Applying the central limit theorem

Patrick Breheny

March 7

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 [1]
- 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?

[1] these are estimates, of course, but for the sake of these examples we will take them to be the true population parameters

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: $\text{SE} = \text{SD}/\sqrt{n}$, 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: $z = (x - \mu)/\text{SE}$, 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:

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

- 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 that 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 an individual

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: $\text{SE} = \text{SD}/\sqrt{n}$
  - #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: $\mu + z(\text{SE})$

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

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

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

Example #3: Solution

- With $n = 10$, 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

$$211 - 1.96(6.5) = 198.2$$
$$211 + 1.96(6.5) = 223.8$$

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 ensure 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

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

Example #6: Solution

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

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

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

$$n = 325.1$$

- 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 ensure 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 $1.645(\mathrm{SE}) = 10$, so

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

$$n = 57.3$$

- Thus, we would sample 58 people

Summary

- A sampling distribution describes the possible values that the outcomes (estimates) of an experiment will take on, and are at the heart of answering the question of how reliable/trustworthy/generalizable the results of a study are
- The central limit theorem tells us what the (approximate) sampling distribution of a sample average will be, and it can be used to:
    - Calculate the probabilities that the sample average will lie within a certain range
    - Calculate values that have a 95% probability of containing the sample average
    - Calculate the sample size necessary in order to achieve the desired precision