PRIME INTUIT
Finishing School

What are some frequently used Percentiles?

Quartiles

What are Quartiles:

1, 2, 3, 7, 9	19, 28, 34, 36, 37	38, 41, 43, 45, 47	51, 53, 55, 57, 59

First 25% of data

Second 25% of data

Third 25% of data

Fourth 25% of data

Q1
25th Percentile

Q2
50th Percentile

Q3
75th Percentile

Quartiles divides the data into 4 equal parts

PRIME INTUIT
Finishing School

Quartiles Example:

$$L_{25} = \frac{25}{100} \times (50+1) = 12.75$$

$$Q_1 = Y_{25} = x_{12} + 0.75(x_{13} - x_{12})$$

$$= 5$$

Shikhar Dhawan T20I scores (50 Sorted Scores)

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,
24,26,29,30,30,32,33,35,41,42,43,46,47,5
1,52,55,60,72,74,76,80,90,92

$$L_p = \frac{p}{100} \times (n+1)$$

$$L_p = i.f$$

$$y_p = x_i + f(x_{i+1} - x_i)$$

$$L_{50} = \frac{50}{100} \times (50+1) = 25.5$$

$$Q_2 = Y_{50} = x_{25} + 0.5 \times (x_{26} - x_{25})$$

$$= 15.5$$

$$L_{75} = \frac{75}{100} \times (50+1) = 38.25$$

$$Q_3 = Y_{75} = x_{38} + 0.25(x_{39} - x_{38})$$

$$= 42.25$$

PRIME INTUIT
Finishing School

Quartiles Example:

Median:

$$\frac{x_{n+1}}{2} \quad \text{or} \quad \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

is odd n is even

$$L_{50} = \frac{50}{100} \times (n+1) = \frac{n+1}{2} \qquad 0.5n + 0.5$$

$$Q_2 = Y_{50} = x_{i_p} + 0.5 \times (x_{i_p+1} - x_{i_p})$$

Are they same/ Yes

$$y_p = x_{0.5n} + 0.5\left(x_{0.5n+1} - x_{0.5n}\right)$$

Why do the formula look so different ?

PRIME INTUIT Finishing School

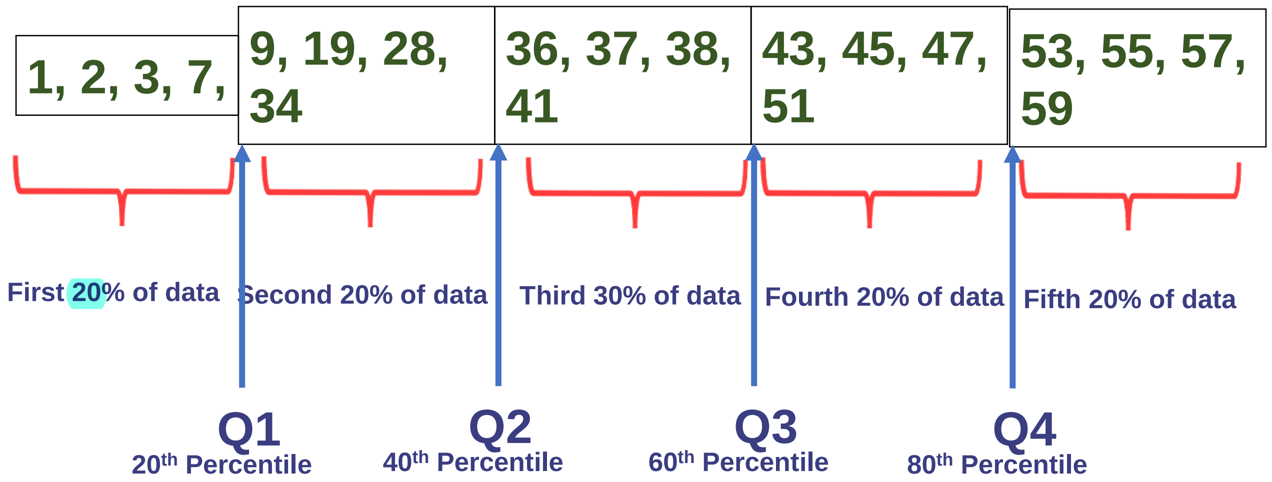
Quartiles Example:

Median is same as Q2

PRIME INTUIT
Finishing School

Quintiles

What are Quintiles:



Quintiles divides the data into 5 equal parts

PRIME INTUIT
Finishing School

Quintiles Example:

Shikhar Dhawan T20I scores (50 Sorted Scores)

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,
24,26,29,30,30,32,33,35,41,42,43,46,47,5
1,52,55,60,72,74,76,80,90,92

$$L_{60} = \frac{60}{100} \times (50+1) = 30.6$$

$$P_3 = Y_{60} = x_{30} + 0.6 * (x_{31} - x_{30})$$

$$= 27.8$$

$$L_{80} = \frac{80}{100} \times (50+1) = 40.8$$

$$P_4 = Y_{80} = x_{40} + 0.8 (x_{41} - x_{40})$$

$$= 46.6$$

Compute other quintiles similarly

PRIME INTUIT
Finishing School

Quintiles

What are Quintiles:

1, 2, 3, 7,	9, 19, 28, 34	36, 37, 38, 41	43, 45, 47, 51	53, 55, 57, 59
First 20% of data	Second 20% of data	Third 30% of data	Fourth 20% of data	Fifth 20% of data

Q1 — 20th Percentile

Q2 — 40th Percentile

Q3 — 60th Percentile

Q4 — 80th Percentile

Quintiles divides the data into 5 equal parts

Quintiles Example:

Shikhar Dhawan T20I scores (50 Sorted Scores)

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,
24,26,29,30,30,32,33,35,41,42,43,46,47,5
1,52,55,60,72,74,76,80,90,92

$$L_{60} = \frac{60}{100} \times (50+1) = 30.6$$

$$P_3 = Y_{60} = x_{30} + 0.6 * (x_{31} - x_{30})$$

$$= \underline{\underline{27.8}}$$

$$L_{80} = \frac{80}{100} \times (50+1) = 40.8$$

$$P_4 = Y_{80} = x_{40} + 0.8 (x_{41} - x_{40})$$

$$= \underline{\underline{46.6}}$$

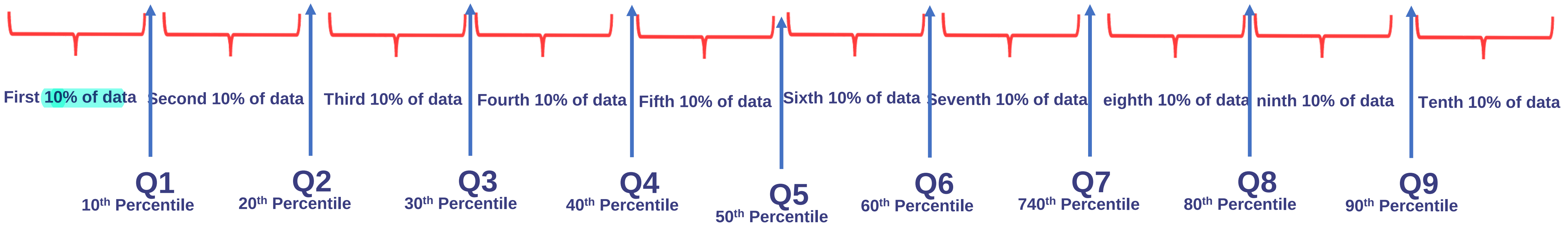
Compute other quintiles similarly

PRIME INTUIT
Finishing School

Deciles

What are Deciles:

1, 2	3, 7	9, 19	28, 34	36, 37	38, 41	43, 45	47, 51	53, 55	57, 59



Deciles divides the data into 10 equal parts

PRIME INTUIT
Finishing School

Deciles Example:

Shikhar Dhawan T20I scores (50 Sorted Scores)

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,
24,26,29,30,30,32,33,35,41,42,43,46,47,5
1,52,55,60,72,74,76,80,90,92

$$L_{30} = \frac{30}{100} * (50+1) = 15.3$$

$$D_3 = Y_{30} = x_{15} + 0.3 * (x_{16} - x_{15})$$

$$= \underline{\underline{5.3}}$$

Compute other deciles similarly

Compute the percentile rank of a value in the data?

Percentile Rank

Compared to other students, how do you rate the performance of the students who scored 44?

44, 43, 37, 68, 55, 46, 19, 59, 34, 46, 51, 62, 47, 52, 44, 28, 36, 56, 65, 60, 55, 66, 54, 48, 62

OR

What is the percentile rank of the student who scored 44

The percentile rank of a value is the percentage of data values that are less than or equal to it

Percentile Rank : Example 1

19, 28, 34, 36, 37, 43, 44, 44, 46, 46, 47, 48, 51, 52, 54, 55, 55, 56, 59, 60, 62, 62, 65, 66, 68

PR_s = Percentile rank of the scores

C_s = number of values less than s

F_s = number of values equal to s

$$PR_s = \frac{C_s + (0.5 * f_s)}{n} * 100$$

= (6 + (0.5*2) / 25) * 100 = 28

$$PR_s = \left[\frac{C_s + (0.5 \times f_s)}{n}\right] \times 100$$

PRIME INTUIT
Finishing School

Percentile Rank : Example 2

Shikhar Dhawan T20I scores (59 Sorted Scores)

0,1,1,1,1,1,2,2,3,3, 3,4,5,5,5, 5,6,6,6,
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,
41,42,43,46,47,51,52,55,60,72,74,76,80,90,92

$$PR_s = \frac{C_s + 0.5 * f_s}{n} * 100$$

$$PR_s = \frac{37 + 0.5 * 2}{59} * 100 = \underline{\underline{64.40}}$$

We typically round it upto the next whole number (65 in this case)

What is the effect of transformation on percentiles?

Transformations

Scaling and Shifting

0F : [22.46, 23.54, 24.26, 27.86, 30.2, 30.74, 34.52, 35.96, 40.46, 44.06, 52.7, 54.68, 56.66, 57.56, 59.54

0C : [-5.3, -4.7, -4.3, -2.3, -1.0, -0.7, 1.4, 2.2, 4.7. 6.7, 11.5, 12.6, 13.7, 14.2, 15.3

Xnew = a * x + C

a = 5/9, c = -160 / 9

PRIME INTUIT
Finishing School

Effect of Transformation on Percentiles

Lp = $\frac{P}{100}(n+1)$ $\qquad$ $y_P = x_{i_P} + f_P\left(x_{i_{P+1}} - x_{i_P}\right)$

Formula for Transformation = $x_{new} = a * x + c$

New Location $L_p^{new} = \frac{P}{100}(n+1)$

$$y_p^{new} = x_{i_P}^{new} + f_P\left(x_{i_{P+1}}^{new} - x_{i_P}^{new}\right)$$

$$y_p^{new} = a * y_P + c$$

PRIME INTUIT
Finishing School

Summary

What is a Percentile

How to compute Percentile?

Frequently used percentiles

How to compute Percentile rank of value?

What is the effect of transformation on percentiles

PRIME INTUIT
Finishing School

What are the measures of spread?

Why do we need measures of spread

Sample A: 61, 61, 62, 62, 63, 63, 64, 64, 65, 65

Mean: 63
Median: 63

Note: All values are very close to the mean & median (low variability in data)

Sample B: 11, 21, 41, 52, 63, 63, 74, 87, 98, 120

Mean: 63
Median: 63

Note: Some values are far from the mean & median (high variability in data)

The measures of centrality don't tell us anything about the spread and variability in the data

Measures of spread (range)

Sample A: 61, 61, 62, 62, 63, 63, 64, 64, 65, 65

Mean: 63
Median: 63

Range: (max value – min value)
= 65 - 61 = 4

Sample B: 11, 21, 41, 52, 63, 63, 74, 87, 98, 120

Mean: 63
Median: 63

Range: (max value – min value)
= 120 - 11 = 109

Range clearly tells us that the second sample has more variability / spread than the first

Measures of spread (range)

Farm yields of Wheat (in Punjab): 40.1, 40.9, 41.8, 44, 46.8, 47.2, 48.6, 49.3, 49.4, 51.9, 53.8, 55.9, 57.3, 58.1, 60.2, 60.7, 61.1, 61.4, 62.8, 633

Range: (max value – min value) = 633 – 40.1 = 592.9

Note: Most values were close to 40.1, the range however gets exaggerated due to one outlier (633)

Just like the mean, the range is very sensitive to outliers !

PRIME INTUIT
Finishing School

$n=20$

Measures of spread (IQR)

with outlier

First 25% of data	Second 25% of data	Third 25% of data	Fourth 25% of data
40.1, 40.9, 41.8, 44, 46.8	47.2, 48.6, 49.3, 49.4, 51.9	53.8, 55.9, 57.3, 58.1, 60.2	60.7, 61.1, 61.4, 62.8, 633

Q1
25th Percentile

Inter Quartile Range

Q3
75th Percentile

Inter Quartile Range (IQR) = Q3 – Q1

Inter Quartile Range (IQR) = 60.575 – 46.9 = 13.675

$$L_{75} = \frac{75}{100} * (20+1) = 15.75$$

$$Q_3 = Y_{75} = x_{15} + 0.75(x_{16} - x_{15}) = 60.575$$

$$L_{25} = \frac{25}{100} * (20+1) = 5.25$$

$$Q_1 = Y_{25} = x_5 + 0.25(x_6 - x_5) = 46.9$$

PRIME INTUIT
Finishing School

n=19 instead of 20

Measures of spread (IQR)

after dropping outliers

40.1, 40.9, 41.8, 44, **46.8**	47.2, 48.6, 49.3, 49.4, 51.9	53.8, 55.9, 57.3, 58.1, **60.2**	60.7, 61.1, 61.4, 62.8, ~~63~~

First 25% of data | Second 25% of data | Third 25% of data | Fourth 25% of data

Q1
25th Percentile

Inter Quartile Range

Q3
75th Percentile

no need to remove 1 &

1 pos removed

Inter Quartile Range (IQR) = Q3 – Q1

Inter Quartile Range (IQR) = 60.575 – 46.9 = 13.675

New IQR = X15 – X5 = 60.2 – 46.8 = 13.4

IQR is Clearly not sensitive to outliers (that is it will not change drastically if we drop the outlier)

n changed

$L_{75} = \frac{75}{100}(19+1) = 15$

$L_{25} = \frac{25}{100}(19+1) = 5$

Measures of spread (Variance)

How different are the values in the data from typical value (mean) in the data ?

Solution: Compute the sum or average deviation of all points from the mean

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

Issue: We already know that sum of deviations from the mean is 0

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

PRIME INTUIT
Finishing School

Measures of Spread (Variance)

Sample A: 61, 61, 62, 62, 63, 63, 64, 64, 65, 65

Mean: 63
Median: 63

Sum of deviations = 0

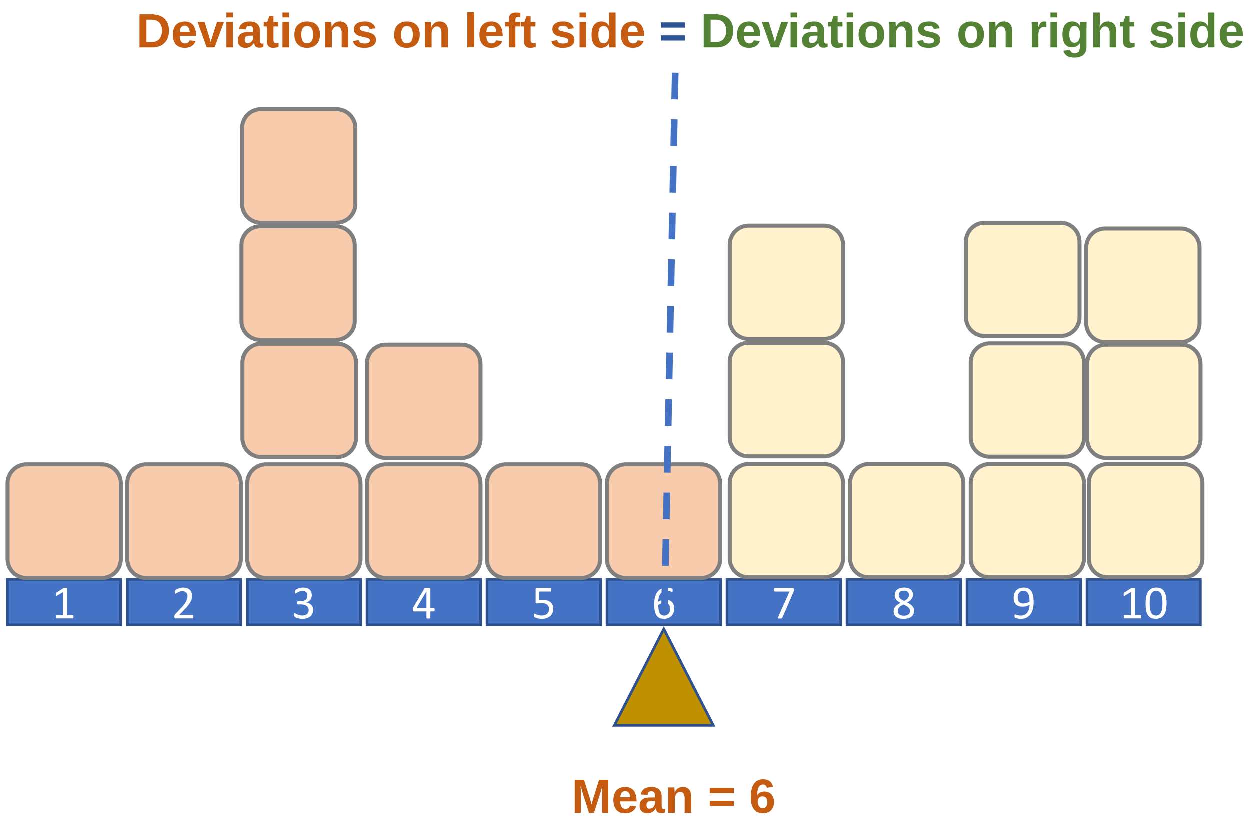
Sample A: 11, 21, 41, 52, 63, 63, 74, 87, 98, 120

Mean: 63
Median: 63

Sum of deviations = 0

The sum of deviations does not tell us anything about the spread or variation of the data

Measures of Spread (Variance)



Summary: We do not care about the sign of the deviation (both positive and negative deviations contribute to the spread in the data and hence we do want them to cancel each other)

Measures of Spread (Variance)

Issue: The sum of deviations from the mean is 0

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})=0$$

Reason: The positive deviations cancel the negative deviations

Solution 1: Use absolute values

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

Solution 2: Use square values (Preferred Solution)

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

PRIME INTUIT
Finishing School

Measures of Spread (Variance)

Variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

If computed from a sample

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

If computed from the entire population

Why is there a difference in the formula?
We will clarify this later once we introduce probability theory

PRIME INTUIT
Finishing School

Measures of Spread (Variance)

2 diff dataset

Variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\bar{x} = 83$ $\quad S^2 = \left(\frac{1}{10-1}\right)20 = 2.22$

$\bar{x} = 83$ $\quad S^2 = \left(\frac{1}{10-1}\right)10244 = 1138$

x	$(x-\bar{x})$	$(x-\bar{x})^2$		x	$(x-\bar{x})$	$(x-\bar{x})^2$
61	-2	4		11	-52	2704
61	-2	4		21	-42	1764
62	-1	1		41	-22	484
62	-1	1		52	-11	121
63	0	0		63	0	0
63	0	0		63	0	0
64	1	1		74	11	121
64	1	1		87	24	576
65	2	4		98	35	1225
65	2	4		120	57	3249
630	**0**	**20**		**630**	**0**	**10244**

PRIME INTUIT
Finishing School

Measures of Spread (Standard deviation)

Observation:

$$S = \sqrt{s^2} = \boxed{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

If computed for a sample

Standard deviation is measured in the same units as the data

Standard deviation = Square root of Varience

$$\sigma = \sqrt{\sigma^2} = \boxed{\sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

If computed from the entire population

Recap of notations

Statistic	Sample (Size n)	Population (Size N)
Mean	$\bar{n}$	$\bar{N}$
Variance	s^2	σ^2
Standard Deviation	s	σ

What we square the Deviations?

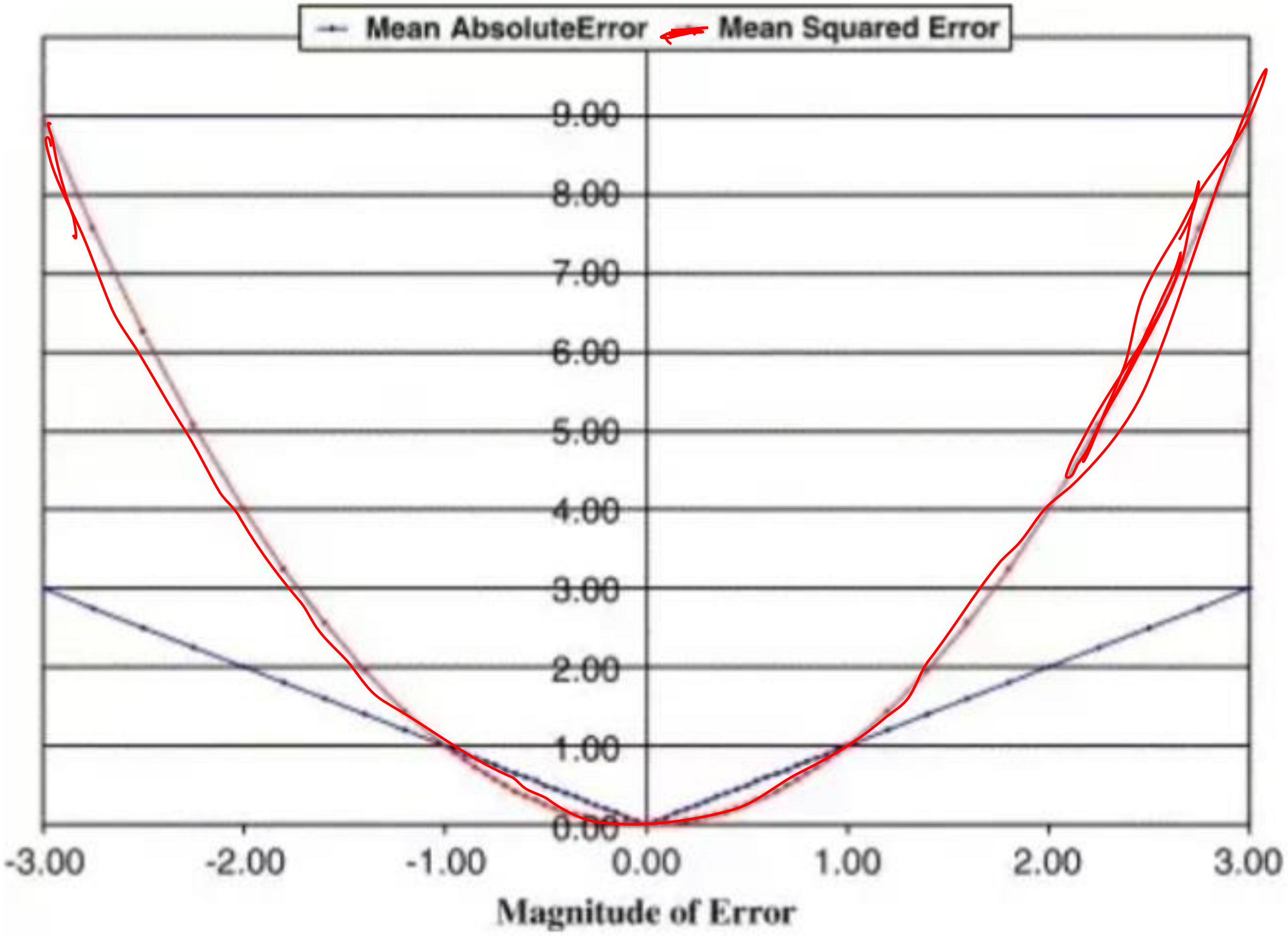
Why we square the Deviations?

Reason 1: The Square function has better properties than the absolute function

1. The square function is a smooth function and hence differentiable everywhere
2. The absolute function is not differentiable at $X_i - X = 0$

Why do we care about differentiability?

In many applications (especially in ML) We need functions which are differentiable



PRIME INTUIT
Finishing School

Why we square the Deviations?

Reason 2: The Square function magnifies the contribution of outliers

Why do we want to magnify the contribution of outliers?

Example: Toxic Content in a fertilizer

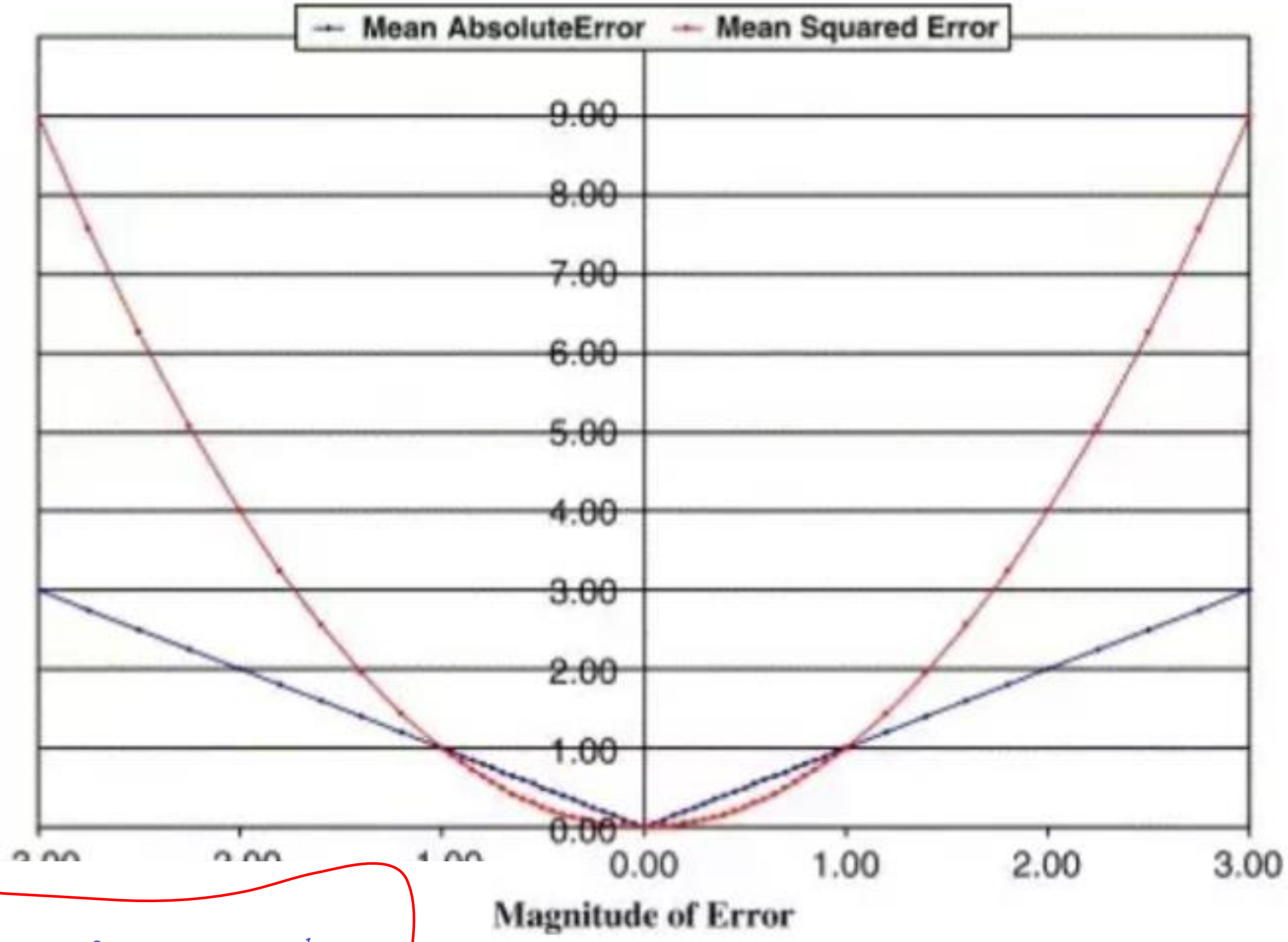
0.1, 0.2, 0.3, 0.3 0.5, 0.1, 0.4, 0.2, 0.6, 10.2

Mean = 1.29

Variance by square = 9.827

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Variance by absolute method = 1.782



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