Two-sample Categorical data: Testing

Patrick Breheny

October 29

Lister's experiment

- In the 1860s, Joseph Lister conducted a landmark experiment to investigate the benefits of sterile technique in surgery
- At the time, it was not customary for surgeons to wash their hands or instruments prior to operating on patients
- Lister developed a new operating procedure in which surgeons were required to wash their hands, wear clean gloves, and disinfect surgical instruments with carbolic acid
- This new procedure was compared to the old, non-sterile procedure and Lister recorded the number of patients in each group that lived or died

Contingency tables

- When the outcome of a two-sample study is binary, the results can be summarized in a 2x2 table that lists the number of subjects in each sample that fell into each category
- Putting Lister's results in this form, we have:

	Survived	
	Yes	No
Sterile	34	6
Control	19	16

- This kind of table is called a *contingency table*, or sometimes a *cross-classification* table

Contingency tables (cont'd)

- Customarily, the rows of a contingency table represent the treatment/exposure groups, while the columns represent the outcomes
- All rows and columns must represent mutually exclusive categories; thus, each subject is located in one and only one cell of the table

Lister's results

- On the surface, Lister's experiment seems encouraging: 46% of patients who received conventional treatment died, compared with only 15% of the patients who were operated on using the new sterile technique
- However, if we calculate (separate, exact) confidence intervals for the proportion who die from each type of surgery, they overlap:
  - Sterile: (6%,30%)
  - Control: (29%,63%)

Lister experiment with balls and urns

- As we've said several times, however, since we're interested in the difference between the two groups, we should analyze that difference directly
- As it turns out, there is a rather elegant, exact way to test for a difference between the two groups
- Consider representing the Lister experiment using balls and urns: there is one ball for each patient, colored red if the patient died and blue if the patient survived

Lister experiment with balls and urns (cont'd)

- Under the null hypothesis, the two groups are identical; thus, we may consider both groups as being drawn from the same urn
- Thus, consider putting all the balls into a single urn (which would contain 53 blue balls and 22 red balls) and drawing out 40 balls that we arbitrarily declare the "sterilized" group
- How often would we see something as extreme or more extreme than only 6 of these balls being red – i.e., only 6 out of 40 patients dying?

Calculating the p-value for Lister's data

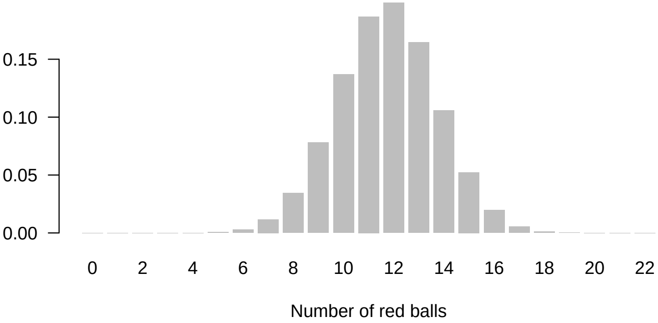
- The probability of drawing 6 red balls is

$$\frac{\binom{53}{34}\binom{22}{6}}{\binom{75}{40}} = 0.003$$

- There are several results as extreme (improbable) or more extreme than this, such as drawing 5 red balls (probability 0.0006) or drawing 18 red balls (probability 0.001)
- Adding up all such probabilities, we obtain $p = 0.005$; this is strong evidence that sterile surgery reduces the probability of death

Hypergeometric distribution

This distribution, with $f(x|n, M, N) = \binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}$, is known as the *hypergeometric* distribution:



Fisher's exact test

- This approach to testing association in a 2x2 table is called *Fisher's exact test*, after R.A. Fisher
- The test may seem somewhat strange in the sense that we are treating the number of patients who survived/died as fixed when we calculate our probability, even though of course it is truly random
- Fisher's rationale was that conditioning on the total number of successes was justified by the fact that we are testing the difference between the two groups, and the total number of successes contains no information about that difference (Fisher called such a statistic "ancillary")
- This was a novel idea in Fisher's day, but the idea of conditioning on ancillary statistics has since become a widely accepted approach to inference

The χ^2 test

- Fisher's exact test involves a fair amount of calculation; what did people do before computational resources were so abundant?
- As you can imagine from looking at the hypergeometric distribution, it is possible to approach this problem using the normal distribution to obtain approximate results
- Indeed, even before Fisher, another famous statistician (Karl Pearson) invented an approximate test for categorical data
- Pearson's invention, the χ^2-test, is one of the earliest (1900) and still most widely used statistical tests

The χ^2 distribution and hypothesis testing

- As we've seen several times in the course, z-tests have the general pattern: random variable minus its expected value divided by the standard error, which we take to be approximately normal
- Squaring this quantity, we have

$$\frac{(O-E)^2}{\mathrm{Var}(O)} \sim \chi^2_1,$$

where O is the observed value of the random variable and E is its expected value

- Note that this test statistic is naturally two-sided, in that both "left" and "right" extremes of the original z-test translate into large values of the χ^2 test statistic

The motivation behind a χ^2-test

- So essentially, the χ^2 test is simply the squared version of the z-test
- However, working with a squared quantity lends itself naturally to combining discrepancies between observed and expected over all cells in a table
- By adding up the $(O-E)^2/\mathrm{Var}(O)$ values over each cell in a table, we obtain a total measure of disagreement between the actual counts and the counts we would expect under the null hypothesis

The χ^2-statistic

- Letting the subscript i denote the cells of the table, we have the test statistic
$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$
where O_i and E_i are the observed and expected number of times category i occurs/should occur
- You may be wondering about the denominator – why have we replaced $\mathrm{Var}(O)$ with E?
- This can be justified a few ways, the simplest being that for a Poisson random variable (often used to model counts), $\mathrm{Var}(x) = \mathrm{E}(X)$; we'll discuss the Poisson distribution in greater depth next week
- Intuitively, however, it should make sense that as E increases, so does its variability

Distribution under the null

- So we've defined our test statistic and it seems as though it should follow some sort of a χ^2 distribution (being the sum of squared standard normals)
- However, the cells are obviously not independent, so perhaps the test statistic doesn't follow a χ^2_4 distribution
- It turns out to (approximately) follow a χ^2_1 distribution, although this is not an obvious fact – Pearson himself thought it followed a χ^2_3 distribution, and was rather hostile to Fisher's (1922) correction of his original χ^2 derivation, and the two feuded bitterly for many years

The χ^2-test: Lister's experiment

- Let's use the χ^2-test to determine how unlikely Lister's results would have been if sterile technique had no impact on fatal complications from surgery
- First, let's create a table of expected counts based on the null hypothesis
- In the experiment, ignoring group affiliation, 22 out of 75 patients died; thus, under the null, we would expect $22/75 = 29.3\%$ of the patients in each group to die:

	Survived	
	Yes	No
Sterile	28.3	11.7
Control	24.7	10.3

The χ^2-test: Lister's experiment (cont'd)

#2 Calculate the χ^2-statistic:

$$\begin{aligned}\chi^2 &= \frac{(34-28.3)^2}{28.3} + \frac{(6-11.7)^2}{11.7} \\ &+ \frac{(19-24.7)^2}{24.7} + \frac{(16-10.3)^2}{10.3} \\ &= 8.50\end{aligned}$$

#3 The area to the right of 8.50 is $1 - F_{\chi^2_1}(8.50) = .004$

- There is only a 0.4% probability of seeing such a large association by chance alone; again, compelling evidence that sterile surgical technique saves lives

Fisher's exact test and the χ^2-test

- Both Fisher's exact test and the χ^2-test address the same null hypothesis, so it is reassuring that we obtain virtually identical results for the Lister experiment ($p = 0.005$ vs. $p = 0.004$)
- This is often the case for 2x2 tables: the results from Fisher's exact test and the approximate χ^2-test are typically in close agreement
- However, when there are many cells with small E_i numbers, the two can yield very different results
- This is particularly problematic in larger tables

$I \times J$ tables

- The ideas of Fisher's exact test and the χ^2 test may be readily extended to larger tables with an arbitrary number of rows I and columns J
- For Fisher's exact test, we can still represent the experiment using balls and urns and calculate table probabilities, although the result no longer follows a simple hypergeometric distribution
- For the χ^2 test, the test statistic is exactly the same (just summing over more cells), and follows a $\chi^2_{(I-1)(J-1)}$ distribution under the null hypothesis

Example: Frequency of wearing gloves outside the lab

	Some College	4-year Degree	Master's Degree	Ph.D.	Other Prof. Degree
Sometimes	1	0	0	0	0
Rarely	1	3	5	15	0
Never	1	17	13	38	3

- Fisher's Exact Test: $p = .12$
- χ^2-test: $p = .00003$

Fisher's exact test vs. the χ^2-test

- How should you decide to use one versus the other?
- As in the case of one-sample data, with modern computers there is little reason to settle for the approximate answer when the exact answer can be calculated in a fraction of a second
- Nevertheless, χ^2-tests are still widely used, largely due to inertia and tradition, but also because the two generally provide very similar results, especially for 2x2 tables
- It is important to be aware, however, that the χ^2-test can be wildly incorrect when some cells have small E_i values – as a rule of thumb, this starts to become a problem when $E_i < 5$, but becomes extreme when $E_i < 1$

Summary

- Contingency tables cross-classify observations in a study according to group and outcome
- The null hypothesis that group membership is independent of the outcome can be tested using two common approaches:
  - Fisher's exact test is an exact test based on conditioning on the total number of successes and failures
  - The χ^2 test is an approximate (large-sample/central limit theorem) test based on summing up normalized differences in observed and expected counts over all cells in a table
- The two generally yield similar answers, although will start to diverge when the expected cell counts drop below ≈ 5