Applying the central limit theorem

Patrick Breheny

October 21

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 what the distribution of estimates would look like if we were to repeat an experiment over and over, because in reality, the experiment is conducted only once
- If we were to repeat the experiment over and over, we would get different estimates each time, depending on the random sample we drew from the population

Sampling distributions

- To reflect the fact that its distribution depends on the random sample, the distribution of an estimate is called a *sampling distribution*
- These sampling distributions are hypothetical and abstract – we cannot see them or plot them (unless by simulation, as in the coin flipping example from our previous lecture)
- So why do we study sampling distributions?
- The reason we study sampling distributions is to understand how variable our estimates are and whether future experiments would be likely to reproduce our findings
- This in turn is the key to answering the question: "How accurate is my generalization to the population likely to be?"

Introduction

- The central limit theorem is a very important tool for thinking about sampling distributions – it tells us the shape (normal) of the sampling distribution, along with its center (mean) and spread (standard error)
- We will go through a number of examples of using the central limit theorem to learn about sampling distributions, then apply the central limit theorem to our one-sample categorical problems from an earlier lecture and see how to calculate approximate p-value and confidence intervals for those problems in a much shorter way than using the binomial distribution

Sampling distribution of serum cholesterol

- 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
- We are planning to collect a sample of 25 individuals and measure their cholesterol levels
- What is the probability that our sample average will be above 230?

Procedure: Probabilities using the central limit theorem

Calculating probabilities using the central limit theorem is quite similar to calculating them from the normal distribution, with one extra step:

#1 Calculate the standard error: $SE = SD/\sqrt{n}$, where SD is the population standard deviation

#2 Draw a picture of the normal approximation to the sampling distribution and shade in the appropriate probability

#3 Convert to standard units: $z = (x - \mu)/SE$, where μ is the population mean

#4 Determine the area under the normal curve using a table or computer

Example #1: Solution

- We begin by calculating the standard error:

$$\begin{aligned} SE &= \frac{SD}{\sqrt{n}} \\ &= \frac{46}{\sqrt{25}} \\ &= 9.2 \end{aligned}$$

- Note that it is smaller than the standard deviation by a factor of $\sqrt{n}$

Example #1: Solution

- After drawing a picture, we would determine how many standard errors away from the mean 230 is:

$$\frac{230-211}{9.2}=2.07$$

- What is the probability that a normally distributed random variable is more than 2.07 standard deviations above the mean?
- 1-.981 = 1.9%

Comparison with population

- Note that this is a very different number than the percent of the population has a cholesterol level above 230
- That number is 34.0% (230 is .41 standard deviations above the mean)
- The mean of a group is much less variable than individuals
- As Sherlock Holmes says in *The Sign of the Four*: "While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to. Individuals vary, but percentages remain constant. So says the statistician."

Procedure: Central limit theorem percentiles

- We can also use the central limit theorem to approximate percentiles of the sampling distribution:
  - #1 Calculate the standard error: $SE = SD/\sqrt{n}$
  - #2 Draw a picture of the normal curve and shade in the appropriate area under the curve
  - #3 Determine the percentiles of the normal curve corresponding to the shaded region using a table or computer
  - #4 Convert from standard units back to the original units: $\mu + z(SE)$

Percentiles

- We can use that procedure to answer the question, "95% of our sample averages will fall between what two numbers?"
- Note that the standard error is the same as it was before: 9.2
- What two values of the normal distribution contain 95% of the data?
- The 2.5th percentile of the normal distribution is -1.96
- Thus, a normally distributed random variable will lie within 1.96 standard deviations of its mean 95% of the time

Example #2: Solution

- Which numbers are 1.96 standard errors away from the expected value of the sampling distribution?

$$211 - 1.96(9.2) = 193.0$$
$$211 + 1.96(9.2) = 229.0$$

- Therefore, 95% of our sample averages will fall between 193 mg/dl and 229 mg/dl

Example #3

- What if we had only collected samples of size 10?
- Now, the standard error is

$$\begin{aligned} SE &= \frac{46}{\sqrt{10}} \\ &= 14.5 \end{aligned}$$

- Now what is the probability of that our sample average will be above 230?

Example #3: Solution

- Now 230 is only

$$\frac{230 - 211}{14.5} = 1.31$$

  standard deviations away from the expected value
- The probability of being more than 1.31 standard deviations above the mean is 9.6%
- This is almost 5 times higher than the 1.9% we calculated earlier for the larger sample size

Example #4

- What about the values that would contain 95% of our sample averages?
- The values 1.96 standard errors away from the expected value are now

$$211 - 1.96(14.5) = 182.5$$
$$211 + 1.96(14.5) = 239.5$$

- Note how much wider this interval is than the interval (193,229) for the larger sample size

Example #5

- What if we'd increased the sample size to 50?
- Now the standard error is 6.5, and the values

$$
\begin{aligned}
211 - 1.96(6.5) &= 198.2 \\
211 + 1.96(6.5) &= 223.8
\end{aligned}
$$

contain 95% of the sample averages

Summary

n	SE	Interval	Width of interval
10	14.5	(182.5,239.5)	57.0
25	9.2	(193.0,229.0)	36.0
50	6.5	(198.2,223.8)	25.6

The width of the interval is going down by what factor?

Example #6

- Finally, we ask a slightly harder question: How large would 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?
- As we saw earlier, 95% of observations fall within 1.96 standard deviations of the mean
- Thus, we need to get the standard error to satisfy

$$\begin{aligned} 1.96(SE) &= 5 \\ SE &= \frac{5}{1.96} \end{aligned}$$

Example #6: Solution

- The standard error is equal to the standard deviation over the square root of n, so

$$\frac{5}{1.96} = \frac{SD}{\sqrt{n}}$$

$$\sqrt{n} = SD \cdot \frac{1.96}{5}$$

$$n = 325.1$$

- In the real world, we of course cannot sample 325.1 people, so we would sample 326 to be safe

Example #7

- How large would the sample size need to be in order to insure a 90% probability that the sample average will be within 10 mg/dl of the population mean?
- There is a 90% probability that a normally distributed random variable will fall within 1.645 standard deviations of the mean
- Thus, we want $1.645(SE) = 10$, so

$$\frac{10}{1.645} = \frac{46}{\sqrt{n}}$$

$$n = 57.3$$

- Thus, we would sample 58 people

Introduction

- We can use the exact same logic to carry out hypothesis tests for one-sample categorical data
- Consider our cystic fibrosis experiment in which 11 out of 14 people did better on the drug than the 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 standard error

- What about the standard error?
- Recall that the standard deviation of an individual outcome for the binomial distribution is $\sqrt{p_0(1-p_0)}$
- 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 $\pm z$

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

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

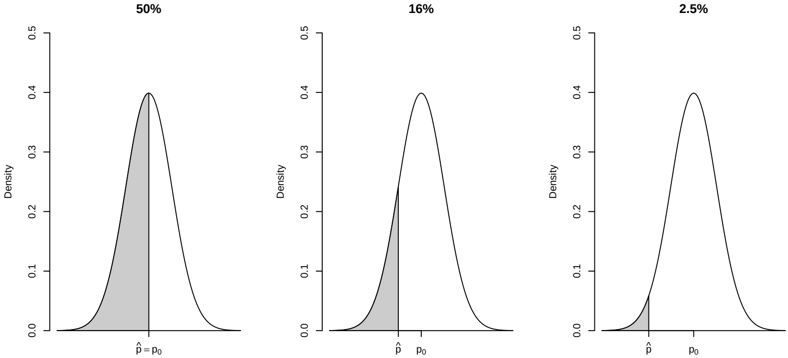
- The p-value of this test is therefore 2(1.6%) = 3.2%

Introduction: confidence intervals

- To find confidence intervals, the logic is similar, although the procedure is different
- Here, we're not assuming that we know where the sampling distribution is centered – we don't assume that p_0 equals anything
- Instead, we're trying to find a range of values for p_0 that are plausible in light of observing $\hat{p}$ – *i.e.*, values of p_0 such that if p_0 were true, it wouldn't be uncommon to see $\hat{p}$

Idea behind confidence intervals

Essentially, this consists of sliding p_0 around until $\hat{p}$ no longer looks like it could reasonably have come from that sampling distribution:



The form of confidence intervals

- Let $z_{x\%}$ be the value such that $\pm z_{x\%}$ contains the middle $x\%$ of the normal distribution (*i.e.*, $z_{95\%} = 1.96$)
- Then in order to get an $x\%$ confidence interval with $(100 - x)\%$ of the errors split equally on both sides, we have to slide the sampling distribution $z_{x\%}$ standard errors to the left of the mean, and then $z_{x\%}$ standard errors to the right of the mean
- This will give us the two endpoints of our confidence interval

Procedure for finding confidence intervals

- Summarizing this line of thought, 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:

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

Standard error

- Note that the standard error calculation is different here than it was for hypothesis tests
- In hypothesis tests, we specified a value of p_0, which in turn determined the standard error:

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

- For confidence intervals, we don't know p_0
- A reasonable thing to do, then, is to estimate SE based on $\hat{p}$:

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

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)

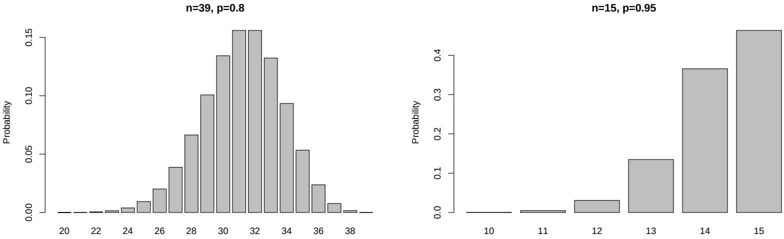
- Suppose we want a 95% confidence interval
- As we noted earlier, $z_{95\%} = 1.96$
- Thus, our confidence interval is:

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

- 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 very close at all 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

Exact vs. approximate intervals (cont'd)

- However, we live in an era with computers, which do the work of finding confidence intervals instantly
- If we can obtain the exact answer, there is no reason to settle for the approximate answer
- That said, in practice, people use and report the approximate approach all the time
- Possibly, this is because the analyst knew it wouldn't matter, but more likely, it's because the analyst learned the approximate approach in their introductory statistics course and doesn't know any other way to calculate a confidence interval