Experimental Designs

Marek J. Drużdżel

Politechnika Białostocka
Wydział Informatyki

m.druzdzel@pb.edu.pl
http://www.wi.pb.edu.pl/~druzdzel/

Overview

- The value of experimentation
- Internal and external validity of studies
- Simple experimental designs
- Truly experimental designs
- Concluding remarks

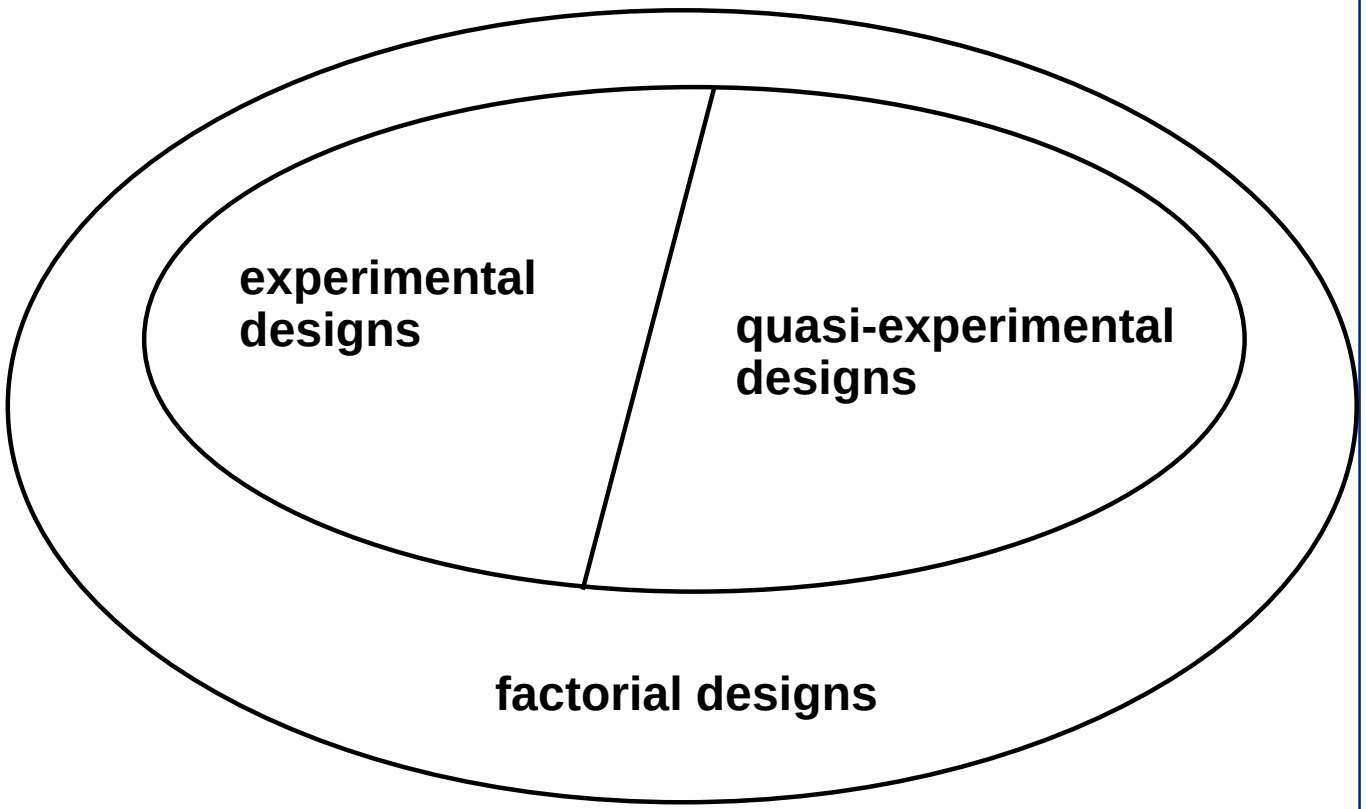
The remaining two classes in this block:

- Factorial designs
- Quasi-experimental designs

Experimental designs roadmap

How are different experimental and quasi-experimental designs related?

Experimental and quasi-experimental designs are two sides of the same coin. Of course, experimental designs are stronger, but they are aimed at the same goal - establishing causality.



Factorial designs just add one more dimension and can be both experimental and quasi-experimental (i.e., one can apply ANOVA to both designs).

How to tie experimentation to real life?

- Usually we have a problem.
- We then formulate a hypothesis.
- Subsequently we test it by an experiment.

The necessity of empirical work

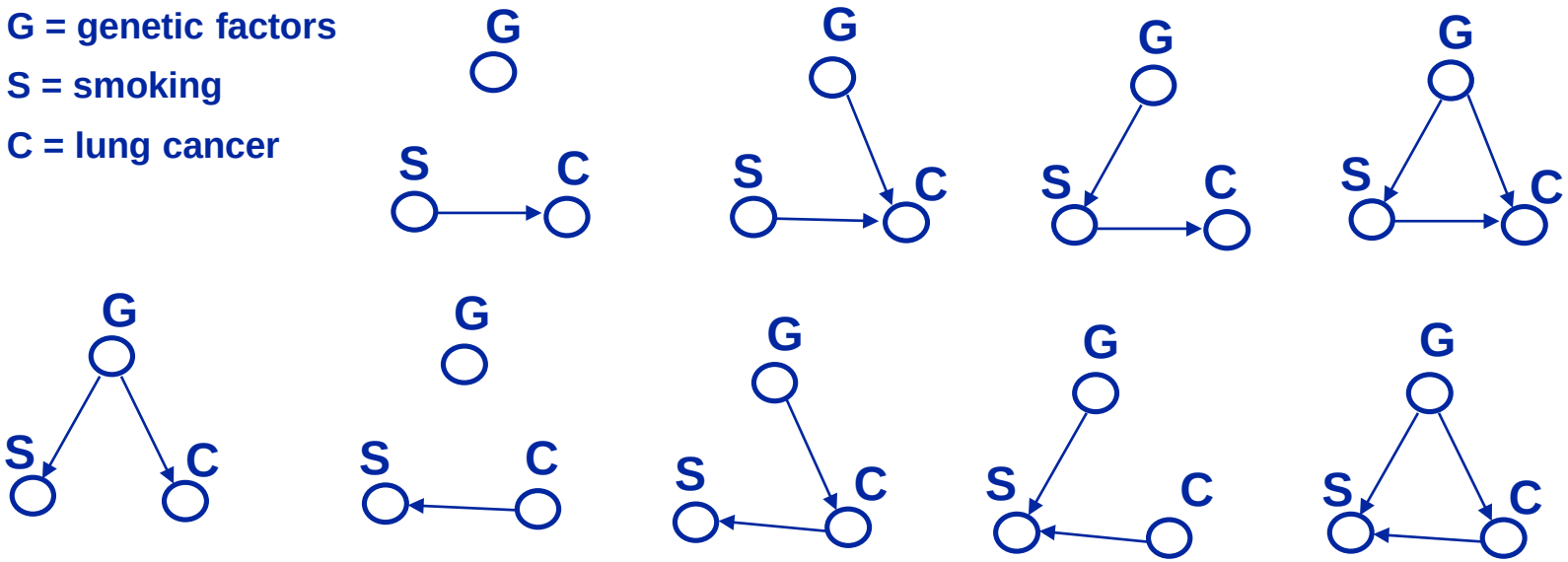
Smoking and lung cancer are correlated.
Can we reduce the incidence of lung cancer by reducing smoking?
In other words: Is smoking a cause of lung cancer?

Each of the following causal structures is compatible with the observed correlation:

G = genetic factors

S = smoking

C = lung cancer



The value of experimentation

It's a good idea to ask oneself the question:

What is the value of experimentation?

Experimentation should not be thought as a panacea to world illnesses, i.e., we should not throw out tradition, conventional common sense wisdom.

A prudent view of experimentation is that it is a necessity: a route to cumulative progress.

This route is difficult: tedious and sometimes disappointing. Following it requires thorough persistence.

Internal and external validity

Validity of an experiment: Are the conclusions valid?

Internal validity: Did the experimental treatments make a difference in this specific experimental instance?

External validity: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Threats to external validity

9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effect of the experimental variable for the untested universe from which the experimental respondents were selected.
10. The interaction effects of selection biases and the experimental variable.
11. Reactive effects of experimental arangements, which should preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental designs.
12. Multiple-treatment interference, likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs or type 8 or 9.

Symbolic notation (Campbell & Stanley)

Two-dimensional encoding of the elements of the design.

Symbols used:

X	Treatment
O	Measurement (testing)
R	Randomization

Time precedence:

Things happen from left to right (e.g., “X O” means that treatment precedes testing).

Multiple groups:

Each group is represented by a different line. Time precedence holds across lines.

Design 1: One-shot Case Study

X O

A single group of subjects is subjected to a treatment. The effect variable is measured at the conclusion of the experiment. A change in the effect variable is attributed to the treatment.

An easy and widely applied design. Many, many problems, which we will discuss in the context of more sophisticated designs.

Design 1: One-shot Case Study

X O

- **A fundamental (commonsensical) problem here is that we are not even able to show probabilistic dependence, a necessary condition for causation. We cannot show that the following inequality holds:**

  $Pr(D|T) \neq Pr(D)$

(D means dependent variable and T means treatment)

- **This is a basic inequality that says that it matters whether we treat to whether we will observe a change in the probability distribution of the dependent variable)**

- **We do not know what the prevalence of D is in the general population. We do not know how the population would react to no treatment, etc.**

Design 2: One-Group Pretest-Posttest Design

$$O_1 \quad X \quad O_2$$

A single group of subjects is tested (O_1), subjected to a treatment, and then tested again (O_2). A change in the effect variable is attributed to the treatment.

- This design fixes the problem of the previous (one-shot case study), because we test the distribution of D in the population before conducting the experiment.
- Unfortunately there are quite a number of problems remaining ☹.

Problems:

History, maturation, testing, instrumentation, and statistical regression.

Threats to internal validity

1. **History**, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

History

$O_1 \quad X \quad O_2$

If O_1 and O_2 were conducted at different times, then the events in between may have caused the difference.
To make a plausible rival hypothesis, such event should have occurred to most of the students in the group under study.

Examples:

1. Influence of Nazi propaganda on views of Americans during the World War II [Collier, 1940, the fall of France as a rival hypothesis].
2. Influence of “tough international politics” on the end of the cold war (can also be attributed to maturation).
3. Leak in terms of old exams in classroom settings.

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. **Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.**
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Maturation

O_1 X O_2

If O_1 and O_2 were conducted at different times, then natural (biological, psychological) processes in between may have caused the difference.

Examples:

1. Testing at the beginning and the end of the school year (or course) neglects the effect of knowing more because of other coursework.
2. Influence of conservative propaganda on views of adults (they become more conservative on the average as they grow older).
3. Influence of "tough international politics" on the end of the cold war (can also be attributed to history).
4. Influence of becoming ripe to lose some pounds on the weight.

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. **Testing, the effects of taking a test upon the scores of a second testing.**
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Testing

$O_1 \quad X \quad O_2$

Test O_1 (pretest) can influence the results of test O_2. The direction of influence can be positive or negative! Usually we know the direction for different types of tests.

Examples:
1. IQ (SAT, GRE?) test results get better with each successive test.
2. Views become more extreme as one sees the questions (seeing hostile or racist statements moves the boundary of what is experienced "socially approved" or "permissible").
3. Influence of screening mammograms on the incidence of breast cancer.

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. **Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.**
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Instrumentation

$O_1 \quad X \quad O_2$

There may be changes in the measuring instrument between the time at which O_1 (pretest) was administered and the time of test O_2. Examples include different graders, changes in grading standards or practice, experience, physical instruments becoming de-calibrated.

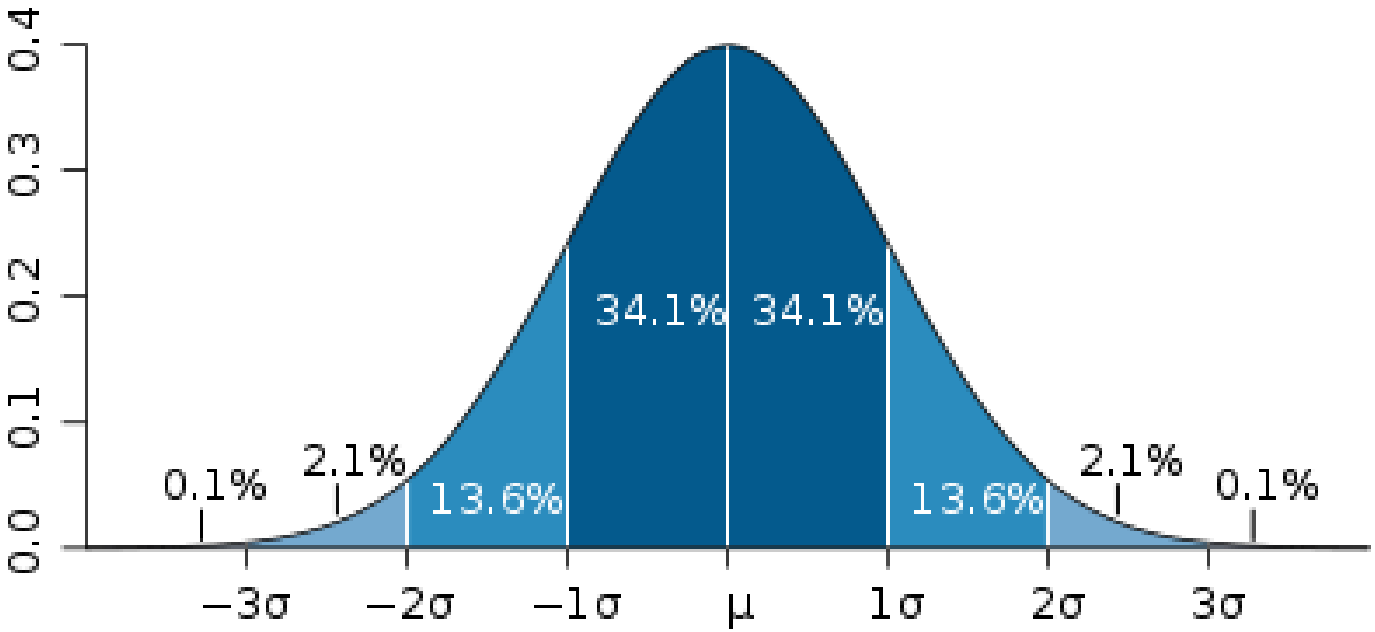
Examples:

1. Interviewer becoming more experienced as the experiment progresses.
2. A teacher lowers her standards as she goes.
3. The system administrator changes the clock speed of the CPU and does not adjust the software, so that our measurements of O_2 are consistently wrong.

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. **Statistical regression, operating where groups have been selected on the basis of their extreme scores.**
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Statistical regression



Source: http://en.wikipedia.org/wiki/Probability_distribution

When we see an unusually high or low result, there is a good chance that it is extreme for the subject. Then, less extreme result is much more likely than more extreme result.
Unusually tall chilren tend to have shorter siblings and parents.
Unusually bright/dumb children tend to have dumb/bright siblings and parents.
Extremely bad/good children behavior is usually followed by normal behavior.

Statistical regression

$O_1 \quad X \quad O_2$

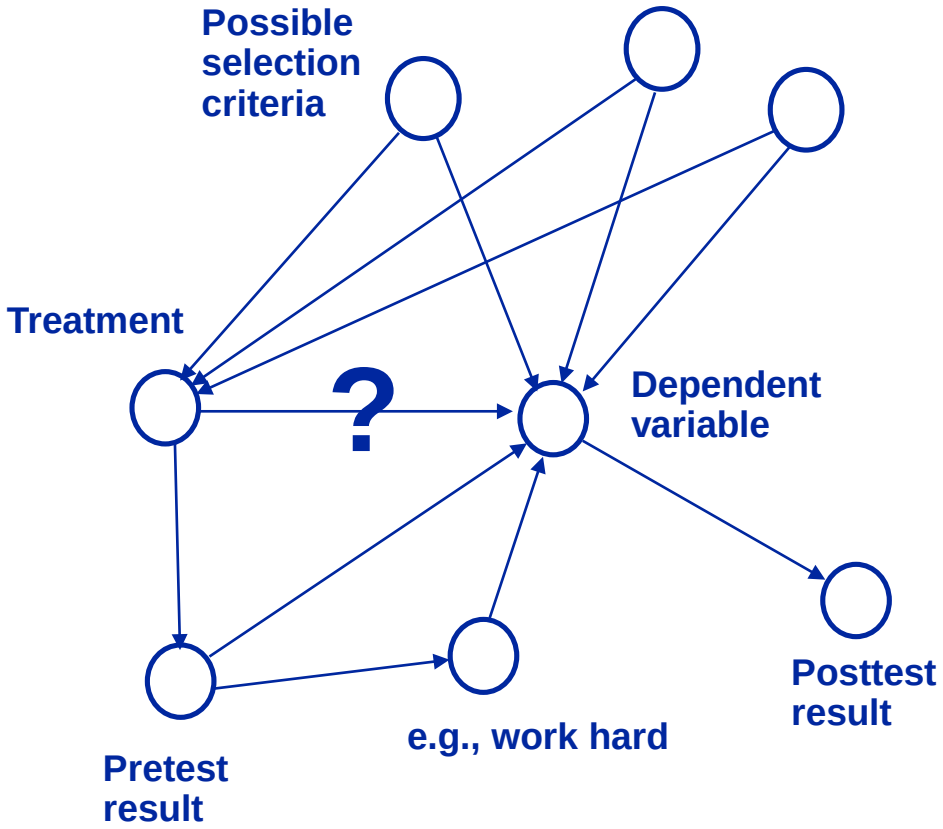
Has to do with selection of subjects: If they are selected from an extreme population, odds are for a spontaneous improvement due to statistical regression.

Examples:
1. Students with the lowest score in a reading test chosen for a remedial reading course.
2. Science program for the "smartest" kids (those who scored the highest on a single test).
3. Network nodes with the highest congestion on a given day chosen for analysis.

Design 2: Threats to internal validity

$O_1 \quad X \quad O_2$

A single group of subjects is tested (O_1), subjected to a treatment, and then tested again (O_2).
A change in the effect variable is attributed to the treatment.



Design 3: The Static-Group Comparison

$$X \quad O_1$$
$$O_2$$

A group of subjects that has experienced X is compared to a group of subjects that did not. A difference between the two groups is attributed to the treatment X.

Problems:

Selection, mortality

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. **Biases resulting in differential selection of respondents from the comparison groups.**
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

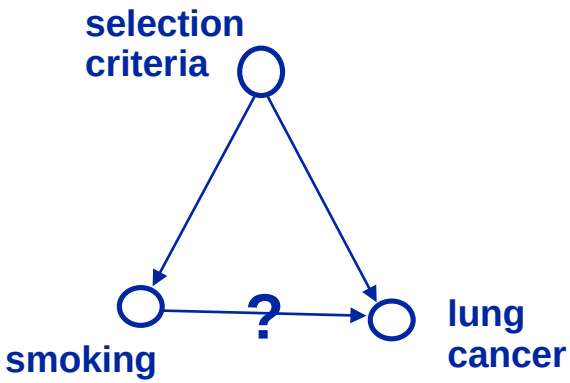
Biases due to differential selection

X O_1
O_2

There are no formal means of certifying that the groups would have been equivalent had it not been for X.

If O_1 and O_2 differ, this difference could well have come about through the differential recruitment of persons making up the groups: the groups might have differed anyway, without occurrence of X.

Examples:
Plenty ...



Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. **Experimental mortality, or differential loss of respondents from the comparison groups.**
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Experimental mortality

X O_1
O_2

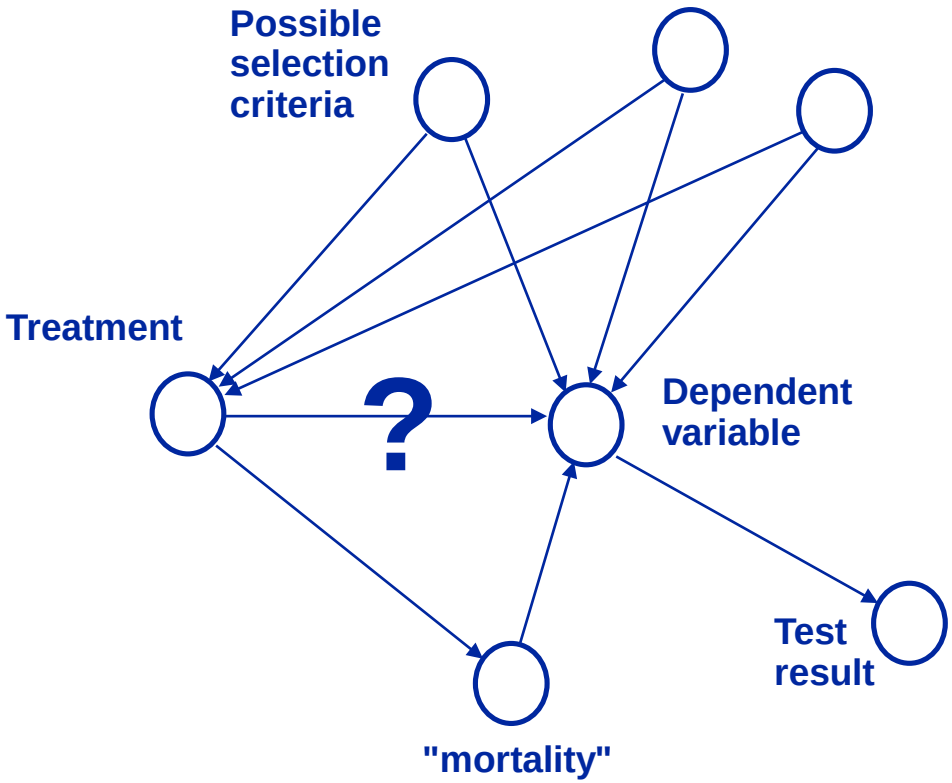
The differences between O_1 and O_2 can come about due to differential drop-out of persons from the groups.
Thus, even if the groups had once been identical, they might differ now not because of any change on the part of individual members, but rather because of selective dropouts of persons from one of the groups.

Examples:
1. Influence of exercise on weight.
2. Influence of smoking on lung cancer.
3. Influence of college education on beauty.

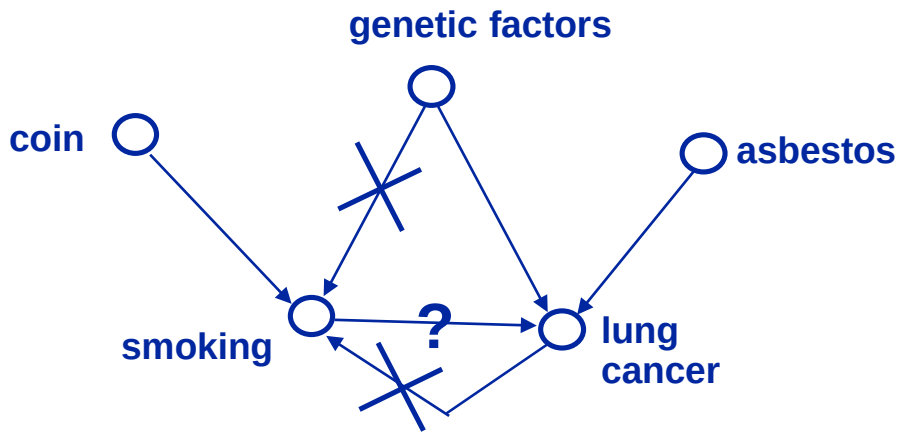
Design 3: Threats to internal validity

$$X \quad O_1$$
$$O_2$$

A group of subjects that has experienced X is compared to a group of subjects that did not. A difference between the two groups is attributed to the treatment X.



Truly experimental designs: Randomization



Solve problems related to selection criteria

- **In a randomized experiment, coin becomes the only cause of smoking.**
- **Smoking and lung cancer will be dependent only if there is a causal influence from smoking to lung cancer.**
- **If Pr(C|S) ≠ Pr(C|~S) then smoking is a cause of lung cancer.**
- **Asbestos will simply cause variability in lung cancer.**

Design 4: The Pretest-Posttest Control Design Group

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Two groups of subjects, equivalence is achieved by randomization. Both groups are tested before and after the experiment. One of the groups undergoes treatment.

How does this design deal with the threats?

Threats to internal validity

1. History, the specific events occurring between the first and second measurement in addition to the experimental variable.
2. Maturation, processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. Testing, the effects of taking a test upon the scores of a second testing.
4. Instrumentation, in which changes in the calibration of a measuring instrument of changes in the observers or scorers used may produce changes in the obtained measurements.
5. Statistical regression, operating where groups have been selected on the basis of their extreme scores.
6. Biases resulting in differential selection of respondents from the comparison groups.
7. Experimental mortality, or differential loss of respondents from the comparison groups.
8. Selection-maturation interaction, etc., which in certain of the multiple-group quasi-experimental designs, such as Design 10, is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Design 4: Dealing with history

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

History is controlled in so far as general historical events that might have produced an O_1-O_2 difference would also produce O_3-O_4 difference. Note that the design does not control for unique intra-session history (a joke, a remark, fire across the street, the experimenter's introductory remarks, etc.). One might try to make the sessions simultaneous, videotape and present the instructions, etc.

Design 4: Maturation and testing

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Maturation and testing are controlled in that they should be manifested equally in experimental and control groups.

Design 4: Instrumentation

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Instrumentation is easily controlled where the conditions for the control of intra-session history are met, particularly where O is achieved by student responses to a fixed instrument, such as printed test.
When observers or interviewers are used, however, we should be very cautious. One trick is to have them do both experimental and control conditions, another is to blind them, yet another is to cross-validate among various graders (record, videotape the interview, preserve the original trace of the experiment).

Design 4: Regression

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Regression is controlled as far as mean differences are concerned, no matter how extreme the group is on pretest scores, if both the experimental and control groups are randomly assigned from the same extreme pool. In such a case, the control group regresses as much as does the experimental group.

Design 4: Selection

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

Selection is ruled out as an explanation of the difference to the extent that randomization has assured group equality. This extent is, of course, stated by our sampling statistics. Thus the assurance of equality is greater for large numbers of random assignments than for small. This assumption may be wrong occasionally, but we control the probability of being wrong.

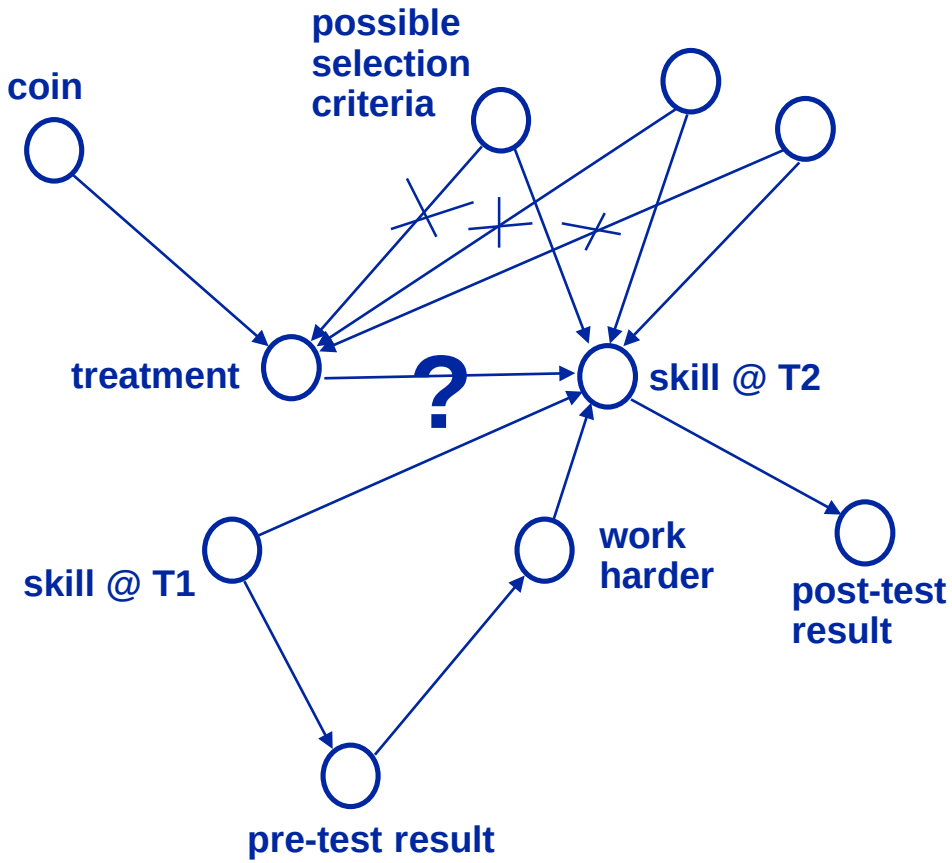
Design 4: Mortality

$$\begin{array}{llll} R & O_1 & X & O_2 \\ R & O_3 & & O_4 \end{array}$$

The data made available by Design 4 make it possible to tell whether mortality offers a plausible explanation of the O_1-O_2 gain. Mortality, lost cases, and cases for which only partial data is available, are troublesome to handle.

Design 4: Threats to internal validity

$$R \quad O_1 \quad X \quad O_2$$
$$R \quad O_3 \quad\quad O_4$$



Two groups of subjects, equivalence is achieved by randomization.
Both groups are tested before and after the experiment.
One of the groups undergoes treatment.

Threats to external validity

9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pre-tested population unrepresentative of the effect of the experimental variable for the untested universe from which the experimental respondents were selected.
10. The interaction effects of selection biases and the experimental variable.
11. Reactive effects of experimental arrangements, which should preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental designs.
12. Multiple-treatment interference, likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs or type 8 or 9.

Threats to external validity

- Can be also called interaction effects, as they involve X and some other variable. They represent a potential specificity of the effect of X to some undesirably limited set of conditions.
- For example, the effects of X observed may be specific to groups warmed up by the pretest. We are logically unable to generalize to the larger un-pretested universe about which we would prefer to be able to speak.
- Problems of external validity are not logically solvable in any neat and conclusive way. Generalization always turns out to involve extrapolation into a realm not represented in one's sample.
- Hume: "… induction or generalization is never fully justified logically …"

Threats to external validity

9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pre-tested population unrepresentative of the effect of the experimental variable for the untested universe from which the experimental respondents were selected.
10. The interaction effects of selection biases and the experimental variable.
11. Reactive effects of experimental arrangements, which should preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental designs.
12. Multiple-treatment interference, likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs or type 8 or 9.

Interaction of testing and X

Subjects may be sensitized to the treatment by a pretest.

Example:

- **Attitude test (testing the degree of anti-Semitism), a subsequent movie (treatment) that deals with the problem (e.g., “Gentlemen’s Agreement”), and a post-treatment test**

Threats to external validity

9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effect of the experimental variable for the untested universe from which the experimental respondents were selected.
10. **The interaction effects of selection biases and the experimental variable.**
11. Reactive effects of experimental arrangements, which should preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental designs.
12. Multiple-treatment interference, likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs or type 8 or 9.

Interaction of selection and X

Design 4 controls for possible differences in effect between the treatment and control groups, but still the effect may be limited to the group from which all subjects have been selected.

Example:

- **“All psychology is the science of behavior of psychology undergraduate students.”**

Threats to external validity

9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effect of the experimental variable for the untested universe from which the experimental respondents were selected.
10. The interaction effects of selection biases and the experimental variable.
11. **Reactive effects of experimental arrangements, which should preclude generalization about the effect of the experimental variable upon persons being exposed to it in non-experimental designs.**
12. Multiple-treatment interference, likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable. This is a particular problem for one-group designs or type 8 or 9.

Reactive arrangements

Artificiality of the experimental settings and the subject's knowledge that she is participating in an experiment.

- **Subjects want to please the experimenter.**
- **They try (often unconsciously) to guess the experimenter's intent and address it.**
- **Example: The Hawthorne effect (people who know that they are being watched or studied, they change their behavior)**
- **Often we cannot avoid them. We should still continue our experimentation.**
- **One way to solve this is to use non-reactive measures (concealing what is being measured, one way mirrors, hidden cameras, etc.), another is to use non-invasive, naturalistic observation.**

Design 5: The Solomon Four-Group Design

R	O_1	X	O_2
R	O_3		O_4
R		X	O_5
R			O_6

Four groups of subjects, equivalence is achieved by randomization. Two groups are tested before the experiment and all four are tested and after the experiment. Two of the groups undergo treatment.

Addresses the problem of interaction of testing and treatment (a threat to external validity). (Allows for compensation, i.e., computing the effect of pretest.)

Design 6: The Posttest-Only Control Group Design

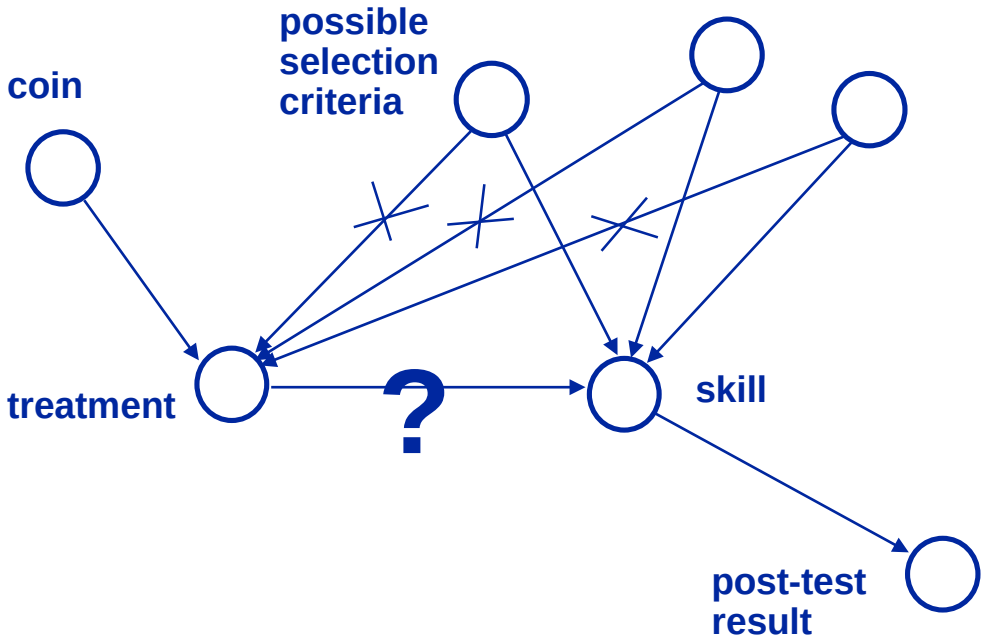
$$\begin{array}{ccc} R & X & O_1 \\ R & & O_2 \end{array}$$

Pretest is not actually essential to true experimental designs. Randomization, if done well, ensures us that the groups are equal with a high probability. Additional advantages of Design 5 may be not worth the cost.

Two groups of subjects, equivalence is achieved by randomization. Both groups are tested only after the experiment. One of the groups undergoes treatment.

Design 6: The Posttest-Only Control Group Design

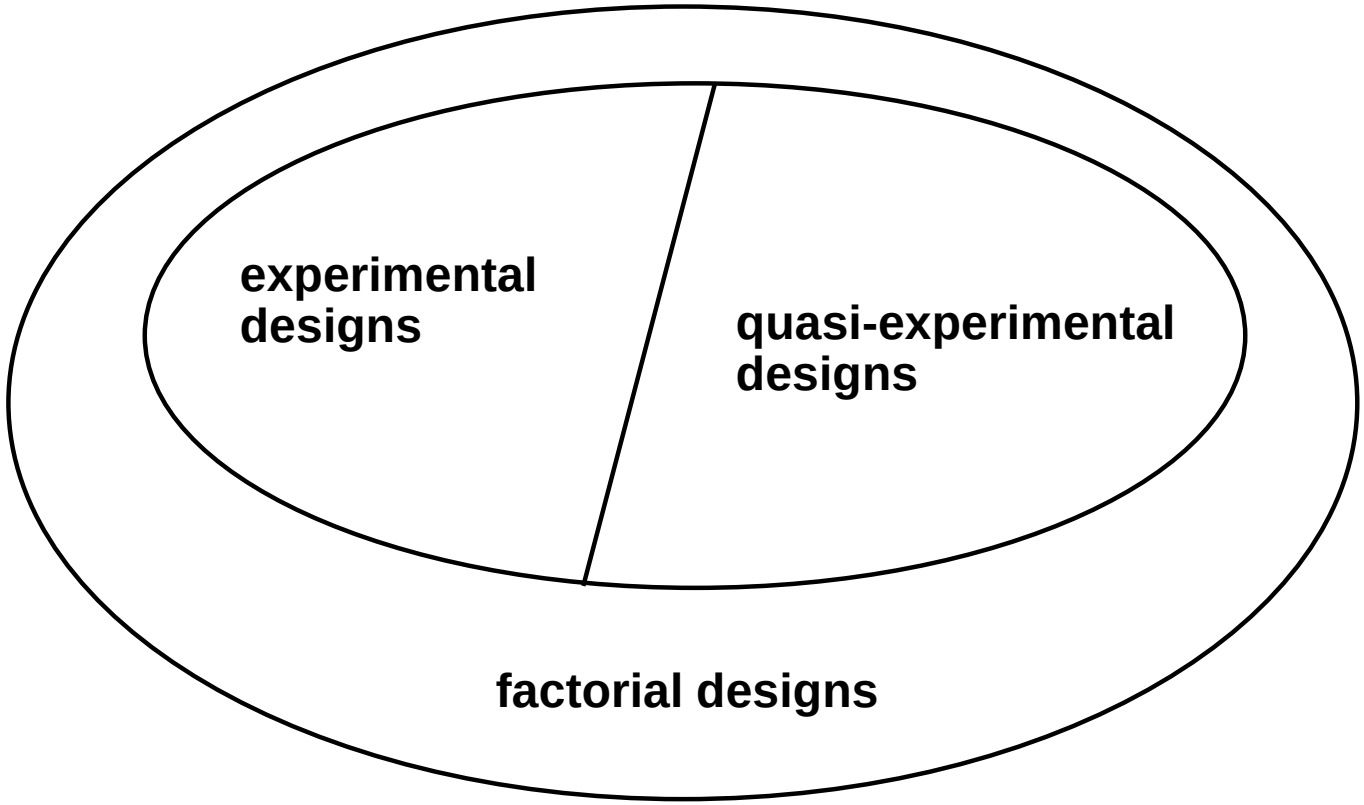
R X O_1
R O_2



Two groups of subjects, equivalence is achieved by randomization.
Both groups are tested only after the experiment.
One of the groups undergoes treatment.

Experimental designs roadmap

How are different experimental and quasi-experimental designs related?



Concluding Remarks

- The threats are just a checklist. You can suffer from them even in a proper design that is supposed to be robust to them (e.g., an event like a fire, joke, whatever, that happens to one of your groups only). The tables in C&S and +, - signs do not apply to all settings, but it is nice to be aware of the possible threats.
- How to tie all the designs to real life? You usually go in the following direction: problem –> hypothesis –> experiment
- It's a good idea not to feel bound by the existing experimental approaches. Be creative, it pays. As long as you make a good argument, people will buy your design. Each of these designs was introduced at some point and there is no guarantee that they are all that there is.
- The main purpose of this course is to make you a critical recipient of research results and to make you see what is going on in a design.