BIOS 4120: Introduction to Biostatistics
Breheny

Lab #2

In today's lab we will do a short review of what we learned last week with R, working our way through a restaurant example. Then we will review some of the concepts from lecture, predominantly confounding and randomization. Lastly we will work through an example of a hypothesis test.

1 R Review Exercise

Please open up RStudio, save a lab2.R file in your class folder, then work through the problems below. In this exercise, feel free to use last week's lab code as a guide.

1) Read in the tips.txt dataset, located on Prof. Breheny's website, into your R environment. Store it as *tips*.

2) Use the function head(*tips*), you should see the output below.

```
  TotBill  Tip Sex Smoker Day  Time Size
1   18.29 3.76   M    Yes Sat Night    4
2   16.99 1.01   F     No Sun Night    2
3   10.34 1.66   M     No Sun Night    3
4   21.01 3.50   M     No Sun Night    3
5   23.68 3.31   M     No Sun Night    2
6   24.59 3.61   F     No Sun Night    4
```

3) What was the largest bill recorded? What was the smallest? What was the average?

4) What was the largest tip recorded? What was the smallest? What was the average?

Bonus Questions

5) Calculate a *tippct* variable by dividing the tip by the total bill. Calculate the maximum, minimum, and mean of this variable. Do these answers surprise you? Why or why not?

6) In this sample, what proportion of people asked for the smoking section?

7) In this sample, was the restaurant frequented more by women or men?

2 A Note About Rates

The table below, taken from the class notes, shows a difference in rates between 2 populations. Make a note to yourself that this rate has nothing to do with the sample size. Some people tend to make the mistake, which would be reasonable if the numbers in the 3rd column were counts, but since they represent rates, we are in a sense controlling for the sample size already.

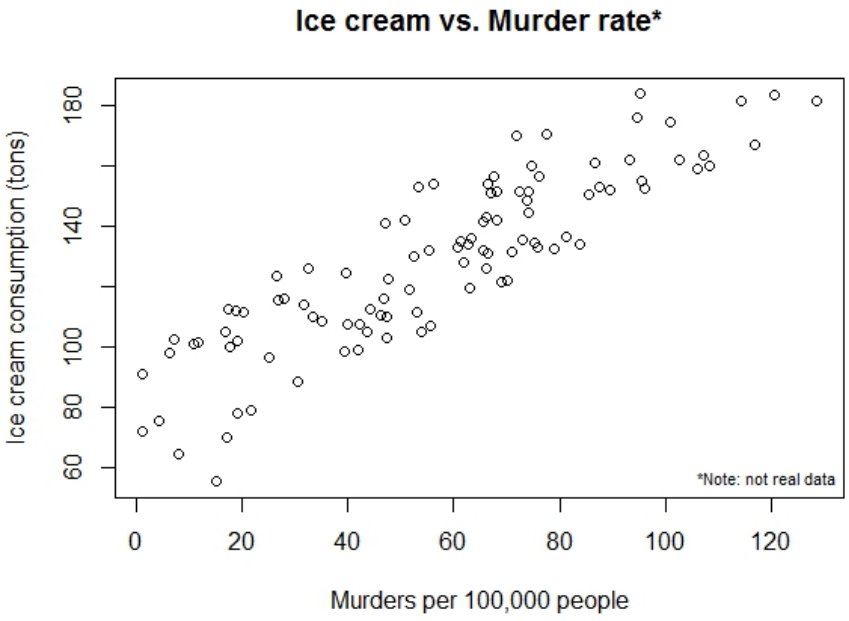
	Size of group	Polio cases per 100,000 children
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

3 Confounding

You may be familiar with the mantra "Correlation does not imply causation". Well the reason this is the case is because of confounding. There are some fun examples of this, and it is a mess to deal with in the social sciences. In experiments, however, we have the randomization process that can circumvent the issue (so long that no selection bias has crept its way into the randomization process).

Example 1: Ice cream and murders

It has been shown that ice cream sales are highly positively correlated with how many murders occur in a given month, that is, as ice cream sales increase, murders increase as well. Consider the plot below.



1) Why might this be?

2) Can you think of a way we may be able to solve whether or not ice cream consumption causes more homicides using the randomization process? (There are many possible answers here)

4 Hypothesis Testing - "Null Until Proven Alternative"

In class, you learned that there are a lot of wrong ways to think about p-values. The courtroom is a helpful example that illustrates the correct usage of p-values and hypothesis tests.

Look at it in terms of "innocent until proven guilty": As the person analyzing data, you are the judge. The hypothesis test is the trial, and the null hypothesis is the defendant. The alternative hypothesis is like the prosecution, which needs to make its case *beyond a reasonable doubt* (say, with 95% certainty).

If the evidence presented doesn't prove the defendant is guilty beyond a reasonable doubt, you still have not proved that the defendant *is innocent*. But based on the evidence, you can't reject that *possibility*.

So how would that verdict be announced? It enters the court record as "Not guilty."

That phrase is perfect: "Not guilty" doesn't mean the defendant is innocent, because that has not been proven. It just means the prosecution couldn't prove its case to the necessary, "beyond a reasonable doubt" standard. It failed to convince the judge to abandon the assumption of innocence.

If you follow that rationale, then you can see that "failure to reject the null" is just the statistical equivalent of "not guilty." In a trial, the burden of proof falls to the prosecution. When analyzing data, the entire burden of proof falls to the sample data you've collected. Just as "not guilty" is not the same thing as "innocent," neither is "failing to reject" the same as "accepting" the null hypothesis.

This method of thinking about hypothesis tests will come in handy when we start formally testing our own hypotheses.

Source: http://blog.minitab.com/blog/understanding-statistics/things-statisticians-say-failure-to-reject-the-null-hypothesis

This concludes Lab 2 ■