Statistical and Critical Thinking

Colby Community College

Definition

A **parameter** is a numerical measurement describing some characteristic of a population.

Definition

A **parameter** is a numerical measurement describing some characteristic of a population.

Definition

A **statistic** is a numerical measurement describing some characteristic of a sample.

Definition

A **parameter** is a numerical measurement describing some characteristic of a population.

Definition

A **statistic** is a numerical measurement describing some characteristic of a sample.

Hint

Parameter and population both start with a "P."
Statistic and sample both start with a "S."

Example 1

There are $17,246,372$ high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days. (Based on data in "Texting While Driving and Other Risky Motor Vehicle Behaviors Among US High School Students" by Olsen, Shults, Eaton, *Pediatrics*, Vol 131, No 6.)

Example 1

There are $17,246,372$ high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days. (Based on data in "Texting While Driving and Other Risky Motor Vehicle Behaviors Among US High School Students" by Olsen, Shults, Eaton, *Pediatrics*, Vol 131, No 6.)

What are the Parameters?

Example 1

There are $17,246,372$ high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days. (Based on data in "Texting While Driving and Other Risky Motor Vehicle Behaviors Among US High School Students" by Olsen, Shults, Eaton, *Pediatrics*, Vol 131, No 6.)

What are the Parameters?

The population size $17,246,372$.

Example 1

There are $17,246,372$ high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days. (Based on data in "Texting While Driving and Other Risky Motor Vehicle Behaviors Among US High School Students" by Olsen, Shults, Eaton, *Pediatrics*, Vol 131, No 6.)

What are the Parameters?

The population size $17,246,372$.

What are the Statistics?

Example 1

There are $17,246,372$ high school students in the United States. In a study of 8505 U.S. high school students 16 years of age or older, 44.5% of them said that they texted while driving at least once during the previous 30 days. (Based on data in "Texting While Driving and Other Risky Motor Vehicle Behaviors Among US High School Students" by Olsen, Shults, Eaton, *Pediatrics*, Vol 131, No 6.)

What are the Parameters?

The population size $17,246,372$.

What are the Statistics?

The sample size 8505 and the percentage of the sample 44.5% that said they texted while driving.

Definition

Quantitative or **numerical** data consist of numbers representing counts or measurements.

Definition

Quantitative or **numerical** data consist of numbers representing counts or measurements.

Definition

Categorical or **qualitative** data consist of names or labels (not numbers).

Definition

Quantitative or **numerical** data consist of numbers representing counts or measurements.

Definition

Categorical or **qualitative** data consist of names or labels (not numbers).

Note

The names and labels in categorical data are sometimes coded with numbers. When a number is used as a name it is **not** quantitative.

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Quantitative data.

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Quantitative data.

The genders (male / female) of subjects enrolled in a clinical trial are?

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Quantitative data.

The genders (male / female) of subjects enrolled in a clinical trial are?

Categorical Data.

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Quantitative data.

The genders (male / female) of subjects enrolled in a clinical trial are?

Categorical Data.

Identification numbers $1, 2, 3, \ldots, 25$ are assigned randomly to 25 subjects in a clinical trail. The identification numbers are?

Example 2

The ages (in years) of subjects enrolled in a clinical trail are?

Quantitative data.

The genders (male / female) of subjects enrolled in a clinical trial are?

Categorical Data.

Identification numbers $1, 2, 3, \ldots, 25$ are assigned randomly to 25 subjects in a clinical trail. The identification numbers are?

Categorical Data. (The numbers don't measure or count anything.)

Definition

Discrete data result when the data values are quantitative and the number of values is finite, or countable.

Definition

Discrete data result when the data values are quantitative and the number of values is finite, or countable.

Example 3

Each of several physicians plans to count the number of physical examinations given during the next full week.

Definition

Discrete data result when the data values are quantitative and the number of values is finite, or countable.

Example 3

Each of several physicians plans to count the number of physical examinations given during the next full week.

Example 4

Casino employees plan to roll a fair die until the number 5 is rolled, and they count the number of rolls required to get a 5. (It is possible, but unlikely, that a 5 could never be rolled.)

Definition

Continuous data result from infinitely many possible quantitative values, where the collection of values is not countable.

Definition

Continuous data result from infinitely many possible quantitative values, where the collection of values is not countable.

Example 5

When a typical patient has blood drawn as part of a routine examination, the volume of blood drawn is between 0 mL and 50 mL.

Definition

Continuous data result from infinitely many possible quantitative values, where the collection of values is not countable.

Example 5

When a typical patient has blood drawn as part of a routine examination, the volume of blood drawn is between 0 mL and 50 mL.

Note

If your data consists of only integers, then it is generally discrete.
If your data consists of real numbers, then it is generally continuous.

Levels of Measurement

The level of measurement of data helps us decide which statistical procedure to use.

Levels of Measurement

The level of measurement of data helps us decide which statistical procedure to use.

Ratio

There is a natural zero starting point and rations make sense. Examples are heights, lengths, distances, and volumes.

Levels of Measurement

The level of measurement of data helps us decide which statistical procedure to use.

Ratio

There is a natural zero starting point and rations make sense. Examples are heights, lengths, distances, and volumes.

Interval

Differences are meaningful, but there is no natural zero starting point and ratios are meaningless. An example is body temperature.

Levels of Measurement

The level of measurement of data helps us decide which statistical procedure to use.

Ratio

There is a natural zero starting point and rations make sense. Examples are heights, lengths, distances, and volumes.

Interval

Differences are meaningful, but there is no natural zero starting point and ratios are meaningless. An example is body temperature.

Ordinal

Data can be arranged in order, but differences either can't be found or are meaningless. An example is ranks of colleges in U.S. News & World Report.

Levels of Measurement

The level of measurement of data helps us decide which statistical procedure to use.

Ratio

There is a natural zero starting point and rations make sense. Examples are heights, lengths, distances, and volumes.

Interval

Differences are meaningful, but there is no natural zero starting point and ratios are meaningless. An example is body temperature.

Ordinal

Data can be arranged in order, but differences either can't be found or are meaningless. An example is ranks of colleges in U.S. News & World Report.

Nominal

Categories only. Data cannot be arranged in order. An example is eye color.

Definition

The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in some meaningful order, such as low to high.

Definition

The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in some meaningful order, such as low to high.

Example 6

A survey has responses of "yes", "no", and "undecided"

Definition

The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in some meaningful order, such as low to high.

Example 6

A survey has responses of "yes", "no", and "undecided"

Example 7

For an item on a survey, respondents are given a choice of possible answers, and they are coded as follows:

1 is coded as "I agree"

2 is coded as "I disagree"

3 is coded as "I don't care"

4 is coded as "I refuse to answer"

The numbers 1,2,3, and 4 don't count or measure anything.

Definition

Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences between data values either cannot be determined or are meaningless.

Definition

Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences between data values either cannot be determined or are meaningless.

Example 8

A college professor assigns grades of A, B, C, D, or F. These grades can be arranged in a meaningful order, but grades are very individualized so the difference between two students grades cannot be calculated.

Definition

Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences between data values either cannot be determined or are meaningless.

Example 8

A college professor assigns grades of A, B, C, D, or F. These grades can be arranged in a meaningful order, but grades are very individualized so the difference between two students grades cannot be calculated.

Example 9

A survey asks people what they felt their blood pressure was. The possible answers are "Low", "Normal", "High." These responses can be arranged in a meaningful order, but the differences between "Low" and "High" doesn't make sense.

Definition

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. *Data at this level do not have a natural zero starting point at which none of the quantity is present.*

Definition

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. *Data at this level do not have a natural zero starting point at which none of the quantity is present.*

Example 10

Body temperatures of 98.2°F and 98.9°F are examples of data at this interval level of measurement. The values are ordered, and we can calculate their difference. There is no natural starting point. (The value 0°F is an arbitrary choice and doesn't represent the complete absence of heat.)

Definition

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. *Data at this level do not have a natural zero starting point at which none of the quantity is present.*

Example 10

Body temperatures of 98.2°F and 98.9°F are examples of data at this interval level of measurement. The values are ordered, and we can calculate their difference. There is no natural starting point. (The value 0°F is an arbitrary choice and doesn't represent the complete absence of heat.)

Example 11

The years 1492 and 1776 can be arranged in order, and the difference of 284 years is meaningful. But, time doesn't not have a natural starting point that represents "no time."

Definition

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and there is a natural starting point which indicates that none of the quantity is present.

Definition

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and there is a natural starting point which indicates that none of the quantity is present.

Example 12

Heights of 180cm and 90cm for a high school student and a preschool student are at the ratio level of measurement. (The starting point is 0cm and 180cm is twice as tall as 90cm.)

Definition

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and there is a natural starting point which indicates that none of the quantity is present.

Example 12

Heights of 180cm and 90cm for a high school student and a preschool student are at the ratio level of measurement. (The starting point is 0cm and 180cm is twice as tall as 90cm.)

Example 13

The times of 50 minutes and 100 minutes for a math class are at the ratio level of measurement. (The starting point is 0 minutes and 100 minutes is twice as long as 50 minutes.)

Definition

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Definition

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Definition

Data science involves applications of statistics, computer science, and software engineering, along with some other relevant fields, such as sociology or finance.

Definition

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Definition

Data science involves applications of statistics, computer science, and software engineering, along with some other relevant fields, such as sociology or finance.

Example 14

- Google provides live traffic maps by recording and analyzing GPS data collected from the smartphones of people traveling in their vehicles.

Definition

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Definition

Data science involves applications of statistics, computer science, and software engineering, along with some other relevant fields, such as sociology or finance.

Example 14

- Google provides live traffic maps by recording and analyzing GPS data collected from the smartphones of people traveling in their vehicles.
- Walmart has a sales database with more than 2.5 petabytes (2,500,000 gigabytes) of data. For online sales, Walmart developed the Polaris search engine that increased sales by 10% to 15%.

Missing Data

When dealing with large data sets, it is inevitable that some data will go missing. Missing data does not necessarly invalidate the data set.

Missing Data

When dealing with large data sets, it is inevitable that some data will go missing. Missing data does not necessarly invalidate the data set.

Definition

A data value is **missing completely at random** if the likelihood of its being missing is independent of it's value or any of the other values in the data set. Thats is, any data values is just as likely to missing as any other data value.

Missing Data

When dealing with large data sets, it is inevitable that some data will go missing. Missing data does not necessarly invalidate the data set.

Definition

A data value is **missing completely at random** if the likelihood of its being missing is independent of it's value or any of the other values in the data set. Thats is, any data values is just as likely to missing as any other data value.

Example 15

When using a keyboard to manually enter ages of survey respondents, the operator is distracted by a text message and makes the mistake of failing to enter the age of 37 years.

Definition

A data value is **missing not at random** if the missing value is related to the reason that it is missing.

Definition

A data value is **missing not at random** if the missing value is related to the reason that it is missing.

Example 16

A survey asks each respondent to enter his or her annual income, but respondents with very low income skip this question because they find it embarrassing.

Correcting for Missing Data

1. One very common method for dealing with missing data is to delete all subjects having and missing values.

Correcting for Missing Data

1. One very common method for dealing with missing data is to delete all subjects having and missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased.

Correcting for Missing Data

1. One very common method for dealing with missing data is to delete all subjects having and missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased.
   - If the data are missing not at random, deleting subjects having missing values can easily bias the results.

Correcting for Missing Data

1. One very common method for dealing with missing data is to delete all subjects having and missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased.
   - If the data are missing not at random, deleting subjects having missing values can easily bias the results.
2. We can "impute" missing data values by substituting values for them. There are different methods of determining the replacement values.

Correcting for Missing Data

1. One very common method for dealing with missing data is to delete all subjects having and missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased.
   - If the data are missing not at random, deleting subjects having missing values can easily bias the results.
2. We can "impute" missing data values by substituting values for them. There are different methods of determining the replacement values.

Note

When analyzing sample data with missing values you must always try to determine why they are missing and decide whether it makes sense to treat the remaining values as being representative of the population.