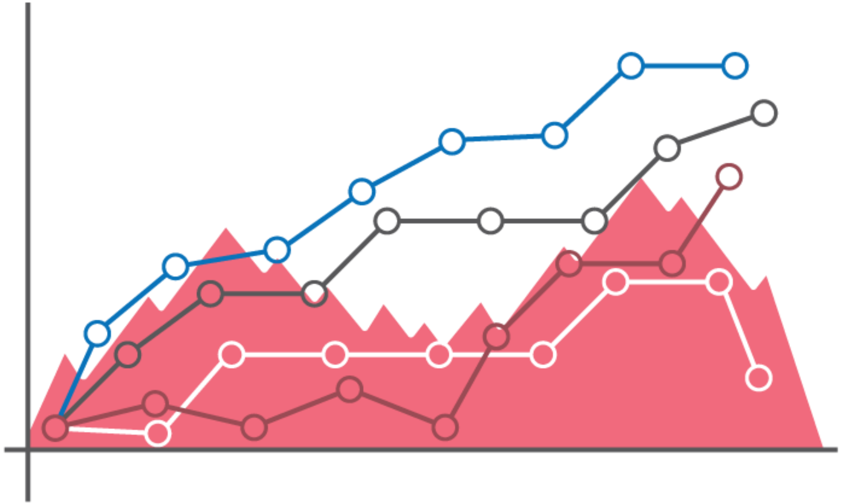
edureka!



MODULE-2
BASIC PROBABILITY, CONDITIONAL PROBABILITY AND BAYESIAN INFERENCE

Objectives

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

→ Understand the rules of probability

→ Learn about Disjoint and Independent events

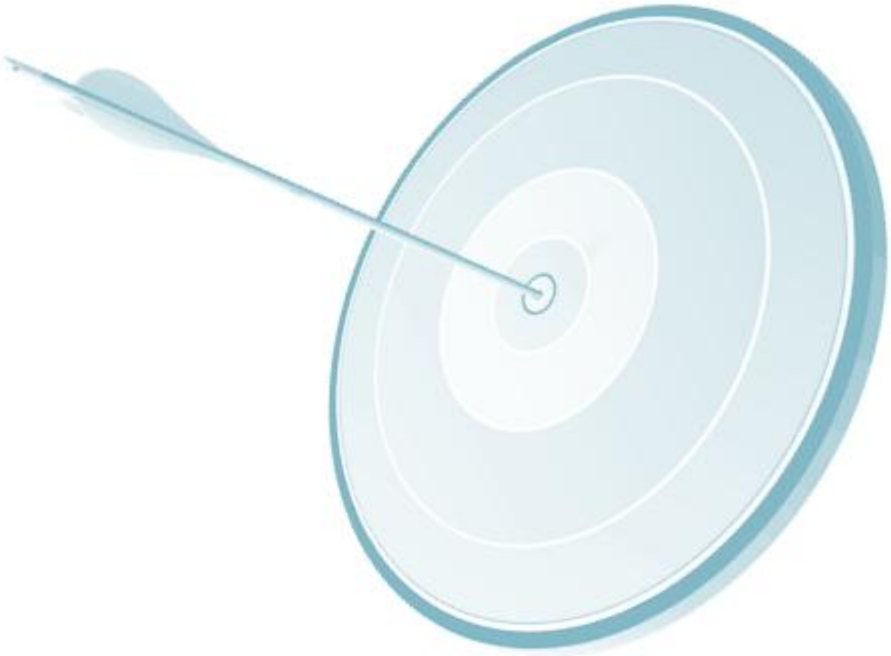
→ Understand the concept of probability

- » Marginal Probability
- » Joint
- » Conditional

→ Implement these Probabilities on a case-study

→ Learn and implement Bayes' Theorem

→ Implement Bayes’ theorem on a case-study



Course Topics

→ Module 1

- » Statistics and Basic Probability

→ Module 2

- » **Basic Probability, Conditional Probability and Bayesian Inference**

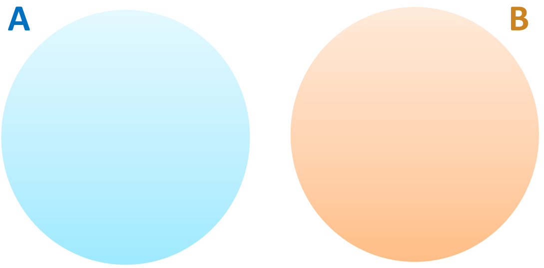
→ Module 3

- » Probability Distributions and Regression Modeling

Probability

→ Disjoint events do not have any common outcomes

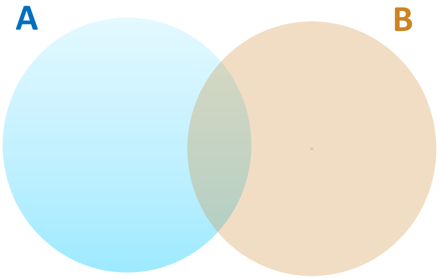
» The outcome of a ball delivered cannot be a sixer and a wicket
» A single card drawn from a deck cannot be a king and a queen
» A man cannot be dead and alive



$$P(A \ and \ B) = 0$$

→ 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



$$P(A \ and \ B) \neq 0$$

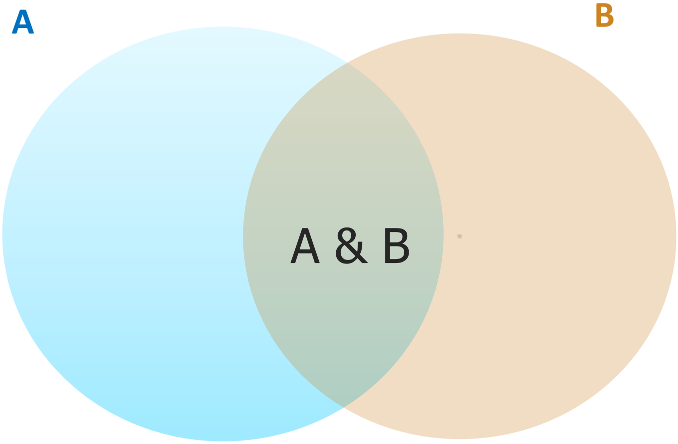
Probability

edureka!

General addition rule

$$P(A + B) = P(A) + P(B) - P(A \; and \; B)$$

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



edureka!

Sample space

$\rightarrow$ 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

edureka!

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

Probability

Independence

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

1st Ball		2nd Ball
4	→	6

$P(4runs) = 0.5 \qquad P(6runs) = 0.5$

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

1st Draw

→

2nd Draw

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

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

Probability

edureka!

Independence

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

Referred as A 'given' B

Probability

Product rule for independent events

$\rightarrow$ If A and B are independent, $P(A \ and \ B) = P(A) \times P(B)$

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

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

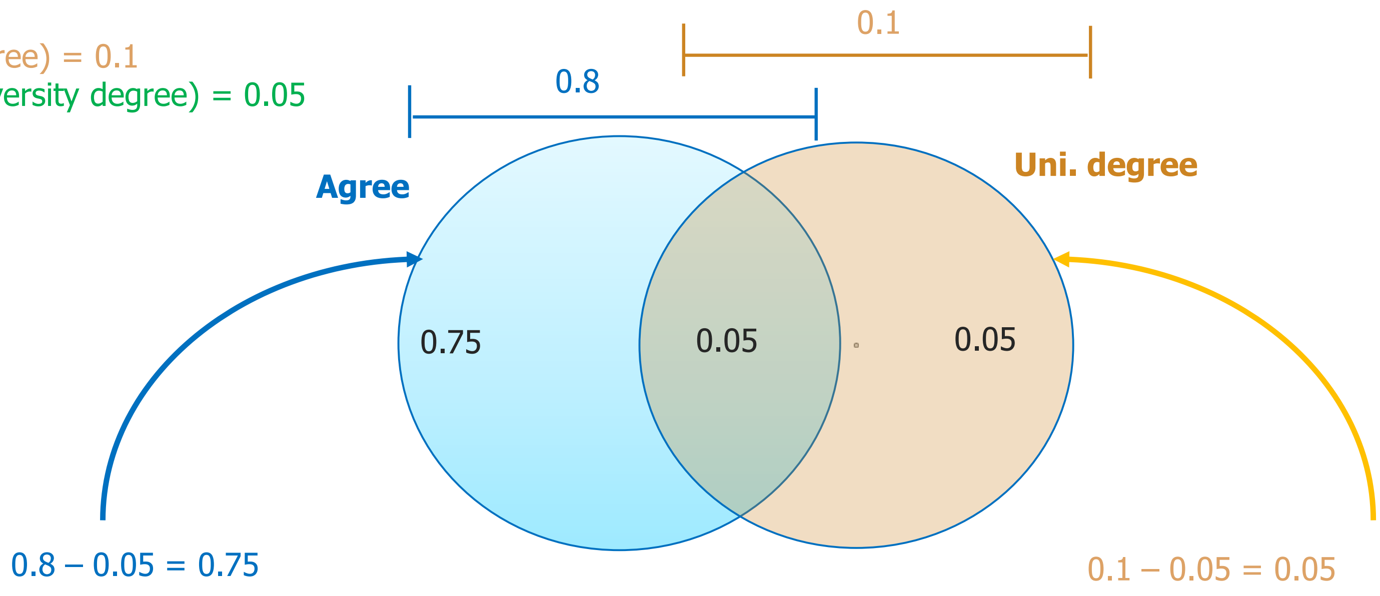
   P(agree) = 0.80
   P(University degree) = 0.10
   P(agree and university degree) = 0.05 $\neq 0$ $\longrightarrow$ not disjoint

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



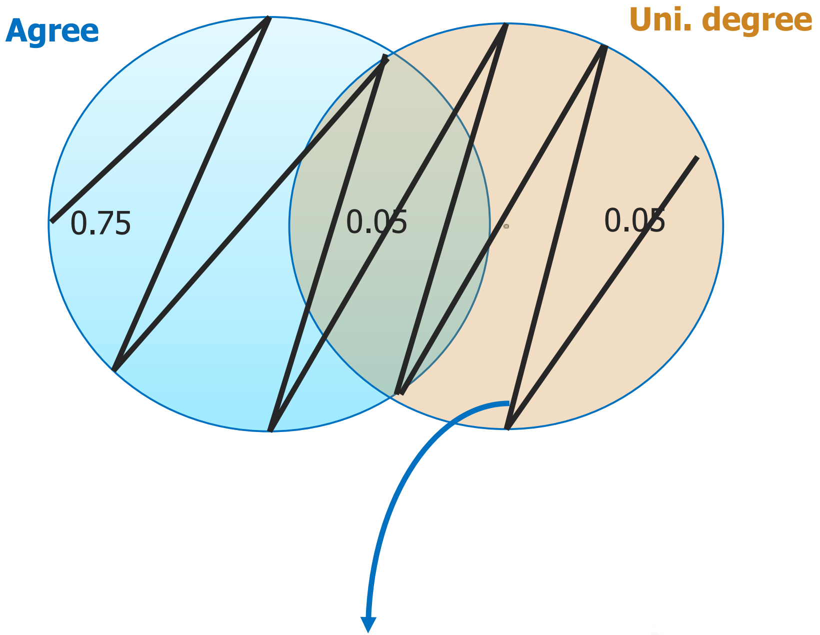
3. What is a probability that a randomly drawn person has a university degree or agrees with the statement about duration time?

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

General addition rule

$$P(A + B) = P(A) + P(B) - P(A \; and \; B)$$

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



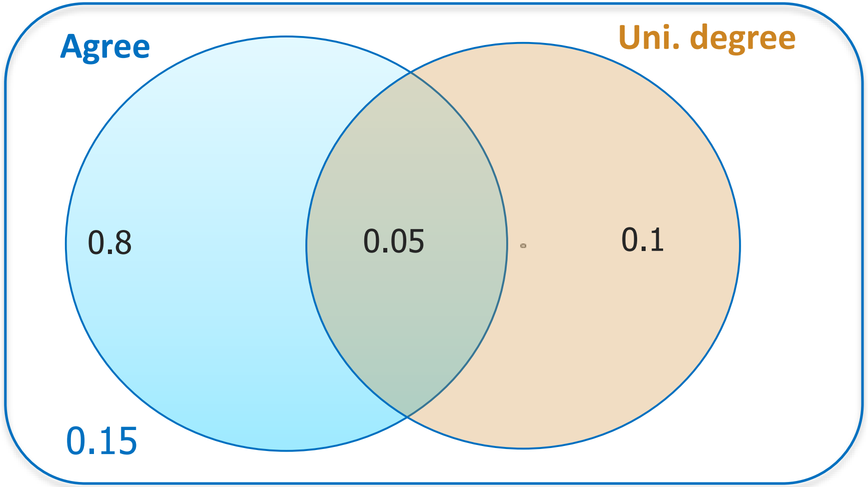
0.75 + 0.05 + 0.05 = 0.85

edureka!

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

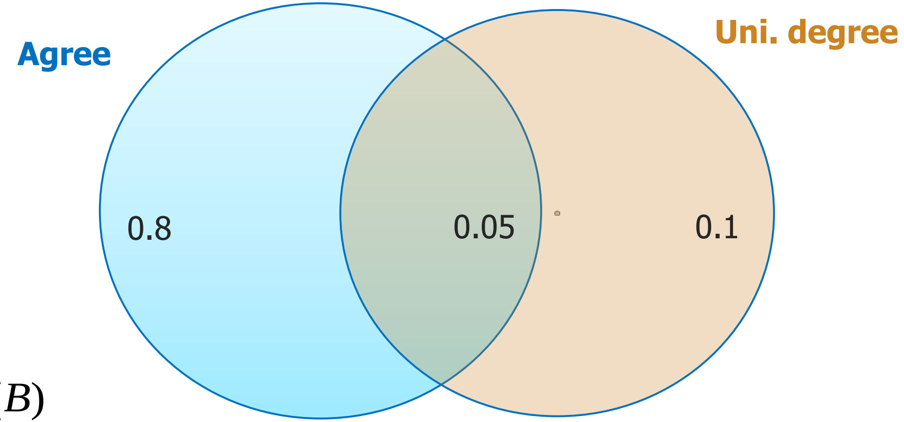
P(agree) = 0.8
P(University degree) = 0.1
P(agree and university degree) = 0.05
P(agree or uni. degree) = 0.85

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, $P(A\ and\ B) = P(A) \times P(B)$

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 $\longrightarrow$ not independent

As L.H.S $\neq$ R.H.S that means values are not independent

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

EDUREKA TRAINING STUDY
TRAINING AND SALARY PACKAGE OF CANDIDATES Study examining salary package and training undergone by candidates
Sample: 60 candidates without training and 45 candidates with edureka training
Study Design: - Assessment of training with salary package

edureka!

Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

Probability and Statistics

edureka!

What is the probability that a candidate has undergone Edureka training?

Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

edureka!

Marginal

What is the probability that a candidate has undergone Edureka training?

$$P(Edu.Training) = 45/105 \approx 0.42$$

Results:		**Training**		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

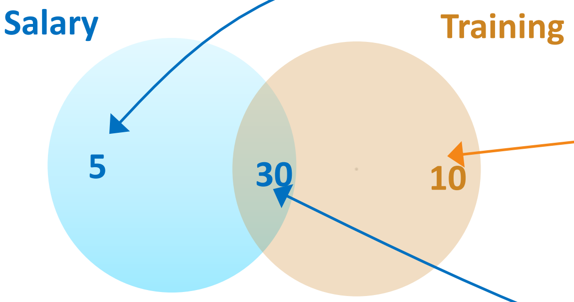
Probability and Statistics

What is the probability that a candidate has attended Edureka training and also has good package?

Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

Joint

What is the probability that a candidate has attended Edureka training and also has good package?



Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

$$P(Good\ Package\ \&\ Edu.Training) = 30/105 \approx 0.28$$

What is the probability that a candidate has a good package given that he has not undergone training?

Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

Conditional

What is the probability that a candidate has a good package given that he has not undergone training?

Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

$$P(Good\ Package \mid Without\ Edureka) = 5/60 \approx 0.08$$

Probability and Statistics

edureka!

$$Bayes'\,theorem:$$

$$P(A\,|\,B) = \frac{P(A\ and\ B)}{P(B)}$$

Results:		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

$$P(Good\ Package\,|\,Without\ Edureka) = \frac{P(Good\ Package\ \&\ Without\ Edureka)}{P(Without\ Edureka)}$$

$$= \frac{5/105}{60/105} = \frac{5}{60} \approx 0.08$$

Probability and Statistics

A study estimates that 24.8% of people in the world are lazy, 23.6% of people do not have a job, and 6.1% fall into both categories.

Based on this information, what percent of people are lazy given that they do not have a job?

A study estimates that 24.8% of people in the world are lazy, 23.6% of people do not have a job, and 6.1% fall into both categories.

Based on this information, what percent of people are lazy given that they do not have a job?

$$\boxed{P(Lazy \mid No.Job) = ?}$$

$$\frac{P(Lazy\ \&\ No.Job)}{P(No.job)} = \frac{0.061}{0.236} \approx \boxed{0.26}$$

$$Bayes'\,theorem:$$

$$P(A \mid B) = \frac{P(A\ and\ B)}{P(B)}$$

edureka!

Product rule for independent events

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

Bayes' theorem

$$P(A \mid B) = \frac{P(A \ and \ B)}{P(B)}$$

General product rule

$$P(A \ AND \ B) = P(A \mid B) \times P(B)$$

Probability and Statistics

A study shows that 15.9% people in India drive above the specified speed limits.

Assume that the government is planning to put some driving speed detectors and levy a huge penalty on drivers who drives above the speed limit.

For those who drive faster than the speed limit, the detector is 89.7% accurate. For those who do not drive more than the speed limit, the detector is 82.6% accurate.

If an individual who is driving a car has been identified by the speed detector as driving overspeed, what is the probability that he is actually driving over-speed?

Probability and Statistics

A study shows that 15.9% people who drive in India actually drive above the specified speed limits.

Assume that the government is planning to put some driving speed detectors and levy a huge penalty on drivers who driver above the speed limit. For those who drive faster than the speed limit, the detector is 89.7% accurate.

For those who do not drive more than the speed limit, the detector is 82.6% accurate. If an individual who is driving a car has been identified by the speed detector as driving over-speed, what is the probability that he is actually driving over-speed?

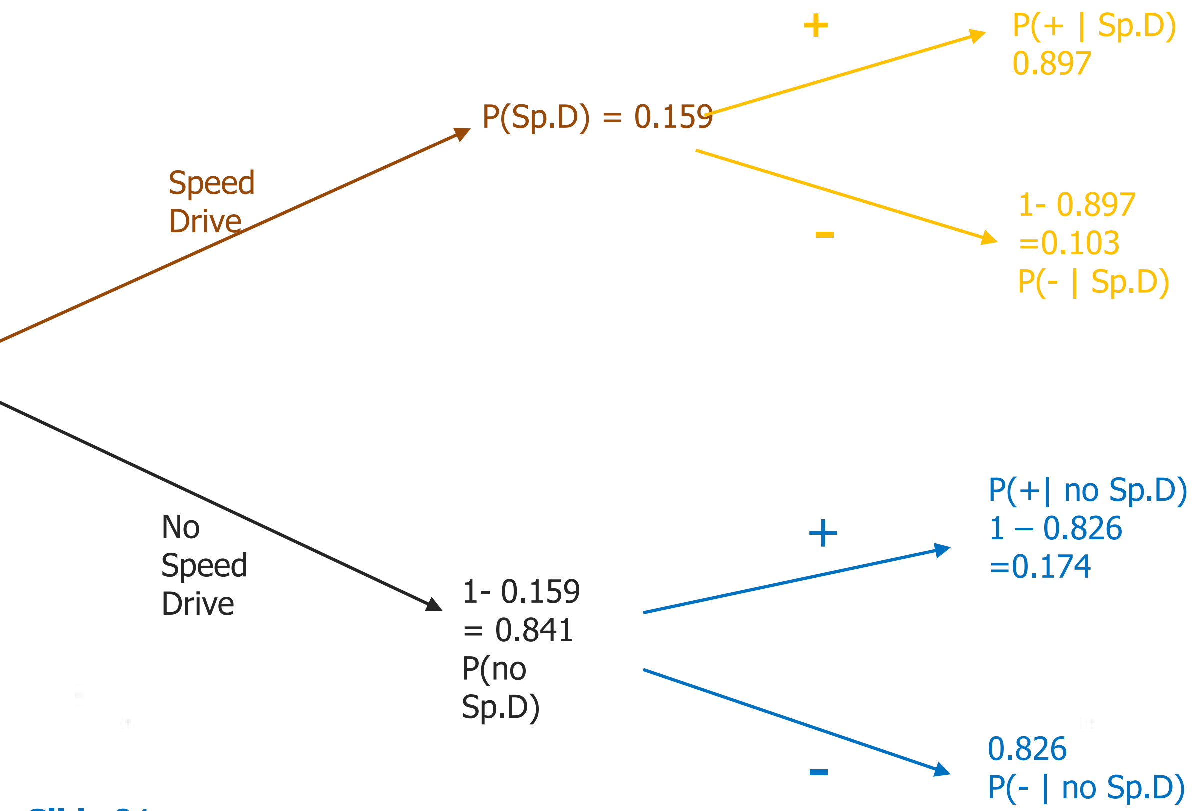
P(over-speed) = 0.159

P(+ /overspeed) = 0.897, P(- / no overspeed) = 0.826

Tree diagram!

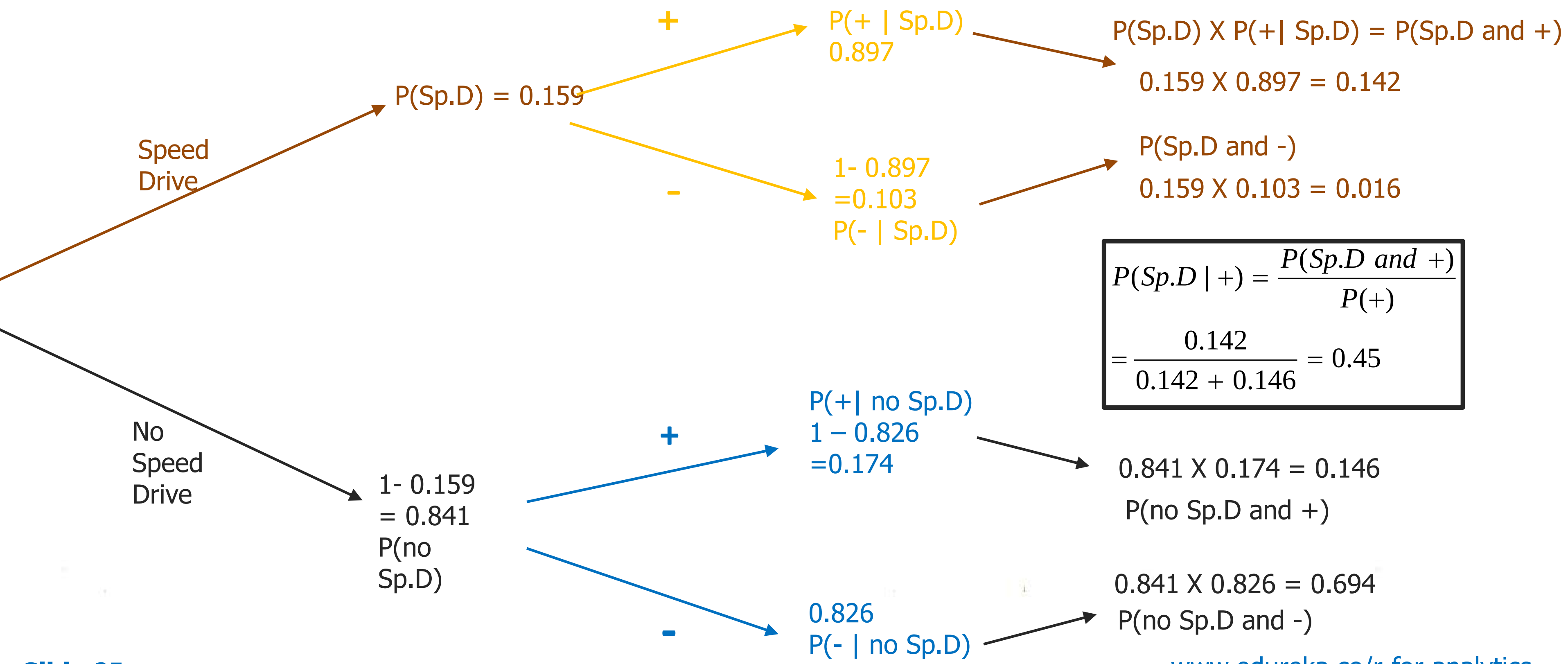
P(overspeed /+) = ?

Probability and Statistics



Probability and Statistics

edureka!



$$P(Sp.D \mid +) = \frac{P(Sp.D \ and \ +)}{P(+)}$$

$$= \frac{0.142}{0.142 + 0.146} = 0.45$$

Probability and Statistics

A study shows that 15.9% people who drive in India actually drive above the specified speed limits.

Assume that the government is planning to put some driving speed detectors and levy a huge penalty on drivers who driver above the speed limit. For those who drive faster than the speed limit, the detector is 89.7% accurate.

For those who do not drive more than the speed limit, the detector is 82.6% accurate. If an individual who is driving a car has been identified by the speed detector as driving over-speed, what is the probability that he is actually driving over-speed?

$$\boxed{\begin{aligned} P(Sp.D \mid +) &= \frac{P(Sp.D \text{ and } +)}{P(+)} \\ &= \frac{0.142}{0.142 + 0.146} = 0.45 \end{aligned}}$$

Project – Part 2

Variable Descriptions in the Data

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	TaxiOu
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-2 Problem Statement

edureka!

1. What is the probability that a flight which is landing/taking off is “WN” Airlines **(marginal probability)**

2. What is the probability that a flight which is landing/taking off is either “WN” or “AA” Airlines **(disjoint events)**

3. What is the joint probability that a flight is both “WN” and travels less than 600 miles **(joint probability)**

4. What is the conditional probability that the flight travels less than 2500 miles given that the flight is “AA” Airlines **(conditional probability)**

5. What is the joint probability of a flight getting cancelled and is supposed to travel less than 2500 miles given that the flight is “AA” Airlines **(joint + conditional probability)**

Agenda for Next Class

→ Probability Distributions & Regression Modeling

- » Normal Distribution
- » Binomial Distribution
- » Linear Regression Model and Analysis

QUESTIONS

Survey

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!