CIs for Differences

Stat 145

Looking at differences

There are many situations where we might be interested in the difference between

- two groups or subpopulations (men vs women, smokers vs non-smokers)
- two treatments in an experiment
- two points in time (pre-test vs post-test, first-year students vs graduates)
- etc.

Paired Design vs "Two-Sample" Designs

Numerical Response When our response is numerical, we have two possible ways to compute differences:

- In a **"two-sample" design**, we **compare one whole group to the other group**.
  - Two variables for each case
    - one to tell us which group the case is in
    - one for the response variable
  - Example: diet (A or B) and amount of weight lost (in pounds) recorded for each person.
  - In this situation we will take an average for each group and compare the the two averages.
- In a **paired design**, we **compare case by case**.
  - Two variables for each case
    - Two comparable (numerical) variables
  - Example: a pre-test score and post-test score for each person.
  - In this situation we take a difference (or ratio) for each subject, and the compute the average difference (or ratio).

It is usually easy to tell which situation you are in if you keep track of the variables:

- 2-sample: 1 numerical variables and one categorical grouping variable for each case.
- paired: 2 numerical variables for each case

If you are wondering what to do if we have 3 or more groups to compare, that's an excellent question, but you'll have to wait a little while for an answer.

Categorical Response It is also possible to have a paired design with a categorical response, but the analysis methods are basically the same because either way we are dealing with two categorical variables.

Paired or 2-sample?

1. For each situation below, determine whether you would use a paired design or a 2-sample design and what your variables would be. Also say whether it would be possible to do it the other way and what you would have to change to do it the other way.

a. You want to know how much higher students' pulse rates are while taking a test vs during a normal class day.

b. You want to compare BMI (body mass index) for smokers and non-smokers.

c. You want to know who gains more weight in the first five years of marriage, men or women?

d. You want to know whether students perform better on tests if they are printed on blue paper vs white paper.

e. You want to know whether left-handed people are taller on average than right-handed people.

2. What is the potential advantage of a paired design?

Analysing Paired Designs

Analyzing a paired design involves two steps:

- **Step 1:** Convert the two numerical variables into one new variable by taking a difference or a ratio.
- **Step 2:** Analyze the new variable just like any other numerical variable.

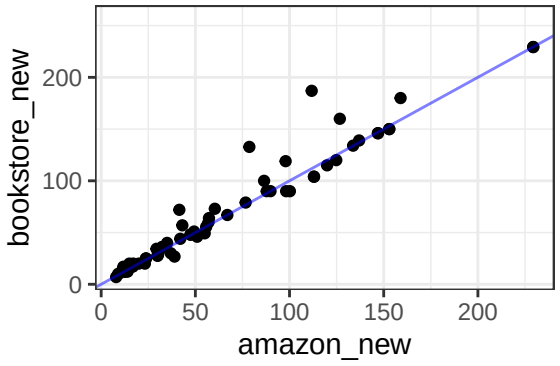
`mutate()`

For the first step, we will learn a new R command for adding a variable to a data set: `mutate()`. As an example, let's look at textbook prices to see if it is cheaper to buy books on Amazon than in the UCLA campus bookstore. This is a paired design because we want to compare prices for the same book to each other.

Let's begin by taking a look at our data

```
library(openintro)
gf_point(bookstore_new ~ amazon_new, data = ucla_textbooks_f18)  %>%
  gf_abline(slope = ~1, intercept = ~0, color = "blue", alpha = 0.5)
```

```
## Warning: Removed 133 rows containing missing values (geom_point).
```



It looks like the bookstore prices are sometimes better, but not as often or by as much as the Amazon prices are better. It looks like the price swings are larger for more expensive books. (Note also the 133 books are missing at least one price. We will remove those below.)

```
Books <-
  ucla_textbooks_f18 %>%
  # create the new variable with mutate()
  mutate(price_diff = bookstore_new - amazon_new) %>%
  # remove rows with missing price data
  filter(!is.na(price_diff))
```

3. Create a plot showing the price differences in the sample.

4. On average, how much more expensive are books in the bookstore?
5. Create a bootstrap distribution for the price difference and use it to create a 95% confidence interval for the mean price difference.
6. Now let's try using price ratio instead of price difference.
    a. Create a new variable called `price_ratio` that contains the price ratio rather than the price difference.
    b. Now create a bootstrap distribution and a 95% confidence interval for the price ratio.
    c. How would you interpret this confidence interval for your friend who has not taken statistics?
7. If you were doing this study, which would you use, price difference or price ratio? Why?

Confidence intervals for a difference in proportions.

Going to a play?

A group of students were given one of two hypothetical scenarios. In the version 1, they were planning to go to a play, but when they arrived at the theater, they realized they had lost $20. In the version 2, they realized that had lost the the ticket (which costs $20 to replace). In both cases, the students were asked if they would buy a ticket for the play.

Let's use the data in `LittleSurvey` to create a confidence interval for the difference in the proportion of people who would buy a ticket if they had lost cash and the proportion of people who would buy a ticket if they had lost a ticket ($p_{cash} - p_{ticket}$).

```
library(fastR2)     # LittleSurvey lives in this package
library(pander)

df_stats(play ~ playver, data = LittleSurvey, counts) %>% pander()
```

response	playver	n_no	n_yes
play	v1	61	103
play	v2	69	44

```
df_stats(play ~ playver, data = LittleSurvey, props) %>% pander()
```

response	playver	prop_no	prop_yes
play	v1	0.372	0.628
play	v2	0.6106	0.3894

```
diffprop(play ~ playver, data = LittleSurvey)
```

```
## diffprop
##    0.239
```

8. Create a bootstrap distribution for the difference in proportions and and use it to create a 95% confidence interval for the difference in proportions.

    [Hint: What three changes do you need to make to the previous line of R code?

SAT Prep

Imagine the following hypothetical study.

In an experiment to see how much an SAT Prep course helps, 2000 students are randomly assigned to two groups. One group receives the prep course, the other does not. Each group takes the SAT twice. The prep course group has their prep course between the two SAT tests. In each group, some students do better and some do worse, but in both groups most do better. The average improvement for each group is displayed in the table below

Prep Course	No Course
42.7	38.5

9. Answer the following questions about this study

    a. Why does this study have students take the SAT test twice?
    b. Is this a paired design? Explain.
    c. Is this a 2-sample design? Explain.
    d. The 95% confidene interval for the difference in mean performance was $(1.04, 7.36)$. Sketch what you think the bootstrap distribution looks like.
    e. Suppose we calculated a p-value from the same data. What would the null and alternative hypotheses be? Approximately what would the p-value be? Sketch what you think the randomization distribution looks like.
    f. Would you pay \$3000 for this prep course? \$300? \$30? \$3? Explain.

Body Temperature

10. Use the `BodyTemp50` data set in `Lock5withR` to answer the following questions. This data set was discussed in

    - Shoemaker, "What's Normal: Temperature, Gender and Heartrate", *Journal of Statistics Education*, Vol. 4, No. 2 (1996)

    Professor Shoemaker taught at Calvin, so I supspect that the subjects of this study are Calvin students (from the 1990s).

    a. How many men and how many women were in this study? Do you think that is a coincidence or part of the study design?
    b. Create a 95% confidence interval for the difference in mean body temperature for men and women. Should you resample within groups?
    c. Create a 95% confidence interval for the difference in mean pulse for men and women. Should you resample within groups?
    d. Create a 95% confidence interval for the correlation between pulse and body temperature. Should you resample within groups? What does this interval tell you about the association between pulse and body temperature in college-age adults?
    e. Compute a p-values for the null hypothesis that the correlation between pulse and body temperature is 0. Is the p-value consistent with what you found out from the confidence interval?