One-sample categorical data: approximate inference

Patrick Breheny

October 6

Introduction

- It is relatively easy to think about the distribution of data – heights or weights or blood pressures: we can see these numbers, summarize them, plot them, etc.
- It is much harder to think about things like the distribution of the sample mean, because in reality the experiment is conducted only once and we only see one mean
- The distribution of the mean is more of a hypothetical concept describing what would happen if we were to repeat the experiment over and over

Sampling distributions

- Consider a study to determine the average cholesterol level in a certain population; if we were to repeat this study many times, we would get different estimates each time, depending on the random sample we drew
- To reflect the fact that its distribution depends on the random sample, the distribution of an estimate (such as the sample mean) is called a *sampling distribution*
- Sampling distributions are of fundamental importance to the long-run frequency approach to statistical inference and essential for carrying out hypothesis tests and constructing confidence intervals
- In a broader sense, we study sampling distributions to understand how reproducible a study's findings are, and in turn, how accurate its generalizations are likely to be

Sampling distributions (cont'd)

- For independent one-zero outcomes, the sampling distribution was simple enough that we could derive it exactly and describe it with a simple formula
- For most other outcomes, however, this is not possible and we often rely instead on the central limit theorem to provide the sampling distribution – as we've seen, this is not exact, but usually a very good approximation

Applying the central limit theorem

- To get a sense of how useful the central limit theorem is, let's return to our hypothetical study to determine an average cholesterol level
- According the National Center for Health Statistics, the distribution of serum cholesterol levels for 20- to 74-year-old males living in the United States has mean 211 mg/dl, and a standard deviation of 46 mg/dl (these are estimates, of course, but for the sake of this example we will take them to be the true population parameters)
  - We collect a sample of size 25; what is the probability that our sample average will be above 230?
  - We collect a sample of size 25; 95% of our sample averages will fall between what two numbers?
  - How large does the sample size need to be in order to insure a 95% probability that the sample average will be within 5 mg/dl of the population mean?

Introduction

- We can use this same line of thinking to develop hypothesis tests and confidence intervals
- We'll begin by revisiting one-sample categorical data because
  - It's the simplest scenario
  - We can compare our new simple-yet-approximate results to the exact hypothesis tests and confidence intervals that we obtained earlier based on the binomial distribution

One-zero (Bernoulli) distribution: mean and variance

- To use the central limit theorem, we need the population mean and variance
- For a single one-zero outcome (known as the *Bernoulli* distribution), its mean is π as we showed in the previous lecture
- **Theorem:** For a Bernoulli random variable X, $\text{Var}(X) = \pi(1 - \pi)$

Hypothesis testing

- Now we're ready to carry out a hypothesis test based on the central limit theorem
- Consider our cystic fibrosis experiment in which 11 out of 14 people did better on the drug than the placebo; expressing this as an average, $\hat{\pi} = 11/14 = .79$ (i.e., 79% of the subjects did better on drug than placebo)
- Under the null hypothesis, the sampling distribution of the percentage who did better on one therapy than the other will (approximately) follow a normal distribution with mean $\pi_0 = 0.5$
- The notation π_0 refers to the hypothesized value of the parameter π under the null

The standard error

- What about the standard error (i.e., the standard deviation of $\hat{\pi}$)?
- Recall that $\text{SE} = \text{SD}/\sqrt{n}$, so for a Bernoulli random variable,

$$SE = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$
$$= \frac{1}{2\sqrt{n}}$$

- For the cystic fibrosis experiment, under the null $\text{SE} = 0.134$

Approximate test for the cystic fibrosis experiment

- To calculate a p-value, we need the probability that $\hat{\pi}$ is more extreme than $11/14$ given that the true probability is $\pi_0 = 0.5$
- By the central limit theorem, under the null

$$\frac{\hat{\pi} - \pi_0}{\text{SE}} \dot{\sim} \text{N}(0, 1)$$

- Thus,

$$\begin{aligned} z &= \frac{.786 - .5}{.134} \\ &= 2.14 \end{aligned}$$

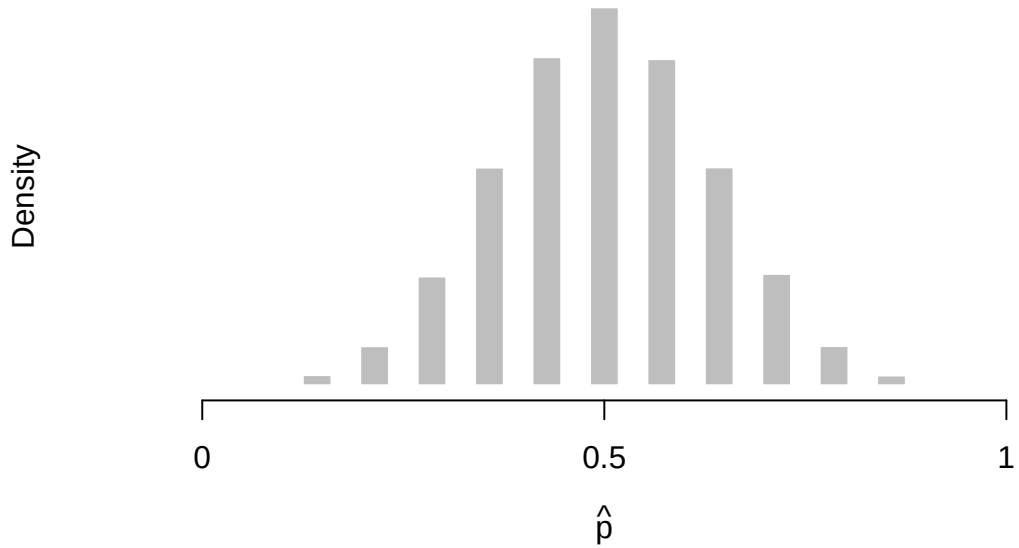
  and the p-value of this test is therefore $2(1 - \Phi(2.14)) = .032$
- In other words, if the null hypothesis were true, there would only be about a 3% chance of seeing the drug do this much better than the placebo

Terminology

- Hypothesis tests revolve around calculating some statistic (known as a *test statistic*) from the data that, under the null hypothesis, you know the distribution of
- In this case, our test statistic is z: we can calculate it from the data, and under the null hypothesis, it follows a normal distribution
- Tests are often named after their test statistics: the testing procedure we just described is called a *z-test*

Accuracy of the approximation

- So the z-test indicates moderate evidence against the null; recall, however, that we calculated a p-value of 6% from the (exact) binomial test, which is more in the "borderline evidence" region
- With a sample size of just 14, the distribution of the sample average is still fairly discrete, and this throws off the normal approximation by a bit:



Introduction: confidence intervals

- Now let's turn our attention to confidence intervals
- As usual, this is a harder problem – hypothesis testing was straightforward because under the null, we knew π_0 and therefore we know the standard error
- This is not true in trying to determine a confidence interval – the SE depends on π, which we don't know
- There are two common approaches to dealing with this problem, known as the *Wald* interval and the *score* interval; we will discuss both

Wald approach: Main idea

- In the Wald approach, we use $\hat{\pi}$ to estimate SE
- The idea behind this approach is that uses our "best guess" about π to obtain a "best guess" for the SE
- Otherwise, however, this approach does not directly account for the fact that SE depends on π

Wald approach for CF study

- For the CF study,

$$\begin{aligned} \text{SE} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= \sqrt{\frac{0.786(1-0.786)}{14}} \\ &= 0.110 \end{aligned}$$

- Now, by the central limit theorem,

$$\frac{\hat{\pi} - \pi}{0.110} \dot{\sim} \text{N}(0, 1)$$

and we can solve for π to obtain a confidence interval

Wald approach for CF study (cont'd)

- For the standard normal distribution,

$$\Phi^{-1}(0.975) = 1.96$$
$$\Phi^{-1}(0.025) = -1.96$$

- Thus,

$$0.95 = P(-1.96 < Z < 1.96) \approx P\left(-1.96 < \frac{\hat{\pi} - \pi}{0.110} < 1.96\right),$$

and

$$[\hat{\pi} - 1.96(0.110), \hat{\pi} + 1.96(0.110)] = [57.1\%, 100.0\%]$$

is an approximate 95% confidence interval for π

Wald formula

- Let z_α denote the value that contains the middle $100(1-\alpha)$ percent of the standard normal distribution
- We can summarize the Wald interval with the formula $\hat{\pi} \pm z_\alpha \text{SE}$, where $\text{SE} = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$
- As we will see, this is actually a very common form for confidence intervals (estimate plus/minus a multiple of the standard error), although the multiplier and standard error formulas change depending on what we are estimating

Score approach: Main idea

- The score approach also uses the central limit theorem to create approximate confidence intervals, but does so in a different manner than the Wald approach
- The score approach works very similarly to the Clopper-Pearson interval, except that instead of inverting the binomial test, we invert the CLT-based test from earlier
- This amounts to solving the quadratic formula

$$\frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} = z_\alpha$$

for π

Score approach: Formula

- In other words, the endpoints of the score interval are given by

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

  where $a = 1 + z_\alpha^2/n$, $b = -z_\alpha^2/n - 2\hat{\pi}$, and $c = \hat{\pi}^2$ (although I certainly don't expect you to remember this formula)
- For the cystic fibrosis study, the 95% CI is $[52.4\%, 92.4\%]$
- The score approach lies somewhat in between the Wald and Clopper-Pearson approaches: still based on a CLT approximation to the true sampling distribution, but accounting for the fact that SE varies with π

Cystic fibrosis study

- Let's take a look at how the three confidence intervals (binomial, wald, score) compare for the three studies we've discussed previously
- For the cystic fibrosis study (x=11, n=14), we have:
  - Binomial: [49.2, 95.3]
  - Wald: [57.1, 100.0]
  - Score: [52.4, 92.4]
- The score interval isn't too bad, but the Wald interval is pretty far off

Infant survival, 25 weeks

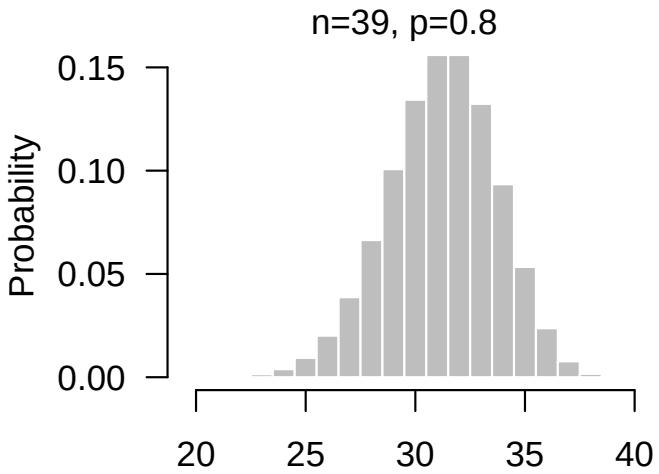
- Sometimes, the agreement is much better; for the infant survival data at 25 weeks (x=31,n=39), we have:
  - Binomial: [63.6, 90.7]
  - Wald: [66.8, 92.2]
  - Score: [64.5, 89.2]
- Here all three intervals are reasonably close, although the score interval is again closer to the binomial interval

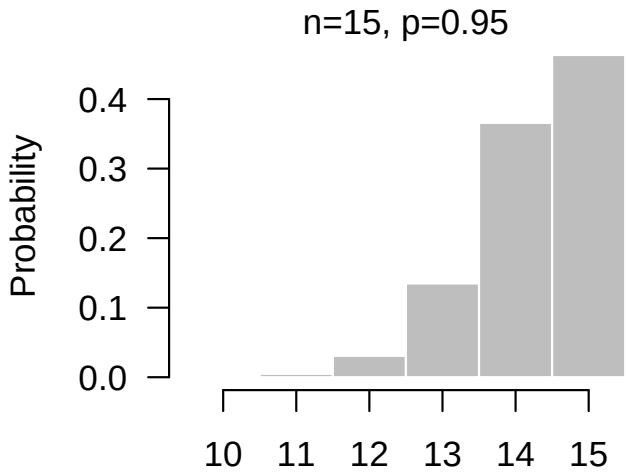
Infant survival, 25 weeks

- And sometimes, the Wald interval fails completely; for the infant survival data at 22 weeks (x=0,n=29), we have:
  - Binomial: [0, 11.9]
  - Wald: [0, 0]
  - Score: [0, 11.7]
- The Wald interval is clearly useless in this scenario

Accuracy of the normal approximation

- The real sampling distribution is binomial, but when n is reasonably big and p isn't close to 0 or 1, the binomial distribution looks a lot like the normal distribution, so the normal approximation works pretty well
- When n is small and/or p is close to 0 or 1, the normal approximation doesn't work very well:





Exact vs. approximate intervals

- When n is large and p isn't close to 0 or 1, it doesn't really matter whether you choose the approximate or the exact approach
- The approximate approaches are easy to do by hand, although in the computer era, this is often not important in real life
- Keep in mind, however, that the Clopper-Pearson interval is "exact" in the sense that it is based on the exact sampling distribution, but as we saw in lab, does not produce exact $1 - \alpha$ coverage

Summary

- A sampling distribution is the distribution of an estimate based on a sample from a population
- Know how to use the CLT to approximate sampling distributions
- Know how to use the CLT to carry out approximate tests for one-sample categorical data
- Wald CI: $\hat{\pi} \pm z_\alpha \text{SE}$, where $\text{SE} = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$, although this approximation can be very poor at times
- Score CI: Based on inverting the CLT-based test; still approximate, but better than Wald