Overview of Data Collection Principles Applications Solutions

Lt Col Ken Horton Professor Bradley Warner

20 August, 2020

Exercises

1. **Generalizability and causality**. Identify the population of interest and the sample in the studies described below, these are the same studies from the prevous lesson. Also comment on whether or not the results of the study can be generalized to the population and if the findings of the study can be used to establish causal relationships.

a. Researchers collected data to examine the relationship between pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[1]

The population of interest is all births. The sample consists of the 143,196 births between 1989 and 1993 in Southern California. If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

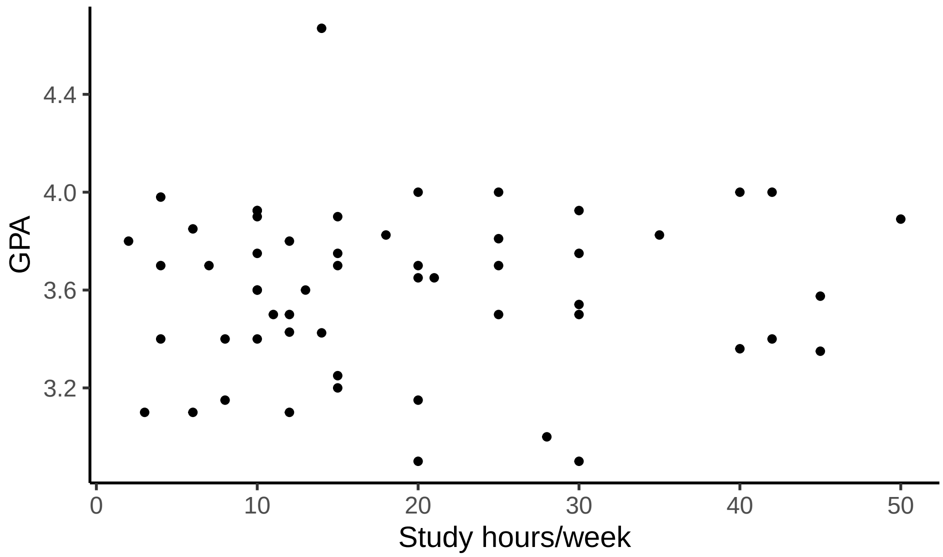
b. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[2]

The population is all 18-69 year olds diagnosed and currently treated for asthma. The sample is the 600 adult patients aged 18-69 years diagnosed and currently treated for asthma. Since the sample is not random (voluntary) the results cannot be generalized to the population at large. However, since the study is an experiment, the findings can be used to establish causal relationships.

[1] B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: Epidemiology 11.5 (2000), pp. 502–511.

[2] J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: Thorax 58 (2003).

2. **GPA and study time**. A survey was conducted on 55 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about their GPA and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.

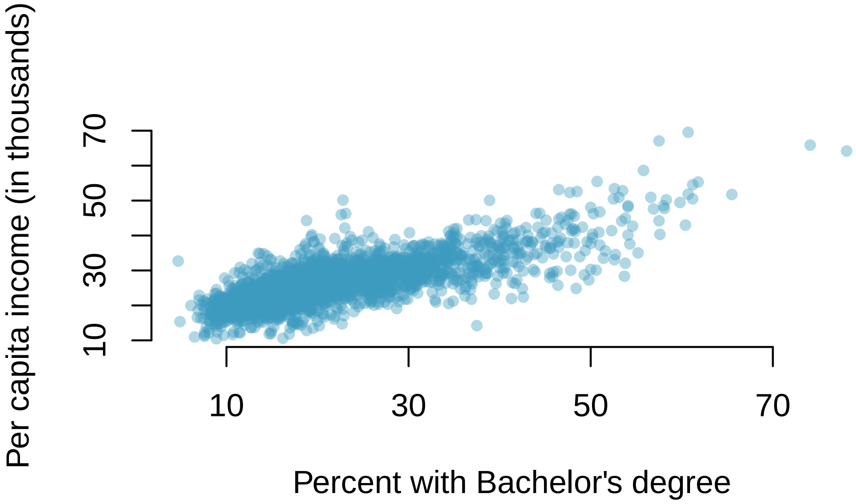


a. What is the explanatory variable and what is the response variable?
b. Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
c. Is this an experiment or an observational study?
d. Can we conclude that studying longer hours leads to higher GPAs?

Solutions

a. The explanatory variable is the number of study hours per week, and the response variable is GPA.
b. There is a somewhat weak positive relationship between the two variables, though the data become more sparse as the number of study hours increases. One responded reported a GPA above 4.0, which is clearly a data error. Also, there are a few respondents who reported unusually high study hours (60 and 70 hours/week). It should also be noted that the variability in GPA is much higher for students who study less than those who study more, also might be due to the fact that there aren't many respondents who reported studying higher hours.
c. This is an observational study.
d. Since this is an observational study, we cannot conclude that there is a causal relationship between the two variables even though there appears to be an association.

3. **Income and education** 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 discuss unusual observations, if any.
c. Can we conclude that having a bachelor's degree increases one's income?

Solutions

a. The explanatory variable is percent of population with a bachelor's degree and the response variable is per capita income (in thousands).
b. There is a strong positive linear relationship between the two variables. As the percentage of population with a bachelor's degree increases the per capita income increases as well. There are very few counties where more than 60% of the population have a bachelor's degree and very few countries that have a more than $50,000 in per capita income.
c. This is an observational study so we cannot make a causal statement based on the results. However, we can say that having a higher percentage of population with bachelor's degree is associated with a higher per capita income.

File Creation Information

- File creation date: 2020-08-20
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.7.0
- `tidyverse` package version: 1.3.0