

Module 1 Solutions

Xihan Qian

Apr.25, 2022

1. Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (setosa, versicolor and virginica). There were 40 flowers from each species in the data set. [*Openintro-statistics*]

(a) How many cases are included in the data?

Solution: There are three species with 40 in each; hence $40 \times 3 = 120$ cases were included in the data.

(b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

Solution: Four continuous numerical variables are included: sepal length, sepal width, petal length, and petal width.

(c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Solution: One categorical variable, species, with three levels: setosa, versicolor, and virginica.

2. A small company took a survey on all of its 70 employees to determine what proportion participates in volunteer activities. In a statistics class, four students made the following statements: [*AP Stats 2012 Practice Exam*]

Student A: The company should not use this data because this is an observational study.

Student B: The company can use this result to prove that working for the company causes employees to participate in volunteer activities.

Student C: The company did not select a random sample of employees, so the survey will not provide the company with any useful information.

Student D: The company can determine the proportion because this survey was already a census of all employees.

Which of the students' response is true?

Solution: Student D.

3. In a survey, a random sample of 1,155 Americans was asked the question: "After an average work day, about how many hours do you spend doing other activities?" The average time spent relaxing was found to be 1.50 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

- (a) One American in the sample.
- (b) Number of hours spent relaxing after an average work day.
- (c) 1.50.
- (d) Average number of hours all Americans spend relaxing after an average work day. [*Openintro-statistics*]

Solution: (a) Observation.

- (b) Variable.
- (c) Sample statistic (mean).
- (d) Population parameter (mean).

4. Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has gained an entirely new reputation as a dietary supplement. In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.

- (a) What type of study is this?
- (b) What are the experimental and control treatments in this study?
- (c) Has blocking been used in this study? If so, what is the blocking variable?
- (d) Has blinding been used in this study?
- (e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large. [*Openintro-statistics 1.37*]

Solution: (a) Experiment.

- (b) Treatment: 25 grams of chia seeds twice a day, control: placebo.
- (c) Yes, gender.
- (d) Yes, single blind since the patients were blinded to the treatment they received.
- (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

5. A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). [*Openintro-statistics 1.29*]

- (a) What is the response variable?
- (b) What is the explanatory variable? What are its levels?
- (c) What is the blocking variable? What are its levels?

Solution: (a) Exam performance.

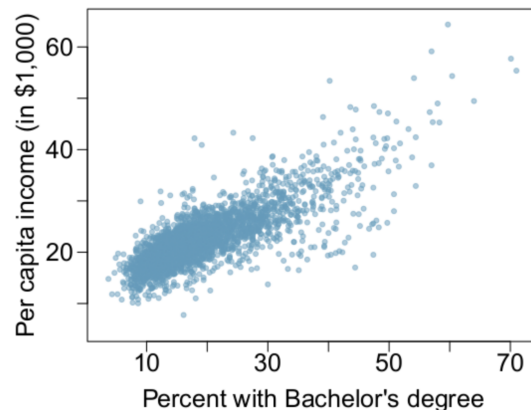
- (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).
- (c) Sex: male, female.

6. Briefly outline a design for a blind randomized control trial using your classmates as participants to determine preference for the taste of Coke or Pepsi. [variation of *Openintro-statistics* 1.33]

One possible Solution: Need randomization and blinding.

- (1) Prepare two cups for each participant, one containing Coke and the other containing Pepsi. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (Coke) and B (Pepsi). (Be sure to randomize A and B for each trial!)
- (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much they liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double-blind experiment.

7. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.



- (a) What are the explanatory and response variables?
- (b) Describe the relationship between the two variables. Make sure to identify unusual observations, if any.
- (c) Can we conclude that having a bachelor's degree increases one's income?

Solution: (a) The explanatory variable is Percent with a Bachelor's Degree, while the response variable is Per Capita Income (in \$1000s).

- (b) There is a strong positive relationship between these two variables. There doesn't seem to be any extremely unusual observations from the scatterplot.
- (c) You cannot conclude that having a bachelor's degree increases one's income because correlation does not imply causation.

8. An article titled Risks: Smokers Found More Prone to Dementia states the following:

"Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and

health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.
[*Openintro-statistics*]

Solution: No, this is an observational study.