Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Basic Probability Concepts, Probability Distributions

Dr. Sridhar Pappu
Executive VP – Academics, INSOFE

December 10, 2017


FLASHBACK
I REMEMBER IT LIKE IT WAS YESTERDAY...
CMCOTT. 04/13/12 #138

MAXIMUM SECURITY

$50.7 BILLION SPENT FOR DEFENCE DEVELOPMENT IN 2016 PLACES INDIA AMONG WORLD'S TOP FIVE DEFENCE SPENDERS

INDIA IS ahead of Saudi Arabia and Russia's expenditure

THE US, China and the UK remain the top three defence spenders ahead of India's fourth place

$46.6 bn **INDIA SPENT** $46.6 billion last year, as per a report released on Monday

THE REPORT said that India is set to overtake UK's budget by 2018

$1.6 trillion The worldwide outlook shows that global defence spending rose by 1 per cent to $1.6 trillion this year, against 0.6 per cent in 2015.

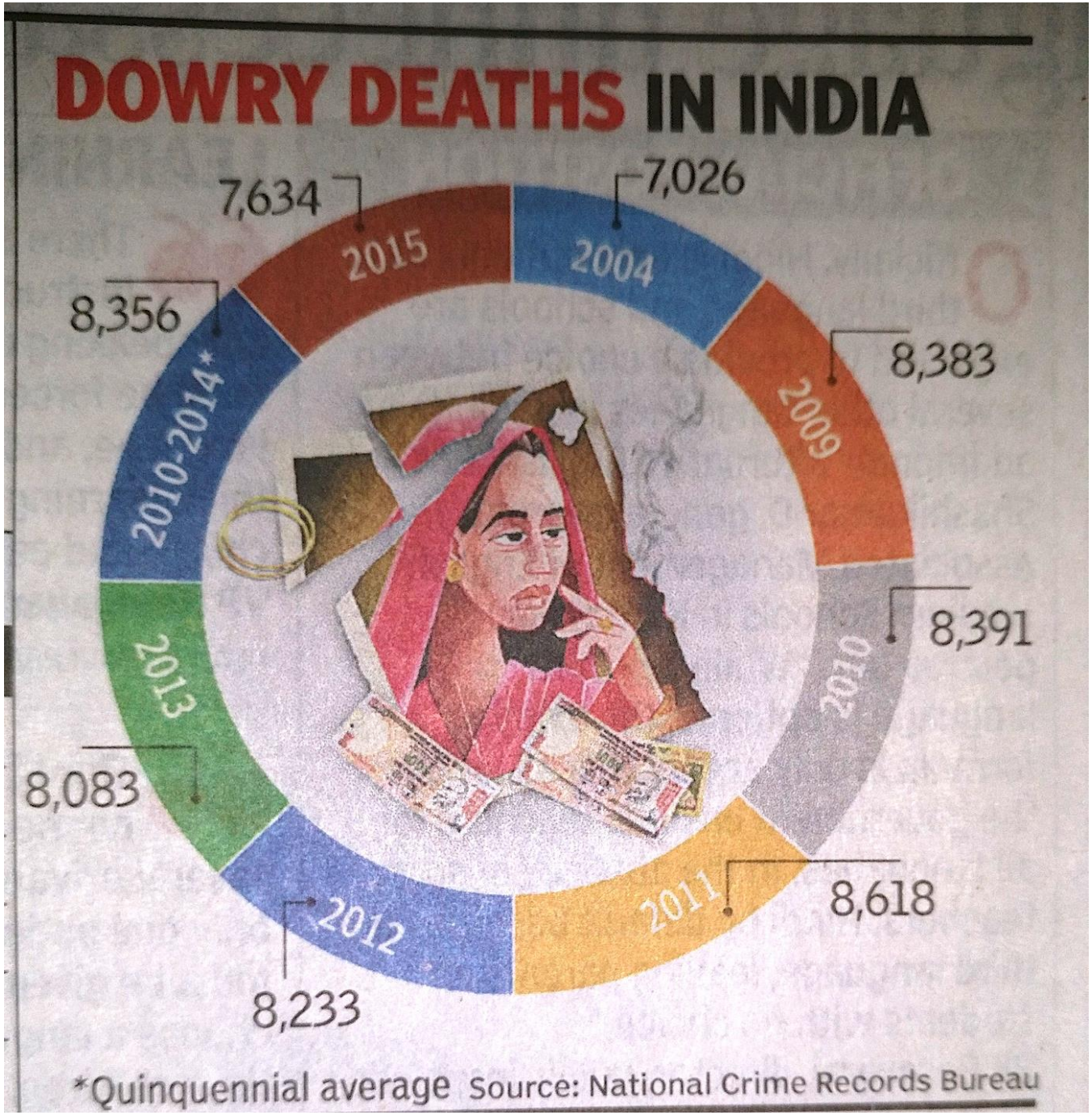
DEFENCE EXPENDITURE

US $622 bn | China $191.7 bn | UK $5.8 bn

Saudi $48.68bn

Russia $48.44 bn

Over the next three years, India will re-emerge as a key growth market for defence suppliers – Craig Caffrey, principal analyst for Asia-Pacific at 'IHS Janes'


DOWRY DEATHS IN INDIA
7,634
2015
7,026
2004
8,356
2010-2014*
8,383
2009
8,391
2010
8,083
2013
8,233
2012
2011
8,618
*Quinquennial average
Source: National Crime Records Bureau

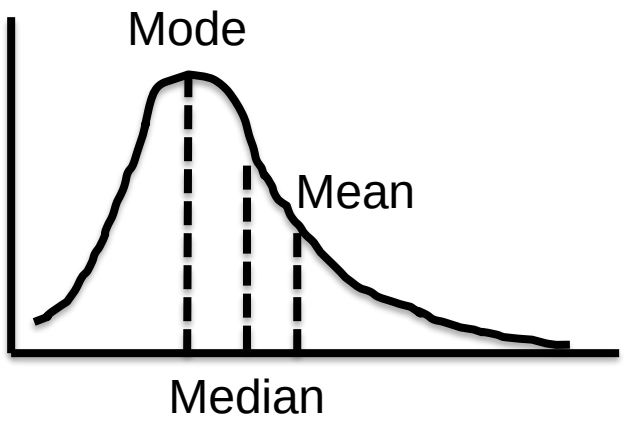
Data Types – Recent Interview Question

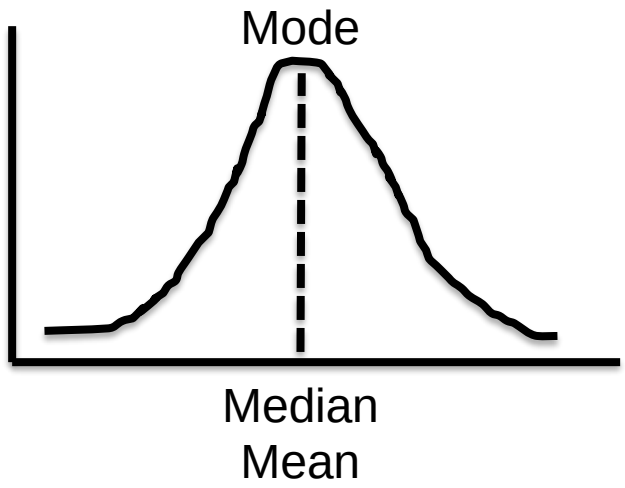
A sample of 400 Bangalore households is selected and several variables are recorded. Which of the following statements is correct?

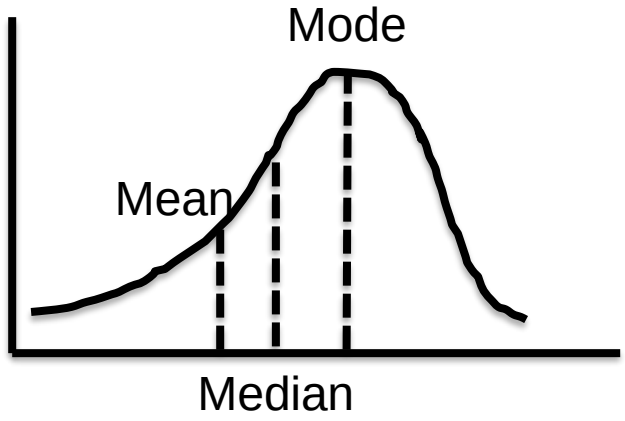
- Socioeconomic status (recorded as “low income”, “middle income”, or “high income”) is nominal level data
- The number of people living in a household is a discrete variable
- The primary language spoken in the household is ordinal level data (recorded as “Kannada”, “Tamil”, etc)

The Central Tendencies

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







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

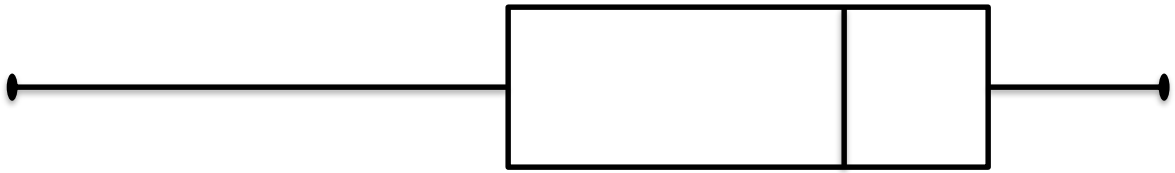
Measures of Spread – Recent Interview Question

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

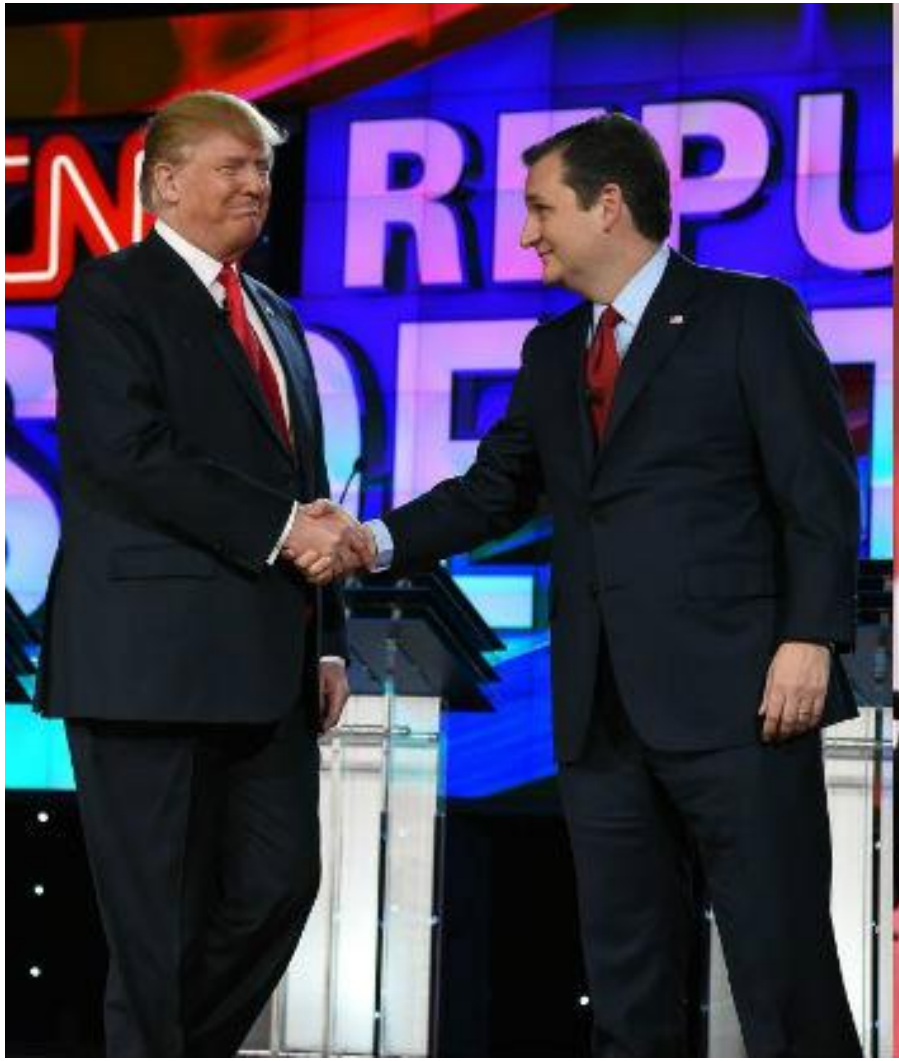
Measures of Spread (Dispersion)

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, 5^{th} decile 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	3553
Cleaners	3571
Computers	3247
Telephone	2448
Financial	2433
Beer, Wine and Liquor	2050
Candy	1137
Toys	699

Independent or Mutually Exclusive?



Donald Trump and Ted Cruz were Republican Party candidates.

Hillary Clinton and Bernie Sanders were Democratic Party candidates.

Independent or Mutually Exclusive?

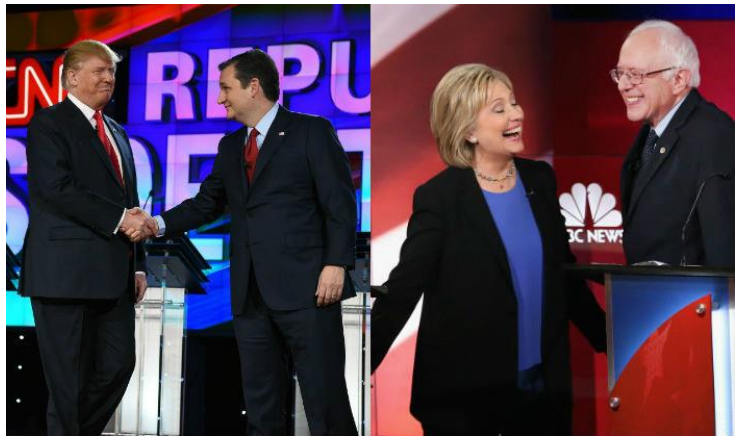
Event A: Trump winning Republican nomination

Event B: Cruz winning Republican nomination

Event C: Clinton winning Democratic nomination

Event D: Sanders winning Democratic nomination

What kinds of events are the below scenarios?



Event A and Event B	*Mutually Exclusive*
Event C and Event D	*Mutually Exclusive*
Event A and Event C	*Independent*

Independent or Mutually Exclusive?

Assuming no other candidates are left in the fray and there is a neck-to-neck contest within each party, what is:

P(A and B)	*0*	P(A or B)	*½ + ½ = 1*
P(A and C)	*½ * ½ = ¼*	P(A or C)	*½ + ½ - ¼ = ¾*

PROBABILITY BASICS

Probability - Types

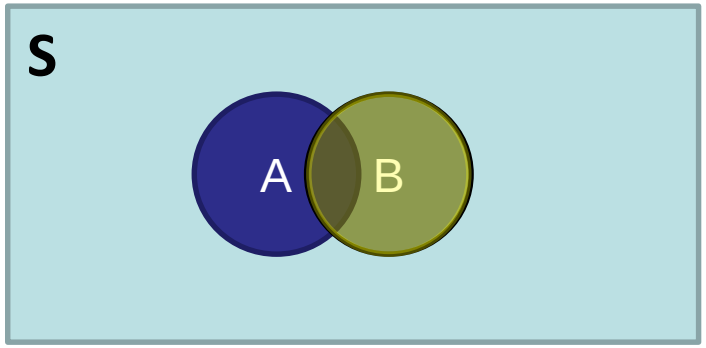
Conditional Probability

		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(No | Middle-Aged) = 0.586/0.690 = 0.85

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

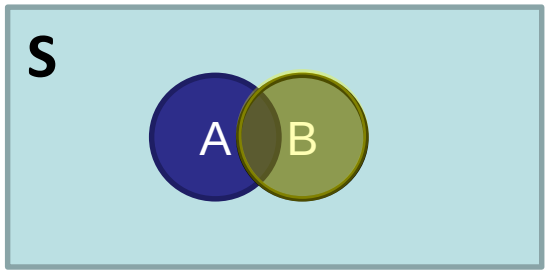
P(Middle-Aged | No) = 0.586/0.816 = 0.72 (Order Matters)

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

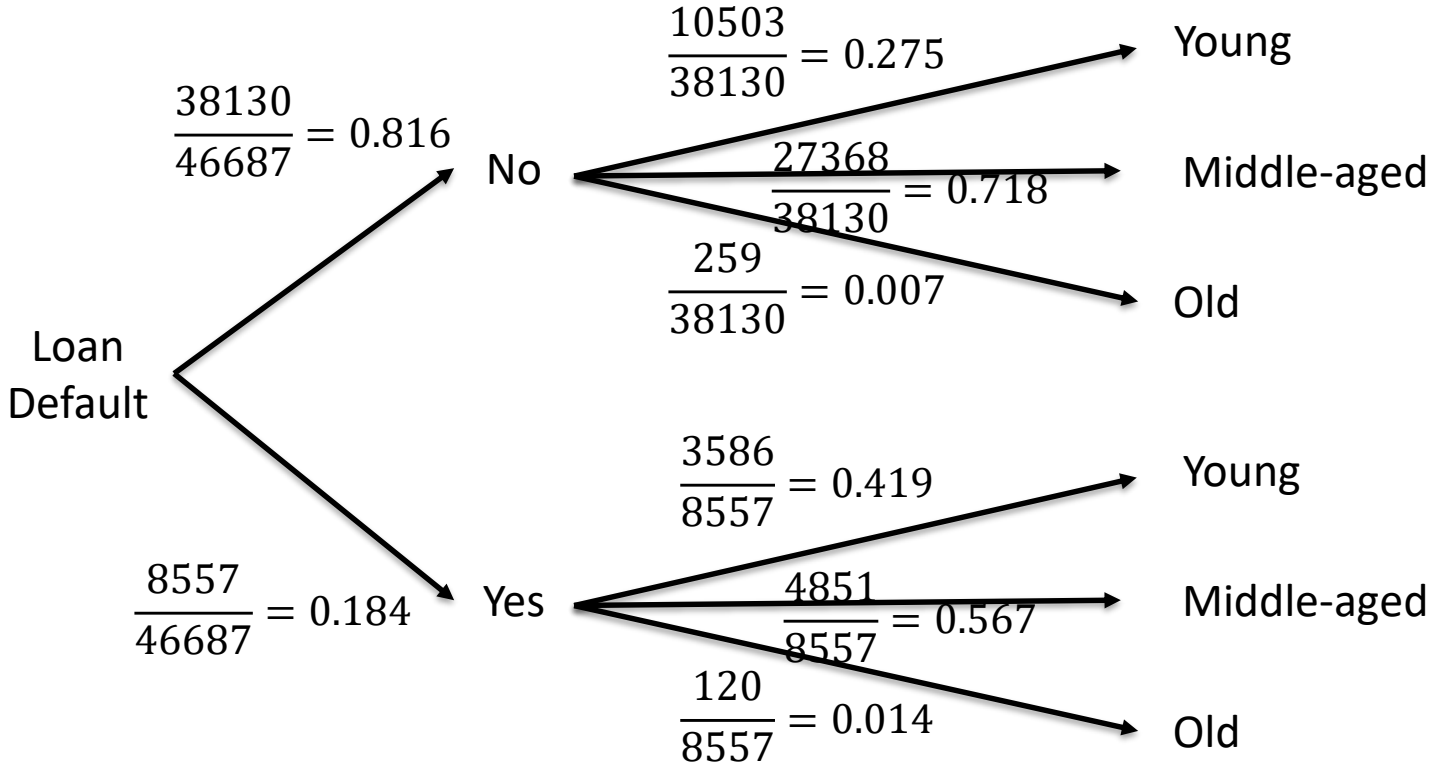
		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000



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

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

What happens when A and B are INDEPENDENT?

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

Equating, we get

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

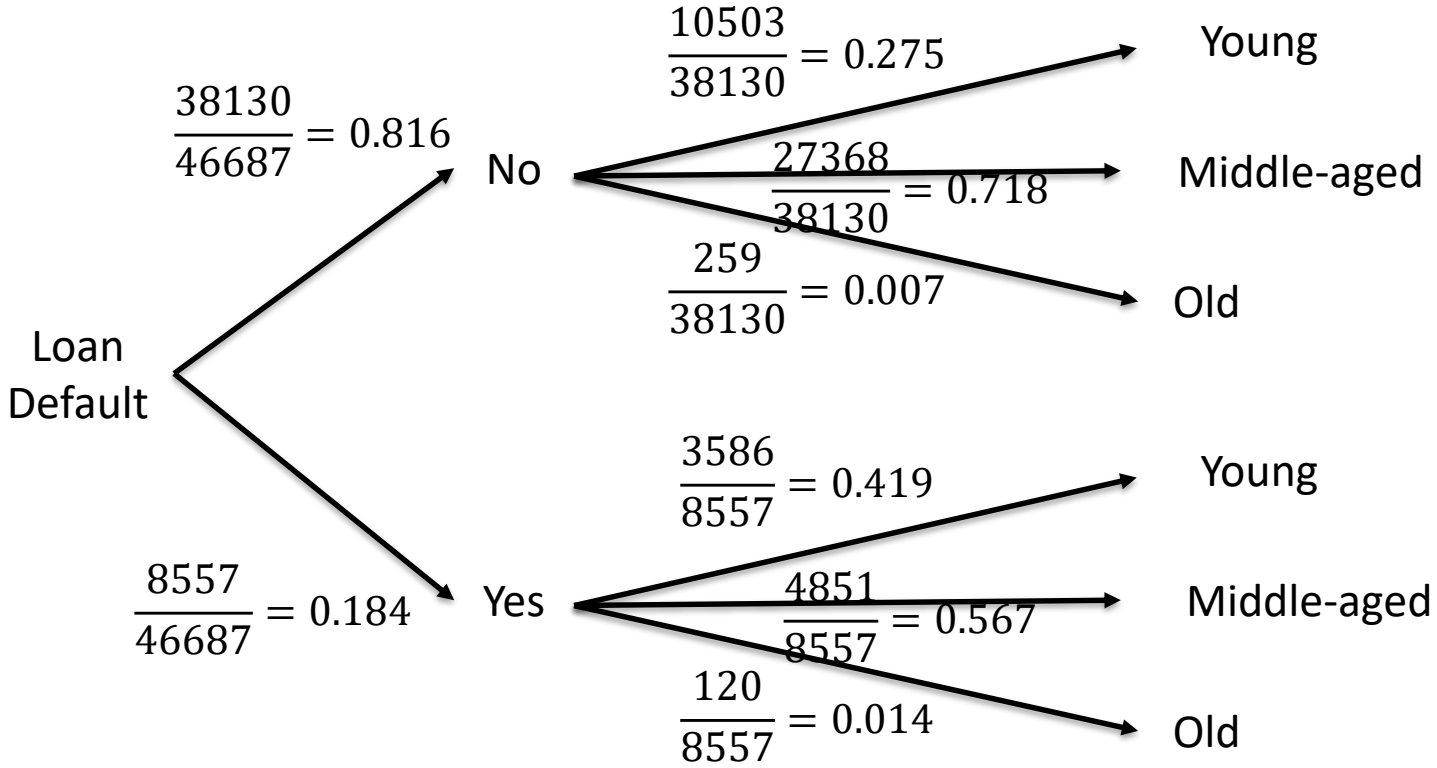
$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

Probability - Types

Conditional Probability – 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

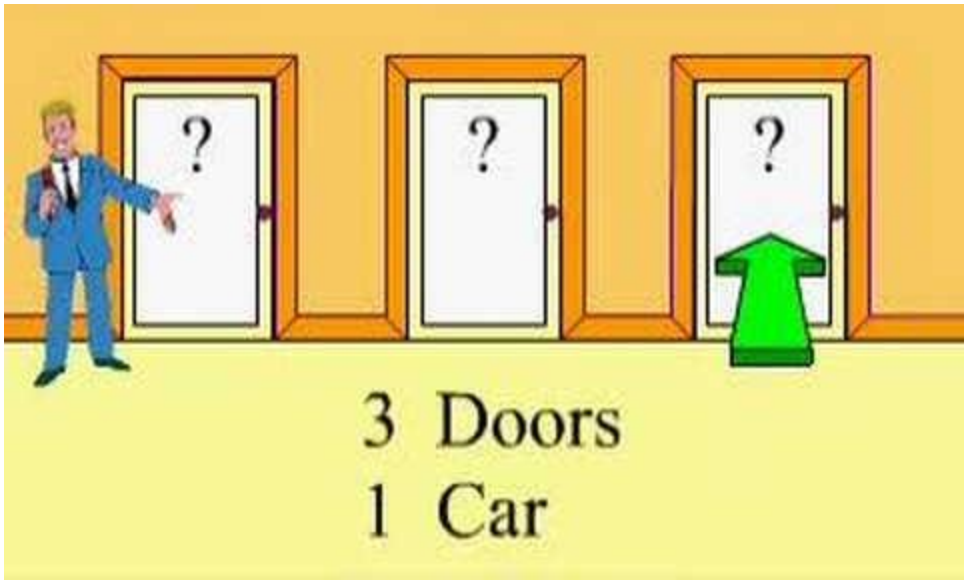
$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$



Now find
P(No | Young)

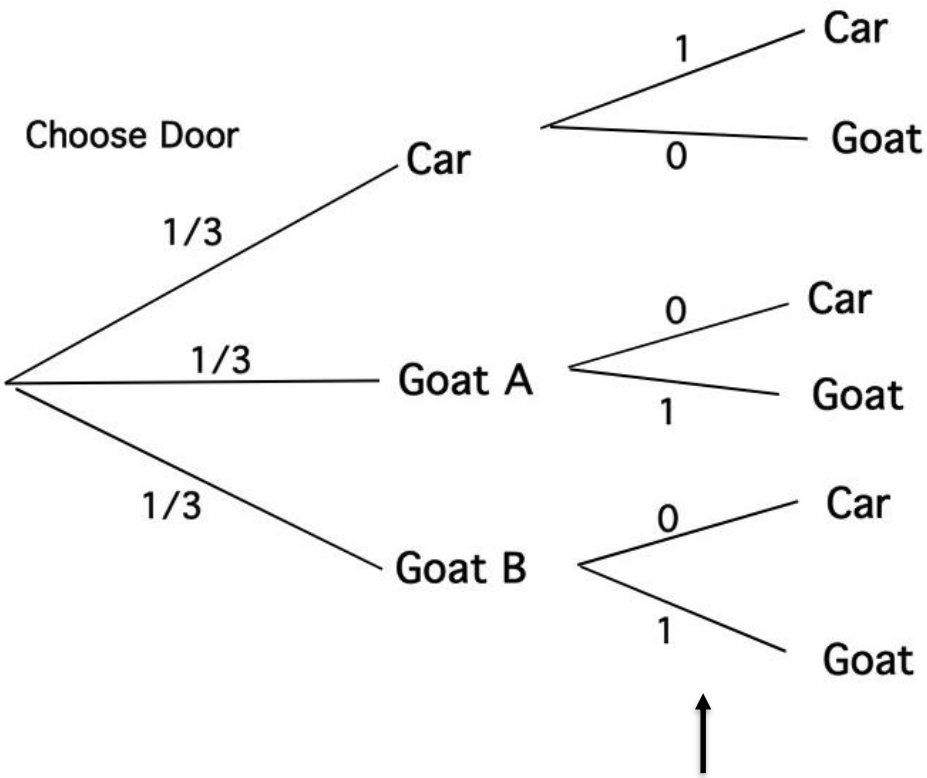
Probability - Types

Monty Hall Problem - Intuitive

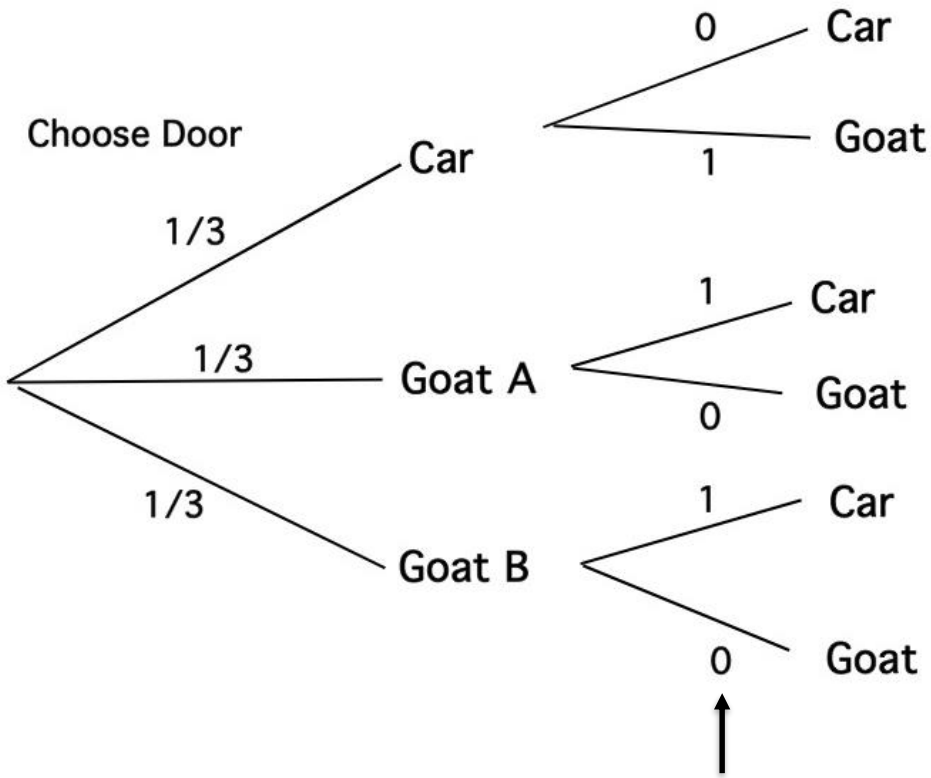


Probability - Types

Monty Hall Problem – Probability Tree



Stay on
Same Door

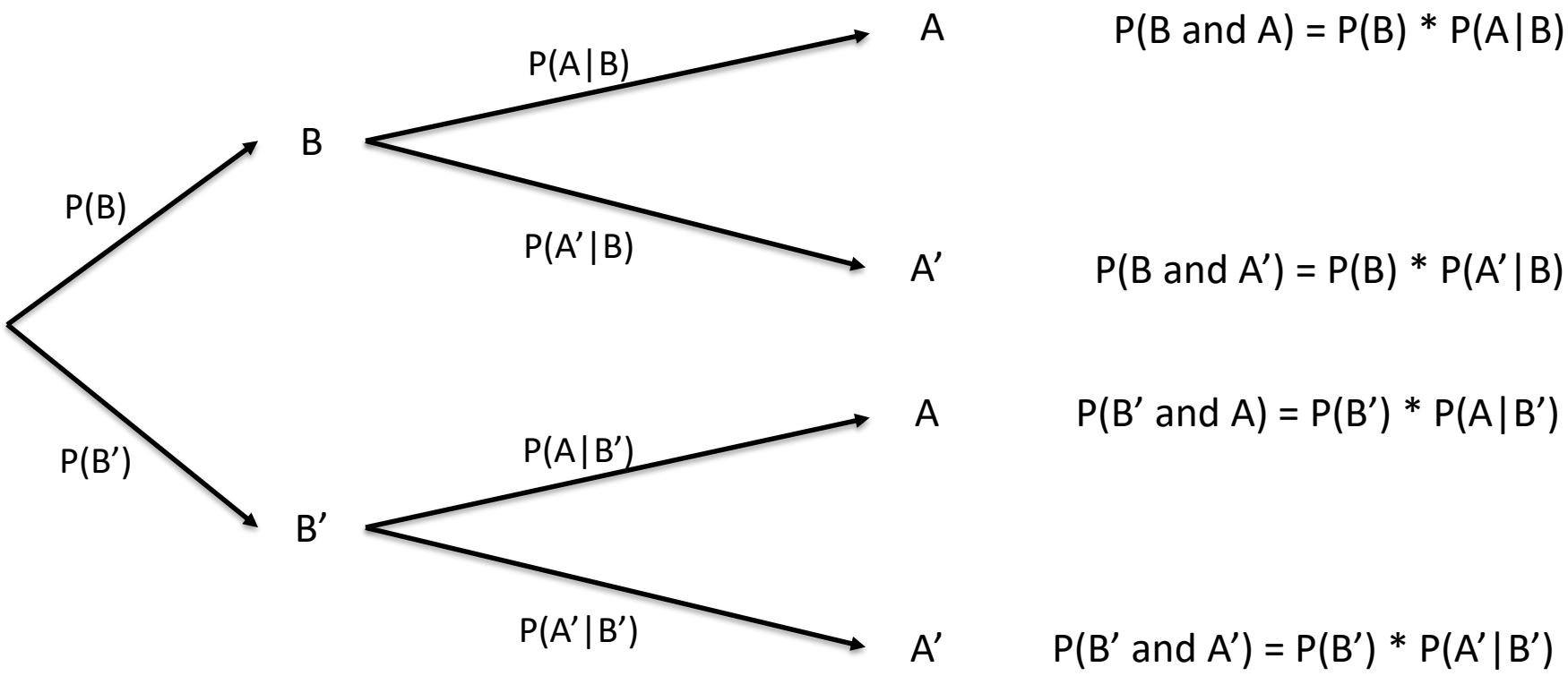


Switch
Door

CSE 7315c

Probability - Types

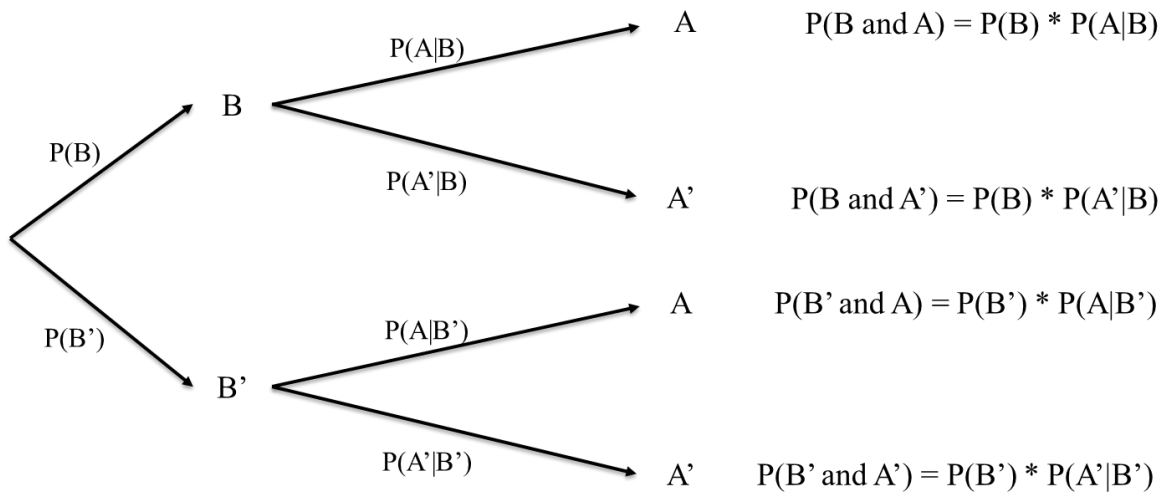
Generalized Probability Tree



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 (*aka* Prior Probability)

P(Test positive | Cancer) = 0.85 (*aka* Likelihood)

P(Test negative | No cancer) = 0.925

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

P(Test Positive) *aka* Evidence

$$Posterior\ Probability = \frac{Prior\ Probability * Likelihood}{Evidence}$$

Bayes' Theorem

Case – Clinical trials

P(Cancer) = 0.005 (*aka* Prior Probability)

P(Test positive | Cancer) = 0.85 (*aka* Likelihood)

P(Test negative | No cancer) = 0.925

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

P(Test Positive) *aka* Evidence

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

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

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

P(Free | No spam) = 0.001

P(Spam | Free) = ?

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

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

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

A slight detour

HOW GOOD IS YOUR CLASSIFICATION?

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?

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?

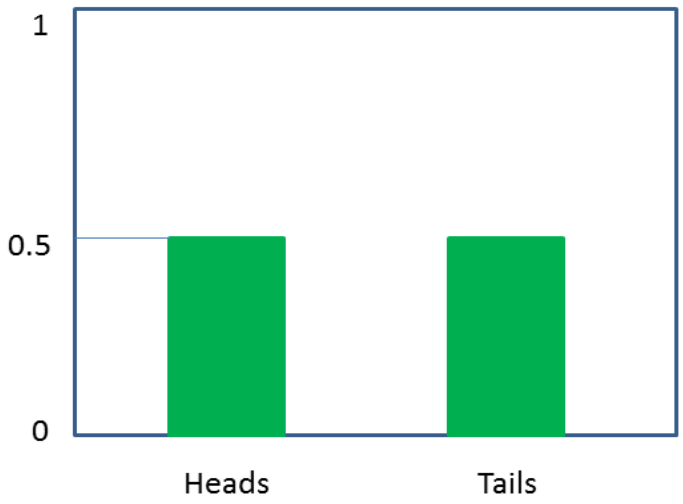
Analyzing attributes

PROBABILITY DISTRIBUTIONS

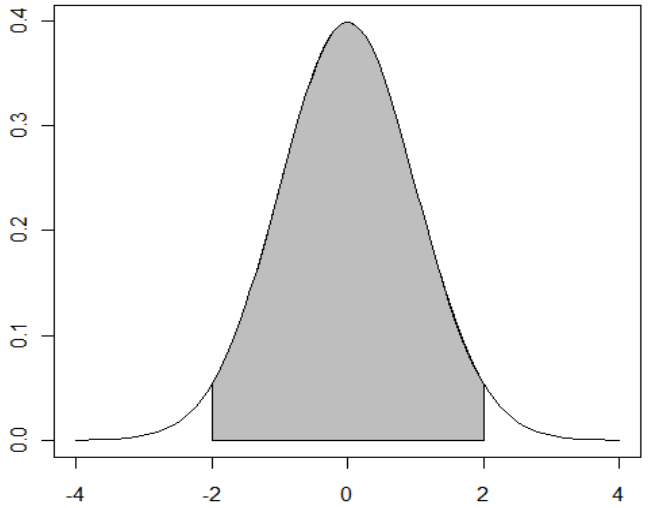
Random variable

- 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

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**

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

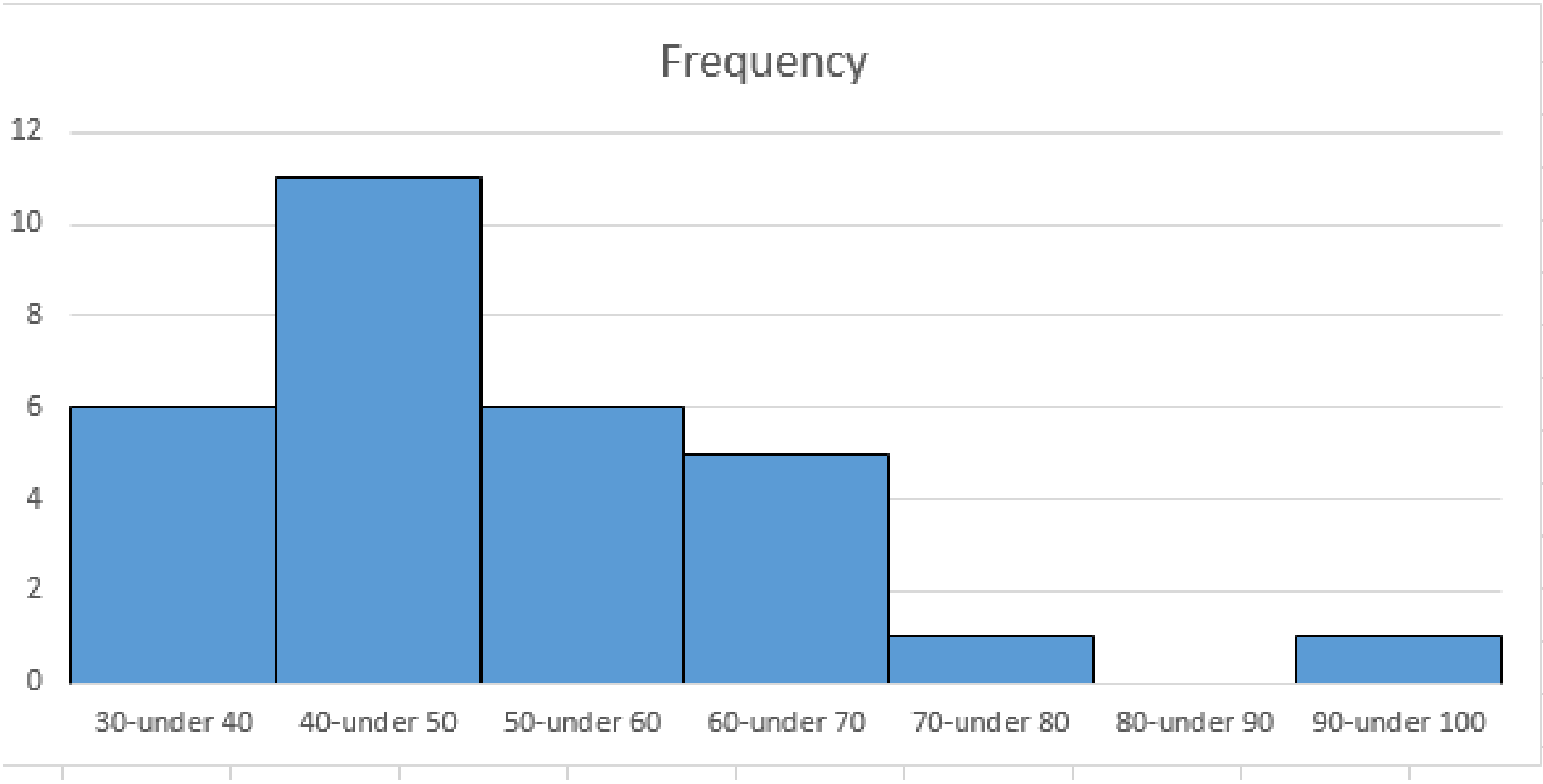
Total passengers enplaned and deplaned, passengers in transit counted once

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

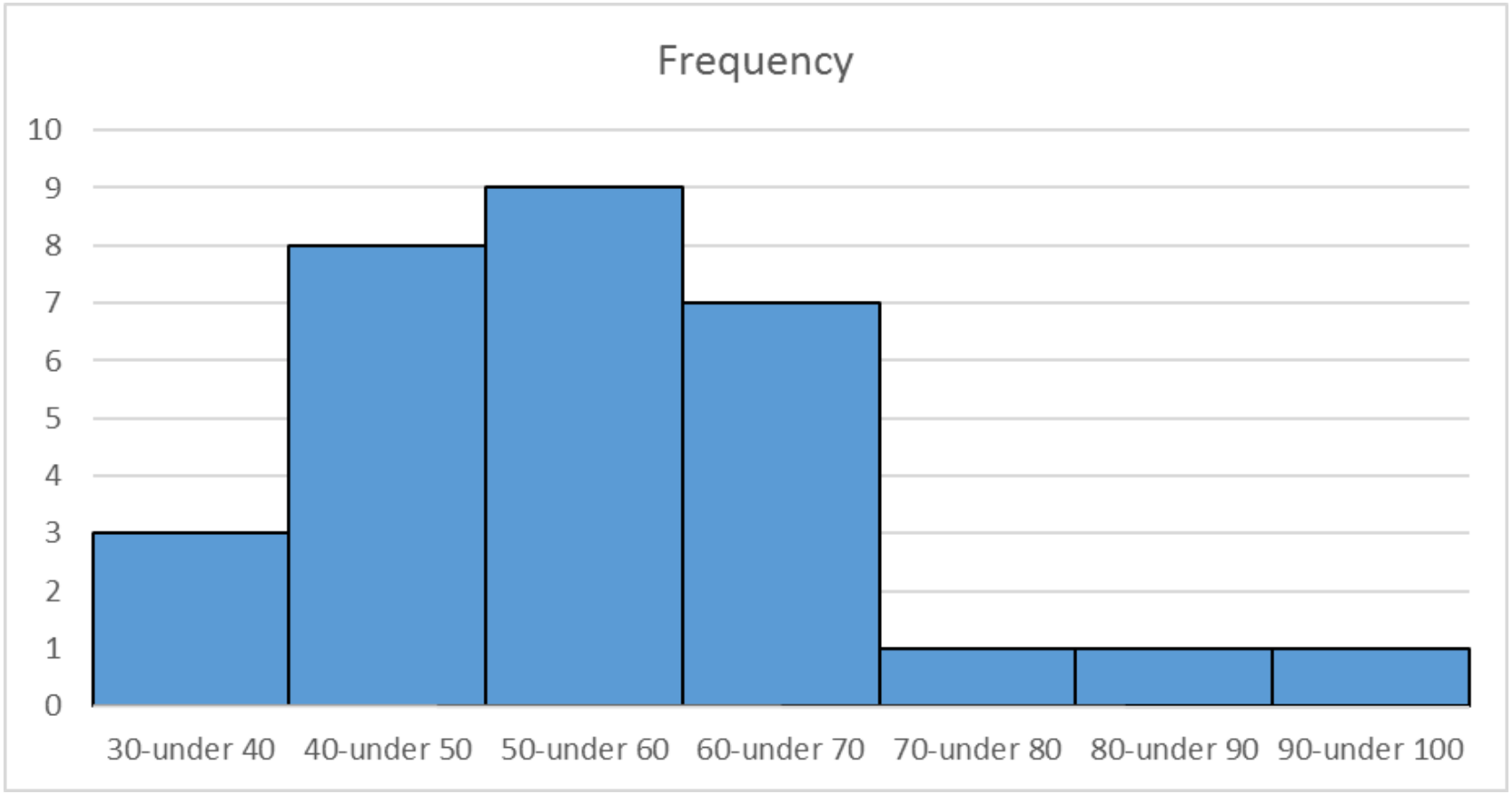
Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final
Last accessed: November 22, 2014



Histogram

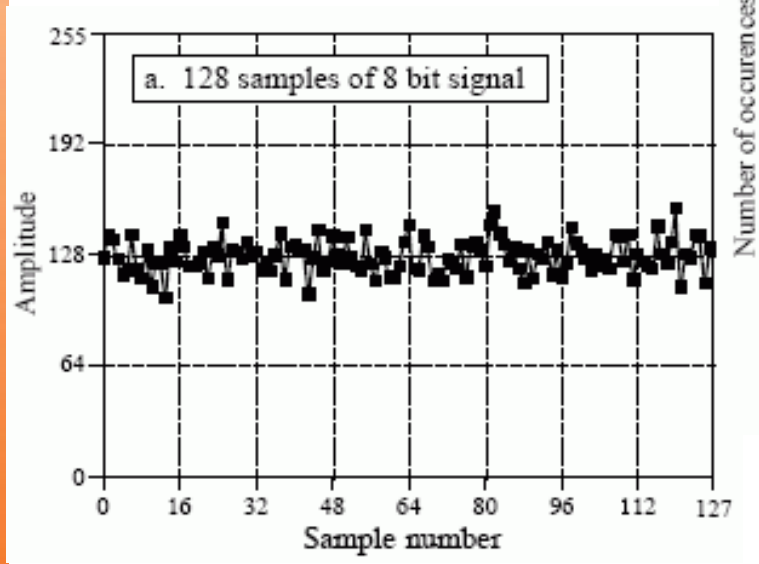
Annual traffic data for 30 busiest airports in the world – 2013

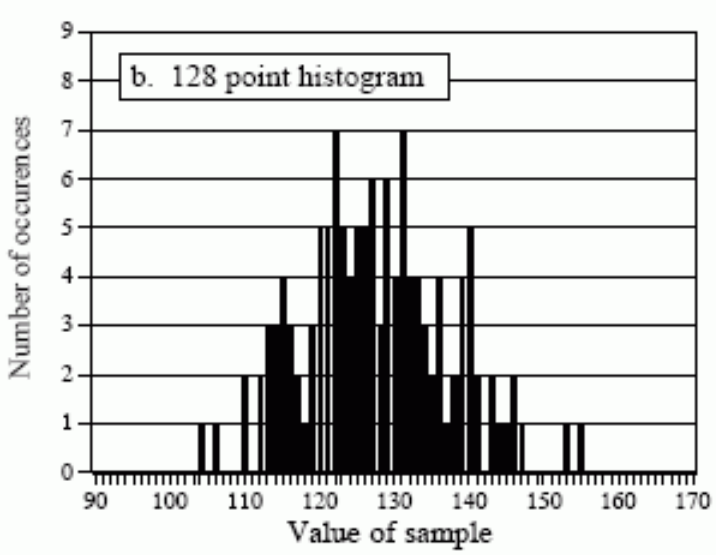
Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final
Last accessed: February 04, 2016

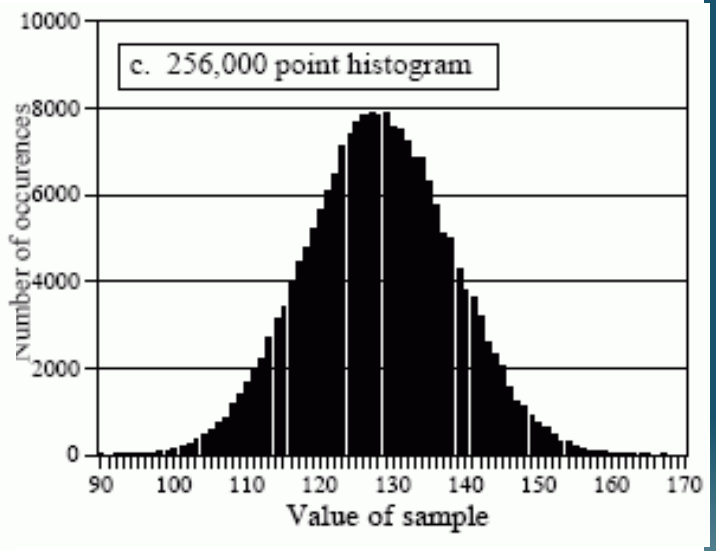


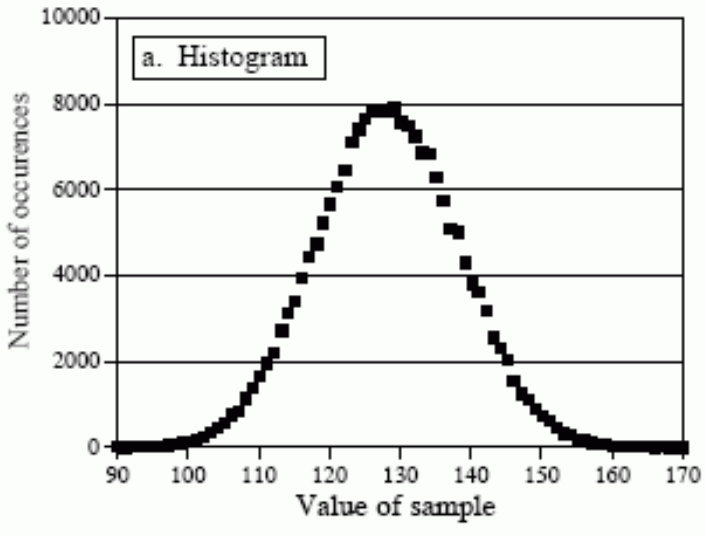
Histogram, PMF and PDF

Signal from an 8-bit analog-to-digital converter attached to a computer, e.g., 0-255 mV converted to digital numbers between 0 and 255.



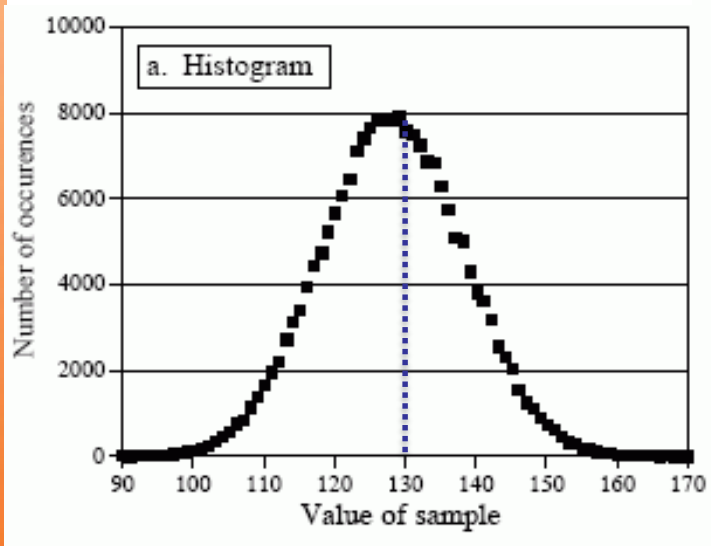


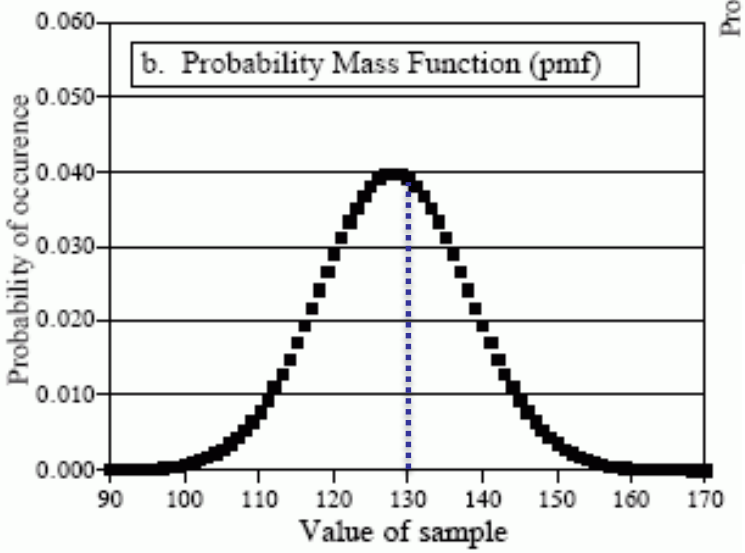


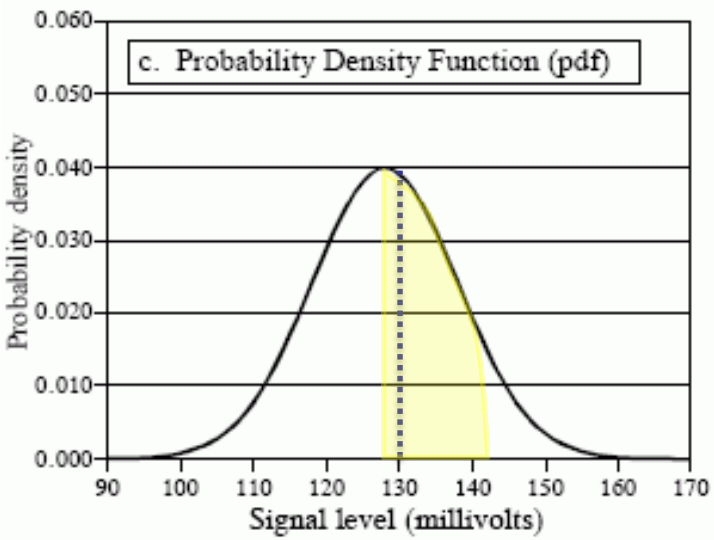


Histogram, PMF and PDF

Signal from an 8-bit analog-to-digital converter attached to a computer, e.g., 0-255 mV converted to digital numbers between 0 and 255.







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

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

Probability Distributions of Winnings and Income

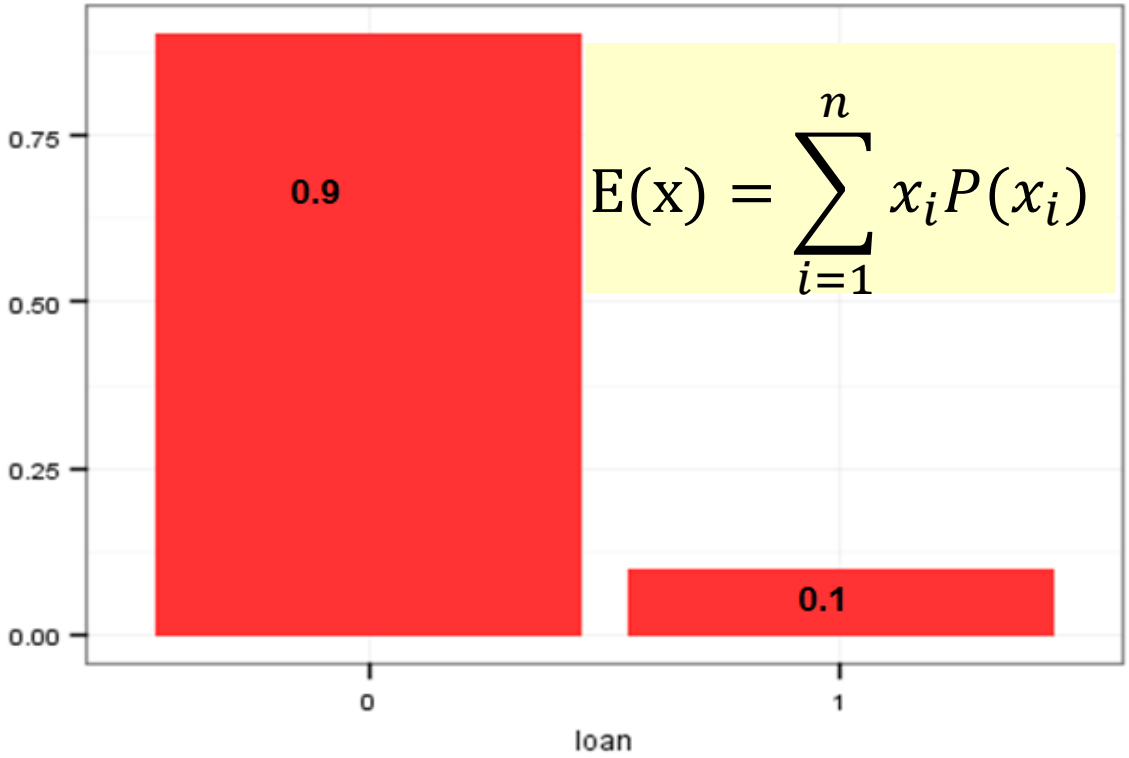
Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	-$1	$4	$9	$14	$19

Probability	0.43	0.04	0.43	0.09
Income (BHD)	100	345	1000	9833

Why do you need a probability distribution?

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

Expectation: Discrete



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

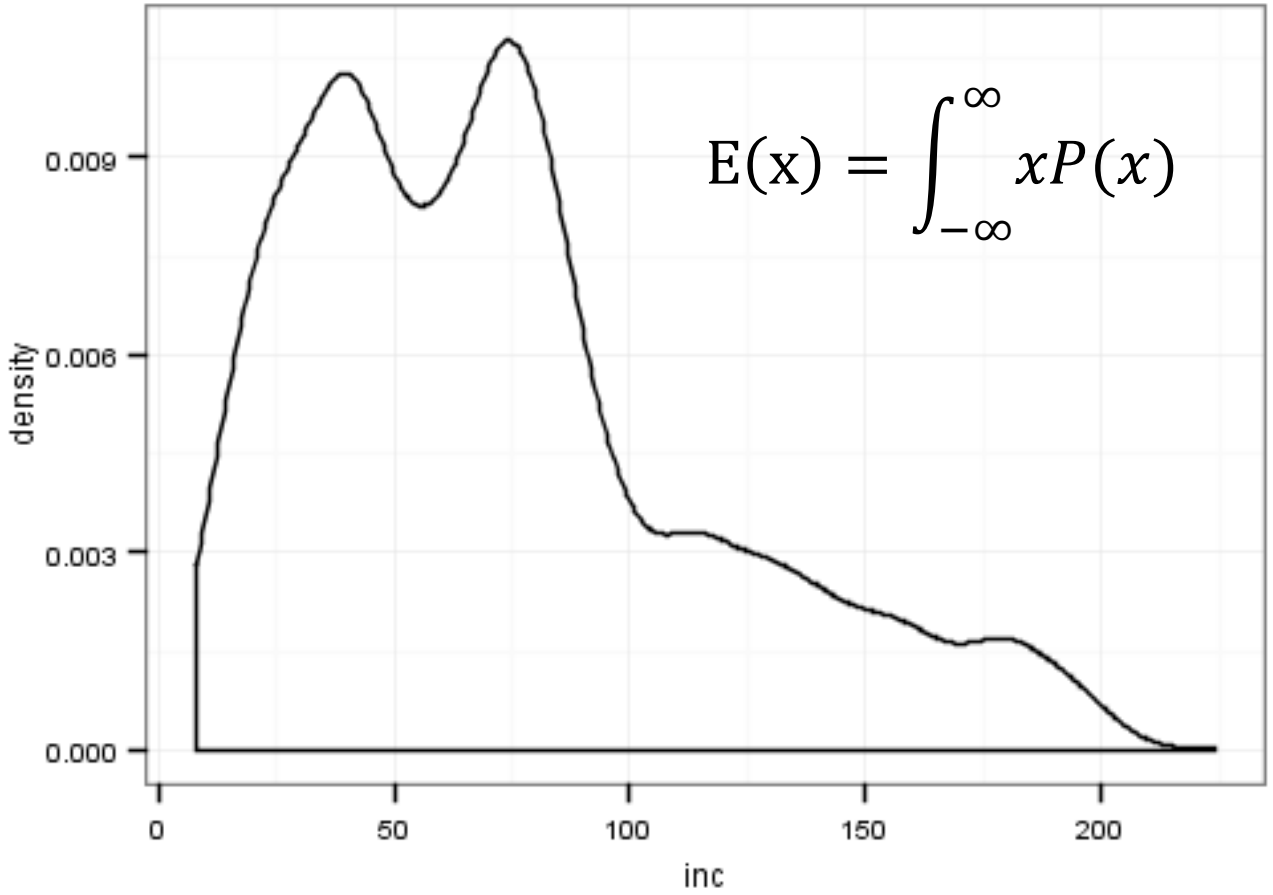
Recall anything like this?

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2
Probability	0.43	0.04	0.43	0.09

Mean, $\mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{100X10+345X1+1000X10+9833X2}{10+1+10+2} = 1348$

Expectation, $E(X) = 100 * 0.43 + 345 * 0.04 + 1000 * 0.43 + 9833 * 0.09 = 1348$

Expectation: Continuous



$$\mathrm{E(x)} = \int_{-\infty}^{\infty} xP(x)$$

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) = -0.77 (calculate and verify)

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

So, why play?

There is **VARIANCE**.

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

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 as μ 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.

If I have a portfolio of 30% TCS, 50% Wipro and 20% Ranbaxy stocks, the expected return of my portfolio is

E(Portfolio) = 0.3 E(TCS) + 0.5 E(Wipro) + 0.2 E(Ranbaxy)

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
$E(Y-E(Y))^2 = a^2\ E(X-E(X))^2$

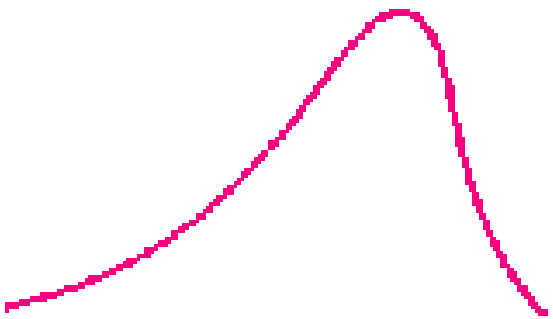
However, the left hand side is Variance of Y and RHS is Variance of X

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

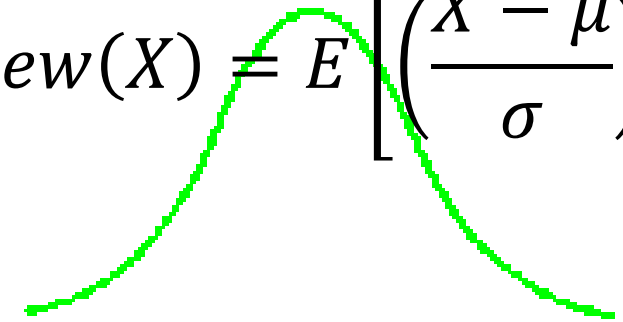
Understanding the shape of a PDF - Skewness

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

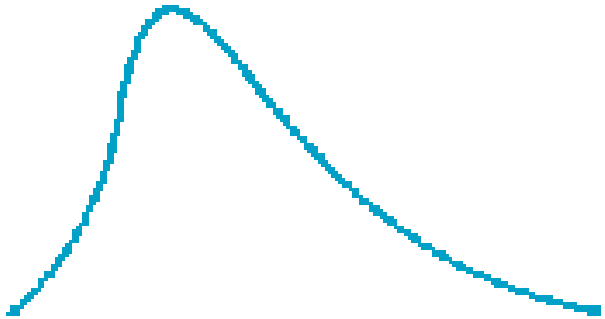
$$skew(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



Negatively (left) skewed distribution



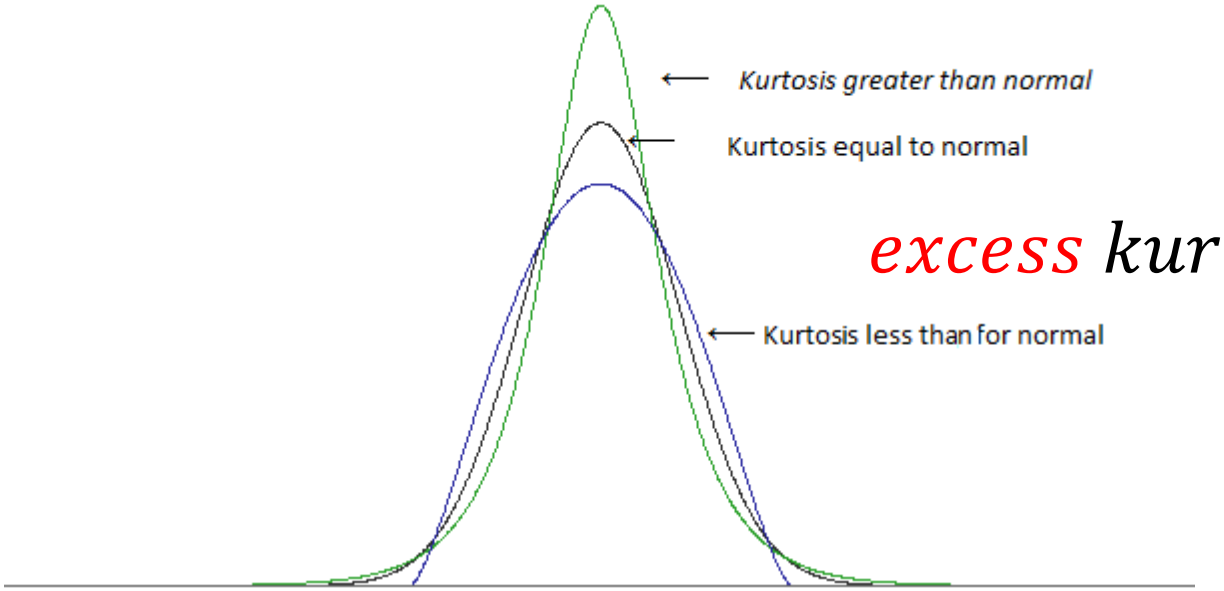
Normal distribution



Positively (right) skewed distribution

Understanding the shape of a PDF - Kurtosis

A measure of the 'tailed'ness of the data distribution as compared to a normal distribution. Negative kurtosis means a distribution with light tails (fewer extreme deviations from mean (or outliers) than in normal distribution). Positive kurtosis means a distribution with heavy tails (more outliers than in normal distribution).



$$excess\ kurt(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

Image Source: http://stats.stackexchange.com/questions/84158/how-is-the-kurtosis-of-a-distribution-related-to-the-geometry-of-the-density-fu
Last accessed: March 31, 2017

Describing a Distribution – Summary of Moments

Measure	Formula	Description
Mean (μ)	$E(X)$	Measures the centre of the distribution of X
Variance (σ^2)	$E[(X-\mu)^2]$	Measures the spread of the distribution of X about the mean
Skewness	$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$	Measures asymmetry of the distribution of X
Kurtosis (excess)	$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]-3$	Measures 'tailed'ness of the distribution of X and useful in outlier identification

CSE 7315c

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, 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: 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?

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.

Deviations from Mean and Median - Excel

The MEDIAN minimizes sum of absolute deviations.

The MEAN minimizes sum of squared deviations.

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.

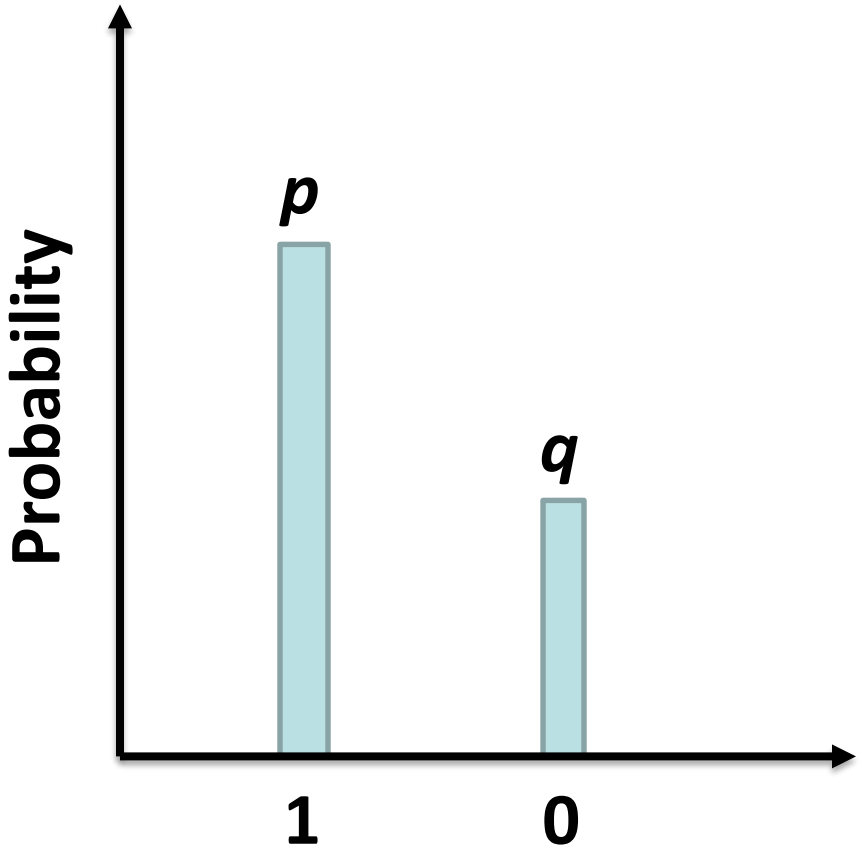
SOME COMMON DISTRIBUTIONS

Bernoulli

There are two possibilities (loan taker or non-taker) with probability *p* of success and *1-p* of failure

- Expectation: *p*
- Variance: *p(1-p)* or *pq*, where *q=1-p*

Bernoulli



$$Expectation, E(X) = \sum x_i P(x_i)$$

$$= 1 * p + 0 * q = p$$

$$Variance, Var = \sum (x_i - \mu)^2 P(x_i)$$

$$= (1-p)^2 * p + (0-p)^2 * (1-p)$$
$$= p(1-p)$$

Geometric Distribution

Number of independent and identical Bernoulli trials needed to get ONE success, e.g., number of people I need to call for the first person to accept the loan.

Geometric Distribution

PMF*, $P(X = r) = q^{r-1}p$ — *(r-1)* failures followed by ONE success.

$P(X > r) = q^r$ — Probability you will need more than *r* trials to get the first success.

CDF**, $P(X \leq r) = 1 - q^r$ — Probability you will need *r* trials or less to get your first success.

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{q}{p^2}$$

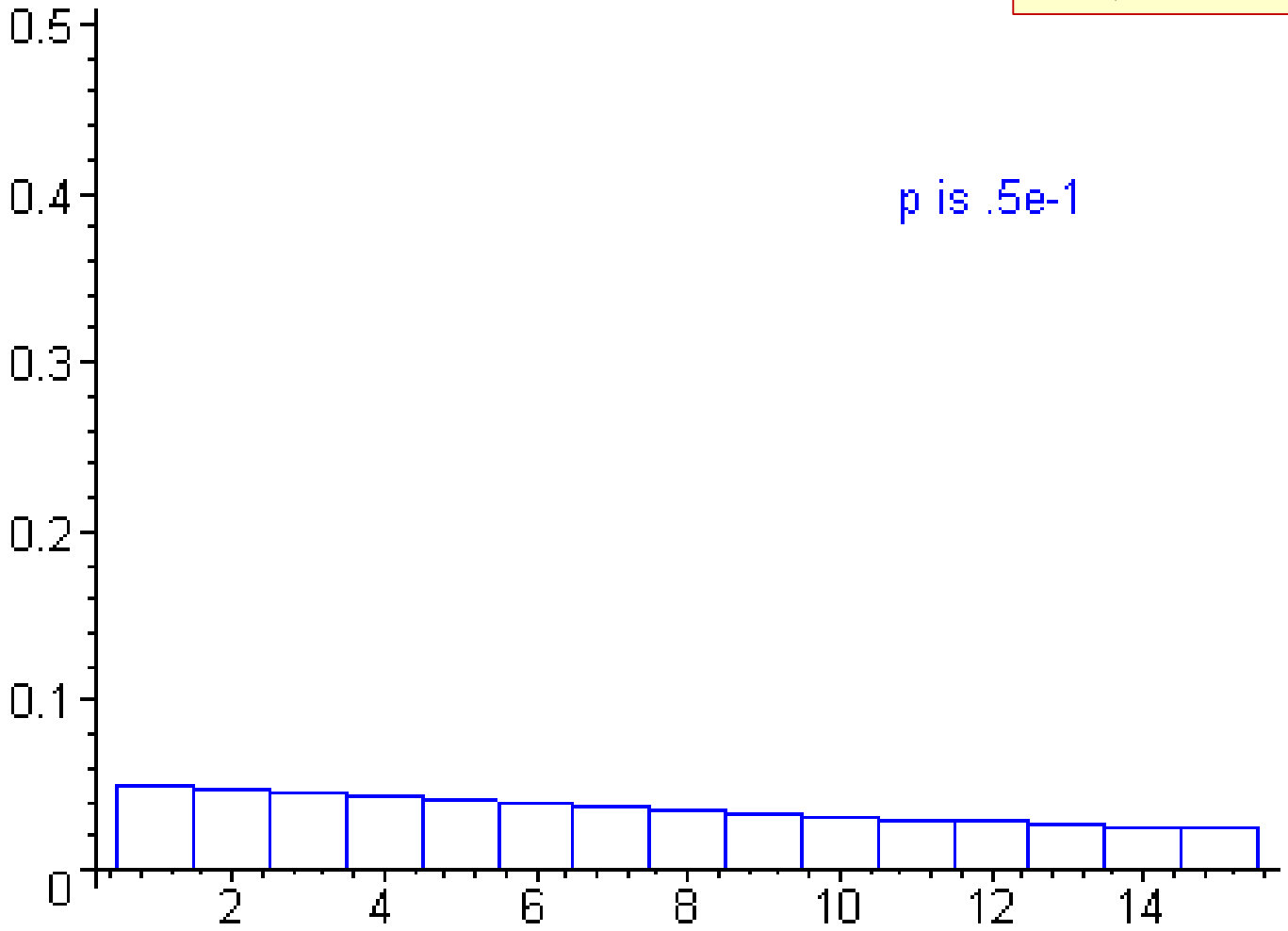
* Probability Mass Function ** Cumulative Distribution Function

Geometric Distribution

- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- The main thing you are interested in is how many trials are needed in order to get the first successful outcome.

X~Geo(p)

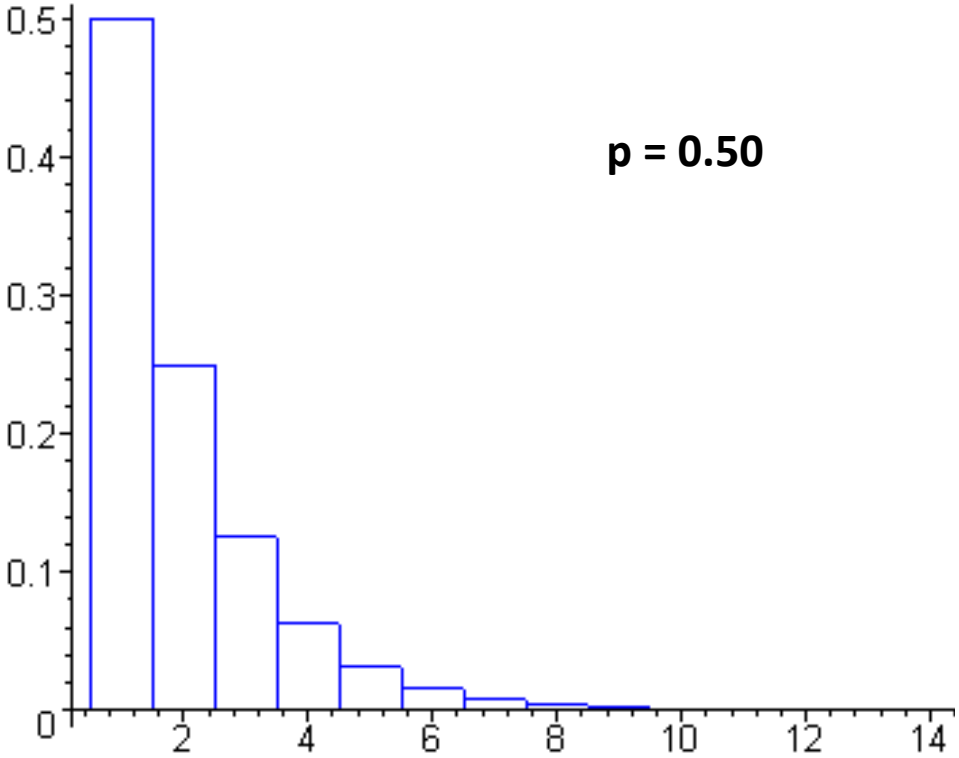
$$P(X = r) = q^{r-1}p$$



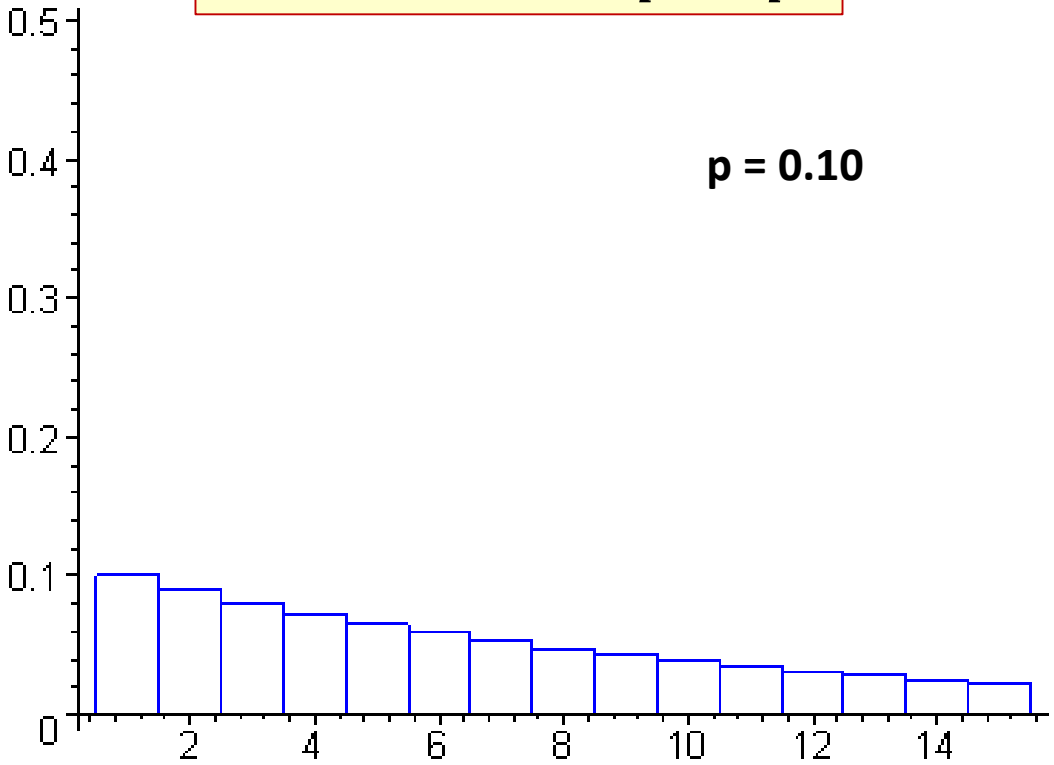
Ref: http://personal.kenyon.edu/hartlaub/MellonProject/Geometric2.html
Last accessed: June 12, 2015

X~Geo(p)

$$P(X = r) = q^{r-1}p$$



p = 0.50



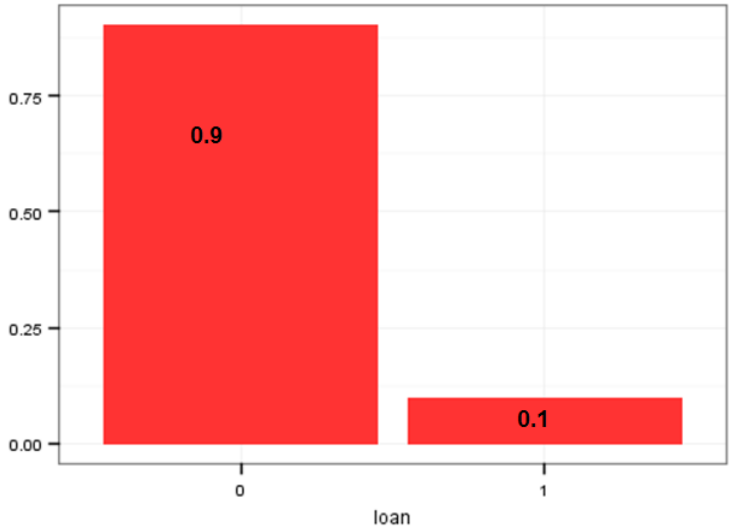
p = 0.10

Ref: http://personal.kenyon.edu/hartlaub/MellonProject/Geometric2.html
Last accessed: December 09, 2017

Binomial Distribution

If I randomly pick 10 people, what is the probability that I will get exactly

- 0 loan takers = 0.9^{10}
- 1 loan taker = $10 * 0.1^{1} * 0.9^{9}$
- 2 loan takers = $C_2^{10} * 0.1^{2} * 0.9^{8}$



Binomial Distribution

If there are two possibilities with probability *p* for success and *q* for failure, and if we perform *n* trials, the probability that we see *r* successes is

PMF, $\mathrm{P}(\mathrm{X} = \mathrm{r}) = C_r^n p^r q^{n-r}$

CDF, $\mathrm{P}(\mathrm{X} \leq \mathrm{r}) = \sum_{i=0}^{r} C_i^n p^i q^{n-i}$

Binomial Distribution

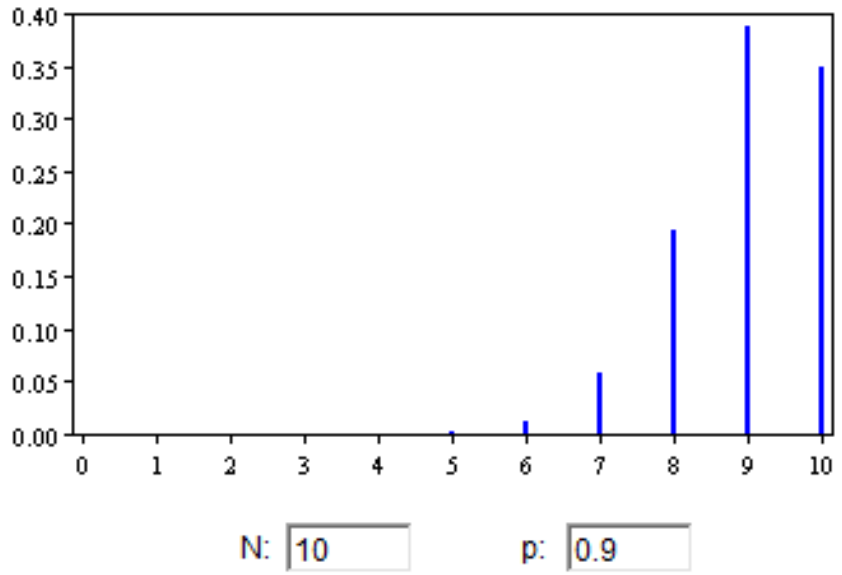
$$E(X) = np \qquad Var(X) = npq$$

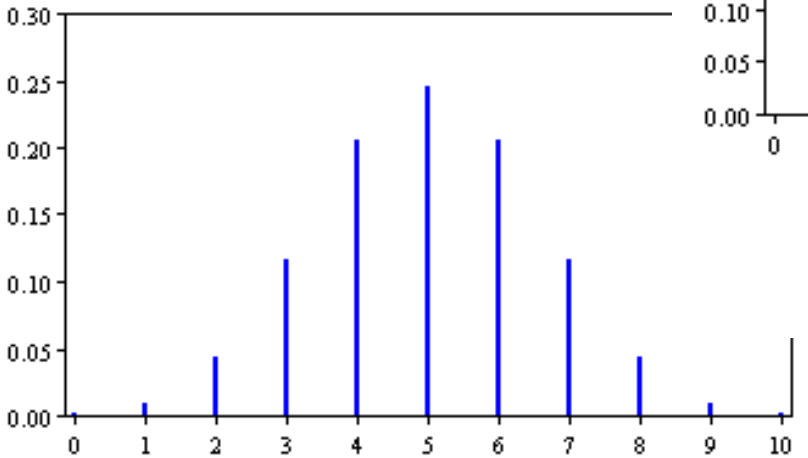
When to use?

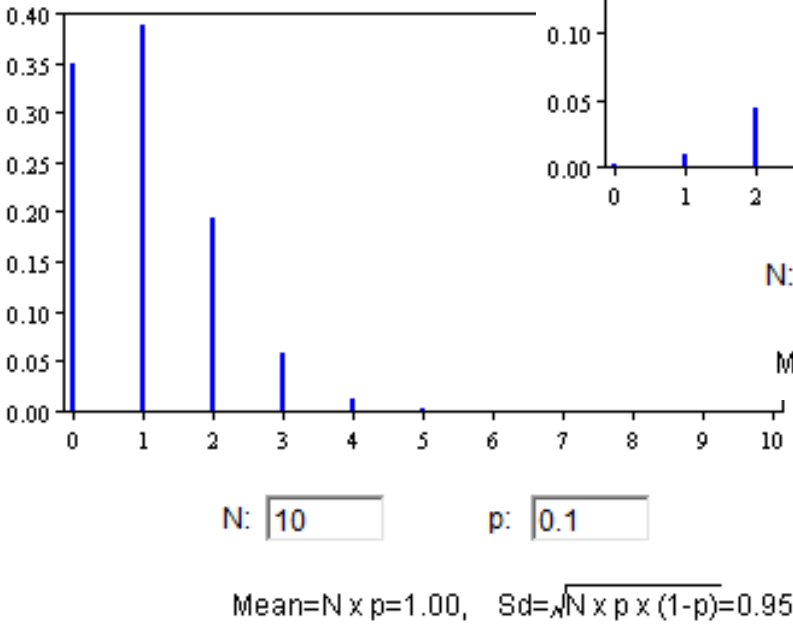
- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- There are a finite number of trials, and you are interested in the number of successes or failures.

X~B(n,p)

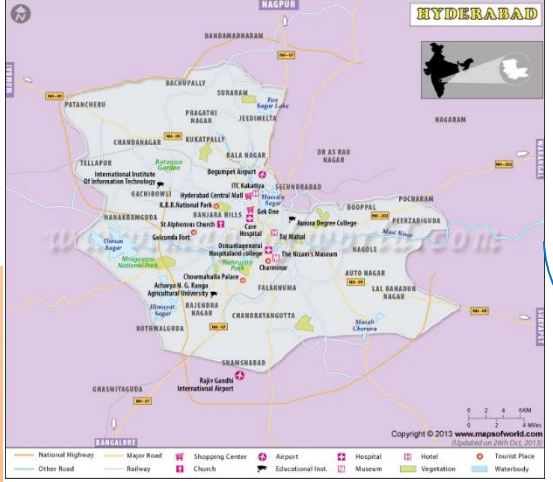
$$P(X = r) = C_r^n p^r q^{n-r}$$



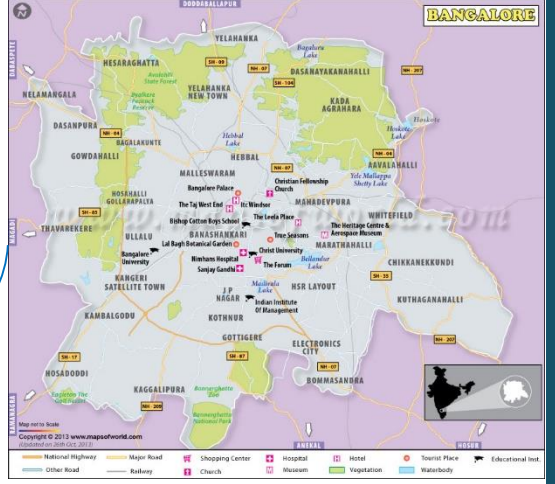




Ref: http://onlinestatbook.com/2/probability/binomial_demonstration.html
Last accessed: December 09, 2017 on Safari



INSOFE
Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

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

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.