Inspire…Educate…Transform.

Statistics and Probability Fundamentals

Basic Statistical Concepts, Measures

Dr. Anand Jayaraman

ajayaraman@gmail.com

Mar 25, 2017

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)

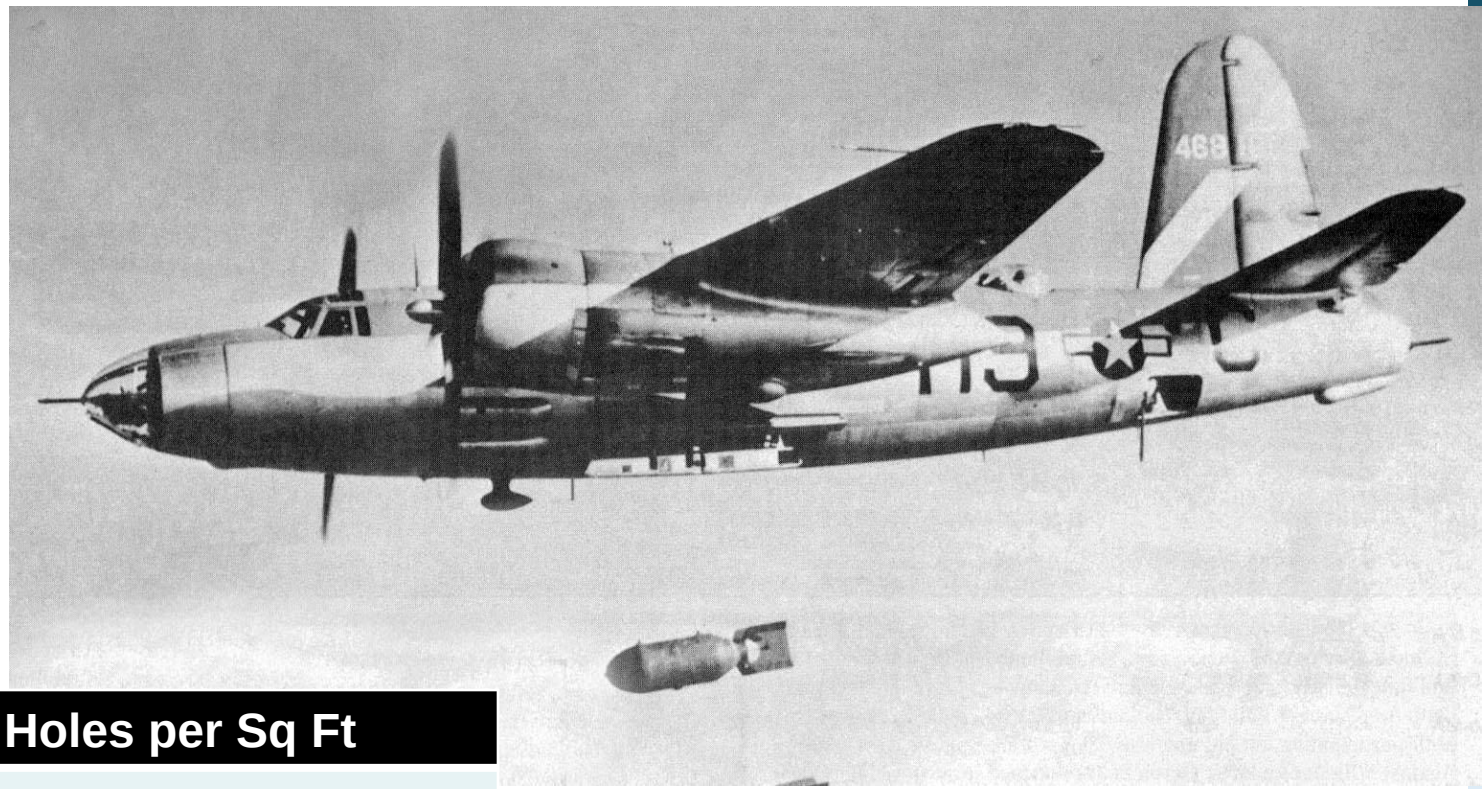
First thoughts on Maths

When am I going to use this?

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

Aircraft Armor Conundrum



Section of Plane	Bullet Holes per Sq Ft
Engine	1.1
Fuselage	1.73
Fuel system	1.6
Rest of the plane	1.8

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

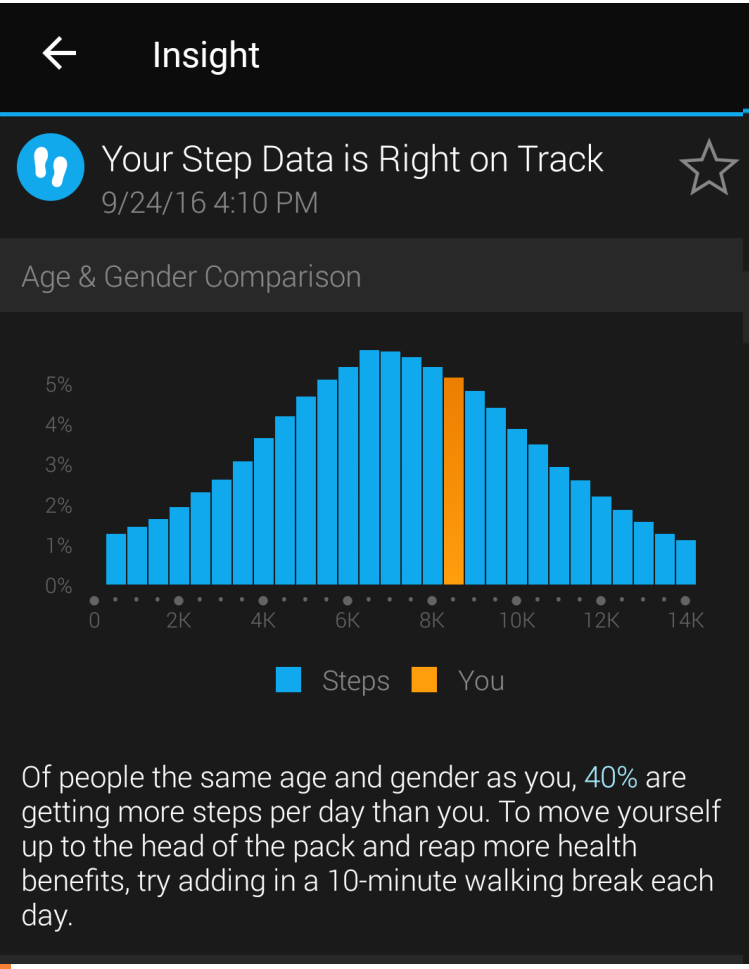
H. G. Wells

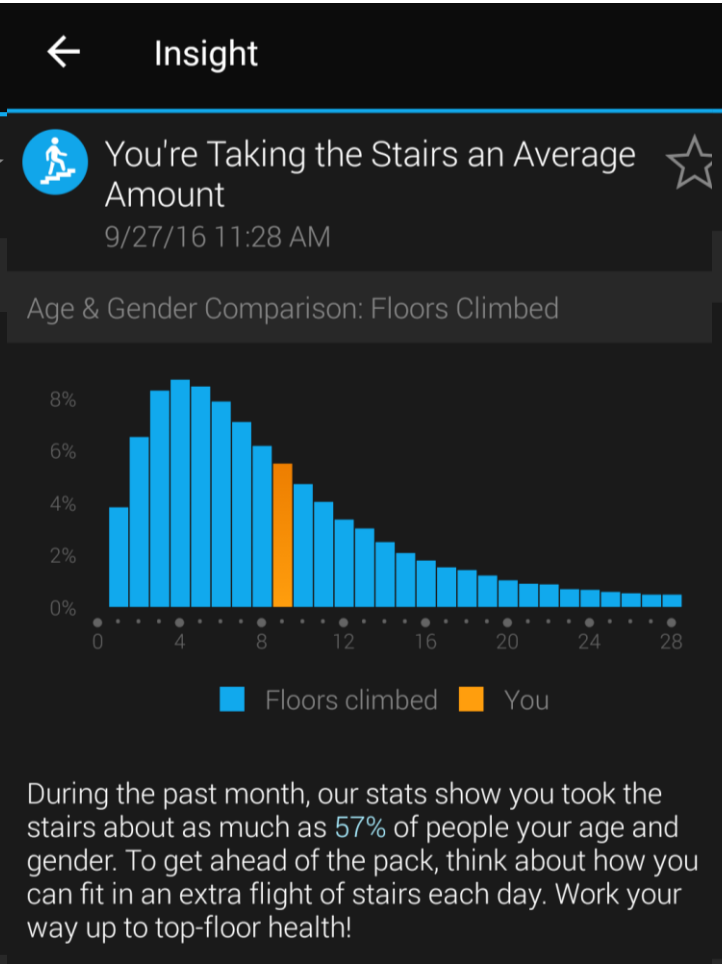
(1903)

Modern Life..

- Abundance of data
- Sometimes contradictory evidence
- Often, making sense of data is hard
- Mathematical sense essential to survive and flourish

Looking at Data





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

- Mark Twain / Benjamin Disraeli

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/

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

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

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

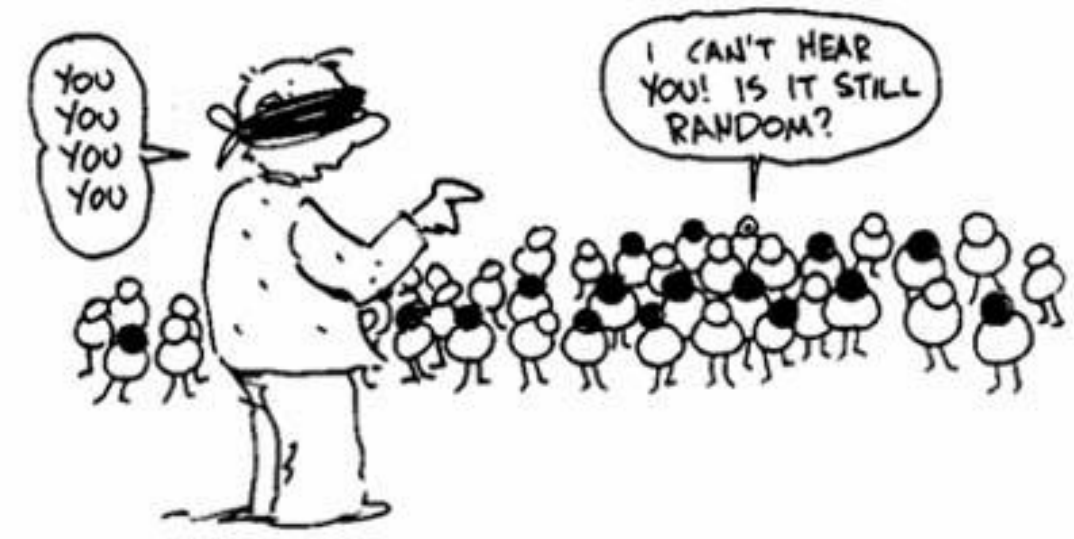
Data Gathering – Sampling Techniques

- Convenience Sampling

- Eg: Online polls, Asking your best friends etc

Data Gathering – Sampling Techniques

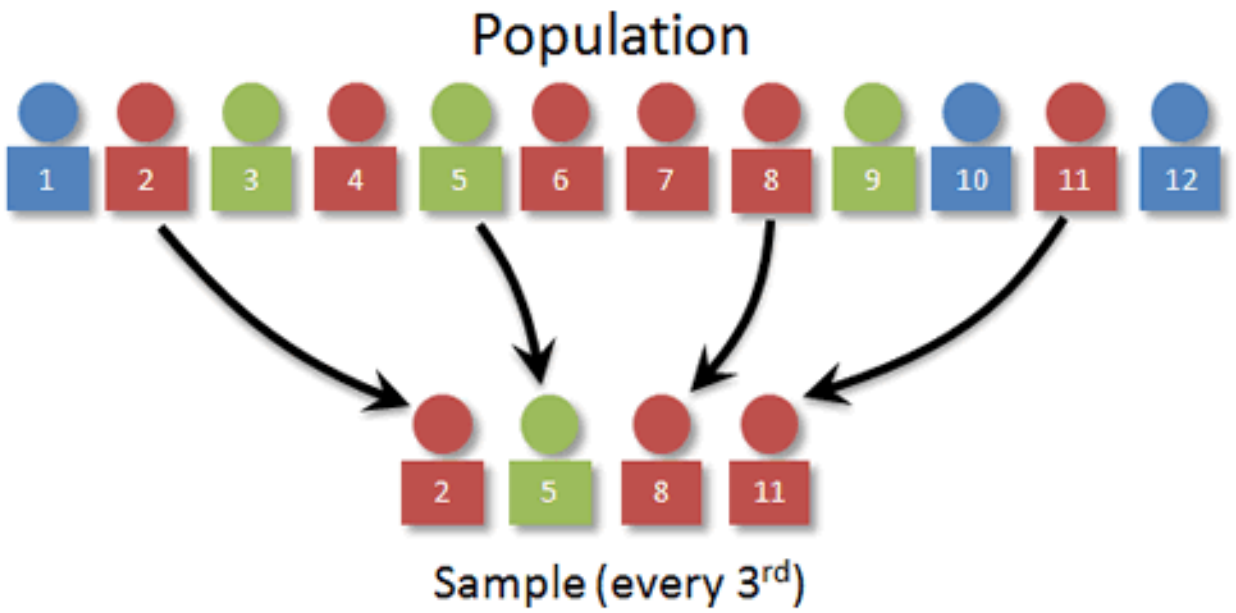
- Random Sampling



Each member has an equal chance of being selected.

Sampling Techniques

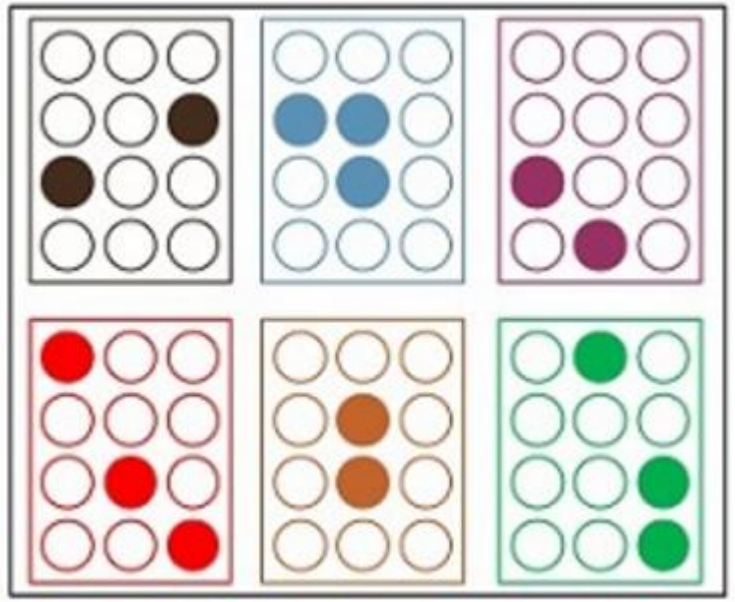
- Systematic Random Sampling



Example: Supermarket chooses every 10th or 15th customer entering the supermarket and conduct the survey.

Sampling Techniques

- Stratified Sampling
  - Divide the data into several relevant strata and then sample from each strata
- Eg: For getting an opinion on demonetization, one choice of strata might be state-wise analysis. We get 20 random volunteers from each and every state.



Sampling Techniques

- Cluster Sampling
  - Divide the population in to groups or clusters. Then select a one or a few clusters and survey **everyone** from the chosen subset.

Misusing statistics

A few examples -

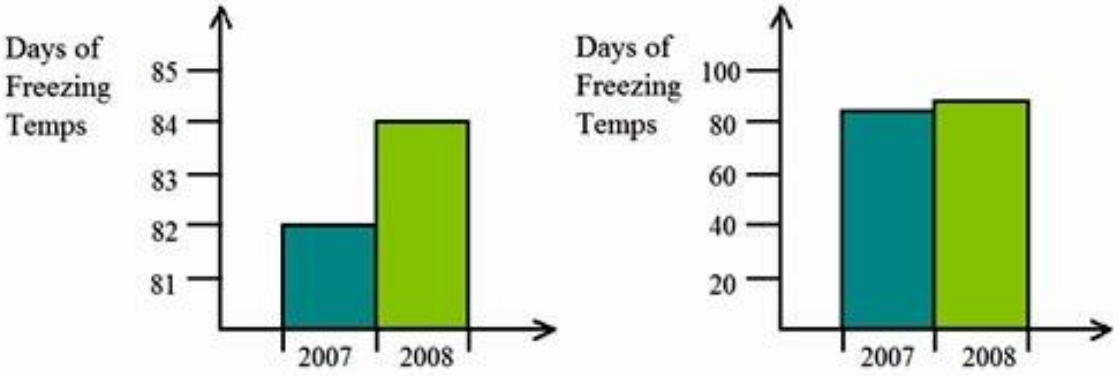
- Bad samples
  - How many of you think that Data Science is a promising career choice?
- Small samples
  - "3 out of 4 dentists agree that Colgate is the best toothpaste"
- Detached statistics
  - "Lays chips is 30% more tasty" Compared to what?
- Poor definition
  - How is "30% more tasty" defined/measured?

Misusing statistics

- Changing the subject
  - "During my administration, the expenses increased by a mere 3%"
  - "My opponent's administration caused an increased expenditure of 100 crores!"
- Semi-attached Figure (Implied connections)
  - "Waking up early is a good habit. Over 70% of the CEOs of the fortune 500 companies are early risers"
- Loaded Questions or Leading Questions
  - "Do you support PM's demonetization action to eliminate black money?"
  - https://www.youtube.com/watch?v=G0ZZJXw4MTA

Misleading Graphs

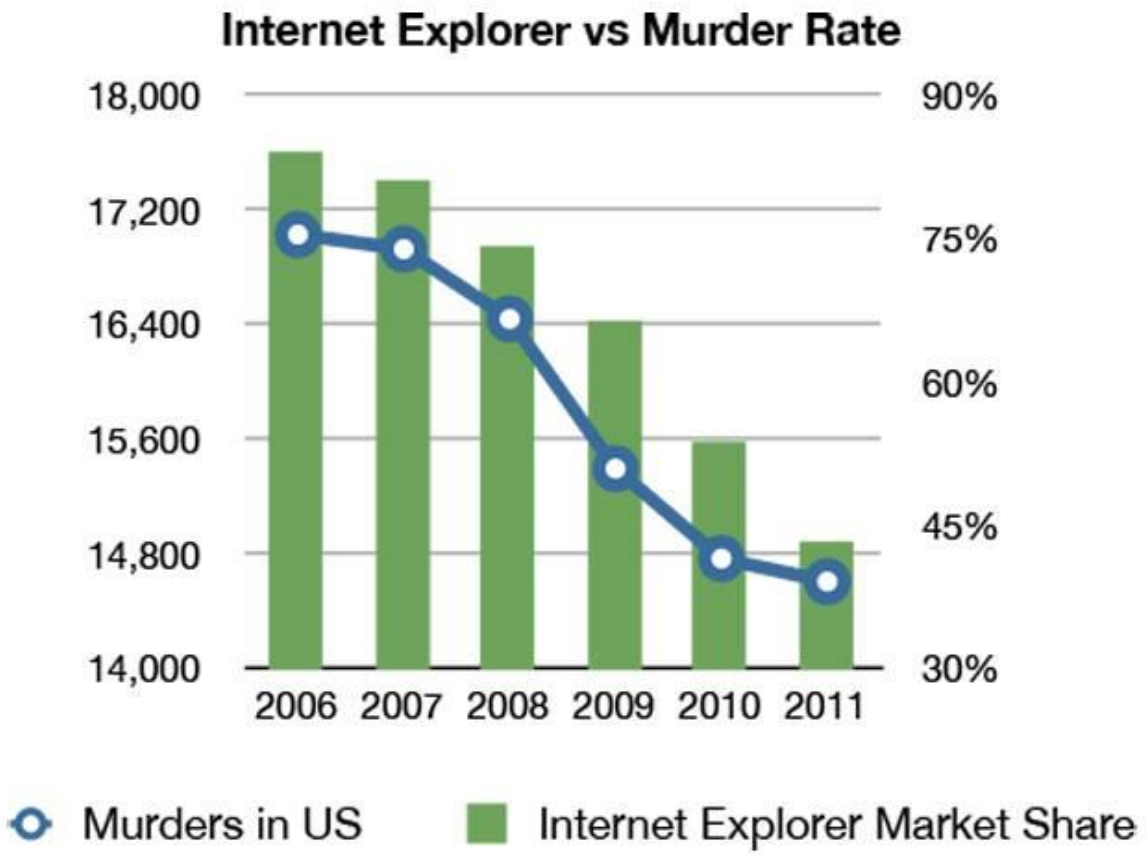
Compare the two graphs



Both show exactly the same data. However, the graph on the left makes the change appear to be much larger than it really is because the numbers on the vertical axis do not start at 0. Each vertical mark on the left graph represents 1 and each mark on the right represents 20 (the scale changes).

Image Source: http://www.his.washk12.org/

Correlation vs Causation

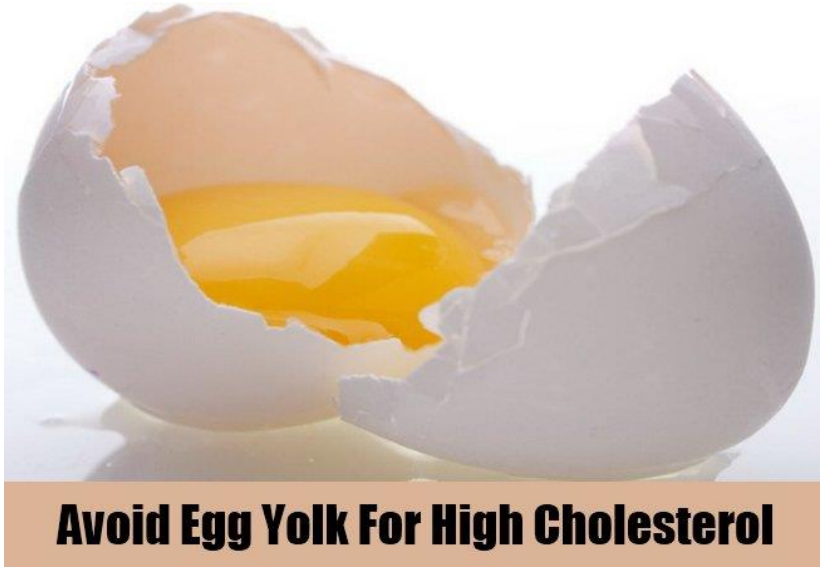


Catching Statistical Errors

Ask these questions:

- Who says so?
- How do they know?
- Watch out for stated & unstated data
- Is the sample large enough?
- Did somebody change the subject?
- Does it make sense?

The Cholesterol Confusion



Source: http://www.searchhomeremedy.com/10-high-cholesterol-foods-to-avoid/

GOOD NEWS

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 *Internation Business Times*.

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.

The 70s, 80s and 90s

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

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 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 AHA/ACC (American Heart Association / American College of Cardiology) report. Cholesterol is not a nutrient of concern for overconsumption."

The Dietary Guidelines Advisory Committee will, in response, no 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-cholersterol foods in their diets to avoid heart disease and strokes.

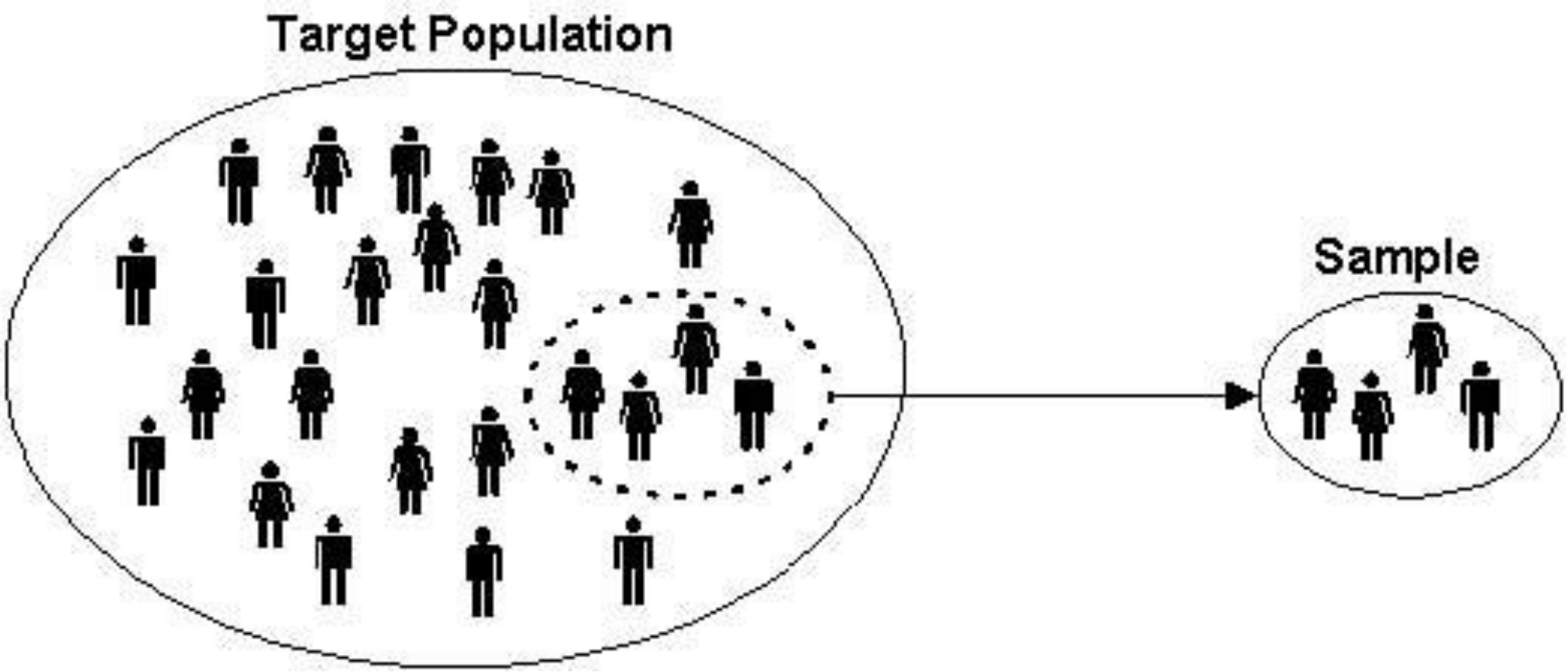
But nutritionists and scientists have long been campaigning for the U-turn, which started with introducing "good cholesterols" back into the 'safe zone'.

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*

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

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

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

Parameter and Statistic

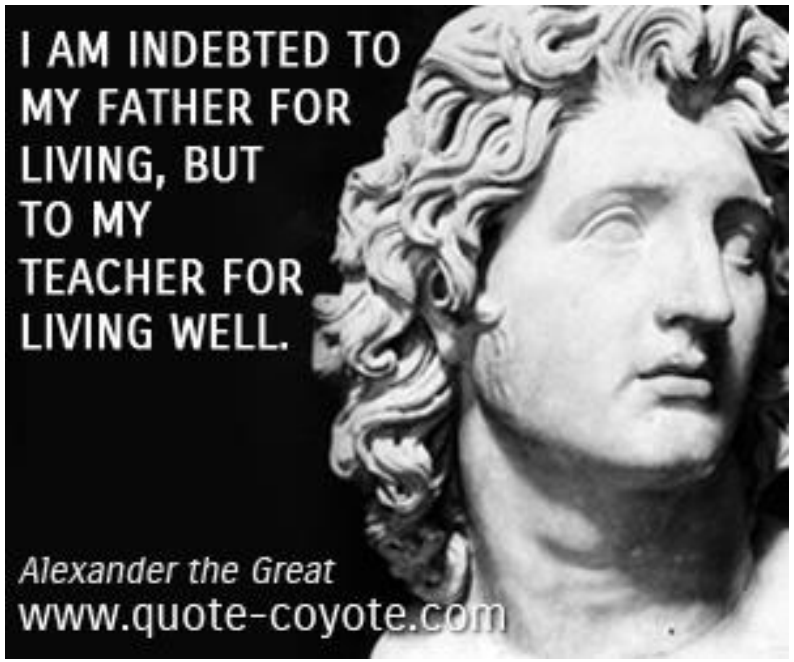
Parameter: A descriptive measure of the **population**.

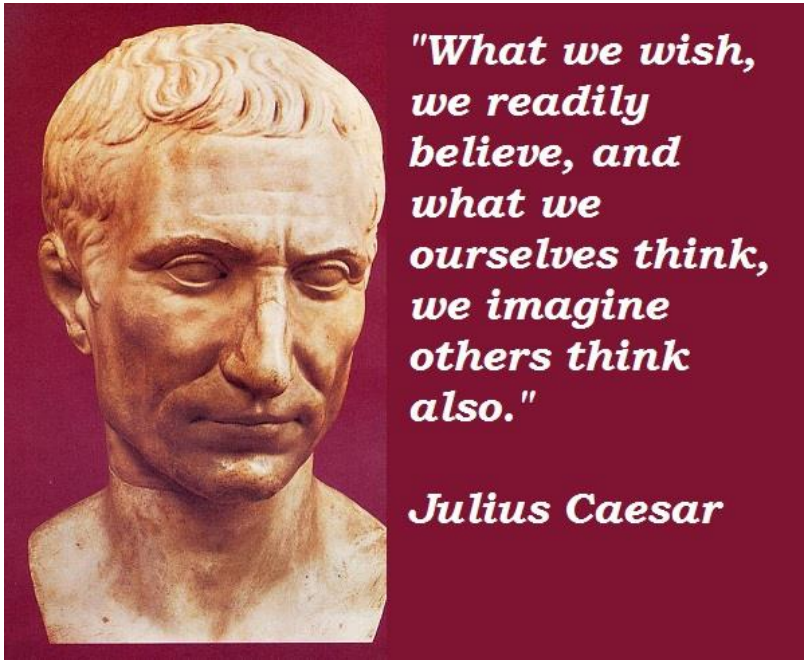
For example, population mean, population variance, population standard deviation, etc.

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

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

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

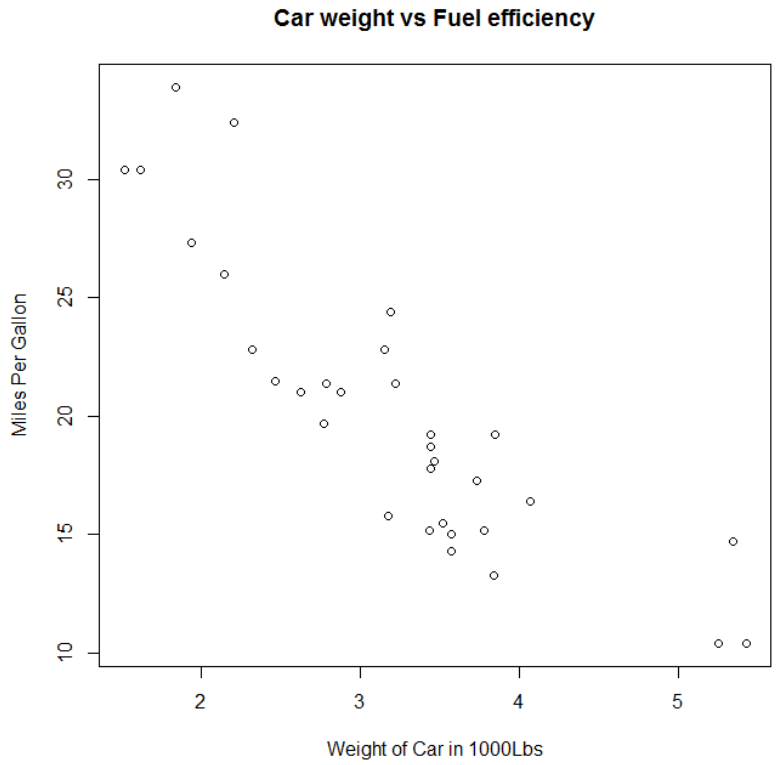
Variables and Data

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

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.

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

Temparature: 14C vs 28C

Quantitative Data - Ratio

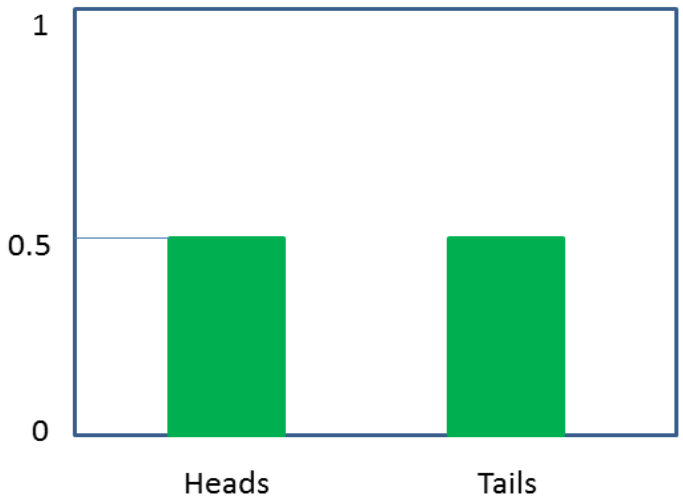
Ratio level data is similar to Interval level data, with the key difference – there is a natural zero point.

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

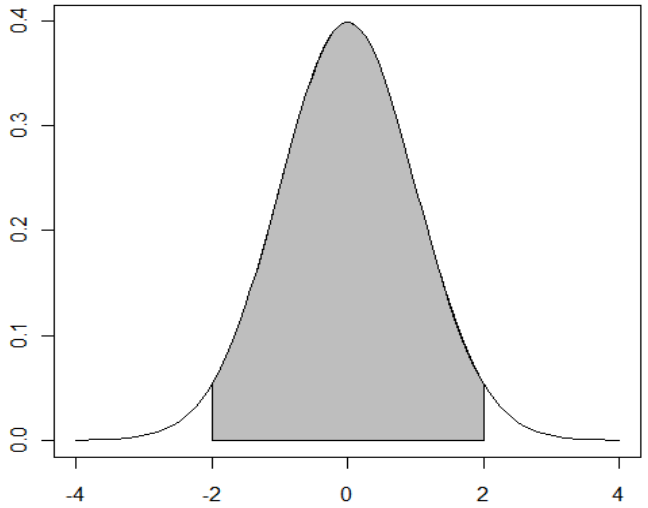
Summary of Levels of Data Measurement

- 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

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

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

Source: http://www.montecitoheightsstudios.com/yoga-for-teens
Last accessed: June 05, 2015

The Central Tendencies

Power workout class composition

Age (years)	13	15	17	90
Frequency, f	4	6	3	1

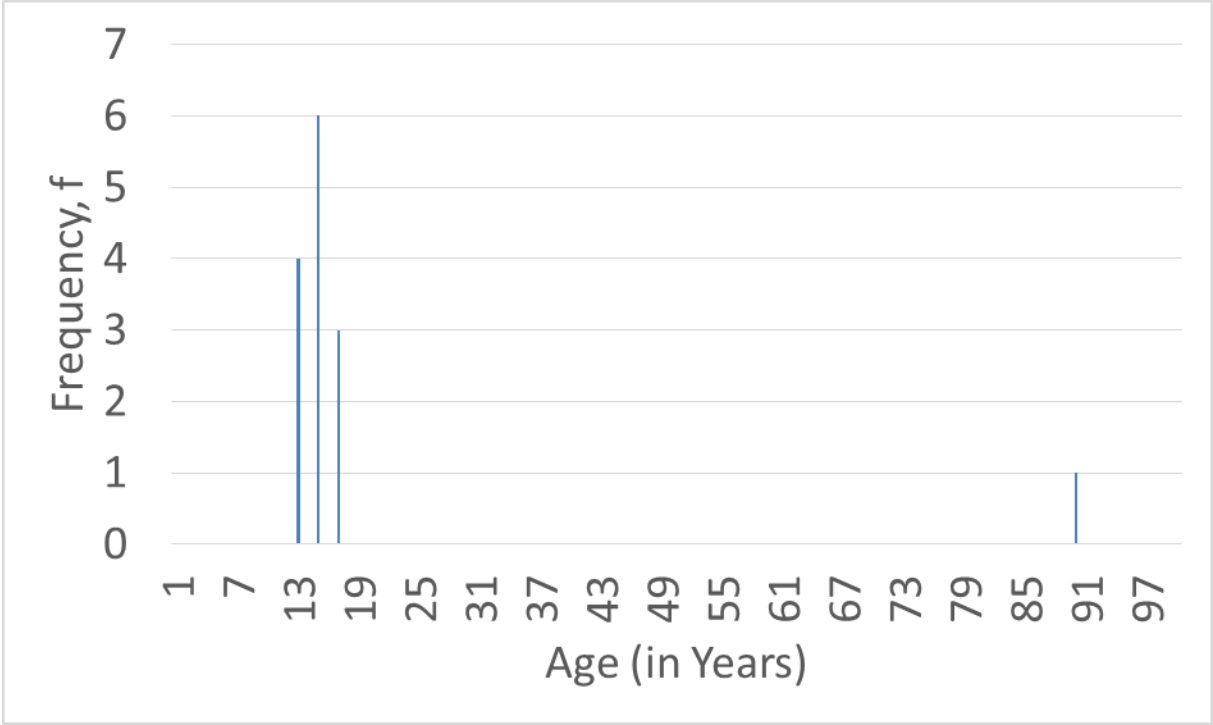
Mean, $\mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{13X4+15X6+17X3+90X1}{4+6+3+1} = 20$

But nobody in Sai's age group ☹.

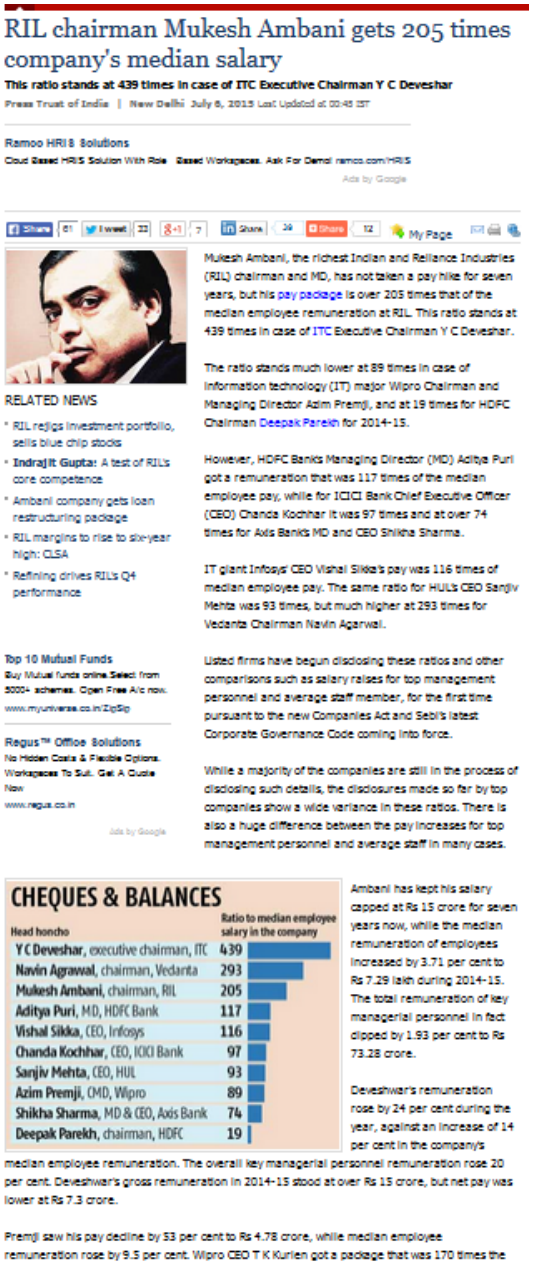
The Central Tendencies

Power workout class composition

Age (years)	13	15	17	90
Frequency, f	4	6	3	1



The Central Tendencies



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, 2015 Last Updated at 00:45 IST

Ramco HRIS Solutions
Cloud Based HRIS Solution With Role Based Workspaces. Ask For Demo! ramco.com/HRIS
Ads by Google

Share 61 Tweet 22 +1 7 Share 20 Share 12 My Page

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 Navin Agarwal.

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.

RELATED NEWS

- RIL rejigs investment portfolio, sells blue chip stocks
- Indrajit Gupta: A test of RIL's core competence
- Ambani company gets loan restructuring package
- RIL margins to rise to six-year high: CLSA
- Refining drives RIL's Q4 performance

Top 10 Mutual Funds
Buy Mutual funds online.Select from 5000+ schemes. Open Free A/c now.
www.myuniverse.co.in/ZipSip

Regus™ Office Solutions
No Hidden Costs & Flexible Options. Workspaces To Suit. Get A Quote Now
www.regus.co.in

Ads by Google

CHEQUES & BALANCES

Head honcho	Ratio to median employee salary in the company
Y C Deveshar, executive chairman, ITC	439
Navin Agrawal, chairman, Vedanta	293
Mukesh Ambani, chairman, RIL	205
Aditya Puri, MD, HDFC Bank	117
Vishal Sikka, CEO, Infosys	116
Chanda Kochhar, CEO, ICICI Bank	97
Sanjiv Mehta, CEO, HUL	93
Azim Premji, CMD, Wipro	89
Shikha Sharma, MD & CEO, Axis Bank	74
Deepak Parekh, chairman, HDFC	19

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.

Deveshwar'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. Deveshwar'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

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

Power workout class composition

Age (years)	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, 15, 17, 17, 17, 90

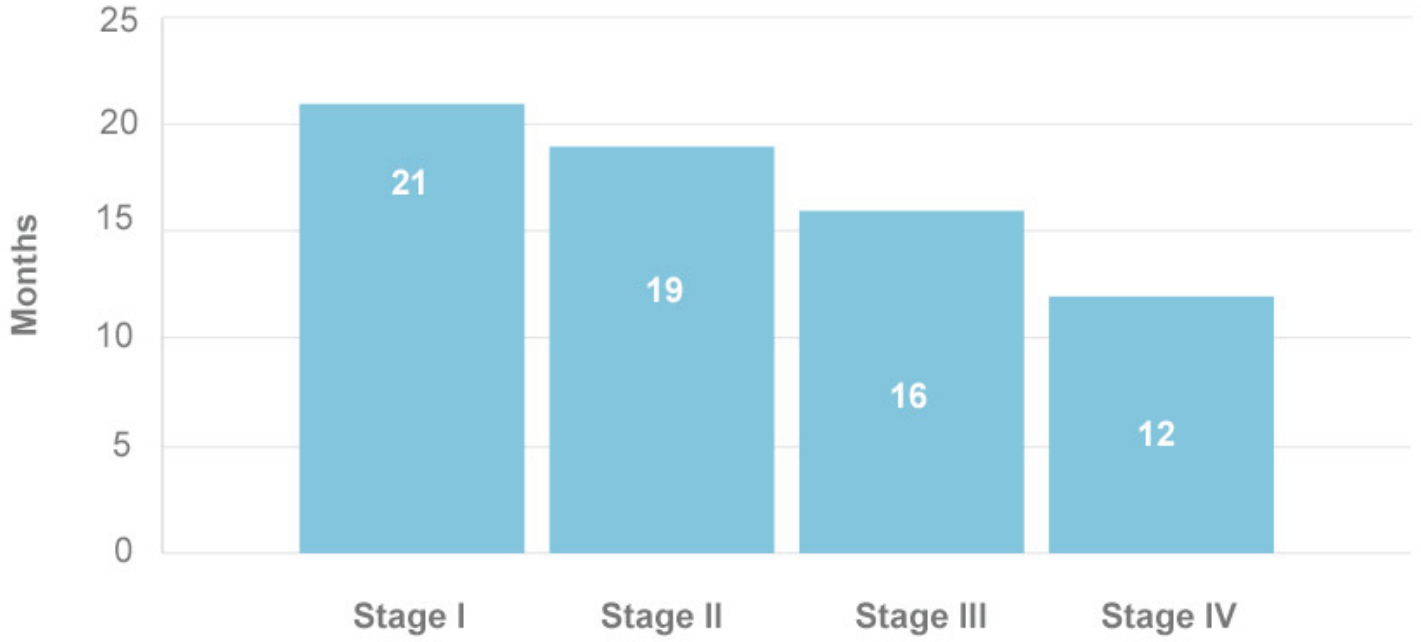
The Central Tendencies

Survival by Stage

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

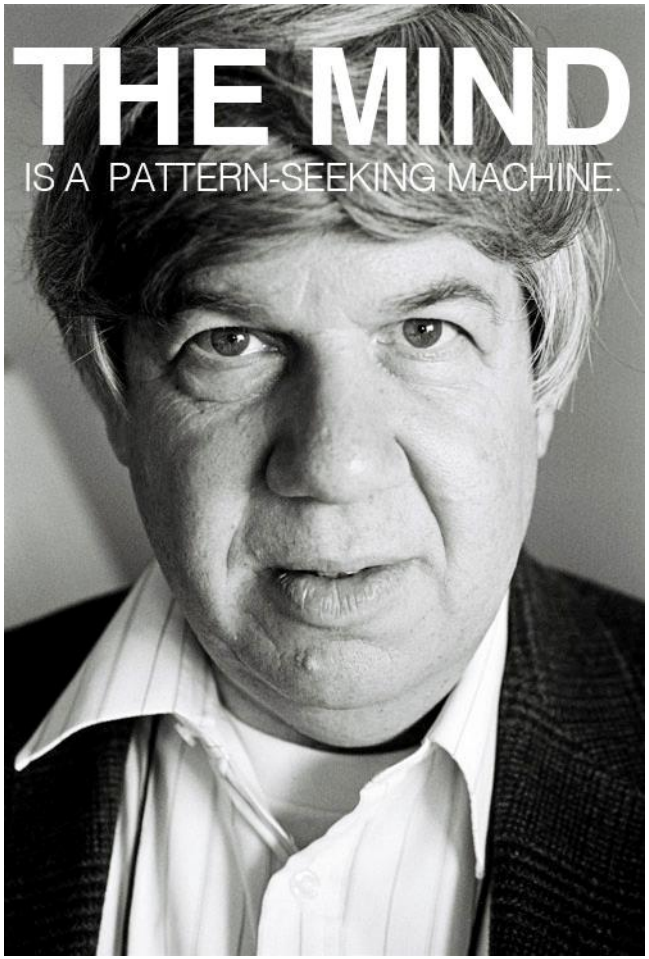
Patients diagnosed at stage I have the best outlook, while survival is worst at stage IV.

Median



Source: http://www.asbestos.com/mesothelioma/statistics.php
Last accessed: April 05, 2016

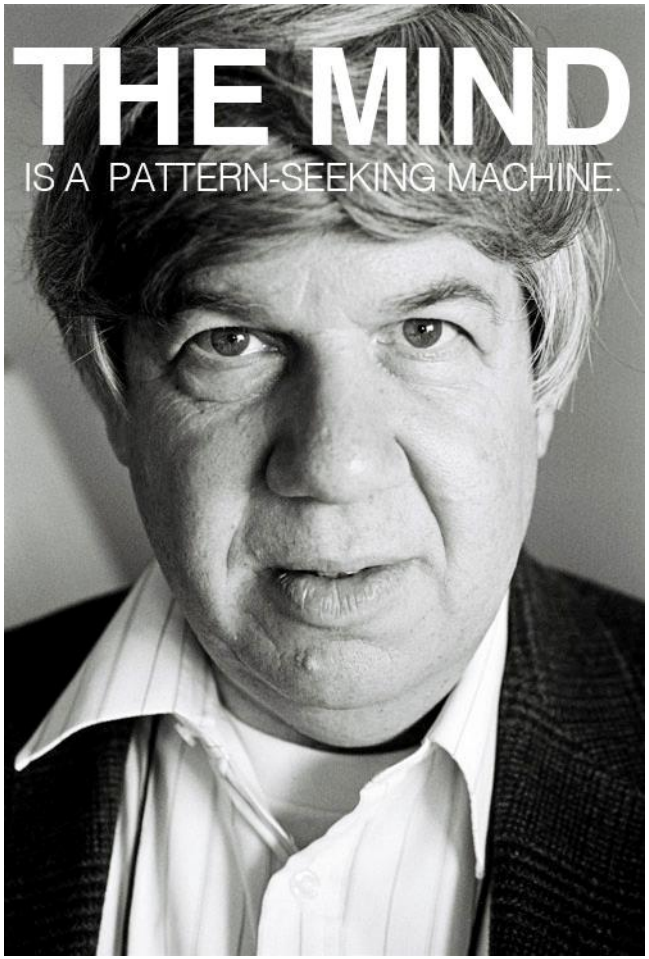
“The Median Isn’t the Message” by Stephen Jay Gould



“…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
Last accessed: April 06, 2016

The Median Isn't the Message by Stephen Jay Gould



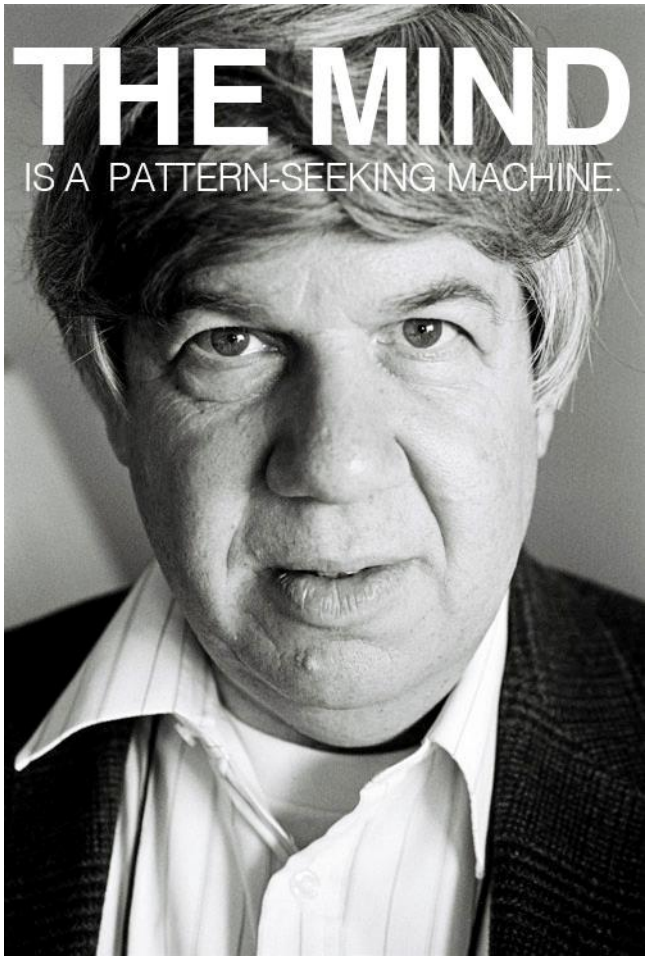
"…What does "median mortality of eight months" signify in our vernacular? I suspect that most people, without training in statistics, would read such a statement as "I will probably be dead in eight months" - the very conclusion that must be avoided, since it isn't so, and since attitude matters so much…"

"…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
Last accessed: April 06, 2016

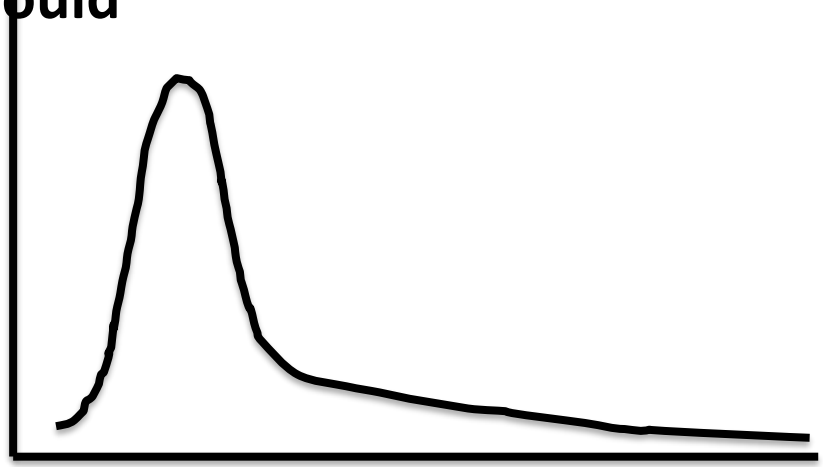
The Median Isn't the Message by Stephen Jay Gould



"…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
Last accessed: April 06, 2016

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.

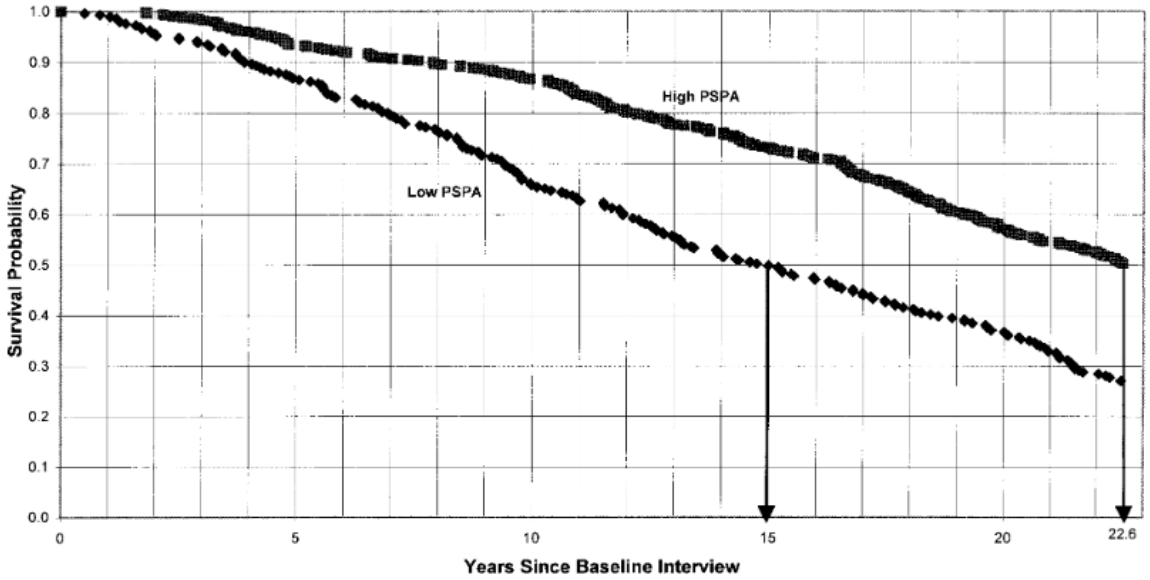


Figure 1. Influence of positive self-perceptions of aging (PSPA) on survival. Arrow indicates median survival.

Yale Study

CMU Study

Source: http://www.huffingtonpost.com/entry/positive-people-live-long_b_774648.html?section=india
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?

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

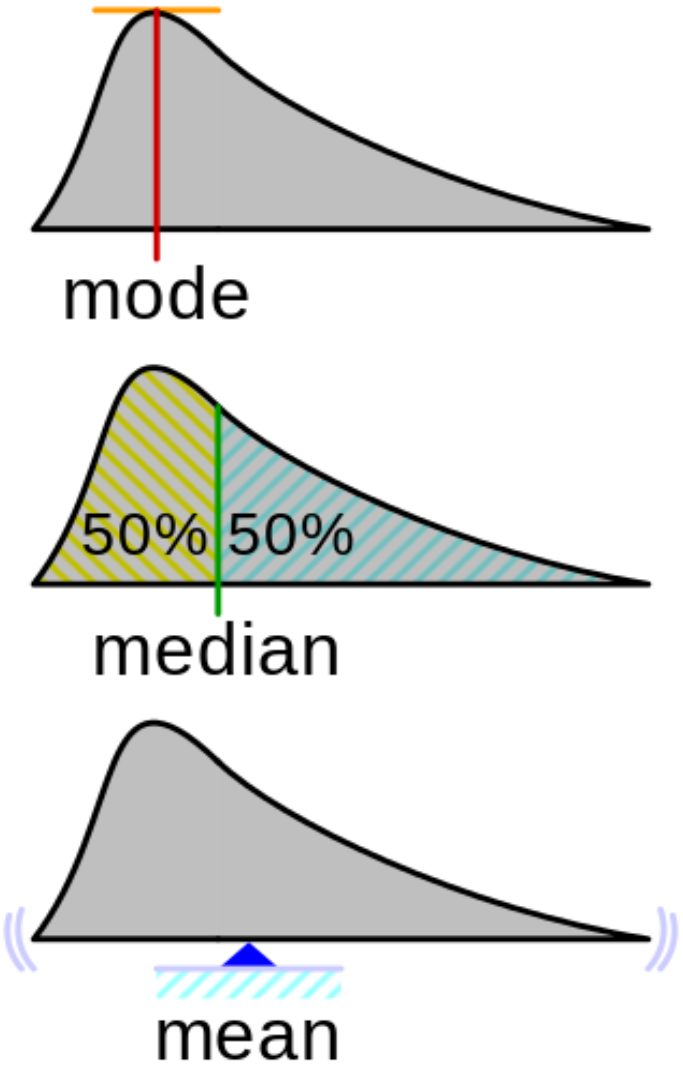
What is the Mode – the most frequently occurring data point?

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

Median = 10

3 3 6 7 7 10

First Quartile = 6.5

10 10 10 11 13 30

Third Quartile = 10.5

Quartiles

- Quartiles : division of the data set into 4 regions

If we have *n* data-points then the Quartile boundaries are given by

Lower quartile (25th percentile, Q1) =($\frac{1*(n-1)}{4} + 1$)th

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

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

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

Percentiles

- Percentile: divide the data set into 100 regions
- pth percentile = = $\left(\frac{p*(n-1)}{100}+1\right)$th

```
> aa
 [1]  3  3  6  7  7 10 10 10 11 13 30
> quantile(aa)
  0%  25%  50%  75% 100%
 3.0  6.5 10.0 10.5 30.0
> quantile(aa,0.3)
30%
  7
> quantile(aa,0.15)
15%
4.5
```

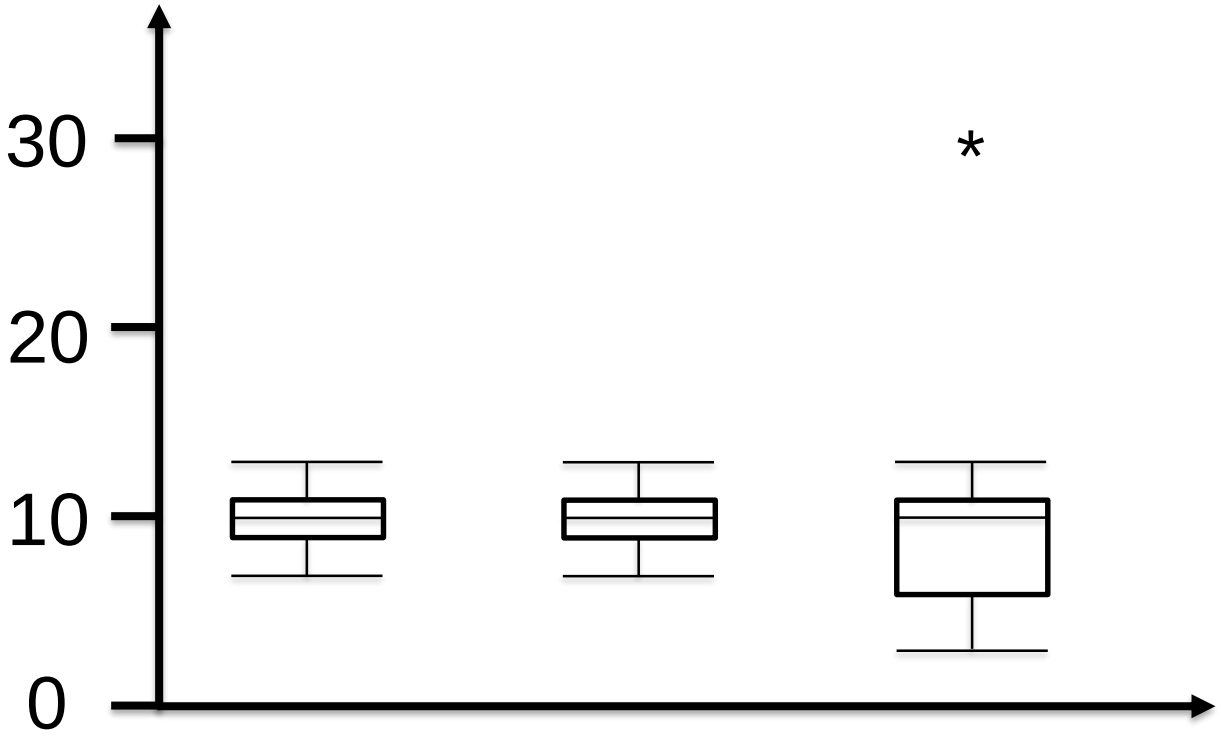
Box-Whisker Plot

- The Box and Whisker plot allows you to visualize the spread in the data easily
- Steps
  - Compute the Q1, Median and Q3 for the data. Compute IQR=Q3-Q1
  - The Box of the plot is drawn from Q3 to Q1 (50% of data is contained within the box)
  - The Whiskers are a maximum of 1.5*IQR from the top and the bottom of the box.
  - If there are no data points at 1.5*IQR, then pick an actual data point within the range of the Whiskers
  - Points lying outside the 1.5*IQR from the box ends are considered as Outliers.

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Box and whisker diagram or Box plot

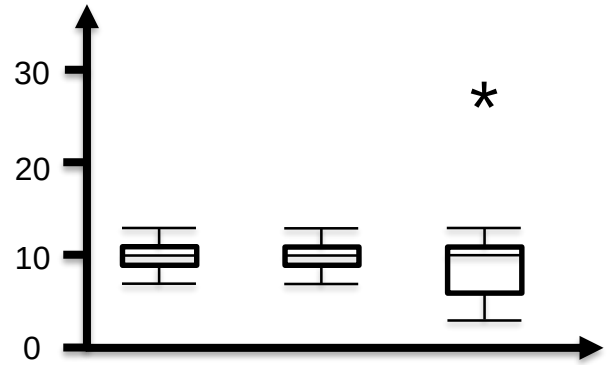


https://www.khanacademy.org/video/constructing-a-box-and-whisker-plot

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

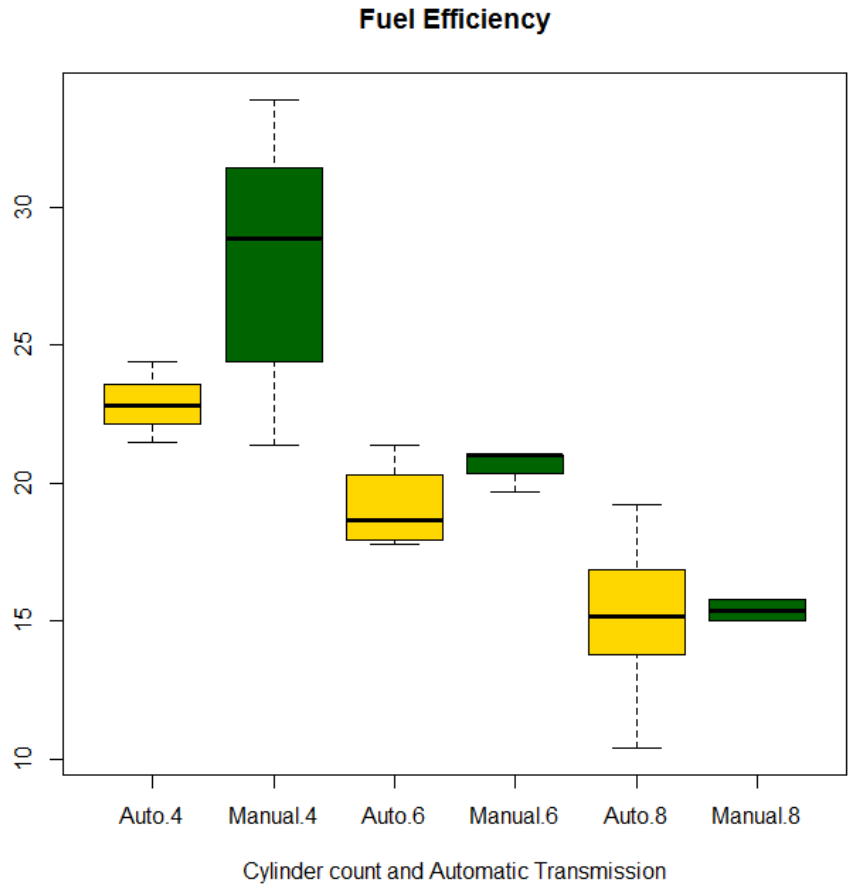
Box and whisker diagram or Box plot



Tukey fences

Name	Formula	Player 1	Player 2	Player 3
Lower Hinge	Q1 = 1st Quartile	9	9	6.5
Mid Line	Q2 = 2nd Quartile = Median	10	10	10
Upper Hinge	Q3 = 3rd Quartile	11	11	10.5
Body of the box	IQR = Q1 - Q3	2	2	4
Step	1.5* IQR	3	3	6
	Lower Hinge - 1 Step	6	6	0.5
	Upper Hinge + 1 Step	14	14	16.5
Lower Fence	Smallest Actual Data Inside Fence	7	7	3
Upper Fence	Largest Actual Data Inside Fence	13	13	13
Outliers	Value beyond the Fence			30

Fuel Efficiency vs (Transmission/ Cylinder count)



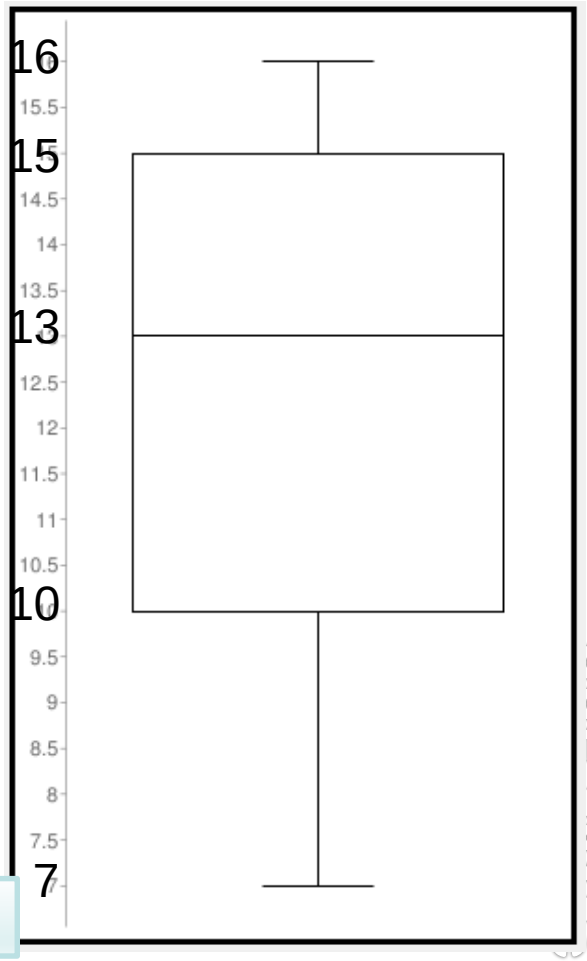
Source: MTCARS dataset

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 — True
- Atleast 75% of the students are 10 years old or older — True
- There is only one 16 year old at the party — Can't say
- The youngest kid is 7 years old — True
- Exactly half the kids are older than 13 — Can't say



Outlier detection – Excel and Box Plot Steps

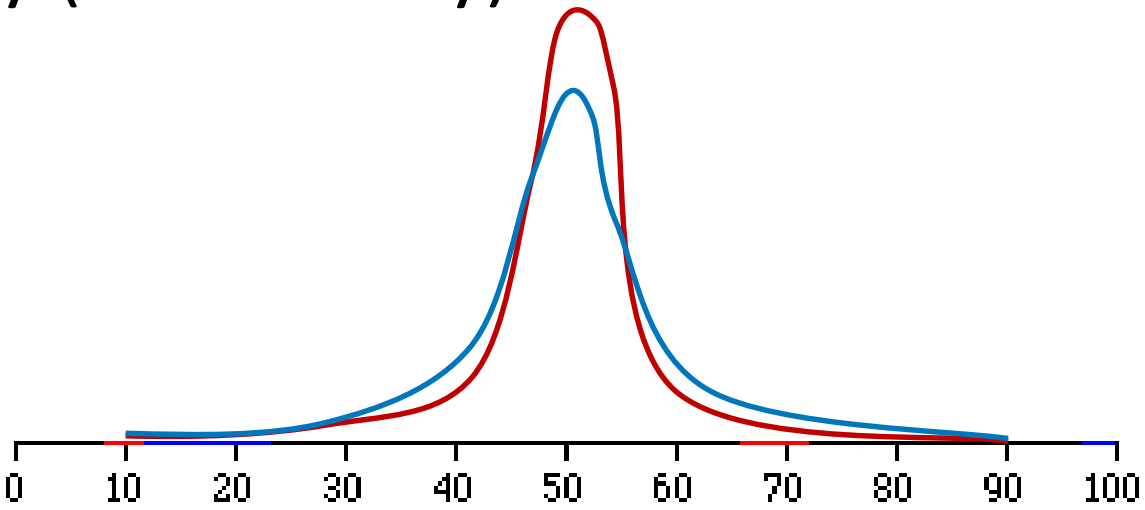
Hadlum vs Hadlum case

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

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

Measuring Variability and Spread

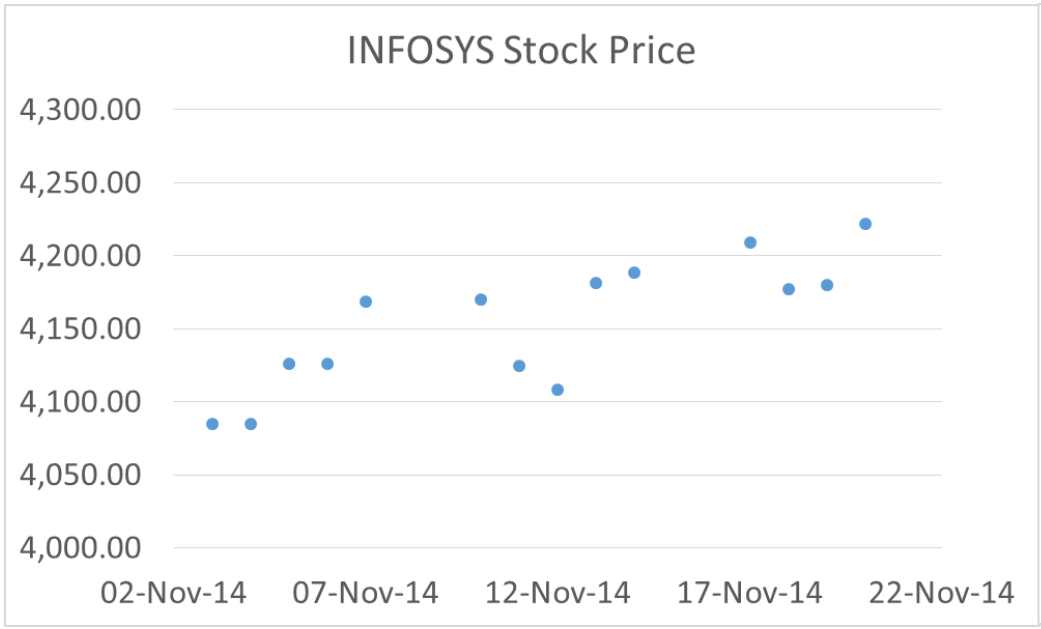
Range and IQR give the spread but still do not describe variability (consistency).

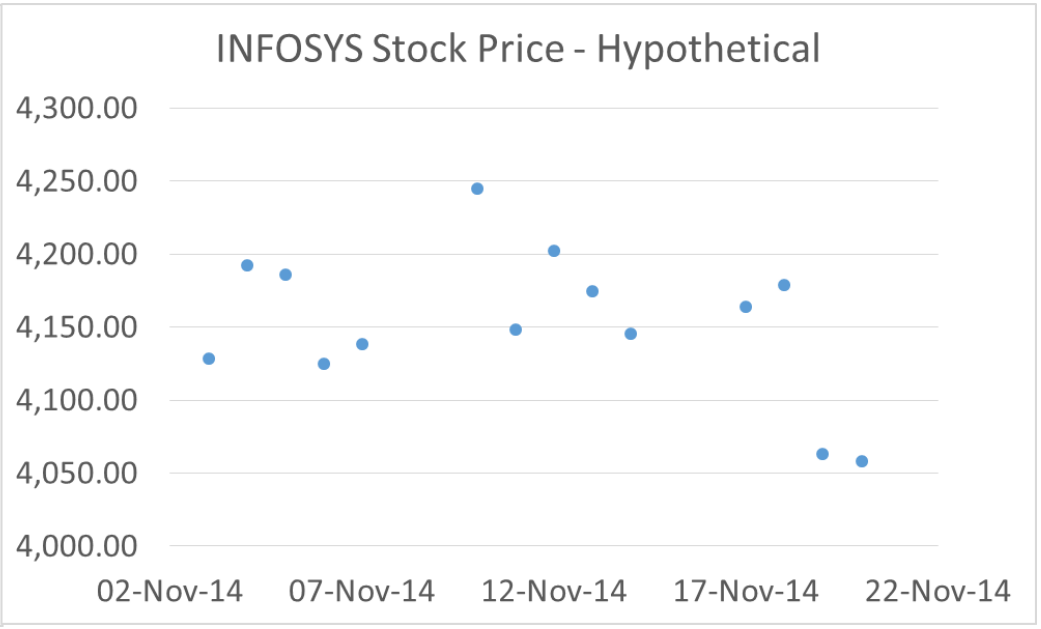


Average distance from the mean?

3 3 6 7 7 10 10 10 11 13 30

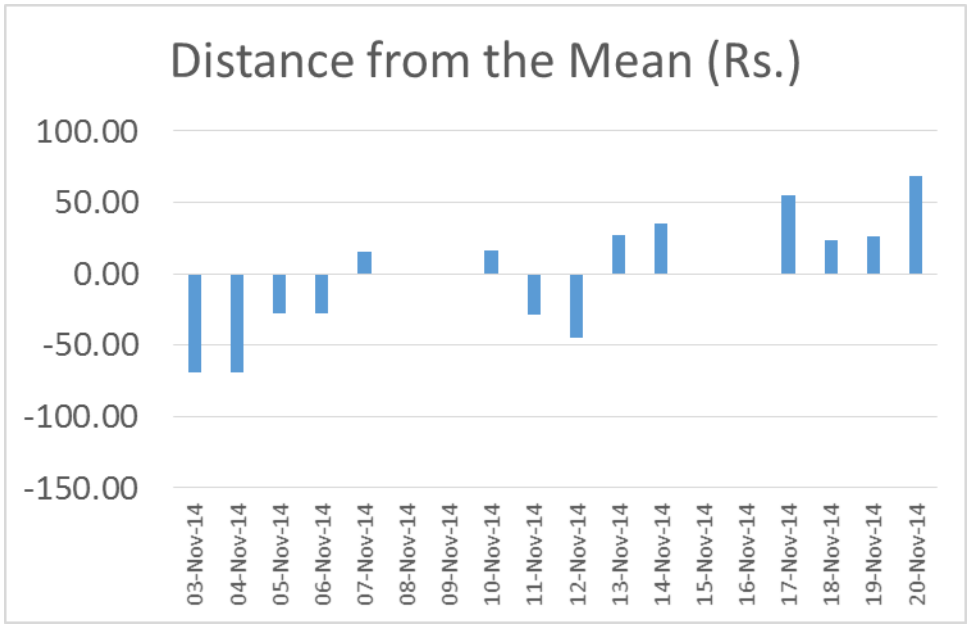
Measures of Spread – Mean Distance, Mean Absolute Deviation or Standard Deviation - Excel

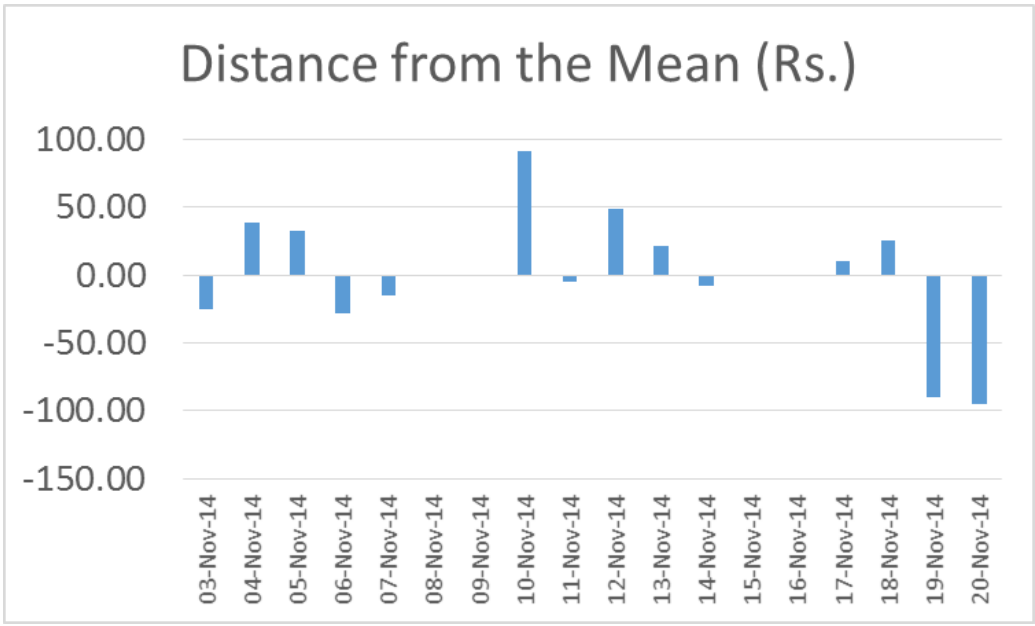




Data Source: https://in.finance.yahoo.com/q/hp?s=INFY.BO

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.

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

Measuring Variability and Spread

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

3 3 6 7 7 10 10 10 11 13 30

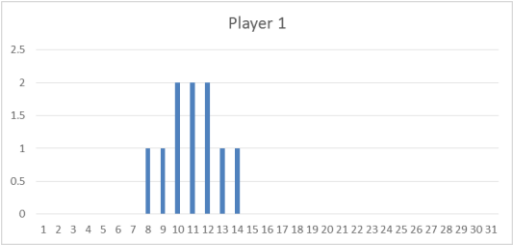
Units are squared, which is not intuitive.

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

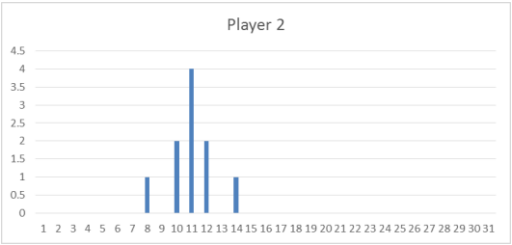
Measuring Variability and Spread

Calculate standard deviation for each player.

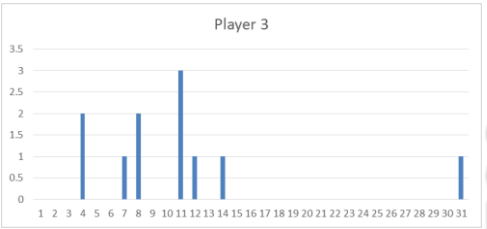
Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1



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.

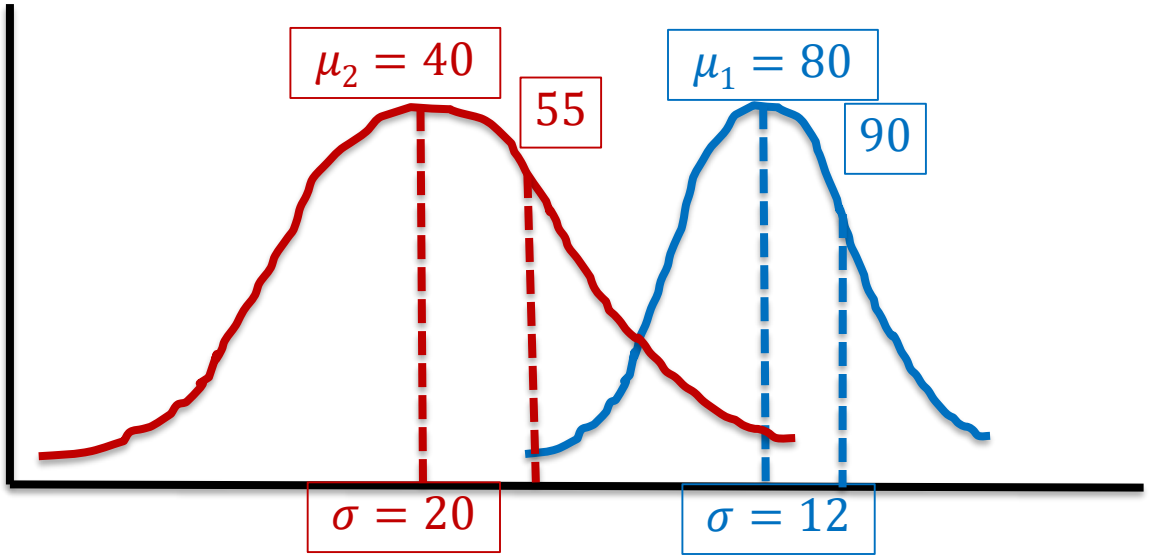
Measuring Variability and Spread

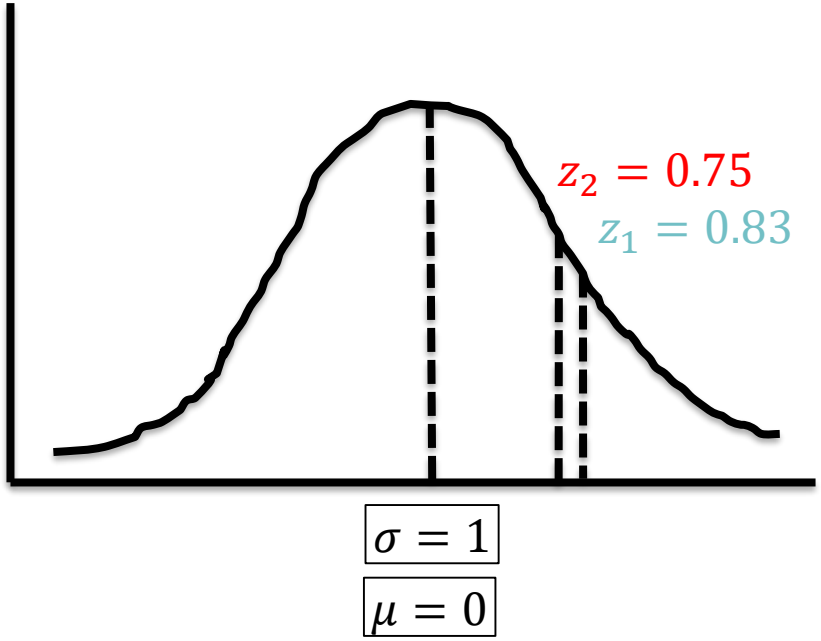
Imagine 2 players with different abilities: one has an average of 80% with 12% Stdev and the other 40% with 20% Stdev.

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

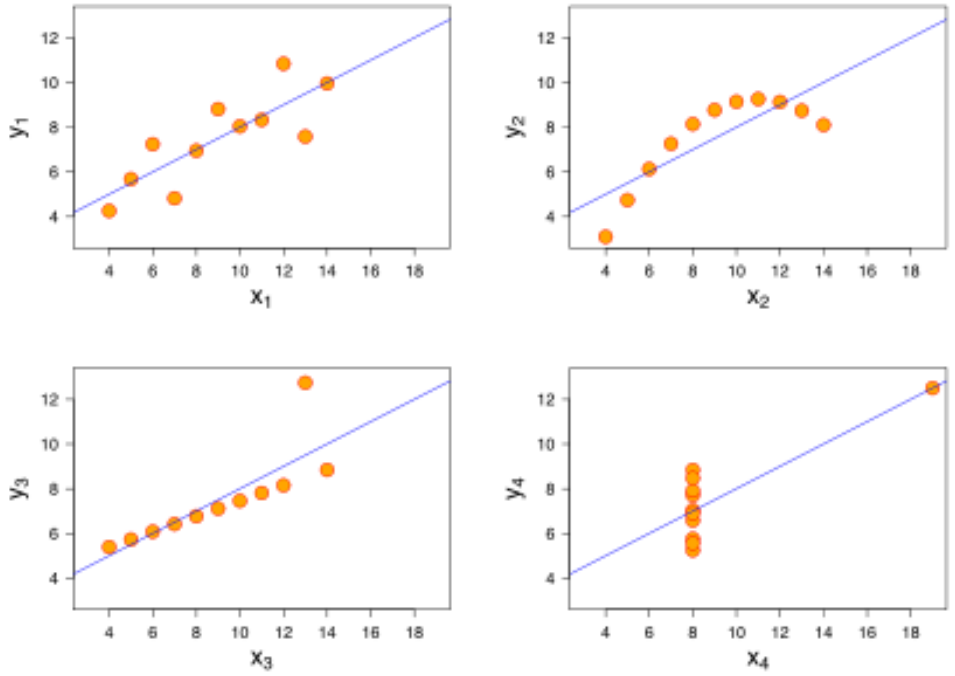




Measuring Variability and Spread

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)



CSE 7315c

Inspire...Educate...Transform.

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