Lecture 03: BI Basic - Statistics

UNIVERSITY of GREENWICH

Alliance with FPT Education

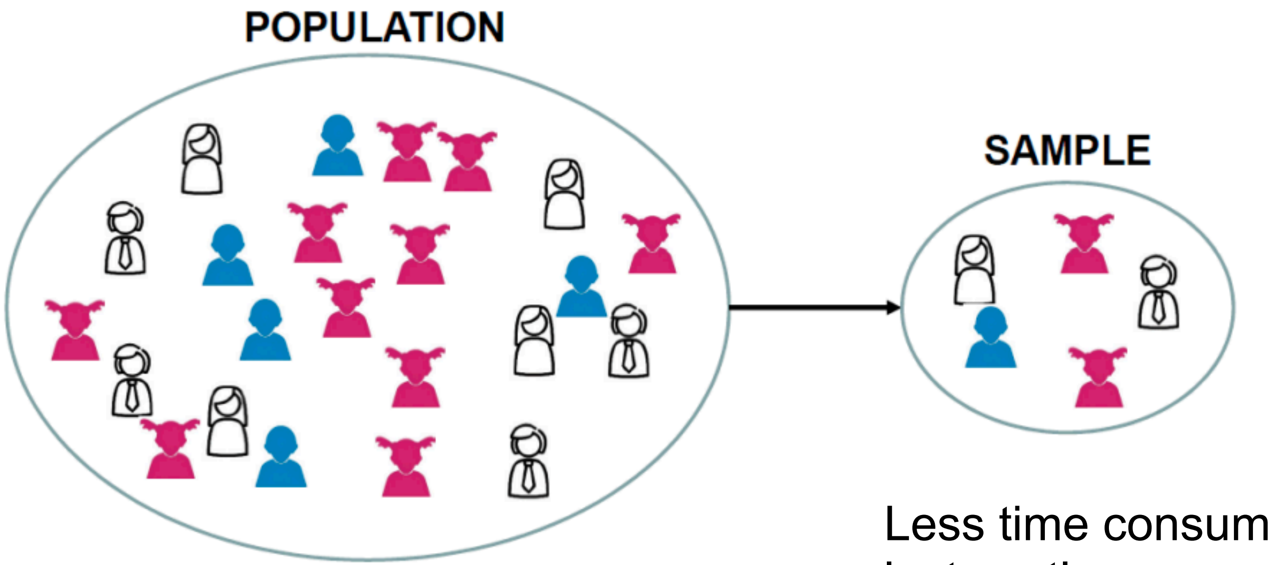
Pearson BTEC

TABLE OF CONTENTS

- Population & Sample
- Types of data
- Levels of measurements
- One/Two variables
- Central techniques
- Variability techniques

Population vs sample

- Population: Collection of all items of interest, denoted N
- Sample: A subset of population, denoted n



Less time consuming,
lest costly

Hard to define, hard to
observe

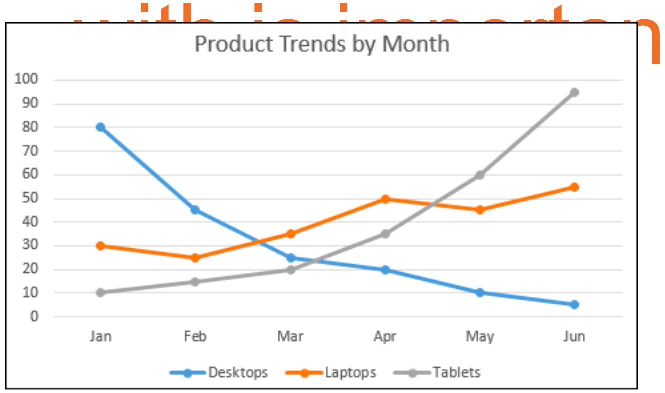
Choosing a sample

- A sample must be both random and representative for an insight to be precise
- Randomness: random sample is collected when each member of the sample is chosen from the population strictly by chance.
- Representativeness: A representative sample is a subset of the population that accurately reflects the members of the entire population.

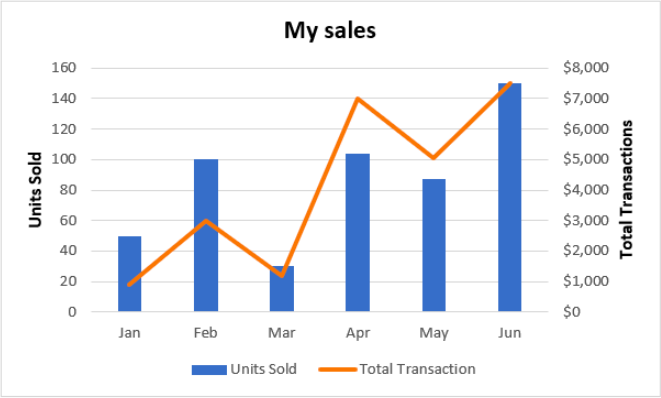
- Example: Doing a survey on students of FPT university by going to the canteen and ask students in the canteen. Is it random or representative?

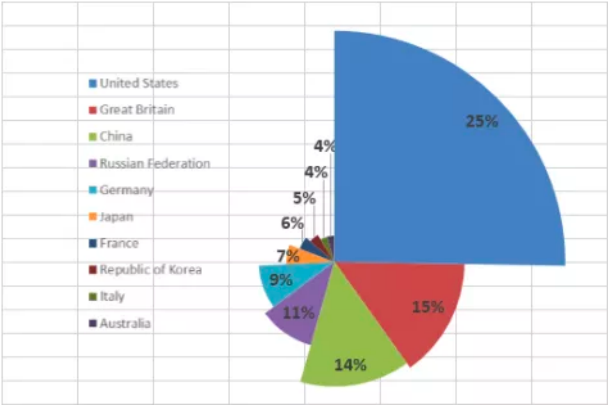
Types of data

- Before we can start analyzing we have to get acquainted with the types of variables
- Different types of variables require different types of statistical and visualization approaches.
- Therefore to be able to classify the data you are working with is important.



TechOnTheNet.com





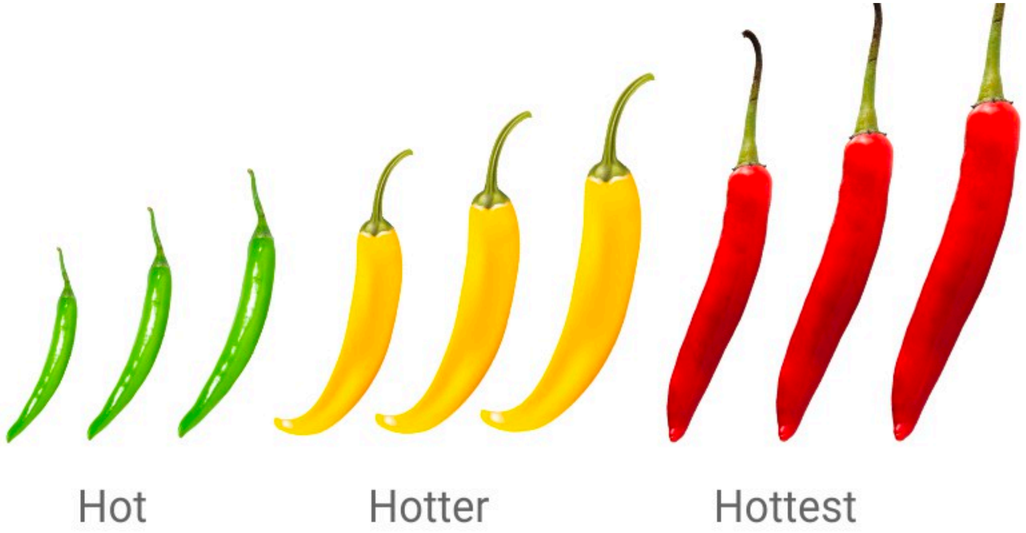
Types of data

- Categorical data:
  - Types of cars
  - Yes/No answers

- Numerical data:
  - Discrete: countable. For example: number of children, grades of assignments
  - Continuous: infinite, impossible to count. For example: weight / height of a person.

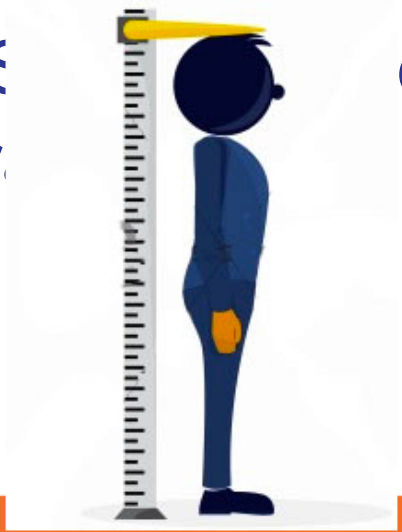
Level of measurements

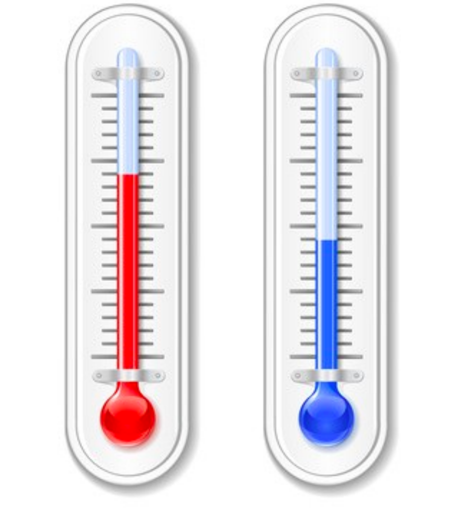
- Qualitative data: nominal or ordinal
  - Nominal data: unordered categories. For example: type of cars, colors of hair
  - Ordinal data: can be ordered. For example: categories of chili, satisfaction of customer



Level of measurements

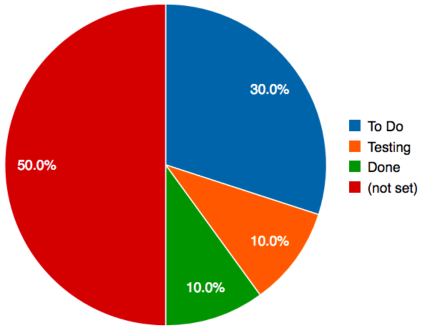
- Quantitative data: interval and ratio both are numbers but ratio has “true” zero
  - Interval data: data is like ordinal except we can say the intervals between each value are equally split. Example Celsius temperature
  - Ratio: S erval b ero. Example: Kevin temper





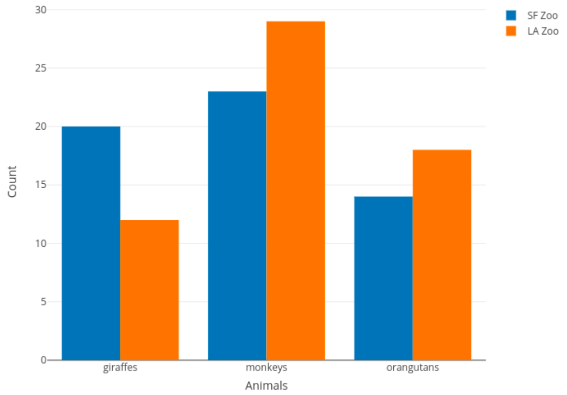
Graphical representation

- The way data should be represented in a graph or chart depends on the level of measurement



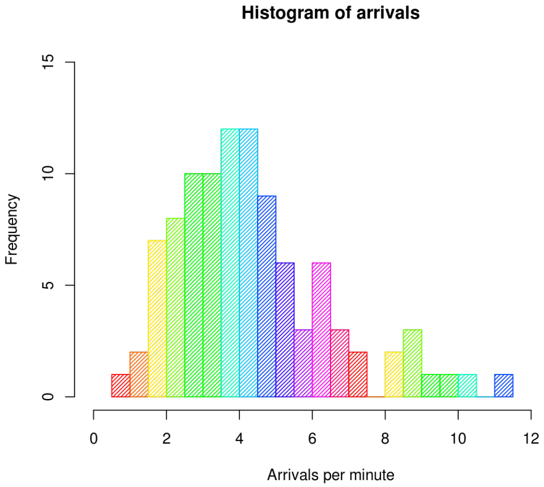
Pie Chart

Nominal



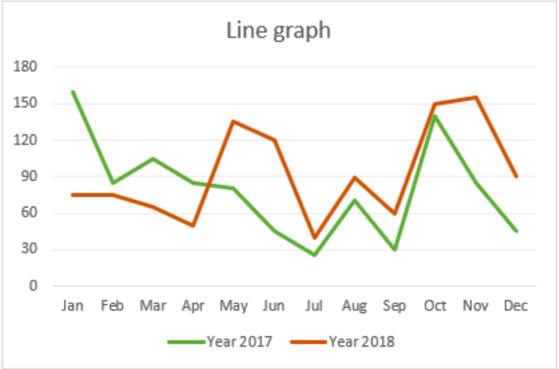
Bar Chart

Nominal, Ordinal, Interval, Ratio



Histogram

Interval, Ratio

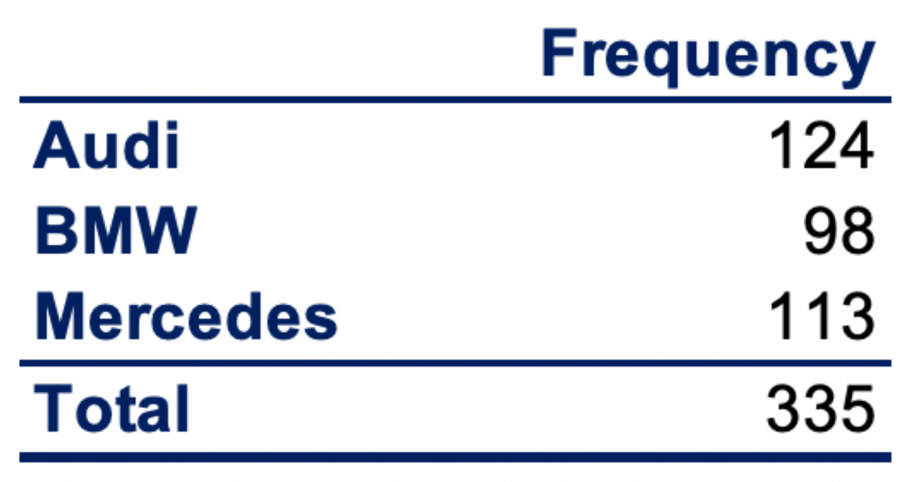


Line chart

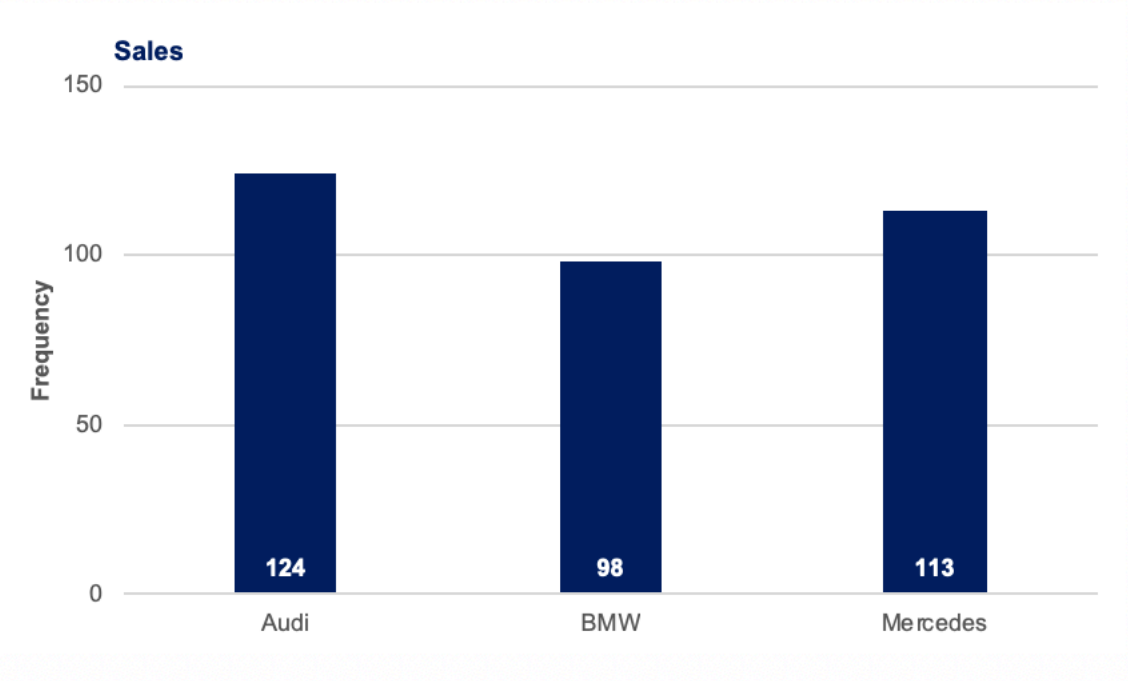
Interval, Ratio

One variable techniques

- Frequency distribution table
- Bar charts



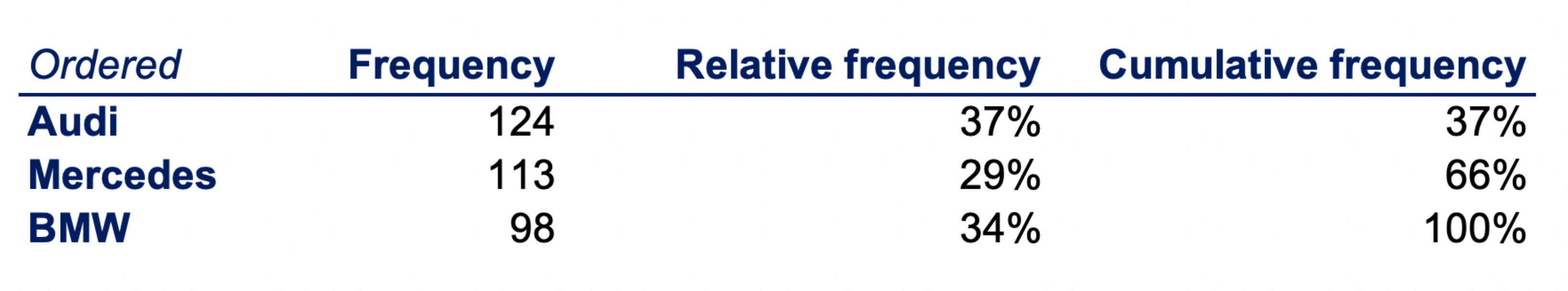
	Frequency
Audi	124
BMW	98
Mercedes	113
Total	335



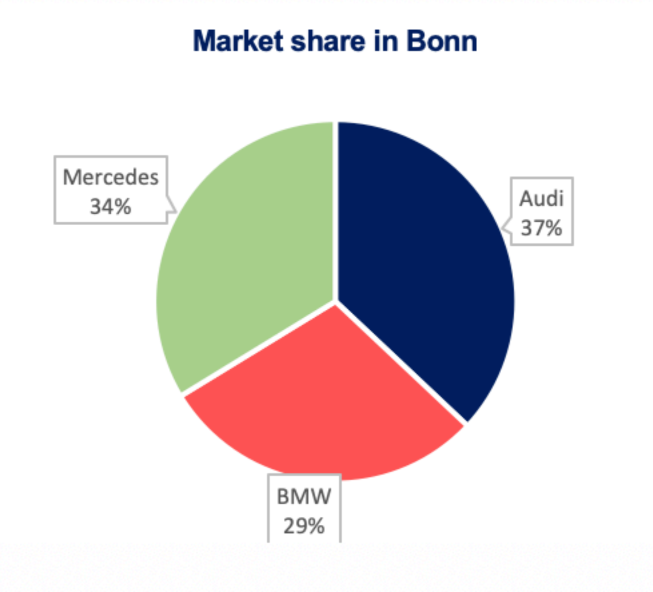
Type of car: variable
Audi, BMW, ..: value of variable

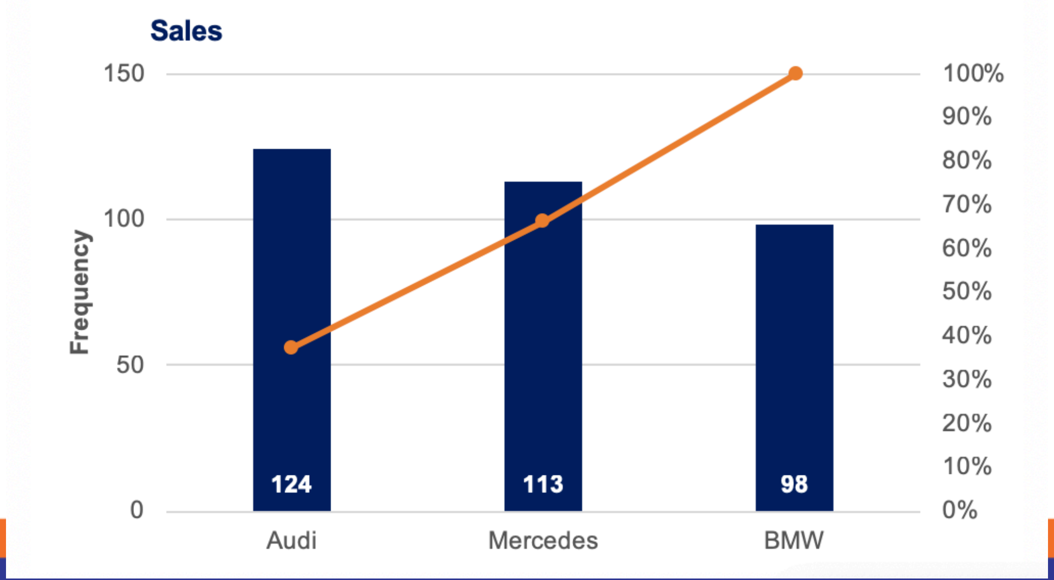
One variable techniques

- Pie charts
- Pareto charts



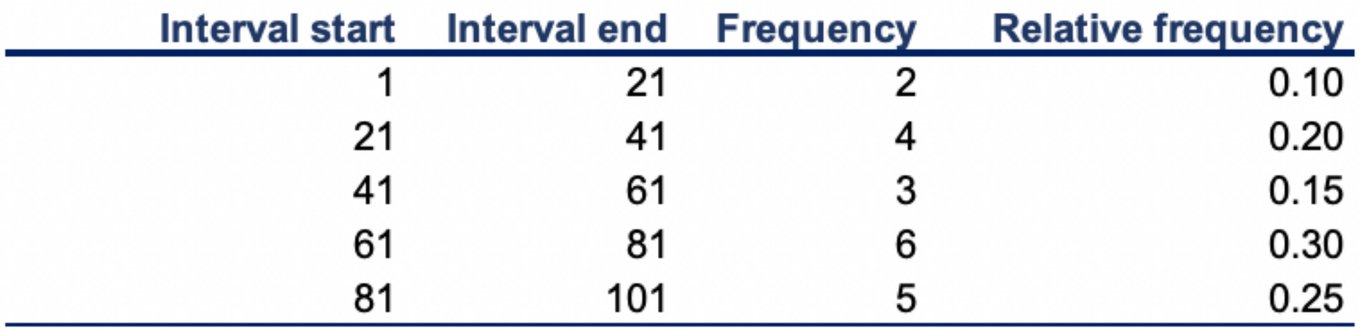
Ordered	**Frequency**	**Relative frequency**	**Cumulative frequency**
Audi	124	37%	37%
Mercedes	113	29%	66%
BMW	98	34%	100%



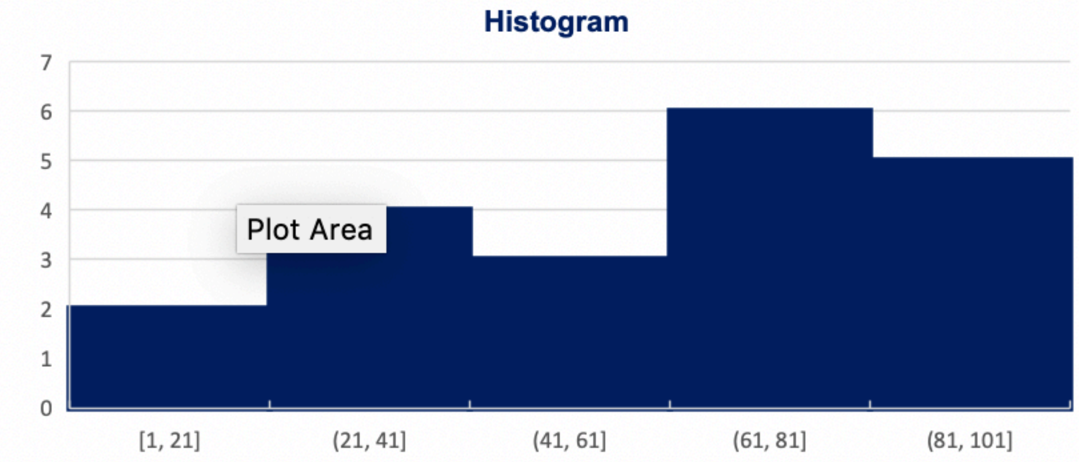


One variable techniques

- Frequency distribution table (may need to set interval)
- Histogram

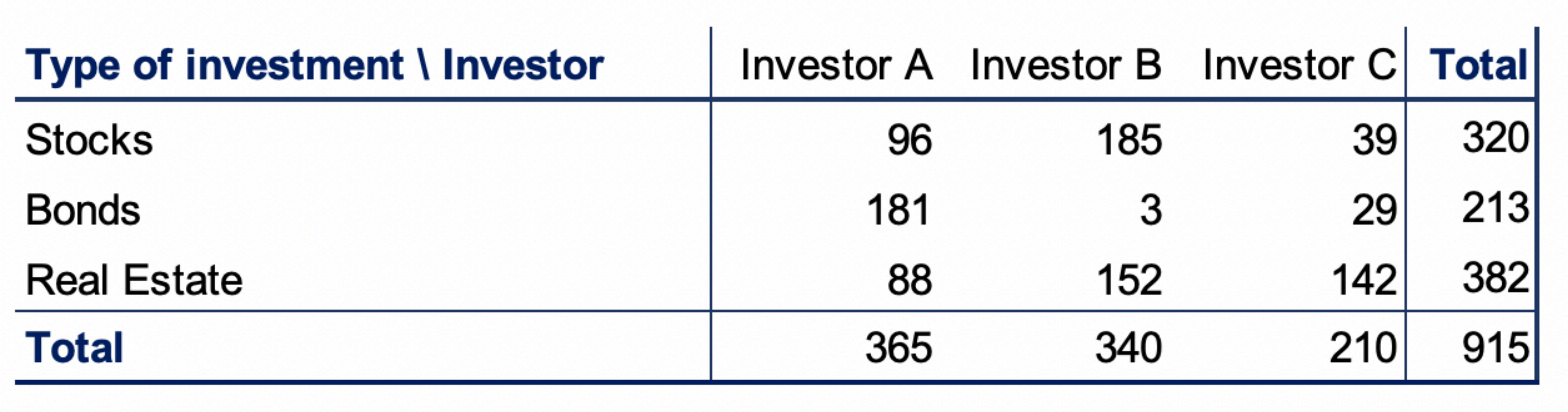


Interval start	Interval end	Frequency	Relative frequency
1	21	2	0.10
21	41	4	0.20
41	61	3	0.15
61	81	6	0.30
81	101	5	0.25



Two variables techniques

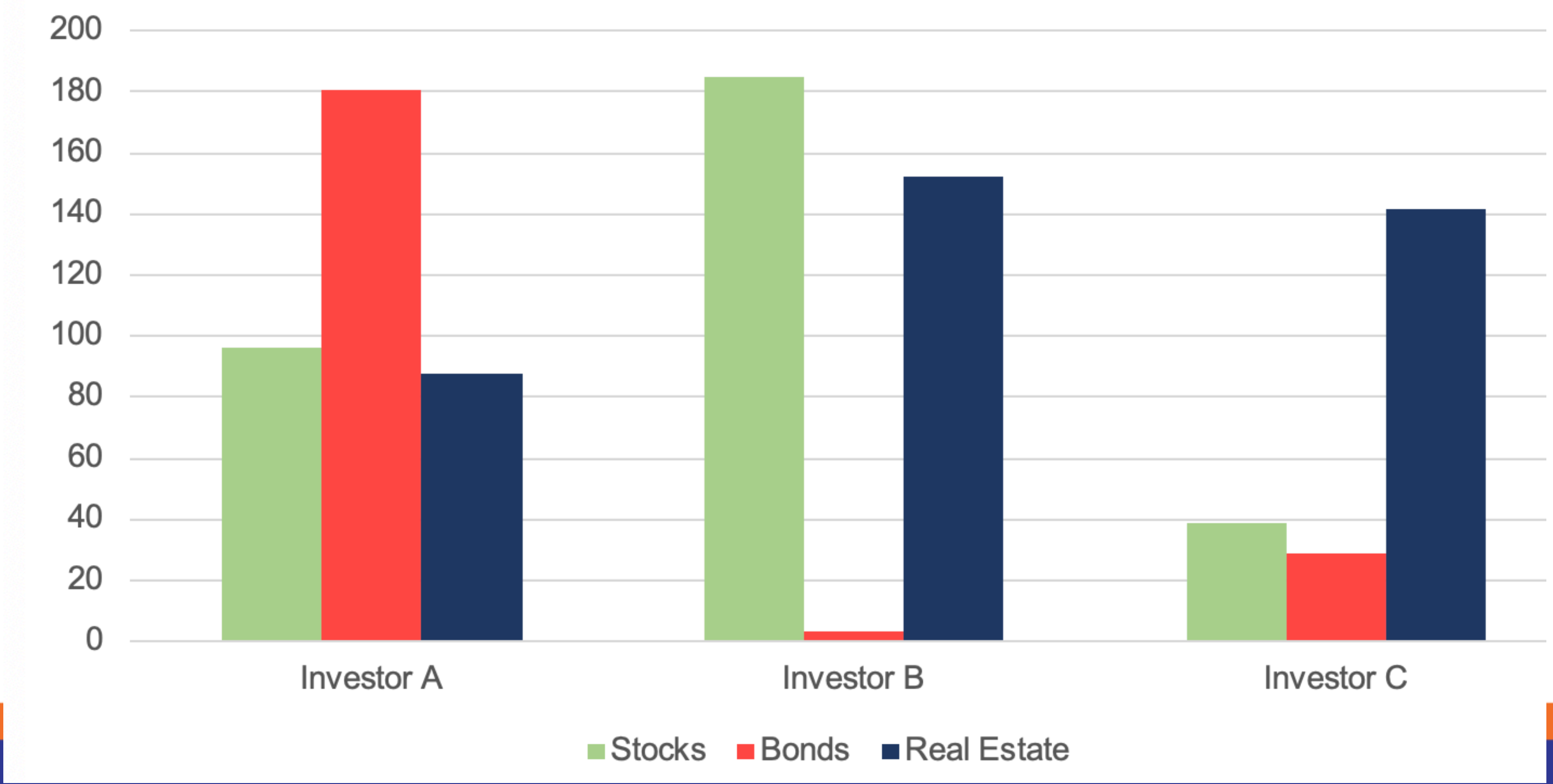
- Cross tables and side-by-side bar charts



Type of investment \ Investor	Investor A	Investor B	Investor C	Total
Stocks	96	185	39	320
Bonds	181	3	29	213
Real Estate	88	152	142	382
Total	365	340	210	915

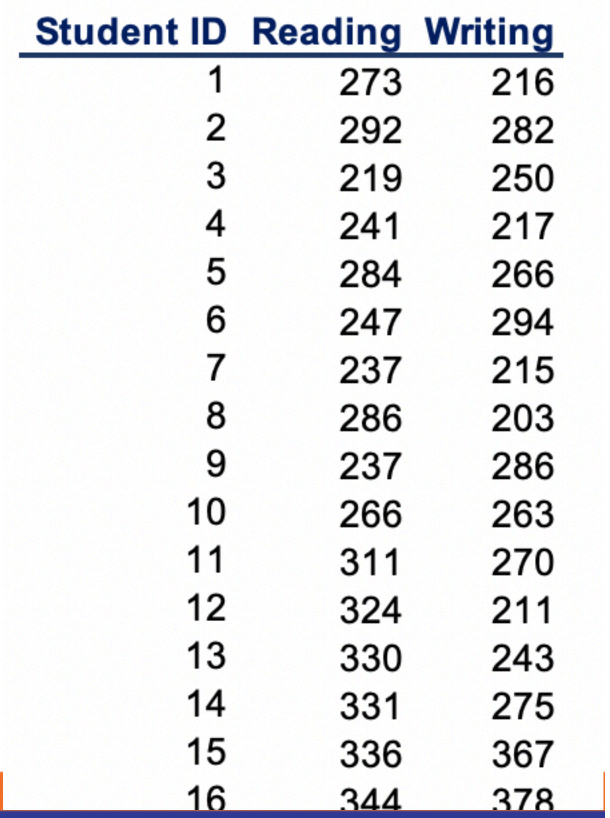
Two variables techniques

- Cross tables and side-by-side bar charts

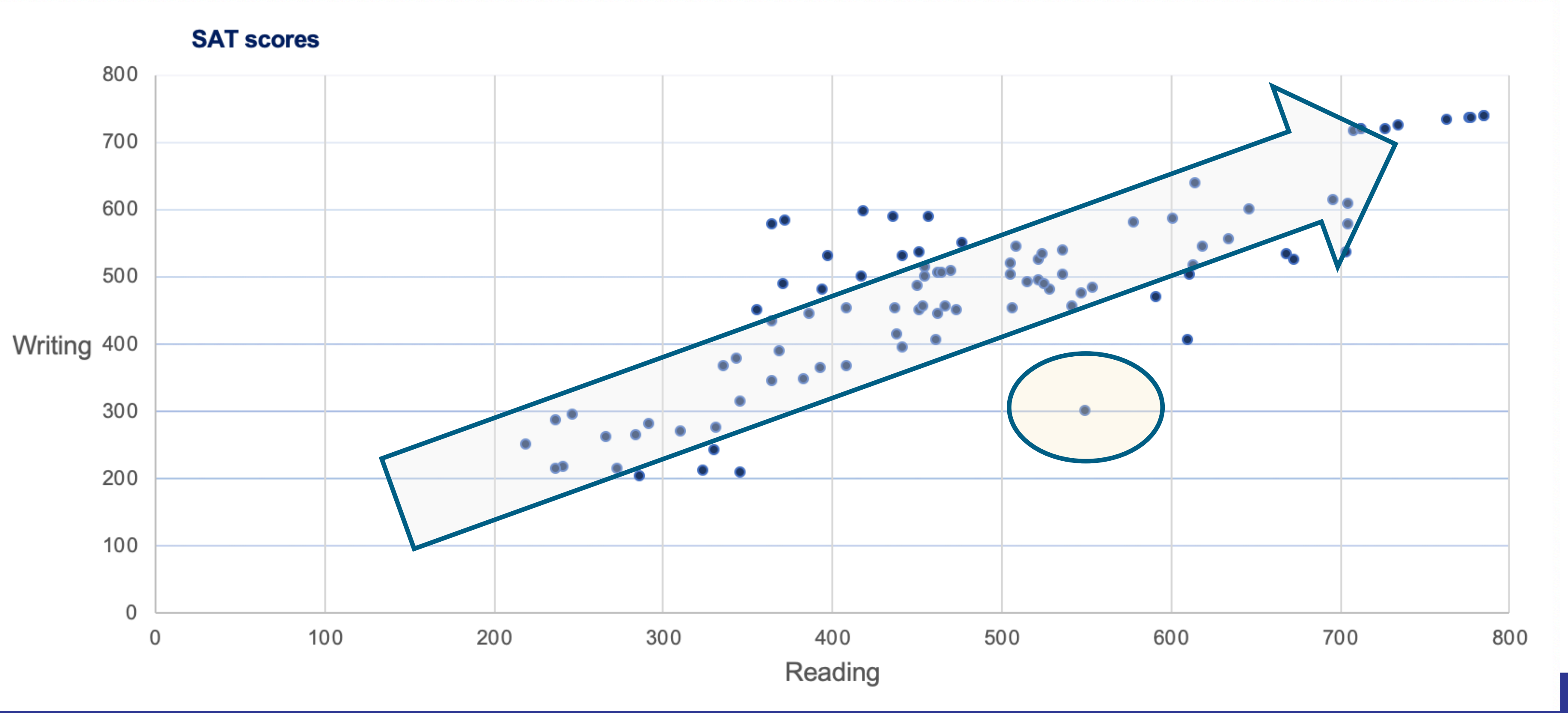


Two variables techniques

- Scatter plot



Student ID	Reading	Writing
1	273	216
2	292	282
3	219	250
4	241	217
5	284	266
6	247	294
7	237	215
8	286	203
9	237	286
10	266	263
11	311	270
12	324	211
13	330	243
14	331	275
15	336	367
16	344	378

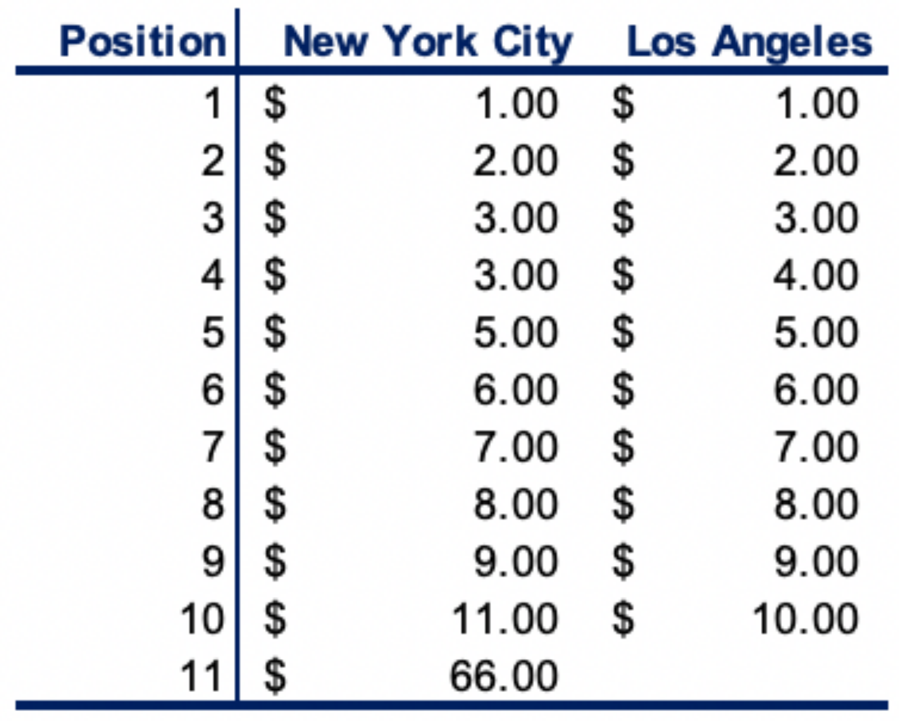


Central techniques

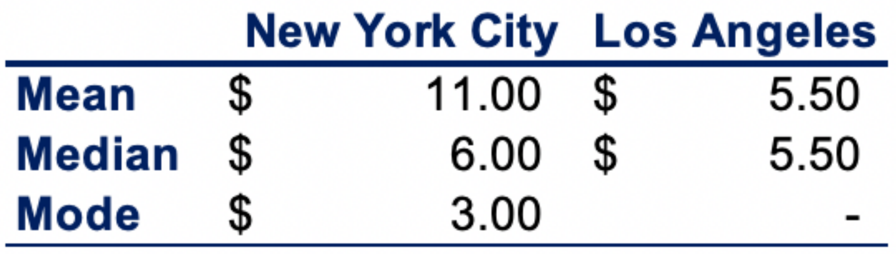
- Mean: the most popular techniques for central tendency

$$\text{mean} = \frac{\text{sum of data}}{\#\text{ of data points}}$$

- Median: The median is the middle point i a dataset.
- Mode: is the most commonly occurring data point in a dataset.



Position	New York City		Los Angeles	
1	$	1.00	$	1.00
2	$	2.00	$	2.00
3	$	3.00	$	3.00
4	$	3.00	$	4.00
5	$	5.00	$	5.00
6	$	6.00	$	6.00
7	$	7.00	$	7.00
8	$	8.00	$	8.00
9	$	9.00	$	9.00
10	$	11.00	$	10.00
11	$	66.00		

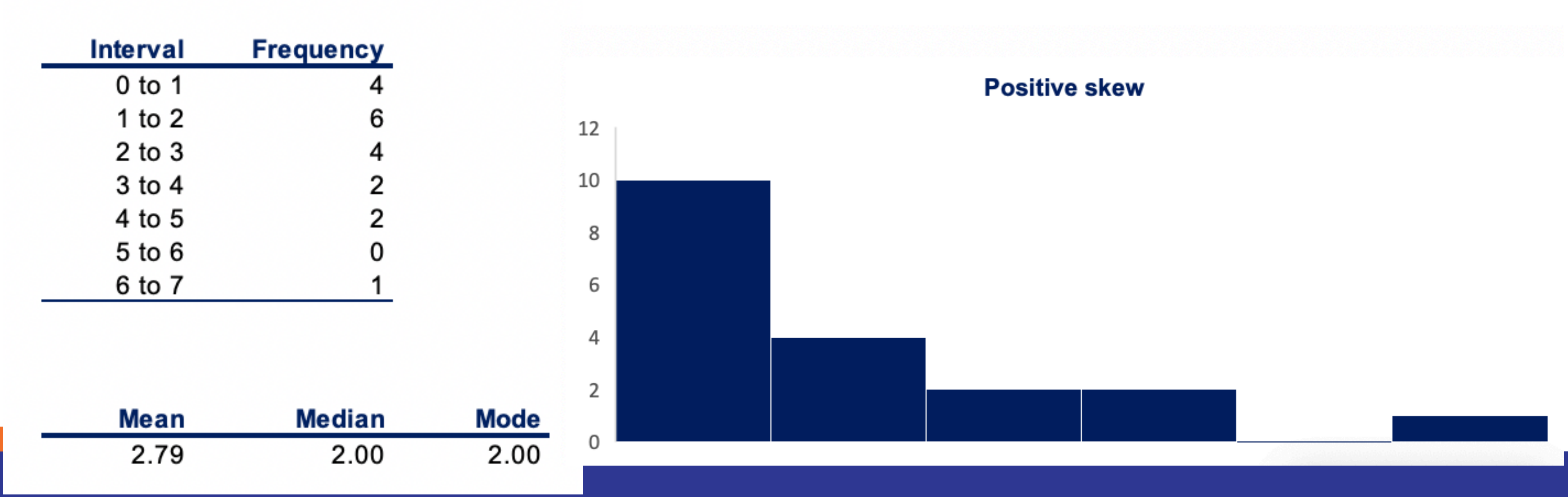


	New York City		Los Angeles	
Mean	$	11.00	$	5.50
Median	$	6.00	$	5.50
Mode	$	3.00		-

Mean (NY) is too high? Why?
Even Median (NY) is still expensive, it it true?
What is the most common price?

Central techniques

- Skewness: indicate whether data is concentrated on one side

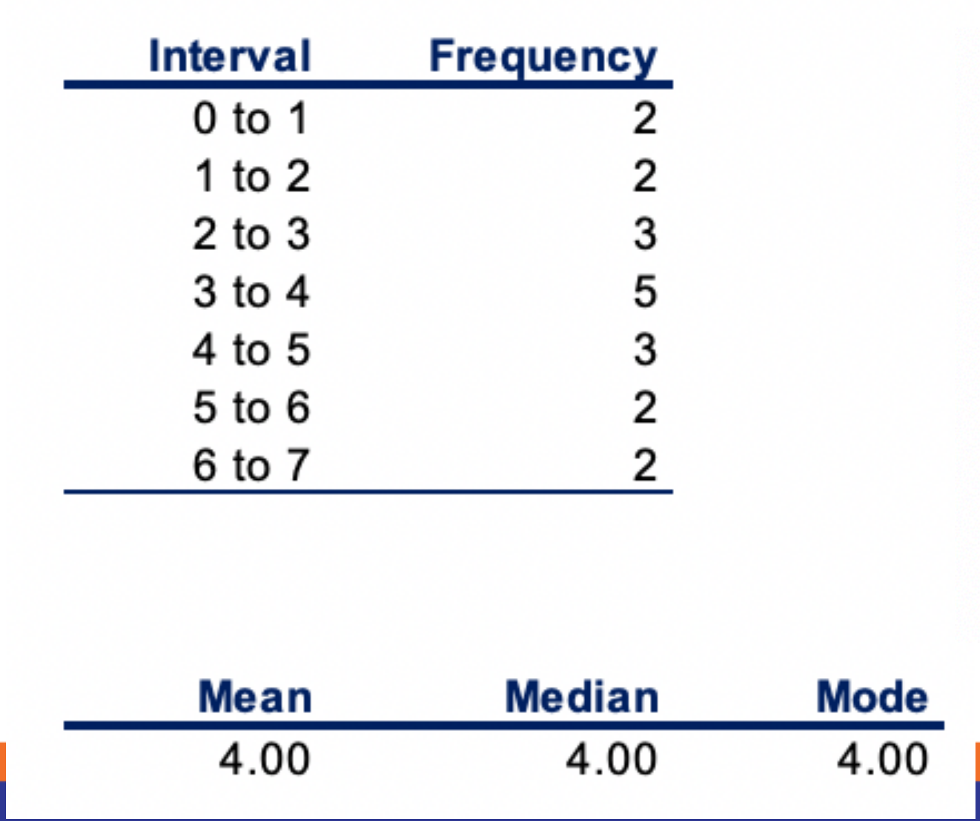


Interval	Frequency
0 to 1	4
1 to 2	6
2 to 3	4
3 to 4	2
4 to 5	2
5 to 6	0
6 to 7	1

Mean	Median	Mode
2.79	2.00	2.00

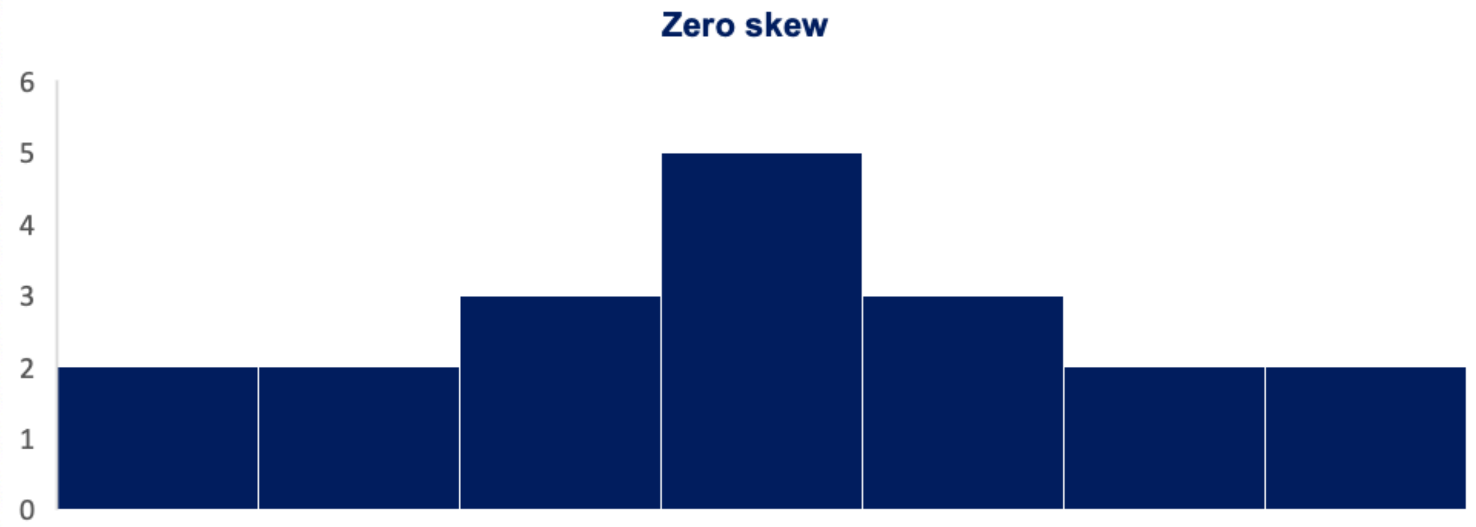
Central techniques

- Zero skewness: mean = median: the distribution is



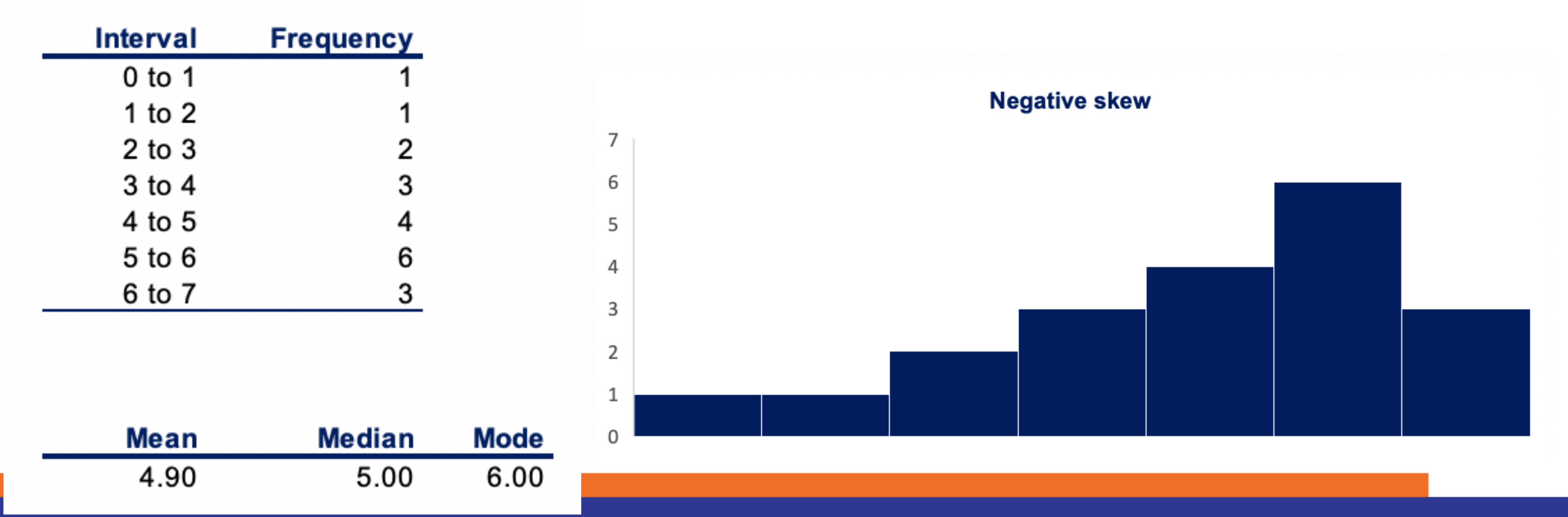
Interval	Frequency
0 to 1	2
1 to 2	2
2 to 3	3
3 to 4	5
4 to 5	3
5 to 6	2
6 to 7	2

Mean	Median	Mode
4.00	4.00	4.00



Central techniques

- Negative skewness: mean < median: outliers are on the



Interval	Frequency
0 to 1	1
1 to 2	1
2 to 3	2
3 to 4	3
4 to 5	4
5 to 6	6
6 to 7	3

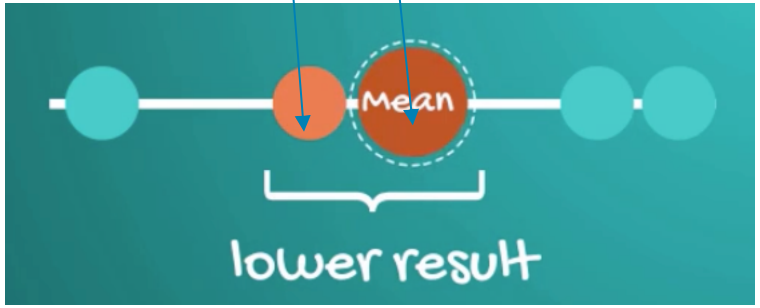
Mean	Median	Mode
4.90	5.00	6.00

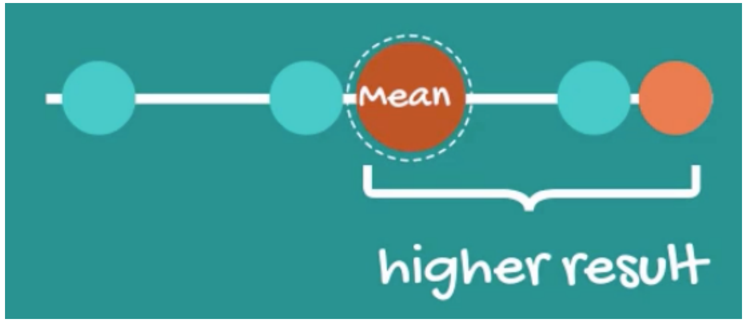
Variability techniques

- Variance: measures the dispersion of a set of data points around their mean value population variance

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

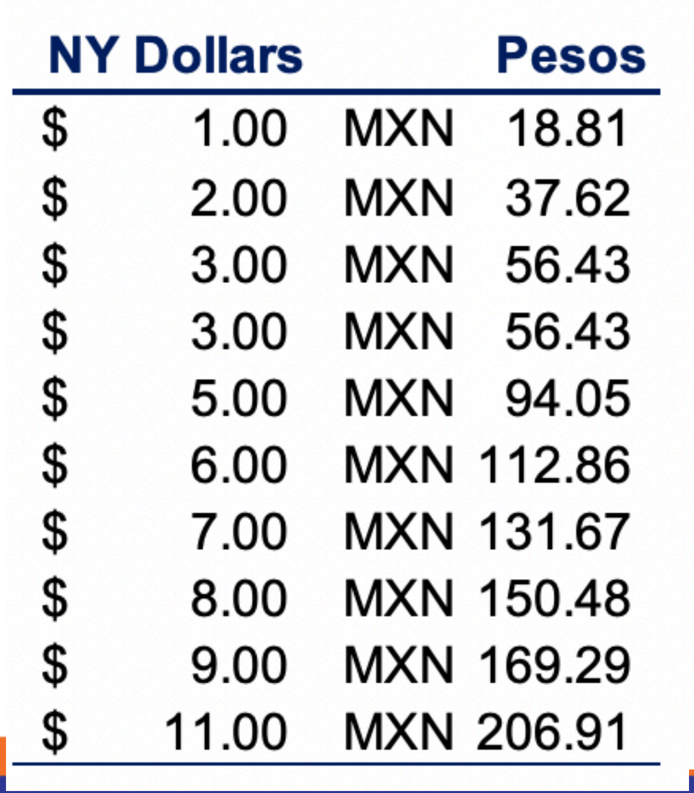




- Why square: make sure distance is positive and amplify the differences

Variability techniques

- Standard deviation: more meaningful than variance, easy to observe



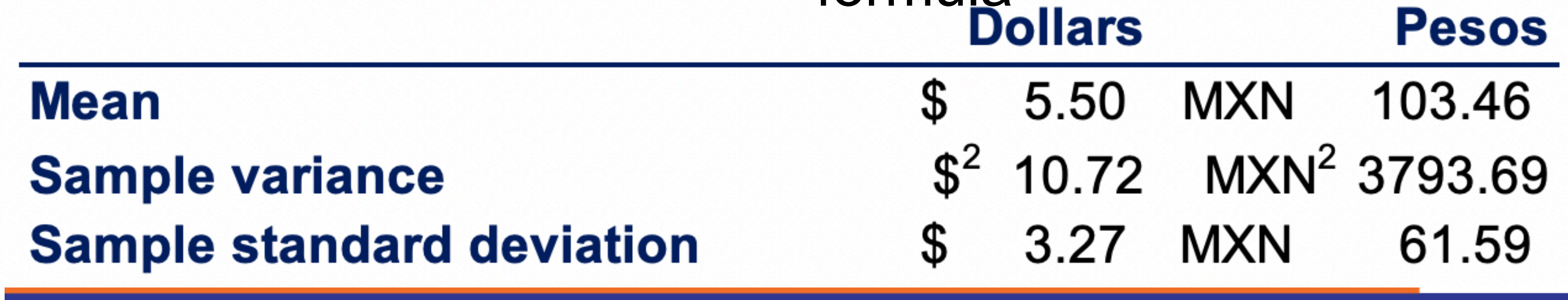
NY Dollars			Pesos
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

$$\sigma = \sqrt{\frac{\Sigma\,(x - u)^2}{N}}$$

Population formula

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

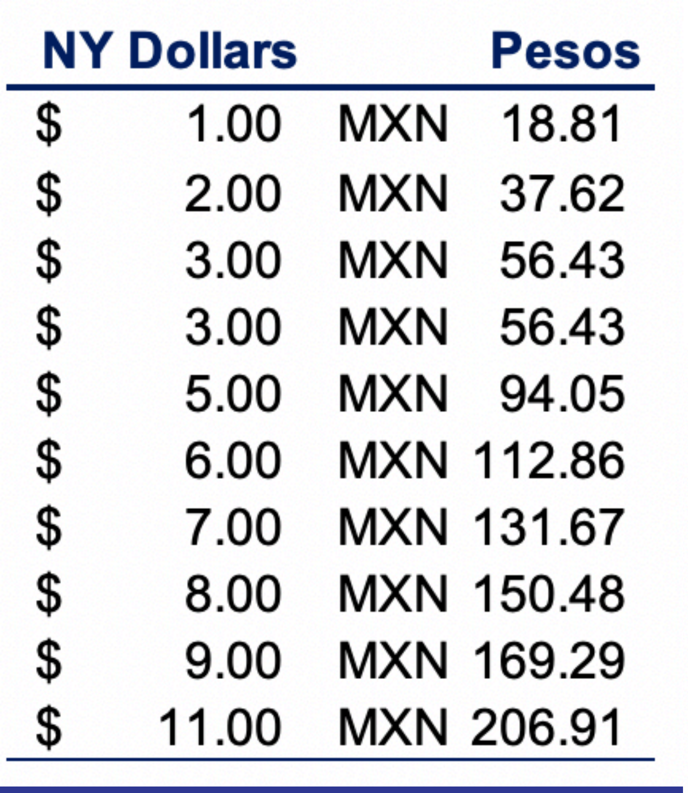
Sample formula



		Dollars		Pesos
Mean	$	5.50	MXN	103.46
Sample variance	$\$^2$	10.72	MXN^2	3793.69
Sample standard deviation	$	3.27	MXN	61.59

Variability techniques

- The coefficient of variation (CV) measures the dispersion of data points in a data series around the mean: no unit, sometimes using %



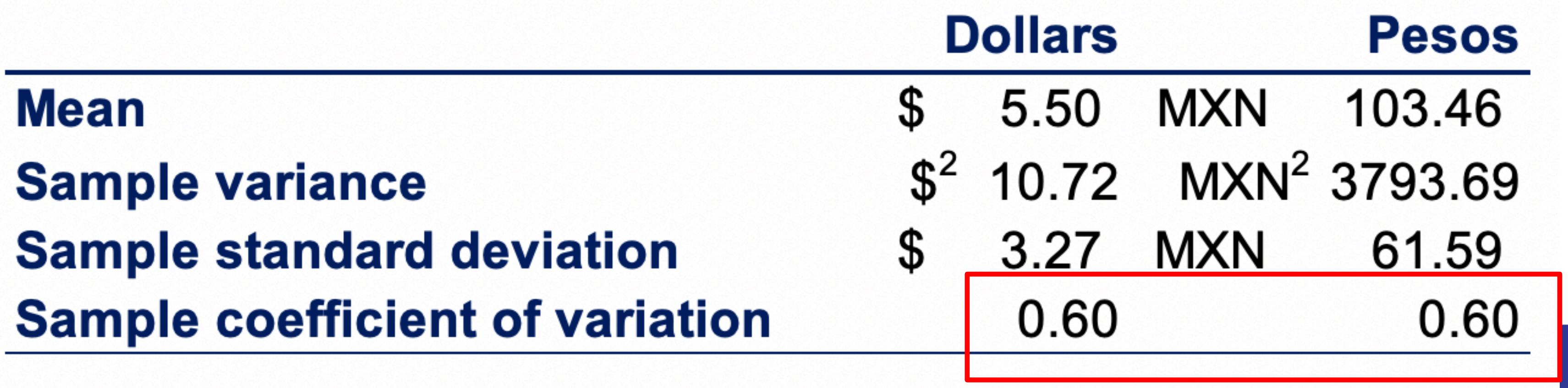
NY Dollars		Pesos	
$	1.00	MXN	18.81
$	2.00	MXN	37.62
$	3.00	MXN	56.43
$	3.00	MXN	56.43
$	5.00	MXN	94.05
$	6.00	MXN	112.86
$	7.00	MXN	131.67
$	8.00	MXN	150.48
$	9.00	MXN	169.29
$	11.00	MXN	206.91

$$CV = \frac{\sigma}{\mu}$$

Population formula

$$CV = \frac{s}{\bar{x}}$$

Sample



	Dollars		Pesos	
Mean	$	5.50	MXN	103.46
Sample variance	$^2	10.72	MXN^2	3793.69
Sample standard deviation	$	3.27	MXN	61.59
Sample coefficient of variation		0.60		0.60