Probability & Probability Distribution

GROUND RULES

- Come prepared for these sessions by watching the video lectures.
  - -Concepts will be covered in the videos.
  - -Hands-On Application will be covered in Mentor Sessions.

- Submit all assignments on time.

- Let’s be punctual & respect each other’s time.

Great Learning

LEARNING OBJECTIVE OF THIS MODULE

- **Probability & Probability Distribution**
- Hypothesis Testing
- ANOVA

LEARNING OBJECTIVES OF THIS SESSION - APPLICATION OF INFERENTIAL STATISTICS

- Probability
- Bayes' Theorem
- Normal Distributions

TRY ANSWERING THE FOLLOWING

- If a dice is thrown twice, what is the probability of getting 6 in both throws

- How to calculate mean and standard deviation

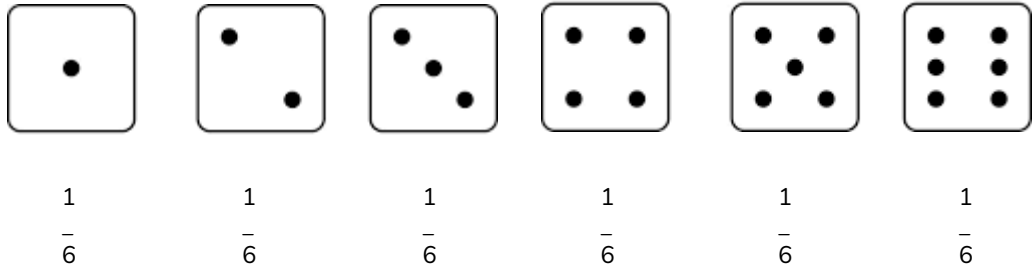
- In case of Normal Distribution what percentage of data will lie in the range of $\mu \pm 2\sigma$?

Probability

Probability refers to chance or likelihood of a particular event taking place. It ranges between 0 to 1

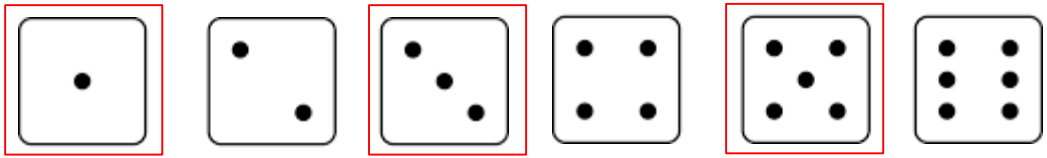
Example: If a die is rolled, what is the probability of getting an odd number?

Possible Outcomes



$\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$

Getting an odd number



$\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$

$$\frac{3}{6} = \frac{1}{2}$$

Mutually Exclusive and Independent Events

Mutually exclusive events

Two events A and B are said to be mutually exclusive if the occurrence of A precludes the occurrence of B. The probability of both the events occurring together is zero.

$$P(A \text{ and } B) = 0$$

Examples:

1. Tossing a coin is a mutually exclusive event because either you will get a head or a tail. You can never get head and tail simultaneously while tossing a coin.
2. While turning to right or left, either you will turn left or right not both.

Independent Events

Two events A and B are said to be independent if the occurrence of A is in no way influenced by the occurrence of B. Likewise occurrence of B is in no way influenced by the occurrence of A.

Example: Getting a head by tossing a coin and getting 5 by rolling a die

Joint, Marginal and Conditional probability

A survey was conducted with 1000 people in New York to determine people's favourite sports. The options were Cricket, Football and others. The following table contains the response gathered:

	Male	Female	Total
Cricket	240	150	390
Football	200	50	250
Others	100	260	260
Total	540	460	1000

Probability distribution

	Male	Female	Total
Cricket	0.24	0.15	0.39
Football	0.2	0.05	0.25
Others	0.1	0.26	0.26
Total	0.54	0.46	1

Joint Probability: It is the probability of two events occurring together at the same time. **Ex.** What is the probability of someone being female and liking football?

P (female and football) = 0.05

Joint, Marginal and Conditional probability

Marginal Probability: It is the probability of an event irrespective of the outcome of another variable. **Ex.** Probability of being a male:

$$P(male) = 0.54$$

which completely ignores the sport.

Conditional Probability: It is the probability of an event occurring given that another event has occurred. **Ex.** What is the probability that a person would like to play cricket given that the person is male.

$$P\ (Cricket \mid Male) = P\ (Cricket, Male) / P\ (Male)$$

$$= 0.24 / 0.54$$

$$= 0.44$$

Bayes Theorem

Example: You are a data scientist in a company and you have to asses that - of the Span Detector created by your team. If the spam detector puts a mail it in the spam folder what is the probability that it was actually a spam? The following details are given to you:

- 3% of the mail you receive is spam
- When a mail is spam, the spam detector detects it with the 99% accuracy
- 0.2% of the time when the mail is not spam, it will mark it as spam

Solution: According to Bayes' theorem

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Or $P(Spam|Detected) = P(Detected|Spam) * P(Spam) / P(Detected)$

Or $P(Spam|Detected) = 0.99 * 0.03 / P(Detected)$

We don't know P(B), that is P(Detected), but we can calculate it using:

$$P(B) = P(B|A) * P(A) + P(B|not\ A) * P(not\ A)$$

Or $P(Detected) = P(Detected|Spam) * P(Spam) + P(Detected|not\ Spam) * P(not\ Spam)$

Bayes Theorem

We know P(Detected|not Spam), which is 0.2 percent and we can calculate P(not Spam) as 1 – P(Spam); for example:

- P(not Spam) = 1 – P(Spam)
- P(not Spam) = 1 – 0.03 = 0.97

Therefore, we can calculate P(Detected) as:

- P(Detected) = 0.99 * 0.03 + 0.002* 0.97
- P(Detected) = 0.03164

That is, about 3 percent of all emails are detected as spam, regardless of whether they are spam or not.

Now we can calculate the answer as:

- P(Spam|Detected) = 0.99 * 0.03 / 0.03164
- P(Spam|Detected) = 0.939

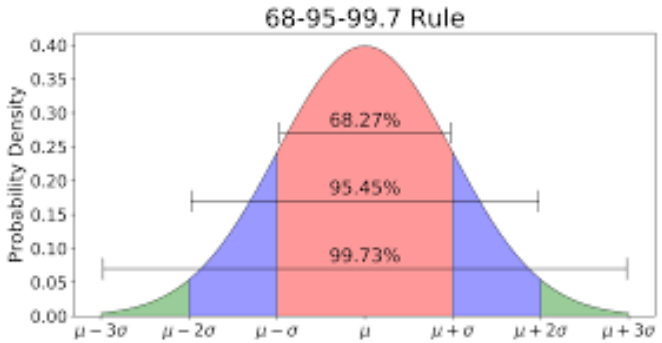
That is, if an email is in the spam folder, there is a **93.9** percent probability that it is, in fact, spam.

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 deviation is 1
- It has a zero skewness
- Mean = Median = Mode



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.

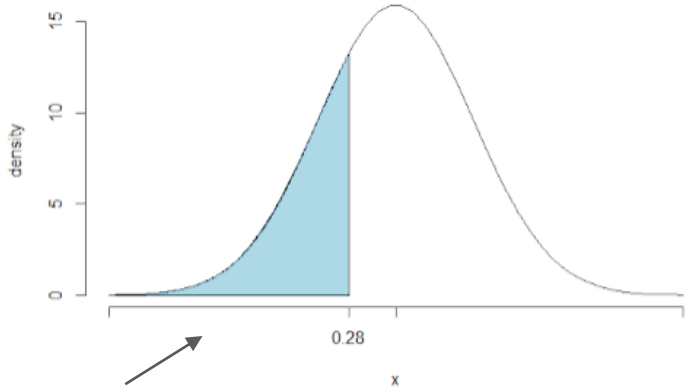
Ex. Given that I am looking for a man who is smaller than 95% of all other men, what size does the subject have to be?

Normal Distribution in Python

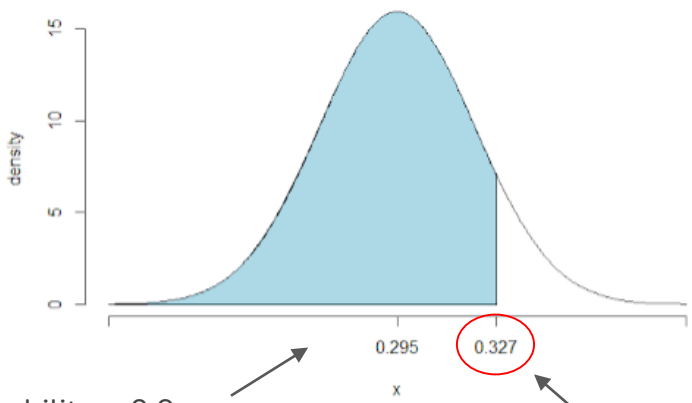
The mean weight of a morning breakfast cereal pack is 0.295 kg with a standard deviation of 0.025 kg. The random variable - weight of the pack follows a normal distribution.

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

Real Life Example - Female Shoe Sales

This data present on the right shows the distribution of Female Shoe Sales in USA in 1998.

It can be used by footwear companies to produce the footwears in similar proportions and hence minimize inventory and maximize profits.

https://thoughtburner.org/tag/normal-distribution/

BY THE ALUMs

I used to devote an hour a day from my busy schedule to watch videos and practice Hand-on almost everyday. This helped me make the best use of the mentored learning session because I could clarify my doubts after understanding the core concepts from the videos itself.

P. Venkatesh Iyer

Great Learning

ANY QUESTIONS

Excelerate CareerPrep:
Career services by Great Learning.

The CareerPrep is an organized way to learn about all of the topics asked in interviews for Data Science jobs.

- Complete Excelerate career prep and learn how to crack an interview with ease.
- Understand the different kind of roles available in the data science ecosystem along with their skill set requirements
- Build a resume that has a higher probability of getting shortlisted by recruiters
- Prepare for different data science roles by reviewing 400+ questions asked in actual interviews
- Understand and implement frameworks to solve guesstimates, logical and aptitude problems often used in interviews
- Learn from the experience of alumni who have successfully transitioned
- Speed up your path towards a transition to a Data science role.

Complete Career Prep today to speed up your path towards DS transition. You'll have a 40% better chance of getting hired than other candidates.

Note: CareerPrep is mandatory step in unlocking the job board.

Great Learning

HAPPY LEARNING