upGrad
Campus

upGrad

Course : Statistics for Business

Lecture On : Inferential Statistics

Instructor :

Session Agendas

In this session we will discuss and learn the concepts of Inferential Statistics.

Introduction to Probability

The two main counting principles through which you can calculate the probability for the outcomes:

1. **Permutations**: A permutation is a way of arranging a selected group of objects in such a way that the order is of significance.

   When there are 'n' objects to be arranged among 'r' spaces, the permutation value is given by the formula $$\frac{n!}{(n-r)!}$$

2. **Combinations:** When you just have to choose some objects from a larger set, and the order is of no significance, then the rule of counting that you use is called combination.

   If you want to choose 'r' objects from a larger set of 'n' objects, then the number of ways in which you can do that is given by the formula

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

What is probability?

$$Probability = \frac{No\ of\ desired\ outcomes}{Total\ no\ of\ possible\ outcomes}$$

Probability values have the following two major properties:

- **Probability values always lie in the range of 0-1:**
  The value is 0 in the case of an impossible event and 1 in the case of a sure event.

- **The probabilities of all outcomes for an experiment always sum up to 1.**

Important terms related to probability

Experiment: Essentially, any scenario for which you want to compute the probabilities for that scenario to be considered an experiment. It is of two types:

- Deterministic: Outcome is the same every time.
- Random: Outcome can take many possible values.

Sample Space: A sample space is nothing but a list of all possible outcomes of a random experiment. It is denoted by S = {all the possible outcomes}.

Event: It is a subset, i.e., a part of the sample space that you want to be true for your probability experiment.

Types of Events

Independent Events: If you have two or more events and the occurrence of one event has no bearing whatsoever on the occurrence/s of the other event/s, then all the events are said to be independent of each other.

Disjoint or Mutually Exclusive Events: Now, two or more events are mutually exclusive when they do not occur at the same time; i.e., when one event occurs, the rest of the events do not occur.

Complement rule for probability

It states that if A and A' are two events that are mutually exclusive/ disjoint and are complementary / negation of each other, then

$$P(A) + P(A') = 1$$

Rules of Probability

Addition Rule: When you have the individual probabilities of two events A and B, denoted by P(A) and P(B), the addition rule states that the probability of the event that either A or B will occur is given by

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

where $P(A \cup B)$ denotes the probability that either the event A or B occurs.

P(A) denotes the probability that the event A occurs

P(B) denotes the probability that the event B occurs

$P(A \cap B)$ denotes the probability that both the events A and B occur simultaneously.

Multiplication Rule: When an event A is not dependent on event B and vice versa, they are known as independent events. And the multiplication rule allows us to compute the probabilities of both of them occurring simultaneously, which is given as

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

Comparison between Addition Rule and Multiplication Rule

- The addition rule is generally used to find the probability of multiple events when either of the events can occur at that particular instance.
- The multiplication rule is used to find the probability of multiple events when all the events need to occur simultaneously.

Basics of Probability

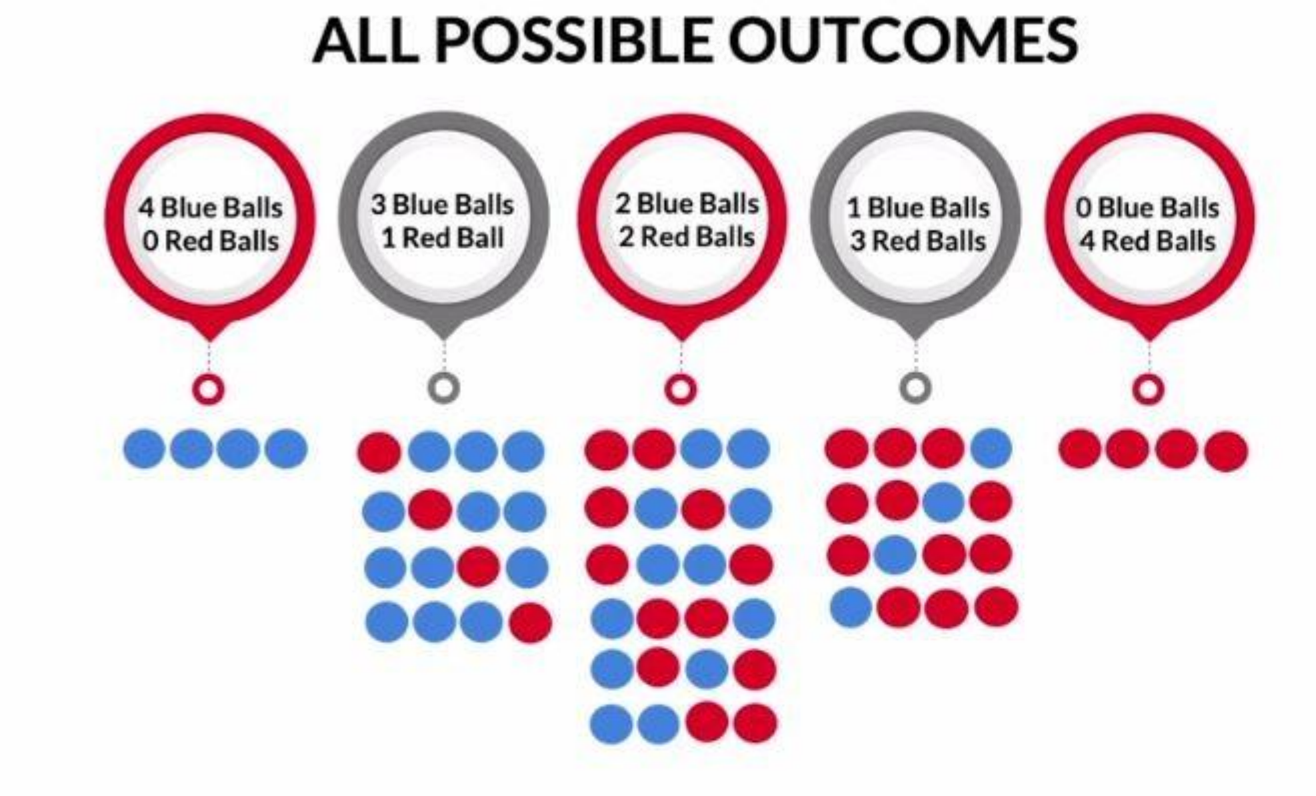
Question:

1. A bag with 2 balls -- one blue and one red
2. You are required to pick a ball front he bag put it back in.
3. Repeat this process 4 times.
4. If you get a red ball on all four occasions, you win 200 rupees.
5. If you get any other combination, you lose 10 rupees.

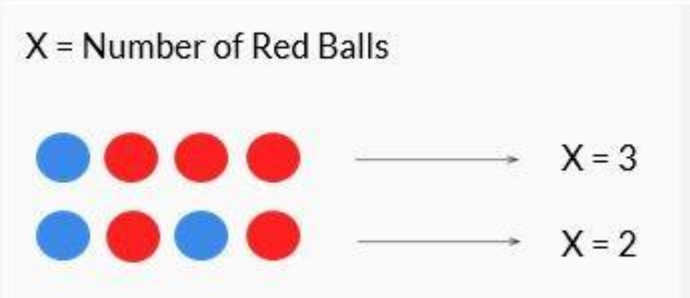
Solution: The approach:

1. Find possible combinations
2. Find probability of each combination
3. Use the probability to estimate profit/loss per player

1.


ALL POSSIBLE OUTCOMES
4 Blue Balls
0 Red Balls
3 Blue Balls
1 Red Ball
2 Blue Balls
2 Red Balls
1 Blue Balls
3 Red Balls
0 Blue Balls
4 Red Balls

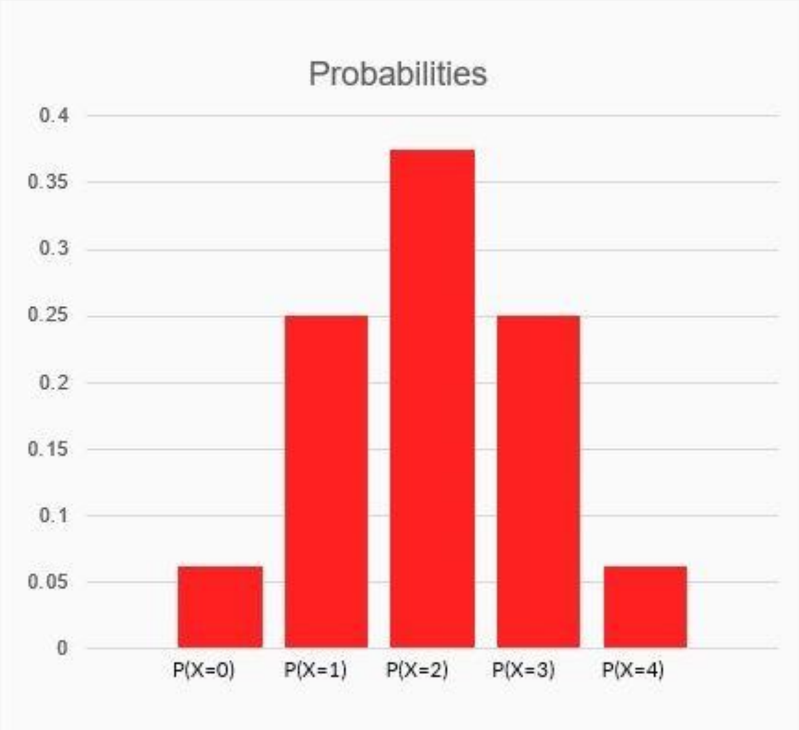
Quantify outcome using random variable:



Finding the probability distribution, which is a distribution giving us the probability for all possible values of X.

x	P(x)
0	0.0625
1	0.25
2	0.375
3	0.25
4	0.0625

In bar chart form:



Finding the expected value for X, the money won by a player after playing the game once:

The expected value (EV) for X was calculated using the formula:

$$EV(X) = x_1 * P(X = x_1) + x_2 * P(X = x_2) + \dots\dots\dots\dots\dots + x_n * P(X = x_n)$$

Another way of writing this is

$$EV(X) = \sum_{i=1}^{i=n} x_i * P(X = x_i)$$

Calculating the answer this way, we find the expected value to be **+2.**

In conclusion:

If we conduct the experiment (play the game) infinite times, the average money won by a player would be ₹2. Hence, we decided that we should either decrease the prize money or increase the penalty to make the expected value of X negative. A negative expected value would imply that, on average, a player would be expected to lose money and the house would profit.

Probability Distributions

Uniform distribution: It is a discrete probability distribution, where the probability of each outcome is exactly the same.

Eg: The most basic example, in this case, is rolling a die. When you roll an unbiased die, the chances of getting any of the numbers are equally probable.

A discrete uniform distribution is a probability distribution that has 'n' discrete outcomes, and the probability of each of these outcomes is the same, i.e., **1/n.**

Cumulative probability: Cumulative probability of X, denoted by F(X), which is the probability that the random variable X takes a value less than or equal to x.

$$F(X) = P(X \leq x)$$

$$\text{E.g. } F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

Cumulative Probability Distribution
Continuous Distribution

Eg:

x	$F(x) = P(X \le x)$
50	0.167
55	0.250
60	0.367
65	0.467
70	0.550
75	0.650
80	0.733
85	0.833
90	0.933
95	0.967
100	1.000

X = Weight of an employee

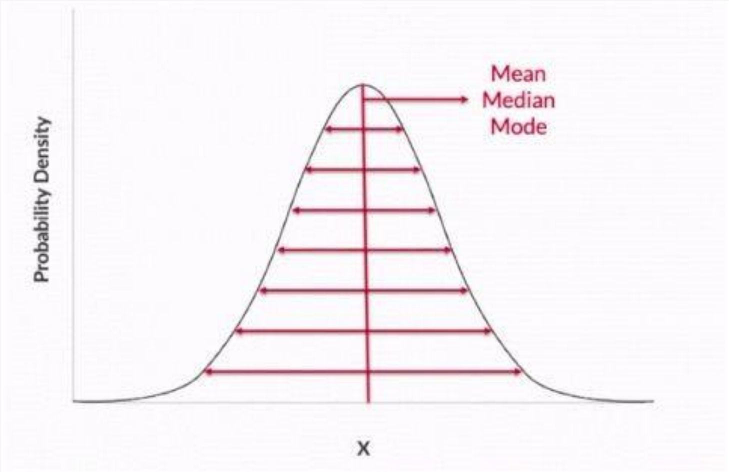
$P(X \le 60) = 0.367$

$$\begin{aligned} P(60 \le X \le 65) &= P(X \le 65) - P(X \le 60) \\ &= 0.467 - 0.367 \\ &= 0.1 \end{aligned}$$

Continuous Probability Distribution
Normal Distribution

Normal distribution is a probability distribution that is symmetric about the mean, showing that 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.

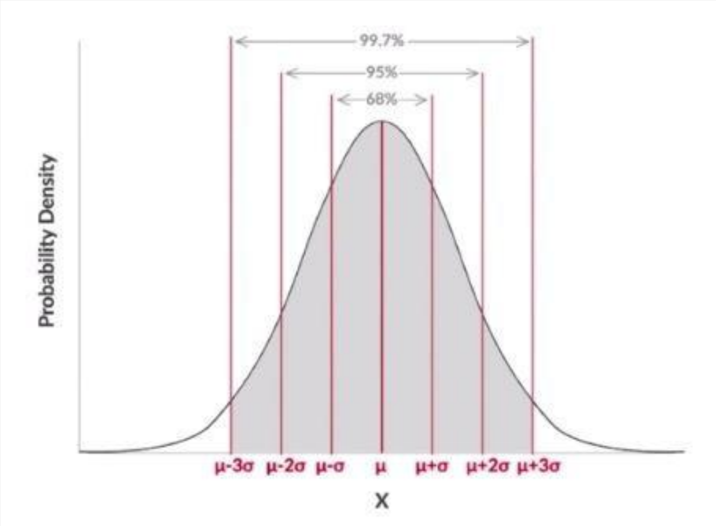
- A normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.



1-2-3 Rule

This rule states that there is a:

- 68% probability of the variable lying within 1 standard deviation of the mean.
- 95% probability of the variable lying within 2 standard deviations of the mean.
- 99.7% probability of the variable lying within 3 standard deviations of the mean.

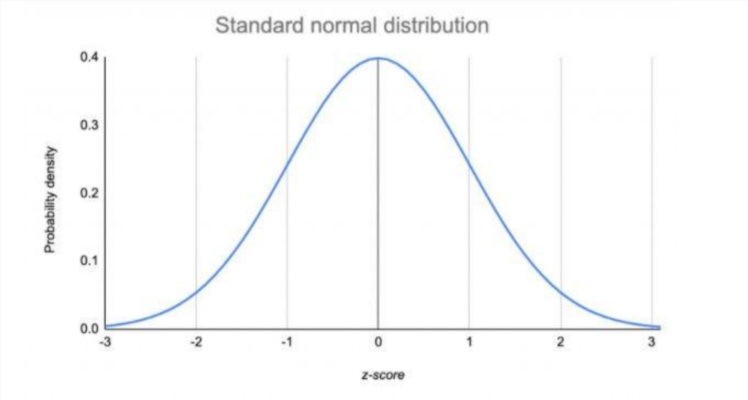


Standard normal distribution

The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.

Any normal distribution can be standardized by converting its values into z-scores. Z-scores tell you how many standard deviations from the mean each value lies.

Converting a normal distribution into a z-distribution allows you to calculate the probability of certain values occurring and to compare different data sets.



Samples and Sampling

Sample is the subset of the population. The process of selecting a sample is known as sampling. Number of elements in the sample is the sample size.

Sampling

Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N} (X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n} (X_i - \bar{X})^2}{n-1}$

Sampling Distribution

A **sampling distribution** is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

- A sampling distribution is a statistic that is arrived out through repeated sampling from a larger population.
- It describes a range of possible outcomes that of a statistic, such as the mean or mode of some variable, as it truly exists a population.
- The majority of data analyzed by researchers are actually drawn from samples, and not populations.

Properties of Sampling Distribution

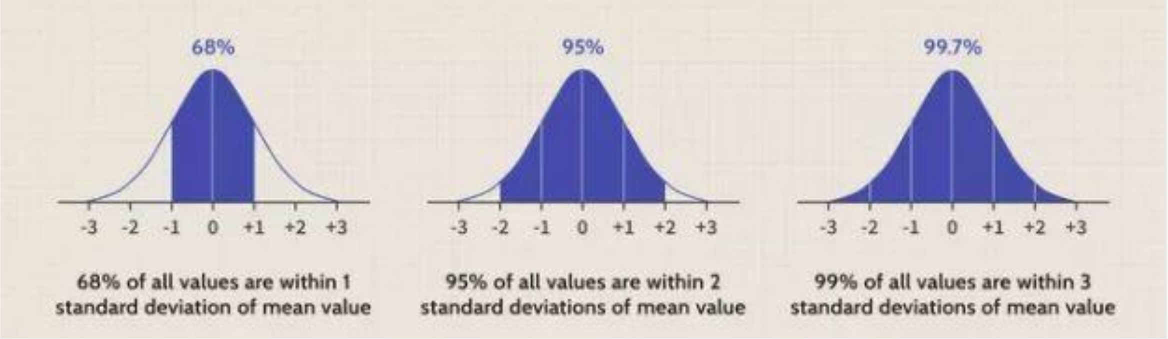
The sampling distribution will follow these three properties:

- Sampling distribution's mean $(\mu_{\bar{X}})$ = Population mean (μ)
- Sampling distribution's standard deviation (Standard error) = $\frac{\sigma}{\sqrt{n}}$, where σ is the population's standard deviation and n is the sample size
- For **n > 30**, the sampling distribution becomes a **normal distribution**

Central Limit Theorem (CLT)

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.
- 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 more accurately.



Terminologies

- The probability associated with the claim is called the **Confidence Level.**
- The maximum error made in the sample mean is called the **Margin of Error.**
- The final interval of values is called the **Confidence Interval**.

Generalised Formula for Estimating Population Mean

Sample Mean ($\bar{X}$)
Sample Standard Deviation (S)
Sample Size (n)

Confidence Interval (y% confidence level) =
$(\bar{X} - \frac{Z^*S}{\sqrt{n}}, \bar{X} + \frac{Z^*S}{\sqrt{n}})$

Where Z^* is the Z-score associated with y% confidence level

Confidence Level	Z^*
90%	± 1.65
95%	± 1.96
99%	± 2.58

Steps for estimating Population Parameters using CLT

The three steps to follow are as follows:

- First, take a sample of size n.
- Then, find the mean $\bar{X}$ and standard deviation S of this sample.
- Now, you can say that for y% confidence level, the confidence interval for the population mean mu is given by

$$\left(\bar{X} - \frac{Z^* S}{\sqrt{n}}, \bar{X} + \frac{Z^* S}{\sqrt{n}}\right).$$

Any Queries?

Thank You!