One-sample categorical data

Patrick Breheny

February 26

One-sample categorical data

- The binomial distribution plays a central role in the analysis of *one-sample categorical data*
- For example, a study at Johns Hopkins estimated the survival chances of infants born prematurely by surveying the records of all premature babies born at their hospital in a three-year period
- In their study, they found 39 babies who were born at 25 weeks gestation, 31 of which survived at least 6 months

One-sample categorical data (cont'd)

- This type of study has **one sample** of 39 babies
- If some of these babies had received one type of therapy and the rest a different kind of therapy, and we were interested in comparing the two therapies, then we would have two samples
- The outcome of this study is **categorical**, in that a baby either survived for 6 months or it didn't
- If we had instead decided to measure lung function or weight or some continuous measure of health, we would have continuous data
- As we will see, recognizing how many samples there are, and what kind of data the outcome is, plays a central role in the proper way to analyze that study

Generalization to the population

- The Johns Hopkins study observed that $31/39 = 79.5\%$ of babies survive after being born at 25 weeks gestation
- The goal of the study was not to audit their hospital's performance, but to estimate the percent of babies in other (comparable) hospitals, in future years (although maybe not too far in the future), that would survive early labor
- This is the generalization they want to make, but how accurate is their percentage?
- Could the actual percent of babies who would survive such an early labor (in other hospitals, in future years) be as high as 95%? As low as 50%?

Confidence interval

- The number of infants who survive will follow a binomial distribution
- Let p denote the true, unknown probability that an infant will survive, and let $\hat{p} = .795$ equal our estimate of that probability based on our sample (this is common notation in statistics to distinguish parameters from estimates)
- In order to build a 95% confidence interval, we need a way to calculate two numbers, (p_L, p_U) that have a 95% probability of containing p
- The most straightforward way of doing this is via hypothesis testing: test all the values of p, and anything we reject doesn't make it into the confidence interval
- Technical point: we're actually going to be doing two tests here, one for p_L and one for p_U, so they need to be carried out at the $\alpha = 0.025$ level

Trial and error

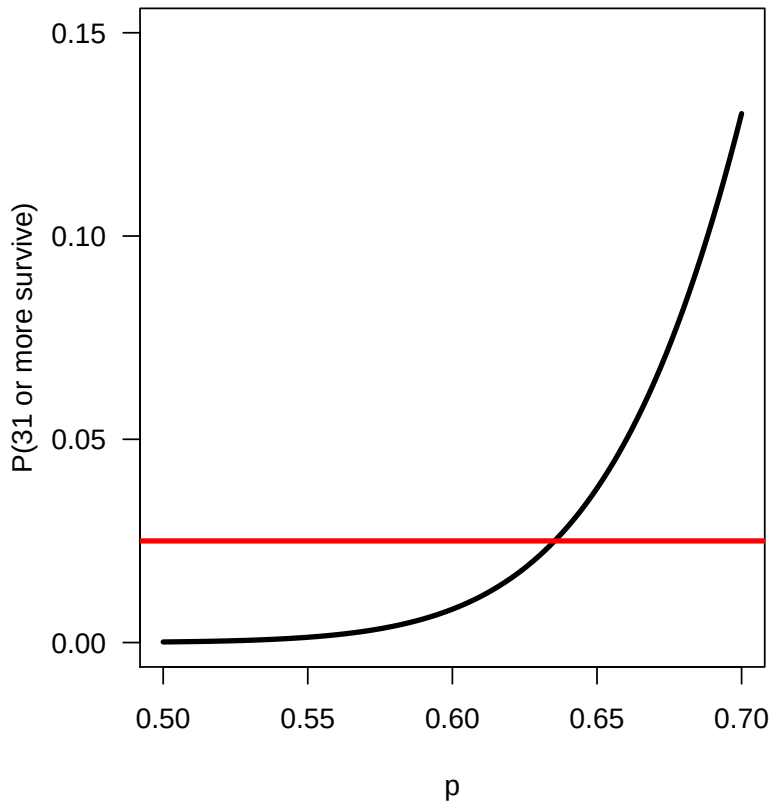
- Let's start by testing whether $p_L = .5$ is reasonable
- If $p = .5$, what is the probability that 31 or more babies (out of 39) would survive?
- Letting X denote the number of babies who survive,

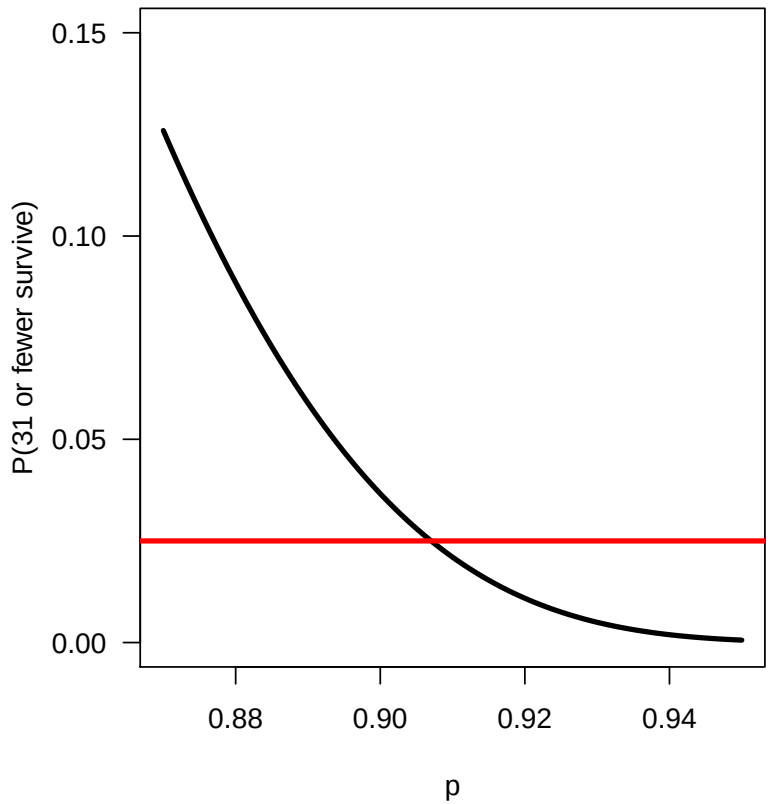
$$\begin{aligned} P(X \geq 31) &= P(X = 31) + P(X = 32) + \ldots \\ &= \frac{39!}{31!8!}.5^{31}(1 - .5)^8 + \ldots \\ &= .000112 + .000028 + \ldots \\ &= .00015 \end{aligned}$$

- This is much lower than 0.025, so we can exclude $p = 0.5$ from our interval

Finding p_L and p_U

This sort of trial and error is tedious to do by hand, but trivial for a computer:





Confidence interval results

- Thus, our confidence interval for the (population) percentage of infants who survive after being born at 25 weeks is (63.5%,90.7%)
- In their study, the Johns Hopkins researchers also found 29 infants born at 22 weeks gestation, none of which survived 6 months
- Applying the same procedure, we obtain the following confidence interval for the percentage of infants who survive after being born at 22 weeks: (0%,11.9%)

Constructing confidence intervals in R

- These intervals are not practical to construct by hand, and I do not expect you to ever attempt it
- It is very useful, however, to know how to calculate these intervals in R:

```
> binom.test(31,39)
95 percent confidence interval:
 0.6353558 0.9070361
> binom.test(0,29)
95 percent confidence interval:
 0.0000000 0.1194449
```

One-sample hypothesis tests

- It is relatively rare to have specific hypotheses in one-sample studies
- One very important exception is the collection of *paired samples*
- In a paired sampling design, we collect n pairs of observations and analyze the difference between the pairs

Hypothetical example: A sunblock study

- Suppose we are conducting a study investigating whether sunblock A is better than sunblock B at preventing sunburns
- The first design that comes to mind is probably to randomly assign sunblock A to one group and sunblock B to a different group
- There is nothing wrong with this design, but we can do better

Signal and noise

- Generally speaking, our ability to make generalizations about the population depends on two factors: *signal* and *noise*
- *Signal* is the magnitude of the difference between the two groups – in the present context, how much better one sunblock is than the other
- *Noise* is the variability present in the outcome from all other sources besides the one you're interested in – in the sunblock experiment, this would include factors like how sunny the day was, how much time the person spent outside, how easily the person burns, etc.
- Hypothesis tests depend on the ratio of signal to noise – how easily we can distinguish the treatment effect from all other sources of variability

Signal to noise ratio

- To get a larger signal-to-noise ratio, we must either increase the signal or reduce the variability
- The signal is usually determined by nature and out of our control
- Instead, we are going to have to reduce the variability/noise
- If our sunblock experiment were controlled, we could attempt such steps as forcing all participants to spend an equal amount of time outside, on the same day, in an equally sunny area, etc.

Person-to-person variability

- But what can be done about person-to-person variability (how easily certain people burn)?
- A powerful technique for reducing person-to-person variability is *pairing*
- For each person, we can apply sunblock A (at random) to one of their arms, and sunblock B to the other arm, and as an outcome, look at the difference between the two arms
- In this experiment, the items that we randomly sample from the population are pairs of arms belonging to the same person

Benefits of paired designs

- What do we gain from this?
- As variability goes down,
  - Confidence intervals become narrower
  - Hypothesis tests become more powerful (smaller p values)

Pairing in observational studies

- Experimenters have come up with all kinds of clever ways to use pairing to cut down on variability:
  - Crossover studies
  - Split-plot experiments
- Pairing is also widely used in observational studies
  - Twin studies
  - Matched studies
- In a matched study, the investigator will pair up ("match") subjects on the basis of variables such as age, sex, or race, then analyze the difference between the pairs
- In addition to increasing power, pairing in observational studies also eliminates (some of the) potential confounding variables

Cystic fibrosis experiment

- As an example of a paired study, we will look at a crossover study of the drug amiloride as a therapy for patients with cystic fibrosis
- Cystic fibrosis is a genetic disease that affects the lungs
- Forced vital capacity (FVC) is the volume of air that a person can expel from the lungs in 6 seconds
- FVC is a measure of lung function, and is often used as a marker of the progression of cystic fibrosis

Design of the cystic fibrosis experiment

- There were 14 people who participated in the study
- Each participant in the trial received both the drug and the placebo (at different times), "crossing over" to receive the other treatment halfway through the trial
- Like all well-designed crossover trials, the therapy (treatment/placebo) that each participant received first was chosen *at random*
- Furthermore, there was a *washout period* during the crossover between the two drug periods

The outcome

- To determine an outcome, the FVC of the patients was measured at the beginning of each treatment period, and again at the end
- The outcome is the reduction in lung function over the treatment period
- So, for example, if a patient's FVC was 900 at the beginning of the drug period and 850 at the end, the reduction is 50
- In the actual study, 11 of the 14 patients did better on the drug than on the placebo
- A hypothesis test informs us whether or not this kind of result could be due to chance alone

The null hypothesis

- The null hypothesis here is that the drug provides no benefit – that whether the patient received drug or placebo has no impact on their lung function
- Under the null hypothesis, then, the probability that a patient does better on drug than placebo is 50% (i.e., $p = 0.5$)
- Essentially, under the null, whether a patient does better on one treatment or another is like flipping a coin

The binomial test

- One way to test this null hypothesis would be to flip a coin 14 times, count the number of heads, and repeat this over and over again to see how unusual "11 heads" is
- However, this is unnecessary, as we already have the binomial distribution to calculate these probabilities for us
- Under the null hypothesis, the number of patients who do better on the drug than placebo (X) will follow a binomial distribution with $n = 14$ and $p = 0.5$
- This approach to hypothesis testing goes by several names, and could be called the *exact test*, the *binomial test*, or the *sign test*
- What we need to do is calculate the p-value: the probability of obtaining results as extreme or more extreme than the one observed in the data, given that the null hypothesis is true

"As extreme or more extreme"

- The result observed in the data was that 11 patients did better on the drug
- But what exactly is meant by "as extreme or more extreme" than 11?
- It is uncontroversial that 11, 12, 13, and 14 are as extreme or more extreme than 11
- But what about 0? Is that more extreme than 11?
- Under the null, $P(11) = 2.2\%$, while $P(0) = .006\%$
- So 0 is more extreme than 11, but in a different direction

One-sided vs. two-sided tests

- Potentially, then, we have two different approaches to calculating this p-value:
  - Find the probability that $X \geq 11$
  - Find the probability that $X \geq 11 \cup X \leq 3$ (the number that is as far away from 7 as 11 is, but in the other direction)
- These are both reasonable things to do, and intelligent people have argued both sides of the debate
- However, the scientific community has for the most part come down in favor of the latter – the so called "two-sided test"
- For this class, all of our tests will be two-sided tests

The binomial test

- Thus, the p-value of the sign test is

$$\begin{aligned} p &= P(X \leq 3) + P(X \geq 11) \\ &= P(X = 0) + \cdots + P(X = 3) + P(X = 11) + \cdots + P(X = 14) \\ &= .006\% + .09\% + .6\% + 2.2\% + 2.2\% + .6\% + .09\% + .006\% \\ &= 5.7\% \end{aligned}$$

- In R:

```
> binom.test(11,14)
p-value = 0.05737
```

- Seeing 11 out of 14 patients do better on one treatment than another is therefore fairly unlikely, but represents only a moderate amount of evidence against the null hypothesis

Cystic fibrosis study: Confidence interval

- Recall that it is not valid to conclude from this that amiloride is equivalent to placebo
- As always, calculating a confidence interval provides more information
- Here, the binomial confidence interval is [49.2, 95.3], indicating that while it is possible the drug has no effect (50% is inside the interval), it is also possible the drug has a huge effect and would benefit 95% of cystic fibrosis patients
- Unfortunately, with only 14 subjects enrolled, this study is fairly inconclusive and the confidence interval is very wide

Summary

- It is possible to calculate exact confidence intervals for one-sample categorical studies using the binomial distribution (but you'd need a computer to do it)
- Pairing is a powerful idea in study design for reducing variability and increasing the power of an experiment
- Often, there is no null hypothesis for one-sample studies; paired studies are an exception
- For one-sample categorical studies, you can calculate exact p-values for testing the null hypothesis using the binomial distribution (you wouldn't need a computer, but a calculator would help)