Test 1

STAT218

2024-02-10

### Instructions

You have 48 hours from the release of this assignment to complete and submit your work. You may refer to all class materials, notes, and textbooks, but **must complete this assignment on your own**. By submitting your work, you are affirming that your work is your own and you have not consulted with anyone else in preparing your answers. Please use the word document provided (download from the class website) and write in your answers below each prompt. You should submit your work via Gradescope; please skip the page matching step and do not match pages to parts of the outline.

The test comprises two parts: concepts, containing short multiple-choice questions; and applications, which requires some data analysis. Revisions will be allowed for the applications part, but not for the concepts part.

## Part I: concepts

This part comprises multiple-choice questions about key concepts discussed in class. You can only answer these once; revisions will not be allowed. To indicate your selection in response to each question, please type your answer explicitly, as shown below:

Example question prompt.

1. correct option
2. incorrect option
3. correct option

Answer: (a) and (c).

In each question you should select every option that applies. Some may have multiple correct options.

1. [L2] Which of the following is an example of an experimental intervention?
   1. Contacting prospective study participants during recruitment
   2. Contacting study participants for follow-up surveys
   3. Offering study participants an incentive, such as a gift card
   4. Allocating different stimuli to study participants

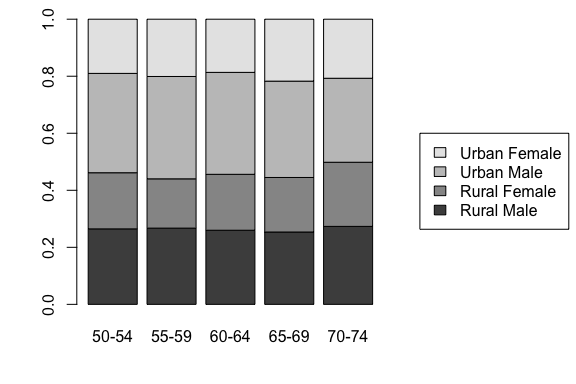
* Answer:

1. [L1] Suppose you wish to conduct a study involving a survey of California residents. Which of the following sampling schemes would identify a random sample from the target population?
   1. Physically visit CVS stores around the state of California and survey shoppers at random
   2. Obtain a list of all residential addresses, use a computer to draw 5,000 at random, and mail your survey to each selected address
   3. Obtain a list of all residential addresses, use a computer to draw 5,000 at random, and recruit volunteers to physically visit each selected address
   4. Generate a large number of random phone numbers with California area codes and send text messages to every generated number.

* Answer:

1. [L1] Suppose you read a study wherein 300 people with a self-reported history of substance abuse and 200 people without any such history are surveyed to determine how many experienced adverse events in childhood, and the study finds that a much higher proportion of those with substance abuse histories report adverse childhood experiences than those without substance abuse histories. Identify the study type.
   1. Retrospective
   2. Prospective
   3. Experiment
   4. Cohort study

The next two questions are based on the figure below, which shows deaths among residents in Virginia in 1940 by age bracket and demographic.



1. [L3] Which age group has the highest death rate?
   1. 50-54
   2. 55-59
   3. 60-64
   4. 65-69
   5. 70-74
   6. can’t tell

* Answer:

1. [L3] For which age group do urban women account for the highest proportion of deaths?
   1. 50-54
   2. 55-59
   3. 60-64
   4. 65-69
   5. 70-74
   6. can’t tell

* Answer:

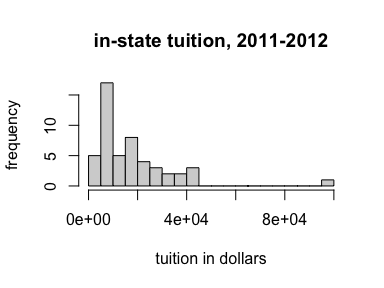
1. [L4] A 95% confidence interval for the diameter in inches of black cherry trees estimated from 31 observations made on felled trees is (12.097, 14.399). Select the correct interpretation of this estimate:
   1. With 95% probability, the diameter of black cherry trees is between 12.097 and 14.399.
   2. With 95% confidence, the diameter of black cherry trees is between 12.097 and 14.399 inches.
   3. With 95% confidence, the mean diameter of black cherry trees is between 12.097 and 14.399.
   4. With 95% confidence, the mean diameter of black cherry trees is between 12.097 and 14.399 inches.

* Answer:

1. [L4] From the interval in the previous question, determine the point estimate for the population mean.
   1. 2.3
   2. 13.2
   3. 14.5
   4. 12.1
   5. can’t tell

* Answer:

The next three questions relate to the following data on in-state tuition in dollars from a sample of 50 public and private colleges. Note that 1e+04 indicates ten thousand dollars, 2e+04 indicates twenty thousand, and so on.



1. [L3] Describe the shape of the distribution.
   1. Left-skewed
   2. Right-skewed
   3. Symmetric
   4. None of the above

* Answer:

1. [L3] The sample mean is 17781 dollars. Identify below the median.
   1. 13852
   2. 17501
   3. 19307
   4. 8000
   5. none of the above

* Answer:

1. [L3] Which measure of spread should be used to describe this data?
   1. Standard deviation
   2. Interquartile range
   3. Range
   4. None of the above

* Answer:

## Part II: applications

In this part you are given data and asked to answer questions pertaining to that data. The datasets, and commands to import them, are available to you in a project called test1 in the class posit workspace. You will need to perform calculations to answer the questions, but do not need to provide codes used to perform the calculations with your answers. When in doubt, look to the lab activities for examples.

### Diets and chick weights

The following data come from a study investigating the early growth of chicks on different diets. In the study, 47 chicks were randomly assigned one of four diets at birth and researchers measured body weight in grams daily. The data below show body weights at 18 days since birth for each chick. The question of interest is: which diet is best?

# read in data  
chick <- read.csv('data/chick.csv')  
  
# preview  
head(chick)

1. [L2] Is this observational or experimental data? Explain your reasoning.
2. [L3] Produce a visualization that compares body weight distributions by diet. For which diet have chicks grown the most? The least? Explain the statistic(s) or features of the distribution you used to make this determination.
3. [L3] Calculate point estimates and standard errors for the mean body weight at 18 days after birth on each diet.
4. [L2] Assume that in the previous question you found that chicks on diet 3 grew the most, regardless of your actual answer. Can you conclude that diet 3 caused the fastest growth? Explain why or why not.
5. [L4] Calculate a 95% confidence interval for the mean body weight of chicks on diet 3 at 18 days after birth. Interpret the interval in context.

### GSS data

The General Social Survey (GSS) is an effort to measure behaviors, opinions, and demographics of Americans and has been conducted annually since 1972. The data below are observations of a small collection of variables for 500 respondents from across several survey years.

# import GSS data  
gss <- read.csv('data/gss.csv')  
  
# preview  
head(gss)

1. [L3] Make a histogram of the weekly hours worked and describe the shape and modality. Choose an appropriate number of breaks.
2. [L3] What is the 20th percentile of weekly hours worked among respondents? What is the 80th percentile?
3. [L3] Compute a point estimate of the mean hours worked. Report the estimate and its standard error.
4. [L4] Produce and interpret an 85% confidence interval for the mean weekly hours worked.
5. [L3] Use an appropriate graphical summary to assess whether mean hours worked seems to differ by class. Explain the plot you produce and interpret any patterns observed.
6. [L3] Make any additional bivariate comparison addressing a question of your choice. State the question in non-technical terms, produce a visualization that conveys the comparison, and interpret any patterns observed.