Parul® University

NAAC A++

Chapter-3: Probability and Types of Testing

Dr. Vinod Patidar

Associate Professor

Department of Computer Science and Engineering

Parul® University

NAAC A++

Content

INDEX

1. Probability

- Probability denotes the possibility of something happening.
- It is a mathematical concept that predicts how likely events are to occur.
- The probability values are expressed between 0 and 1.
- The definition of probability is the degree to which something is likely to occur.

Example-

Tossing a Coin

- When we flip a coin in the air, there are two possible outcomes:

  Heads (H) or

  Tails (T)

- So, the probability of a head to come as a result is 1/2.
- And the probability of a tail to come as a result is 1/2.

Throwing Dice

- When a single die is thrown, there are six possible outcomes:

  1, 2, 3, 4, 5, 6

- The probability of any one of them is 1/6.

Probability Formula

- The probability is the measure of the likelihood of an event to happen.
- It measures the certainty of the event.
- The formula for probability is given by;

$$\mathbf{P(E)} = \frac{\textbf{Number of Favorable Outcomes}}{\textbf{Number of total outcomes}}$$

$$\mathbf{P(E) = \frac{n(E)}{n(S)}}$$

Where,

n(E) = Number of events favorable to event E

n(S) = Total number of outcomes

2. Probability Distributions

- A probability distribution is a statistical function that describes all the possible values and probabilities for a random variable within a given range.
- This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability distribution will be determined by a number of factors.
- The mean (average), standard deviation, skewness, and kurtosis of the distribution are among these factors.
- *Probability Distribution is basically the set of all possible outcomes of any random experiment or event.*

Random Variable

- A Random Variable is a real-valued function whose domain is the sample space of the random experiment.
- It is represented as

  X(sample space) = Real number
- We need to learn the concept of Random Variables because sometimes we are only interested in the probability of the event but also the number of events associated with the random experiment.
- Hence, it is called a random variable and it is generally represented by the letter “X”

Example

- Let us consider an experiment for tossing a coin two times.
- Hence, the sample space for this experiment is

  $S = \{HH, HT, TH, TT\}$

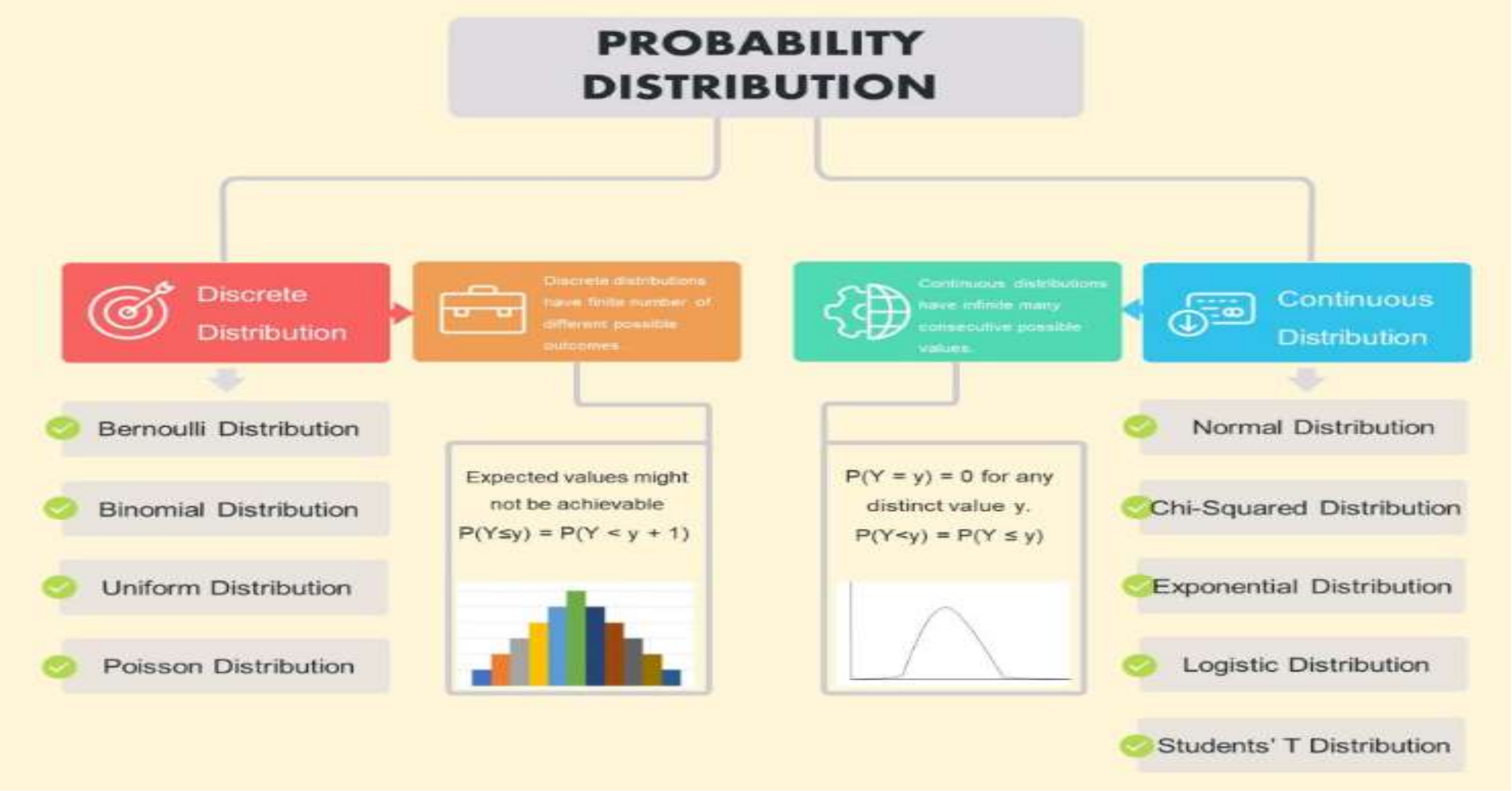
- If X is a random variable and it denotes the number of heads obtained, then the values are represented as follows:

  $X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0$

- Similarly, we can define the number of tails obtained using another variable, say Y.

  (i.e) $Y(HH) = 0, Y(HT) = 1, Y(TH) = 1, Y(TT)= 2$

3. Types of Probability Distribution



PROBABILITY DISTRIBUTION

Discrete Distribution

Discrete distributions have finite number of different possible outcomes

- Bernoulli Distribution
- Binomial Distribution
- Uniform Distribution
- Poisson Distribution

Expected values might not be achievable

$P(Y \le y) = P(Y < y + 1)$

Continuous distributions have infinite many consecutive possible values.

$P(Y = y) = 0$ for any distinct value y.

$P(Y < y) = P(Y \le y)$

Continuous Distribution

- Normal Distribution
- Chi-Squared Distribution
- Exponential Distribution
- Logistic Distribution
- Students' T Distribution

Fig. 3.1: Types of Probability Distribution

Discrete Probability Distribution…

- A discrete distribution describes the probability of occurrence of each value of a discrete random variable.
- Discrete distributions have a finite number of different possible outcomes.

Characteristics of Discrete Distribution

- We can add up individual values to find out the probability of an interval
- Discrete distributions can be expressed with a graph, piece-wise function or table
- In discrete distributions, a graph consists of bars lined up one after the other

Discrete Probability Distribution…

- In the graph, the discrete distributions look like as,

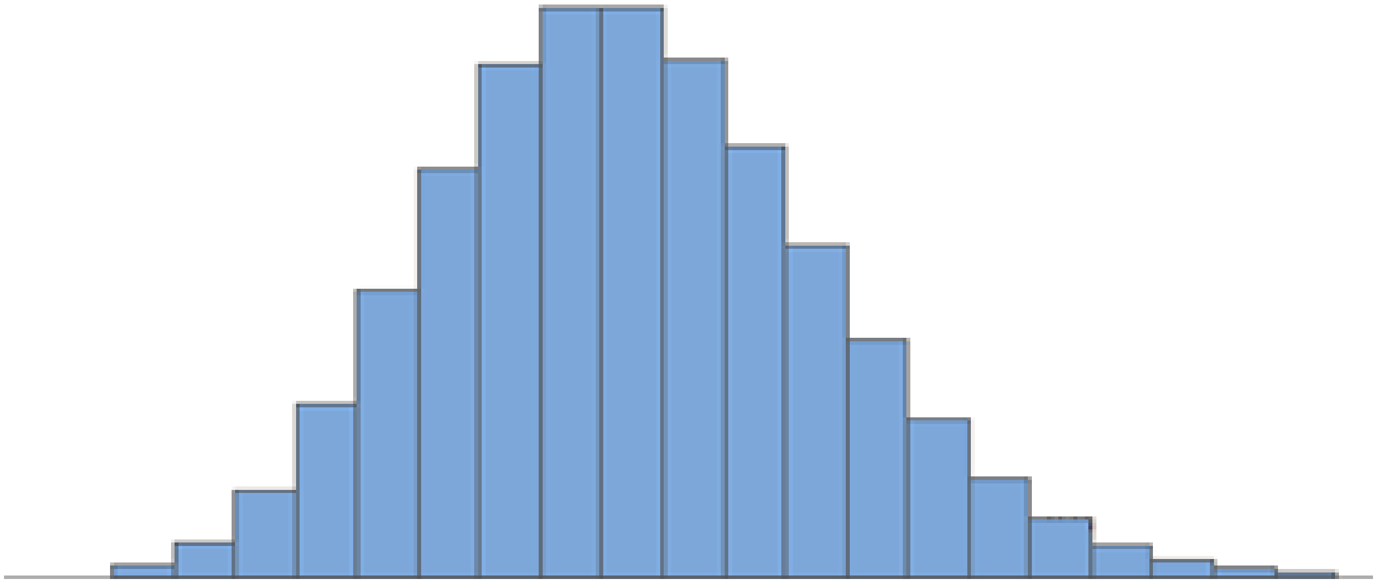


Fig. 3.2 Discrete Distribution

1. Bernoulli's Distribution

- The Bernoulli distribution is a variant of the Binomial distribution in which only one experiment is conducted, resulting in a single observation.
- As a result, the Bernoulli distribution describes events that have exactly two outcomes.
- The Bernoulli random variable's expected value is p, which is also known as the Bernoulli distribution's parameter.
- The experiment's outcome can be a value of 0 or 1. Bernoulli random variables can have values of 0 or 1.

1. Bernoulli's Distribution…

Examples and Uses:

- Guessing a single True/False question.
- It is mostly used when trying to find out what we expect to obtain in a single trial of an experiment.

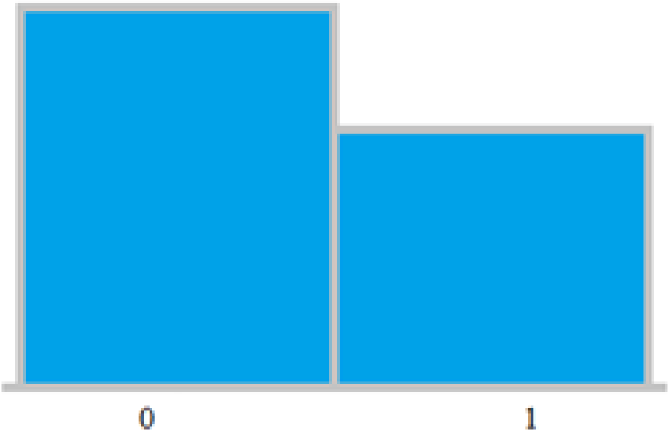


Fig. 3.3 Bernoulli Distribution

2. Binomial Distribution

- The binomial distribution is a discrete distribution with a finite number of possibilities.
- When observing a series of what are known as Bernoulli trials, the binomial distribution emerges.
- A Bernoulli trial is a scientific experiment with only two outcomes: success or failure.

2. Binomial Distribution…

Examples and Uses:

- Simply determine, how many times we obtain a head if we flip a coin 10 times.
- It is mostly used when we try to predict how likelihood an event occurs over a series of trials.

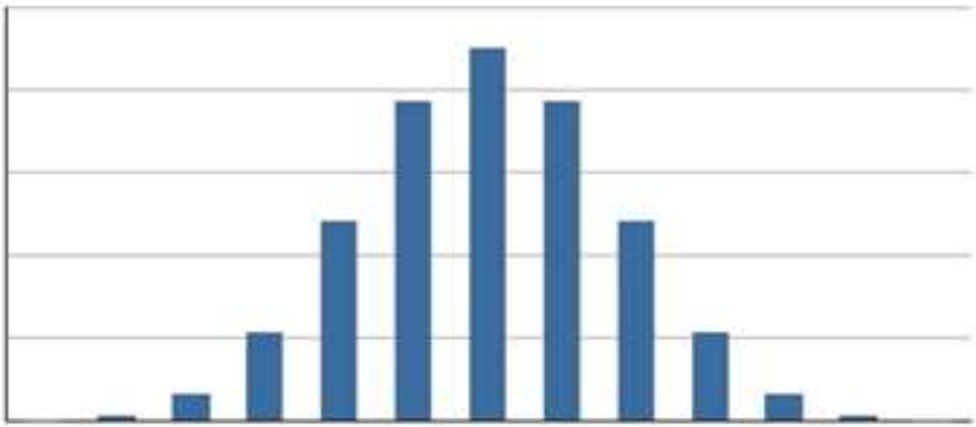


Fig. 3.4 Bernoulli Distribution

3. Uniform Distribution

- In uniform distribution all the outcomes are equally likely.
- In the graph, all the bars are equally tall.
- The expected value and variance have no predictive power.

3. Uniform Distribution…

Examples and Uses:

- Result obtained after rolling a die.
- Due to its equality, it is mostly used in shuffling algorithms

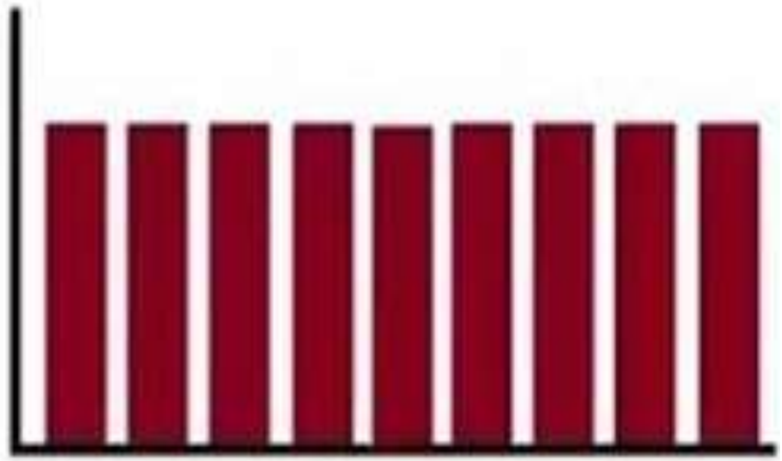


Fig. 3.5 **Uniform** Distribution

4. Poisson Distribution

- A Poisson distribution is a probability distribution used in statistics to show how many times an event is likely to happen over a given period of time.

- To put it another way, it's a count distribution. Poisson distributions are frequently used to comprehend independent events at a constant rate over a given time interval.

- Siméon Denis Poisson, a French mathematician, was the inspiration for the name.

4. Poisson Distribution…

Examples and Uses

- It is used to determine how likelihood a certain event occur over a given interval of time or distance.
- Mostly used in marketing analysis to find out whether more than average visits are out of the ordinary or otherwise.

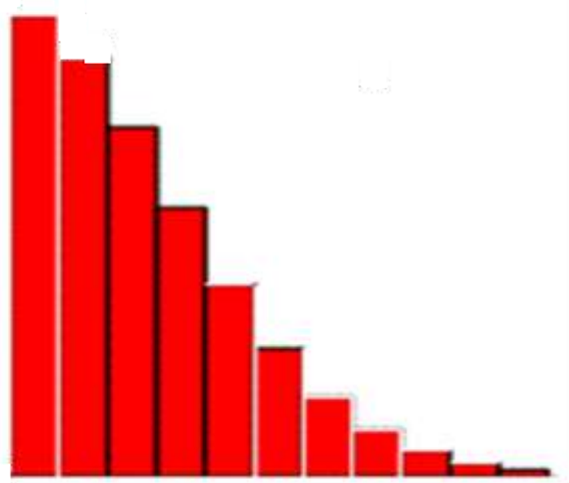


Fig. 3.6 Poisson Distribution

Continuous Probability Distributions

- A continuous distribution describes the probabilities of a continuous random variable's possible values.
- A continuous random variable has an infinite and uncountable set of possible values (known as the range).
- The mapping of time can be considered as an example of the continuous probability distribution.
- It can be from 1 second to 1 billion seconds, and so on.
- The area under the curve of a continuous random variable's PDF is used to calculate its probability.
- As a result, only value ranges can have a non-zero probability.

Continuous Probability Distributions…

- A continuous random variable's probability of equaling some value is always zero.
- In the graph, the continuous distributions look like as,

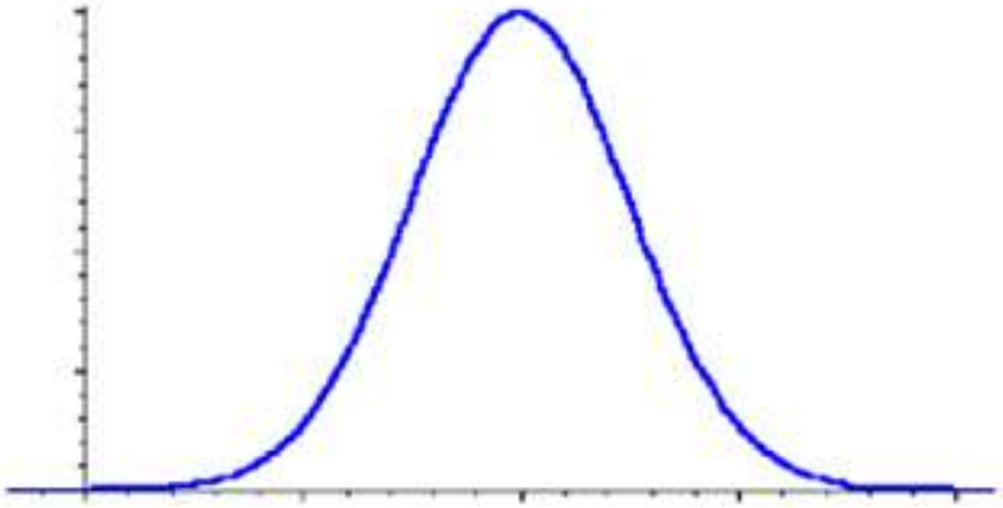


Fig. 3.7 Continuous Probability Distribution

1. Normal Distribution

- Normal Distribution is one of the most basic continuous distribution types. Gaussian distribution is another name for it.
- Around its mean value, this probability distribution is symmetrical.
- It also demonstrates that data close to the mean occurs more frequently than data far from it.
- Here, the mean is 0, and the variance is a finite value.
- In the example, you generated 100 random variables ranging from 1 to 50.

1. Normal Distribution…

Examples and Uses

- Normal distributions are mostly observed in the size of animals in the desert.
- We can convert any normal distribution into a standard normal distribution. Normal distribution could be standardized to use the Z-table.

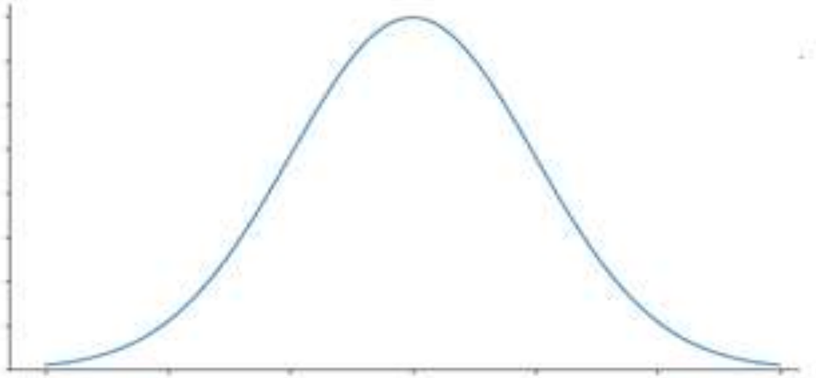


Fig. 3.8 Normal Distribution

2. Chi-Squared Distribution

- Chi-Squared distribution is frequently being used. It is mostly used to test wow (goodness) of fit.
- The graph obtained from Chi-Squared distribution is asymmetric and skewed to the right.
- It is square of the t-distribution.

2. Chi-Squared Distribution…

Examples and Uses:

- It is mostly used to test wow (goodness) of fit.
- It comprises a table of known values for its CDF called the x^2 – table.

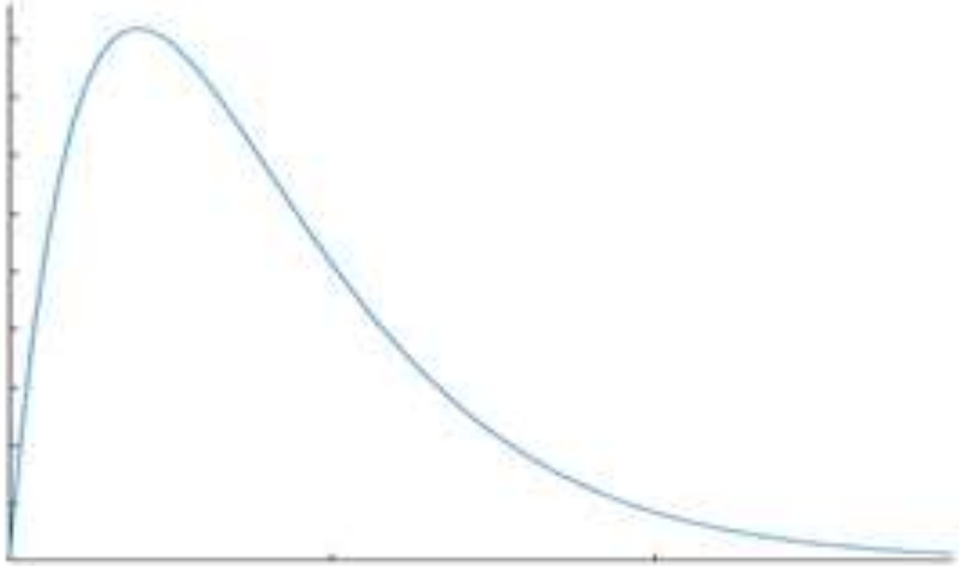


Fig. 3.9 Chi-Squared Distribution

3. Exponential Distribution

- In a Poisson process, an exponential distribution is a continuous probability distribution that describes the time between events (success, failure, arrival, etc.).
- It is usually observed in events that considerably change early on.

3. Exponential Distribution…

Examples and Uses

- It is mostly used with dynamically changing variables, such as online websites traffic.

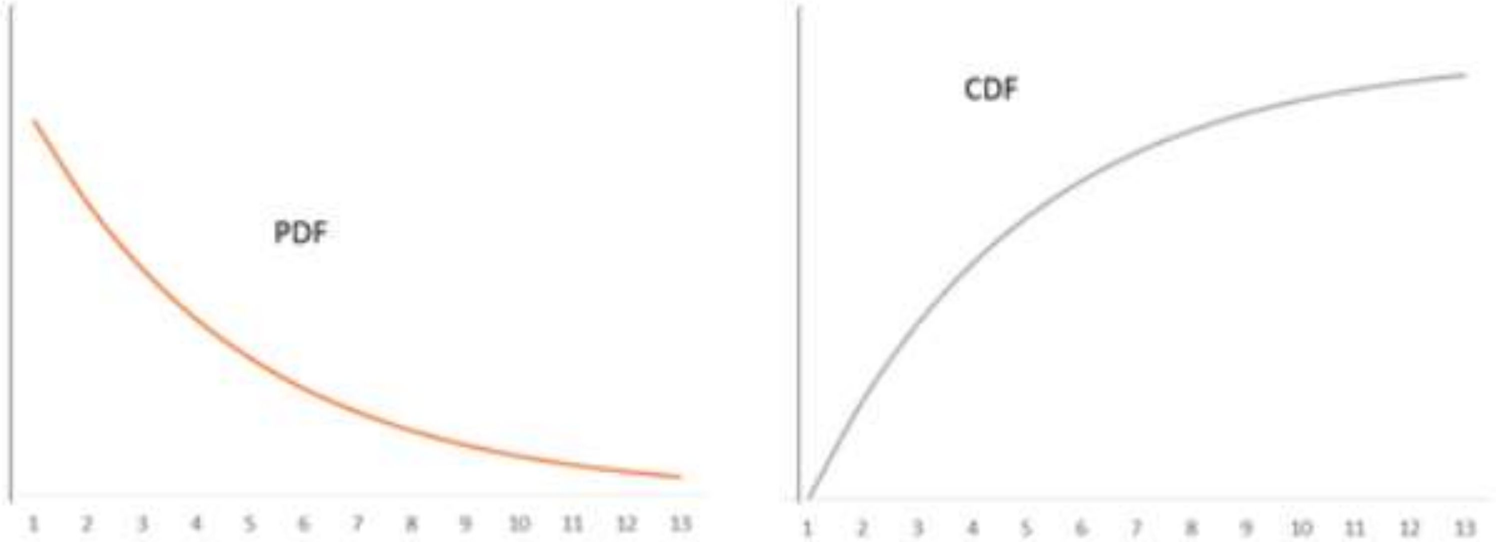


Fig. 3.10 Exponential Distribution

4. Logistic Distribution

- It is used to observe how continuous variable inputs can affect the probability of a binary result.

- The Cumulative Distributed Function picks up when we reach values near the mean.

- The lesser the scale parameter, the faster it reaches values close to 1.

4. Logistic Distribution…

Examples and Uses

- It is mostly used in sports to predict how a player’s or team’s feat can conclude the result of the match.

5. Students’ T Distribution

- Students’ T Distribution or simply called T Distribution is used to estimate population limitation when the sample size is small and population variance is not known.

- A small sample size estimation of a normal distribution.

- Its graph is symmetric and bell-shaped curve, however, it has large tails.

5. Students’ T Distribution…

Examples and Uses

- It is used in examination of a small sample data which usually follows a normal distribution.

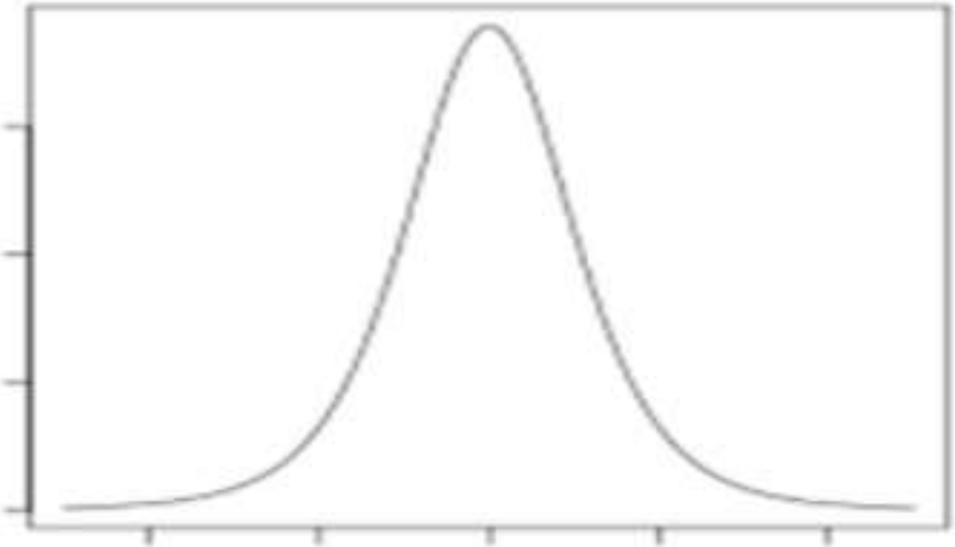
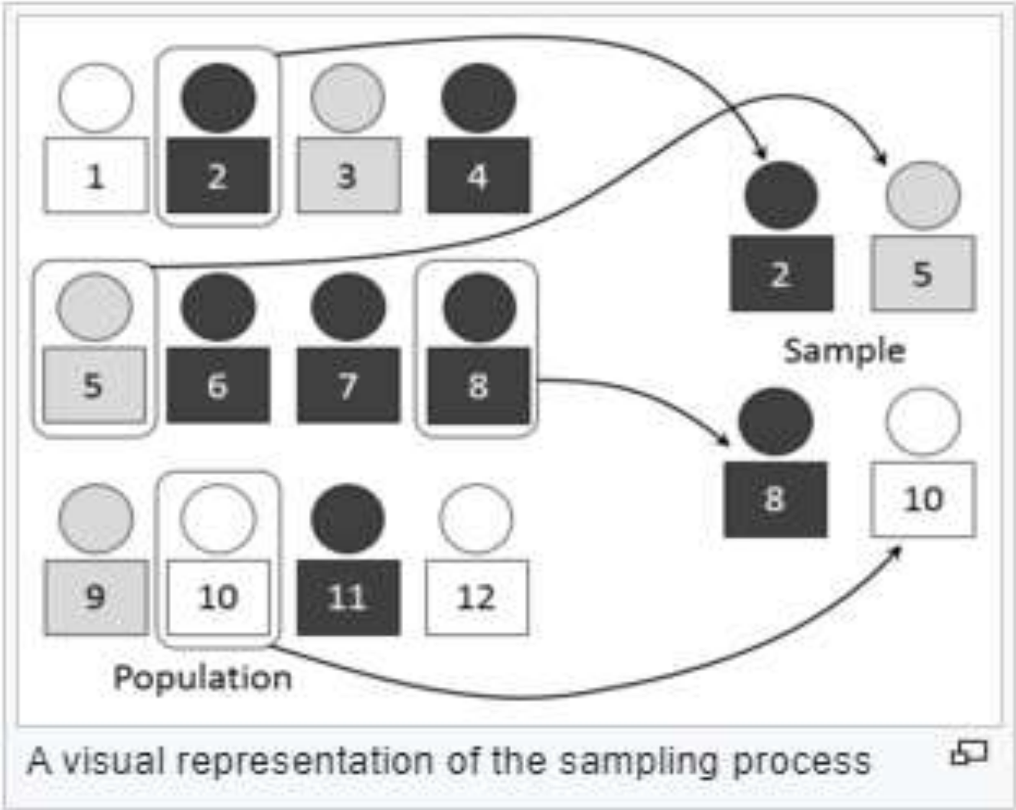


Fig. 3.11 Students’ Distribution

4. Sampling

- Sampling is the selection of a subset or a statistical sample (termed sample for short) of individuals from within a statistical population to estimate characteristics of the whole population.
- The subset is meant to reflect the whole population and statisticians attempt to collect samples that are representative of the population.
- Sampling has lower costs and faster data collection compared to recording data from the entire population, and thus, it can provide insights in cases where it is infeasible to measure an entire population.

Sampling…



A visual representation of the sampling process

Fig. 4.1 Sampling

5. Sampling Distribution

- A sampling distribution is a concept used in statistics.
- It 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.
- This allows entities like governments and businesses to make more well-informed decisions based on the information they gather.
- The sampling distribution depends on multiple factors – the statistics, sample size, sampling process, and the overall population.
- It is used to help calculate statistics such as means, ranges, variances, and standard deviations for the given sample.

How Does it Work?

1. Select a random sample of a specific size from a given population.
2. Calculate a statistic for the sample, such as the mean, median, or standard deviation.
3. Develop a frequency distribution of each sample statistic that you calculated from the step above.
4. Plot the frequency distribution of each sample statistic that you developed from the step above.
5. The resulting graph will be the sampling distribution.

Types of Sampling Distribution

1. Sampling distribution of mean
2. Sampling distribution of proportion
3. T-distribution

Sampling Distribution

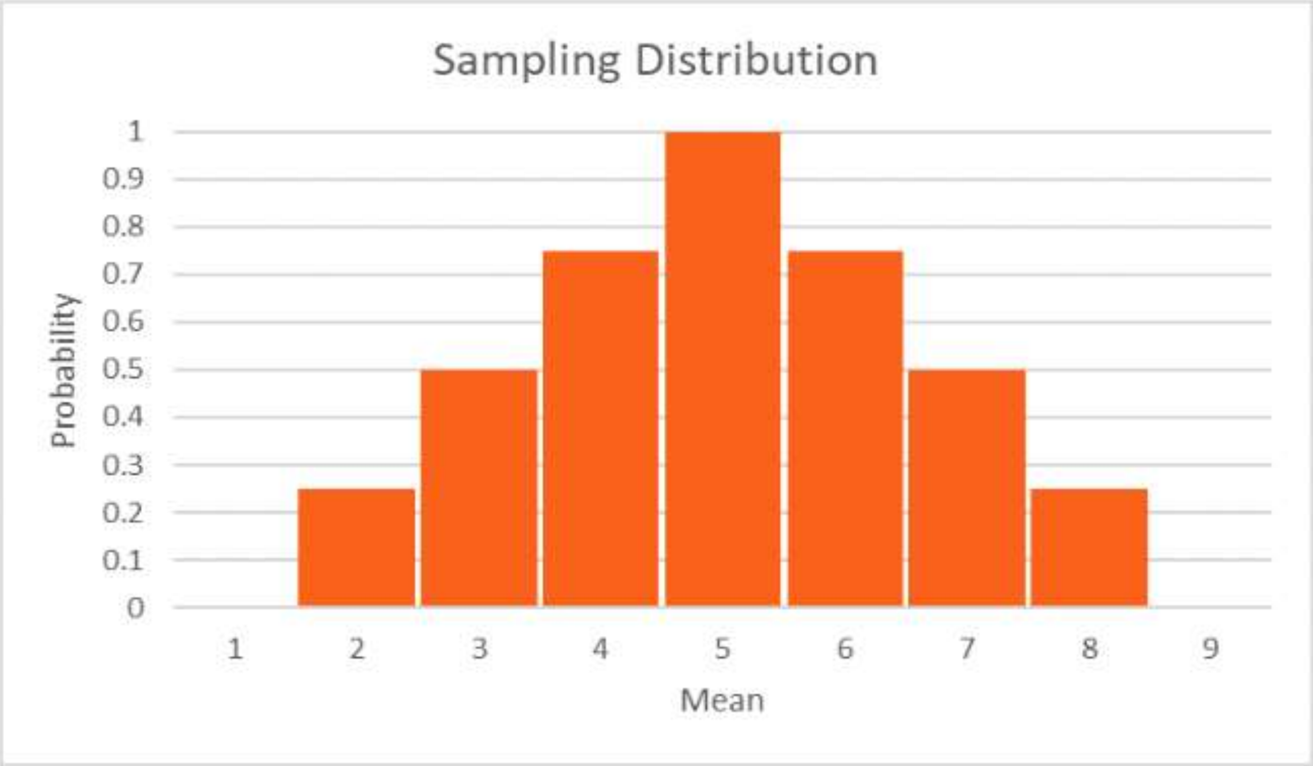


Fig. 5.1 Sampling Distribution

1. Sampling distribution of mean

- As shown from the example above, you can calculate the mean of every sample group chosen from the population and plot out all the data points.
- The graph will show a normal distribution, and the center will be the mean of the sampling distribution, which is the mean of the entire population.

2. Sampling distribution of proportion

- It gives you information about proportions in a population.
- You would select samples from the population and get the sample proportion.
- The mean of all the sample proportions that you calculate from each sample group would become the proportion of the entire population.

3. T-distribution

- T-distribution is used when the sample size is very small or not much is known about the population.

- It is used to estimate the mean of the population, confidence intervals, statistical differences, and linear regression.

6. Hypothesis Testing

- Hypothesis testing involves formulating assumptions about population parameters based on sample statistics and rigorously evaluating these assumptions against empirical evidence.

- Hypothesis testing is a statistical method that is used to make a statistical decision using experimental data.

- Hypothesis testing is basically an assumption that we make about a population parameter. It evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Hypothesis Testing…

- Hypothesis testing is an important procedure in statistics. Hypothesis testing evaluates two mutually exclusive population statements to determine which statement is most supported by sample data

- **Example:** Lets consider an average height in the class is 30 or a boy is taller than a girl. All of these is an assumption that we are assuming, and we need some statistical way to prove these. We need some mathematical conclusion whatever we are assuming is true.

a. Null hypothesis (H0):

- In statistics, the null hypothesis is a general statement or default position that there is no relationship between two measured cases or no relationship among groups. In other words, it is a basic assumption or made based on the problem knowledge.

- **Example:** A company's mean production is 50 units/per day H_0: μ= 50

b. Alternative hypothesis (H1):

- The alternative hypothesis is the hypothesis used in hypothesis testing that is contrary to the null hypothesis.

- **Example:** A company's production is not equal to 50 units/per day H_1: $\mu \neq 50$

Key Terms of Hypothesis Testing

- **Level of significance:** It refers to the degree of significance in which we accept or reject the null hypothesis. 100% accuracy is not possible for accepting a hypothesis, so we, therefore, select a level of significance that is usually 5%. This is normally denoted with α.

- Generally, it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.

Key Terms of Hypothesis Testing…

- **P-value:** The P value, or calculated probability, is the probability of finding the observed/extreme results when the null hypothesis(H_0) of a study-given problem is true.

- If our P-value is less than the chosen significance level then we reject the null hypothesis i.e. accept that our sample claims to support the alternative hypothesis.

Key Terms of Hypothesis Testing…

- **Test Statistic:** The test statistic is a numerical value calculated from sample data during a hypothesis test, used to determine whether to reject the null hypothesis. It is compared to a critical value or p-value to make decisions about the statistical significance of the observed results.

- **Critical value:** The critical value in statistics is a threshold or cutoff point used to determine whether to reject the null hypothesis in a hypothesis test.

Key Terms of Hypothesis Testing…

- **Degrees of freedom:** Degrees of freedom are associated with the variability or freedom one has in estimating a parameter.
- The degrees of freedom are related to the sample size and determine the shape.

Type errors in Hypothesis Testing

- In hypothesis testing, Type I and Type II errors are two possible errors that researchers can make when drawing conclusions about a population based on a sample of data. These errors are associated with the decisions made regarding the null hypothesis and the alternative hypothesis.

- **Type I error:** When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha α

- **Type II errors:** When we accept the null hypothesis, but it is false. Type II errors are denoted by beta β

How does Hypothesis Testing work?

Step 1: Define Null and Alternative Hypothesis

- State the null hypothesis (H_0), representing no effect, and the alternative hypothesis (H_1), suggesting an effect or difference.

- We first identify the problem about which we want to make an assumption keeping in mind that our assumption should be contradictory to one another, assuming Normally distributed data.

How does Hypothesis Testing work?...

Step 2 – Choose significance level

- Select a significance level (α), typically 0.05, to determine the threshold for rejecting the null hypothesis. It provides validity to our hypothesis test, ensuring that we have sufficient data to back up our claims. Usually, we determine our significance level beforehand of the test. The p-value is the criterion used to calculate our significance value.

Step 3 – Collect and Analyze data.

- Gather relevant data through observation or experimentation. Analyze the data using appropriate statistical methods to obtain a test statistic.

How does Hypothesis Testing work?...

Step 4- Calculate Test Statistic

- The data for the tests are evaluated in this step we look for various scores based on the characteristics of data. The choice of the test statistic depends on the type of hypothesis test being conducted.

- There are various hypothesis tests, each appropriate for various goal to calculate our test. This could be a Z-test, Chi-square, T-test, and so on.

How does Hypothesis Testing work?...

- **Z-test:** If population means and standard deviations are known. Z-statistic is commonly used.
- **t-test:** If population standard deviations are unknown and sample size is small than t-test statistic is more appropriate.
- **Chi-square test:** Chi-square test is used for categorical data or for testing independence in contingency tables.
- **F-test:** F-test is often used in analysis of variance (ANOVA) to compare variances or test the equality of means across multiple groups.

7. ANOVA (Analysis of Variance) test

- ANOVA, short for Analysis of Variance, is a statistical method used to see if there are significant differences between the averages of three or more unrelated groups. This technique is especially useful when comparing more than two groups.

- ANOVA works by analyzing the levels of variance within more than two groups through samples taken from each of them.

- An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help us to figure out if we need to reject the null hypothesis or accept the alternate hypothesis.

ANOVA (Analysis of Variance) test…

- Ronald Fisher developed ANOVA in 1918, expanding the capabilities of previous tests by allowing for the comparison of multiple groups at once.
- This method is also referred to as Fisher’s analysis of variance, highlighting its ability to analyze how a categorical variable with multiple levels affects a continuous variable.
- The use of ANOVA depends on the research design. Commonly, ANOVAs are used in three ways: one-way ANOVA, two-way ANOVA and N-way ANOVA.

Example-

- A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. We want to see if one therapy is better than the others.

- A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.

- Students from different colleges take the same exam. We want to see if one college outperforms the other.

How does ANOVA work?

- ANOVA works by analyzing the levels of variance within more than two groups through samples taken from each of them.

- In an ANOVA test we first examine the variance within each group defined by the independent variable – this variance is calculated using the values of the dependent variable within each of these groups. Then, we compare the variance within each group to the overall variance of the group means.

How does ANOVA work?...

- In general terms, a large difference in means combined with small variances within the groups signifies a greater difference between the groups. Here the independent variable significantly varies by dependent variable, and the null hypothesis is rejected.

- On the flip side, a small difference in means combined with large variances in the data suggests less variance between the groups. In this case, the independent variable does not significantly vary by the dependent variable, and the null hypothesis is accepted.

One-way ANOVA

One-way ANOVA is its most simple form – testing differences between three or more groups based on one independent variable.

One-Way ANOVA is a statistical method used when we’re looking at the impact of one single factor on a particular outcome. For instance, if we want to explore how IQ scores vary by country, that’s where One-Way ANOVA comes into play.

For example, comparing the sales performance of different stores in a retail chain.

Two-way ANOVA

One-way ANOVA is used when there are two independent variables.

Two-way ANOVA allows for the evaluation of the individual and joint effects of the variables.

Two-Way ANOVA, also known as factorial ANOVA, allows us to examine the effect of two different factors on an outcome simultaneously.

For example, it could be used to understand the impact of both advertising spend and product placement on sales revenue.

N-way ANOVA

- When researchers have more than two factors to consider, they turn to N-Way ANOVA, where “n” represents the number of independent variables in the analysis.
- **Example:** This could mean examining how IQ scores are influenced by a combination of factors like country, gender, age group, and ethnicity all at once.
- N-Way ANOVA allows for a comprehensive analysis of how these multiple factors interact with each other and their combined effect on the dependent variable, providing a deeper understanding of the dynamics at play.

8. Chi-Square (χ^2) test

- Chi-Square test is a statistical method crucial for analyzing associations in categorical data. Its applications span various fields, aiding researchers in understanding relationships between factors.

- The chi-square test is a statistical test used to determine if there is a significant association between two categorical variables. It is a non-parametric test, meaning it makes no assumptions about the distribution of the data.

Chi-Square (χ^2) test…

- The test is based on the comparison of observed and expected frequencies within a contingency table. The chi-square test helps with feature selection problems by looking at the relationship between the elements.

- It determines if the association between two categorical variables of the sample would reflect their real association in the population.

- It belongs to the family of continuous probability distributions.

Chi-Square (χ^2) test…

- The Chi-Squared distribution is defined as the sum of the squares of the k independent standard random variables given by:

$$x_c{}^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where,

- *c* is degree of freedom
- O_{ij} is the observed frequency in cell *ij*
- E_{ij} is the expected frequency in cell *ij*

Steps to perform Chi-square test

1. **Define-**

Null Hypothesis (H_0): There is no significant association between the two categorical variables.

Alternative Hypothesis (H_1): There is a significant association between the two categorical variables.

2. Create a contingency table that displays the frequency distribution of the two categorical variables.

Steps to perform Chi-square test…

3. Find the Expected values using formula.

$$E_{ij} = \frac{R_i \; X \; C_j}{N}$$

Where,

- R_i : Totals of row i
- C_j : Totals of column j
- N: Total number of Observations

4. Calculate the Chi-Square Statistic

Steps to perform Chi-square test…

5. Degrees of Freedom using formula:

$$df = (m_1)(n_1)$$

Where,

- m corresponds to the number of categories in one categorical variable.
- n corresponds to the number of categories in another categorical variable.

Steps to perform Chi-square test…

6. Accept or Reject the Null Hypothesis: Compare the calculated chi-square statistic to the critical value from the chi-square distribution table for the chosen significance level (e.g., 0.05)

- If χ^2 is greater than the critical value, reject the null hypothesis, indicating a significant association between the variables.

- If χ^2 is less than or equal to the critical value, fail to reject the null hypothesis, suggesting no significant association.

Parul®
University

NAAC
GRADE A++

https://paruluniversity.ac.in/