Inspire…Educate…Transform.

Statistics and Probability Fundamentals

Basic Probability Concepts, Probability Distributions

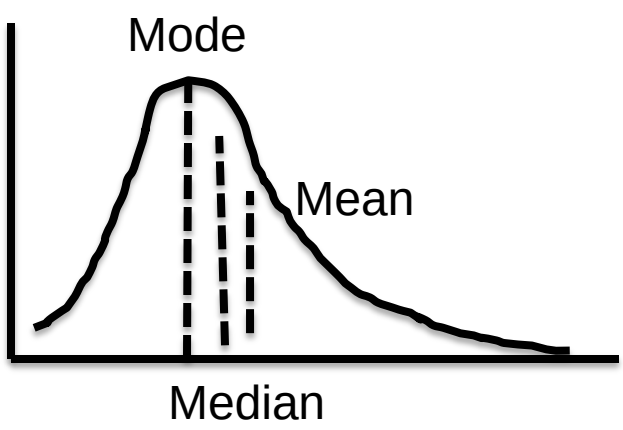
Dr. Anand Jayaraman

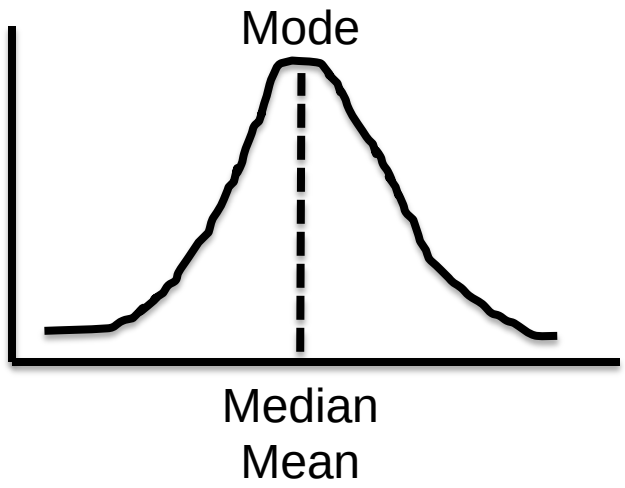
Dec 18, 2016

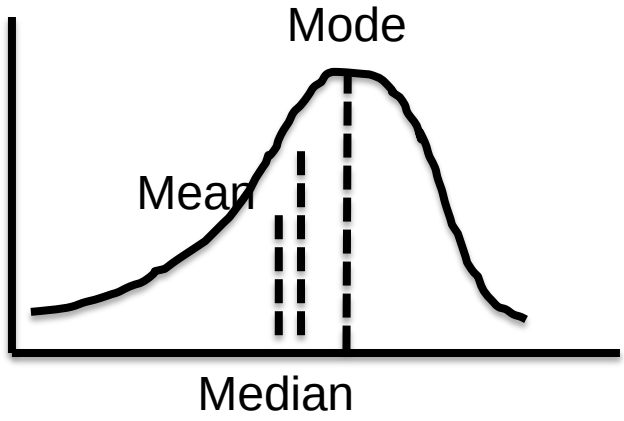
Thanks to Dr.Sridhar Pappu for the material

The Central Tendencies

Identify where the MODE, MEDIAN and MEAN lie in the below distributions.







The Central Tendencies – Recent Interview Question

For the dataset, 13, 4, 7, 10, 8, 5, the median is

- 7.5
- 7
- 5
- 8

Measures of Spread – Recent Interview Question

The spread of the data in a dataset could be studied using

_____

- Interquartile range
- Variance
- Standard Deviation
- Range (max-min)
- All of the above

Measures of Spread – Recent Interview Question

Given the numbers are 68, 83, 58, 84, 100, 64, the second quartile is:

- 74.5
- 75.5
- 75
- 74

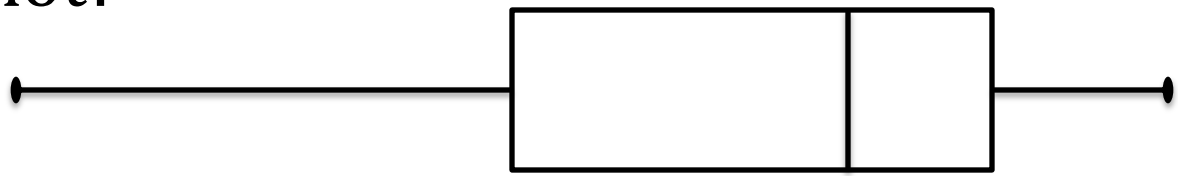
Measures of Spread – Recent Interview Question

Which of the following plot is used to analyze interquartile range

- Scatterplot
- Histogram
- Lineplot
- Boxplot
- All of the above

Measures of Spread – Recent Interview Question

What term would best describe the shape of the given boxplot?



- Symmetric
- Skewed with right tail
- Skewed with left tail
- Normal

Data Types – Recent Interview Question

A sample of 400 Bangalore households is selected and several variables are recorded. Which of the following statements is correct?

- Socioeconomic status (recorded as “low income”, “middle income”, or “high income”) is nominal level data
- The number of people living in a household is a discrete variable
- The primary language spoken in the household is ordinal level data (recorded as “Kannada”, “Tamil”, etc)

Measures of Spread (Dispersion)

We studied Quartiles in depth and mentioned Deciles and Percentiles in passing. However, just as Quartiles divide data into 4 equal parts, Deciles divide it into 10 equal parts and Percentiles into 100 equal parts.

Given the above, find the 25^{th}, 50^{th}, 75^{th} and the 90^{th} percentiles for the top 16 global marketing sectors for advertising spending for a recent year according to *Advertising Age*. Also, find Q2 and IQR. Data in next slide.

Sector	Ad spending (in $ million)
Automotive	22195
Personal Care	19526
Entertainment and Media	9538
Food	7793
Drugs	7707
Electronics	4023
Soft Drinks	3916
Retail	3576
Restaurants	3571
Cleaners	3553
Computers	3247
Telephone	2448
Financial	2433
Beer, Wine and Liquor	2050
Candy	1137
Toys	699

Strategic decisions must be based on hard data

"In God we trust; all others must bring data."

Edward Deming*

*The man behind Japanese post-war industrial revoluti

Present Day

PROBABILITY BASICS

Probability - Applications

- Gaming industry – Establish charges and payoffs
- Manufacturing/Aerospace – Prevent major breakdowns
- Business – Deciding on a business proposal based on probability of success vs cost
- Risk Evaluation – Scenario analysis

Assigning Probabilities

Classical Method – *A priori* or Theoretical

Probability can be determined prior to conducting any experiment.

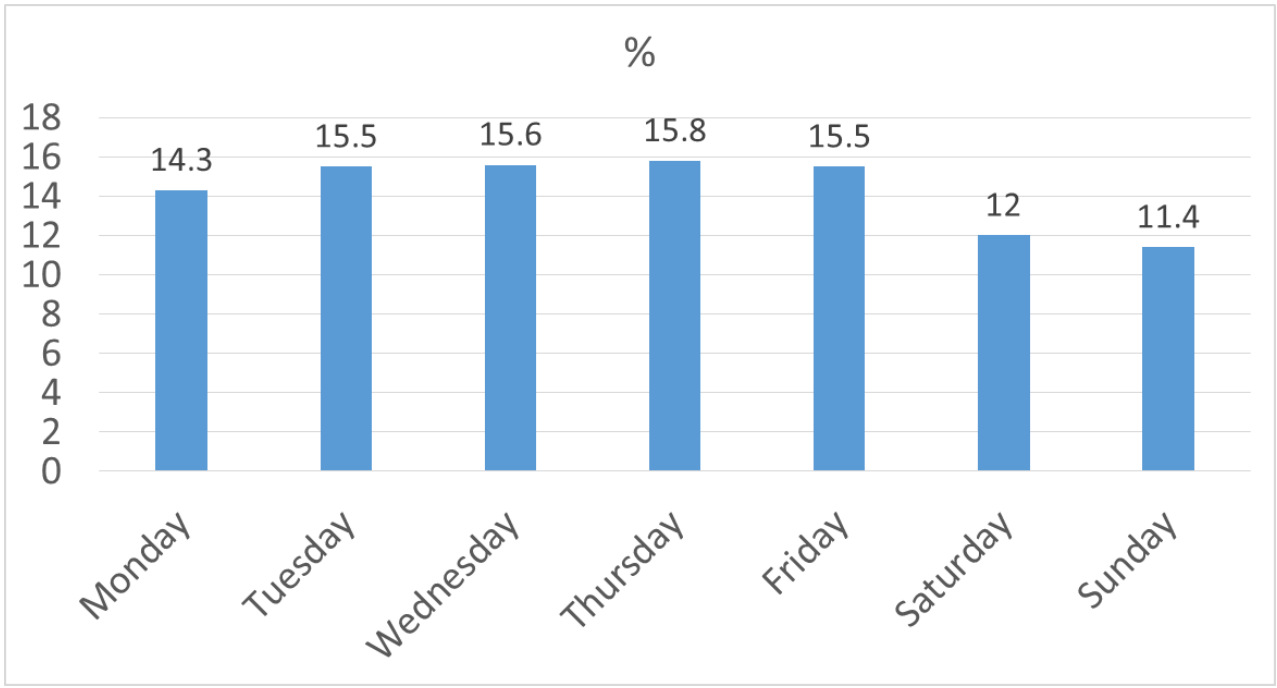
$$P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ event\ occurs}{total\ possible\ \#\ of\ outcomes}$$

Example: Tossing of a fair die

Assigning Probabilities

What is the probability of a baby being born on a Wednesday?

$$1/7 = 14.3\%$$



Data from "Risks of Stillbirth and Early Neonatal Death by Day of Week", by Zhong-Cheng Luo, Shiliang Liu, Russell Wilkins, and Michael S. Kramer, for the Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System. Data of 3,239,972 births in Canada between 1985 and 1998. The reported percentages do not add up to 100% due to rounding.

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Probability can be determined post conducting a thought experiment.

$$P(E) = \frac{\#\ of\ times\ an\ event\ occurred}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occurred}$$

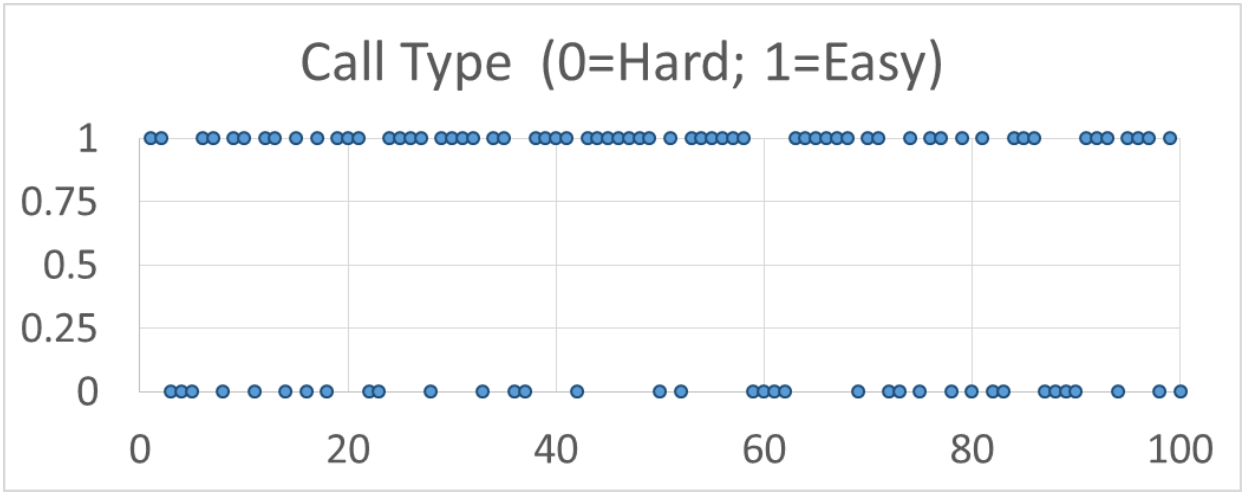
Example: Tossing of a weighted die…well!, even a fair die. The larger the number of experiments, the better the approximation.

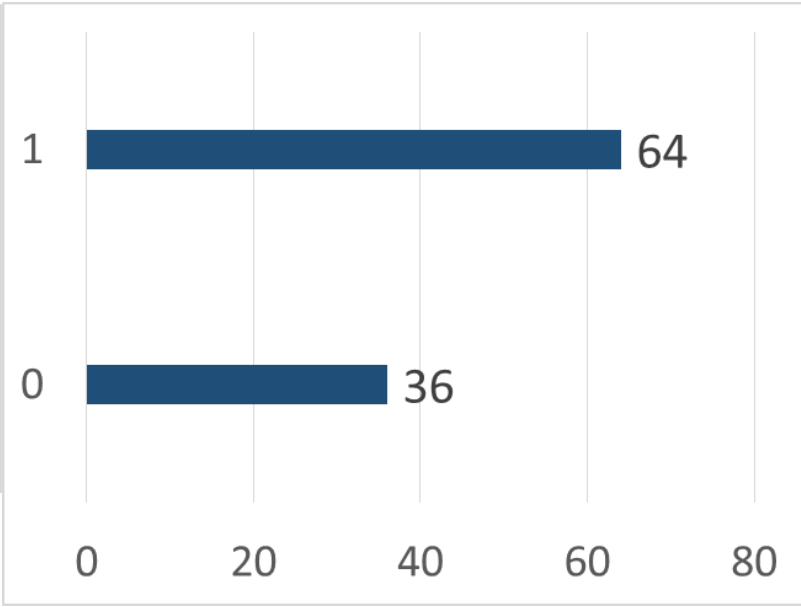
This is the most used method in statistical inference.

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

100 calls handled by an agent at a call centre



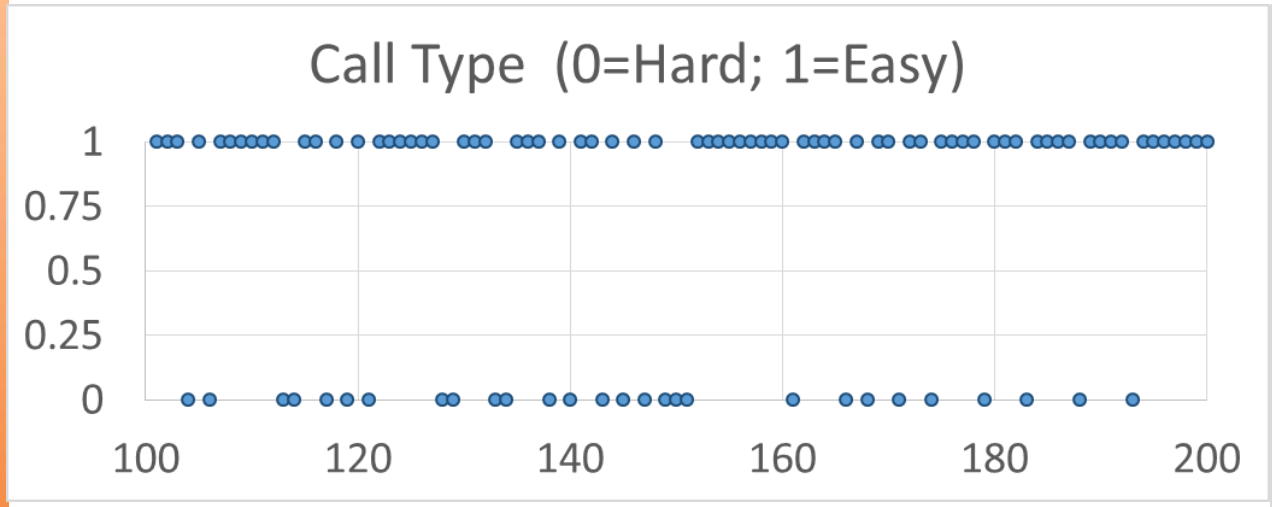


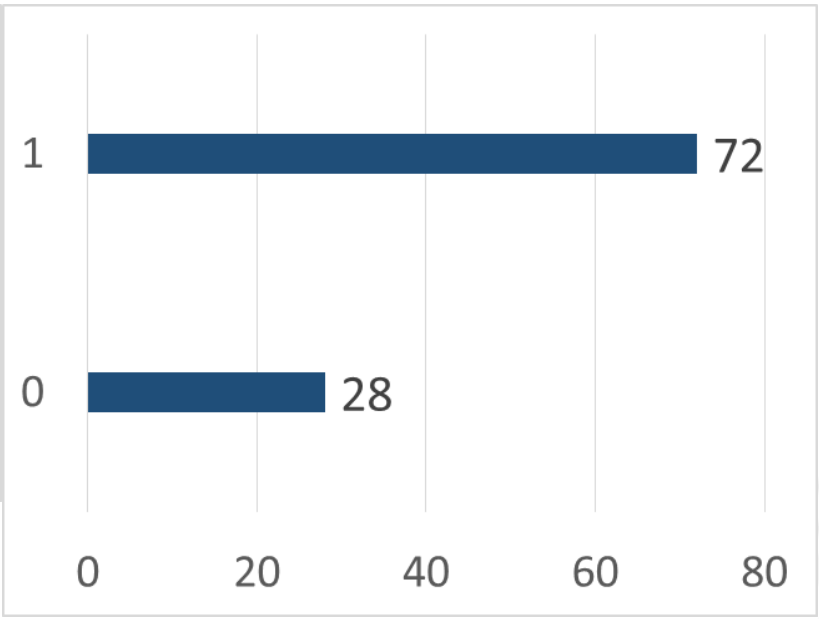
CSE 7315c

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Next 100 calls handled by an agent at a call centre



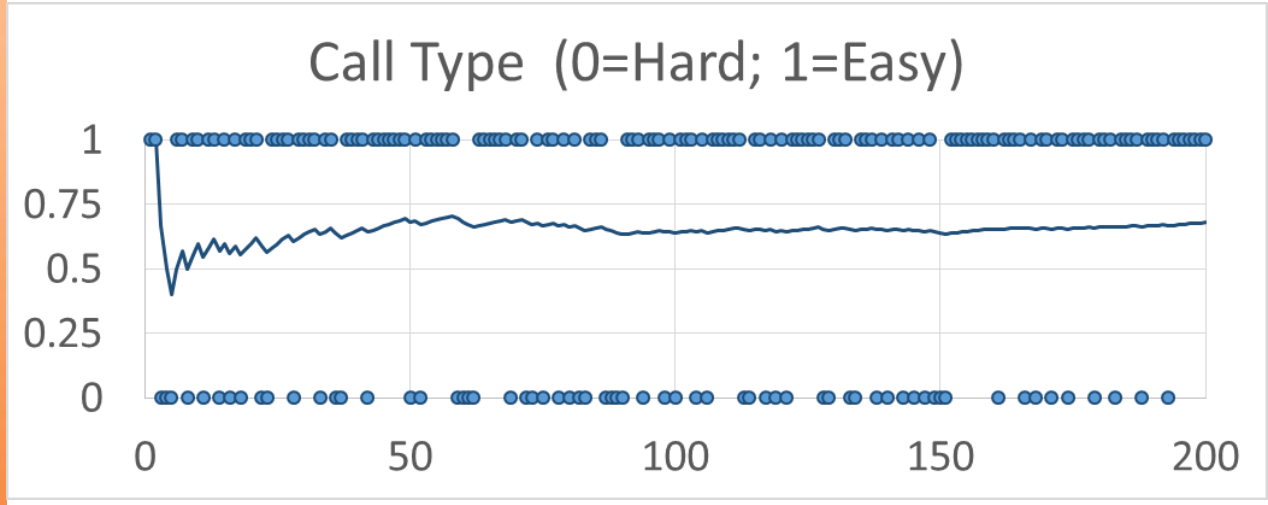


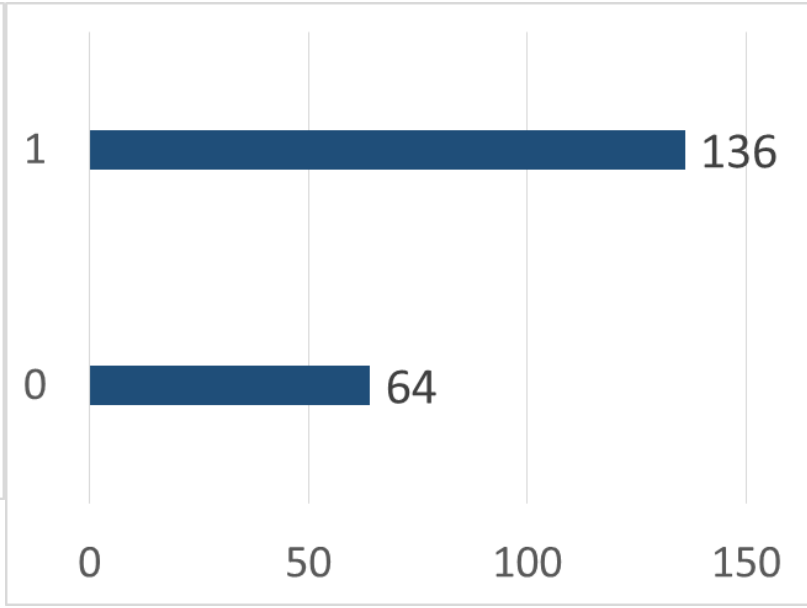
CSE 7315c

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Averages over the long run





P(easy) = 0.7

Assigning Probabilities

Subjective Method

Based on feelings, insights, knowledge, etc. of a person.

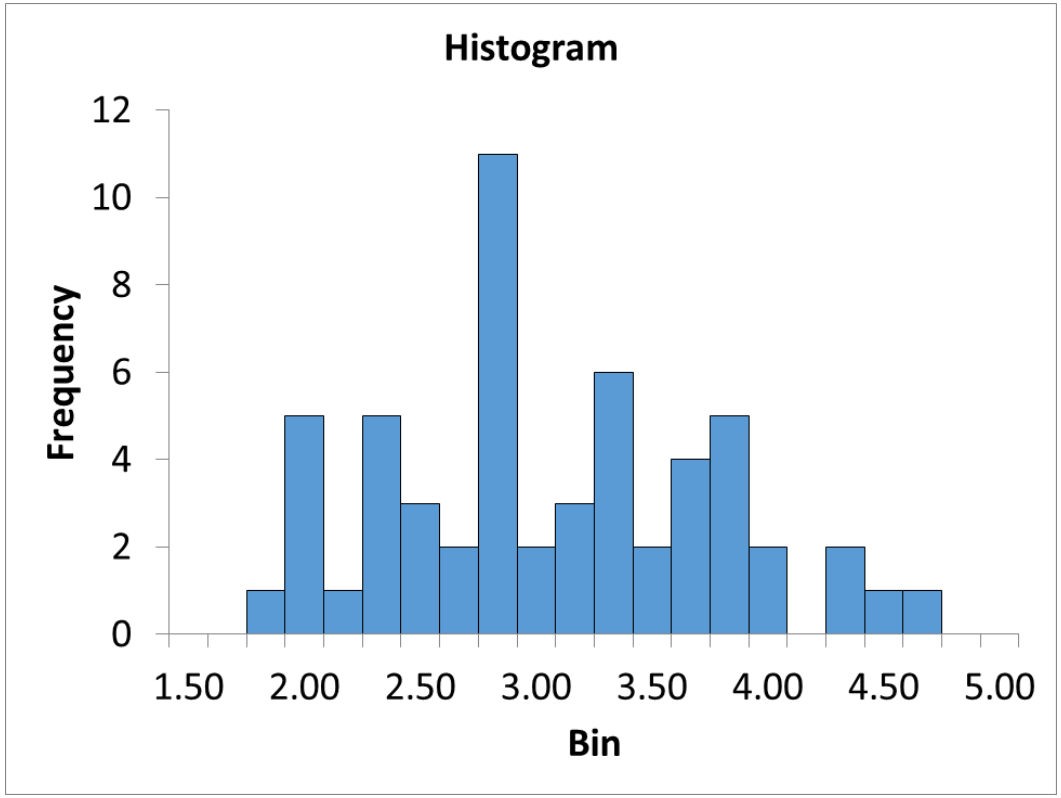
What is the probability of India winning the match tomorrow?

Assigning Probabilities

Subjective Method

2010 rates of growth in US GDP anticipated by 56 economists at the start of 2010.

Does it mean probability of GDP growing by more than 4% is 6/56 = 0.11?



Actual growth 2.5%

Data from: http://projects.wsj.com/econforecast/#ind=gdpa&r=10&e=75 and http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG

Probability - Terminology

Sample Space – Set of all possible outcomes, denoted S.

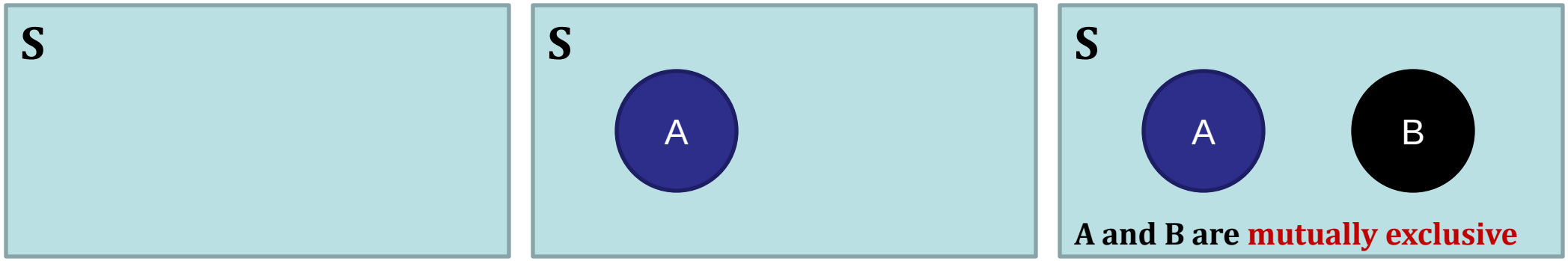
Example:

After 2 coin tosses, the set of all possible outcomes are {HH, HT, TH, TT}

Event – A subset of the sample space.

An Event of interest might be - HH

Probability - Rules



$$P(S) = 1$$

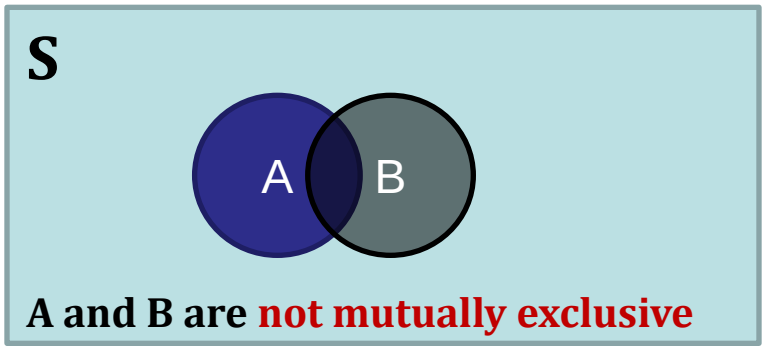
$$0 \leq P(A) \leq 1$$

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

Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.

Mutually Exclusive – If event A happens, event B cannot.

Probability - Rules



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

Example

Event A – Customers who default on loans

Event B – Customers who are High Net Worth Individuals

Probability - Rules

Independent Events – Outcome of event B is not dependent on the outcome of event A.

Probability of customer B defaulting on the loan is not dependent on default (or otherwise) by customer A.

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

If the probability of getting an *easy* call is 0.7, what is the probability that the next 3 calls will be *easy*?

$$P(easy_1 \text{ and } easy_2 \text{ and } easy_3) = 0.7^3 = 0.343$$

Probability - Question

A basketball team is down by 2 points with only a few seconds remaining in the game. Given that:

- Chance of making a 2-point shot to tie the game = 50%
- Chance of winning in overtime = 50%
- Chance of making a 3-point shot to win the game = 30%

What should the coach do: go for 2-point or 3-point shot?

What are the assumptions, if any?

Probability - Types

Contingency table summarizing 2 variables, *Loan Default* and *Age*:

		Age			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**No**	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Probability - Types

Convert it into probabilities:

		Age			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**No**	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability - Types

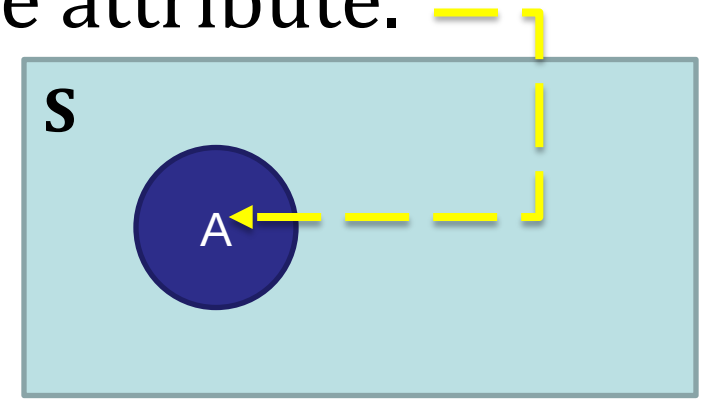
Marginal Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability describing a single attribute.

P(No) = 0.816

P(Old) = 0.008



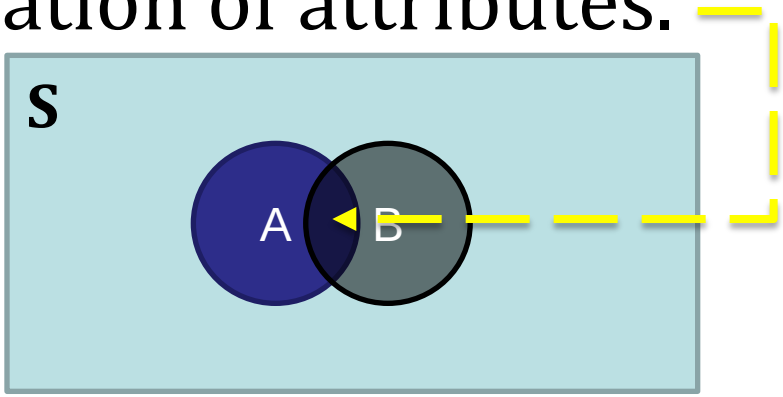
Probability - Types

Joint Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability describing a combination of attributes.

P(Yes **and** Young) = 0.077

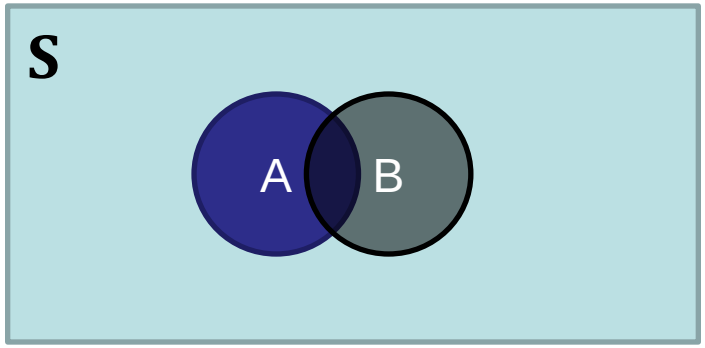


Probability - Types

Union Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

P(Yes **or** Young) = P(Yes) + P (Young) – P(Yes and Young) = 0.184 + 0.302 – 0.077 = 0.409



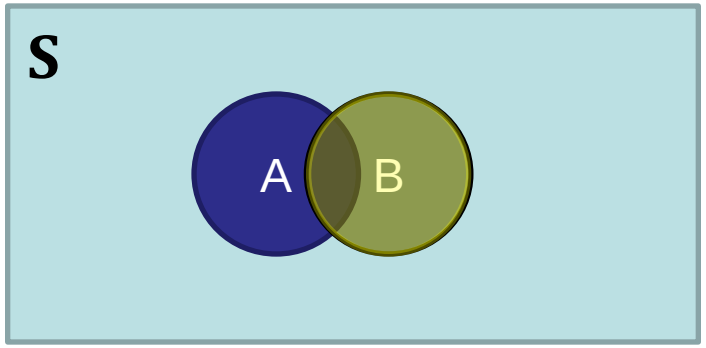
Probability - Types

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

Probability of *A* occurring **given that** *B* has occurred.

The sample space is restricted to a single row or column. This makes rest of the sample space irrelevant.



Probability - Types

Conditional Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

What is the probability that a person will not default on the loan payment **given** she is middle-aged?

$$P(\text{No} \mid \text{Middle-Aged}) = 0.586/0.690 = 0.85$$

Note that this is the ratio of **Joint Probability** to **Marginal Probability**, i.e., $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

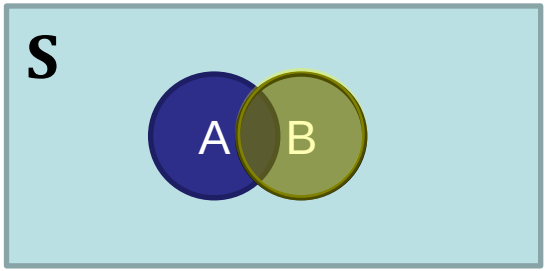
$$P(\text{Middle-Aged} \mid \text{No}) = 0.586/0.816 = 0.72 \text{ (Order Matters)}$$

Probability - Types

Conditional Probability – Visualizing using Probability Tables and Venn Diagrams

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

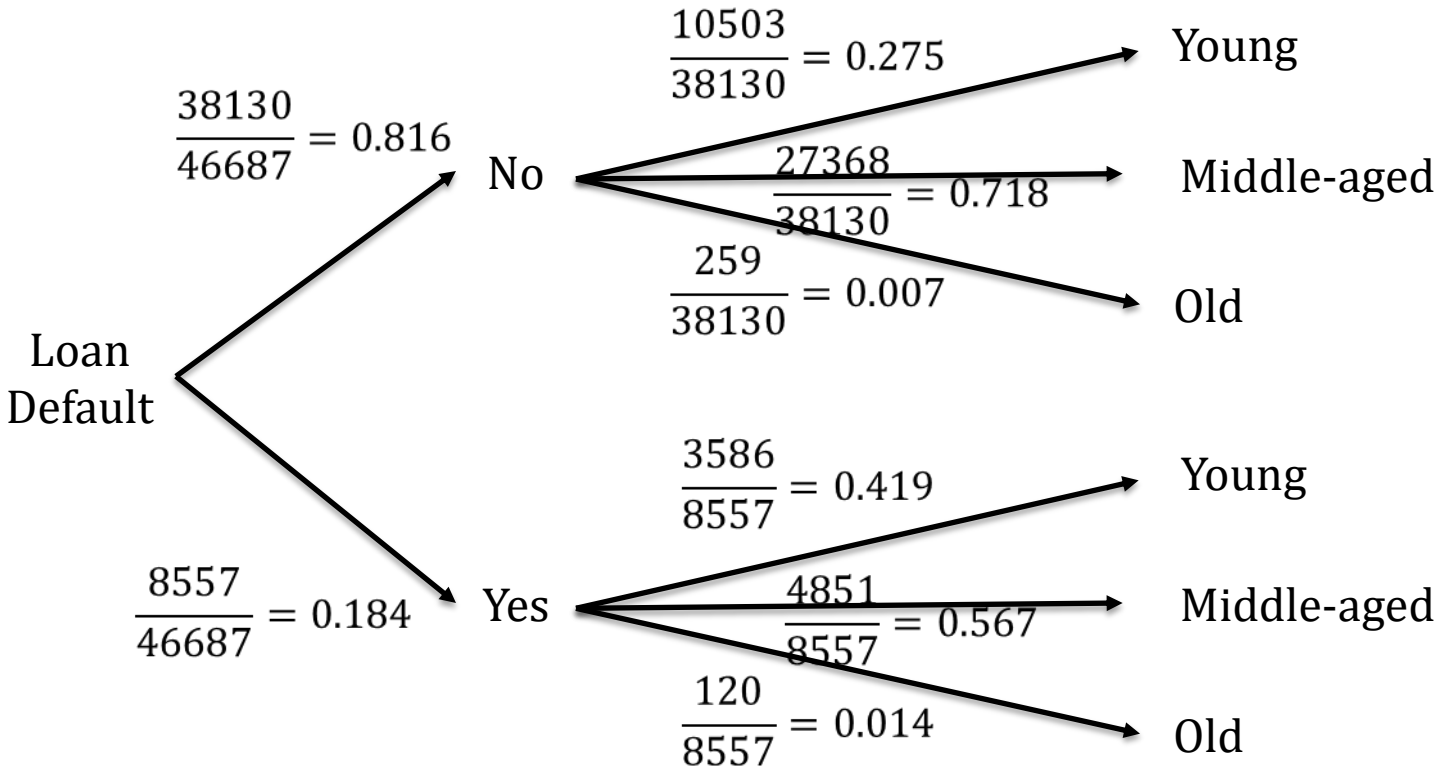
		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000



Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Numbers)				Age (Probabilities)			
		Young	Middle-aged	Old	Total	Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130	0.225	0.586	0.005	0.816
	Yes	3,586	4,851	120	8,557	0.077	0.104	0.003	0.184
	Total	14,089	32,219	379	46,687	0.302	0.690	0.008	1.000



Find

- P(Old and Yes)
- P(Yes and Old)
- P(Old)
- P(Yes)
- P(Old | Yes)
- P(Yes | Old)
- P(Young | No)

CSE 7315c

Probability - Types

Attention Check

Identify the type of probability in each of the below cases:

1. P(Old and Yes)
2. P(Yes and Old)
3. P(Old)
4. P(Yes)
5. P(Old | Yes)
6. P(Yes | Old)
7. P(Young | No)
8. P(Middle-aged or No)
9. P(Old or Young)

		Age (Probabilities)			
		Young	**Middle-aged**	**Old**	**Total**
Loan Default	**No**	0.225	0.586	0.005	**0.816**
	Yes	0.077	0.104	0.003	**0.184**
	Total	**0.302**	**0.690**	**0.008**	**1.000**

1 and 2: **Joint**; 3 and 4: **Marginal**; 5, 6 and 7: **Conditional**; 8 and 9: **Union**

Probability - Types

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \Rightarrow P(A \text{ and } B) = P(B) * P(A|B)$$

Similarly

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \Rightarrow P(A \text{ and } B) = P(A) * P(B|A)$$

Equating, we get

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

$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Probability - Types

Conditional Probability

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

Older people make up only 1.4% of the loan defaulters. Now, given that the probability that someone defaults on a loan is 0.184, find the probability that a older person defaults on the loan. Older people make up only 0.8% of the clientele.
P(Yes | Old) = ?

$$P(Yes|Old) = \frac{P(Yes) * P(Old|Yes)}{P(Old)}$$

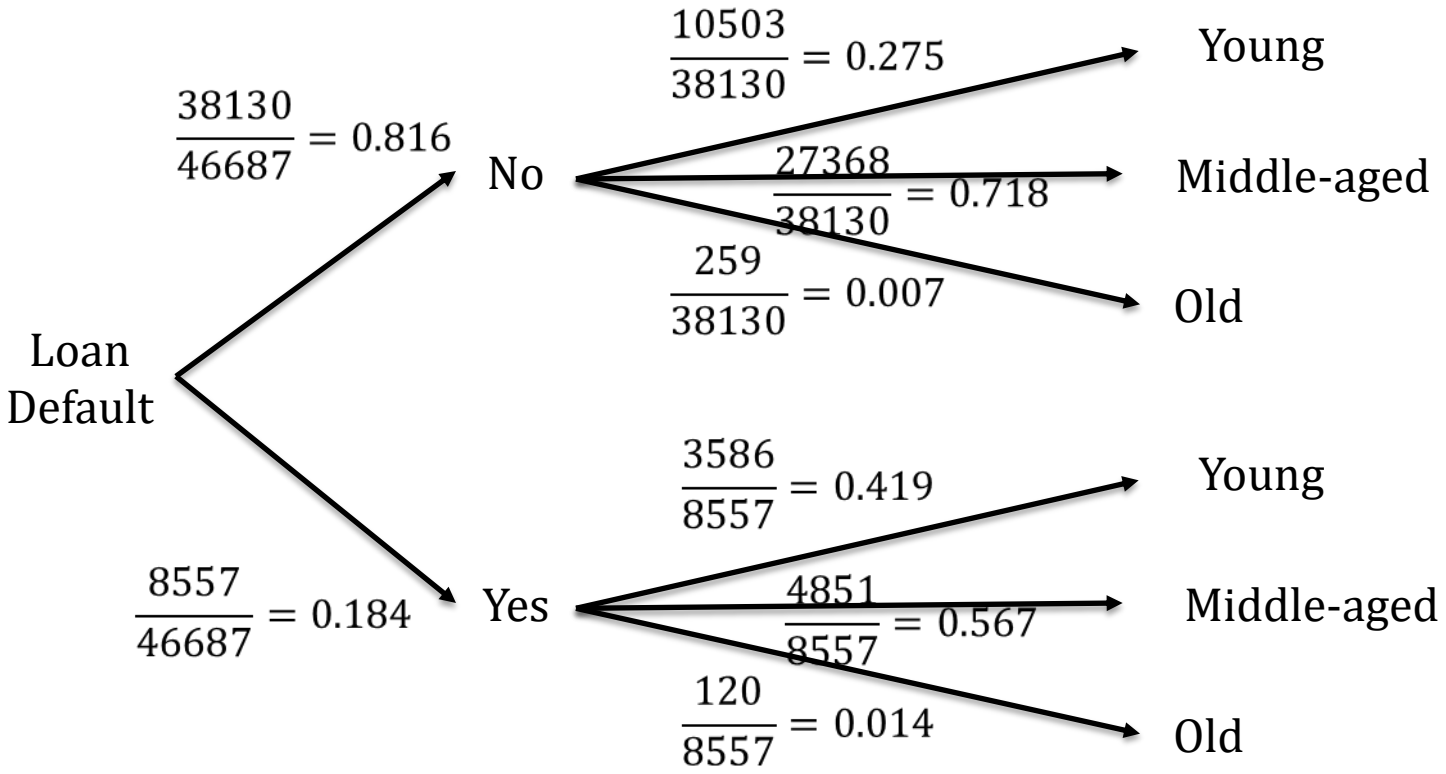
$$P(Yes|Old) = \frac{0.184*0.014}{0.008} = 0.32$$

Probability - Types

Conditional Probability – Visualizing using Probability Trees

		Age (Probabilities)			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

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

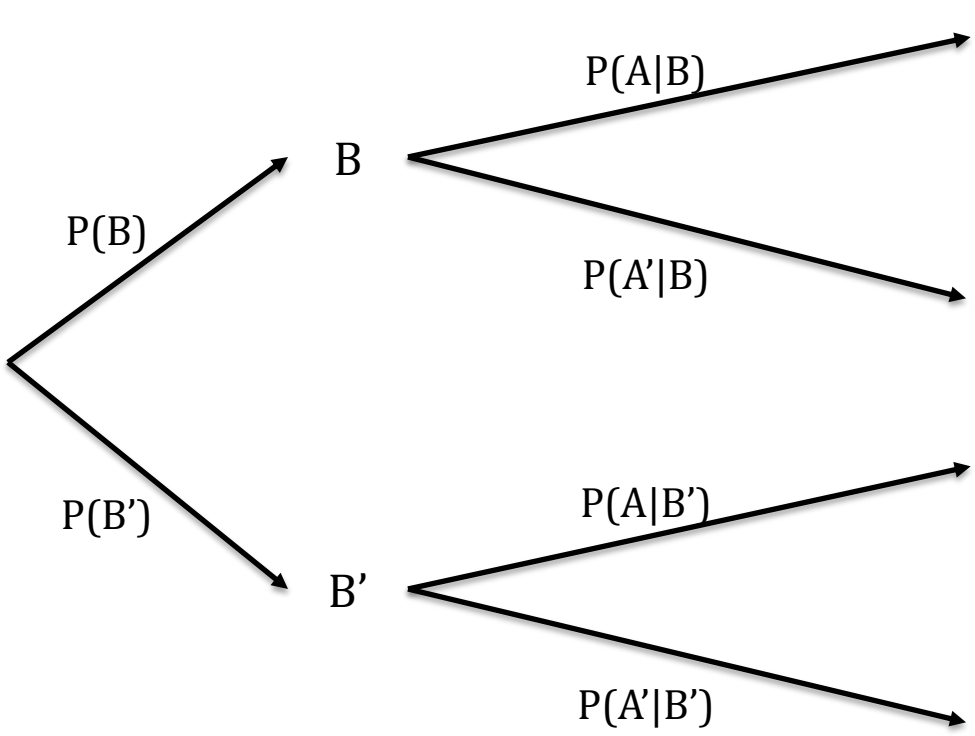


Now find P(Yes | Old)

CSE 7315c

Probability - Types

Generalized Probability Tree



P(B and A) = P(B) * P(A|B)

P(B and A') = P(B) * P(A'|B)

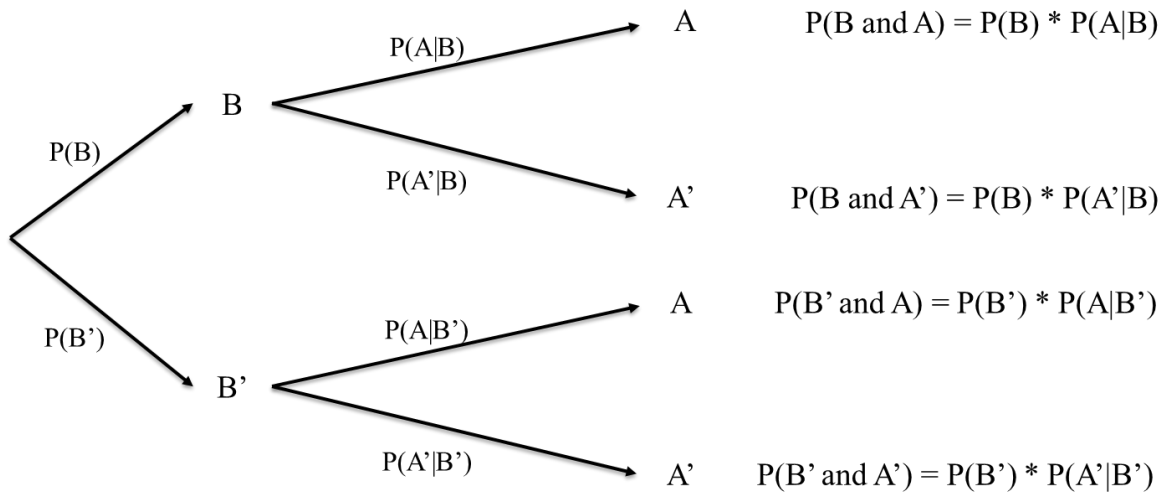
P(B' and A) = P(B') * P(A|B')

P(B' and A') = P(B') * P(A'|B')

State each probability in English; note B' means "not B".

Probability - Types

Conditional Probability -> Bayes' Theorem



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

Note B' means "not B"

Bayes' Theorem

Bayes' Theorem allows you to find reverse probabilities, and to allow **revision of original probabilities** with new information.

Case – Clinical trials

Epidemiologists claim that probability of breast cancer among Caucasian women in their mid-50s is 0.005. An established test identified people who had breast cancer and those that were healthy. A new mammography test in clinical trials has a probability of 0.85 for detecting cancer correctly. In women without breast cancer, it has a chance of 0.925 for a negative result. If a 55-year-old Caucasian woman tests positive for breast cancer, what is the probability that she in fact has breast cancer?

Bayes' Theorem

Case – Clinical trials

P(Cancer) = 0.005

P(Test positive | Cancer) = 0.85 (*aka* Prior Probability)

P(Test negative | No cancer) = 0.925

P(Cancer | Test positive) = ? (*aka* Posterior or Revised Probability)

$$P(Cancer|Test+) = \frac{P(Cancer) * P(Test+|Cancer)}{P(Test+|Cancer) * P(Cancer) + P(Test+|No\ cancer) * P(No\ cancer)}$$

$$= \frac{0.005 * 0.85}{0.85 * 0.005 + 0.075 * 0.995} = \frac{0.00425}{0.078875} = 0.054$$

Homework

Draw a Probability Table and a Probability Tree for the above case.

Bayes' Theorem

Case – Spam filtering

Apache SpamAssassin™

Latest News

2015-04-30: SpamAssassin 3.4.1 has been released! Highlights include:

- improved automation to help combat spammers that are abusing new top level dc
- tweaks to the SPF support to block more spoofed emails;
- increased character set normalization to make rules easier to develop and stop sp
- continued refinement to the native IPv6 support; and
- improved Bayesian classification with better debugging and attachment hashing.

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word "free" appears in 20% of the mails marked as spam, i.e., P(Free | Spam) = 0.20. Assuming 0.1% of non-spam mail includes the word "free" and 50% of all mails received by the user are spam, find the probability that a mail is spam if the word "free" appears in it.

Bayes' Theorem

Case – Spam filtering

P(Spam) = 0.50

P(Free | Spam) = 0.20 (*aka* Prior Probability)

P(Free | No spam) = 0.001

P(Spam | Free) = ? (*aka* Posterior or Revised Probability)

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

$$= \frac{0.5 * 0.2}{0.2 * 0.5 + 0.001 * 0.5} = \frac{0.1}{0.1005} = 0.995$$

This helps the spam filter automatically classify the messages as spam.

A slight detour

HOW GOOD IS YOUR CLASSIFICATION?

CSE 7315c

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

$$Recall\ (Sensitivity) = \frac{952}{1478} = 0.644$$

$$Precision = \frac{952}{1119} = 0.851$$

$$Accuracy = \frac{952 + 3025}{952 + 3025 + 526 + 167} = \frac{3977}{4670} = 0.852$$

$$Specificity = \frac{3025}{3025 + 167} = \frac{3025}{3192} = 0.948$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * 0.851 * 0.644}{0.851 + 0.644} = \frac{1.096}{1.495} = 0.733$$

Which measure(s) is/are more important?

CSE 7315c

Confusion Matrix

Breast cancer detection		Predicted		Total
		Positive	Negative	
Actual	Positive	852	126	978
	Negative	67	1025	1092
Total		919	1151	2070

$$Recall\ (Sensitivity) = \frac{852}{978} = 0.871$$

$$Precision = \frac{852}{919} = 0.927$$

$$Accuracy = \frac{852 + 1025}{852 + 1025 + 126 + 67} = \frac{1877}{2070} = 0.907$$

$$Specificity = \frac{1025}{1025 + 67} = \frac{1025}{1092} = 0.939$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * 0.871 * 0.927}{0.871 + 0.927} = \frac{1.615}{1.798} = 0.898$$

Which measure(s) is/are more important?

Confusion Matrix – Recent Interview Question

You have been tasked to build a classifier for cancer diagnosis. It is of high importance that patients with cancer can be diagnosed wrongly as negative but patients without cancer should NEVER be diagnosed as positive.

Which of the following classification models would you prefer?
(Assuming: Positives = Cancer, Negatives = Not cancer)

Options:

- True Positive Rate [which is = True Positive / Actual Positive]
- True Negative Rate [which is = True Negative / Actual Negative]
- Precision [which is = True Positive / Predicted Positive]
- Total Accuracy [which is = (True Positive + True Negative) / Total Population]

Analyzing attributes

PROBABILITY DISTRIBUTIONS

Histogram

A series of contiguous rectangles that represent the frequency of data in given class intervals.

How many class intervals?

Rule of thumb: 5-15 (not too many and not too few)

Freedman-Diaconis rule:

$$No.of\ bins = \frac{(max - min)}{2 * IQR * n^{\frac{-1}{3}}},$$

$$where\ the\ denominator\ is\ the\ bin - width$$

Histogram - Excel

Annual traffic data for 30 busiest airports in the world – 2013 and 2011

Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final and http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final

Last accessed: February 04, 2016

1	Passenger Traffic 2011 FINAL (Annual)			
2	Last Update: 8 July 2013			
3	Passenger Traffic			
4	Total passengers enplaned and deplaned, passengers in transit counted once			
5	Rank	City (Airport)	Total Passengers	% Change
6	1	ATLANTA GA, US (ATL)	92389023	3.5
7	2	BEIJING, CN (PEK)	78675058	6.4
8	3	LONDON, GB (LHR)	69433565	5.4
9	4	CHICAGO IL, US (ORD)	66701241	-0.1
10	5	TOKYO, JP (HND)	62584826	-2.5
11	6	LOS ANGELES CA, US (LAX)	61862052	4.7
12	7	PARIS, FR (CDG)	60970551	4.8
13	8	DALLAS/FORT WORTH TX, US (DFW)	57832495	1.6
14	9	FRANKFURT, DE (FRA)	56436255	6.5
15	10	HONG KONG, HK (HKG)	53328613	5.9
16	11	DENVER CO, US (DEN)	52849132	1.7
17	12	JAKARTA, ID (CGK)	51533187	16.2
18	13	DUBAI, AE (DXB)	50977960	8
19	14	AMSTERDAM, NL (AMS)	49755252	10
20	15	MADRID, ES (MAD)	49653055	-0.4
21	16	BANGKOK, TH (BKK)	47910904	12
22	17	NEW YORK NY, US (JFK)	47644060	2.4
23	18	SINGAPORE, SG (SIN)	46543845	10.7
24	19	GUANGZHOU, CN (CAN)	45040340	9.9
25	20	SHANGHAI, CN (PVG)	41447730	2.1
26	21	SAN FRANCISCO CA, US (SFO)	40927786	4.3
27	22	PHOENIX AZ, US (PHX)	40591948	5.3
28	23	LAS VEGAS NV, US (LAS)	40560285	2
29	24	HOUSTON TX, US (IAH)	40128953	-0.9
30	25	CHARLOTTE NC, US (CLT)	39043708	2.1
31	26	MIAMI FL, US (MIA)	38314389	7.3
32	27	MUNICH, DE (MUC)	37763701	8.8
33	28	KUALA LUMPUR, MY (KUL)	37704510	10.6
34	29	ROME, IT (FCO)	37651222	3.9
35	30	ISTANBUL, TR (IST)	37406025	16.3

Passenger Traffic 2013 FINAL (Annual)

Last Update: 22 December 2014

Passenger Traffic

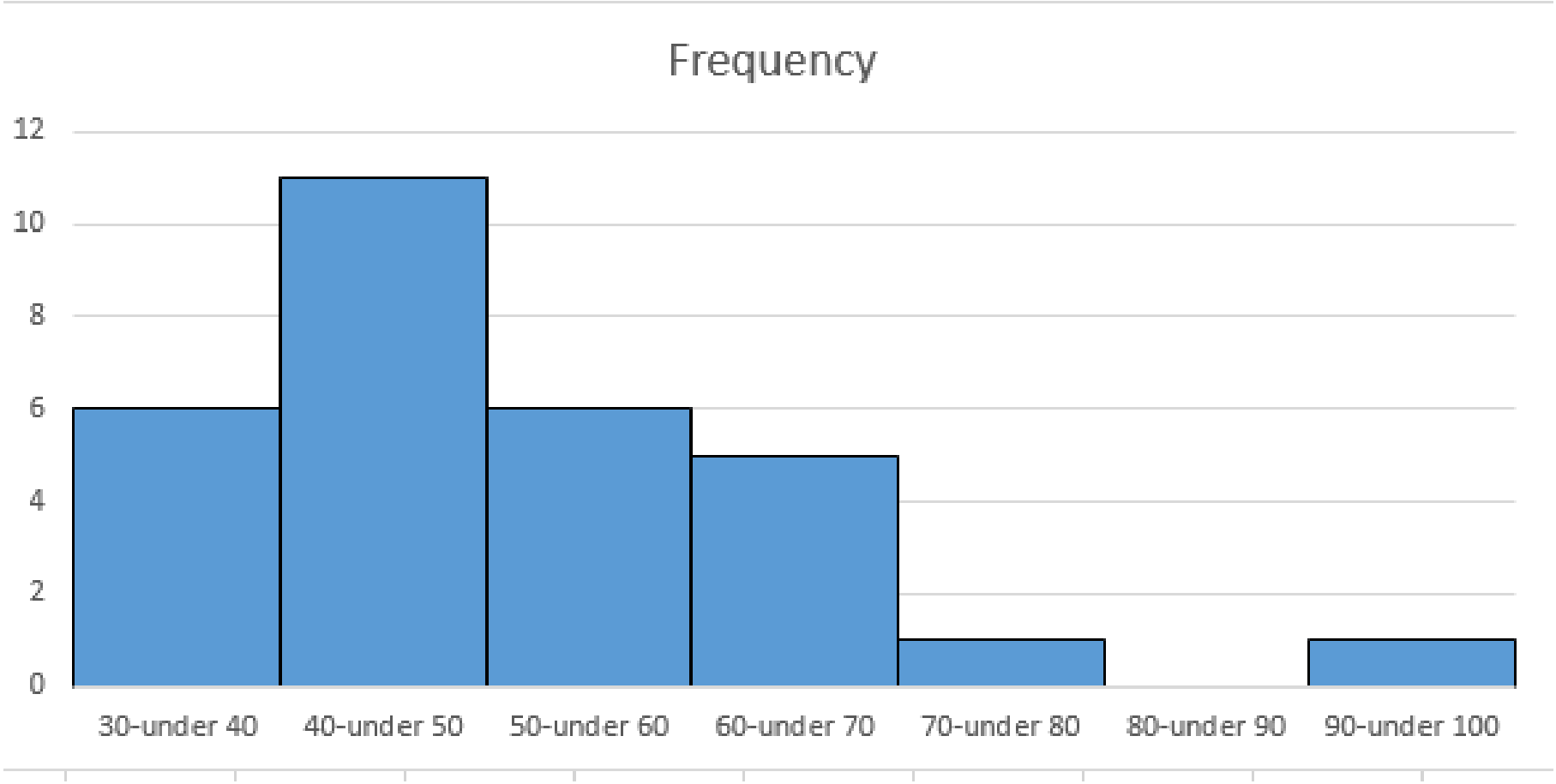
Total passengers enplaned and deplaned, passengers in transit counted once

Rank	City (Airport)	Passengers 2013	Passengers 2012	% Change
1	ATLANTA GA, US (ATL)	9,44,31,224	9,55,13,828	-1.1
2	BEIJING, CN (PEK)	8,37,12,355	8,19,29,359	2.2
3	LONDON, GB (LHR)	7,23,68,061	7,00,38,804	3.3
4	TOKYO, JP (HND)	6,89,06,509	6,67,95,178	3.2
5	CHICAGO IL, US (ORD)	6,67,77,161	6,66,29,600	0.2
6	LOS ANGELES CA, US (LAX)	6,66,67,619	6,36,88,121	4.7
7	DUBAI, AE (DXB)	6,64,31,533	5,76,84,550	15.2
8	PARIS, FR (CDG)	6,20,52,917	6,16,11,934	0.7
9	DALLAS/FORT WORTH TX, US (DFW)	6,04,70,507	5,86,20,160	3.2
10	JAKARTA, ID (CGK)	6,01,37,347	5,77,72,864	4.1
11	HONG KONG, HK (HKG)	5,95,88,081	5,60,61,595	6.3
12	FRANKFURT, DE (FRA)	5,80,36,948	5,75,20,001	0.9
13	SINGAPORE, SG (SIN)	5,37,26,087	5,11,81,804	5
14	AMSTERDAM, NL (AMS)	5,25,69,200	5,10,35,590	3
15	DENVER CO, US (DEN)	5,25,56,359	5,31,56,278	-1.1
16	GUANGZHOU, CN (CAN)	5,24,50,262	4,83,09,410	8.6
17	BANGKOK, TH (BKK)	5,13,63,451	5,30,02,328	-3.1
18	ISTANBUL, TR (IST)	5,13,04,654	4,51,23,758	13.7
19	NEW YORK NY, US (JFK)	5,04,23,765	4,92,91,765	2.3
20	KUALA LUMPUR, MY (KUL)	4,74,98,127	3,98,87,866	19.1
21	SHANGHAI, CN (PVG)	4,71,89,849	4,48,80,164	5.1
22	SAN FRANCISCO CA, US (SFO)	4,49,45,760	4,43,99,885	1.2
23	CHARLOTTE NC, US (CLT)	4,34,57,471	4,12,28,372	5.4
24	INCHEON, KR (ICN)	4,16,79,758	3,91,54,375	6.4
25	LAS VEGAS NV, US (LAS)	4,09,33,037	4,07,99,830	0.3
26	MIAMI FL, US (MIA)	4,05,62,948	3,94,67,444	2.8
27	PHOENIX AZ, US (PHX)	4,03,41,614	4,04,48,932	-0.3
28	HOUSTON TX, US (IAH)	3,97,99,414	3,98,91,444	-0.2
29	MADRID, ES (MAD)	3,97,17,850	4,51,76,978	-12.1
30	MUNICH, DE (MUC)	3,86,72,644	3,83,60,604	0.8

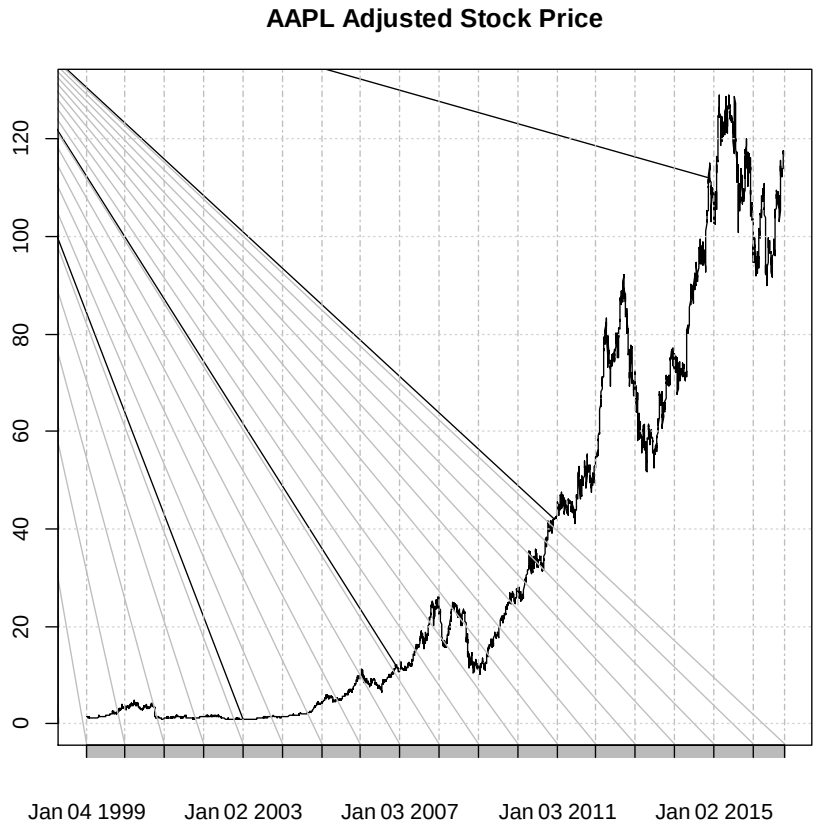
Histogram

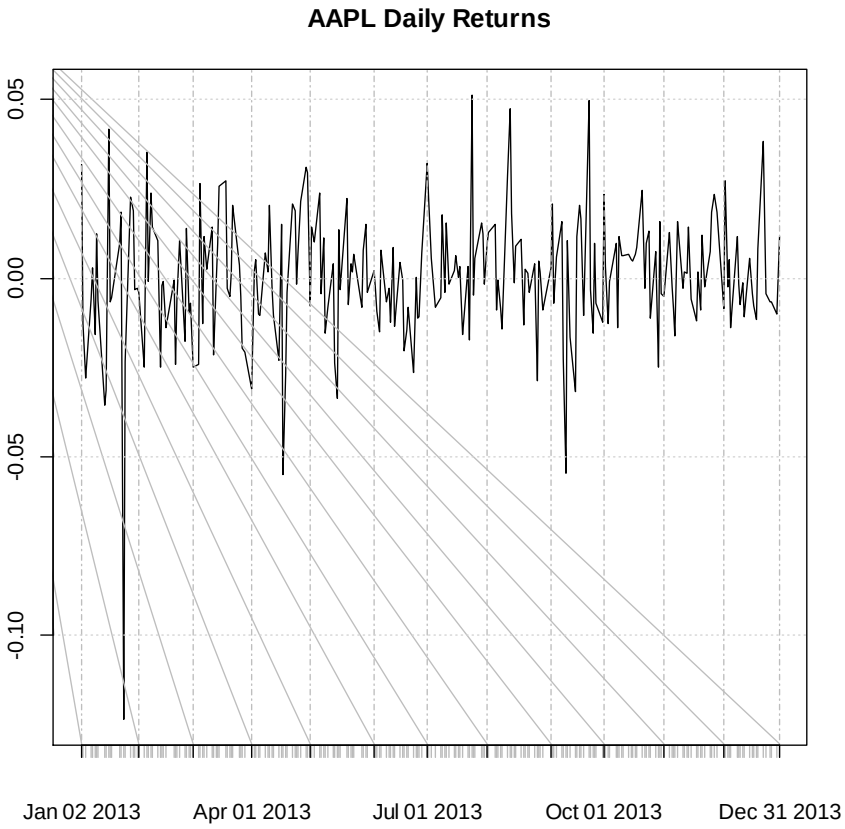
Annual traffic data for 30 busiest airports in the world – 2011

Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final
Last accessed: November 22, 2014



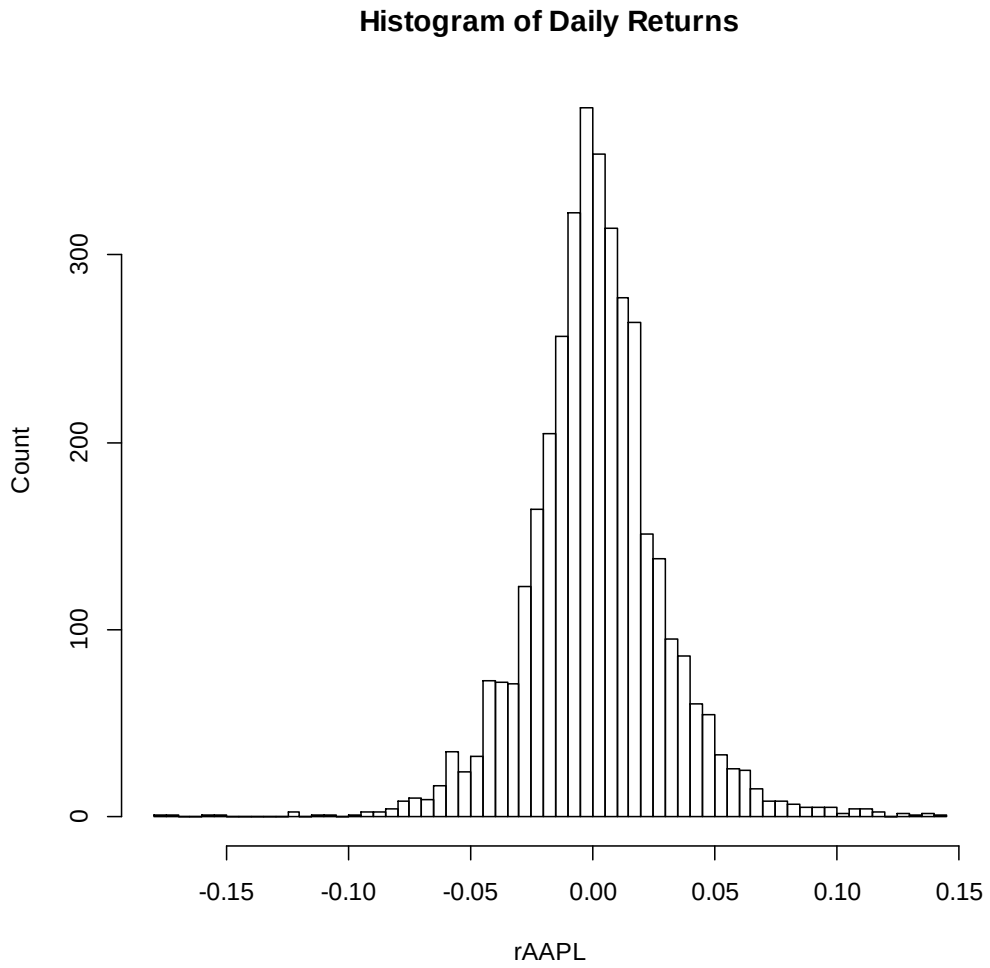
Stock Returns





CSE 7315c

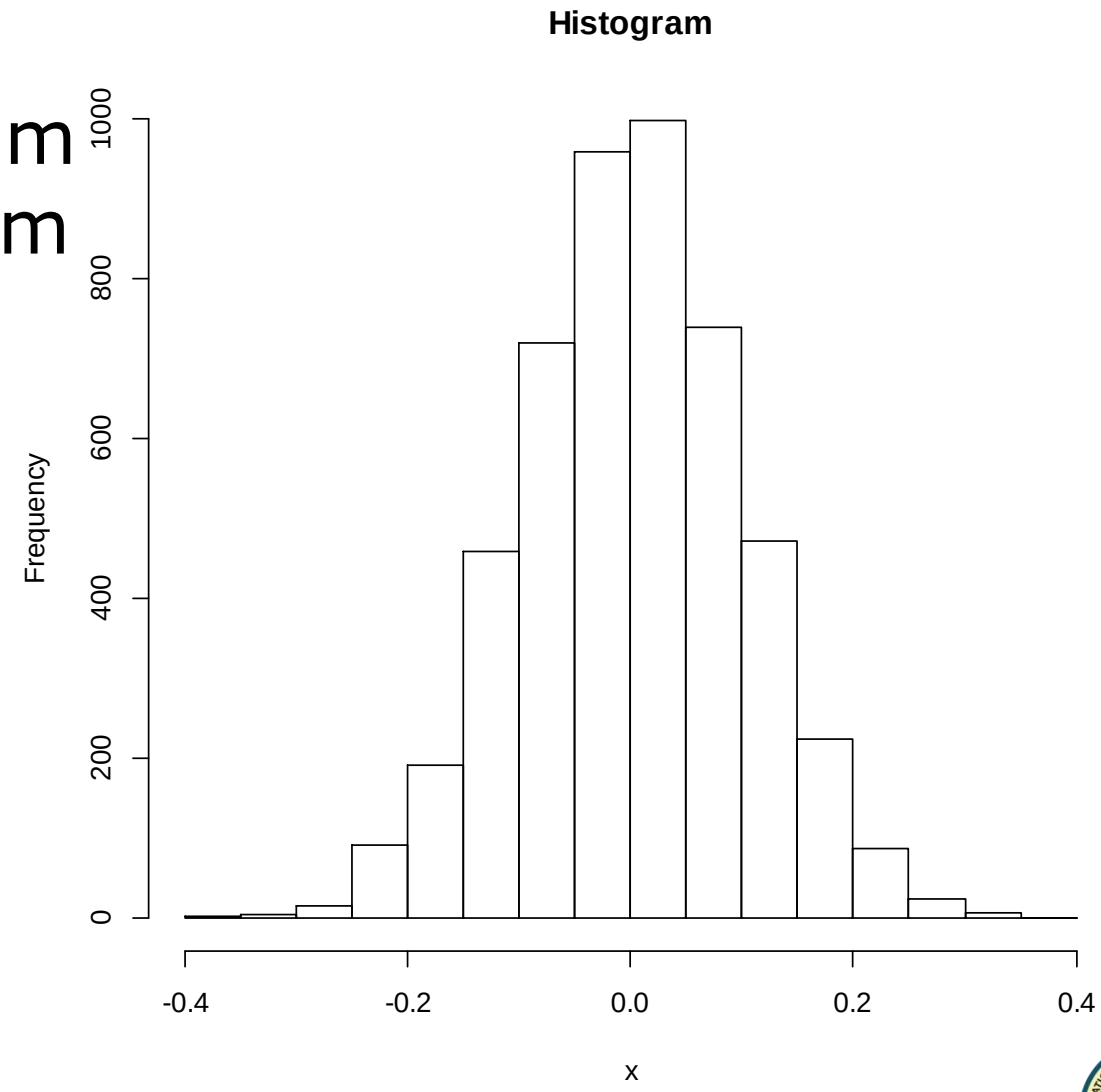
Histogram of Stock Returns



CSE 7315c

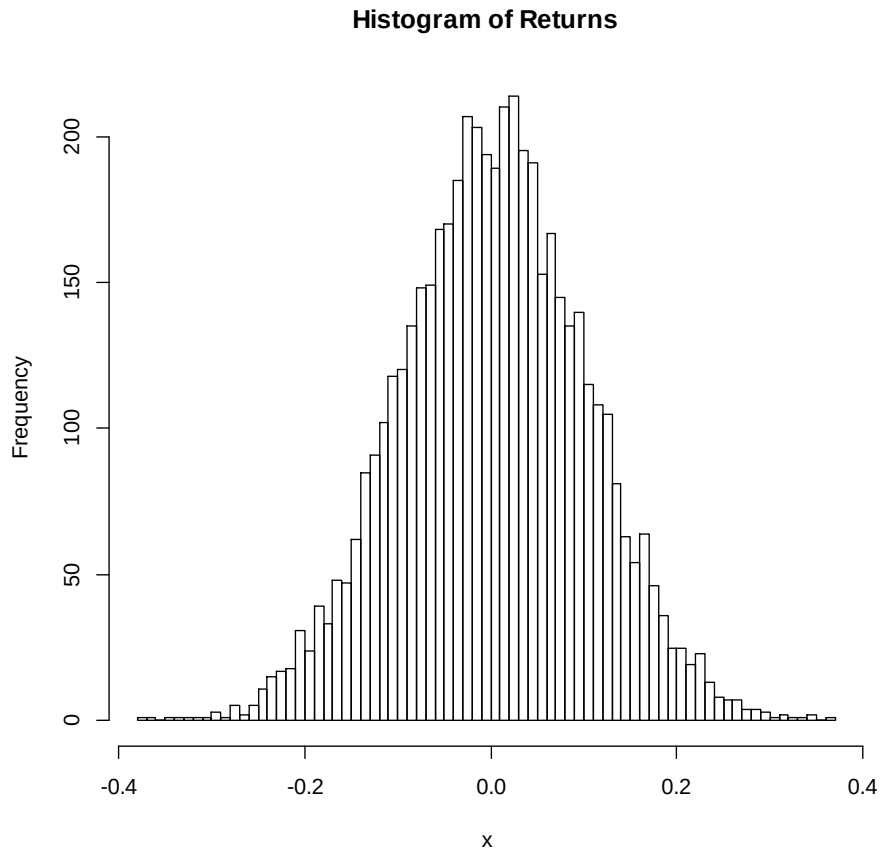
Histogram of Stock Returns

- Consider a histogram of stock returns from 5000 days



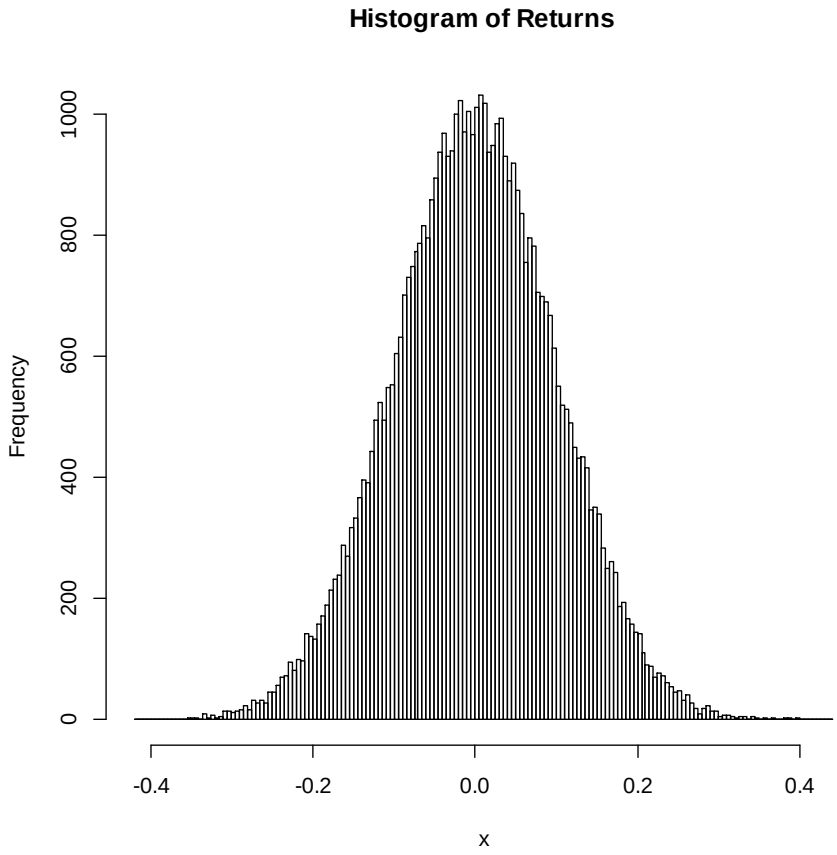
Histogram of Stock Returns

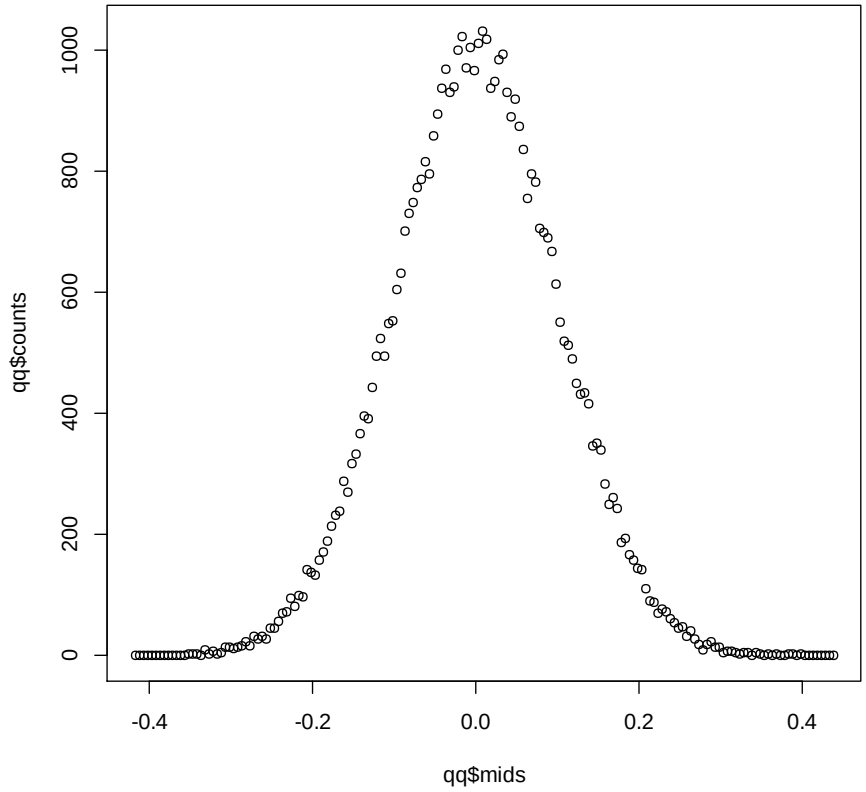
- The same histogram with larger number of bins



Histogram of Stock Returns

50000 data points with 200 bins

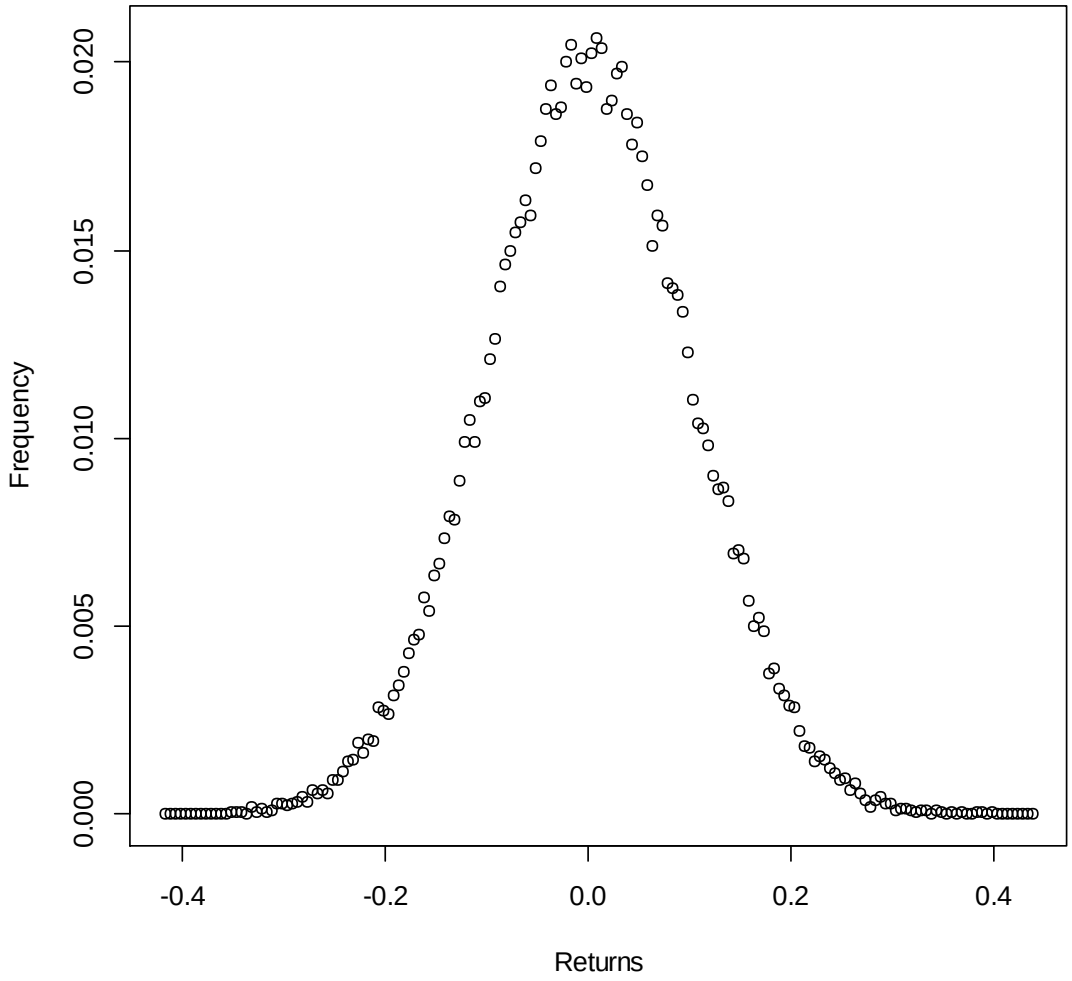




CSE 7315c

Histogram / Probability Distribution Function

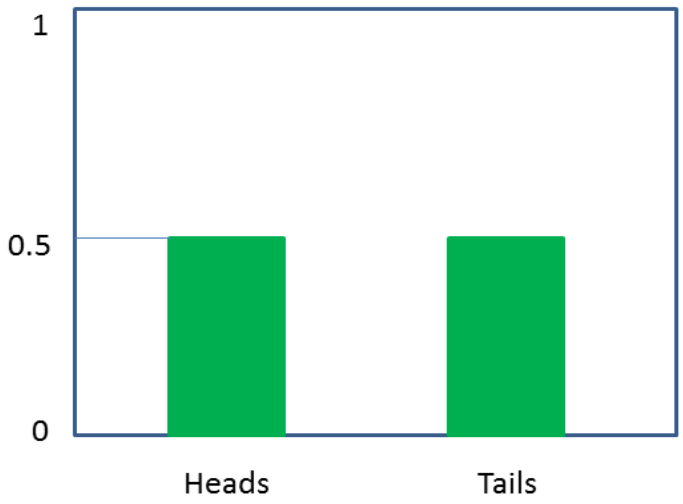
Convert the counts to frequency by dividing by 50000



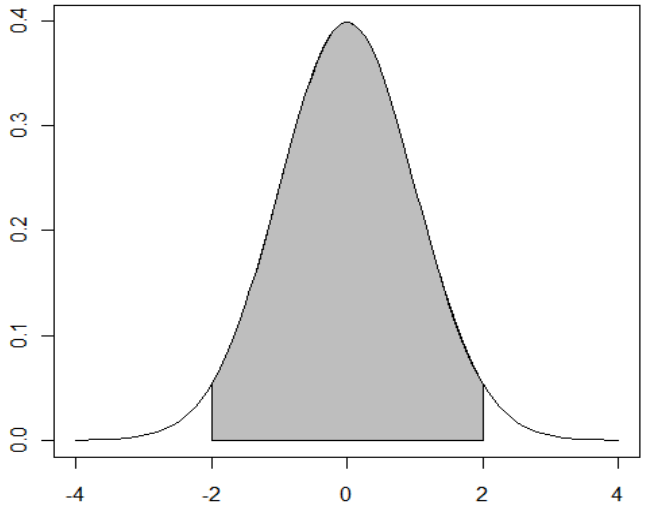
Random variable

- A variable that can take multiple values with different probabilities.
- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

Discrete and Continuous



Countable



Measurable

CSE 7315c

Can any function be a probability distribution?

Discrete Distributions	Continuous Distributions
Probability that X can take a specific value x is $P(X = x) = p(x)$.	Probability that X is between two points a and b is $P(a \leq X \leq b) = \int_a^b f(x)dx$.
It is non-negative for all real x.	It is non-negative for all real x.
The sum of $p(x)$ over all possible values of x is 1, i.e., $\sum p(x) = 1$.	$\int_{-\infty}^{\infty} f(x)dx = 1$
Probability Mass Function	**Probability Density Function**

Possible Outcome	$	Cherry	Lemon	Other
Probability of Outcome	0.1	0.2	0.2	0.5

Cost: $1 for each game

Winning combinations:

 = $20

 = $15 (any order)

 = $10

 = $5

CSE 7315c

Probability of Winnings Combinations

Possible Outcome	$	Cherry	Lemon	Other
Probability of Outcome	0.1	0.2	0.2	0.5

Probability of Winning combinations:

= 0.1*0.1*0.1 =0.001

= 3*(0.1*0.1*0.2)=0.006

= 0.2 * 0.2 * 0.2 = 0.008

= 0.2* 0.2* 0.2 = 0.008

No win probability= ?

=1 – (Win something)
= 1- (0.001+0.006 +0.008 +0.008)

=0.977

CSE 7315c

Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	-$1	$4	$9	$14	$19

Cost: $1 for each game

Winning combinations:


= $20


= $15 (any order)


= $10


= $5

CSE 7315c

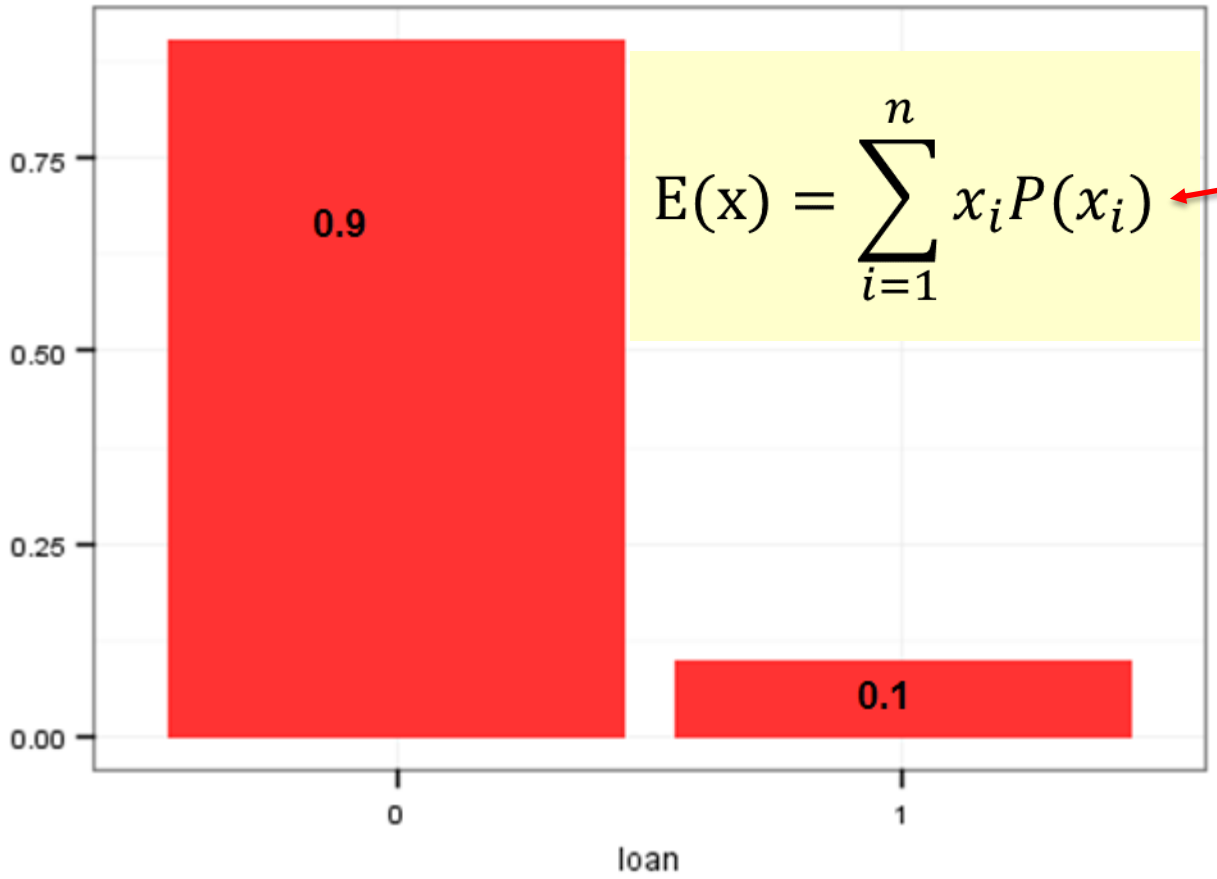
Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	-$1	$4	$9	$14	$19

Why do you need a probability distribution?

Once a distribution is calculated, it can be used to determine the EXPECTED outcome.

Expectation: Discrete



$$E(x) = \sum_{i=1}^{n} x_i P(x_i)$$

Recall anything like this?

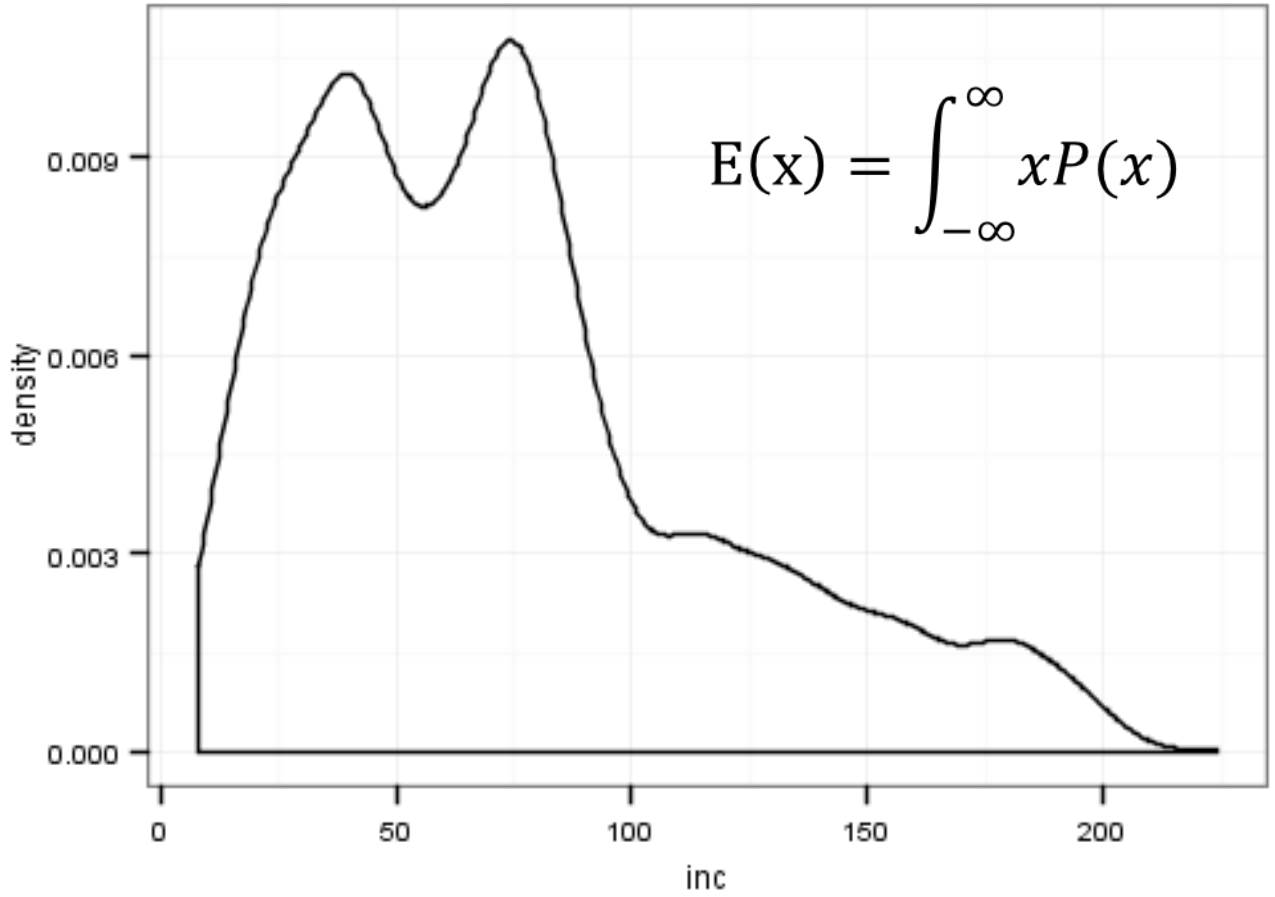
Yoga class composition

Age (years)	13	15	17
Frequency, f	1	3	2

$$\text{Mean, } \mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{13X1+15X3+17X2}{1+3+2} = 13 * \frac{1}{6} + 15 * \frac{3}{6} + 17 * \frac{2}{6} = 15.3$$

Recall Assigning Probabilities using Empirical or Frequentist Method

Expectation: Continuous



Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
P(X=x)	0.977	0.008	0.008	0.006	0.001
x	-$1	$4	$9	$14	$19

EXPECTATION, $\mathrm{E}(X) = \mu = \Sigma x P(X = x)$

E(X) = (-$1)*0.977 + $4*0.008 + $9*0.008 + $14*0.006 + $19*0.001
= -$0.77

This is the amount of $ expected to be "gained" on each pull of the lever.

So, why play?

It never makes sense to play the Slot machine or the Lottery.

Until it does!

Massachusetts State Lottery

match all 6 numbers	1 in 9.3 million	variable jackpot
match 5 of 6	1 in 39,000	$4,000
match 4 of 6	1 in 800	$150
match 3 of 6	1 in 47	$5
match 2 of 6	1 in 6.8	free lottery ticket

Cost of the ticket = $2

Jackpot value = Atleast $1Million

It does not make sense to play!

E(x)= ($1 million / 9.3 million) + ($4,000 / 39,000) + ($150 / 800) + ($5 / 47) + ($2 / 6.8) = 79.8 cents.

Massachusetts State Lottery

- RollDay - When the Jackpot increases to $2M, then prize money for Match 5 also increases

Prize	Chance of winning	Expected number of winners	Roll-down allocation	Roll-down per prize
match 5 of 6	1 in 39,000	12	$600,000	$50,000
match 4 of 6	1 in 800	587	$1.4m	$2,385
match 3 of 6	1 in 47	10,000	$600,000	$60

Expected value on the roll day changes dramatically.
E(x) = $5.53

See: http://www.theatlantic.com/business/archive/2016/02/how-mit-students-gamed-the-lottery/470349/

Variance of the Distribution

- The Width/Spread of the distribution
- **VARIANCE**, $Var(X) = E(X - \mu)^2 = \Sigma(x - \mu)^2 P(X = x)$
- $\sigma = \sqrt{Var(X)}$

Simplifying the Formula

$$E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2]$$

$$= E[X^2] - 2\mu E[X] + \mu^2$$ (we get this from previous formula as m is just a number)

$$= E[X^2] - 2\mu^2 + \mu^2$$

$$= E[X^2] - \mu^2 = E[X^2] - [E(X)]^2$$

Expectation Properties

E(X+Y) = E(X) + E(Y) e.g., Playing a game each on 2 slot machines with different probabilities of winning. This is called Independent Observation.

E(aX+b) = aE(X)+E(b) = aE(X) + b e.g., values x have been changed. This is called Linear Transformation.

* Not all central tendencies posses this nice property

Variance Properties

- Var(X+a) = Var(X) (Variance does not change when a constant is added)
- Var(X+Y) = Var(X) + Var(Y) for Independent Observations
- Var(X-Y) = Var(X) + Var(Y)

Variance Properties

$Var(aX) = a^2 Var(X)$ for Linear Transformation

Say, $Y = aX$

$E(Y) = a\,E(X)$ (from the previous set of relations)
$Y - E(Y) = a(X - E(X))$

Squaring both sides and taking expectations
$E(Y-E(Y))^2 = a^2\,E(X-E(X))^2$

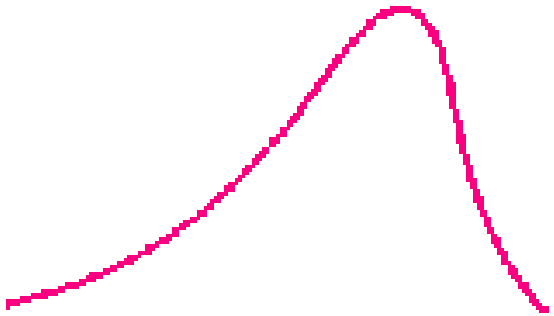
However, the left hand side is Variance of Y and RHS is Variance of X

$Var(Y) = a^2 Var(X)$ or $Var(aX) = a^2 Var(X)$

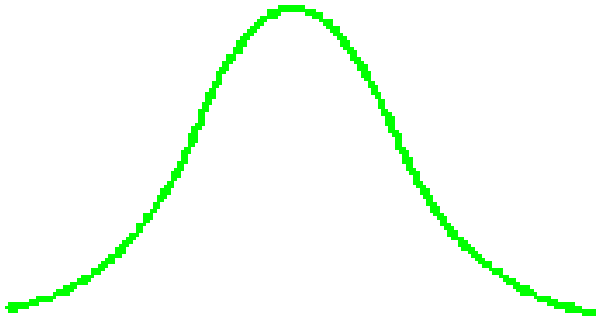
Understanding the shape of a PDF - Skewness

- A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means median is less than mean.

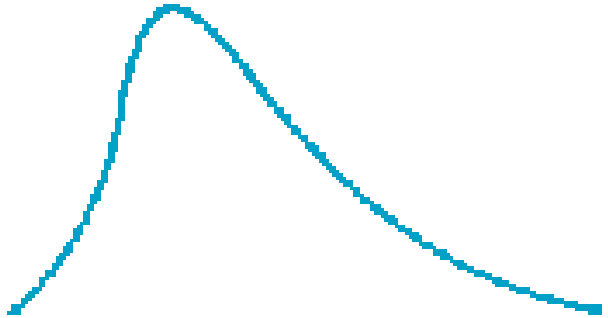
Understanding the shape of a PDF - Skewness



Negatively (left) skewed distribution



Normal distribution



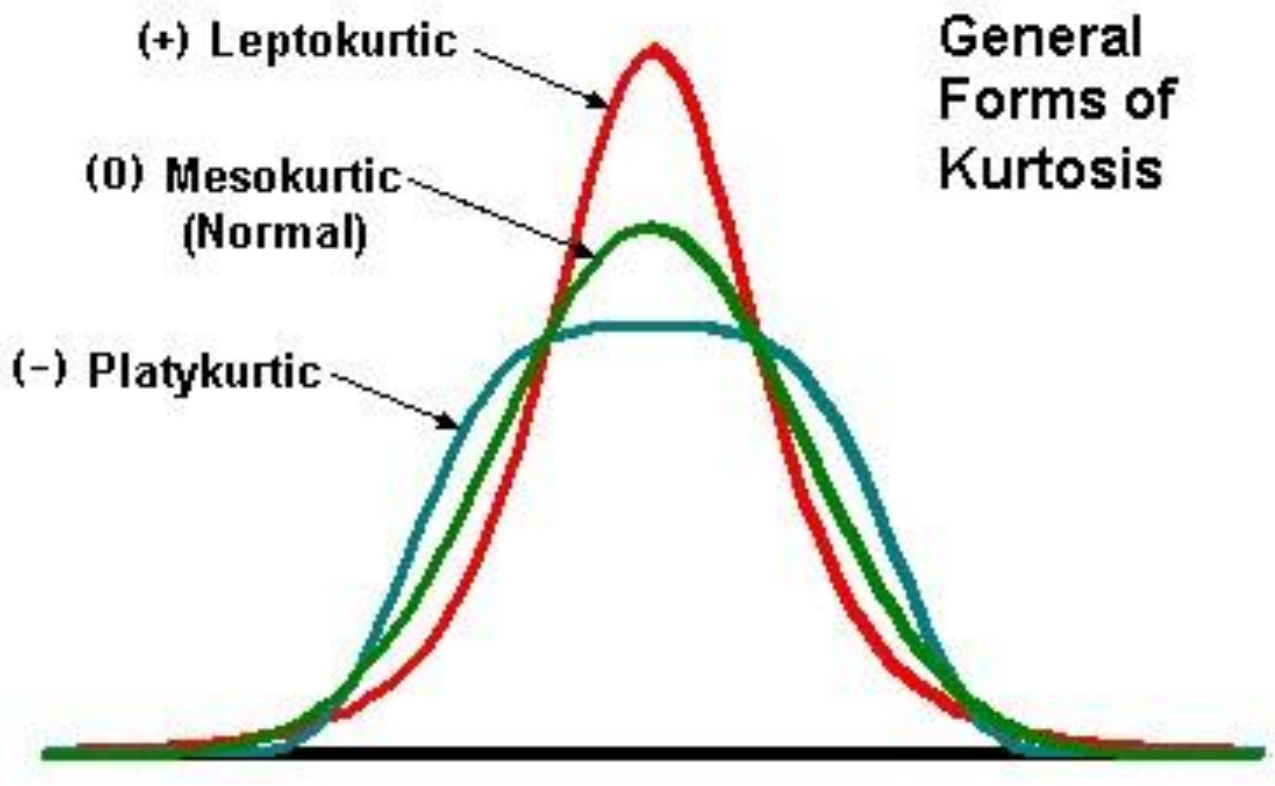
Positively (right) skewed distribution

CSE 7315c

Understanding the shape of a PDF - Kurtosis

- A measure of the ‘peaked’ness of the data distribution. Negative kurtosis means a flat distribution. Positive kurtosis means a peaked distribution.

Understanding the shape of a PDF - Kurtosis



Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Guide to Airline Fees in India

cleartrip

	Change fee (Domestic)	Change fee (International)	Cancellation fee (Domestic)	Cancellation fee (International)	No show charges (Domestic)	No show charges (International)
Indigo	Rs 1000 / passenger / sector	Rs 1,850 / passenger / sector	Rs 1,000 / passenger / sector	Rs 1,850 / passenger / sector	No refund	No refund
Jet Airways	Rs 250 - 997 (Premiere) Rs 500 - 1050 (Economy)	Rs 5,500 to NIL (depending on fare class)	Rs 500 - 997 (Premiere) Rs 750 - 1,050 (Economy)	Rs 8,000 to NIL (depending on fare class)	Rs 1,500 to NO REFUND (depending on fare class)	Rs 8,000 to NIL (depending on fare class)
JetKonnect	Rs 250 - 997 (Premiere) Rs 500 - 1050 (Economy)	NA	Rs 500 - 997 (Premiere) Rs 750 - 1,050 (Economy)	NA	Rs 1,500 to NO REFUND (depending on fare class)	NA
Spicejet	Rs 950 / passenger / sector	Rs 1,750 / passenger / sector	Rs 950 / passenger / sector	Rs 1,750 / passenger / sector	No refund	No refund
GoAir	Rs 950 (GoSmart) NIL (GoFlexi & GoBusiness)	NA	Rs 950 (GoSmart) Rs 350 (GoFlexi) NIL (GoBusiness, >24 hrs) Rs 750 (GoBusiness, <24 hrs)	NA	12 month credit shell for PSF + service tax	NA
Air India	Rs 750 - NIL (Economy, based on fare class); NIL (Executive / First Class)	Rs 5,000 - NIL (Economy) Rs 7,500 - NIL (Executive) Rs 5,000 - NIL (First class)	Rs 500 to NO REFUND (Economy) Rs 200 (Executive / First)	No refund (Economy Web Specials) Rs 5,000 - NIL (Economy) Rs 14,000 - NIL (Executive) Rs 5,000 - NIL (First class) + Rs 300 Refund Administration fee (all classes)	Rs 1,500 to NO REFUND (Economy); Rs 200 (Executive / First class)	Rs 5,000 - NIL (Economy) Rs 14,000 - NIL (Executive) Rs 5,000 - NIL (First class) + Rs 300 Refund Administration fee (all classes)
Kingfisher	Rs 950 (Kingfisher Red); Rs 500-950 (Kingfisher, Kingfisher First)	NA	Rs 950 (Kingfisher Red) Rs 500 - 100% of Base Fare (Kingfisher, Kingfisher First)	NA	NO REFUND (Kingfisher Red, Kingfisher); Rs 1,000 + Cancellation / change fee (Kingfisher First)	NA

Data sourced from airline websites, accurate as of 18 September 2012.
Always check fare rules before booking. Visit airline website for more details.

3

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Kingfisher Airlines* would like to maximize revenues by ensuring no empty seats on its flight between Bengaluru and Hyderabad. They intentionally wish to overbook the flights based on the historical data of no-shows on this sector.

You have been hired as a statistical consultant to help formulate a solution.

*currently not in business

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

The frequency distribution of "No-Shows" from 200 randomly selected flights on this sector is:

# of No-Shows	1	2	3	4	5	6	Total
Frequency	70	40	10	20	20	40	200

What is your advice for Kingfisher on the number of seats they should overbook on this sector?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

What is the Random Variable in this problem?

Random variable, X is the # of No-Shows.

What is the PMF for the frequency distribution seen in the sample?

# of No-Shows	1	2	3	4	5	6	Total
Frequency	70	40	10	20	20	40	200

X	1	2	3	4	5	6
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

What is the Expectation?

X	1	2	3	4	5	6
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 1 * 0.35 + 2 * 0.20 + 3 * 0.05 + 4 * 0.10 + 5 * 0.10 + 6 * 0.20 = 3$$

So, you'd advise Kingfisher to overbook 3 seats on this sector, which is the **mean** of the data in the sample.

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 1: Kingfisher tells you that it will pay you Rs 500 for your consulting and Rs 1500 as bonus for each correct prediction (prediction must be exactly correct, no more no less). Will you still go with the **mean**?

X	1	2	3	4	5	6
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 1 * 0.35 + 2 * 0.20 + 3 * 0.05 + 4 * 0.10 + 5 * 0.10 + 6 * 0.20 = 3$$

So, will you advise Kingfisher to overbook 3 seats on this sector, which is the **mean** of the data in the sample?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 1

What is the probability distribution of your earnings if you went with the expected value (or the mean)?

X (Your earnings)	500	500	2000	500	500	500
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 500 * (0.35 + 0.20 + 0.10 + 0.10 + 0.20) + 2000 * 0.05 = Rs\ 575$$

How much would you earn in other cases?

Would you still stick to Mean or switch to Median or Mode?

CSE 7315c

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 2

Instead of a binary state for your earnings, if Kingfisher offers to pay you Rs 2000 for the consulting minus Rs 125 for each under or overbooked seat, what will be your advice now?

X (Your earnings)	2000	1875	1750	1625	1500	1375
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X)$$
$$= 2000 * 0.35 + 1875 * 0.20 + 1750 * 0.05 + 1625 * 0.10 + 1500 * 0.10 + 1375 * 0.20$$
$$= Rs\ 1750$$

How much would you earn in other cases?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 3

Instead of penalizing based on absolute magnitude of the prediction error, if Kingfisher offers to pay you Rs 2500 for the consulting minus Rs 75 times the square of the prediction error (penalizing larger errors more), what will be your advice now?

X (Your earnings)	2500	2425	2200	1825	1300	625
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X)$$
$$= 2500 * 0.35 + 2425 * 0.20 + 2200 * 0.05 + 1825 * 0.10 + 1300 * 0.10 + 625 * 0.20$$
$$= Rs\ 1907.50$$

How much would you earn in other cases?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Conclusion

For the same dataset, depending on the business problem, Mode was the best option in Scenario 1, Median in Scenario 2 and Mean in Scenario 3.

Moral of the story

- You should look at data carefully in the context of the business domain and problem.
- You must inculcate statistical way of thinking in all you do.
- Statistics don't lie; Statisticians may.
- In God we Trust; all others must bring data.

Useful Resources

- Conditional probability explained visually https://www.khanacademy.org/video/conditional-probability2
- Bayes Theorem : https://youtu.be/E4rlJ82CUZI

- Creating a histogram: https://www.khanacademy.org/video/histograms-intro

- Probability Distribution Functions
- https://www.khanacademy.org/video/discrete-probability-distribution
- https://www.khanacademy.org/video/probability-density-functions

International School of Engineering

Plot 63/A, Floors 1&2, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering