# Homework 1 Answer Key

**DATA1220-55, Fall 2024**

Sarah E. Grabinski

2024-10-14

## Objectives

The material covered in Homework 1 was drawn from Chapter 1, *Introduction to Data*, of the textbook OpenIntro Statistics available on Canvas. This chapter discusses the basics of how to best collect, analyze, and draw conclusions from data. In addition, Homework 1 introduced the use of RStudio to create publication-ready data analyses. The objectives of Homework 1 were as follows.

- Help Sarah get to know you and better understand the specific needs of the class
- Register for Campuswire, where you will earn many of your participation points
- Identify different populations and sampling strategies
- Describe the reliability, validity, and generalizability of different types of data
- Perform a basic data analysis
  - Identify variable types
  - Visualize and interpret data
  - Communicate results
- Become familiar with creating markdown documents with R code in RStudio

## Problem 1 - Survey

### *Survey*

***A Google Forms survey was sent to your JCU email. I estimate it will take 5-10 minutes to complete.***

***Points: 5***

- All students completed the survey, so all students received 5 points.

### *Campuswire*

***Instructions for registering to our class Campuswire forum were sent to your JCU email. I have also posted our first discussion topic. Interacting with it will earn you participation credit. Make sure to check your notification settings so you don't miss anything you want to interact with!***

***Points: 5***

- Students who signed up for Campuswire *and* interacted with the first discussion post received 5 points.
- Students who signed up for Campuswire but did *not* interact with the first discussion post received 2.5 points.

## Problem 2 - Interpreting Studies

### *The Studies*

***Researchers in the UK wanted to answer the question of how much crime there was in Britain and whether it was going up or down. They used 2 different approaches to gather data for their investigation, but they need help determining the validity of their approach.***

### *Data Set 1*

***The Crime Survey for England and Wales is a survey in which approximately 38,000 people are questioned about their experiences with crime. People surveyed are 16 years of age or older and were not living in communal residences. Answers are self-reported.***

### *Data Set 2*

***UK Police keep administrative records of crimes they have investigated. Police use internal definitions of crimes and their discretion when creating these records.***

### *Questions*

***1. In 1 sentence each, describe the study population of the data sets. (Points: 2)***

- Students received 2 points if…
  - Their answer for data set 1 described the study population as all people 16 years of age or older not living in communal residences in England and Wales.
  - Their answer for data set 2 described the study population crimes committed in the UK and/or victims of that crime.
- Students lost 0.25 points if…
  - They described the study population for data set 1 as 38,000 individuals rather than all individuals and/or
  - They did not indicate the age or residence restrictions for data set 1
  - They did not indicate mention crimes or victims of crimes for data set 2
  - They did not mention the geographical source (e.g. England & Wales, UK) for the data sets. *Geographical location is important to include when describing populations.*
- Students received at least 1 point if they made a good-faith attempt to answer the question.
- Students who did not attempt to answer the question received 0 points.

***2. In 1 sentence each, describe the sampling strategy of the data sets. (Points: 2)***

- Students received 2 points if…
  - Their answer for data set 1 indicated that these were voluntary responses and/or that no sampling strategy was indicated.
  - Their answer for data set 2 indicated that no sampling strategy was used, as they have access to all the data that is available, also known as a census.
- Students lost 0.25 points if…

  - they described the sampling strategy for data set 1 as a random sample. *Never assume data was (randomly) sampled unless explicitly told so!*
  - they described any sampling strategy for data set 2 besides a census. If you're not using a *subset* of the total data available, then you're not using a *sampling strategy.*

- I also accepted convenience sampling for data set 2, however it is not technically correct. Although it is correct that using records like this is convenient, when you use *all* the available data, you aren't actually doing any sampling.
- Students received at least 1 point if they made a good-faith attempt to answer the question.
- Students who did not attempt to answer the question received 0 points.

***3. In 1 sentence each, describe the sampled population of the data set. (Points: 2)***

- Students received 2 points if…
  - Their answer for data set 1 described the sampled population as 38,000 people 16 years of age or older not living in communal residences in England and Wales.
  - Their answer for data set 2 indicated that the sampled population was victims of crimes reported to police in the UK where police made a record of that report (i.e. people who reported crimes the police thought were worth investigating) or the records of the crimes themselves.
- Students lost 0.25 points if…
  - they failed to mention the geographical source (e.g. England, UK, Wales) of either data set.
  - they failed to mention the sample size for data set 1
- Students received at least 1 point if they made a good-faith attempt to answer the question.
- Students who did not attempt to answer the question received 0 points.

***4. In 1 sentence, describe the target population of the study (Points: 1)***

- Students received 1 point if their answer indicated that the target population was crime in Britain (or UK, or England/Wales) and/or its potential victims, as described under ***The Studies*** introducing the motivation for the research.
- Students lost 0.25 points if…

  - they described the target populations for data sets 1 and 2 separately, as a single motivation for the research was described under ***The Studies***.
  - they described a *plausible* target population for this research but not the specific target population described under ***The Studies***.

- Students received at least 0.5 points if they made a good-faith attempt to answer the question.
- Students who did not attempt to answer the question received 0 points.

5. ***In a short paragraph (3-6 sentences), please describe… (Points: 3)***

   a) ***the reliability of each data set***

   b) ***the validity of each data set***

   c) ***if conclusions based on each data set from the study population are generalizable to the target population***

- Students received all 3 points if they…
  - Indicated data set 1 was self-report, which affects its reliability
  - Indicated data set 2 was also self-report (i.e. crimes themselves are self-reported, police define what counts as a crime, police choose which reported crimes to document and how), which affects its reliability
  - Indicated data set 1 is missing data from non-responders, which affects whether or not its a valid representation of the study population if there's a high non-response rate.
  - Indicated data set 2 is missing unreported crimes and/or reported crimes that police don't document, which could exclude a meaningful amount of crime and affects whether or not its a valid representation of the experiences of crime in the study population
  - Mentioned the large sample size of data set 1 or that we have all available data for set 2 when discussing their validity and/or generalizability to the target population
- Students lost 0.25 points if…
  - they failed to note that data set 1 and/or 2 contains self-report data, which affects their reliability.
  - they failed to discuss the impact of data which may be missing from data set 1 and/or 2, which affects their validity
  - they failed to discuss the size of the data sets when describing their generalizability

- Students received at least 1.5 points if they made a good-faith attempt to answer the question.
- Students who did not attempt to answer the question received 0 points.
- I did not deduct points for this, but when considering how generalizable a study is, it's important to consider who the data is NOT generalizable to.
  - The conclusions from data set 1 were not generalizable to people under the age of 16 or those in communal residences, because they were not part of the study population
    - The conclusions from data set 1 were not generalizable to people who are victims of crime in the UK who are NOT native residents, because they were not part of the study population
  - The conclusions from data set 2 were not generalizable to those likely to be victims of crimes that are underreported or not well documented/investigated by police, because they were not part of the study population

## Problem 3 - Interpreting Data

### *The Data*

***The Child Health and Development Studies investigate a range of topics. One study, in particular, considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area.***

### Questions

***1. Read in the .csv document (Points: 1)***

- Students received 1 point if the output in their HTML document indicates that they properly loaded the `.csv` file.

***2. Print a summary of the data. (Points: 1)***

- Students received 1 point if the output in their HTML document contains the numerical summary of the data created by the `describe()` function from the `Hmisc` package.

***3. Complete the data dictionary. (Points: 3)***

- Students received all 3 points if they…

  - Correctly identified `bwt`, `gestation`, `age`, `height`, and `weight` as numeric variables
  - Specified whether numeric variables were continuous or discrete
  - Correctly identified `parity` and `smoke` as categorical variables
  - Described `parity` and/or `smoke` as binary or nominal categorical variables
- Students lost 0.25 points if…
  - they didn't attempt to identify numerical variables as continuous or discrete.
  - they didn't attempt to identify categorical variables as binary, nominal, or ordinal.
  - they identified `parity` or `smoke` as numerical variables
- ***NOTE:*** Students did *not* lose points for indicating that `case` was a numeric variable, but its important to remember to treat variables that contain unique identifiers (e.g. ID #'s) as categorical variables and not as numbers.
- Students received at least 1.5 point if they made a good-faith attempt to complete the table.
- Students who did not attempt to complete the table received 0 points.

***4. Add the name of an explanatory (i.e. independent) variable to the x-axis of the plot and a response (i.e. dependent) variable to the y-axis of the plot. In 1 sentence, describe what you see. (Points: 3)***

- Students received all 3 points if they…
  - Changed both the x and y variables shown in the plot as requested in the instructions
  - Successfully embedded the plot in their HTML document
  - Appropriately described the relationship between the two variables as positive, negative, or independent
- Students lost 0.25 points if…
  - they did not change the variables as requested in the instructions
  - they accurately described the relationship in their plot but failed to use any of the words "independent", "positive", or "negative" to describe it.
- Students received at least 1.5 points if they did any one of the things listed.
- Students who did not attempt the question received 0 points.

***5. BONUS: Add features such as titles, axis labels, colors, shapes, etc. to enhance your data visualization (Points available: 2)***

- Students received 0.5 additional points for adding a regression line to help visualize the relationship between the two variables
- Students received 0.5 additional points for adding text features like titles and labels
- Students received 0.5 additional points for adding aesthetic features like colors and shapes
- Students received 0.5 additional points for adding anything to the plot which was not included in one of the previous lectures

***6. Render your document as an HTML file (Points: 2)***

- Students received 2 points if their document was properly rendered and submitted as a `.html` file.
- Students received 1 point if there was still a problem with their `.html` document that caused it to look meaningfully different from the example on Canvas

## Last Update

This document was last updated on 2024-10-14.