Test 1

STAT218

2024-02-10

## Part I: concepts

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: D

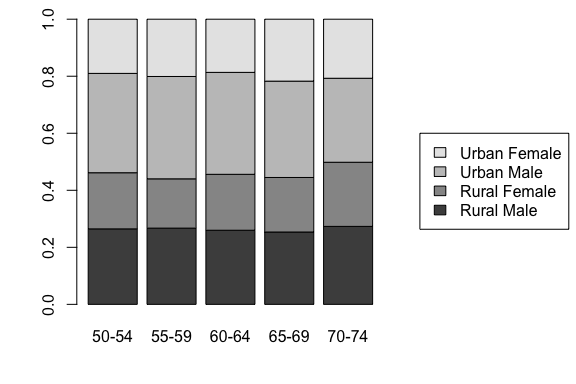
1. [L1] Suppose you wish to conduct a study involving a survey of California residents. Which of the following sampling schemes would identify a random sample from the target population?
   1. Physically visit CVS stores around the state of California and survey shoppers at random
   2. Obtain a list 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: B, C

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

* Answer: A

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: F

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: D

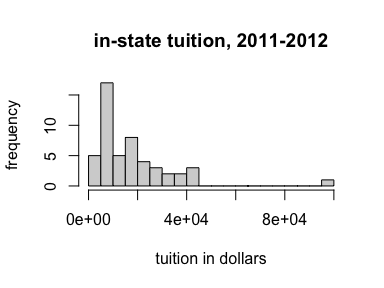
1. [L4] A 95% confidence interval for the diameter in inches of black cherry trees estimated from 31 observations made on felled trees is (12.097, 14.399). Select the correct interpretation of this estimate:
   1. With 95% probability, the diameter of black cherry trees is between 12.097 and 14.399.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

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: B

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: B

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: A

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: B

## Part II: applications

### 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?

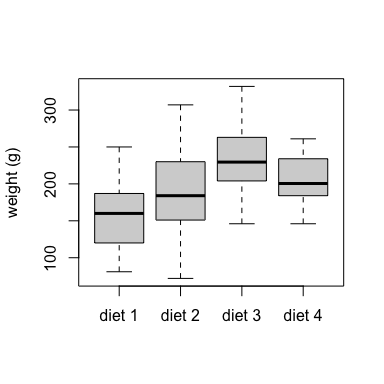
# read in data  
chick <- read.csv('data/chick.csv')  
  
# preview  
head(chick)

1. [L2] Is this observational or experimental data? Explain your reasoning.

Experimental. Researchers allocated diets at random to the chicks, which is an experimental intervention in the relevant sense.

1. [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.

boxplot(weight ~ diet, data = chick, xlab = '', ylab = 'weight (g)')



Based on the boxplots, chicks grew the most on diet 3 and the least on diet 1. This is most easily seen by comparing medians (central bar in each box), but also reflected in the other quartiles.

1. [L3] Calculate point estimates and standard errors for the mean body weight at 18 days after birth on each diet.

library(tidyverse)  
chick |>  
 group\_by(diet) |>  
 summarize(avg.weight = mean(weight),  
 se = sd(weight)/sqrt(n())) |>  
 pander::pander()

| diet | avg.weight | se |
| --- | --- | --- |
| diet 1 | 158.9 | 11.94 |
| diet 2 | 187.7 | 20.03 |
| diet 3 | 233.1 | 18.21 |
| diet 4 | 202.9 | 10.61 |

1. [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.

Yes. Since the diets were randomly allocated, no additional factors can be confounded with diet. Therefore, estimated differences in growth are due to the effect of diet.

1. [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.

With 95% confidence, the mean body weight of a chick on diet 3 at 18 days after birth is estimated to be between 191.9043713 and 274.2956287 grams.

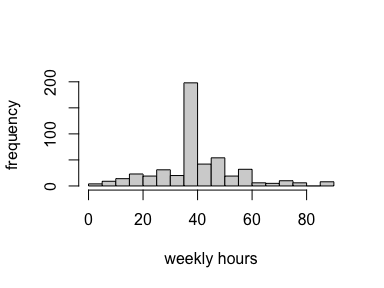
### GSS data

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.

hist(gss$weekly.hours.worked, breaks = 20,  
 main = '', xlab = 'weekly hours', ylab = 'frequency')



The breaks chosen should make apparent the relatively high frequency of ~40 hour weeks compared with all other hours.

1. [L3] What is the 20th percentile of weekly hours worked among respondents? What is the 80th percentile?

quantile(gss$weekly.hours.worked, probs = c(0.2, 0.8)) |>   
 pander::pander()

| 20% | 80% |
| --- | --- |
| 30.8 | 50 |

1. [L3] Compute a point estimate of the mean hours worked. Report the estimate and its standard error.

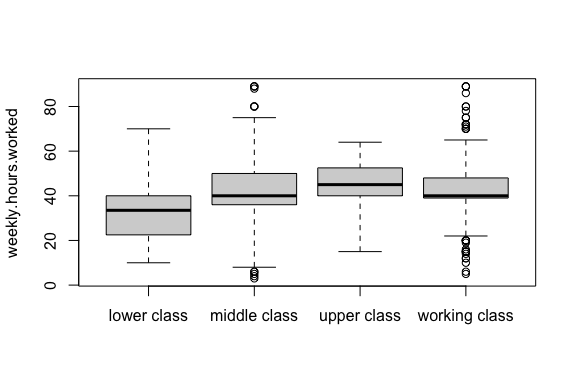
The mean weekly hours worked is estimated to be 41.382 (*SE* = 0.6627689).

1. [L4] Produce and interpret an 85% confidence interval for the mean weekly hours worked.

With 85% confidence, the mean weekly hours worked is estimated to be between 40.4264524 and 42.3375476.

1. [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.

boxplot(weekly.hours.worked ~ class,   
 data = gss,   
 range = 2,  
 xlab = '')



The boxplot shows some slight differences by class. The nost notable difference is lower class compared with all others, as the interquartile range for lower class hours worked is almost entirely below the central value of 40. Most likely, class definitions include hours worked either directly or indirectly (*e.g.,* through accounting for income), so the observed differences are not especially straightforward to interpret or necessarily meaningful.

1. [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.