Goodness of fit

S520

November 15, 2016

First: Read Trosset chapter 13.2 for the general set-up for chi-squared tests. There are two ways to do the chi-squared test: the likelihood ratio version and Pearson's version. If you're going to do more statistics, I recommend the likelihood ratio version, because likelihood turns out to be a fundamental idea in both frequentist and Bayesian statistics. (Having said that, we're totally going to gloss over the idea and leave it to you to read about it Trosset if you're interested.) On the other hand, Pearson's is more interpretable, and harder to mess up.

Fully-specified models

Roll a die 20,000 times.

```
observed = c(3407, 3631, 3176, 2916, 3448, 3422)
```

Our null hypothesis is that the die is fair. From this, we can write down the counts of ones, twos, etc. that we'd expect if this null were true.

```
expected = rep(20000/6, 6)
```

Now we choose one of two chi-squared tests: the likelihood ratio version or Pearson's version. In both cases, degrees of freedom is number of categories minus 1, which gives 5. Let's do the LR first:

```
G2 = 2 * sum(observed * log(observed/expected))
1 - pchisq(G2, df=5)
```

```
## [1] 0
```

Now Pearson's:

```
X2 = sum((observed - expected)^2 / expected)
1 - pchisq(X2, df=5)
```

```
## [1] 0
```

In both cases, the P-value is basically zero. This die is unfair.

Partially-specified models

The binomial

We observe the genders of the first three children in families with 3 or more kids in Denmark.

- 0 girls: 23,236
- 1 girl: 58,529
- 2 girls: 53,908
- 3 girls: 18,770

Does the number of girls follow a binomial distribution? We treat the data as the result of a random process, and do a chi-squared test of the null hypothesis that the data comes from a binomial.

Firstly: To use a binomial, you need to know the parameters n and p. n is easy: there are 3 children per family. What about p, the probability of a girl? The null hypothesis is silent, so we need to estimate p from the data. Of all the children in the data set, what proportion are girls? Do the division:

```
families = 18770 + 53908 + 58529 + 23236
girls = 3*18770 + 2*53908 + 58529
p.girl = girls / (3*families)
```

Now do a chi-squared test. We have the observed data, now compare to the expected:

```
observed = c(23236, 58529, 53908, 18770)
expected = c(families*dbinom(0, 3, p.girl),
             families*dbinom(1, 3, p.girl),
             families*dbinom(2, 3, p.girl),
             families*dbinom(3, 3, p.girl))
```

We could also get the expected counts using this line of code:

```
expected = families * dbinom(0:3, 3, p.girl)
```

Now calculate the two version of the chi-square statistic. Find P-values by comparing to a chi-squared distribution. Degrees of freedom is categories minus 1 minus number of estimated parameters, which is two.

```
# Likelihood ratio
G2 = 2 * sum(observed * log(observed/expected))
1 - pchisq(G2, df=2)
```

```
## [1] 0
```

```
# Pearson's X^2
X2 = sum((observed - expected)^2/expected)
1 - pchisq(X2, df=2)
```

```
## [1] 0
```

The P-value is basically zero. The data isn't consistent with a binomial model.

Why not? Compare the observed and the expected:

```
data.frame(girls = 0:3, differences = round(observed - expected))
```

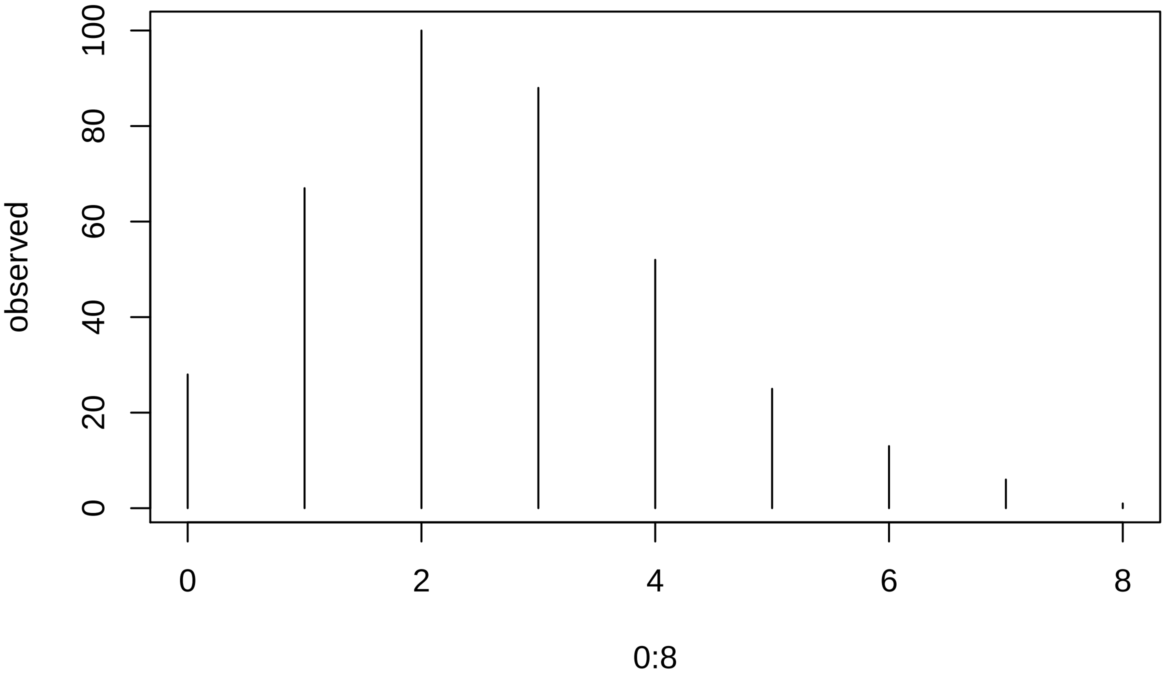
```
##   girls differences
## 1     0        1590
## 2     1       -1548
## 3     2       -1672
## 4     3        1631
```

In the real data, there are more families with 0 and 3 girls, and fewer with both boys and girls. In other worlds, boys run in some families, girls in others. The binomial assumes each child's sex is independent, so it can't account for this.

The Poisson

We have data on the number of goals scored (by both teams together) in each game of the 2000-01 English Premier League soccer season. Draw a picture:

```
# Observed number of games with 0 to 8 goals
observed = c(28, 67, 100, 88, 52, 25, 13, 6, 1)
plot(0:8, observed, type="h")
```



The **Poisson** distribution is a common probability model for count data when there's no hard maximum (as there is for the binomial.) The Poisson model is rarely literally true (except for certain physical processes.) But we can still perform a goodness-of-fit test. If the P-value is small, then the Poisson may not be adequate to describe the data.

The parameter of the Poisson distribution is λ ("lambda"), which is the expected value – here, the expected number of goals in a game. The null doesn't specify λ, so we need to estimate it from the data. To estimate a population average, we find the sample average.

```
games = sum(observed)
goals = sum((0:8)*observed)
ave = goals/games
```

Now we need to find expected counts. This is a bit tricky, because the Poisson distribution has no upper bound, while (this version of) the chi-squared test requires a finite number of categories. The rule of thumb we'll use is an expected count of at least five in each category. First, we'll try way too many categories: from 0 to 20 goals.

```
expected = games * dpois(0:20, ave)
data.frame(goals=0:20, expected=round(expected, 1))
```

```
##    goals expected
## 1      0     27.9
## 2      1     72.9
## 3      2     95.2
## 4      3     82.8
## 5      4     54.0
## 6      5     28.2
## 7      6     12.3
## 8      7      4.6
## 9      8      1.5
## 10     9      0.4
## 11    10      0.1
## 12    11      0.0
## 13    12      0.0
## 14    13      0.0
## 15    14      0.0
## 16    15      0.0
## 17    16      0.0
## 18    17      0.0
## 19    18      0.0
## 20    19      0.0
## 21    20      0.0
```

Now we can see we can get an expected count of 5 or more in each category by making the last category "7 or more" (giving 8 categories in total.) This requires a bit of fiddling – our vector "observed" currently uses nine categories, so we need to collapse the last two together.

```
observed = c(28, 67, 100, 88, 52, 25, 13, 7)
expected = rep(NA, 8)
expected[1:7] = games * dpois(0:6, ave)
expected[8] = games * (1 - ppois(6, ave))
sum(expected)
```

```
## [1] 380
```

Now do your chi-squared test of choice with $8 - 1 - 1 = 6$ degrees of freedom.

```
G2 = 2 * sum(observed * log(observed/expected))
1 - pchisq(G2, df=6)
```

```
## [1] 0.9545531
```

```
X2 = sum((observed - expected)^2/expected)
1 - pchisq(X2, df=6)
```

```
## [1] 0.9557758
```

This time the P-value is not small. The data is consistent with the null hypothesis of a Poisson distribution.

Testing indepedence

Handedness by sex

We have data on the handedness (right, left, or ambidextrous) of a representative sample of men and women.

- Right-handed: 934 men, 1070 women
- Left-handed: 113 men, 92 women
- Ambidextrous: 20 men, 8 women

Are sex and handedness independent? Our null hypothesis is independence. However, this isn't a fully-specified model. We need to estimate a bunch of things. To get expected counts for all six categories, at a bare minimum we need:

- Proportion of right-handers
- Proportion of left-handers
- Proportion of men

From these, we easily get the proportion who are ambidextrous and the proportion who are women. Then from independence, we can get the proportion of right-handed men (proportion of right-handers times proportion of men,) the proportion of ambidextrous women, and so on. Finally, turn the proportions into expected counts by multiplying by the total sample size.

```r
observed = c(934, 1070, 113, 92, 20, 8)
N = sum(observed)
RH = (934 + 1070) / N
LH = (113 + 92) / N
ambi = (20 + 8) / N
men = (934 + 113 + 20) / N
women = c(1070 + 92 + 8) / N
expected = c(RH*men*N, RH*women*N, LH*men*N,
             LH*women*N, ambi*men*N,
             ambi*women*N)
```

Check: Are all the expected counts at least five?

```r
data.frame(observed, round(expected, 1))
```

```
##   observed round.expected..1.
## 1      934              955.9
## 2     1070             1048.1
## 3      113               97.8
## 4       92              107.2
## 5       20               13.4
## 6        8               14.6
```

Now calculate the chi-squared statistic, and compare to a distribution with $6 - 1 - 3 = 2$ degrees of freedom. (In general, when testing independence for a table of counts with r rows and c columns, degrees of freedom is $(r - 1) \times (c - 1)$. See Trosset p. 343.)

Now do the test:

```r
G2 = 2 * sum(observed * log(observed/expected))
1 - pchisq(G2, df=2)
```

```
## [1] 0.002528078
```

```r
X2 = sum((observed - expected)^2/expected)
1 - pchisq(X2, df=2)
```

```
## [1] 0.002731055
```

No, handedness and sex are not independent.