Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Basic Statistical Concepts, Central Tendencies and Measures of Variability, Probability Basics

Prof. Anuradha Sharma

December 02, 2018

MATERIAL CONTENT FROM Dr. SRIDHAR PAPPU

MOTIVATION

Continuous improvement is better than delayed perfection.
Mark Twain – American Writer, Humorist

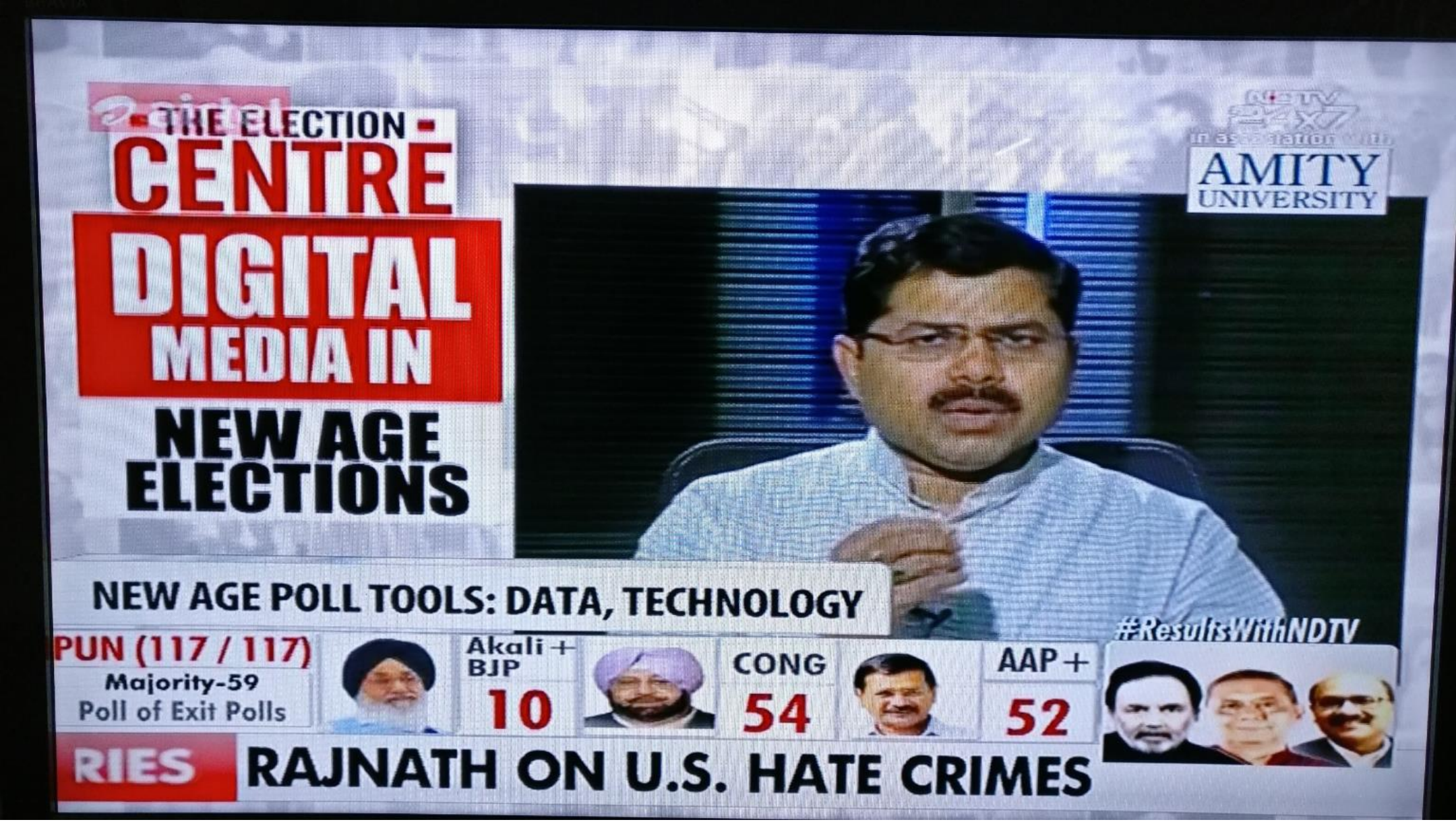
What gets measured, gets managed.
Peter Drucker – Management Guru

Statistics is the technology of finding the invisible and measuring the unmeasurable.
Dr C.R. Rao – Mathematician and Statistician

$1 invested in statistics helped us to gain back $1.08 in revenue. **John F Welch Jr past CEO of General Electric (GE)**

Why Study Statistics?

Statistics are part of your daily life and are all around you.


THE ELECTION
CENTRE
DIGITAL
MEDIA IN
NEW AGE
ELECTIONS
AMITY
UNIVERSITY
NEW AGE POLL TOOLS: DATA, TECHNOLOGY
#ResultsWithNDTV
PUN (117 / 117)
Majority-59
Poll of Exit Polls
Akali + BJP
10
CONG
54
AAP +
52
RIES
RAJNATH ON U.S. HATE CRIMES

There are three kinds of lies: lies, damned lies, and statistics.

- Mark Twain / Benjamin Disraeli

Patient: “Will I survive this risky operation?”

Surgeon: “Yes, I’m absolutely sure that you will survive the operation.”

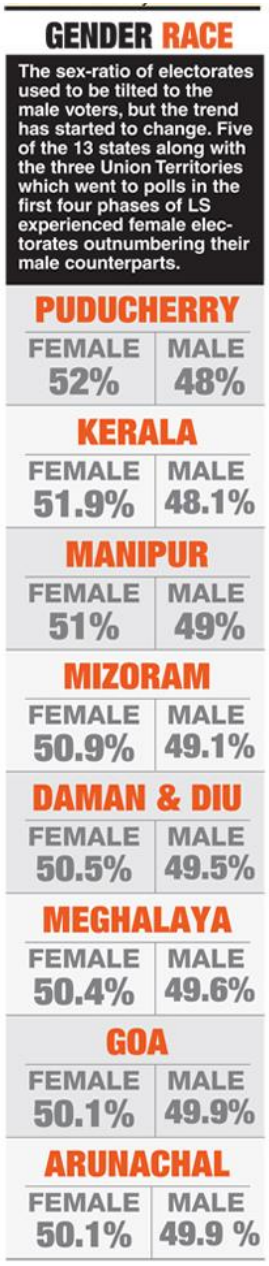
Patient: “How can you be so sure?”

Surgeon: “9 out of 10 patients die in this operation, and yesterday patient who died was my ninth patient”

Why Study Statistics?

Statistics don't lie but Statisticians will in any of the following situations:

- Data Gathering
- Data Understanding
- Data Analysis/Interpretation
- Data Presentation



LIARLIAR PANTS ON FIRE!

Problem #1: Data Gathering

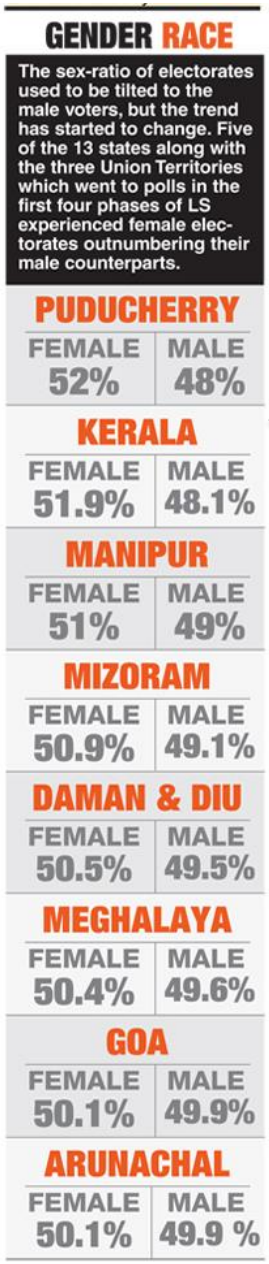
Schedule Reference		Parliamentary Constituency			
		Sl.	PC No.	PC Name	Type
Schedule no:	**7**	1	1	Daman & Diu	GEN
No of PCs going to poll	**1**				
Issue of Notification:	02 Apr 14 (Wed)				
Last Date for filing Nominations:	09 Apr 14 (Wed)				
Scrutiny of Nominations:	10 Apr 14 (Thu)				
Last date for withdrawal of Candidature:	12 Apr 14 (Sat)				
Date of Poll	**30 Apr 14 (Wed)**				
Counting of Votes:	**16 May 14 (Fri)**				
Date before which the election shall be completed	28 May 14 (Wed)				

Source: http://eci.nic.in/eci_main1/GE2014/Schedule/DD.htm
Last accessed: October 24, 2014

By April 24, when Puducherry went to polls, 6 phases (not 4) were completed, and 19 States and 5 UTs had completed polling (not 13 and 3, respectively; Daman & Diu went to polls on April 30).

Source: http://epaper.deccanchronicle.com/articledetailpage.aspx?id=474880;
Last accessed: April 27, 2014

Problem #2: Data Understanding

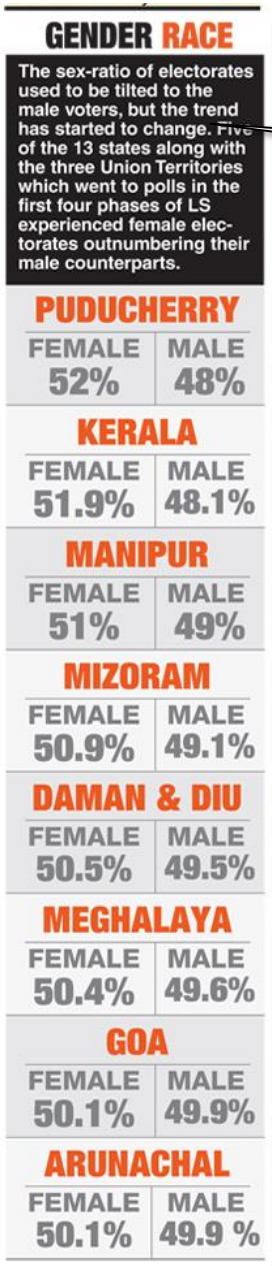


The ratios reflect the ratios of registered voters.

	Registered Voters			Voted in 2014 General Elections		
State/UT	Male	Female	% Female	Male	Female	% Female
Puducherry	432048	469309	52.07	351360	388657	52.52
Kerala	11734258	12592391	51.76	8678185	9297708	51.72
Manipur	871431	902894	50.89	685427	727210	51.48
Mizoram	346219	355951	50.69	216167	217034	50.1
Daman & Diu	57011	54816	49.02	42378	44855	51.42
Meghalaya	777639	789602	50.38	524774	553284	51.32
Goa	528308	532469	50.2	395766	421234	51.56
Arunachal Pradesh	379627	379760	50.01	289291	307665	51.54

Data from http://pib.nic.in/newsite/PrintRelease.aspx?relid=105116 and http://pib.nic.in/newsite/efeatures.aspx?relid=104195.

Source: http://epaper.deccanchronicle.com/articledetailpage.aspx?id=474880;
Last accessed: April 27, 2014



Problem #3: Data Analysis/Interpretation

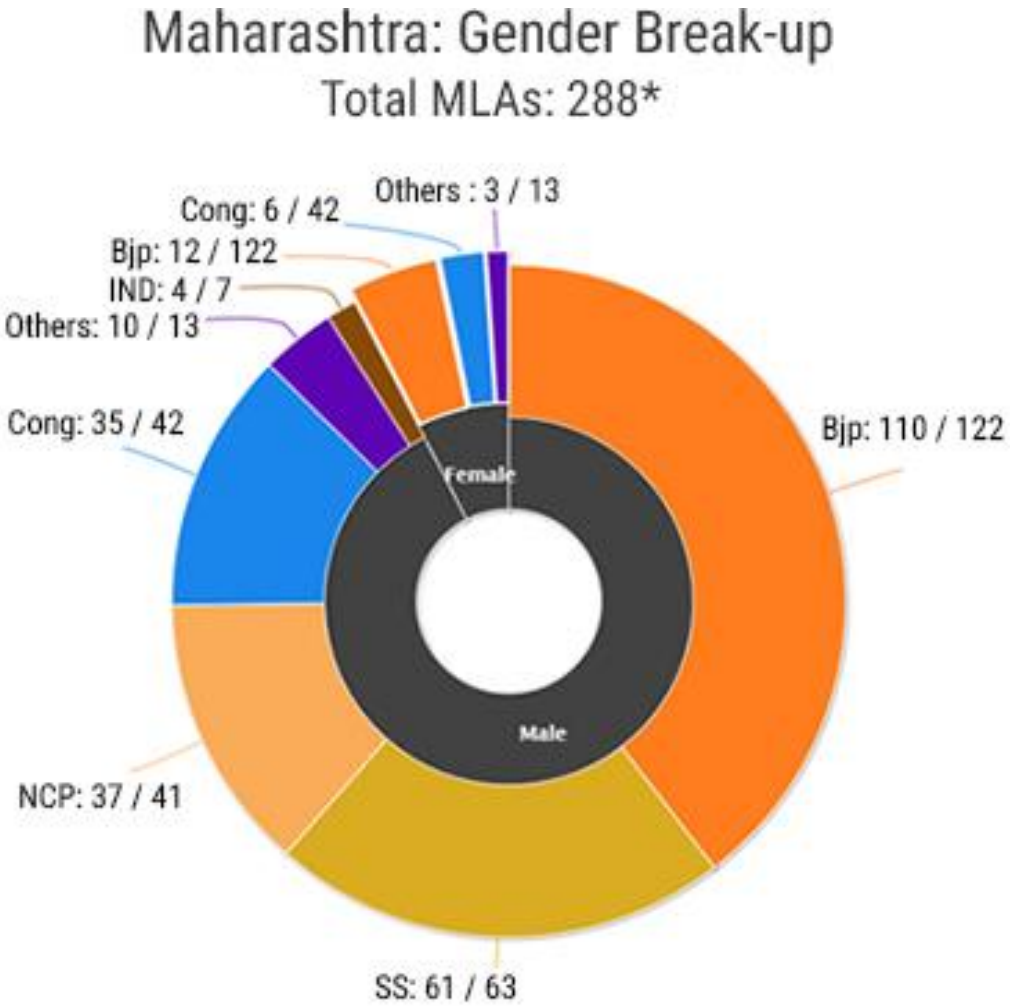
The sex-ratio of electorates used to be tilted to the male voters, but the trend has started to change.

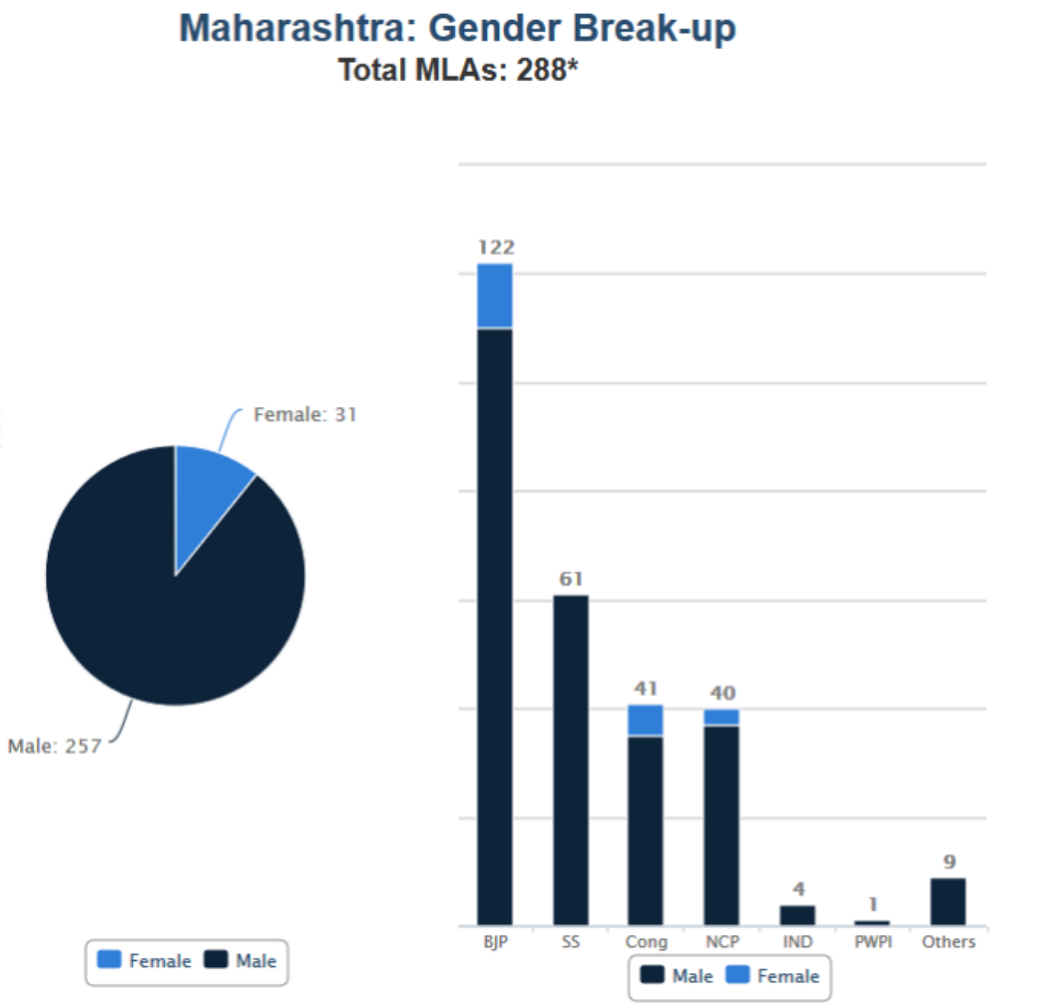
	Male			Female			Male	Female	Female-Male			
State/UT	2006-08	2011-13	2014	2006-08	2011-13	2014	Sparklines		2006-08	2011-13	2014	Sparklines
Puducherry	84.48	83.97	81.32	86.29	86.97	82.81			1.81	3	1.49	
Kerala	73.17	75.08	73.96	71.08	74.78	73.84			-2.09	-0.3	-0.12	
Manipur	85.88	76.94	78.66	86.82	81.36	80.54			0.94	4.42	1.88	
Mizoram	78.77	80.3	62.44	81.24	82.2	60.97			2.47	1.9	-1.47	
Daman & Diu			74.33			81.83					7.5	
Meghalaya	88.62	85.17	67.48	89.36	88.44	70.07			0.74	3.27	2.59	
Goa	69.7	78.86	74.91	70.3	84.57	79.11			0.6	5.71	4.2	
Arunachal Pradesh			76.2			81.02					4.82	

Data from http://pib.nic.in/newsite/PrintRelease.aspx?relid=105116 and http://pib.nic.in/newsite/efeatures.aspx?relid=104195.

Source: http://epaper.deccanchronicle.com/articledetailpage.aspx?id=474880;
Last accessed: April 27, 2014

Problem #4: Data Presentation





Source: http://www.ndtv.com/elections/assembly-cabinet/maharashtra
Last accessed: October 24, 2014

Problem #4: Data Presentation

COPS IN SOUP

Fearing law, traffic violators abandon seized vehicles

There are 3,421 unclaimed vehicles lying in city traffic police stations

ANUSHA PUPPALA | DC
HYDERABAD, JAN. 14

After several special drives and vehicles seizures by Hyderabad traffic police, there are 3,421 vehicles lying in the city traffic police stations, including 2,995 two-wheelers and 67 four wheelers. These vehicles are abandoned because many violators prefer leaving their vehicles instead of attending counselling sessions, facing court procedures and possible imprisonment for drunk driving.

Apart from different traffic violations, these vehicles are also seized

VEHICLE AUCTIONS CONDUCTED BY CITY POLICE

Phase	Scrap vehicles	Serviceable vehicles	Total vehicles	RTA fixed price	Realised price
1	2,596	2,596		₹45, 65,800	₹55, 76,000
2	488	488		₹29, 67,300	₹40, 69,900
3	641	199	840	₹23, 22,600	₹33, 79,500
4	774	774		₹9, 52,800	₹12, 20,000
5	435	1,435		₹31, 50,700	₹33, 75,000
6	2,315	2,315		₹4,55,9500	₹50, 12,000
7	684	320	1004	₹26,78,900	₹41,65,600
8	1,028	1,028		₹25,11,900	₹28,12,000
Total	9,473	1,007	10,480	₹2,37,09,500	₹2,96,10,000

Source: Deccan Chronicle, Hyderabad edition, December 9, 2016 and January 15, 2018

Cops shift focus, 50% dip in cases

Traffic cops chase helmet violations but see cut in cases filed against other culprits

DC CORRESPONDENT
HYDERABAD, DEC. 8

The number of cases booked for triple riding, cell phone driving and signal jumping have come down by over 50 per cent this year compared to previous year.

The drop, however, is neither due to enhanced enforcement by the traffic cops nor because of improved compliance to rules by motorists.

The likely cause of the dip is, among others, is the traffic cops' focus on drives against violation of specific rules like helmet rule violation. Consequently, the number of cases booked for helmet violation jumped from 1.34 lakh in 2015 to over 17 lakh in 2016.

During August and September this year, the traffic police was going slow on enforcement due to heavy rains. That was followed in November by demonitisation.

Indian Road Safety Federation Chief Functionary Mr Vinod Kumar Kanumala, said the drop cannot be taken as a benchmark to declare that violation of traffic rules has come down or compliance to rules has increased.

"A scientific study has to be done to check if violations are repeated by motorists, how many are first time violators and if they are given counseling. How the counselling helps needs to be seen," he said.

Social worker TS Gupta said the number of helmet violation cases would go up by 100 per cent if traffic cops were to intensify the drive with many bikers driving without helmets.

"Even the triple riding cases would be in large numbers if traffic cops enforce the rule strictly," he said.

Deputy Commissioner of Police (Traffic) Mr A V Ranganath admitted that a fall in number of cases booked this year compared to previous year cannot be attributed to only enforcement or improvement in rule compliance.

CASES BOOKED IN 2015 AND IN 2016 UP TO DECEMBER 5

	2015	2016
Triple riding:	73,549	31,704
Cell Phone Driving:	27,342	10,015
Signal Jumping:	51,725	16,105
Not wearing helmet:	1,34,092	17,20,169
Without number plate:	4,934	4351

'Cancer may strike due to bad luck, not lifestyle'

Random Changes In DNA During Cell Division Cause Nearly Two-Third Of All Cancers In Humans, Finds Study

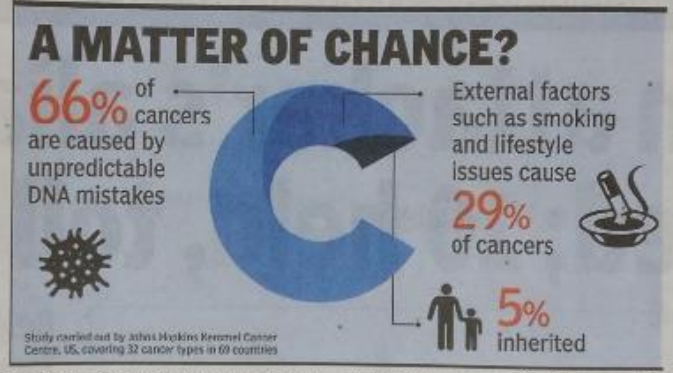
Subodh.Varma@timesgroup.com

Why does cancer strike some people and not others? New research shows that random changes or 'mistakes' in DNA when cells are dividing cause nearly two-third of all cancers in humans. These changes are neither caused by external factors like smoking or exposure to harmful chemicals, nor by hereditary factors. They are chance events occurring at the molecular level. In other words, cancer can strike anybody.

This upends prevailing wisdom that cancer is mostly a lifestyle related disease caused by external or environmental factors like smoking, harmful chemicals and conditions like obesity. While all these are valid and important risk increasing factors, random chance may be the real driver, if one goes by this new research.

Different types of cancers have different origins, according to the study. For example, in pancreatic cancers, 77% are due to random DNA copying errors, 18% to environmental factors, such as smoking, and the remaining 5% to heredity. In other cancer types, such as those of the prostate, brain or bone, more than 95% of the mutations are due to random copying errors.

Lung cancer is most likely to be caused by environmental factors, mostly smoking. About 65% of all the mutations are due to smoke and 35% due to DNA copying errors. Inherited factors have negligible role. The study involved a statistical analysis of cancer data from 69 countries including India, representing 4.8 billion people, more than half of the world's population. It was done by scientists from Johns Hopkins Kimmel Cancer Center at Baltimore, US, and published in the peer reviewed journal Science on March 24.



Human bodies grow by constant division of cells, starting from the first cell formed by fusion of the male sperm with the female egg. Every time a cell divides into two, the genetic code carrying DNA is copied. What the scientists are saying is that mistakes occur in this copying process that accumulate over time and ultimately cause cancer. "These copying mistakes are a potent source of cancer mutations that historically have been scientifically undervalued, and this new work provides the first estimate of the fraction of mutations caused by these mistakes," said the paper's lead author Cristian Tomasetti.

The researchers studied all 32 cancer types and estimated that 66% of cancer mutations result from copying errors, 29% can be attributed to lifestyle or environmental factors, and the remaining 5% are inherited. They found a strong correlation between cancer incidence and normal cell divisions among 17 cancer types, regardless of the countries' environment or stage of economic development. This means that lifestyle factors like smoking or exposure to toxic chemicals are also very important factors causing nearly a third of cancers.

"We need to continue to encourage people to avoid environmental agents and lifestyles that increase their risk of developing cancer mutations," co-author Bert Vogelstein emphasised.

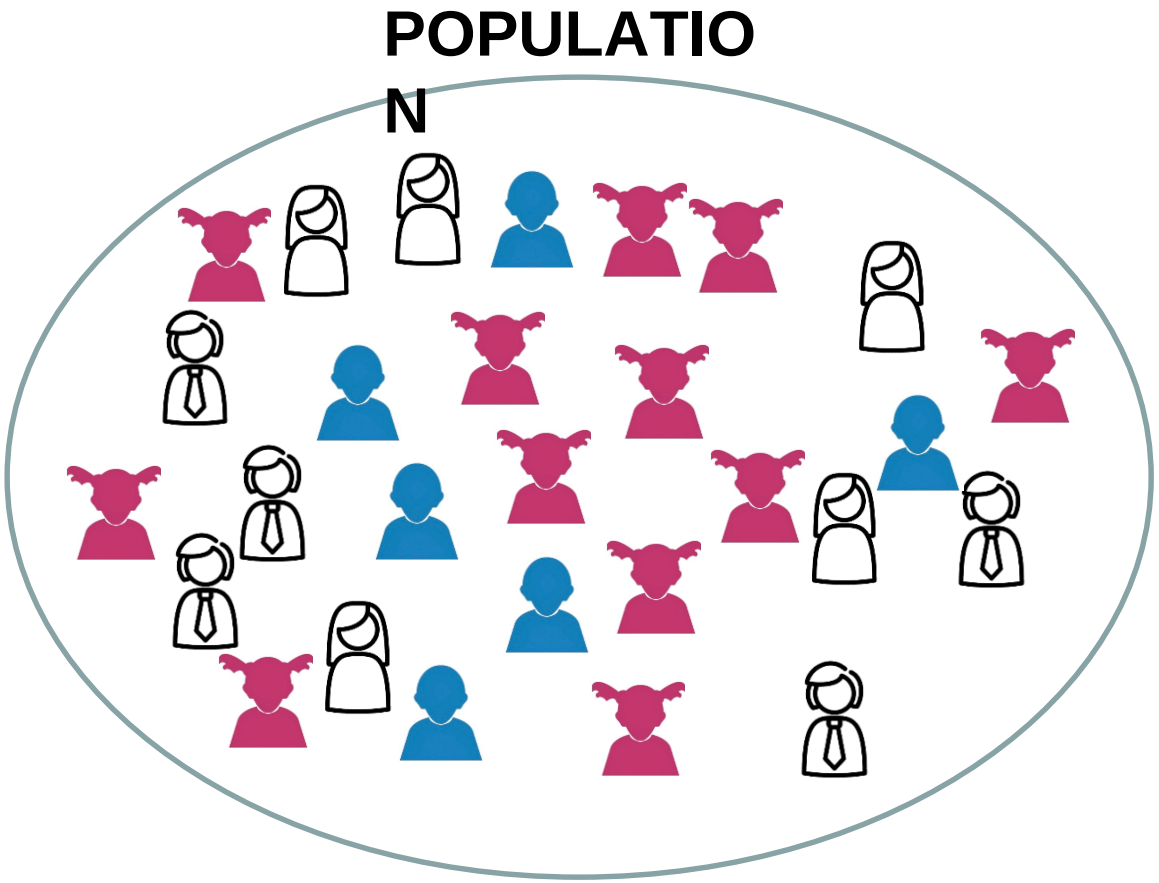
Why Study Statistics?

Statistics don't lie but Statisticians will in any of the following situations:

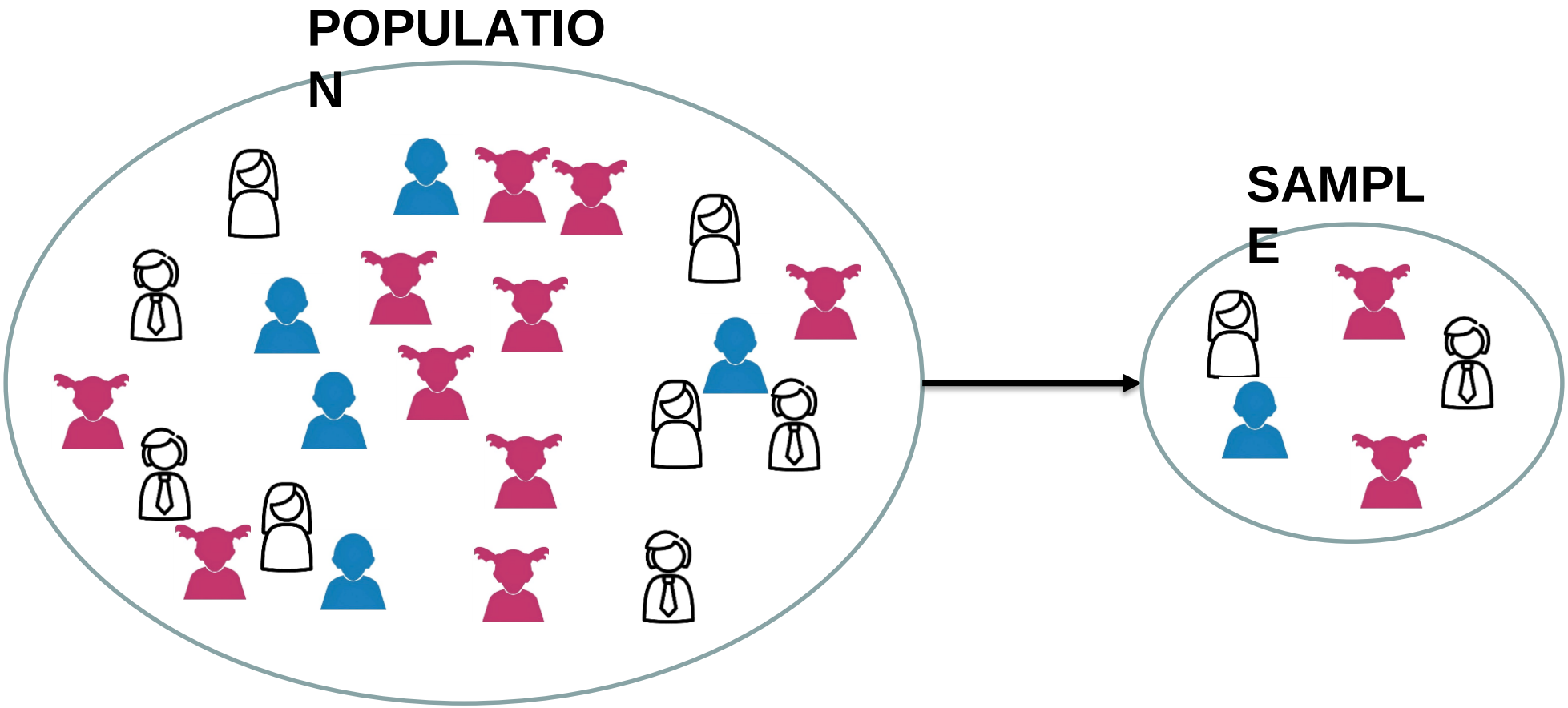
- Data Gathering
- Data Understanding
- Data Analysis/Interpretation
- Data Presentation

BASIC STATISTICAL TERMINOLOGY

Population and Sample



Population and Sample



Name	Ht.	Hometown	Class
Cheyanne Bustle	5'0"	Prestonburg, KY	Fr.
Jaclyn Fyffe	5'3"	Richmond, KY	Fr.
Brooke Gibbs	4'11"	Pineville, KY	So.
Michelle Malavasi	4'10"	Heredia, Costa Rica	So.
Madison Mullin	5'2"	Georgetown, KY	Fr.
Dallas Pringle	5'2"	Reno, NV	Fr.
Chelsee Ramos	5'2"	Madison, WI	Jr.
Sydney Shelton	4'10"	Scottsville, KY	Jr.
Ashley Wettstain	5'0"	Owensboro, KY	Fr.
Madison Yee	5'2"	San Marcos, CA	So.

Source: http://www.ukathletics.com/trads/cheer-roster.html
Last accessed: October 7, 2014

No.	Name	Pos.	Cl.-Exp.	Ht.	Hometown/High School/Last College
0	Jennifer O'Neill	PG	SR-3L	5-6	Bronx, N.Y./Saint Michael Academy
2	Ivana Jakubcova	C	JR-JC	6-6	Bratislava, Slovakia//Murray State College
3	Janee Thompson	PG	JR-2L	5-7	Chicago, Ill./Whitney Young
5	Kyvin Goodin-Rogers	F	SO-HS	6-1	Lebanon, Ky./Marion Co.
12	Jelleah Sidney	F/C	SR-2L	6-2	Queens Village, N.Y./Saint Michael Academy/Chipola JC
13	Bria Goss	G	SR-3L	5-10	Indianapolis, Ind./Ben Davis
15	Linnae Harper	G	SO-1L	5-8	Chicago, Ill./Whitney Young
24	Jaycee Coe	G	FR-HS	5-11	Gainesboro, Tenn./Jackson Co.
25	Makayla Epps	G	SO-1L	5-10	Lebanon, Ky./Marion Co.
35	Alexis Jennings	F/C	FR-HS	6-2	Madison, Ala./Sparkman
45	Alyssa Rice	C	FR-HS	6-3	Reynoldsburg, Ohio/Reynoldsburg
50	Azia Bishop	F/C	SR-3L	6-3	Toledo, Ohio/Start

Source: http://www.ukathletics.com/sports/w-baskbl/mtt/kty-w-baskbl-mtt.html
Last accessed: October 7, 2014

Source: http://www.dailymail.co.uk/news/article-2742468/Tall-small-s-basketball-Ladies-Kentucky-Wildcats-team-tower-cheerleaders.html
Last accessed: October 7, 2014

One day there was a fire in a wastebasket in the office of the Dean of Sciences. In rushed a physicist, a chemist, and a statistician.

The physicist immediately starts to work on how much energy would have to be removed from the fire to stop the combustion.

The chemist works on which reagent would have to be added to the fire to prevent oxidation.

While they are doing this, the statistician is setting fires to all the other wastebaskets in the office.

“What are you doing?” the others demand. The statistician replies, “Well, to solve the problem, you obviously need a larger sample size.

Census and Survey

Census: Gathering data from the **whole population** of interest.

For example, elections, 10-year census, etc.

Survey: Gathering data from the **sample** in order to make conclusions about the population.

For example, opinion polls, quality control checks in manufacturing units, etc.

Census and Survey

VOTER PULSE

KCR orders massive surveys ahead of polls

MP, MLA performance to be analysed

CH.V.M. KRISHNA RAO | DC
HYDERABAD, JAN. 21

Chief Minister K. Chandrasekhar Rao has, for the first time, engaged three private agencies to conduct separate large-scale surveys on the performance of the TRS government, the party and the MLAs/MPs across the state.

In all the earlier surveys, there were around 250 to 300 respondents from each Assembly segment, which totals 35,000 respondents. This time, the surveys will gather opinions from 3,000 persons from each Assembly segment, taking the total sample to around 3.5 lakh for each survey agency and all the three agencies put together it will cross one million.

These three simultaneous surveys will gather information on the performance of various government schemes, such as whether they are reaching people or not, what more they need from the government etc.

- Three agencies will survey 3,000 persons from each Assembly segment.
- Each agency will survey 3.5 lakh in all.
- Total sample to cross 1 million.

Page 6: Crores to be spent on surveys

Survey is key for tickets

DC CORRESPONDENT
HYDERABAD, JAN. 21

Chief Minister K. Chandrasekhar Rao has, for the first time, engaged three private agencies to conduct separate large-scale surveys on the performance of the TRS government, the party and the MLAs/MPs across the state.

Second will be an assessment on the performance of the TRS party — whether it is active or not, whether the party leaders are visiting their constituencies, whether any change is required in party policies etc.

Third, the survey will assess the performance of each MLA/ MP as to their functioning, tours of their constituencies, attending to the problems of people, whether they are capable of winning from the segment again or not.

Sources said crores of rupees are being spent on conducting these surveys. The survey reports are likely to reach the Chief Minister by January end. "Based on the reports, the Chief Minister will plan his strategy on government, party, and MLA/MP's future performance. He had announced earlier that 99 per cent of the present MLAs/MPs will be fielded again in the coming elections, but now a decision will be made after the survey report comes in," said a key source.

- **THOUGH KCR** had earlier announced that 99 per cent of the present MLAs/MPs will be fielded again in the coming elections, the decision on party tickets will be made after the survey report comes in

Source: Deccan Chronicle, Hyderabad edition, December 8, 2017 and January 22, 2018

Parameter and Statistic

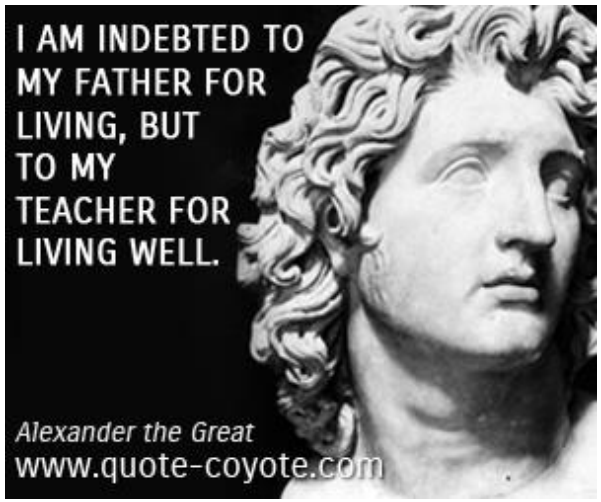
Parameter: A descriptive measure of the **population**.

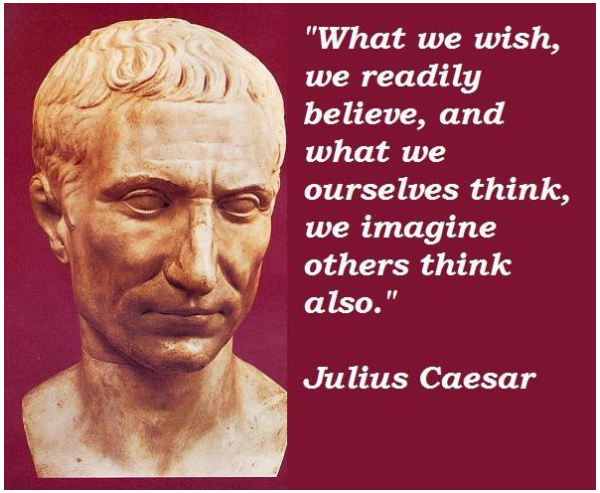
For example, population mean, population variance, population standard deviation, etc.

Statistic: A descriptive measure of the **sample**.

For example, sample mean, sample variance, sample standard deviation, etc.

Parameter and Statistic





Greek – Population Parameter

Mean – μ

Variance – σ^2

Standard Deviation - σ

Roman – Sample Statistic

Mean – $\bar{x}$

Variance – s^2

Standard Deviation - s

Y= ax+b or Y = αx+ β

(α, β – Parameter)

(a, b Sample)

Descriptive and Inferential Statistics

- Descriptive Statistics – Data gathered about a group (sample or population) to reach conclusions about the same group. Ex. Mean, Median …..

- Inferential Statistics – Data gathered from a sample and the statistics generated to reach conclusions about the population from which the sample is

1 Diabetes is a huge problem in India.

The prevalence of diabetes increased tenfold, from 1.2% to 12.1%, between 1971 and 2000.

Noncommunicable Diseases in the Southeast Asia Region, Situation and Response, World Health Organization, 2011. http://apps.searo.who.int/PDS_DOCS/B4793.pdf

It is estimated that 61.3 million people aged 20-79 years live with diabetes in India (2011 estimates). This number is expected to increase to 101.2 million by 2030.

David R. Whiting, et al. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030, Diabetes Research and Clinical Practice, Volume 94, Issue 3, December 2011, Pages 311-321, http://www.sciencedirect.com/science/article/pii/S0168822711005912)

And, 77.2 million people in India are said to have pre-diabetes.

Anjana RM, Pradeepa R, Deepa M, Datta M, Sudha V, Unnikrishnan R, et al. "Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research-India Diabetes (ICMR-INDIAB) study" Diabetologia 54.12 (2011): 3022-7. NCBI. Web. March 2013.

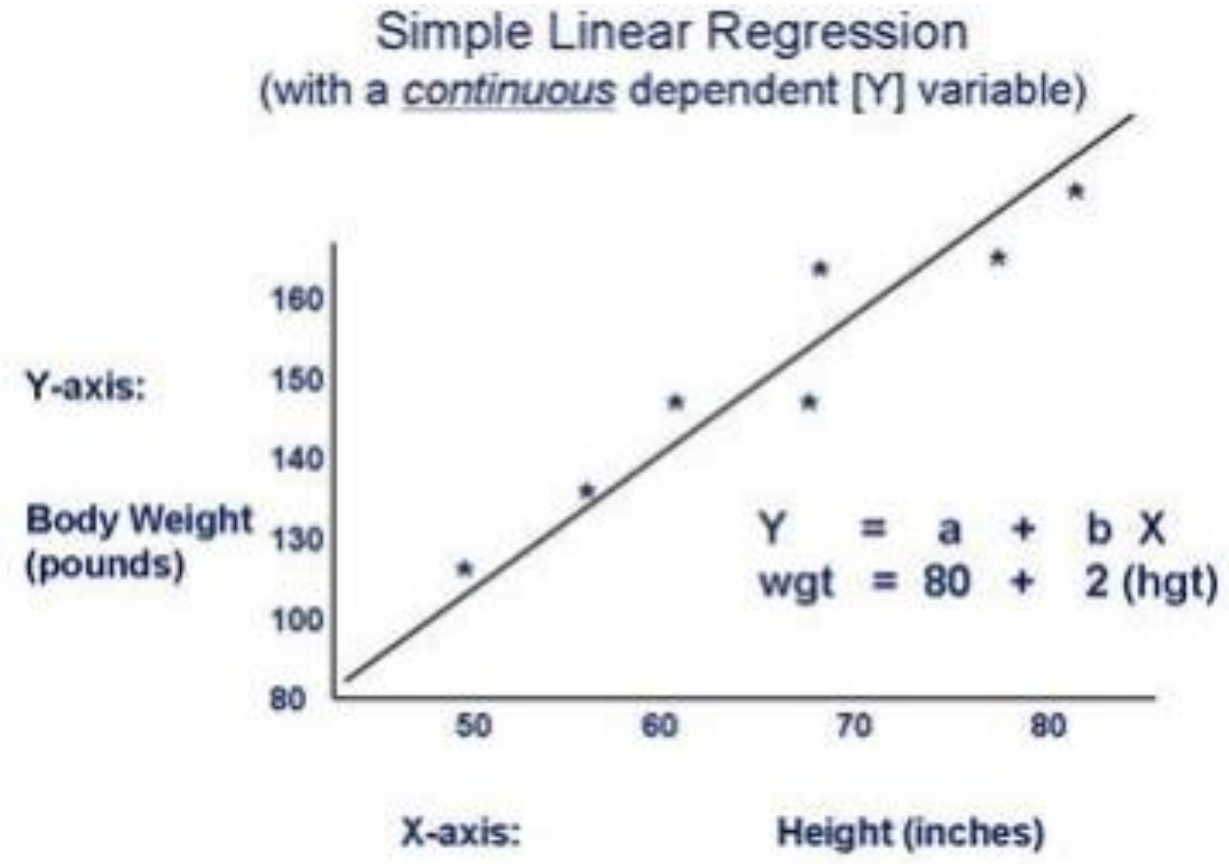
Source: http://www.arogyaworld.org/wp-content/uploads/2010/10/ArogyaWorld_IndiaDiabetes_FactSheets_CGI2013_web.pdf
Last accessed: November 25, 2015

Variables and Data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no
32	blue-collar	single	primary	no	23	yes	yes	unknown	5	may	160	1	-1	0	unknown	no
25	services	married	secondary	no	50	yes	no	unknown	5	may	342	1	-1	0	unknown	no
40	retired	married	primary	no	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no

Variables – Dependent and Independent

- Dependent variables on y-axis and Independent on x-axis.
- Dependent variable (what I want to predict) also called Target variable or Class variable.



Data: numeric and categorical

18 kg

27 kg

Sources: http://banglanews24.com/en/files/2013August/SM/Gold-sm20130830024804.jpg, http://myoor.com/wp-content/uploads/2014/01/gold.jpg and http://im.rediff.com/cricket/2014/feb/01india1.jpg
Last accessed: November 22, 2014

Categorical Data (Qualitative)

Nominal

Examples

- Employee ID
- Gender
- Religion
- Ethnicity
- Pin codes
- Place of birth
- Aadhaar numbers

Ordinal

Examples

- Mutual fund risk ratings
- Fortune 50 rankings
- Movie ratings

Product rating Online sites

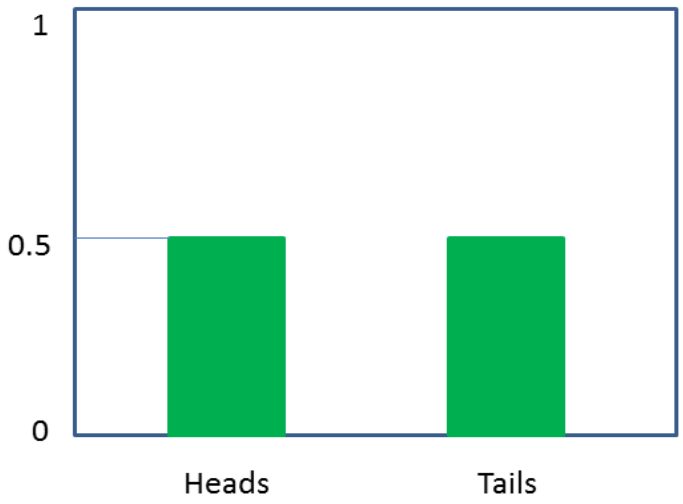
While there is an order, difference between consecutive levels are not always equal.

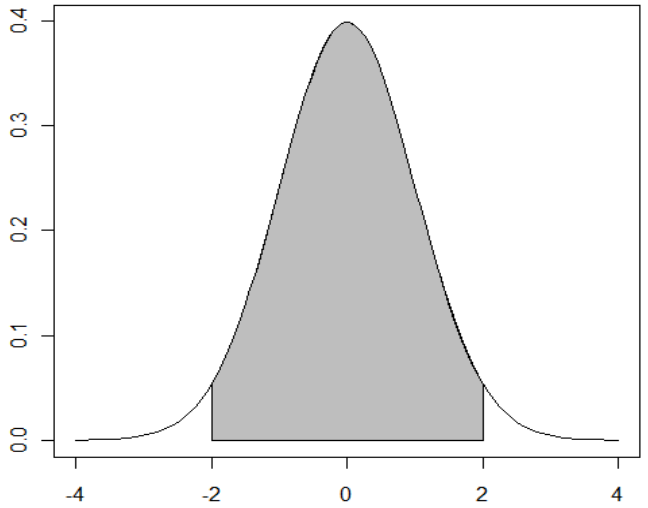
Numeric Data (Quantitative)

Examples

- Height
- Weight
- Time
- Volume
- Number of iPads sold
- Number of complaints received at the call centre
- Number of employees
- Percentage return on a stock
- Rupee change in stock price

Discrete and Continuous





Countable

Measurable

Discrete or Continuous?

BREAK

Time between customer arrivals at a retail outlet	Continuous
Sampling 100 voters in an exit poll and determining how many voted for the winning candidate	Discrete
Lengths of newly designed automobiles	Continuous
No. of customers arriving at a retail outlet during a five-minute period	Discrete
No. of defects in a batch of 50 items	Discrete

DESCRIBING DATA THROUGH STATISTICS

The Central Tendencies - Mean

45, 34, 100, 33 – Runs Scores by a Batsman in four T20 Matches

Runs	**45**	**34**	**100**	**33**
Match	1	2	3	4

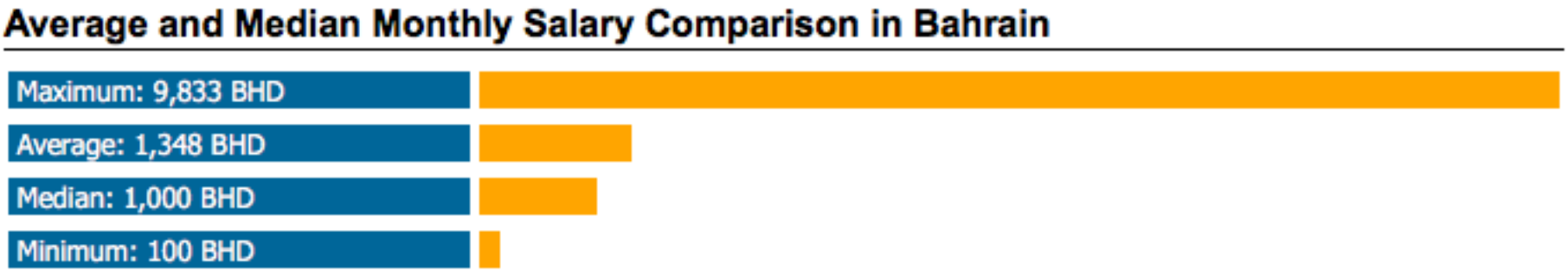
$$\text{Mean, } \mu = \frac{\Sigma x}{n} = \frac{45+34+100+33}{4} = 53$$

60, 60, 70, 45, 34, 45, 45, 81 – Runs Scored by a Batsman in eight T-20 Matches

Runs	**60**	**70**	**45**	**34**	**81**
Frequency	2	1	3	1	1

- $\text{Mean, } \mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{60X2+70X1+45X3+34X1+81X1}{2+1+3+1+1} = 55$

The Central Tendencies - Mean



Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2

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

Source: http://www.salaryexplorer.com/salary-survey.php?loc=17&loctype=1
Last accessed: May 17, 2016

The Central Tendencies

NEWS 18 .COM | Home | Politics | Assam NRC | India | Opinion | Movies | Tech | Auto | Buzz | Videos | Cricket

Google CEO Sundar Pichai Earns Over Rs 3.52 Crore Per Day As Salary

Google CEO Sundar Pichai received a salary of USD 650,000 last year, slightly less than the USD 652,500 he earned in 2015.

PTI | Updated:April 29, 2017, 3:05 PM IST

gle CEO Sundar Pichai Earns Over Rs 3.52 Crore Per Day As... | Crimes Through Fake News Are Unacceptable, Find Tech Soluti... | Tesla Quarterly Report

The Central Tendencies - Median

Median: Arrange data in increasing order and find the mid-point $\frac{(n+1)}{2}$.

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2

100,100,100,100,100,100,100,100,100,100,

345,1000,1000,1000,1000,1000,1000,1000,

1000,1000,1000,9833,9833

n = 23 – Sample Size

Median =

Median = 12th place whose value is = "1000"

The Central Tendencies - Excel

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2

The Central Tendencies - Median

Median: Arrange data in increasing order and find the mid-point .

34, 45, 45, 45, 60, 60, 70, 81,

n = 8

$$\text{Median} = \frac{(8+1)}{2} = \frac{9}{2} = 4.5$$

Median = 4.5 which means Median is between 4th and 5th value

$$\text{Median} = \frac{(45+60)}{2} = \mathbf{52.5}$$

Source: http://www.business-standard.com/article/companies/ambani-gets-205-times-ril-s-median-pay-115070500340_1.html
Last accessed: July 7, 2015

The Central Tendencies - Mode

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2

What is the Mode – the most frequently occurring data point?

Mode - 1000 and 100 (Bimodal)

The Central Tendencies

Mean and Median need not be in the dataset but Mode has to be in it.

Mode is also the only average that works with categorical data.

The Central Tendencies

The management of Good Heart Inc. wants to give all its employees a raise. They are unable to decide if they should give a straight Rs 2000 to everyone or to increase salaries by 10% across the board. The mean salary is Rs 50,000, the median is Rs 20,000 and the mode is Rs 10,000.

How do these central tendencies change in both cases?

The Central Tendencies

The management of “BigSteel” wants to decrease the production of number of steel sheets. They are unable to decide if they should cut 3,000 per week or cut by 20% production of steel sheets. The mean production of steel sheets is 80,000 per week , the median is 30,000 and the mode is 15,000 per week

How do these central tendencies change in both cases?

Measuring Variability and Spread

Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

Points scored per game	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency, f	1	1	2	3	2	1	1

Points scored per game	**7**	**9**	**10**	**11**	**13**
Frequency, f	1	2	5	2	1

Points scored per game	**3**	**6**	**7**	**10**	**11**	**13**	**30**
Frequency, f	2	1	2	3	1	1	1

Mean = Median = Mode = 10 for all 3.

Measuring Variability and Spread

Range = Max - Min

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	3	2	1	1

Range = 13-7 = 6

Points scored per game	7	9	10	11	13
Frequency, f	1	2	5	2	1

Range = 13-7 = 6

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

Range = 30-3 =27

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

3 3 6 7 7 10 10 10 11 13 30 n=11

Lower quartile (25th percentile, Q1) = th

Middle quartile = Median = th

Upper quartile (75th percentile, Q3) = th

Interquartile range, IQR = Q3-Q1 (central 50% of data)

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

3 3 6 7 7 10 10 10 11 13 30 n=11

Lower quartile (25th percentile, Q1) =

Middle quartile = Median =

Upper quartile (75th percentile, Q3) = = 9th value =“11”

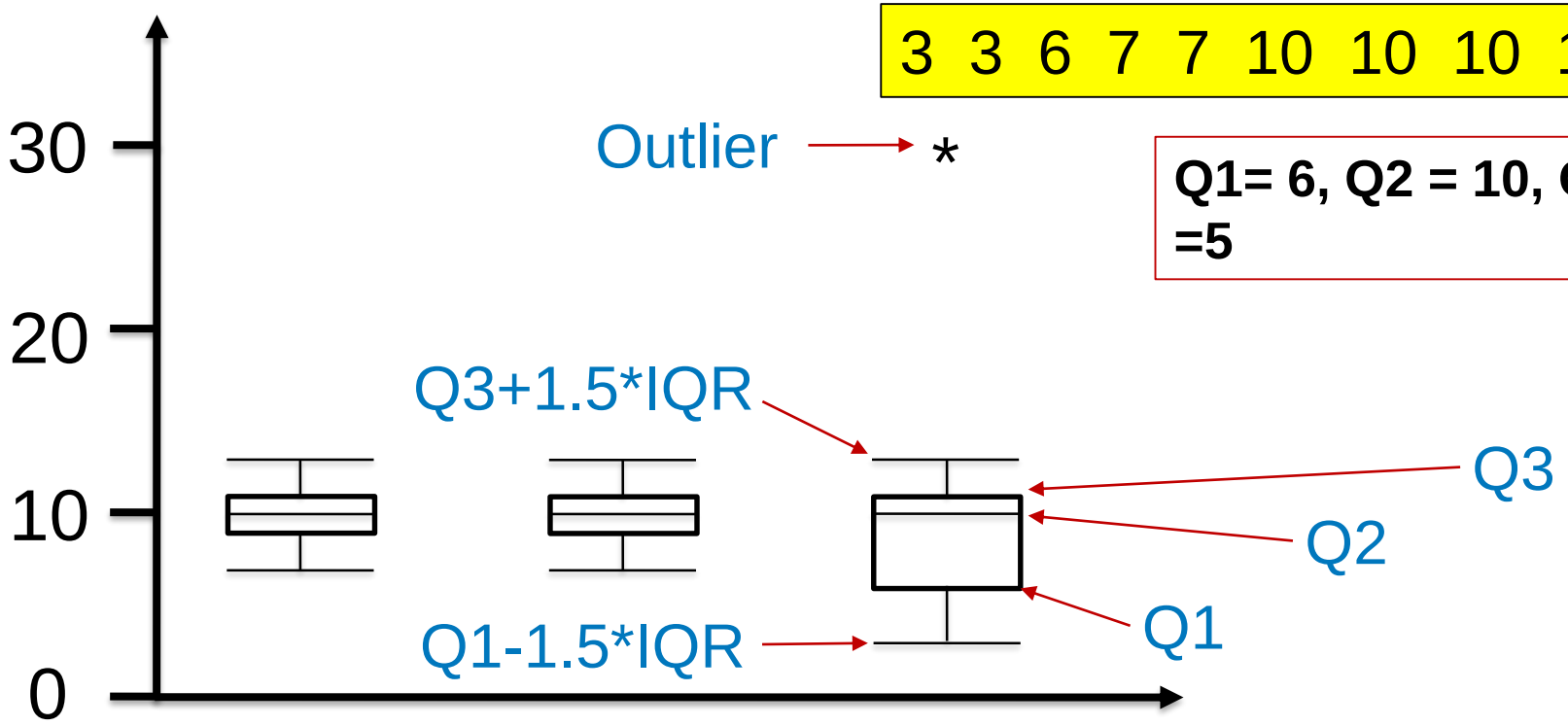
Interquartile range, IQR = Q3-Q1 =9th value-3rd value = “11”-“6”= 5

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Box and whisker diagram or Box plot

3 3 6 7 7 10 10 10 11 13 30



Q1= 6, Q2 = 10, Q3=11, IQR =5

Calculated value of Upper Whisker is 18.5. Value in **data set closet** to 18.5 but **less than** 18.5 **if the calculated value is not present in the data set – In our example it is 13**

Calculated value of Lower Whisker is **-1.5** Value in **data set** closet to -1.5 but **more than** -1.5 **if the calculated value is not present In the data set – In our example it is 3**

Outlier detection – Excel and Box Plot Steps

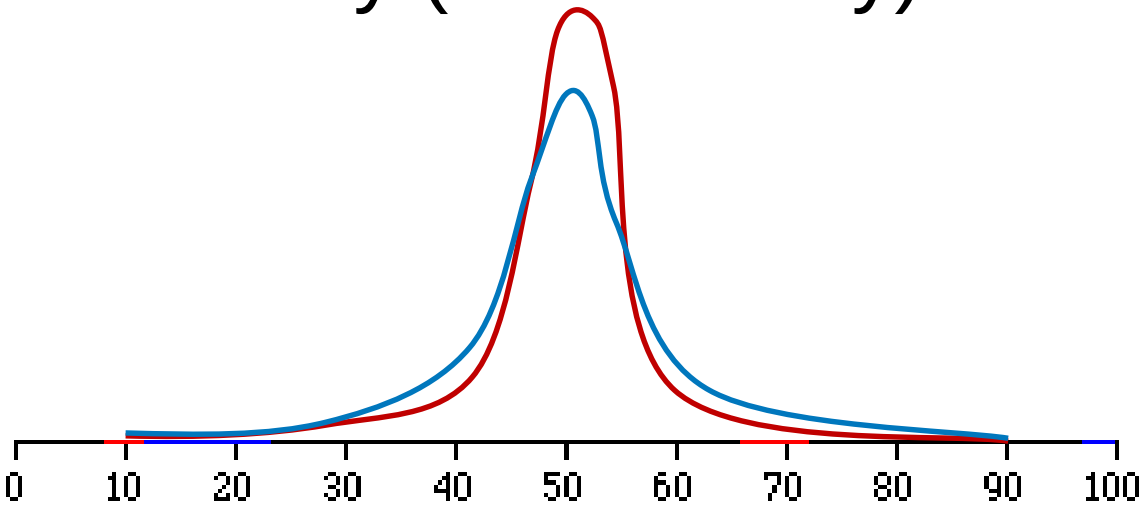
Hadlum vs Hadlum case (**Excel Sheet Pregnancy**)

Source: http://www.alphamom.com/legacy/pregnancy-calendar/week36.jpg
Last accessed: November 01, 2014

Source: http://3.bp.blogspot.com/-0YwIRjLMWr0/T4DqOwVClgI/AAAAAAAAAgg/Yjf-ttkQLSg/s1600/fishy.jpg
Last accessed: November 01, 2014

Measuring Variability and Spread

Range and IQR give the spread but still do not describe variability (consistency).



Average distance from the mean?

3 3 6 7 7 10 10 10 11 13 30

Measures of Spread – Mean Distance, Mean Absolute Deviation or Standard Deviation – Excel Sheet Deviations

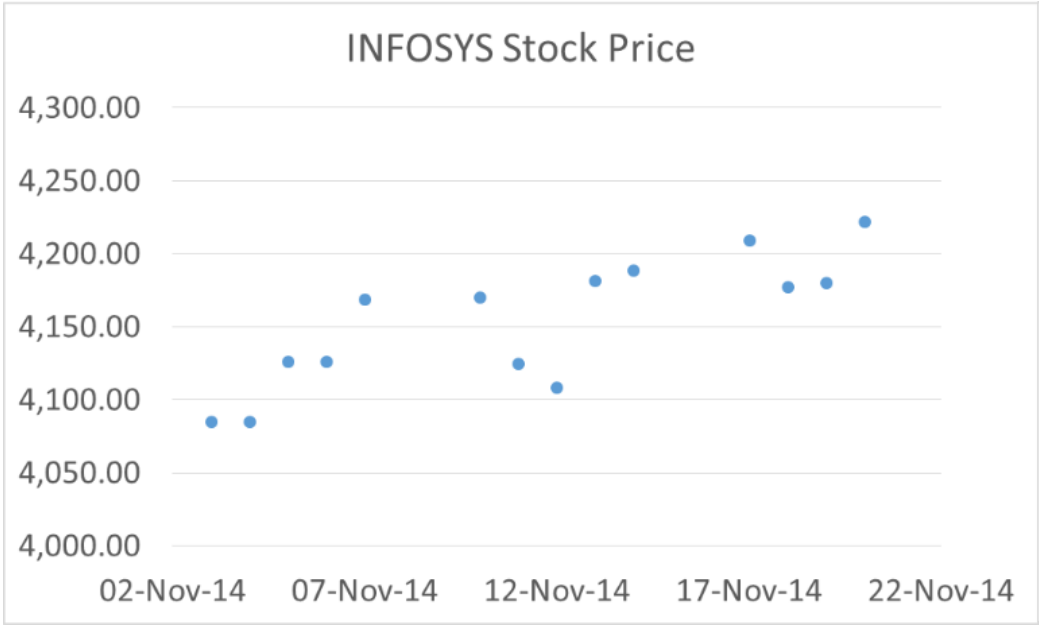


Fig:1 Real data collected over 20 days

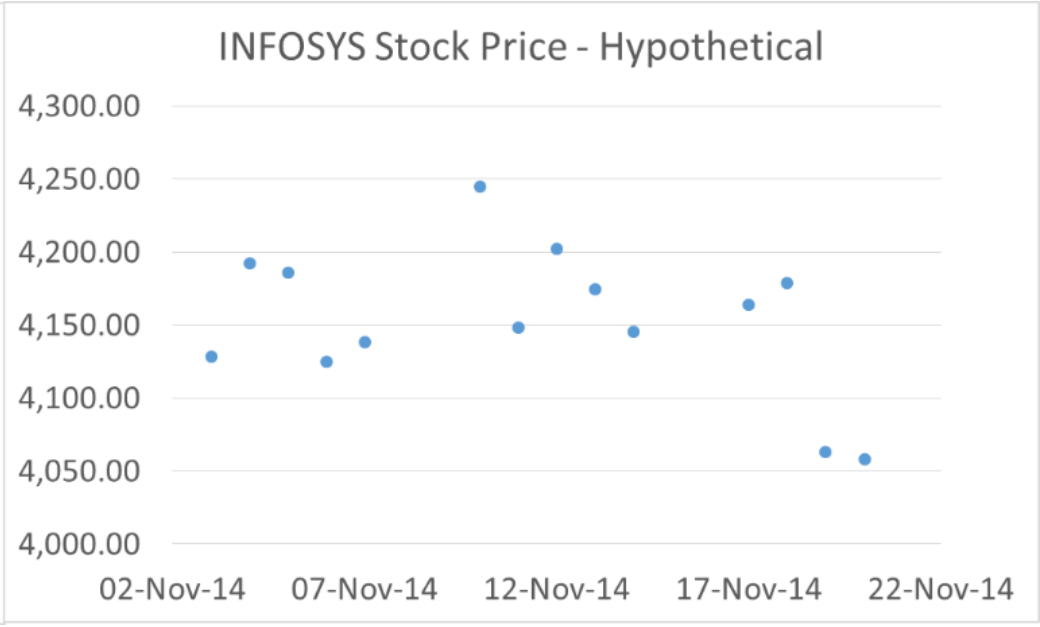
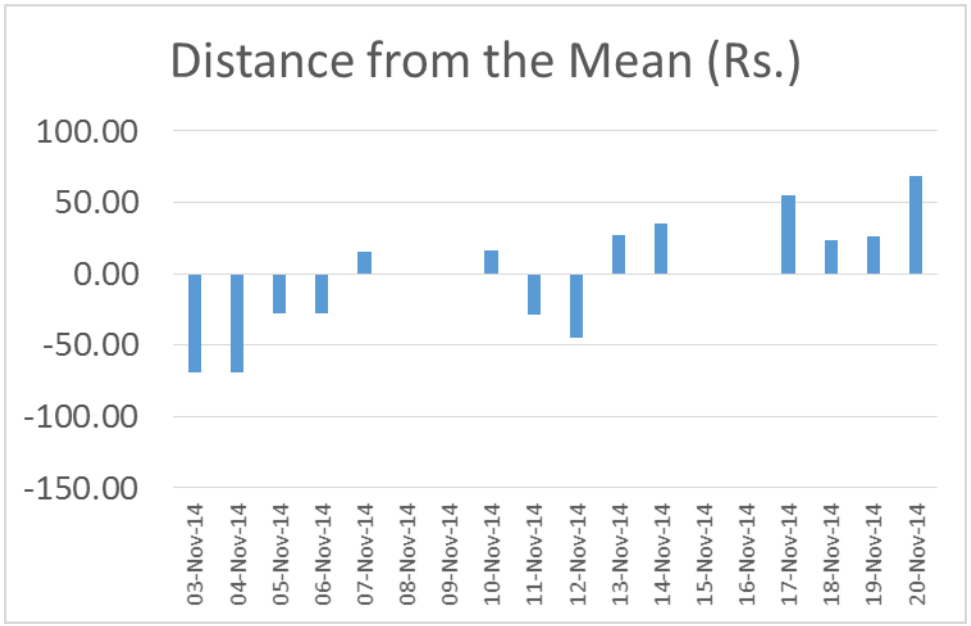
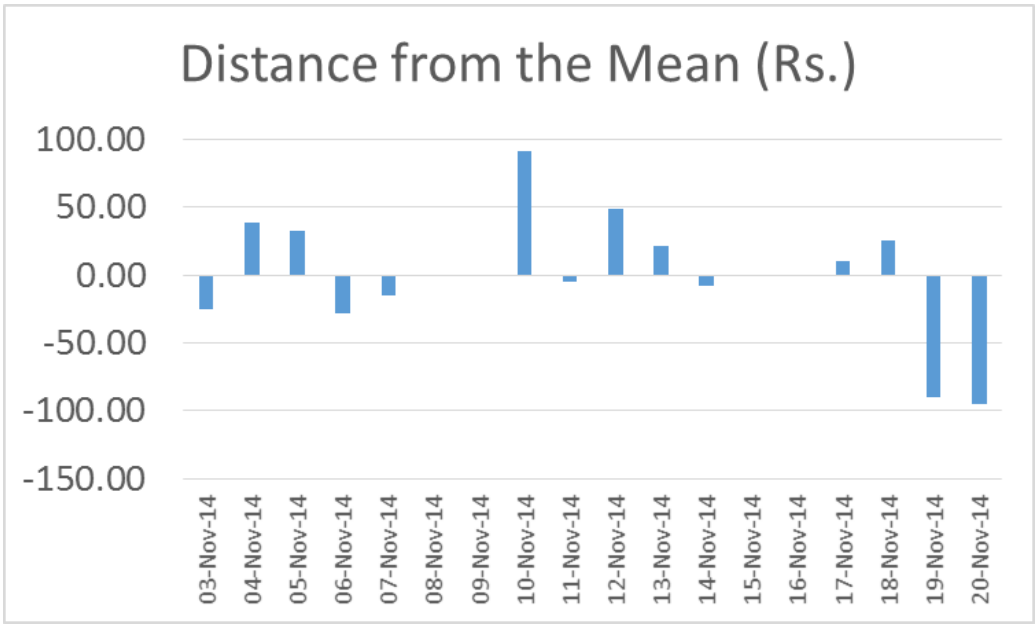


Fig:2 Assumed data for 20 days

ata Source: https://in.finance.yahoo.com/q/hp?s=INFY.BO

Measures of Spread – Mean Distance, Mean Absolute Deviation or Standard Deviation - Excel





- Mean Distance in both cases = 0
- Mean Absolute Deviation in both cases = 38.17
- Std Dev is 42.54 in the first case and 48.80 in the second.

ata Source: https://in.finance.yahoo.com/q/hp?s=INFY.BO

Measuring Variability and Spread

Variance = (Derive)

3 3 6 7 7 10 10 10 11 13 30

Units are squared, which is not intuitive.

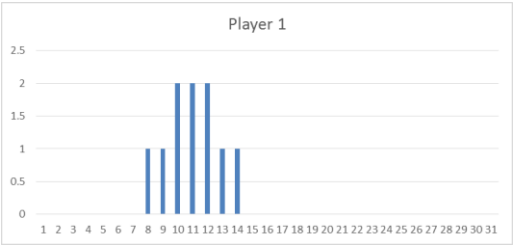
Standard Deviation,

Measuring Variability and Spread

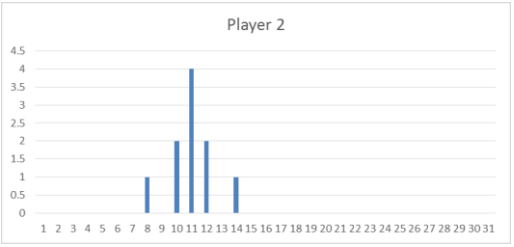
BREAK

Calculate standard deviation for each player.

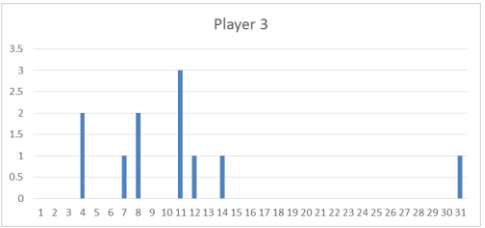
Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	3	2	1	1



Points scored per game	7	9	10	11	13
Frequency, f	1	2	5	2	1



Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1



Mean is 10

1.65, 1.41, 7.02 – Standard Deviation

Player 3 is the least reliable.

Measuring Variability and Spread

What happens to Standard Deviation if Good Heart Inc. gave all employees a Rs 2000 raise?

What happens to Standard Deviation if Good Heart Inc. gave all employees a 10% raise?

No change.

Increases by 1.1 times.

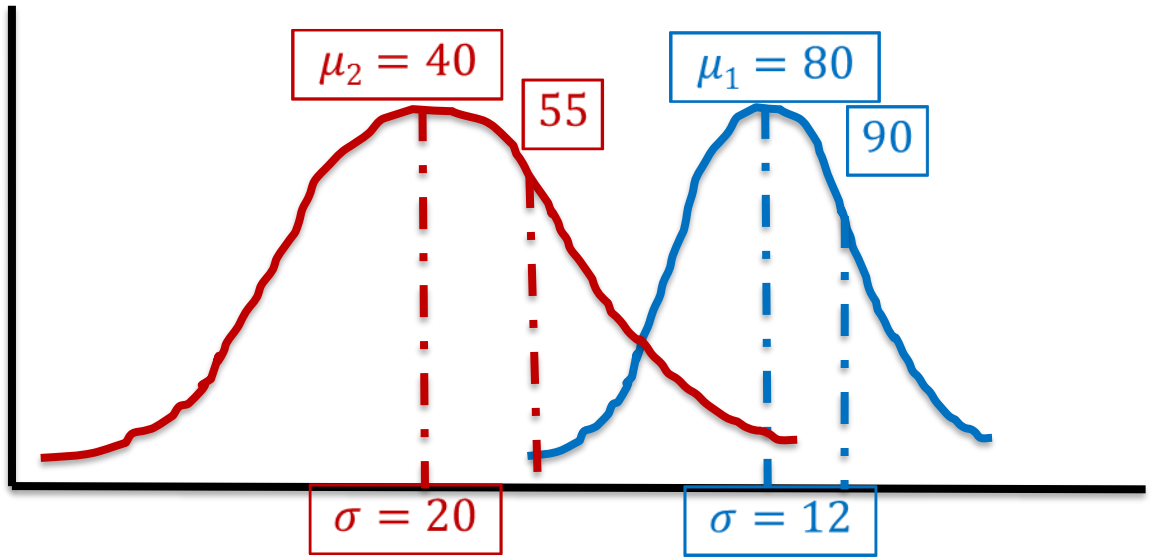
Measuring Variability and Spread

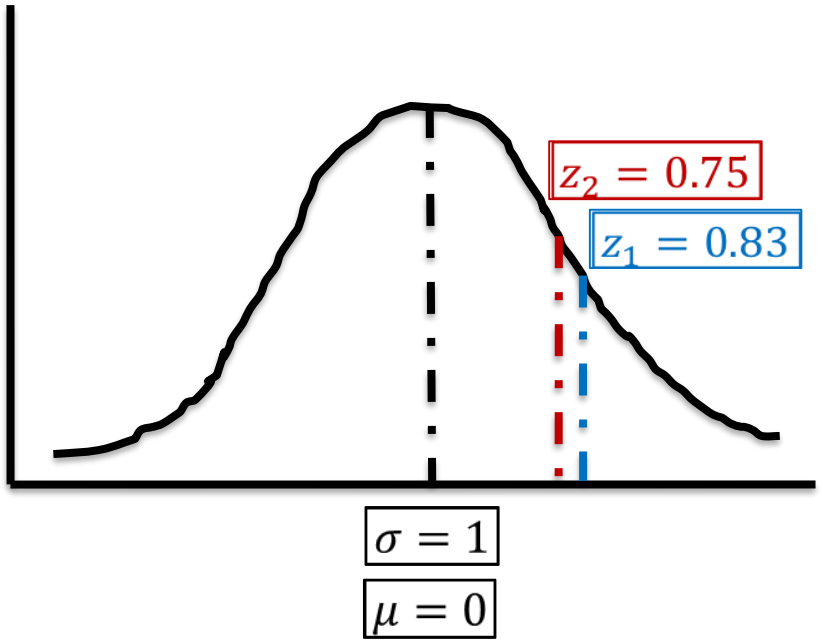
Imagine 2 players with different abilities: one has an average of 80% with 12% Stdev and the other 40% with 20% Stdev.

In a particular practice session, the first one scores 90% of the time and the second 55%. Who did best against their PERSONAL track record?

Measuring Variability and Spread

Standard score, , # of stdevs from the mean





- Standard score, $z = \frac{x-\mu}{\sigma}$, # of stdevs from the mean

Z – Score

Before entering Eng. Degree or Administrate services students have to take CET or UPSE exams. The following are the mean and standard deviation of the two exams

Exam	Mean (μ)	Standard Deviation (σ)
CET	150	12
UPSE	46	5

Prakash took both the exams and scored 164 in CET and 54 in UPSE. Which exam did he do relatively better ?

$$Z_{CET} = \frac{x - \mu}{\sigma}$$

$$Z_{CET} = \frac{164 - 150}{12}$$

$$Z_{CET} = 1.16$$

$$Z_{UPSE} = \frac{x - \mu}{\sigma}$$

$$Z_{UPSE} = \frac{54 - 46}{5}$$

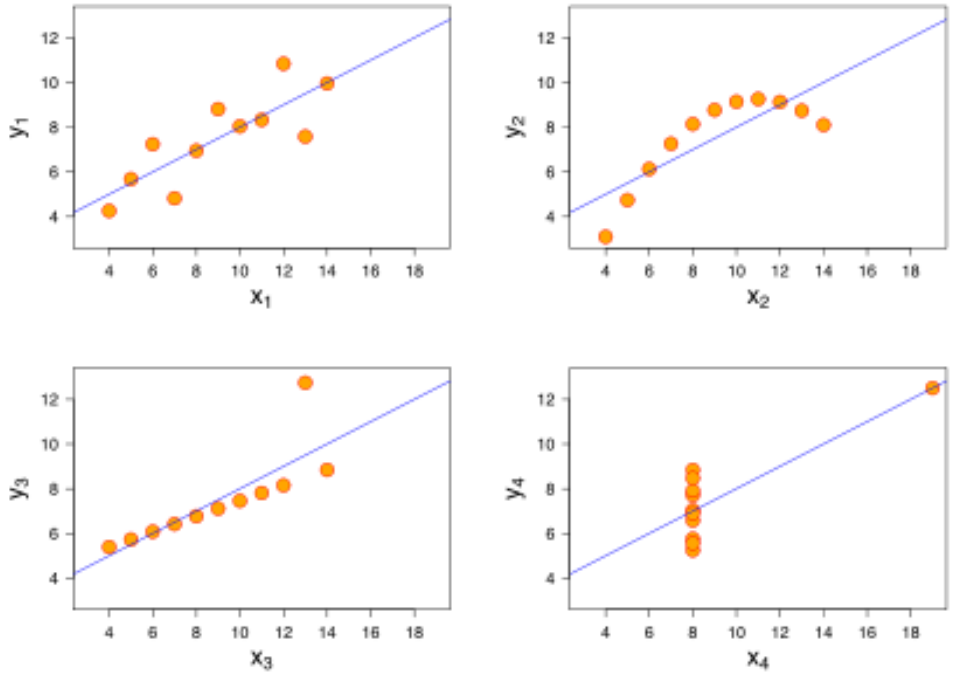
$$Z_{UPSE} = 1.6$$

Prakash did slightly better on the UPSE exam

Measuring Variability and Spread

Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.1	10	7.46	8	6.6
8	6.95	8	8.1	8	6.77	8	5.8
13	7.58	13	8.7	13	12.7	8	7.7
9	8.81	9	8.8	9	7.11	8	8.8
11	8.33	11	9.3	11	7.81	8	8.5
14	9.96	14	8.1	14	8.84	8	7
6	7.24	6	6.1	6	6.08	8	5.3
4	4.26	4	3.1	4	5.39	19	13
12	10.8	12	9.1	12	8.15	8	5.6
7	4.82	7	7.3	7	6.42	8	7.9
5	5.68	5	4.7	5	5.73	8	6.9

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)



"It's easy to lie with statistics. It's hard to tell the truth without statistics."

- Andrejs Dunkels – Teacher and Author

PROBABILITY BASICS

Sholay

Probability vs Statistics

- Probability – Predict the likelihood of a future event
- Statistics – Analyse the past events

- Probability – What will happen in a given ideal world?
- Statistics – How ideal is the world?

Probability vs Statistics

Probability is the basis of inferential statistics.

Probability - Applications

Table 1. Life table for the total population: United States, 2003

Age	Probability of dying between ages x to $x+1$	Number surviving to age x	Number dying between ages x to $x+1$	Person-years lived between ages x to $x+1$	Total number of person-years lived above age x	Expectation of life at age x
	q_x	l_x	d_x	L_x	T_x	e_x
0–1	0.006865	100,000	687	99,394	7,743,016	77.4
1–2	0.000469	99,313	47	99,290	7,643,622	77.0
2–3	0.000337	99,267	33	99,250	7,544,332	76.0
3–4	0.000254	99,233	25	99,221	7,445,082	75.0
4–5	0.000194	99,208	19	99,199	7,345,861	74.0
5–6	0.000177	99,189	18	99,180	7,246,663	73.1
6–7	0.000160	99,171	16	99,163	7,147,482	72.1

Insurance industry uses probabilities in actuarial tables for setting premiums and coverages.

Probability - Applications

Gaming industry – Establish charges and payoffs

HR – Does a company have biased hiring policies?

Manufacturing/Aerospace – Prevent major breakdowns

Assigning Probabilities

Classical Method – *A priori* or Theoretical

Probability can be determined prior to conducting any experiment.

Example: Tossing of a fair die

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Probability can be determined post conducting a thought experiment.

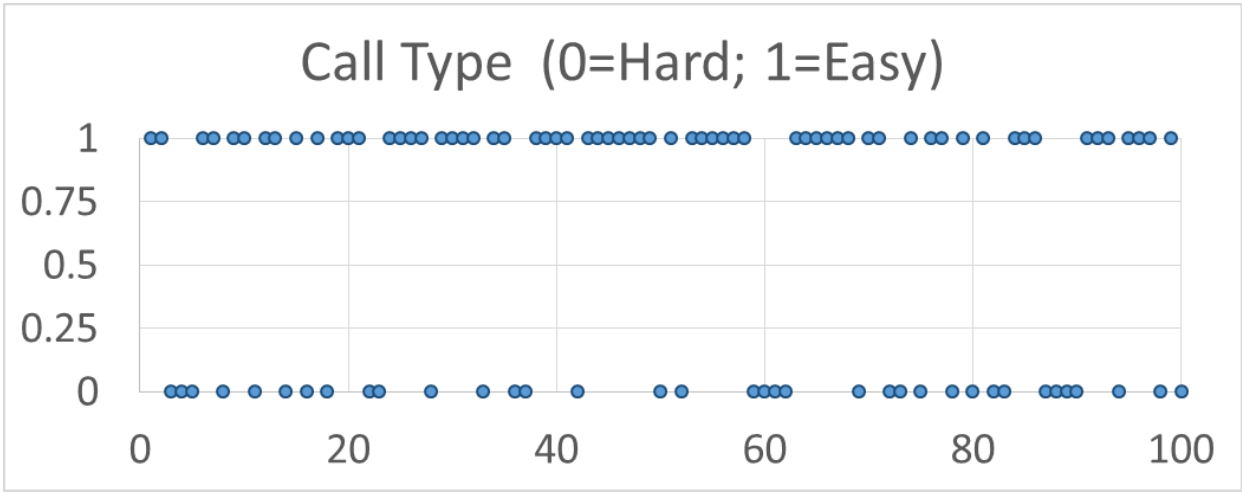
Example: Tossing of a weighted die…well!, even a fair die. The larger the number of experiments, the better the approximation.

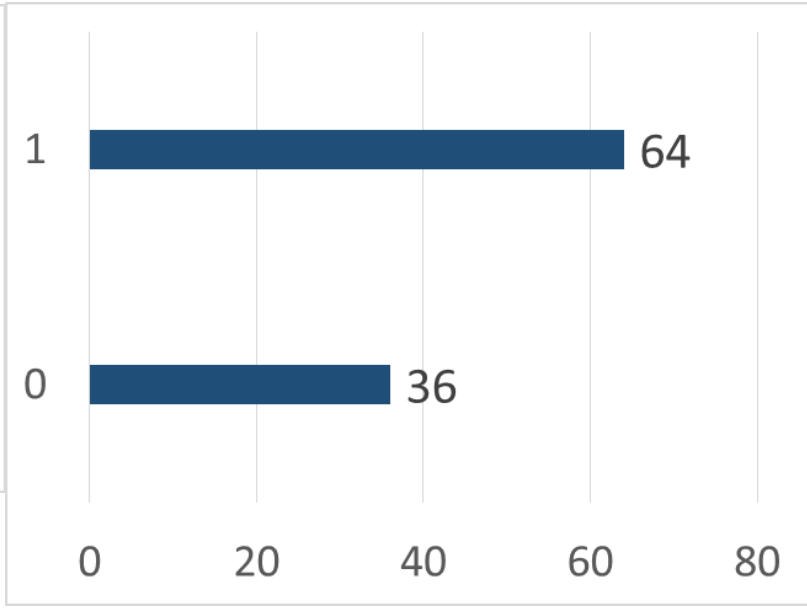
This is the most used method in statistical inference.

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

100 calls handled by an agent at a call centre

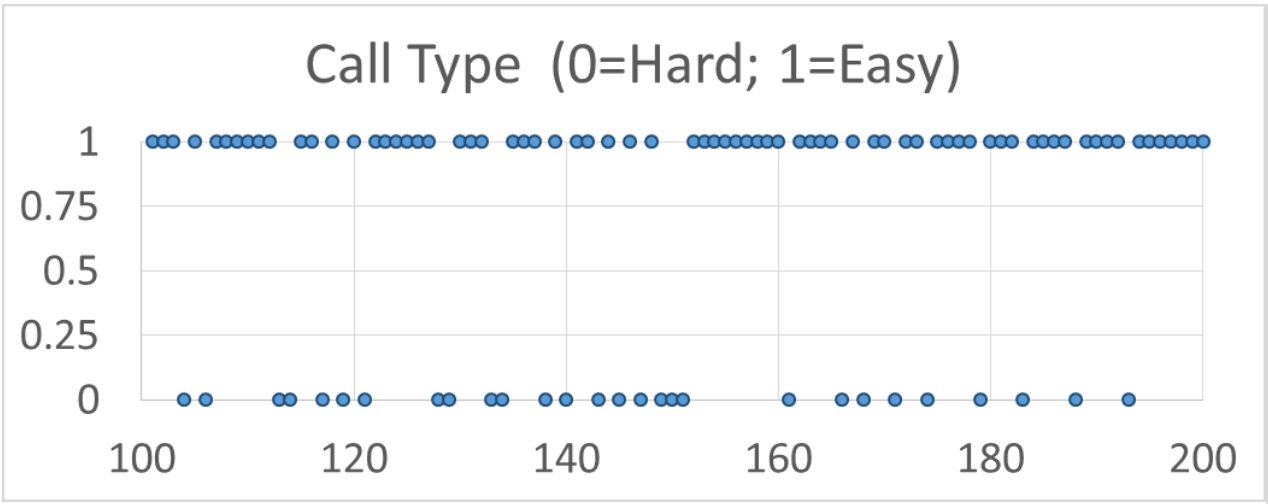


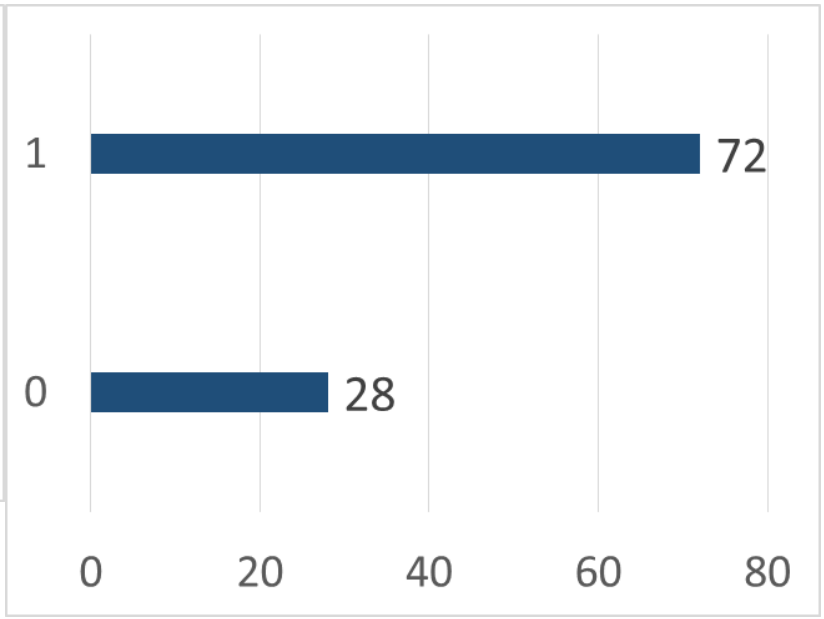


Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Next 100 calls handled by an agent at a call centre

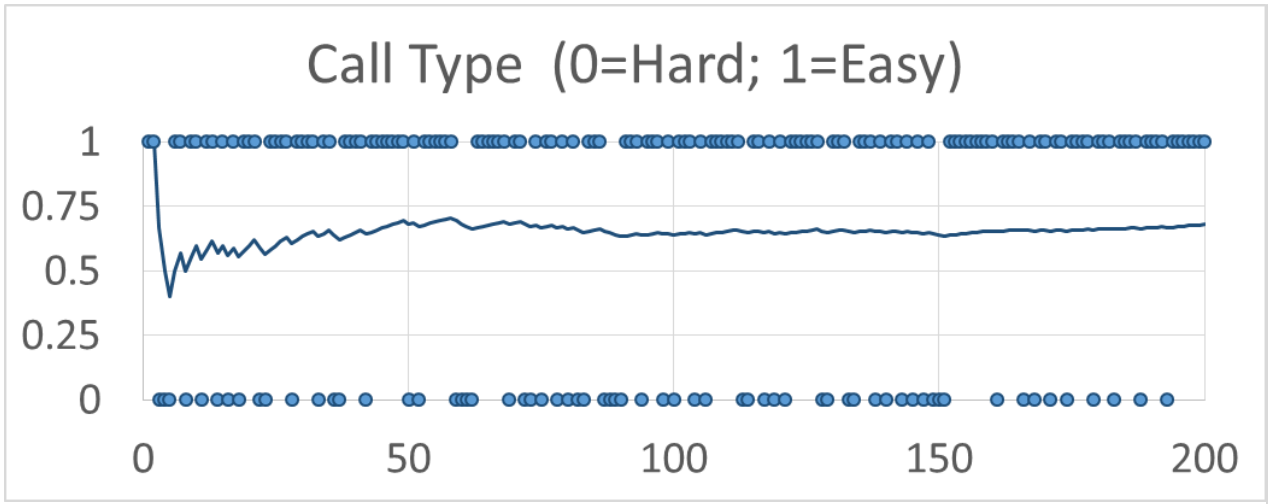


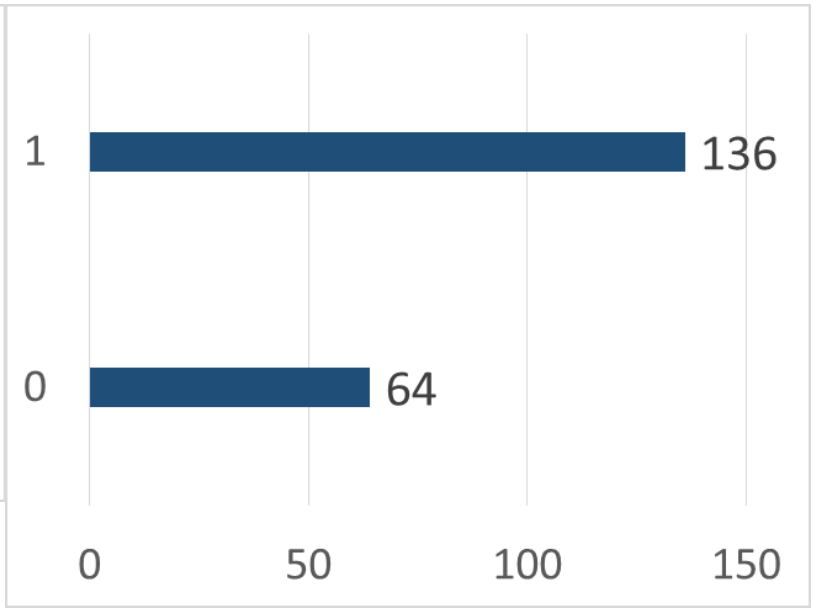


Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

Averages over the long run





P(easy) ~= 0.7

Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist

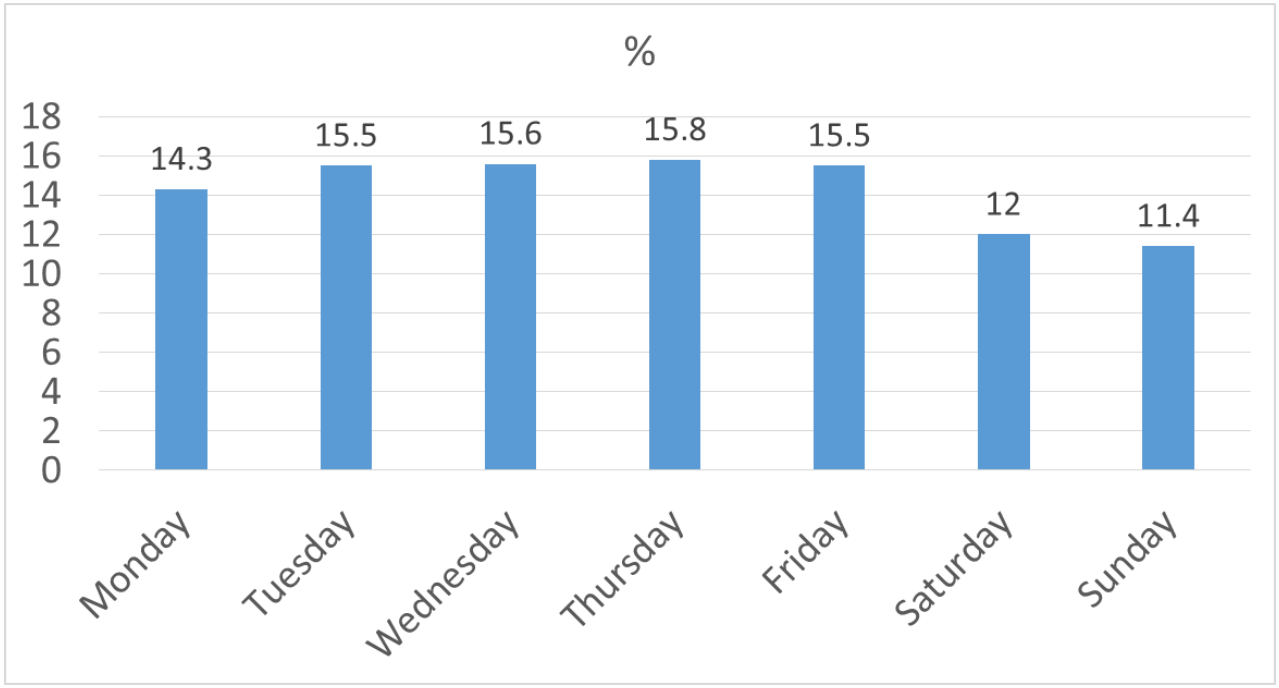
What is the probability of having a monthly income of 1000 BHD?

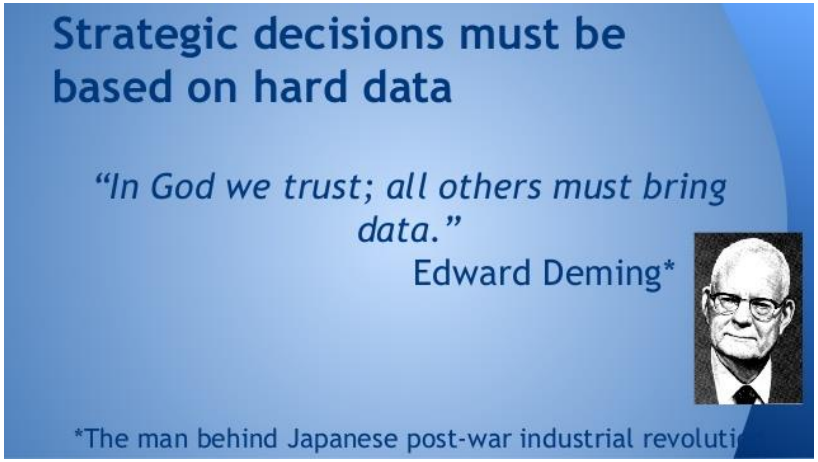
10/23 = 0.43

INCOME(BHD)	FREQUENCY
100	10
345	1
1000	10
9833	2

Assigning Probabilities

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





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.

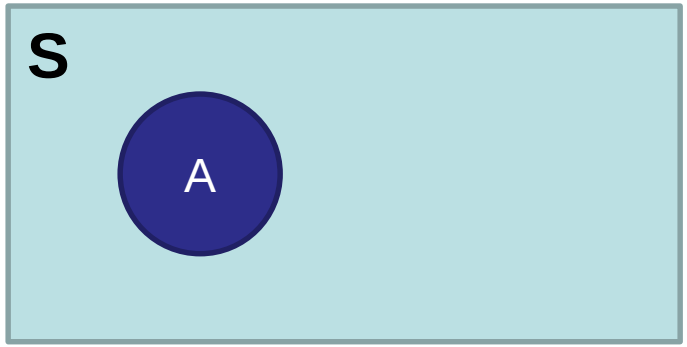
Probability - Terminology

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

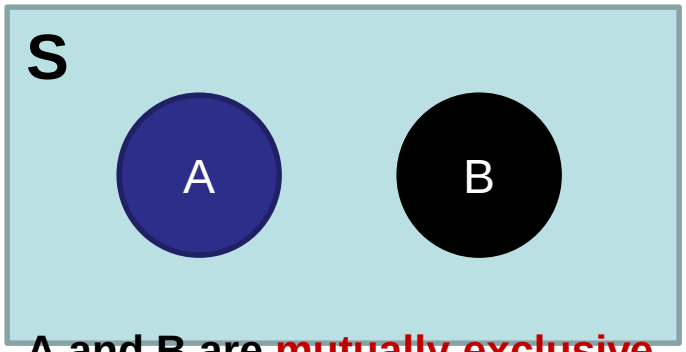
Event – A subset of the sample space.

Probability – Rules - Mutually Exclusive

$$\mu = 0$$



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



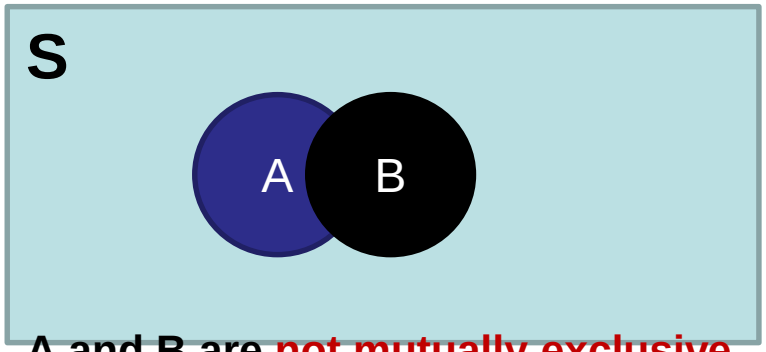
A and B are **mutually exclusive**

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

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



A and B are not mutually exclusive

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

Example

Event A – Customers who default on loans

Event B – Customers who are High Net Worth Individuals

Probability – Rules – Independent Events

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.

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

Probability - Rules

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

$$P(easy_1 \; and \; easy_2 \; and \; easy_3) = 0.7^3 = 0.343$$

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 - 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?

Ans: **Team goes for 2 point shot then**

P(winning the game) = P(2 Point shot) * P(winning in overtime)

P(winning the game) =1/2 * 1/2 = ¼ =0.25

Team goes for 3 point shot then

P(winning the game) = 0.30 –**3 POINT SHOT IS BETTER**

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

P(Young and Not Defaulting on the loan) = 10503/46687 = 0.225
P(Old and Defaulting on loan) = 120/46687 = 0.003

Probability - Types

Convert it into probabilities:

		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

P(Young and Not Defaulting on the loan) = 10503/46687 = 0.225
P(Old and Defaulting on loan) = 120/46687 = 0.003
P(Yes) = 8557/46687 =0.184
P(Young) = 14089/46687 = 0.302

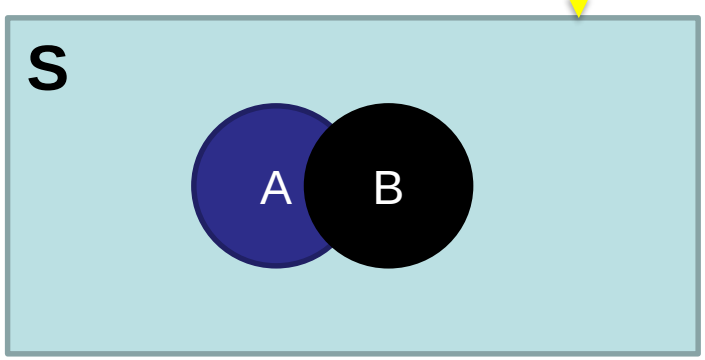
Probability - Types

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

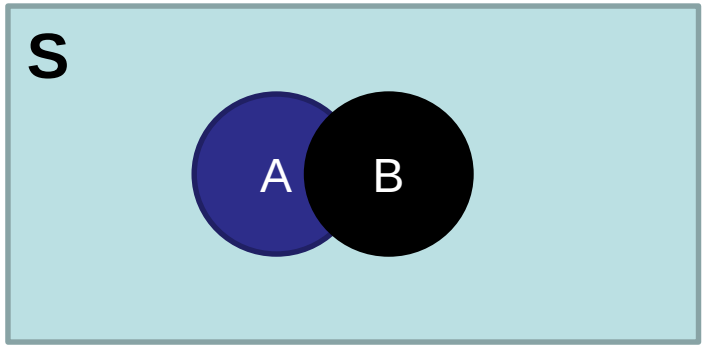


Probability - Types

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

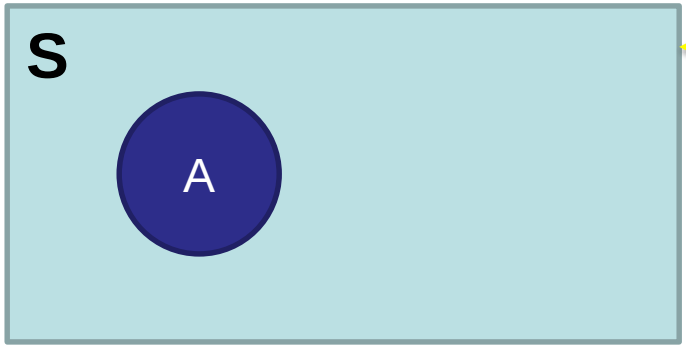
Marginal Probability

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.536	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.000

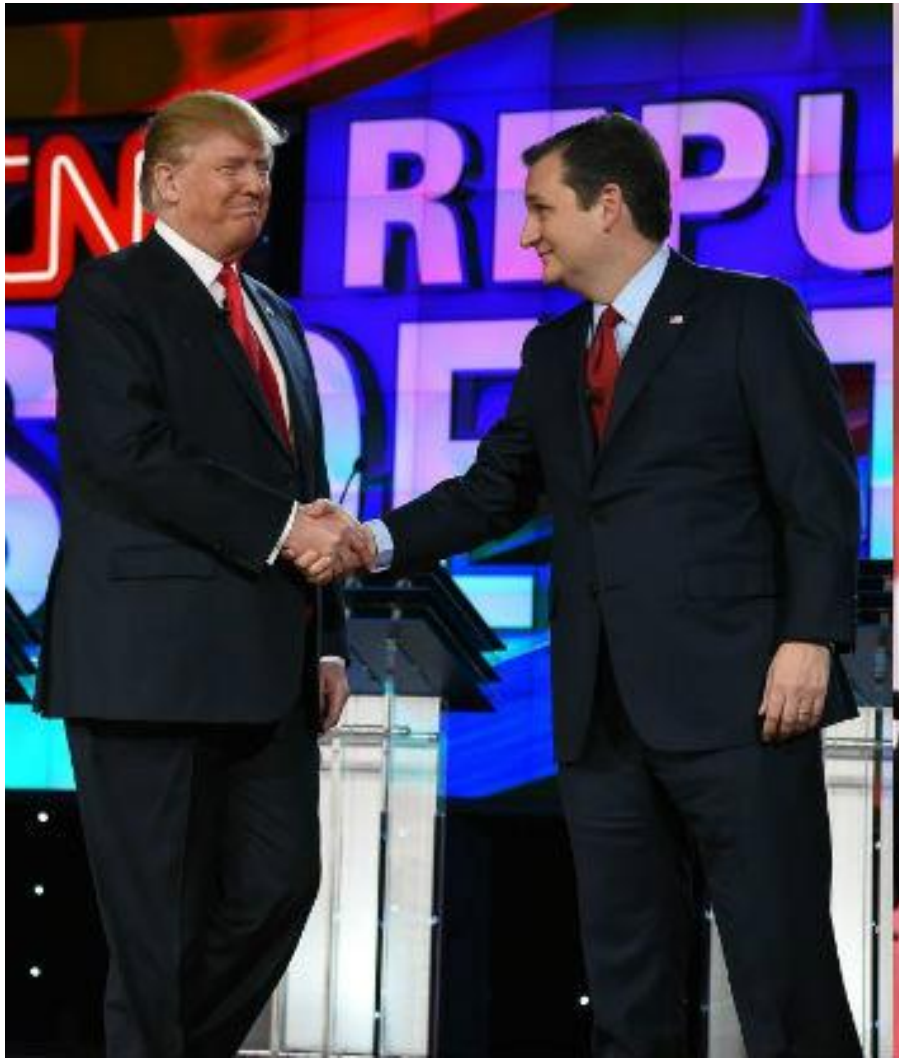
Probability describing a single attribute.

P(No) = 0.816

P(Old) = 0.008



Independent or Mutually Exclusive?



Donald Trump and Ted Cruz were Republican Party candidates.

Hillary Clinton and Bernie Sanders were Democratic Party candidates.

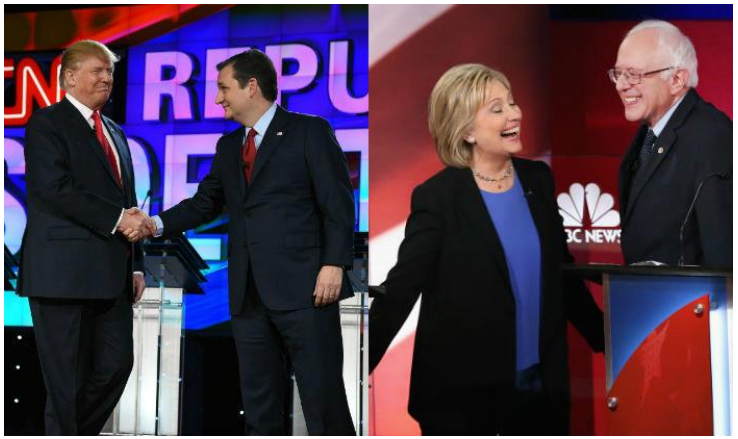
Independent or Mutually Exclusive?

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?

Event A: Trump winning Republican nomination

Event B: Cruz winning Republican nomination

Event C: Clinton winning Democratic nomination

Event D: Sanders winning Democratic nomination

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(B and A) *0*

P(A and C) *½ * ½ = ¼*

P(C and A) *½ * ½ = ¼*

P(A or B) *½ + ½ = 1*

P(B or A) *½ + ½ = 1*

P(A or C) *½ + ½ - ¼ = ¾*

P(C or A) *½ + ½ - ¼ = ¾*

Probability - Types

- Joint Probability
  - P(A and B) = P(A)*P(B)
- Union Probability
  - P (A or B) = P(A) + P(B) – P(A and B)
- Marginal Probability - Probability of a Single Attribute
  - Only one P(A), P(B)
- Conditional Probability

Day 1 : Recall

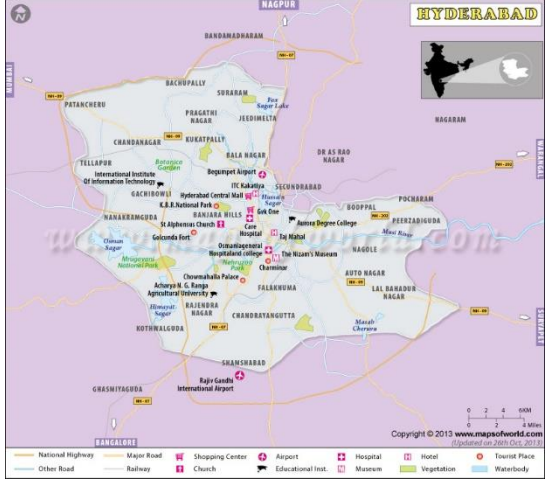
End of Day 1

- What is statistics :
  - Data Gathering, Understanding, Analysis and Presentation
- Population and Sample, Census and Survey
- Parameter (Greek symbols) and Statistic(Roman or Latin symbols)
- Descriptive and Inferential Statistics
- Variable(Dependent & Independent) & Data(Numerical & Categorical)
- Discrete and Continuous
- Central Tendencies –
  - Mean, Median, Mode, Range
  - Quartile (Lower, Middle, Upper) and Inter Quartile Range (IQR)
- Box Plot
- Measures of Spread
  - Mean Distance, Mean Absolute Deviation (MAD), Variance Standard Deviation & Standard score (Z)

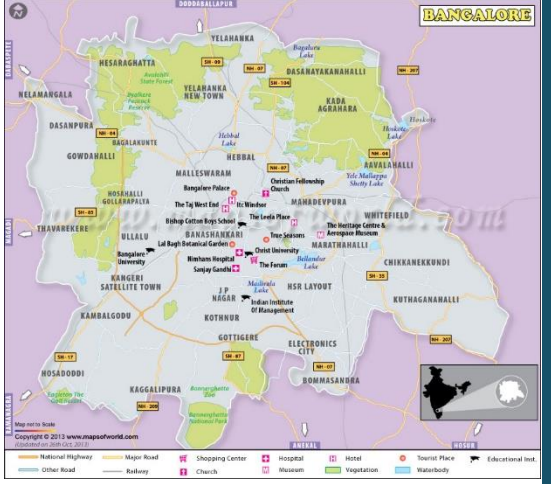
Day 2 : Recall

End of Day 1

- Probability Basics and Types
- Intro to Probability
- Differences between probability and Statistics
- Probability classification
  - Classical vs Frequentist
- Types of Probability
  - Joint Probability
  - Union Probability
  - Marginal Probability



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.