Bernoulli Distribution

27 March 2023 16:06

$p \rightarrow$ prob of success

$1-p \rightarrow$ prob of failure

$\rightarrow p$

experiment

Bernoulli distribution is a probability distribution that models a binary outcome, where the outcome can be either success (represented by the value 1) or failure (represented by the value 0). The Bernoulli distribution is named after the Swiss mathematician Jacob Bernoulli, who first introduced it in the late 1600s.

coin toss $\rightarrow$

H (1) T (0)

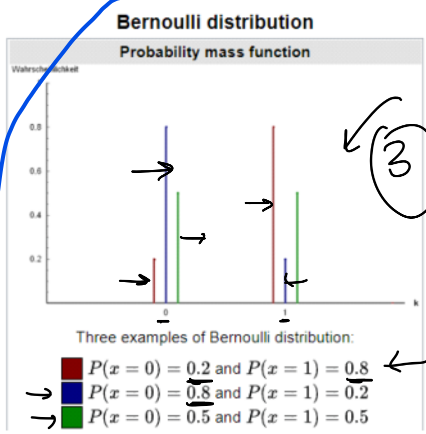
binary outcome

The Bernoulli distribution is characterized by a single parameter, which is the probability of success, denoted by p. The probability mass function (PMF) of the Bernoulli distribution is:

$$\boxed{pmf} = P(X=x) = \boxed{p^{x}(1-p)^{1-x}}$$

$$P(X=1) = \left(\tfrac{1}{2}\right)^{1}\left(\tfrac{1}{2}\right)^{0} = \tfrac{1}{2}$$

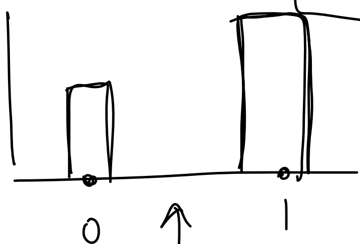
$$P(X=0) = \left(\tfrac{1}{2}\right)^{0}\left(1-\tfrac{1}{2}\right)^{1-0} = \boxed{1 \times \tfrac{1}{2}}$$



3

rolling a dice getting a 5

5 $\left(\tfrac{1}{6}\right)$ ~5 $\left(\tfrac{5}{6}\right)$



machine learning

The Bernoulli distribution is commonly used in machine learning for modelling binary outcomes, such as whether a customer will make a purchase or not, whether an email is spam or not, or whether a patient will have a certain disease or not

Binomial distribution

explaination - x a random variable in our case . we assume head be an success even therefore randome variable x will be 1. . p^x denotes probability of succes (head). and (1 - p) probability of failure(tails) rest put the values in the formula

Binomial Distribution

27 March 2023 16:36

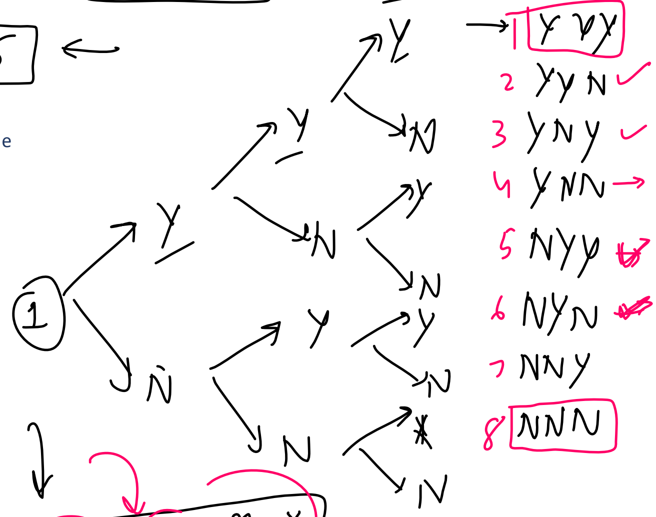
(p) $\boxed{(n, p)}$ ← $n = 1 \rightarrow$ binomial → bernoulli

Bernoulli trial / n times ⑤ → Binomial

Binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials with two possible outcomes (often called "success" and "failure"), where the probability of success is constant for each trial. The binomial distribution is characterized by two parameters: the number of trials n and the probability of success p.

$\boxed{\text{feedback}}$ → 10 students

$\boxed{0.5}$ ←



The Probability of anyone watching this lecture in the future and then liking it is 0.5. What is the probability that:

1. No-one out of 3 people will like it — $\frac{1}{8}$
1. 1 out of 3 people will like it — $\frac{3}{8}$
1. 2 out of 3 people will like it — $3/8$
1. 3 out of 3 people will like it — $1/8$

2 2 2 → $\boxed{2^3}$ →

Binomial

$$\boxed{P(X=x) = {}^{n}C_{x}\, p^{x} (1-p)^{n-x}}$$

$n \rightarrow$ # of trials

$p \rightarrow$ prob of success

$x \rightarrow$ desired result.

PDF Formula:

$${}^{3}C_{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{1}$$

$$\frac{3!}{2!\,1!} \times \frac{1}{8} \rightarrow$$

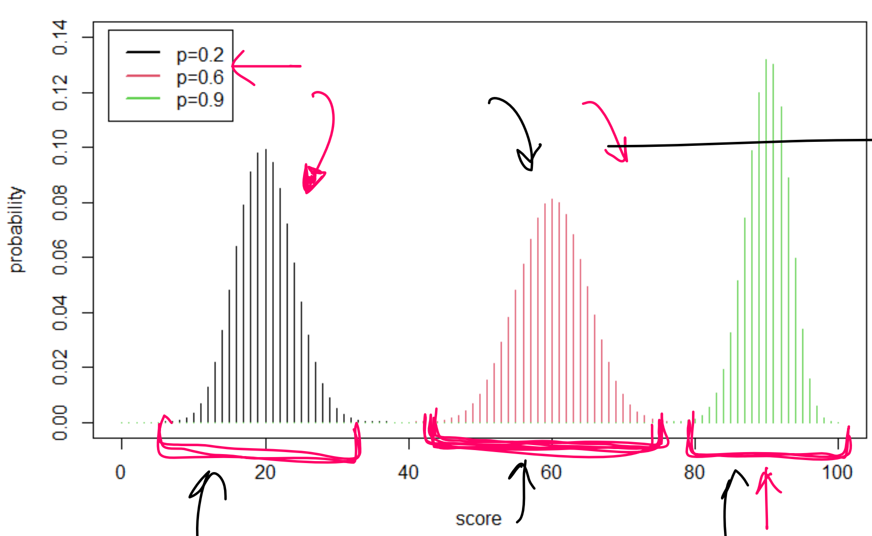
$$3 \times \frac{1}{8} = \boxed{\frac{3}{8}}$$

$\boxed{H\ T\ H\ T\ H\ T\ \_\ \_\ 10}$

$\boxed{H\ \_\ \_\ \_\ \_\ \_}$

Graph of PDF:

Binomial distribution with different probabilities of success



p(s) increase ↑

refer to the code and explaination in the folder , you will understand than why when p = 0.5 is centered and also with different values

Criteria:

Criteria:

1. The process consists of n trials
2. Only 2 exclusive outcomes are possible, a success and a failure.
3. P(success) = p and P(failure) = 1-p and it is fixed from trial to trial
4. The trials are independent.

1. **Binary classification problems**: In binary classification problems, we often model the probability of an event happening as a binomial distribution. For example, in a spam detection system, we may model the probability of an email being spam or not spam using a binomial distribution.

2. **Hypothesis testing**: In statistical hypothesis testing, we use the binomial distribution to calculate the probability of observing a certain number of successes in a given number of trials, assuming a null hypothesis is true. This can be used to make decisions about whether a certain hypothesis is supported by the data or not.

3. **Logistic regression**: Logistic regression is a popular machine learning algorithm used for classification problems. It models the probability of an event happening as a logistic function of the input variables. Since the logistic function can be viewed as a transformation of a linear combination of inputs, the output of logistic regression can be thought of as a binomial distribution.

4. **A/B testing**: A/B testing is a common technique used to compare two different versions of a product, web page, or marketing campaign. In A/B testing, we randomly assign individuals to one of two groups and compare the outcomes of interest between the groups. Since the outcomes are often binary (e.g., click-through rate or conversion rate), the binomial distribution can be used to model the distribution of outcomes and test for differences between the groups.

not possible to reach every member of population , hence draw samples from population then based on sample we try to infer for the population . so in sampling distribution we do is we draw out set of samples multiple times , the times can be multiple but the size of sample must be described.

lets say humne 100 bar 50- 50 logo ko nikala , total samples = 100*50 = 5000, now un 50 - 50 har ek samples ka mea nikala, now you have 100 sample means and that is called sampling distirbution of mean

Sampling Distribution

27 March 2023 17:10

Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as the sample mean or sample proportion) computed from multiple independent samples of the same size from a population.

population Vs sample

140 Cr

5000 → sample

25k, 100k, 200k, 15k, 0k

distributions

variance

India → sare log → pop → salary

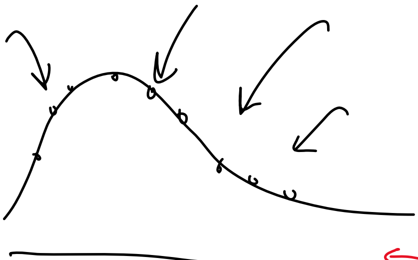
sample → (50) →

x x x . . . . . x → 50 people → $\bar{x}_1$

x x x x x . . x → 50 peop → $\bar{x}_2$

x x x x . . . . x → $\bar{x}_3$

$\bar{x}_{100}$



dist

100 times

$\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_{100}$

sampling distribut of sampe mean

Why Sampling Distribution is important?

Sampling distribution is important in statistics and machine learning because it allows us to estimate the variability of a sample statistic, which is useful for making inferences about the population. By analysing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis tests, and make predictions about the population based on the sample data.

→ central limit theorm

explaination of central limit theorm - when you plot the distribution of sampling distribution of samples mean , you will get the normal distribution regardless of the type of population distribution , even if it is log normal , exp , uniform or any random , but when you plot the sampling distribution of samples you will get the normal distribution but there are certain assumptions

Note - the mean of population distribution and sampling distribtion will be same , and the variance of population is sigma^2 , the variance of sampling distribution will be sigma^2/n where n is the sample size.

refer to the notebooks in this folder to have a btter understanding

Central Limit Theorem

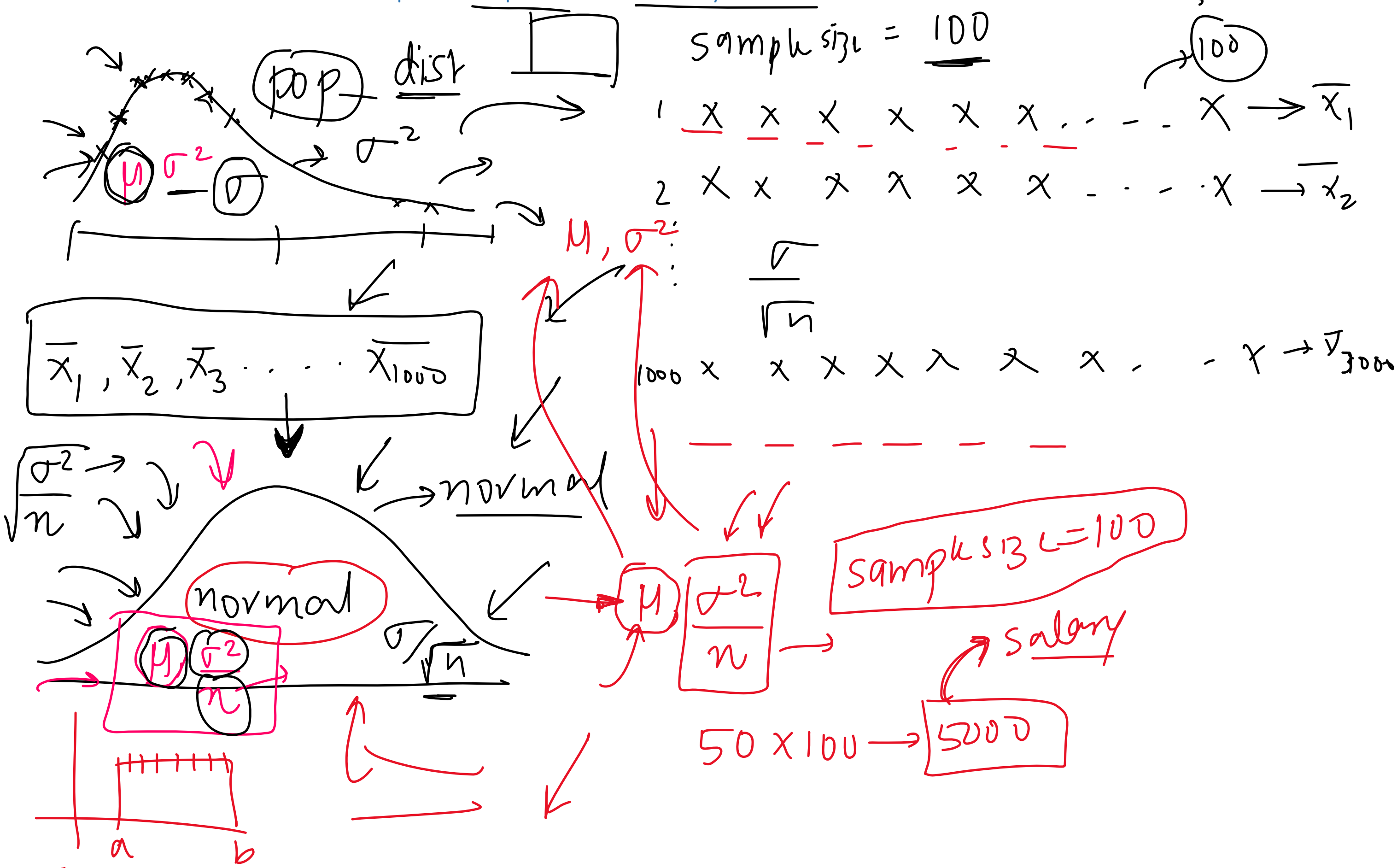
27 March 2023 17:10

interval $\boxed{\mu = 32.584}$ → fare $\mu = 32.584$

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables.

The conditions required for the CLT to hold are:

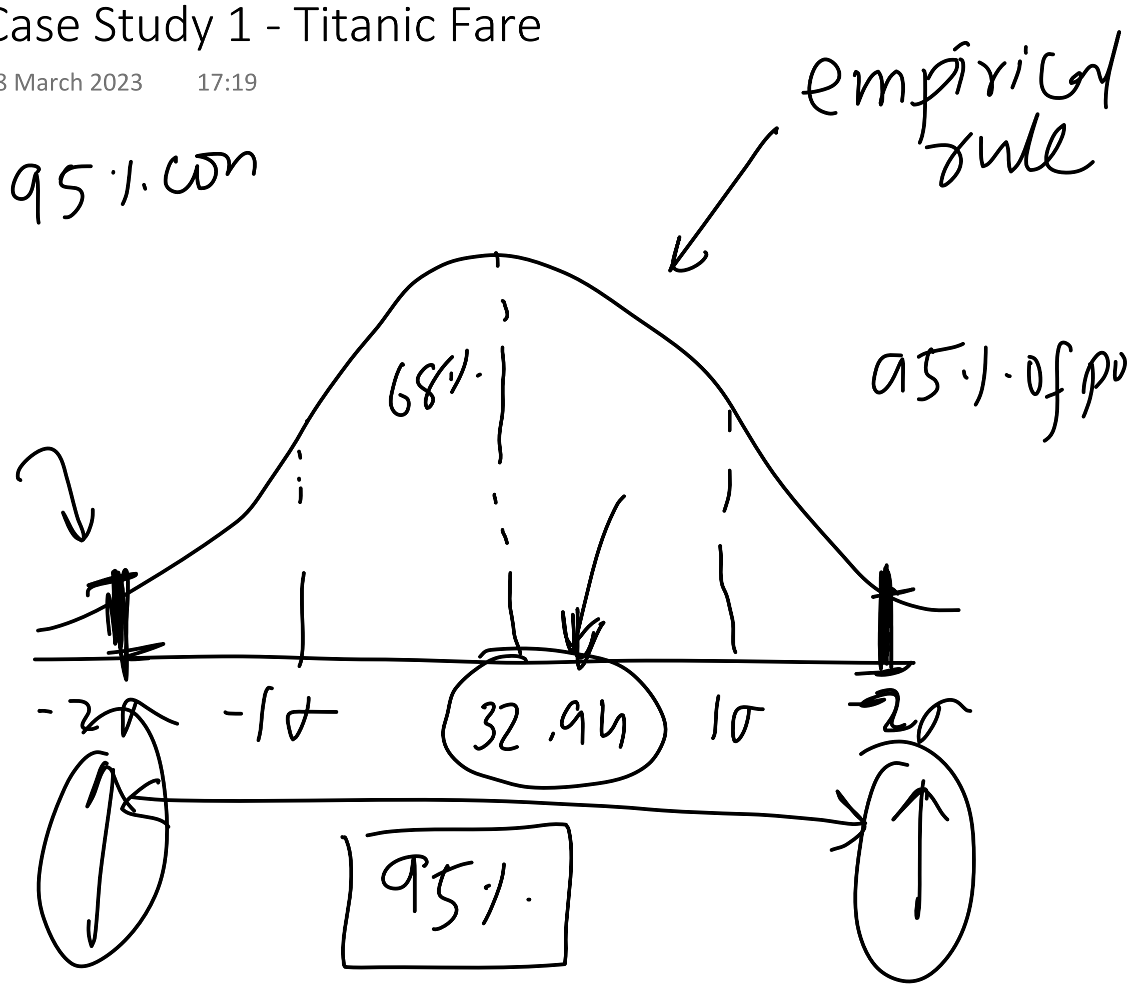
1. The sample size is large enough, typically greater than or equal to 30.
2. The sample is drawn from a finite population or an infinite population with a finite variance.
3. The random variables in the sample are independent and identically distributed.



The CLT is important in statistics and machine learning because it allows us to make probabilistic inferences about a population based on a sample of data. For example, we can use the CLT to construct confidence intervals, perform hypothesis tests, and make predictions about the population mean based on the sample data. The CLT also provides a theoretical justification for many commonly used statistical techniques, such as t-tests, ANOVA, and linear regression.

Case Study 1 - Titanic Fare

28 March 2023 17:19



Explaination - rather than saying exact point , we can use intervals , (roughly you can say confidence interval). so basically we know samplnig distribution will follow the emperical rule , so by this rule we know that approximate 95 percent will fall within 2 std , so we can choose those 2 values for the interval and say that we are approximate 95 percent sure than mean will be in this interval.

Case Study - What is the average income of Indians

28 March 2023 15:49

Step-by-step process:

1. Collect multiple random samples of salaries from a representative group of Indians. Each sample should be large enough (usually, n > 30) to ensure the CLT holds. Make sure the samples are representative and unbiased to avoid skewed results.

2. Calculate the sample mean (average salary) and sample standard deviation for each sample.

3. Calculate the average of the sample means. This value will be your best estimate of the population mean (average salary of all Indians).

4. Calculate the standard error of the sample means, which is the standard deviation of the sample means divided by the square root of the number of samples.

5. Calculate the confidence interval around the average of the sample means to get a range within which the true population mean likely falls. For a 95% confidence interval:

   lower_limit = average_sample_means - 1.96 * standard_error
   upper_limit = average_sample_means + 1.96 * standard_error

6. Report the estimated average salary and the confidence interval.

Python code

Remember that the validity of your results depends on the quality of your data and the representativeness of your samples. To obtain accurate results, it's crucial to ensure that your samples are unbiased and representative.