BIOS 4120: Introduction to Biostatistics
Breheny

Lab #10

In today's lab, we'll be learning about some very useful R functions to help out with t-tests, Fisher's exact test, and chi-squared tests.

1 *t*-tests in R

The *t.test()* function is incredibly useful. For this section refer to the *lipids.txt* dataset and the *cystic-fibrosis.txt* dataset.

```
lipids <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lipids.txt")
attach(lipids)
cf <- read.delim("http://myweb.uiowa.edu/pbreheny/data/cystic-fibrosis.txt")
attach(cf)
```

One sample continuous case: Consider TRG. Suppose we'd like to know if this sample's mean TRG level is statistically significantly different from a "healthy level" of 120. We can do this by hand (see Alzheimer's example from last week's lab), or we can do one quick line of code in R:

```
t.test(TRG, mu = 120)
```

Paired data case: Consider the cystic fibrosis data. Recall that this was a crossover study and that the data is paired. A "paired t-test" is exactly the same thing as a one-sample t-test done on the differences between the paired observations. So, you could use R to create a plug in the difference between the two into t.test() as above… Or, R actually lets you specify that you want a paired test done in a parameter:

```
t.test(Drug,Placebo,paired=TRUE)
t.test(Drug - Placebo) ## Same thing
```

Note that this gives you the same p-value and confidence intervals that we calculated in class (3/24 notes). You could once again specify the `conf.level` option if you wanted confidence intervals other than the default 95% ones.

Finally, note the `paired=TRUE` option in R. If you don't tell R to do a paired t-test, it will assume that the two sets of numbers you're giving it come from two separate samples and do a two-sample t-test (which we will discuss in the coming weeks). This represents a substantial loss of power because you are failing to take advantage of the paired design. Note that if you carry out this analysis, the p-value increases from 0.04 to 0.17.

2 Power and sample size calculation in R

The R function *power.t.test()* is very useful, and it is quite versatile. It takes all but one of the following parameters, then calculates whichever one you've omitted.

- delta = the expected difference in means
- sd = the expected standard deviation
- n = the sample size in each group
- power = the power level sought in the study

Optional: type = "paired"

Consider the cystic fibrosis example again. Say we'd like to use this data as a pilot study for a more powerful analysis. How many patients would we need in our NEW study to achieve a power of .9? We can then use the mean of the differences in the pilot group (136.5) and the standard deviation of them (223.17) as our initial estimates, and then all we need is the *power.t.test()* function:

```
power.t.test(delta = -136.5, power = .9, sd = 223.17,type = "paired")
```

Notice the n is omitted here; this is because we are telling R that "n" is what we'd like it to compute for us. Alternatively, if we knew we could recruit 50 people in each group, we could use the parameters below to have R calculate the power we'd achieve.

```
power.t.test(delta = -136.5, n = 50, sd = 223.17,type = "paired")
```

3 Fisher's Exact test and the Chi-Squared test

Example 1: Lister's experiment

Recall from class Lister's experiment testing the benefits of sterilization in surgery. This data can be found on the course website under *lister.txt*.

```
lister <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lister.txt")
attach(lister)
(tab <- table(lister))
```

```
         Outcome
Group     Died Survived
  Control   16       19
  Sterile    6       34
```

Note the `table()` function gives us a contingency table here, exactly like the one we saw in class today. Once you have your data in this form, R makes it easy to calculate either a chi-squared test or a Fisher's exact test.

```
chisq.test(tab, correct = FALSE)
fisher.test(tab)
```

Note that *fisher.test()* provides a confidence interval and estimate for the "odds ratio". We'll be talking more about these values in the coming weeks, but for now just note that if the confidence interval includes 1, we know that we will fail to reject the null hypothesis that the two treatments have the same effect.

Note also that *chisq.test()* required a "`Correct = FALSE`" statement. This is just because R by default makes a little adjustment/correction to the chi-square test, and we need the statement to shut that off. It isn't necessarily wrong to make the adjustment, but the answer you get won't agree with the approach we discussed (or will discuss very soon) in class.

Example 2: Breast cancer and age at first labor

As you may recall from earlier labs, with categorical data, people often don't report every single observation – nor do they have to. The number of subjects that fell into each category is all you need to know. So, for example, consider the results of the CDC's study of the relationship between breast cancer and the age at which a woman gave birth to her first child:

	Cancer	
	No	**Yes**
Before age 25	4475	65
25 or older	1597	31

So, is there an association between age at first labor (before/after 25) and cancer?

In R it’s pretty easy to manually enter this data by hand:

```
(tab <- cbind(c(4475, 1597), c(65, 31)))
fisher.test(tab)
```

The `cbind()` function binds lists of numbers together by column (there is also an *rbind* function if you would prefer to enter the numbers by row). If you look at tab and experiment with *cbind*, it should be pretty clear how it works. Alternatively, you can use *matrix()*.

Exercises:

- Consider the contingency table below, observing students’ note-taking habits and whether they passed a difficult comprehension test. Run a test to determine if people who took notes by hand did any different than the people who took them on a laptop.

	Passed comprehension test	
	No	**Yes**
Laptop	23	44
By hand	20	56

- Consider a retrospective study where researchers wanted to determine if marijuana is a gateway drug. They gathered a random sample of people and handed out questionnaires, the relevant responses are below. Perform a chi-squared test to determine if there is any association between marijuana use and the use of other drugs.

Used other illegal substances	**Used marijuana**	
	No	**Yes**
Yes	0	2
No	57	33

- Perform Fisher’s exact test on the same data. Does this agree with the chi-squared test? If not, which results do you trust more? Why?