VI. 20 Chance Errors in Sampling

- Estimators versus Target Parameters

  Example. In polling for voter opinion, we obtain a sample and use the fraction in the sample who have a certain opinion as an estimate of the true (but unknown) fraction in the target population who have that opinion (e.g., "like Obama's health care plan"). The true fraction is the "parameter" being estimated by the sample estimator.

  Issue. The estimator varies from sample to sample. For example, if the population size is 6,672 and the sample size is 100, there are about \( \frac{10^{200}}{10^{100}} \) different samples.

  \[
  \text{sample estimator} = \text{target parameter} + \text{chance error}
  \]

  The histogram of all these possible values obtained over all the possible samples has an expected value and an SD. Also, a normal approximation applies here.

  We will discuss this for the case that the target parameter is the fraction (or \( \% \)) of a population that has a certain property (e.g., income > $K$).
Expected Value and Standard Error

Sample fraction = population fraction + chance error

A chance variable, depending on which sample is drawn.

With a simple random sample, the expected value for the sample % equals the population %.

Example. Population % is 27%. Then Expected Value of Sample % is 27%.

With a simple random sample (drawn with replacement), the SE of the sample % equals

\[ \sqrt{\frac{(\text{population} \%) (1 - \text{population} \%)}{\text{(sample size)}}} \]

Example, continued. Let sample size be 240. Then

\[ \text{SE of sample} % = \sqrt{\frac{(27\%)(73\%)}{240}} \]

\[ = \sqrt{\frac{0.27(0.73)}{240}} \]

\[ = 0.0287 \text{ (or } 2.87\%) \]

In sum: If the true population fraction is \( \hat{p} \), and the sample size is \( n \), and the sample estimator is \( \hat{p} \), then

\[ E(\hat{p}) = p \]

\[ \text{SE}(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \]

NOTE. The SE decreases as \( \frac{1}{\sqrt{n}} \) as \( n \) increases.
• Using the Normal Approximation

As the sample size increases, the histogram of possible \( \hat{p} \) values taken over all possible samples follows a normal curve: \( \mathcal{N}(p, \frac{p(1-p)}{n}) \).
(Central Limit Theorem, again)

\( \hat{p} = 0.57 \)

Illustration. Suppose that 57% of voters in a given population plan to vote for Candidate A in a forthcoming election. A pollster draws a simple random sample of size \( n = 1500 \). (Assume—as an approximation—that the sampling is with replacement.)

(i) The sample proportion \( \hat{p} \) has expected value 0.57

(ii) The sample proportion \( \hat{p} \) has SE

\[
\sqrt{\frac{0.57 \times 0.43}{1500}} = 0.013 \quad (1.3\%)
\]

(iii) What is the chance that \( \hat{p} \) happens to fall between 0.57 ± 0.013 (0.557, 0.583) ?

Answer. This is the interval (expected value ± 1 SE).

By the Normal approximation, the chance is \( \approx 68\% \).
Correction Factor for Sampling Without Replacement

\[
SE(\hat{p}) = \sqrt{\frac{N-n}{N-1}} \times SE(\hat{p}) \quad \text{drawing without replacement}
\]

\[
= \sqrt{\frac{N-n}{N-1}} \left( \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)
\]

"correction factor"

N = population size

Note: For N "large", this factor becomes very close to 1.

Q. What does this tell us about the relative importance of the population size versus the sample size, as regards sampling variability of \( \hat{p} \)?

A. Except when \( \frac{n}{N} \) is large, it is the sample size \( n \) that mostly affects the variability of \( \hat{p} \).

Q. Is sampling without replacement more efficient than sampling with replacement, for estimation of \( p \)?

A. Yes. The SE of \( \hat{p} \) is smaller when based on sampling without replacement.