

Data is the basis for everything we do in statistics. Every method we use in this course starts with the collection of data. **Observational Studies** and **Experiments** are the two basic types of statistical studies that we use to obtain data.

### Observational Studies

**Observational Study:** The subject is observed. **No attempt to modify or treat the subject in any way is allowed.** You can measure and record information about the characteristics being observed.

The most basic Observational Study occurs when a subject or subjects are observed at the current **point in time**. They are observed, measurements of the trait being observed are taken and recorded based on **the current point in time**. This type of study is called a **Cross Sectional Study**. It is a view of a cross section of people taken at a **single point in time**.

Data on the subjects may be collected **from the past** by reviewing old records, interviews or other studies previously performed and recorded. This type of study is called a **Retrospective Study**. **There are some drawbacks to this type of study.** The person doing the study must rely on the data that is available from past records. A topic that is of interest today may not have been of interest in the past so no data may have been collected. You have no control over how the data was collected or what groups the data was collected from.

A **Longitudinal Study** is a study where the subject or subjects are observed at regular intervals over a period of time and measurements of a selected trait are recorded at each time period **starting from the present time**. Longitudinal Studies are very common in the social sciences to help discover trends that may develop over time. They are also common in drug testing. The measurements of the effects of a drug taken over a long period of time will be a longitudinal study. The study is conducted to try and determine if the trait being observed is changing over time.

An example of a **Longitudinal Study** would be to observe a group of people who subscribe to the Sacramento Bee now to see if they continue to subscribe to the Sacramento Bee over the next 5 years. You may use this study to help determine if the number of people that subscribe to the Sacramento Bee is diminishing as time goes on. You can think of this kind of study as a series of a cross sectional studies of a the same group taken at regular intervals over a period of time.

### Retrospective Study versus Longitudinal Study

A researcher has much more control if they start a study in the current time period and select the traits to be measured and the group to be measured. This makes a **Longitudinal Study** desirable but it will take a long time a long time to collect the data. One advantage of a **Retrospective Study** is that if useful data can be found the results of the study can be done right away.

## Experiments

**Experiment:** The subject is given a **treatment** and then the subject is observed to see what effect the treatment has on a specific trait.

The basis for an experiment is that a subject or group of subjects are given some kind of **treatment**. This treatment could be a drug, a physical activity they are required to perform, an educational program that they take part in or any other action that they perform or is performed on them. After the time period prescribed for the treatment **the trait being studied is measured**. The experimenter will then try to determine if the treatment had an effect on that specific trait. A **Clinical Trial** is a special type of experiment that involves a drug or medical treatment.

There are a few components that must be included in an experiment for the measurements to be considered valid.

1. The subjects for the experiment must be chosen through a **Random Sample**.
2. The subjects must be randomly divided into two groups, the **Experimental Group** (or treatment group) and the **Control Group** (or Placebo Group).
  - A) The **Experimental Group** (or treatment group) is the group of subjects that will be given the real treatment.
  - B) The **Control Group** (or Placebo Group) is a group of subjects that are given a “fake” treatment called a **placebo**. The placebo treatment is a treatment that looks just like the real thing but is in fact **the same as not being treated**. If the experimental group is given a pill then the placebo group would be given a “sugar pill” that looks just like the real pill. That way the control group thinks it is being treated but in fact they are not. The control groups results are then compared with the treatment groups results. If the treatment is effective then the results should be better for the treatment group than the control group. If the results of the treatment and control group are close to the same than the treatment can be determined to be ineffective.

## Blinding

If the **Control Group** (or Placebo Group) knew that they were being given a “fake” treatment then the results would not be valid. It is important that the **Control Group** (or Placebo Group) be “blind” to the fact that they are not getting the real treatment. The “fake treatment” must be given so that the **Control Group** (or Placebo Group) thinks it is getting the real treatment.

If the person giving the treatment knew that the treatment was fake they may act in a way that let the subjects know they were in the **Control Group** (or Placebo Group). That may create a problem with the results if the control group knew that the treatment was “fake”. A nurse may take great care in getting an accurate measurement for the treatment group because she knows that they are getting the “real” drug but make a half hearted effort for the **Control Group** because she knows they are not really being treated. If that happened then the results would not be valid.

## Double Blind Experiments

If both the persons conducting the treatment and the subjects being treated are “blind” to who is in the **Experimental Group** (or treatment group) and who is in the **Control Group** (or Placebo Group) then the experiment is considered a **Double Blind Experiment**.

The best experimental design is for all the subjects to be from a random sample. The Experimental Group and the Control Group must be placed at random into each group. Neither the subjects being treated or the persons treating the subjects should have a knowledge who is in either group.

## Samples

The basis for every Experiment or Observational Study is a random sample of subjects taken from an entire population of possible subjects. The results of an experiment based on a random sample allows us to take the numerical properties about the **sample** and use those numerical properties to **infer what numerical properties might be true** about the **entire population**.

## Non Sampling Error

A **Non Sampling Error** occurs in the collection phase of the experiment. The data was recorded incorrectly. The sample data was collected from a sample that was not random. The data was collected but the measuring device was not working correctly. Some data was lost during the collecting phase or recording phase. These types of errors create errors in the data itself. Any sample with this type of error cannot be used and must be ignored.

## Sampling Error

**(the difference in sample and population parameters)**

The numerical properties about a sample cannot be expected to be the same as the numerical properties of the entire population. The difference between the numerical properties about the **sample** and the numerical properties of the **entire population** is called **Sampling Error**.

A **Sampling Error** does not mean that an actual error was made in the techniques used in the experiment. It is a simple acknowledgment that the numerical properties about a **sample cannot be expected** to be an exact match for the numerical properties of the **entire population**. This type of error cannot be corrected or eliminated but the size of the error can be reduced with large sample sizes.

## Common Errors in the Collection of Data

There are many possible errors in the creation of a sample, the collection of the data and the conclusions made about the results of the experiment or study.

**Self Selected Sample:** It is common to try and collect information by asking many people to respond to a survey. The data from those that do respond may then be used to make a conclusion. No matter how many people respond this is a serious error. There is no control over who responds. The people who do respond to the survey **self select** that they wanted to provide the information. This self interest tends to slant a survey towards people who are strongly for or against a topic.

**Self Interest Study:** A Self Interest Study is any study where the **person conducting the study** has a something to gain from the outcome of the study. The person who designs the study, selects the random subjects, collects the response and states the results of the study should have not be in a position to benefit in any way from the outcome. The results of such studies are always in question.

**Samples without basic controls:** Many surveys do not prevent a person from responding to the survey multiple times. There must be a way to verify that a person does not respond more than once. This is a serious error. Many surveys cannot guarantee that the person selected to sample is indeed the person who completes the survey. There must be a way to verify that the person is who gives the information is who they say they are. Many people deliberately report information on a survey that is not accurate. They are trying to influence the survey by reporting data that will support a given point of view. There must be a way to verify that the data collected is accurate. **It is important that the researcher select the subjects themselves, verify that each response is tied to a single subject and that the data given is an accurate response.**

**Small Sample Size:** The size of the random sample is very important. It is very hard to expect that a small number of subjects from a large population can represent the entire population. There is no simple answer as to how large the sample should be. A much deeper discussion of this topic will be given in the later chapters of this course. At this point in the course it is best to say **“the larger the sample size the better”**. As the sample size gets closer and closer to the population size the sampling error is reduced.

**Not a Random Sample:** It is beyond the scope of this course to cover the complexity of what a true random sample is. For this course we will say that a random sample is made in a manner that allows any one subject in the population to have **the same (or equal) chance at being selected**. One way to do this is to put the names of every subject in the population on a separate piece of paper and place them in a bag. After shaking the bag for a long time one name is selected at a time. This will be considered to be a random sample. Putting the names of every subject in the population into an Excel spreadsheet with numbered cells and using a random number generator to select name at random can also be considered to be a random sample. There are other techniques that will produce a random sample. For the limited purposes of this course we will simply state that a random sample was taken and limit the description as to how the random sample was collected.

**Loaded Question: Loaded Questions** are questions that are asked in a manner that makes the desired response far more likely to happen than any other response. This is a common source of sampling error. When any study is published the exact question that was asked to the subjects should be attached. The way a question is asked and the wording used can have a large impact on the subjects response. If you were asked to respond to the question "do you support oil drilling off the coast of Santa Barbara that may cause major damage to the environment and cause many fish and birds to die" it would be very hard to respond in a positive way. If you were asked "do you support drilling an oil well off the coast of Santa Barbara to relive the natural seeping of oil that is causing damage to the fish and bird life" you may be far more positive in your reply. Avoid asking questions in such a manner that the response is predictable based on the way the question is asked.

**Missing Data:** Once a response is collected it must be included in the total responses. If you have a response from 100 subjects then the data should reflect all 100 responses. It is not acceptable to leave out a response. Some researchers have been caught "losing" data that did not help their desired outcome. It is a temptation to do this if the results of the study do not support the conclusion you desire. At that point you may be tempted to find a reason to remove some responses and then use the new data set to reach the conclusion you desired. In the last few years several prominent researchers have lost their research positions at major institutions for doing this very thing.

## Accurate Numbers

The foundation of every conclusion we make about the sample or population is based on the data collected. It is easy to see that we must put an extremely high importance on the data being accurate. **Accurate Numbers** are numbers that count or measure **exactly what the true count or measure is**. A person shows a bartender a driver's license that states their age as 21. This is an accurate number only if that person is really 21 years old. If, in fact, they are really 18 then that number is not accurate no matter what the document claims. Every effort should be made to verify that the numbers recorded do count or measure **exactly what the true count or measure is**.

You should make an attempt, within reason, to verify the accuracy of the number. What should be done if you feel that a number collected for your study is not accurate? That is a difficult question. If you decide that the number may not be accurate you may exclude the data from the results, but you must **make a clear note of that action and include the reason for this action**. This will allow readers of the study to decide for themselves if they want to use the results of the study with the deleted numbers

A sign at the entrance to the City of Folsom states the population as 72,591 and has done so for since 2008. Wikipedia states that the city has a total area of 21.7 square miles of land and 2.4 square miles of water. These numbers seem very impressive. If the sign stated 72,000 then you may think that the number was just an estimate. The number of digits in 72,591 or the decimal place in 21.7 may make people feel the number is more accurate. This is not the case. The number of digits or decimal places cannot determine the accuracy of a number. The accuracy of a number is based on whether the number counts or measures **exactly what the true count or measure is**.

## Precise Numbers

**The number of decimal places in the number** is a measure of how precise a number is. If a balance in the chemistry laboratory states the mass of to the first decimal place (i.e. 2.3 grams) then we say the balance is precise to the tenth of a gram. If a balance in the chemistry laboratory states the mass of to the second decimal place (i.e. 2.36 grams) then we say the balance is precise to the hundredth of a gram. Each additional decimal place that the balance reports increases the precision of the device. The college has a balance that is precise to 4 decimal places. It must have a glass case to prevent the air currents in the room from having an effect on the measurement.

## Accurate versus Precise

It is important to understand the difference between precise and accurate. If the scale is precise to the third decimal place then measurement will be stated as 2.543 grams. If the scale has been damaged and it is not able to record the true measurement then the precise numbers are useless. The most important thing is to collect accurate data. The device that measures the material needs to be as precise as the purpose of the study requires. Do not let a precise number with lots of decimal places impress you unless you are sure the number is accurate. The college pays a company each summer to examine each balance in the laboratory and check for the accuracy of the measurements. The company cannot improve the precision of the balance, that is based on the design of the balance. The company can use a known mass to test the accuracy of the measurement and correct the balance if it is no longer accurate.

## Causality and Correlation

### Causality

**Causality** is the relationship between an action and an outcome in such a way that the **action is the CAUSE of the outcome**. We often call this “cause and effect” and say the cause precedes and is responsible for the effect.

Law and Order detective Lenny Brisco finds the dead body of a 60 year old man laying on the ground in a large pool of blood. Lenny finds that a nasty pending divorce, a balcony on the 30th floor and lots of alcohol on the balcony are a part of the story. Lenny sends the body to the morgue for a coroner to perform an autopsy. The purpose of the autopsy is to determine the **cause of death**. To Lenny it seems simple, the large amount of blood on the ground indicated that the person had bled to death. The coroner points out that the man’s body was smashed and broken from falling off a 30 story building and had no traces of alcohol in his system. The coroner states that **the fall from 30 stories caused the mans instant death**. The blood was just part of what happened after the body hit the pavement and died from the impact.

### Correlation

An event is **correlated** with an outcome if when **the event is present then the outcome is also present**. The event and the outcome may both be present but the event may not be the cause of the outcome. In the case of the dead body Lenny Brisco found we know that **the coroner stated that the fall caused the persons death**. What about the pool of blood? A pool of blood has a correlation with a fall from a large height. Every body that falls that far will have a pool of blood present but the bleeding will not have caused the death. **All the events surrounding the death are correlated to the death by time and location**. The divorce, the alcohol, the balcony, the fall, the pool of blood and the smashed body are all correlated to the death. In this case **the impact from the fall was the only cause of death and the other things are simply correlated with the death**.

### Correlation does not imply Causality

Correlation is a matter of two events being connected by a common factor. Two events can be correlated with each other by happening at the same time, occurring at the same location, or both being related to a third event. It is a common error to note that that two events happened at the same time and think that one **caused** the other. One event may be correlated with a second event but may not be the cause of the second event. Many people make the mistake of seeing two events related by a common factor and concluding that one event was the cause of the second event. **This is often not the case**.

Events that are correlated in strange or obscure ways seem interesting to many people. They want to see if one event caused the other event. The search for the cause of an event is much more complicated than just stating that the events are correlated by a common factor. You must prove that the one event was the cause of the other event. If you can show that **one event was the cause of the other event** then you have proven causality.

## Confounding

The desire to determine the exact cause of an event is one of the reasons we perform experiments. **Confounding** occurs when **more than one action may have an affect on an outcome**. The cause for the high level of unemployment at this time in our nation is not agreed upon by economists, or government leaders. It seems clear that there are many causes. A study or experiment can be designed to determine which of the many causes is **most responsible** for the outcome.

We know that smoking is a factor in lung cancer. Contact with pesticides also can also be a factor. Genetics can also be a factor. It is clear that no single factor is the cause of all lung cancer. This is the reason that cigaret companies fight lawsuits brought against them by smokers who are dying of lung cancer. Cigarette companies use the effect of confounding as a defense in lawsuits against them to show that factors other than smoking can cause lung cancer.

The design of a study or experiment should consider that **there may be more than one factor that could be the cause of a given outcome**. Try to design the experiment in such a manner that only one factor is being tested at a time. The other possible factors are not allowed to have an effect on the outcome. It may require sever experiments to study all the possible factors. In this manner, the several factors that can affect an outcome can be studied and measured.

## Hidden or Lurking Variables

It's hard to believe but detective Lenny Brisco finds another dead body of another 60 year old man laying on the ground in a pool of blood. Lenny finds that a nasty pending divorce, a balcony on the 30th floor and lots of alcohol on the balcony are also a part of this story. Lenny sends the body to the morgue for a coroner to perform an autopsy. The purpose of the autopsy is to determine the **cause of death**. To Lenny It seems simple. The events surrounding this death seem to be the same as the first death. Lenny thinks that the cause of this death will be the fall just as it was with the first death. What Lenny cannot know is that there are facts in this case that he is not aware of that will make the cause of death differ from the first death.

The divorce, the alcohol, the balcony, the fall, the pool of blood and the smashed body are all correlated to the death. An autopsy shows that the person had been poisoned two hours before the fall and that **the poison was the cause of the death**. The fall was staged to make it look like a suicide or an accident due to heavy drinking. The **poison was the cause of death** and the divorce, the alcohol, the balcony, the fall, the pool of blood and the smashed body are **are simply correlated with the death**.

The detective Lenny Brisco reads the autopsy report and disagrees with the stated cause of the deaths. He says that **the cause of both deaths were the affairs** that both of the 60 year old men were having with their 18 year old secretaries. The poison, the falls and the pools of blood were just a simple correlation with the affair. Lenny says that "Heaven has no fury like a women spurned"

## Lurking Variables

Read the following events and the conclusion that is made as to what factor may have been the cause of the event. Try to find a second factor that **better explains** the cause of the event. We call the second variable that better explains the cause a lurking variable.

A) A study found that the patients at nursing homes are about 80% women.

**Conclusion:** Women prefer nursing homes more than men do.

The preference of women is not the cause that 80% of the patients at nursing homes are women. Women live longer than men do. The percentage of men and women may have started out the same at age 70 but over time the men die off and the women continue to live. As time goes on the nursing home has a mix of 70 year old men and women but most of the older patients are women. The longevity of women causes the percentage of women to be high.

B) A study found that students that did not work had a higher GPA than those that did work.

**Conclusion:** Quitting work will cause your GPA to rise.

The reason that students who do not work have a higher GPA is based in the fact that they use the extra time that was spent working doing more homework. If you do not do more homework when you quit your job you cannot expect your GPA to go up. The amount of time spent on your studies for the class causes the high GPA.

C) A study found that people who drive a Mercedes Benz to work make more money than those that do not.

**Conclusion:** Driving a Mercedes Benz to work will cause you to have a high income.

Driving a Mercedes Benz to work is not the cause of the high income. The people who already have a high income can afford an expensive car. The skill or education that leads to a good paying job is the cause of the high income. The car just follows the education and the job.