

Tabular data: χ^2 and such

Lada Adamic
November 16, 2006
SI 544

1 Hablas vous R Sprache?

Sometimes (OK, actually pretty much all the time), the data is going to be in some other format, excel, SPSS, Stata, you name it, it will be in everything but a simple tab delimited file. Not to worry, there is an R library to help you read data in different formats. It is called “foreign”, and you could load it like so:

```
> library(foreign)
```

Now we can get a nice big juicy survey data set. This one is from the Pew Internet Project, which calls up thousands of people and asks them about their internet use. This particular survey (and there are many, many others) asked people about spyware, adware, and their computer use in general. Even though I downloaded this data straight off the web (after giving my name, phone, & email), it’s not really OK for me to just post it out there. So I will put it on cTools under Resources/datasets, you can download it and plunk it into your local directory (I’ve named my directory “pew”), and then load it like so:

```
> spyware = read.spss("pew/Spyware.sav")
```

This is a rather big survey, with lots and lots of questions (makes you wonder how so many people sat through the whole thing), and each question here is labeled cryptically as something like e.g. Q54D. If we examine that column we have:

```
> summary(tmp$Q54D)
```

```
Yes, have done this to avoid unwanted software No, havent done this to avoid unwanted software
                622                                593
(VOL) Have never done this activity                Dont know/Refused
                101                                20
                NA's
                665
```

In order to figure out what the question was, we need to look at the accompanying documentation in MS Word “May-June 2005 Spyware Topline.doc”. It tells us that question 54 D was:

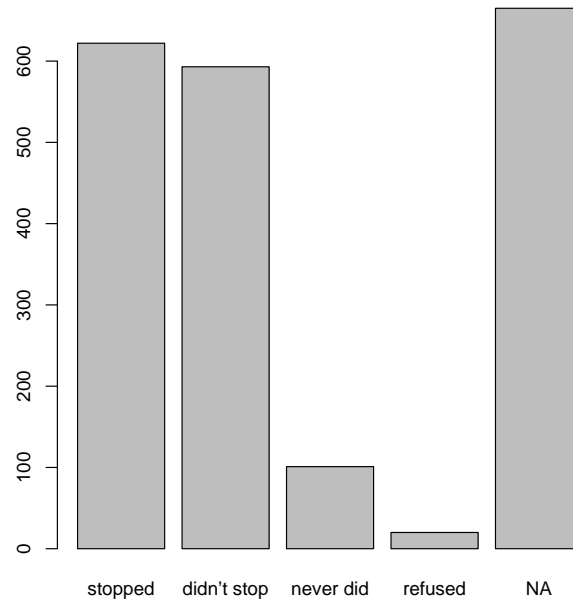
Q54 Have you, personally, done any of the following to avoid getting unwanted software programs on your computer?

d Stopped visiting particular websites

We can make a barplot of the responses. Since the column names are rather long, I renamed them.

```
> barplot(summary(tmp$Q54D),
+ names.arg=c("stopped","didn't stop","never did","refused","NA"),
+ main="Stopped visiting particular websites to avoid spyware" )
```

Stopped visiting particular websites to avoid spyware



2 tabulating data

Often times with survey data, you would like to tabulate the responses, not just by a single variable, but by considering several factors at once, for example, whether the person started avoiding websites and whether the respondents said their computers were previously infected by spyware (Q39).

As you may know, certain software programs - sometimes called “spyware” - can be installed on a person’s computer without their explicit consent, either by ”piggy-backing” onto a file or program the person downloads from the internet or just by visiting a particular website. These programs can keep track of a person’s internet habits and the sites they visit, and can transmit this information back to a central source.

As far as you know, have you ever had one of these “spyware” programs on your home computer?

Fortunately, R’s `table()` function will tabulate things for us:

```
> t1 = table(spyware$Q39,spyware$Q54D)
> colnames(t1) = c("stopped","didn't stop","never did","refused")
> rownames(t1) = c("had spyware","didn't have spyware","no computer","don't know")
> t1
```

| | stopped | didn't stop | never did | refused |
|---------------------|---------|-------------|-----------|---------|
| had spyware | 256 | 160 | 16 | 4 |
| didn't have spyware | 306 | 373 | 63 | 5 |
| no computer | 32 | 33 | 18 | 6 |
| don't know | 28 | 27 | 4 | 5 |

Notice that the missing values were automatically excluded. Now let’s just narrow our focus to the users who visited websites, had a computer and said whether they were aware of having been infected with spyware or not.

```
> t1s = t1[1:2,1:2]
> t1s
```

```
                stopped didn't stop
had spyware      256         160
didn't have spyware 306         373
```

3 χ^2 aka chi-squared

Cool, so now we've got ourselves a little table. We can run a chi-squared test on it. What are we trying to find out? Well, whether having had their computers infected by spyware influenced (of course causality strictly open to interpretation) some people to change their browsing habits to avoid certain websites.

```
> chisq.test(t1s)
```

```
Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  t1s
X-squared = 27.3612, df = 1, p-value = 1.688e-07
```

Yup it did! χ^2 is large, and the p-value is correspondingly small. The bit about Yates' continuity correction means that instead of having

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (1)$$

We have

$$\chi_{Yates}^2 = \sum \frac{(|O - E| - 0.5)^2}{E} \quad (2)$$

This has the effect of making the χ^2 statistic a bit smaller, and so the p-value a bit larger (that is less significant). This correction has a greater effect when the data is rather sparse (few entries in each cell) because you are subtracting 0.5, which is more significant if the deviation is just by 1 or 2 counts.

So we have evidence that users change their behavior based on whether they have had their computers infected.

What would have happened if we didn't narrow down the scope of the chi-squared test to only the 4 most numerous table cells?

```
> chisq.test(t1)
```

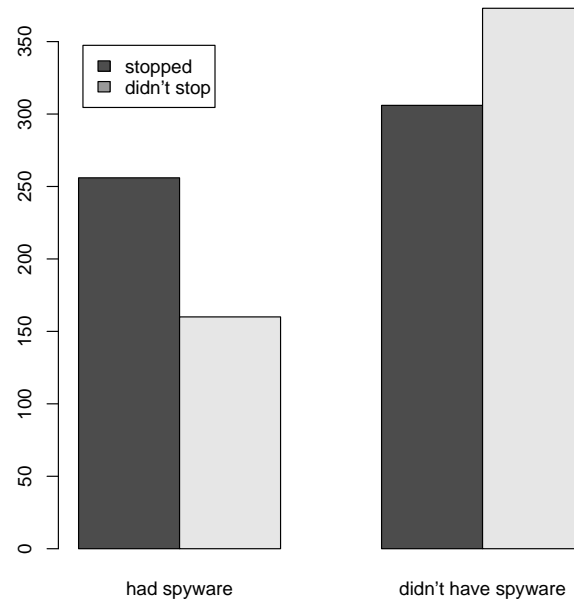
```
Pearson's Chi-squared test
```

```
data:  t1 X-squared = 99.5581, df = 9, p-value < 2.2e-16
```

```
Warning message: Chi-squared approximation may be incorrect in:
chisq.test(t1)
```

We get a very high chi-squared because now we have even more bins which have a chance of being different, but note the warning. The chi-squared approximation may be incorrect because some of the expected values were less than 5. What to do? Well, you can omit those responses from the equation, as we have done, or you can turn to more exact tests such as fisher's test (which we will discuss shortly). However, fisher's test, while exact, can be excessively computationally intensive if some of the table cells have large counts, as is the case here.

Incidentally, one does not need to tabulate the data ahead of time for R, if all the cells would end up sufficiently populated, one can just pass the data in and R can figure out the factors:



```
> chisq.test(spyware$Q39,spyware$Q54D)
```

Pearson's Chi-squared test

```
data:  spyware$Q39 and spyware$Q54D
X-squared = 99.5581, df = 9, p-value < 2.2e-16
```

Warning message:

```
Chi-squared approximation may be incorrect in: chisq.test(spyware$Q39, spyware$Q54D)
```

We can see that this has the exact same result.

4 Other tests for tabular data

You may remember the binomial distribution, corresponding to the number of successful outcomes in n trials where the probability of success at each individual trial is p . Well, instead of trying to figure out the one-tail probability by using the `pbinom()` function, we can just use `binom.test`.

So what if we were to observe 10 girl births out of 30. Is there something going on there? Could the true probability still have been 0.487?

```
> binom.test(10,30,0.487)
```

Exact binomial test

```
data: 10 and 30 number of successes = 10, number of trials = 30,
p-value = 0.1021 alternative hypothesis: true probability of success
is not equal to 0.487 95 percent confidence interval:
0.1728742 0.5281200
```

```
sample estimates: probability of success
0.3333333
```

So 10/30 is still something that could occur just by chance even if the probability of boy and girl births are about equal.

What if we tried this with just a regular test of proportions?

```
> prop.test(10,30,0.487)
```

```
1-sample proportions test with continuity correction
```

```
data: 10 out of 30, null probability 0.487
X-squared = 2.2538, df = 1, p-value = 0.1333
alternative hypothesis: true p is not equal to 0.487
95 percent confidence interval:
 0.1793758 0.5286259
sample estimates:
      p
0.3333333
```

Our estimate would not be as accurate. The p-value is a bit higher and the confidence interval for the estimate of p , the binomial probability, is ever so wider.

5 Fisher's exact test

Do you remember the gender and dieting example from the Wikipedia page on Fisher's exact test? You have 12 men and 12 women. 9 of the women are dieting, but only one of the men is. What is the probability that you would observe this by chance? Let's do this in R.

```
> dietbygender = matrix(c(1,9,11,3),nrow = 2,byrow=T)
> rownames(dietbygender) = c("dieting","not dieting")
> colnames(dietbygender) = c("men","women")
> dietbygender
      men women
dieting      1      9
not dieting 11      3
> fisher.test(dietbygender)
```

```
Fisher's Exact Test for Count Data
```

```
data: dietbygender
p-value = 0.002759
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0006438284 0.4258840381
sample estimates:
odds ratio
0.03723312
```

```
> barplot(dietbygender, beside=T)
```

So there's no way that so many more men than women are dieting by chance. Could we have used the χ^2 test instead? Actually, we could have, because we would expect ≥ 5 people in each of the cells:

```
> chisq.test(dietbygender)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: dietbygender
```

```
X-squared = 8.4, df = 1, p-value = 0.003752
```

The χ^2 test ends up giving us a looser p-value. Here it's not as important, since the result is still significant, but for best results, one should use Fisher's exact test when the samples are small.

Let's stop there.

