PRIME INTUIT
Finishing School

Measure of Centrality and Spread

PRIME INTUIT
Finishing School

Introduction to Descriptive statistics

Descriptive Statistics

- ✓ Different types of data
- ✓ Different types of plots
- ✓ Measure of centrality and Spread

Probability Theory

- ✓ Sample Specs, events, axioms
- ✓ Discrete and continuous RVs
- ✓ Bernoulli, Uniform, Normal dist
- ✓ Sampling strategies

Inferential Statistics

- ✓ Interval Estimators
- ✓ Hypothesis testing (z-test, t-test)
- ✓ ANOVA, Chi-square test
- ✓ Linear Regression

PRIME INTUIT
Finishing School

What questions are we trying to answer

Why do we need measures of spread and centrality?

What are the different measures of centrality?

What are some characteristics of these measures?

What are the measures of centrality look like for different types of distribution?

How do you compute these measures from histograms?

What is the effect of certain transformations on these measures?

PRIME INTUIT
Finishing School

Why do we need measures of spread and centrality

Age	Height	Weight	Cholesterol	Sugar Level	LDL	…….
34	178	82	128	108	88	…….
26	163	76	122	150	130	…….
46	146	69	116	123	102	…….
32	158	60	114	110	98	…….
29	170	85	112	96	87	…….
…….	…….	…….	…….	…….	…….	…….

Imagine millions of such records

Drawing plots can give a good visual summary

In some cases we want an even more succinct summary (based on a few records)

Why do we need measures of spread and centrality – (Recap on Statistics)

What is a population?

What is a Sample?

What is a parameter?

A Parameter is the numeric property of the entire population under study

What is statistics?

A Statistic is any numerical property of a sample for a parameter (Used as an estimate for the corresponding parameter of the population)

PRIME INTUIT
Finishing School

Why do we need measures of spread and centrality – Summary Statistics

Measures of Centrality (mean, mode, median)

Percentiles (quartiles, quintiles, deciles)

Measures of Spread (range, IQR, variance, standard deviation)

PRIME INTUIT
Finishing School

What are the different measures of centrality?

What is the typical value of an attribute in our dataset?

Match #	Runs	Mins	SR	BF	4s	6s	Pos	Dismissal	Oppn	Date	Match ID
0	0	0	0.00	2	0	0	5	Caught	Pakistan	18-12-89	ODI # 593
1	0	2	0.00	2	0	0	5	Caught	New Zealand	1-3-90	ODI # 612
2	36	51	92.3	39	5	0	6	Caught	New Zealand	6-3-90	ODI # 616
3	10	15	63.33	12	0	0	5	Caught	Sri Lanka	25-4-90	ODI # 623
4	20	31	60.00	25	1		7	Run Out	Pakistan	27-4-90	ODI # 625
5	19	38	54.28	35	1	0	4	Bowled	England	18-7-90	ODI # 634
6	31	31	119.23	26	3	1	6	Bowled	England	20-7-90	ODI # 635
7	53	83	129.26	41	7	2	5	Bowled	Sri Lanka	1-12-90	ODI # 646
…	…	…	…	…..	…..	…..	…	…	…	…	…..

How many runs does Sachin Tendulkar typically score in a match?

How many balls would Sachin Tendulkar typically face in a match?

PRIME INTUIT
Finishing School

Measures of centrality (Mean)

Match #	Runs	Mins	SR	BF	4s	6s	Pos	Dismissal	Oppn	Date	Match ID
0	0	0						Caught	Pakistan	18-12-89	ODI # 593
1	0	2						Caught	New Zealand	1-3-90	ODI # 612
2	36	51						Caught	New Zealand	6-3-90	ODI # 616
3	10	15						Caught	Sri Lanka	25-4-90	ODI # 623
4	20	31									ODI # 625
5	19	38									ODI # 634
6	31	31	119.23	26	3	1					ODI # 635
7	53	83	129.26	41	7	2					ODI # 646
…	…	…	…	…..	…..	…..					…..

X1 = 0, x2 =0, x3 = 36, x4 = 10, x5 = 20, x6 = 19, x7 = 31, x8 = 36, x9 = 53, x10 = 30, x11 = 0, ……………x452 = 52

Mean of a sample = $\bar{X}$
Mean of a population = μ

Notation: n data points x1, x2, x3, x4, x5, x6, …………xn

Measures of centrality (Mean)

Mean

X1 = 0, x2 =0, x3 = 36, x4 = 10, x5 = 20, x6 = 19, x7 = 31, x8 = 36, x9 = 53, x10 = 30, x11 = 0, ...x452 = 52

$\bar{X} = \frac{x1+x2+x3+\ldots\ldots\ldots xn}{n}$

$\bar{x}$=(x1 +x2+x3.....xn)/ n

$\bar{x}$ = (0+0+36+10+20.....52)/452

= 40.76

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{+n} x_i$$

PRIME INTUIT
Finishing School

Measures of centrality (Median)

Median

Shikar Dhawan T20I scores (59 scores)

5,32,30,0,1,33,3,11,5,42,26, 9,51,46,2,1,16,60,1,6,23,13, 23,15,2,80,1,6,72,24,47,90, 55,8,35,10,74,4,10,5,3,43,9 2,76,41,29,30,5,14,1,23,3,4 0,36,41,31,19,32,52

Sorted

0,1,1,1,1,1,2,2,3,3,3,4,5,5,5,5,6,6, 8,9,10,10,11,13,14,15,16,19,23,23 ,23,24,26,29,30,30,31,32,32,33,35 ,36,40,41,41,42,43,46,47,51,52,55 ,60,72,74,76,80,90,92

Median is the value which appears at the center of the data when the data is sorted

N = 59 is odd

Center location = (n+1)/2

N=(59+1)/2 = 30

Prime Intuit Finishing School

Measures of centrality (Median)

First 29 elements	Mid point	Last 29 elements
0,1,1,1,1,1,2,2,3,3,3,4,5,5,5, 5,6,6,8,9,10,10,11,13,14,15, 16,19,23	23	23,24,26,29,30,30,31,32,32, 33,35,36,40,41,41,42,43,46, 47,51,52,55,60,72,74,76,80, 90,92

First 29 elements

Mid point
30^{th} element

Last 29 elements

There are equal number of elements on either side of the central location

When n is odd, the median is the value at the central location (or mid-point) which is 23 in this case

Measures of centrality (Median)

What happens when n is even? (Say we had data for 50 TOIs only)

0,1,1,1,1,1,2,2,3,3,4,5,5,5,5,6,6,8,
9,10,10,11,13,14,15,16,19,23,23,2
4,26,29,30,30,32,33,35,41,42,43,4
6,47,51,55,60,72,74,76,80,90,92

Measures of centrality (Median)

0,1,1,1,1,1,2,2,3,3,4,5,5,5,5,
6,6,8,9,10,10,11,13,14

15

16

15.5
median

23,24,26,29,30,30,32,33,35,
41,42,43,46,47,51,55,60,72,
74,76,80,90,92

First 24 elements

2 Mid points
30th element

Last 24 elements

There are 2 midpoints now such that the number of elements on either side is the same

When n is even, the median is the average of the value at the two central locations (or mid-points) which is (15+16)/2 = 15.5 in this case

Finishing School

Measures of centrality (Median) Summary

Data: x1, x2, x3,x4,..............xn

If n is odd: median = $x_{\frac{n+1}{2}}$

If n is even: median = $\dfrac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$

PRIME INTUIT
Finishing School

Measures of centrality (Mode)

Shikar Dhawan T20I scores (59 sorted scores)

0,1,1,1,1,1,2,2,3,3,3,4,5,5,5,5,6,6,8,9,10,10,11,13,14,15,16,19,23,23,23,24,26,29,30,30,31,32,32,33,35,36,40,41,41,42,43,46,47,51,52,55,60,72,74,76,80,90,92

Mode is defined as the element that occurs most frequently in the data set

Mode = 1

1,2,2,2,3,4,5,5,5,5,5,6,6,7,7,12,12,13,14,15,15,15,15,15,17,18,19,19

Mode = 5, 15 (bi-model data)

Multiple Modes: (more than 1 most frequent values)

1,2,3,4,6,8,9,23,24,56,78,76,54,61

No modes: (all values appear exactly once)

PRIME INTUIT
Finishing School

Measures of centrality (Summary)

Mean is the sum of all the elements in the data divided by the total number of elements

Median is the value which appears at the centre of the data when the dat is sorted (slight difference when n is odd vs when n is even)

Mode is the most frequent value appearing in the)

PRIME INTUIT
Finishing School

Characteristics of Measures of centrality

Mean is the center of gravity of the data

Data: x1, x2, x3, x4, x5 . . . . .xn

Mean = $\bar{x}$

"The deviation of a point from the mean is defined as the difference between this point and the mean"

Deviation: $x_{\mathrm{i}} - \bar{x}$

"Sum of the deviations of all points from the mean is Zero"

Characteristics of Measures of centrality

Mean is the **center of gravity of the data**

$$\text{sum of deviations} = \sum_{i=1}^{n}(x_i - \bar{x})$$

$$= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x})$$

$$= \begin{array}{l} (x_1 + x_2 + x_3 + \cdots + x_n) \\ -(\bar{x} + \bar{x} + \ldots n\ times) \end{array}$$

$$= \sum_{i=1}^{n} x_i - n\bar{x}$$

$$= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

"Sum of the deviations of all points from the mean is Zero"

PRIME INTUIT
Finishing School

Characteristics of Measures of centrality

Sum of deviations from the mean = 0

Number of lines as seesaw

Data points as weights on the seesaw

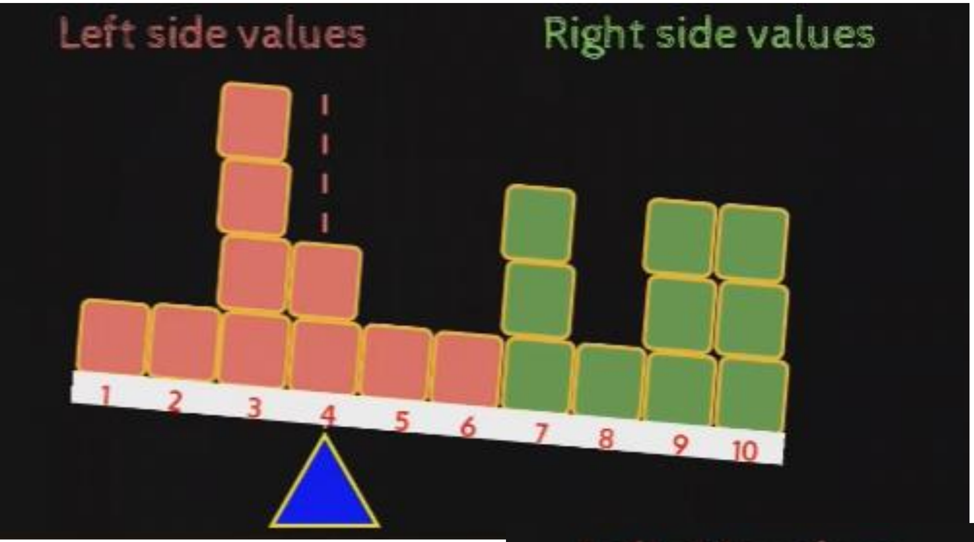
Weight is proportion to deviation from Mean

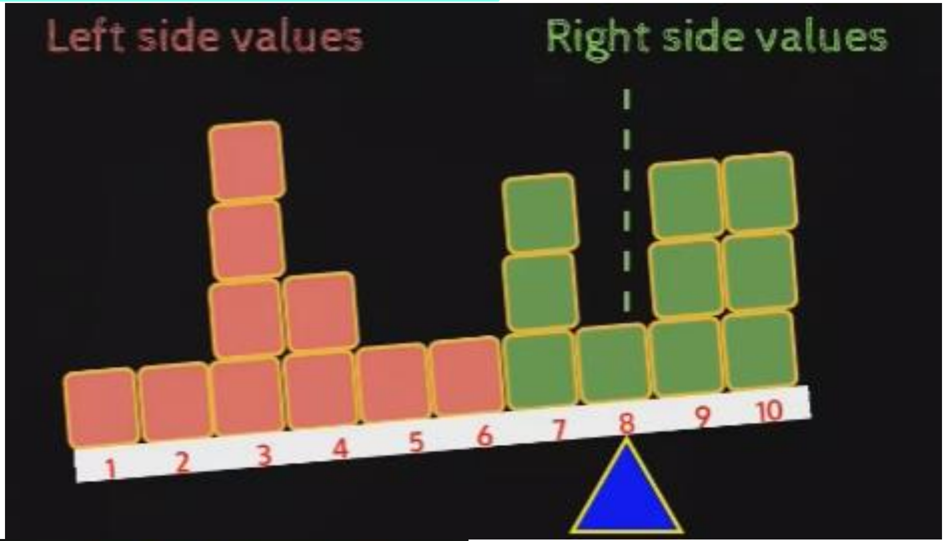
What is the physical interpretation of the above result

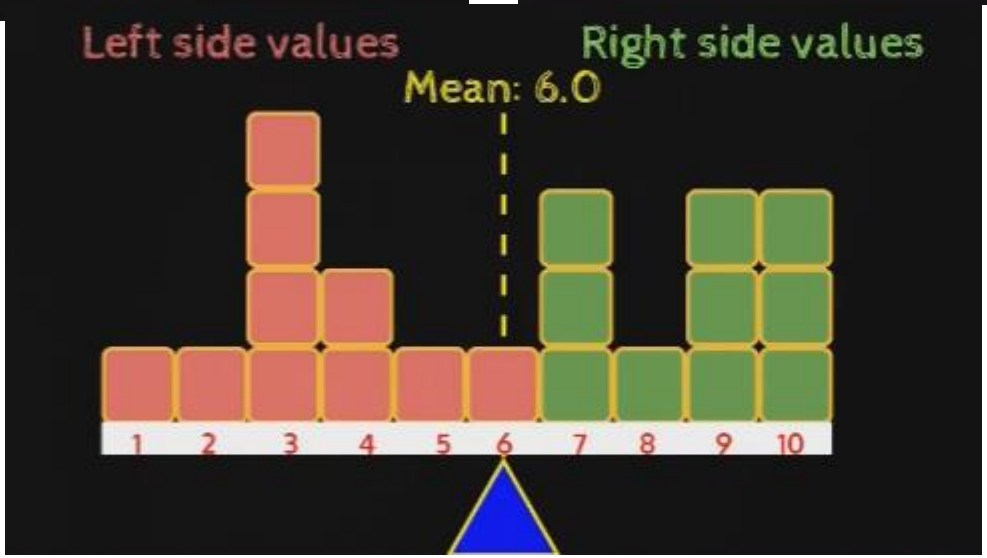
PRIME INTUIT Finishing School

Characteristics of Measures of centrality

Sum of deviations from the mean = 0







The deviations on the left side = The deviations on the right side

The mean is thus also called the center of gravity of the data

Score	Mean	Deviation*
8	9.67	-1.67
25	9.67	+15.33
7	9.67	-2.67
5	9.67	-4.67
8	9.67	-1.67
3	9.67	-6.67
10	9.67	+.33
12	9.67	+2.33
9	9.67	-.67

PRIME INTUIT
Finishing School

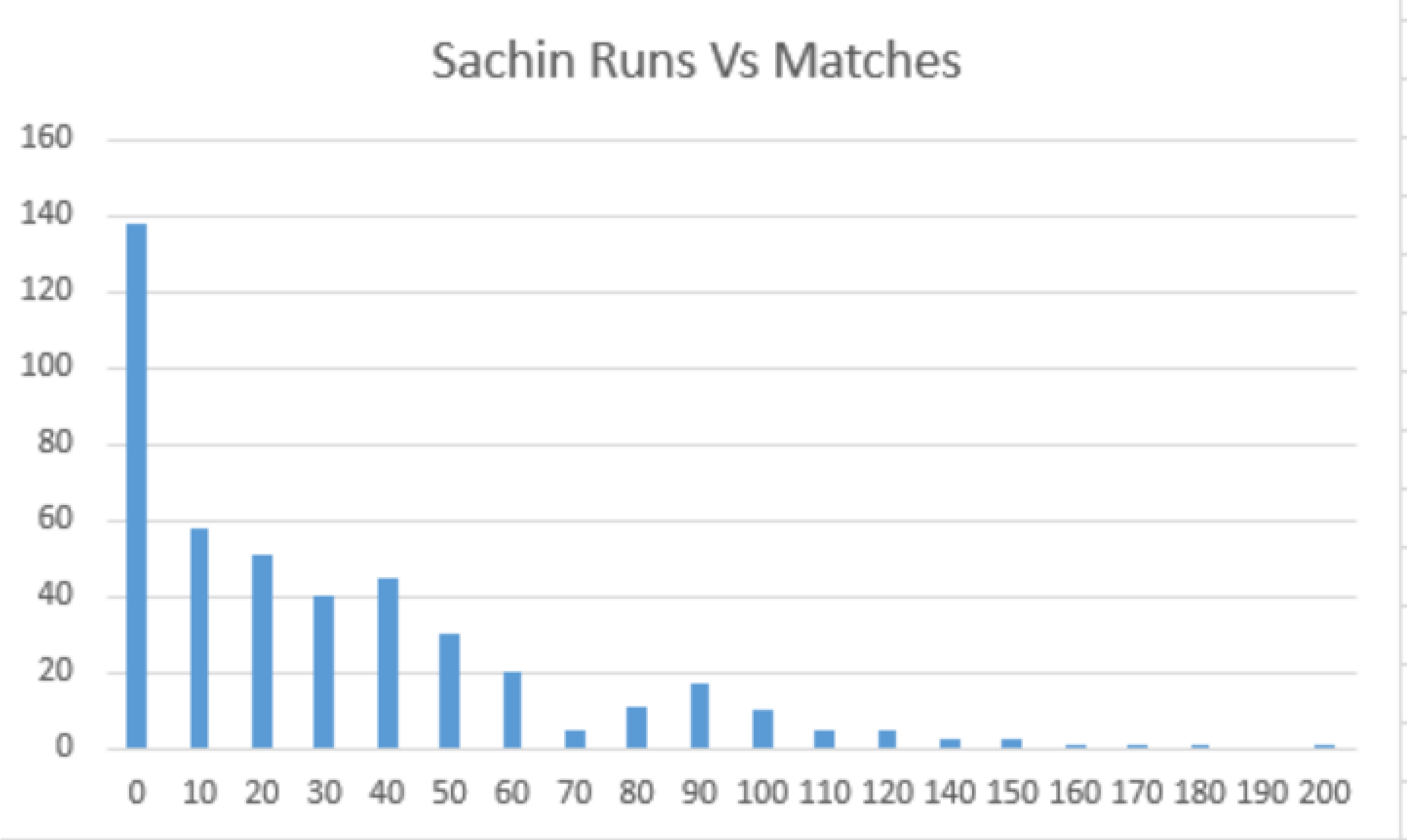
Outside our Learning Environment

https://www.youtube.com/watch?v=wTbrk0suwbg

https://www.youtube.com/watch?v=UwsrzCVZAb8

Sensitivity of Measures of centrality to outliers

Sensitivity of Measures of centrality to outliers



Outlier is a point which is far of from other values in the data set

Sensitivity of Measures of centrality to outliers

diff ↑ in cons

Outlier

Alistair Cook: 2, 7, 7, 10, 14, 16, 37, 39, 244

Joe Root: 1,9,14, 15, 51, 58, 61, 67, 83

Mean = 41.78

Mean = 39.88

Median = 14

Median = 51

Ashes 2017-2018 series (runs Scored)

diff ↓ ⇒ consist

Captain

Except for one high score (outlier). Cook performed poorly whereas Root was more consistent (This is reflected in the median but not in the mean)

Sensitivity of Measures of centrality to outliers

What if we dropped the outlier?

drop Only 244

Outlier

Alistair Cook: 2, 7, 7, 10, 14, 16, 37, 39, ~~244~~

Joe Root: 1,9,14, 15, 51, 58, 61, 67, 83

Old Mean = 41.78
New Mean = 16.5

Mean = 39.88

Median = 14
New Median = 12

Median = 51

Except for one high score (outlier). Cook performed poorly whereas Root was more consistent (This is reflected in the median but not in the mean)

Mean is very sensitive to outlier, where as the median is not so sensitive

Sensitivity of Measures of centrality to outliers

To account for the sensitivity to outliers it often advised to compute the trimmed mean

drop extreme

Alistair Cook: ~~2~~, 7, 7, 10, 14, 16, 37, 39, ~~244~~

Joe Root: ~~1~~,9,14, 15, 51, 58, 61, 67, ~~83~~

Old Mean = 41.78

Trimmed Mean=18.57

Mean = 39.88

Trimmed Mean=39.28

Trimmed mean is computed by dropping the extreme elements from either side (note that we need to drop the same number of elements from both sides**)**

PRIME INTUIT
Finishing School

Sensitivity of Measures of centrality to outliers

9.1, 9.4, 10.5, 10.5, 11.5, 11.7, 12.3, 12.7, 12.8, 13.7, 13.8, 14.9, 15.3, 16.2, 17.5, 17.6, 18.5, 18.6, 19.3, 19.9, 20.8, 23.6, 23.6, 24.3, 24.4, 32.1, 35.3, 45.5, 98.3, 133.1

Students salaries (INR Lakhs) at a top university

Mean = 24.57

Median = 17.5

Trimmed Mean = 18.95
(dropping 2 extreme values on either sides)

Sensitivity of Measures of centrality to outliers (mode)

Shikhar Dhawan T20Is

0,1,1,1,1,1,2,2,3,3,3,4,5,5,5,5,6,6,8,9,10,10,11,13,14,15,16,19,23,23,23,24,26,29,30,30,31,32,32,33,35,36,40,41,41,42,43,46,47,51,52,55,60,72,74,76,80,90,92

Sample:

8,9,11,19,21,23,25,27,31,35, 64,64

Mode is not sensitive to outliers unless the Mode itself is an outlier (which is very a rare case)

Sensitivity of Measures of centrality to outliers (Summary)

Mean is sensitive to outliers, where as median and mode are not

Its is often a good idea to compute trimmed mean by dropping same number of elements from both the extremes