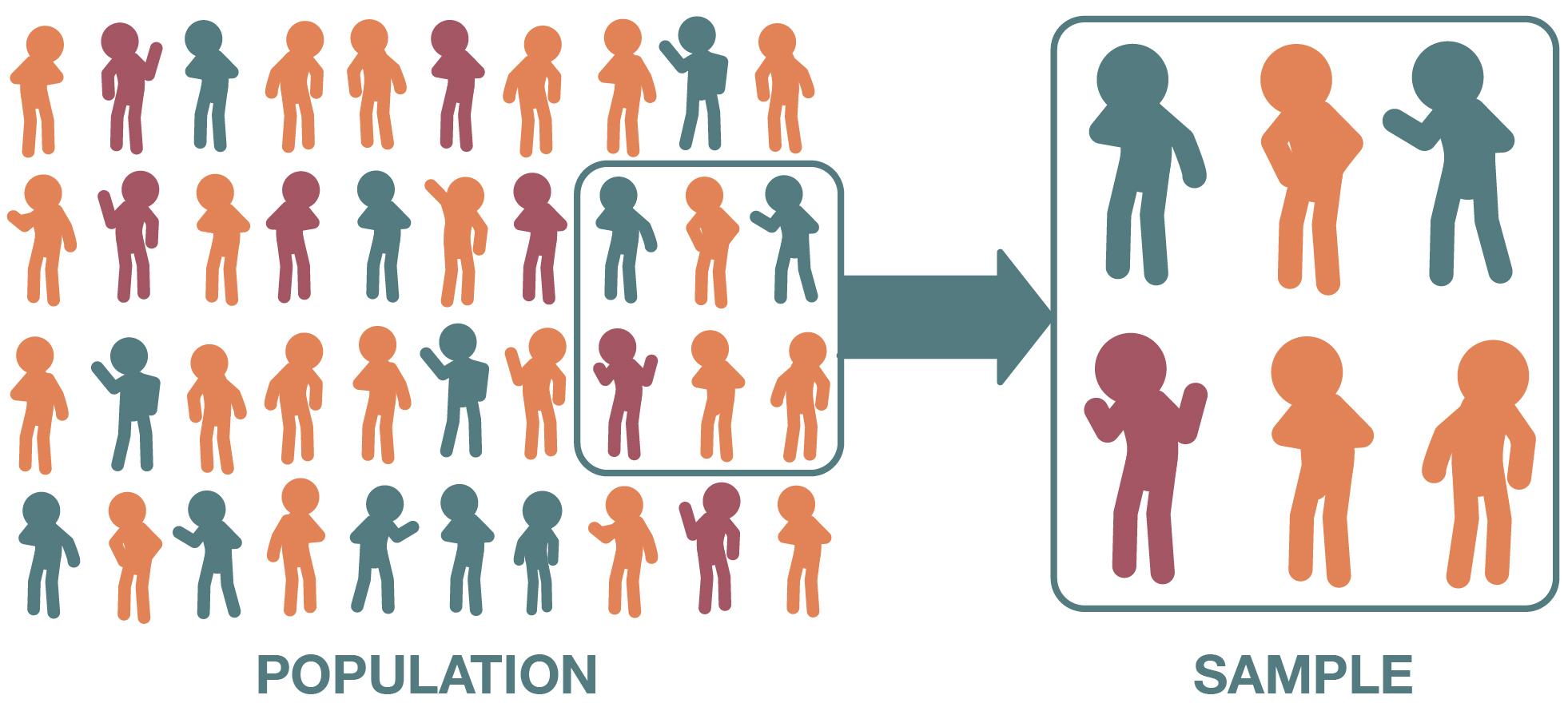
**An Introduction to Analyzing Statistical Data**

**Introduction to Data and Measurement Issues**

**Classifying Variables**

**Statisticians** refer to an entire group that is being studied as a population. Each member of the population is called a unit.

**Population vs. Sample**

**Population** is the total group being studied. A representative group from the population is called sample.

In statistics, we call the actual number of objects as a parameter. Any number that describes the individuals in a sample (length, weight, age) is called a **statistic**. Each statistic is an estimate of a parameter, whose value may or may not be known.

**Errors in Sampling**

* A statistician would report the estimate of the parameter in two ways: **as a point estimate** (e.g., 915) and as an **interval estimate.**
* The difference between the true parameter and the statistic obtained by sampling is called **sampling error.**
* It is also possible that the researcher made mistakes in her sampling methods in a way that led to a sample that does not accurately represent the true population.
* This type of systematic error in sampling is called **bias.**

**Levels of Measurement**

**Nominal measurement**

A nominal measurement is one in which the values of the variable are names.

Ex. The colors of crayons in a 24-crayon box.

**Ordinal measurement**

An **ordinal measurement** involves collecting information of which the order is somehow significant. The name of this level is derived from the use of ordinal numbers for ranking (1st, 2nd, 3rd, etc.).

Ex. High school men soccer players classified by their athletic ability: Superior, Average, Above Average

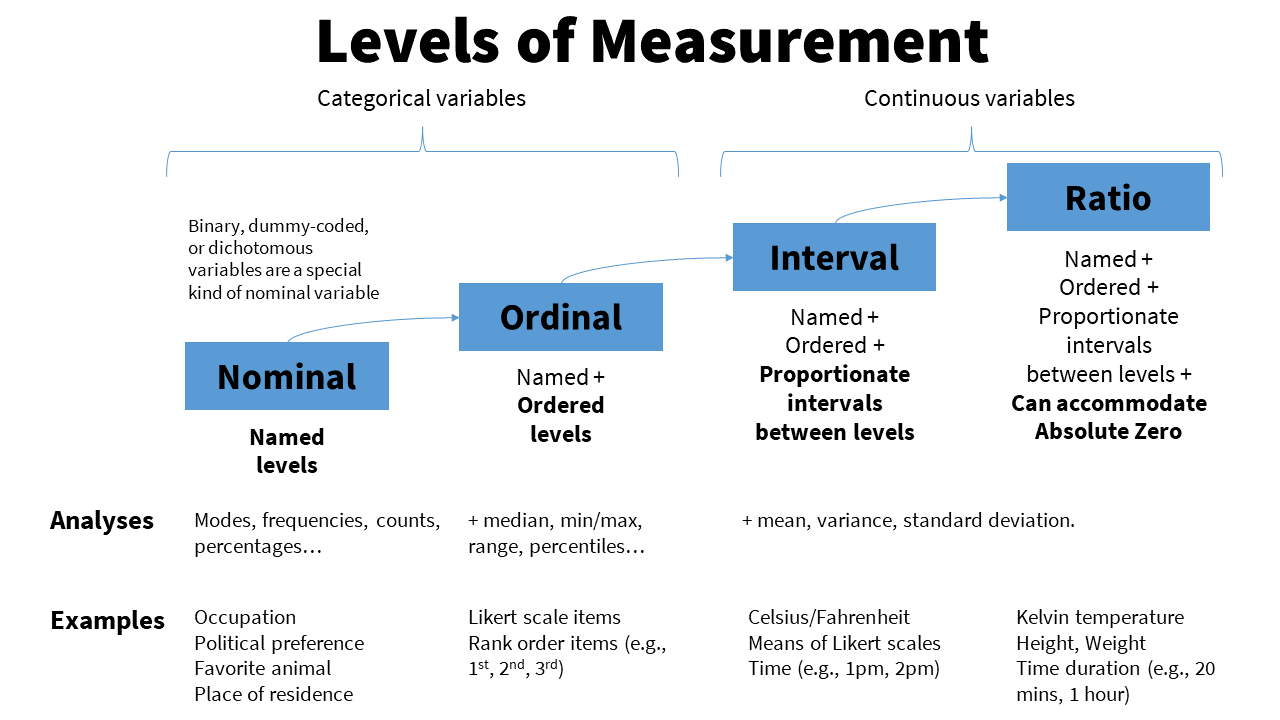
**Interval measurement**

With interval measurement, there is significance to the distance between any two values.

Ex. Time of the day on an Analog clock.

**Ratio measurement**

A **ratio measurement** is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. A variable measured at this level not only includes the concepts of order and interval, but also adds the idea of 'nothingness', or absolute zero.

Ex. Incomes measured in dollars.

**Measures of Central Tendency and Dispersion Measurements of Center**

The students in a statistics class were asked to report the number of children that live in their house (including brothers and sisters temporarily away at college).

The data are recorded below: 1, 3, 4, 3, 1, 2, 2, 2, 1, 2, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 6

Three commonly used measures of center are the mode, the median, and the mean

Okay, let's break down how to understand the average number of children in the students' homes using the mode, median, and mean. Imagine these numbers represent the number of kids in each student's family:

1, 3, 4, 3, 1, 2, 2, 2, 1, 2, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 6

**1. The Mode:** This is the number that appears most often. Look at the list – the number "2" shows up the most times. So, the mode is 2. It's a quick way to see what's most typical.

**2. The Median:** This is the middle number when the data is arranged from least to greatest. First, let's sort the data:

1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5, 6

There are 19 numbers. The middle number is the 10th number. Therefore, the median is 2. The median gives us the exact middle value.

**3. The Mean:** This is what people usually call the "average." To find it, add up all the numbers and then divide by the total number of numbers.

(1 + 1 + 1 + 1 + 2 + 2 + 2 + 2 + 2 + 2 + 2 + 3 + 3 + 3 + 3 + 4 + 4 + 5 + 6) / 19 = 2.63

The mean is approximately 2.63. This tells us the average number of children across all families.

**In short:**

* **Mode:** 2 (most frequent)
* **Median:** 2 (middle value)
* **Mean:** 2.63 (average)

These three measures give slightly different perspectives on the "center" of the data. The mode and median are less sensitive to extreme values (like the "6"), while the mean is affected by them. In this case, they're all close to each other.

**1. Mode**

The **mode** is defined as the most frequently occurring number in a data set. The mode is most useful in situations that involve categorical (qualitative) data that are measured at the nominal level.

**2. Mean**

* Another measure of central tendency is the arithmetic average, or mean.
* This value is calculated by adding all the data values and dividing the sum by the total number of data points.
* The mean is the numerical balancing point of the data set.

**3. Median**

* The median is simply the middle number in an ordered set of data.
* Suppose a student took five statistics quizzes and received the following grades: 80, 94, 75, 96, 90

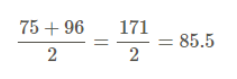
**Outliers and Resistance**

* The mean and the median are so different in this example because there is one grade that is extremely different from the rest of the data.
* In statistics, we call such extreme values outliers.
* The mean is affected by the presence of an outlier; however, the median is not.

**Summary Statistics, Summarizing Univariate Distributions Midrange**

The **midrange** (sometimes called the mid extreme) is found by taking the mean of the maximum and minimum values of the data set.

Consider the following quiz grades: 75, 80, 90, 94, and 96. The midrange would be:

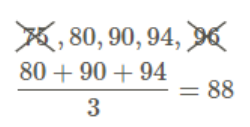


Since it is based on only the two most extreme values, the midrange is not commonly used as a measure of central tendency.

**Trimmed Mean**

To calculate a trimmed mean, you remove the maximum and minimum values and divide by the number of values that remain.

Consider the following quiz grades: 75, 80, 90, 94, 96. A trimmed mean would remove the largest and smallest values, 75 and 96, and divide by 3.



Measures of Spread and Dispersion Range

* One measure of spread is the range.
* The range is simply the difference between the largest value (maximum) and the smallest value (minimum) in the data