One-sample categorical data (approximate)

Patrick Breheny

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

The expected value of the binomial distribution

- To start, we need the mean and standard deviation of the binomial distribution
- The expected value is straightforward: if a trial is repeated n times and each trial has probability π that the event happens, the expected total is $n\pi$
- This makes sense: flip a coin 10 times, you expect 5 heads
- The expected value of the *proportion* (total divided by n) is therefore

  $$\frac{n\pi}{n} = \pi;$$

  note that expected value of the average is the same as the expected value of an individual trial

The standard deviation of the binomial distribution

- Of course, because of variability, you won't always get 5 heads
- The standard deviation of an individual yes/no outcome with probability π is $\sqrt{\pi(1-\pi)}$
- This means that the standard deviation of the binomial distribution is $\sqrt{n\pi(1-\pi)}$
- To continue our example of flipping a coin 10 times, here the SD is $\sqrt{10(0.5)(0.5)} = 1.58$, so we can expect the number of heads to be 5 ± 3 about 95% of the time (by rule of thumb)
- Note that the SD is highest when $\pi = 0.5$ and gets smaller as π is close to 0 or 1 — this makes sense, as if π is close to 0 or 1, the event is more predictable and less variable

The standard error of a proportion

- What about the standard deviation of the proportion (i.e., the standard error)?
- As before,

$$\begin{aligned}\mathrm{SD}(\hat{\pi}) &= \frac{\mathrm{SD}(X)}{n} \\ &= \frac{\sqrt{n\pi(1-\pi)}}{n} \\ &= \sqrt{\frac{\pi(1-\pi)}{n}}\end{aligned}$$

Summary

These derivations can be summarized in the following table:

	Expected value	Standard deviation
Individual	π	$\sqrt{\pi(1-\pi)}$
Total	$n\pi$	$\sqrt{n\pi(1-\pi)}$
Average	π	$\sqrt{\pi(1-\pi)/n}$

Transmission disequilibrium tests

- A common question in genetic epidemiology is whether or not a gene is associated with a certain trait
- One way of testing for this association empirically is called the *transmission disequilibrium test*
- The idea is to find (i.e., sample) parent-child pairs in which the child has the trait of interest and the parent is heterozygous for the gene of interest (i.e., has one copy of each version of the gene)
- As we have discussed, the parent is equally likely to pass on either copy, so if there is no link between the trait and the gene, we would expect 50% of the children to have version "A" and the other 50% to have version "B"

Transmission disequilibrium tests (cont'd)

- However, we have systematically sampled only children with the trait — any children without the trait are not included in the study
- If version "A" causes a child to be more likely to develop the trait of interest, we will find a higher proportion of version "A" in the children in our sample than version "B"
- In other words, if the two are associated, the "transmission" of the gene is distorted away from "equilibrium" (50/50 balance), hence the name of the test

TDTs and the binomial distribution

- Under the null hypothesis that the gene and trait are independent, the number of children in the sample who received the “A” version from their heterozygous parent will follow a binomial distribution with $P(A) = 50\%$
- In statistical shorthand, we would write $H_0 : \pi_0 = 0.5$, where π_0 refers to the hypothesized value of the parameter π under the null
- So the number of children *exactly* follows a binomial distribution, but from the central limit theorem, we also know that the sample proportion *approximately* follows a normal distribution, with expected value 0.5 and standard error $\sqrt{0.5(0.5)/n}$, where n is the number of parent-child pairs

Diabetes study

- In a 1989 study reported in the journal *Genetic epidemiology*, data was collected for 124 parent-child pairs in which the offspring had Type I diabetes and the parent was heterozygous for 5'FP (a flanking polymorphism adjacent to the insulin gene on chromosome 11)
- Among the children, 78 received the "class 1" version of 5'FP from their parent, while the other 46 did not
- In other words, $\hat{\pi} = 78/124 = 62.9\%$ of the children received the class 1 version; is this far enough from 50% to conclude that a departure from equilibrium is present?

Procedure for a z-test

- To test the hypothesis, we need to calculate the probability of seeing a sample proportion as far or farther from 50% than 62.9% is
- We can use our procedure from the previous lecture:
  1. Calculate the standard error: $\mathrm{SE} = \sqrt{\pi_0(1-\pi_0)/n}$
  2. Calculate $z = (\hat{\pi} - \pi_0)/\mathrm{SE}$
  3. Draw a normal curve and shade the area outside $\pm z$
  4. Calculate the area under the normal curve outside $\pm z$

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 diabetes study

- For the diabetes study, $\pi_0 = 0.5$ and $n = 124$
- Therefore,

$$\begin{aligned}\text{SE} &= \sqrt{\frac{\pi_0(1-\pi_0)}{n}} \\ &= \sqrt{\frac{0.5(0.5)}{124}} \\ &= .045\end{aligned}$$

The z-test for the diabetes study (cont'd)

- The test statistic is therefore

$$\begin{aligned} z &= \frac{\hat{\pi} - \pi_0}{\text{SE}} \\ &= \frac{.629 - .5}{.045} \\ &= 2.87 \end{aligned}$$

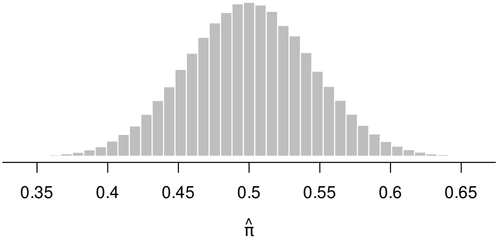
- The p-value of this test is therefore 2(.002) = .004
- In other words, if the null hypothesis were true, there would only be about a 0.4% chance of seeing so many children in our sample with the "class 1" version of the gene; this represents very strong evidence that the gene is associated with diabetes

Accuracy of the approximation

- We could also carry out the exact test:

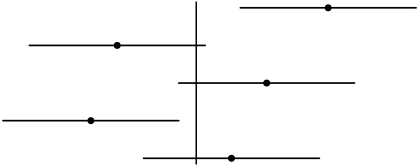
```r
binom.test(78,124)$p.value
# [1] 0.005161225
```

- In this case, the two answers are virtually identical (0.004 vs. 0.005) because for $n = 124$ and $\pi_0 = 0.5$, the binomial distribution looks almost exactly like the normal distribution:



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 π, and we need π 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, $\hat{\pi}$, in calculating the standard error:

$$\text{SE} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

- 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: $\text{SE} = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$

(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:

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

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

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

Example: Survival of premature infants (cont'd)

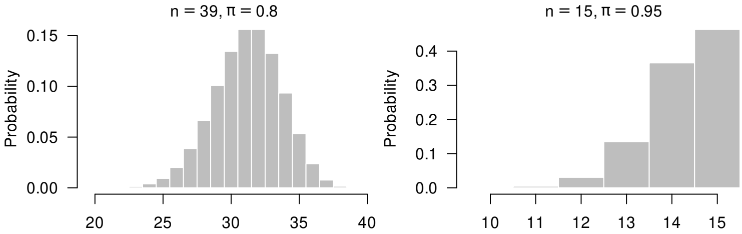
- Suppose we want a 95% confidence interval
- As we noted earlier, $z_{95\%} = 1.96$
- Thus, our confidence interval is:

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

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

Accuracy of the normal approximation

- The central limit theorem approach works reasonably well here: the real distribution is binomial, but when n is reasonably big and π 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?
- $\mathrm{SE} = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = 0$, 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 π 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 really necessary 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