MODULE 1 EXAMPLES

TAs: Nihal Reddy
Email: nireddy@ucsd.edu
OH: Thursdays 6-7pm

Slide Credits: Kira Fleischer

PROBLEM #: KEY TOPICS FROM PROBLEM

Problem setup and description.

Question

Key notes from readings/lectures needed to answer the question

Solution: written with as much detail as we expect you to give on your homework sets

PROBLEM 1: CASES AND VARIABLES

Shawn is interested in dogs and one day he decided to study a data set of three types of dogs: Husky, Bulldog, and Chihuahua that contains their weights, lengths of fur, and heights. There are 60 dogs from each type in the data set.

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

Cases: observational units

Solution: There are three species with 60 in each; hence 60 × 3 = 180 cases were included in the data.

PROBLEM 1: CASES AND VARIABLES

Shawn is interested in dogs and one day he decided to study a data set of three types of dogs: Husky, Bulldog, and Chihuahua that contains their weights, lengths of fur, and heights. There are 60 dogs from each type in the data set.

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

Numerical Variables: takes number values

Continuous Variables: can take any value in a range, and can be infinitely divided

Discrete Variables: values are distinct numbers/counts, and cannot be divided further

Solution: Three continuous numerical variables are included: Length of fur, weights, and heights of the dogs.

PROBLEM 1: CASES AND VARIABLES

Shawn is interested in dogs and one day he decided to study a data set of three types of dogs: Husky, Bulldog, and Chihuahua that contains their weights, lengths of fur, and heights. There are 60 dogs from each type in the data set.

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

Categorical Variables: values represent distinct categories or groups, rather than numerical values

Levels: the actual category

Solution: One categorical variable, types of dog, with three levels: Husky, Bulldog, and Chihuahua.

PROBLEM 2: CENSUS SURVEYS

In a Statistics class at UCSD, the professor took a survey on all 50 students in the class to determine what proportion participates in voluntary work. The professor then asked four students in the class to analyze the results. They made the following statements:

Student A: The professor can use this result to prove that studying in this class causes students to participate in voluntary work.

Student B: The professor can determine the proportion because this survey was already a census of all students.

Student C: The professor should not use this data because this is an observational study.

Student D: The professor did not select a random sample of students, so the survey will not provide the professor with any useful information.

Which of the students' response is true?

PROBLEM 2: CENSUS SURVEYS

In a Statistics class at UCSD, the professor took a survey on all 50 students in the class to determine what proportion participates in voluntary work. The professor then asked four students in the class to analyze the results. They made the following statements:

Student A: The professor can use this result to prove that studying in this class causes students to participate in voluntary work.

Correlation does not imply causation. Even if the data showed a high proportion of students participating in voluntary work, it doesn’t mean that studying in this class causes this participation. Other factors could influence both studying and voluntary work, and without a controlled experiment, causation cannot be established.

Thus, Student A is incorrect.

PROBLEM 2: CENSUS SURVEYS

In a Statistics class at UCSD, the professor took a survey on all 50 students in the class to determine what proportion participates in voluntary work. The professor then asked four students in the class to analyze the results. They made the following statements:

Student B: The professor can determine the proportion because this survey was already a census of all students.

Since the survey was conducted with all 50 students in the class, it represents a complete census. This means the professor has accurate data on the participation in voluntary work for the entire population of interest (the students in that class). Therefore, the proportion can be calculated directly from the data collected.

Thus, Student B is correct.

PROBLEM 2: CENSUS SURVEYS

In a Statistics class at UCSD, the professor took a survey on all 50 students in the class to determine what proportion participates in voluntary work. The professor then asked four students in the class to analyze the results. They made the following statements:

Student C: The professor should not use this data because this is an observational study.

While it’s true that the data collected is observational (the professor is observing responses without manipulating conditions), this does not invalidate the data for determining proportions. Observational studies can provide useful information about characteristics within a population.

Thus, Student C is incorrect.

PROBLEM 2: CENSUS SURVEYS

In a Statistics class at UCSD, the professor took a survey on all 50 students in the class to determine what proportion participates in voluntary work. The professor then asked four students in the class to analyze the results. They made the following statements:

Student D: The professor did not select a random sample of students, so the survey will not provide the professor with any useful information.

In this case, the survey is a census of all students in the class rather than a random sample. While it's true that randomness is important for generalizing findings beyond the sampled group, having data from all students means that the professor has complete information about this particular population (the class), which is still useful for understanding that group, even if it doesn't generalize to other classes or populations.

Thus, Student D is incorrect.

PROBLEM 3: OBSERVATIONS, VARIABLES, SAMPLE STATISTICS, AND POPULATION PARAMETERS

In order to better understand its employees' time schedules after work, a random sample of 1500 employees of a large company were asked the question: "After an average work day, about how many hours do you spend relaxing?" The average time spent relaxing was found to be 2.5 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 employee in the sample.

(b) Number of hours spent relaxing after an average work day.

(c) 2.5.

(d) Average number of hours all employees spend relaxing after an average work day.

Observation: single data point

Variable: attribute being measured or categorized

Sample Statistic: value calculated based on the observed sample

Population Parameter: value that describes a characteristic for the entire population

Solution: (a) Observation.

(b) Variable.

(c) Sample statistic (mean).

(d) Population parameter (mean).

PROBLEM 4: EXPERIMENTS, BLOCKING, AND BLINDING

A lot of people have trouble sleeping. In a study, a group of scientists want to see if a newly developed drug is efficient for helping these people sleep. They recruited 40 people older than 50 years old with trouble sleeping and divided them randomly into two groups: control or treatment. They also recruited 40 people younger than 50 years old with trouble sleeping, and they randomly placed half of these participants into the control group and the other half into the treatment group. One group was given one sleeping pill once a week, and the other was given a placebo. The subjects all volunteered to be a part of the study. After 10 weeks, the scientists found no significant difference between the groups in terms of quality of sleep.

(a) What type of study is this?

Solution: Experiment: there is a treatment and control group that are being studied.

PROBLEM 4: EXPERIMENTS, TREATMENTS, CONTROLS, BLOCKING, AND BLINDING

A lot of people have trouble sleeping. In a study, a group of scientists want to see if a newly developed drug is efficient for helping these people sleep. They recruited 40 people older than 50 years old with trouble sleeping and divided them randomly into two groups: control or treatment. They also recruited 40 people younger than 50 years old with trouble sleeping, and they randomly placed half of these participants into the control group and the other half into the treatment group. One group was given one sleeping pill once a week, and the other was given a placebo. The subjects all volunteered to be a part of the study. After 10 weeks, the scientists found no significant difference between the groups in terms of quality of sleep.

(b) What are the experimental and control treatments in this study?

Treatment: subjects who receive the drug

Control: subjects who receive no treatment, in order to serve as a baseline

Solution: Treatment: one pill once a week, control: placebo.

PROBLEM 4: EXPERIMENTS, TREATMENTS, CONTROLS, BLOCKING, AND BLINDING

A lot of people have trouble sleeping. In a study, a group of scientists want to see if a newly developed drug is efficient for helping these people sleep. They recruited 40 people older than 50 years old with trouble sleeping and divided them randomly into two groups: control or treatment. They also recruited 40 people younger than 50 years old with trouble sleeping, and they randomly placed half of these participants into the control group and the other half into the treatment group. One group was given one sleeping pill once a week, and the other was given a placebo. The subjects all volunteered to be a part of the study. After 10 weeks, the scientists found no significant difference between the groups in terms of quality of sleep.

(c) Has blocking been used in this study? If so, what is the blocking variable?

Blocking: method to reduce effects of potential confounding variables by splitting subjects into different blocks (groups) based on certain characteristics.

Solution: Yes, age.

PROBLEM 4: EXPERIMENTS, TREATMENTS, CONTROLS, BLOCKING, AND BLINDING

A lot of people have trouble sleeping. In a study, a group of scientists want to see if a newly developed drug is efficient for helping these people sleep. They recruited 40 people older than 50 years old with trouble sleeping and divided them randomly into two groups: control or treatment. They also recruited 40 people younger than 50 years old with trouble sleeping, and they randomly placed half of these participants into the control group and the other half into the treatment group. One group was given one sleeping pill once a week, and the other was given a placebo. The subjects all volunteered to be a part of the study. After 10 weeks, the scientists found no significant difference between the groups in terms of quality of sleep.

(d) Has blinding been used in this study?

Blinding: method to reduce bias where either just the participants do not know if they are in the treatment group or control group (single blind), or where both the participants and researchers do not know who is in the treatment and control group (double blind).

Solution: Yes, single blind since the patients were blinded to the treatment they received.

PROBLEM 4: EXPERIMENTS, TREATMENTS, CONTROLS, BLOCKING, AND BLINDING

A lot of people have trouble sleeping. In a study, a group of scientists want to see if a newly developed drug is efficient for helping these people sleep. They recruited 40 people older than 50 years old with trouble sleeping and divided them randomly into two groups: control or treatment. They also recruited 40 people younger than 50 years old with trouble sleeping, and they randomly placed half of these participants into the control group and the other half into the treatment group. One group was given one sleeping pill once a week, and the other was given a placebo. The subjects all volunteered to be a part of the study. After 10 weeks, the scientists found no significant difference between the groups in terms of quality of sleep.

(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.

Causal Statements: can only be made about experiments

Generalizations: can only be made if random samples are used

Solution: 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.

PROBLEM 5: RESPONSE AND EXPLANATORY VARIABLES, BLOCKING

A study is designed to test whether different types of locations would have different effects on people's appetite. The researcher believes that locations might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are by the ocean, indoors, and in amusement parks.

(a) What is the response variable?

Response Variable (aka dependent variable): main variable of interest that is suspected to be influenced by another variable (the explanatory variable)

Solution: people's appetite.

PROBLEM 5: RESPONSE AND EXPLANATORY VARIABLES, BLOCKING

A study is designed to test whether different types of locations would have different effects on people's appetite. The researcher believes that locations might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are by the ocean, indoors, and in amusement parks.

(b) What is the explanatory variable? What are its levels?

Explanatory Variable (aka independent variable): variable that is observed or manipulated to determine its effect on the response variable

Levels: the actual categories of the variable

Solution: Locations: by the ocean, indoors, and in amusement parks.

PROBLEM 5: RESPONSE AND EXPLANATORY VARIABLES, BLOCKING

A study is designed to test whether different types of locations would have different effects on people's appetite. The researcher believes that locations might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are by the ocean, indoors, and in amusement parks.

(c) What is the blocking variable? What are its levels?

Blocking Variable: groups with similar characteristics

Levels: the actual categories

Solution: Sex: male, female.

PROBLEM 6: BLIND RANDOMIZED CONTROL TRIAL

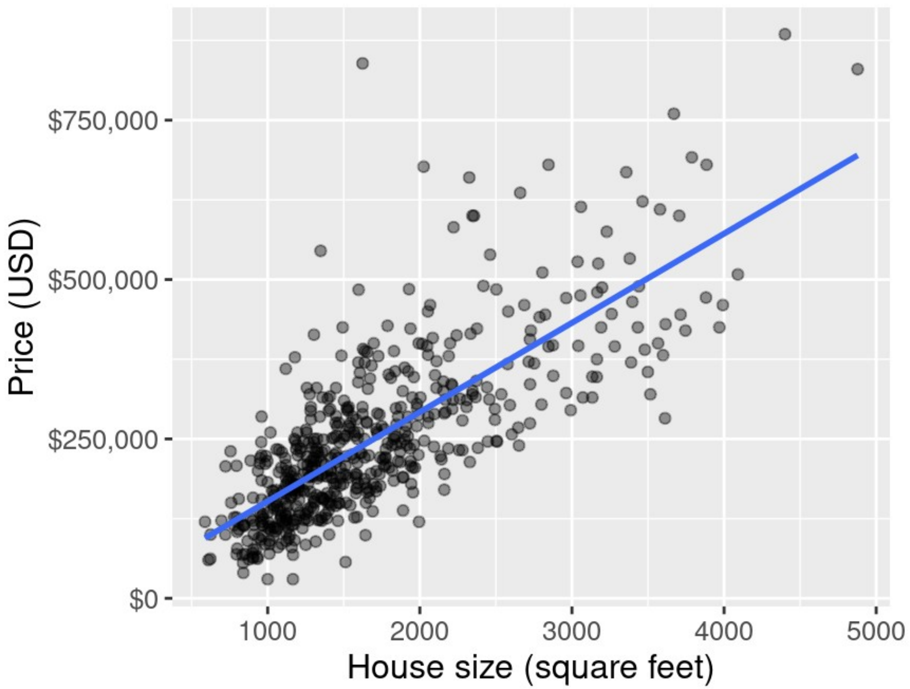
Briefly outline a design for a blind randomized control trial using your classmates as participants to determine preference for the taste of Mountain Dew or Diet Mountain Dew.

One possible Solution: (1) Prepare two cups for each participant, one containing Mountain Dew and the other containing Diet Mountain Dew. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (Mountain Dew) and B (Diet Mountain Dew). (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.

PROBLEM 7: SCATTERPLOT

The scatterplot below shows the relationship between the sale price versus house size for the full Sacramento housing data.



(a) What are the explanatory and response variables?

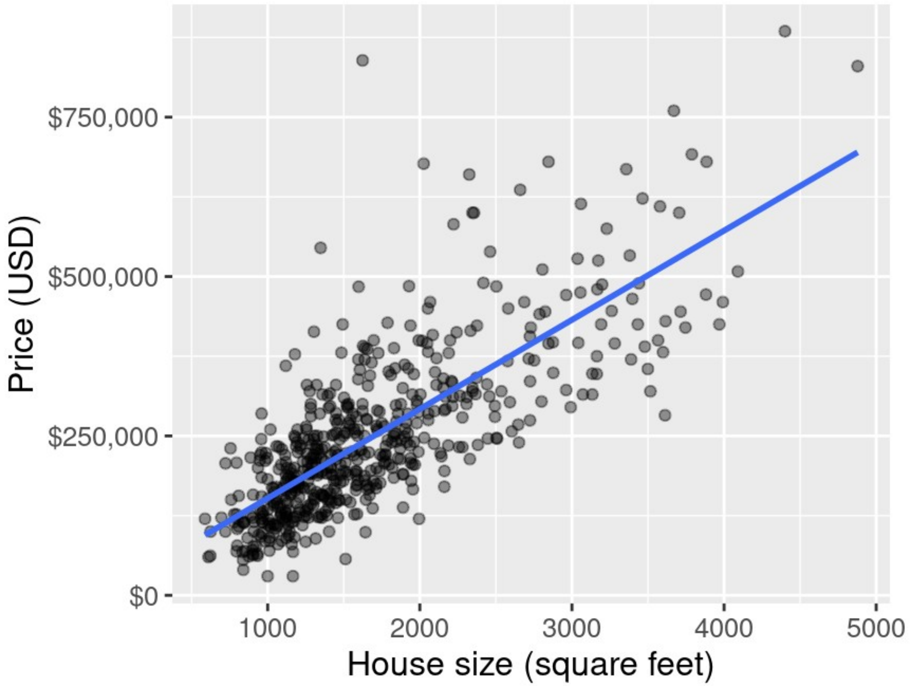
Explanatory Variable: typically on the x-axis

Response Variable: typically on the y-axis

Solution: The explanatory variable is the house size (in square feet), while the response variable is the price (in USD).

PROBLEM 7: SCATTERPLOT

The scatterplot below shows the relationship between the sale price versus house size for the full Sacramento housing data

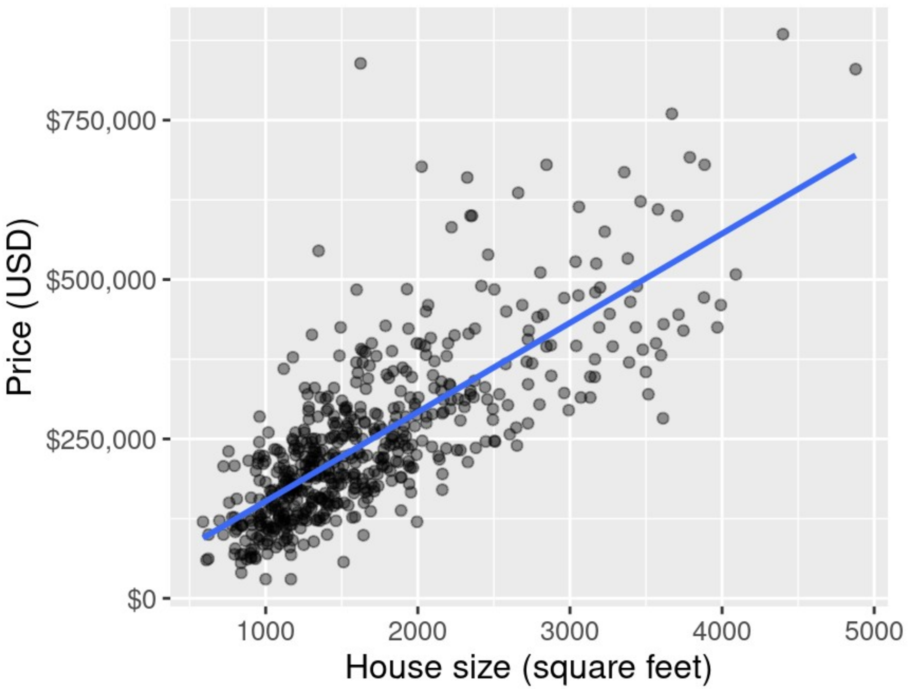


(b) Describe the relationship between the two variables. Make sure to identify unusual observations, if any.

Solution: There is a strong positive relationship between these two variables. There doesn't seem to be any extremely unusual observations from the scatterplot except for an influential point with coordinates ≈(1600, 850000).

PROBLEM 7: SCATTERPLOT

The scatterplot below shows the relationship between the sale price versus house size for the full Sacramento housing data



(c) Can we conclude that if the house size is bigger, then they also tend to have a higher price?

Correlation ≠ Causation

Solution: You cannot conclude this because correlation does not imply causation.

PROBLEM 8: DRAWING CONCLUSIONS FROM STUDIES

Suppose researchers want to investigate whether having physical activities would increase one's appetite. In an article they published, it states the following:

"Researchers analyzed data from 1,000 people who participated voluntarily. Among the participants, some always have physical activities before taking meals, while others don't. Physical activity was measured using a multi-sensor actigraph along with a self rated scale. Appetite was measured using the Council on Nutrition Appetite Questionnaire. The researchers followed up with these participants for 18 months and found that about 20% of these people had improvements in appetites. Among these, some had significant improvements while others didn't. After adjusting for other factors, the researchers concluded that people who take some form of physical activities were 35% more likely to score better than those who didn't."

Based on this study, can we conclude that doing physical activities would improve one's appetite? Explain your reasoning.

Conclusions can only be drawn from experiments. Experiments must have treatment and control groups.

Solution: No, this is an observational study.