

Inspire...Educate...Transform.

# **Statistics and Probability Fundamentals**

## **Basic Statistical Concepts, Measures**

**Dr. Anand Jayaraman**

Dec 17, 2016

Thanks to Dr.Sridhar Pappu for the material

I have a great subject [statistics] to write upon, but feel keenly my literary incapacity to make it easily intelligible without sacrificing accuracy and thoroughness.

-Sir Francis Galton

(1822-1911)

CSE 7315C



# First thoughts on Maths

$$C(x) = \frac{\sum_{k=1}^n p_k x^k}{\sum_{k=1}^n p_k}$$

$$y = \phi(x) = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{t^2}{2}} dt$$

$$S(x,t) = \frac{2}{\pi} \int_0^t \frac{\sin at}{t} dt$$

$$P(\eta_{\infty} < x) = F(x)$$

$$\lim_{n \rightarrow \infty} \frac{(n!)^c}{(2n)!} = e^{-2x}$$

$$S_n = \ln n! / A_n$$

$$W_k = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P(\eta_{\infty} < x) = \sup_{t \geq 0} P(\eta_t < x | F = x)$$

$$|A_n| = \frac{n!}{2} \left| \int_{|x|>A} f(x) \log_2 \frac{1}{f(x)} dx \right| < \varepsilon$$

$$\int dG_n(x) \geq \frac{1}{2} \quad \forall n > \infty$$

$$f_{n+1}(t) = \int_0^t f_n(u) f_n(t-u) du = \frac{2^{n+1} t^n e}{n!}$$

$$\log \varphi(t) = i g t - c |t|^2 \left[ 1 + i \beta \frac{k}{|t|} \omega(t, \alpha) \right] B(t)$$

$$\int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = F(x) \left( \frac{1}{\sqrt{2\pi}} \right)^{-1} \quad |\Psi_5(t)| = \left| \int_{-\infty}^{\infty} e^{itx} dx \right|$$

$$\prod_m = \prod_r \prod_{m-r}$$

$$|X \cup Y| = |X| + |Y| - |X \cap Y| \quad \lim_{n \rightarrow \infty} \frac{1}{n}$$

$$f: X \rightarrow X \cap W$$

$$Q(A) = \int_A Q(\omega) dP \quad C(x) = -\log_2 \left( \frac{\sum_{k=1}^r p_k^x \log_2 \frac{1}{p_k^x}}{\sum_{k=1}^n p_k^x} - \left( \frac{\sum_{k=1}^r p_k^x \log_2 \frac{1}{p_k^x}}{\sum_{k=1}^n p_k^x} \right)^2 \right)$$

$$Q\left(e^{-x} \sqrt{\frac{1-q}{nq}} - 1\right) = x \sqrt{\frac{q(1-q)}{n}} + O\left(\frac{1}{n}\right)$$

$$\prod_{k=1}^r \left[ g_k \left( \frac{x}{\sqrt{n}} \right) \right]^{N_k} = e^{-\frac{x^2}{2}}$$



$$(t/y) = \frac{2e^{\frac{y^2}{2}}}{\sqrt{2\pi}} \left( \frac{e^{-\frac{y^2}{2}} du}{(1 - \frac{y^2}{u^2})^{\frac{3}{2}}} \right) \quad DN = \sum_{n=1}^N \frac{E_n}{n}$$

$$U_{n-c}^+ = \binom{2n}{n} - \binom{2n}{n-c}$$

$$\int_{-\infty}^{\infty} \varphi(t) dt = \frac{\sinh h}{th} [\varphi(t) e^{-itx} + \varphi(t+ix)]$$

$$\geq \frac{n!}{\prod_{k=1}^r n_k(k)!} \quad \frac{1}{m} \Psi(t) = \Psi\left(c \left(\frac{n}{m}\right) t\right)$$

$$Q_h(d) = \frac{p_d^h}{\sum_{j=1}^n p_j^h} \quad P(T_2 =$$

$$Q = F^{-1}(q) \quad \frac{|h|}{\log \log n^2} \leq 1 \quad \Psi(t) = 1 - \sqrt{1 - e^{2it}}$$

$$fg(u_i) = f\left(\sum_{j=1}^{\dim V_2} a_{ji} v_j\right) = \sum_{j=1}^{\dim V_2} a_{ji} \left(\sum_{k=1}^{\dim V_3} b_{kj} w_k\right) \frac{(2\ell)}{2^{2\ell}} \approx \frac{1}{\sqrt{16\ell}}$$

$$P_{j,k}^{(m)} = \sum_{l=0}^{\infty} P_{jl}^{(r)} P_{lk}^{(m-r)} \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} \operatorname{Re} \left\{ \varphi(t) \frac{e^{-ita} - e^{+ibt}}{it} \right\} dt$$

$$P(|\operatorname{Im} z| > \varepsilon) \leq \frac{C_q}{\log N}$$

$$\lim_{n \rightarrow \infty} \left( \int_{\mathbb{R}} u_n(x) \log_2 \frac{1}{|x|} dx \right) = \left( \int_{\mathbb{R}} f(x) \log_2 \frac{1}{|x|} dx \right)$$

CSE 7315C

# When am I going to use this?



CSE 7315c



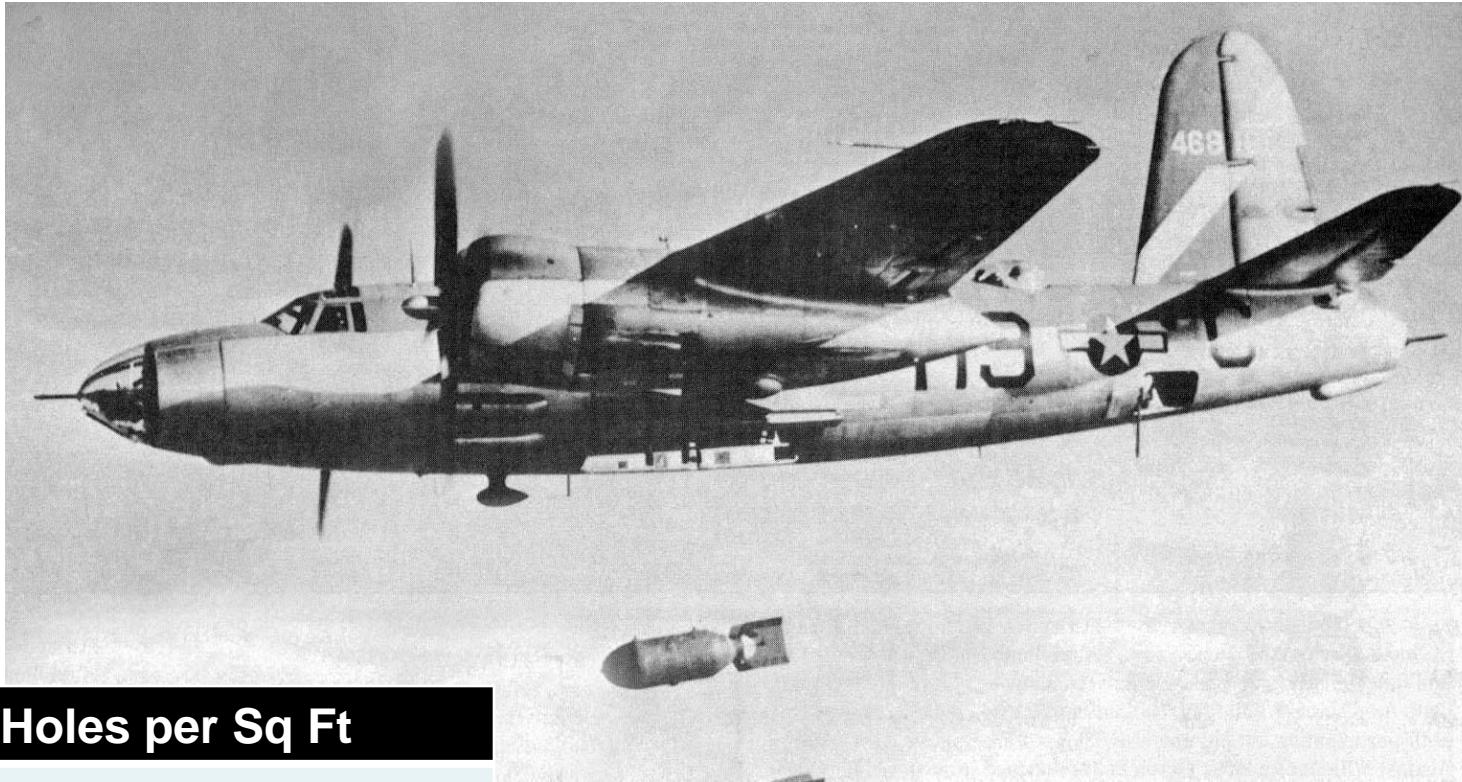
# Early example of Data analytics - WW-II

- Statistical Research Group -group of Applied Mathematicians aiding with war effort
- Recommendations made on everything - best trajectory of fighter plane to keep enemy aircraft in sight, optimal mixture of ammunition, strategic bombing etc

CSE 7315C



# Aircraft Armor Conundrum



Section of Plane	Bullet Holes per Sq Ft
Engine	1.1
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

CSE 7315C



Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

- H. G. Wells

CSE 7315C



# Modern Life..

- Abundance of data
- Sometimes contradictory evidence
- Making sense of data hard
- Mathematical sense essential to survive and flourish

CSE 7315C



# Looking at Data

← Insight

← Insight

← Insight



Your Step Data is Right on Track

9/24/16 4:10 PM

Age & Gender Comparison



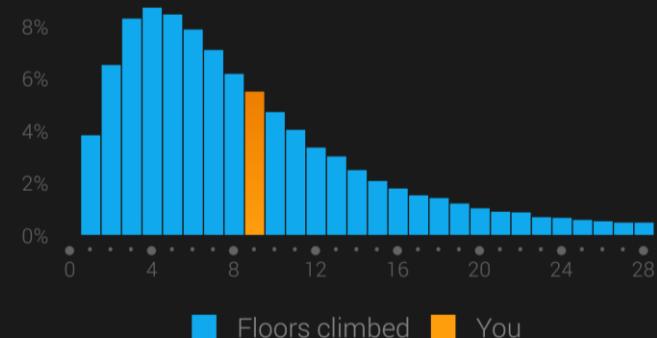
Of people the same age and gender as you, 40% are getting more steps per day than you. To move yourself up to the head of the pack and reap more health benefits, try adding in a 10-minute walking break each day.



You're Taking the Stairs an Average Amount

9/27/16 11:28 AM

Age & Gender Comparison: Floors Climbed



During the past month, our stats show you took the stairs about as much as 57% of people your age and gender. To get ahead of the pack, think about how you can fit in an extra flight of stairs each day. Work your way up to top-floor health!



Are You Getting Enough Sleep?

10/17/16 2:15 AM

Age & Gender Comparison



Looks like you're operating on less sleep than you may need. In fact, about 73% of people your age and gender have gotten more shut-eye than you over the past month. Sleep boosts your cardiovascular health and helps your brain process information, so make sure you're getting enough.



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

- Mark Twain / Benjamin Disraeli

CSE 7315C



# Misleading statistics



Wisconsin Gov. Scott Walker waves as he walks off-stage after addressing the Conservative Political Action Conference in National Harbor on Thursday. Credit: Associated Press

Wisconsin Republican Party says more than half the nation's job growth in June came from Wisconsin

<http://www.politifact.com/wisconsin/statements/2011/jul/28/republican-party-wisconsin/wisconsin-republican-party-says-more-than-half-nat/>

CSE 7315C



# Full Truth

- June 2011
  - US labor dept statistics – 18000 jobs in June
  - Wisconsin Added 9500 jobs

Wisconsin the fastest growing state in US?

- California Added 28800 jobs
- Texas 35000 jobs!
- Many states lost jobs
- Moral:
  - Do not trust percentages when negative numbers are involved

# Overabundance of Data

- Millions of calls being made every minute
- Billions of web-pages and page views
- Hundreds of thousands of sick patients
- Millions of cellphones manufactured

CSE 7315C



Statistics is “A telescope that allows us to study the large terrain and make it accessible to our unaided vision”

CSE 7315C



# Statistics – Big Picture

Statistics provides a way of organizing data to extract information on a wider and objective basis than relying on personal experience

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

CSE 7315C





## GENDER RACE

The sex-ratio of electorates used to be tilted to the male voters, but the trend has started to change. Five of the 13 states along with the three Union Territories which went to polls in the first four phases of LS experienced female electorates outnumbering their male counterparts.

### PUDUCHERRY

FEMALE	MALE
52%	48%

### KERALA

FEMALE	MALE
51.9%	48.1%

### MANIPUR

FEMALE	MALE
51%	49%

### MIZORAM

FEMALE	MALE
50.9%	49.1%

### DAMAN & DIU

FEMALE	MALE
50.5%	49.5%

### MEGHALAYA

FEMALE	MALE
50.4%	49.6%

### GOA

FEMALE	MALE
50.1%	49.9%

### ARUNACHAL

FEMALE	MALE
50.1%	49.9 %

# Problem #1: Data Gathering

Schedule Reference	Parliamentary Constituency			
	SI.	PC No.	PC Name	Type
Schedule no:	7	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](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).

CSE 7315C

Source: <http://epaper.deccanchronicle.com/articledetailpage.aspx?id=474880>;

Last accessed: April 27, 2014

**GENDER RACE**

The sex-ratio of electorates used to be tilted to the male voters, but the trend has started to change. Five of the 13 states along with the three Union Territories which went to polls in the first four phases of LS experienced female electorates outnumbering their male counterparts.

**PUDUCHERRY**

FEMALE	MALE
52%	48%

**KERALA**

FEMALE	MALE
51.9%	48.1%

**MANIPUR**

FEMALE	MALE
51%	49%

**MIZORAM**

FEMALE	MALE
50.9%	49.1%

**DAMAN & DIU**

FEMALE	MALE
50.5%	49.5%

**MEGHALAYA**

FEMALE	MALE
50.4%	49.6%

**GOA**

FEMALE	MALE
50.1%	49.9%

**ARUNACHAL**

FEMALE	MALE
50.1%	49.9 %

# Problem #2: Data Understanding

The ratios reflect the ratios of registered voters.



State/UT	Registered Voters			Voted in 2014 General Elections		
	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

CSE 7315C



**GENDER RACE**

The sex-ratio of electorates used to be tilted to the male voters, but the trend has started to change. Five of the 13 states along with the three Union Territories which went to polls in the first four phases of LS experienced female electorates outnumbering their male counterparts.

**PUDUCHERRY**

FEMALE	MALE
52%	48%

**KERALA**

FEMALE	MALE
51.9%	48.1%

**MANIPUR**

FEMALE	MALE
51%	49%

**MIZORAM**

FEMALE	MALE
50.9%	49.1%

**DAMAN & DIU**

FEMALE	MALE
50.5%	49.5%

**MEGHALAYA**

FEMALE	MALE
50.4%	49.6%

**GOA**

FEMALE	MALE
50.1%	49.9%

**ARUNACHAL**

FEMALE	MALE
50.1%	49.9 %

# Problem #2: Data Analysis/Interpretation

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

State/UT	Male			Female			Male	Female	Female-Male			
	2006-08	2011-13	2014	2006-08	2011-13	2014			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>.

CSE 7315C

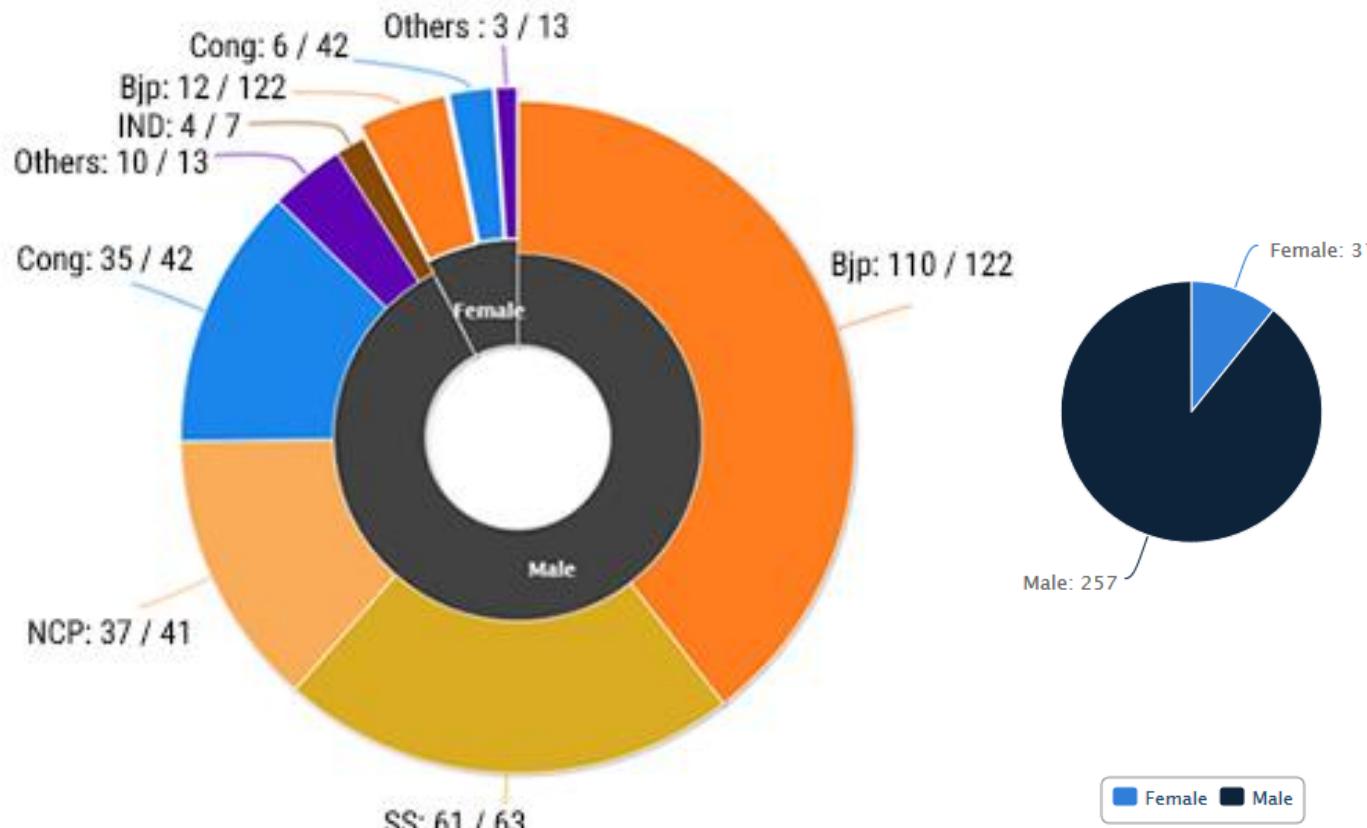
Source: <http://epaper.deccanchronicle.com/articledetailpage.aspx?id=474880>;

Last accessed: April 27, 2014

# Problem #4: Data Presentation

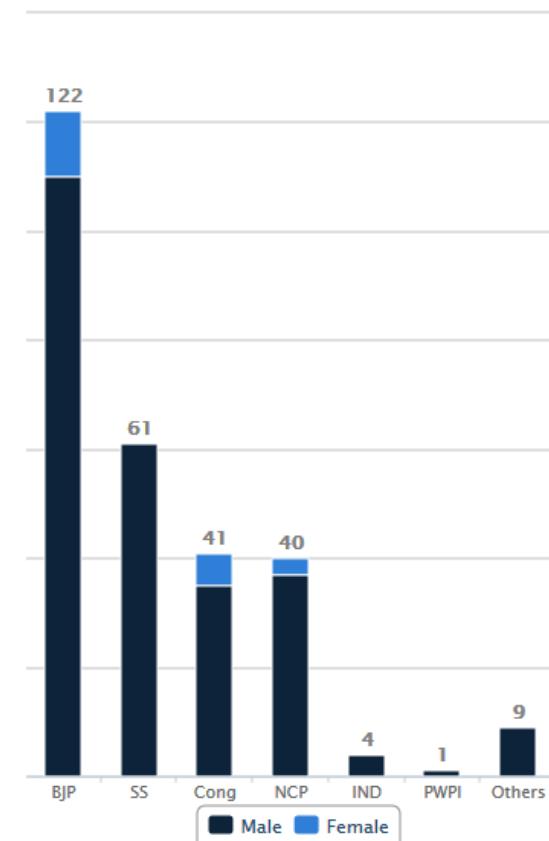
## Maharashtra: Gender Break-up

Total MLAs: 288\*



## Maharashtra: Gender Break-up

Total MLAs: 288\*



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

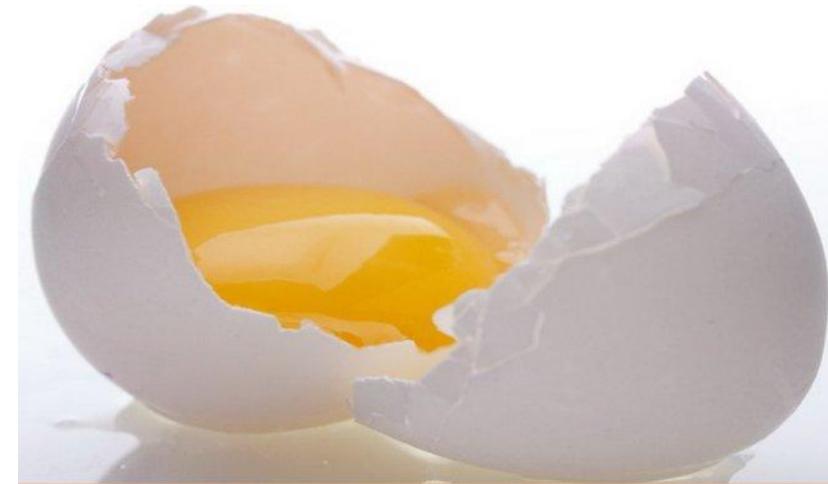
# The Cholesterol Confusion



**Avoid Dairy Products For High Cholesterol**



**Avoid Meat, Fish For High Cholesterol**



**Avoid Egg Yolk For High Cholesterol**

Source: <http://www.searchhomremedy.com/10-high-cholesterol-foods-to-avoid/>

CSE 7315C



# Cholesterol not a threat

## ■ US to remove high-cholesterol food from 'naughty' list

**Washington, May 26:** United States officials have finally given the green light for a U-turn on previous warnings on cholesterol, which has been on the "naughty" list of nutrients for nearly 40 years. Health officials have been warning people to stay away from high-cholesterol foods since the 1970s to avoid heart disease and clogged arteries.

However, after a study, eggs, butter, full-fat dairy products, nuts, coconut oil and meat have now been classified as "safe" and have been officially removed from the "nutrients of concern" list, reported the *International Business Times*.

The US Department of Agriculture, which is responsible for updating the guidelines every five years, stated in its findings for 2015: "Previously, the Dietary Guidelines for Americans recommended that cholesterol intake be limited to no more than 300

### FOODIES' DELIGHT

Butter, full-fat dairy products, nuts, coconut oil and meat have now been classified as "safe" and have been officially removed from the "nutrients of concern" list.



mg/day. The 2015 DGAC will not bring forward this recommendation because available evidence shows no appreciable relationship between consumption of dietary cholesterol and serum (blood) cholesterol, consistent with the



### The 70s, 80s and 90s

were the 'non fat' years, with the US government warning people to limit the amount of high-cholesterol foods in their diets.

longer warn people against eating high-cholesterol foods and will instead focus on sugar as the main substance of dietary concern.

The 70s, 80s and 90s were the 'non fat' years, with the US government warning people to limit the amount of high-cholesterol foods in their diets to avoid heart disease and strokes.

But nutritionists and scientists have long been campaigning for the U-turn, which started with introducing "good cholesterol" back into the 'safe zone'.

US cardiologist Dr Steven Nissen said: "It's the right decision. We got the dietary guidelines wrong. They've been wrong for decades."

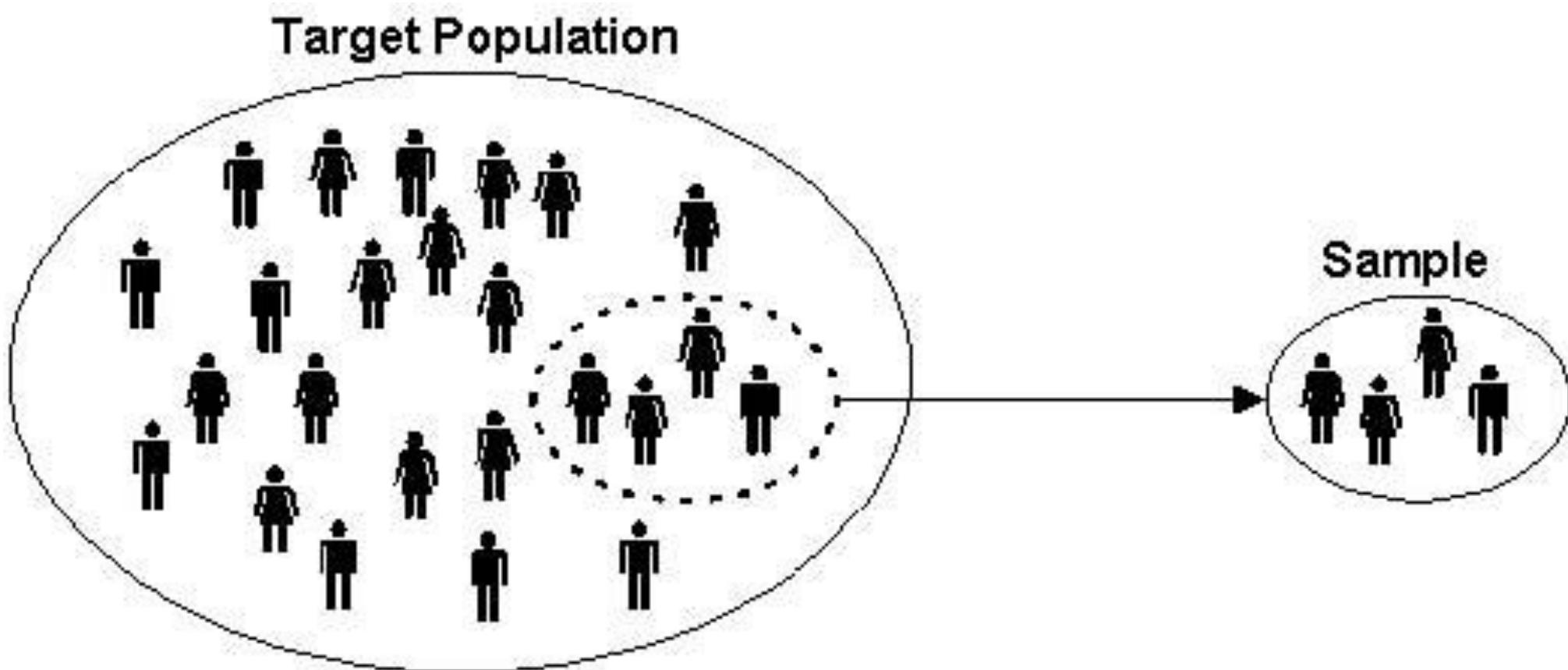
Dr Chris Masterjohn added: "When we eat more foods rich in this compound, our bodies make less. If we deprive ourselves of foods high in cholesterol — such as eggs, and butter — our body revs up its cholesterol synthesis." — Agencies

CSE 7315C

# BASIC STATISTICAL TERMINOLOGY



# Population and Sample



Source: <http://www.snapsurveys.com/blog/wp-content/uploads/2011/08/target-population.jpg>  
Last accessed: October 7, 2014

CSE 7315C



# 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.



# Height of Women in a University



© Alyssa Rice / Twitter

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

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.
<u>Chelsee Ramos</u>	5'2"	Madison, WI	Jr.
<u>Sydney Shelton</u>	4'10"	Scottsville, KY	Jr.
Ashley Wettstain	5'0"	Owensboro, KY	Fr.
<u>Madison Yee</u>	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	Kwin 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/ky-w-baskbl-mtt.html>

Last accessed: October 7, 2014

CSE  
73156



# 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.

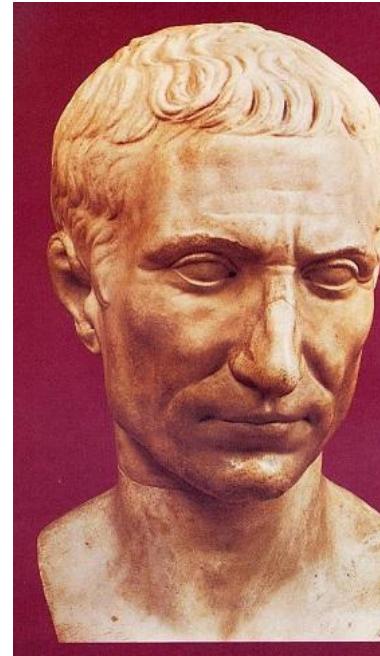


# Parameter and Statistic

I AM INDEBTED TO  
MY FATHER FOR  
LIVING, BUT  
TO MY  
TEACHER FOR  
LIVING WELL.



Alexander the Great  
[www.quote-coyote.com](http://www.quote-coyote.com)



*"What we wish,  
we readily  
believe, and  
what we  
ourselves think,  
we imagine  
others think  
also."*

**Julius Caesar**

## Greek – Population Parameter

Mean –  $\mu$

Variance –  $\sigma^2$

Standard Deviation -  $\sigma$

## Roman – Sample Statistic

Mean –  $\bar{x}$

Variance –  $s^2$

Standard Deviation -  $s$

CSE 7315c

# Descriptive and Inferential Statistics

- Descriptive Statistics – Data gathered about a group to reach conclusions about the same group.
- Inferential Statistics – Data gathered from a sample and the statistics generated to reach conclusions about the population from which the sample is taken. Also known as Inductive Statistics.

**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](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](http://www.arogyaworld.org/wp-content/uploads/2010/10/ArogyaWorld_IndiaDiabetes_FactSheets CGI2013_web.pdf)

Last accessed: November 25, 2015



CSE 7315C

# Variables and Data

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

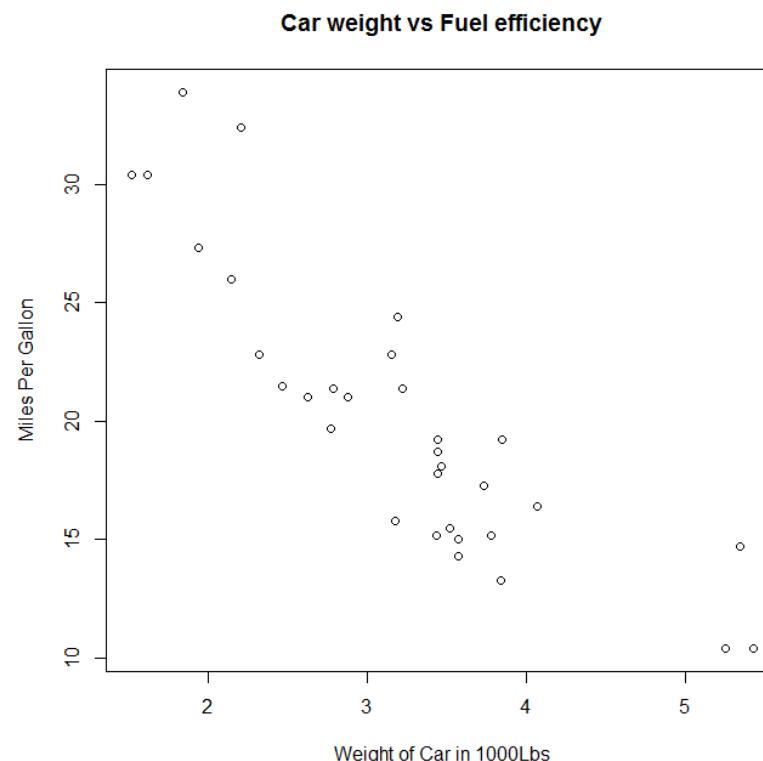
CSE7315C

Source: MTCARS dataset. Data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)



# Variables – Dependent and Independent

- Dependent variables on y-axis and Independent on x-axis.
- Dependent variable also called Target variable or Class variable.



Source: MTCARS dataset

# 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

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

CSE 7315C



# Quantitative Data - Interval

Data where ordering is clear and the difference in data values is meaningful.

However, there is no natural zero or origin.

Example: Year 1008 vs 2016

Temperature: 14C vs 28C

# Quantitative Data - Ratio

Ratio level data is similar to Interval level data, with the key difference – there is a natural zero point.

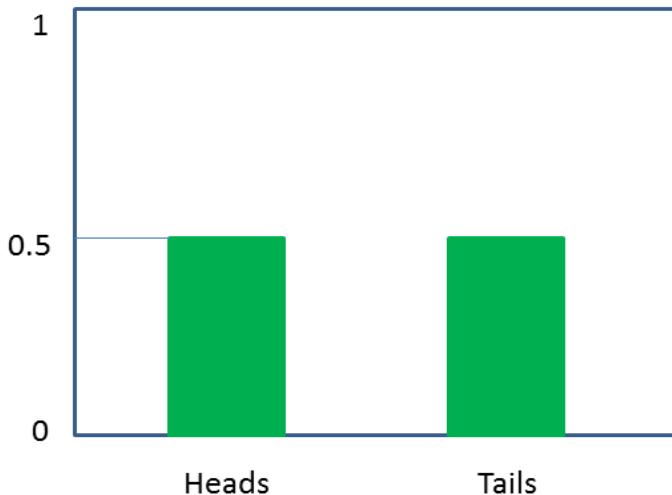
Examples: Weights, Cost of things,  
Number of correct answers in a exam



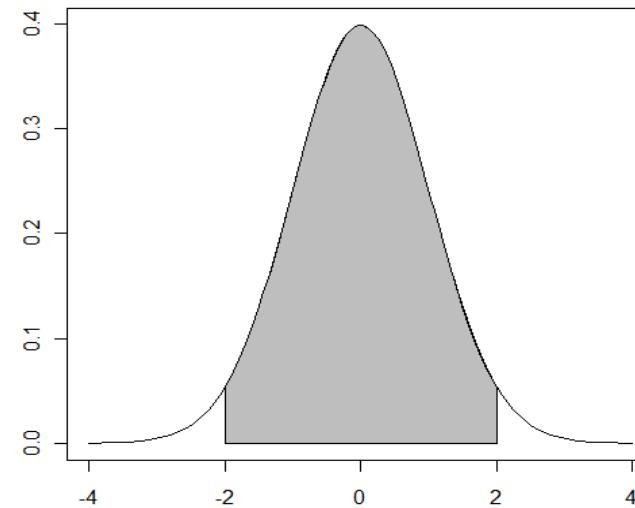
# Summary of Levels of Data Measurement

- Nominal - Categories only
- Ordinal - Categories with some order
- Interval – Meaningful differences, but no zero point
- Ratio – Meaningful differences with a natural starting point

# Discrete and Continuous



Countable



Measurable

# Discrete or Continuous?

Time between customer arrivals at a retail outlet	Continuous
Sampling the volume of liquid nitrogen in a storage tank	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**

CSE 7315C



# The Central Tendencies

Sai wants to join a health club in an activity that has others in the same age group as him. He is 22 years old. Mean ages for Yoga, Power Workout and Swimming classes are:

15 years



20 years



17 years



# The Central Tendencies

## Yoga class composition

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



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

CSE 7315C

Source: <http://www.montecitoheightsstudios.com/yoga-for-teens>

Last accessed: June 05, 2015



# The Central Tendencies

## Power workout class composition

<b>Age (years)</b>	13	15	17	90
Frequency, f	4	6	3	1

$$\text{Mean, } \mu = \frac{\sum x}{n} = \frac{\sum fx}{\sum f} = \frac{13X4+15X6+17X3+90X1}{4+6+3+1} = 20$$

But nobody in **Sai's** age group ☹.



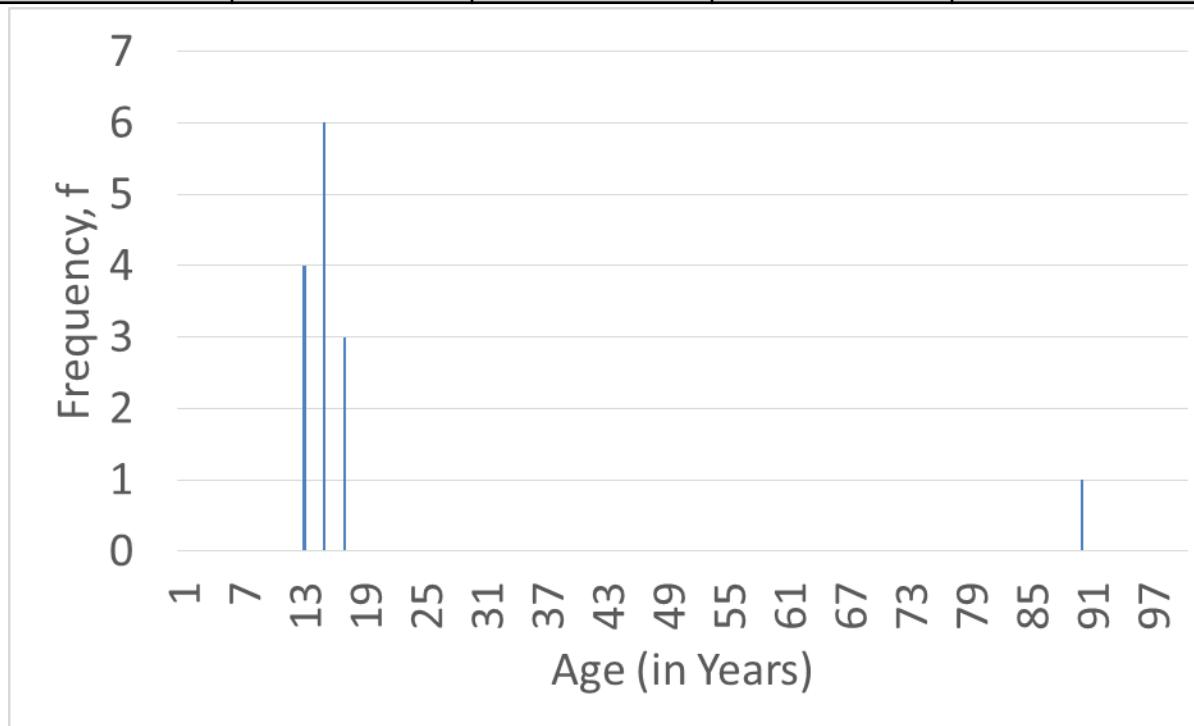
CSE 7315C



# The Central Tendencies

## Power workout class composition

<b>Age (years)</b>	13	15	17	90
Frequency, f	4	6	3	1



# The Central Tendencies

Source: [http://www.business-standard.com/article/companies/ambani-gets-205-times-ril-s-median-pay-115070500340\\_1.html](http://www.business-standard.com/article/companies/ambani-gets-205-times-ril-s-median-pay-115070500340_1.html)

Last accessed: July 7, 2015

RIL chairman Mukesh Ambani gets 205 times company's median salary

This ratio stands at 439 times in case of ITC Executive Chairman Y C Deveshar  
Press Trust of India | New Delhi | July 6, 2013 Last Updated at 00:49 IST

Ramoo HRIS Solutions

Cloud Based HRIS Solution With Role Based Workspaces. Ask For Demo! ramoo.com/HRIS

Ads by Google



Mukesh Ambani, the richest Indian and Reliance Industries (RIL) chairman and MD, has not taken a pay hike for seven years, but his [pay package](#) is over 205 times that of the median employee remuneration at RIL. This ratio stands at 439 times in case of [ITC Executive Chairman Y C Deveshar](#).

The ratio stands much lower at 89 times in case of Information technology (IT) major Wipro Chairman and Managing Director Azim Premji, and at 19 times for HDFC Chairman Deepak Parekh for 2014-15.

However, HDFC Banks Managing Director (MD) Aditya Puri got a remuneration that was 117 times of the median employee pay, while for ICICI Bank Chief Executive Officer (CEO) Chanda Kochhar it was 97 times and at over 74 times for Axis Banks MD and CEO Shikha Sharma.

IT giant Infosys CEO Vishal Sikka's pay was 116 times of median employee pay. The same ratio for HUL's CEO Sanjiv Mehta was 93 times, but much higher at 293 times for Vedanta Chairman Naresh Agarwal.

Top 10 Mutual Funds  
Buy Mutual funds online. Select from 5000+ schemes. Open Free A/c now.  
[www.myuniverse.co.in/ZgSp](http://www.myuniverse.co.in/ZgSp)

Regus™ Office Solutions  
No Hidden Costs & Flexible Options.  
Workspaces To Suit. Get A Quote  
Now  
[www.regus.co.in](http://www.regus.co.in)

Ads by Google

Listed firms have begun disclosing these ratios and other comparisons such as salary raises for top management personnel and average staff member, for the first time pursuant to the new Companies Act and Sebi's latest Corporate Governance Code coming into force.

While a majority of the companies are still in the process of disclosing such details, the disclosures made so far by top companies show a wide variance in these ratios. There is also a huge difference between the pay increases for top management personnel and average staff in many cases.

## CHEQUES & BALANCES



Ambani has kept his salary capped at Rs 15 crore for seven years now, while the median remuneration of employees increased by 3.71 per cent to Rs 7.29 lakh during 2014-15. The total remuneration of key managerial personnel in fact dipped by 1.93 per cent to Rs 73.28 crore.

Deveshar's remuneration rose by 24 per cent during the year, against an increase of 14 per cent in the company's

median employee remuneration. The overall key managerial personnel remuneration rose 20 per cent. Deveshar's gross remuneration in 2014-15 stood at over Rs 15 crore, but net pay was lower at Rs 7.3 crore.

Premji saw his pay decline by 53 per cent to Rs 4.78 crore, while median employee remuneration rose by 9.5 per cent. Wipro CEO T K Kurien got a package that was 170 times the



# The Central Tendencies

## Power workout class composition

<b>Age (years)</b>	13	15	17	90
Frequency, f	4	6	3	1

Data has outliers

Median – the mid-point

13, 13, 13, 13, 15, 15, 15, 15, 15, 17, 17, 17, 17, 90

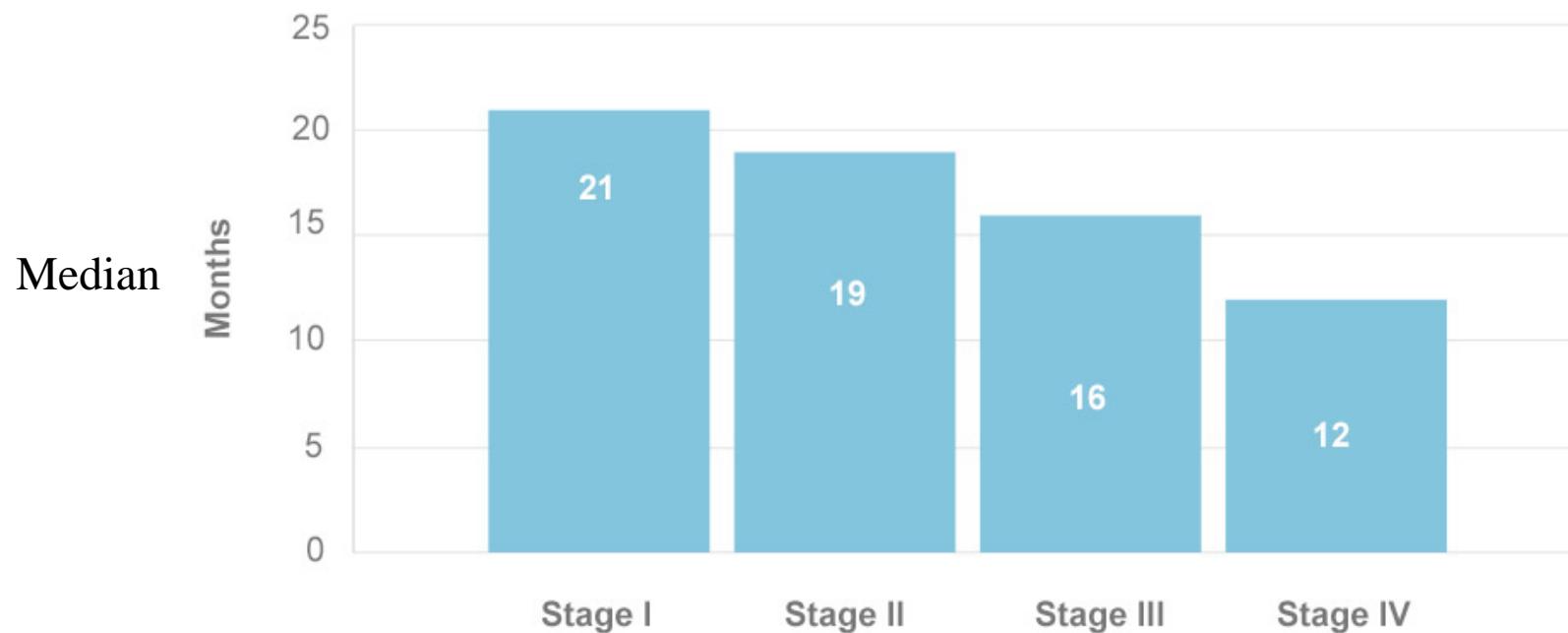


# The Central Tendencies

## Survival by Stage

Doctors use a **four-stage system** to describe how far the cancer has advanced within the body.

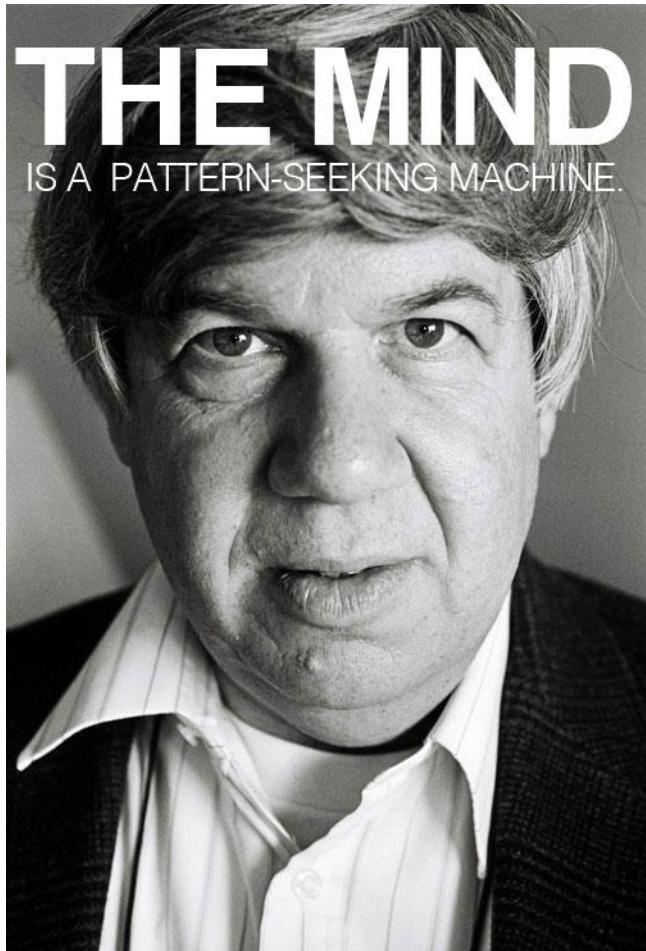
Patients diagnosed at stage I have the best outlook, while survival is worst at stage IV.



Source: <http://www.asbestos.com/mesothelioma/statistics.php>

Last accessed: April 05, 2016

# The Median Isn't the Message by Stephen Jay Gould



“...This is a personal story of statistics, properly interpreted, as profoundly nurturant and life-giving...”

“...The literature couldn't have been more brutally clear: mesothelioma is incurable, with a **median** mortality of only eight months after discovery...”

“...Attitude clearly matters in fighting cancer...”

– Stephen Jay Gould

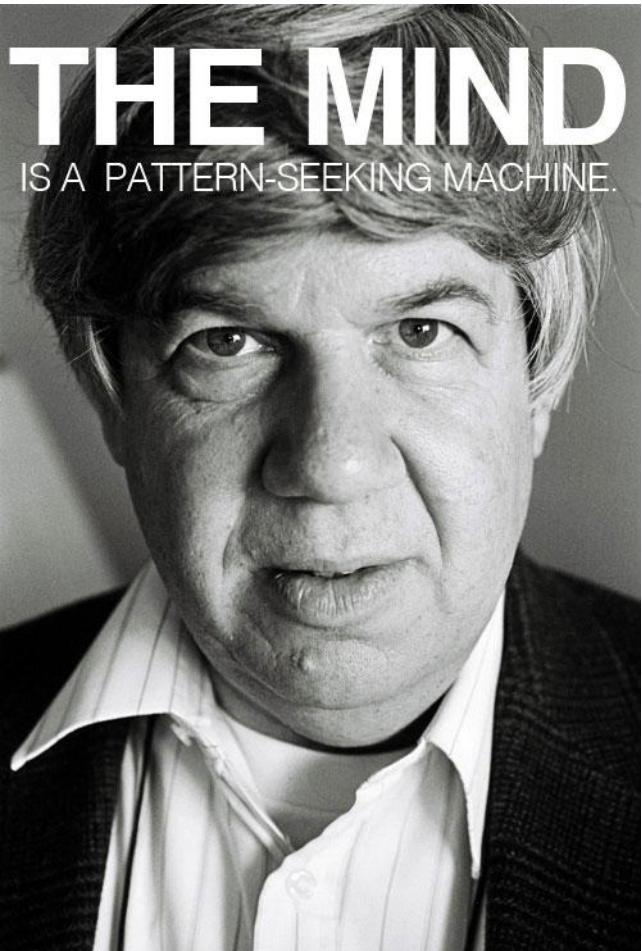
Source: [http://cancerguide.org/median\\_not\\_msg.html](http://cancerguide.org/median_not_msg.html)

Last accessed: April 06, 2016

CSE 7315C



# The Median Isn't the Message by Stephen Jay Gould



“...What does "median mortality of eight months" signify in our vernacular? I suspect that most people, without training in statistics, would read such a statement as "I will probably be dead in eight months" - the very conclusion that must be avoided, since it isn't so, and since attitude matters so much...”

“...But all evolutionary biologists know that variation itself is nature's only irreducible essence. **Variation** is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions...I had to place myself amidst the variation.”

– Stephen Jay Gould

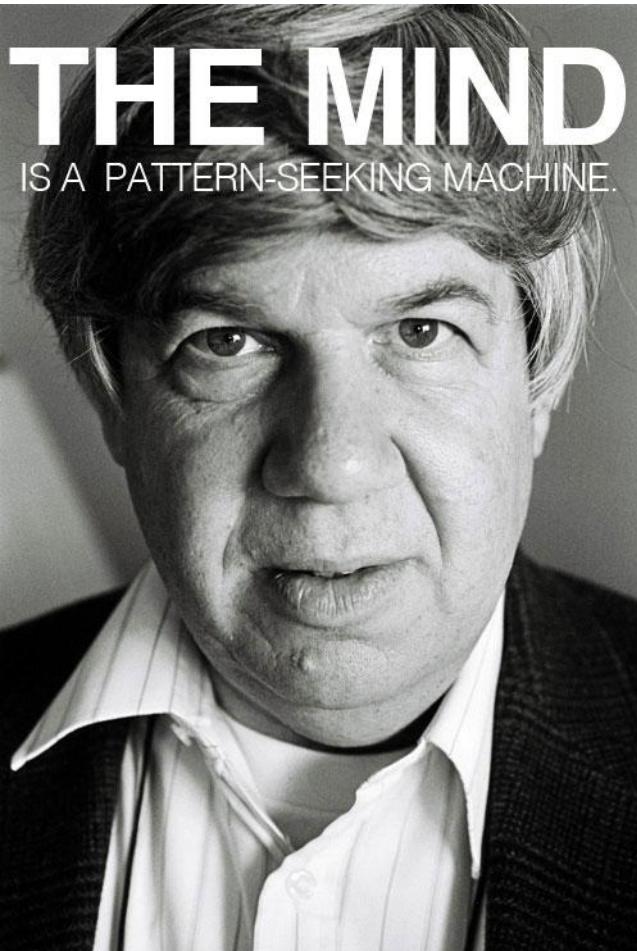
Source: [http://cancerguide.org/median\\_not\\_msg.html](http://cancerguide.org/median_not_msg.html)

Last accessed: April 06, 2016

CSE 7315C

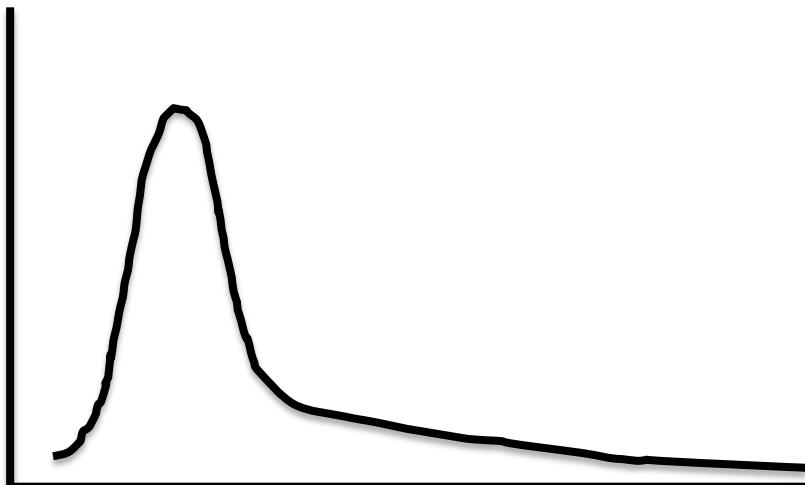


# The Median Isn't the Message by Stephen Jay Gould



“...I immediately recognized that the distribution of variation about the eight-month median would almost surely be what statisticians call “right skewed”...”  
“...The distribution was indeed, strongly right skewed, with a long tail (however small) that extended for several years above the eight month median. I saw no reason why I shouldn't be in that small tail...”

– Stephen Jay Gould



Source: [http://cancerguide.org/median\\_not\\_msg.html](http://cancerguide.org/median_not_msg.html)

Last accessed: April 06, 2016

CSE 7315C



# It's All in the Mind

Psychosomatic Medicine:

November/December 2006 - Volume 68 - Issue 6 - pp 809-815

doi: 10.1097/01.psy.0000245867.92364.3c

Original Articles

## Positive Emotional Style Predicts Resistance to Illness After Experimental Exposure to Rhinovirus or Influenza A Virus

Cohen, Sheldon PhD; Alper, Cuneyt M. MD; Doyle, William J. PhD; Treanor, John J. MD; Turner, Ronald B. MD

### Abstract

**Objective:** In an earlier study, positive emotional style (PES) was associated with resistance to the common cold and a bias to underreport (relative to objective disease markers) symptom severity. This work did not control for social and cognitive factors closely associated with PES. We replicate the original study using a different virus and controls for these alternative explanations.

**Methods:** One hundred ninety-three healthy volunteers ages 21 to 55 years were assessed for a PES characterized by being happy, lively, and calm; a negative emotional style (NES) characterized by being anxious, hostile, and depressed; other cognitive and social dispositions; and self-reported health. Subsequently, they were exposed by nasal drops to a rhinovirus or influenza virus and monitored in quarantine for objective signs of illness and self-reported symptoms.

**Results:** For both viruses, increased PES was associated with lower risk of developing an upper respiratory illness as defined by objective criteria (adjusted odds ratio comparing lowest with highest tertile = 2.9) and with reporting fewer symptoms than expected from concurrent objective markers of illness. These associations were independent of prechallenge virus-specific antibody, virus type, age, sex, education, race, body mass, season, and NES. They were also independent of optimism, extraversion, mastery, self-esteem, purpose, and self-reported health.

**Conclusions:** We replicated the prospective association of PES and colds and PES and biased symptom reporting, extended those results to infection with an influenza virus, and "ruled out" alternative hypotheses. These results indicate that PES may play a more important role in health than previously thought.

BMI = body mass index; CI = confidence interval; NES = negative emotional style; PES = positive emotional style; RV = rhinovirus; TCID = Tissue Culture Infectious Dose.

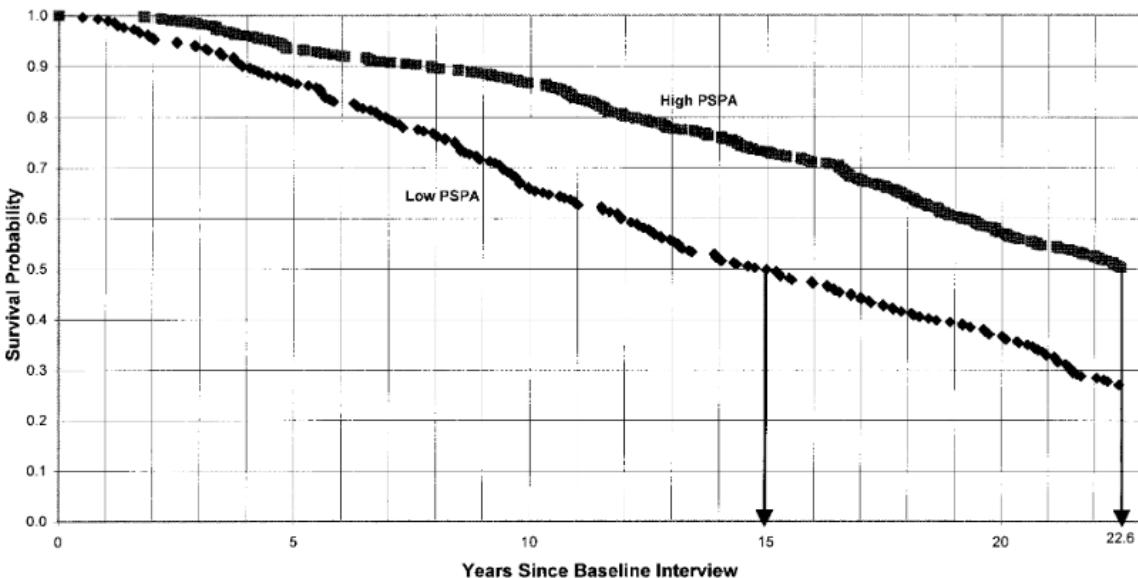


Figure 1. Influence of positive self-perceptions of aging (PSPA) on survival. Arrow indicates median survival.

## Yale Study

## CMU Study

Source: [http://www.huffingtonpost.com/entry/positive-people-live-long\\_b\\_774648.html?section=india](http://www.huffingtonpost.com/entry/positive-people-live-long_b_774648.html?section=india)

Last accessed: April 06, 2016

# The Central Tendencies

Sai is disturbed and wants some relaxation. He joins the swimming class. He didn't understand why they were asking where his kid was...

Age (years)	1	2	3	30	31	32	33
Frequency, f	3	4	3	1	3	2	4

Mean ~ 17 years

Median?

What happens to Median if another kid or adult is added?



CSE 7315C

Source: <http://0.tqn.com/y/pediatrics/1/W/8/a/infant-swim-class.jpg>  
Last accessed: October 25, 2014



# The Central Tendencies

Age (years)	1	2	3	30	31	32	33
Frequency, f	3	4	3	1	3	2	4

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

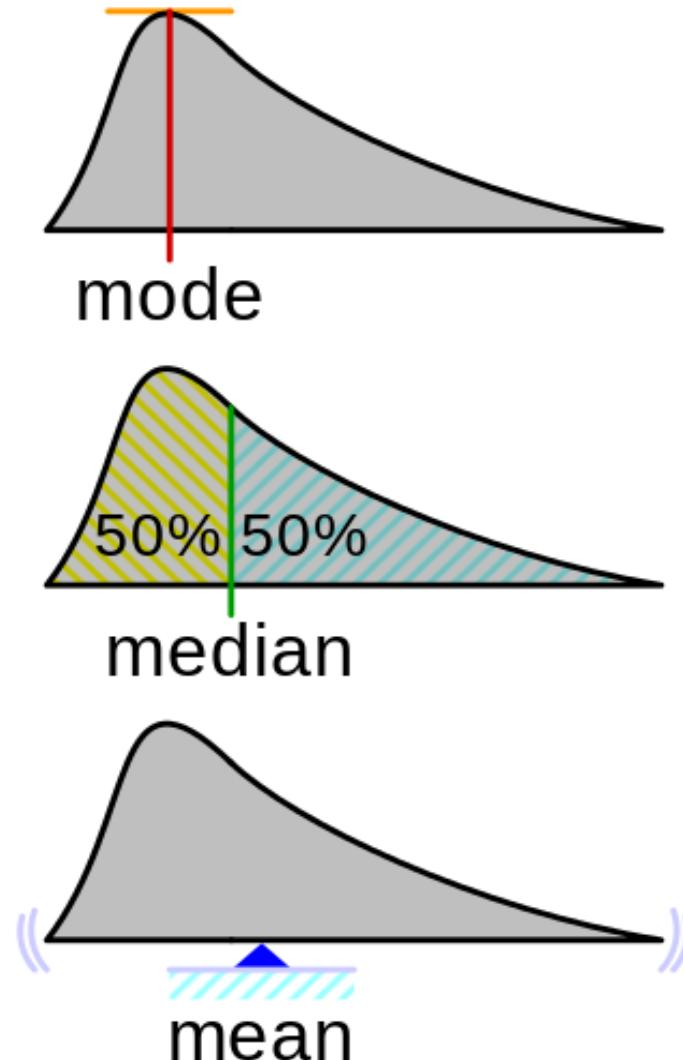
# The Central Tendencies

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

Mode is also the only central-tendency statistic that works with categorical data.



# The Central Tendencies



# 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?



# 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	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	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	2	2	1	1

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

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

# 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

Lower quartile (25<sup>th</sup> percentile, Q1) =  $\frac{(n+1)}{4}$ th

Middle quartile = Median =  $\frac{2*(n+1)}{4}$ th

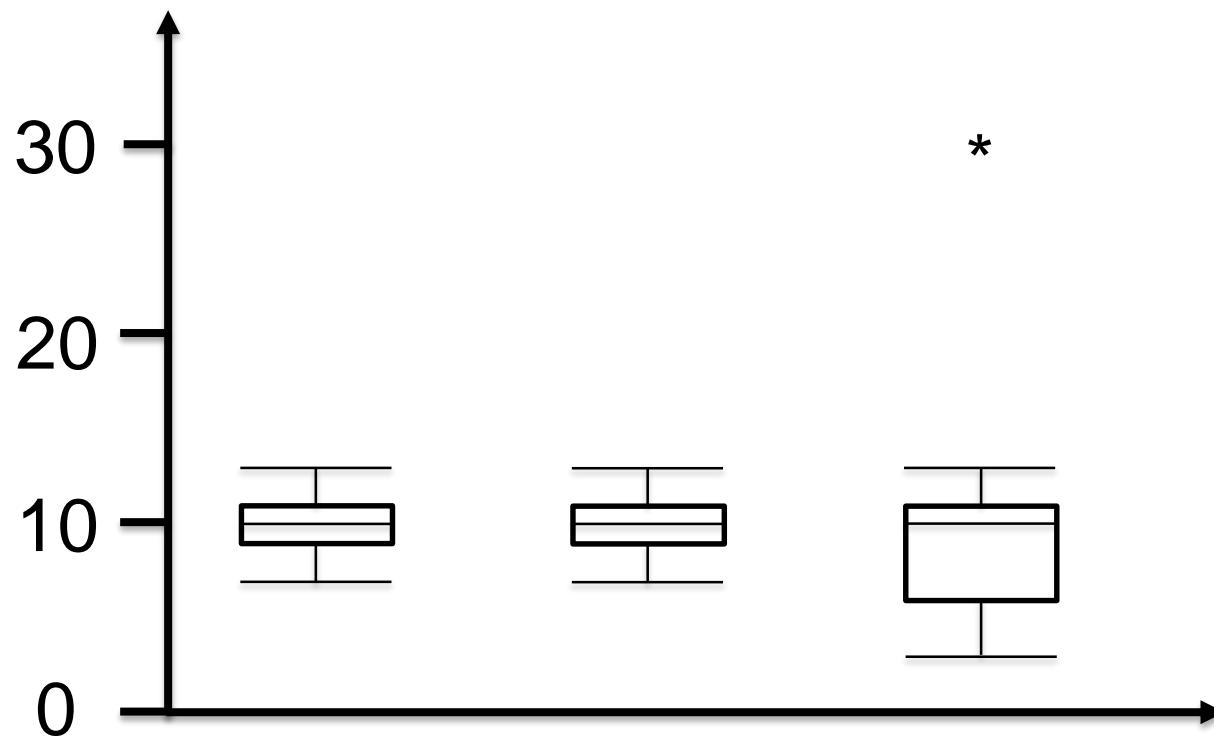
Upper quartile (75<sup>th</sup> percentile, Q3) =  $\frac{3*(n+1)}{4}$ th

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

# Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Box and whisker diagram or Box plot



<https://www.khanacademy.org/video/constructing-a-box-and-whisker-plot>

CSE 7315C



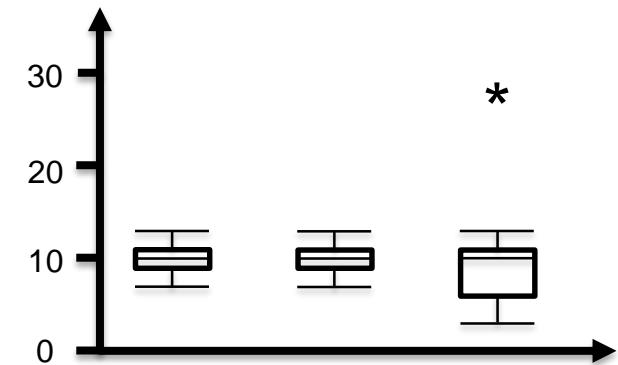
# Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

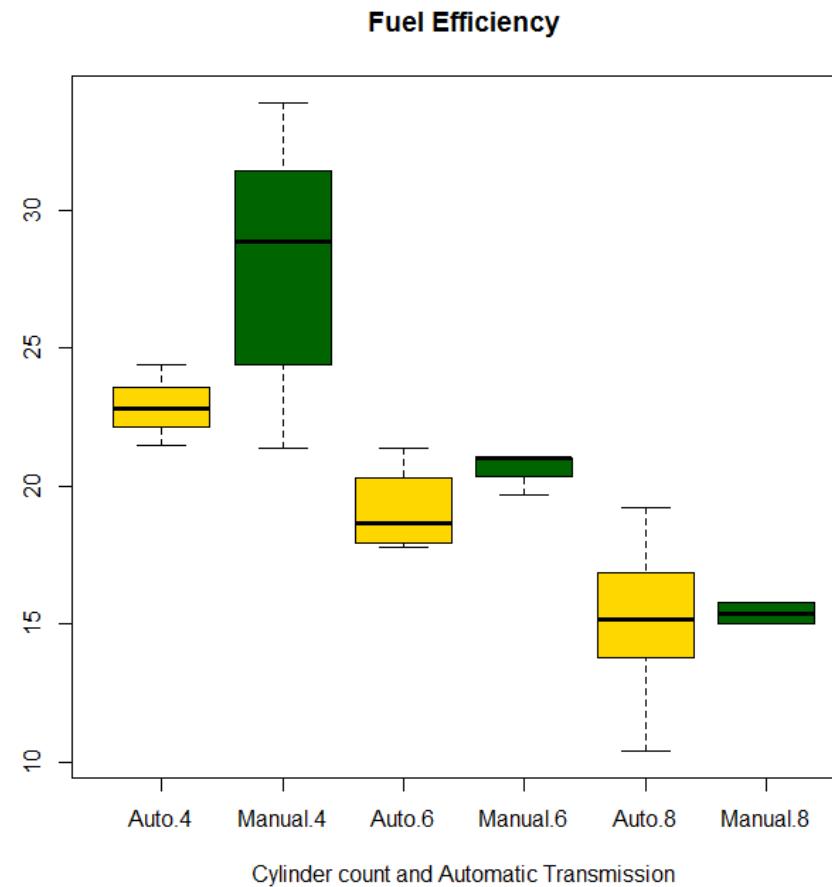
Box and whisker diagram or Box plot

Tukey fences

Name	Formula	Player 1	Player 2	Player 3
Upper Hinge	75th Percentile	11	11	11
Lower Hinge	25th Percentile	9	9	6
H-Spread	Upper Hinge - Lower Hinge (IQR)	2	2	5
Step	$15 \times \text{H-Spread}$ ( $15 \times \text{IQR}$ )	3	3	7.5
Upper Inner Fence	Upper Hinge + 1 Step (75th percentile + $15 \times \text{IQR}$ )	14	14	18.5
Lower Inner Fence	Lower Hinge - 1 Step (25th percentile - $15 \times \text{IQR}$ )	6	6	-15
Upper Outer Fence	Upper Hinge + 2 Steps (75th percentile + $3 \times \text{IQR}$ )	17	17	26
Lower Outer Fence	Lower Hinge - 2 Steps (25th percentile - $3 \times \text{IQR}$ )	3	3	-9
Upper Adjacent	Largest value below Upper Inner Fence	13	13	13
Lower Adjacent	Smallest value above Lower Inner Fence	7	7	3
Outside Value (Outliers)	A value beyond an Inner Fence but not beyond an Outer Fence			
Far Out Value (Extreme Values)	A value beyond an Outer Fence			30



# Fuel Efficiency vs (Transmission/ Cylinder count)



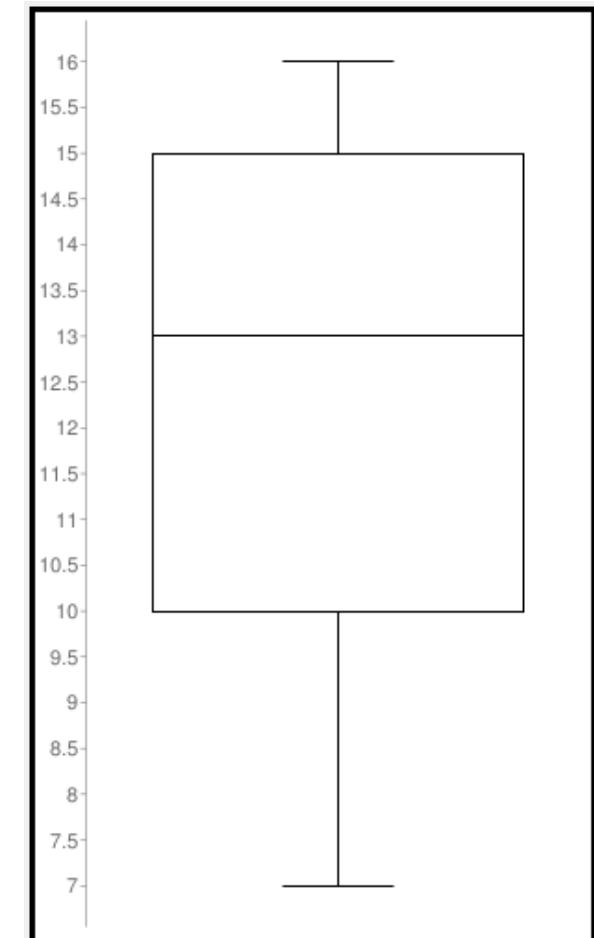
Source: MTCARS dataset

# Interpreting Box-whiskers plot

## Age of kids in a party

Which of the following statements are true?

- All of the students are less than 17 years old
- Atleast 75% of the students are 10 years old or older
- There is only one 16 year old at the party
- The youngest kid is 7 years old
- Exactly half the kids are older than 13



# Outlier detection – Excel and Box Plot Steps

Hadlum vs Hadlum case



Source: <http://www.alphamom.com/legacy/pregnancy-calendar/week36.jpg>

Last accessed: November 01, 2014



Source: <http://3.bp.blogspot.com/-0YwIRjLMWr0/T4DqOwVClgI/AAAAAAAAGg/Yjf-ttkQLSg/s1600/fishy.jpg>

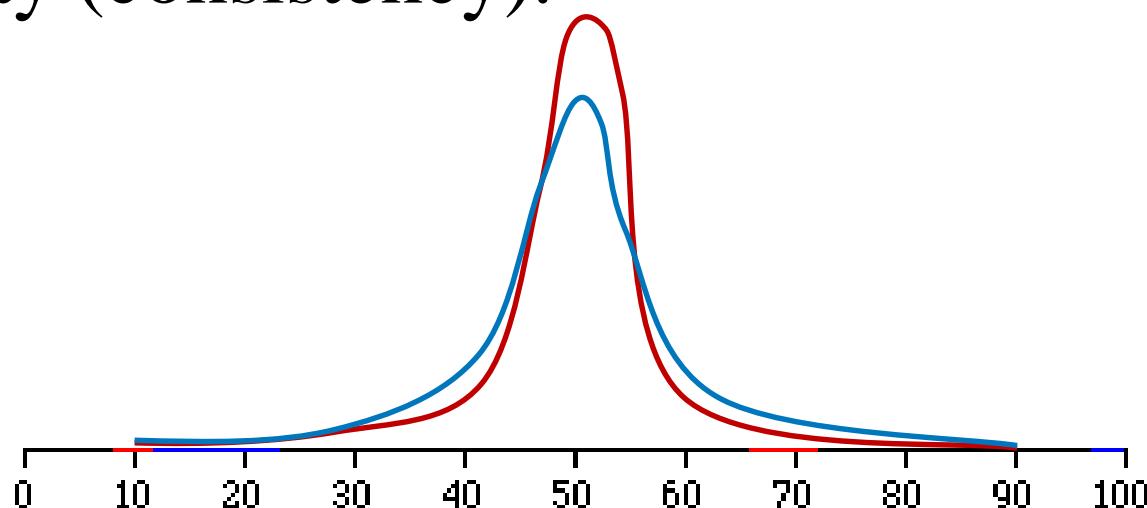
Last accessed: November 01, 2014

CSE 7315C



# Measuring Variability and Spread

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



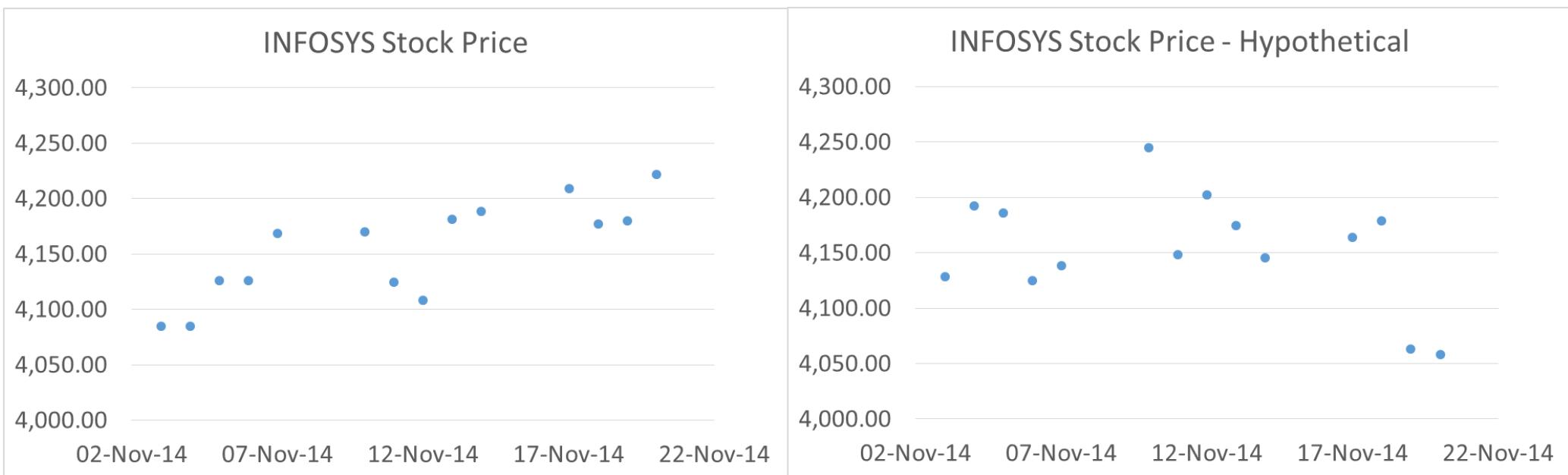
Average distance from the mean?

3 3 6 7 7 10 10 11 13 30

CSE 7315C



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

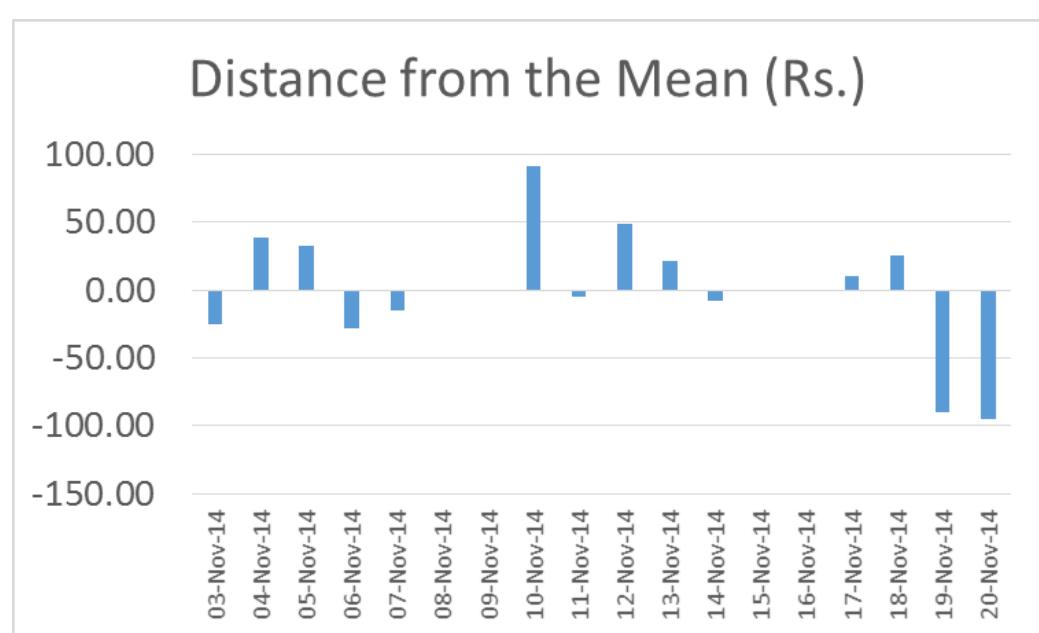
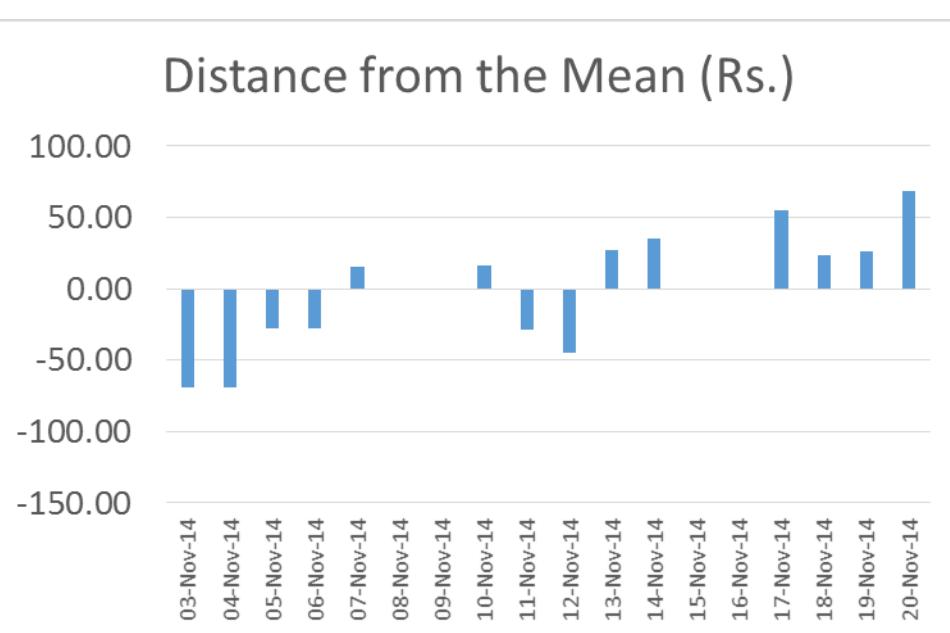


Data Source: <https://in.finance.yahoo.com/q/hp?s=INFY.BO>

CSE 7315C



# 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.

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

# Measuring Variability and Spread

$$\text{Variance} = \frac{\sum(x-\mu)^2}{n} = \frac{\sum x^2}{n} - \mu^2 \text{ (Derive)}$$

3 3 6 7 7 10 10 10 11 13 30

Units are squared, which is not intuitive.

Standard Deviation,  $\sigma = \sqrt{\text{Variance}}$



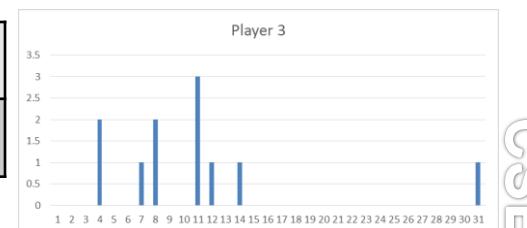
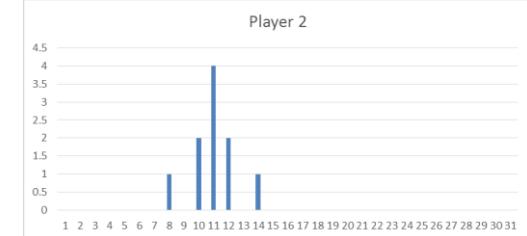
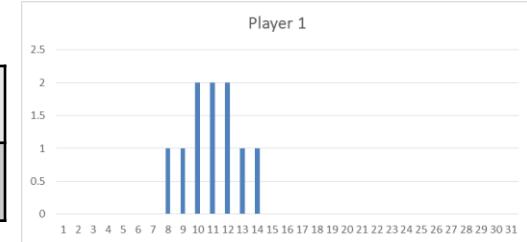
# Measuring Variability and Spread

Calculate standard deviation for each player.

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

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

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



1.73, 1.48, 7.02

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

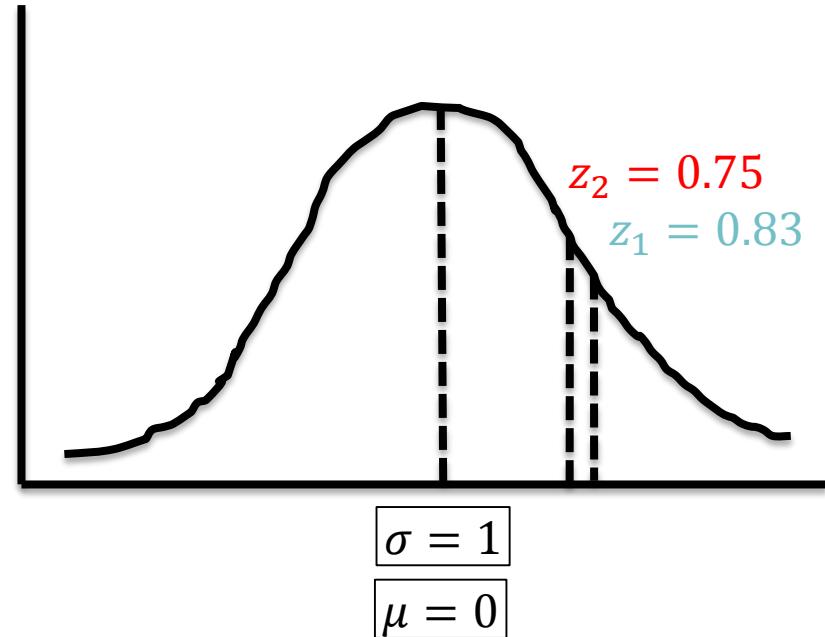
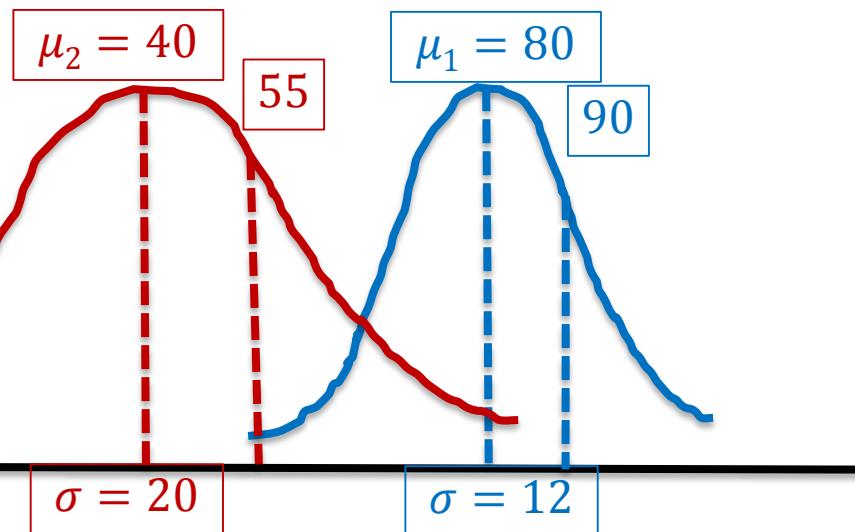
Imagine 2 players with different abilities: one has an average of 80% with 12% Stdev and the other 40% with 20% Stdev.

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

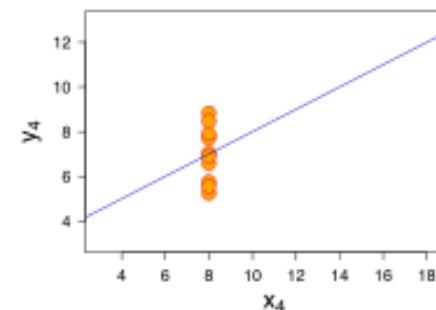
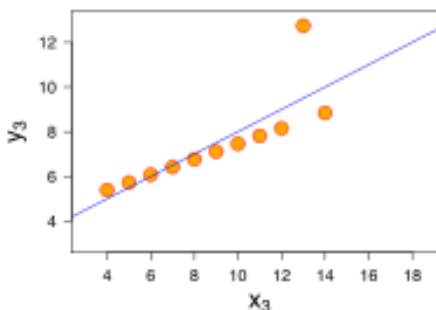
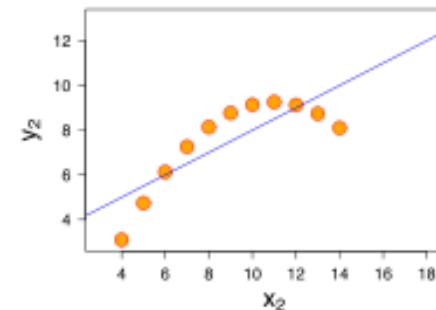
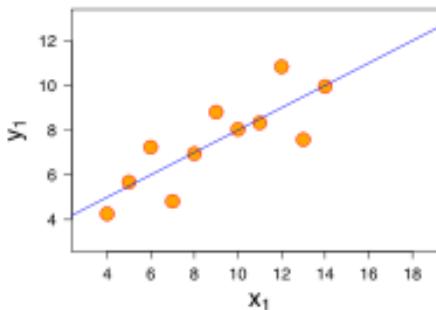
Standard score,  $z = \frac{x-\mu}{\sigma}$ , # of stdevs from the mean



# 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)



# PROBABILITY BASICS

CSE 7315C



# Understanding Probability

- Coin Toss – Probability of Head is  $\frac{1}{2}$
- Weather – Probability of thunderstorm tomorrow is 25%
- Election – Probability of AAP win in Punjab is 50%

# Probability vs Statistics

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

Questions addressed -

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

CSE 7315C



# Probability vs Statistics



Probability is the basis of inferential statistics.

CSE 7315C



# Probability - Applications

8 National Vital Statistics Reports, Vol. 54, No. 14, April 19, 2006

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

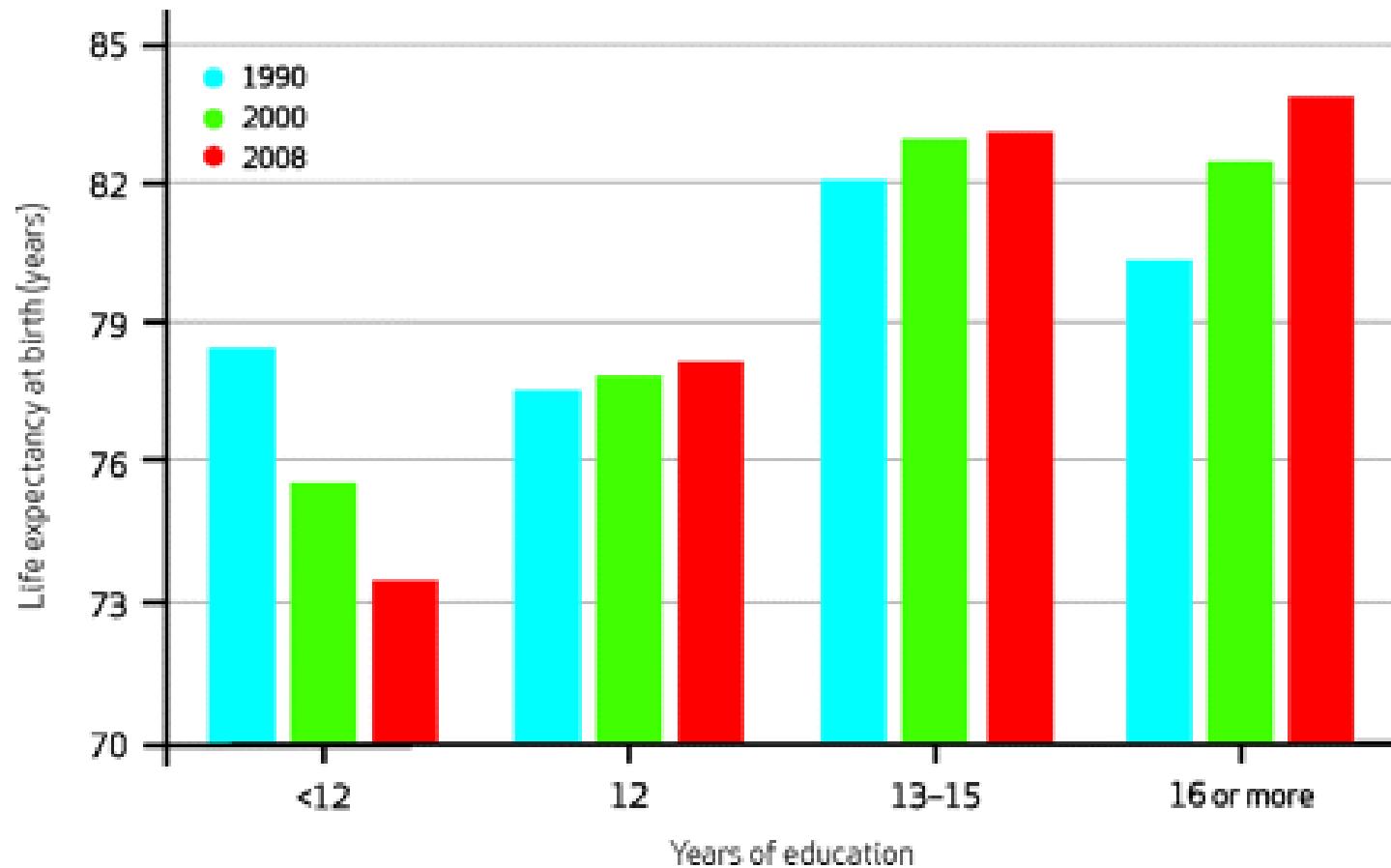
Age	Probability of dying between ages $x$ to $x+1$	Number surviving to age $x$	Number dying between ages $x$ to $x+1$	Person-years lived between ages $x$ to $x+1$	Total number of person-years lived above age $x$	Expectation of life at age $x$
	$q_x$	$l_x$	$d_x$	$L_x$	$T_x$	$e_x$
0-1	0.006865	100,000	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.

CSE 7315C



# Probability - Applications



# Probability - Applications

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



# Assigning Probabilities

## Classical Method – *A priori* or Theoretical

Probability can be determined prior to conducting any experiment.

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

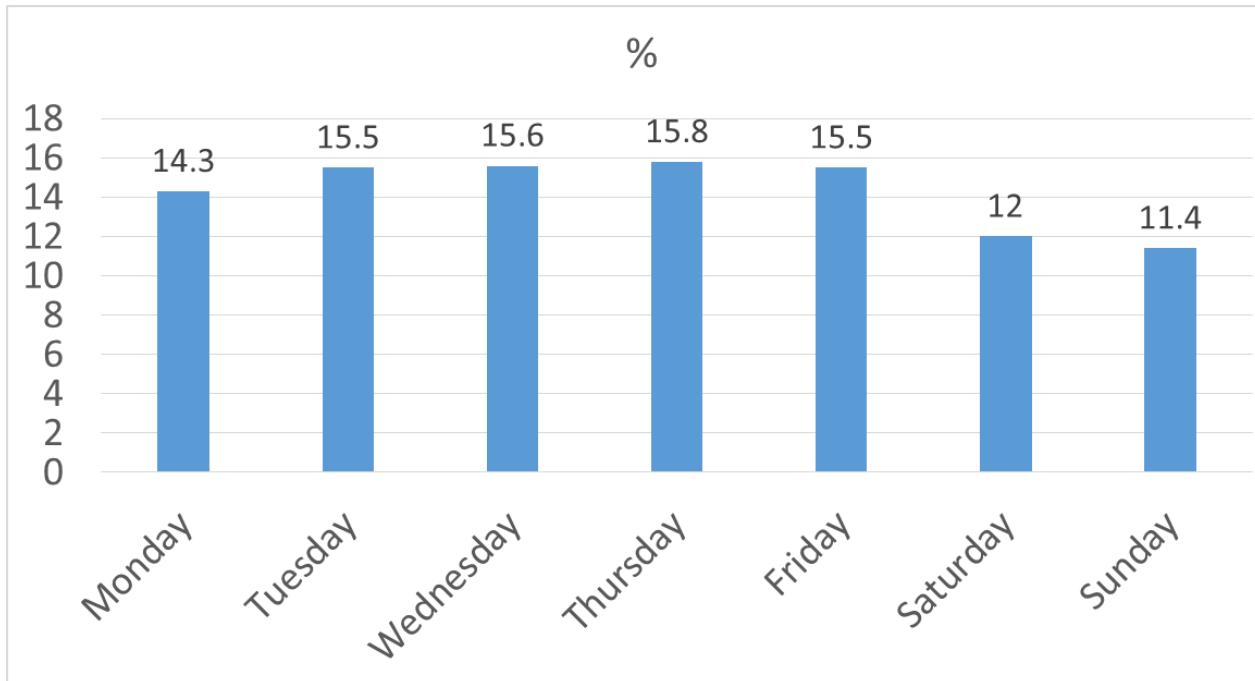
Example: Tossing of a fair die



# Assigning Probabilities

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

$$1/7 = 14.3\%$$



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

# Assigning Probabilities

## Empirical Method – *A posteriori* or Frequentist

Probability can be determined post conducting a thought experiment.

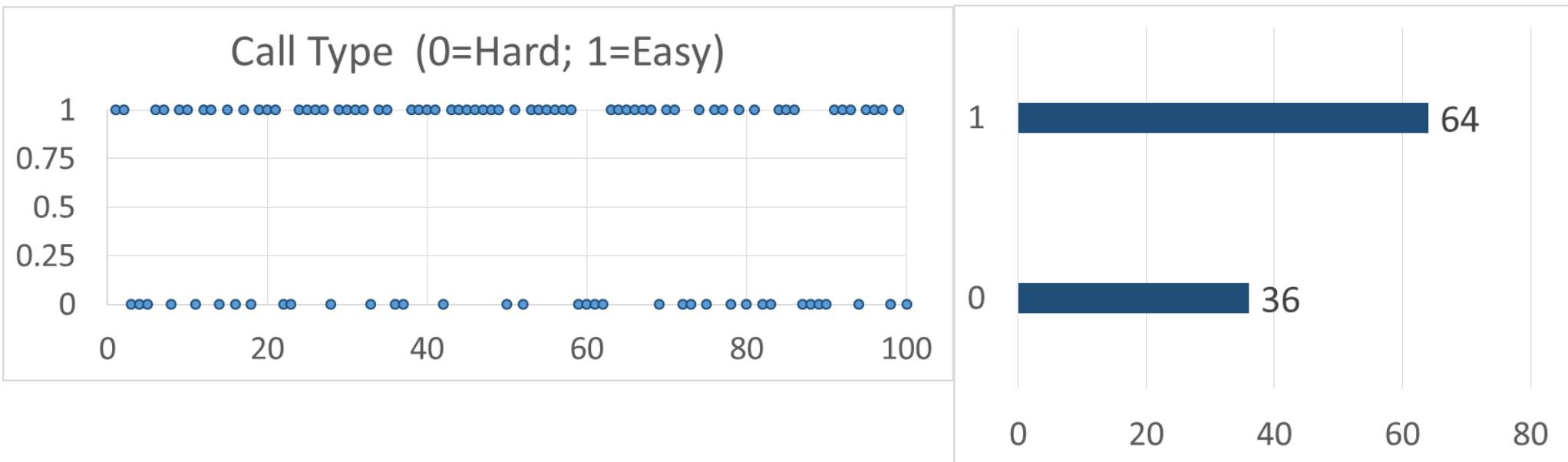
$$P(E) = \frac{\text{\# of times an event occurred}}{\text{total \# of opportunities for the event to have occurred}}$$

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

This is the most used method in statistical inference.

# Assigning Probabilities

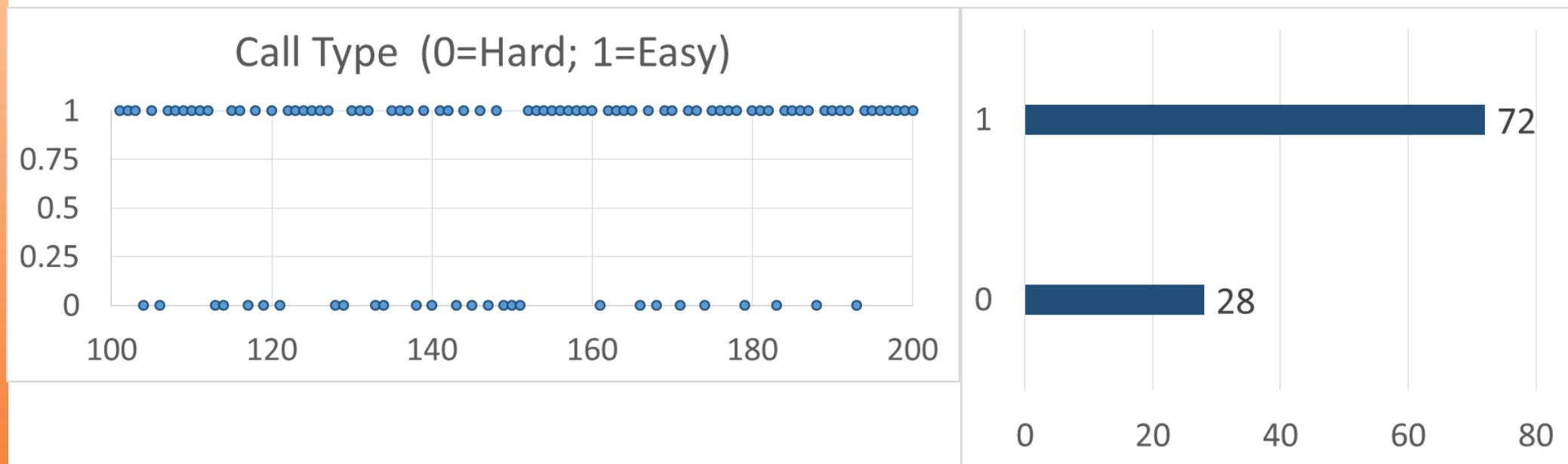
Empirical Method – *A posteriori* or Frequentist  
100 calls handled by an agent at a call centre



# Assigning Probabilities

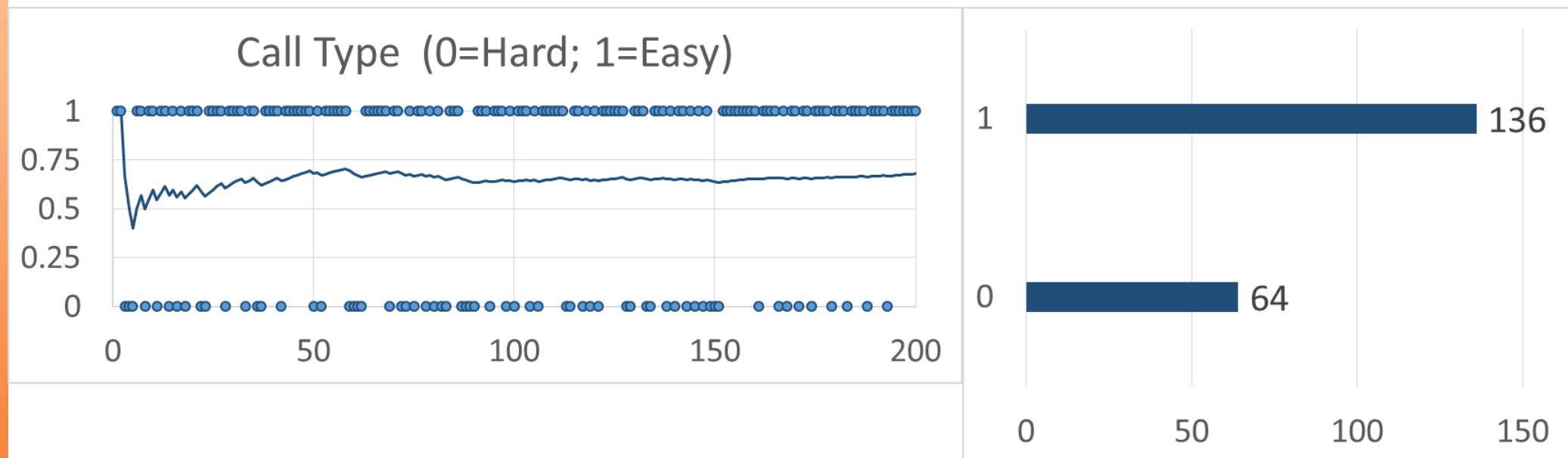
Empirical Method – *A posteriori* or Frequentist

Next 100 calls handled by an agent at a call centre



# Assigning Probabilities

Empirical Method – *A posteriori* or Frequentist  
Averages over the long run



$$P(\text{easy}) = 0.7$$

# Assigning Probabilities

## Subjective Method

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

What is the probability of India winning the match tomorrow?

CSE 7315C

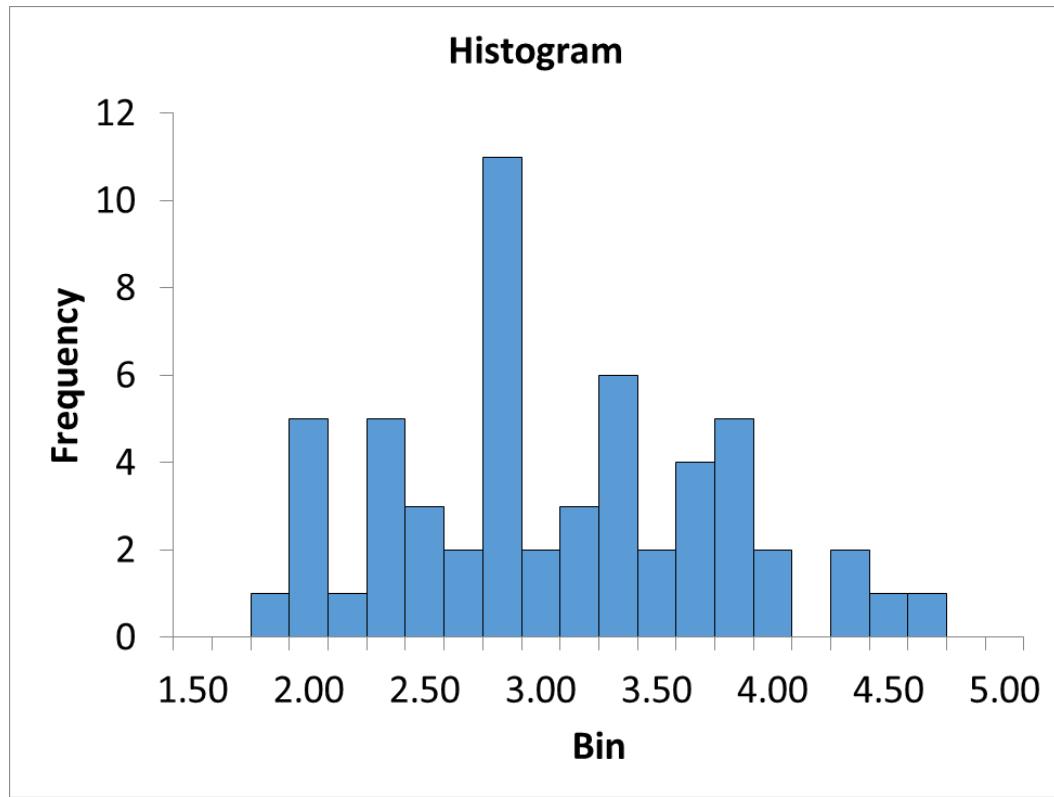


# Assigning Probabilities

## Subjective Method

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

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



Actual growth 2.5%

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

# Probability - Terminology

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

Example:

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

**Event** – A subset of the sample space.

An Event of interest might be - HH



# Probability - Rules

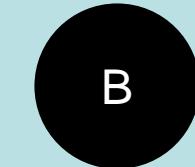
S

S

A

S

A



A and B are **mutually exclusive**

$$P(S) = 1$$

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

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

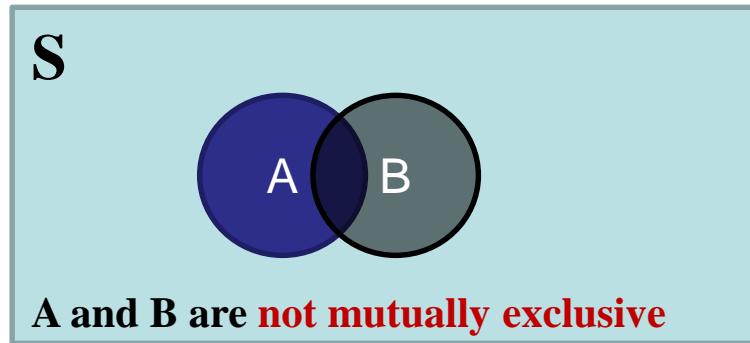
Area of the rectangle denotes sample space, and since probability is associated with area, it cannot be negative.

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

CSE 7315C



# Probability - Rules



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

## Example

Event A – Customers who default on loans

Event B – Customers who are High Net Worth Individuals

# Probability - Rules

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

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

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

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

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

# Probability - Question

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

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

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

What are the assumptions, if any?



## **International School of Engineering**

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

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

For Corporates: +91-9618483483

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

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

Twitter: <https://twitter.com/Insofeedu>

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

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

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

*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.*