greatlearning
Power Ahead

Inferential Statistics

Agenda

- Probability
- Random Variable
- Distributions around us
- Binomial Distribution
- Normal Distribution
- Uniform Distribution
- Central Limit Theorem
- Estimation

Pop Quiz

1. What do you understand by binomial distribution?
2. What are the assumptions of binomial?
3. What is the importance of normal distribution?
4. What is the importance of uniform distribution?
5. Why do we need sampling?
6. What makes the central limit theorem so important?
7. What do you mean by interval estimation ?

Probability

- Probability refers to chance or likelihood of a particular event taking place.
- An event is an outcome of an experiment.
- An experiment is a process that is performed to understand and observe possible outcomes.
- Set of all outcomes of an experiment is called the sample space.

Probability : Example

A company is conducting a telephone survey of randomly selected individuals to get their overall impressions of the year 2020. So far, 1646 people have been surveyed. The frequency distribution shows the results. What is the probability that the next person surveyed has a positive overall impression of 2020?

Response	Number of times
Positive	201
Negative	903
Neither	235
Don't know	307
Sum	1646

P(response is positive) = #Positive Responses / #Total Responses

P(response is positive) = 201/1646 ≈ 0.12

Random Variable

Suppose there are 1000 students in the university. What is the probability that 500 students will pass the upcoming exam?

here is a 50-50 chance that each student will pass or fail

The total number of students who pass can range from 0 to 1000

A random variable assigns a numerical value to each outcome of an experiment. It assumes different values with different probability.

There are two types of Random variables:

1) **Discrete random variable:** It can take finite number of values. For example: Number of employees getting promoted in an organization

2) **Continuous random variable:** It can take infinite number of values in a given range. For example: Speed of an aircraft.

Distributions around us (commonly occuring)

Bernoulli	The outcome of tossing a fair coin
Binomial	The number of non-defective products in a production run.
Uniform	The number of books sold weekly at a bookstore
Normal	IQ distribution of all the seven years old children in the New York

Binomial Distribution

The binomial distribution is the probability distribution of a success or failure outcome of an experiment that is conducted multiple times. **Ex.** Probability of getting a head after tossing a coin 10 times

The assumptions of Binomial distributions are as follows:

1. There are only two possible outcomes (success or failure) for each trial.
2. The number of observations or trials is fixed.
3. Each observation or trial is independent. In other words, none of your trials have an effect on the probability of the next trial.
4. The probability of success (tail or head) is exactly the same from one trial to another.

Uniform Distribution

In a deck of cards, an individual can draw a spade, a heart, a club, or a diamond. So, there is an equal probability of selecting any of the card and all the outcomes are mutually exclusive in nature.

The last digit of the social security number of the goalkeeper for the US World Cup soccer team is about equally likely to be any digit from 0 to 9 and all the numbers are mutually exclusive .

We can say that the probabilities of occurrence are uniformly distributed.

This is what we called as uniform distribution.

Uniform Distribution

There are two types of Uniform Distribution

Discrete Uniform Distribution: Can take a finite number (*m*) of values and each value has equal probability of selection.

For example: Number of refrigerator sold by at electronic shop per day can be uniformly distributed between 30 to 50.

Continuous Uniform Distribution: Can take any value between a specified range.

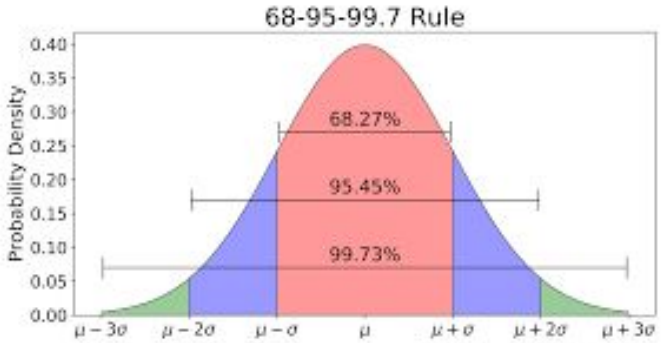
For example: Time taken by a student to finish the assignment can be uniformly distributed between 3 to 5 hours.

Normal Distribution

The normal distribution is the probability distribution that is symmetric about the mean. It is also known as bell curve.

Properties:

- In a normal distribution, mean is zero and standard dev
- It has a zero skewness
- Mean = Median = Mode



Normal distribution

Cumulative distribution function (cdf): It is the area under the curve for the given value.

Ex. What is the chance that a man is less than 165 cm tall?

Percent point function (ppf): It is the inverse of the cdf value.

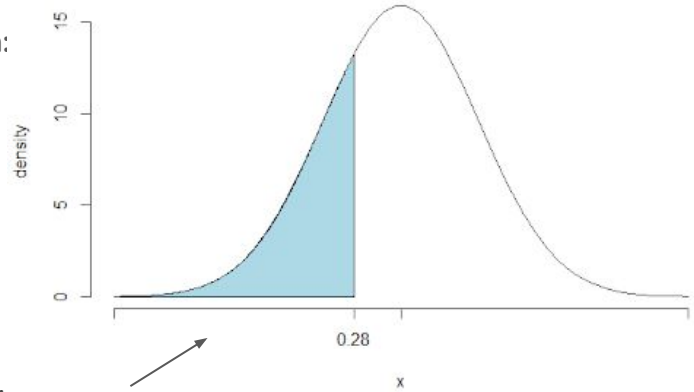
Ex. Given that I am looking for a man who is smaller than 95% of all other men, what size does the subject have to be?

Normal Distribution in Python

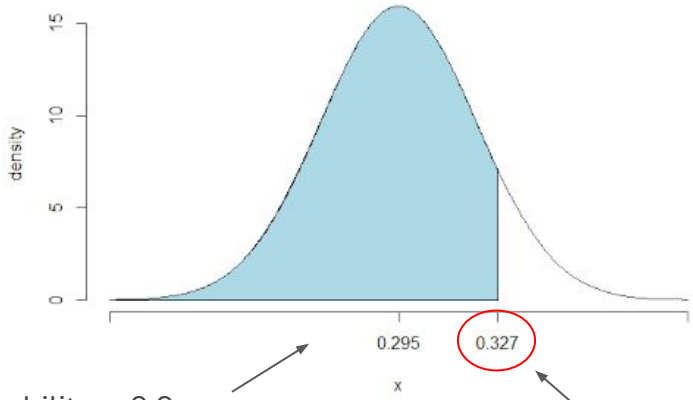
The mean weight of a morning breakfast cereal pack is 0.295 kg with a standard deviation of 0.025 kg. The random variable - weight of the pack follows a normal distribution.

1. What is the probability that the pack weighs less than 0.280 kg? - cdf?
2. Given that I am looking for a pack which weighs higher than at least 90% of the other packs, what is the weight value? - ppf?

Solution:



Probability = 0.274 i.e. 27.4%

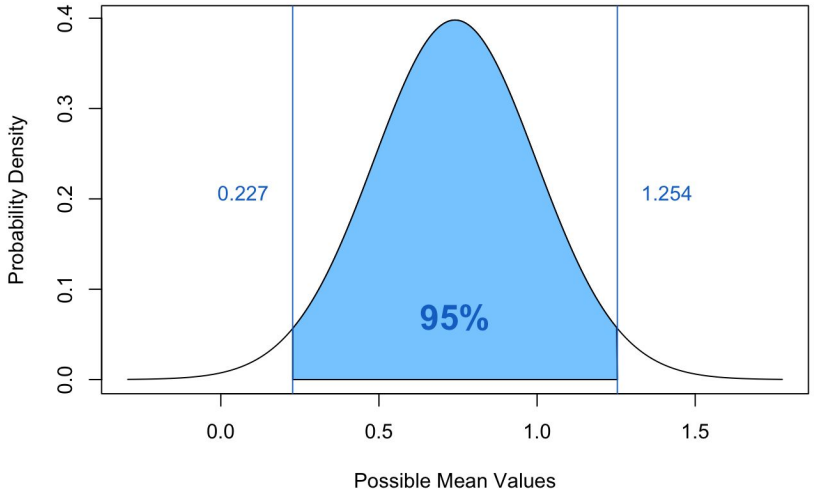


Probability = 0.9 i.e. 90%

Weight value

Sampling Distributions

- **Need for sampling**
  - Given the limited resources and time, it is not always possible to study the population. That's why we choose a sample out of the population to make inference about the population
- **Sampling Distributions**
  - It is a distribution of a particular sample statistic obtained from all possible samples drawn from a specific population



Central Limit Theorem

The sampling distribution of the sample means will approach normal distribution as the sample size gets bigger, no matter what the shape of the population distribution is.

Assumptions

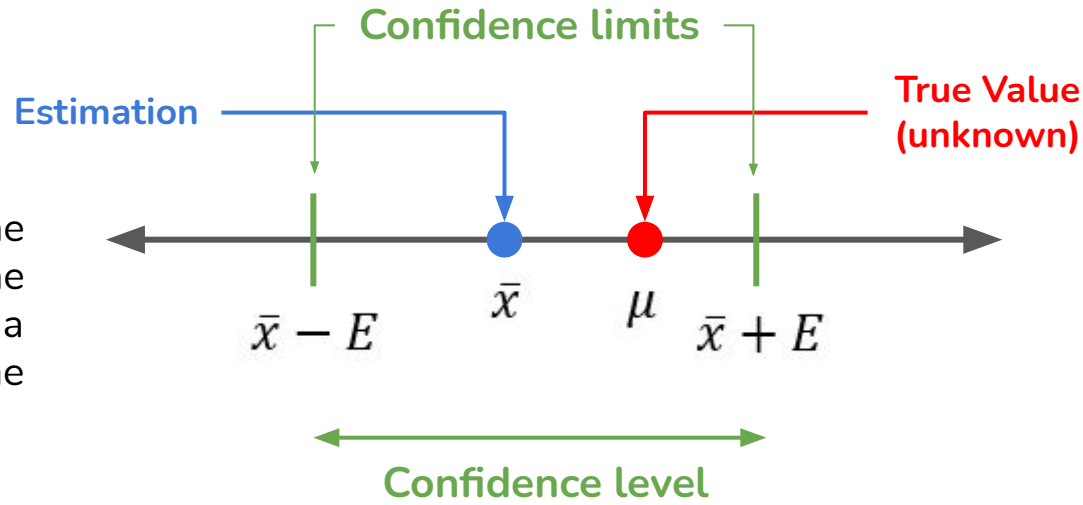
- Data must be **randomly sampled**
- Sample values must be **independent** of each other
- Samples should come from the **same distribution**
- Sample size must be **sufficiently large (≥30)**

Let's see CLT in action by simulation - Link to external site

Confidence interval

Confidence interval provides an interval, or a range of values, which is expected to cover the true unknown parameter.

The upper and lower limits of the interval are determined using the distribution of the sample mean and a multiplier which specifies the 'confidence'



greatlearning
Power Ahead

Happy Learning !