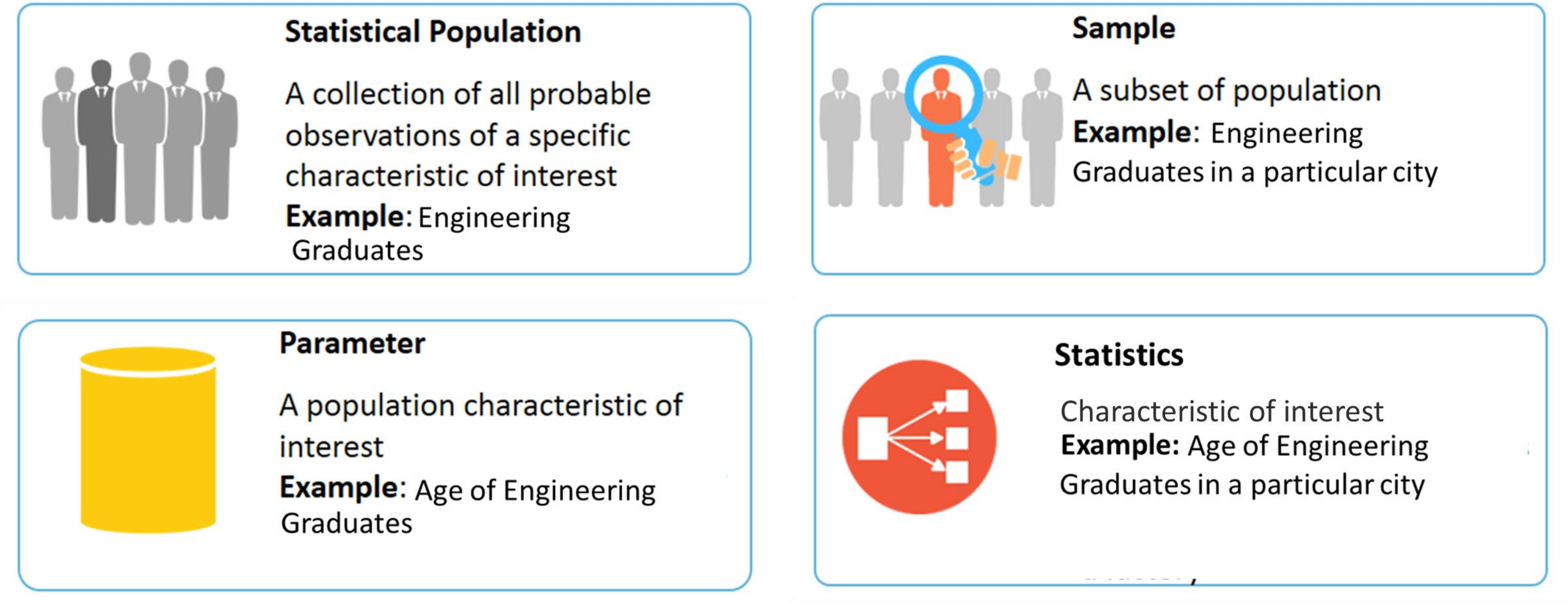
Population and Sample

Population and Sample

- A population is the set of all measurements of interest to the study.
- A sample is a selected subset of measurements of a population to represent the population.



Statistical Population

A collection of all probable observations of a specific characteristic of interest

Example: Engineering Graduates

Sample

A subset of population

Example: Engineering Graduates in a particular city

Parameter

A population characteristic of interest

Example: Age of Engineering Graduates

Statistics

Characteristic of interest

Example: Age of Engineering Graduates in a particular city

Population and Sample

- **Market Share of a Product**
  - For example you need to estimate the market share of a detergent product specifically, say, Tide
  - Population here is the entire population
  - Sample is the a set of Supermarkets/shops
  - Market Share is calculated on the sample, not the population

Sources of Data

- **Primary Data**
  - Surveys
    - Mail: Lowest rate of response, usually the lowest cost
    - Web: Faster response and inexpensive
    - Telephone: Fastest response
    - Personal Interview: Usually focus groups. Most costly. Interviewer effects can be seen
- **Secondary Data**
  - This is the data that has been compiled or published elsewhere
  - Example: Census Data
  - Advantages: It can be gathered quickly and inexpensively
  - Disadvantages: May be outdated. May not be accurate

Errors

- **Response Errors**
  - Subject lies
  - Subject makes a mistake
  - Interviewer makes a mistake
  - Interviewer effects
- **Non Response Errors**
  - If the rate of response is low, then the sample is not representative
  - Might get a biased view of the population

Which is better?

Sample 1

- n = 2000
- Response rate = 90%

Sample 2

- n = 1,000,000
- Response rate = 20%

Which is better?

- Small but representative sample can be useful in making inferences
- A large sample which is unrepresentative, which makes them biased, is useless. There is no way to correct for it
- Therefore, sample 1 is better than sample 2

Types of Data

Types of Data

Categorical Data

- This refers to data that can be classified into separate groups.
- It is also called qualitative data.
- This data represents characteristics.
- For example, gender of a person can be male or female. It can also have numerical values like 1 for male and 0 for female.
- Categorical data can be further classified as nominal or ordinal.

Types of Data

Numerical Data

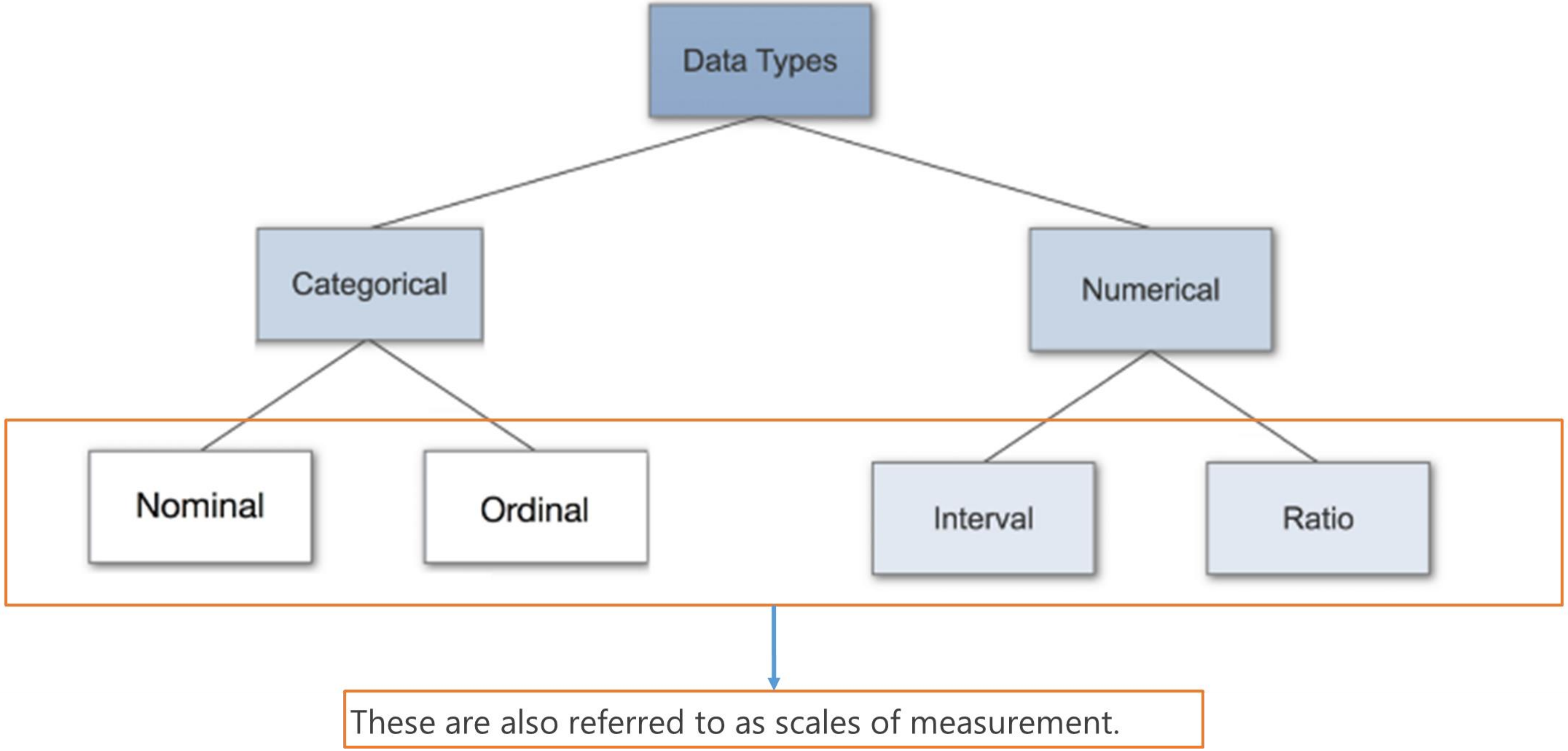
- Data that can be measured is called numerical data.
- It is also called quantitative data.
- Discrete Data:
  - If the values can be clearly separated from each other, then it is discrete data.
  - **Example:** Number of children
- Continuous data
  - **Example:** height of a person

Types of Data

Numerical Data

- One simple way to check if the data is continuous or discrete is to check whether if we can add more decimal points to the data
    - You might say you are 5’11’’ tall. But in actuality you may be 5’11.23432” tall
    - If you say you have 2 children, you cannot have 2.234545 children

Types of Data

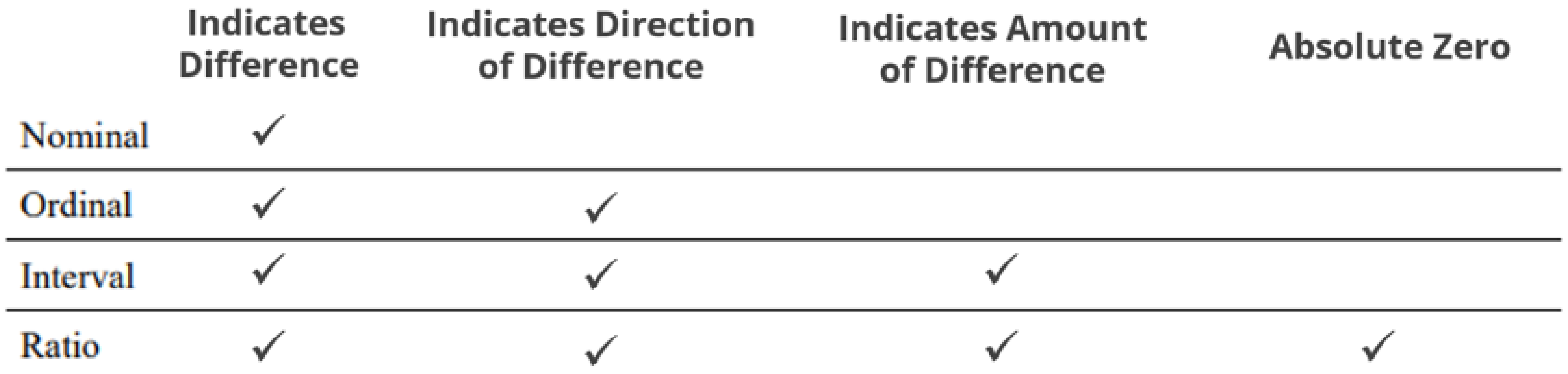


These are also referred to as scales of measurement.

Scales of Measurement

Scales of Measurement

- The differences between the four scales of measurement can be easily understood from the table:



	Indicates Difference	Indicates Direction of Difference	Indicates Amount of Difference	Absolute Zero
Nominal	✓			
Ordinal	✓	✓		
Interval	✓	✓	✓	
Ratio	✓	✓	✓	✓

- It is clear from the table that ratio scale satisfies all the four properties of scales of measurements

What do we do with the data?

Measure of Central Tendency

Measures of Central Tendency

- A measure of central tendency is a summary measure that attempts to describe a whole set of data with **a single value** that represents the **middle or centre** of its distribution.
- There are three main measures of central tendency: the **mean, median, and mode**. Each of these measures describes a different indication of the typical or central value in the distribution.

Comparisons of Measures of Central Tendency

Comparison

- **The “Hotshot” Sales Executive**
  - Kurt works as a sales manager at vsellhomes.com. In the monthly sales review, Kurt reports that he will achieve his quarterly target of $1M.
  - Kurt claims his average deal size is $100,000 and he has 10 deals in his pipeline.
  - At the end of quarter, even after closing 8 deals Kurt fails to meet his target number and falls short by more than $500,000.

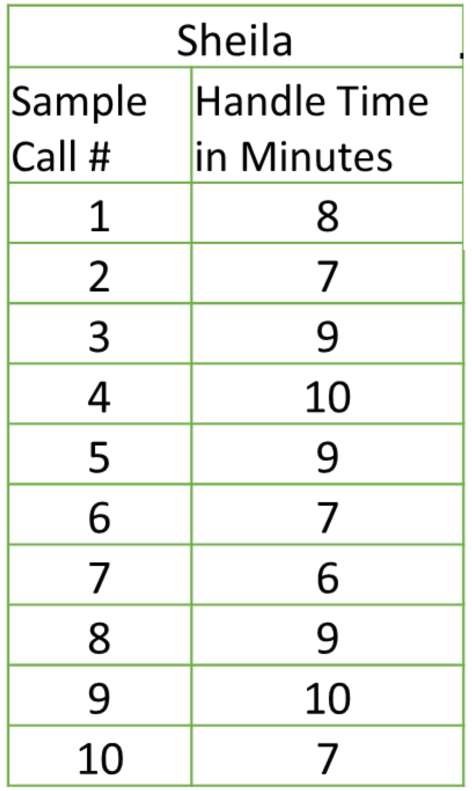
Comparison

- **The Reality of the “Hotshot” Sales Executive**
  - Average deal size in pipeline = $100,000
  - Deal #10 is of significantly higher value than all the
  - other deals and impacts the average calculation
  - Median = $55,000 more realistic measure

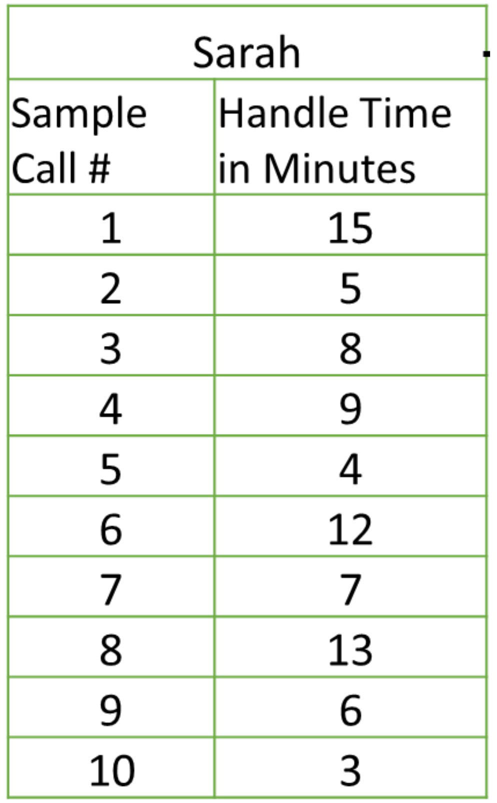
Deal #	Deal Value	Deal Status
1	70,000	Open
2	50,000	Closed
3	55,000	Closed
4	60,000	Closed
5	55,000	Closed
6	50,000	Closed
7	50,000	Closed
8	60,000	Closed
9	50,000	Closed
10	5,00,000	Open

Who is the Better Agent?

- Sheila and Sarah work as customer service agents in vsellhomes.com. During the annual performance review, the manager reviews their call handle times data. Both Sheila and Sarah have an average handle time of 8.2 minutes, which is as per the team's target of <10 minutes.



Sheila	
Sample Call #	Handle Time in Minutes
1	8
2	7
3	9
4	10
5	9
6	7
7	6
8	9
9	10
10	7



Sarah	
Sample Call #	Handle Time in Minutes
1	15
2	5
3	8
4	9
5	4
6	12
7	7
8	13
9	6
10	3

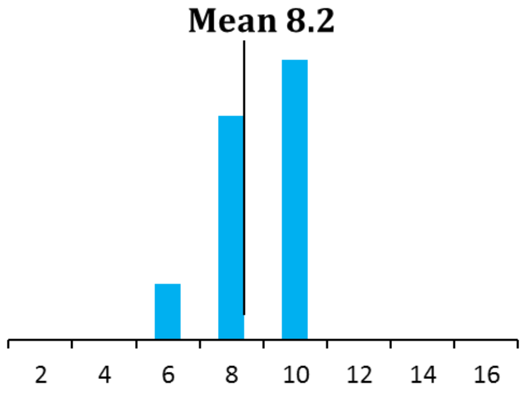
Who is the Better Agent?

- Sheila and Sarah work as customer service agents in vsellhomes.com. During the annual performance review, the manager reviews their call handle times data. Both Sheila and Sarah have an average handle time of 8.2 minutes, which is as per the team's target of <10 minutes.

Sheila	
Sample Call #	Handle Time in Minutes
1	8
2	7
3	9
4	10
5	9
6	7
7	6
8	9
9	10
10	7

Average handle time

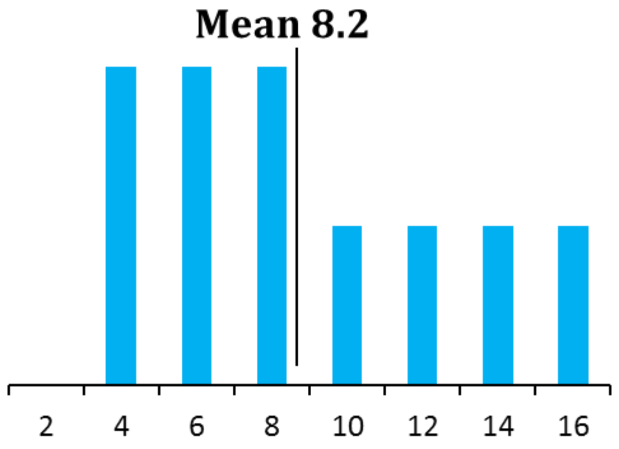
= 8.2minutes



Sarah	
Sample Call #	Handle Time in Minutes
1	15
2	5
3	8
4	9
5	4
6	12
7	7
8	13
9	6
10	3

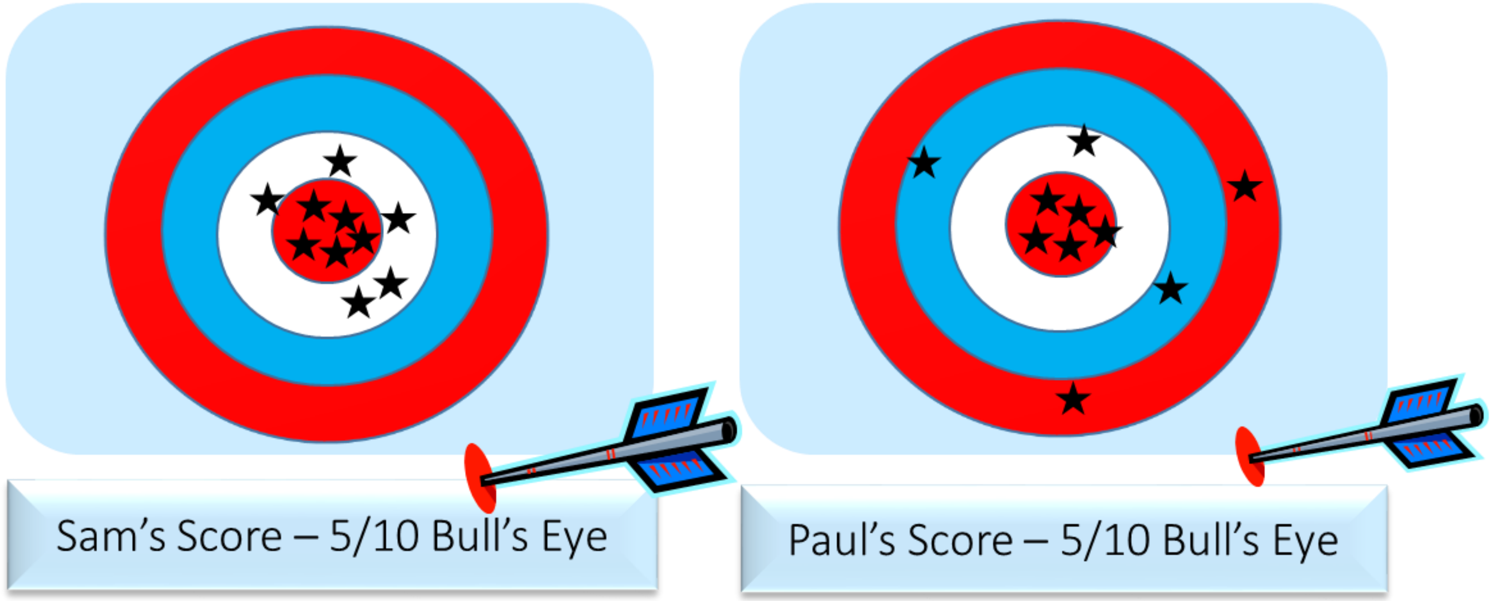
Average handle time

= 8.2 minutes



Bull’s Eye

- Sam and Paul are throwing darts at the local sports bar. A few of their friends start a betting pool. Both Sam and Paul shoot 10 practice shots each so that their friends can decide their bets.



Sam’s Score – 5/10 Bull’s Eye

Paul’s Score – 5/10 Bull’s Eye

- Who is a better bet: Sam or Paul?

Measures Of Dispersion

Measures of Dispersion

- Often, measures of central tendency are not adequate to describe the data completely
- Dispersion is the amount of the spread, or variability in the data
- They roughly describe how consistent is the data. They describe how closely data points are located to one another
- Let us looks at the necessity of measures of dispersion with an example

Range

- It is simply the range on the number line which the dataset covers
- $Range = Max(Data) - Min(Data)$
- Example:
- We have a dataset given as 1 2 3 4 8
- Range here is $8 - 1 = 7$

- Problem here is range is extremely influenced by Outliers

Standard Deviation

- It is probably the most used measure of dispersion
- It measures the “average” deviation about the mean

Variance

- It is the square of standard deviation