Why Statistics and Data representation

Day 1

Do The Math

Chicago, IL
Bangalore, India
www.mu-sigma.com

2016

Content

- Why statistics? [Examples]
- What is Data?
- Building intuition about data – [Data visualization examples/ video]
  - Napoleon Map
  - Florence Nightingale Pie charts
  - Hans Rosling Video
- Data exploration through an example
- Knowing your data
  - Dangers of trusting numerical measures
  - Simpson's paradox

Why statistics? Why do you think it is important?

Who will you recommend? – Part I

- Mr. Shared Power and Mr. Rash Pilot are two workers contending for a promotion
- Each worker can operate six semi-automatic machines simultaneously
- They operated a group of six machines (M1 to M6) for six consecutive hours each. The table below gives the average number of units per hour produced by Mr. Shared Power and Mr. Rash Pilot on each machine
- The results were summarized by the assistant of Mr. Peevee Yen, the consultant from Yaponandon Consultants

Machine #	Avg Production Per Hour	
	Mr Power	Mr. Pilot
1	570.8	571.5
2	550.0	552.0
3	605.0	612.3
4	587.6	590.0
5	542.7	554.3
6	544.2	545.6

Who will you recommend? – Part II

- Mr. Shared Power and Mr. Rash Pilot are two workers contending for a promotion
- Each worker can operate six semi-automatic machines simultaneously
- They operated a group of six machines (M1 to M6) for six consecutive hours each. The table below gives the average number of units per hour produced by Mr. Shared Power and Mr. Rash Pilot on each machine
- The results were summarized by the assistant of Mr. Peevee Yen, the consultant from Yaponandon Consultants

Worker	Shared Power	Rash Pilot
Total Units	14,816	14,707

Facts!

- Both the tables in the previous two slides were created from the same data

	Shared Power						Rash Pilot					
Hour	1	2	3	4	5	6	1	2	3	4	5	6
M1	572	570	568	BD	574	570	572	BD	570	572	BD	572
M2	550	548	BD	552	BD	BD	550	554	554	BD	552	550
M3	606	598	600	612	BD	609	610	615	BD	BD	612	BD
M4	BD	588	586	588	586	590	BD	BD	588	592	BD	590
M5	543	BD	BD	BD	535	550	560	565	558	545	550	548
M6	548	546	545	540	BD	542	544	542	546	548	BD	548

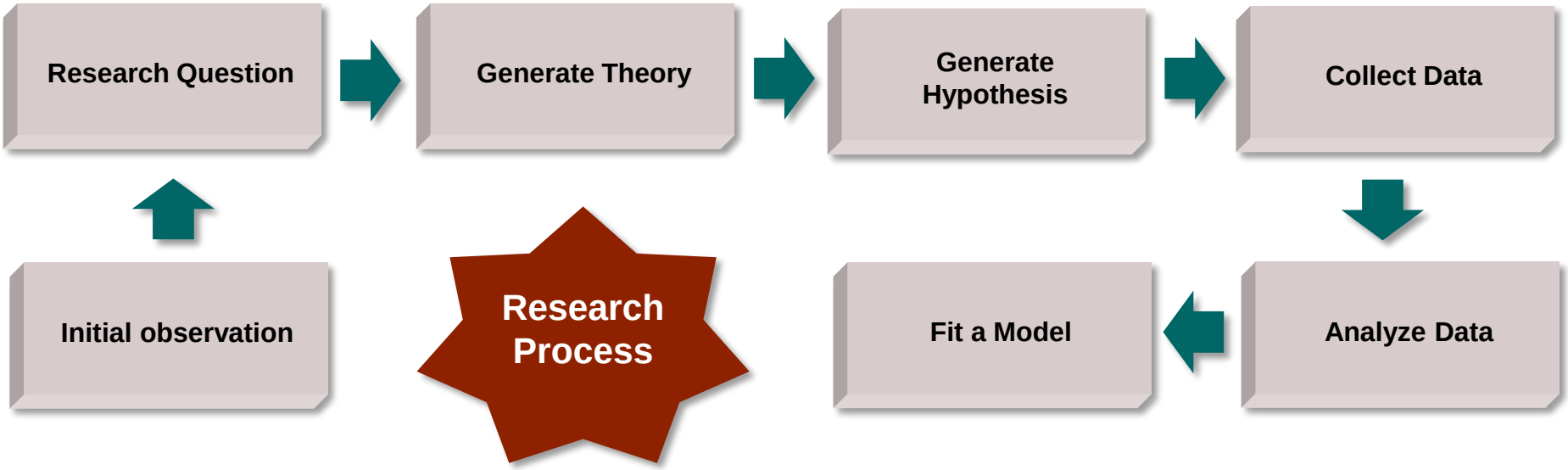
*BD indicates breakdown

- Data can be manipulated to make you decide either way
- Data has to be looked at in context of the question!

Statistics starts with question, not with data/information

- What is average age of Mu Sigman?
- % people coming before 9 AM
- # of plates for breakfast and dinner
- # of cabs required on Friday
- % of error free deliverables to client

The scientific method..



What is Data? Examples of types of data…

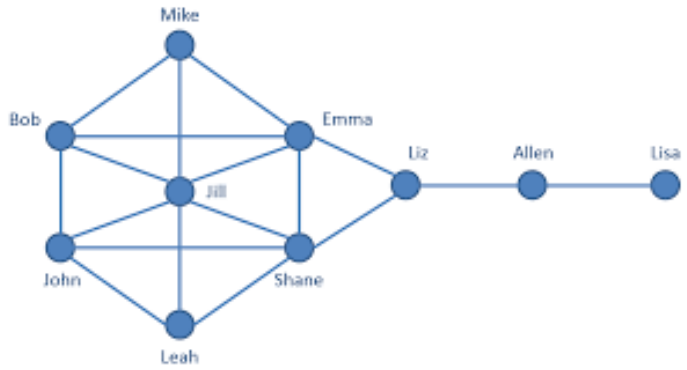
Structured data

Unstructured data

Tabular data

Customer	Avg. Sales	Age
A123	105.4	25
A345	110.2	35
A112	245.3	22
A890	67.9	48
A564	97.3	28
A887	99.1	26

Network data

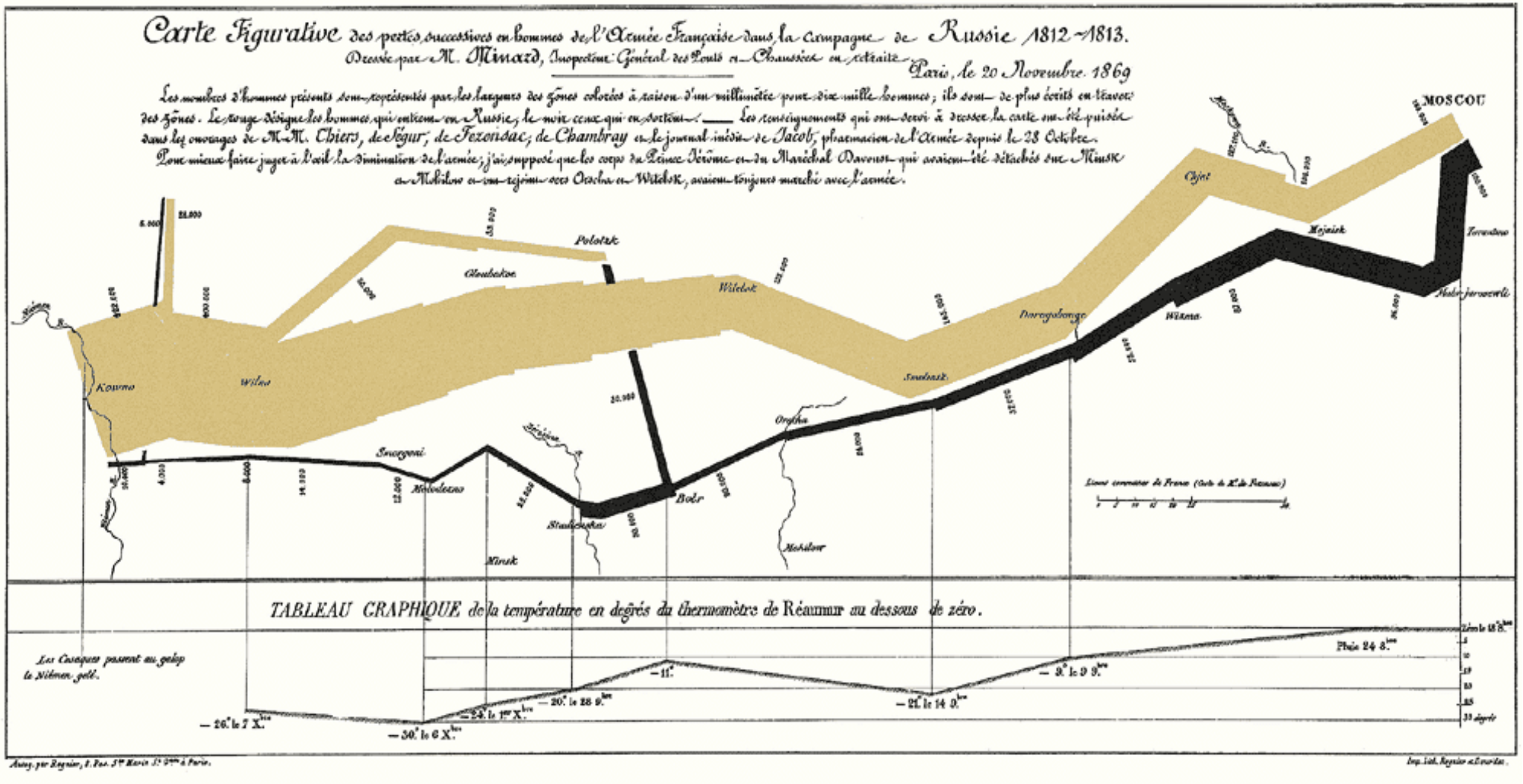


Textual data

The power of data visualization – Napoleon Map

- Dimensions of data
  - Advance and retreat of Groups 1 – 3 of the Napoleon army [Longitude]
  - Temperature that the troops experienced
  - Loss of life over time/ location
  - Map/ Path of the advanced and retreat [Geographically]

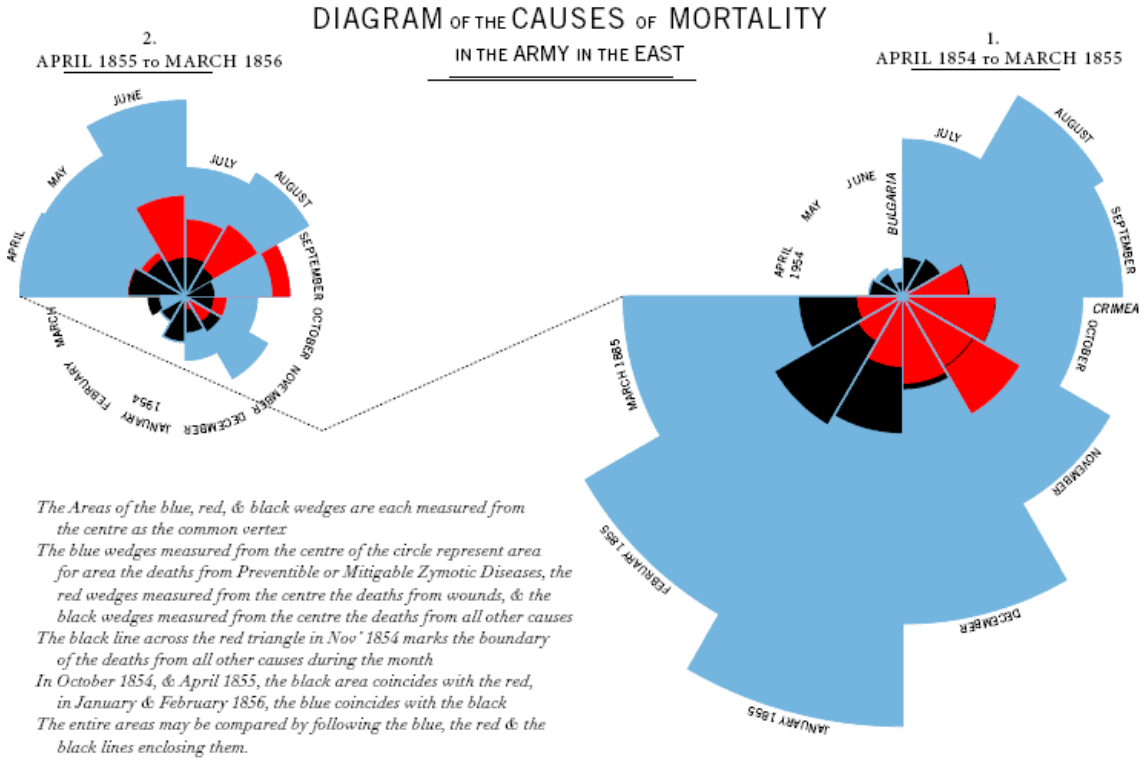
Minard's graphic



The power of data visualization – Nightingale pie charts

Making graphical representations brings home the message

Florence Nightingale



- There are many different ways to represent data
- One of the most imaginative ways was invented by Florence Nightingale (1820-1910) in 1857
- Nightingale Roses were used to represent the cause of death of British soldiers in each month during the Crimean war

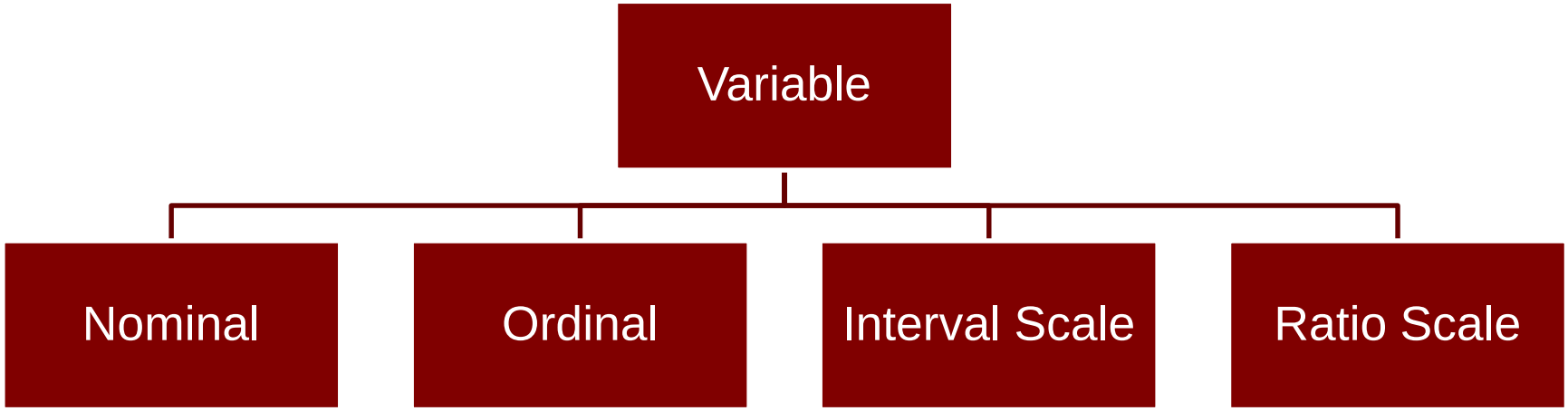
Data visualization – Hans Rosling

Knowing your purpose drives all other decisions in creating your table or graph

- You need to have a purpose statement for every table or graph you create and design the display to serve the purpose
- Ensure that the information is correct and that it is presented in a way that doesn't distort the truth.
- Avoid heavy grids, excess tick marks, redundant representation of simple data, shadows etc.
- Don't hide data…show it. And consider annotations to help explain what data means

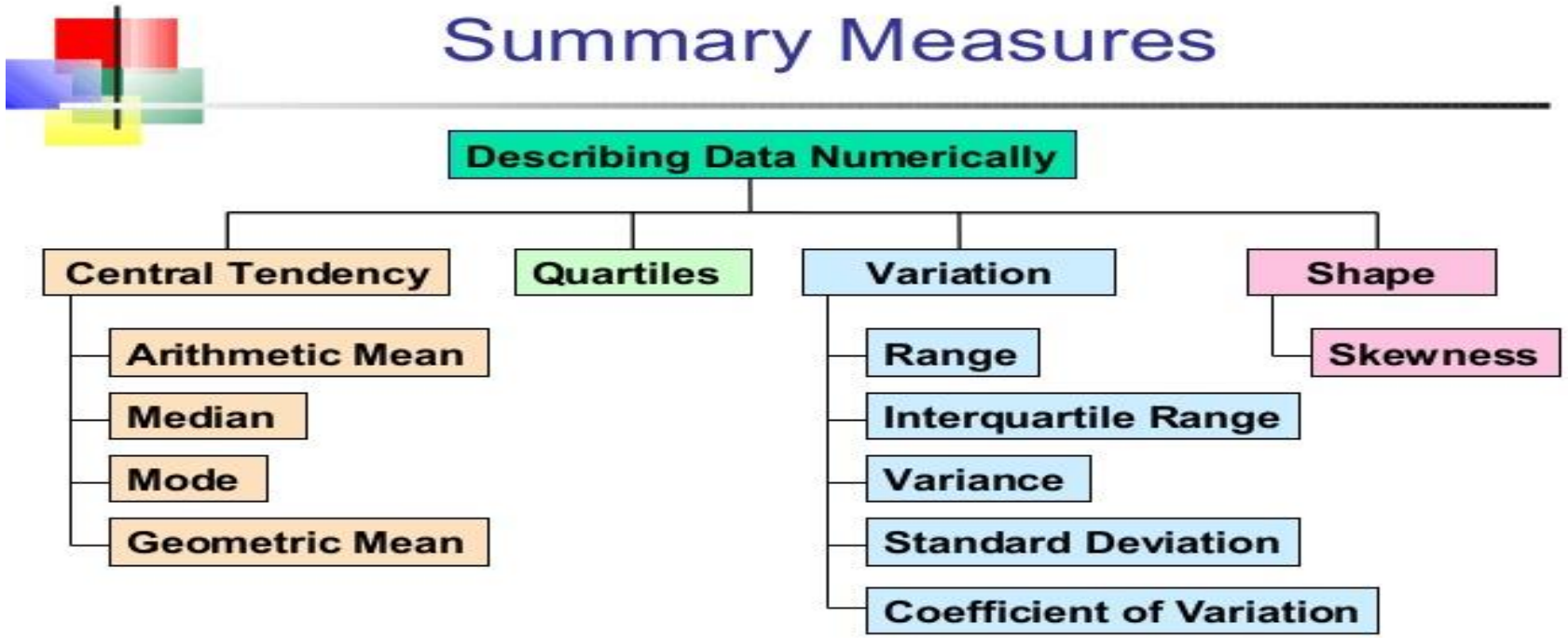
Variables and scales of measurement

Scales of Measurement



16

Summarizing data



Consider a dataset, detailing the salaries of the players of two MLB National League teams – the Washington Nationals and the New York Mets.

Player	Salary (in millions $)	Position	Team
Adam LaRoche	8	Baseman	Washington Nationals
Andres Torres	2.7	Outfielder	New York Mets
Brad Lidge	1	Pitcher	Washington Nationals
Brett Carroll	0.6	Outfielder	Washington Nationals
D.J. Carrasco	1.2	Pitcher	New York Mets
Daniel Murphy	0.5	Baseman	New York Mets
Chris Marrero	0.5	Baseman	Washington Nationals
Craig Stammen	0.5	Pitcher	Washington Nationals
D.J. Carrasco	1.2	Pitcher	New York Mets
Daniel Murphy	0.5	Baseman	New York Mets

The data depicts the salaries of players of Washington Nationals and New York Mets for the Year 2012 (Courtesy: USA Today). Data shown are first 10 points of the dataset.

Player name

Salary

Position

Team

Question 1: Which team allocated their payroll better? [in terms of final results]

Question 2: What is the right payroll allocation strategy to perform better in the league?

How do we get an initial idea about the salaries of the players?

Player	Salary (in millions $)	Position	Team
Adam LaRoche	8	Baseman	Washington Nationals
Andres Torres	2.7	Outfielder	New York Mets
Brad Lidge	1	Pitcher	Washington Nationals
Brett Carroll	0.6	Outfielder	Washington Nationals
D.J. Carrasco	1.2	Pitcher	New York Mets
Daniel Murphy	0.5	Baseman	New York Mets
Chris Marrero	0.5	Baseman	Washington Nationals
Craig Stammen	0.5	Pitcher	Washington Nationals
D.J. Carrasco	1.2	Pitcher	New York Mets
Daniel Murphy	0.5	Baseman	New York Mets

Average player salary

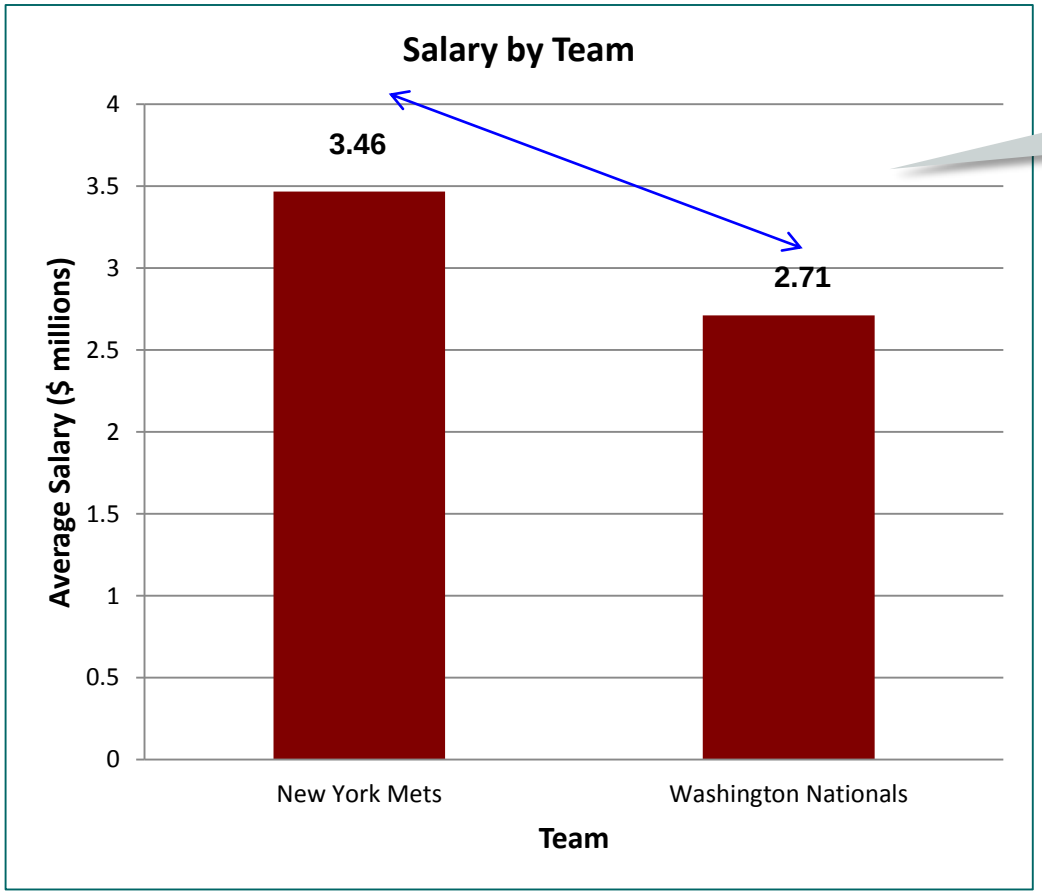
$ 3.06 million

?

Does this help us completely understand the salaries of the players?

We need to look understand the salary spread better in each of the teams to get a more accurate picture

What does the plot of average salary across teams tell us?

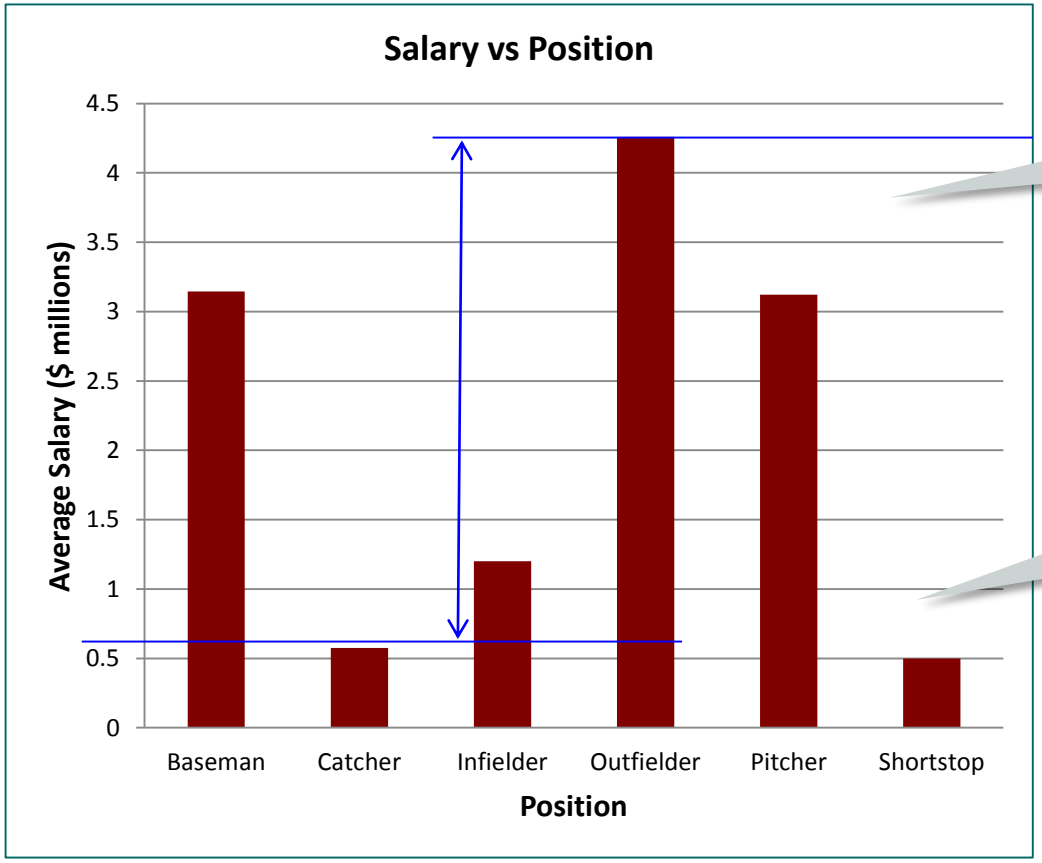


The Mets have a much higher average salary as compared to the Nationals.

This makes sense as the Mets were a much bigger and better known MLB franchise at that point

Aggregations like mean, sum, etc. can summarize across information variables

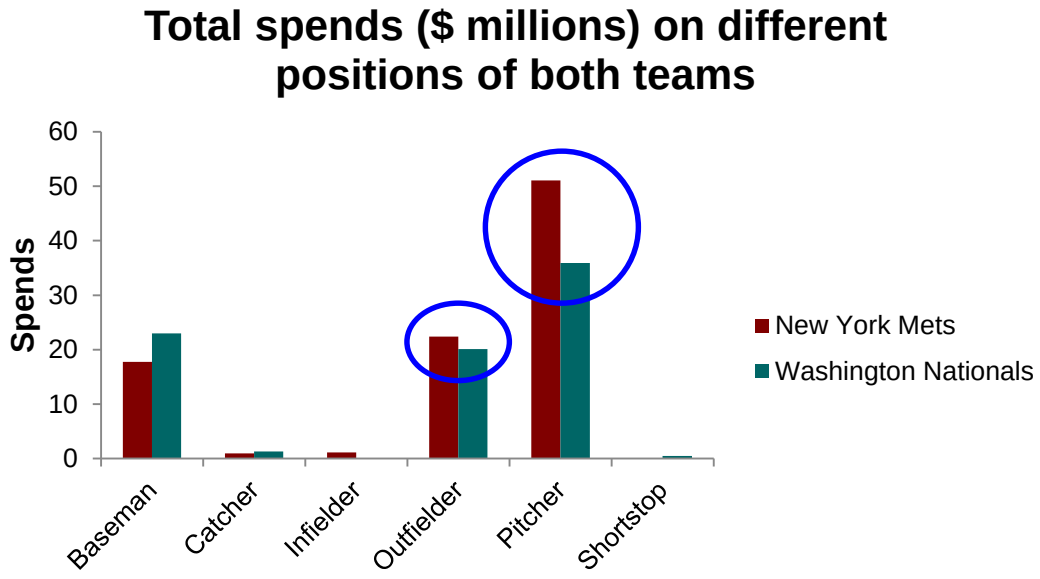
How does the average salary vary across different positions?

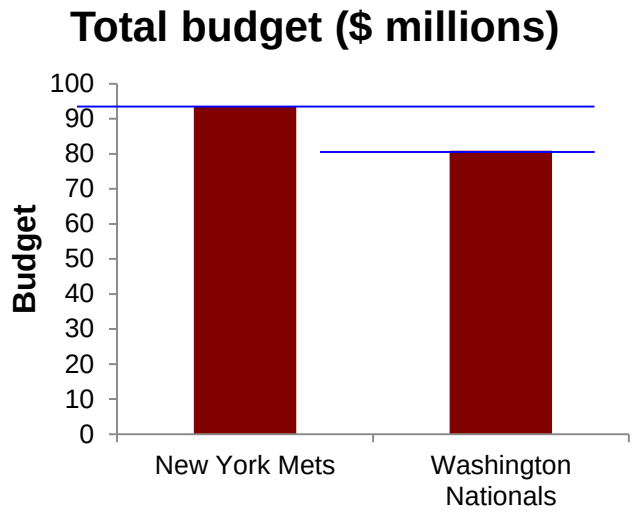


We see that on average, outfielders earn the most

Pitchers and basemen come in second in terms of average salary

Lets look at spends of both the teams across positions



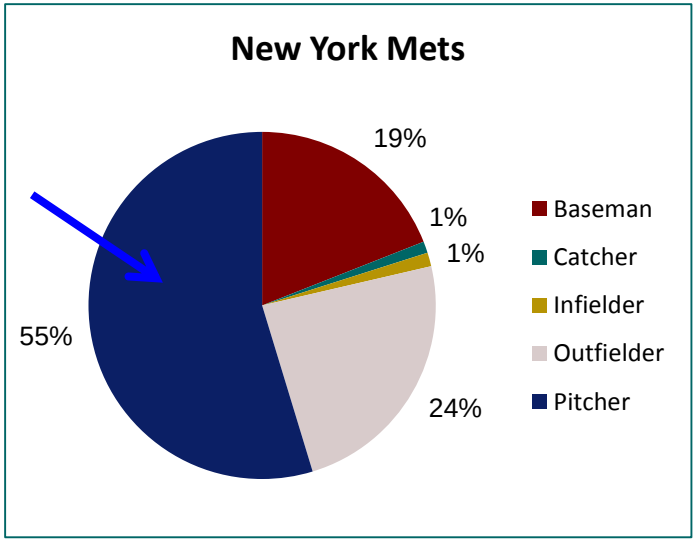


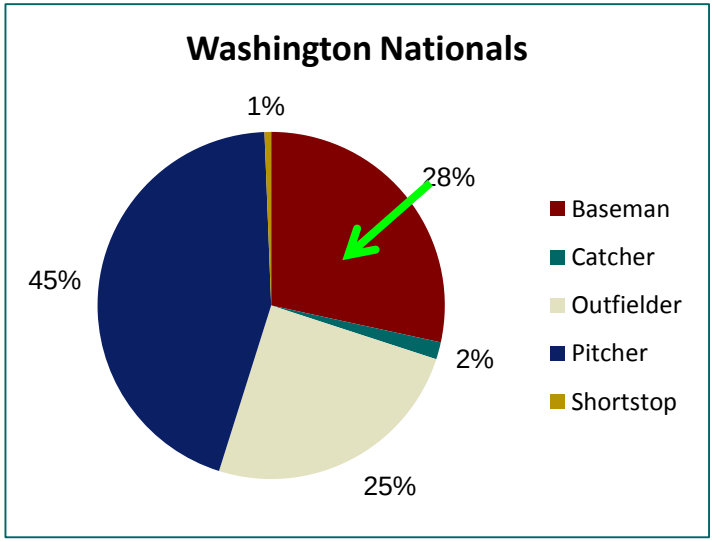
The Mets spend much more on pitchers and outfielders as compared to the Nationals

This might be misleading as the teams might have different budgets

The Mets have a much higher total budget than the Nationals (Difference of almost $ 13 million)

Let's look at the allocation of the payroll budget by both teams





Mets spend a higher proportion on pitchers while the Nationals spend a higher proportion on basemen

Courtesy: USA Today

Cross sections can generate key insights from data in a simple and effective manner

Looking at the results…

NL East	W	L	PCT	GB	HOME	ROAD
Washington Nationals	98	64	.605	—	50–31	48–33
Atlanta Braves	94	68	.580	4	48–33	46–35
Philadelphia Phillies	81	81	.500	17	40–41	41–40
New York Mets	74	88	.457	24	36–45	38–43
Miami Marlins	69	93	.426	29	38–43	31–50

Looking at the 2012 season results (from secondary research) we see that focus on a better infield defence is perhaps a winning strategy

The Nationals have allocated their resources more efficiently, as they ended up winning the league that year

Trend over time saw a huge dip in Mets spending and increase in Nationals spending. Today, the Nationals have a much higher budget/ spend as compared to the Mets

In 1936 US presidential election, *The Literary Digest* predicted that Landon will win with 57.1% of the votes against Roosevelt

- Roosevelt won the election with 60.8% of the popular vote
- Landslide victory with wins in 46 out of 48 states, losing only to Maine and Vermont
- Approximately 10 million questionnaires were mailed to prospective voters, making the *Literary Digest* poll one of the largest ever conducted
- Approximately 2.3 million responded
- Prospective voters were chosen from the subscription list of the magazine, from automobile registration lists, phone lists, club memberships, etc.
- Using similar technique, *The Literary Digest* had predicted correctly in the last four elections

Take a look at this study to evaluate treatments for kidney stones…

Size / Treatment	Treatment A	Treatment B
Overall result	78% (273/350)	**83% (289/350)**

Simpson's paradox

Size / Treatment	Treatment A	Treatment B
Small Stones	Group 1 **93% (81/87)**	Group 2 87% (234/270)
Large Stones	Group 3 **73% (192/263)**	Group 4 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

- The lurking variable here is the severity of the kidney stone which influences the doctor's decision to choose a treatment
- This is an example of how data can lead to the wrong causal conclusions

Why should we not take a naïve view of numbers?

- You can be fooled by numbers
  - Since 1980's researchers have described numerous statistical fallacies and misconceptions in peer reviewed scientific literature
  - Misinterpreted p-values have caused numerous false positives
- You need to learn how to communicate the numbers sincerely
  - Poor statistical education have led scientists to conclude that most published research findings are probably false
  - Errors have massive impact on the real world
  - Clinical trials determine the safety of prescription drugs
  - Marketers and business managers find best ways to sell products
- We have embraced all the statistical tools available to us – however we have failed to embrace the statistical education that is required

Appendix

Wrong turns on red – perils of incorrect inference

Virginia Dept. of Highways and Transportation studied accidents at 20 intersections

Accidents Before Change	Accidents After Change
308	**337**

- In 1973, many cities in US started allowing drivers to **turn right at a red light**
- Several studies were conducted to consider the safety impact of the change
- The studies were not **statistically significant** leading to the conclusion that there was no impact on safety of pedestrians and bicyclists
- These studies were later found to be severely underpowered – a blunder which led to loss of lives due to accidents at accidents

Sally Clark – The case of inexpert witness

- Solicitor Sally Clark was tried in 1999 for the murder of two children (Christopher, 11 weeks & Harry, 8 weeks)
- Clark's first son died suddenly few weeks of his birth in Sep, 1996
- Harry died in similar manner in Dec, 1998
- The prosecution case relied on the statistical evidence presented by paediatrician Professor Sir Roy Meadow, who testified that the chance of two children from an affluent family suffering **sudden infant death** syndrome was 1 in 73 million

Clark was released in 2003 but dies of alcohol poisoning in 2007

SALFORD ADVERTISER THURSDAY JUNE 13, 2002

SALFORD UNI MAN SAYS SALLY CLARK CONVICTION MAY BE WRONG

Maths professor challenges double baby murder case

A SALFORD University Maths professor will challenge evidence used to convict a solicitor of murdering her two baby sons at a conference on cot-deaths next week.

Prof Ray Hill, from Eccles, head of the university's Applied and Discrete Mathematics Research Unit said statistical evidence used to convict Sally Clark, from Wilmslow, in October 2000, was not only quoted out of context and unfairly used to imply guilt, but was actually wrong.

Watching the trial on the TV he became furious and told us: "I shouted at the screen 'that figure's wrong!' They took an estimated figure for the likelihood of one cot death and then just squared it to get this one-in-73 million chance. That's not allowed unless you're sure the events are independent. A bookie wouldn't give you those odds."

He has now studied the Confidential Enquiry into Stillbirths and Deaths in Infancy (CESDI) report, which gives detailed figures on the number of deaths from 1993-1996.

He said: "It seems the chances of two cot deaths in the same family are much higher than the prosecution led the jury to believe."

Prof Hill has written to several national newspapers and is working with Sally Clark's defence team on the campaign to free her.

He will present his full criticism of the evidence at a Developmental Physiology Conference on cot deaths organised by Leicester University on June 28.

The Criminal Cases Review Commission has been looking at the case and is expected to report within the next few weeks. With their report imminent, Sally Clark's defence team and family do not feel it is appropriate to comment.

For more information on the Sally Clark campaign visit www.sallyclark.org.uk

Evidence challenge: Prof Ray Hill (2953-S 02)

Why should we not take a naïve view of numbers?

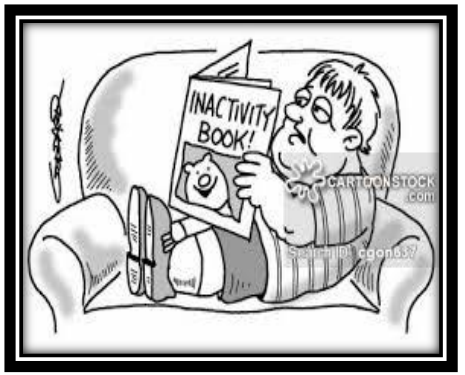
- You can be fooled by numbers
  - Since 1980's researchers have described numerous statistical fallacies and misconceptions in peer reviewed scientific literature
  - Misinterpreted p-values have caused numerous false positives
- You need to learn how to communicate the numbers sincerely
  - Poor statistical education have led scientists to conclude that most published research findings are probably false
  - Errors have massive impact on the real world
  - Clinical trials determine the safety of prescription drugs
  - Marketers and business managers find best ways to sell products
- We have embraced all the statistical tools available to us – however we have failed to embrace the statistical education that is required

Cause and Effect

Let us take an example:-

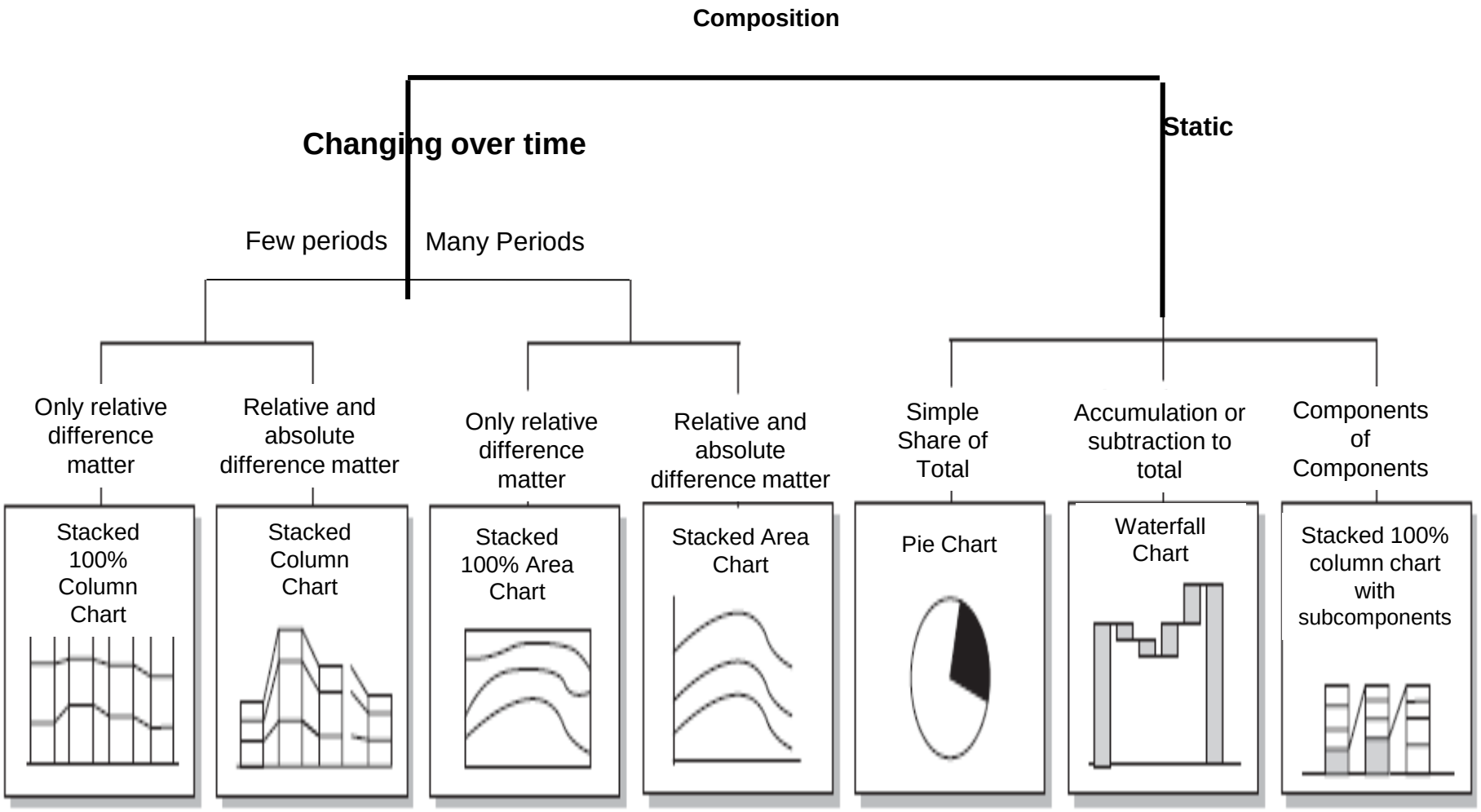
Ram is worried about his body weight. He thought for a while and found out it might be due to the three major reasons:-

1. Work timings
2. Diet
3. Lack of exercise



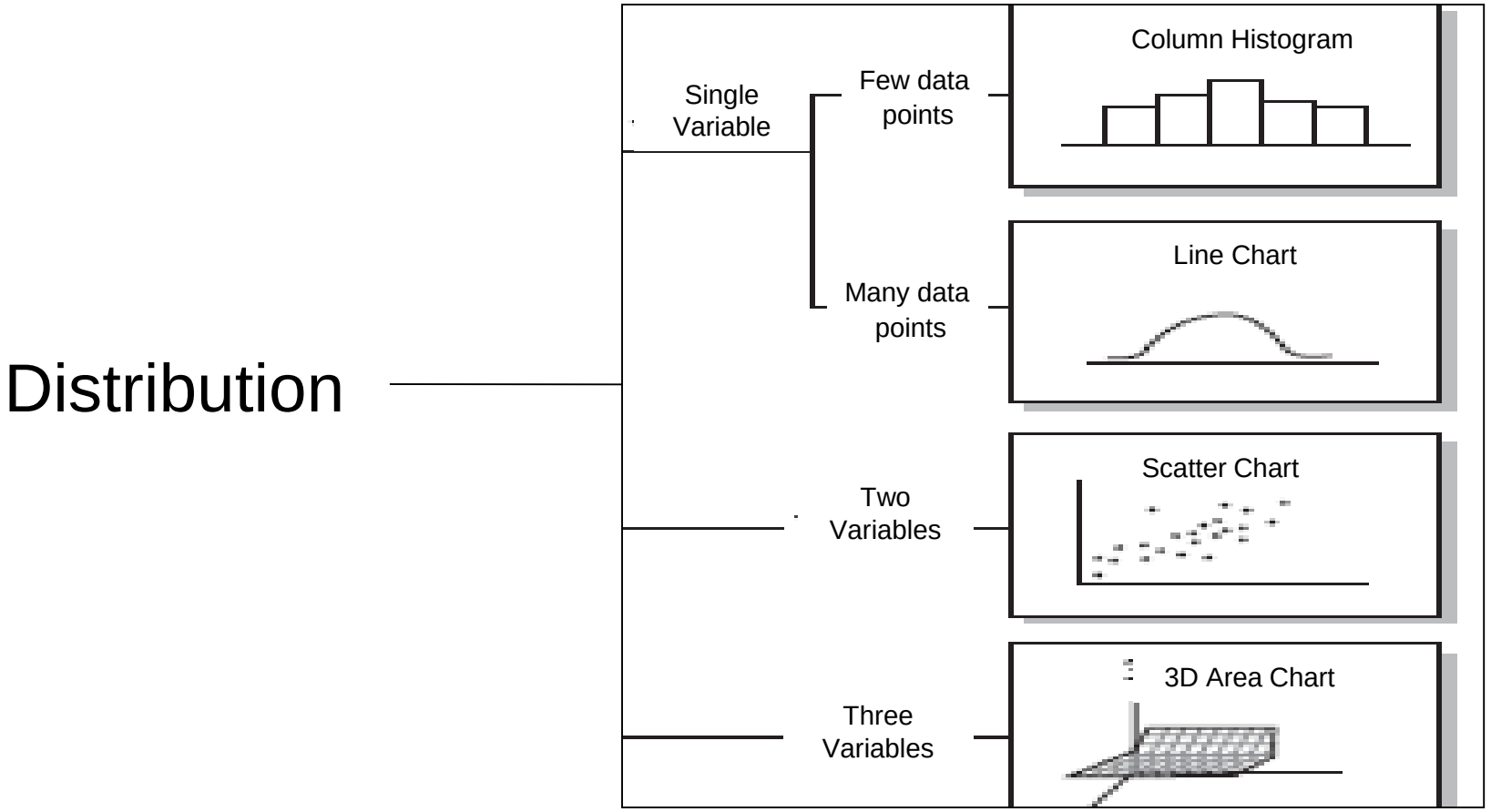
In the above examples, work timings, diet and lack of exercise acts as the independent variables effecting the weight of Ram which is a dependent variable

Choosing the right picture - Composition



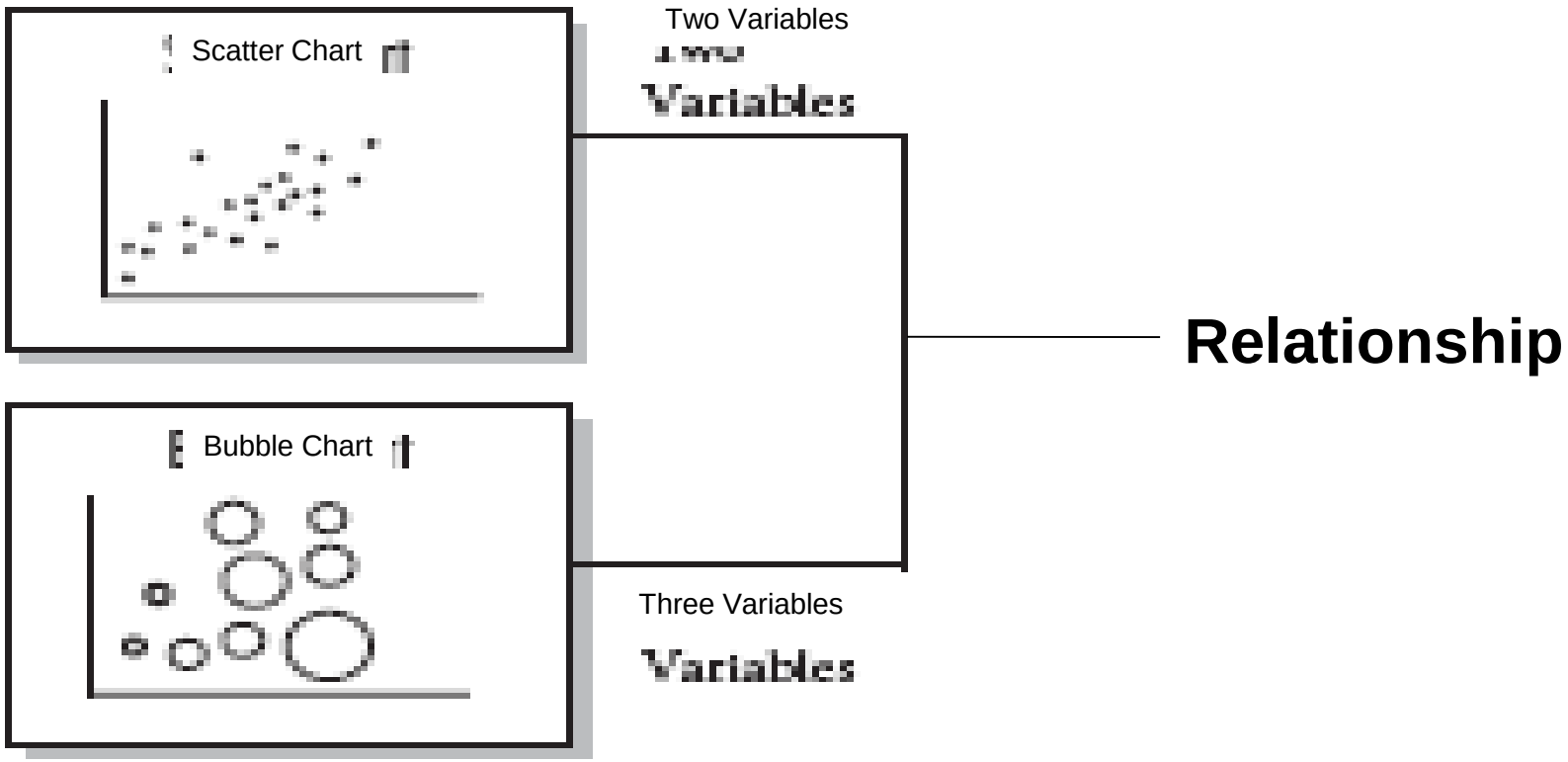
35

Choosing the right picture - Distribution



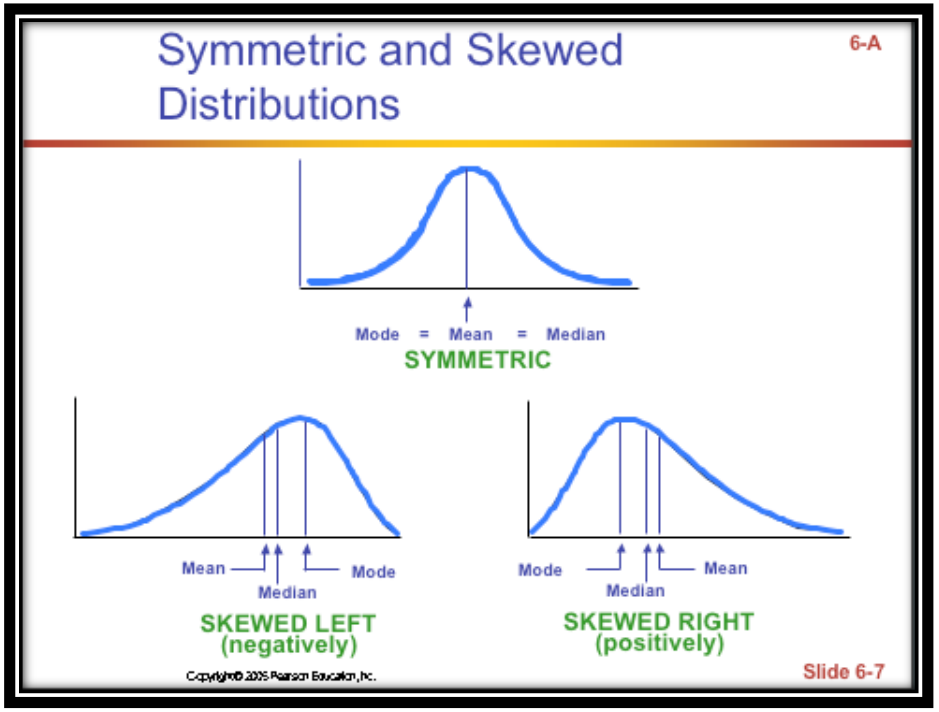
36

Choosing the right picture – Relationship



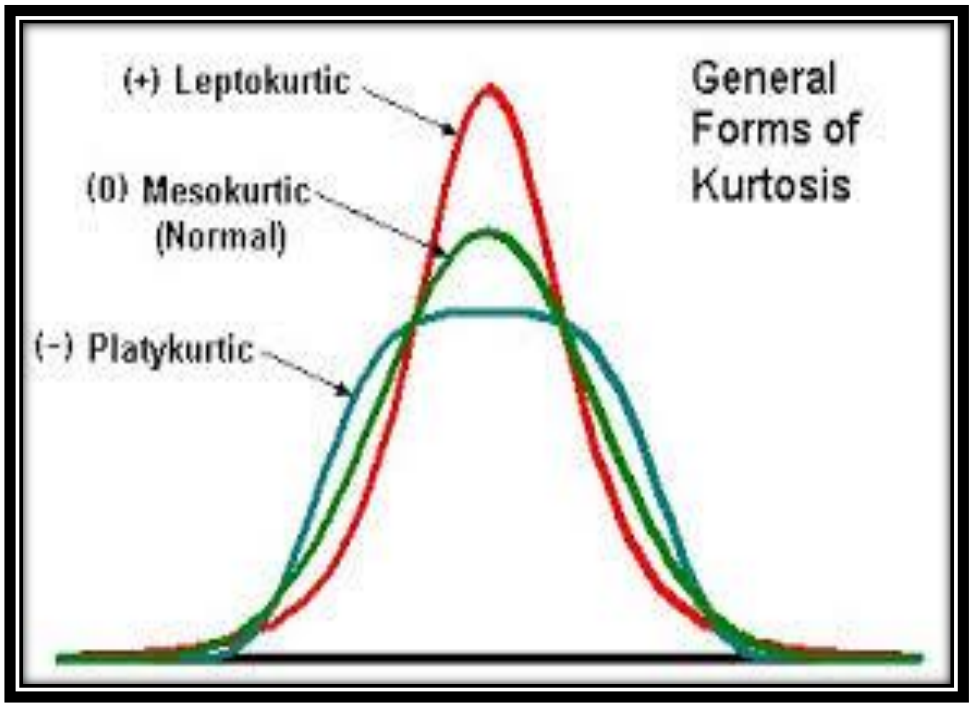
37

Shapes (Skewness and Kurtosis)



Skewness

Kurtosis



Variables – Part I

Example #1: Can blueberries slow down aging?

A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month old rats (equivalent to 60-year old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor tests. Although all supplemented rats showed improvement, those supplemented with blue berry powder showed the most notable improvement.

1. What are the independent variables?
2. What is the dependent variable?

39

Variables – Part II

Example #2: How bright is right?

An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

1. What are the independent variables?
2. What is the dependent variable?