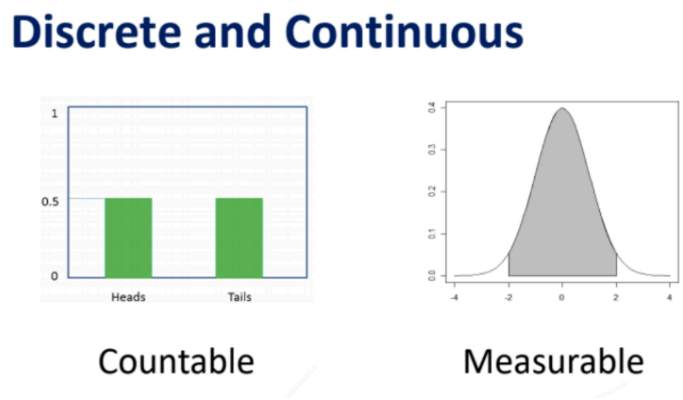
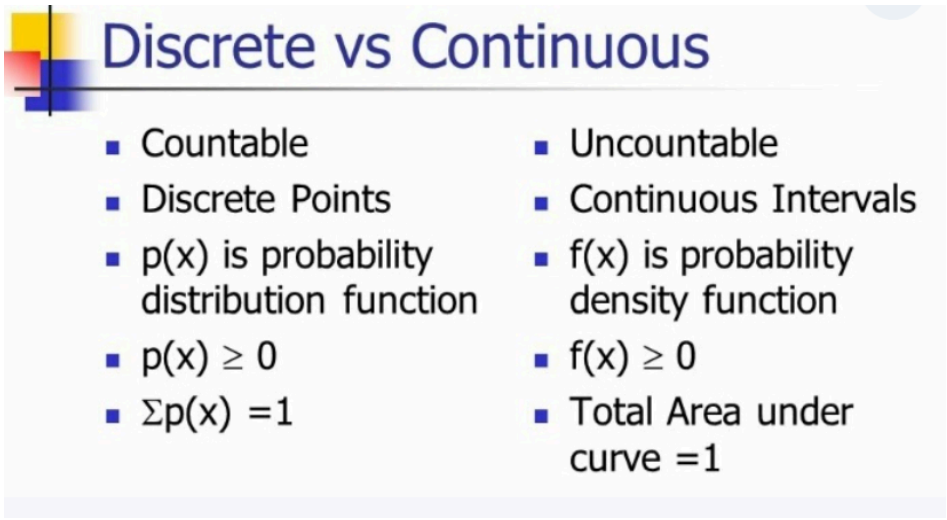
Random Variable

A variable that can take multiple values with different probabilities

Two types of Random Variable - Discrete , Continuous



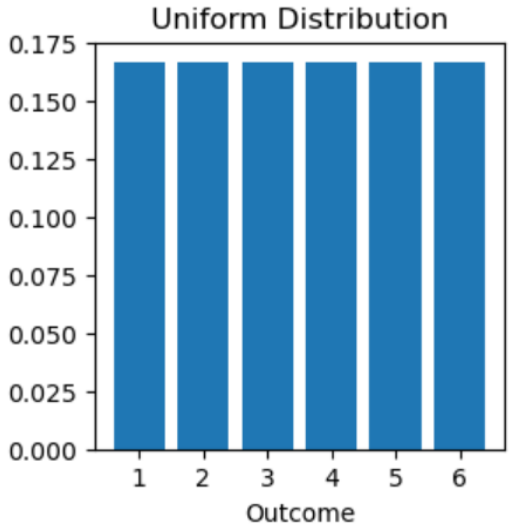


The probability distribution function for a random variable describes how the probabilities are distributed over the values of the random variable. For Discrete,it's called **Probability Mass Function** and for continuous it's called **Probability Density Function.**

Discrete Distribution Examples

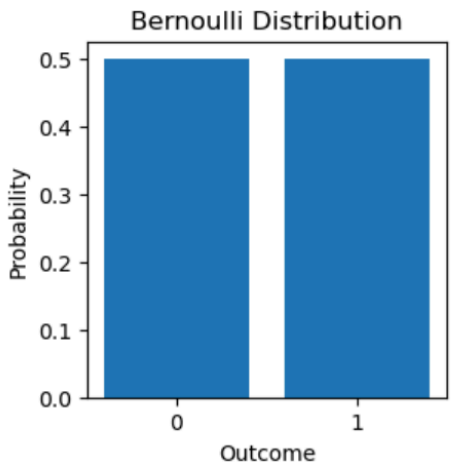
Uniform Distribution:

- The uniform distribution is a probability distribution where each possible outcome has equal probability.
- In other words, it represents a situation where all outcomes are equally likely.
- For example, rolling a fair six-sided die would result in a uniform distribution because each side has a 1/6 chance of occurring.



Bernoulli Distribution

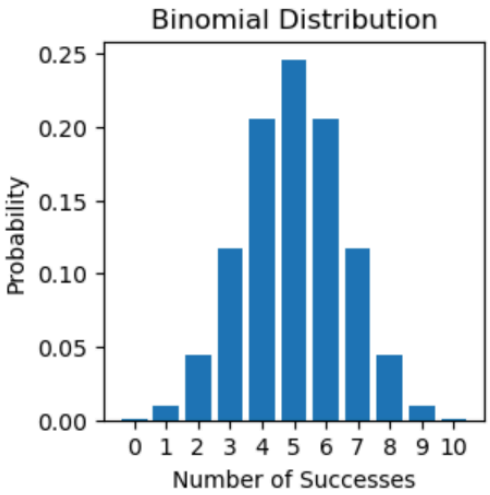
- It represents a single trial of a binary experiment, where there are only two possible outcomes (e.g., success or failure).
- Examples include flipping a coin (success = heads, failure = tails) or testing if a light bulb works (success = working, failure = not working).



Binomial Distribution

A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.

- The outcomes need not be equally likely.
- Each trial is independent.
- A total number of n identical trials are conducted.
- The probability of success and failure is the same for all trials. (Trials are identical.)



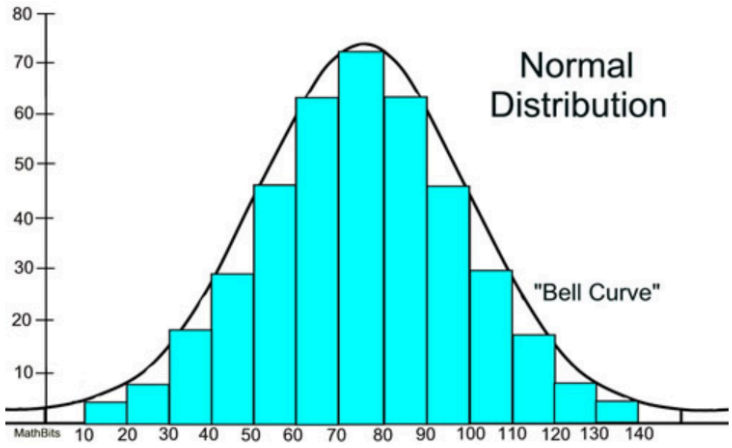
Continuous Distribution

Normal Distribution

- Normal distribution represents the behaviour 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.

Characteristics of a Normal Distribution:

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



CENTRAL LIMIT THEOREM

Theorem - The central limit theorem in statistics states that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

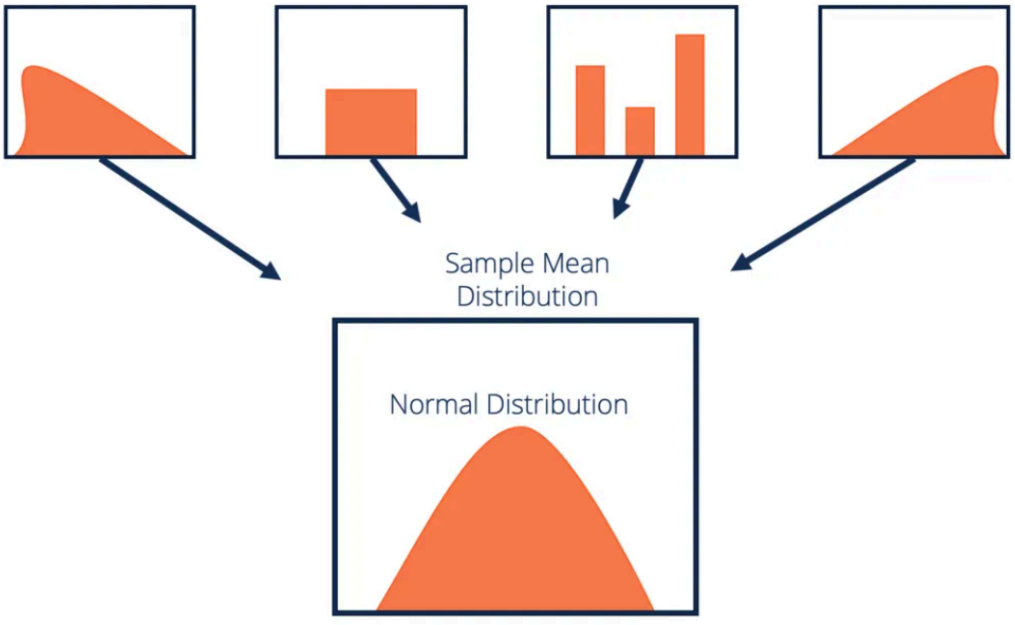
(In Layman's term – even if the data is not normally distributed, the mean of the distribution is normal distribution provided the sample size is large).

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger.

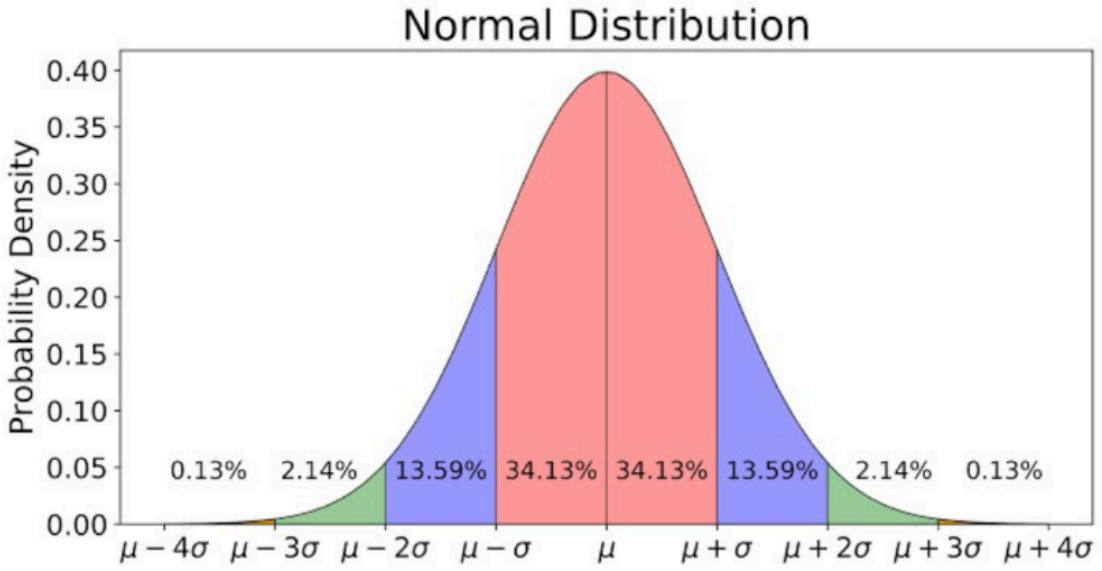
- Sample sizes equal to or greater than 30 are considered sufficient for the CLT to hold.

- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.

- A sufficiently large sample size can predict the characteristics of a population accurately.



Empirical Rule

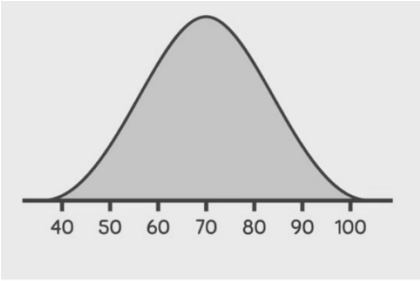


The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

- 68.26% of data falls within the first standard deviation from the mean.
- 95.44% fall within two standard deviations from the mean.
- 99.72% fall within three standard deviations from the mean.
- 0.28% of the data lies outside three standard deviations mean (μ), And this part of the data is considered as outliers.

The rule is also called the 68-95-99 Rule or the Three Sigma Rule.

Q1: The Normal distribution has a standard deviation of 10 and mean 70. Approximately what area is between 70 and 90?



Ans: 47.72

Q2. The mean life of a tire is 30,000 km. The standard deviation is 2000 km.Then, 68% of all tires will have a life between _____ km and _____ km.

Q3. In 2019 base salary of NYC employees between $1000 and $150000.

Mean =$ 73,555.88
SD=$ 27,505.98
What proportion of the population makes at most $128,567.80?

Ans: Approximately 97.5% of the population makes at most $128,567.80.

Q4.What proportion of the population makes more than $101,061.90?
Mean =$ 73,555.88
SD=$ 27,505.98

Ans: Approximately 15.86% of the population makes more than $101,061.90

Q5 . One student score 90 marks in Mathematics and 75. Marks in English. Consider the average class score of Mathematics is 85, SD is 5. Average class score of English is 60 , SD is 5. Which performance is better: Mathematics or English?

In Maths, he scored better than 84% of students and in English he score better than 99.5% of the students.

Q6. In a Garden the heights of plants are normally distributed, the mean of the plants are 22.2 inches and the SD of 4.5 inches. Estimate the percentage of plants that are less than
13.2 inches tall.

Ans: Approximately 2.27% of the plants are less than 13.2 inches tall in the garden.

Q7.A competency test has scores with a mean of 80 and a standard deviation of 10. A histogram of the data shows that the distribution is normal. Use the Empirical Rule to find the percentage of scores between 60 and 90.

Ans: Approximately 81.85% of the scores are between 60 and 90

Q8. The mean June midday temperature in Delhi is 36°C and the standard deviation is 3°C Assuming this data is normally distributed, how many days in June would you expect the midday temperature to be between 39°C and 42°C?

Ans: Approximately 13.59% of June will be between 39°C and 42°C
ie 4.077 days

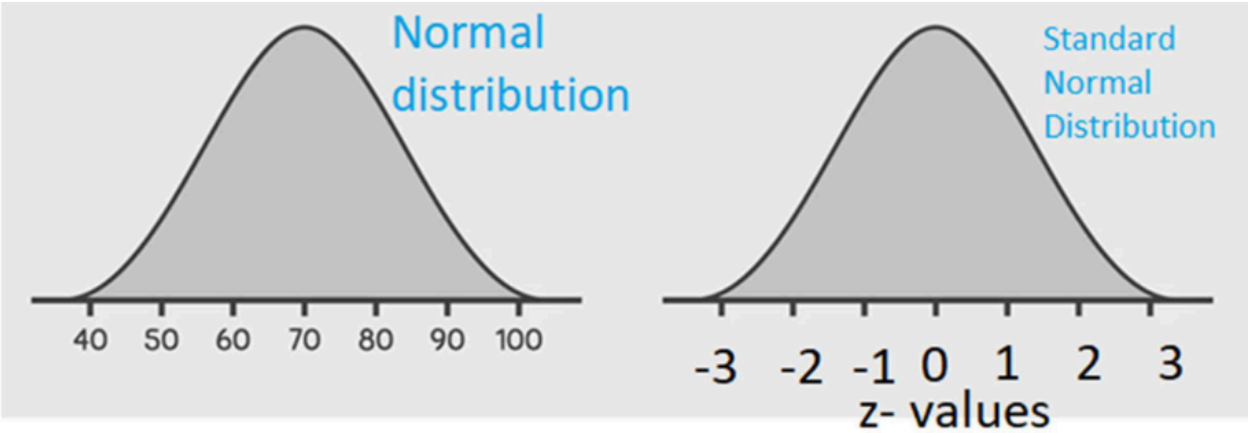
Z-Score

Z Score gives how many standard deviations away from mean a value is. However, to understand the probability associated with it, we need to refer to Z-Table.

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

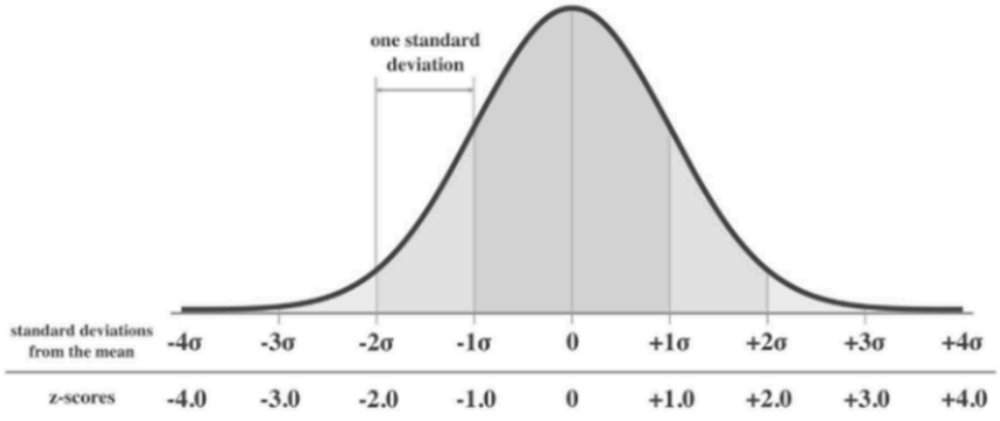
μ = Mean
σ = Standard Deviation

Q1. For a recent final written statistics exam for a “Data scientist” job selection process, the mean was 70 with a standard deviation of 10. If you scored 76 marks. What is your percentile or (area in the Normal distribution)?.



Mean value 70 Z score = (70-70)/10=0
80 marks Z score = (80-70)/10=1
60 marks Z score = (60-70)/10=-1
76 marks Z score =(76-70)/10=0.60
In the z table the value of 0.60 is 0.7257.
This is the value of area under curve or the percentile.

In Standard Normal Distribution mean = 0 and std = 1.
We can convert any normal distribution to standard normal distribution

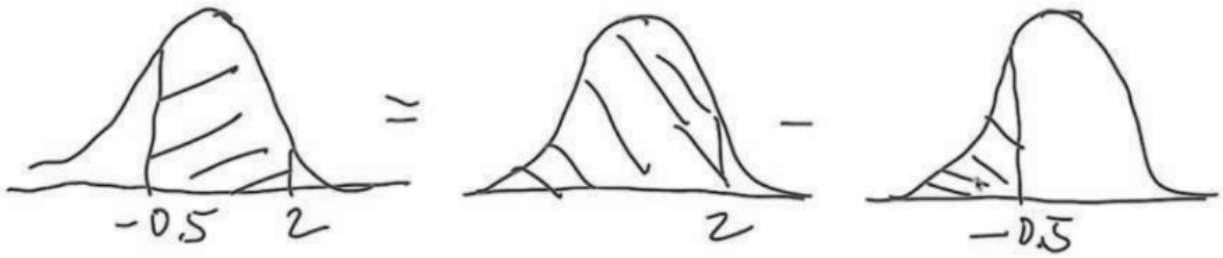


2 Types of problems :
Type 1: Comparison of 2 different Normally Distributed values (Z-Score is enough)
Type 2: Finding the probability or percentage of values. (Need Z-table)

Q1. Fathers height follows normal distribution with a mean of 68.3 inches and a SD of 1.8 inches. What percentage of fathers have heights between 67.4 and 71.9?

Ans:
Step 1: Standardize
Z(67.4) = 67.4-68.3/1.8 = -0.5
Z(71.9) = 71.9-68.3/1.8 = 2

Step 2:
Area to left of the point z ie -0.5 = 0.3085
Area to left of the point z ie 2 = 0.9772



Percentage of fathers have heights between 67.4 and 71.9 = 0.9772-0.3085

= 0.668

Q2. What is the 30 percentile of the Father's height?

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

$\mu =$ Mean
$\sigma =$ Standard Deviation

Z = -0.5217
Mean = 68.3, sd = 1.8 ,x = (-0.5217*1.8) + 68.3

Q3. One student score 80 marks in Mathematics and 75 Marks in English. At this point we can say he has performed excellent in Math as compare to English.
Consider the average class score of Mathematics is 90, SD is 5. Average class score of English is 60 , SD is 5. Verify the performance?

Ans:- lets calculate the Z values
Zm=(80-90)/5= -2
Ze=(75-60)/5=3

3 is the higher value and it's closer to +3

Q4. What proportion of students are between 5.81 feet & 6.3 feet height. Given Mean=5.5, sd=0.5 feet.

Z1=(5.81-5.5)/0.5= 0.619
Z2=(6.3-5.5)/0.5= 1.599

Step 2:
Area to left of the point z ie 0.619 = 0.7291
Area to left of the point z ie 1.599 = 0.9772

Percentage of students heights between 5.81 and 6.3 = 0.9772-0.7291
= 0.2481 / 24.81%

Q5. Mean height of Gurkhas is 146 cm with Sd of 3 cm . What is the probability of
(a)Height greater than 152 cm.
(b)Height between 140 and 150 cm.

Ans:

(a) Z score (152) = (152-146)/3 = 2
Area to the right = 1 - 0.9772 = 0.0228

Calculate Z scores
(b) Z score (140) = (140 -146)/3 = -2
Z score (150) = (150 -146)/3 = 1.333

Step 2:
Area to left of the point z ie -2 = 0.0228
Area to left of the point z ie 1.333 = 0.9082

Height between 140 and 150 cm = 0.9082 - 0.0228 = 0.8860 / 88.6%

Q6. According to the Center for Disease Control, heights for U.S. adult females and males are approximately normal.
Females: mean of 64 inches and SD of 2 inches
Males: mean of 69 inches and SD of 3 inches
Find the probability of a randomly selected U.S. adult female being taller than 65 inches.