One-sample categorical data

Patrick Breheny

September 15

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
- It is reasonable to assume that the number of babies who survived follows a binomial distribution

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 θ 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 be as high as 95%? As low as 50%?

Hypothesis testing

- Suppose we wanted to test the hypothesis that $\theta = 0.5$; we need to calculate the probability of seeing results as extreme or more extreme than what we saw, under the given assumption $\theta = 0.5$
- This is a pretty straightforward given the binomial formula:

$$\begin{aligned} P(X \geq 31|\theta = 0.5) &= \sum_{x=31}^{39} P(X = x|\theta = 0.5) \\ &= \sum_{x=31}^{39} \binom{39}{x} 0.5^{39} \\ &= .00015 \end{aligned}$$

What about a confidence interval?

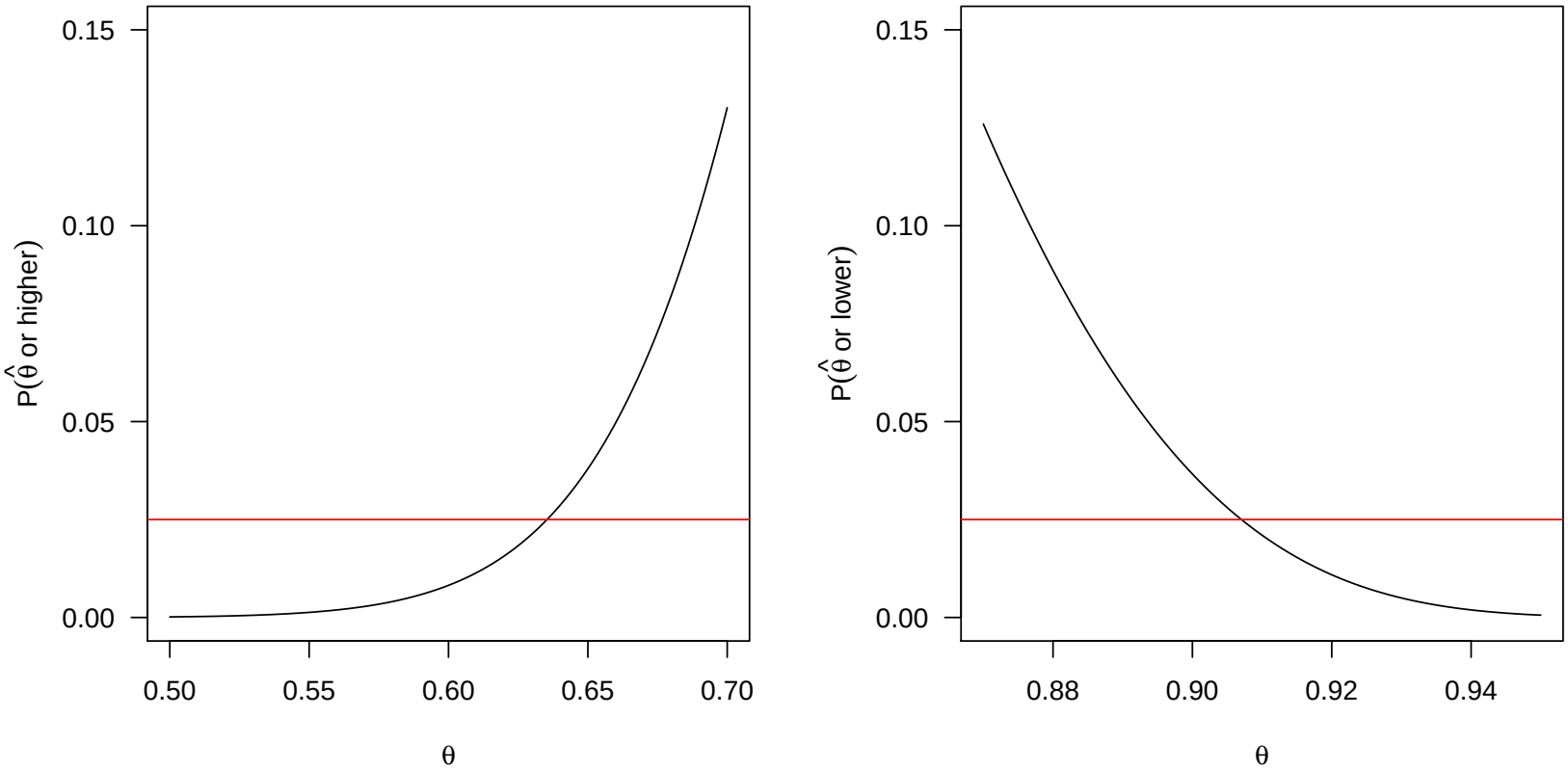
- Of course, this is an odd hypothesis test because there's nothing special about $\theta = 0.5$
- What we really want is a confidence interval (θ_L, θ_U) for the range of values of θ that are consistent with the data
- It's certainly not obvious here how you would go about finding a confidence interval directly
- However, recall the connection between hypothesis testing and confidence intervals; if we can test values of θ, we can use that to make a confidence interval

Setting up the problem

- Let's suppose we're constructing a 95% confidence interval; that means we get to make a mistake 5% of the time
- In other words, in carrying out our hypothesis tests, we get to have a 5% type I error rate
- To construct the confidence interval, however, we have to carry out two sets of tests: one to find θ_L and another to find θ_U
- If each set has a type I error rate of 5%, the resulting confidence interval will miss the mark 10% of the time; thus, each set of tests has to set $\alpha = 0.025$
- Remark: In principle, we could have $\alpha_L = 0.01$ and $\alpha_U = 0.04$ or something and that would still technically be a 95% confidence interval, but it would be hard to justify this

Finding θ_L and θ_U

It would be rather time-consuming to find θ_L and θ_U by hand, but easy with a computer, as we will see in lab later:



This procedure is known as the *Clopper-Pearson interval*

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%]

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 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., $\theta = 0.5$)
- Essentially, under the null, whether a patient does better on one treatment or another is like flipping a coin, and we want to know how unusual ("extreme") getting 11 heads is

The binomial test

- We don't actually have to flip coins, of course, because we 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 $\theta = 0.5$
- 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 $X \geq 11 \cup X \leq 3$ (under the null, $P(X = 3)$ and $P(X = 11)$ are equal)
- These are both reasonable things to do, and intelligent people have argued both sides of the debate
- However, the scientific community has come down in favor of the latter – the so called "two-sided test"
- In this class, our tests will be two-sided tests (with a few exceptions, such as when we're using a test to construct a confidence interval)

The binomial test

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

$$\begin{aligned} p &= P(X \leq 3) + P(X \geq 11) \\ &= P(X = 0) + \cdots + P(X = 3) + P(X = 11) + \cdots + P(X = 14) \\ &= .00006 + .0009 + .006 + .022 + .022 + .006 + .0009 + .00006 \\ &= 0.057 \end{aligned}$$

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

Confidence interval

- Of course, it is also worth calculating a confidence interval here for the percentage of patients who would do better on amiloride
- In this case, the interval is [0.49, 0.95]
- So perhaps we can't entirely rule out the possibility that the drug has no effect, but it's also possible that the drug has a considerable effect on lung function and would benefit 95% of cystic fibrosis patients

Summary

- The binomial distribution can be directly applied to carry out hypothesis tests for one-sample studies with binary outcomes
- By inverting the hypothesis test, we can construct confidence intervals
- 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