SfEDA selected exercises - Lab 2

Jorge M. Mendes and M. Helena Baptista

2021

Exercise 1 - Statistical inference I

For each question, indicate whether it is best assessed by using a confidence interval, a hypothesis test, or whether statistical inference is not relevant to answer it. If inference is not relevant, explain why.

1. What percent of US voters support instituting a national kindergarten through 12th grade math curriculum?
2. Do basketball players hit a higher proportion of free throws when they are playing at home than when they are playing away?
3. Do a majority of adults riding a bicycle wear a helmet?
4. On average, were the 23 players on the 2010 Canadian Olympic hockey team older than the 23 players on the 2010 US Olympic hockey team?
5. What proportion of people using a public restroom wash their hands after going to the bathroom?
6. On average, how much more do adults who played sports in high school exercise than adults who did not play sports in high school?
7. In 2010, what percent of the US Senate voted to confirm Elena Kagan as a member of the Supreme Court?
8. What is the average daily calorie intake of 20-year-old males?

Exercise 2 - Does Vitamin C Cure the Common Cold?

A study conducted on a college campus tested to see whether students with colds who are given large doses of vitamin C recover faster than students who are not given the vitamin C. The pp-value for the test is 0.003.

1. Given the pp-value, what is the conclusion of the test: Reject H_0H0 or do not reject H_0H0?
2. Results of statistical inference are only as good as the data used to obtain the results. No matter how low a pp-value is, it has no relevance (and we can’t trust conclusions from it) if the data were collected in a way that biases the results. Describe an inappropriate method of collecting the data for this study that would bias the results so much that a conclusion based on the pp-value is very unreliable.
3. Describe a method of collecting the data that would allow us to interpret the pp-value appropriately and to extend the results to the broader student population.

4. Assuming the data were collected as you describe in part 3., use the pp-value to make a conclusion about vitamin C as a treatment for students with a common cold.

Exercise 3 - Effect of Smoking on Pregnancy Rate

Studies have concluded that smoking while pregnant can have negative consequences, but could smoking also negatively affect one’s ability to become pregnant? A study collected data on 678 women who had gone off birth control with the intention of becoming pregnant. Smokers were defined as those who smoked at least one cigarette a day prior to pregnancy. We are interested in the pregnancy rate during the first cycle off birth control. The results are summarized in the following table:

```
##              Smoker     Non-smoker Sum

## Pregnant           38.0      206.0      244.0

## Not pregnant       97.0      337.0      434.0

## Sum               135.0      543.0      678.0
```

We wish to estimate the difference in the proportion who successfully get pregnant, between smokers and non-smokers.

1. Find the best point estimate for the difference in proportions.
2. Use *R* or other technology to find and interpret a 90% confidence interval for the difference in proportions. Is it plausible that smoking has no effect on pregnancy rate?

Exercise 4 - Standard Error for Proportion of Hollywood Movies that are Action Movies

Previously the dataset *HollywoodMovies2011*, which contains information on all the 136 movies that came out of Hollywood in 2011. Thirty-two of the Hollywood movies that year were classified as action movies.

1. What proportion of Hollywood movies in 2011 were action movies? Use the correct notation with your answer.
2. Use *R* to generate a sampling distribution for the proportion of action movies for sample proportions of size $n=30$n=30. Give the shape and center of the sampling distribution and give the standard error.

Exercise 5 - Bootstrap Distributions for Intervals vs Randomization Distributions for Tests

What is the expected center of a bootstrap distribution generated to find a confidence interval?

Distribution of sample statistics with a mean approximately equal to the mean in the original distribution and a standard deviation known as the standard error

What is the expected center of a randomization distribution generated to test a hypothesis?

What makes a randomization distribution different is that it is constructed given that the null hypothesis is true. The randomization distribution will be centered on the value in the null hypothesis.

Exercise 6 - Average Size of a Performing Group in the Rock and Roll Hall of Fame

From its founding through 2012, the Rock and Roll Hall of Fame has inducted 273 groups or individuals, and 181 of the inductees have been performers while the rest have been related to the world of music in some way other than as a performer. The full dataset is available at **RockandRoll**. Some of the 181 performer inductees have been solo artists while some are groups with a large number of members. We are interested in the average number of members across all groups or individuals inducted as performers.

1. What is the mean size of the performer inductee groups (including individuals)? Use the correct notation with your answer.
2. Use *R* to create a graph of all 181 values. Describe the shape, and identify the two groups with the largest numbers.
3. Use *R* to generate a sampling distribution for the mean size of the group using samples of size $n=10$n=10. Give the shape and center of the sampling distribution and give the standard error.

Exercise 7 - What Proportion Watch the Super Bowl?

The Super Bowl is the final championship game in the National Football League in the US, and is one of the most watched television events of the year. In February 2012, just before Super Bowl XLVI, a random sample 62 of 1807 American adults were asked if they plan to watch the Super Bowl. A 95% confidence interval for the proportion planning to watch is 0.61 to 0.65.

1. What is the population? What is the sample?
2. Interpret the confidence interval in context.

3. Approximately what is the best point estimate and margin of error for the estimate?

Exercise 8 - Proportion of a Country’s Population with Access to the Internet

One of the variables in the *AllCountries* dataset gives the percent of the population of each country with access to the Internet. This information is available for all 199 countries (ignoring a few with missing values). We are interested in the average percent with Internet access.

1. What is the mean percent with Internet access across all countries? What is the standard deviation of the values? Use the correct notation with your answers.
2. Which country has the highest Internet access rate, and what is that percent? Which country has the lowest Internet access rate, and what is that percent? What is the Internet access rate for your country?
3. Use *R* or other technology to generate a sampling distribution for the mean Internet access rate using samples of size $n=10$n=10. Give the shape and center of the sampling distribution and give the standard error.

Exercise 9 - What Is an Average Budget for a Hollywood Movie?

The dataset *HollywoodMovies2011*, which contains information on all the 136 movies that came out of Hollywood in 2011.

1. Find the mean and standard deviation for the budgets (in millions of dollars) of all 2011 Hollywood movies. Use the correct notation with your answer.
2. Use *R* to generate a sampling distribution for the sample mean of budgets of 2011 Hollywood movies using a sample size of $n=20$n=20. Give the shape and center of the sampling distribution and give the standard error.

Exercise 10 - False Positives in Lie Detection

Is lie detection software accurate? A recent study was conducted in order to test the accuracy of a commonly used method of lie detection. The researchers are specifically interested in how lie detectors perform when an individual is stressed. A sample of 48 participants were gathered and attached to the lie detection device. They were asked to read a deceptive material passage out loud while receiving an electric shock (to add stress). The lie detection software inaccurately reported deception in 57% of the cases.

A bootstrap distribution shows an estimated standard error of 0.07.

1. Give a point estimate for the population parameter of interest.

2. Give a 95% confidence interval for this population parameter.
3. Comment on the accuracy of this lie detector. Do you think results from this lie detector should hold up in court?

Exercise 11 - How Common are False Positives in Lie Detection?

In previous exercise, we learn that when 48 stressed individuals read a truthful passage while being hooked up to a lie detector, the lie detection software inaccurately reported deception by 27 of them. Does this sample provide evidence that lie detection software will give inaccurate results more than half the time when used in situations such as this? State the null and alternative hypotheses.
Use *R* to create a randomization distribution, find a pp-value, and give a clear conclusion in context.

Exercise 12 - Cell Phones and Cancer

Does heavy cell phone use increase the incidence of brain tumors? A study of cell phone use among 10,000 participants found that “the 10% who used their phones most often and for the longest time had a 40% higher risk of developing some form of brain cancer than those who didn’t use a mobile phone”. Nonetheless, the results were not statistically significant. Epidemiologists Saracci and Samet write that the results ``tell us that the question of whether mobile-phone use increases risks for brain cancers remains open."

Based on this study, describe whether each statement below is plausible for this population:

1. Heavy cell phone use has no effect on developing brain cancer.
2. Heavy cell phone use is associated with an increased risk of brain cancer.
3. Heavy cell phone use causes an increased risk of brain cancer.

Exercise 13 - Infections in Childbirth

The Centers for Disease Control and Prevention (CDC) conducted a randomized trial in South Africa designed to test the effectiveness of an inexpensive wipe to be used during childbirth to prevent infections. Half of the mothers were randomly assigned to have their birth canal wiped with a wipe treated with a drug called chlorohexidine before giving birth, and the other half to get wiped with a sterile wipe (a placebo).

The response variable is whether or not the newborns develop an infection. The CDC hopes to find out whether there is evidence that babies delivered by the women getting the treated wipe are less likely to develop an infection.

1. Define the relevant parameter(s) and state the null and alternative hypotheses.
2. What is/are the sample statistic(s) to be used to test this claim?
3. If the results are statistically significant, what would that imply about the wipes and infections?

4. If the results are not statistically significant, what would that imply about the wipes and infections?

Exercise 14 - Intensive Care Unit (ICU) Admissions

The dataset *ICUAdmissions*, contains information about a sample of patients admitted to a hospital Intensive Care Unit (ICU). For each of the research questions below, define any relevant parameters and state the appropriate null and alternative hypotheses.

1. Is there evidence that mean heart rate is higher in male ICU patients than in female ICU patients?
2. Is there a difference in the proportion who receive CPR based on whether the patient’s race is white or black?
3. Is there a positive linear association between systolic blood pressure and heart rate?
4. Is either gender over-represented in patients to the ICU or is the gender breakdown about equal?
5. Is the average age of ICU patients at this hospital greater than 50?

Exercise 15 - Flaxseed and Omega-3

Flaxseed and Omega-3 Studies have shown that omega-3 fatty acids have a wide variety of health benefits. Omega-3 oils can be found in foods such as fish, walnuts, and flaxseed. A company selling milled flaxseed advertises that one tablespoon of the product contains, on average, at least 3800mg of ALNA, the primary omega-3.

1. The company plans to conduct a test to ensure that there is sufficient evidence that its claim is correct. To be safe, the company wants to make sure that evidence shows the average is higher than 3800 mg. What are the null and alternative hypotheses?
2. Suppose, instead, that a consumer organization plans to conduct a test to see if there is evidence against the claim that the product contains an average of 3800 mg per tablespoon. The consumer organization will only take action if it finds evidence that the claim made by the company is false and the actual average amount of omega-3 is less than 3800 mg. What are the null and alternative hypotheses?

Exercise 16 - Statistical Inference II

In the following cases, indicate whether the analysis involves a statistical test. If it does involve a statistical test, state the population parameter(s) of interest and the null and alternative hypotheses.

1. Polling 1000 people in a large community to determine the average number of hours a day people watch television
2. Polling 1000 people in a large community to determine if there is evidence for the claim that the percentage of people in the community living in a mobile home is greater than 10%

3. Utilizing the census of a community, which includes information about all residents of the community, to determine if there is evidence for the claim that the percentage of people in the community living in a mobile home is greater than 10%
4. Testing 100 right-handed participants on the reaction time of their left and right hands to determine if there is evidence for the claim that the right hand reacts faster than the left
5. Testing 50 people in a driving simulator to find the average reaction time to hit the brakes when an object is seen in the view ahead
6. Giving a Coke/Pepsi taste test to random people in New York City to determine if there is evidence for the claim that Pepsi is preferred
7. Using the complete voting records of a county to see if there is evidence that more than 50% of the eligible voters in the county voted in the last election

Exercise 17 - Posture and Pain

Research shows that people adopting a dominant pose have reduced levels of stress and feel more powerful than those adopting a submissive pose. Furthermore, it is known that if people feel more control over a situation, they have a higher tolerance for pain. Putting these ideas together, a recent study describes three experiments investigating how posture might influence the perception of pain.

1. In the first experiment, 89 participants were told that they were in a study to examine the health benefits of doing yoga poses at work. All participants had their pain threshold measured both before and after holding a yoga pose for 20 seconds. The pain threshold was measured by inflating a blood pressure cuff until participants said stop: The threshold was measured in mmHg and the difference in before and after thresholds was recorded. Participants were randomly divided into two groups: One group ($n=45$n=45) was randomly assigned to strike a dominant pose (moving limbs away from the body) while the other group ($n=44$n=44) was assigned to strike a submissive pose (curling the torso inward). The mean change in pain threshold for the group striking a dominant pose was 14.3 with a standard deviation of 39.8, while the mean change in pain threshold for the group striking a submissive pose was −6.1 with a standard deviation of 40.4. Does the experiment provide evidence that a dominant pose increases one’s mean tolerance of pain more than a submissive pose?
2. Prior research has shown that a person will assume a pose complementary to the pose of a peer or colleague: assuming a more submissive pose if the peer has a dominant pose and vice-versa. In the second experiment, 30 participants were told they were participating in a study on relaxation methods and randomly divided into two groups of size 15. Each participant took turns describing nature photographs with a peer who was part of the study and was secretly told to strike either a dominant or submissive posture during the interactions. Pain thresholds were measured in the same way as in the first experiment. Mean difference in pain threshold was −13.8 with a standard deviation of 27.1 for the group with a dominant peer and 4.2 with a standard deviation of 22.9 for the group with a submissive peer. Does the experiment provide evidence that mean pain tolerance is higher if one’s interaction partner is submissive? The data do not have any significant outliers.
3. As part of the experiment described in part 2., participants were also given a handgrip strength test both before and after the interaction with the peer, and the difference in handgrip strength was measured in newtons. Mean change in handgrip strength for those with a dominant interaction partner is −45.3 newtons with a standard deviation of 46.5 while for those with a submissive partner mean change was −6.8 with a standard deviation of 31.0. The data do not have any very large outliers. Find a 90% confidence interval for the

difference in means and interpret the result. Based on the confidence interval, do you believe that there is a significant difference in mean change in handgrip strength between those with a submissive partner and those with a dominant partner?

4. Since reducing the perception of pain is a goal in health care, what are the implications of these studies for health care professionals?

Exercise 18 - Infection in Dialysis Patients

The data show the time to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment.

2,5,6,7,7,8,12,13,15,15,17,22,22,23,24,27,30,34,39,53,54,63,96,113,119,130,132,141,149,152,152, 185,190,292,402,447,511,536

There are 38 patients, and the data give the first observation for each patient. The five number summary for these data is:

```
## [1]   2  15  46 149 536
```

1. Identify any outliers in the data. Justify your answer.
2. Draw the boxplot.
3. Find a 99% confidence interval for the mean time to infection for these patients. Give the best estimate, the margin of error, and give and interpret the confidence interval.
4. Is it reasonable to find a patient with a time to infection of 24 days? How about 152 days?
5. Is it reasonable to find the mean time to infection in the population is 24 days? How about 152 days?