Quantitative Analysis

DATA 201: Thinking With Data
Winter 2021

Jonathan Hudson, Ph.D
Instructor
Department of Computer Science
University of Calgary

Thursday, February 25, 2021

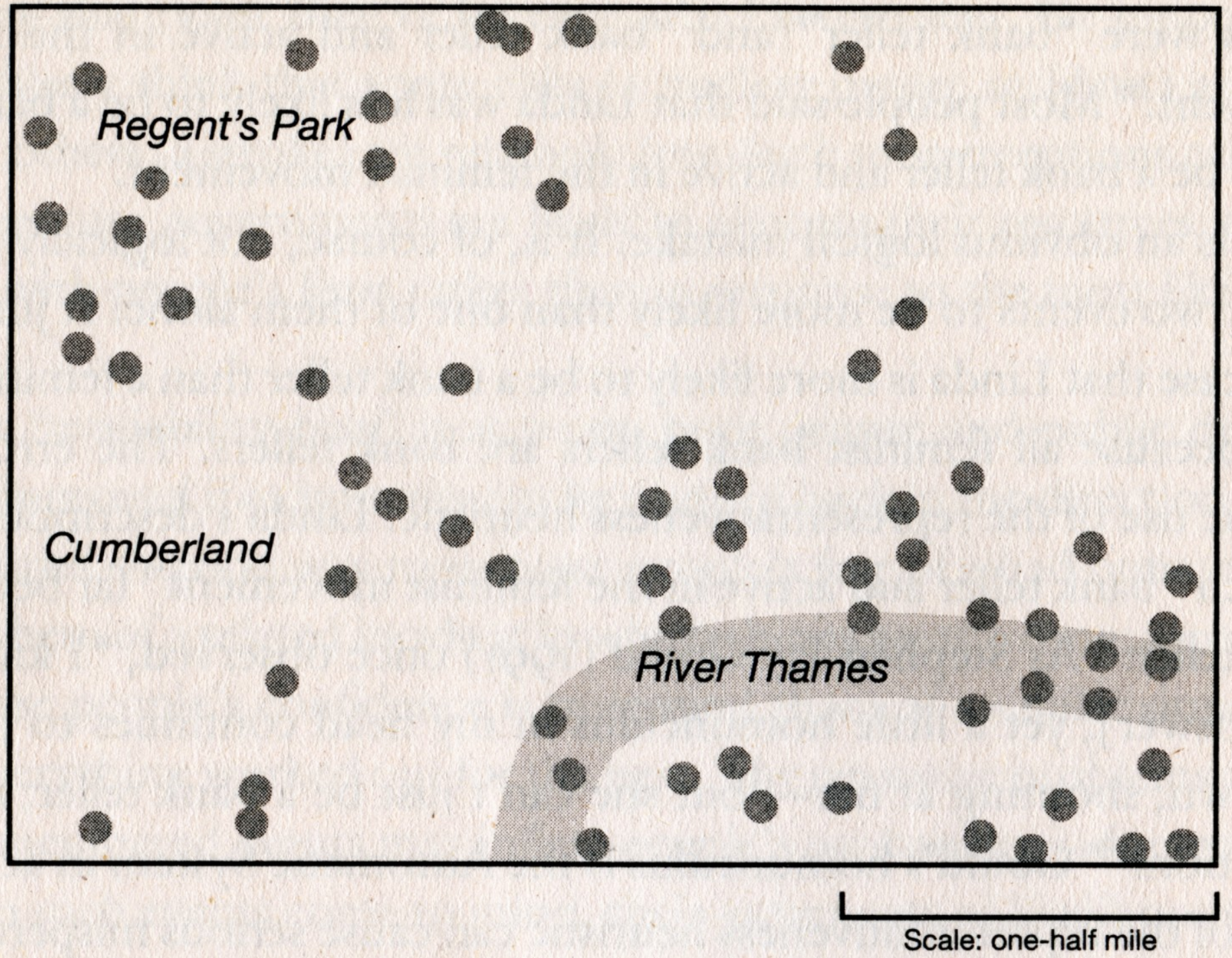
UNIVERSITY OF CALGARY

What is Statistics?

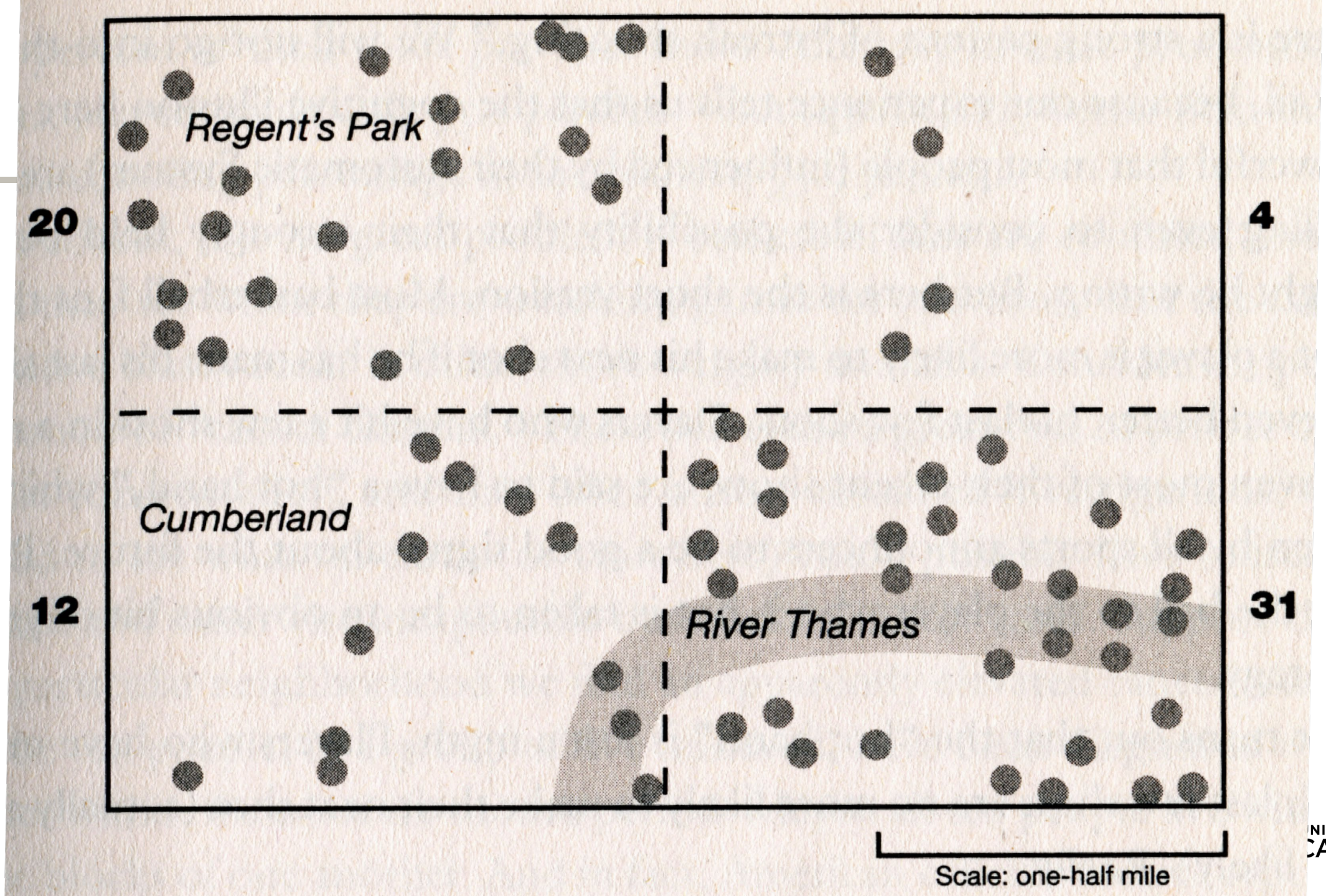
Statistics is the study of the collection, analysis, interpretation, presentation and organization of data.

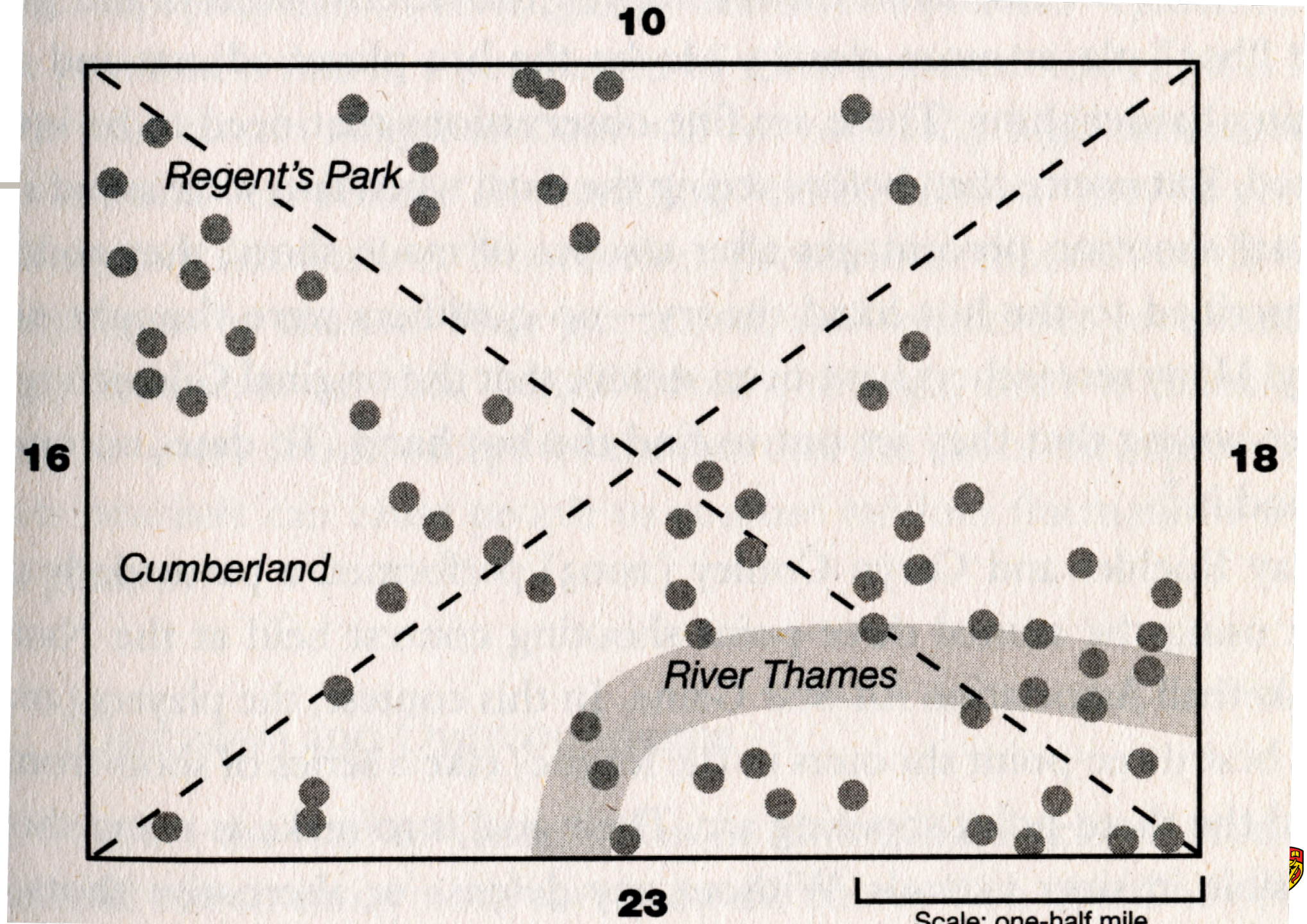
– Dodge, Y. (2006) The Oxford Dictionary of Statistical Terms, OUP.

Why Statistics?



UNIVERSITY OF CALGARY





We have questions and want answers, but our intuition is not always right.

Data Analysis

Exploratory Data Analysis

- the process of gathering evidence much like detective work

Confirmatory Data Analysis

- the process of evaluating evidence is comparable to a court trial

Exploratory Data Analysis

Exploratory Data Analysis

- Understanding data and finding interesting things from the data
- Visualizations can help

Confirmatory Data Analysis

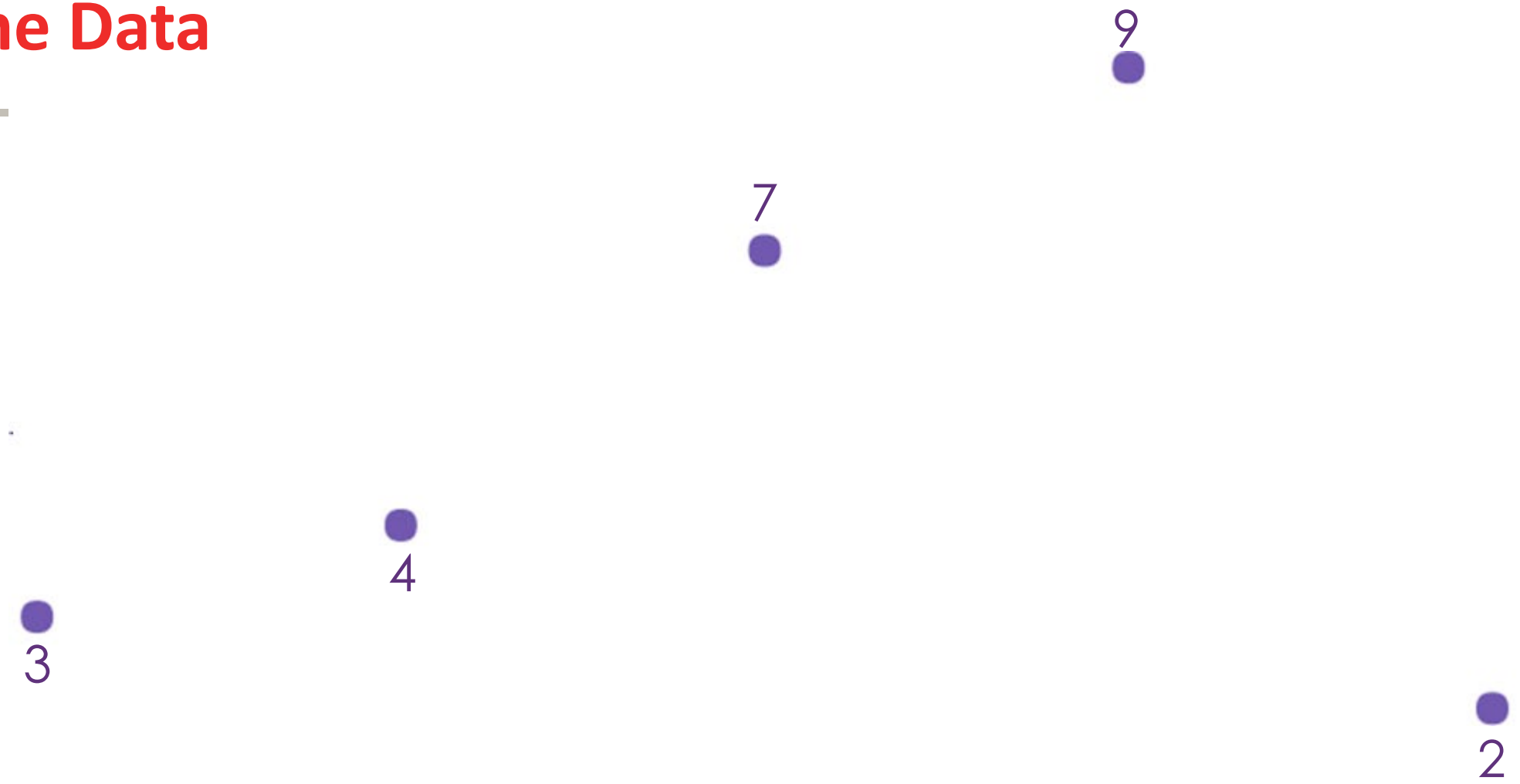
Confirmatory Data Analysis

- Testing hypotheses

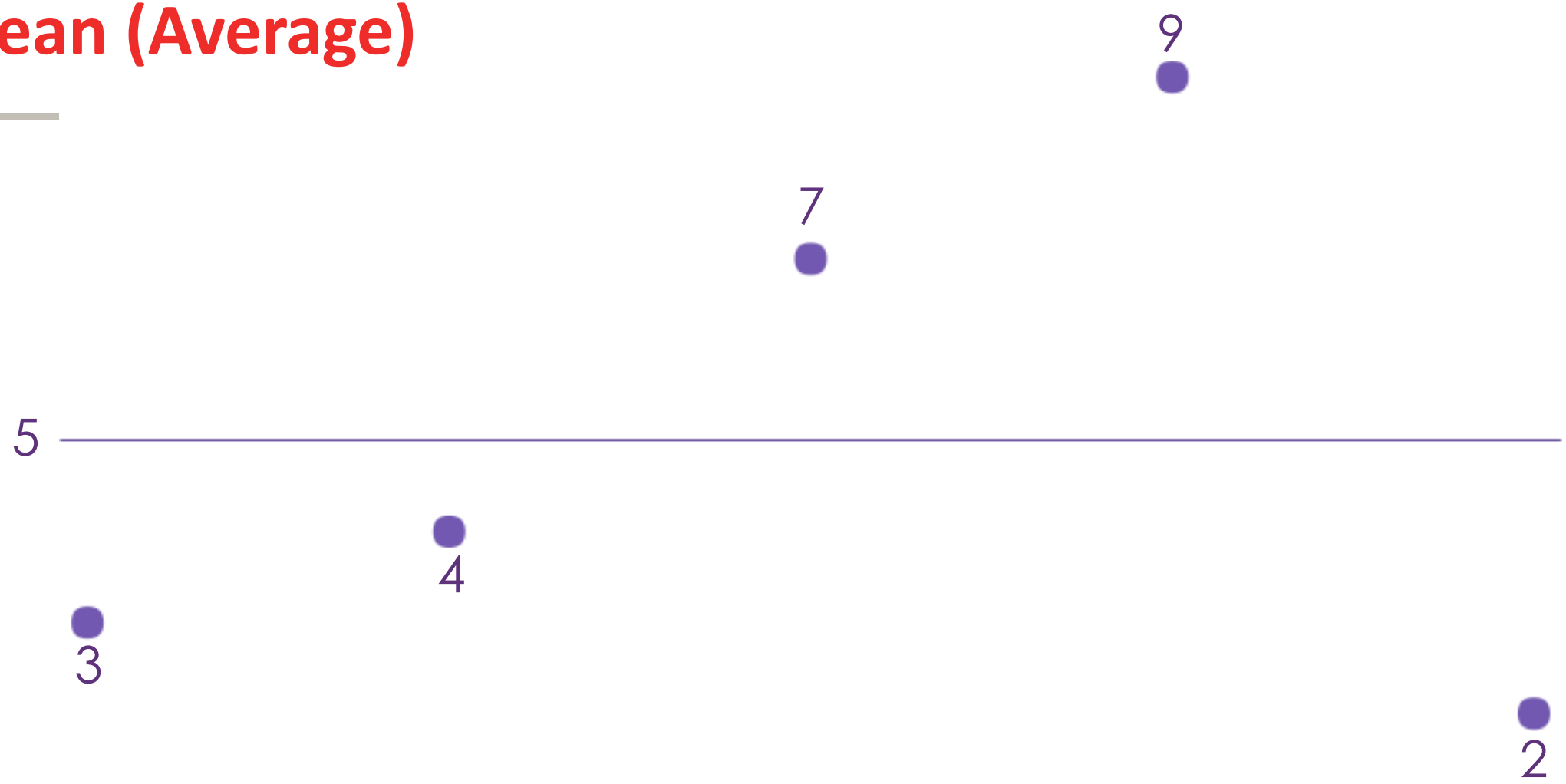
Basic Statistical Terms

- Mean
- Variance
- Standard Deviation

Some Data

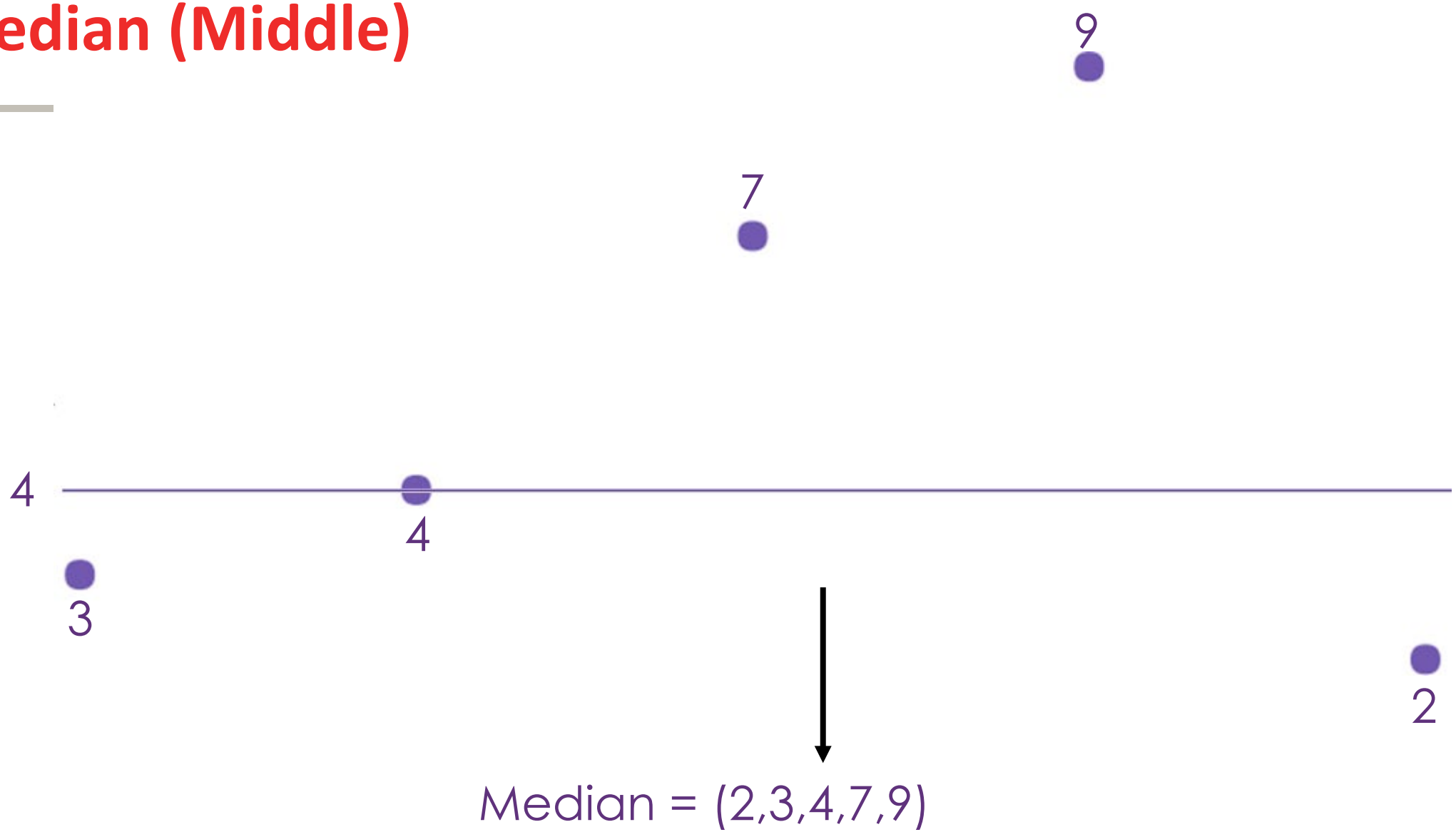


Mean (Average)

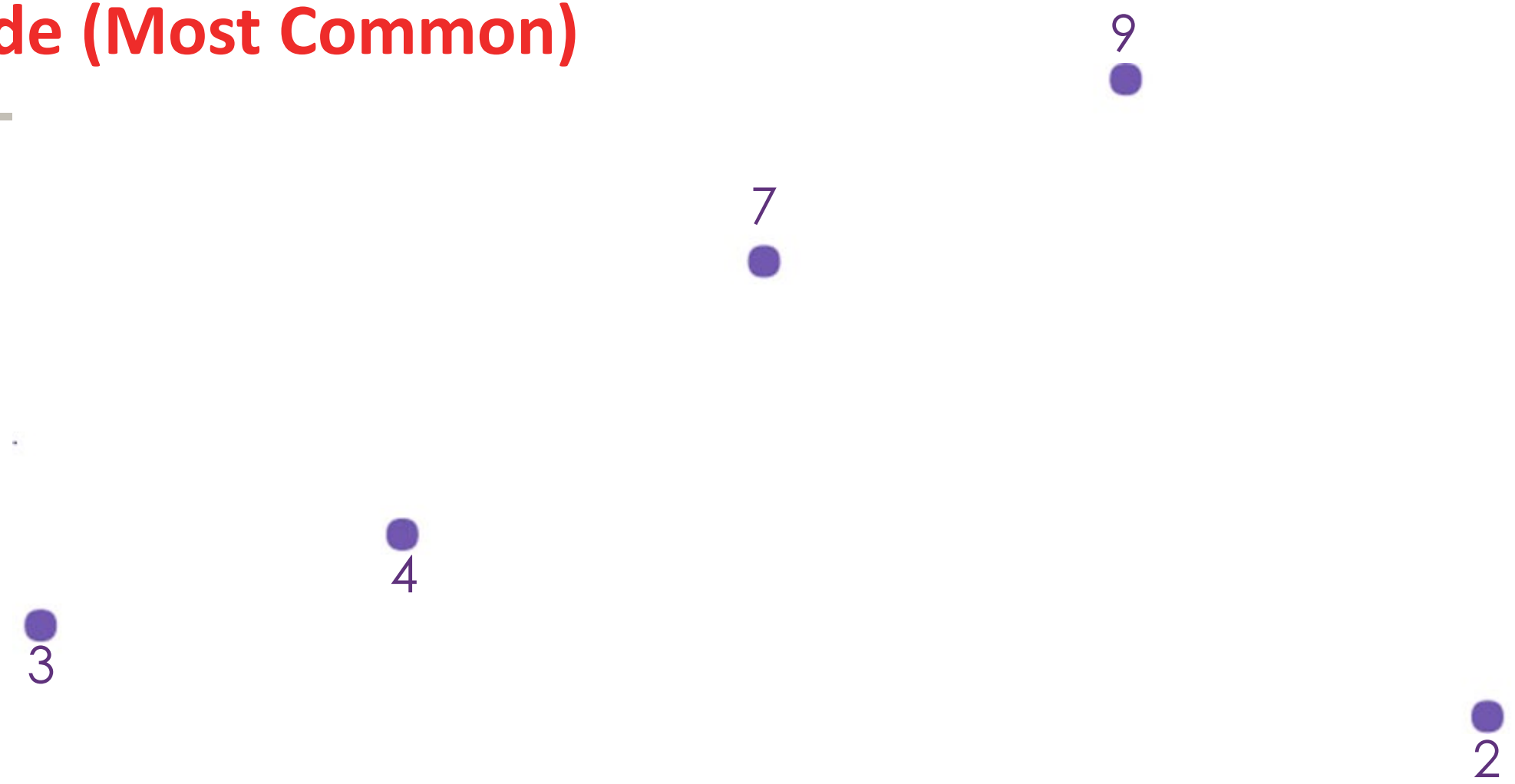


mean = (3+4+7+9+2) / 5

Median (Middle)

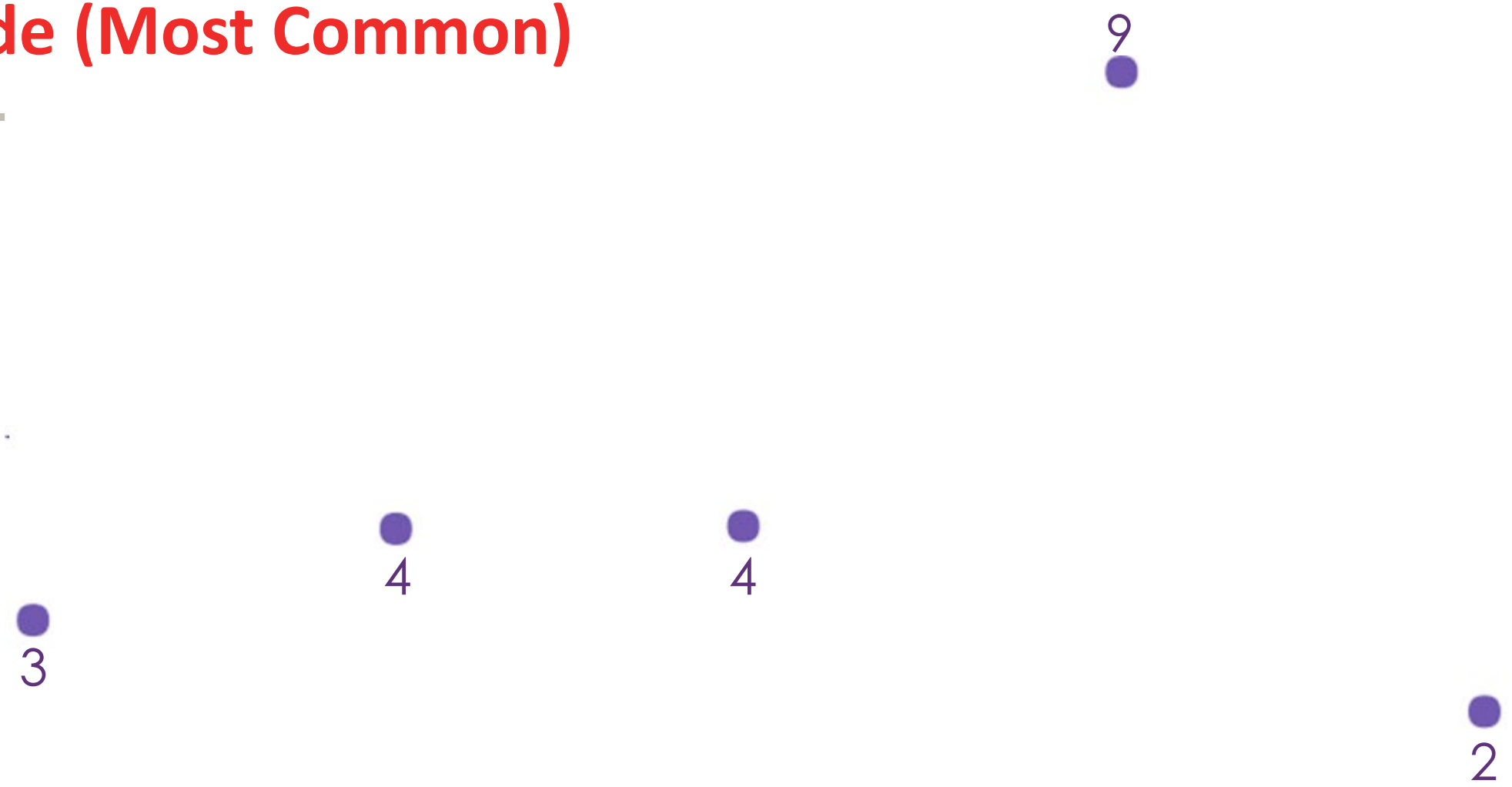


Mode (Most Common)



Mode = (2,3,4,7,9) =
None or All

Mode (Most Common)

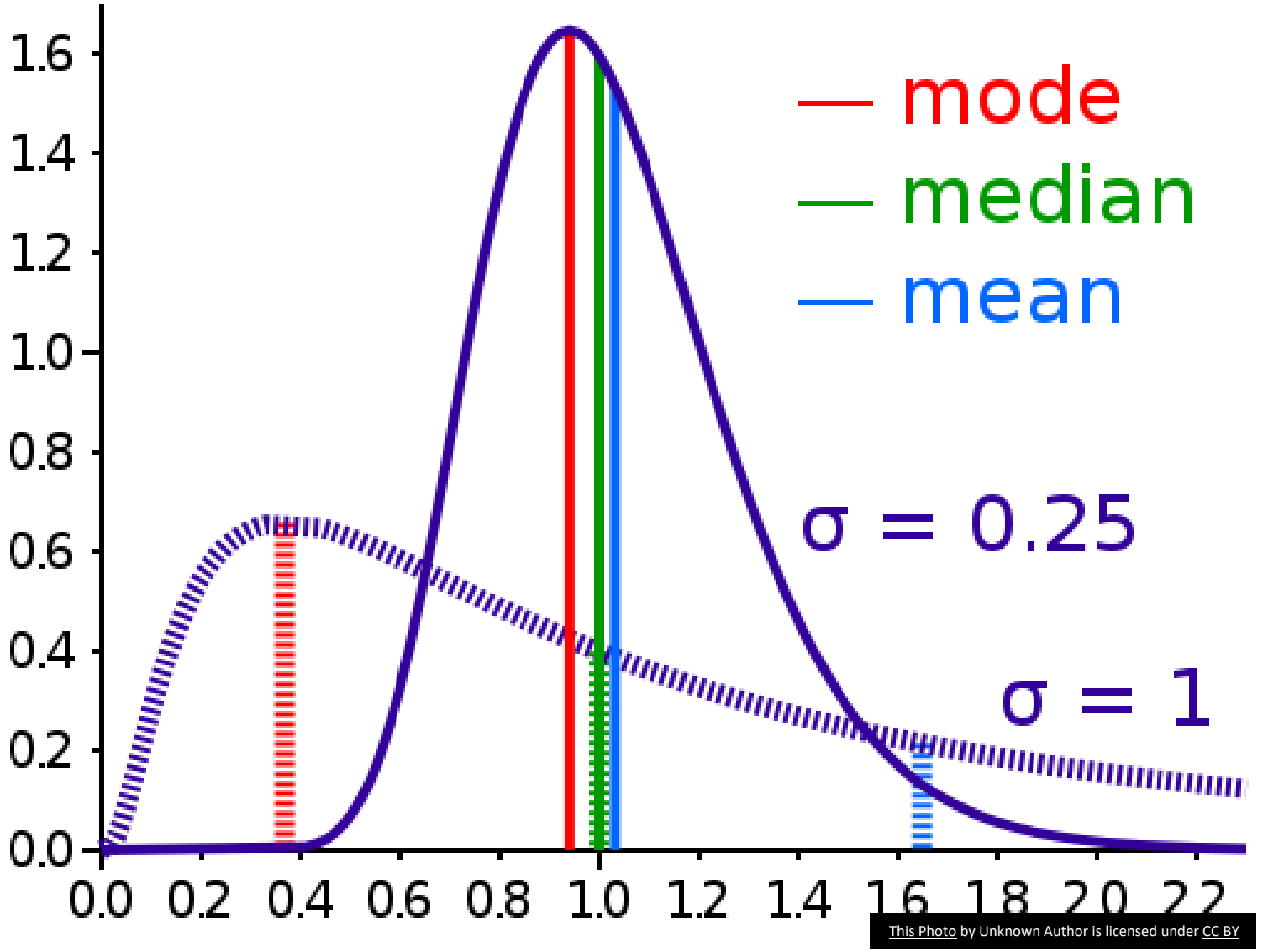


Mode = (2,3,4,4,9) = 4

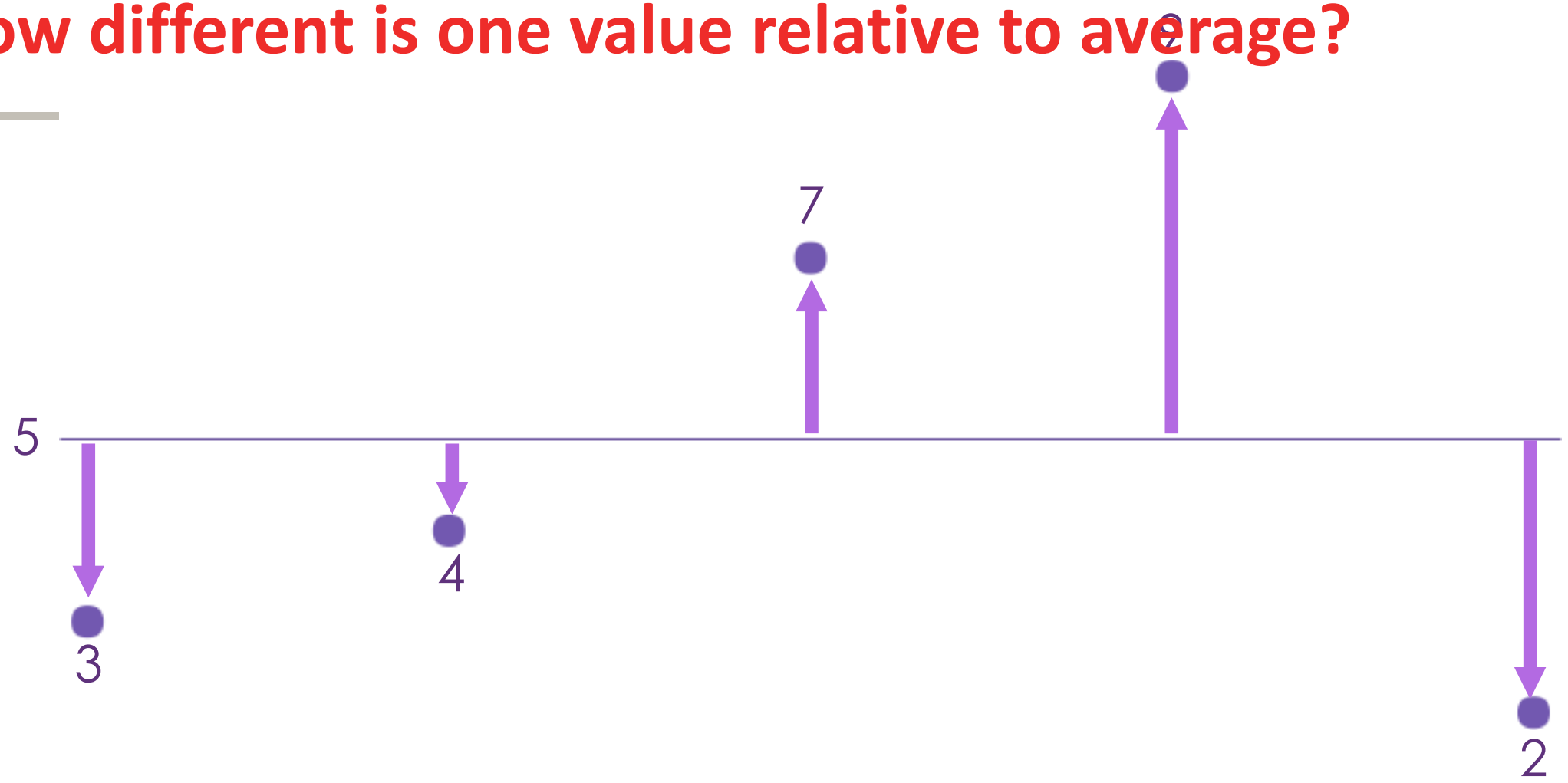
Why would we use mode over mean?

- Mean -> Good for continuous and symmetrical data (clustered around a point)
- Median -> When data has weird distribution this can avoid outlier influence
  - Ex when Bill Gates walks into the bar the average income in the bar goes up but the median is unchanged (ordinal data -> sortable)
- Mode -> Great for nominal data
  - (not ordered or relative and can't do math on it)
  - Could be used for other data but generally less useful that other 'central tendency measures'
  - Ex. USRIs report to instructor most common answer as 1-7 on the likert scale of strongly disagree to strongly agree (i.e. the mode of answers)
- When data is normally distributed (They are all the same!)

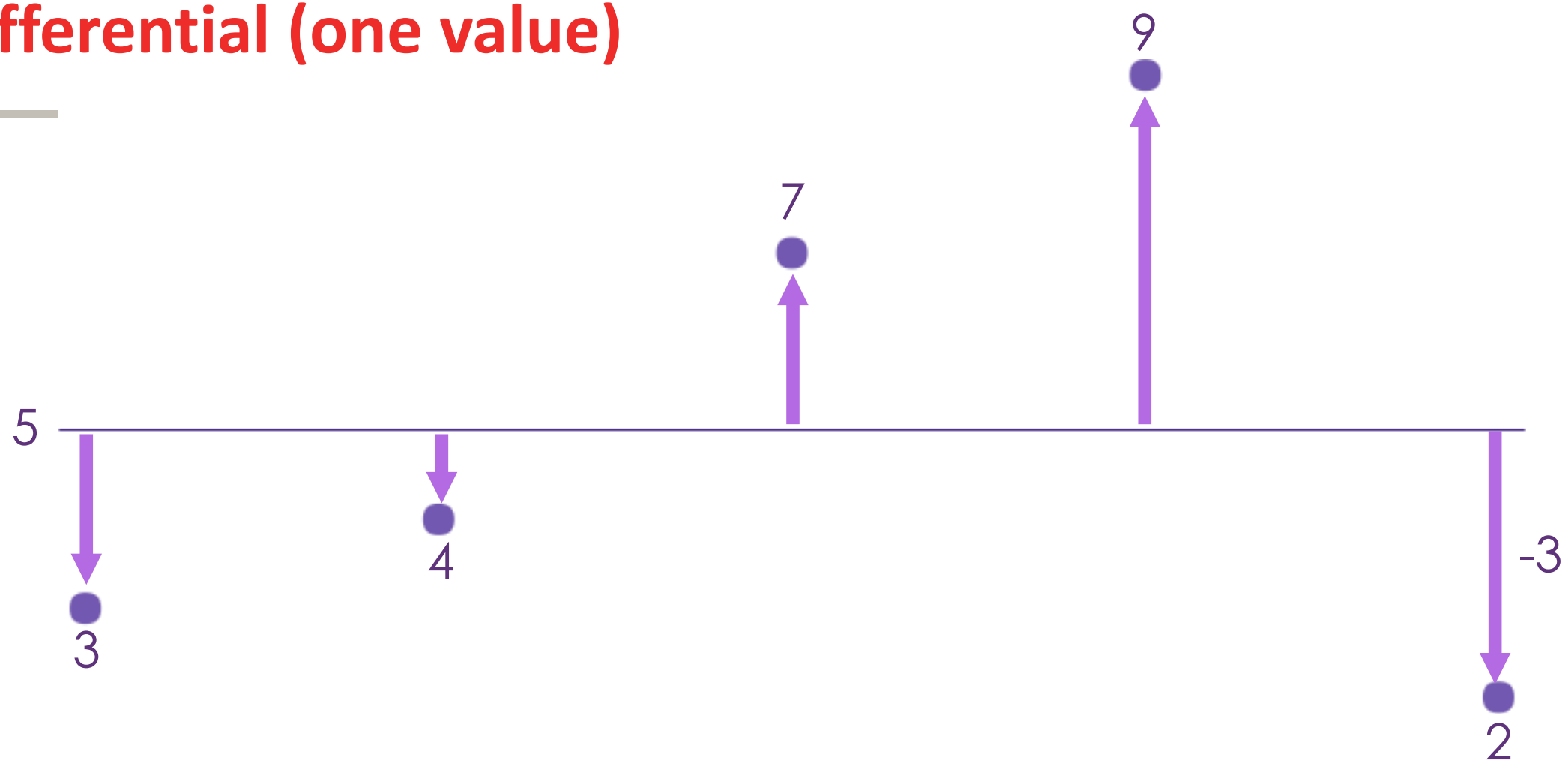
Mean
Median
Mode



How different is one value relative to average?



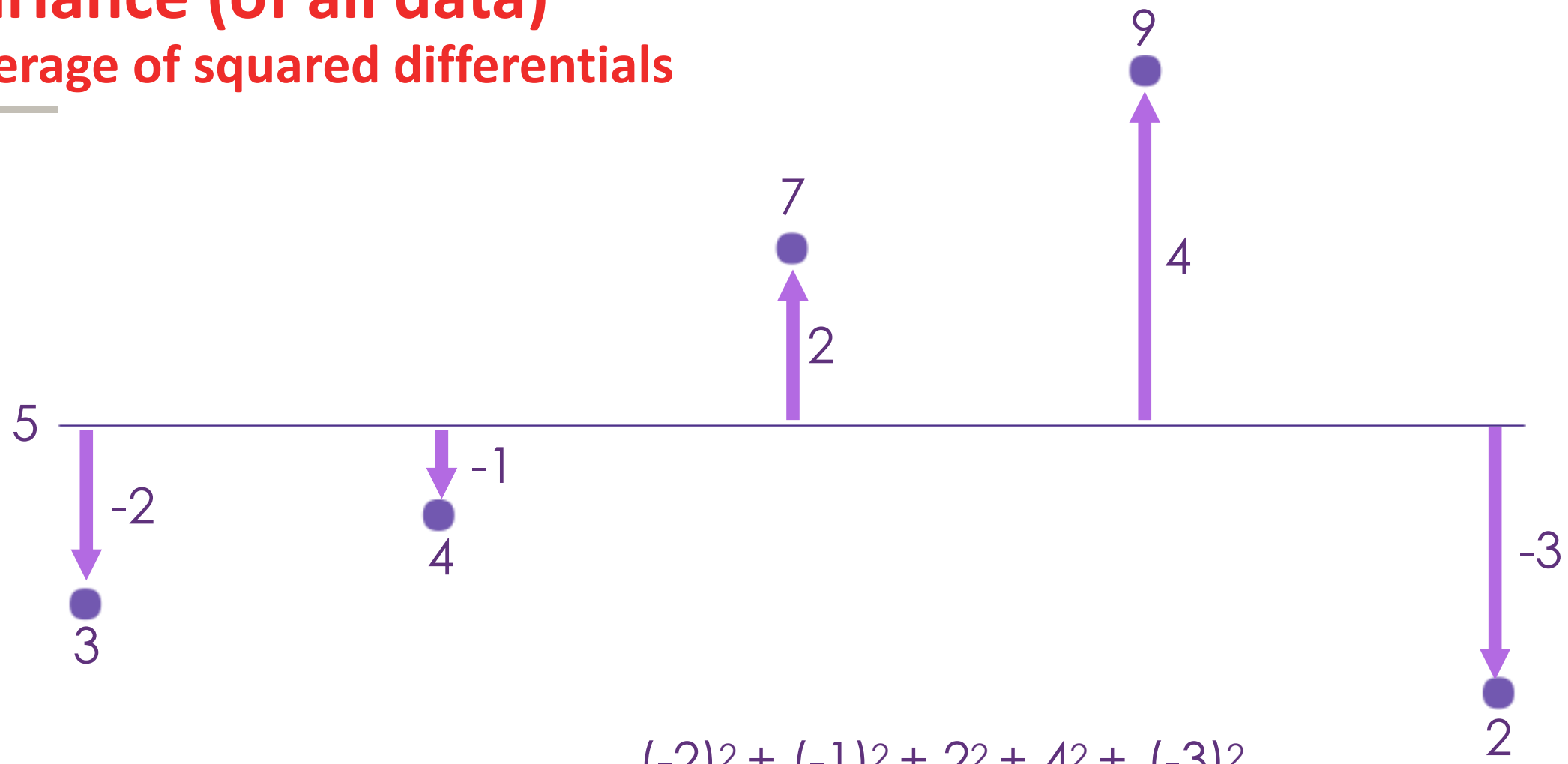
Differential (one value)



diff = (value - mean) = 2 - 5 = -3

Variance (of all data)

Average of squared differentials

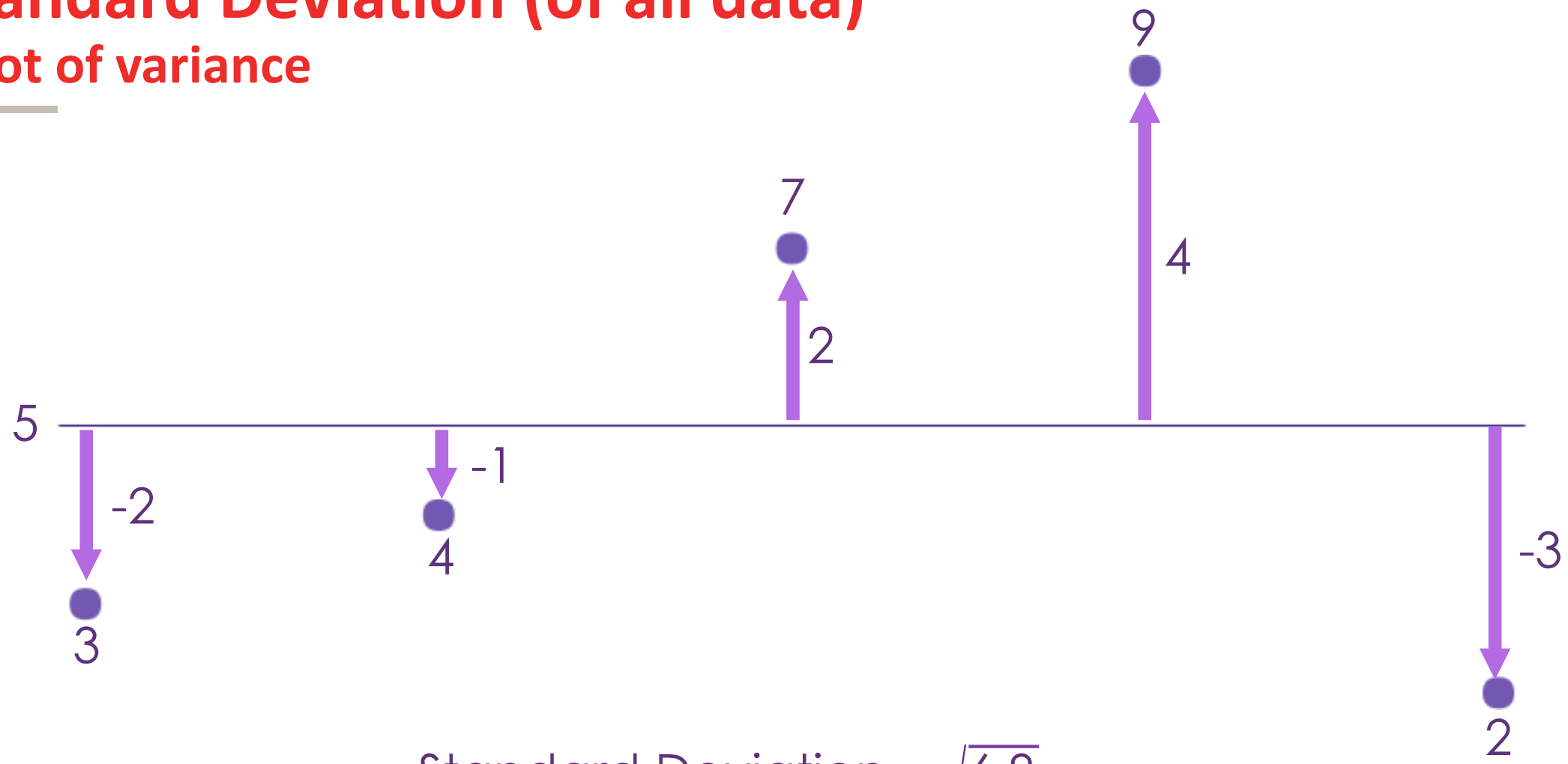


$$\text{Variance} = \frac{(-2)^2 + (-1)^2 + 2^2 + 4^2 + (-3)^2}{5}$$

$$= 6.8$$

Standard Deviation (of all data)

Root of variance



Standard Deviation = $\sqrt{6.8}$

= 2.6

Basic Statistical Terms

- Mean — average
- Variance — squared deviations of individual data points from the mean
- Standard Deviation — square root of the variance

Different Data Sets

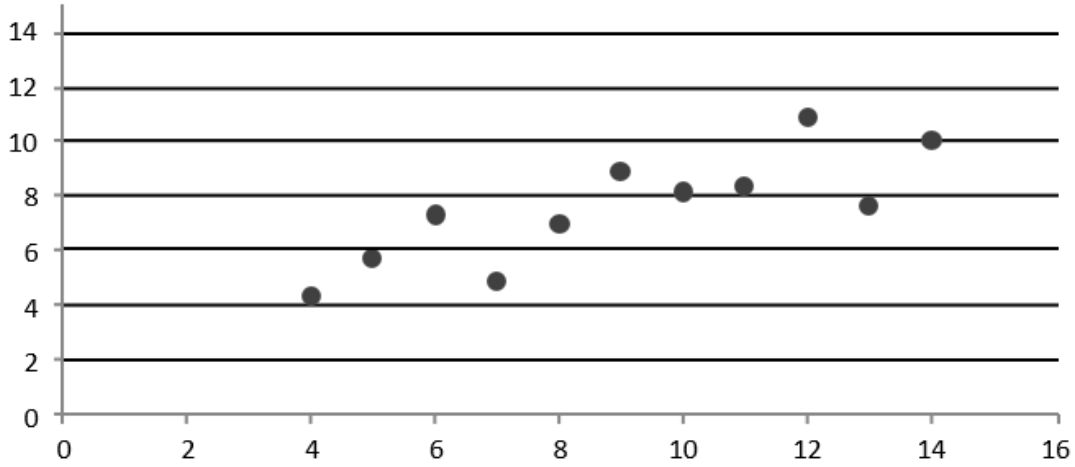
X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

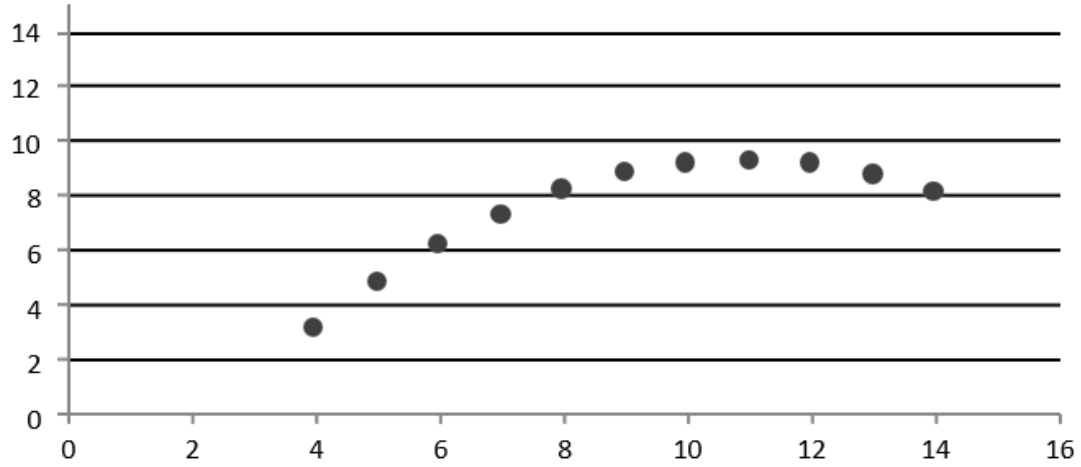
X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

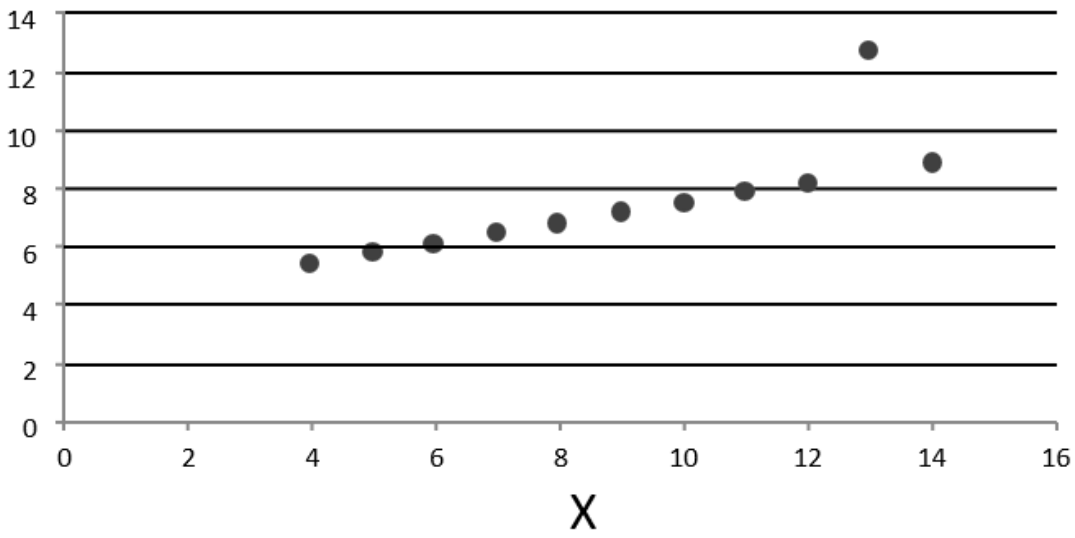
X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

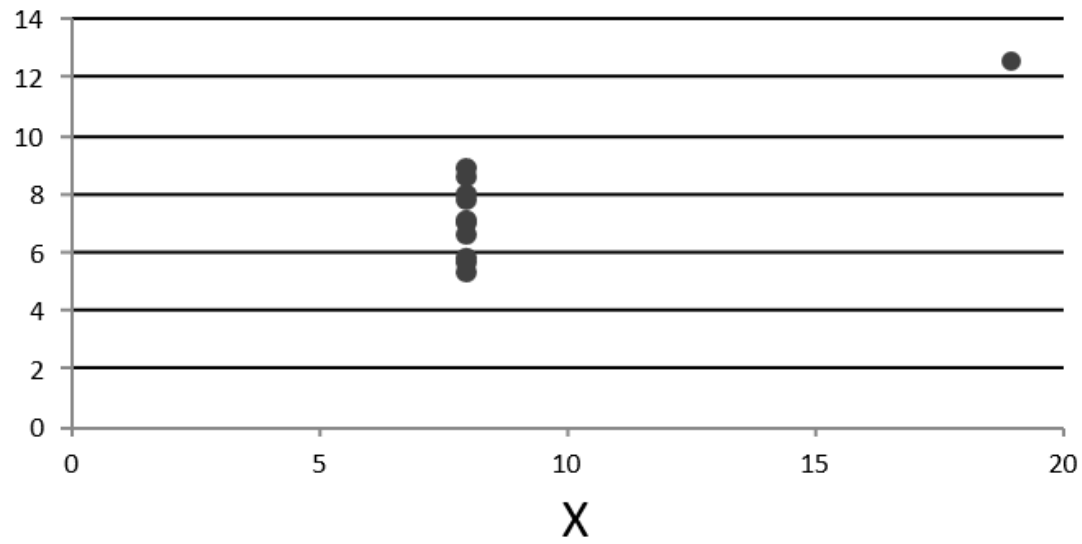
X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Anscombe 1973



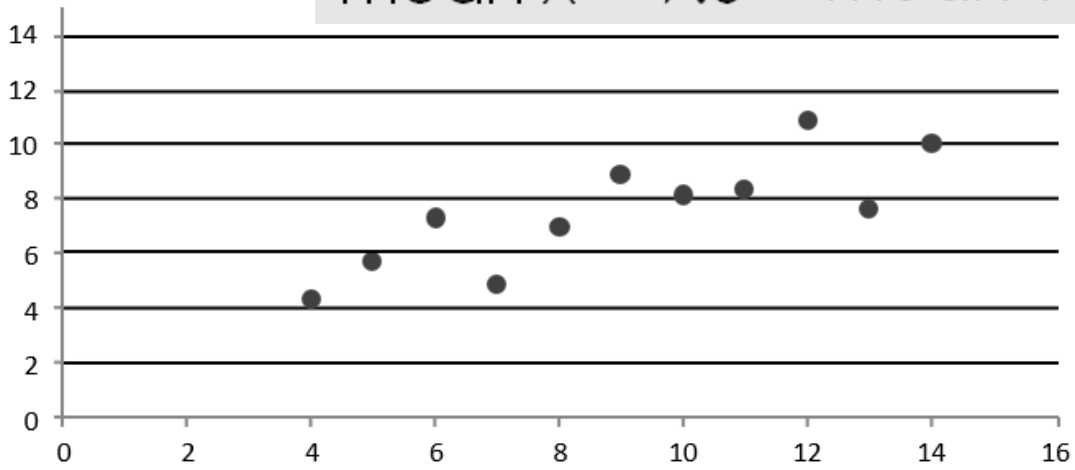


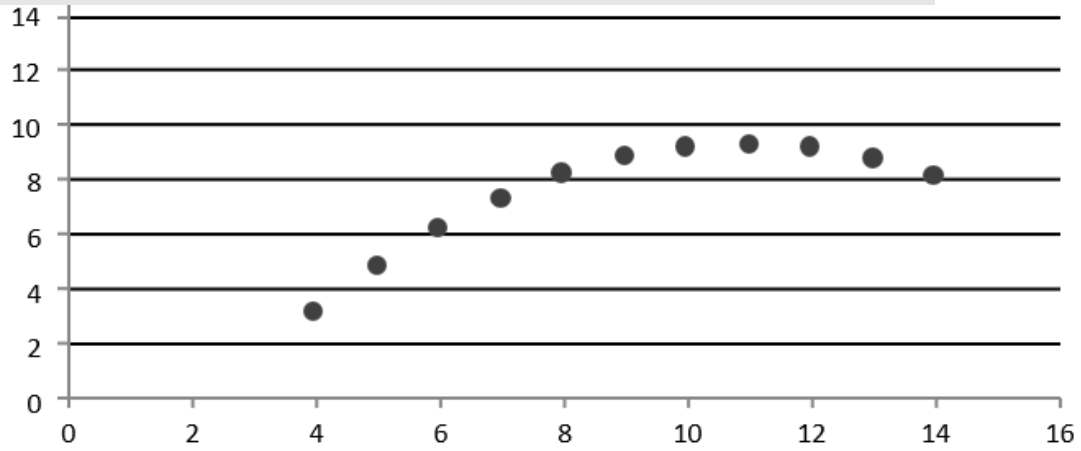


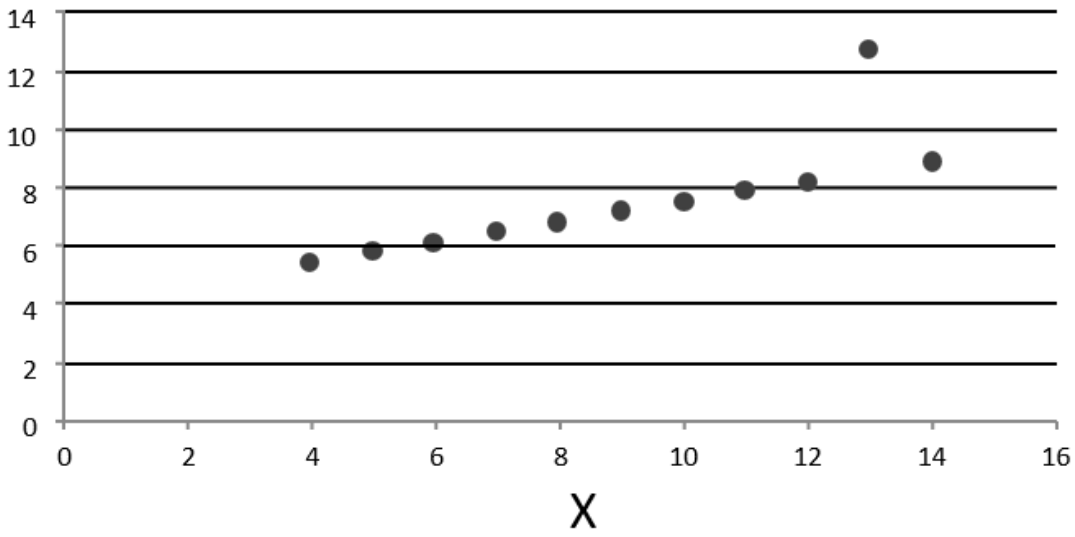


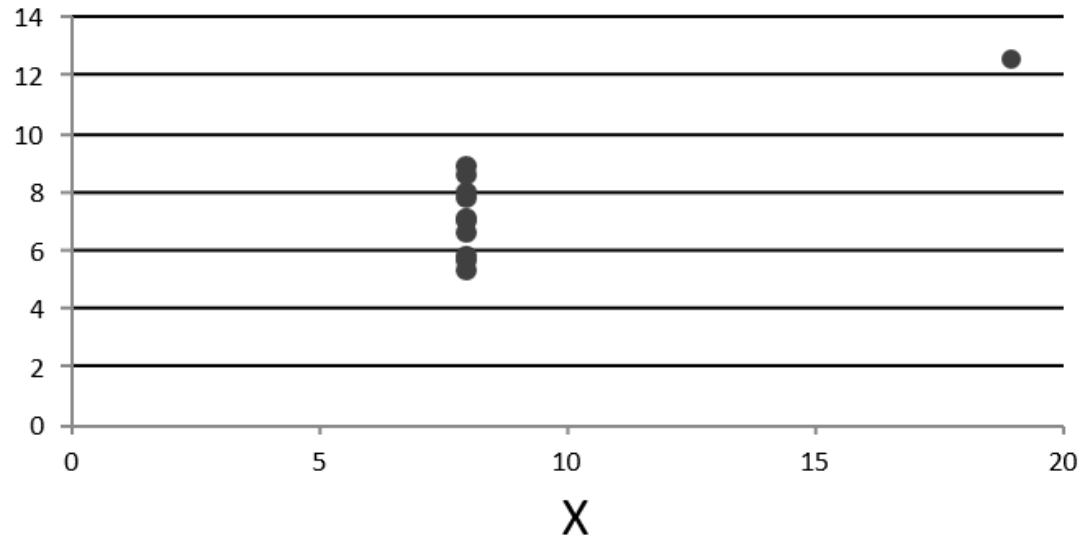
Anscombe 1973

mean X = 9.0 mean Y = 7.5 sd X = 3.317 sd Y = 2.03







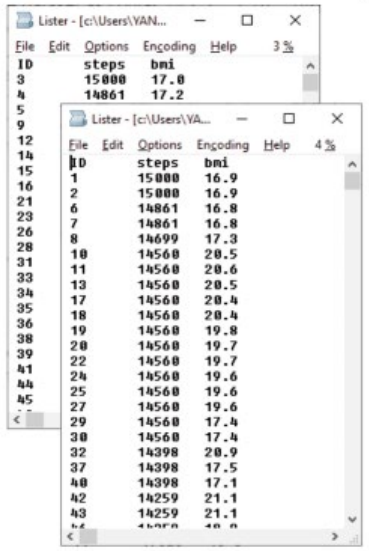


Anscombe 1973

A hypothesis is a liability

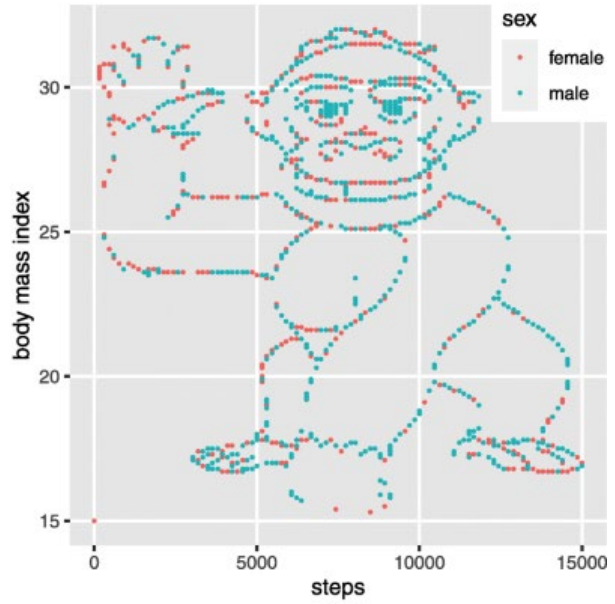
- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02133-w
- students without a specific hypothesis were almost five times more likely to discover the gorilla when analyzing this dataset

a



ID	steps	bmi
1	15000	16.9
2	15000	16.9
6	14861	16.8
7	14861	16.8
8	14699	17.3
10	14560	20.5
11	14560	20.6
13	14560	20.5
17	14560	20.4
18	14560	20.4
19	14560	19.8
20	14560	19.7
22	14560	19.7
24	14560	19.6
25	14560	19.6
27	14560	19.6
29	14560	17.4
30	14560	17.4
32	14398	20.9
37	14398	17.5
40	14398	17.1
42	14259	21.1
43	14259	21.1

b



c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Statistics

- Descriptive Statistics — gives information that describes the data in some manner

- Inferential Statistics — uses descriptive statistics to estimate population parameters

Descriptive Statistics

Descriptive Statistics

- Measures of Centre
- Measures of Spread

Descriptive Statistics

- Measures of Centre

  Central Tendency

  1. Mean — average
  2. Median — the middle value (in a sorted set of data points)
  3. Mode — most frequently occurring value

Income

UNIVERSITY OF CALGARY

Measures of Centre

Central Tendency

1. Mean — average
2. Median — the middle value (in a sorted set of data points)
3. Mode — most frequently occurring value

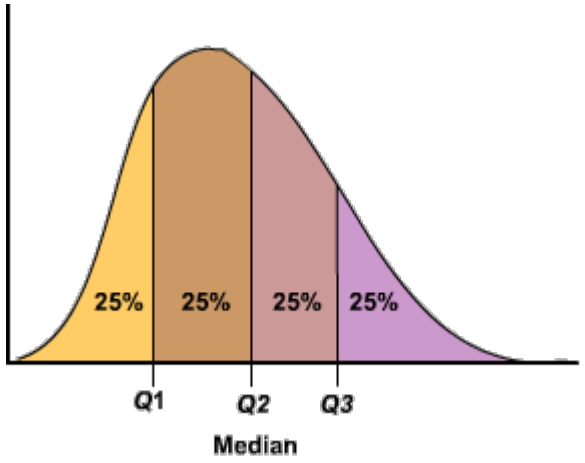
Measures of Spread

How data points are deviated from the average of a distribution

1. Variance — squared deviations of individual data points from the mean
2. Standard Deviation — square root of the variance
3. Range — difference between max and min
4. Interquartile Range (IQR) — difference between Q3 and Q1

Box Plot

Assume no outliers

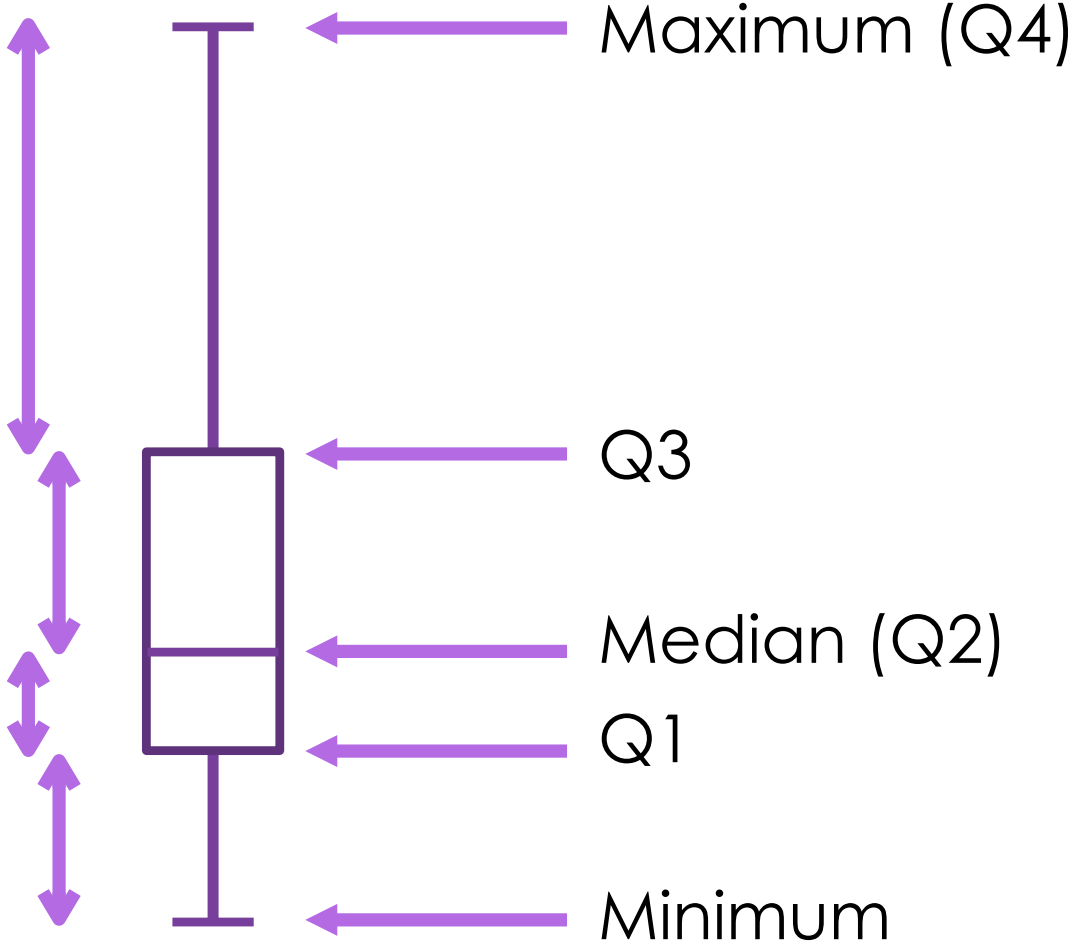


4th quartile

3rd quartile

2nd quartile

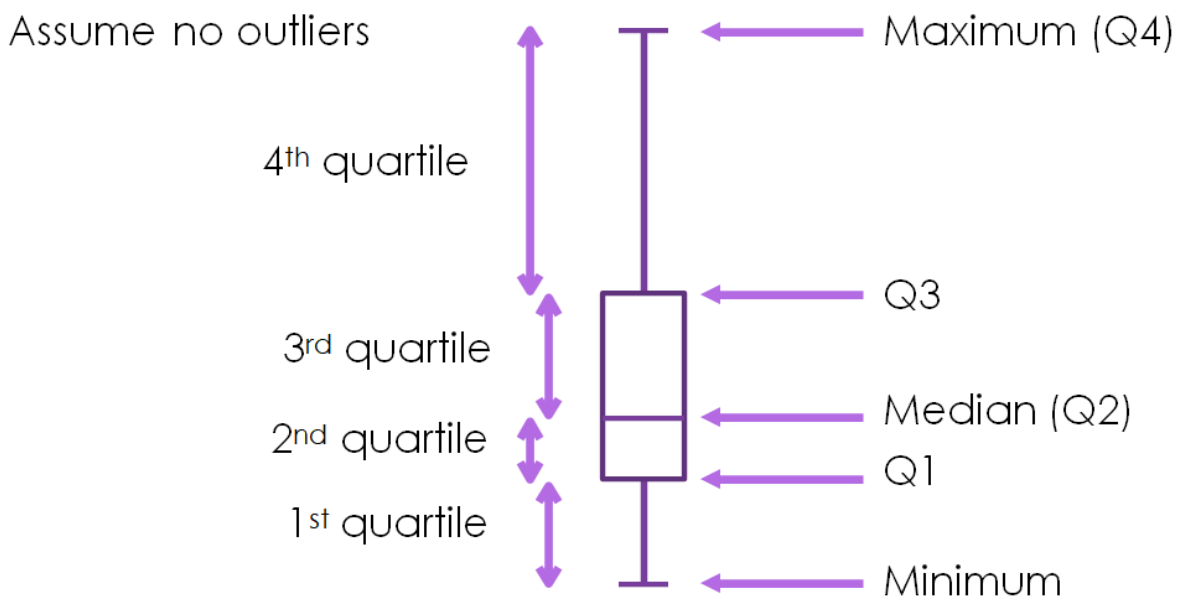
1st quartile



Measures of Spread (Box Plot)

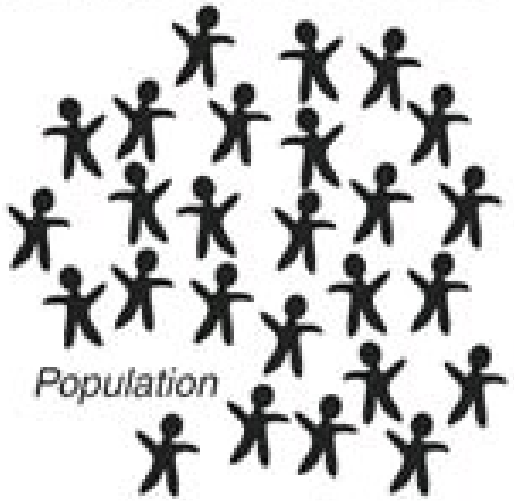
How data points are deviated from the average of a distribution

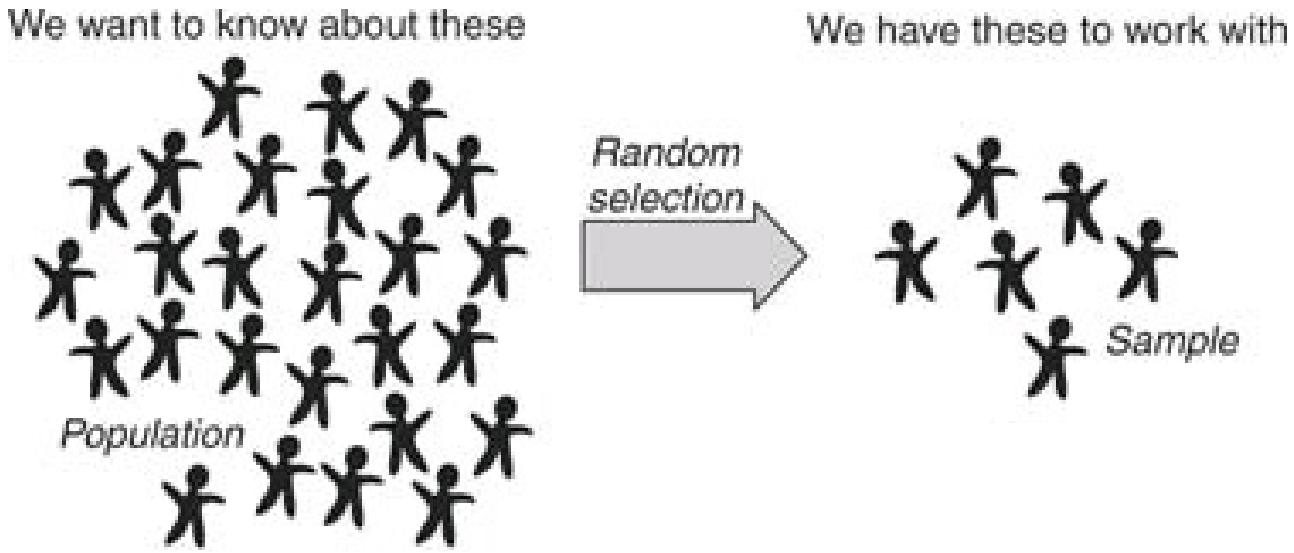
1.
2.
3. Range — difference between max and mir
4. Interquartile Range (IQR) — difference between Q3 and Q1

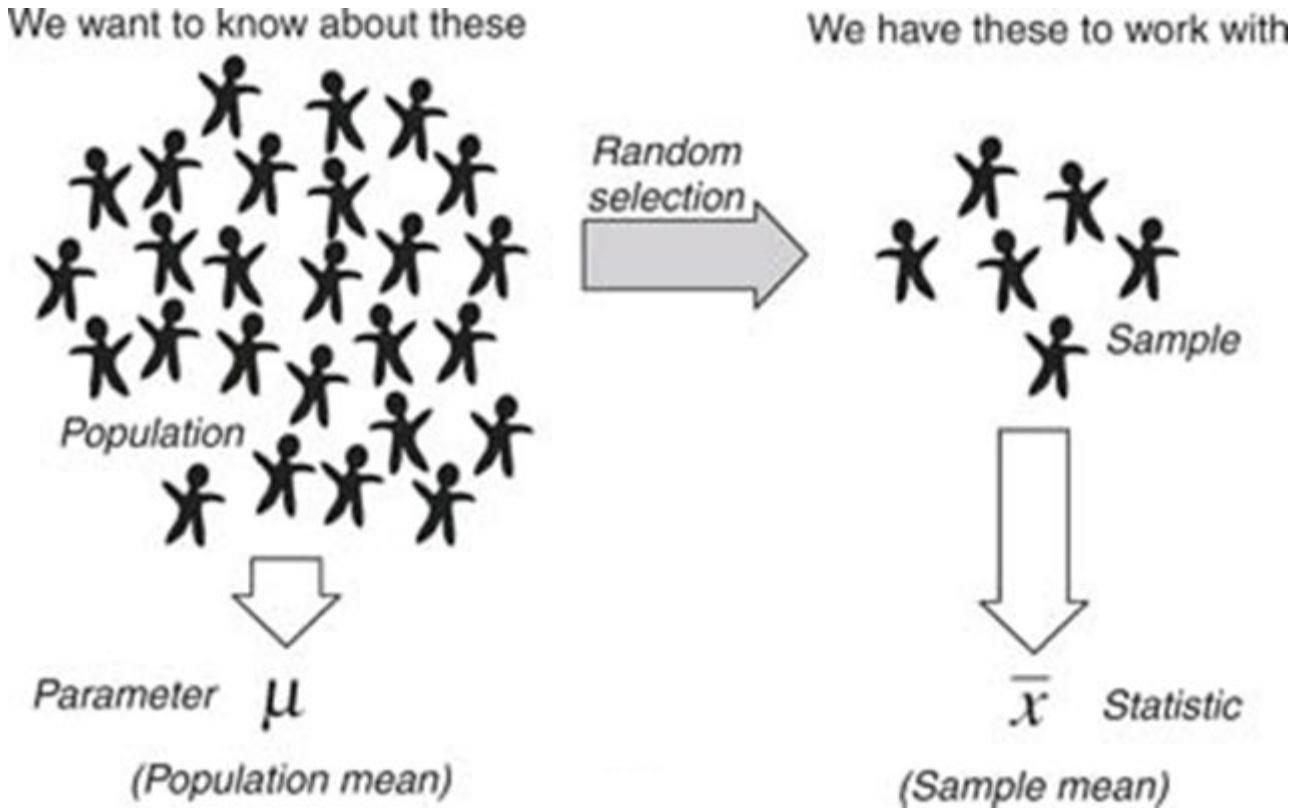


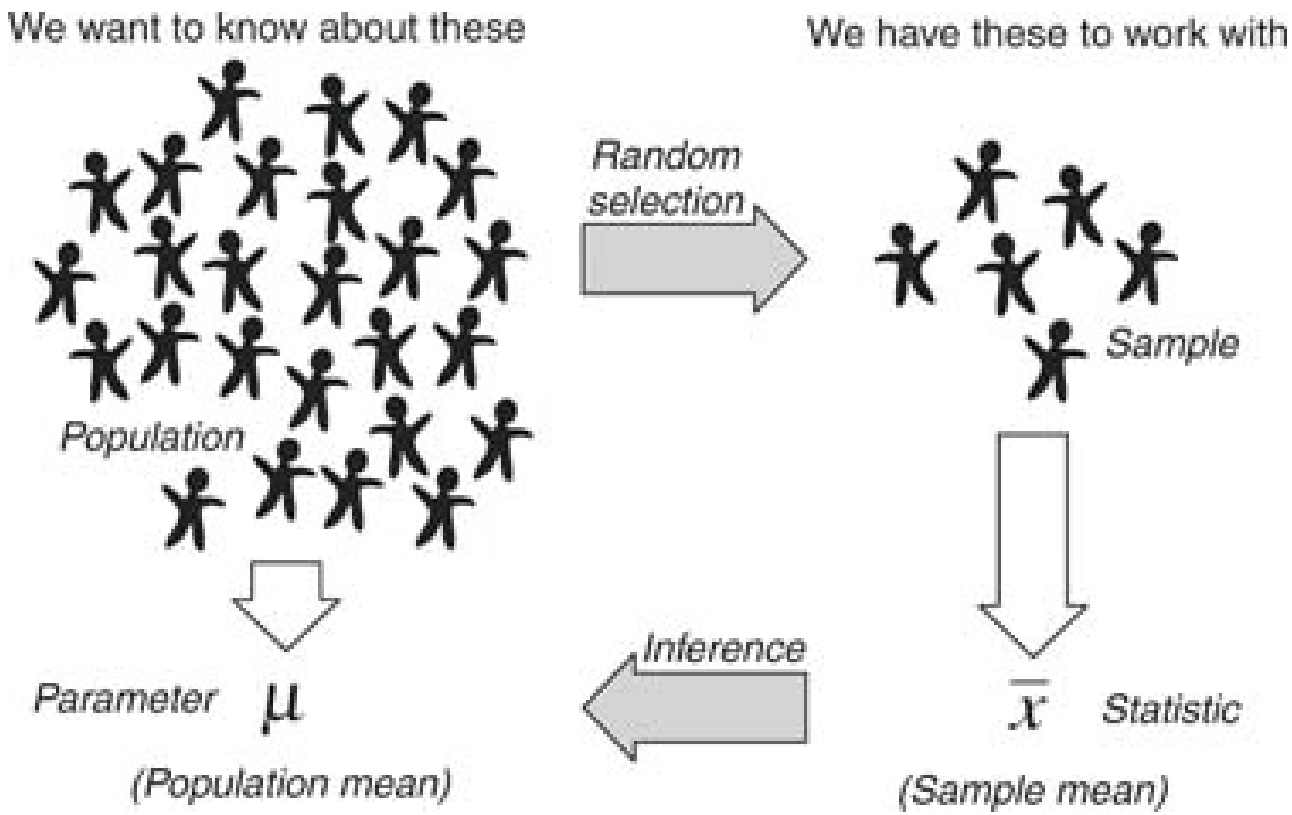
Inferential Statistics

We want to know about these

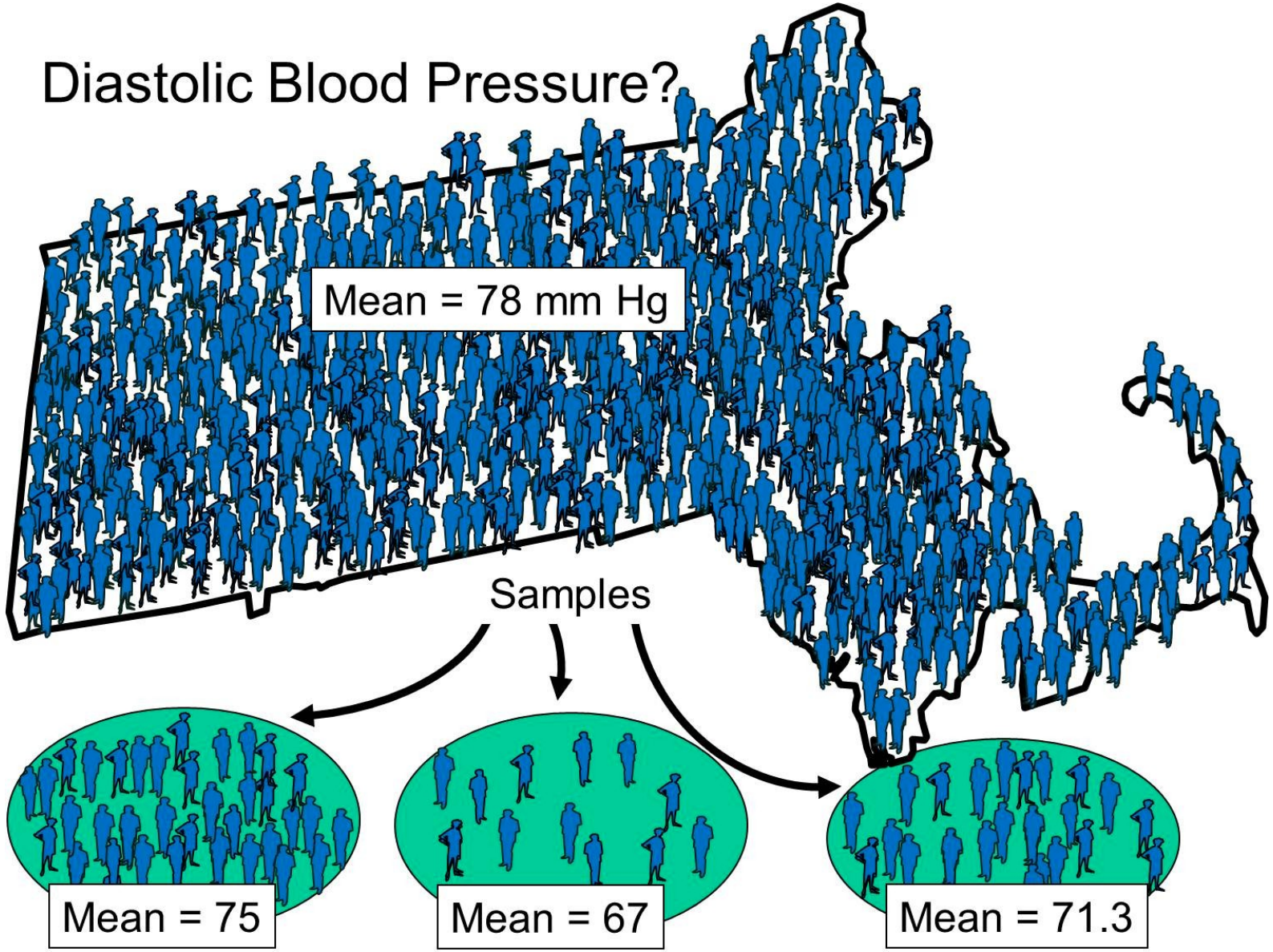


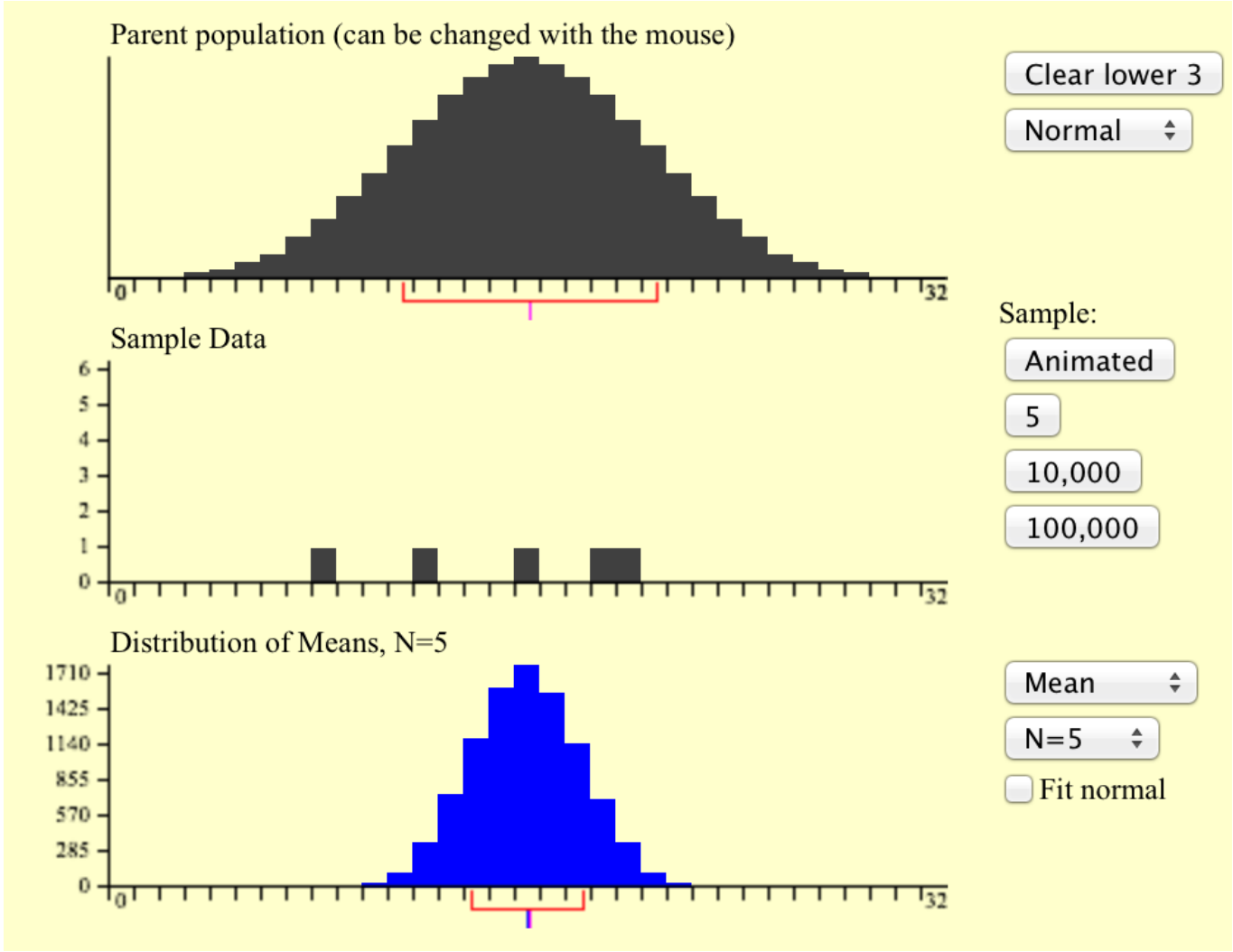






Lisa Sullivan



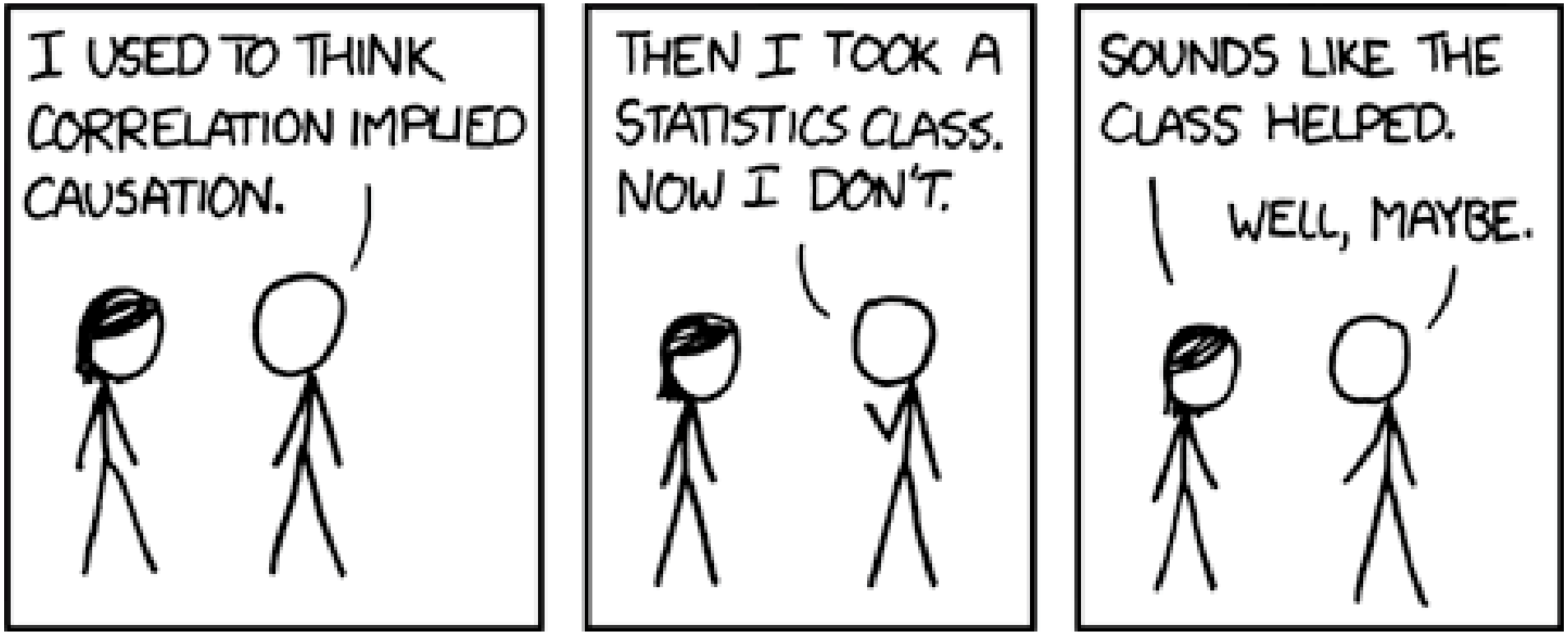


http://onlinestatbook.com/stat_sim/sampling_dist/

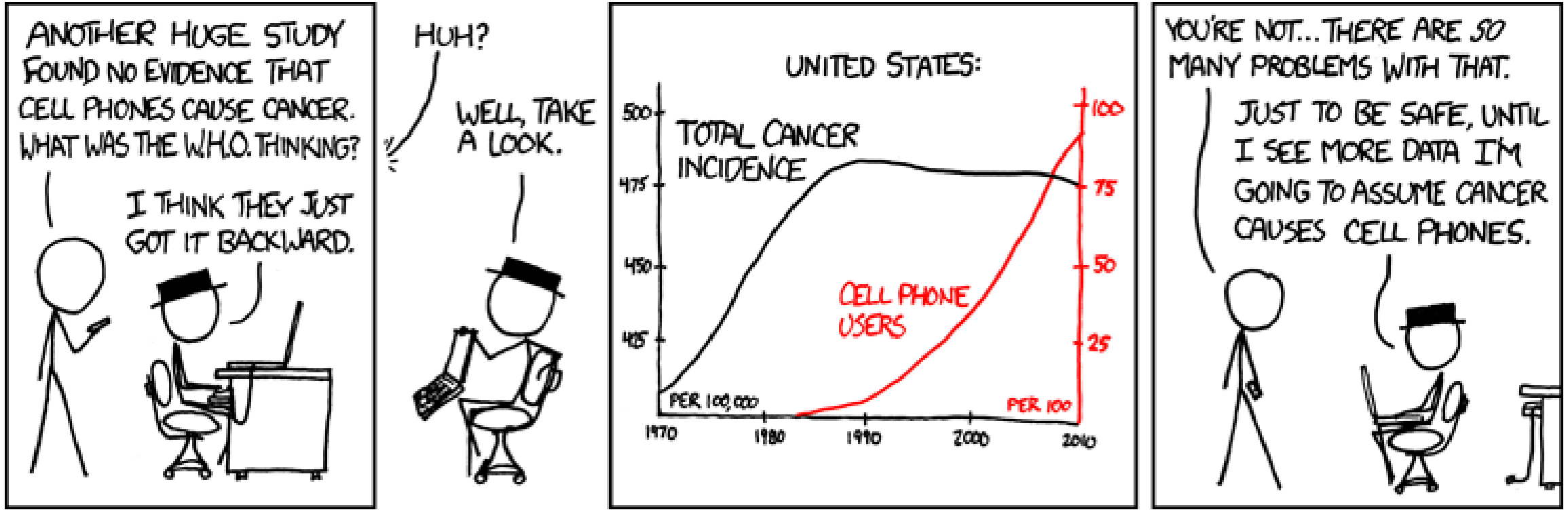
Spurious Correlations

http://tylervigen.com

Correlation?



Causation?



Onward to …
Qualitative Analysis

Jonathan Hudson
jwhudson@ucalgary.ca
https://pages.cpsc.ucalgary.ca/~jwhudson/

UNIVERSITY OF CALGARY