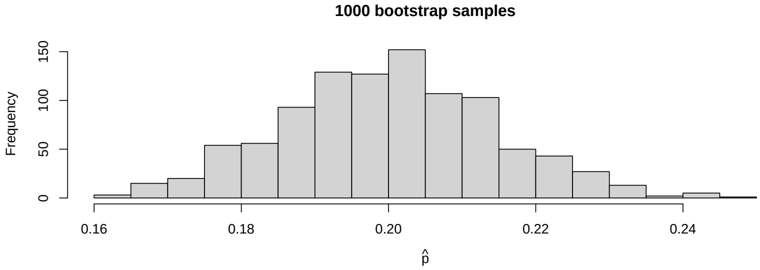
Confidence Intervals

Part 2 - Normal Approximations

Ryan Miller

Normal Distributions

We've now seen several *bootstrap distributions* and you may have noticed they tend to be "bell-shaped":



This is not a coincidence, it's backed up by statistical theory

Grinnell College
Statistics

Normal Distributions

Bootstrap distributions can be characterized by the curve:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- This curve defines the **Normal Distribution**
  - μ is the center (mean) of the distribution
  - σ is the standard deviation of the distribution
  - We use the shorthand $N(\mu,\sigma)$ to express a normal distribution, for example: N(3,1) is a curve centered at 3 with a standard deviation of 1
- You don't need to know the formula for the normal curve, though you should know that it depends on μ and σ

Normal Approximation

- When calculating a confidence interval estimate, we can use a *normal approximation* instead of bootstrapping
  - To do this, we need the distribution's mean and standard deviation (since any normal curve is entirely by μ and σ)
  - Thus, the approximation will be $N(\text{estimate}, \text{SE})$
    - We saw the bootstrap distribution was centered around the estimate from the original sample
    - We generated bootstrap samples and bootstrap statistics to find SE, but is there another way?

Central Limit Theorem

- The **Central Limit Theorem** (CLT), one of the most well-known results in statistics, provides a mathematical expression for the *SE* of many commonly used descriptive statistics
  - We'll first look at a CLT result for *one proportion*:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

In words, the sample proportion, $\hat{p}$, follows a normal distribution with a mean of p and standard deviation of $\sqrt{\frac{p(1-p)}{n}}$, thus providing a normal approximation of the sampling distribution

Using the CLT (one proportion)

Central Limit theorem gives us:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

- Thus, $SE = \sqrt{\frac{p(1-p)}{n}}$ when estimating a *single proportion*
- We don't know p, but $\hat{p}$ is our *best estimate*, together these suggest the 95% confidence interval:

$$\hat{p} \pm 2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

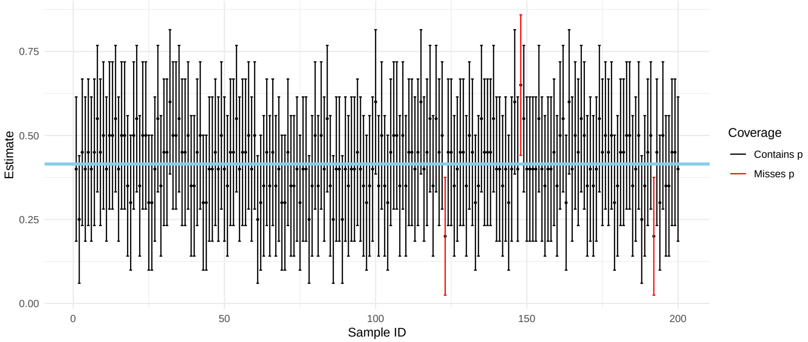
Confidence Interval Coverage

The phrase "95% confidence" describes the long-run success rate of *the procedure* used to calculate the interval. So let's apply the procedure from the previous slide to many random samples of size $n = 20$ from a population with $p = 0.415$:

Sample ID	Sample proportion	Calculation	95% CI
1	0.4	0.4 +/- 2* 0.11	(0.181,0.619)
2	0.25	0.25 +/- 2* 0.097	(0.056,0.444)
3	0.45	0.45 +/- 2* 0.111	(0.228,0.672)
4	0.4	0.4 +/- 2* 0.11	(0.181,0.619)
5	0.45	0.45 +/- 2* 0.111	(0.228,0.672)
6	0.4	0.4 +/- 2* 0.11	(0.181,0.619)

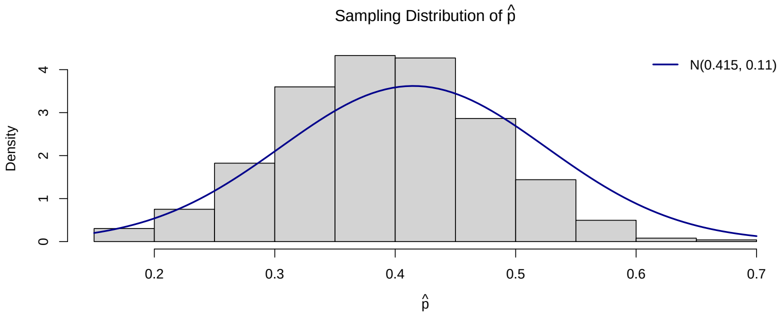
Confidence Interval Coverage

When we apply this procedure 200 times, only 3 intervals fail to capture the true p, suggesting the procedure is valid (but perhaps slightly conservative):



Confidence Interval Coverage

A long-run success rate that is slightly above 95% makes sense, as a normal approximation of the sampling distribution is decent but not perfect:



Grinnell College
Statistics

Practice

In a study conducted by Johns Hopkins University researchers investigated the survival of babies born prematurely. They searched their hospital's medical records and found 39 babies born at 25 weeks gestation (15 weeks early), 31 of these babies went on to survive at least 6 months. With your group:

1. Use a normal approximation to construct a 95% confidence interval estimate for the true proportion of babies born at 25 gestation that are expected to survive.
2. An article on Wikipedia suggests 70% of babies born at 25 weeks gestation survive. Is this claim consistent with the Johns Hopkins study?

Practice - Solution

1. $\hat{p} = 31/39 = 0.795$, using the normal approximation provided by CLT, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.795(1-0.795)}{39}} = 0.065$; this suggests the 95% CI:
$$0.795 \pm 2 * 0.065 = (0.668, 0.922)$$
2. Yes, 0.70 is contained in the 95% confidence interval, suggesting it is a plausible value of the population parameter.

Grinnell College
Statistics

Sufficiently Large?

The normal approximation suggested by the Central Limit Theorem is only accurate when n is sufficiently large

- For a single proportion, "sufficiently large" also depends upon the value of p
- A common rule of thumb for whether this normal approximation of $\hat{p}$ is reasonable requires:

1. $n * p \geq 10$
2. $n * (1 - p) \geq 10$

If *either* of these conditions isn't met you should consider an alternative (our lab will introduce *exact binomial confidence intervals*)

Practice

- With your group, check the conditions of the normal approximation used in the Johns Hopkins example
- Then, use StatKey to generate a bootstrap distribution for the proportion of babies who survived
  - Compare the *SE* found using bootstrapping to the *SE* calculated using the CLT normal approximation
  - Compare the shape of the bootstrap distribution to that of the normal curve

Practice - Solution

- The normal approximation conditions are *not* met, $n * (1 - p) = 8$ for these data
- The SE of the bootstrap distribution is very similar to that calculated using the CLT formula
- The shape of the bootstrap distribution is slightly left skewed
  - This is partly because 1 represents a hard upper-bound for a single proportion
- In this scenario you might consider a *percentile bootstrap confidence interval* to be the most reliable option

Using the CLT (difference in proportions)

For a difference in proportions (ie: *risk difference*), provided n_1 and n_2 are sufficiently large, CLT suggests:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

In this scenario, checking the conditions of this normal approximation is a little more tedious:

1. $n_1 * p_1 \geq 10$
2. $n_1 * (1 - p_1) \geq 10$
3. $n_2 * p_2 \geq 10$
4. $n_2 * (1 - p_2) \geq 10$

Practice

A 2010 study took two groups of doctors that had similar error rates before the study and randomly assigned half of them to use an electronic prescription form, while the other half continued using written prescriptions. After 1 year, the error rate of each group was recorded:

	Error	Non-errors	Total
Electronic	254	3594	3848
Hand-written	1478	2370	3848

1) Find the 95% confidence interval for difference proportions (e-prescriptions minus hand-written prescriptions)
2) Is it plausible that both electronic and handwritten forms have the same error rate?

Practice - Solution

1. $\hat{p}_e - \hat{p}_{hw} = 254/2848 - 1478/3848 = 0.066 - 0.384 = -0.318$; while $SE = \sqrt{\frac{0.066(1-0.066)}{3848} + \frac{0.384(1-0.384)}{3848}} = 0.009$; thus the 95% Confidence Interval is given by:

$$-0.318 \pm 2 * 0.009 = (-0.336, -0.300)$$

2. No, the 95% confidence interval does not contain zero, implying the error rates being equal for both forms is not plausible

Confidence Levels that aren't 95%

Recall that confidence intervals have the form:

$$\text{Estimate} \pm c * SE$$

- Because areas under the normal curve are known, we aren't limited by the 68-95-99 rule when it comes to determining a meaningful value for c
- The **standard normal** distribution has a mean of 0 and standard deviation of 1
  - We can use cut-points in this distribution to achieve *any* confidence level that we'd like
  - One place we can do this is the "Theoretical Distribution" menu on StatKey

Practice

Recall that in the handwritten vs. electronic prescription study:
$\hat{p}_e - \hat{p}_{hw} = 254/3848 - 1478/3848 = 0.066 - 0.384 = -0.318$ and
$SE = \sqrt{\frac{0.066(1-0.066)}{3848} + \frac{0.384(1-0.384)}{3848}} = 0.009$

1) Use StatKey to find the appropriate normal quantile (ie: c) for constructing an 80% confidence interval
2) Use this value to construct an 80% confidence interval for effect of handwritten vs. electronic prescriptions

Practice - Solution

1) $c = 1.282$
2) $-0.318 \pm 1.282 * 0.009 = (-0.330, -0.306)$

Conclusion

We're now able to create *confidence interval estimates* for two population parameters using normal approximations:

1. A *single proportion*, we estimate p using an interval of the form:

$$\hat{p} \pm c * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

2. A *difference in proportions* (risk difference), we estimate $p_1 - p_2$ using an interval of the form:

$$\hat{p}_1 - \hat{p}_2 \pm c * \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

These formulas are only reliable when the sample is *sufficiently large*. Exact approaches should be used for small samples.