One-sample categorical data (approximate)

Patrick Breheny

March 13

Introduction

- As we have seen, the central limit theorem can be used to derive the (approximate) sampling distribution of the average
- In the previous lecture, we saw how we could use that fact to calculate the probability that the sample average will be over a certain value, or to find an interval that has a 95% probability of containing the sample average
- In this lecture, we will see how we can use the same line of thinking to develop hypothesis tests and confidence intervals for one-sample categorical data
- We will then compare these results to the exact hypothesis tests and confidence intervals that we obtained earlier based on the binomial distribution

Hypothesis testing

- Let's begin with hypothesis testing, and 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, $\hat{p} = 11/14 = .79$; 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 $p_0 = 0.5$
- The notation p_0 refers to the hypothesized value of the parameter p under the null

The standard error

- What about the standard error?
- Recall that the standard deviation of an individual outcome for the binomial distribution is $\sqrt{p(1-p)}$
- Therefore, under the null hypothesis, the standard deviation is $\sqrt{p_0(1-p_0)} = \sqrt{1/4} = 1/2$
- Thus, the standard error is

$$\begin{aligned} SE &= \sqrt{\frac{p_0(1-p_0)}{n}} \\ &= \frac{1}{2\sqrt{n}} \end{aligned}$$

Procedure for a z-test

To summarize this line of thinking into a procedure:

#1 Calculate the standard error: $SE = \sqrt{p_0(1-p_0)/n}$

#2 Calculate $z = (\hat{p} - p_0)/SE$

#3 Draw a normal curve and shade the area outside $\pm z$

#4 Calculate the area under the normal curve outside $\pm z$

Note that these are the exact same four steps we had in the previous lecture for calculating the probability that the sample average fell into a certain range

Terminology

- Hypothesis tests revolve around calculating some statistic from the data that, under the null hypothesis, you know the distribution of
- This statistic is called a *test statistic*, since it's a statistic that the test revolves around
- 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*

The z-test for the cystic fibrosis experiment

- For the cystic fibrosis experiment, $p_0 = 0.5$
- Therefore,

$$\begin{aligned} SE &= \sqrt{\frac{p_0(1-p_0)}{n}} \\ &= \sqrt{\frac{0.5(0.5)}{14}} \\ &= .134 \end{aligned}$$

The z-test for the cystic fibrosis experiment (cont'd)

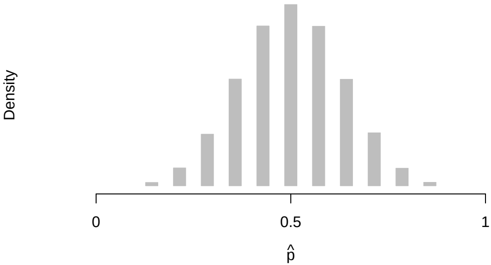
- The test statistic is therefore

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{SE} \\ &= \frac{.786 - .5}{.134} \\ &= 2.14 \end{aligned}$$

- The p-value of this test is therefore 2(.016) = .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; this represents moderate to substantial evidence against the null hypothesis

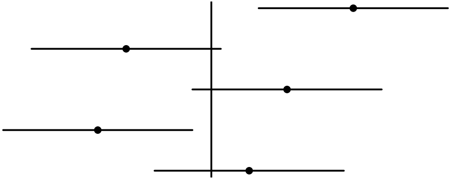
Accuracy of the approximation

- Recall, however, that we calculated a p-value of 6% from the (exact) binomial test, which is more in the moderate-to-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

- Similarly, the procedure for finding an interval with a 95% probability of containing the sample mean is closely related to constructing 95% confidence intervals:



- In other words, if the truth $\pm 1.96\ \mathrm{SE}$ has a 95% probability of containing the sample mean, then the sample mean $\pm 1.96\ \mathrm{SE}$ has a 95% probability of containing the truth

The standard error

- The major conceptual difference from the hypothesis test is that we don't know p, and we need p in order to calculate the standard error.
- The simplest and most natural thing to do (and what we will do in this class) is to use the observed value, $\hat{p}$, in calculating the standard error:

$$\mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- However, as we will see, this approximation does not always work so well, and it is perhaps worth being aware of the fact that other, better approximations have also been developed

Procedure for finding confidence intervals

Writing all this out as a procedure, the central limit theorem tells us that we can create $x\%$ confidence intervals by:

#1 Calculate the standard error: $SE = \sqrt{\hat{p}(1-\hat{p})/n}$

#2 Determine the values of the normal distribution that contain the middle $x\%$ of the data; denote these values $\pm z_{x\%}$

#3 Calculate the confidence interval:

$$(\hat{p} - z_{x\%}SE, \hat{p} + z_{x\%}SE)$$

Example: Survival of premature infants

- Let's return to our example from a few weeks ago involving the survival rates of premature babies
- Recall that 31/39 babies who were born at 25 weeks gestation survived
- The estimated standard error is therefore

$$\begin{aligned} SE &= \sqrt{\frac{.795(1-.795)}{39}} \\ &= 0.0647 \end{aligned}$$

Example: Survival of premature infants (cont'd)

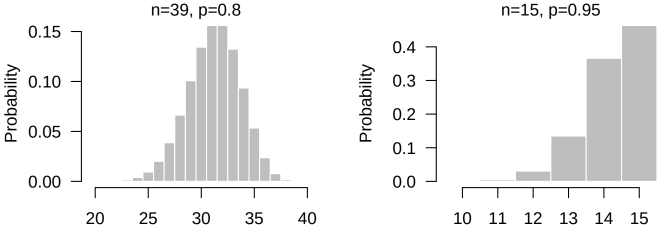
- Suppose we want a 95% confidence interval
- As we noted earlier, $z_{95\%} = 1.96$
- Thus, our confidence interval is:

$$(79.5 - 1.96(6.47), 79.5 + 1.96(6.47)) = (66.8\%, 92.2\%)$$

- Recall that our exact answer from the binomial distribution was (63.5%,90.7%)

Accuracy of the normal approximation

- Thus, we see that the central limit theorem approach works reasonably well here
- 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
- Other times, the normal approximation doesn't work very well:



Example: Survival of premature infants, part II

- Recall that the Johns Hopkins researchers also observed 0/29 infants born at 22 weeks gestation to survive
- What happens when we try to apply our approximate approach to find a confidence interval for the true percentage of babies who would survive in the population?
- $SE = \sqrt{\hat{p}(1-\hat{p})/n} = 0$, so our confidence interval is (0,0)
- This is an awful confidence interval, not even close to the exact one we calculated earlier: (0%, 12%)

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 advantage of the approximate approach is that it's easy to do by hand
- In comparison, finding exact confidence intervals by hand is quite time-consuming
- However, we live in an era with computers, which do the work of finding confidence intervals instantly, so in the real world there is no reason to settle for the approximate answer

Summary

- Know how to calculate by hand (with the aid of a table):
  - Approximate p-values for one-sample categorical data based on the central limit theorem
  - Approximate confidence intervals for one-sample categorical data based on the central limit theorem
- Although these are not useful in the real world (because computers can calculate exact answers), they are an instructive place to start, as many of the statistical methods we will use in the rest of the course will rely on central limit theorem approximations and follow very similar procedures