C. How close will estimator be to true value?

1. Motivation:
   a. Give an interval in which we believe true parameter will lie.
   b. Example
      i. Suppose we want to estimate what proportion $\pi$ of patients
         with a certain characteristic will benefit from a certain therapy
      ii. Suppose we choose 10 patients to give a treatment to, and
          determine whether it helped or not.
      iii. Suppose all 10 were helped. What values of $\pi$ are consistent
           with these results?
      iv. Any value but 0 could have produced the data, but those values
          near zero would seldom produce such data.
   c. Goal: Eliminate not only values that could not have produced
      data but also those that would produce such data very seldom;
      what is left over is called a confidence interval.
   d. Will look like data stuff $\leq$ parameter $\leq$ different data stuff,
      and attach a confidence level, telling us how often it fails.

2. Construction for population mean, when sample mean sampling
distribution approx. normal:

a. Suppose that we want an interval that will cover the correct value of the parameter 95% of the time.

b. Start with relationship \( P(-1.96 \leq Z \leq 1.96) \), where

\[
Z = \left( \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \right).
\]

c. Then \( P(-1.96 \leq (\bar{X} - \mu)/(\sigma / \sqrt{n}) \leq 1.96) = 95\% \)

d. Then \( P(-1.96\sigma / \sqrt{n} \leq \bar{X} - \mu \leq 1.96\sigma / \sqrt{n}) = 95\% \)

e. Then \( P(\mu - 1.96\sigma / \sqrt{n} \leq \bar{X} \leq \mu + 1.96\sigma / \sqrt{n}) = 95\% \) gives exactly the opposite of what we need

f. Then \( P(1.96\sigma / \sqrt{n} \geq \mu - \bar{X} \geq -1.96\sigma / \sqrt{n}) = 95\% \)

g. Then \( P(\bar{X} + 1.96\sigma / \sqrt{n} \geq \mu \geq \bar{X} - 1.96\sigma / \sqrt{n}) = 95\% \)

h. Other confidence levels may be constructed by substituting something for 1.96.

i. Figure out how often you insist on being right

- 99%? 90%?

ii. Subtract from 1 to find out how often you’re going to allow yourself to be wrong

- 1%? 10%?

iii. Divide by 2, since we’re allowing the same probability of over-
and under-shooting.

iv. Find number in normal table with this much probability above it.

i. Example: Population mean Log blood lead for all participants

\[ \bar{X} = 0.8788, \sigma = 0.2778, n = 243 \]

ii. 95% Confidence interval is \((0.8438, 0.9137)\).
   \[ z = 1.96. \]

iii. 99% Confidence interval is \((0.8329, 0.9247)\).
   \[ z = 2.57. \]

iv. 90% Confidence interval is \((0.8495, 0.9081)\).
   \[ z = 1.645. \]

j. Example: Population mean Log blood lead for home owners

i. \( \bar{X} = 0.6998, \sigma = 0.2536, n = 38 \)

ii. 95% Confidence interval is \((0.6192, 0.7805)\).

3. Interpretation: Rule holds with specified probability, but the interval it gives doesn’t have a probability attached; see Fig. 34/.

   a. Instead we substitute the term confidence.

4. Case of sample proportion \( p \) to estimate \( \pi \)

   a. We would know standard deviation \( \sigma \) if we knew \( \pi : \sqrt{\pi (1 - \pi)} \)
b. Could use \( \sqrt{p(1 - p)} \)

c. Could use sample variance of underlying \( X_i \)

d. Try this at home: \( p(1 - p) = \sum_j (X_j - p)^2 / n \) rather than \( \sum_j (X_j - p)^2 / (n - 1) \); doesn’t really matter which you use, and \( p(1 - p) \) is easier.

e. \( \sqrt{p(1 - p) / n} \) is standard error of sample proportion.

f. Use if \( 5 < np \), \( 5 < n(1 - p) \)
5. Harder case: When we don’t know standard deviation $\sigma$ of underlying population:

a. Subcase when $n$ large enough so that sample standard deviation should be very close to true: Substitute sample standard deviation $s$ for population standard deviation $\sigma$.

b. Note that sample standard deviation $s/n$ is no longer really the standard deviation of anything.

i. It is an estimate of the standard deviation of the sampling distribution of the estimator useful for confidence intervals.

ii. Called *standard error*.

c. Note that confidence interval gets narrower as $n$ increases.

i. Can calculate how large sample must be to give us a small enough confidence interval.

ii. Let $B =$ half width of CI

iii. $B = z\sigma/\sqrt{n}$

iv. $n = z^2\sigma^2/B^2$; maybe have to guess $\sigma$

v. Ex., to estimate the average age of a population whose standard deviation is 10 to within 2 years, with 95% confidence, $n = (1.96)^2(10^2)/2^2 = 96.04$; round up.