Two-sample Categorical data: Testing

Patrick Breheny

March 26

Separate vs. paired samples

Despite the fact that paired samples usually offer a more powerful design, separate samples are much more common, for a variety of reasons:

- It is often impossible to pair samples: for example, in the Salk vaccine trial, a child must either be vaccinated or unvaccinated – there is no way to receive both
- Even if theoretically possible, it is often impractical: for example, in studies of chronic diseases, we have to wait years to observe whether or not a person suffers a heart attack or develops breast cancer
- Furthermore, in observational studies, we have no control over which subjects end up in which group

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*, also known as a *cross-classification* table or sometimes just "cross table"

Contingency tables (cont'd)

- Often, the rows of a contingency table represent the treatment/exposure groups, while the columns represent the outcomes, although this is not universal
- 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%)

Differences between groups

- It's nice to know about the actual separate proportions that died for each type of surgery, but what we really want to know in this experiment is whether there is a difference between the two treatments
- So, rather than ask two separate questions, each using part of the data and addressing part of the question of interest, a more powerful approach is to focus all of the analysis on the one primary question of interest

Setting up a hypothesis test

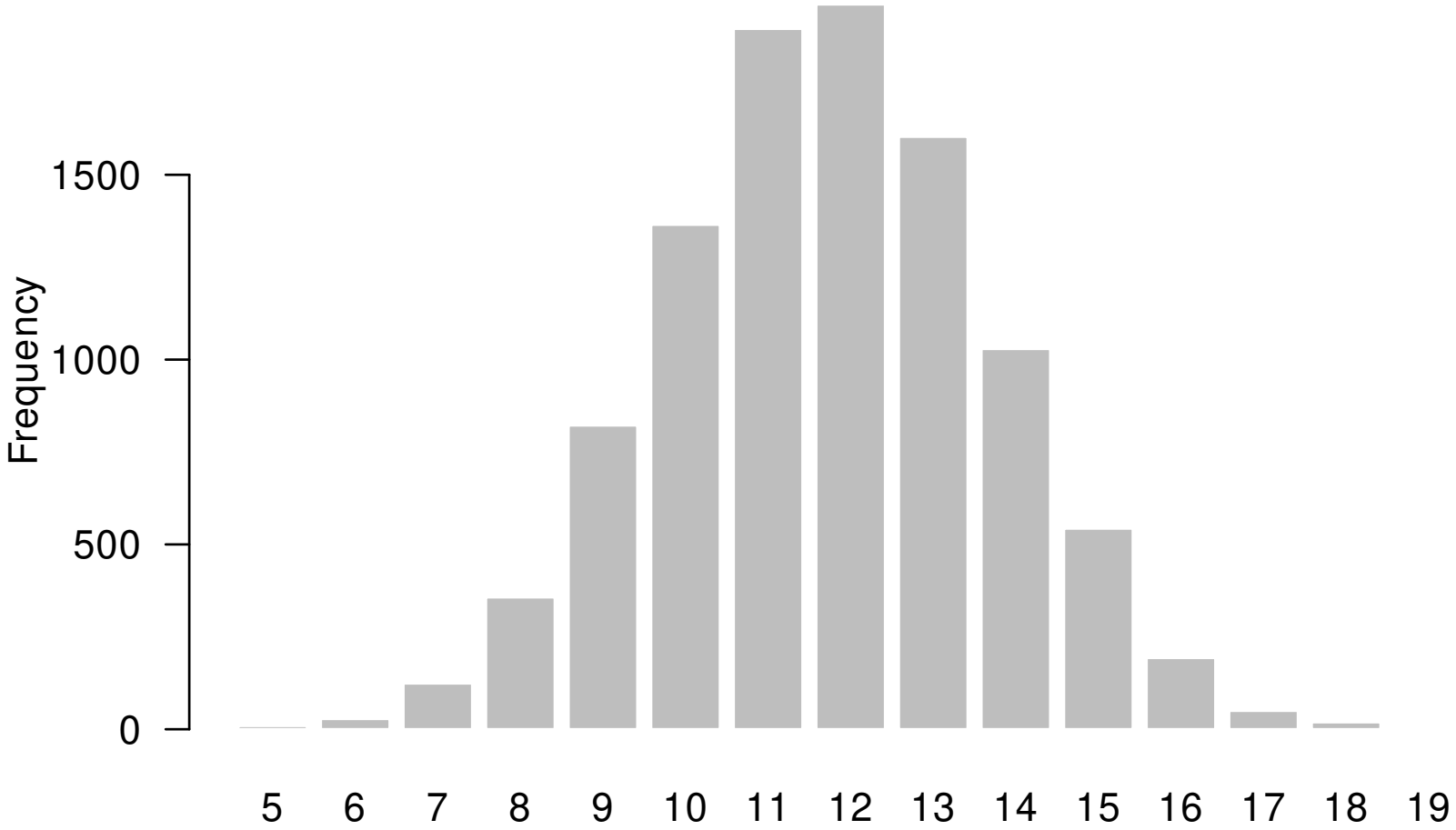
- Let's think about what a p-value is: the probability of seeing results as extreme or more extreme than what we saw, if the null hypothesis were true
- The null hypothesis here is that sterilization has no impact on the probability that a patient lives or dies – *i.e.*, that it doesn't make any difference which type of surgery the patients received
- If that were true, then it is as if we only really had one group, and everyone lived or died with equal probability regardless of the sterility of their surgery

Setting up a hypothesis test

- Thus, consider putting all the patients' outcomes into a single urn without considering the type of surgery they received (*i.e.*, this urn would contain 22 balls with "died" written on them and 53 balls with "survived" written on them)
- Our sample of 40 patients who received sterile surgery would be like randomly drawing 40 balls out of the urn
- How often would we see something as extreme or more extreme than only 6 out of 40 patients dying?

Performing the experiment

Making these draws 10,000 times, I got the following results:



Calculating a p-value from the experiment

- When I drew 40 balls from the combined urn, I only drew 6 balls 28 times (out of 10,000)
- The only results "as extreme or more extreme" than 6 were:

Number of "died" balls:	5	6	18	19
Number of times drawn:	9	28	19	2

- So I obtained a result as extreme or more extreme than the observed value a total of 58 times out of 10,000
- From this experiment, then, I would calculate a p-value of $58/10000 = 0.0058$

Fisher's exact test

- This approach to testing association in a 2x2 table is called *Fisher's exact test*, after R.A. Fisher, probably the most influential statistician of the 20th century
- Although we did the experiment using a simulation, Fisher worked out the exact probabilities that would result from this experiment
- For Lister's data, the exact probability (in lab, we will use a computer to get this result) is 0.0050; this is generally in agreement with the simulation, although by chance we ended up with slightly more extreme tables in this particular simulation

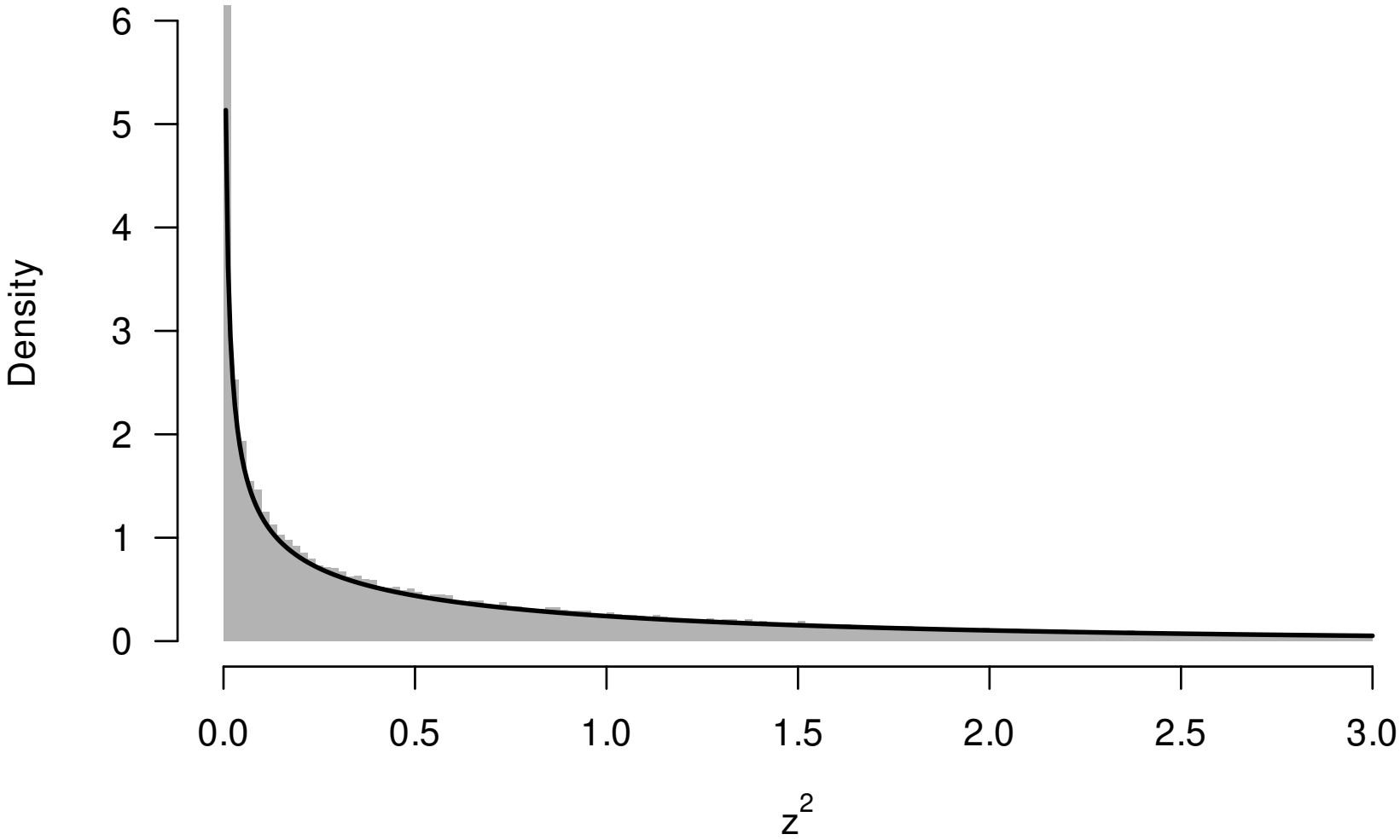
The χ^2 test

- As you might imagine, Fisher's exact test is too complicated to realistically carry out by hand in most situations
- Looking at the bar plot of the simulation results, however, it would seem possible to obtain an approximate answer using the normal distribution
- 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 most widely used statistical tests

The χ^2 curve

- The χ^2 test involves a curve we haven't seen yet called the χ^2 curve
- Before we get to the test, let's take a quick look at where the χ^2 curve comes from
- Suppose that we generated a lot of random observations from the normal distribution; the histogram of these observations would look like the normal curve
- Now suppose that we took those observations, squared them, and then made a histogram

The χ^2 curve (with 1 degree of freedom)



The χ^2 curve and hypothesis testing

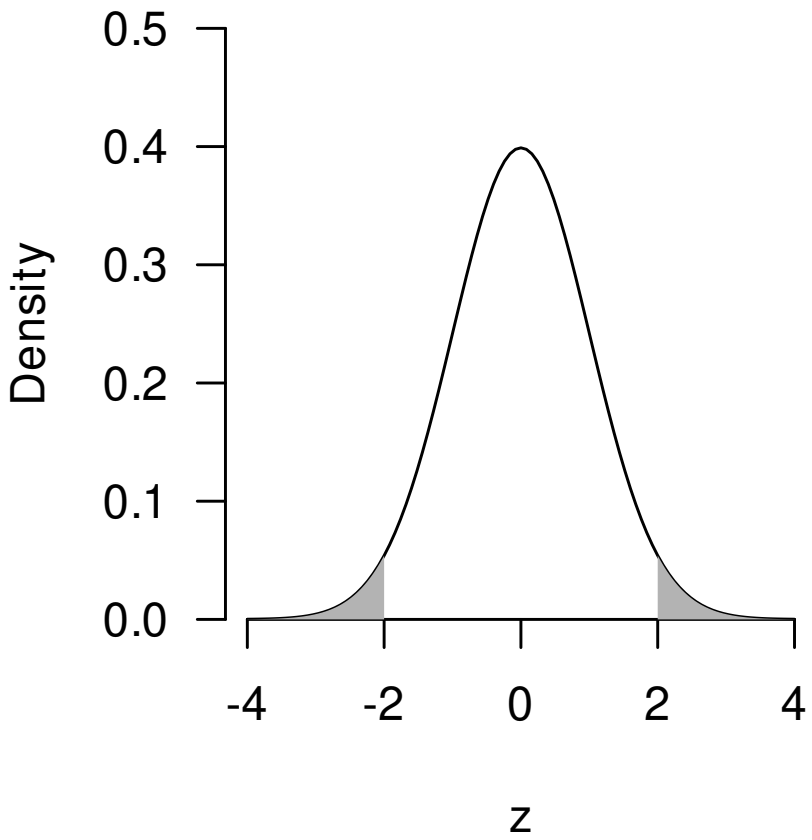
- What are the implications for hypothesis testing?
- Well, suppose we were performing a z-test: normally, we would calculate a test statistic z and find the area under the normal curve outside $\pm z$
- But what if we squared z instead?

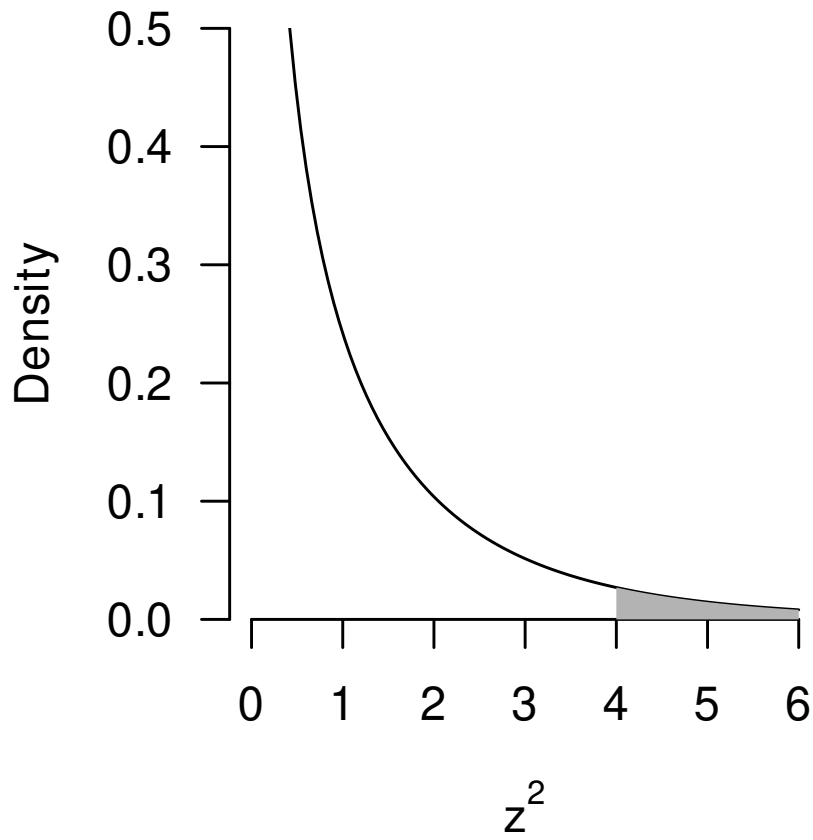
$$z^2 = \frac{(\hat{\mu} - \mu_0)^2}{\text{SE}^2}$$

The χ^2 curve and hypothesis testing (cont'd)

- Now, the area under the normal curve above $+z$ will lie above $+z^2$ ... and so will the area under the normal curve below $-z$
- The area to the right of z^2 now contains both tails of the original normal curve
- To summarize: regardless of whether $\hat{\mu}$ is far above μ_0 or far below μ_0, z^2 will be large and we will naturally get a two-sided test
- So, an alternative way to calculate p-values for z-tests is to find the area to the right of z^2 on the χ^2 curve – and don't double it

Graphical representation





The motivation behind a χ^2-test

- So essentially, the χ^2 test is simply the squared version of the z-test
- The fact that this test statistic is naturally two-sided makes it easy to compare the observed number of times each category occurs with the number of times it would be expected to occur under the null hypothesis, and then sum up these results over each of the cells in the table
- The more disagreement there is between observed and expected results, the further we will be in the right-hand tail of the χ^2 curve, and the lower our p-value will be

The χ^2-statistic

- Specifically, this is done by calculating the χ^2-statistic: letting the subscript i denote the possible categories,

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

  where O_i and E_i are the observed and expected number of times category i occurs/should occur
- The numerator should look familiar: the difference between an observed value and its expected value under the null
- The denominator, on the other hand, looks weird
- However, it just so happens that when you're counting occurrences of a category, $\text{SD} \approx \sqrt{E}$, so SD^2 is approximately equal to the expected value

The χ^2-test procedure

The procedure for performing a χ^2-test is as follows:

(1) Create a table of expected counts based on the null hypothesis

(2) Calculate the χ^2-statistic:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

(3) Determine the area to the right of χ^2 on the χ^2 curve

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

(a) Create a table of expected counts based on the null hypothesis

(b) Calculate overall event rate, ignoring group affiliation: In the experiment, 22 out of 75 patients died, so under the null, we would expect $22/75 = 29.3\%$ of patients in each group to die

(c) Multiply that rate by the sample size in each group:

	Survived	
	Yes	No
Sterile	28.3	11.7
Control	24.7	10.3

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

- Calculate the χ^2-statistic:

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

- The area to the right of 8.50 is $1 - .996 = .004$
- There is only a 0.4% probability of seeing such a large association by chance alone; this is compelling evidence that sterile surgical technique saves lives

Fisher's exact test and the χ^2-test

- So far, everything we've talked about is testing the same null hypothesis with the same data, so one would expect similar results
- Indeed, the p-values are very similar:
  - Experiment: $p = 0.006$
  - Fisher's Exact Test: $p = 0.005$
  - χ^2-test: $p = 0.004$

Fisher's exact test vs. the χ^2-test

- 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, as can happen in larger tables, the two can yield very different results
- For example, the next slide presents data from an occupational health study comparing the frequency of wearing gloves outside the lab vs. education level

Study: Frequency of wearing gloves outside the lab

	Wear gloves outside lab		
Education	Some	Rarely	Never
Some College	1	1	1
4-year Degree	0	3	17
Master's	0	5	13
Ph.D.	0	15	38
Other Degree	0	0	3

- Fisher's Exact Test: $p = 0.12$
- χ^2-test: $p = 0.00003$

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

- Know what a contingency table is and how to construct one
- Know how to carry out a χ^2-test, and in particular, how to construct a table of expected counts under the null hypothesis
- The χ^2-test is based on an approximation; there is an exact alternative (Fisher's Exact Test) that produces exact p-values
  - Complex to perform by hand
  - Trivial with modern computers