

Chapter 1: Introduction to Statistics

Section 1 – 1: Descriptive Statistics:

The first 3 chapters of this course will develop the concepts involved with **Descriptive Statistics**.

Descriptive Statistics is the act of
collecting, organizing, displaying and summarizing
information.

Collecting Information:

Statistics starts with the collecting of information. The act of **collecting information** can be as simple as recording the age of every student in your current statistics class. It can be far more complicated than that. If you wanted to find out the age of every student at Folsom Lake College for the Spring 2011 semester several problems must be considered. Should you count every student who is enrolled on the first day of the fall semester? Should you count the students who enroll during the first weeks of class? What about students who drop in the first few weeks or never show up? You must decide on how to **clearly describe** the people you are going to include. **Collecting data begins with a clear precise description of the information you want to collect.**

The actual collecting of the data can also pose a problem. Will the college make the data available to you? If not, how will you go about collecting the data? Even if you can and do collect the students ages, **can you trust** the information? People can and do give information that is untrue. Not everyone lists their age truthfully.

Information can be recorded incorrectly or hard to read. For example, a student writes down their age as 78. Was the student honest in declaring their age? Was the age of 78 a misprint. The seven in the 78 could be correct but it also could have been an 18 year old whose age was written poorly and looks like a seven.

The act of **collecting information** is really an art. It is very important to be very specific in the description of who, what, when or where the information describes. It is also very important to collect and record the information in a manner that helps ensure that the information reflects an honest record. Verifying the accuracy of the data as it is collected and recorded can be a major cost in the process.

The methods of collecting and verifying information are not within the scope of this course. The student is asked to consider that all information given in this course has been collected and verified with methods that are consistent with accepted practices.

An Example of Problems in Collecting Information:

State, county and local church records in Texas contain records of a Joe Hardy, a Joseph Hardy, a Joe E. Hardy and a J. Edward Hardy all born in 1891. These same records show the birth of a Julia Baskin, born in 1894, a Julia Baskin born in 1893, and a Elisa Julia Baskin born in 1894. The various local records also show several combinations of these people marrying each other in either 1919 or 1920. The original records were handwritten. They were later stored on micro film and now exist in several digital formats. This information has been kept for all these years as an official record of two people who lived, married, gave birth and died in the towns of Oklahoma and northern Texas.

1. **Is the information** that was collected, stored and protected all these years **correct**?
2. The original collection of a piece of information is called the **first generation record**. Each time that the same information is rewritten, copied, or compiled along with other information starts a new generation for that piece of information. Does information become more accurate as the generations increase?
3. **What would a reasonable person do with the information** mentioned above if they were compiling their own record of this information?
3. If someone believes that there were only 2 people reflected in all these documents, **how should they correct the information**? Should they edit the original first generation documents? Should they create a new generation of records with a record of the birth, marriage and death using the dates they think are correct?

Organizing Information:

Many methods exist to **organize information**. A common way to organize information is to place the information into a list. The list may be put into a table format. Putting the data into a table in an **Excel spreadsheet** allows the list to be sorted by alphabetic or numerical means to increase the usefulness of the data.

Sorting data after it has been put into a table may be the most common way to organize data today. Sorting data into groups based on common attributes is very useful. Sorting data into groups based on the month of the year may help you see trends over time. Sorting data into groups based on the state of birth may help you see trends based on geographic location. Sorting data into groups based year of birth may help you see trends based on age.

Database programs like **Access or Oracle** allow for more complicated searches of the information. These programs allow you to search for all of the information that has several common attributes. These programs also allow the user to control the way the results of the search is displayed. Database programs like these are becoming the standard for the storage and display of large amounts of information.

Summarizing Information

It is very common to view **small amounts** of data in a **sorted list**. The use of a sorted list becomes a problem when there is a large amount of data. It is also common for large data sets to have many repeating values. When large data sets with repeating values occurs, it is common to summarize the data into two types of tables.

A **Frequency Table** takes all the repeats of a value and **lists each different value once** and records how often (how frequent) each of the different values occurs.

A Frequency Table

x = number of cars owned by your family	Frequency of x
0	1
1	2
2	7
3	6
4	4

When the data contains a large number of different values then collecting the data into groups or **classes** based on a **range of values** allows the data to be viewed in a smaller table that summarizes the information. This type of table is called a **Class Frequency Table**.

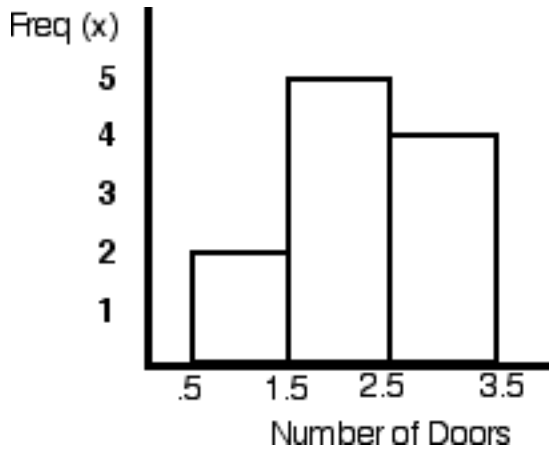
A Class Frequency Table

x: age of child	Frequency of x
0 – 4	2
5 – 9	8
10 – 14	5
15 – 19	10
20 – 24	5

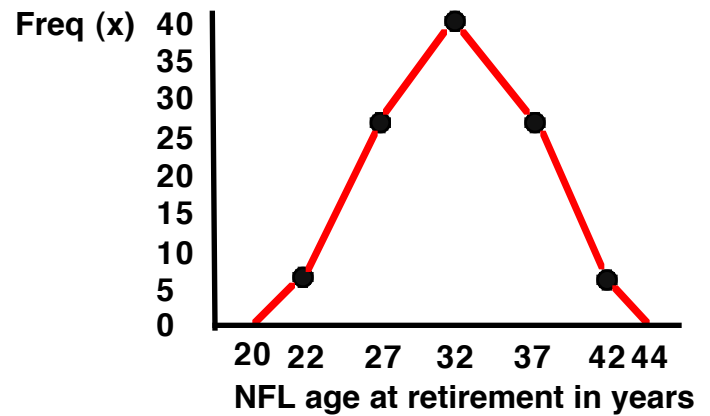
Displaying Information

After data has been put in a table and grouped into a summary table it is common to use a graph to provide a **visual view of the data**. The most common graphs used are **bar graphs, line graphs or pie charts**. Excel and other software programs provide these three graphs as well as a wide variety of other types of graphs. Each type of graph provides a different way to view data.

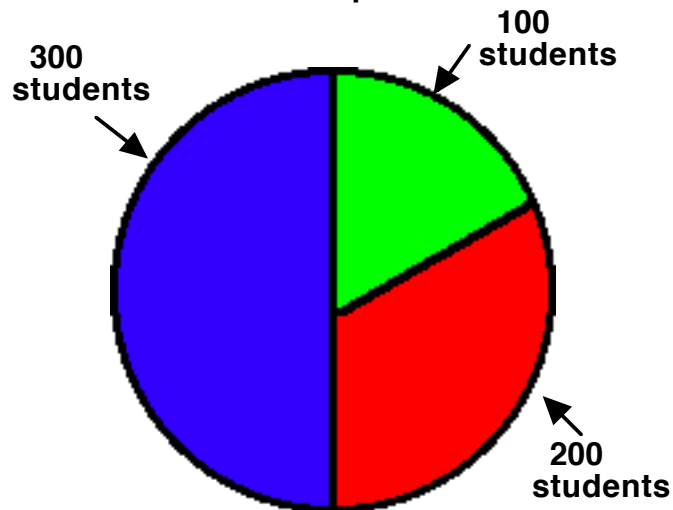
Bar Graph



Line Graph



Circle Graph



**Favorite Color between
Red, Blue and Green
based on 600 students**

Population

A **Population** is the **entire collection** of things or people that will be studied. A populations can consist of a very few members or have a very large number of members. If you are recording the height of every person who played in the NBA this season then that population would be much larger then if you were recording the height of every member of your family. The key idea is that the population must include **every member that meets the description of that population**.

Sample

A **Sample** is a collection that consists of only **part of the entire population**. If the **population is defined to be every person in this statistics class** then a **sample of that population could consist of only the students that sit in the front row**. If the **population is defined to be every person who was in the school play** then a **sample of that population could consist of only the people whose names begin with the letter A**. Each set is a sample because each set represents only **part of the entire population**.

Population versus Sample

A Population is the entire collection of things that will be studied.

A Sample is part of an entire population.

Population (P) or Sample (S) Examples

Classify the following information as being found from a **population (P)** or **Sample (S)**.

- A) ____ I tested 2 students in a class of 30 and they were both nearsighted.
- B) ____ The Major League Baseball Teams in California are, Oakland Athletics, San Diego Padres, the San Francisco Giants, the Los Angeles Dodgers and the Los Angeles Angels.
- C) ____ I selected Mary and Bob out of the 40 students in my math class to grade their homework.
- D) ____ Sue selected 16 of the 50 people at the Bookstore to see how long it took them to buy their books.

Answers:

- A) S B) P C) S D) S

Census

A **Census** is a collection of **information about a characteristic** that was collected from **the entire population**.

When we collect **information about every member of the population the information collected is called a census**. If the population is defined to be every person in a math class that took Test One then a census of that population would be **all the actual test scores for every student** in the class that took Test One. The same population could have a census taken of the **time it took each student to complete the test**. The same population produced two different sets of data and each set of data was a census of the entire population of class members.

Examples

Yes (Y) or No (N). Was a **Census used** to find the following information.

- A) ____ Sue reported that all 3 of her children used contacts.
- B) ____ Sam asked 25 of the 40 students in his class if they are left handed.
- C) ____ Mary observed that every person enrolled in her class had met the prerequisite.
- D) ____ John watched 155 people out of the 4300 people at the Mall and recorded how long it took each of the 155 people to find a parking place.

Answers:

- A) N B) N C) Y D) N

It is very hard to take a real Census

Every 10 years the government takes what it calls a population **census** to determine the number of people that live in the country on a state by state basis. **Is this a real census?** There are several reasons that this is not an actual census. It takes so long to collect the data that people being born and people dying cannot be correctly accounted for. People move in and out of the country, as well as from state to state during the census process so that data never includes all the population. To be a Census **every member of the population** must have it's information collected. You cannot skip a few people. This means that many collections of data that claim to be a census are really a sample of the entire population and not a census.

Parameter **(Numerical Measurements of a Population)**

A **Parameter** is a **measurement** describing a **numerical property** about a **population**.

There are **three numerical Parameters** about a **population** that will be of interest in this course:

1. Population Average: The **average value** for all the **numerical values** in the **population**.
2. Population Proportion: What percent (or **proportion**) of the population **has a given attribute**.
3. Population Variance: How much the **numerical values** in the population **vary from the average**.

It is often difficult, or impossible to find the values for the population parameters because collecting the population data may prove too difficult or expensive. It is possible to take a sample that consists of part of the entire population.

Statistic **(Numerical Measurements of a Sample)**

A **Statistic** is a **measurement** describing a **numerical property** about a **sample**.

There are **three numerical Statistics** about a **sample** that will be of interest in this course:

1. Sample Average: The **average value** for the **numerical values** in the **sample**.
2. Sample Proportion: What percent (or **proportion**) of the sample **has a given attribute**.
3. Sample Variance: How much the **numerical values** in the sample **vary from the average**.

Even though all the sample values come from the population, the sample does not contain all of the population values. It is easy to see that the **numerical properties** about a **sample will not be the same as the numerical properties** about a **population**.

Parameter versus **Statistic**

A **Parameter** is a **measurement** describing a **numerical property** about a **population**.

A **Statistic** is a **measurement** describing a **numerical property** about a **sample**.

Classify the following as a **Parameter (P)** or a **Statistic (S)**.

- A) ____ 12 students in the graduating class of 400 FLC are transferring to U.C. Davis
- B) ____ I asked every student enrolled in my class if they had done their homework and 45% of them said that they had not.
- C) ____ Every head coach in the NFL is male.
- D) ____ During the midnight opening of the last Harry Potter film 89 people out of the 300 in attendance had a wand with them.

Answers:

- A) S B) P C) P D) S

Inferential Statistics

Inferential statistics involves the use of methods that allow us to take **numerical properties about a sample** of the population and then use those numerical properties to **infer what numerical properties** about the **entire population MIGHT be true**. **Inferential Statistics** also involves methods that produce a measure of **how reliable the inference about the entire population is**.

Samples must represent the entire population

The sample data is the basis for inferring what the population data MAY be. For this reason it is extremely critical that the sample data represents the entire population. A sample should have all the different components of the population represented in about the **same proportion** as they exist in the population. If the sample does not represent the population very closely then any inference about the population based on that sample will not be dependable.

If the population being examined is the students in this statistics class then the sample taken must be taken so it represents all those students. If the sample contained all females while the population contained an equal number of male and female students then that sample would not represent the population very well. If the population is all the registered voters in Folsom California and a sample of those voters contains only people over the age of 65 then that sample would not represent the population very well.

Random Sample

A sample is considered to be a **Random Sample** if every individual member of the population has an **equal chance of being selected**

A sample must be taken from the population in such a way that the selection process produces a **Random Sample**. No inferential statistics techniques exist that can produce dependable results about the population based on a sample whose selection process did not produce a **Random Sample**. If the selection process **DOES NOT** produce a **Random Sample then the inference process has no value**. In almost every statistical procedure used in this course the first requirement is that the sample data collected is from a random sample as defined above. In this course we will not state the method used to get a random sample for each problem. We will simply state that the data collected represents a random sample.

A simplistic way to think of creating a random sample is to put **every member of the population into a bag** and **shake the bag** until all the members are well mixed. Then reach in a **take out as many members of the population as needed** for your random sample.

The concept of a truly random selection is far more complicated than one may think. Higher level math is required to discuss the concept. Major developments in this area cause some researchers to argue that there is no such thing as any sample being truly random. These researchers say that within every random data set there are areas that are very orderly and thus any sample taken will never be truly random. Even the random number generators used by Intel and gambling casinos have been shown to not be as random as previously thought. It is beyond the scope of this course to discuss even a rudimentary level discussion of this concept. An advanced statistics course in experimental design can be taken that will present many methods that can be used to help ensure a random sample.

Example 1: An Inference about a Population Mean (Average)

I want to know the average number of units taken by college students in California during the Fall 2012 semester. Due to the cost of collecting all the information and the fact that students are adding and dropping classes it would not be possible to collect accurate and up to date population information. I collect a **random sample** of 4000 college students in California during the Fall 2012 semester and record how many units they were enrolled in. The **average** number of units **for the 4000** sampled students was found to be 12.5 units.

Would the 12.5 units taken by the students in the sample be the exact average for **all the students** enrolled in California during the Fall 2010 semester? It should not surprise you if I said that you **CANNOT use a sample** to get an **exact value about the population**. Information based on 4000 students cannot be expected to be the same as information based on the entire the entire population. **Inferential statistics** allow us to take the fact that the **sample of 4000** students were enrolled in an average of 12.5 units and use that sample average to **infer what numerical properties MIGHT be true** about the **entire population**. **Inferential Statistics involves** methods also produce a measure of **how reliable the conclusions about the population parameters are**.

The methods of **Inferential statistics** allow us make the following statement.

I am 95% confident that the average number of units taken by all the college students in California during the Fall 2012 semester falls between 11.8 and 13.2 units

Example 2: An Inference about a Population Proportion

I want to know the percent (or proportion) of the population of registered voters the United States that support a bill in Congress that would sell the state of Alaska to Canada. Due to the cost of collecting all the population information and the fact that some voters are changing their decisions every day based on TV adds and talk shows it would not be possible to collect accurate and up to date population information. I collect a random sample of 500 people who are registered voters in the United States and record if they are in favor of selling Alaska to Canada. 56% of the people sampled said they were in favor of selling Alaska. 85% of those sampled said that they would be in favor if Sara Palin was part of the deal. (It's a joke)

The methods of **Inferential statistics** allow us make the following statement

I am 99% confident that the percent of registered voters in the United States that support a bill in Congress that would sell the state of Alaska to Canada falls between 52% and 60%.

Example 3: An Inference about a Population Standard Deviation

I want to know how much the actual volume of Dr. Pepper varies from the reported volume of 20 oz. Due to the cost of collecting all the population information and the fact that bottles are constantly being produced and used it would not be possible to collect accurate and up to date population information. I collect a **random sample of 50** 20 oz. bottles of Dr. Pepper and personally record the actual volume and then drink each sample. It is a tough job but someone has to do it. I then find how much the volume of these 50 bottles varies from the reported 20 ounces.

The methods of **Inferential statistics** allow us make the following statement

I am 98% confident that the amount the volume of Dr. Pepper varies from the reported volume of 20 oz. falls between **.1 oz. and .3 oz.**

Chapters 1 to 3 in this course will involve the study of descriptive statistics.

Chapters 5 to 9 in this course will involve the study of Inferential statistics.