Probability

Agenda

Key Takeaways-

- What is Probability?
- Recap to probability, key terminologies, properties etc.
- Conditional Probability
- Population vs. Sample
- PDF and CDF
- Why do we need Probability distribution?
- Various types of Probability distribution
- Normal distribution

Probability Recap

Chance

Chance is a possibility of happening something.

Examples :

- Is it possible to have a black colour Sun?

  Not possible at all.
- Is it possible to have February as next month if the current month is January?

  Yes, it is possible.
- Is it possible to cancel today's session?

  May or may not be possible but there are some chances.

So, in case 1 and case 2 there is a surety in possibility of occurrence whereas case 3 is having uncertain possibility of occurrence.

Probability

- Probability is a measure of likelihood of an event to occur.
- It can be used to quantify the chance of occurence of an outcome based on the number of times it has occured.
  For e.g. Assume a coin is tossed 5 times and 3 times it turns into head then we can tell what is the chance that head will occur in the next toss.

Computing Probability -

Suppose a random experiment is repeated n times and the Event A occurs n(A) times then probability of Event A is

$$P(A) = \frac{n(A)}{n} \quad \text{such that } P(A) \geq 0$$

Key Terminologies

- Experiment : Process of observation of an activity
- Outcomes of an Experiment : The results of an observation
- Random Experiment : An experiment whose outcome cannot be predicted.

  Examples -

  Tossing a coin

  Rolling a die

  Drawing a card from the deck

  Getting OTPs
- Sample Space : Set of all possible outcomes of a random experiment

  Further, it can be classified as -

  Discrete Sample Space :

  Number of people attending this session

  Continuous Sample Space :

  Price of a house can be between 1 rupee to 50 cr.

Types of Event

Event is basically a subset of sample space.

E.g. Getting 2 in rolling a die

Or getting an even number in rolling a die

Dependent Events	Independent Events
Occurrence of one event affects the occurrence of another event	Occurrence of one event does not affect the occurrence of another event
E.g. If a larger set of smartphones have some defects then the company's stock values get affected.	E.g. Sales of a particular smartphone in USA and production of Tea in Assam are likely to be independent.

Quiz 1

The CO(Carbon Monoxide) strength in the air for a metro city varies from 80 to 104 PPM.
S = {80.2, 80.9, 81.2, 81.8, 82.6, 84.4, 87.3, 86.1, …………………...103., 103.9}

Identify the type of sample space for the above experiment.

1. Discrete Sample Space
2. Continuous Sample Space
3. None of these two

Counting Techniques

To count the possible number of outcomes in a random experiment, we might need counting techniques.

The most commonly used counting techniques are -

1. Product Rule
2. Sum Rule
3. Combination
4. Permutation

Axioms

For an Event A in a finite sample space S, the probability P(A) that is a real number should follow the following axioms.

1. $P(A) \geq 0$
2. $P(S) = 1$
3. If two events *A* and *B* are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$ where $P(A \cap B) = \oint$

Elementary Properties

Following properties are obtained using the axioms.

1. $P(\bar{A}) = 1 - P(A)$
2. P(∮) = 0
3. $P(A) < P(B)$ if $A \subset B$
4. $P(A) \leq 1$
5. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Computing Probability

Example 1 - Consider a box having 5 blue balls, 3 green balls and 2 red balls. If one ball is drawn randomly then what is the probability that it will be a green ball?

Prob. of drawing a green ball $= \dfrac{(Total\ number\ of\ ways\ a\ green\ ball\ can\ be\ drawn)}{Total\ possible\ outcomes} = \dfrac{3}{10} = 0.3$

Example 2 - In rolling a fair die, what is the probability of getting an even number?

Prob. of getting an even no. $= \dfrac{(Total\ number\ of\ times\ an\ even\ no.\ occurs)}{Total\ possible\ outcomes} = \dfrac{3}{6} = 0.5$

Conditional Probability

Conditional probability is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) has already occurred.

Example - A die is rolled. If the the outcome is an even number, then what is the probability that the number is 2?

Here, Event B = Outcome is an even number = {2,4,6}
Event A = {2}

$$P(\textit{Outcome being } 2 \mid \textit{given that outcome is even})$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\textit{No. of elements in } A \cap B}{\textit{No. of elements in } B} = \frac{1}{3}$$

Population vs Sample

Assume University X has 200k enrolled students in all its affiliated colleges across various programmes. On its golden jubilee, the university wants to gift t-shirts to all its students.

So they need to order t-shirts of different sizes like S, M, L, XL etc. But they don't know how many quantities of each size should be ordered?

Approach 1 - Collect data of all 200k students.

But cost of such data collection is very high and time consuming.

Here, assume they know(by any means) that people with height ≥ 180 cm, tends to wear XL size t-shirts. Similarly, with height [165 cm, 180 cm] tends to wear L size and so on.

Approach 2 - Collect height of 1000 random students and estimate their mean and standard deviation and find

P(height ≥ 180 cm) = 7.2%

P(165 cm ≥ height ≤ 180 cm) = 33.4%

Population vs Sample

- Population - When the complete data required for some observation/analysis is collected and experimented.

Downside - The complete data collection is time consuming and costly process.

- Sample - When the limited data is collected/analyzed.

PDF and CDF

For a continuous random variable X, we have

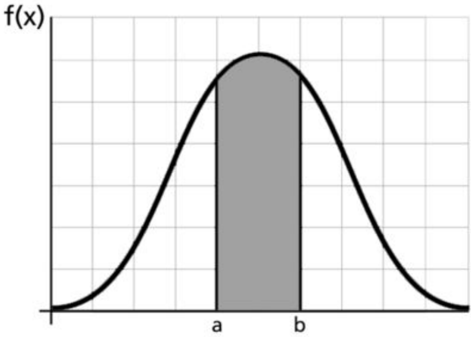
- Probability Density Function (PDF)
- Cumulative Density Function (CDF)

Probability Density Function (PDF)

If X is a continuous random variable, then the *pdf* of X is a function, f(x), such that for any two numbers, a and b with a≤b

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

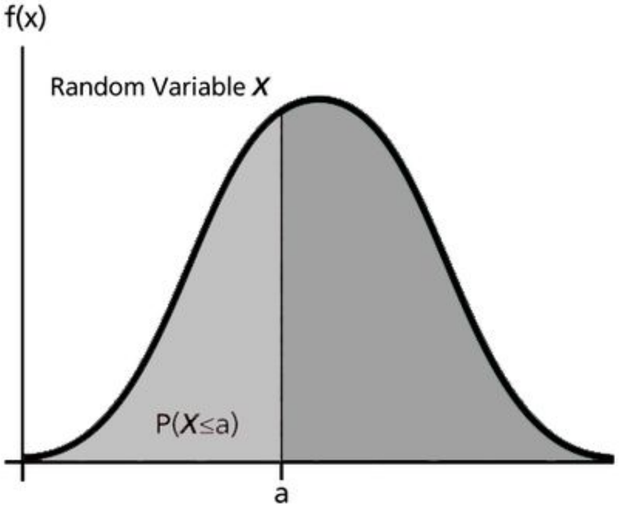
That is, the probability that X takes on a value in the interval [a,b] is the area under the density function from a to b.



Cumulative Density Function (CDF)

The *cdf* is a function, F(x), of a random variable X, and is defined for a number x by:

$$F(x) = P(X \le x) = \int_{0}^{x} f(s)\, ds$$



- The cdf represents the cumulative values of the pdf, whereas PDF is the derivative of CDF.

Probability

Day 2

Agenda

Key Takeaways-

- Continue with Normal distribution
- Normality Check
- Chebyshev inequality
- Log Normal distribution
- Uniform distribution
- Student’s t distribution
- Chi-Square distribution

Quiz 2

In normal distribution, the height of the curve decreases and the spread increases for a -

A. Larger value of mean

B. Smaller value of variance

C. Larger value of variance

D. Smaller value of mean

Hypothesis Testing

In statistics, Hypothesis is an assumption/statement/claim about the population, about the distribution of data or whether one set of results are different from another set of results.

Hypothesis testing is the process of verifying the claim/assumption by collecting enough evidences about the claim/assumption. As a result of hypothesis testing either we accept or reject the claim/assumption made.

Few examples where we need to determine -

- If the coin is biased towards the head or not?
- Whether the data follows normal distribution or not?
- Whether two samples are drawn from the same underlying distribution or not?
- Is there any difference in heights of class 1 and 2 students?

Hypothesis Testing

There are two competing hypotheses in hypothesis testing.

- Null Hypothesis (H_0) - It is often called the default assumption. By default it is assumed to be true.
- Alternate Hypothesis (H_1) - It is contradictory/opposite to the null hypothesis. If there are enough evidences to reject the null hypothesis, then the alternate hypothesis is accepted.

To accept/reject the null hypothesis, we calculate the probability of finding the observed data (test statistic) given that the null hypothesis is true.

So if the p-value is -

Significantly low (< the significance level)	Reject the Null Hypothesis
Significantly high(≥ the significance level)	Fail to reject the Null Hypothesis so accept the Alternate Hypothesis

Steps involved in Hypothesis Testing

1. Define the null hypothesis, alternate hypothesis and the level of significance.
2. Calculate the p-value (probability of finding the observed data/test statistic) assuming the null hypothesis to be true.
3. Conclude whether to accept or reject the null hypothesis based on the p-value
   - If p-value < significance level, then reject the null hypothesis
   - If p-value ≥ significance level, then accept the null hypothesis

4. State the conclusion.

Examples - Hypothesis Testing

Determine if the coin is biased towards head or not based on the observational data from the experiments.

Case 1 - Got 5 heads in flipping the coin 5 times.

Null Hypothesis - Coin is not biased towards the head.
Alternate Hypothesis - Coin is biased towards the head.

p(obs. | H_0) = ½ * ½ * ½ * ½ *½ = 1/32 = 0.03%

Since p-value < 0.05 so reject the null hypothesis and accept the alternate one.

Examples - Hypothesis Testing

Determine if the coin is biased towards head or not based on the observational data from the experiments.

Case 2 - Got 3 heads in flipping the coin 3 times.

Null Hypothesis - Coin is not biased towards the head.
Alternate Hypothesis - Coin is biased towards the head.

p(obs. | H_0) = ½ * ½ * ½ = 1/8 = 0.125%

Since p-value ≥ 0.05 so accept the null hypothesis.

Quiz 3

Choose the correct statement(s).

A. There are fixed number of trails in binomial distribution.

B. As we move far from the mean in normal distribution, the probability decreases.

C. QQ plot can be used only for normality test.

D. All of the above.

Student’s t distribution

Helpful in estimating population mean using a sample when the sample size is small and population standard deviation is not known.

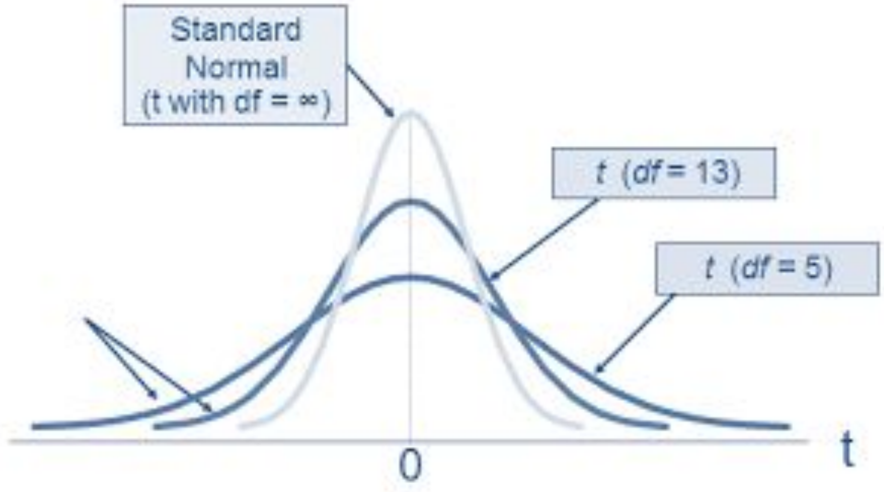
“***Student's t-distribution*** *(or simply the* ***t-distribution****) is any member of a family of continuous probability distributions that arise when estimating the mean of a normally-distributed population in situations where the sample size is small and the population's standard deviation is unknown.”* - Wikipedia

Consider t distribution in the cases when -

- Sample size is small (30 or less)
- Population standard deviation is not known
- The population is approximately normally distributed

Normal vs t-distribution

- Both are symmetric in nature.
- t-distribution is lower at the mean and higher at the tails compared to normal distribution.
- As the sample size increases, the flatness of t-distribution decreases and it approximates



Copyright ©. All rights reserved.

Degree of freedom

Degree of freedom indicates the number of independent observations in the data.

Example -

Assume two sample values x & y and their mean is 10. Since (x+y)/2 = 10 so x+y = 20

Now if x takes a value 6 then y is no longer free to take any value except 14. So here in a sample of two, only one is free to take any value i.e. number of independent observations are 1.

Similarly, for 3 sample values x, y & z if mean = 10 then (x+y+z) = 30 Now if x takes a value 10 and y takes 5 then z needs to be equal to 15. So only two are free to take any value.

Therefore, degree of freedom = (Sample size) - 1

Properties of t-distribution

- The mean of the distribution = 0
- Its probability distribution is a bell-shaped curve.
- The variance of the distribution = r/(r-2), where r = degree of freedom
- As the degree of freedom increases, the t-distribution starts converging to normal distribution.

Chi-Square distribution

If a random variable X is normally distributed with mean μ and standard deviation σ, then

$$V = (\frac{X-\mu}{\sigma})^2 = \chi^2$$

Where, *V* is a chi-square distributed random variable with one degree of freedom.

*“The **chi-square distribution** (also **chi-squared** or **χ^2-distribution**) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. “* - Wikipedia

If Z_1, ..., Z_k are independent, s $$Q = \sum_{i=1}^{k} Z_i^2$$ andom variables, then the sum of their squares,

Properties of Chi-Square distribution

- The mean of the distribution = degrees of freedom (k).
- Its variance of the distribution is twice the degrees of freedom, σ2 = 2k
- It is positively skewed and as the degree of freedom increases the skewness decreases.
- As the degree of freedom increases, the chi-square distribution starts converging to normal distribution.