

# WHAT IS STATISTICS?



What are *data*? (*data* is the plural of the singular *datum*)

What are *data*? (*data* is the plural of the singular *datum*)

*Data* are collections of observations such as measurements, genders, survey responses, and so on.



What is a *population*?



What is a *population*?

A *population* is the complete collection of items or individuals that is intended to be studied.



What is a *sample*?



What is a *sample*?

A *sample* is any proper subset of a population.



What is a *random sample*?



What is a *random sample*?

A *random sample* is a sample in which each element of the population has an equally likely chance of being selected.



What is a *random sample*?

A *random sample* is a sample in which each element of the population has an equally likely chance of being selected.

We take random samples in an attempt to prevent *bias*.



What is *bias*?



What is *bias*?

*Bias-*

1. A particular tendency or inclination, especially one that prevents unprejudiced consideration of a question.
2. *Statistics*: A systematic as opposed to a random distortion of a measure as a result of a sampling procedure.



What is a *voluntary response sample*?



What is a *voluntary response sample*?

A *voluntary response sample* is one in which each individual can decide whether or not to participate.



What is a *voluntary response sample*?

A *voluntary response sample* is one in which each individual can decide whether or not to participate.

**Example:** The results at [ratemyprofessors.com](http://ratemyprofessors.com) represent a voluntary response sample since no one is forced to make a comment. They volunteer the information.



What is a *voluntary response sample*?

A *voluntary response sample* is one in which each individual can decide whether or not to participate.

**Example:** The results at [ratemyprofessors.com](http://ratemyprofessors.com) represent a voluntary response sample since no one is forced to make a comment. They volunteer the information.

*Voluntary response samples* are not necessarily valid since they are very susceptible to bias.



What is a *census*?



What is a *census*?

A *census* is a collection of data from every member of a population.



What is a *parameter*?



What is a *parameter*?

A *parameter* is a numerical characteristic of a population.



What is a *statistic*?



What is a *statistic*?

A *statistic* is a numerical characteristic of a sample.

What is the *science of statistics*?

What is a the *science of statistics*?

The *science of statistics* consists of planning studies and experiments, obtaining data, analyzing the collected data, and drawing conclusions based on the data.



What is *descriptive statistics*?



What is *descriptive statistics*?

*Descriptive statistics* attempts to summarize and describe data.



What is *inferential statistics*?



What is *inferential statistics*?

*Inferential statistics* attempts to extrapolate from the data and determine if the results statistically significant.

In doing any statistical study, the following factors should be considered:

In doing any statistical study, the following factors should be considered:

- *Context of the data*

In doing any statistical study, the following factors should be considered:

- *Context of the data*

- What is the context of the data? Numbers without context are meaningless. For example, are we looking at the performance of two individuals? Are we looking at before and after scores for a single individual?

In doing any statistical study, the following factors should be considered:

- *Context of the data*

- What is the context of the data? Numbers without context are meaningless. For example, are we looking at the performance of two individuals? Are we looking at before and after scores for a single individual?
- What statistical procedures are suggested by the context?



In doing any statistical study, the following factors should be considered:

- *Source of the data*



In doing any statistical study, the following factors should be considered:

- *Source of the data*

- Is the data source objective or biased? For example, a tobacco company does a study to “prove” that cigarettes are safe.



In doing any statistical study, the following factors should be considered:

- *Source of the data*

- Is the data source objective or biased? For example, a tobacco company does a study to “prove” that cigarettes are safe.
- Are there problems with the data that might affect the outcome of analysis? For example, the presence of outliers.



In doing any statistical study, the following factors should be considered:

- *Source of the data*

- Is the data source objective or biased? For example, a tobacco company does a study to “prove” that cigarettes are safe.
- Are there problems with the data that might affect the outcome of analysis? For example, the presence of outliers.

**Outlier:** An *outlier* is an unusual element of data, generally an observation that is numerically distant from the rest of the data.



(1.2)



In doing any statistical study, the following factors should be considered:

- *Source of the data*

- Is the data source objective or biased? For example, a tobacco company does a study to “prove” that cigarettes are safe.
- Are there problems with the data that might affect the outcome of analysis? For example, the presence of outliers.
- The world is often run by bad data.



In doing any statistical study, the following factors should be considered:

- *Sampling method*



In doing any statistical study, the following factors should be considered:

- *Sampling method*

- *Were there problems in the data collection procedure that might affect the outcome?*



In doing any statistical study, the following factors should be considered:

- *Sampling method*

- *Were there problems in the data collection procedure that might affect the outcome?*
- *Is it a random sample?*



In doing any statistical study, the following factors should be considered:

- *Sampling method*

- *Were there problems in the data collection procedure that might affect the outcome?*
- *Is it a random sample?*
- *Is it a voluntary response sample?*

In doing any statistical study, the following factors should be considered:

- *Conclusions*

In doing any statistical study, the following factors should be considered:

- *Conclusions*

- *What conclusions can we draw from the data?*

In doing any statistical study, the following factors should be considered:

- *Conclusions*

- *What conclusions can we draw from the data?*
- *Can we express the conclusions in plain English for those who are not well versed in statistics?*



In doing any statistical study, the following factors should be considered:

- *Conclusions*

- *What conclusions can we draw from the data?*
- *Can we express the conclusions in plain English for those who are not well versed in statistics?*
- *Have we been careful not to confuse correlation with causality?*



In doing any statistical study, the following factors should be considered:

- *Conclusions*

- *What conclusions can we draw from the data?*
- *Can we express the conclusions in plain English for those who are not well versed in statistics?*
- *Have we been careful not to confuse correlation with causality?*

**Example:** In general, **correlation** means that there is a mathematical relationship between two variables. For instance, as a person's height increases, their weight also increases. However, will gaining weight **cause** you to grow taller?



In doing any statistical study, the following factors should be considered:

- *Conclusions*

- *What conclusions can we draw from the data?*
- *Can we express the conclusions in plain English for those who are not well versed in statistics?*
- *Have we been careful not to confuse correlation with causality?*

**Example:** Or, consider this. As the temperature increases, the number of drowning deaths also increases (since more people go swimming). However, does temperature cause drowning, or does drowning cause the temperature to go up?



In doing any statistical study, the following factors should be considered:

- *Conclusions*

- *What conclusions can we draw from the data?*
- *Can we express the conclusions in plain English for those who are not well versed in statistics?*
- *Have we been careful not to confuse correlation with causality?*

**Example:** A recent scientific study showed that people who walk faster tend to live longer. Do you suspect that there is a causal relationship here? What is it? Do we know without doing other studies that causality is involved?



In doing any statistical study, the following factors should be considered:

- *Practical implications*



In doing any statistical study, the following factors should be considered:

- *Practical implications*
  - *Are the results statistically significant?*



In doing any statistical study, the following factors should be considered:

- *Practical implications*
  - *Are the results statistically significant?*

**Example:** For now, think of **statistical significance** as meaning simply the occurrence of something that is unlikely to have been due to chance. In other words, we measure the statistical significance of an event in terms of whether it was a probable or an improbable outcome.



In doing any statistical study, the following factors should be considered:

- *Practical implications*
  - *Are the results statistically significant?*

**Example:** For instance, if you flip a “fair” coin 100 times and get heads each time, then is highly improbable. Thus, you would deem the result statistically significant, and you would question your assumption that it was a “fair” coin.



In doing any statistical study, the following factors should be considered:

- *Practical implications*
  - *Are the results statistically significant?*
  - *Do the results have practical significance?*



In doing any statistical study, the following factors should be considered:

- *Practical implications*

- *Are the results statistically significant?*
- *Do the results have practical significance?*

**Example:** When the number of records or subjects is large, even small changes have a tendency to be **statistically significant**. However, even if a change in the average math SAT score from, say, 550 to 551 turns out to be statistically significant, would you really consider that result to have practical significance?



In doing any statistical study, the following factors should be considered:

- *Practical implications*
  - *Are the results statistically significant?*
  - *Do the results have practical significance?*

**Example:** On the other hand, when the number of records or subjects is small, even “large” changes may be statistically insignificant. Nonetheless, while a change in enrollment of 20 students at a small, private school may not be statistically significant, the difference in revenue that results may be practically significant.



What is *quantitative data*?



What is *quantitative data*?

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



What is *quantitative data*?

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

**Example:** The heights of students in a classroom give us *numerical data*. We can, thus, compute things such as the difference between two heights or an average height.



What is *categorical data*?



What is *categorical data*?

*Categorical or qualitative or attribute data or nonnumerical data* consist of names or labels that are not numbers representing counts or measurements.



What is *categorical data*?

*Categorical or qualitative or attribute data or nonnumerical data* consist of names or labels that are not numbers representing counts or measurements.

**Example:** The names of students in a classroom give us *nonnumerical (categorical) data*. Thus, questions such as, “What is *Bob minus Fred*?” are meaningless.



What is *discrete data*?



What is *discrete data*?

*Discrete data* results when the number of possible values is either a finite number or a “countable” number (0 or 1 or 2, and so on). The values of discrete data are separated from one another.



What is *discrete data*?

*Discrete data* results when the number of possible values is either a finite number or a “countable” number (0 or 1 or 2, and so on). The values of discrete data are separated from one another.

**Example:** The number of students in a classroom is *discrete*. You can have 5 students or you can have 6 students, but you can't have  $5 \frac{1}{2}$  students in class.



What is *continuous data*?



What is *continuous data*?

*Continuous data* results from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruption, or jumps. Continuous data exist along a continuum.



What is *continuous data*?

**Continuous data** results from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruption, or jumps. Continuous data exist along a continuum.

**Example:** The heights of students in a classroom is **continuous**. Between any two height values it's always theoretically possible for someone to have a height that's in between. Possible height values exists along a continuum.



What is *the nominal level of measurement*?



What is *the nominal level of measurement*?

*The nominal level of measurement* is characterized by names, labels, and categories only. The data is not ordered in any way such as from low to high.



What is *the nominal level of measurement*?

*The nominal level of measurement* is characterized by names, labels, and categories only. The data is not ordered in any way such as from low to high.

**Example:** If we collect information only on the type of cell phone a person has (I,e, Verizon, Blackberry, Android, iPhone, etc.), then we will have *nominal data*.



What is *the ordinal level of measurement?*



What is *the ordinal level of measurement*?

Data are at *the ordinal level of measurement* if the context provides a natural order to the data, but differences between data values either cannot be determined or are meaningless.



What is *the ordinal level of measurement*?

Data are at *the ordinal level of measurement* if the context provides a natural order to the data, but differences between data values either cannot be determined or are meaningless.

**Example:** If we collect information on the “star” rating of restaurants, then we will have *ordinal data*.



What is *the ordinal level of measurement*?

Data are at *the ordinal level of measurement* if the context provides a natural order to the data, but differences between data values either cannot be determined or are meaningless.

**Example:** Another example of *ordinal data* is percentile ranking on a standardized test like the *SAT*.



What is *the interval level of measurement*?



What is *the interval level of measurement*?

*The interval level of measurement* is like the ordinal level except that the difference between any two data values is meaningful. However, at this level of data measurement, an absolute or natural zero does not exist.



What is *the interval level of measurement*?

*The interval level of measurement* is like the ordinal level except that the difference between any two data values is meaningful. However, at this level of data measurement, an absolute or natural zero does not exist.

**Example:** A standard example of *interval data* is temperature measured in degrees Fahrenheit. For example, it takes the same amount of energy to raise the temperature from 50° to 60° as it does from 80° to 90°, but we can't say that 60° is twice as hot as 30° since we aren't measuring temperature with respect to an absolute zero.



What is *the ratio level of measurement*?



What is *the ratio level of measurement*?

*The ratio level of measurement* is like the interval level with the addition that an absolute zero point exists. Consequently, with respect to this absolute zero point, we can talk in terms of one data element being twice as large as another data element.



What is *the ratio level of measurement*?

*The ratio level of measurement* is like the interval level with the addition that an absolute zero point exists. Consequently, with respect to this absolute zero point, we can talk in terms of one data element being twice as large as another data element.

**Example:** If we measure temperature in *Kelvins*, then we have *ratio data* since an absolute zero point of reference exists. Thus, it does make sense to say that *60 Kelvins* is twice as hot as *30 Kelvins*.



What is *the ratio level of measurement*?

*The ratio level of measurement* is like the interval level with the addition that an absolute zero point exists. Consequently, with respect to this absolute zero point, we can talk in terms of one data element being twice as large as another data element.

**Example:** More common examples of *ratio data* would be height and weight. Since each of those scales has an absolute zero, it makes sense to say that one person is twice as tall as another or weighs twice as much as another.



What is a *voluntary response sample*?



What is a *voluntary response sample*?

A *voluntary response sample* is one in which the respondents themselves decide whether to be included. Because of the bias this creates, we can never be certain whether the results of a voluntary response sample are accurate or not.



Does *correlation imply causality?*



Does *correlation imply causality*?

*Never conclude that correlation proves causality.* Correlation is a measure of the strength of the relationship between two variables. However, just because two variables are related does not mean that one causes the other. ***Correlation does not imply causality.*** However, given that caveat, there may be situations where causality is a reasonable or educated guess.



What should we remember about *reported versus measured results*?



What should we remember about *reported versus measured results*?

Beware of ***reported versus measured results***. An example of “reported” data would be when you ask someone their weight instead of measuring it.

What should we remember about *percentages*?

What should we remember about *percentages*?

*Beware of incorrect uses of percentages.* Percent means “per one hundred.” Keep that in mind. Also, one college administrator I used to work with in Texas would characterize any increase over zero as a “100% increase.” That is not correct. For example, if an amount increases by 100%, then it doubles. However, if we double zero, then it’s still zero. I often hear statements like this: “Inflation went up 4% for each of the past three years. That’s 12% increase.” Actually, it’s a 12.4864% increase.

What should we remember about *loaded questions*?

What should we remember about *loaded questions*?

***Beware of loaded questions.*** “Are you in favor of liberal political policies” versus “Are you in favor of progressive political policies?”  
“Do you support welfare” versus “Do you support assistance to the poor?”

What should we remember about *question order*?

What should we remember about *question order*?

Be aware that even ***question order can affect responses***. The response to the first question might condition your response to the second question. “Should political leaders be elected by majority vote?”  
“Should we eliminate the Electoral College?”

What should we remember about *questions that ask too many things?*

What should we remember about *questions that ask too many things*?

***Avoid questions that ask too many things.*** For example, do you support increasing spending on social security, Medicare, and the military?



What should we remember about *bad data*?



What should we remember about *bad data*?

*The world is full of bad data!* Errors can occur in either the collection or the recording of data.

What should we remember about *nonresponses*?

What should we remember about *nonresponses*?

*Be aware of how many **nonresponses** there are in the data.* For example, some people refused to respond to recent census questionnaires.

What should we remember about *missing data*?

What should we remember about *missing data*?

*Be aware of the impact of missing data.* For example, the recent census results may be affected by an inability to collect data on some homeless people, immigrants or low income minorities.

What should we remember about *precise numbers*?

What should we remember about *precise numbers*?

*Be wary of precise numbers.* Another problem that can occur in statistics with percents is displaying too many digits. For example, uncertainty often exists when it comes to real world measurements, and in statistics we are often dealing with “best estimates” of various quantities. Consequently, when there is some uncertainty in our data, but we display our results to several decimal places, we may be suggesting a higher level of accuracy than is actually present. Often we may round results to two or three decimal places.

What is an *observational study*?

What is an *observational study*?

In an *observational study* we observe and measure specific characteristics, but we don't attempt to modify the subjects being studied.

What is an *experiment*?

What is an *experiment*?

In an *experiment*, we apply some *treatment* and then proceed to observe its effects on the subjects.

What is a *probability sample*?

What is a *probability sample*?

A *probability sample* involves selecting members from a population in such a way that each member of the population has a known, but not necessarily the same, chance of being selected.



What is a *random sample*?



What is a *random sample*?

A *random sample* is a probability sample where members from the population are selected in such a way that each individual member in the population has the same chance of being selected.



What is a *simple random sample of size  $n$* ?



What is a *simple random sample of size  $n$* ?

In a *simple random sample of size  $n$* , every possible sample of size  $n$  has the same chance of being selected. If we pick a state at random and select its two senators, then this is a random sample where each senator has a 1 in 50 chance of being selected, but it is not a simple random sample of size 2 since not every possible sample of size 2 can be selected.



What is a *systematic sample*?

(1.5)



What is a *systematic sample*?

In *systematic sampling*, we arrange the elements of the population in some order, select some starting point, and then select every  $k$ th item from that point on.



What is a *systematic sample*?

In *systematic sampling*, we arrange the elements of the population in some order, select some starting point, and then select every  $k$ th item from that point on.

**Example:** If I select every third person in the class for participation in a survey, then I have taken a *systematic sample*.



What is a *stratified sample*?

(1.5)



What is a *stratified sample*?

In *stratified sampling*, we subdivide the population into two or more distinct subgroups (or *strata*), and then we draw a sample from each subgroup.



What is a *stratified sample*?

In *stratified sampling*, we subdivide the population into two or more distinct subgroups (or *strata*), and then we draw a sample from each subgroup.

**Example:** If I divide the class into two groups based on gender and take a sample from each group, then I will have created a *stratified sample*.



## What is a *stratified sample*?

In *stratified sampling*, we subdivide the population into two or more distinct subgroups (or *strata*), and then we draw a sample from each subgroup.

**Example:** Furthermore, if I'm smart about it, then I'll construct a *proportional stratified sample*. In other words, if the class contains twice as many males as females, then I should be certain that my sample also contains twice as many males as females.



What is a *cluster sample*?

(1.5)



What is a *cluster sample*?

In *cluster sampling*, we first divide the population into sections (or *clusters*), then randomly select some of the clusters, and then we sample all the members of the selected clusters.



What is a *cluster sample*?

In *cluster sampling*, we first divide the population into sections (or *clusters*), then randomly select some of the clusters, and then we sample all the members of the selected clusters.

**Example:** If five math classes are selected at random from the population of all SCC math classes and given a survey, then we have created a *cluster sample*. The *clusters* in this case are the individual math classes.



## What is a *cluster sample*?

In *cluster sampling*, we first divide the population into sections (or *clusters*), then randomly select some of the clusters, and then we sample all the members of the selected clusters.

**Example:** If five math classes are selected at random from the population of all SCC math classes and given a survey, then we have created a *cluster sample*. The *clusters* in this case are the individual math classes.

The primary purpose of random, systematic, stratified, and cluster sampling is to eliminate bias!



What is a *convenience sample*?

(1.5)



What is a *convenience sample*?

In *convenience sampling*, we simply use whatever data or results that are easy to get to.



## What is a *convenience sample*?

In *convenience sampling*, we simply use whatever data or results that are easy to get to.

**Example:** At major universities, Psychology Departments would like to do experiments involving adults of all ages. However, freshmen psychology students are often the only subjects that are *convenient*. Furthermore, since the students often get to pick which experiments they will participate in, the result is also a *voluntary response sample*.

What is *multistage sampling*?

What is *multistage sampling*?

In *multistage sampling*, the sample is selected in stages, and a different sampling method might be used for each stage.



What is a *cross-sectional study*?



What is a *cross-sectional study*?

In a *cross-sectional study*, data are observed, measured, and collected at one point in time.



What is a *cross-sectional study*?

In a *cross-sectional study*, data are observed, measured, and collected at one point in time.

**Example:** If we do a survey at a particular point in time to determine a political leader's popularity, that is a *cross-sectional study*.



What is a *retrospective study*?



What is a *retrospective study*?

In a *retrospective* or *case-control study*, data are collected from the past using, for example, existing records and interviews.



What is a *retrospective study*?

In a *retrospective* or *case-control study*, data are collected from the past using, for example, existing records and interviews.

**Example:** If we analyze historical trends in inflation, that would be a *retrospective study*.



What is a *prospective study*?



What is a *prospective study*?

In a *prospective* or *longitudinal* or *cohort study*, data are collected in the future from groups sharing common factors called *cohorts*.



What is a *prospective study*?

In a *prospective* or *longitudinal* or *cohort study*, data are collected in the future from groups sharing common factors called *cohorts*.

**Example:** If we give vaccinations to one group and withhold them from another, and then collect data on longevity several years later, that is a *prospective* or *longitudinal study*.

What is *randomization*?

What is *randomization*?

*Randomization* is used when subjects are assigned to different groups through a process of random selection.

What is *replication*?

What is *replication*?

*Replication* is the repetition of an experiment on more than one subject.



What is the *placebo effect*?

(1.5)



What is the *placebo effect*?

When an untreated subject shows improvement, this is called the *placebo effect*.



What is *blinding*?



What is *blinding*?

*Blinding* is a technique in which the subject doesn't know whether he or she is receiving a treatment or a placebo.



What is a *double-blind experiment*?

(1.5)



What is a *double-blind experiment*?

In a *double-blind experiment* neither “doctors” nor “patients” know which treatment group the patient belongs to.



What is a *confounding variable*?



What is a *confounding variable*?

*Confounding* occurs in an experiment when you are not able to distinguish among the effects of different variables. It is highly desirable to eliminate any *confounding variables* that might impact the interpretation of the results.



## What is a *confounding variable*?

*Confounding* occurs in an experiment when you are not able to distinguish among the effects of different variables. It is highly desirable to eliminate any *confounding variables* that might impact the interpretation of the results.

**Example:** Suppose you are trying to determine if high stress levels make it more difficult for a person to get a good night's sleep. However, all your subjects with high stress also drink more coffee than usual (as a result of the stress). Then it is hard to determine if the sleep problems are due to the stress or due to the coffee or a combination of both. In this example, coffee drinking is a *confounding variable*.



## What is a *confounding variable*?

*Confounding* occurs in an experiment when you are not able to distinguish among the effects of different variables. It is highly desirable to eliminate any *confounding variables* that might impact the interpretation of the results.

**Example:** If we give a single group of patients a new, experimental drug and if they improve, then their improvement might be due to the *placebo effect*. In this case, the *placebo effect* is a *confounding variable*. This is why we want to also have a control group and do a double-blind study.

What is a *completely randomized experimental design*?

What is a *completely randomized experimental design*?

In a *completely randomized experimental design*, subjects are assigned to different treatment groups through a process of *random selection*.

What is a *randomized block design*?

(1.5)

What is a *randomized block design*?

A *block* is a group of subjects that are similar. In a *randomized block design*, blocks are defined, and then subjects are randomly assigned within the different blocks.

What is a *randomized block design*?

A *block* is a group of subjects that are similar. In a *randomized block design*, blocks are defined, and then subjects are randomly assigned within the different blocks.

**Example:** A simple example would be to create two *blocks*, male and female, and assign subjects randomly within the two different groups.

What is a *rigorously controlled design*?

What is a *rigorously controlled design*?

In a *rigorously controlled design*, subjects are carefully assigned to treatment groups based upon given characteristics.

What is a *rigorously controlled design*?

In a *rigorously controlled design*, subjects are carefully assigned to treatment groups based upon given characteristics.

**Example:** We might, for instance, control the effect of age by making sure that subjects of all ages are assigned to the different treatment groups.

What is a *matched pairs design*?

What is a *matched pairs design*?

In a *matched pairs design*, exactly two treatment groups are created, and subjects in the two groups are matched in pairs based upon similar characteristics.

What is a *matched pairs design*?

In a *matched pairs design*, exactly two treatment groups are created, and subjects in the two groups are matched in pairs based upon similar characteristics.

**Example:** If we have two treatment groups and if we make sure that for every person in one group there is a person of the same age and gender in the other group, then we have a *matched pairs design*.



What is *sampling error*?

(1.5)



What is *sampling error*?

A *sampling error* is a difference due to random chance between the sample result and the true population result.



What is *nonsampling error*?

(1.5)



What is *nonsampling error*?

A *nonsampling error* occurs when the sample data are incorrectly collected, recorded, or analyzed.