12.2: The χ2 test of independence (or association)

GUARDBOT1:	Halt!
GUARDBOT2:	Be you robot or human?
LEELA:	Robot…we be.
FRY:	Uh, yup! Just two robots out roboting it up! Eh?
GUARDBOT1:	Administer the test.
GUARDBOT2:	Which of the following would you most prefer? A: A puppy, B: A pretty flower from your sweetie, or C: A large properly-formatted data file?
GUARDBOT1:	Choose!

– Futurama, "Fear of a Bot Planet

The other day I was watching an animated documentary examining the quaint customs of the natives of the planet *Chapek 9*. Apparently, in order to gain access to their capital city, a visitor must prove that they're a robot, not a human. In order to determine whether or not visitor is human, they ask whether the visitor prefers puppies, flowers or large, properly formatted data files. "Pretty clever," I thought to myself "but what if humans and robots have the same preferences? That probably wouldn't be a very good test then, would it?" As it happens, I got my hands on the testing data that the civil authorities of *Chapek 9* used to check this. It turns out that what they did was very simple… they found a bunch of robots and a bunch of humans and asked them what they preferred. I saved their data in a file called `chapek9.Rdata`, which I can now load and have a quick look at:

```
load( "./rbook-master/data/chapek9.Rdata" )
str(chapek9)
```

```
## 'data.frame':    180 obs. of  2 variables:
##  $ species: Factor w/ 2 levels "robot","human": 1 2 2 2 1 2 2 1 2 1 ...
##  $ choice : Factor w/ 3 levels "puppy","flower",..: 2 3 3 3 3 2 3 3 1 2 ...
```

Okay, so we have a single data frame called `chapek9`, which contains two factors, `species` and `choice`. As always, it's nice to have a quick look at the data,

```
head(chapek9)
```

```
##   species choice
## 1   robot flower
## 2   human   data
## 3   human   data
## 4   human   data
## 5   robot   data
## 6   human flower
```

and then take a `summary()`,

```
summary(chapek9)
```

```
##   species      choice
##  robot:87   puppy : 28
##  human:93   flower: 43
##             data  :109
```

In total there are 180 entries in the data frame, one for each person (counting both robots and humans as "people") who was asked to make a choice. Specifically, there's 93 humans and 87 robots; and overwhelmingly the preferred choice is the data file. However, these summaries don't address the question we're interested in. To do that, we need a more detailed description of the data. What we want to do is look at the `choices` broken down *by* `species`. That is, we need to cross-tabulate the data (see Section 7.1). There's quite a few ways to do this, as we've seen, but since our data are stored in a data frame, it's convenient to use the `xtabs()` function.

```
chapekFrequencies <- xtabs( ~ choice + species, data = chapek9)
chapekFrequencies
```

```
##         species
## choice   robot human
##   puppy     13    15
##   flower    30    13
##   data      44    65
```

That's more or less what we're after. So, if we add the row and column totals (which is convenient for the purposes of explaining the statistical tests), we would have a table like this,

	Robot	Human	Total
Puppy	13	15	28
Flower	30	13	43
Data file	44	65	109
Total	87	93	180
which actual	ly would	be a nice	way to report the descriptive statistics for this data set. In any case, it's quite clear that the vast majority of the humans chose the data file, whereas the robots tended to be a lot more even in their preferences. Leaving aside the question of *why* the humans might be more likely to choose the data file for the moment (which does seem quite odd, admittedly), our first order of business is to determine if the discrepancy between human choices and robot choices in the data set is statistically significant.

12.2.1 Constructing our hypothesis test

How do we analyse this data? Specifically, since my *research* hypothesis is that "humans and robots answer the question in different ways", how can I construct a test of the *null* hypothesis that "humans and robots answer the question the same way"? As before, we begin by establishing some notation to describe the data:

	Robot	Human	Total
Puppy	O_{11}	O_{12}	R_1
Flower	O_{21}	O_{22}	R_2
Data file	O_{31}	O_{32}	R_3
Total	C_1	C_2	N

In this notation we say that O_{ij} is a count (observed frequency) of the number of respondents that are of species j (robots or human) who gave answer i (puppy, flower or data) when asked to make a choice. The total number of observations is written N, as usual. Finally, I've used R_i to denote the row totals (e.g., R_1 is the total number of people who chose the flower), and C_j to denote the column totals (e.g., C_1 is the total number of robots).[176]

So now let's think about what the null hypothesis says. If robots and humans are responding in the same way to the question, it means that the probability that "a robot says puppy" is the same as the probability that "a human says puppy", and so on for the other two possibilities. So, if we use P_{ij} to denote "the probability that a member of species j gives response i" then our null hypothesis is that:

H0:	All of the following are true:
	$P_{11}=P_{12}$ (same probability of saying puppy)
	$P_{21}=P_{22}$ (same probability of saying flower) and
	$P_{31}=P_{32}$ (same probability of saying data).

And actually, since the null hypothesis is claiming that the true choice probabilities don't depend on the species of the person making the choice, we can let P_i refer to this probability: e.g., P_1 is the true probability of choosing the puppy.

Next, in much the same way that we did with the goodness of fit test, what we need to do is calculate the expected frequencies. That is, for each of the observed counts O_{ij}, we need to figure out what the null hypothesis would tell us to expect. Let's denote this expected frequency by E_{ij}. This time, it's a little bit trickier. If there are a total of C_j people that belong to species j, and the true probability of anyone (regardless of species) choosing option i is P_i, then the expected frequency is just:

$$E_{ij} = C_j \times P_i$$

Now, this is all very well and good, but we have a problem. Unlike the situation we had with the goodness of fit test, the null hypothesis doesn't actually specify a particular value for P_i. It's something we have to estimate (Chapter 10) from the data! Fortunately, this is pretty easy to do. If 28 out of 180 people selected the flowers, then a natural estimate for the probability of choosing flowers is 28/180, which is approximately .16. If we phrase this in mathematical terms, what we're saying is that our estimate for the probability of choosing option i is just the row total divided by the total sample size:

$$\hat{P}_i = \frac{R_i}{N}$$

Therefore, our expected frequency can be written as the product (i.e. multiplication) of the row total and the column total, divided by the total number of observations:[177]

$$E_{ij} = \frac{R_i \times C_j}{N}$$

Now that we've figured out how to calculate the expected frequencies, it's straightforward to define a test statistic; following the exact same strategy that we used in the goodness of fit test. In fact, it's pretty much the *same* statistic. For a contingency table with r rows and c columns, the equation that defines our X_2 statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

The only difference is that I have to include two summation sign (i.e., $\sum$) to indicate that we're summing over both rows and columns. As before, large values of X^2 indicate that the null hypothesis provides a poor description of the data, whereas small values of X^2 suggest that it does a good job of accounting for the data. Therefore, just like last time, we want to reject the null hypothesis if X^2 is too large.

Not surprisingly, this statistic is X^2 distributed. All we need to do is figure out how many degrees of freedom are involved, which actually isn't too hard. As I mentioned before, you can (usually) think of the degrees of freedom as being equal to the number of data points that you're analysing, minus the number of constraints. A contingency table with r rows and c columns contains a total of r×c observed frequencies, so that's the total number of observations. What about the constraints? Here, it's slightly trickier. The answer is always the same

$$\text{df} = (r-1)(c-1)$$

but the explanation for *why* the degrees of freedom takes this value is different depending on the experimental design. For the sake of argument, let's suppose that we had honestly intended to survey exactly 87 robots and 93 humans (column totals fixed by the experimenter), but left the row totals free to vary (row totals are random variables). Let's think about the constraints that apply here. Well, since we deliberately fixed the column totals by Act of Experimenter, we have c constraints right there. But, there's actually more to it than that. Remember how our null hypothesis had some free parameters (i.e., we had to estimate the P_i values)? Those matter too. I won't explain why in this book, but every free parameter in the null hypothesis is rather like an additional constraint. So, how many of those are there? Well, since these probabilities have to sum to 1, there's only r−1 of these. So our total degrees of freedom is:

$$\begin{aligned}
\text{df} &= (\text{number of observations}) - (\text{number of constraints}) \\
&= (rc) - (c + (r-1)) \\
&= rc - c - r + 1 \\
&= (r-1)(c-1)
\end{aligned}$$

Alternatively, suppose that the only thing that the experimenter fixed was the total sample size N. That is, we quizzed the first 180 people that we saw, and it just turned out that 87 were robots and 93 were humans. This time around our reasoning would be slightly different, but would still lead is to the same answer. Our null hypothesis still has r−1 free parameters corresponding to the choice probabilities, but it now *also* has c−1 free parameters corresponding to the species probabilities, because we'd also have to estimate the probability that a randomly sampled person turns out to be a robot.[178] Finally, since we did actually fix the total number of observations N, that's one more constraint. So now we have, rc observations, and (c−1)+(r−1)+1 constraints. What does that give?

$$\begin{aligned}
\text{df} &= (\text{number of observations}) - (\text{number of constraints}) \\
&= rc - ((c-1) + (r-1) + 1) \\
&= rc - c - r + 1 \\
&= (r-1)(c-1)
\end{aligned}$$

Amazing.

12.2.2 Doing the test in R

Okay, now that we know how the test works, let's have a look at how it's done in R. As tempting as it is to lead you through the tedious calculations so that you're forced to learn it the long way, I figure there's no point. I already showed you how to do it the long way for the goodness of fit test in the last section, and since the test of independence isn't conceptually any different, you won't learn anything new by doing it the long way. So instead, I'll go straight to showing you the easy way. As always, R lets you do it multiple ways. There's the `chisq.test()` function, which I'll talk about in Section @ref(chisq.test, but first I want to use the `associationTest()` function in the `lsr` package, which I think is easier on beginners. It works in the exact same way as the `xtabs()` function. Recall that, in order to produce the contingency table, we used this command:

```
xtabs( formula = ~choice+species, data = chapek9 )
```

```
##         species
## choice   robot human
##   puppy     13    15
##   flower    30    13
##   data      44    65
```

The `associationTest()` function has exactly the same structure: it needs a `formula` that specifies which variables you're cross-tabulating, and the name of a `data` frame that contains those variables. So the command is just this:

```
associationTest( formula = ~choice+species, data = chapek9 )
```

```
##
##      Chi-square test of categorical association
##
## Variables:   choice, species
##
## Hypotheses:
##    null:        variables are independent of one another
##    alternative: some contingency exists between variables
##
## Observed contingency table:
##         species
## choice   robot human
##   puppy     13    15
##   flower    30    13
##   data      44    65
##
## Expected contingency table under the null hypothesis:
##         species
## choice   robot human
##   puppy   13.5  14.5
##   flower  20.8  22.2
##   data    52.7  56.3
##
## Test results:
##    X-squared statistic:  10.722
##    degrees of freedom:  2
##    p-value:  0.005
##
## Other information:
##    estimated effect size (Cramer's v):  0.244
```

Just like we did with the goodness of fit test, I'll go through it line by line. The first two lines are, once again, just reminding you what kind of test you ran and what variables were used:

```
Chi-square test of categorical association

Variables:   choice, species
```

Next, it tells you what the null and alternative hypotheses are (and again, I want to remind you not to get used to seeing these hypotheses written out so explicitly):

```
Hypotheses:
   null:        variables are independent of one another
   alternative: some contingency exists between variables
```

Next, it shows you the observed contingency table that is being tested:

```
Observed contingency table:
        species
choice   robot human
  puppy     13    15
  flower    30    13
  data      44    65
```

and it also shows you what the expected frequencies would be if the null hypothesis were true:

```
Expected contingency table under the null hypothesis:
        species
choice   robot human
  puppy   13.5  14.5
  flower  20.8  22.2
  data    52.7  56.3
```

The next part describes the results of the hypothesis test itself:

```
Test results:
   X-squared statistic:  10.722
   degrees of freedom:  2
   p-value:  0.005
```

And finally, it reports a measure of effect size:

```
Other information:
   estimated effect size (Cramer's v):  0.244
```

You can ignore this bit for now. I'll talk about it in just a moment.

This output gives us enough information to write up the result:

Pearson's $\chi 2$ revealed a significant association between species and choice ($\chi^2(2)=10.7, p<.01$): robots appeared to be more likely to say that they prefer flowers, but the humans were more likely to say they prefer data.

Notice that, once again, I provided a little bit of interpretation to help the human reader understand what's going on with the data. Later on in my discussion section, I'd provide a bit more context. To illustrate the difference, here's what I'd probably say later on:

The fact that humans appeared to have a stronger preference for raw data files than robots is somewhat counterintuitive. However, in context it makes some sense: the civil authority on Chapek 9 has an unfortunate tendency to kill and dissect humans when they are identified. As such it seems most likely that the human participants did not respond honestly to the question, so as to avoid potentially undesirable consequences. This should be considered to be a substantial methodological weakness.

This could be classified as a rather extreme example of a reactivity effect, I suppose. Obviously, in this case the problem is severe enough that the study is more or less worthless as a tool for understanding the difference preferences among humans and robots. However, I hope this illustrates the difference between getting a statistically significant result (our null hypothesis is rejected in

favour of the alternative), and finding something of scientific value (the data tell us nothing of interest about our research hypothesis due to a big methodological flaw).

12.2.3 Postscript

I later found out the data were made up, and I'd been watching cartoons instead of doing work.
