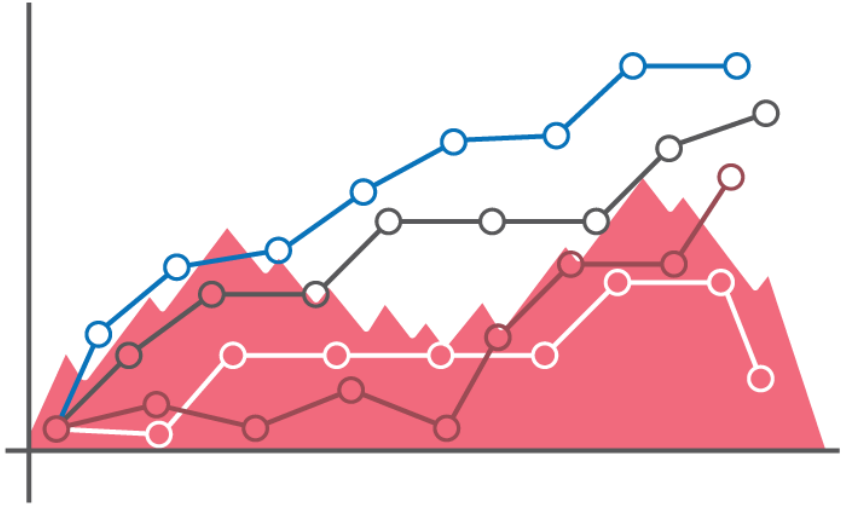
edureka!



MODULE-1
INTRODUCTION TO STATISTICS AND BASIC PROBABILITY

Objectives

edureka!

At the end of this module, you will be able to:

→ Understand the various terminologies of probability,

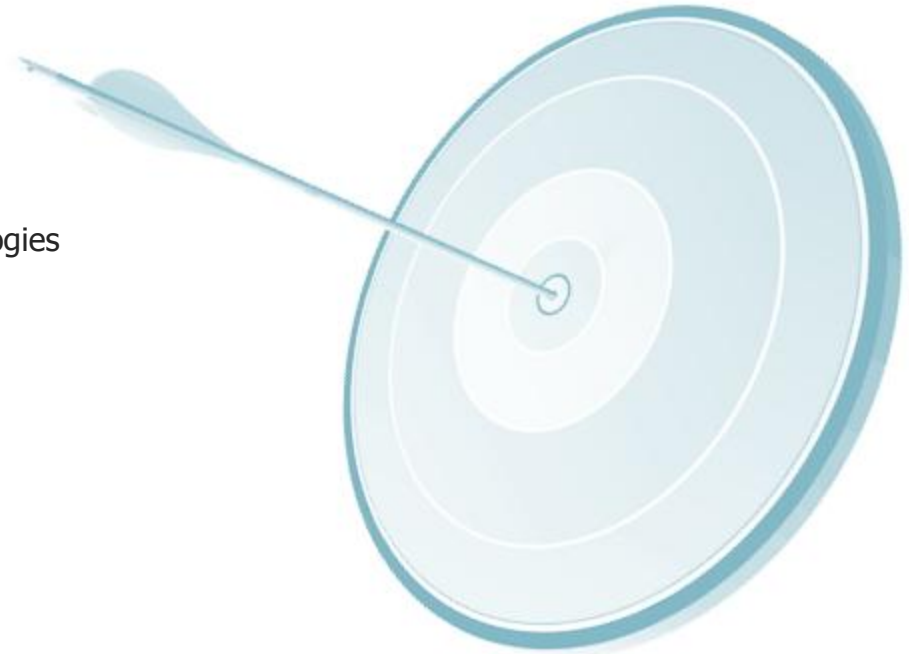
- » Skewness
- » Modality
- » Measures of Center
- » Measures of Spread etc.

→ Understand the relationship between these terminologies

→ Understand the rules of probability

→ Learn about Disjoint and Independent events

→ Analyze airlines dataset to gather insights



Course Topics

edureka!

- Module 1
  - **Statistics and Basic Probability**
- Module 2
  - Conditional Probability and Bayesian Inference
- Module 3
  - Probability Distributions and Regression Modeling

Few Terminologies

edureka!

Famous Sampling Bias

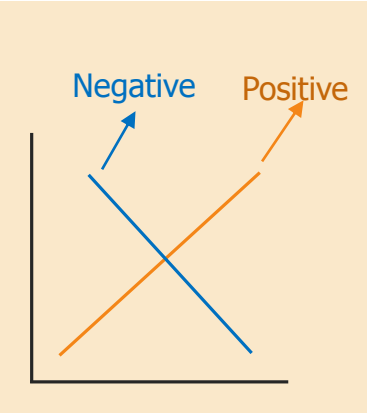
edureka!

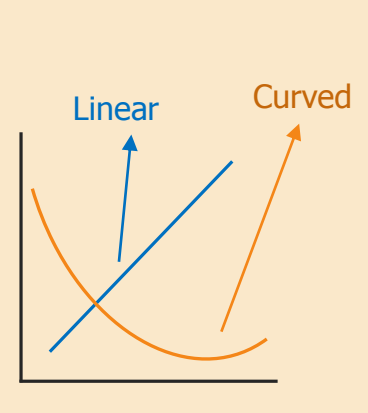
In United States presidential elections of 1936, the Democratic candidate, Franklin D. Roosevelt won over the Republican candidate Alf Landon by 62% of the votes.

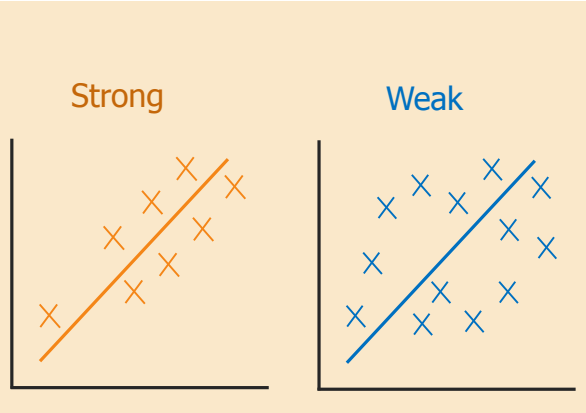
Alf Landon lost by 43% of the votes

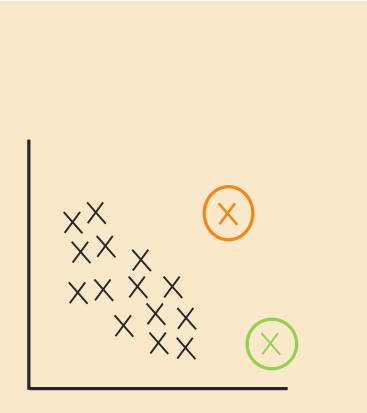
Exploratory Analysis

edureka!





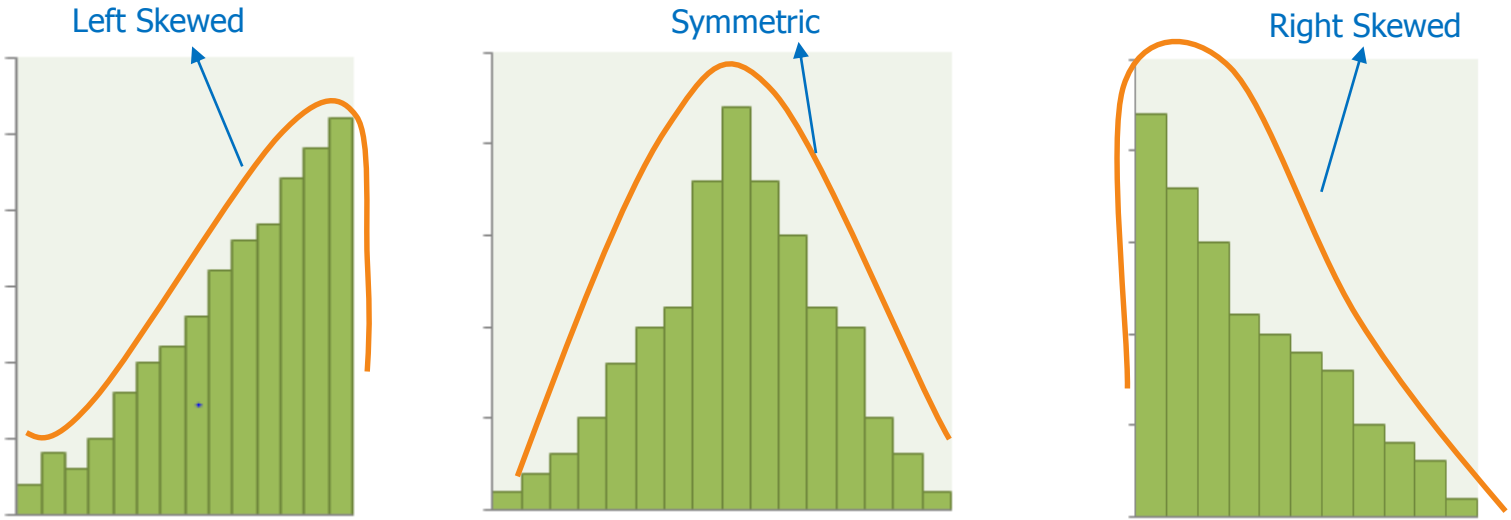




Few Terminologies

edureka!

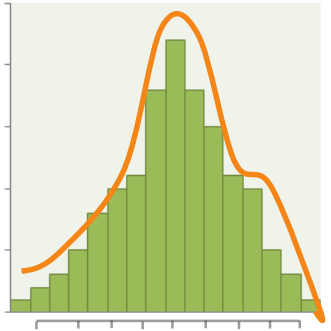
Skewness is a measure of the asymmetry of the distribution



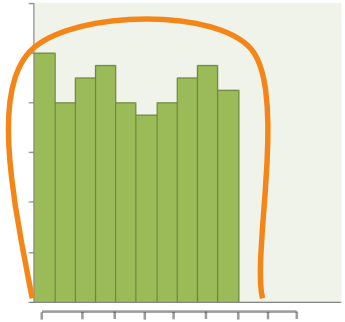
Few Terminologies

edureka!

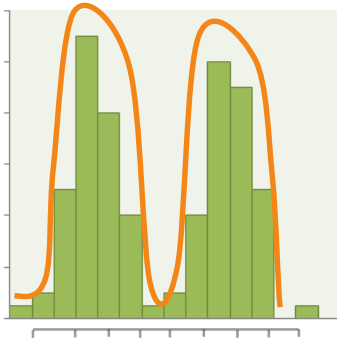
Unimodal



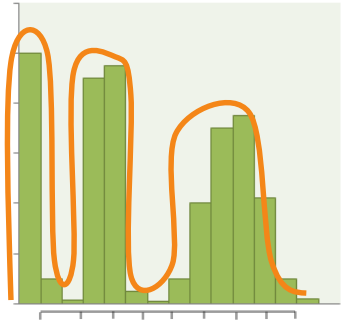
Uniform



Bimodal



Multimodal



Measures of Center

edureka!

Mean	Median	Mode
Arithmetic average $\overline{x}$ mean	Midpoint of the distribution (50th percentile)	Most frequent observation

Annie's Introduction

Annie's Question

edureka!

Calculate the mean, mode and median for the below 10 students exam scores:

98,35,67,85,56,78,45,88,98,92

Annie's Answer

edureka!

Ans. Mean : 74.2
Mode : 98
Median: 81.5

Measures of Spread

edureka!

- → Variance
- → Standard deviation
- → Range
- → Inter-quartile range

Variance

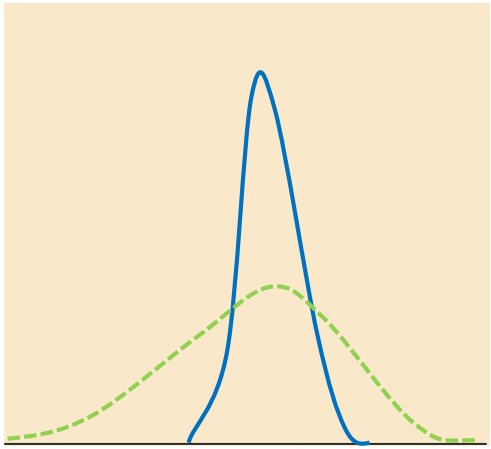
→ Variance measures how far a set of numbers is spread out

$$s^2 = \frac{\sum_{1}^{n=1} (x_i - \bar{x})^2}{n}$$

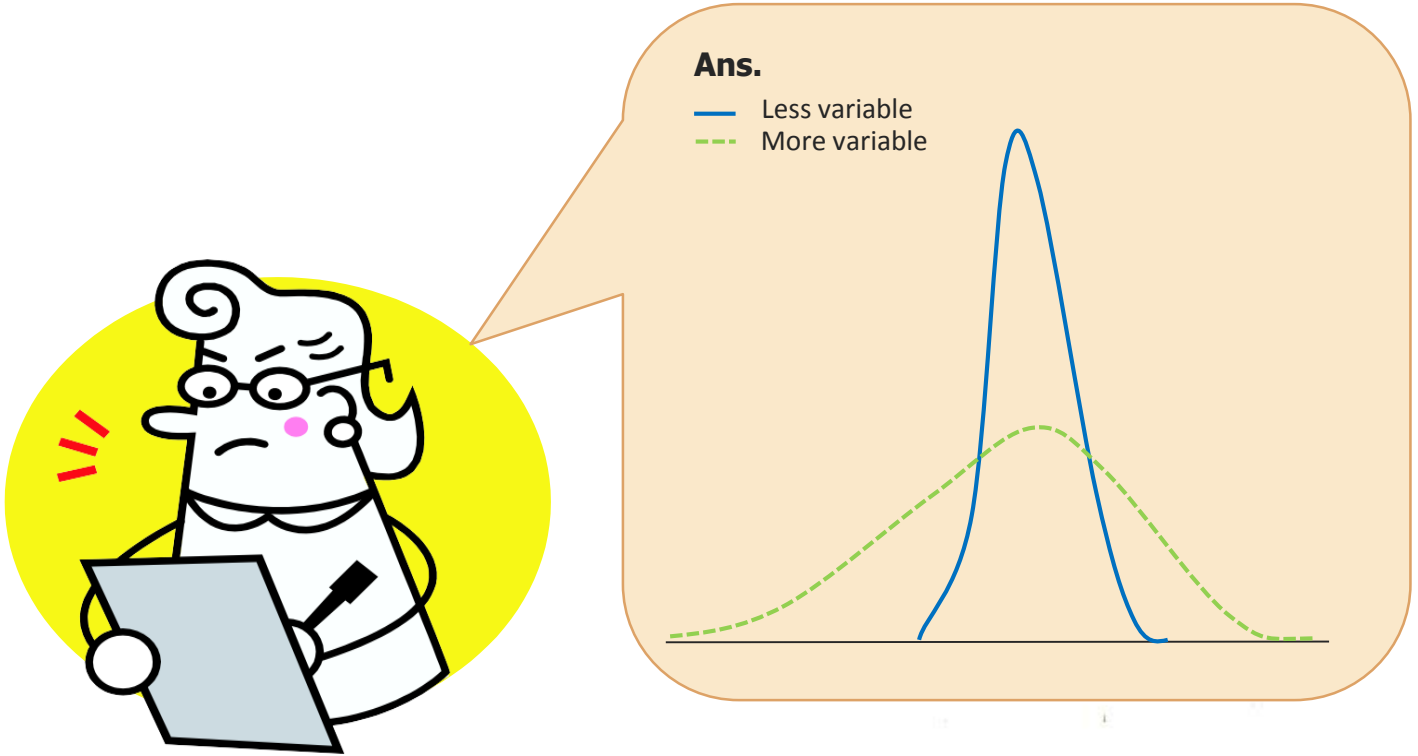
Annie's Question

Which among the following 2 curves are more variable

1. The one in blue solid line
2. The one in green dotted line



Annie's Answer



Standard Deviation

edureka!

→ The standard deviation measures the amount of variation or dispersion from the average

→ Represented as:
  » SD, σ

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{1}^{n=1} (x_i - \bar{x})^2}{n}}$$

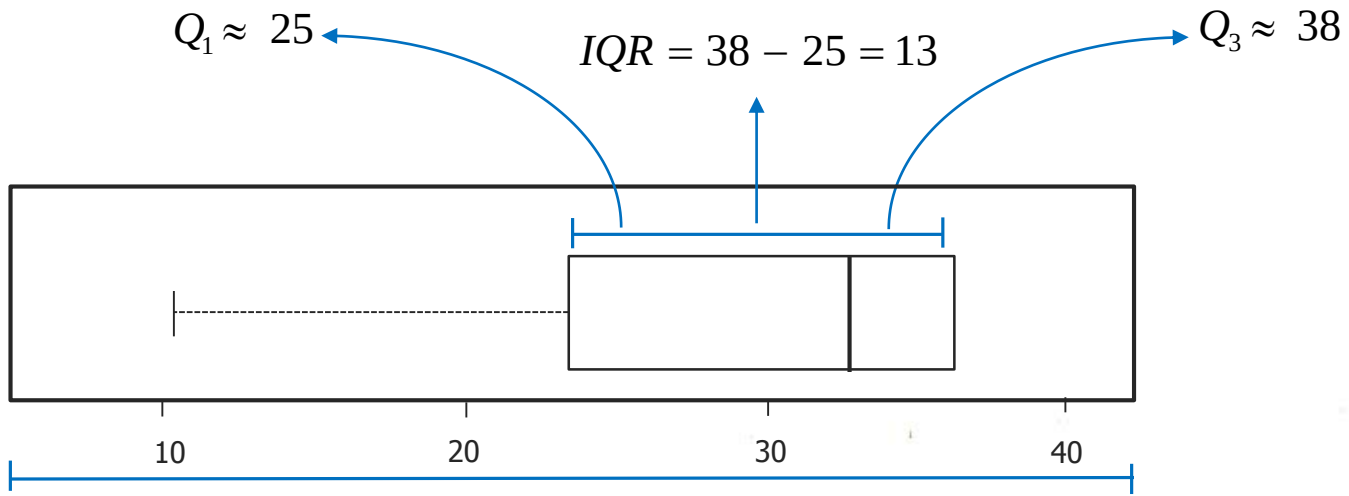
→ Thus, standard deviation is the square root of variance

Interquartile Range

edureka!

→ The interquartile range (IQR), also called the midspread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles

$$IQR = Q3 - Q1$$



Probability

edureka!

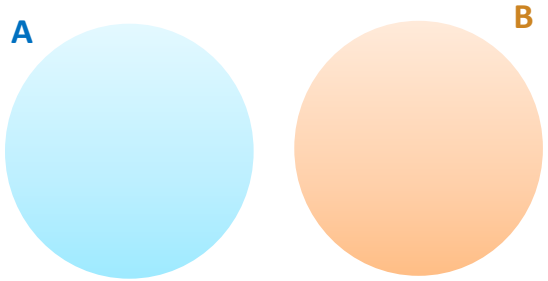
- → Probability rules
- → Conditional probability
- → Probability distributions
- → Binomial
- → Normal

Probability

edureka!

→ Disjoint events do not have any common outcomes

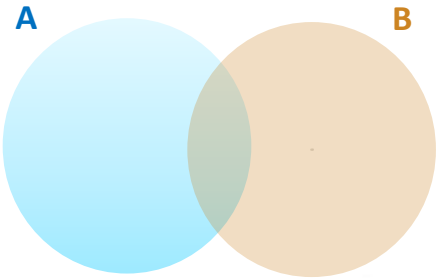
- » The outcome of a ball delivered cannot be a sixer and a wicket
- » A single card drawn from a deck cannot be a king and a queen
- » A man cannot be dead and alive



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

→ Non-disjoint events can have common outcomes

- » A student can get 100 marks in statistics and 100 marks in probability
- » The outcome of a ball delivered can be a no ball and a six

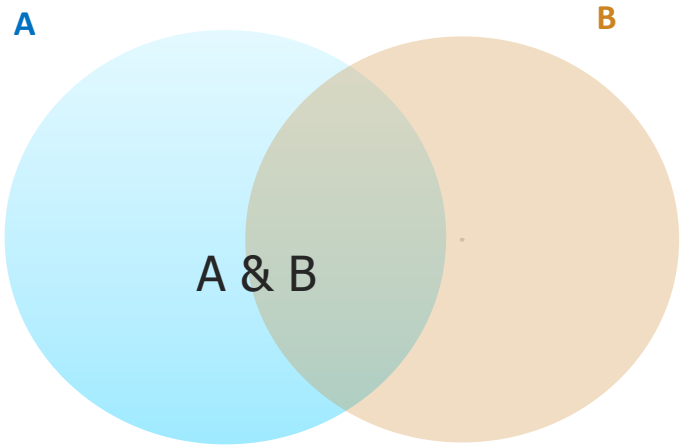


$$P(A \text{ and } B) \neq 0$$

General addition rule

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

Note: When A and B are disjoint, P(A and B) = 0,
Hence P(A or B) = P(A) + P(B)



Sample space

→ The sample space of an experiment or random trial is the set of all possible outcomes or results of that experiment

Example: A coin is tossed 2 times, what is the sample space for the outcomes of these tosses?

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

Probability distributions

→ A probability distribution assigns a probability to each measurable subset of the possible outcomes of a random experiment

One Toss	Head	Tail
Probability	0.5	0.5

Two tosses	Head-head	Tail-tail	Head-tail	Tail-head
Probability	0.25	0.25	0.25	0.25

Rules

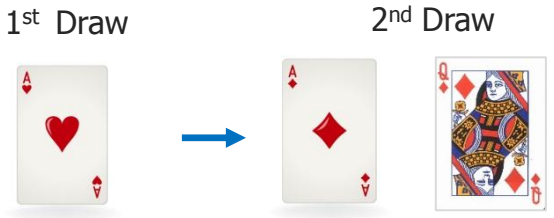
1. The outcomes listed must be disjoint
2. Each probability must be between 0 and 1
3. The probabilities must sum to 1

Independence

→ Two processes are independent if the occurrence of one does not affect the probability of the other.

$P(4\,runs) = 0.5 \quad P(6\,runs) = 0.5$

Outcomes of two balls (assume for simplicity 4 or 6 as the sample space) in a cricket match are independent



$P(A) = \frac{3}{51} \quad P(Q) = \frac{4}{51}$

Outcomes of two draws (say Ace and Queen without replacement) are dependent

Independence

→ $P(A|B) = P(A)$, then A and B are independent

Referred as A 'given' B

Product rule for independent events

→ If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

Example: Naptha has 2 kids, what is the probability of both the kids being female?

Probability

edureka!

Edureka has done a survey about its course.

The most recent phase of the survey that polled 100,000 participants estimates that a 80% of the population agree with the statement "The duration of the courses conducted by Edureka is just right".

The survey also estimates that 10% people have university degree, and that 5% of people fit both criteria.

P(agree) = 0.80
P(University degree) = 0.1
P(agree and university degree) = 0.05

Probability

edureka!

1. Are agreeing with the statement "Duration of courses is just right" and having a university degree disjoint events?

Probability

edureka!

1. Are agreeing with the statement "Men should have more right to a job than women" and having a university degree disjoint events?

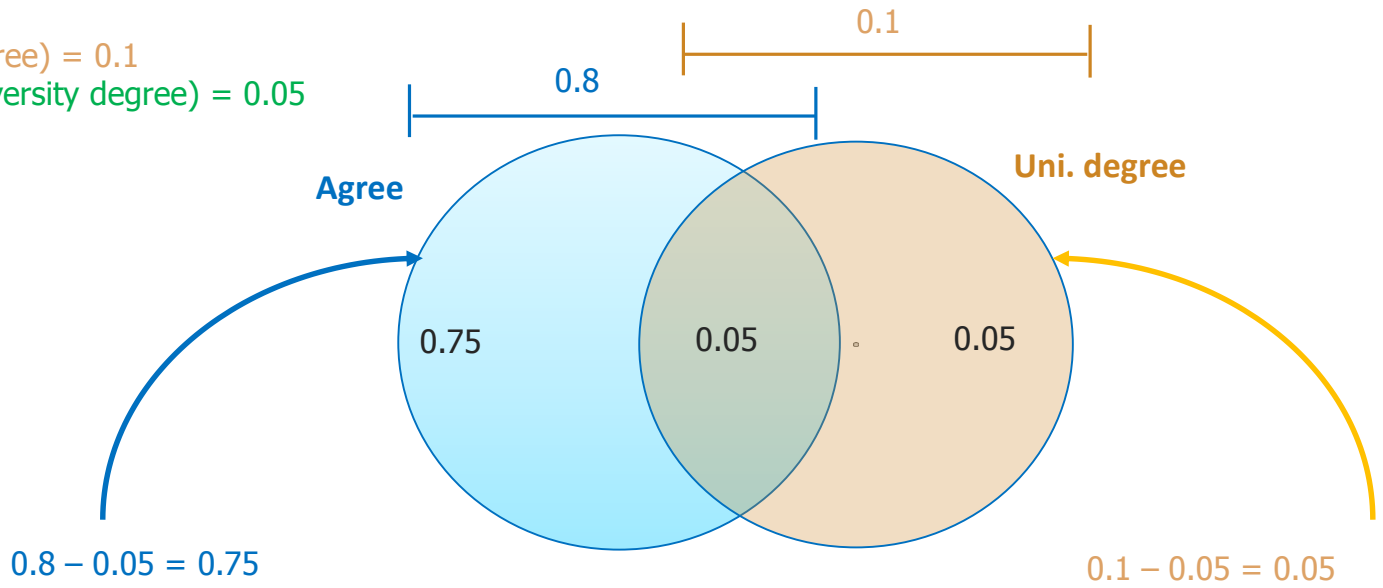
P(agree) = 0.80
P(University degree) = 0.10
P(agree and university degree) = 0.05 ≠0 ⟶ not disjoint

2. Draw a Venn diagram summarizing the variables and their associated probabilities.

P(agree) = 0.8
P(University degree) = 0.1
P(agree and university degree) = 0.05

2. Draw a Venn diagram summarizing the variables and their associated probabilities.

P(agree) = 0.8
P(University degree) = 0.1
P(agree and university degree) = 0.05



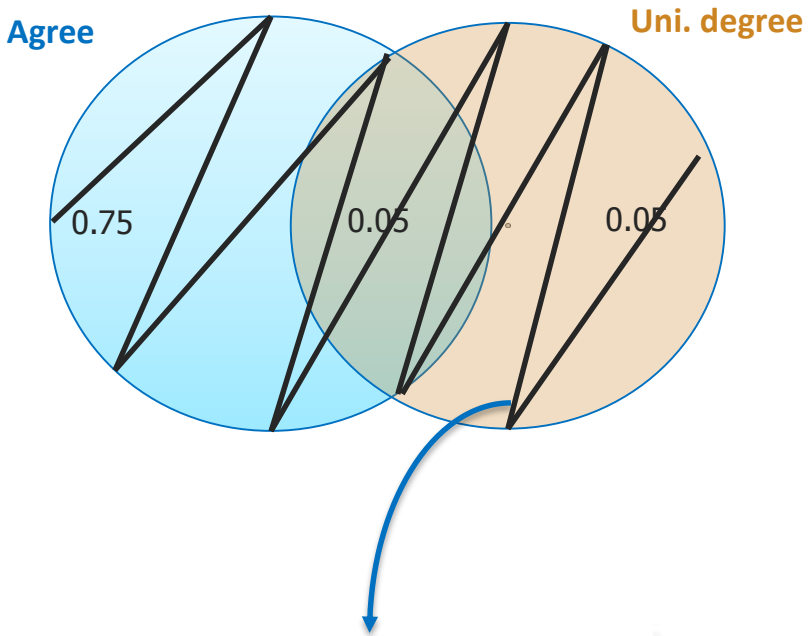
3. What is a probability that a randomly drawn person has a university degree or agrees with the statement about duration time?

P(agree) = 0.8
P(University degree) = 0.1
P(agree and university degree) = 0.05

General addition rule

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

P(agree or uni. degree) = 0.362
= P(agree) + P(uni. degree) – P(agree & uni. degree)
= 0.8 + 0.1 – 0.05
= 0.85

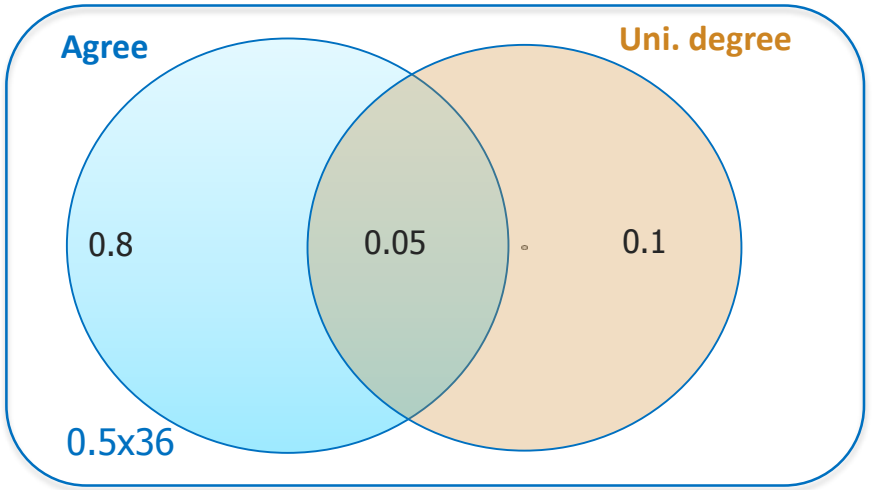


0.326 + 0.036 + 0.102 = 0.464

4. What percent of the population do not have a university degree and disagree with the statement about duration time of lectures

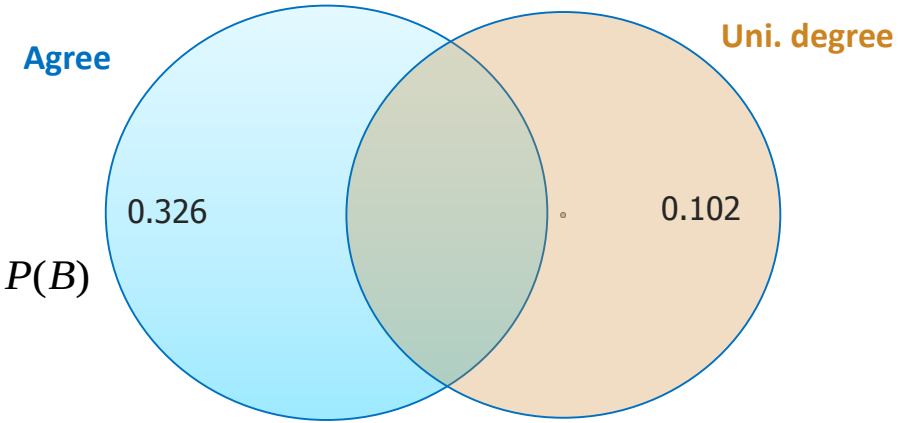
P(agree) = 0.8
P(University degree) = 0.1
P(agree and university degree) = 0.05
P(agree or uni. degree) = 0.85

P(neither agree nor uni. degree)
= 1 – P(agree or uni. degree)
= 1– 0.85
= 0.15



5. Does it appear that the event that someone agrees with the statement about duration is independent of the event that they have a university degree?

P(agree) = 0.8
P(University degree) = 0.1
P(agree and university degree) = 0.05
P(agree or uni. degree) = 0.85



If A and B are independent, $P(A \text{ and } B) = P(A) \times P(B)$

Let's check whether the statement is true or not.

P(neither agree & uni. degree) = P(agree) X P(uni. degree)

$= 0.036 \neq 0.05$ ⟶ not independent

As L.H.S ≠ R.H.S that means values are not independent

Statistics

edureka!

Disjoint Events

Two events cannot happen at the same time

Independent Events

outcome of one provides no useful information about the outcome of the other

Project – Part 1

About the Data

edureka!

The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. There are around 29 attributes.

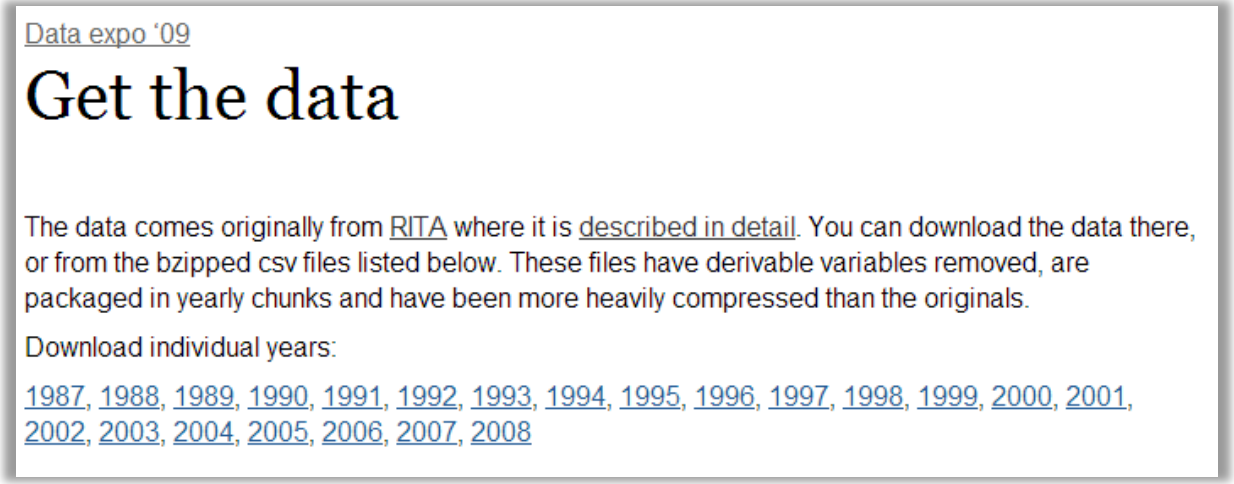
The data used in this project is real and is based on collection over more than 20 years.

The total number of records in this dataset is roughly around 120 million rows.

How to get the Data?

edureka!

The data originally comes from http://stat-computing.org/dataexpo/2009/the-data.html
You will see a screenshot like this in that site


Data expo '09

Get the data

The data comes originally from RITA where it is described in detail. You can download the data there, or from the bzipped csv files listed below. These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals.

Download individual years:

1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008

You can download the data for each year by clicking the appropriate link in the above website (Remember the size is going to be more than 12GB).

Variable Descriptions in the Data

edureka!

In order to understand the data, one has to follow the following variable descriptions:

Serial No	Variable	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number

Variable Descriptions in the Data (Contd.)

edureka!

Serial No	Variable	Description
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)

Variable Descriptions in the Data (Contd.)

edureka!

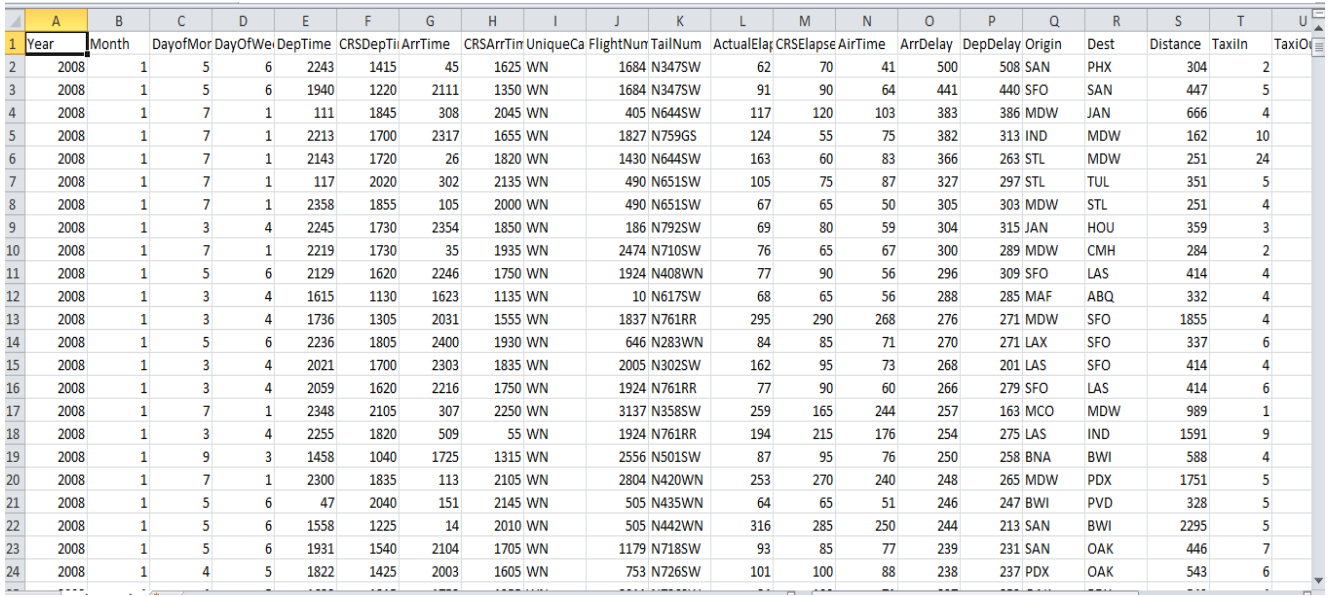
Serial No	Variable	Description
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

Snapshot of the Dataset

edureka!

You can take any of the years and try to solve the following problems.

A screenshot containing the 25 first lines may look like this:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Year	Month	DayofMor	DayOfWe	DepTime	CRSDepTi	ArrTime	CRSArrTin	UniqueCa	FlightNun	TailNum	ActualEla	CRSElapse	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiO
2	2008	1	5	6	2243	1415	45	1625	WN	1684	N347SW	62	70	41	500	508	SAN	PHX	304	2	
3	2008	1	5	6	1940	1220	2111	1350	WN	1684	N347SW	91	90	64	441	440	SFO	SAN	447	5	
4	2008	1	7	1	111	1845	308	2045	WN	405	N644SW	117	120	103	383	386	MDW	JAN	666	4	
5	2008	1	7	1	2213	1700	2317	1655	WN	1827	N759GS	124	55	75	382	313	IND	MDW	162	10	
6	2008	1	7	1	2143	1720	26	1820	WN	1430	N644SW	163	60	83	366	263	STL	MDW	251	24	
7	2008	1	7	1	117	2020	302	2135	WN	490	N651SW	105	75	87	327	297	STL	TUL	351	5	
8	2008	1	7	1	2358	1855	105	2000	WN	490	N651SW	67	65	50	305	303	MDW	STL	251	4	
9	2008	1	3	4	2245	1730	2354	1850	WN	186	N792SW	69	80	59	304	315	JAN	HOU	359	3	
10	2008	1	7	1	2219	1730	35	1935	WN	2474	N710SW	76	65	67	300	289	MDW	CMH	284	2	
11	2008	1	5	6	2129	1620	2246	1750	WN	1924	N408WN	77	90	56	296	309	SFO	LAS	414	4	
12	2008	1	3	4	1615	1130	1623	1135	WN	10	N617SW	68	65	56	288	285	MAF	ABQ	332	4	
13	2008	1	3	4	1736	1305	2031	1555	WN	1837	N761RR	295	290	268	276	271	MDW	SFO	1855	4	
14	2008	1	5	6	2236	1805	2400	1930	WN	646	N283WN	84	85	71	270	271	LAX	SFO	337	6	
15	2008	1	3	4	2021	1700	2303	1835	WN	2005	N302SW	162	95	73	268	201	LAS	SFO	414	4	
16	2008	1	3	4	2059	1620	2216	1750	WN	1924	N761RR	77	90	60	266	279	SFO	LAS	414	6	
17	2008	1	7	1	2348	2105	307	2250	WN	3137	N358SW	259	165	244	257	163	MCO	MDW	989	1	
18	2008	1	3	4	2255	1820	509	55	WN	1924	N761RR	194	215	176	254	275	LAS	IND	1591	9	
19	2008	1	9	3	1458	1040	1725	1315	WN	2556	N501SW	87	95	76	250	258	BNA	BWI	588	4	
20	2008	1	7	1	2300	1835	113	2105	WN	2804	N420WN	253	270	240	248	265	MDW	PDX	1751	5	
21	2008	1	5	6	47	2040	151	2145	WN	505	N435WN	64	65	51	246	247	BWI	PVD	328	5	
22	2008	1	5	6	1558	1225	14	2010	WN	505	N442WN	316	285	250	244	213	SAN	BWI	2295	5	
23	2008	1	5	6	1931	1540	2104	1705	WN	1179	N718SW	93	85	77	239	231	SAN	OAK	446	7	
24	2008	1	4	5	1822	1425	2003	1605	WN	753	N726SW	101	100	88	238	237	PDX	OAK	543	6	

Module-1 Problem Statement

edureka!

1. Check the skewness of Distance travelled by airlines
2. Calculate the mean, median and quantiles of the distance travelled by US Airlines (US)
3. Check the standard deviation of distance travelled by American Airlines (AA)
4. Draw a boxplot of UniqueCarrier with Distance
5. Draw the direction of relationship between ArrDelay and DepDelay by drawing a scatterplot

What’s within the LMS

Agenda for Next Class

edureka!

→ Conditional Probability & Bayesian Inference

- » Terms
- » Definitions
- » Examples
- » Concepts & Applications

QUESTIONS

Survey

edureka!

Your feedback is important to us, be it a compliment, a suggestion or a complaint. It helps us to make the course better!

Please spare few minutes to take the survey after the webinar.

edureka!