

We will see vocabulary and concepts that follow us through the semester.

Introduction to the Practice of Statistics (Section 1.1)

Definition: Statistics: the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of the confidence in our conclusions.

Definitions: The entire group to be studied is the **population**. This could be a group of people but it could also be all the cars on a highway or the crayons in a box. An **individual** is a person or object that is a member of the population. Since it's often cumbersome to ask our question of every individual (called a **census**), we will sample to get answers. A **sample** is a subset of the population that is actually studied. We then infer about the entire population based on what we learned from the sample.

Definitions: Descriptive Statistics versus Inferential Statistics:

A **statistic** is a numerical summary of a *sample*. **Descriptive statistics** consist of organizing and summarizing data. It uses numerical summaries, tables, and graphs.

Inferential statistics uses methods that take the result from a sample and extend it to the population. It also provides a measure of the reliability of the result. A **parameter** is a numerical summary of the *population*. It is this information that statistics can help us estimate by looking at the sample.

Qualitative versus Quantitative Variables:

Definition: Variable: A **variable** is a characteristic of the individuals within the population. Examples include hair color, sex, income, IQ, amount of disposable income, number of televisions, etc. Variables are called that because they vary from individual to individual. That variation is a big deal in statistics.

Definition: Qualitative, or categorical, variables allow for some classification of individuals based on some attribute or characteristic.

Definition: Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

expl 1: Consider the following variables. Which are qualitative and which are quantitative?

- | | |
|---|--|
| a.) Temperature | d.) Zip code |
| b.) Favorite music group | e.) Length of movie |
| c.) Number of cell phones owned by a family | f.) Number of hours per night a college student sleeps |

Definitions: Discrete versus Continuous Variables:

Both are types of quantitative variables. The difference here is what kinds of values can be obtained from the question we ask.

Definition: Discrete variables have either a finite number of possible values or a countable number of possible values. The term “countable” means you got the number from counting something, such as 0, 1, 2, 3, etc. A discrete variable *cannot* take on every possible value between any two possible values.

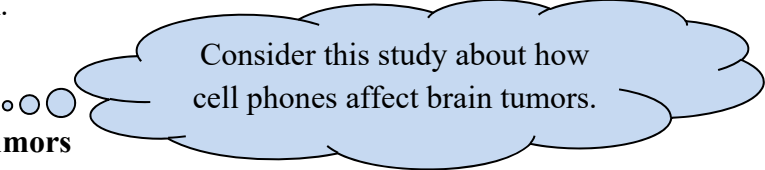
Definition: Continuous variables have an infinite number of possible values that are *not* countable. A continuous variable may take on every possible value between any two values.

expl 2: For the quantitative variables from the last example, determine if each is discrete or continuous. (We do not do this with the qualitative variables so they are crossed off.)

- | | |
|---|--|
| a.) Temperature | d.) Zip code |
| b.) Favorite music group | e.) Length of movie |
| c.) Number of cell phones owned by a family | f.) Number of hours per night a college student sleeps |

Observational Studies versus Designed Experiments (Section 1.2)

You'll hear that coffee or wine reduces heart disease. You'll hear cell phones cause cancer. You'll hear breastfeeding improves IQ. We'll study these observational studies and designed experiments to determine if we can trust them.



Consider this study about how cell phones affect brain tumors.

EXAMPLE: Cellular Phones and Brain Tumors

Researchers Joachim Schüz and associates wanted “to investigate cancer risk among Danish cell phone users who were followed for up to 21 years.” To do so, they kept track of 420,095 people whose first cell phone subscription was between 1982 and 1995. In 2002, they recorded the number of people out of the 420,095 people who had a brain tumor and compared the rate of brain tumors in this group to the rate of brain tumors in the general population. They found no significant difference in the rate of brain tumors between the two groups. The researchers concluded “cellular telephone was not associated with increased risk for brain tumors.” (Source: Joachim Schüz et al. “Cellular Telephone Use and Cancer Risk: Update of a Nationwide Danish Cohort,” *Journal of the National Cancer Institute* 98(23): 1707-1713, 2006)

Definitions: Explanatory and Response Variables: Remember a variable is just some characteristic of the individuals in the population, like age or cell phone use or cancer rate. The **explanatory variable** is the characteristic that we think may influence another characteristic, which is the **response variable**.

expl 3: In the above study, what is the explanatory variable? What is the response variable?

The above study is an example of an **observational study**. The researchers simply observed the participants. They did *not* intervene in their subjects' lives and instruct them to do anything differently.

Three Types of Observational Studies:

1. Cross-sectional Studies Observational studies that collect information about individuals at a specific point in time, or over a very short period of time.

2. Case-control Studies These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals who have certain characteristics are matched with those who do not.

3. Cohort Studies A cohort study first identifies a group of individuals to participate in the study (the cohort). The cohort is then observed over a long period of time. Over this time period, characteristics about the individuals are recorded. Because the data is collected over time, cohort studies are **prospective**.

Here is another group of researchers approaching the same problem.

EXAMPLE: Cellular Phones and Brain Tumors

Researchers Joseph L. Roti and associates examined “whether chronic exposure to radio frequency (RF) radiation at two common cell phone signals—835.62 megahertz, a frequency used by analogue cell phones, and 847.74 megahertz, a frequency used by digital cell phones—caused brain tumors in rats. The rats in group 1 were exposed to the analogue cell phone frequency; the rats in group 2 were exposed to the digital frequency; the rats in group 3 served as controls and received no radiation. The exposure was done for 4 hours a day, 5 days a week for 2 years. The rats in all three groups were treated the same, except for the RF exposure. After 505 days of exposure, the researchers reported the following after analyzing the data. “We found no statistically significant increases in any tumor type, including brain, liver, lung or kidney, compared to the control group.” (Source: M. La Regina, E. Moros, W. Pickard, W. Straube, J. L. Roti Roti. “The Effect of Chronic Exposure to 835.62 MHz FMCW or 847.7 MHz CDMA on the incidence of Spontaneous Tumors in Rats.” Bioelectromagnetic Society Conference, June 25, 2002.)

This is an example of a **designed experiment**. Unlike an observational study, the researcher in an experiment *manipulates* the explanatory variable to determine how varying the explanatory variable will affect the response variable.

Which is better? An observational study or designed experiment?

They both have advantages and disadvantages.

expl 4: Can you think of any other factors that could affect the response variable of tumor rates in humans?

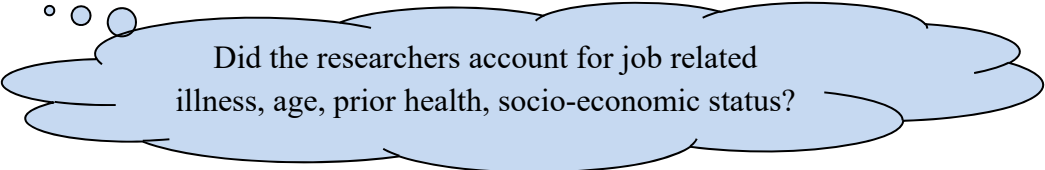
An important distinction: Observational studies versus Designed experiments:

Observational studies do *not* allow a researcher to claim causation, only association. Let's suppose that the first cancer study (researchers Joachim Schüz and associates from page 3) *did find* an association between cell phones and brain tumors. It would be *wrong* to take that finding and say that cell phones *caused* any differences in brain tumors. We only would see an association and *not* be able to nail down what actually *caused* the increase in tumors. Designed experiments, which we will look into further, can be used to show that one thing *causes* another.

Definitions: Confounding in a study occurs when the effects of two or more explanatory variables are *not* separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables *not* accounted for in the study.

A **confounding variable** is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

A **lurking variable** is an explanatory variable that was *not* considered in a study, but that affects the value of the response variable in the study. In addition, lurking variables are typically related to any explanatory variables considered in the study.



Did the researchers account for job related illness, age, prior health, socio-economic status?

Definition: Census: As opposed to a sample, a **census** is a list of *all* individuals in a population along with certain characteristics (the variables of interest) of each individual.

Simple Random Sampling (Section 1.3)

We often cannot ask every individual in a population the question we want answered. So we sample a small group of the population. But you cannot just choose any group. For instance, if you pick your five good friends to question about the video games they like, you might find they do *not* represent the whole population (perhaps all college students, or all people in your age group, or all Americans) well. We must sample with that in mind.

Definition: Random sample: A **random sample** is one where the individuals from the population all have the same chance of being selected for the sample. We will pick them randomly so that is true.

A random sample will represent the population. Because the sample is selected randomly, it can be assumed to be a microcosm of the population.

Definition: Simple random sample (srs): A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an equally likely chance of occurring. The sample is then called a **simple random sample (srs)**.

There are other methods that will be discussed in later sections.

Definition: Frame: A **frame** is a list of the individuals in a population. We will think of them as labeled with the numbers 1, 2, 3, ... N . (Here, N is the number of individuals in the population as described above.)

To obtain a simple random sample, we will start with the frame and use a random number generator to pick a sample.

expl 5: The thirty people listed below are members of a club. The leaders of the club wish to survey its members on possible future trips. Obtain a simple random sample of ten of their members. (Notice the first nine are labeled as 01, 02, etc. Why do you think that is?)

01. Trey	06. Morgan	11. Stefanie	16. London	21. Molly	26. George
02. Amy	07. Bill	12. Joel	17. Julia	22. Tom	27. Robert
03. Marge	08. Jill	13. Penn	18. Chealon	23. Kevin	28. Lora
04. Elise	09. Steve	14. Savvy	19. Jenny	24. Linda	29. Penny
05. Bob	10. Mindy	15. Sawyer	20. Fred	25. Kristen	30. Josh

Your calculator has a random number generator built into it. On the TI calculators, press the **MATH** button and arrow over to **PRB** (stands for probability). Select **1: rand**. The “rand” will appear on the screen; just start pressing **ENTER** to start generating random numbers.

Since we have more than nine people to choose from, each person is assigned a two-digit identifying number.

My calculator gave me the following.

.9435974025
 .908318861
 .1466878292
 .5147019505
 .4058096418
 .7338123112
 .0439919875
 .2209784733
 .0062633066

Mark off these numbers in two-digit increments. Start at the beginning. When a number is greater than 30, ignore it. When a number falls in the range, 01 – 30, we select that person for the random sample.

What do you do when you get a number twice?

Write the sample of ten members here.

This is called a **sample without replacement**. Once you survey a person, it does not make sense to survey them again so they are skipped if their number comes up again. In other words, they are not replaced back into the pool once they are selected.

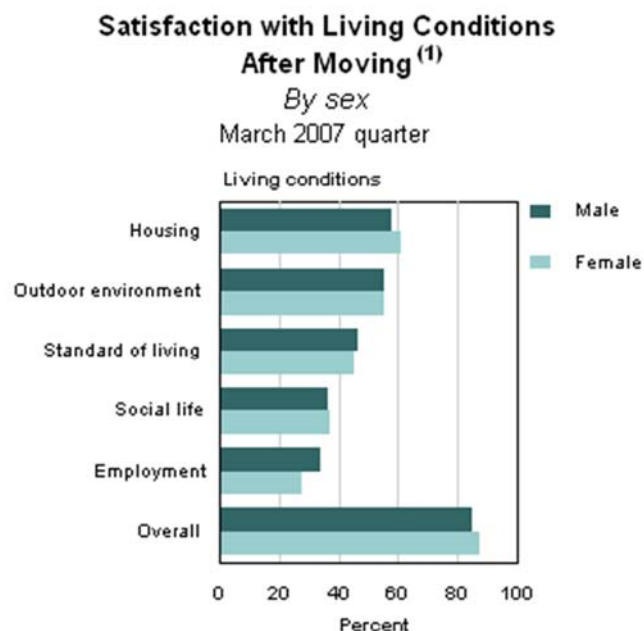
Other Sampling Methods (Section 1.4)

Simple random sampling is *not* the only way to sample properly. There are other methods that are less costly in money, time, or other resources. In fact, these other methods may capture the true picture of the population far better than simple random sampling.

Consider this graph that shows the discrepancies between how men and women feel about a recent move.

(Source:
http://archive.stats.govt.nz/browse_for_stats/population/Migration/internal-migration/benefits-of-moving-men-women.aspx)

If we were to perform a simple random sample of Americans, we would get an aggregated view of how Americans think. Asking women and men separately shows the differences between the sexes. This is an example of a **stratified random sample**.



(1) Movers who stated that their current living conditions were better or much better than before they moved.

Definition: Stratified Random Sample: A **stratified random sample** is obtained by separating the population into non-overlapping groups called **strata** and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

Stratum is singular, strata is plural.

These are very helpful if you have small subgroups within a population, like transgendered people or Native Americans, who may get overlooked otherwise.

One big advantage of this technique is that fewer individuals need to be surveyed to get the same information as a simple random sample. You can also report on the differences among groups.

Another technique eliminates the need for a frame for the whole population. Do you remember what a **frame** is?

Definition: Systematic sample: A **systematic sample** is obtained by selecting every k^{th} individual from the population. The first individual selected is a random number between 1 and k .

An example of this would be to survey every fifth house as you walk down a block.

Sometimes this can go awry. I remember hearing of a study where they decided beforehand to survey every third household. The survey team found themselves in a neighborhood of three-story walk-up apartment buildings, each with one apartment per floor. Do you see why this might cause trouble?

Systematic surveys are useful for (voting) exit polls and surveys of customers, among other uses. Advantages over an “srs” include less cost and ease of use.

Yet another good technique is the cluster sample.

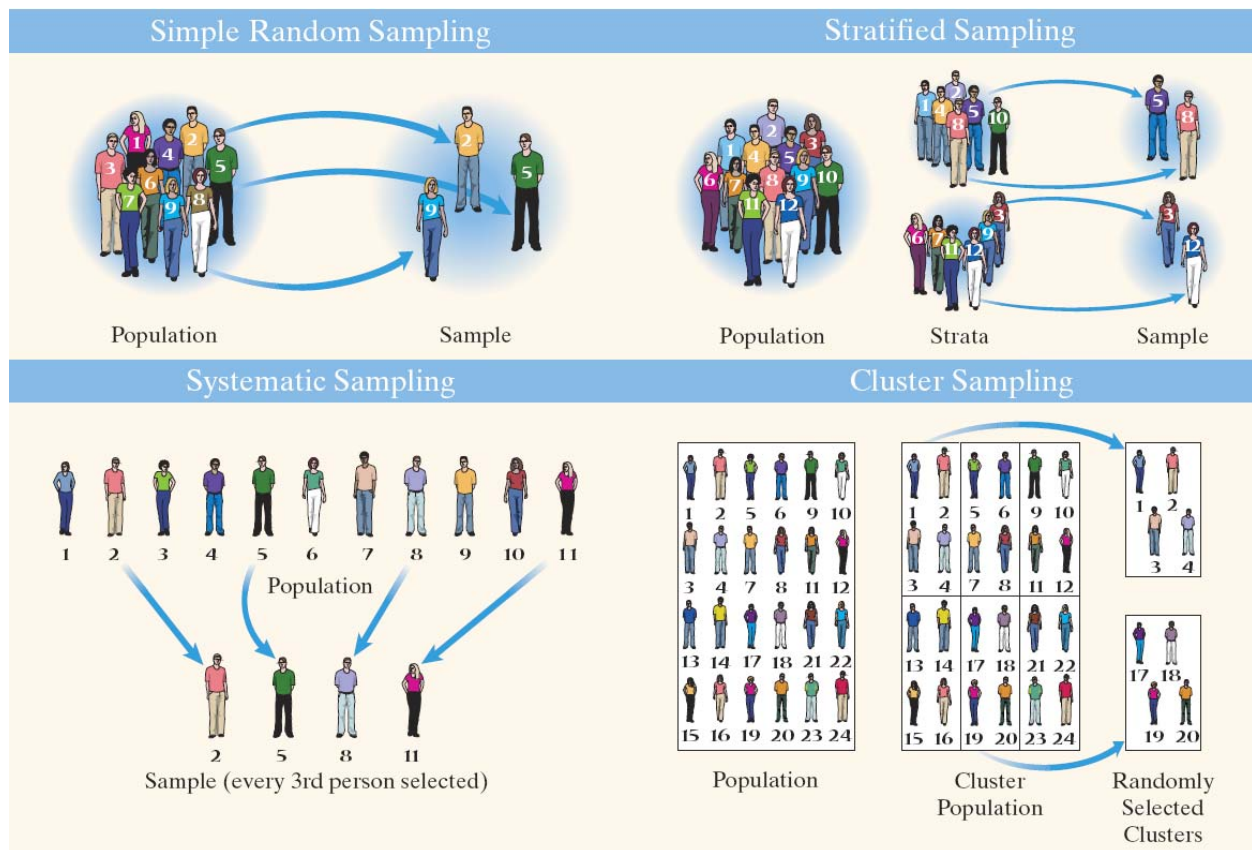
Definition: Cluster sample: A **cluster sample** is obtained by selecting *all* individuals within a randomly selected collection or group of individuals (clusters).

Here, the technique involves determining clusters (groups of individuals) that you will randomly choose from. Once you have determined the clusters you want included, you survey *every* individual in those clusters.

EXAMPLE Obtaining a Cluster Sample:

A school administrator wants to obtain a sample of students in order to conduct a survey. She randomly selects 10 classes and administers the survey to *all* the students in the selected classes.

Here is a useful illustration of the differences among the four techniques we have studied.



Multistage sampling occurs when a sample uses more than one of these techniques.

Bad Sampling Methods:

We are bombarded by polls that claim to show the views of fellow Americans. Here is one I found online.



(source: <https://knss.radio.com/poll-results?page=1>)

It seems that every radio and TV show has running polls. The respondents are **self-selected**. What do you think that means?

These are examples of **Voluntary Samples** and are *not* valid. The respondents are *not* chosen randomly and so the results *cannot* be used to infer about the larger population. That does not stop many radio hosts from making that claim.

Another type of bad sample is called the **Convenience Sample**. This is what it sounds like. The individuals in the sample were chosen simply because they were convenient. You have probably been asked to fill out such a survey. They are rampant in colleges as professors give them out for their own research.

Bias in Sampling (Section 1.5)

A study can be biased toward one view or another. Sometimes the bias is intentional but often not.

Definition: Bias: If the results of the sample are *not* representative of the population, then the sample has **bias**.

There are three sources of bias. They are

1. Sampling bias
2. Nonresponse bias
3. Response bias

One famous (but still influential) example of a biased study is Andrew Wakeham. In 1998, he and twelve colleagues published a *Lancet* article saying the MMR (measles, mumps, rubella) vaccine “may predispose to behavioral regression and pervasive developmental disorder in children”. Despite the small sample size of 12, the study was widely publicized. In fact, later studies showed vaccines and autism to be linked only because they both occur in early childhood. Unfortunately, the damage was done and the anti-vaxxer movement gained a lot of momentum. Under scrutiny, it was found that study subjects were hand-picked. In 2004, ten of the researchers retracted their interpretation of the data. It came to light that Wakeham had been employed by lawyers working for parents suing vaccine-producing companies and that Wakeham had a patent for a single measles vaccine, from which he stood to profit once the combined vaccine was taken off the market.

1. Definition: Sampling bias means that the technique used to obtain the individuals to be in the sample tends to favor one part of the population over another.

Definition: Undercoverage results in sampling bias. It occurs when the proportion of one segment of the population is lower in a sample than it is in the population. This can happen when you leave out major groups in your sample, like women or transgendered people or vegetarians.

Phone surveys can suffer from this because not everyone has a phone (homeless people come to mind) and many will not pick up the phone if they do not recognize who is calling. If these people are different than those who do get sampled, then we have undercoverage.

Have you ever gotten a survey in your email and ignored it? If so, you are included in the next type of bias.

2. Definition: Nonresponse bias exists when individuals selected to be in the sample who do *not* respond to the survey have different opinions from those who do.

All surveys suffer from nonresponse. However, nonresponse can be improved through the use of callbacks or rewards/incentives. Samples often attach a small reward for participation to ease this bias.

3. Definition: Response bias exists when the answers on a survey do *not* reflect the true feelings of the respondent. This can happen in many different ways.

Types of Response Bias

1. Interviewer error
2. Misrepresented answers
3. Wording of questions
4. Order of questions or words

Let's look at some examples of these biases.

expl 6: (Wording of questions) In the early 1990's, Gallup asked Americans whether they supported the US bombing Serbian forces in Bosnia. In this survey, 35% of respondents supported the idea. The very same day, ABC News asked whether Americans would support the US, along with its allies in Europe, bombing Serbia forces in Bosnia. In this survey, 65% supported the idea. Explain the difference in the wording of the question. What does this suggest?

expl 7: (Misrepresented answers) Ask a group of people how many push-ups they can do and then ask them to actually do them. How accurate do you think the survey would be?

expl 8: (Order of questions or words) Consider a survey given in 1980 that contained **both** of the questions below.

a.) Do you think the US should let Communist reporters from other countries come in here and send back to their papers the news as they see it?

b.) Do you think a Communist country such as Russia should let American newspaper reporters come in and send back to America the news as they see it?

If you were taking this survey, how would you respond? What if your survey had the two questions in reverse order? Would that change your opinion?

In fact, this survey was given in two different forms, one with question *a* first and one with question *b* first.

When question *a* was asked first, 54.7% of respondents answered “yes” to question *a* and 63.7% then answered “yes” to question *b*.

But when they asked question *b* first, 81.9% answered “yes” to *b* and 74.6% then answered “yes” to *a*.

Why do you think that is?

To summarize:	When question <i>a</i> was asked first...	When question <i>b</i> was asked first...
Percent who said “yes” to question <i>a</i> *	55%	75%
Percent who said “yes” to question <i>b</i> *	64%	82%

*I rounded these to whole number percents for ease of discussion.

Reputable surveys will often word their questions so they can rotate the options. This helps to eliminate **response bias**. For instance, consider the question “Do you favor or oppose the reduction of estate taxes?” They will ask the same question in half the surveys but phrase it as “oppose or favor”.

Definition: Data-entry error: Although not technically a result of response bias, **data-entry error** will lead to results that are *not* representative of the population. Once data are collected, the results may need to be entered into a computer, which could result in input errors. Or, a respondent may make a data entry error. For example, 39 may be entered as 93. It is imperative that data be checked for accuracy.

As we will see, most surveys have errors. We try to minimize them. Even censuses have errors.

Definition: Nonsampling errors are errors that result from sampling bias, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a complete census of the population.

This is opposed to **Sampling error** which is an error that results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population. These errors are *not* mistakes per se. It is expected and part of the analysis we do.

Worksheet: Sampling questions:

This worksheet focuses on questions that you should ask of any survey you read about. Who carried out the survey? How was the sample selected? And more... Answering these questions helps you determine how much bias is a factor. The worksheet uses a study that explored how HIV/AIDS is viewed in the African American population.

The Design of Experiments (Section 1.6)

A well-designed observational study can give us good information but we can never be sure that the explanatory variable is actually *causing* the change in the response variable. Perhaps the effect is simply a coincidence. An experiment will allow us to determine **causality**.

Definitions: An **experiment** is a controlled study conducted to determine the effect of varying one or more explanatory variables (or **factors**) has on a response variable. Any combination of the values of the factors is called a **treatment**.

The **experimental unit** (or **subject**) is a person, object or some other well-defined item upon which a treatment is applied.

A **completely randomized design** is one in which each experimental unit is randomly assigned to a treatment.

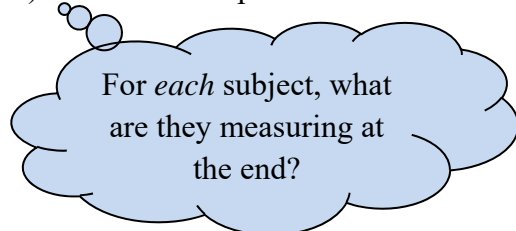
Since we randomly assign individuals to each treatment, we can say more definitively that the explanatory variable (for instance, the amount of radiation from a cell phone) did, in fact, cause any change we notice in the response variable (tumor growth rate). We still have to be careful in the design of our experiment. Let's delve into some details.

expl 9: Lipitor is a cholesterol lowering drug made by Pfizer. In the Collaborative Atorvastatin Diabetes Study (CARDS), the effect of Lipitor on cardiovascular disease was assessed in 2838 subjects, ages 40 to 75, with type 2 diabetes but without prior history of cardiovascular disease. In this **placebo-controlled, double-blind experiment**, subjects were split roughly in half and randomly assigned to either Lipitor 10 mg daily or a placebo. The subjects were followed for four years. The researchers counted the subjects in each group who experienced a major cardiovascular event, such as a stroke or heart attack. This experiment found that Lipitor did reduce the number of cardiovascular events (83 events in the Lipitor group versus 127 events in the placebo group) and deaths (61 in Lipitor group versus 82 in placebo group).

a.) What are the treatments?

b.) The group that gets the placebo is called the **control group**. What is a placebo? Why would we want half the subjects to take a placebo?

c.) What is the response variable? Is it a qualitative or quantitative variable?



d.) A **blind study** is one where the subjects do not know which treatment (Lipitor or placebo) they are receiving. What do you think **double-blind** means? Why would we want this?

Steps of Designing an Experiment: ○ ○ ○

We will *not* be designing our own experiments.

Step 1: Identify the explicit problem to be solved. This is often referred to as the **claim**. Identify the response variable and the population to be studied.

Step 2: Determine the factors that affect the response variable. Once the factors are identified, it must be determined which factors are to be fixed at some predetermined level (the control), which factors will be manipulated, and which factors will be uncontrolled.

Step 3: Determine the number of experimental units. As a general rule, choose as many experimental units as time and money allow. Techniques exist for determining sample size, provided certain information is available.

Step 4: Determine the level(s) of the factors (explanatory variables). There are two ways to deal with the factors: control and randomize.

1. **Control:** There are two ways to control the factors.
 - a) Set the level of a factor at one value throughout the experiment (if you are *not* interested in its effect on the response variable).
 - b) Set the level of a factor at various levels (if you are interested in its effect on the response variable). The combinations of the levels of all varied factors constitute the treatments in the experiment.
2. **Randomize:** Randomize the experimental units to various treatment groups so that the effects of variables whose level cannot be controlled is minimized. The idea is that randomization “averages out” the effect of uncontrolled explanatory variables.

Step 5: Conduct the Experiment.

- a) **Replication** occurs when each treatment is applied to more than one experimental unit. This helps to assure that the effect of a treatment is not due to some characteristic of a single experimental unit. It is recommended that each treatment group have the same number of experimental units.
- b) Collect and process the data by measuring the response variable. Any difference in the value of the response variable is assumed to be a result of differences in the level of the treatment.

Step 6: Test the claim. This is the subject of inferential statistics. Inferential statistics is a process in which generalizations about a population are made on the basis of results obtained from a sample. Provide a statement regarding the level of confidence in the generalization. Methods of inferential statistics are presented later in the text.

Definition: Placebo effect: To know you are receiving any treatment can be powerful medicine. The **placebo effect** refers to how people will improve even though the treatment they received has no actual efficacy. It is truly mind over matter. Using placebos in experiments will help offset this effect in the data.

Definition: Hawthorne effect: This refers to when a person changes their behavior because they know they are being observed. Imagine trying to figure the percentage of people who wash their hands after using a public restroom by sitting next to the sink with a clipboard. How would you account for that?