Experimental Design, I

9.07

3/18/2004

Goal of experiments (and thus experimental design)

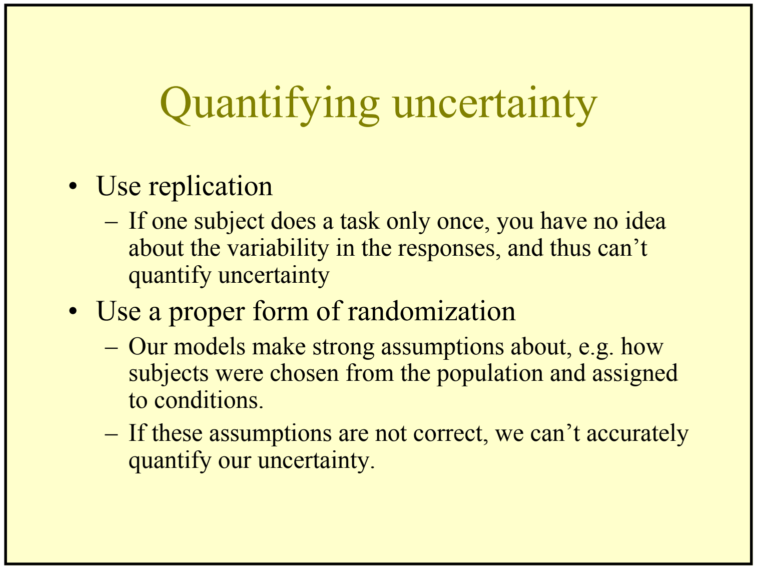
- Determine whether a relationship is likely to exist between
  - One or more independent variables (factors) and a dependent variable, or
  - Two or more dependent variables
- Minimize the possibility that the results you get might be due to a hidden confounding factor
- Maximize the power of your test for this relationship, while keeping the probability of a Type I error to a minimum
- Quantify your uncertainty in the results
- Wide range of applicability of the results

Minimizing confounding factors

- Confounding = a difference between the “treatment” and comparison groups, other than the “treatment”, which affects the responses under study (the dependent variables).
- From homework: does dressing well cause you to do better on the SAT? Confounding factor = income.
- We’ll talk about experimental designs that are better or worse as far as confounding factors.

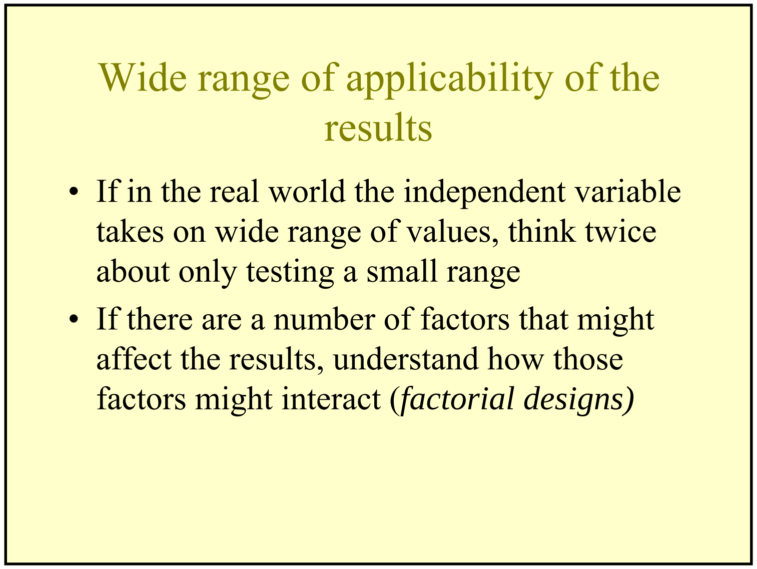
Power

- Probability of detecting a relationship when one in fact exists.
- Increasing power:
  - ~~Increase α (tradeoff between Type I & Type II errors)~~
  - Increase n (*replication*)
  - Increase the “signal-to-noise ratio,” i.e. if possible, increase the size of the effect by either increasing the raw effect size ($m_1 - m_2$), or *decreasing the irrelevant variability*.
  - Do a better statistical test.



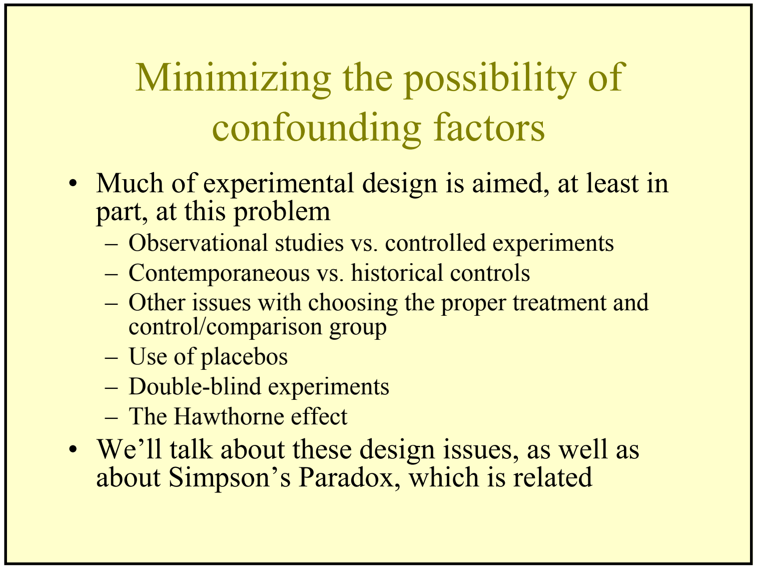
Quantifying uncertainty

- Use replication
  - If one subject does a task only once, you have no idea about the variability in the responses, and thus can't quantify uncertainty
- Use a proper form of randomization
  - Our models make strong assumptions about, e.g. how subjects were chosen from the population and assigned to conditions.
  - If these assumptions are not correct, we can't accurately quantify our uncertainty.



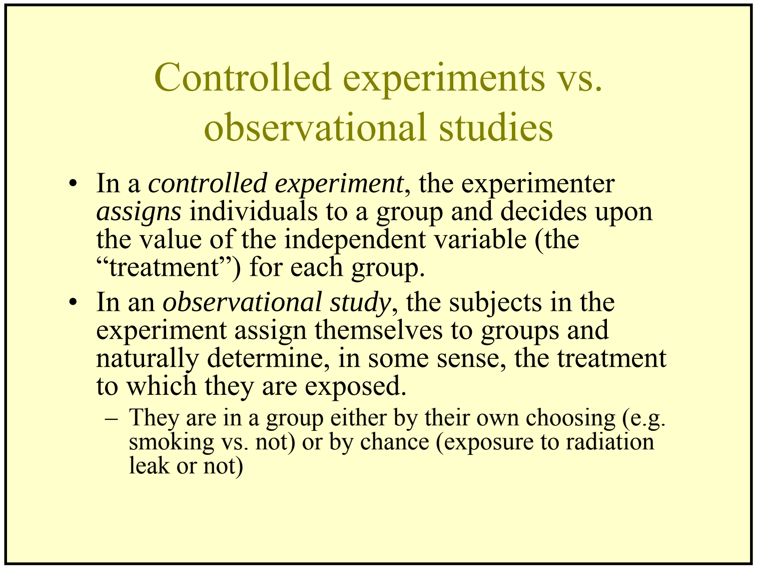
Wide range of applicability of the results

- If in the real world the independent variable takes on wide range of values, think twice about only testing a small range
- If there are a number of factors that might affect the results, understand how those factors might interact (*factorial designs)*



Minimizing the possibility of confounding factors

- Much of experimental design is aimed, at least in part, at this problem
  - Observational studies vs. controlled experiments
  - Contemporaneous vs. historical controls
  - Other issues with choosing the proper treatment and control/comparison group
  - Use of placebos
  - Double-blind experiments
  - The Hawthorne effect
- We'll talk about these design issues, as well as about Simpson's Paradox, which is related



Controlled experiments vs. observational studies

- In a *controlled experiment*, the experimenter *assigns* individuals to a group and decides upon the value of the independent variable (the "treatment") for each group.
- In an *observational study*, the subjects in the experiment assign themselves to groups and naturally determine, in some sense, the treatment to which they are exposed.
  - They are in a group either by their own choosing (e.g. smoking vs. not) or by chance (exposure to radiation leak or not)

Controlled experiment vs. control group

- Controlled experiment defined as in last slide. The investigator *controls* into which group each subject falls, and *controls* the conditions under which each subject is tested.
- A *control*, or *control group*, is a particular kind of comparison group which does not receive some treatment (training, medication, e.g.) when the other groups do.
  - Not every controlled experiment has a control, per se. We will tend to talk about treatment and control groups here, because it makes the story simpler.

Examples

- Study whether a new vaccine works. There are two groups: a treatment group that receives a vaccine, and a control group that does not.
- Subjects sign up for the study, and those that consent to take the vaccine get the vaccine. Those that do not consent are studied as part of the control group.
- Controlled experiment, or observational study?

Suppose, in this study, that vaccinated subjects get the disease less frequently than unvaccinated

- Does this mean the vaccine works, or might there be confounding factors?
- Perhaps poor people are less likely to agree to the vaccine, and are also more likely to get the disease.
  - This would lead to the stated result – unvaccinated subjects would be more likely to get the disease not necessarily because of the vaccine, but because they were more likely to get the disease to begin with.

Observational studies and confounding factors

- A problem with observational studies is that there may be some factor that both influences which group a subject ends up in, and the response of that subject to the experiment.

Another example

- Is smoking a risk factor for mental illness?
- Follow a group of smokers, and a group of non-smokers, look at their mental health after 20 years.
- A researcher finds that the smokers were more likely to later have a serious mental illness.
- What are possible confounding factors due to this being an observational study?

Controlled experiments

- Investigator controls which subject get which experimental condition, by assigning subjects to one group (e.g. treatment group) or another (e.g. control group)
- Assignment to groups is intended to make sure both groups are similar in all ways except the experimental manipulation

It’s not always ethical to do a controlled experiment

- In the smoking and mental illness example, you can’t really force people to smoke, given what we know about its harmful and addictive effects.
- Similarly, it’s not ethical to expose people to a harmful radiation leak, to test whether it affects their performance on a motor task. And you shouldn’t expose people to a traumatic situation just so you can study PTSD.
- Unfortunately, this is not to say that people haven’t done experiments like this. We have human/animal subjects boards to try to keep this sort of behavior to a minimum.

Controlled experiments vs. observational studies

- Sometimes, you’re just stuck with an observational study, for either practical or ethical reasons
- Observational studies can make good “pre-experimental” designs, i.e. it may be easy to do an observational study, and it may suggest whether or not it’s worth doing a controlled experiment to further investigate the issue

Choice of control/comparison group

- Contemporaneous: studied at the same time as the other group(s).
- Historical: compare with results from other studies in the past.
  - When we did examples of one-sample t-tests, many of these were comparing with historical controls.
  - E.G. Compare performance of students with new training to take the SAT to historical performance on the SAT.

Contemporaneous vs. historical comparisons

- It's generally best, when possible, to use contemporaneous rather than historical controls/comparisons.
- Historical controls may differ in some respects from the treatment groups, other than the treatment.
  - Changes due to the passage of time
  - The historical data may have been collected in a different way, which makes comparison questionable

Examples of problems with historical controls

- Students testing with new SAT training did better than historical controls. Did training help SAT performance?
  - Maybe, but maybe the SAT was just easier this year than in the years used as control
- We give a new polio vaccine to a bunch of people, and observe that there were fewer cases this year than last. Did the polio vaccine protect against polio?
  - Maybe, but number of cases fluctuates a lot from year to year. Maybe this was a good year.

A poor controlled experiment

- Want to study the effect of a new liver transplant surgery
- The doctor chooses a set of good candidates for the surgery; others get a traditional non-surgical treatment
- The patients who receive treatment have a lower mortality rate than the ones who receive the traditional treatment
- Do you recommend the surgery?

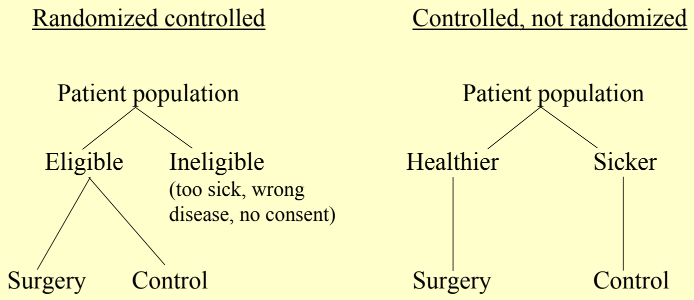
A poor controlled experiment

- Do you recommend the surgery? Not necessarily.
- The doctor may have chosen the surgery candidates based on their health at the time of the study. He may have chosen patients who were healthy enough to undergo the surgery.
- The control group was less healthy to begin with, and this may explain their higher mortality rate.
- Assignment to groups was intended to make sure both groups are similar, aside from the experimental manipulation. Here this did not occur, because the doctor was biased in his selection of the two groups.

Randomization

- Experimental subjects ("units") should be assigned to treatment groups at random.
  - – *Randomized controlled experiment*

Experiments



Randomized controlled

Patient population

Eligible

Ineligible (too sick, wrong disease, no consent)

Surgery

Control

Controlled, not randomized

Patient population

Healthier

Sicker

Surgery

Control

How to randomize

- At random does not mean haphazardly.
- People are notoriously bad at doing things "randomly"
- One needs to explicitly randomize using
  - A computer, or
  - Coins, dice or cards.

Why randomize?

- Avoid bias
  - For example: the first surgery candidates you find may be basically healthier
- Control the role of chance
  - Randomization allows the later use of probability theory, and so gives a solid foundation for statistical analysis.

Why randomize, II

- You would like your groups to be as similar as possible, in everything but the treatment you wish to test
- If you knew everything about factors that might influence your experiment, and could assign subjects to conditions in an unbiased way, then randomization wouldn't be as necessary
  - Matched-sample designs attempt to do things more systematically
- But often there are both known and unknown factors that are potential confounds

Randomization and Confounding

- Randomization is supposed to have the effect of distributing confounders, *both known and unknown,* between the different conditions of the experiment
- E.G. maybe (unbeknownst to the researcher) mood affects success with either the standard treatment or with surgery. By randomizing, we hope to have roughly the same number of good-mood patients in each group, even if we didn't think to control for mood.

Randomization of subjects

- To help assure that groups are similar, subjects are randomly assigned to experimental conditions
  - *Randomized controlled experiment*
- Randomization does not assure that the groups are the same: still need to assess whether they are

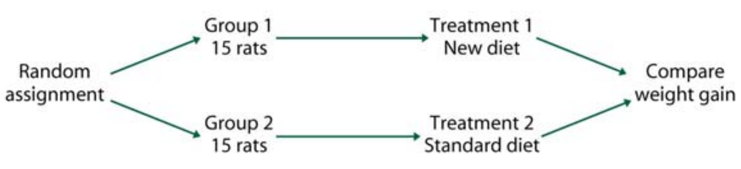
Randomization designs

- Full randomization
- Blocking

Full randomization

- Randomly assign subjects to conditions in the experiment

Full random assignment to condition



Random assignment

Group 1
15 rats

Group 2
15 rats

Treatment 1
New diet

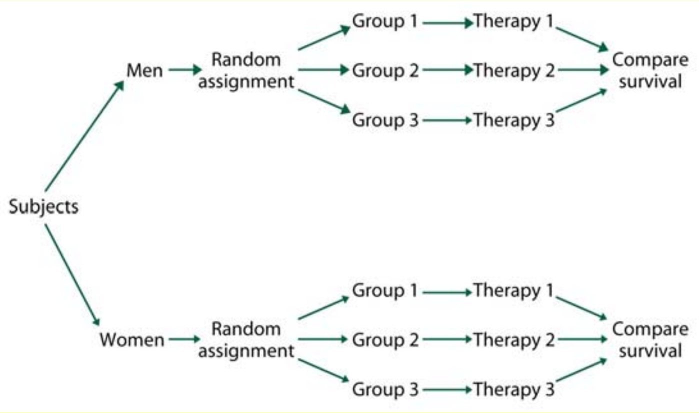
Treatment 2
Standard diet

Compare weight gain

Blocking (aka “stratification”)

- If you have factors that you suspect will have an effect on your results, you can also *block* those factors, and take them into account in your analysis of the data
- Ensure that for each level of a given factor, you have equal numbers of subjects in each condition.
- E.G. perhaps gender has an effect on survival rate in treatment for liver problems
  - Make sure you test both men and women in equal numbers in each condition

Blocked design



Subjects

Men → Random assignment → Group 1 → Therapy 1 / Group 2 → Therapy 2 / Group 3 → Therapy 3 → Compare survival

Women → Random assignment → Group 1 → Therapy 1 / Group 2 → Therapy 2 / Group 3 → Therapy 3 → Compare survival

Why equal numbers?

- As we'll see later this semester, a number of statistical tests are more well behaved if you have equal numbers of measurements for each group.
- You can make your test more powerful by designing your experiment with equal numbers per group.
- We saw a bit of this already: the less conservative two-sample t-test was robust to the equal variance assumption if $n_1 = n_2$

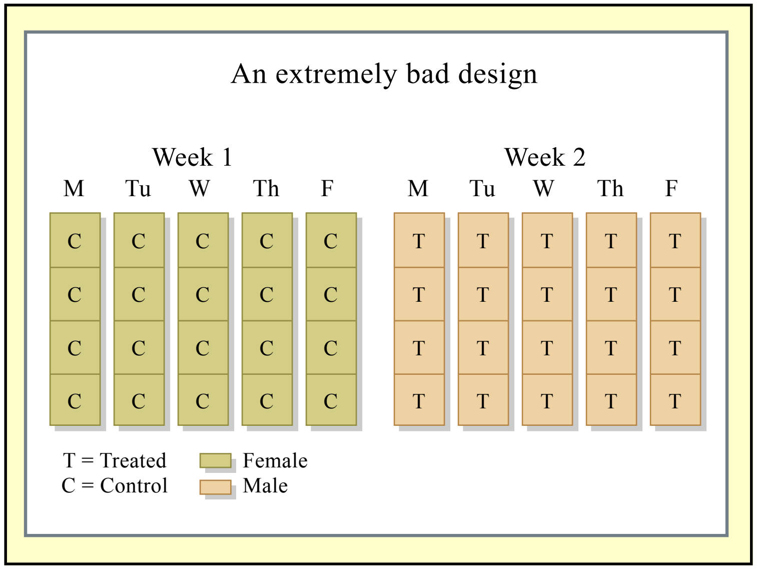
Terminology

- Try not to be confused – "block" is used in other ways, when talking about experiments.
- For instance, suppose a subject does, say, 1000 trials in an experiment, in groups of 200. If each group of 200 consists of trials in a single condition (as opposed to a mix of conditions), then we say that the trials are "blocked".
  - A A A A B B B B C C C C D D D D … vs.
  - A B C D D A B C C D A B B C D A

Example of randomization and blocking

- 20 male mice and 20 female mice.
- Half to be treated; the other half left untreated.
- Can only work with 4 mice per day.

Question: How to assign individuals to treatment groups and to days, if you think gender matters, and you think on what day you test them might matter.

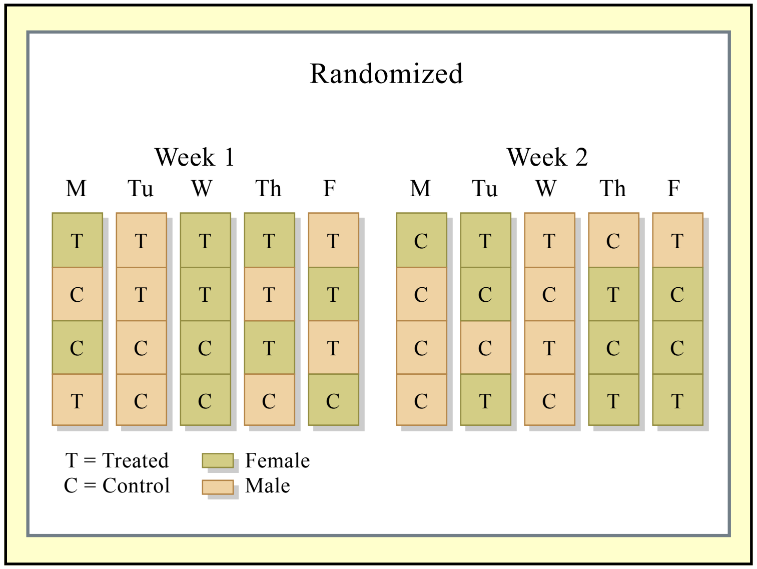


An extremely bad design

Week 1					Week 2				
M	Tu	W	Th	F	M	Tu	W	Th	F
C 🟩	C 🟩	C 🟩	C 🟩	C 🟩	T 🟧	T 🟧	T 🟧	T 🟧	T 🟧
C 🟩	C 🟩	C 🟩	C 🟩	C 🟩	T 🟧	T 🟧	T 🟧	T 🟧	T 🟧
C 🟩	C 🟩	C 🟩	C 🟩	C 🟩	T 🟧	T 🟧	T 🟧	T 🟧	T 🟧
C 🟩	C 🟩	C 🟩	C 🟩	C 🟩	T 🟧	T 🟧	T 🟧	T 🟧	T 🟧

T = Treated 🟩 Female
C = Control 🟧 Male

Figure by MIT OCW.

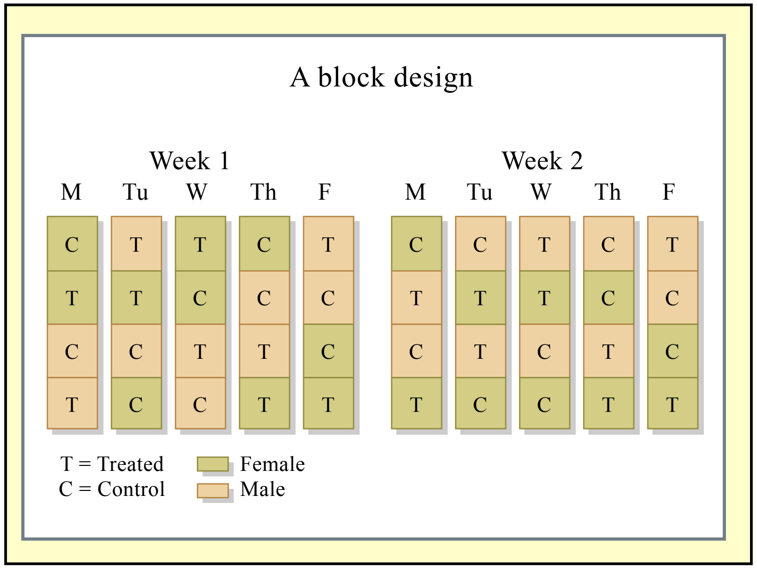


Randomized

Week 1					Week 2				
M	Tu	W	Th	F	M	Tu	W	Th	F
T 🟩	T 🟧	T 🟩	T 🟩	T 🟧	C 🟩	T 🟩	T 🟧	C 🟧	T 🟧
C 🟧	T 🟧	T 🟩	T 🟧	T 🟩	C 🟧	C 🟩	C 🟧	T 🟩	C 🟩
C 🟩	C 🟧	C 🟩	T 🟩	T 🟧	C 🟧	C 🟧	T 🟧	C 🟩	C 🟩
T 🟧	C 🟧	C 🟩	C 🟧	C 🟩	C 🟧	T 🟩	C 🟧	T 🟩	T 🟩

T = Treated 🟩 Female
C = Control 🟧 Male

Figure by MIT OCW.

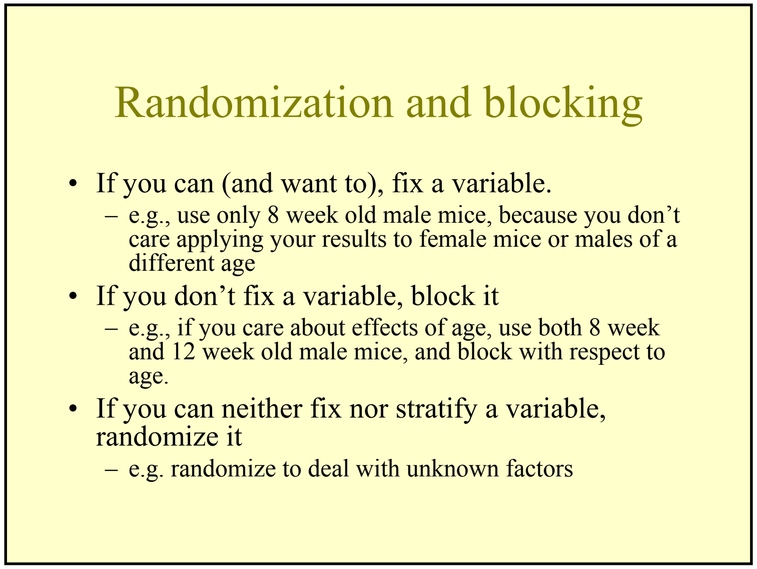


A block design

Week 1					Week 2				
M	Tu	W	Th	F	M	Tu	W	Th	F
C 🟩	T 🟧	T 🟩	C 🟩	T 🟧	C 🟩	C 🟧	T 🟧	C 🟧	T 🟧
T 🟩	T 🟩	C 🟩	C 🟧	C 🟧	T 🟧	T 🟩	T 🟩	C 🟩	C 🟧
C 🟧	C 🟧	T 🟧	T 🟧	C 🟩	C 🟧	T 🟧	C 🟧	T 🟧	C 🟩
T 🟧	C 🟩	C 🟧	T 🟩	T 🟩	T 🟩	C 🟩	C 🟩	T 🟩	T 🟩

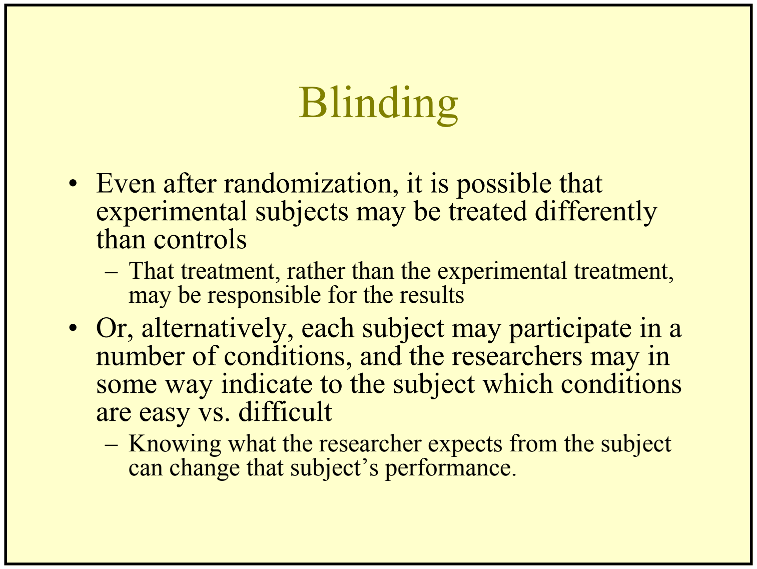
T = Treated 🟩 Female
C = Control 🟧 Male

Figure by MIT OCW.



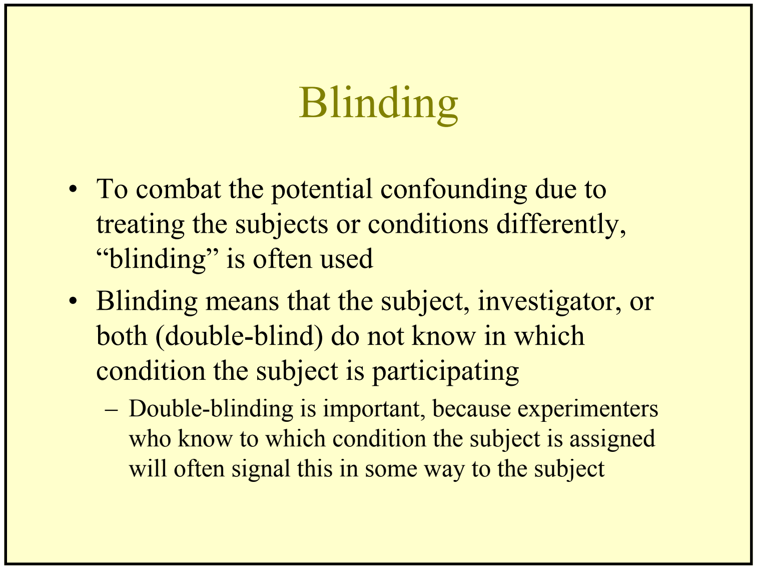
Randomization and blocking

- If you can (and want to), fix a variable.
  - e.g., use only 8 week old male mice, because you don't care applying your results to female mice or males of a different age
- If you don't fix a variable, block it
  - e.g., if you care about effects of age, use both 8 week and 12 week old male mice, and block with respect to age.
- If you can neither fix nor stratify a variable, randomize it
  - e.g. randomize to deal with unknown factors



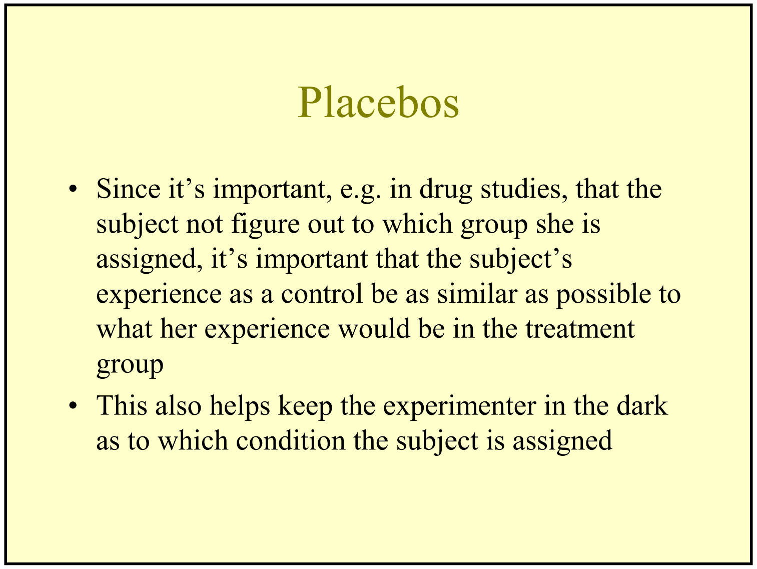
Blinding

- Even after randomization, it is possible that experimental subjects may be treated differently than controls
  - That treatment, rather than the experimental treatment, may be responsible for the results
- Or, alternatively, each subject may participate in a number of conditions, and the researchers may in some way indicate to the subject which conditions are easy vs. difficult
  - Knowing what the researcher expects from the subject can change that subject’s performance.



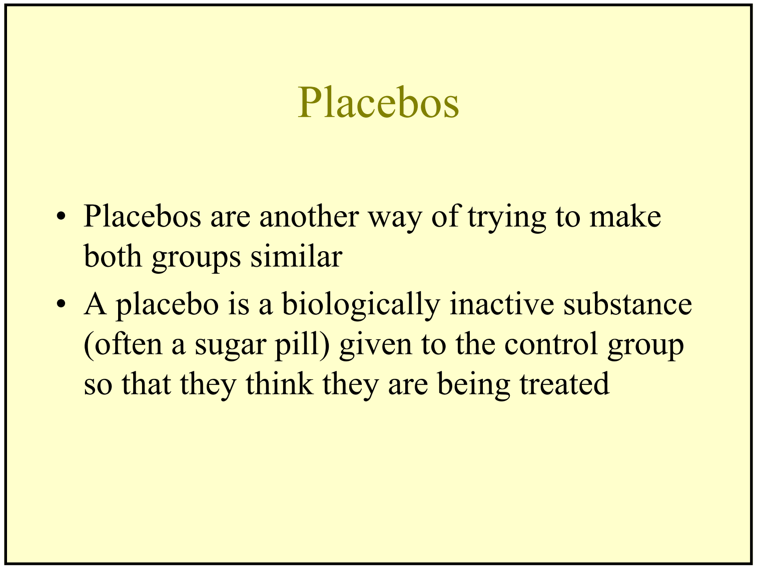
Blinding

- To combat the potential confounding due to treating the subjects or conditions differently, “blinding” is often used
- Blinding means that the subject, investigator, or both (double-blind) do not know in which condition the subject is participating
  - Double-blinding is important, because experimenters who know to which condition the subject is assigned will often signal this in some way to the subject



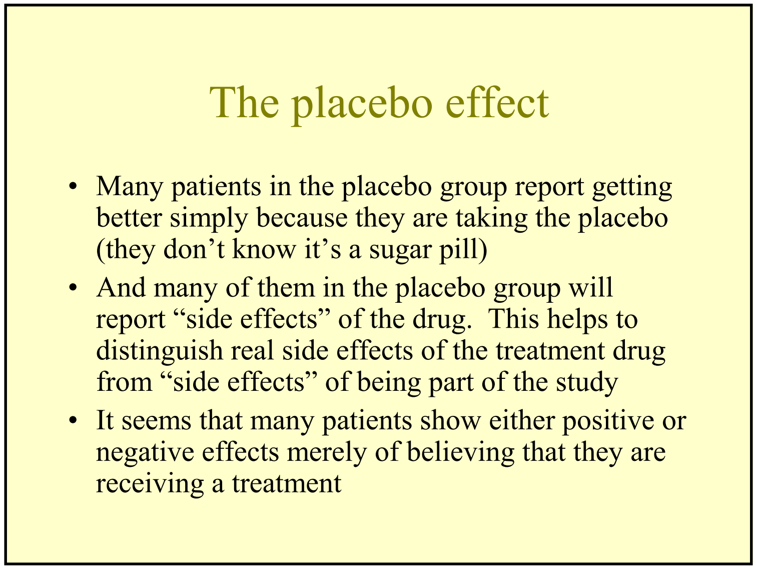
Placebos

- Since it’s important, e.g. in drug studies, that the subject not figure out to which group she is assigned, it’s important that the subject’s experience as a control be as similar as possible to what her experience would be in the treatment group
- This also helps keep the experimenter in the dark as to which condition the subject is assigned



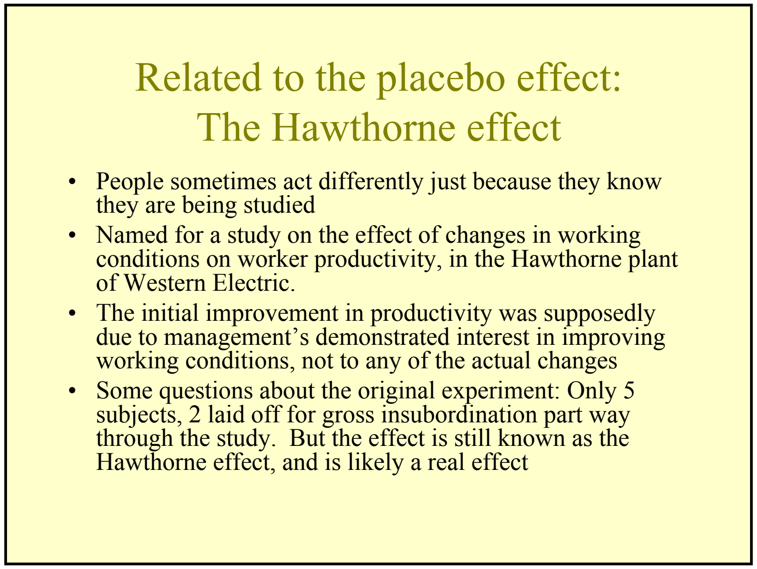
Placebos

- Placebos are another way of trying to make both groups similar
- A placebo is a biologically inactive substance (often a sugar pill) given to the control group so that they think they are being treated



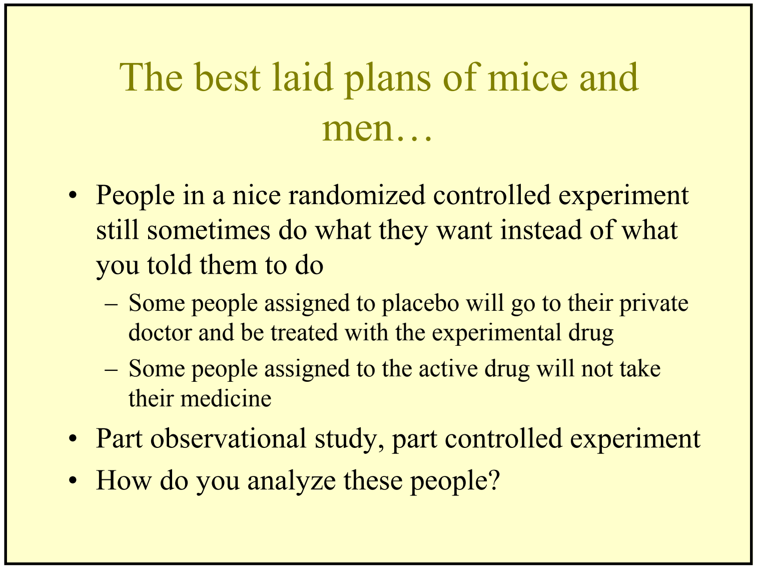
The placebo effect

- Many patients in the placebo group report getting better simply because they are taking the placebo (they don’t know it’s a sugar pill)
- And many of them in the placebo group will report “side effects” of the drug. This helps to distinguish real side effects of the treatment drug from “side effects” of being part of the study
- It seems that many patients show either positive or negative effects merely of believing that they are receiving a treatment



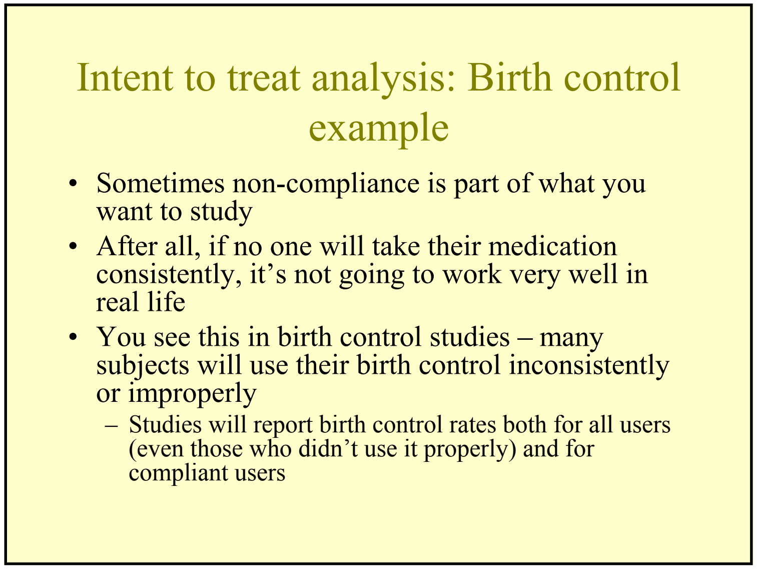
Related to the placebo effect: The Hawthorne effect

- People sometimes act differently just because they know they are being studied
- Named for a study on the effect of changes in working conditions on worker productivity, in the Hawthorne plant of Western Electric.
- The initial improvement in productivity was supposedly due to management’s demonstrated interest in improving working conditions, not to any of the actual changes
- Some questions about the original experiment: Only 5 subjects, 2 laid off for gross insubordination part way through the study. But the effect is still known as the Hawthorne effect, and is likely a real effect



The best laid plans of mice and men…

- People in a nice randomized controlled experiment still sometimes do what they want instead of what you told them to do
  - Some people assigned to placebo will go to their private doctor and be treated with the experimental drug
  - Some people assigned to the active drug will not take their medicine
- Part observational study, part controlled experiment
- How do you analyze these people?



Intent to treat analysis: Birth control example

- Sometimes non-compliance is part of what you want to study
- After all, if no one will take their medication consistently, it’s not going to work very well in real life
- You see this in birth control studies – many subjects will use their birth control inconsistently or improperly
  - Studies will report birth control rates both for all users (even those who didn’t use it properly) and for compliant users

Intent to Treat Analysis

- People should be analyzed in the groups that they were assigned to
  - If assigned to placebo, analyze as placebo
  - If assigned to active drug, analyze as active drug, even if you have evidence that they did not take the drug

Intent to Treat Analysis

- An individual assigned to a particular intervention group is included in that group's outcome statistics even if he/she never receives the intervention
- Preserves the full value of randomization

An example

- Study effect of the Atkins diet in pregnant women on fetal birth weight
- Many women don't stick with the diet

Group	#	Avg. weight
Treatment		
Compliant	500	5.5 lbs
No-compliant	1000	6.0 lbs
Total	1500	5.8 lbs
Control	1500	6.2 lbs

- Is there evidence that diet reduces birth weight? Yes.
- Which two numbers indicate this?

  5.8 lbs and 6.2 lbs.
- You cannot conclude this from 5.5 lbs and 6.2 lbs – subjects chose whether or not to be compliant, not the experimenter. Some confounding factor might cause women both to be compliant dieters, and to have lower birth weight babies.

Another example of analyzing this kind of data

- Coronary Drug Project
- Men given clofibrate vs. placebo
- Outcome was 5-year mortality rate

Group	Clofibrate	Placebo
Adherers	15%	15%
Non-adherers	25%	28%
Total	20%	21%

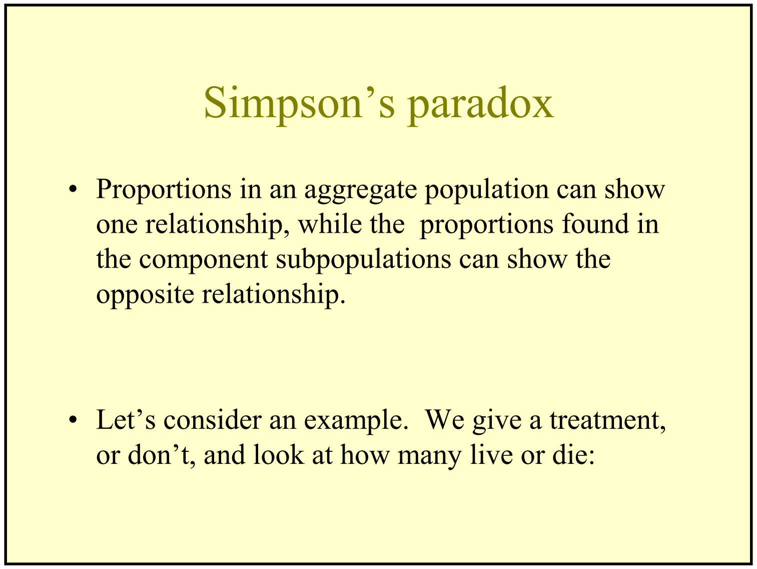
- At first glance, seems like evidence the drug works: 15% vs. 25%.
- But this is just an observational comparison. Perhaps adherers differ in some relevant way from non-adherers, other than drug treatment.
- Among adherers, compare 15% (drug) to 15% (placebo). (They don't know which drug they're getting.)
  - Just being an adherer gives you a lower mortality rate – perhaps you care more about your health.
  - Clofibrate seems to have no effect.

Minimizing the possibility of confounding factors

- Much of experimental design is aimed, at least in part, at this problem
  - Observational studies vs. controlled experiments
  - Contemporaneous vs. historical controls
  - Other issues with choosing the proper treatment and control/comparison group
  - Use of placebos
  - Double-blind experiments
  - The Hawthorne effect
- We'll talk about these design issues, as well as about Simpson's Paradox, which is related

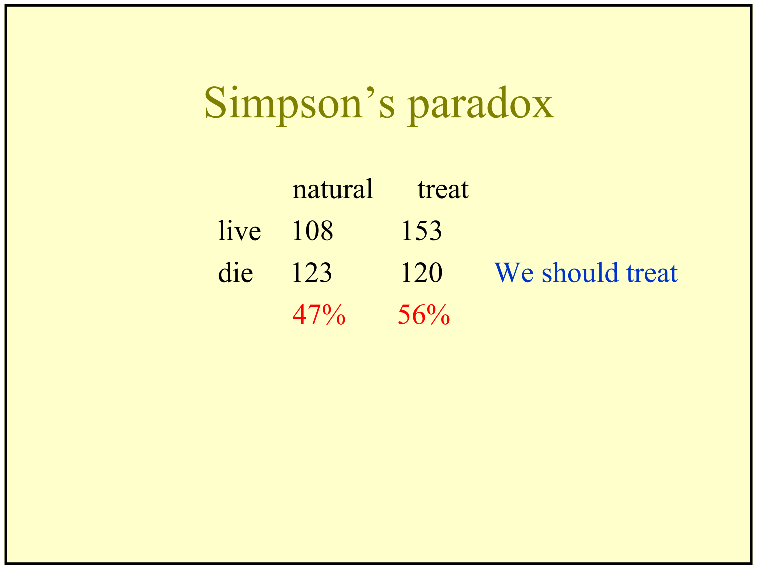
More on hidden variables

- Simpson's Paradox:
  - When data from several groups are combined, the direction of an association can reverse



Simpson's paradox

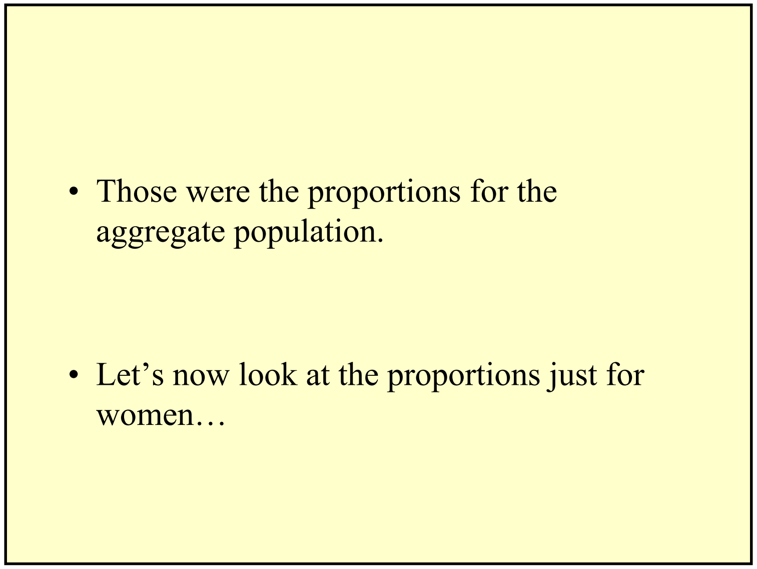
- Proportions in an aggregate population can show one relationship, while the proportions found in the component subpopulations can show the opposite relationship.
- Let's consider an example. We give a treatment, or don't, and look at how many live or die:



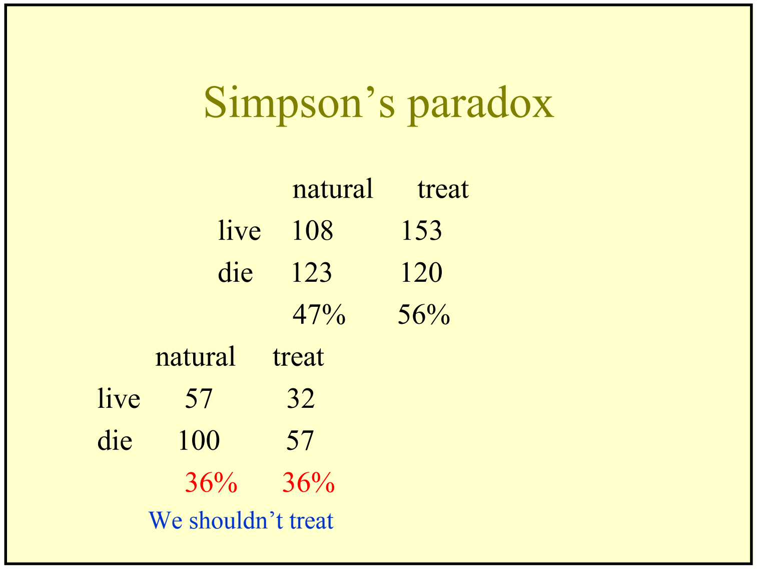
Simpson's paradox

	natural	treat
live	108	153
die	123	120
	47%	56%

We should treat



- Those were the proportions for the aggregate population.
- Let's now look at the proportions just for women…

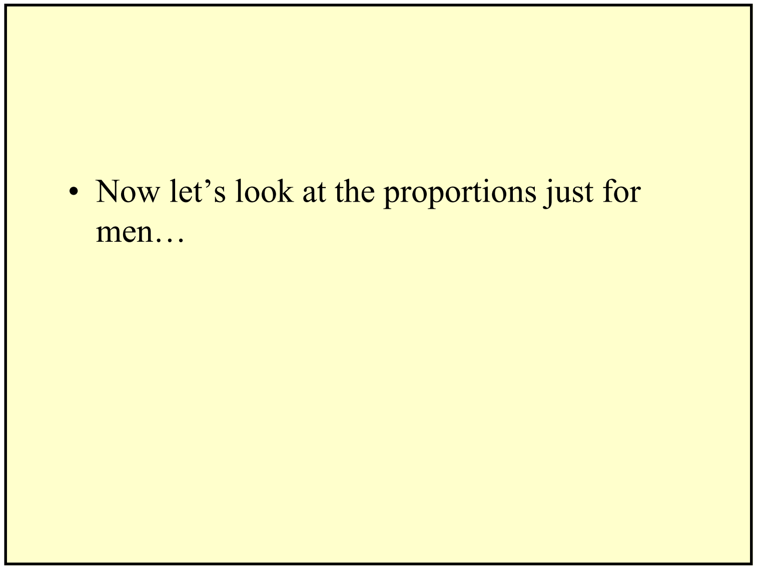


Simpson's paradox

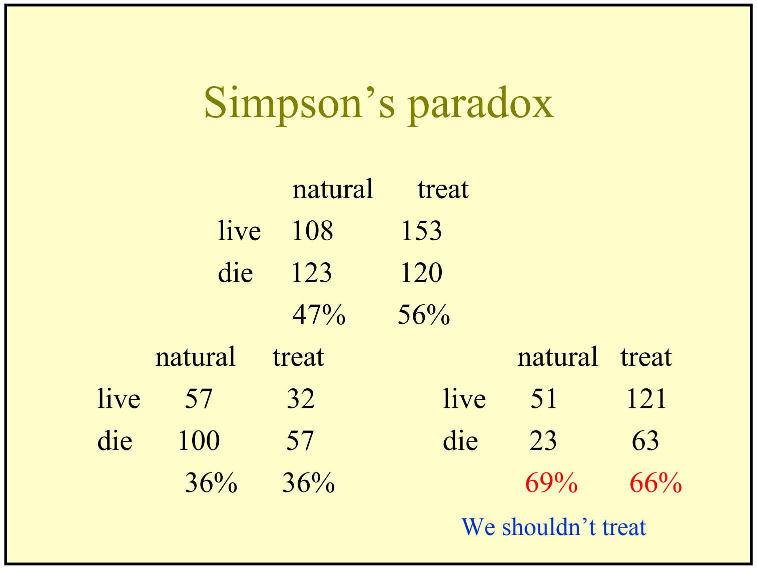
	natural	treat
live	108	153
die	123	120
	47%	56%

	natural	treat
live	57	32
die	100	57
	36%	36%

We shouldn't treat



- Now let's look at the proportions just for men…



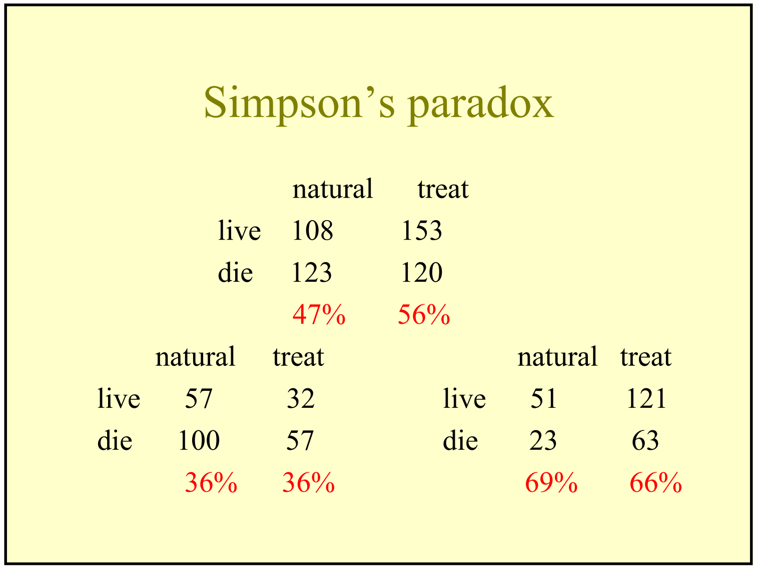
Simpson's paradox

	natural	treat
live	108	153
die	123	120
	47%	56%

	natural	treat
live	57	32
die	100	57
	36%	36%

	natural	treat
live	51	121
die	23	63
	69%	66%

We shouldn't treat

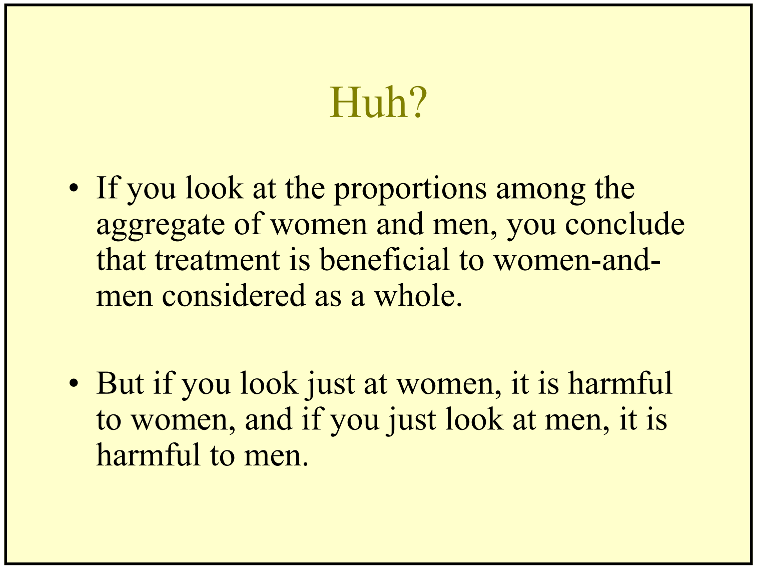


Simpson's paradox

	natural	treat
live	108	153
die	123	120
	47%	56%

	natural	treat
live	57	32
die	100	57
	36%	36%

	natural	treat
live	51	121
die	23	63
	69%	66%



Huh?

- If you look at the proportions among the aggregate of women and men, you conclude that treatment is beneficial to women-and-men considered as a whole.
- But if you look just at women, it is harmful to women, and if you just look at men, it is harmful to men.

How did this happen?

	natural	treat
live	108	153
die	123	120
	47%	56%

	natural	treat
live	57	32
die	100	57
	36%	36%

	natural	treat
live	51	121
die	23	63
	69%	66%

And most of them were not treated

Prognosis for women is not good in general

How did this happen?

	natural	treat
live	108	153
die	123	120
	47%	56%

	natural	treat
live	57	32
die	100	57
	36%	36%

	natural	treat
live	51	121
die	23	63
	69%	66%

And most of them were treated

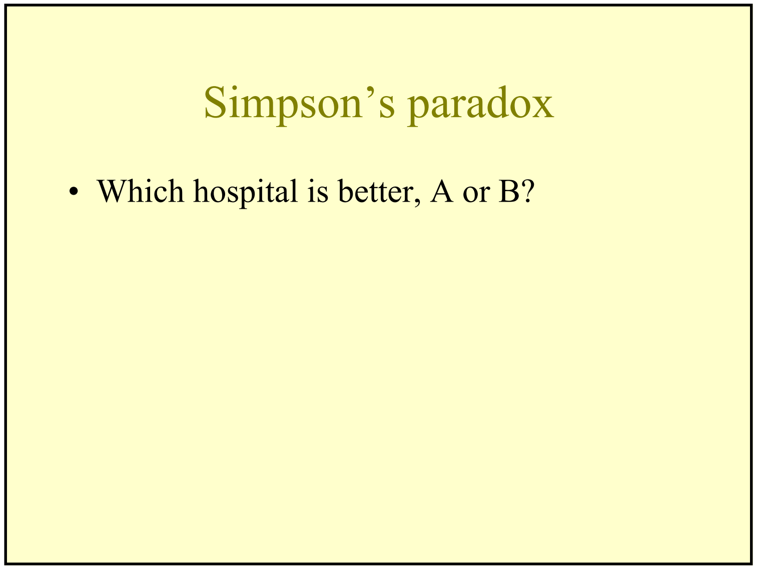
Prognosis for men is good in general

How does this happen

- Putting more women into the no-treatment group, when they have a poor prognosis, biased the aggregate results against no-treatment.
- Putting more men into the treatment group, when they have a good prognosis, biased the aggregate results against treatment.

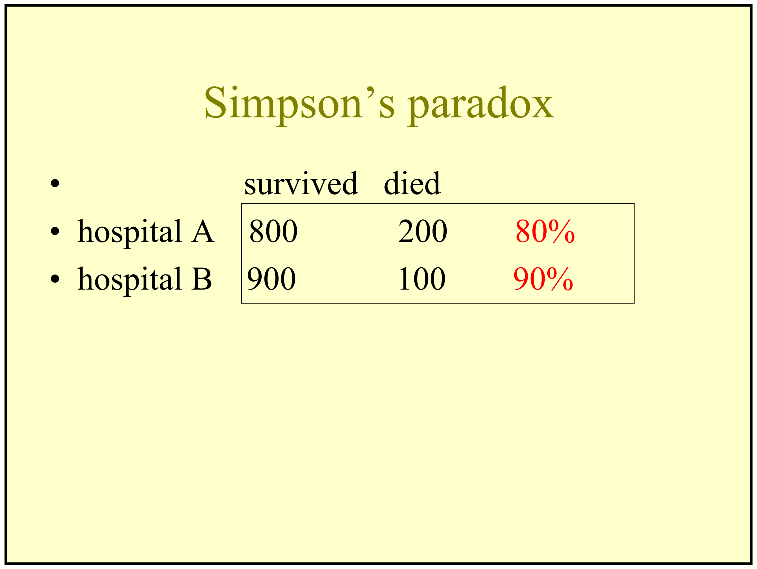
Simpson’s paradox

- Any statistical relationship between two variables can be reversed by looking at additional factors in the analysis.
- Let’s look at one more example…



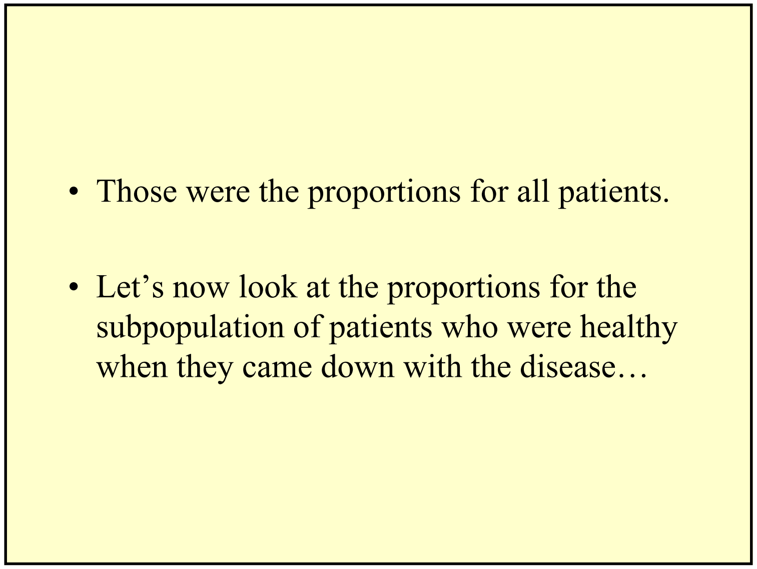
Simpson’s paradox

- Which hospital is better, A or B?

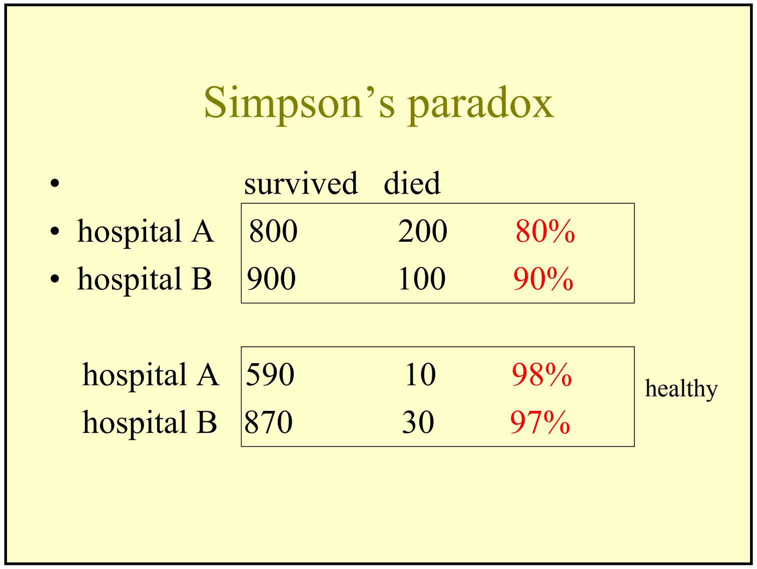


Simpson’s paradox

	survived	died	
hospital A	800	200	80%
hospital B	900	100	90%



- Those were the proportions for all patients.
- Let’s now look at the proportions for the subpopulation of patients who were healthy when they came down with the disease…



Simpson’s paradox

	survived	died	
hospital A	800	200	80%
hospital B	900	100	90%

hospital A	590	10	98%	healthy
hospital B	870	30	97%	

- Let’s look at the proportions for the subpopulation of patients who were already sick/feeble when they caught the disease…

Simpson’s paradox

	survived	died		
hospital A	800	200	80%	
hospital B	900	100	90%	
hospital A	590	10	98%	healthy
hospital B	870	30	97%	
hospital A	210	190	53%	sick
hospital B	30	70	30%	

- Hospital A takes more patients who were sick when they caught the disease (and thus had a poorer prognosis).
- This biases the aggregate results against Hospital A, even though it does better on both subpopulations.

- Ponder this.
- Have a good Spring Break!