

**Deep Thoughts**  
Unit 3: Collecting Data

1. Alzheimer's disease results in a loss of cognitive ability beyond what is expected with typical aging. A local newspaper published an article with the following headline;

**Study Finds Strong Association Between Smoking and Alzheimer's**

The article reported that a study tracked medical histories of 21,123 men and women for 23 years. The article stated that, for those who smoke at least two packs of cigarettes a day, the risk of developing Alzheimer's disease was 2.57 times the risk for those who did not smoke.

- (a) Identify the explanatory and response variable in the study.

**Solution:** The explanatory variable is the person's degree of cigarette smoking. The response variable is whether the person develops Alzheimer's disease during the course of the study.

- (b) Is the study described an observational study or an experiment.

**Solution:** This is an observational study because the people in the study were not assigned to a certain degree of cigarette smoking. Rather, the degree of cigarette smoking for each person was passively observed and recorded, not manipulated by researchers.

- (c) Exercise status (regular weekly exercise versus no regular weekly exercise) was mentioned in the article as a possible confounding variable. Explain how exercise status could be a confounding variable in the study.

**Solution:** A confounding variable is one that is related to the explanatory variable and possibly influences the response variable. In this case it seems plausible that people who exercise more regularly might be more health conscious, therefore, less likely to smoke cigarettes than people who do not exercise regularly. Similarly, it's possible that people who exercise more regularly are less likely to develop Alzheimer's disease than people who do not exercise regularly. If both of these relationships turn out to be true, then smoking cigarettes would be associated with developing Alzheimer's disease due to the association of both variables with exercise, even if there were no cause-and-effect relationship between smoking cigarettes and developing Alzheimer's disease.

Need to:

1. Provides a reasonable explanation that exercise status is related to smoking status.
2. States that exercise status might influence where the person develops Alzheimer's disease.

2. An apartment building has nine floors and each floor has four apartments. The building owner wants to install new carpeting in eight apartments to see how well it wears before she decides whether to replace the carpet in the entire building.

The figure below shows the floors of the apartments in the building with their apartment numbers. Only nine apartments indicated with an asterisk have children in the apartment.

11*	12	21	22*	31	32	
	1st Floor		2nd Floor		3rd Floor	
14		24	23*	34	33	
41	42	51*	52	61	62	* = Children in the apartment
	4th Floor		5th Floor		6th Floor	
44	43	54	53	64	63	
71	72	81	82	91	92*	
	7th Floor		8th Floor		9th Floor	
74*	73*	84*	83	94	93*	

- (a) For convenience, the apartment building owner wants to use a cluster sampling method, in which the floors are clusters, to select the eight apartments. Describe the process for randomly selecting eight different apartments using this method.

**Solution:** The following two-step procedure can be used to select eight apartments:

1. Generate a random integer between 1 and 9, inclusive, using a calculator, a computer program, or a table of random digits. Select all four apartments on the floor corresponding to the selected integer.
2. Generate another integer between 1 and 9, inclusive. Repeat this process until a different integer from step 1 has been selected. Again, select all four apartments on the floor corresponding to the second selected integer.

The cluster sample consists of the eight apartments on the two randomly selected floors. MUST INCLUDE:

1. Using 10 random digits rather than nine.
2. Failing to explicitly deal with the issue of potentially repeated random numbers.

- (b) An alternative sampling method would be to select a stratified random sample of eight apartments, where the strata are apartments with children and apartments with no children. A stratified random sample of size eight might include two randomly selected apartments with children and six randomly selected apartments with no children. In the context of this situation, give one statistical advantage of selecting such a stratified sample as opposed to a cluster sample of eight apartments using the floors as clusters.

**Solution:** Because the amount of wear on the carpets in apartments with children could be different from the wear on the carpets in apartments without children, it would be advantageous to have apartments with children represented in the sample. The cluster

sampling procedure in part (a) could produce a sample with no children in the selected apartments; for example, a cluster sample of the apartments on the third and sixth floors would consist entirely of apartments with no children. Stratified random sampling, where the two strata are apartments with children and apartment without children, gauruntees a sample that includes apartment with and without children, which, in turn, would yield sample data that are representative of both types of apartments.

MUST INCLUDE:

1. The amount of carpet wear is different for apartments with and without children.
2. The stratified random sample ensures that some apartments with children will be selected.

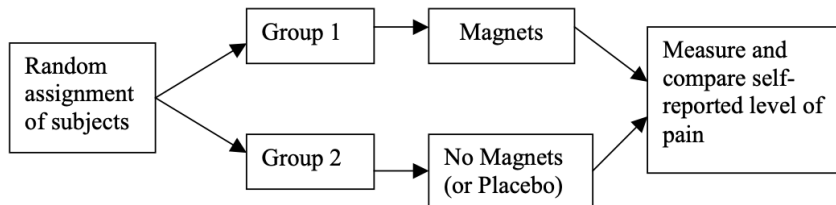
3. A preliminary study conducted at a medical centre in St. Louis has shown that treatment with small, low-intensity magnets reduces the self reported level of pain in polio patients. During each session, a patient rested on an examining table in the doctor's office while the magnets, embedded in soft pads, were strapped to the body at the site of pain. Sessions continued for several weeks, after which pain reduction was measured.

A new study is being designed to investigate whether magnets also reduce pain in patients suffering from herniated disks in the lower back. One hundred male patients are available for the new study.

- (a) Describe an appropriate design for the new study. Your discussion should address treatments used, methods of treatment assignment, and what variables would be measured. Do not describe how the data would be analyzed.

**Solution:** Must Include:

1. Two Treatments: magnets and no magnets (or magnets and placebo). Subjects in the no magnet group would be handled in the same way as the magnet group, but there would be no magnets embedded in the pads used.
2. There must be random assignment of subjects to treatments. How the randomization would be carried out does not need to be specified, but it must be clear what is being randomized.
3. variable measured: Self-reported level of pain of reduction in pain.



- (b) Would you modify the design above if, instead of 100 male patients, there were 50 male and 50 female patients available for the study? If so, how would you modify your design? If not, why not?

**Part (b):** Either one of the following approaches is acceptable.

1. Saying yes and indicating how they would alter the design: Separating the subjects into the two gender groups and then randomizing subjects to treatments within each group. This may also be described using a diagram, as shown below, but the blocking factor and randomization must be clearly indicated.

OR

2. Saying no and describing why. For example, indicating that the randomization in (a) should equalize the effects of gender in the two groups or assuming gender does not have a strong effect and since the sample size is large

OR

providing a good explanation for *why* gender does not have a strong differential effect on the outcome.

**Solution:** Part (b)

4. An administrator at a large university wants to conduct a survey to estimate the proportion of students who are satisfied with the appearance of the university buildings and grounds. The administrator is considering three methods of obtaining a sample of 500 students from the 70,000 students at the university.

- (a) Because of financial constraints, the first method the administrator is considering consists of taking a convenience sample to keep the expenses low. A very large number of students will attend the first football game of the season, and the first 500 students who enter the football stadium could be used as a sample. Why might such a sampling method be biased in producing an estimate of the proportion of students who are satisfied with the appearance of the buildings and grounds?

**Solution:** The first 500 students who enter the football stadium were not likely to be representative of the population of all students at the university. In other words, these 500 students were likely to differ systematically from the population with regard to many variables. For example, these 500 students might have more school pride than the population of students as a whole, which might be related to their opinions about the appearance of university buildings and grounds. Perhaps their school pride is related to having more positive opinions about the appearance of university buildings and grounds, in which case the sample proportion of students who were satisfied would be biased toward overestimating the population proportion of students who were satisfied

- (b) Because of the large number of students at the university, the second method the administrator is considering consists of using a computer with a random number generator to select a simple random sample of 500 students from a list of 70,000 student names. Describe how to implement such a method.

**Solution:** Obtain a list of all 70,000 students at the university. Assign an identification number from 1 to 70,000 to each student.  
Then use a computer to generate 500 random integers between 1 and 70,000 without replacement. The students whose ID numbers correspond to those numbers were then selected for the sample.

- (c) Because stratification can often provide a more precise estimate than a simple random sample, the third method the administrator is considering consists of selecting a stratified random sample of 500 students. The university has two campuses with male and female students at each campus. Under what circumstance(s) would stratification by campus provide a more precise estimate of the proportion of students who are satisfied with the appearance of the university buildings and grounds than stratification by gender?

**Solution:** Stratifying by campus would be more advantageous than stratifying by gender provided that opinions about appearance of university buildings and grounds between the two campuses differ more than the opinions about appearance of university buildings and grounds between the two genders.

## Deep Thoughts

### Collecting Data

1. Before beginning a unit on frog anatomy, a seventh-grade biology teacher gives each of the 24 students in the class a pretest to assess their knowledge of frog anatomy. The teacher wants to compare the effectiveness of an instructional program in which students physically dissect frogs with the effectiveness of a different program in which students use computer software that only simulates the dissection of a frog. After completing one of the two programs, students will be given a posttest to assess their knowledge of frog anatomy. The teacher will then analyze the changes in the test scores (score on posttest minus score on pretest).
  - (a) Describe a method for assigning the 24 students to two groups of equal size that allows for a statistically valid comparison of the two instructional programs.

**Solution: Completely Randomized Design:** Each student will be assigned a unique random number using a random number generator on a calculator, statistical software, or a random number table. The assigned numbers will be listed in ascending order. The students with the lowest 12 numbers in the ordered list will receive the instructional program that requires physically dissecting frogs. The students with the highest 12 numbers will receive the instructional program that uses computer software to simulate the dissection of a frog.

**OR**

**Randomized Block Design:** Students will be paired or placed into blocks of size two, based on having similar pretest scores. So, the first block will contain the two students with the two lowest pretest scores, the second block will contain the two students with the third- and fourth-lowest pretest scores, and so on, with the last block containing the two students with the two highest pretest scores. In each block, the students will be assigned a unique random number using a random number generator on a calculator, statistical software, or a random number table. The student in each block with the lower random number will receive the instructional program that requires physically dissecting frogs, and the student with the higher random number will receive the instructional program that uses computer software to simulate the dissection of a frog.

Essentially correct (E) if a proper method of randomization is described that (1) creates two groups of equal size; AND (2) assigns the named treatments to the groups in a manner that knowledgeable statistics users would employ to assign the students to the two instructional groups.

- (b) Suppose the teacher decided to allow the students in the class to select which instructional program on frog anatomy (physical dissection or computer simulation) they prefer to take, and 11 students choose actual dissection and 13 students choose computer simulation. How might that self-selection process jeopardize a statistically valid comparison of the changes in the test scores (score on posttest minus score on pretest) for the two instructional programs? Provide a specific example to support your answer.

**Solution:**

By not randomizing and allowing the students to self-select, there is a potential for changes to occur in the differences between pretest and posttest scores for a particular group because of the characteristics of students who choose a particular instructional method, not because of the instructional method itself. For example, suppose frog-loving students already know a lot about frog anatomy; one would therefore expect these students to be less likely to show a large change between the pretest and posttest scores. Suppose the frog-loving students tend to select the computer simulation method (perhaps because they do not like the notion of dissecting the frogs they love). The possible low change between pretest and posttest scores for the computer simulation group might then be attributed to the students already knowing a lot about frog anatomy beforehand, not to the instructional method itself. The frog dissection group might see a larger change in scores because the students entering this group are those with the lower pretest scores (less prior knowledge) and who are thus more likely to show greater improvement between pretest and posttest scores.

Essentially correct (E) if (1) the example gives a reasonable characteristic of the self-selected students in the study; AND (2) explains how this characteristic could be associated with changes in the differences between the pretest and posttest scores.

2. Agricultural experts are trying to develop a bird deterrent to reduce costly damage to crops in the United States. An experiment is to be conducted using garlic oil to study its effectiveness as a nontoxic, environmentally safe bird repellent. The experiment will use European starlings, a bird species that causes considerable damage annually to the corn crop in the United States. Food granules made from corn are to be infused with garlic oil in each of five concentrations of garlic: 0 percent, 2 percent, 10 percent, 25 percent, and 50 percent. The researchers will determine the adverse reaction of the birds to the repellent by measuring the number of food granules consumed during a two-hour period following overnight food deprivation. There are forty birds available for the experiment, and the researchers will use eight birds for each concentration of garlic. Each bird will be kept in a separate cage and provided with the same number of food granules.

- (a) For the experiment, identify: the treatments, the experimental units, and the response that will be measured.

**Solution:** The treatments are the different concentrations of garlic in the food granules. Specifically, there are five treatments: 0 percent, 2 percent, 10 percent, 25 percent and 50 percent.

The experimental units are the birds (starlings), each placed in an individual cage.

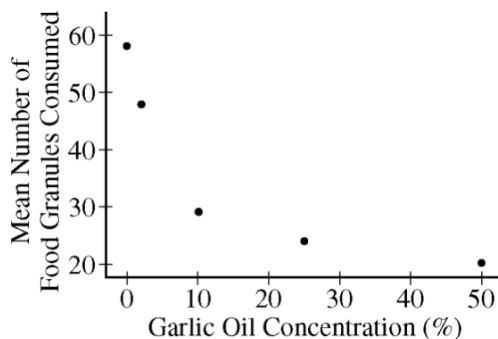
The response is the number of food granules consumed by the bird.

Essentially correct (E) if the student correctly identifies all three subparts: the treatments, the experimental units and the response that will be measured.

- (b) After performing the experiment, the researchers recorded the data shown in the table below: Construct a graph of the data that could be used to investigate the appropriateness

Garlic oil concentration	0%	2%	10%	25%	50%
Mean number of food granules consumed	58	48	29	24	20
Number of birds	8	8	8	8	8

of a linear regression model for analyzing the results of the experiment. Based on your graph, do you think a linear regression model is appropriate? Explain.



**Solution:**



The curved pattern in this scatterplot reveals that a linear regression model would not be appropriate for modelling the relationship between these variables.

Essentially correct (E) if the student produces a correct graph (a reasonable scatterplot or residual plot with correct labels and scales) and then concludes, based on one or more features of the graph, that the pattern of the relationship does not appear to be linear.