AP Statistics – Lesson Notes - Chapter 13: Experiments and Observational Studies

Observational Studies

In unit 2 (ch7,8,9,10) we looked at ***associations*** between data variables.

These data were collected by simply observing situations which were occurring. There was no attempt to control what was happening. This kind of data collection is called an **observational study**.

There are two general categories of observational studies:

Retrospective Study: Looking at information from the past. The subjects are identified after events have already occurred, and the results are studied.
Example: If studying whether learning a musical instrument improves academic performance, we may collect information from a high school senior class.

Prospective Study: The subjects are identified in advance and a **survey** is conducted once or over a period of time.
Example: We might identify a group of young students who state they will learn a musical instrument and a group who will not and track these groups' academic performance over time.

Observational studies are fine for revealing associations and strengths of associations. But in unit 2 we discovered that an association, even a strong one, does not imply causation. We do not know, for certain, that changes in one variable are *causing* the observed changes in the other variable. There may be ***lurking variables*** or other reasons which explain the association.

The only way to establish a causal relationship is to use an experiment.

Some experiment terms...

Subject, participant, experimental unit: One individual object or person to which treatment is applied and response data is measured.

Group: A collection of experimental units.

Factor: An explanatory variable whose levels are controlled.

Treatment: Applying different levels of the factor to a group.

Response variable: The variable which is measured to determine the effect of the treatment.

Activity - An experiment to test the effectiveness of a drug

Our company has created a new drug that is supposed to improve the general health of all people who use it. Using the drug over a 6 month period is expected to result in a significant increase in scores on a health survey.

Subjects: *60 people*

Factors (what we are investigating): 1 factor: drug
(2 levels: drug, no drug)

Groups/Treatments: Group 1: 30 people / Take the drug for 6 months
Group 2: 30 people / Do not take the drug

Response Variable: The score each person has on a health survey taken at the end of the 6 month period.

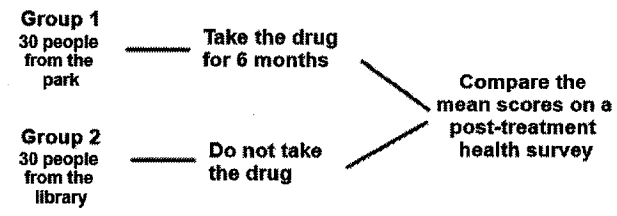
(We will compare the means of the scores of the two groups. If the mean score of the drug group is significantly higher, then the drug is working.)

We need to find some subjects. There is a nearby library with a park beside it. Both are good places to ask people to participate.

We will ask people leaving the park and the library if they would participate and will offer them a 10 year supply of the drug for free if it is found to be effective as incentive.

We set up tables at the park and the library and sign up 30 people at each location.

To keep it simple, let's use the people we signed up at the park as group 1 (who will receive the drug) and the people we signed up at the library as group 2 (who will not take the drug):



This is called an experiment design diagram

Activity - An experiment to test the effectiveness of a drug

Virtual version of this activity - please open a tab on your web browser and browse to www.mrfelling.com/experiment

Read the explanation on the virtual activity screen.

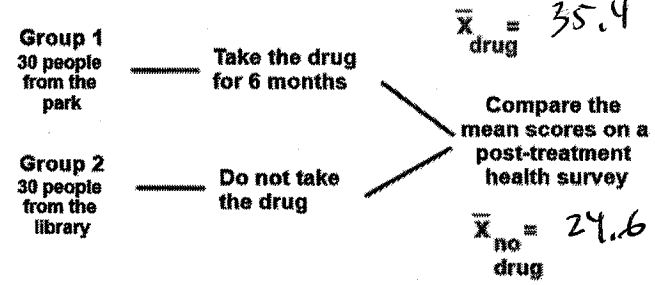
The program simulates giving each person in group 1 (from the park) the drug and each person in group 2 (from the library) no drug, and then after some time assessing their health with a health survey (higher score = better health).

1) Enter the health scores for Group 1 into L1, and the health scores for Group 2 into L2.

2) Find the mean scores for each group (1-Var Stats L1, 1-Var Stats L2)

3) Write down your two means on paper.

Would you say this experiment provides evidence that the drug is effective?



$$\text{improvement} = \bar{x}_{drug} - \bar{x}_{no\ drug}$$

$$35.4 - 24.6$$

$$= 10.8$$

(11)

Activity - An experiment to test the effectiveness of a drug

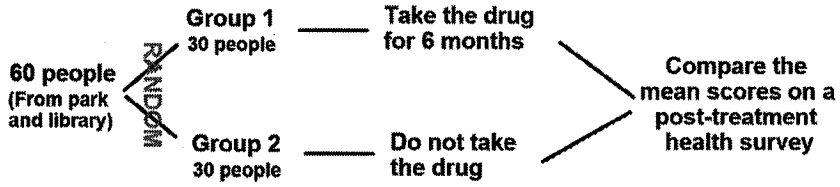
Other research groups also investigated this drug, but they found that there was no significant difference in the means, so they concluded that the drug was ineffective, and, in fact, it does turn out this drug is completely ineffective.

So what happened? Why did our experiment come to the wrong conclusion?

It turns out we do have potential uncontrolled factors between the people other than whether or not they took the drug.

The good news is there is a simple fix - we just assign people randomly to the two groups.

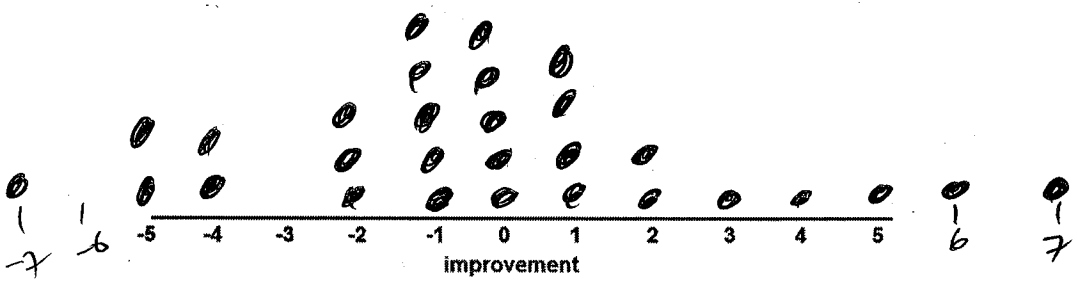
Next, we'll have conduct the experiment again, but this time with the park and library people all grouped together, and then we'll random select which group they are placed into:



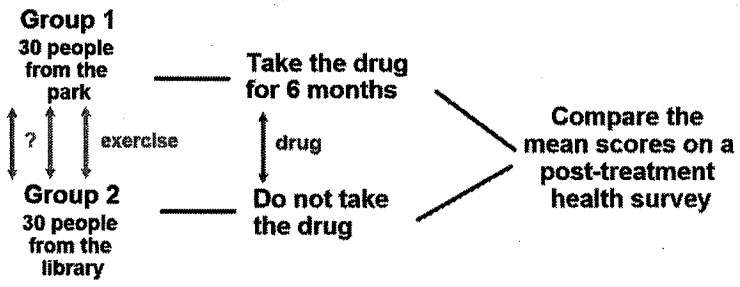
Press the 'Part 2' button the bottom to get the results of the new experiment, but the group health scores into lists and find the mean health score for these new groups.

Compute the 'improvement' number for your 2nd experiment, and type it in the chat (round to the nearest integer):

$$\text{improvement} = \bar{x}_{\text{drug}} - \bar{x}_{\text{no drug}}$$



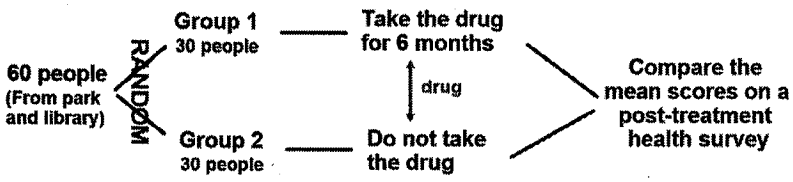
In the first experiment, there was evidence of a difference between the groups...



...but there are other difference between the groups besides taking the drug, such as 'exercise'.

We would say the variables (or factors) 'drug' and 'exercise' are confounded.

In the second experiment, we used random assignment to even out differences between the people other than the drug...



Completely randomized design

So now we can say that the effect we saw in the health scores was caused by the drug. This is the only way to establish cause-and-effect, with a well-designed experiment.

Not only will this remove the effect of exercise, it will also control for any other differences between the people (including things we aren't aware of).

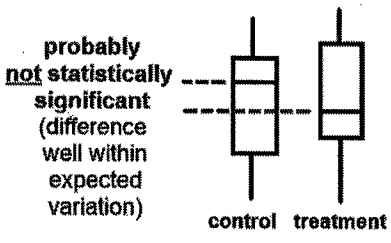
Is the difference in the response variable significant?

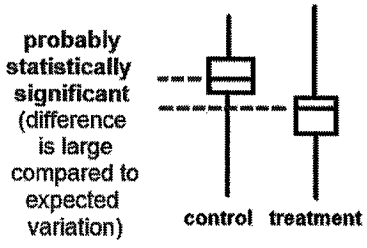
The whole point of an experiment is to observe a change in the response variable associated with the variation of a single factor. But some variation is inevitable, so how do we determine if a response variation is significant?

Later, we will be able to answer this more precisely using probability models and concepts in 2nd semester, but for now we say we have **convincing evidence** or a result is **statistically significant if the observed difference is too large for us to believe that it happened by natural variation or chance.**

We conducted a simulation of multiple experiments to see the expected variation in difference. It isn't always possible or practical to do this, so sometimes we just compare the results between the groups to the variations within the groups.

If instead of computing the 'improvement' (difference in means) we could also just compare the two data sets against the variation in the data sets:



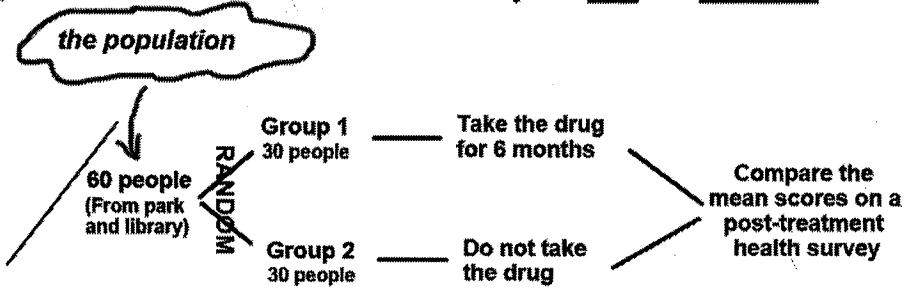


Only a well-designed experiment can conclude cause-and-effect

For a study to be called an experiment, technically, only one thing is required:

- **Researchers apply a treatment to multiple groups.**

But a "well-designed" experiment takes further measures to reduce the impact of the two enemies of statistical analysis - bias and variability



Controlling bias: Use an appropriate sampling technique to select the subjects from the population under study. This allows the conclusion to be applied as broadly as possible.

To reduce variability...

1) **Random assignment of subjects to more than one group.**

Random assignment to groups controls for (removes the variability of) differences between the subjects (known and unknown).

Note: If one group receives no treatment sometimes this is called a 'control group' but this is *not* required, you just must have any two or more groups.

2) **Control of the factors.**

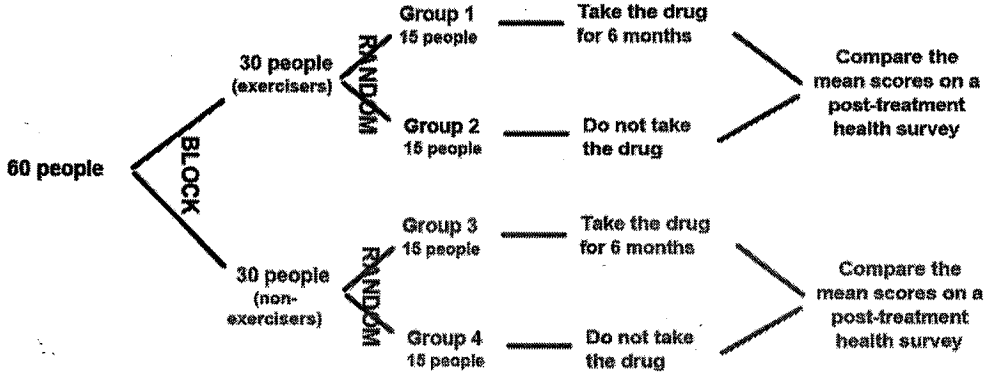
At least one factor must be under control and *imposed* as a treatment by the experimenters on the subjects.

3) **Replication. Two different meanings of replication, both important:**

- ***Replication of treatment*:** Randomization is how we control for differences in the subjects we don't know about. **There must be *enough subjects in each group*** for the 'averaging out' to work.
- ***Replication of experiment*:** Because experiments can sometimes randomly have unusual results, **the *entire experiment should be replicated*,** preferably by different researchers.

What if we know about differences in the subjects which may affect results?

If we know about differences in the *subjects* we can divide the subjects by these differences first. This is called **blocking** and the resulting experiment is called a **block-design experiment**. We run parallel experiments on each block. Here, we would "block on exercise"...



Completely randomized block design

Note: comparison is done separately within each block.

With a block design, we might find out that the drug works, but maybe only for people who are not exercising.

Block Designs

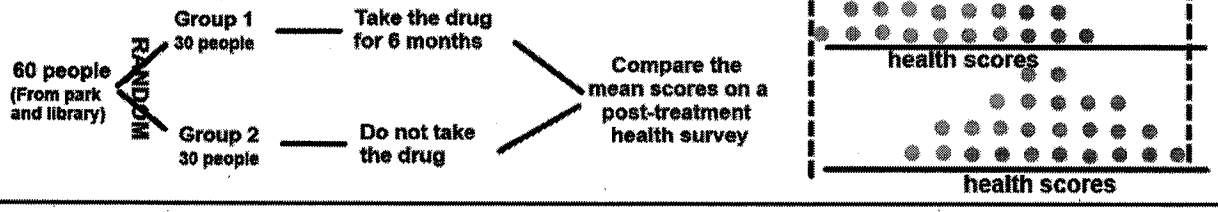
- Blocking is the same idea for experiments as stratifying is for sampling:

  Block == Experiment
  Stratified == Observational Study Sampling

- Block design and stratified sampling are done to handle *known* differences in the subjects. (Randomization is done to handle *unknown* differences).

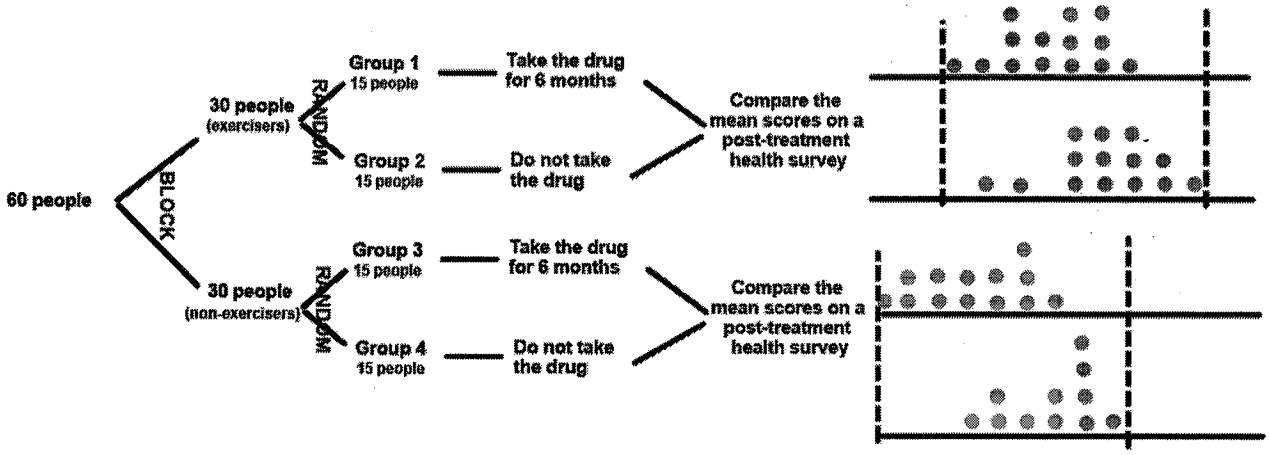
- Blocking also usually reduces natural sampling variation...we are effectively 'removing' the factor we are blocking on.

Without blocking...



With blocking...

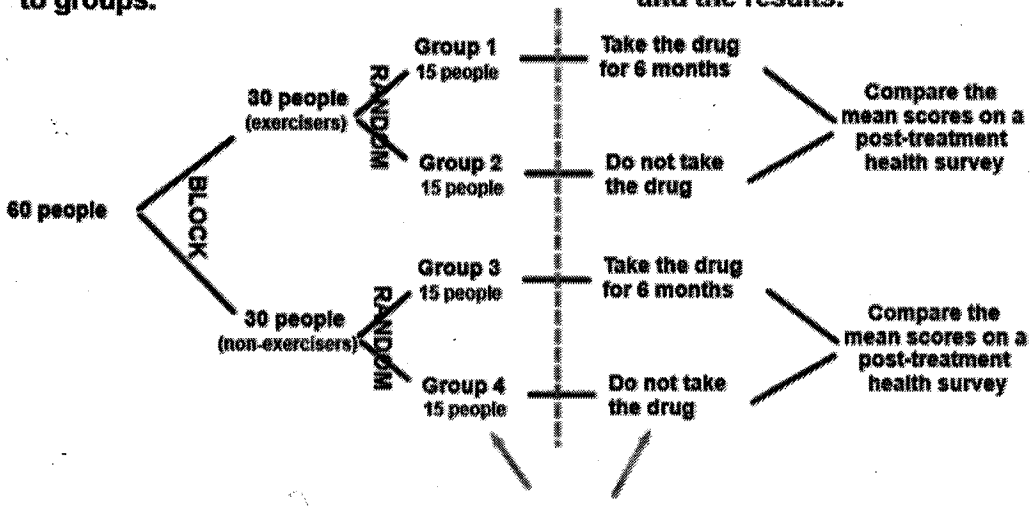
There may be less variability in results within each block, so it can be easier to find difference (evidence) 'convincing'



Experiment Diagrams

This half of the diagram is about the *subjects* and how they are assigned to groups.

This half of the diagram is about the *treatments* imposed on the groups and the results.



Note: groups and treatments are always shown separately

Controlling more than one factor

Let's say we wanted to investigate two new drug's ability to improve health scores. Maybe the first drug (drug A) comes in two dosages (10mg, 20mg), and the second drug (drug B) comes in only one dosage (50 mg):

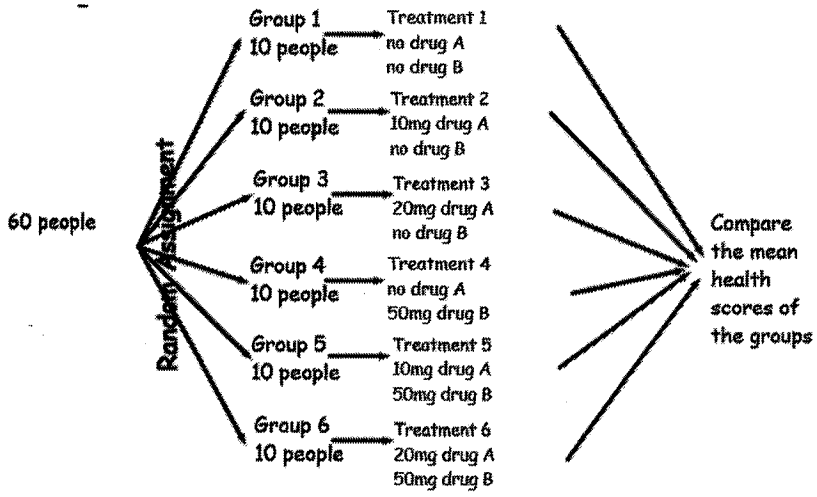
2 Factors:	Levels
Drug A:	*Not given, 10mg, 20mg* (3 levels)
Drug B:	*Not given, 50mg* (2 levels)

We need to impose every combination of these factors as separate treatments with a separate group for each:

	Drug A: not given	Drug A: 10mg	Drug A: 20mg
Drug B: not given	group 1	group 2	group 3
Drug B: 50mg	group 4	group 5	group 6

Because we need to combine each level of each factor with each level of every other factor, the number of treatments (and groups) is:

$$\#\,treatments = (\#levels\ \ factor\ 1)(\#levels\ \ factor\ 2)(\#levels\ \ factor\ 3)...$$



This is called a **completely randomized two-factor experiment.**

Blinding and Placebos

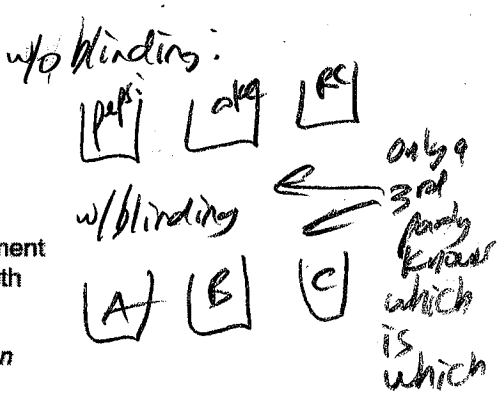
Humans are notoriously susceptible to errors in judgment. When we know what treatment was assigned, it's difficult not to let that knowledge influence our assessment of the response (even subconciously). Subjects (if human) and researchers are both subject to the potential of bias in this way.

Example: Subjects are asked to taste test various colas, but the brands aren't hidden so their brand-loyalty biases the results.
Example: A researcher prefers a particular colas and subconsciously uses different body language when presenting subjects with these colas for evaluation.

To eliminate these issues subject and/or researchers can be prevented from knowing which experimental units are assigned to which groups. This is called **blinding.**

Single-blind: When one class (subject or researcher) are blinded.
Example: Researcher knows which cola is which, but brand is hidden from subject.

Double-blind: When everyone in both classes (subject and researcher) are blinded.
Example: A 3rd party prepares the cola samples so both the researcher and subject do not know the brands. Codes are used and only revealed after the results of the experiment are final.



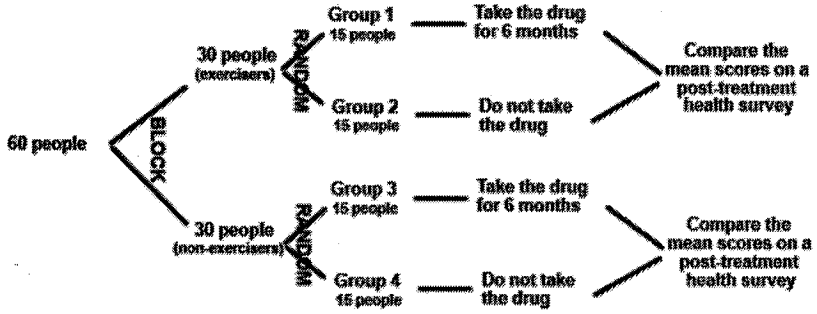
Experiments often include a group which receives no treatment. This may be detectable by the subject if the subject is human (for example, if the study is about a new medication and the subject is given nothing, then they know they are in the 'control' group).

Humans may have reasons (conscious, or subconscious) to want an experiment to have a particular result and if they can detect that they are in the control group, that can bias the results. So the subject can be given a 'fake treatment' which replicates the experience of receiving the treatment without actually doing anything. This fake treatment is called a **placebo**.

For example, in our park/library drug experiment, the 'no drug' group could have been given a pill with inert ingredients as a placebo.

In fact, it is not unusual for subjects treated with a placebo to show a result. Frequently 20% or more of subjects in medical studies report effects in variables which are subjective in nature, such as reduction in pain, improved range of motion, greater alertness, etc. when given placebos. This is called the **placebo effect**.

What other improvements can we make to the drug/health score experiment, besides blocking on exercise?

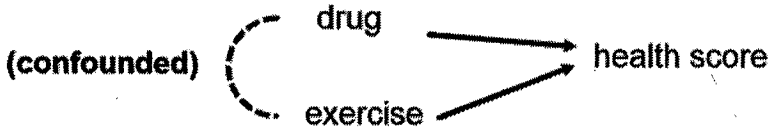


- **Blinding/placebo:** We could give the 'no drug' subjects a placebo making this a single-blind experiment.
- **Bias in sample:** The sample is a convenience sample, and offering incentive may induce one part of the population to participate more than others.
- **Pre-, Post-treatment tests:** Should we give the health survey before and after treatment and look at improvement? This would reduce subject to subject variation.

Terminology: Confounding vs. Lurking Variables

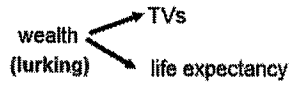
In the experiment with the park and library drug trials, we originally thought there was one factor: drug. But there was another, hidden, factor: exercise. 'Exercise' isn't a lurking variable, though, because while it is controlling health score, it isn't really controlling the drug variable....

Instead, the situation is more like this...both drug and exercise affect health score...



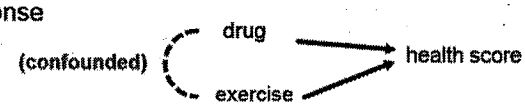
In this situation, we say that the variables 'drug' and 'exercise' are **confounded**, because we can't separate the two effects on the response variable.

Lurking variable: When one variable causes two other variables to change together, making them appear associated.



(Used by our textbook, not an official term - should avoid using this on the AP Exam)

Confounded variables: When the effect of multiple explanatory variables on a response variable can't be separated.
(Official term, used on the AP Exam)



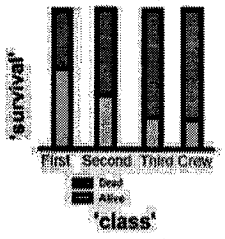
If you are unsure about the situation, you can just say 'other' variables to avoid misusing the terminology.

Terminology: Other issues and cautions

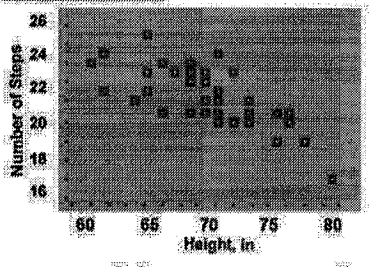
Association/Relationship: General term meaning there appears to be some relationship between variables. (Official term, used on AP Exam)
Appropriate for observational studies and experiments.

Can be used with any combination of categorical and numerical variables:

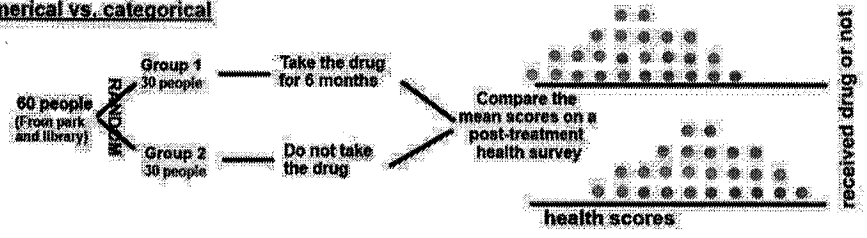
categorical vs. categorical



numerical vs. numerical

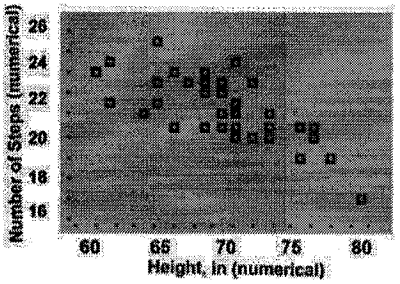


numerical vs. categorical



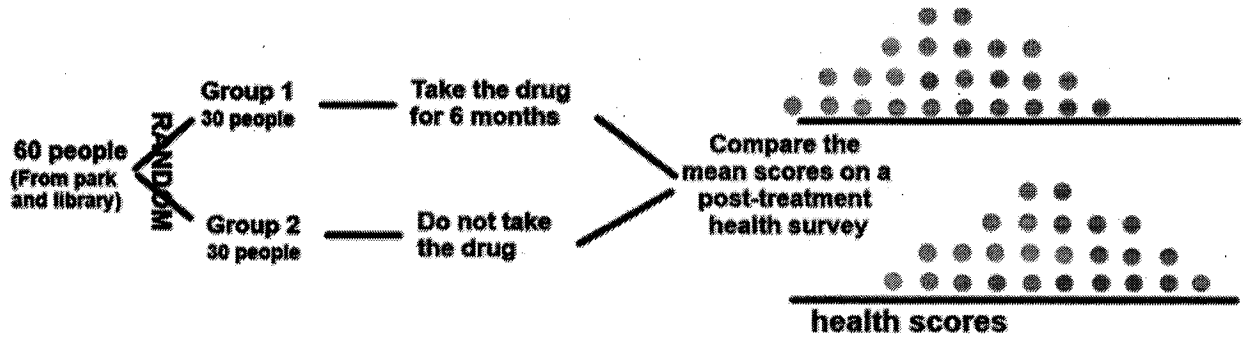
Correlation: Precise term describing the strength and direction of a linear relationship (usually taken to mean the correlation coefficient, *r*)
(Official term, used on AP Exam)
Appropriate for observational studies and experiments.

Can only be used with two numerical variables where it is possible to calculate a correlation coefficient:



r = - 0.74

Causation (Cause-and-Effect): Changing the explanatory variable is not only 'associated' with changes in the response variable, it ***causes*** changes in the response variable. (Official term, used on AP Exam)
Appropriate only for experiments.



If the experiment is well-designed, it is appropriate to make conclusions like 'the drug causes health scores to increase'.

Experimental Units: The individuals that receive the treatment and produce a data value in the response variable data set. (Official term, used on AP Exam)
Often, questions will ask you to identify the experimental units, which can be tricky. Here are two examples from AP exams...

A biologist is interested in studying the effect of growth-enhancing nutrients and different salinity (salt) levels in water on the growth of shrimps. The biologist has ordered a large shipment of young tiger shrimps from a supply house for use in the study. The experiment is to be conducted in a laboratory where 10 tiger shrimps are placed randomly into each of 12 similar tanks in a controlled environment. The biologist is planning to use 3 different growth-enhancing nutrients (A, B, and C) and two different salinity levels (low and high).

(a) The length of each shrimp will be measured before and after exposure to the nutrient and salinity environment, growth for each shrimp is computed from subtracting the lengths: growth = $\text{length}_{\text{after}} - \text{length}_{\text{before}}$.

Identify the treatments, experimental units and response variable of the experiment.

Treatments: The tank environment combinations of nutrients and salinity levels are the treatments.

Experimental Units: The experimental units are **the individual shrimp.**

Response variable: The response variable is the growth of each individual shrimp (change in length).

(b) The mean length of shrimp in each tank will be measured before and after exposure to the nutrient and salinity environment, growth for each tank is computed from subtracting the means:
growth = $\text{mean length}_{\text{after}} - \text{mean length}_{\text{before}}$.

Identify the treatments, experimental units and response variable of the experiment.

Treatments: The tank environment combinations of nutrients and salinity levels are the treatments.

Experimental Units: The experimental units are **the tanks, each containing 10 shrimp.**

Response variable: The response variable is the increase in mean length in each tank (change in mean length).