Simpson's paradox

Simpson's paradox

We will not be covering most of the material from chapter 6.

Simpson's paradox

We will not be covering most of the material from chapter 6. But it is useful to be aware of Simpson's paradox.

Simpson's paradox

We will not be covering most of the material from chapter 6. But it is useful to be aware of Simpson's paradox.

Fact 1. [Simpson's paradox] *It is possible for one individual to outperform another in every category measured yet to not perform as well in the aggregate.*

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines,*

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix.*

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix. At their hubs we have the following data:*

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix. At their hubs we have the following data:*

*	Alaska OT	Alaska delayed	AW OT	AW delayed
PHX	221	12	4840	415
SEA	1841	305	201	61

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix. At their hubs we have the following data:*

*	Alaska OT	Alaska delayed	AW OT	AW delayed
PHX	221	12	4840	415
SEA	1841	305	201	61

Calculate the on-time percentage of each airline at each airport.

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix. At their hubs we have the following data:*

*	Alaska OT	Alaska delayed	AW OT	AW delayed
PHX	221	12	4840	415
SEA	1841	305	201	61

Calculate the on-time percentage of each airline at each airport. Calculate the on-time percentage over both airports.

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix. At their hubs we have the following data:*

*	Alaska OT	Alaska delayed	AW OT	AW delayed
PHX	221	12	4840	415
SEA	1841	305	201	61

Calculate the on-time percentage of each airline at each airport. Calculate the on-time percentage over both airports. Explain what you see.

Example 2. *Let us look at flight delays for two of our local carriers, Alaska Airlines and America West Airlines, the former of which has a hub in Seattle, the latter in Phoenix. At their hubs we have the following data:*

*	Alaska OT	Alaska delayed	AW OT	AW delayed
PHX	221	12	4840	415
SEA	1841	305	201	61

Calculate the on-time percentage of each airline at each airport. Calculate the on-time percentage over both airports. Explain what you see.

It is simple to see how Simpson's paradox works if we look at a simple enough example.

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades.

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades. Dick has HW count 10% and the final exam 90%,

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades. Dick has HW count 10% and the final exam 90%, and Jane has HW count 90% and the final exam count 10%.

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades. Dick has HW count 10% and the final exam 90%, and Jane has HW count 90% and the final exam count 10%. They get A and A-, respectively, on their HW,

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades. Dick has HW count 10% and the final exam 90%, and Jane has HW count 90% and the final exam count 10%. They get A and A-, respectively, on their HW, and C and C- on their final exams, respectively.

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades. Dick has HW count 10% and the final exam 90%, and Jane has HW count 90% and the final exam count 10%. They get A and A-, respectively, on their HW, and C and C- on their final exams, respectively. So Dick has scored better on both.

It is simple to see how Simpson's paradox works if we look at a simple enough example. Suppose Dick and Jane both take MA 243 and (somehow!) negotiate negotiate different weightings to compute their final grades. Dick has HW count 10% and the final exam 90%, and Jane has HW count 90% and the final exam count 10%. They get A and A-, respectively, on their HW, and C and C- on their final exams, respectively. So Dick has scored better on both. But he ends up with a C+ and Jane with a B+.

Some useful terminology, if thinking in terms of percentages: there are two ways to take an average, a *weighted average* which depends on the sample sizes, and a "straight" average of percentages

Some useful terminology, if thinking in terms of percentages: there are two ways to take an average, a *weighted average* which depends on the sample sizes, and a "straight" average of percentages (which really is not so straight).

Some useful terminology, if thinking in terms of percentages: there are two ways to take an average, a *weighted average* which depends on the sample sizes, and a "straight" average of percentages (which really is not so straight). The weighted average is the one which calculates the true percentage, but it is susceptible to Simpson's paradox.

Some useful terminology, if thinking in terms of percentages: there are two ways to take an average, a *weighted average* which depends on the sample sizes, and a "straight" average of percentages (which really is not so straight). The weighted average is the one which calculates the true percentage, but it is susceptible to Simpson's paradox. A "straight" average behaves predictably in this way, but the final answer depends on the categories by which the data has been broken down.

Looking at how data is produced

Looking at how data is produced

So far we have taken data as given and analyzed it.

Looking at how data is produced

So far we have taken data as given and analyzed it.

For single variables have found mean, median, quartiles, and seen the standard deviation. If the variable is normally distributed, we can find answers to more detailed questions about percentiles.

Looking at how data is produced

So far we have taken data as given and analyzed it.

For single variables have found mean, median, quartiles, and seen the standard deviation. If the variable is normally distributed, we can find answers to more detailed questions about percentiles.

For many variables, we have taken them two at a time and compared them through scatterplots.

Looking at how data is produced

So far we have taken data as given and analyzed it.

For single variables have found mean, median, quartiles, and seen the standard deviation. If the variable is normally distributed, we can find answers to more detailed questions about percentiles.

For many variables, we have taken them two at a time and compared them through scatterplots. Using the value r and the regression line, we have looked for positive and negative correlations.

Looking at how data is produced

So far we have taken data as given and analyzed it.

For single variables have found mean, median, quartiles, and seen the standard deviation. If the variable is normally distributed, we can find answers to more detailed questions about percentiles.

For many variables, we have taken them two at a time and compared them through scatterplots. Using the value r and the regression line, we have looked for positive and negative correlations.

But so far we have not questioned how good our data is.

But so far we have not questioned how good our data is. That is an important question since data analysis, like any analysis, obeys the maxim “garbage in, garbage out.”

But so far we have not questioned how good our data is. That is an important question since data analysis, like any analysis, obeys the maxim “garbage in, garbage out.” We now start to talk about generating reliable data.

Sampling

Sampling

Suppose you have a question you wish to answer about a large population. For example,

Sampling

Suppose you have a question you wish to answer about a large population. For example,

- What percent of Americans think George Bush is doing a good job?

Sampling

Suppose you have a question you wish to answer about a large population. For example,

- What percent of Americans think George Bush is doing a good job?
- What percentage of Americans know that the Earth orbits the sun in a year?
- What percent of people with headaches are helped by aspirin?

Sampling

Suppose you have a question you wish to answer about a large population. For example,

- What percent of Americans think George Bush is doing a good job?
- What percentage of Americans know that the Earth orbits the sun in a year?
- What percent of people with headaches are helped by aspirin?

Gathering data to answer these questions can be problematic.

Gathering data to answer these questions can be problematic.

In all of these cases, there are simply too many people to find out the answer for each one.

Gathering data to answer these questions can be problematic.

In all of these cases, there are simply too many people to find out the answer for each one. We will see that *carefully* taking data from a subset, called "sampling," is the best we can do and can sometimes answer these questions well.

Gathering data to answer these questions can be problematic.

In all of these cases, there are simply too many people to find out the answer for each one. We will see that *carefully* taking data from a subset, called "sampling," is the best we can do and can sometimes answer these questions well.

In addition, for the third case, there is a question of how to measure to what extent aspirin "helps."

Gathering data

Gathering data

There are two basic methods of gathering data:

Definition 3. *In an* observational *study one observes individuals and measures variables of those individuals.*

Gathering data

There are two basic methods of gathering data:

Definition 3. *In an* observational *study one observes individuals and measures variables of those individuals.*

If the study is to lead to conclusions about the overall population there are a two things that must be considered:

Gathering data

There are two basic methods of gathering data:

Definition 3. *In an* observational *study one observes individuals and measures variables of those individuals.*

If the study is to lead to conclusions about the overall population there are a two things that must be considered:

- Does the sample of individuals reflect the overall population?

- Is the measure of the interesting variables accurate?

- Is the measure of the interesting variables accurate?

Example 4. *Phoning 1000 randomly chosen residential phone numbers during the workday, one asks for the answer to two variables*, age, *and* how many hours of TV the subject watches per day*. If the phone is not answered, one calls the next number.*

- Is the measure of the interesting variables accurate?

Example 4. *Phoning 1000 randomly chosen residential phone numbers during the workday, one asks for the answer to two variables*, age, *and* how many hours of TV the subject watches per day*. If the phone is not answered, one calls the next number. What are possible problems?*

- Is the measure of the interesting variables accurate?

Example 4. *Phoning 1000 randomly chosen residential phone numbers during the workday, one asks for the answer to two variables*, age, *and* how many hours of TV the subject watches per day*. If the phone is not answered, one calls the next number. What are possible problems?*

Understanding what problems may arise in collecting data is an artform best learned in the discipline or setting in which you are working.

- Is the measure of the interesting variables accurate?

Example 4. *Phoning 1000 randomly chosen residential phone numbers during the workday, one asks for the answer to two variables*, age, *and* how many hours of TV the subject watches per day*. If the phone is not answered, one calls the next number. What are possible problems?*

Understanding what problems may arise in collecting data is an artform best learned in the discipline or setting in which you are working.

Definition 5. *In an* experimental *study one treats a group of individuals in a particular way with the goal of discovering the effect of that treatment.*

Definition 5. *In an* experimental *study one treats a group of individuals in a particular way with the goal of discovering the effect of that treatment.*

Experimental studies can be fraught with difficulties.

Definition 5. *In an* experimental *study one treats a group of individuals in a particular way with the goal of discovering the effect of that treatment.*

Experimental studies can be fraught with difficulties.

If this study is to lead to conclusions about the efficacy of a treatment one must

Definition 5. *In an* experimental *study one treats a group of individuals in a particular way with the goal of discovering the effect of that treatment.*

Experimental studies can be fraught with difficulties.

If this study is to lead to conclusions about the efficacy of a treatment one must

- Make certain that there is a group of untreated individuals (a *control* group) with which to compare the treated indivuals.

Definition 5. *In an* experimental *study one treats a group of individuals in a particular way with the goal of discovering the effect of that treatment.*

Experimental studies can be fraught with difficulties.

If this study is to lead to conclusions about the efficacy of a treatment one must

- Make certain that there is a group of untreated individuals (a *control* group) with which to compare the treated indivuals.

- One also wants to make sure that there isn't a lurking variable distinguising between treated and untreated individuals.

- One also wants to make sure that there isn't a lurking variable distinguising between treated and untreated individuals.

Example 6. *The "placebo effect" and other bias in medical studies, and the need for "double-blind" protocols.*

- One also wants to make sure that there isn't a lurking variable distinguising between treated and untreated individuals.

Example 6. *The "placebo effect" and other bias in medical studies, and the need for "double-blind" protocols.*

We will discuss this further later.

Summarizing our discussions of data collection:

To get good conclusions, one must be careful about how one is gathering data and

WHAT MIGHT BE *BIAS*ING THE SAMPLE

WHAT MIGHT BE *BIAS*ING THE SAMPLE

Example 7. *A research firm phones 100 clients of acupuncturists to ask if their treatment has improved their health. What can one learn from this experiment about the efficacy of acupuncture?*

WHAT MIGHT BE *BIAS*ING THE SAMPLE

Example 7. *A research firm phones 100 clients of acupuncturists to ask if their treatment has improved their health. What can one learn from this experiment about the efficacy of acupuncture? (Efficacy vs. satisfaction)*

WHAT MIGHT BE *BIAS*ING THE SAMPLE

Example 7. *A research firm phones 100 clients of acupuncturists to ask if their treatment has improved their health. What can one learn from this experiment about the efficacy of acupuncture? (Efficacy vs. satisfaction)*

WHAT MIGHT BE *BIAS*ING THE SAMPLE

Example 7. *A research firm phones 100 clients of acupuncturists to ask if their treatment has improved their health. What can one learn from this experiment about the efficacy of acupuncture? (Efficacy vs. satisfaction)*

If you *really* want to measure acupuncture's efficacy, you need to take a group of people and randomly assign half of them to get treated by acupuncture, and half to be untreated (as a control group if you want to compare

acupuncture to no treatment) or treated by western medicine (as a control group if you wanted to compare acupuncture to western medicine). This study would be even better if the patients didn't *know* whether they were being treated via acupuncture or in the control group.

acupuncture to no treatment) or treated by western medicine (as a control group if you wanted to compare acupuncture to western medicine). This study would be even better if the patients didn't *know* whether they were being treated via acupuncture or in the control group.

Example 8. *A local Eugene TV station asks callers to phone in during the news to say if they favor changing the time of the Eugene Celebration Parade so as not to conflict with the Ducks game. What can one learn from this poll?*

This called a *voluntary response sample*.

To draw a conclusion about the population of Eugene, you would want to poll a randomly selected group of Eugenians. This kind of data can be quite hard to get.

Example 9. *Poll at poll.excite.com: "Do you think the Iraq war was worth it?" (Yes or No).*

- *yes: 46%*
- *no: 48%*
- *not sure: 4%*

Example 9. *Poll at poll.excite.com: "Do you think the Iraq war was worth it?" (Yes or No).*

- *yes: 46%*
- *no: 48%*
- *not sure: 4%*

16,926 answers.

What is population? What is sample size? Is this sample random?

Example 9. *Poll at poll.excite.com: "Do you think the Iraq war was worth it?" (Yes or No).*

- *yes: 46%*
- *no: 48%*
- *not sure: 4%*

16,926 answers.

What is population? What is sample size? Is this sample random? What criticisms should we have of this sample?

What general conclusion can we draw?

What general conclusion can we draw?

To excite's credit, the following text appears below the poll results:

"The excite poll. . . is a voluntary poll for our users, and is not scientifically projectable to any other population."

It is interesting to surf for various polling sites to see their methodologies.

It is interesting to surf for various polling sites to see their methodologies. One particularly interesting discussion for lay people is at: `http://mysterypollster.typepad.com/`.

Simple random sample

Simple random sample

Generally one wants to avoid a voluntary response sample, and one wants to avoid a *convenience* sample (chosen for the convenience of the person conducting the study).

A *simple random sample* is one made by choosing individuals at random from the entire population.

Simple random sample

Generally one wants to avoid a voluntary response sample, and one wants to avoid a *convenience* sample (chosen for the convenience of the person conducting the study).

A *simple random sample* is one made by choosing individuals at random from the entire population.

Choose simple random sample from classroom.

Simple random sample

Generally one wants to avoid a voluntary response sample, and one wants to avoid a *convenience* sample (chosen for the convenience of the person conducting the study).

A *simple random sample* is one made by choosing individuals at random from the entire population.

Choose simple random sample from classroom.Get random numbers from between 1 and 120, and read off until I get 10 people.

Simple random sample

Generally one wants to avoid a voluntary response sample, and one wants to avoid a *convenience* sample (chosen for the convenience of the person conducting the study).

A *simple random sample* is one made by choosing individuals at random from the entire population.

Choose simple random sample from classroom.Get random numbers from between 1 and 120, and read off until I get 10 people. Is this random from class? Is it

random from students at the university?

random from students at the university?

To be systematic, take the following steps.

- Label the individuals in your population numerically. (Example: students in class labeled 1-120).

random from students at the university?

To be systematic, take the following steps.

- Label the individuals in your population numerically. (Example: students in class labeled 1-120).
- Decide how many digits are required to hit all possible labels. (In our case, 3).

random from students at the university?

To be systematic, take the following steps.

- Label the individuals in your population numerically. (Example: students in class labeled 1-120).
- Decide how many digits are required to hit all possible labels. (In our case, 3).
- Start someplace in Table B reading off groups of that many digits. Discard numbers outside the range for your population. (In our example, if we start at line

129, we get $007, 051, 087, 045$, etc.)

This last step can be tedious and can involve discarding a lot of numbers. Another possibility is to obtain a randomly generated sequence of numbers within the range indexing your population, say from `http://www.random.org/`.

129, we get $007, 051, 087, 045$, etc.)

This last step can be tedious and can involve discarding a lot of numbers. Another possibility is to obtain a randomly generated sequence of numbers within the range indexing your population, say from `http://www.random.org/`.

Some polls use basically this methodology.

129, we get $007, 051, 087, 045$, etc.)

This last step can be tedious and can involve discarding a lot of numbers. Another possibility is to obtain a randomly generated sequence of numbers within the range indexing your population, say from `http://www.random.org/`.

Some polls use basically this methodology. But even when getting a random sample across a population, some polls will "correct" for making sure different demographics are appropriately represented.

129, we get $007, 051, 087, 045$, etc.)

This last step can be tedious and can involve discarding a lot of numbers. Another possibility is to obtain a randomly generated sequence of numbers within the range indexing your population, say from `http://www.random.org/`.

Some polls use basically this methodology. But even when getting a random sample across a population, some polls will "correct" for making sure different demographics are appropriately represented. Neither

method seems entirely satisfactory, but that is probably because of issues inherent in polling complex issues with simple questions over populations which are much too small.

Designing experiments

Designing experiments

In an experiement you want to do something (*treat*) a population and try to understand what the effect of that treatment is.

Designing experiments

In an experiement you want to do something (*treat*) a population and try to understand what the effect of that treatment is.

Example 10. *You wish to understand the effect of pricing items so that their prices end in $9. You prepare a catalog in which some items are priced ending in $9 and others ending in $4. You mail the catalog to 10,000 people and measure sales of items ending in $4 and in $9 to draw conclusions.*

Designing experiments

In an experiement you want to do something (*treat*) a population and try to understand what the effect of that treatment is.

Example 10. *You wish to understand the effect of pricing items so that their prices end in $9. You prepare a catalog in which some items are priced ending in $9 and others ending in $4. You mail the catalog to 10,000 people and measure sales of items ending in $4 and in $9 to draw conclusions.*

What are the weaknesses of such an experiment?

In order to address this weakness, consider the following alternate experiment.

Example 11. *You wish to understand the effect pricing items so that their prices end in \$9. You prepare a catalog in which some items are priced ending in \$9. You prepare a second catalog in which the same items are priced one dollar less, so the price ends in \$8.*

What are the weaknesses of such an experiment?

In order to address this weakness, consider the following alternate experiment.

Example 11. *You wish to understand the effect pricing items so that their prices end in \$9. You prepare a catalog in which some items are priced ending in \$9. You prepare a second catalog in which the same items are priced one dollar less, so the price ends in \$8.*

Out of your group of 10000, you pick 5000 people randomly to send one catalog to, and you send the other

catalog to the other 5000 people. You keep track of orders for these specific numbers from these two groups.

catalog to the other 5000 people. You keep track of orders for these specific numbers from these two groups.

How does this experiment design address the weaknesses of the previous example?

catalog to the other 5000 people. You keep track of orders for these specific numbers from these two groups.

How does this experiment design address the weaknesses of the previous example?

In this case you have a *control group* for your hypothesis that prices ending in $9 encourage purchase.