To Do

Read Chapter 8

Do Chapter 8 Problems 1-6

Detailed information about Final Exam is posted on Learn.

My exam office hours are posted on Learn.

Today's Class

1) **What does the statement "X causes Y " mean?**

2) **Relationships and Causation**
(Confounding and Lurking Variates)

3) **How to establish causation in experimental studies - the importance of randomization.**

4) **Observational studies and establishing causation.**

Cause and Effect

Definition:

***x* has a causal effect on *Y* if, when all other factors that affect *Y* are held constant, a change in *x* induces a change in a property of the distribution of *Y* (e.g. E(*Y*), P(*Y* > c)).**

Unfortunately this definition is impractical since we cannot hold all other factors that affect *y* constant. (We may not even know what all the factors are!)

The definition serves as an ideal that should be used to conduct studies in order to show that a causal relationship exists.

Some reasons two variates can be related:

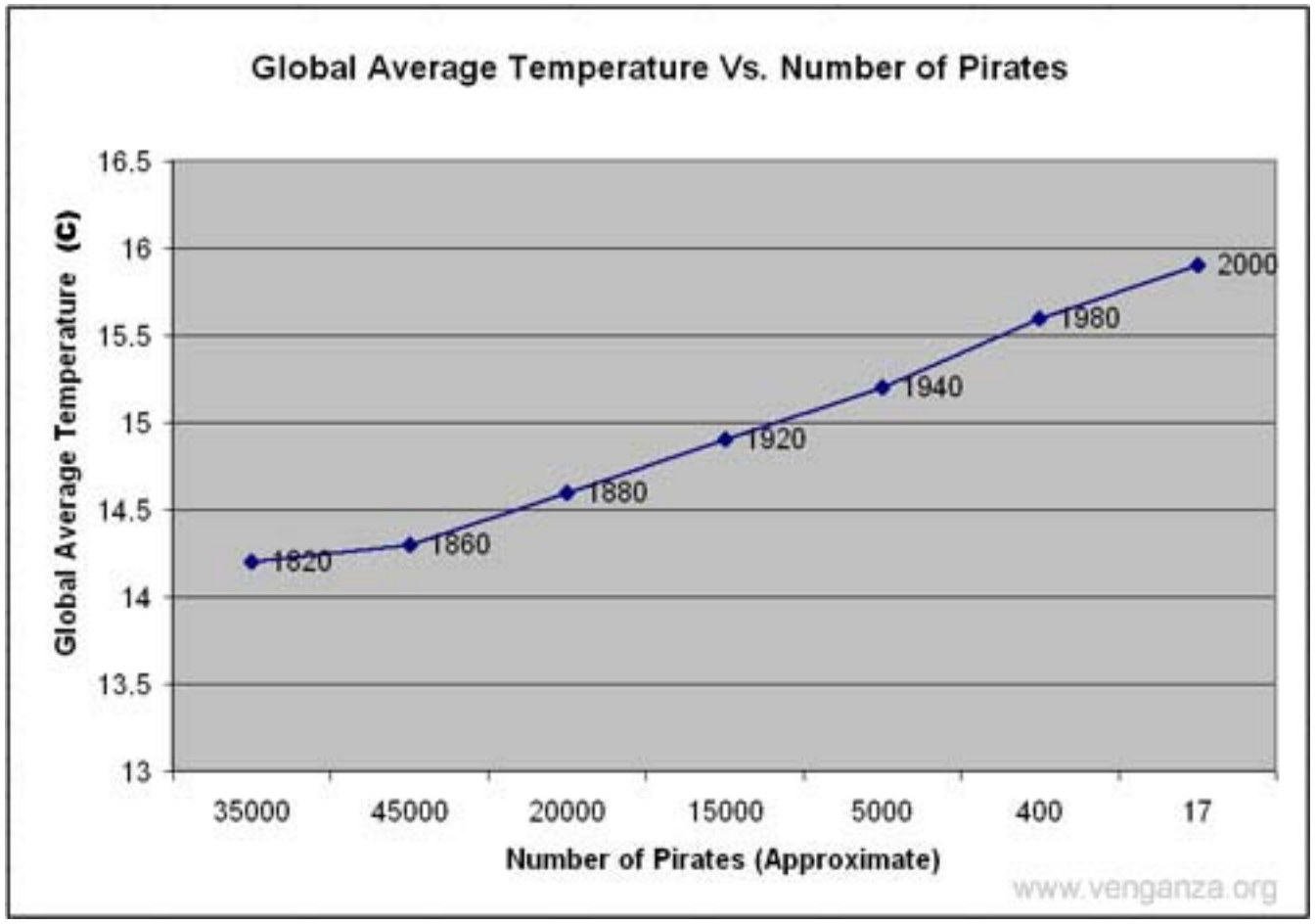
1) The explanatory variate is the direct cause of the response variate.
2) The response variate is causing a change in the explanatory variate.
3) The explanatory variate is a contributing but not sole cause of the response variate.
4) Both variates are changing with time.
5) The association may be due to coincidence.
6) Confounding variates may exist.
7) Both variates may result from a common cause.

Reason 4: Both variates are changing over time.

Nonsensical associations often result from correlating two variates that are both changing over time.

For example a strong correlation between number of women in the workforce and number of Christmas trees sold in Canada between 1930 and the present would be observed since both of these numbers are increasing with time.

Global Warming and Pirates



Reason 5: The association may be nothing more than coincidence.

For example, suppose a new office building opened and within a year there was an unusually high rate of brain cancer among workers in the building.

Suppose someone calculated that the odds of having that many cases in one building were only 1 in 10,000.

We might suspect that something wrong in the environment was causing people to develop brain cancer.

The problem with this reasoning is that it focuses on the odds of seeing such a rare event in that particular building in that particular city. It fails to take into account that there are thousands of new office buildings.

If the odds really were only 1 in 10,000, we should expect to see this phenomenon just by chance in about 1 of every 10,000 buildings.

It would be unusual if we did not occasionally see clusters of diseases as chance occurrences.

Reason 6: Confounding variates may exist.

Definition:

Two variates are confounded if their effects on a third variate cannot be separated.

Thus both variates may help cause the change in the third variate, but there is no way to establish how much is due to one and how much is due to the other.

For example, minority students in the U.S. have lower average scores on college entrance exams such as the SAT than do white students.

But minorities (again on the average) grew up in poorer households and attended poorer schools than did white students.

The effects of social and economic conditions on the test scores are mixed together in a way that makes any cause and effect conclusion impossible.

Reason 7: Both variates may result from a common cause.

An association between two variates may be observed because both variates are responding to changes in some unobserved variate or variates.

These variates are sometimes referred to as lurking variates.

Lurking Variate - Example

For example, a much used explanation by the tobacco companies for the association between smoking and lung cancer is that smoking behaviour and lung cancer are both responses to a genetic predisposition.

Sex Bias in Graduate Admissions at Berkeley University - 1973

	Admitted	Denied	Total
Male	**1158**	**1493**	**2651**
Female	**557**	**1278**	**1835**

PJ Bickel, EA Hammel, and JW O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley" *Science*, **187** (1975), 393-404.

Sex Bias in Graduate Admissions at Berkeley University - 1973

	Percent Admitted	Percent Denied
Male	43.7	56.3
Female	30.4	69.6

Looks like a serious case of sex bias.
Be careful!

Sex Bias in Graduate Admissions at Berkeley University - 1973

	Males		Females	
Program	No. applicants	% admitted	No. applicants	% admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Explanation:

1) The first 2 programs were easy to get into. Over 50% of the men applied to these two.

2) The other 4 programs were much harder to get into. Over 50% of the women applied to these four.

The Dangers of Lurking Variables

Program is a lurking variate.

Furthermore, it is related to the sex (male or female) of an applicant, so we cannot ignore it in trying to see if there is a relationship between sex and admission rates.

Remark: The feature illustrated in this example is called Simpson's Paradox.

Simpson's Paradox

For events A, B_1, B_2 and $C_1, C_2, \ldots, C_k$, it is possible to have $P(A \mid B_1 \cap C_i) > P(A \mid B_2 \cap C_i)$ for each $i = 1,2,\ldots,k$ and yet $P(A \mid B_1) < P(A \mid B_2)$.

Note that

$$P(A \mid B_1) = \sum_{i=1}^{k} P(A \mid B_1 \cap C_i) P(C_i \mid B_1)$$

and similarly for $P(A \mid B_2)$, so they depend on $P(C_i \mid B_1)$ and $P(C_i \mid B_2)$.

In the Berkeley example

A = {person is admitted}, B_1= {person is female}, B_2= {person is male},

C_i ={person applies to Program i}, $i = 1,2,\ldots,k$.

Cause and Effect

Can we conclude that your STAT 230 grade causes your STAT 231 grade?

Can we conclude that your country of hometown causes your program?

Can we conclude that your handspan causes your foot length?

Establishing Cause and Effect

Given the number of possible explanations for association between two variates, how do we ever establish that there is actually a causal connection?

Ideally, in establishing a causal connection, we would change nothing in the environment except the suspected causal variate and then measure the result on the suspected response variate.

Establishing Cause and Effect

The best method -- indeed, the only compelling method -- of establishing a causal connection is to conduct a carefully designed experiment in which the effects of possible confounding or lurking variates are controlled.

Problem: How to “control” all the possible variates if we don’t even know what they are?

Establishing Cause and Effect

Problem:

It is not possible to “control” all the variates since in many cases it is not ethical and in other cases we don’t even know what all the variates are that should be controlled.

Solution:

Randomization

The Importance of Randomization

If we have a large sample, and we use proper randomization, we can assume that the levels of confounding variates will be about equal in the two treatment groups.

Randomization also reduces the chances that an observed association is due to lurking variates.

Recall how the vitamin D/influenza A experiment was conducted.

Aspirin and the Risk of Stroke

Suppose 500 people at high risk of stroke, have agreed to take part in a clinical trial to assess whether aspirin lowers the risk of stroke.

These people are representative of a population of high risk individuals.

The study is conducted by giving some people aspirin and some people a placebo, and then comparing the two groups in terms of the number of strokes observed.

Aspirin and the Risk of Stroke

Other variates such as age, sex, weight, existence of high blood pressure, and diet also may affect the risk of stroke.

These variates obviously vary substantially across persons and cannot be held constant or otherwise controlled.

Randomization is used to deal with this problem.

The Power of Randomization

The distribution of confounding variates such as dietary factors and blood pressure among the subjects in the aspirin group and among the placebo group will be similar due to the randomization.

The distribution of any other lurking variates we have not even considered will also be similar in both groups.

If a lower risk in the aspirin group is observed then it can only be due to taking aspirin.

Smoking and Lung Cancer

One way the lung cancer/smoking controversy (that raged for a long time and which still continues) could have been settled is to conduct the following experiment:

Select a large number of people at birth, forcing half of them to smoke and the other half to abstain and observe the subjects until they died.

Such an experiment would settle the issue since smoking or not smoking would be imposed on the subjects independent of their heredity or their lifestyle.

Smoking and Lung Cancer

Alternatively a smaller matched pairs experiment could be conducted in which identical twins are selected.

One twin is forced to smoke and the other is forced to abstain and then they are forced to live identical lives and are observed until they die. Obviously it would be unethical to conduct either experiment.

Example

See Example 8.2.2 in Course Notes.

Designing a good experiment is not always easy!

Establishing Causation in Observational Studies

In observational studies controlling variates and using randomization is not possible. Establishing causation in observational studies is much more difficult and requires at least the following four features:

1) The association between the two variates must be observed in many studies of different types among different groups. This reduces the chance that an observed association is due to a defect in one type of study or a peculiarity in one group of subjects

Establishing Causation in Observational Studies

2) The association must continue to hold when the effects of plausible confounding variates are taken into account.

3) There must be a plausible scientific explanation for the direct influence of one variate on the other variate, so that a causal link does not depend on the observed association alone.

4) There must be a consistent response, that is, one variate always increases (decreases) as the other variate increases.

Smoking and Lung Cancer

The claim that cigarette smoking causes lung cancer meets these criteria.

A strong association has been observed in numerous studies in many countries.

Many possible sources of confounding variates have been examined in these studies and have not been found to explain the association.

For example, data about nonsmokers who are exposed to secondhand smoke contradicts the genetic hypothesis.

Smoking and Lung Cancer

Animal experiments have demonstrated conclusively that tobacco smoke contains substances that cause cancerous tumors.

Therefore there is a known pathway by which smoking causes lung cancer.

The lung cancer rates for ex-smokers decrease over time since smoking cessation.

The evidence for causation here is about as strong as non-experimental evidence can be.

Section 8.4 - Clofibrate Study

Example of an experimental study followed by an observational study.