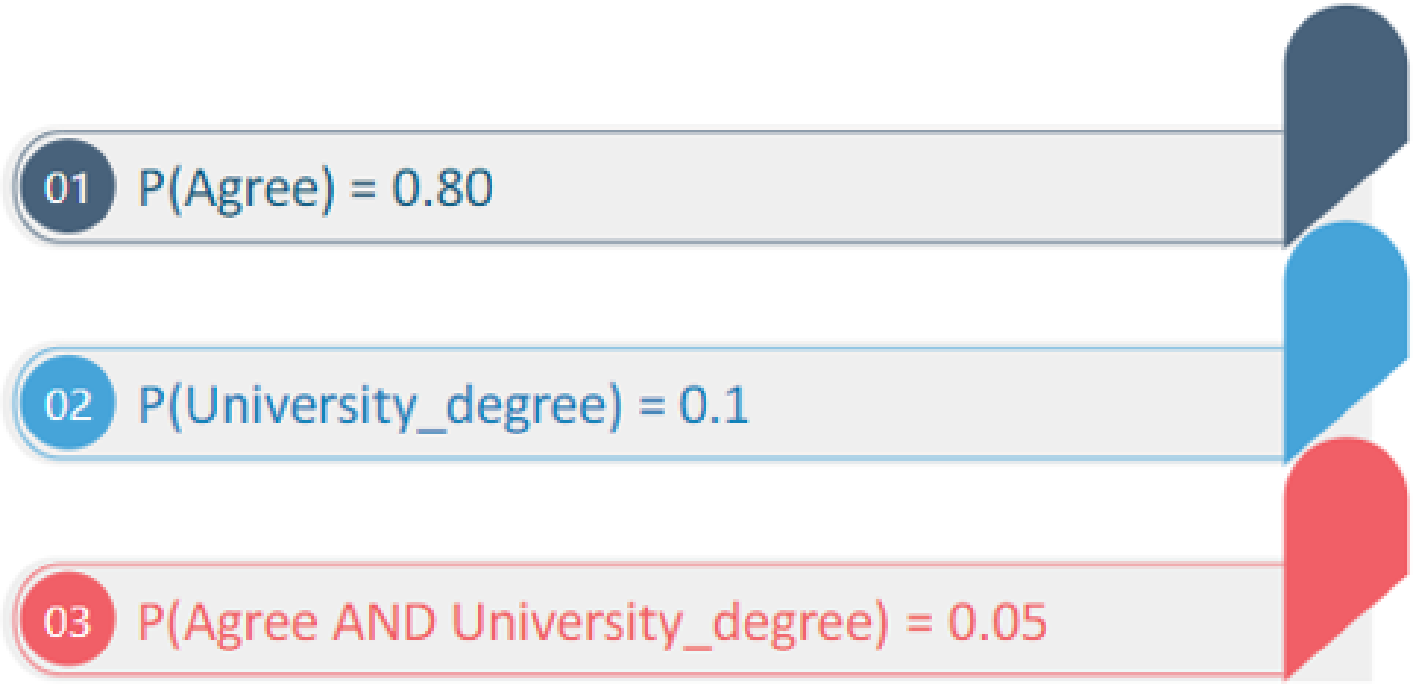
The survey also estimates that 10% people have university degree

03

The survey estimates that 5% of the total population fits both criteria

Use Case – Interpreting the Scenario

- Based on this scenario, three events & their probabilities can be interpreted as:



Use Case Analysis – 1 (Check for Disjoint)

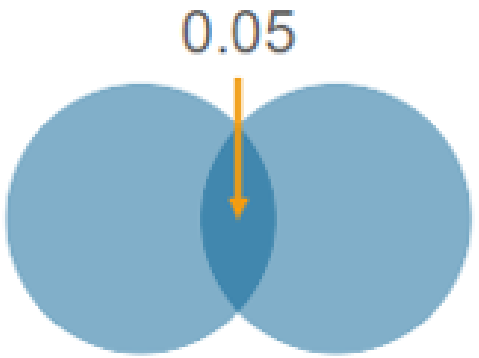
1. **Event 1**: Students agreeing with the statement "Duration of the course is just right"

   Event 2: Students having a university degree

Check, if both the events are disjoint or not:

Solution:

- P(Agree) = 0.80
- P(University_degree) = 0.10
- **P(Agree AND University_degree) = 0.05 ≠ 0**



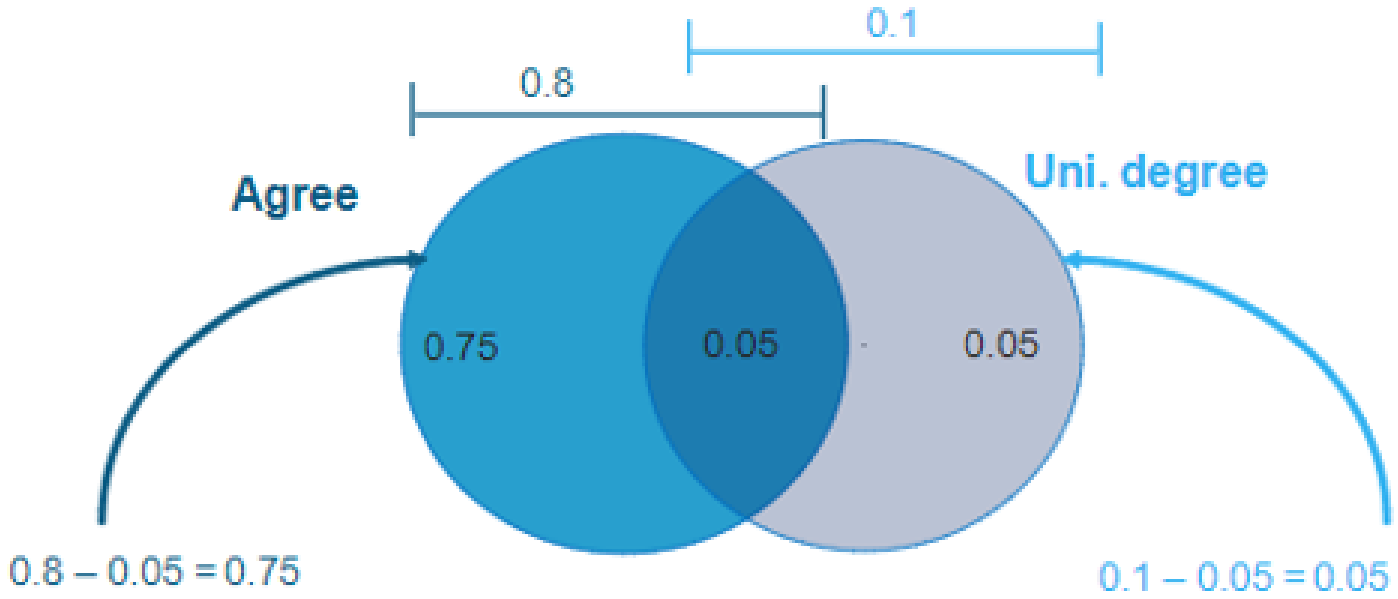
Since, the probability of agreeing with the statement "Duration of the course is just right" and having a university degree is 0.05 which is not equal to 0. Hence, the events are not Disjoint (i.e. Joint)

Use Case Analysis – 2 (Venn Diagram)

2. Draw a Venn diagram summarizing the variables and their associated probabilities:

- P(Agree) = 0.8
- P(University_degree) = 0.1
- P(Agree AND University_degree) = 0.05

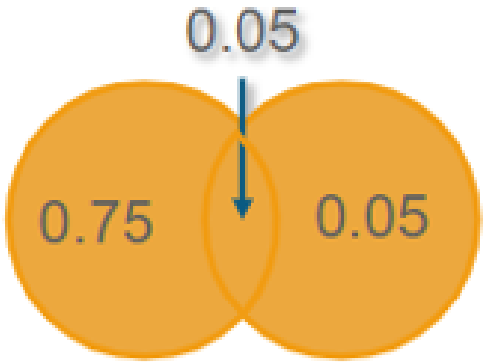
Solution:



Use Case Analysis – 3 (Logical OR)

3. What is the probability that a randomly drawn person has a university degree or agrees with the statement about duration of the course?

- P(Agree) = 0.8
- P(University_degree) = 0.1
- P(Agree AND University_degree) = 0.05



General Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$

P(Agree OR University_degree) = P(Agree) +

P(University_degree) - P(Agree AND University_degree)

= 0.8+0.1-0.05

= 0.85

Use Case Analysis – 4 (Complementary Set)

4. What percent of the population does not have a university degree and also disagrees with the statement about duration of the course?

- P(Agree) = 0.8
- P(University_degree) = 0.1
- P(Agree OR University_degree) = 0.85

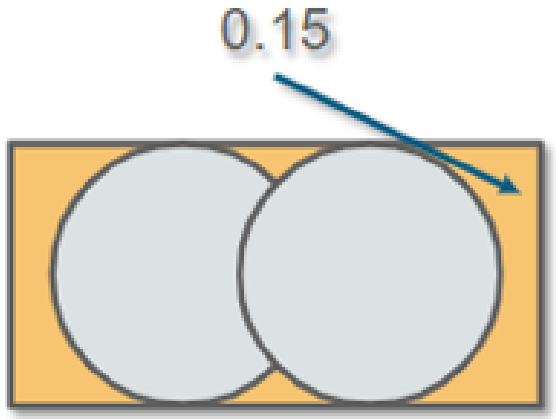
Solution:

P (Neither Agree Nor University_ degree)

= 1 – P(Agree OR University_degree)

= 1 – 0.85

= 0.15



Marginal, Joint and Conditional Probability

- **Marginal probability**: the probability of an event occurring (p(A)), it may be thought of as an unconditional probability. It is not conditioned on another event.
- Example: the probability that a card drawn is red (p(red) = 0.5).
- Another example: the probability that a card drawn is a 4 (p(four)=1/13).

- **Joint probability**: p(A and B). The probability of event A and event B occurring.
- It is the probability of the intersection of two or more events.
- The probability of the intersection of A and B may be written p(A ∩ B).
- Example: the probability that a card is a four and red =p(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).
  - P(A ∩ B) =P(A) * P(B) = (4/52) * (26/52) = 1/26

- **Conditional probability**: p(A|B) is the probability of event A occurring, given that event B occurs.
- Example: given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

Joint Probability Example

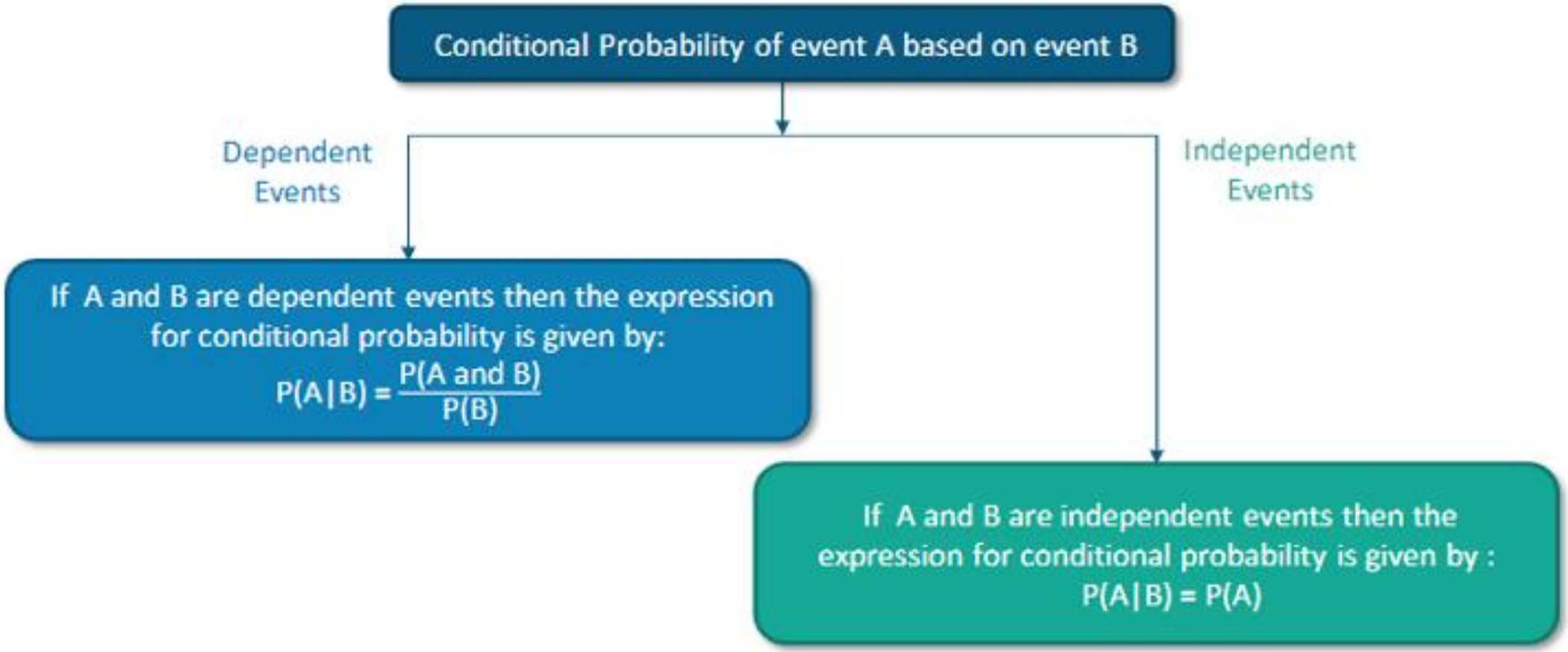
- **Joint probability** is the likelihood of more than one event occurring at the same time
- Joint probability is calculated by multiplying the probability of event A, expressed as P(A), by the probability of event B, expressed as P(B)
- For example
  - the probability that the number five will occur twice when two dice are rolled at the same time
  - Since each dice has six possible outcomes, the probability of a five occurring on each dice is 1/6 or 0.1666
  - P(A)=0.1666
    P(B)=0.1666

    P(A,B)=(0.1666 x 0.1666)=0.02777

Conditional Probability

- P(A|B) is the conditional probability of event A occurring, given that event B occurs
- If Events A and B are not independent, then P(A|B) = P(A and B) / P(B)
- Example:
  - What is the probability that two cards drawn at random from a deck of playing cards will both be aces?
  - Once the first card chosen is an ace, the probability that the second card chosen is also an ace is called the conditional probability of drawing an ace
  - Symbolically, we write this as: P(ace on second draw | an ace on the first draw)
  - P(ace on first draw) = 4/52
  - P(ace on second draw | an ace on the first draw) = 3/51
  - P(ace on first draw & ace on second draw) = (4/52) *(3/51) = (1/13)*(3/51)

Expressions for Conditional Probability



Conditional Probability (e.g. term deposit data)

Probability Experiment

If a customer has university degree, what will be probability that he is a male

education	gender		
	female	male	All
basic.4y	2104	2072	4176
basic.6y	1169	1123	2292
basic.9y	3092	2953	6045
high.school	4735	4780	9515
illiterate	7	11	18
professional.course	2647	2596	5243
university.degree	6142	6026	12168
unknown	902	829	1731
All	20798	20390	41188

Event A: Customer is male

Event B: Customer has university degree

Event C: Customer is male and has university degree

Event occurrence:

A: 20390 **B:** 12168 **C(A ∩ B) :** 6026 **S:** 41188

$$P(A \cap B) = \frac{6026}{41188} = 0.146$$

$$P(B) = \frac{12168}{41188} = 0.295$$

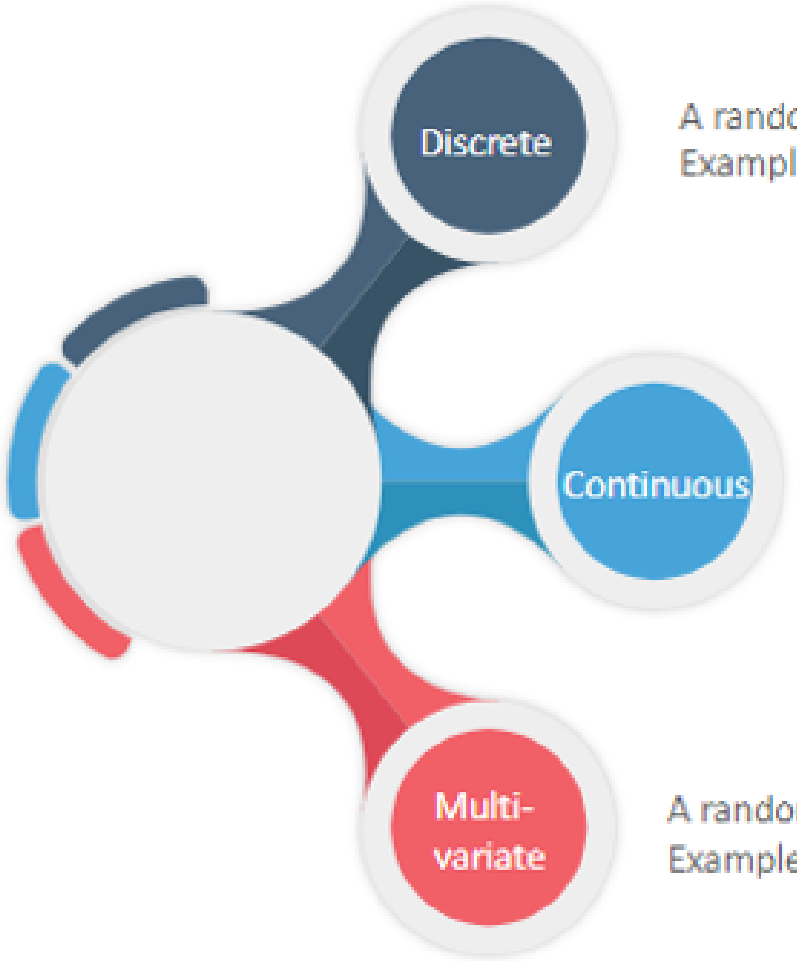
$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.146}{0.295} = 0.495$$

Introduction to Random Variables

- A Random Variable is a variable whose possible values are numerical outcomes of a random function
- It is usually denoted as X

- **Example** – Consider the event of tossing a coin twice
- The sample space becomes = {HH, HT, TH, TT}
- Let X be a random variable representing the phenomenon of obtaining a Head
- The possible values of X are 2,1,0

Types of Random Variables



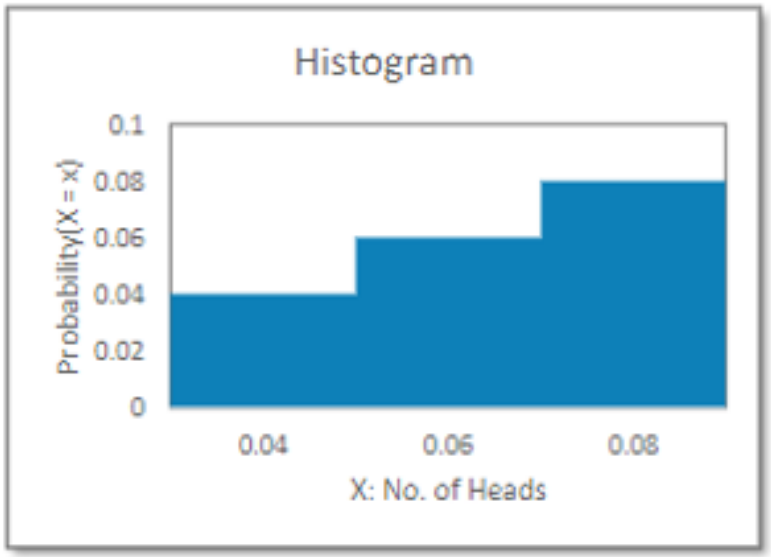
A random variable which takes a finite number of values
Example - Number of cars sold by a dealer on a particular day

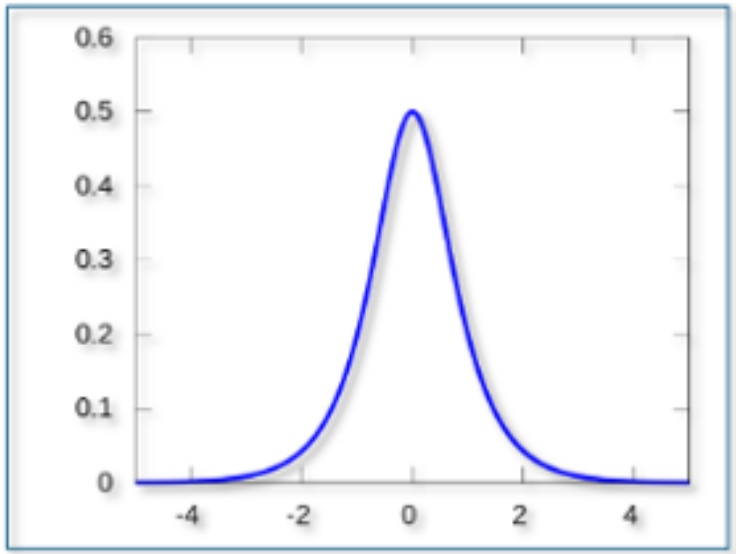
A random variable which takes any value in an interval
Example – Weight of a person in kilograms

A random variable which is represented by more than one random quantities
Example – Patient details containing Age, Height, Weight, BMI etc.

Probability Distribution

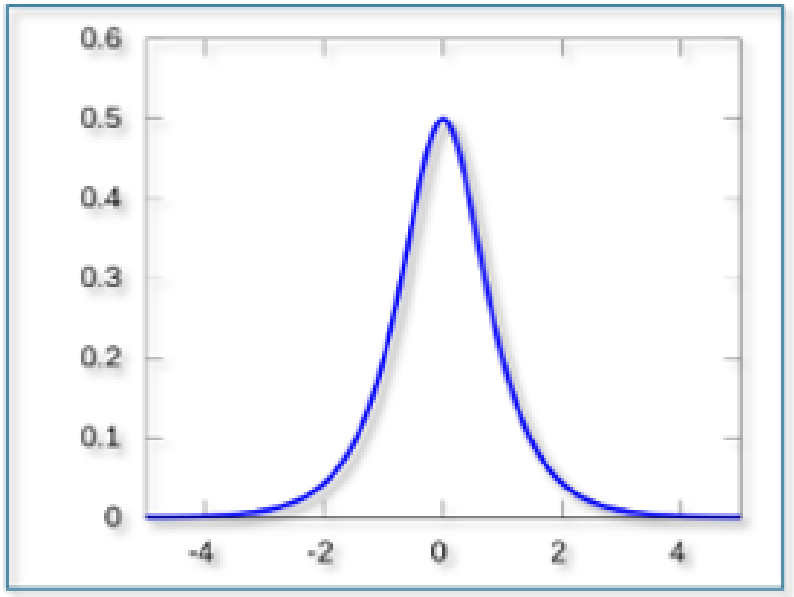
- Probability distribution is used to described a random variable
- It describes how the probabilities are distributed over the possible values of that random variable
- It can then be plotted on a graph with the help of Probability Distribution Functions





Probability Density Function

- The function describing the probabilities of a continuous random variable is called a Probability Density Function or PDF
- It can be represented as follows:



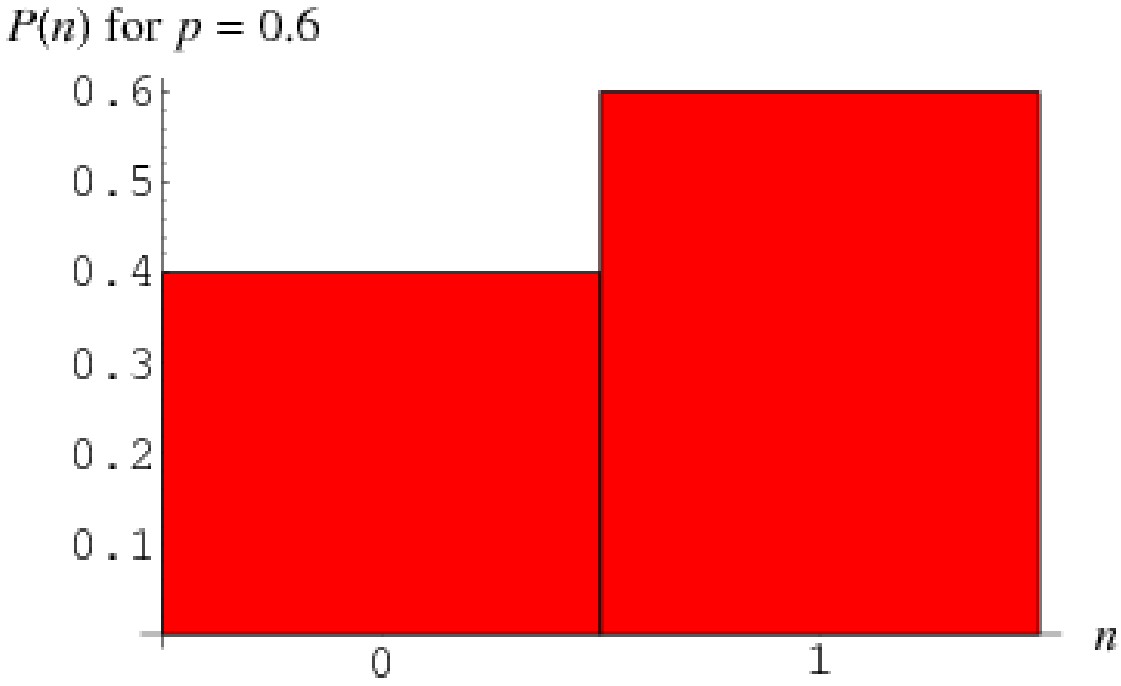
Types of Probability Distribution Functions

- For Continuous Variables
  - Uniform
  - Exponential
  - Normal

- For Discrete Variables
  - Bernoulli
  - Binomial
  - Geometric

Bernoulli Distribution

- The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by n=0 and n=1
  - in which n=1 ("success") occurs with probability p
  - and n=0 (“failure") occurs with probability q=1-p
  - where 0<p<1
- The distribution of heads and tails in coin tossing is an example of a Bernoulli distribution with p=q=1/2
- The Bernoulli distribution is the simplest discrete distribution, and is the building block for other more complicated discrete distributions



Binomial Distribution

- A binomial distribution can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.
- The binomial is a type of distribution that has two possible outcomes (the prefix “bi” means two, or twice).
- For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail
- Experiment repeated ‘N’ times
- **Q. A coin is tossed 10 times. What is the probability of getting exactly 6 heads?**
- P(x=6) = ${}_{10}C_6$ * 0.5^6 * 0.5^4
- = 210 * 0.015625 * 0.0625
- = 0.205078125

$$b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n-x}$$

Where:

b = binomial probability

x = total number of “successes” (pass or fail, heads or tails etc.)

P = probability of a success on an individual trial

n = number of trials

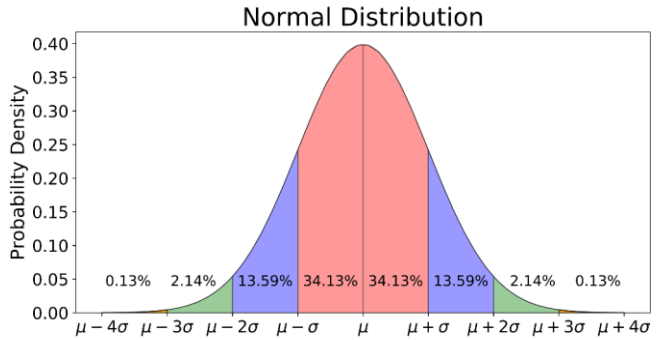
because ${}_nC_x = n!/x!(n-x)!$

$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$

Normal Distribution

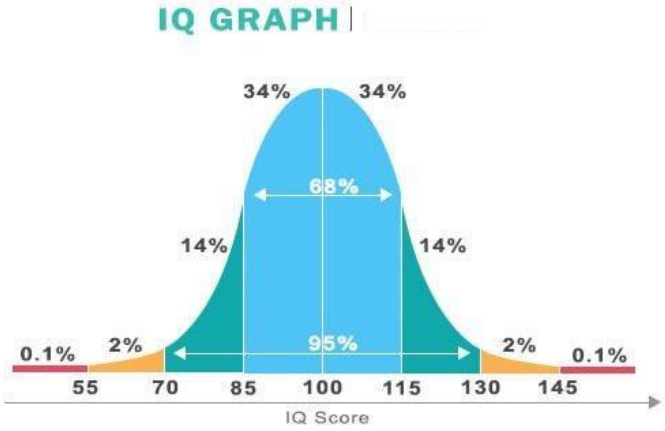
- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean
  - The data near the mean are more frequent in occurrence than data far from the mean
  - In graph form, normal distribution will appear as a **Bell curve**
  - The standard normal distribution has two parameters: the mean and the standard deviation
  - The curve never touches the x axis. It gets increasingly closer
- The normal distribution is the most common type of distribution assumed in lot of natural phenomena and in other types of statistical analyses

- The equation for Normal Distribution: $$y = \frac{e^{-(X-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$



Normal Distribution - examples

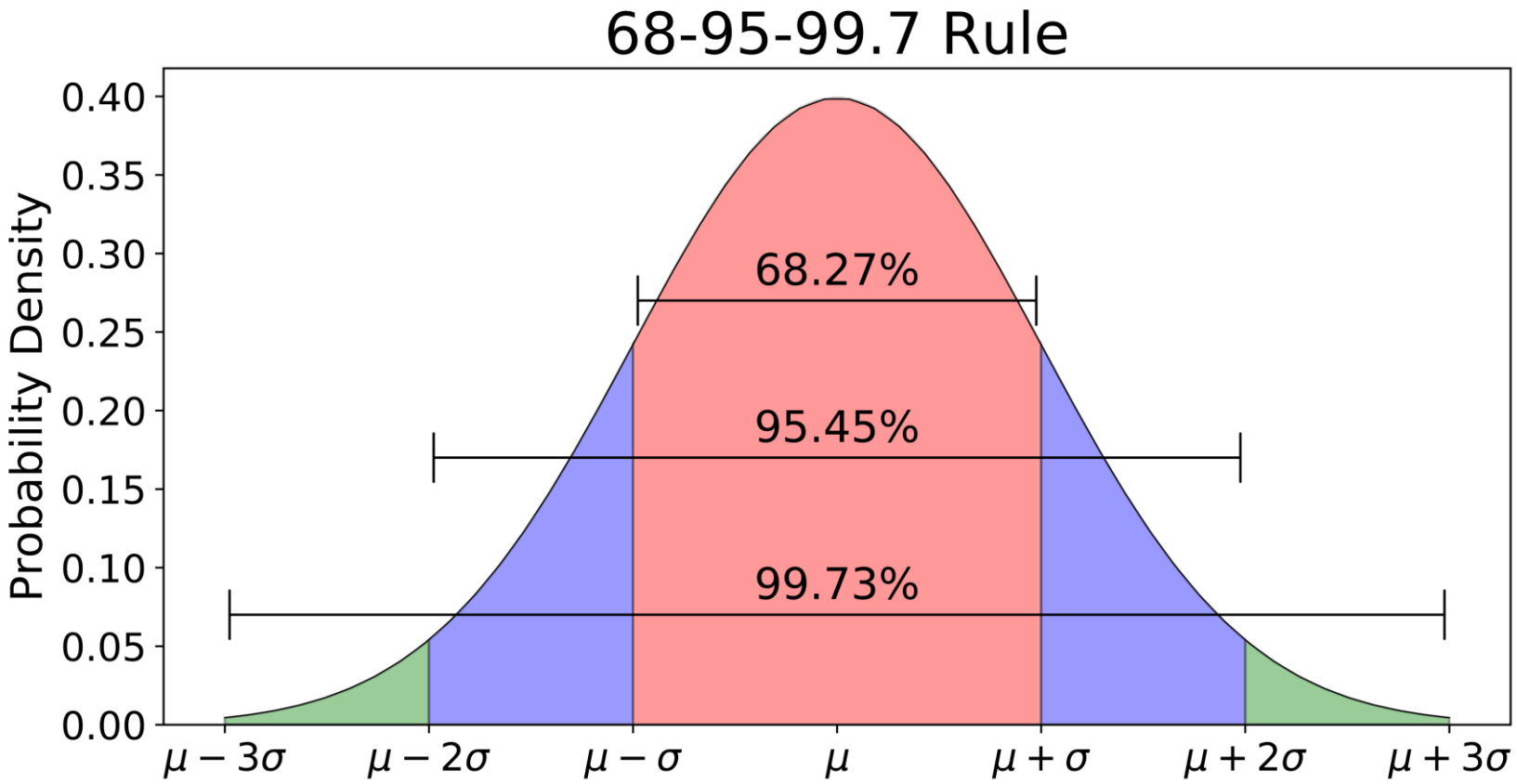
- A lot of natural phenomena exhibit normal distribution. Some examples:
  - **Height of the population** is the example of normal distribution. Most of the people in a specific population are of average height. The number of people taller and shorter than the average height people is almost equal, and a very small number of people are either extremely tall or extremely short
  - School authorities find the **average academic performance** of all the students, and in most cases, it follows the normal distribution curve.
  - **IQ of a particular population** is a normal distribution curve; where IQ of a majority of the people in the population lies in the normal range



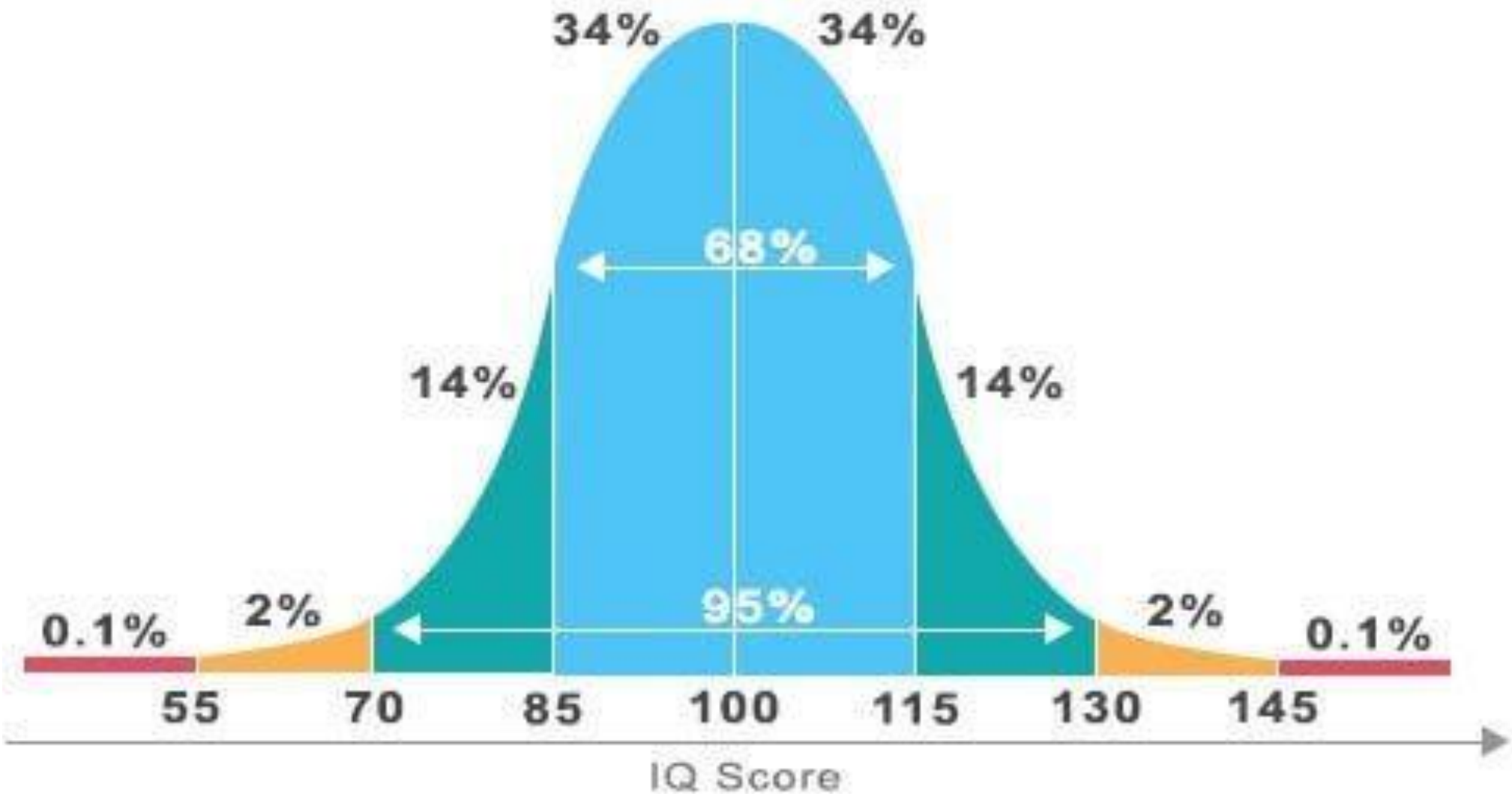
Normal Distribution (68-95-99.7 Rule)

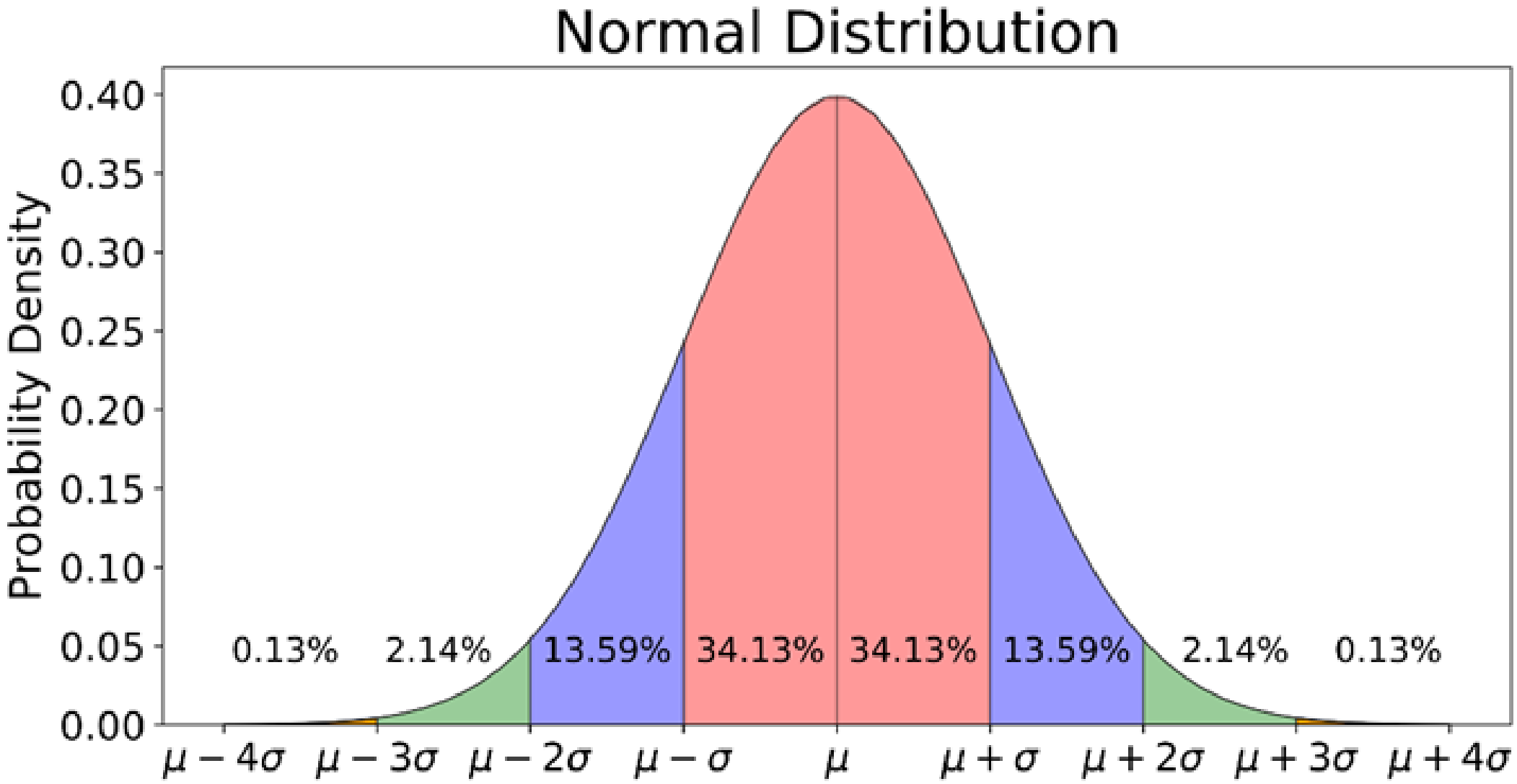
For a normal distribution:

- 68% of the observations are within +/- one standard deviation of the mean
- 95% are within +/- two standard deviations
- and 99.7% are within +- three standard deviations



IQ GRAPH

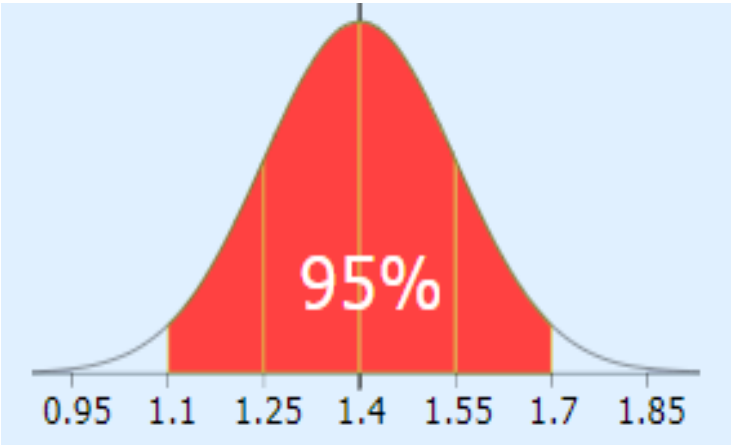



Normal Distribution
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
0.13%
$\mu - 4\sigma$
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$
$\mu + 4\sigma$

Normal Distribution & Std. Deviation - example

- Example:
- 95% of students at school are between **1.1m and 1.7m** tall
- Mean Height of students is **1.4m**
- Assuming this data is normally distributed can you calculate the standard deviation?

- 95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:
  - 1 standard deviation = (1.7m-1.1m) / 4 = 0.6m / 4
  - = 0.15m



z-score

- The z-score (z-value) is the number of standard deviations that a particular X value is away from the mean. The formula for finding the z value is:

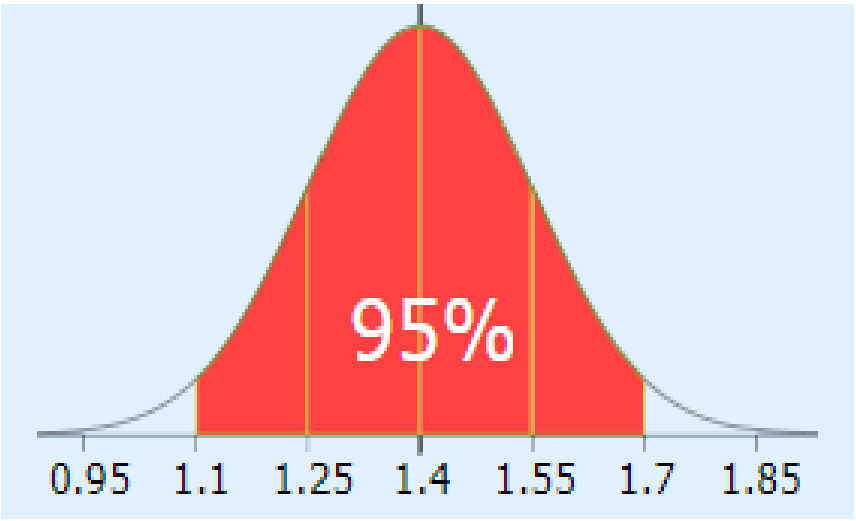
$$z = \frac{\text{value - mean}}{\text{standard deviation}}$$

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

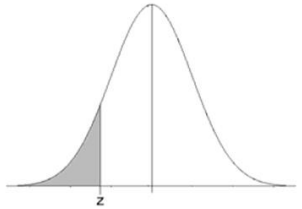
Normal Distribution & z-score example

- Example: In that same school one of your friends is 1.85m tall.

- *How far is 1.85 from the mean?*
  - It is 1.85 - 1.4 = **0.45m from the mean**
- *How many standard deviations is that?*
  - The standard deviation is 0.15m, so:
  - 0.45m / 0.15m = **3 standard deviations**
- You can also see on the bell curve that 1.85m is **3 standard deviations** from the mean of 1.4, so:
  - Your friend's height has a "z-score" of 3.0



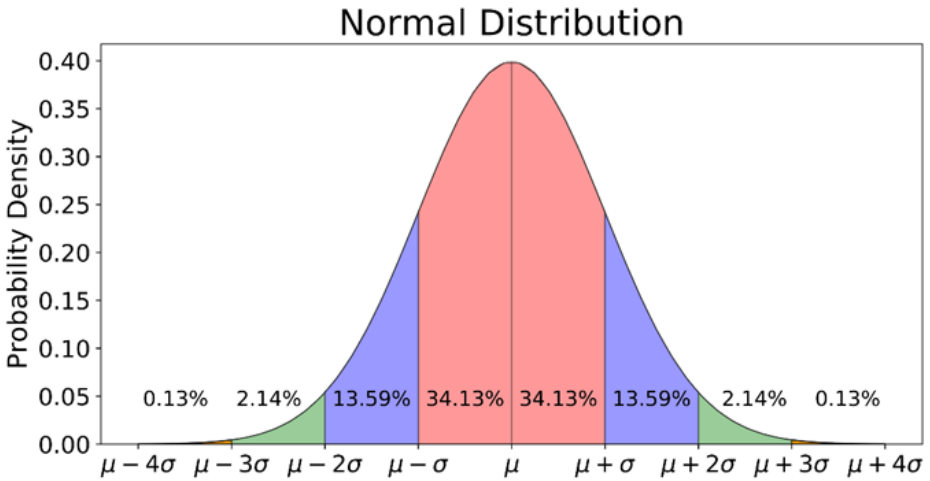
Standard Normal Cumulative Probability Table



Cumulative probabilities for NEGATIVE z-values are shown in the following table:

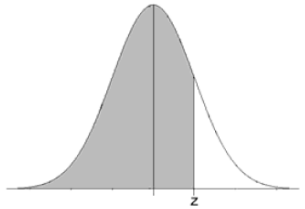
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

- z = -2.0
  - Cumulative Probability=0.0228
- z = -1.0
  - Cumulative Probability=0.1587
- z = -1.96
  - Cumulative Probability=0.025



https://www.dummies.com/education/math/statistics/how-to-use-the-z-table/

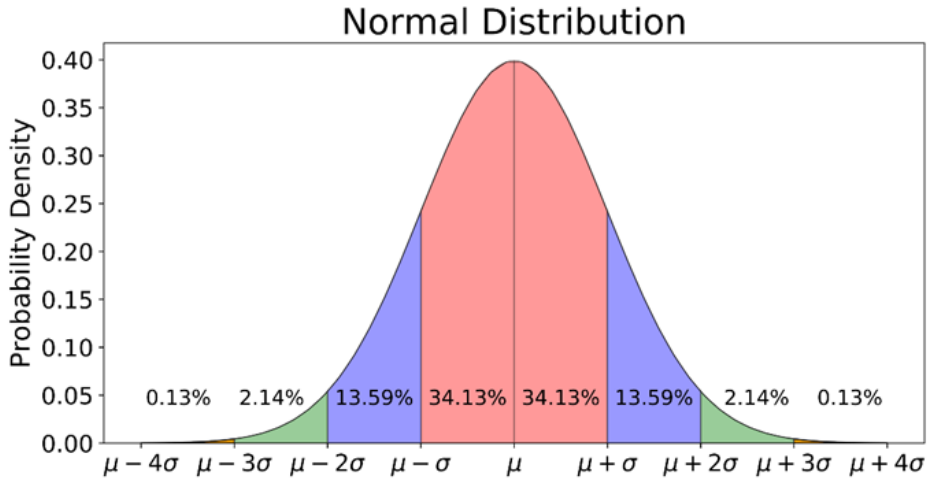
Standard Normal Cumulative Probability Table



Cumulative probabilities for POSITIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

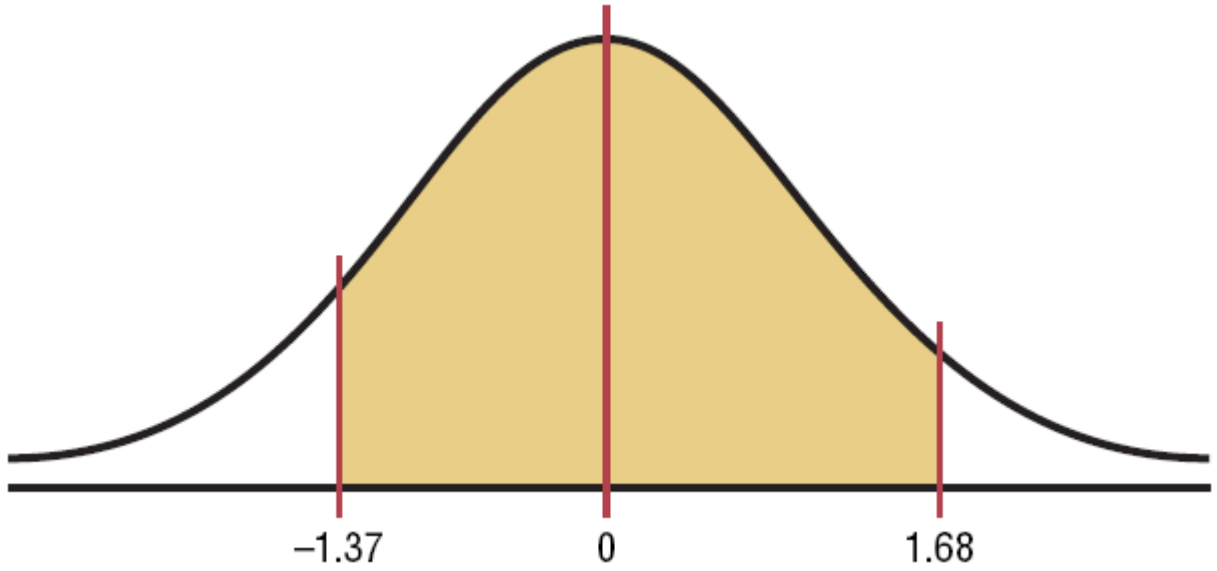
- z = 2.0
  - Cumulative Probability=0.9772
- z = 1.0
  - Cumulative Probability=0.8413
- z = 1.96
  - Cumulative Probability=0.9750



https://www.dummies.com/education/math/statistics/how-to-use-the-z-table/

Example : Area under the Curve

Find the area between z = 1.68 and z = -1.37.



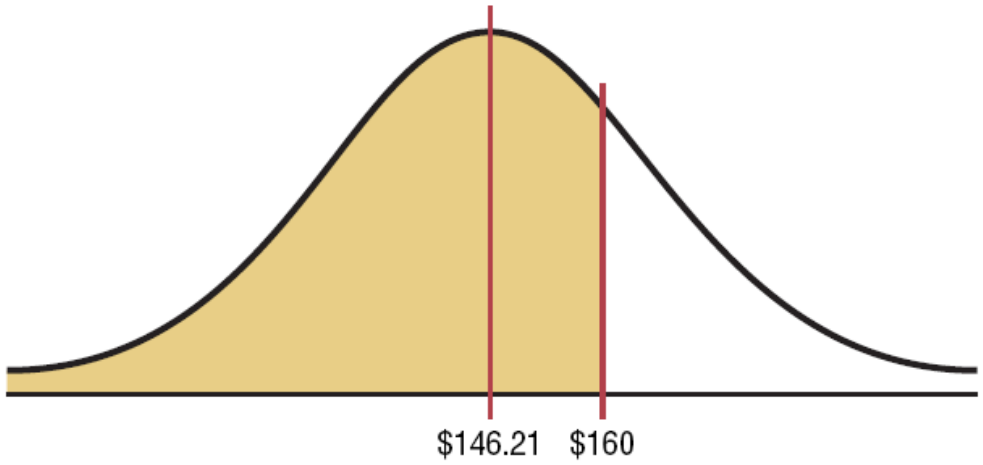
- For z = 1.68, cumulative probability = 0.9535
- For z = -1.37, cumulative probability = 0.0853
- Area under the curve, between z=1.68 and z=-1.37 will be
  - 0.9535 – 0.0853
  - = 0.8683

Example : Holiday Spending

- A survey by the National Retail Federation found that women spend on average $146.21 for the Christmas holidays.
- Assume the standard deviation is $29.44.
- Find the percentage of women who spend less than $160.00.
- Assume the variable is normally distributed.

- Lets find the value of z for $160

$$z = \frac{X - \mu}{\sigma} = \frac{160.00 - 146.21}{29.44} = 0.47$$



- From the table, area to the left of z = 0.47 is 0.6808
- Thus, 68.08% of the women spend less than $160

Normal Distribution

Consider the data representing the cost of a cup of coffee (USD) collected from different coffee shops in an area

[1,1,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5,5,5,5,5,5,5,6,6,6,6,7,7,7,7,8,8,8,8,9,9,9,10,10]

Mean of the above data is **5.4**
Median is **5.0**
Mode is **5.0**

Note: In the case of a perfect normal distribution, the mean, median and mode are equal. In the above example, we might not get a perfectly symmetrical bell curve !

Standardizing a Normal Distribution

Normally distributed data into a standard normal distribution

[1,1,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5,5,5,5,5,5,5,6,6,6,6,7,7,7,7,8,8,8,8,9,9,9,10,10]

Mean of the above data is **5.4**
Standard Deviation is **2.37**

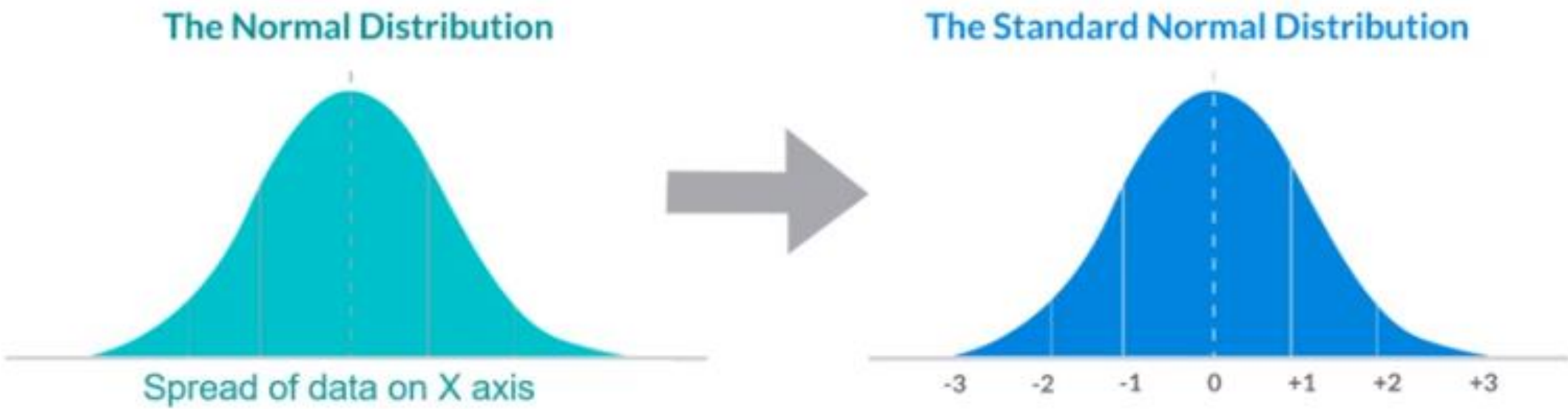
Every data point in the above data set is subtracted from the mean and is divided by the standard deviation. The result obtained is called as a **Z-score**

$$Z = \frac{X - \mu}{\sigma}$$

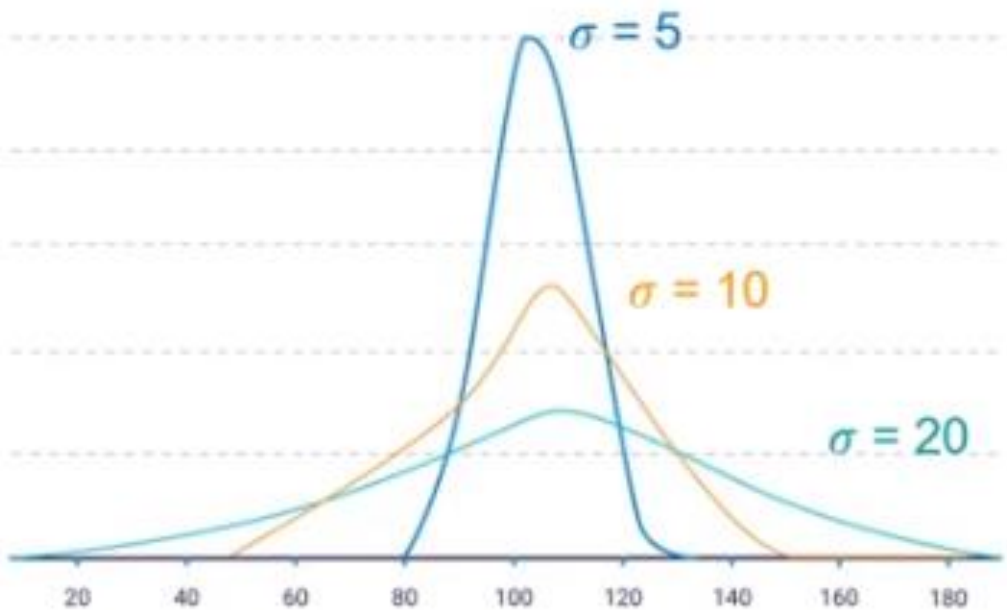
X can take each value in the data set
μ (Mu) is the mean (**5.4**) and σ (sigma) is standard deviation (**2.37**)

Standardization

Converting observations of normally distributed data into a standard scale using the Z score is called standardization. The standard normal distribution is symmetric at a Z score of zero.



Effect of Standard Deviation on Distribution



Standard Deviation of Normal Distribution

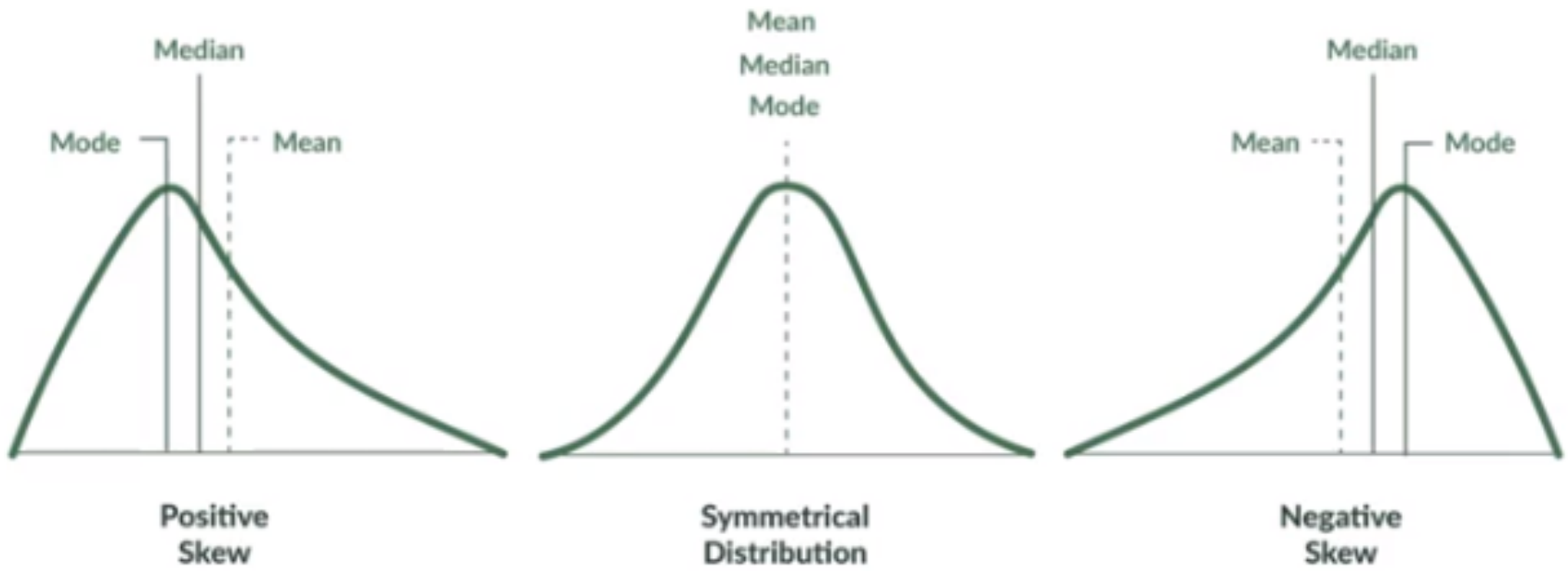
It determines the spread of the distribution

If the SD is more, then the distribution curve is wide, if the SD is small then the distribution curve is narrow

It is denoted using $\sigma(Sigma)$

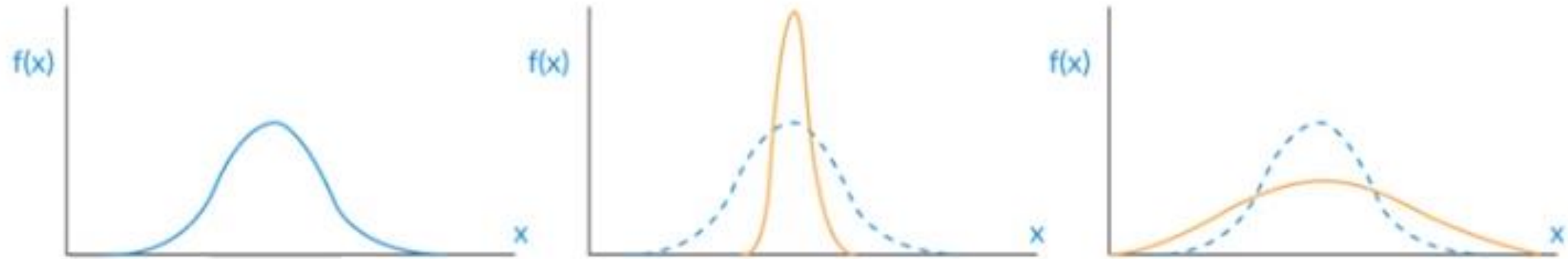
Skewness

Skewness is a measure of symmetry, precisely lack of symmetry. A distribution is symmetric when it looks the same to the left and right of the centre



Kurtosis

Kurtosis is a measure of whether data are heavy-tailed or light-tailed with respect to normal distribution



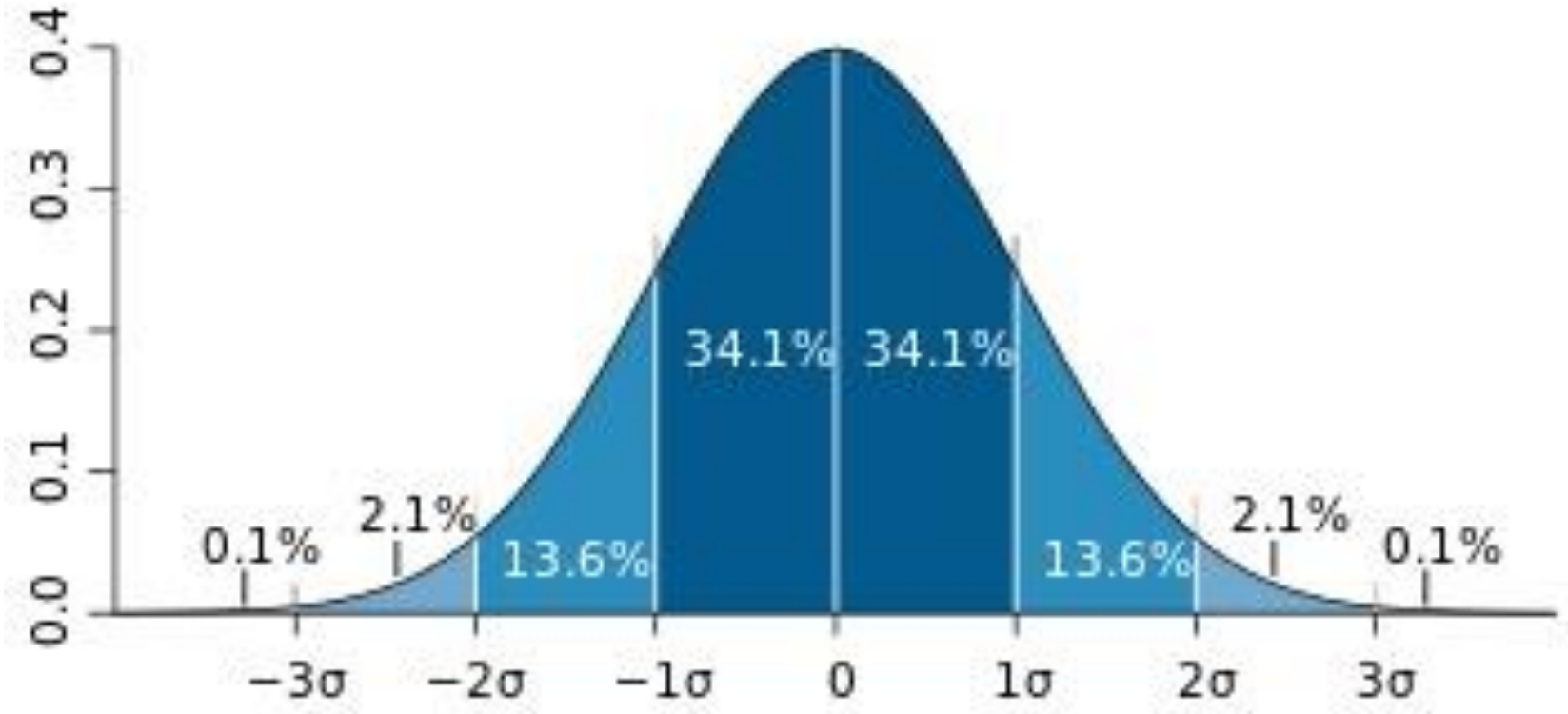
Zero kurtosis
Gaussian distribution

Positive kurtosis

Negative kurtosis

The kurtosis mainly speaks about the pointedness of the probability distribution curve

Standard Normal distribution –> mean = 0



Standard Scaling – scales a feature, where mean is zero

Z-score = (x-Mu)/Sigma (std deviation Away from mean)

Percentages for every half of a standard deviation